BAB III
METODE PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain survei deskriptif dan eskplanatori secara bersamaan. Desain survei deskriptif dimaksudkan untuk mendeskripsikan kondisi ketenagakerjaan dan insiden mismatch pendidikan di Indonesia. Adapun desain survei eksplanatori bertujuan untuk menguji hubungan antar variabel penelitian secara ekonometrik. Tujuan dari penelitian ini adalah untuk menguji teori terkait dengan hubungan atau pengaruh antara variabel-variabel bebas terhadap variabel terikat. Disebut menguji teori karena hubungan atau pengaruh antar variabel dalam penelitian ini dibangun berdasarkan teori serta penelitian-penelitian empiris. Hubungan antara variabel terikat dan bebas dalam penelitian ini selanjutnya disusun dalam model-model penelitian yang berasal dari model teoritis.

Penelitian ini memiliki 7 model teoritis yang pada intinya membahas tentang determinan mismatch pendidikan dan dampak dari mismatch tersebut. Model 1 sampai dengan 3 dalam penelitian ini membahas tentang determinan mismatch, yaitu determinan *overeducation*, *undereducation*, dan vertikal mismatch. Sementara itu, model 4 hingga 7 berkaitan dengan pengaruh mismatch vertikal (*overeducation* dan *undereducation*) dan horizontal terhadap pendapatan individu, pengangguran, pertumbuhan ekonomi, dan produktivitas tenaga kerja. Dalam menguji model-model tersebut, penelitian ini menggunakan data dari Survei Angkatan Kerja Nasional (Sakernas).

3.2. Objek Penelitian, Sumber Data, Populasi dan Sampel

Terdapat 2 objek dalam penelitian ini yaitu tenaga kerja dan indikator makro ekonomi Indonesia. Untuk menguji hipotesis 1 hingga 4, objek penelitiannya adalah tenaga kerja Indonesia, sedangkan hipotesis 5 hingga 7 objeknya adalah indikator makro ekonomi di tingkat Provinsi. Sumber data yang digunakan untuk menguji hipotesis 1 hingga 4 tersebut adalah Survei Angkatan Kerja Nasional (Sakernas) tahun 2022 yang dilakukan oleh Badan Pusat Statistik

(BPS). Sementara itu, sumber data untuk menguji hipotesis 5 hingga 7 adalah Sakernas serta Survei Ekonomi Nasional (Susenas) dari BPS tahun 2012 hingga 2022.

Populasi untuk menguji hipotesis 1 hingga 4 dalam penelitian ini adalah seluruh penduduk usia kerja (di atas 15 tahun) di Indonesia pada tahun 2022. Berdasarkan data Sakernas 2022, total populasi yang merupakan penduduk usia kerja tersebut adalah sebanyak 209,420,383 jiwa. Sakernas 2022 merepresentasikan populasi ini ke dalam 752,688 responden sebagai sampel. Akan tetapi, penelitian ini akan menggunakan sampel Sakernas 2022 ini dengan jumlah yang berbeda-beda. Dalam mengkalkulasi insiden *mismatch* vertikal misalnya, total sampel yang akan digunakan penelitian ini hanya sebanyak 498,451. Jumlah tersebut merupakan jumlah angkatan kerja dengan kategori bekerja minimal 1 jam per-minggu, melakukan kegiatan untuk memperoleh penghasilan, membantu kegiatan usaha/pekerjaan oranglain, dan bekerja namun sedang tidak aktif.

Setiap tahunnya, BPS merilis Sakernas periode Februari dan Agustus. Survei gelombang terakhir yang dilakukan BPS untuk Sakernas adalah pada bulan Agustus 2022. Namun, setiap gelombangnya, data Sakernas menggunakan responden baru, atau berbeda dengan gelombang sebelumnya. Oleh karena itu, data Sakernas sulit untuk dijadikan data panel, meskipun dapat dijadikan *pooled cross sectional* data. Tetapi, salah satu keunggulan Sakernas adalah menyediakan data terbaru dengan jumlah responden besar.

Sakernas merinci 2 estimasi sampel yaitu tingkat kabupaten/kota dan tingkat provinsi. Sakernas menggunakan metode multistage sampling yang secara operasional dengan *two stages one phase stratified sampling* dengan 3 tahapan. Tahap pertama, dilakukan metode *probability proportional to size* atau pengambilan sampel yang sebanding dengan ukuran sehingga sampel dipilih secara proporsional dengan ukuran total populasi. Tahap kedua, memilih n blok sensus sesuai dengan strata di setiap wilayah dan strata lapangan usaha. Tahap ketiga, dilakukan pemilihan rumah tangga hasil pemutakhiran dengan metode *systematic sampling*.

Meskipun menggunakan *two stages sampling*, tetapi BPS hanya menyediakan 1 jenis variabel bobot (weight). Hal ini karena bobot sampling

babak pertama dan kedua sudah disatukan. Hal inilah yang membuat metode pengambilan sampel BPS untuk data Sakernas ini disebut sebagai *two stages one phase stratified sampling* (dua babak dalam 1 fase stratifikasi sampel). Penggunaan metode tersebut diimplementasikan ke dalam 3 tahap pengambilan sampel. Penggunaan 3 tahap pengambilan sampel untuk tingkat kabupaten/kota yang dilakukan Sakernas ini membuatnya memiliki sampling dan *non sampling error* yang sangat rendah. BPS mengklaim bahwa *sampling* dan *non sampling error*-nya hanya berada pada batas kurang dari 5%.

Untuk Sakernas 2022, total sampel yang diperoleh adalah 752.690. Namun demikian, tidak seluruh total sampel tersebut akan dipergunakan dalam penelitian ini. Hal ini karena total sampel Sakernas masih mencakup responden yang masih sekolah, menganggur, setengah pengangguran, dan aktivitas lainnya. Hingga saat ini, Sakernas masih digunakan untuk menyusun beragam data mikro (tingkat regional) maupun makro (tingkat nasional) terkait dengan indikator-indikator ketenagakerjaan Indonesia.

Diantara indikator ketenagakerjaan yang didapatkan dari Sakernas antara lain data tingkat pengangguran terbuka, jumlah angkatan kerja, tingkat setengah pengangguran, dan lainnya. Berdasarkan hal tersebut, data dari Sakernas dapat secara langsung diterapkan ke dalam populasi tingkat kabupaten/kota maupun provinsi. Hal ini merupakan salah satu kelebihan utama dari data Sakernas. Meskipun tentunya, Sakernas juga memiliki sejumlah kekurangan, diantarannya adalah konsistensi datanya relatif masih dipertanyakan karena seringkali terjadi perubahan dalam variabel yang diukur (Mugijayani, 2020).

Terlepas dari adanya kekurangan tersebut, penelitian ini tetap menggunakan Sakernas sebagai database dalam mengestimasi *mismatch* pendidikan. Hal ini karena Sakernas merupakan salah satu database primer yang digunakan untuk mengukur banyak indikator ketenagakerjaan Indonesia. Meskipun disebutkan memiliki kekurangan, namun Sakernas telah cukup banyak digunakan oleh riset-riset yang dipublikasikan pada jurnal internasional bereputasi. Smith et al. (2002) misalnya, menggunakan data Sakernas untuk menganalisis dampak krisis terhadap pasar tenaga kerja. Kemudian, Suryahadi et al. (2005), Osterreich (2013), Pratomo & Manning (2022), dan Caraka et al.

(2021) juga menggunakan data Sakernas. Selain itu, mengacu pada Dong (2016), Sakernas relatif merupakan sumber data yang lebih baik daripada database survei sejenis yaitu Indonesian Familiy Live Survei (IFLS) untuk melakukan analisis yang berkaitan dengan upah tenaga kerja. Hal ini karena Sakernas memiliki ukuran sampel yang jauh lebih besar daripada IFLS.

Penelitian ini menggunakan data Sakernas dalam bentuk data cross sectional dan data panel. Data cross sectional untuk menganalisis model-model terkait dengan determinan *mismatch* vertikal dan horizontal, serta dampaknya terhadap pendapatan. Data cross sectional ini hanya akan menggunakan Sakernas 2022. Sementara itu, data Sakernas yang akan dijadikan data panel adalah Sakernas periode Februari 2012 hingga 2022 dengan level penyajian Provinsi. Total Provinsi yang akan diteliti adalah sebanyak 33, sehingga total observasi adalah 363 (33 Provinsi x 11 tahun periode pengamatan). Dengan menggunakan sebelas gelombang Sakernas tersebut, penelitian ini berupaya untuk menganalisis dampak *mismatch* pendidikan terhadap tingkat pengangguran, pertumbuhan ekonomi, dan produktivitas tenaga kerja (model 5, 6, dan 7). Untuk lebih jelasnya, data-data yang akan digunakan dalam penelitian ini adalah sebagaimana berikut:

Tabel 3.1.

Jumlah Populasi dan Sampel Penelitian

Uraian	Total Sampel/Observasi	Total Populasi	Jenis Data	Sumber Data
Statistik Kondisi Ketenagakerjaan	752,688	209,420,383	Cross Sectional	Sakernas 2022
Insiden *Mismatch* Vertikal	498,451	135,296,713	Cross Sectional	Sakernas 2022
Insiden *Mismatch* Horizontal	193,000	56,254,773	Cross Sectional	Sakernas 2022
Model 1	192,524	56,029,408	Cross Sectional	Sakernas 2022
Model 2	192,524	56,029,408	Cross Sectional	Sakernas 2022
Model 3	192,524	56,029,408	Cross Sectional	Sakernas 2022
Model 4	752,688	209,420,383	Cross Sectional	Sakernas 2022
Model 5	363	-	Panel	Sakernas 2012 - 2022
Model 6	363	-	Panel	Sakernas 2012 - 2022
Model 7	363	-	Panel	Sakernas 2012 - 2022

Selain menelaah model-model penelitian yang disusun, penelitian ini juga memberikan gambaran mengenai kondisi ketenagakerjaan Indonesia berdasarkan Sakernas 2022. Untuk memberi gambaran umum kondisi ketenagakerjaan

tersebut, penelitian ini menggunakan seluruh sampel penelitian yang tersedia dalam Sakernas 2022. Jumlah sampel sebanyak 752,688 adalah seluruh penduduk usia kerja (minimal 15 tahun). Tujuan utama menggambarkan kondisi ketenagakerjaan ini adalah untuk memberikan argumentasi faktual dari temuan-temuan penelitian dari setiap model penelitian.

Rumusan masalah pertama dalam penelitian ini adalah tentang gambaran insiden *mismatch* vertikal dan horizontal ketenagakerjaan Indonesia. Untuk menjawab rumusan masalah tersebut, penelitian ini melakukan estimasi insiden *mismatch* vertikal dengan menggunakan 498,451 sampel dengan jumlah populasi sebanyak 135,296,713. Jumlah ini merupakan jumlah angkatan kerja Indonesia dengan status bekerja minimal 1 jam per-minggu sebanyak 464,443 sampel, melakukan kegiatan untuk memperoleh penghasilan sebanyak 4,373 sampel, membantu kegiatan usaha/pekerjaan oranglain sebanyak 20,761, dan bekerja namun sedang tidak aktif sebanyak 10,874 sampel.

Dalam mengestimasi insiden *mismatch* vertikal, penelitian ini mengekslusikan sejumlah kelompok sampel dengan jumlah yang cukup signifikan. Total data sampel yang digunakan untuk mengestimasi insiden horizontal hanya sebanyak 193,000 dengan ukuran total populasi sebanyak 56,254,773 orang. Ini karena sampel yang digunakan untuk insiden horizontal *mismatch* hanyalah sampel dengan tingkat pendidikan minimal SMA/Sederajat. Mengidentifikasi horizontal *mismatch* memerlukan informasi tentang bidang/jurusan studi, sedangkan individu dengan tingkat pendidikan di bawah SMA/Sederajat belum memiliki bidang/jurusan studi. Jumlah sampel ini juga akan digunakan untuk mengestimasi model 1 hingga 4.

Untuk model 5, 6, dan 7, data yang digunakan bukan lagi bersifat cross sectional dengan format survei, melainkan data panel. Data panel tersebut tetap berasal dari data Sakernas, tetapi Sakernas dari tahun 2012 hingga 2022. Dari data Sakernas setiap tahun tersebut, penelitian ini melakukan estimasi indeks *mismatch* vertikal dan horizontal untuk tiap Provinsi. Oleh karenanya, akan terdapat sebanyak 363 jumlah observasi atau data yang akan digunakan untuk mengestimasi model-model tersebut.

3.3. Instrumen Penelitian

Penelitian ini menggunakan data sekunder yang berasal dari database Sakernas yang dilakukan BPS. Instrumen yang digunakan untuk mengumpulkan data Sakernas tersebut menggunakan kuesioner. Kuesioner tersebut ditanyakan kepada para responden melalui tatap muka secara langsung. Adapun kuesioner yang dipergunakan oleh BPS untuk data Sakernas dapat diakses secara online melalui laman https://silastik.bps.go.id.

Dalam penelitian ini, data-data yang dihasilkan Sakernas akan diolah untuk mengukur variabel-variabel penelitian. *Mismatch* pendidikan misalnya, dalam penelitian ini diukur dengan menggunakan pendekatan normatif (job analysis), realized match (VV), dan mode. Pendekatan job analysis dilakukan dengan mencocokan tingkat pendidikan dengan KBJI tahun 2014 yang mengadaptasi ISCO 2008. Minimal pendidikan yang ditetapkan dalam ISCO 2008 untuk setiap golongan pekerjaan dapat terlihat sebagaimana berikut:

Tabel 3.2.

Golongan Pekerjaan dan Tingkat Pendidikan ISCO 2008 dan ISCED 11

ISCO 2008 and ISCED 11 Levels of Education			*Skill* level 1		*Skill* Level 2		*Skill* Level 3		*Skill* Level 4	
	No Schooling	*Early Childhood Education*	*1 Primary Education*	*2 Lower Secondary Education*	*3 Upper Secondary Education*	*4 Post Secondary non-tertiary Education*	*5 Short-cycle tertiary education*	*6 Bachelor or Equivalent level*	*7 Master or equivalent level*	*8 Doctoral or equivalent level*
Managers	UE	UE	UE	UE	UE	UE	M	M	M	M
Professionals	UE	UE	UE	UE	UE	UE	UE	M	M	M
Technicians and Associate Professionals	UE	UE	UE	UE	UE	UE	M	OE	OE	OE
Clerical Support Workers	UE	UE	UE	M	M	M	OE	OE	OE	OE
Services and Sales Workers	UE	UE	UE	M	M	M	OE	OE	OE	OE
Skilled Agricultural, Forestry and Fishery Workers	UE	UE	UE	M	M	M	OE	OE	OE	OE
Craft and Related Trades Workers	UE	UE	UE	M	M	M	OE	OE	OE	OE
Plant and Machine Operators and Assemblers	UE	UE	UE	M	M	M	OE	OE	OE	OE
Elementary Occupation	UE	UE	M	OE	OE	OE	OE	OE	OE	OE

Keterangan : UE adalah *undereducation*, M adalah matched, dan OE adalah *overeducation*. Sebagai contoh, menurut ISCO dan ISCED, untuk bekerja sebagai manajer, skill minimal yang diperlukan adalah level 3 dengan tingkat pendidikan 5 yaitu short-cycle tertiary education (pendidikan tinggi siklus pendek) atau tingkat diploma, baik itu diploma I, II, maupun III

Jika mengacu pada tabel diatas, maka golongan okupasi manajer dan profesional cenderung tidak akan mengalami *overeducation*. Akan tetapi, mengacu pada ISCO 2008, dijelaskan bahwa *skill* level untuk manajer adalah level 3 ditambah 4. Suatu pekerjaan yang memerlukan *skill* level 3 jika mengacu pada ISCO 2008 yakni melibatkan kinerja yang kompleks yang berkaitan dengan tugas teknis dan praktis yang memerlukan pengetahuan faktual, teknis, dan prosedural yang luas dalam bidang khusus.

Sementara itu, *skill* level 4 melibatkan pekerjaan yang memerlukan pemecahan masalah yang kompleks dan pengambilan keputusan berbasis pengetahuan teoretis dan faktual yang luas dalam bidang khusus. Atas dasar itu, untuk menjadi seorang manajer di suatu perusahaan misalnya, *skill* level 3 saja tidaklah cukup, melainkan harus juga memiliki *skill* level 4. Jika dikoversi ke dalam KKNI, *skill* level 4 yang dipaparkan dalam ISCO setara dengan KKNI level 8 dalam KKNI. Akan tetapi, tingkat pendidikan serta *skill* yang diperlukan oleh seorang manajer juga berkaitan dengan perusahaan tempatnya bekerja.

Manajer perusahaan-perusahaan besar, misalnya seperti manajer di BUMN, dinilai lebih memerlukan *skill* yang lebih tinggi daripada manajer toko. Di sisi lain, Sakernas tidak memberikan informasi yang spesifik mengenai jenis perusahaan manajer, sehingga manajer BUMN tidak dapat dibedakan dengan manajer toko. Berdasarkan hal tersebut, jika tingkat *skill* untuk manajer seluruhnya ditetapkan level 8 KKNI, maka berpotensi terjadi bias dalam pengukuran *mismatch* vertikal serta pengaruhnya terhadap pendapatan. Hal ini karena pendapatan atau gaji dari seorang manajer juga ditentukan oleh jenis perusahaanya. Seperti yang dicontohkan, pendapatan manajer di suatu perusahaan BUMN dengan gaji manajer toko retail, misalnya di suatu Alfamart tentu dapat jauh berbeda.

Berdasarkan fakta tersebut, maka penelitian ini berpatokan pada tingkat pendidikan terendah untuk kriteria *matched* pada tabel 3.2 dalam mengukur *mismatch* pendidikan berdasarkan pendakatan normatif. Untuk pekerjaan manajer, ditetapkan tingkat pendidikan yang diperlukan adalah setara D3, sedangkan untuk profesional, ditetapkan tingkat pendidikan setara S-1. Gambaran mengenai

penentuan kriteria minimal pendidikan dalam penelitian ini adalah sebagaimana berikut:

Tabel 3.3.

Required education Untuk Menentukan *Mismatch* Vertikal (Normatif)

Kode KBJI	Golongan Pokok	*Skill* Level ISCO 2008	Level KKNI Minimal	*Required education* (RE)	Tahun RE
0	Tentara Nasional Indonesia (TNI) dan Kepolisian Negara RI	2	3	SMA/Sederajat	12
1	Manajer	3	5	D3	15
2	Profesional	3	6	S1	16
3	Teknisi dan Asisten Profesional	3	5	D3	15
4	Tenaga Tata Usaha	2	2	SMA/Sederajat	12
5	Tenaga Usaha Jasa dan Tenaga Penjualan	2	2	SMA/Sederajat	12
6	Pekerja Terampil Pertanian, Kehutanan, dan Perikanan	2	2	SMA/Sederajat	12
7	Pekerja Pengolahan, Kerajinan, dan YBDI	2	2	SMA/Sederajat	12
8	Operator dan Perakit Mesin	2	2	SMA/Sederajat	12
9	Pekerja Kasar	1	1	Sekolah Dasar	6

Selanjutnya, untuk menentukan *mismatch* horizontal, penelitian ini menggunakan 2 metode yaitu dengan pendekatan normatif dan statistik dengan menggunakan nilai modus (mode) sebagai *required field*. Pada pendekatan normatif, penelitian ini akan mencocokan sektor pekerjaan dengan *narrow field* ISCED 11 yang berjumlah 57 bidang studi. Namun demikian, dikarenakan data Sakernas memiliki pengkodean tersendiri untuk setiap bidang studinya, maka kode bidang studi dalam Sakernas tersebut akan terlebih dahulu dikonversi ke dalam ISCED 11. Gambaran mengenai kecocokan jurusan/bidang pendidikan dengan kode rumpun ilmu ISCED dan kode sektor lapangan pekerjaan dalam Sakernas untuk menentukan horizontal *mismatch* adalah sebagaimana yang ditunjukan dalam lampiran 1. Sementara itu, kode-kode ISCED 11 yang dirilis pada tahun 2013 secara lebih rinci dapat terlihat dari lampiran 2.

Sektor-sektor lapangan usaha berdasarkan KBLI 2014 yang digunakan dalam penelitian ini untuk menentukan horizontal *mismatch* adalah sebagaimana berikut:

Tabel 3.4.

Sektor-Sektor Lapangan Usaha Berdasarkan KBLI

No	Kode KBLI 2014	Sektor
1	A	Pertanian, Kehutanan & Perikanan
2	B	Pertambangan & Penggalian
3	C	Industri Pengolahan
4	D	Pengadaan Listrik, Gas, Uap/Air Panas & Udara Dingin
5	E	Treatment Air, Treatment Air Limbah, Treatment & Pemulihan
6	F	Konstruksi
7	G	Perdagangan Besar & Eceran; Reparasi & Perawatan Mobil & Sepeda Motor
8	H	Pengangkutan & Pergudangan
9	I	Penyediaan Akomodasi & Penyediaan Makan Minum
10	J	Informasi & Komunikasi
11	K	Aktivitas Keuangan & Asuransi
12	L	Real Estate
13	M, N	Jasa Profesional & Perusahaan
14	O	Administrasi Pemerintahan, Pertahanan & Jaminan Sosial Waj
15	P	Pendidikan
16	Q	Aktivitas Kesehatan Manusia & Aktivitas Sosial
17	R, S, T, U	Jasa Lainnya

Saat ini, sebenarnya telah terdapat pembaruan dalam KBLI yaitu KBLI 2020. Dalam KBLI 2020, terdapat 21 sektor lapangan pekerjaan. Perbedaan antara KBLI 2014 dan 2020 terletak pada klasifikasi sektor M, N dan R, S, T, U. Dalam KBLI 2020, sektor M dan N dipisahkan, sektor M adalah Aktivitas Profesional, Ilmiah Dan Teknis, sedangkan sektor N adalah Aktivitas Penyewaan dan Sewa Guna Usaha Tanpa Hak Opsi, Ketenagakerjaan, Agen Perjalanan dan Penunjang Usaha Lainnya.

Selanjutnya, dalam KBLI 2020, sektor R, S, T, dan U juga seluruhnya dipisahkan. Sektor R adalah Kesenian, Hiburan Dan Rekreasi, S adalah Aktivitas Jasa Lainnya, T adalah Aktivitas Rumah Tangga Sebagai Pemberi Kerja; Aktivitas Yang Menghasilkan Barang Dan Jasa Oleh Rumah Tangga yang Digunakan untuk Memenuhi Kebutuhan Sendiri, dan U adalah Aktivitas Badan Internasional Dan Badan Ekstra Internasional Lainnya.

Dalam KBLI 2014, sektor M dan N disatukan menjadi Jasa Profesional dan Perusahaan, sedangkan sektor R, S, T, dan U disatukan menjadi Jasa Lainnya. Meskipun terdapat perbedaan antara KBLI 2014 dan 2020, namun isi sektor-

sektor lapangan pekerjaan dalam M, N, serta R, S, T, dan U tetaplah sama. Atas dasar itu, KBLI 2014 tetap berkesusaian dengan KBLI 2020. Dalam hal ini, Sakernas 2022 yang dipergunakan sebagai data penelitian ini masih menggunakan KBLI 2014, oleh karenanya, penelitian ini menggunakan KBLI 2014 untuk mengklasifikasi kesesuaian sektor lapangan pekerjaan dengan bidang pendidikan untuk menentukan horizontal *mismatch*.

3.4. Prosedur Penelitian

Prosedur penelitian ini dilakukan dalam 3 langkah yaitu pembentukan model, operasionalisasi variabel, serta pengecekan kekebalan model. Pada tahapan pembentukan model, penelitian ini menjelaskan bentuk-bentuk model penelitian yang berupa persamaan matematis. Model-model tersebut disusun berdasarkan model teoritis yang telah dipaparkan pada BAB II. Sementara itu, operasionalisasi variabel dalam penelitian ini bertujuan untuk memberikan gambaran mengenai konsep teoritis, jenis-jenis variabel penelitian, metode pengukuran variabel, serta jenis data penelitian atas variabel tersebut.

Dalam penelitian ini, terdapat 7 model penelitian yang akan diuji yaitu: determinan *overeducation* (model 1), determinan *undereducation* (model 2), determinan *mismatch* horizonal (model 3), pengaruh *mismatch* vertikal dan horizontal terhadap pendapatan (model 5). Kemudian, pengaruh *mismatch* vertikal dan horzontal terhadap terhadap tingkat pengangguran (model 5), pengaruh *mismatch* vertikal dan horizontal terhadap tingkat pertumbuhan ekonomi (model 6), serta pengaruh *mismatch* vertikal dan horizontal terhadap produktivitas tenaga kerja (model 7).

3.4.1. Model 1 : Determinan *Overeducation*

Berdasarkan model teoritis yang telah ditentukan, model pertama dalam penelitian ini adalah mengenai determinan *overeducation*. Pada model 1 ini, *overeducation* diukur dengan menggunakan 3 metode yaitu normatif, RM atau VV, dan mode. Bentuk dari model 1 adalah sebagaimana berikut:

$$\mathrm{DOVER_i} = \alpha + \beta_j \mathrm{X_{ij}} + \varepsilon_i \text{ (model 1)}$$

dimana, $\mathrm{DOVER_i}$ adalah *overeducation* yang diukur secara dummy atau katagorikal (diisi 1 jika mengalami *overeducation*, dan diisi 0 jika tidak).

Sementara itu, X adalah vector dari variabel-variabel bebas yang mempengaruhi $DOVER_i$. Dimana, $DOVER_i$ berasal dari asumsi jika $S\alpha > Sr = So$ (*overeducation*), $S\alpha = Sr$ = matched, dan $S\alpha < Sr = Su$ (*undereducation*). Dalam hal ini, Sα adalah *school attained* atau tahun sekolah yang dicapai/diselesaikan, Sr adalah *school require*d atau tahun sekolah yang diperlukan untuk mendapatkan pekerjaan, So adalah *school overeducation* atau surplus schooling, dan Su adalah *school undereducation*, atau *deficit schooling*. Sementara itu, $\beta_j X_{ij}$ adalah variabel-variabel bebas yang berupa karakteristik individu, karakteristik institusi pendidikan, karakteristik pekerjaan dan pemberi kerja, dan karakteristik spasial. Informasi lebih detail mengenai variabel-variabel bebas tersebut dapat dilihat dalam tabel 3.2.

Penggunaan *overeducation* sebagai variabel dummy dalam penelitian ini mengacu pada penelitian Verdugo & Verdugo (1989). Selanjutnya, dikarenakan variabel *overeducation* merupakan variabel kategorikal (dummy), maka model 1 akan diestimasi dengan menggunakan metode Logit. Oleh karena itu, bentuk dari $DOVER_i$ adalah odds rasio (Li) sebagaimana berikut :

$$L_i DOVER = \ln\left(\frac{P_i}{1 - P_i}\right)$$

dimana, P_i adalah probabilitas atau kemungkinan terjadinya *overeducation* individu i. Sehingga L_i atau $P_i/1\text{-}P_i$ adalah odds rasio atau peluang seseorang mengalami *overeducation* atau tidak. Nilai odds rasio tersebut selanjutnya dijadikan bentuk logaritma natural (ln), sehingga nilai OV_i berubah menjadi L_i dalam bentuk Logit (log of the odds ratio).

Setelah variabel *overeducation* diubah ke dalam bentuk Logit, maka secara teknis, persamaan model 1 menjadi :

$$L_i DOVER = \ln\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta_j X_{ij} + \varepsilon_i$$

Variabel-variabel X_{ij} dalam model 1 ini terdiri atas variabel-variabel bebas yang tersedia dalam tabel 3.5 tentang operasionalisasi variabel.

3.4.2. Model 2 : Determinan *Undereducation*

Model 2 adalah tentang determinan dari *undereducation*. Sama halnya dengan model 1, model 2 ini juga *undereducation* dengan pendekatan variabel

dummy. Model 2 ini akan menggunakan data Sakernas 2022 dengan bentuk model sebagaimana berikut:

$$\mathrm{DUNDER_i} = \alpha + \beta_j \mathrm{X_{ij}} + \varepsilon_{it} \quad \text{(model 2)}$$

dimana, DUNDER adalah dummy variabel dari *undereducation* yang diukur secara kategorikal (1 jika mengalami *undereducation*, 0 jika tidak). Untuk menentukan *undereducation* ini, dilakukan penghitungan atas asumsi sebagaimana berikut: jika $S\alpha > Sr = So$ (*overeducation*), $S\alpha = Sr$ = matched, dan $S\alpha < Sr = Su$ (*undereducation*). Dalam konteks ini, Sα adalah schooling attained atau tahun sekolah yang dicapai, sedangkan Sr adalah tahun sekolah yang diperlukan (required schooling). Oleh karena itu, DUNDER akan diisi 1 jika $S\alpha < Sr$. Adapun ε adalah error term.

Dalam model 2, X_{ij} adalah variabel-variabel bebas yang berupa karakteristik individu, karakteristik institusi pendidikan, karakteristik pemberi kerja, dan karakteristik spasial. Sementara itu, untuk menentukan berapa nilai S^r, penelitian ini akan digunakan metode normatif, metode statistikal pendekatan mode dan mean (VV). Selanjutnya, dikarenakan variabel *undereducation* bersifat kategorikal atau dummy, maka model 2 ini juga akan dianalisis dengan menggunakan regresi logistik (Logit).

Atas dasar itu, UE akan diubah ke dalam bentuk Logit, sehingga persamaannya adalah sebagaimana berikut:

$$\mathrm{L_i DUNDER} = \ln\left(\frac{\mathrm{P_i}}{1-\mathrm{P_i}}\right) = \alpha + \beta_j \mathrm{X_{ij}} + \varepsilon_i$$

dimana, LiDUNDER adalah nilai Logit atau log odds rasio dari *undereducation*. Nilai a adalah konstanta, sedangkan Xij adalah variabel-variabel eksplanatori yang berupa karakteristik individu, karakteristik institusi pendidikan, karakteristik pekerjaan dan pemberi kerja, dan karakteristik spasial. Secara lebih lengkap, variabel-variabel Xij terdapat pada bagian operasionalisasi variabel

3.4.3. Model 3 : Determinan Horizontal *Mismatch*

Model 3 dalam penelitian ini menjelaskan tentang determinan horizontal *mismatch*. Model 3 dalam penelitian ini adalah sebagaimana berikut:

$$\mathrm{HMM_i} = \alpha + \beta_j \mathrm{X_{ij}} + \varepsilon_i \text{ (model 3)}$$

dimana, HMM adalah horizontal *mismatch* yang diukur secara katagorikal (diisi 1 jika mengalami horizontal *mismatch*, dan diisi 0 jika tidak). Selanjutnya, i adalah individual ke i, ε = terms error. Adapun X_{ij} adalah variabel-variabel bebas yang terdapat dalam operasionalisasi variabel

Pengukuran HMM dalam model 3 ini dilakukan dengan 2 metode yaitu metode normatif dan metode statistikal pendekatan mode. Sama halnya dengan model 1 dan 2, dikarenakan variabel terikat dalam model 3 ini bersifat kategorikal, maka model 3 ini juga akan diestimasi dengan menggunakan metode Logit.

3.4.4. Model 4 : Dampak *Mismatch* Pendidikan Terhadap Pendapatan

Terdapat 4 model yang akan diuji dalam model 4 ini yaitu model 4a, 4b, 4c, dan 4d. Model 4a mengestimasi dampak *mismatch* vertikal pendidikan terhadap pendapatan dengan pendekatan ORU. Dikarenakan menggunakan pendekatan ORU, maka pengaruh *mismatch* vertikal pendidikan terhadap pendapatan adalah tingkat pengembalian dari *overeducation*, *required education*, dan *undereducation*. Dari model ORU ini, akan ditambahkan sejumlah variabel bebas lain untuk mengontrol keragaman individu.

Model 4a ini mengacu pada model ORU yang dikembangkan oleh Duncan & Hoffman (1981) dengan menambahkan variabel-variabel kontrol. Bentuk model 4a adalah sebagaimana berikut:

$$\mathrm{LnIncome_i} = \beta_1 \mathrm{YOVER_i} + \beta_2 \mathrm{YREQ_i} + \beta_2 \mathrm{YUNDER_i} + \beta_j X_{ij} + \varepsilon_i \text{ (model 4a)}$$

dimana, $\mathrm{LnIncome_i}$ adalah pendapatan yang berupa gaji bulanan yang didapatkan oleh individu i dalam bentuk logaritma natural, sedangkan YOVER, YREQ, dan YUNDER merupakan jumlah tahun sekolah/pendidikan yang diperlukan, tahun *overeducation*, dan tahun *undereducation*. Dalam penelitian Duncan & Hoffman (1981), YOVER, YREQ, dan YUNDER ini dituliskan sebagai S^r, S^o, dan S^u. Penelitian ini menuliskan YOVER, YREQ, dan YUNDER agar dapat menegaskan bahwa *mismatch* vertikal pendidikan yang digunakan dalam model ORU ini adalah berupa tahun lebih, tahun yang sesuai, dan tahun kurang pendidikan.

Nilai YOVER adalah YEDUC – YREQ, jika YEDUC> YREQ. Dalam hal ini, YEDUC adalah jumlah tahun pendidikan yang dicapai oleh individu. Sementara itu, nilai YUNDER adalah YREQ – YEDUC, jika YEDUC < YREQ.

Koefisien dari YOVER, YREQ, dan YUNDER tersebut akan menunjukkan tingkat pengembalian (*rate of return*) dari *overeducation*, *required education*, dan *undereducation*. Sementara itu, X_{ij} adalah vektor yang merupakan variabel-variabel kontrol (termasuk di dalamnya konstanta). Dalam konteks ini, X_{ij} adalah variabel-variabel bebas yang berupa usia (age), pengalaman (experience), pengalaman kuadrat (experience^2), wilayah pedesaan (rural), migrasi (migration), dan pekerja paruh waktu (half timer), sedangkan ε_i adalah error term.

Selanjutnya, dalam model 4b, penelitian ini juga akan menguji pengaruh *mismatch* vertikal pendidikan terhadap pendapatan individu dengan menggunakan pendekatan Verdugo & Verdugo (1989). Dalam pendekatan ini, *overeducation* dan *undereducation* diposisikan sebagai variabel dummy. Persamaan pendapatan dari Verdugo & Verdugo (1989) ini merupakan modifikasi dari model ORU dengan mengganti S^o, S^r, dan S^u dengan variabel dummy overducation (DOVER) dan *undereducation* (DUE). Tujuan utama dari model ini adalah untuk menelaah wage penalty dan wage premium secara lebih jelas. Dengan mengacu pada pendekatan Verdugo & Verdugo (1989) tersebut, maka model 4b dalam penelitian ini adalah:

$$\mathrm{LnIncome_i} = \beta_1 \mathrm{EDUC_i} + \beta_2 \mathrm{DOVER_i} + \beta_3 \mathrm{DUNDER_i} + \beta_j \mathrm{X_{ij}} + \varepsilon_i \text{ (model 4b)}$$

dimana, $\mathrm{LnIncome_i}$ adalah logaritma natural dari pendapatan individu i. EDUC adalah tahun sekolah yang diraih oleh individu i, sedangkan DOVER dan DUNDER merupakan *overeducation* dan *undereducation* yang diukur secara kategorikal (dummy variabel). Adapun Xij adalah variabel-variabel kontrol yang berupa usia (age), pengalaman (experience), pengalaman kuadrat (experience^2), wilayah pedesaan (rural), migrasi (migration), dan pekerja paruh waktu (half timer).

Adapun model 4c dalam penelitian ini adalah untuk menelaah bagaimana pengaruh horizontal *mismatch* terhadap pendapatan individu. Dalam penelitian ini, horizontal *mismatch* tidak disatukan dalam vertikal *mismatch* atas pengaruhnya terhadap pendapatan individu. Hal ini karena horizontal *mismatch* dalam penelitian ini diasumsikan hanya terjadi pada tenaga kerja dengan tingkat pendidikan minimal SMA/sederajat. Padahal, dalam model 4a dan 4b akan melibatkan seluruh sampel tenaga kerja. Akibatnya, jika horizontal *mismatch* tidak dipisahkan, makan jumlah observasi sampel yang diestimasi akan menjadi jauh

lebih rendah. Selain itu, dapat juga terjadi persoalan bias seleksi yang sulit diantisipasi.

Bentuk model 4c dalam penelitian ini adalah sebagaimana berikut:

$$\text{LnIncome}_i = \beta_1 \text{HMM}_i + \beta_2 \text{EDUC}_i + \beta_j X_{ij} + \varepsilon_i \text{ (model 4c)}$$

dimana, LnIncome_i adalah logaritma natural dari pendapatan individu i. EDUC adalah tahun sekolah yang diraih oleh individu i, sedangkan HMM adalah variabel horizontal *mismatch* yang diukur dengan menggunakan metode normatif serta metode mode. Adapun Xij adalah variabel-variabel kontrol yang berupa usia (age), pengalaman (experience), pengalaman kuadrat (experience^2), wilayah pedesaan (rural), migrasi (migration), dan pekerja paruh waktu (half timer).

Model 4a, 4b, dan 4c ini akan diestimasi dengan menggunakan metode Heckman selection sample model (Heckit model). Hal ini dilakukan untuk mengontrol kemungkinan bias seleksi dari sampel dengan status tidak bekerja. Selain itu, metode Heckit ini juga dapat mengantisipasi kemungkinan adanya masalah endogenitas. Penjelasan lebih lengkapnya dituliskan dalam sub bab 3.5.3.

Terakhir, model tentang pengaruh *mismatch* pendidikan terhadap pendapatan individu yang juga akan diuji adalah model 4d. Model 4d ini akan mengekslusikan sampel dengan jenis pekerjaan militer dan pekerja kasar (kode 0 dan 9 dalam KBJI). Selain itu, model 4d ini akan juga mengekslusikan sampel dengan tingkat pendidikan di bawah Perguruan Tinggi. Dengan demikian, tingkat pendidikan terendah sampel dalam model 4a ini adalah diploma 3 (15 tahun). Meskipun lulusan diploma 1 dan 2 juga termasuk lulusan perguruan tinggi, tetapi database Sakernas 2022 tidak merincinya, sehingga tidak dapat teridentifikasi.

Tujuan utama dibentuknya model 4d adalah untuk menguji dampak *mismatch* pendidikan terhadap pendapatan individu pada pekerjaan dengan tingkat pendidikan tinggi. Penelitian ini mengasumsikan bahwa tenaga kerja lulusan perguruan tinggi akan menjalani pekerjaan yang memerlukan tingkat pendidikan tinggi juga. Dimana, model 4d ini akan dibagi menjadi 2 model yaitu 4d1 dan 4d2. Model 4d1 menggunakan pendekatan ORU, sedangkan model 4d2 menggunakan pendekatan Verdugo & Verdugo (1989).

Dalam model 4d1 maupun 4d2 ini, vertikal dan horizontal *mismatch* akan disatukan karena jumlah sampelnya tidak mengalami perbedaan. Bentuk model 4d1 adalah:

$$\text{LnIncomehour}_i = \beta_1 \text{YUNDER}_i + \beta_2 \text{YOVER}_i + \beta_3 \text{YREQ}_i + \beta_4 \text{HMM}_i + \beta_j X_{ij} + \varepsilon_i \quad \text{(model 4d1)}$$

dimana, LnIncomehour adalah logaritma natural dari pendapatan per-jam, YUNDER, YOVER, dan YREQ adalah tahun kurang pendidikan, lebih, dan tahun yang diperlukan. Koefisien YOVER menunjukkan tingkat pengembalian dari *overeducation*, YUNDER berarti return to *undereducation*, dan YREQ adalah return to *required education*. HMM adalah horizontal *mismatch* yang dibuat dalam variabel dummy, diukur dengan metode normatif. Xj adalah variabel-variabel untuk mengontrol heterogenitas individu yakni usia (age), pengalaman (experience), pengalaman kuadrat (experience^2), wilayah pedesaan (rural), migrasi (migration), dan pekerja paruh waktu (half timer).

Adapun model 4d2 adalah sebagaimana berikut:

$$\text{LnIncomehour}_i = \beta_1 \text{EDUC}_i + \beta_2 \text{DOVER}_i + \beta_3 \text{DUNDER}_i + \beta_4 \text{HMM}_i + \beta_j X_{ij} + \varepsilon_i \quad \text{(model 4d2)}$$

dimana, LnIncomehour adalah logaritma natural dari pendapatan per-jam. EDUC adalah tahun pendidikan yang diraih individu, DOVER dan DUNDER adalah variabel dummy dari *overeducation* dan *undereducation*. HMM adalah variabel dummy dari horizontal misamtch yang diukur dengan metode normatif. Xj adalah variabel-variabel kontrol yang sama dengan pada model 4d1.

Pada model 4d1 dan 4d2, *mismatch* vertikal pendidikan diukur dengan metode VV atau realized match. Seseorang dinyatakan *undereducation* apabila tingkat pendidikannya kurang dari 1 standar deviasi rata-rata pendidikan (kelompok okupasi/pekerjaan) yang diperlukan. Sebaliknya, jika tingkat pendidikannya lebih dari 1 standar deviasi rata-rata pendidikan yang diperlukan, maka akan dinyatakan *overeducation*. Jika tidak masuk ke dalam kriteria itu, maka nilai untuk *undereducation* dan *overeducation* akan menjadi 0.

Model 4d hanya mengukur *mismatch* vertikal pendidikan dengan metode VV karena sampel tenaga kerja minimal pendidikan perguruan tinggi cenderung jauh lebih rendah dengan full sampel data Sakernas 2022. Sebaran tingkat

pendidikannya juga relatif tidak terlampau jauh (dibuktikan dengan nilai range yang rendah), sehingga akan lebih akurat jika menggunakan metode VV daripada yang lain (normatif atau mode).

3.4.5. Model 5 : Dampak *Mismatch* Pendidikan Terhadap Pengangguran

Model 5 dalam penelitian ini adalah tentang pengaruh *mismatch* pendidikan terhadap pengangguran. Data yang akan digunakan untuk mengestimasi model ini adalah data panel tingkat provinsi yang berasal dari Sakernas dan Susenas tahun 2012 hingga 2022. Tingkat pengangguran dalam penelitian ini merupakan indikator makro tingkat regional. Oleh karena itu, objek data yang diobservasi bukan lagi individu, melainkan wilayah provinsi. Dengan demikian, *mismatch* vertikal pendidikan (*overeducation* dan *undereducation*) serta *mismatch* horizontal harus disusun dalam bentuk indeks regional (provinsi).

Meskipun data Sakernas adalah data sampling yang didapatkan dari setiap individu di berbagai wilayah di Indonesia, tetapi data ini dapat secara langsung diterapkan ke dalam populasi. Hal ini karena Sakernas menyediakan variabel pembobot (weight) yang dapat secara langsung diberlakukan ke dalam populasi. Dengan demikian, jumlah persentase *overeducation* dan *undereducation* yang didapatkan dari data Sakernas sudah menunjukkan persentase yang sebenarnya. Bahkan, data-data ketenagakerjaan yang tersedia di BPS seperti TPT, persentase penduduk bekerja, persentase angkatan kerja, dan indikator-indikator ketenagakerjaan lainnya juga berasal dari data Sakernas.

Tingkat pengangguran dalam penelitian ini diukur secara akumulatif dalam tingkat regional. BPS mengukur tingkat pengangguran ini melalui tingkat pengangguran terbuka (TPT) dengan formula berikut ini:

$$\text{TPT} = \frac{\text{PP}}{\text{PAK}} \text{x} 100\%$$

dimana, TPT adalah persentase tingkat pengangguran terbuka, PP adalah jumlah pengangguran (dalam orang), dan PAK adalah jumlah angkatan kerja (jumlah orang).

Data TPT ini bersifat regional pada provinsi, sehingga ukuran dari *mismatch* vertikal maupun horizontal juga harus disusun dalam tingkat regional. Terdapat 3 jenis data yang dapat digunakan sebagai ukuran dari *mismatch* vertikal dan horizontal. Pertama, *overeducation* dan *undereducation* diposisikan sebagai

rata-rata tahun lebih dan kurang pendidikan di tingkat provinsi. Kedua, *overeducation*, *undereducation*, dan HMM dibuat ke dalam jumlah tenaga kerja yang kemudian dikonversi ke dalam bentuk logaritma. Ketiga, *overeducation*, *undereducation*, dan HMM merupakan nilai persentase. Akan tetapi, dikarenakan BPS tidak menyediakan data *mismatch* dalam tingkat regional, maka penelitian ini melakukan agregasi atas *mismatch* pendidikan yang berasal dari data Sakernas 2012 hingga 2022. Ketiga ukuran *mismatch* tersebut diukur dengan menggunakan metode normatif.

Dengan demikian, model 5 ini akan menghasilkan sebanyak 9 model yang bermula dari model 5a, 5b, dan 5c. Model 5a, terdiri atas 5a1, 5a2, dan 5a3, sedangkan model 5b terdiri atas model 5b1, 5b2, dan 5b3. Adapun model 5c terdiri atas model 5c1, 5c2, dan 5c3. Model 5a memposisikan *overeducation* dan *undereducation* sebagai rata-rata tahun lebih dan tahun kurang pendidikan di suatu provinsi. Model 5b menggunakan nilai logaritma dari jumlah tenaga kerja yang mengalami *undereducation*, *overeducation*, dan *well match*ed. Sementara itu, model 5c menggunakan nilai persentase dari *undereducation*, *overeducation*, dan *well match*ed.

Variabel *overeducation* dalam model 5a1 yang diposisikan sebagai rata-rata tahun lebih pendidikan di tingkat provinsi dihitung dengan cara berikut ini:

$$\mathrm{YOVER_{it}} = \frac{\Sigma \mathrm{OVER_{emp,it}}}{\Sigma \mathrm{EMP_{it}}}$$

dimana, YOVER adalah rata-rata jumlah tahun pendidikan berlebih di provinsi i pada tahun t. $\mathrm{OVER_{emp}}$ adalah jumlah tahun pendidikan berlebih setiap tenaga kerja di provinsi i pada tahun t, sedangkan EMP adalah tenaga kerja keseluruhan.

Sementara itu, *undereducation* yang diposisikan sebagai rata-rata tahun kurang pendidikan di tingkat provinsi dihitung dengan cara:

$$\mathrm{YUNDER_{it}} = \frac{\Sigma \mathrm{UNDER_{emp,it}}}{\Sigma \mathrm{EMP_{it}}}$$

dimana, YUNDER adalah rata-rata jumlah tahun pendidikan kurang di provinsi i pada tahun t. $\mathrm{UNDER_{emp}}$ adalah jumlah tahun pendidikan kurang setiap tenaga kerja di provinsi i pada tahun t, sedangkan EMP adalah tenaga kerja keseluruhan.

Adapun YREQ pada tingkat provinsi diestimasi sebagaimana berikut:

$$\mathrm{YREQ_{it}} = \mathrm{EDUC_{it}} + \mathrm{YUNDER_{it}} - \mathrm{YOVER_{it}}$$

dimana, EDUC adalah rata-rata tahun pendidikan yang dicapai pada provinsi i tahun t. Rata-rata tahun pendidikan yang dicapai ini kemudian ditambah dengan nilai pendidikan kurang (YUNDER) dan dikurangi pendidikan berlebih (YOVER).

Dikarenakan YUNDER, YOVER, dan YREQ akan saling berkorelasi satu sama lain, maka estimasinya harus dipisahkan. Dengan demikian, model 5a1 adalah sebagaimana berikut:

$$TPT_{it} = \alpha + \beta_1 YUNDER_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \quad \text{(model 5a1)}$$

dimana, TPT adalah tingkat pengangguran terbuka, YUNDER merupakan rata-rata tahun pendidikan kurang di setiap provinsi i pada tahun t. HMM merupakan persentase tenaga kerja yang mengalami horizontal *mismatch* di suatu provinsi. Adapun Xij adalah variabel-variabel untuk mengontrol heterogenitas antar provinsi yaitu LogDI, LogFDI, GINI, dan Inflasi. HMM dalam model di atas, didapatkan dari formula berikut:

$$HMM_{it} = \frac{\Sigma HMM_{emp,it}}{\Sigma EMP_{emp,it}}$$

dimana, HMM adalah persentase tenaga kerja horizontal *mismatch* tingkat regional. HMM didapatkan dari jumlah pekerja yang mengalami horizontal *mismatch* (dalam orang), dan EMP adalah jumlah tenaga kerja dengan tingkat pendidikan minimal SMA/sederajat.

Model 5a1 mengestimasi pengaruh YUNDER dan HMM terhadap TPT, sedangkan untuk mengestimasi pengaruh YOVER dan HMM terhadap TPT adalah sebagaimana berikut:

$$TPT_{it} = \alpha + \beta_1 YOVER_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \quad \text{(model 5a2)}$$

dimana, YOVER adalah rata-rata tahun lebih pendidikan di tingkat provinsi. Xij variabel-variabel untuk mengontrol heterogenitas antar provinsi yaitu LogDI, LogFDI, GINI, dan INLFASI.

Selanjutnya, untuk menelaah pengaruh tahun pendidikan yang diperlukan (*required education*) dalam konteks makro, bentuk modelnya adalah:

$$TPT_{it} = \alpha + \beta_1 YREQ_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \quad \text{(model 5a3)}$$

dimana, YREQ merupakan rata-rata tahun pendidikan yang diperlukan untuk mendapatkan pekerjaan dalam konteks agregatif. YREQ ini juga dapat

mengindikasikan tingkat perkembangan teknologi dalam pasar tenaga kerja. Hal ini karena semakin tinggi YREQ, berarti kebutuhan tenaga kerja berpendidikan tinggi juga menjadi semakin besar. Tingginya kebuthan tenaga kerja berpendidikan tinggi tersebut mencirikan bahwa industri memerlukan tingkat tekknologi yang tinggi.

Adapun bentuk dari model 5b1 adalah sebagaimana berikut:

$$\mathrm{TPT_{it}} = \alpha + \beta_1 \mathrm{LogUnder_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \quad \text{(model 5b1)}$$

dimana, LogUnder merupakan nilai logaritma dari jumlah total tenaga kerja yang mengalami *undereducation*. HMM adalah persentase horizontal *mismatch*, sedangkan Xj terdiri atas LogDI, LogFDI, GINI, dan Inflasi.

$$\mathrm{TPT_{it}} = \alpha + \beta_1 \mathrm{LogOver_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \quad \text{(model 5b2)}$$

dimana, LogOver adalah nilai logaritma dari jumlah total tenaga kerja yang mengalami *overeducation* pada provinsi i di tahun t. Sedangkan variabel-variabel Xj yaitu LogDI, LogFDI, GINI, dan Inflasi.

Bentuk dari model 5b3 yang menelaah nilai logaritma dari tenaga kerja yang *well match*ed (*required education*) terhadap tingkat pengangguran adalah sebagaimana berikut:

$$\mathrm{TPT_{it}} = \alpha + \beta_1 \mathrm{LogMatch_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \quad \text{(model 5b3)}$$

dimana, LogMatch adalah nilai logaritma dari jumlah tenaga kerja yang dinyatakan *well match*ed di suatu Provinsi i pada tahun t. HMM adalah persentase tenaga kerja yang mengalami horizontal *mismatch* yang dibandingkan dengan total tenaga kerja. Xj adalah variabel-variabel untuk mengontrol heterogenitas yaitu LogDI, LogFDI, GINI, dan Inflasi.

Sebagai upaya memastikan model yang terbaik, penelitian ini juga mengestimasi pengaruh *undereducation*, *overeducation*, dan matched terhadap TPT berdasarkan nilai persentasenya. Penelitian ini menetapkan persentase *undereducation* sebagaimana berikut:

$$\mathrm{UNDERPERCENT} = \frac{\Sigma \mathrm{UE}}{\Sigma \mathrm{EMP}}$$

dimana, UNDERPRECENT adalah indeks *undereducation* tingkat regional, UE adalah jumlah *undereducation* (dalam orang), dan EMP adalah jumlah orang yang bekerja. Dengan demikian, formula untuk menghitung OVERPERCENT adalah:

$$\text{OVERPERCENT} = \frac{\Sigma \text{OV}}{\Sigma \text{EMP}}$$

dimana, OVERPERCENT adalah indeks *overeducation* tingkat regional (dalam persen), OV adalah jumlah *overeducation* (dalam orang), dan EMP adalah jumlah orang yang bekerja. Adapun indeks Match (persentase tenaga kerja yang well-matched) adalah:

$$\text{MATCHPERCENT} = \frac{\Sigma \text{MATCH}}{\Sigma \text{EMP}}$$

dimana, MATCHPERCENT adalah indeks *well match* dalam persen, MATCH adalah jumlah tenaga kerja yang dinyatakan *well match*, sedangkan EMP adalah jumlah orang yang bekerja.

Dengan memposisikan undereduation sebagai persentase tenaga kerja yang mengalami *undereducation* di suatu Provinsi, maka model 5c1 adalah:

$$\text{TPT}_{it} = \alpha + \beta_1 \text{UNDERPERCENT}_{it} + \beta_2 \text{HMM}_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 5c1)}$$

dimana, TPT_{it} adalah tingkat pengangguran terbuka di Provinsi i pada waktu t, sedangkan UNDERPERCENT adalah persentaese *undereducation* yang terjadi di Provinsi i pada tahun t jika dibandingkan dengan *overeducation* dan *well match*. HMM adalah persentase tenaga kerja yang mengalami horizontal *mismatch* apabila dibandingkan dengan total tenaga kerja. Sementara itu, $\beta_j X_{itj}$ merupakan variabel-variabel untuk mengontrol heterogenitas antar individu Provinsi yaitu LogDI, LogFDI, GINI, dan inflasi. Sedangkan ε_{it} adalah error

Selanjutnya, model yang menjelaskan pengaruh *overeducation* terhadap tingkat pengangguran dengan memposisikan *overeducation* sebagai persentase tenaga kerja yang mengalami *overeducation* di suatu Provinsi adalah:

$$\text{TPT}_{it} = \alpha + \beta_1 \text{OVERPERCENT}_{it} + \beta_2 \text{HMM}_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 5c2)}$$

dimana, OVERPERCENT adalah persentase tenaga kerja yang mengalami *overeducation* di Provinsi i pada tahun t jika dibandingkan dengan *undereducation*, dan *well match*. Adapun model 5c3 adalah sebagaimana berikut:

$$\text{TPT}_{it} = \alpha + \beta_1 \text{MATCHPERCENT}_{it} + \beta_2 \text{HMM}_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 5c3)}$$

dimana, MATCHPERCENT adalah persentase tenaga kerja yang dinyatakan *well match* di Provinsi i, pada tahun t.

Mismatch vertikal maupun horizontal yang diestimasi dalam model 5 ini seluruhnya diestimasi dengan menggunakan metode normatif. Alasan utamanya adalah karena metode tersebut cenderung lebih mudah diimplementasikan serta ukurannya bersifat umum. Maksudnya adalah, tingkat pendidikan yang diperlukan berdasarkan metode normatif dapat berlaku di setiap sektor lapangan kerja. Sebagai contoh, tingkat pendidikan untuk menjadi seorang profesional di DKI Jakarta akan sama dengan tingkat pendidikan yang diperlukan untuk menjadi profesional di Bali, maupun provinsi-provinsi lainnya.

3.4.6. Model 6 : Dampak *Mismatch* Pendidikan Terhadap Pertumbuhan Ekonomi

Model 6 dalam penelitian ini menelaah pengaruh *mismatch* pendidikan, baik itu vertikal maupun horizontal terhadap pertumbuhan ekonomi. Data yang akan digunakan dalam menganalisis model 6 ini adalah data Sakernas 2012 sampai dengan 2022. Dalam model ini, pertumbuhan ekonomi yang dimaksud adalah pertumbuhan ekonomi di tingkat regional (provinsi) dengan menggunakan proksi real growth GDP. Sementara itu, indeks *overeducation*, *undereducation*, dan horizontal *mismatch*, diukur sebagaimana yang dilakukan pada model 5.

Pertumbuhan ekonomi dalam penelitian ini diukur dengan menggunakan formula berikut ini:

$$\mathrm{GROWTH_{it}} = \frac{\mathrm{RealGDP_{it} - RealGDP_{it-1}}}{\mathrm{RealGDP_{it}}} \mathrm{x} 100\%$$

dimana, Growth adalah growth GDP di Provinsi i, pada tahun t. Real GDP adalah pertumbuhan GDP yang diukur.

Sebagaimana halnya model 5, model 6 ini juga akan diuji menjadi 9 model. Pada model 6a1, 6a2, dan 6a3, *undereducation*, *overeducation*, dan *required education* akan diposisikan sebagai tahun kurang, lebih atau tahun pendidikan yang diperlukan. Hal ini mengacu pada model ORU tetapi diterapkan dengan menggunakan data panel pada tingkat analisis makro. Bentuk modelnya adalah sebagaimana berikut:

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{YUNDER_{it}} + \beta_2 \mathrm{YOVER_{it}} + \beta_3 \mathrm{YREQ_{it}} + \beta_4 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6a)}$$

dimana, $GROWTH_{it}$ adalah pertumbuhan ekonomi provinsi i pada waktu t yang diukur dengan menggunakan real Growth GDP, sedangkan YUNDER, YOVER, dan YREQ adalah rata-rata tahun kurang, lebih dan tahun pendidikan yang diperlukan. Sementara itu, HMM adalah persentase horizontal *mismatch* pada suatu provinsi i di tahun t. Sementara itu, $\beta_j X_{itj}$ merupakan variabel-variabel untuk mengontrol heterogenitas individu (dalam hal ini adalah LogDI, LogFDI, GINI, dan Inflasi), adapun ε_{it} adalah error.

Namun demikian, model 6a di atas akan mengalami persoalan multikolinieritas karena hampir pasti terjadi korelasi yang sangat tinggi antara YUNDER, YOVER, dan YREQ. Hal ini karena YUNDER adalah YREQ – YEDUC, jika YEDUC < YREQ. YOVER adalah YREQ – YEDUC, jika YEDUC> YREQ, sedangkan YREQ adalah YEDUC + YUNDER – YOVER. Atas dasar itu, model 6a di atas akan diestimasi secara bertahap dengan sebagaimana berikut:

$$GROWTH_{it} = \alpha + \beta_1 YUNDER_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 6a1)}$$

Dalam model 6a1 di atas, nilai dari YUNDER adalah rata-rata tahun kurang pendidikan secara agregatif dalam suatu provinsi i pada tahun t. Sementara itu, HMM adalah nilai persentase tenaga kerja dengan tingkat pendidikan minimal SMA/sederajat yang mengalami horizontal *mismatch* di suatu provinsi i pada tahun t. HMM tersebut dikalkukasi dengan menggunakan metode normatif. Metode ini mencocokan antara jurusan pendidikan yang diidentidikasi menggunakan ISCED dengan sektor jurusan yang teridentifikasi berdasarkan satu digit KBLI.

Untuk menganalisis tingkat pengembalian dari tahun lebih pendidikan (*overeducation*), modelnya adalah sebagaimana berikut:

$$GROWTH_{it} = \alpha + \beta_1 YOVER_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 6a2)}$$

Sama halnya dengan model 6a1, HMM dalam model 6a2 juga merupakan nilai persentase tenaga kerja yang mengalami horizontal *mismatch* dengan dibandingkan dengan total tenaga kerja. HMM juga ditentukan dengan menggunakan metode normatif. Model 6a3 adalah sebagai berikut:

$$GROWTH_{it} = \alpha + \beta_1 YREQ_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 6a3)}$$

Model 6a3 menguji pengaruh *required education* yang diukur sebagai tahun yang diperlukan untuk mendapatkan pekerjaan pada suatu provinsi. Tinggi rendahnya YREQ di suatu Provinsi menunjukkan kondisi permintaan tingkat pendidikan tenaga kerja. Jika YREQ tinggi, maka tenaga kerja yang diperlukan oleh sektor lapangan pekerjaan di Provinsi tersebut cenderung merupakan tenaga kerja berpendidikan tinggi.

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{LogUnder_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6b1)}$$

Model 6b1 menelaah pengaruh *undereducation* yang nilainya adalah logaritma dari jumlah tenaga kerja yang mengalami *undereducation* di suatu Provinsi. Variabel-variabel kontrol yang digunakan tetap sama, yaitu LogDI, LogFDI, GINI, dan INFLASi.

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{LogOver_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6b2)}$$

LogOver adalah nilai logaritma dari jumlah tenaga kerja yang mengalami *overeducation* di suatu Provinsi i pada tahun t.

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{LogMatch_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6b3)}$$

dimana, LogMatch merupakan nilai logaritma dari jumlah tenaga kerja yang match pada suatu Provinsi i pada waktu t. Model 6b3 ini untuk mengetahui bagaimana gambaran pengaruh tingkat match terhadap pertumbuhan ekonomi. Adapun model 6c1 yang memposisikan *mismatch* sebegai rata-rata persentase di tiap Provinsi adalah sebagaimana berikut:

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{UNDERPERCENT_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6c1)}$$

UDERPERCENT merupakan *undereducation* yang dibentuk dalam persentase. Bentuk persentase-nya adalah desimal. HMM adalah horizontal *mismatch* yang diukur dalam bentuk persentase.

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{OVERPERCENT_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6c2)}$$

Model 6c2 berfokus untuk menguji pengaruh *overeducation* yang diposisikan dalam bentuk persentase terhadap pertumbuhan ekonomi.

$$\mathrm{GROWTH_{it}} = \alpha + \beta_1 \mathrm{MATCHPERCENT_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 6c3)}$$

Model 6c3 menguji bagaimana pengaruh persentase tenaga kerja yang match terhadap pertumbuhan ekonomi.

3.4.7. Model 7 : Dampak *Mismatch* Pendidikan Terhadap Produktivitas Tenaga Kerja

Model 7 menelaah tentang dampak dari *mismatch* pendidikan, baik itu vertikal maupun horizontal terhadap produktivitas tenaga kerja. Sama halnya dengan model 5 dan 6, *mismatch* pendidikan diukur dengan menggunakan indeks *mismatch* pendidikan regional yang berasal dari Sakernas 2012 hingga 2022. Sementara itu, produktivitas dalam penelitian ini merupakan ukuran dari total output yang diproduksi oleh per-unit tenaga kerja. Output tersebut dapat diukur dengan menggunakan pendapatan, sedangkan inputnya adalah jam kerja. Oleh karenanya, pengukuran produktivitas tenaga kerja dalam penelitian ini yakni:

$$\mathrm{LP} = \frac{\text{Rata} - \text{Rata Pendapatan Tenaga Kerja}}{\text{Rata} - \text{Rata Jam Kerja Tenaga Kerja}}$$

dimana, LP adalah labour productivity di tingkat Provinsi. Sehingga, rata-rata pendapatan tenaga kerja didapatkan dari rata-rata pendapatan seluruh tenaga kerja di suatu Provinsi. Nilai tersebut selanjutnya dibagi dengan rata-rata jam kerja tenaga kerja per-Bulan. LP dalam penelitian ini menunjukkan berapa banyak pendapatan yang dihasilkan oleh tenaga kerja untuk setiap jam kerjanya.

Sama halnya dengan model 5 dan 6, model 7 dalam penelitian ini juga akan menghasilkan 9 model. Model 7a1, 7a2, dan 7a3 memposisikan *undereducation*, *overeducation*, dan required match sebagai surplus pendidikan, defisit pendidikan, dan tahun pendidikan yang diperlukan. Model 7b1, 7b2, 7b3 memposisikan *mismatch* vertikal sebagai

Bentuk dari model 7a1 dalam penelitian ini adalah sebagaimana berikut:

$$\mathrm{LP_{it}} = \alpha + \beta_1 \mathrm{YUNDER_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 7a1)}$$

dimana, LP adalah produktivitas tenaga kerja pada wilayah regional i dan waktu t, sedangkan YUNDER adalah indeks defisit pendidikan, dan HMM adalah persentase horizontal *mismatch*. Adapun Xj merupakan variabel-variabel untuk mengontrol heterogenitas (dalam hal ini adalah LogDI, LogFDI, GINI, INFLASI, TPT, dan SEKFOR).

Dalam model 7a2 dan 7a3, YUNDER diganti dengan YOVER dan YREQ. Adapun bentuk dari model 7b1 adalah:

$$\mathrm{LP_{it}} = \alpha + \beta_1 \mathrm{LogUnder_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \varepsilon_{it} \text{ (model 7b1)}$$

Model 7b2 dan 7b3 mengganti LogUnder dengan LogOver dan LogMatch. Sementara itu, bentuk model 7c1 adalah:

$$LP_{it} = \alpha + \beta_1 UNDERPERCENT_{it} + \beta_2 HMM_{it} + \beta_j X_{itj} + \varepsilon_{it} \text{ (model 7c1)}$$

Model 7c2 dan 7c3 menggantikan UNDERPERCENT dengan OVERPERCENT dan MATCHPERCENT. Untuk variabel-variabel kontrol yang digunakan, seluruhnya sama untuk model 7 ini yaitu LogDI, LogFDI, GINI, INFLASI, TPT, dan SEKFOR.

3.4.8. Operasionalisasi Variabel

Variabel terikat dalam model 1 hingga 3 dalam penelitian ini adalah *overeducation*, *undereducation*, dan horizontal *mismatch*. Ketiga variabel tersebut diukur secara kategorikal dengan menggunakan dummy variabel. Metode pengukuran *overeducation* dalam model 1 ini akan menggunakan 3 pendekatan yaitu dengan pendekatan normatif, pendekatan statistik dengan nilai mean dan mode. Sementara itu, variabel-variabel bebas dalam model 1 hingga 3 tersebut adalah variabel yang masuk ke dalam karakteristik individu, karakteristik insitusi pendidikan, karakteristik pemberi kerja, dan karakteristik wilayah. Variabel-variabel terikat yang digunakan dalam model 1 hingga 3 tersebut selanjutnya juga akan digunakan sebagai variabel kontrol dalam menguji model 4 hingga 7.

Operasionalisasi variabel yang menjelaskan tentang definisi operasional, metode pengukuran, dan jenis data dari variabel-variabel dalam penelitian ini dapat dilihat dari tabel sebagaimana berikut:

Tabel 3.5.

Operasionalisasi Variabel Penelitian

Kode	Variabel	Definisi	Pengukuran	Jenis Data
DV.1	*Overeducation* (DOVER)	Suatu kondisi dimana tingkat pendidikan yang dicapai lebih tinggi dari tingkat pendidikan yang diperlukan untuk mendapatkan pekerjaan	1. Normatif. Diukur sebagai variabel dummy (nilai 1 jika *overeducation*, dan 0 jika tidak). Individu dinyatakan mengalami *overeducation* jika tingkat pendidikan yang dicapainya lebih besar daripada tingkat pendidikan yang diperlukan (*required education* atau dilambangkan juga dengan S^r). Penentuan S^r dilakukan dengan metode normatif atau job analysis dengan menyocokan jenis pekerjaan dengan KBJI 2014 dan ISCO 2008 (lihat tabel	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
			3.2). 2. Metode VV. Diukur sebagai variabel dummy (nilai 1 jika *overeducation*, dan – jika tidak). Individu dinyatakan mengalami *overeducation* jika tingkat pendidikan yang dicapainya lebih besar daripada tingkat pendidikan yang diperlukan (*required education* atau dilambangkan juga dengan S^r). Penentuan S^r dilakukan dengan metode statistik dengan pendekatan Verdugo & Verdugo (1989) (VV). Nilai S^r adalah nilai rata-rata pendidikan sekelompok individu pada jenis pekerjaan/jabatan tertentu. Jika terdapat individu yang tahun capaian pendidikannya lebih besar dari 1 standar deviasi/2 dari S^r, maka dinyatakan *overeducation*. 3. Mode (modal procedure). Diukur sebagai variabel dummy (nilai 1 jika *overeducation*, dan – jika tidak). Individu dinyatakan mengalami *overeducation* jika tingkat pendidikan yang dicapainya lebih besar daripada tingkat pendidikan yang diperlukan (*required education* atau dilambangkan juga dengan S^r). Penentuan S^r dilakukan dengan metode statistik dengan pendekatan Kiker et al. (1997). Nilai S^r adalah nilai modus (mode) pendidikan sekelompok individu pada jenis pekerjaan/jabatan tertentu. Jika nilai tahun capaian pendidikan individu lebih tinggi daripada nilai modus, maka individu tersebut mengalami *overeducation*	
DV.2	*Undereducation* (DUE)	Kondisi dimana tingkat pendidikan yang dicapai lebih rendah dari pekerjaan yang didapatkan	1. Normatif. Diukur sebagai variabel dummy (nilai 1 jika *undereducation*, dan 0 jika tidak). Individu dinyatakan mengalami *undereducation* jika tingkat pendidikan yang dicapainya lebih rendah daripada tingkat pendidikan yang diperlukan (*required education* atau dilambangkan juga dengan S^r). Penentuan S^r dilakukan dengan metode normatif atau job analysis dengan menyocokan jenis pekerjaan dengan KBJI 2014 dan ISCO 2008 (lihat tabel 3.2). 2. Metode VV. Diukur sebagai variabel dummy (nilai 1 jika *undereducation*, dan 0 jika tidak). Penentuan S^r dilakukan dengan metode statistik dengan pendekatan Verdugo & Verdugo (1989) (VV). Nilai S^r adalah nilai rata-rata pendidikan sekelompok	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
			individu pada jenis pekerjaan/jabatan tertentu. Jika terdapat individu yang tahun capaian pendidikannya lebih rendah dari 1 standar deviasi/2 dari S^r, maka dinyatakan *undereducation*. 3. Mode (modal procedure). Diukur sebagai variabel dummy (nilai 1 jika *undereducation*, dan – jika tidak). Penentuan S^r dilakukan dengan metode statistik dengan pendekatan Kiker et al. (1997). Nilai S^r adalah nilai modus (mode) pendidikan sekelompok individu pada jenis pekerjaan/jabatan tertentu. Jika nilai tahun capaian pendidikan individu lebih rendah daripada nilai modus, maka individu tersebut mengalami *overeducation*	
DV.3	*Mismatch* Horizontal (HMM)	Kondisi dimana pekerjaan yang didapatkan tidak sesuai dengan bidang pendidikan yang diraih	1. Normatif. Diukur dengan menggunakan metode normatif. Hasil dari pengukuran akan menjadi variabel dummy dengan nilai 1 jika bidang pendidikan yang diraih tidak sesuai dengan sektor lapangan usaha pekerjaan yang didapatkan. Jika tidak mencapai kondisi tersebut, maka akan diberi nilai 0 2. Mode (modal procedure). Diukur berdasarkan nilai modus sektor dari kelompok individu pada suatu bidang ilmu. Nilai modus sektor yang dihasilkan akan dijadikan sebagai dasar untuk menentukan individu mengalami horizontal *mismatch* ataukah tidak	Nominal
DV.4	YUNDER	Tahun kurang pendidikan (defisit pendidikan)	YUNDER = YREQ – EDUC jika EDUC < YREQ. Untuk menentukan YREQ ini, penelitian ini akan menggunakan metode normatif, VV, dan mode.	Diskret
DV.5	YOVER	Tahun lebih pendidikan (surplus pendidikan)	YOVER = EDUC – YREQ jika EDUC > YREQ. Untuk menentukan YREQ ini, penelitian ini akan menggunakan metode normatif, VV, dan mode.	Diskret
DV.6	YREQ	Tahun pendidikan yang diperlukan untuk mendapatkan atau menjalani suatu pekerjaan	YREQ dalam penelitian ini akan diestimasi dengan metode normatif, VV, dan mode.	Diskret
DV.7	Pendapatan Individu (Ln_Wage)	Total pendapatan dari gaji bersih yang didapatkan dalam 1 bulan	Nilai logaritma dari total pendapatan gaji bulanan	Rasio
DV.8	Tingkat Pengangguran (UNEM)	Persentase jumlah pengangguran terhadap jumlah angkatan kerja	Tingkat Pengangguran Terbuka	Rasio
DV.9	Pertumbuhan Ekonomi	Real GDP growth	Persentase perubahan real GDP	Rasio

Kode	Variabel	Definisi	Pengukuran	Jenis Data
	(GROWTH_GDP)			
DV.10	Produktivitas Tenaga Kerja (LP)	Produktivitas tenaga kerja merupakan ukuran dari total volume output (diukur dengan pendapatan) yang dihasilkan oleh per-unit tenaga kerja (diukur dari jumlah orang yang bekerja atau jam kerja) selama periode waktu tertentu	Perbandingan antara total output dengan total jam kerja tenaga kerja	Rasio
IV.1	Karakteristik Individu (Personal Characteristics)			
IV.1.1	Kelompok Usia 1 (Age 1)	Kelompok usia responden dari 15 – 24 tahun (Kelompok Usia Muda)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 15-24 tahun, 1 = responden usia 15 – 24 tahun	Nominal
IV.1.2	Kelompok Usia 2 (Age 2)	Kelompok usia responden dari 25 – 34 tahun (Kelompok Pekerja Muda)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 25-34 tahun, 1 = responden usia 25-34tahun	Nominal
IV.1.3	Kelompok Usia 3 (Age 3)	Kelompok usia responden dari 35 – 44 tahun (Kelompok Paruh Baya)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 34-44 tahun, 1 = responden usia 34-44 tahun	Nominal
IV.1.4	Kelompok Usia 4 (Age 4)	Kelompok usia responden dari 45 – 54 tahun (Kelompok Pra Pensiun)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 45-54 tahun, 1 = responden usia 45-54tahun	Nominal
IV.1.5	Kelompok Usia 5 (Age 5)	Kelompok usia responden dari 55 – 64 tahun (Kelompok Pensiun)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 55-64 tahun, 1 = responden usia 55-64 tahun	Nominal
IV.1.6	Kelompok Usia 6 (Age 6)	Kelompok usia responden dari 65 tahun ke atas (Kelompok Usia Lanjut)	Diukur sebagai dummy variabel dengan nilai 0 = responden usia di luar 65 tahun ke atas, 1 = responden usia 65 tahun ke atas	Nominal
IV.1.7	Kelompok Pengalaman 1 (Experience 1)	Responden dengan pengalaman kerja 1 hingga 5 tahun. Adapun pengalaman ini didapatkan dari total tahun masa kerja dalam pekerjaan saat ini (tahun 2022 di kurangi tahun masuk pekerjaan saat ini)	Diukur sebagai dummy variabel dengan nilai 0 = bukan termasuk kelompok responden dengan pengalaman kerja antara 1 hingga 5 tahun, 1 = responden dengan pengalaman kerja antara 1 hingga 5 tahun	Nominal
IV.1.8	Kelompok Pengalaman 2 (Experience 2)	Responden dengan pengalaman kerja 6 hingga 10 tahun	Diukur sebagai dummy variabel dengan nilai 0 = bukan termasuk kelompok responden dengan pengalaman kerja antara 6 hingga 10 tahun, 1 = responden dengan pengalaman kerja antara 6 hingga 10 tahun	Nominal
IV.1.9	Kelompok Pengalaman 3	Responden dengan	Diukur sebagai dummy variabel	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
	(Experience 3)	pengalaman kerja 11 hingga 20 tahun	dengan nilai 0 = bukan termasuk kelompok responden dengan pengalaman kerja antara 11 hingga 20 tahun, 1 = responden dengan pengalaman kerja antara 11 hingga 20 tahun	
IV.1.10	Kelompok Pengalaman 4 (Experience 4)	Responden dengan pengalaman kerja di atas 21 tahun	Diukur sebagai dummy variabel dengan nilai 0 = bukan termasuk kelompok responden dengan pengalaman kerja lebih dari 21 tahun, 1 = responden dengan pengalaman kerja lebih dari 21 tahun	Nominal
IV.1.11	Jenis Kelamin (Female)	Jenis kelamin responden yang menggunakan laki-laki (male) sebagai basis	Diukur sebagai dummy variabel dengan nilai 0 = Laki-Laki, 1 = Perempuan	Nominal
IV.1.12	Pelatihan (Training)	Keikutsertaan individu dalam pelatihan	Diukur secara dummy dengan nilai 0 = Tidak pernah mengikuti pelatihan, dan 1= pernah mengikuti pelatihan atau kursus tertentu	Nominal
IV.1.13	Pekerjaan Sampingan (Side_Job)	Individu yang memiliki pekerjaan lebih dari 1	Diukur secara dummy dengan nilai 0 = Tidak memiliki pekerjaan sampingan, 1 = Memiliki pekerjaan sampingan	Nominal
IV.1.14	Pekerja Paruh Waktu (Half_Employment)	Individu yang setengah pengangguran (jam kerja kurang dari 35 jam per-minggu)	Diukur secara dummy dengan nilai 0 = Jam kerja lebih dari 35 jam per-minggu, 1 = Jam kerja kurang dari 35 jam per-minggu	Nominal
IV.2	Karakteristik Institusi Pendidikan (*Education Institutions Characteristics*)			
IV.2.1	Institusi Negeri (Public_Educ)	Institusi pendidikan negeri	Diukur secara dummy dengan nilai 0 = institusi pendidikan di luar negeri, 1 = institusi pendidikan negeri	Nominal
IV.2.2	Institusi Swasta (Private_Educ)	Institusi pendidikan swasta	Diukur secara dummy dengan nilai 0 = institusi pendidikan negeri, kedinasan, dan lainnya, 1 = institusi pendidikan swasta	Nominal
IV.2.3	Institusi Kedinasan (Agency_Educ)	Institusi pendidikan Kedinasan	Diukur secara dummy dengan nilai 0 = institusi pendidikan di luar kedinasan, 1 = institusi pendidikan kedinasan	Nominal
IV.2.4	Institusi Lainnya (Other_Educ)	Institusi pendidikan lainnya	Diukur secara dummy dengan nilai 0 = institusi pendidikan negeri, swasta, dan kedinasan, 1 = institusi pendidikan lainnya	Nominal
IV.2.5	Pemagangan (Internship)	Institusi pendidikan yang mewajibkan peserta didiknya mengikuti pemagangan	Diukur secara dummy dengan nilai 0 = tidak pernah mengikuti magang, 1 = jika pernah mengikuti magang	Nominal
	Rumpun Ilmu (2 Digit pertama ISCED 11) Berisi 39 Bidang Ilmu/Jurusan			
IV.2.6	Bidang Ilmu ISCED 2 Digit yang berisi bidang ilmu yang termasuk ke dalam rumpun Pendidikan, Seni dan Humaniora, Ilmu Sosial, Jurnalistik dan	Individu yang kualifikasi pendidikannya berada pada rumpun ilmu yang sesuai dengan kode 2 digit ISCED	Diukur secara dummy dengan nilai 0 = bukan merupakan bidang/jurusan rumpun ilmunya, 1 = jika bidang keahlian atau pendidikannya merupakan bidang/jurusan rumpun ilmu sesuai ISCED 11	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
	Komunikasi, Ilmu Bisnis, administrasi, dan hukum, ilmu alam, matematika, dan statistik, ilmu informasi dan komunikasi, hingga rumpun ilmu jasa			
IV.3	Karakteristik Pekerjaan dan Pemberi Kerja/Tempat Kerja (*Jobs and Workplace Characteristics*)			
	Status Pekerjaan (Job Status)			
IV.3.1	Berusaha Sendiri (*Self employed*)	Individu dengan status pekerjaan berusaha sendiri (wiraswasta mandiri)	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam kategori berusaha sendiri, 1 = jika termasuk ke dalam kategori berusaha sendiri	Nominal
IV.3.2	Berusaha dibantu pekerja tidak tetap/pekerja keluarga/tidak dibayar (*Doing business assisted by temporary workers/family workers/unpaid*)	Individu yang berusaha atau berbisnis dengan dibantu oleh karyawan keluarga yang tidak dibayar	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam kategori berusaha dengan dibantu pekerja tidak tetap atau tidak dibayar, 1 = jika termasuk ke dalam kategori berusaha dengan dibantu pekerja tidak tetap atau tidak dibayar	Nominal
IV.3.3	Berusaha dibantu pekerja tetap dan dibayar (*Doing business assisted by regular and paid workers*)	Individu yang berusaha atau berbisnis dengan mempekerjakan pekerja tetap yang dibayar secara rutin	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam kategori berusaha dengan dibantu pekerja tetap dan dibayar, 1 = jika termasuk ke dalam kategori berusaha dengan dibantu pekerja tetap dan dibayar	Nominal
IV.3.4	Buruh/karyawan/pegawai (*Laborer/employees*)	Individu yang bekerja sebagai buruh atau karyawan swasta atau pegawai negeri	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam kategori buruh/karyawan/pegawai, 1 = jika termasuk ke dalam kategori kategori buruh/karyawan/pegawai	Nominal
IV.3.5	Pekerja bebas pertanian (*Agricultural free labour*)	Status pekerjaan non formal sebagai pekerja lepas sektor pertanian, kehutanan, maupun perikanan	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam pekerja bebas pertanian, 1 = jika termasuk ke dalam kategori pekerja bebas bidang pertanian	Nominal
IV.3.6	Pekerja bebas non pertanian (*Non agricultural free labour*)	Status pekerjaan non formal sebagai pekerja lepas sektor lain selain pertanian, kehutanan, maupun perikanan	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam pekerja bebas non pertanian, 1 = jika termasuk ke dalam kategori pekerja bebas bidang non pertanian	Nominal
IV.3.7	Pekerja keluarga/ tidak dibayar (*Family employee/unpaid worker*)	Individu yang status pekerjaannya membantu bisnis atau usaha keluarganya dengan tidak dibayar	Diukur secara dummy dengan nilai 0 = jika bukan termasuk ke dalam kategori pekerja keluarga/tidak dibayar, 1 = jika termasuk ke dalam kategori pekerja keluarga/tidak dibayar	Nominal
	Kebijakan Ketenagakerjaan (Employment Policy)			
IV.3.8	Jaminan Kesehatan (*Health_insurance*)	Perusahaan/pemberi kerja memberikan jaminan kesehatan kepada karyawan	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak memberikan jaminan kesehatan, 1 = jika perusahaan/pemberi kerja memberikan jaminan kesehatan	Nominal
IV.3.9	Program Pensiun (*pension_benefit*)	Perusahaan/pemberi kerja memberikan program pensiun	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak memberikan program pensiun, 1 = jika	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
		kepada karyawan	perusahaan/pemberi kerja memberikan program pensiun	
IV.3.10	Kebijakan Cuti (*Leave_entitlements*)	Perusahaan/pemberi kerja memberikan hak cuti tanpa memotong gaji karyawan	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak memberikan hak cuti tanpa memotong gaji, 1 = jika perusahaan/pemberi kerja memberikan hak cuti tanpa memotong gaji	Nominal
IV.3.11	Kontrak Kerja Karyawan (*Employment Contract*)	Status karyawan yang memiliki kontrak kerja/perjanjian kerja	Diukur secara dummy dengan nilai 0 = jika tidak memiliki kontrak kerja/perjanjian kerja, 1= jika memiliki kontrak kerja/perjanjian kerja	Nominal
	Jenis Institusi Pemberi Kerja (*Type of Employeer/Workplace Institution*)			
IV.3.12	Institusi Negeri (*Public Institution*)	Individu yang bekerja di institusi negeri/pemerintah	Diukur secara dummy dengan 0 = bukan merupakan institusi negeri/pemerintah, 1 = bekerja di bawah institusi negeri/pemerintah	Nominal
IV.3.13	Institusi Swasta (*Private institution*)	Individu yang bekerja di institusi swasta	Diukur secara dummy dengan 0 = bukan merupakan institusi swasta, 1 = bekerja di bawah institusi swasta	Nominal
IV.3.14	Pekerja Mandiri/Wiraswasta (*Self employed*)	Individu yang bekerja secara mandiri (wiraswasta mandiri)	Diukur secara dummy dengan 0 = bukan merupakan pekerja mandiri, 1 = bekerja secara mandiri / wiraswasta mandiri	Nominal
IV.3.15	Bisnis Rumah Tangga/Keluarga (*Household business*)	Individu yang bekerja di institusi negeri/pemerintah	Diukur secara dummy dengan 0 = bukan merupakan bisnis rumah tangga/keluarga, 1 = bekerja di bisnis rumah tangga/keluarga	Nominal
IV.3.16	Institusi Lain-Lain (*Others institution*)	Individu yang bekerja di institusi lainnya (organisasi nirlaba, internasional, dan lain-lain)	Diukur secara dummy dengan 0 = bukan merupakan institusi lainnya (organisasi nirlaba internasional, dan lain-lain), 1 = bekerja di bawah institusi institusi lainnya (organisasi nirlaba internasional, dan lain-lain),	Nominal
IV.3.17	Institusi tidak dikenal/tidak teridentifikasi (*Unidentified Institution*)	Individu yang bekerja di institusi yang bukan merupakan kategori	Diukur secara dummy dengan 0 = bukan merupakan 5 kategori institusi sebelumnya, 1 = bekerja di bawah institusi yang tidak teridentifikasi ke dalam 5 kategori institusi sebelumnya	Nominal
	Jenis transportasi ke tempat kerja (*type of transportation to work*)			
IV.3.18	Kendaraan Pribadi (*Private vehicle*)	Individu yang menggunakan kendaraan pribadi ke tempat kerja	Diukur secara dummy dengan 0 = jika individu menggunakan transportasi lain selain kendaraan pribadi ke tempat kerja, 1 = jika individu menggunakan kendaraan pribadi ke tempat kerja (motor atau mobil)	Nominal
IV.3.19	Kendaraan Umum (*Mass transportation*)	Individu yang menggunakan moda transportasi masal ke tempat kerja	Diukur secara dummy dengan 0 = jika individu menggunakan transportasi lain selain kendaraan umum ke tempat kerja, 1 = jika individu menggunakan kendaraan umum ke tempat kerja	Nominal
IV.3.20	Angkutan online (*Online transportation*)	Individu yang menggunakan angkutan online (ojek online) ke tempat kerja	Diukur secara dummy dengan 0 = jika individu menggunakan transportasi lain selain angkutan online ke tempat kerja, 1 = jika individu menggunakan	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
			angkutan onlie ke tempat kerja (motor atau mobil)	
IV.3.21	Tidak menggunakan transportasi ke tempat kerja (*Walk to work*)	Individu yang ke tempat kerja dengan berjalan kaki dari rumah	Diukur secara dummy dengan 0 = jika individu menggunakan alat transportasi, 1 = jika individu tidak menggunakan alat transportasi ke tempat kerja (berjalan kaki_	Nominal
	Sistem Pembayaran Gaji (*Salaries payment system*)			
IV.3.22	Bulanan (*Montly*)	Individu yang dibayar gajinya secara bulanan	Diukur secara dummy dengan 0 = jika individu dibayar bukan bulanan, 1 = jika individu dibayar/digaji secara bulanan rutin	Nominal
IV.3.23	Mingguan (*Weekly*)	Individu yang dibayar gajinya secara mingguan	Diukur secara dummy dengan 0 = jika individu dibayar bukan mingguan, 1 = jika individu dibayar/digaji secara mingguan	Nominal
IV.3.24	Harian (*Daily*)	Individu yang dibayar gajinya secara harian	Diukur secara dummy dengan 0 = jika individu dibayar bukan harian, 1 = jika individu dibayar/digaji secara harian	Nominal
IV.3.25	Per-Jam (*Hourly*)	Individu yang dibayar gajinya secara per-jam kerja	Diukur secara dummy dengan 0 = jika individu dibayar bukan per-jam kerja, 1 = jika individu dibayar/digaji secara per-jam kerja	Nominal
IV.3.26	Borongan (*Whosale*)	Individu yang dibayar gajinya secara borongan	Diukur secara dummy dengan 0 = jika individu dibayar bukan secara borongan, 1 = jika individu dibayar/digaji secara borongan	Nominal
IV.3.27	Per-hasil kerja (*Per-Unit*)	Individu yang dibayar gajinya per-hasil unit kerja	Diukur secara dummy dengan 0 = jika individu dibayar bukan per-unit/hasil kerja, 1 = jika individu dibayar/digaji secara per-unit/hasil kerja	Nominal
IV.3.28	Komisi (*Commision*)	Individu yang dibayar gajinya dalam bentuk komisi	Diukur secara dummy dengan 0 = jika individu dibayar bukan dalam bentuk komisi, 1 = jika individu dibayar/digaji dalam bentuk komisi	Nominal
	Sistem Pembukuan Bisnis (*Business Accounting System*)			
IV.3.29	Tanpa Pembukuan (*No accounting record*)	Perusahaan/pemberi kerja tidak melakukan pembukuan keuangan usahanya (usaha perorangan kecil/mikro)	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja melakukan pembukuan usaha, 1 = jika perusahaan/pemberi kerja melakukan pembukuan usaha	Nominal
IV.3.30	Pembukuan Sederhana (*Simple Accounting*)	Perusahaan/pemberi kerja melakukan pembukuan keuangan sederhana (usaha EMKM/ETAP)	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak melakukan pembukuan usaha atau melakukan pembukuan lengkap, 1 = jika perusahaan/pemberi kerja melakukan pembukuan usaha secara sederhana (SAK EMKM/ETAP)	Nominal
IV.3.31	Pembukuan Lengkap (*Complete Accounting System*)	Perusahaan/pemberi kerja melakukan pembukuan keuangan lengkap (perusahaan besar)	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak melakukan pembukuan usaha atau melakukan pembukuan sederhana, 1 = jika perusahaan/pemberi kerja melakukan pembukuan usaha lengkap	Nominal

Kode	Variabel	Definisi	Pengukuran	Jenis Data
			(PSAK)	
IV.3.32	Pembukuan tidak jelas (*unclear accounting*)	Perusahaan/pemberi kerja tidak memiliki pembukuan yang jelas	Diukur secara dummy dengan nilai 0 = jika perusahaan/pemberi kerja tidak melakukan pembukuan usaha, melakukan pembukuan sederhana, atau melakukan pembukuan lengkap 1 = jika perusahaan/pemberi kerja melakukan pembukuan dengan sistem yang tidak jelas	Nominal
	Karakteristik Spasial (Spatial Characteristics)			
IV.3.33	Perkotaan (Urban)	Responden yang berada di wilayah perkotaan	Diukur secara dummy dengan nilai 0 = jika responden berada atau bermukim di wilayah pedesaan, 1 = jika responden berada atau bermukim di wilayah perkotaan	Nominal
IV.3.34	Pedesaan (Rural)	Responden yang berada di wilayah perkotaan	Diukur secara dummy dengan nilai 0 = jika responden berada atau bermukim di wilayah perkotaan, 1 = jika responden berada atau bermukim di wilayah pedesaan	Nominal
IV.3.35	Migrasi (Migration)	Responden yang pernah melakukan migrasi	Diukur secara dummy dengan nilai 0 = jika responden tidak pernah berpindah tempat tinggal ke luar kabupaten/kota/provinsi, 1 = jika responden pernah berpindah tempat tinggal ke luar kabupaten/kota/provinsi	Nominal
IV.3.36	Mobilitas Spasial (Spatial_Mobility)	Responden bekerja di luar wilayah kabupaten/kota/provinsi tempat tinggalnya serta melakukan aktivitas pulang pergi	Diukur secara dummy dengan nilai 0 = jika responden bekerja di wilayah kabupaten/kota/provinsi yang sama dengan tempat tinggalnya, 1 = jika responden bekerja di luar wilayah kabupaten/kota/provinsi tempat tinggalnya dan melakukan pulang pergi (komuting atau menggunakan kendaraan pribadi)	Nominal
IV.3.37	Pengalaman Kerja di Luar Negeri (Migrant Experience)	Responden pernah bekerja (memiliki pengalaman bekerja) di luar negeri	Diukur secara dummy dengan nilai 0 = jika responden tidak pernah bekerja di luar negeri, 1 = jika responden pernah bekerja (memiliki pengalaman bekerja) di luar negeri	Nominal

3.4.9. Pengecekan Kekebalan Model (Robustness Checks)

Seluruh model dalam penelitian ini dicek kekebalannya. Proses pengecekan kekebalan model tersebut dilakukan dengan beberapa cara, diantarannya adalah dengan melakukan estimasi ulang model dengan estimator atau metode yang lain. Selain itu, cara yang digunakan dalam mengecek kekebalan model adalah dengan mengganti metode pengukuran variabel tertentu, atau dengan mengekslusikan sejumlah sampel sesuai dengan kriteria yang diperlukan.

Metode pengecekan kekebalan model untuk setiap modelnya cenderung berbeda-beda. Dalam mengecek kekebalan model 1, 2, dan 3 misalnya, penelitian ini mengestimasi ulang model-model tersebut dengan menggunakan estimator Probit. Kemudian, untuk mengecek kekebalan model 4, penelitian ini mengekslusikan sampel penelitian yang tidak bekerja serta sampel penelitian yang tingkat pendidikannya di bawah SMA/sederajat dengan Heckit. Untuk model 4 juga, penelitian ini mengestimasi ulang model dengan estimator OLS. Apabila asumsi heteroskedaktisitas dalam asumsi OLS tersebut tidak dapat terpenuhi, maka akan diestimasi dengan metode heteroscedactic linear regression (HLR). Sedangkan untuk menguji kekebalan model 5, 6, dan 7, penelitian ini menggunakan metode System Generalized Method of Moment (Sys GMM).

3.4.9.1. Pengecekan Kekebalan Model 1, 2, dan 3

Penelitian ini mengecek kekebalan model 1 dengan mengestimasi ulang model 1 dengan metode Probit. Selain itu, sampel yang digunakan dalam pengecekan kekebalan model 1 tidak hanya menggunakan sampel tingkat pendidikan minimal SMA/sederajat, tetapi menggunakan seluruh tenaga kerja. Estimasi yang akan dihasilkan oleh metode Probit adalah koefisien (berbeda dengan Logit yang estimasinya adalah odds rasio). Jika pengaruh suatu variabel *independent* hasil estimasi Probit ini mendukung odds rasio yang dihasilkan oleh Logit, maka pengaruhnya dinyatakan robust. Dalam mengecek kekebalan model 2, penelitian ini juga menggunakan metode Probit. Sampel yang digunakannya juga sama dengan pengecekan kekebalan model 1, yaitu dengan seluruh tenaga kerja.

Selanjutnya, penelitian ini melakukan pengecekan kekebalan model 3 dengan mengekslusikan sampel penelitian yang tingkat pendidikan di bawah perguruan tinggi. Diekslusikannya kelompok sampel tersebut karena *skill* yang dimiliki lulusan SMA/Sederajat (termasuk SMK dan MAK) cenderung belum spesifik dan masih berada pada KKNI level 2. Lulusan SMA jurusan IPS misalnya, belum dapat menawarkan *skill* yang spesifik kepada calon pemberi kerja, sehingga tidak memiliki kemampuan untuk menolak atau memilih suatu pekerjaan pada suatu sektor pekerjaan yang sesuai dengan kualifikasi

pendidikannya. Kondisi ini dapat memicu terjadinya bias dalam mengestimasi determinan penentu horizontal *mismatch*.

Setelah sampel diesklusikan sehingga hanya merupakan lulusan perguruan tinggi (diploma, sarjana, magister, dan doktoral), selanjutnya akan diestimasi dengan metode Probit. Metode Probit ini digunakan karena variabel horizontal *mismatch* hanya dapat bersifat kategorik (dichotomous). Dalam metode Probit, probabilitas dari seseorang mengalami horizontal *mismatch* atau tidak ditentukan oleh asumsi standar normal cumulative distribution function (CDF), berbeda dengan Logit yang menggunakan asumsi logistic CDF. Alasan dipilihnya metode Probit untuk mengecek kekebalan model 3 ini adalah karena setelah diekslusikan, terdapat kemungkinan yang besar data dapat berdistribusi normal.

Selain dengan metode Probit, pengecekan keebalan model 3 juga akan menggunakan metode Logit namun dengan output estimasi dalam bentuk koefisien (bukan odds rasio). Apabila odds rasio yang dihasilkan dari model 3 tidak kontradiktif dengan koefisien yang dihasilkan dari Probit dan Logit, maka pengaruh variabel *independent* terhadap horizontal *mismatch* dapat dinyatakan robust. Dalam pengecekan kekebalan model 3 ini, akan dilakukan penambahan variabel *independent* yaitu sektor lapangan kerja sesuai dengan kode KBLI.

3.4.9.2. Pengecekan Kekebalan Model 4

Pengecekan kekebalan model 4 dalam penelitian ini dilakukan dengan mengesklusikkan sampel tenaga kerja yang tidak bekerja serta yang tingkat pendidikannya di bawah SMA/sederajat. Setelah diesklusikan, masih terdapat bias seleksi karena masih terdapat tenaga kerja dengan status setengah pengangguran atau half timer (paruh waktu). Oleh karenanya, sampel tersebut tetap akan diestimasi dengan menggunakan metode Heckit dengan persamaan seleksi tenaga kerja penuh waktu (full timer).

Selain itu, pengecekan kekebalan model 4 ini juga akan menggunakan analisis OLS. Analisis OLS ini tergolong merupakan estimator yang paling akurat karena memiliki status BLUE (best linear unbias estimator). Akan tetapi, terdapat asumsi ketat yang harus dipenuhi yaitu normalitas, heteroskedaktisitas, autokorelasi, dan multikoliniearitas. Dari keempat asumsi tersebut, asumsi yang cenderung sulit dipenuhi dalam persamaan pendapatan adalah heteroskedaktisitas.

Hal ini karena pendapatan tenaga kerja yang diungkapkan tenaga kerja relatif terlalu bervariasi. Apabila asumsi heteroskedaktisitas tersebut tidak dapat tercapai, maka estimator yang akan digunakan adalah HLR.

3.4.9.3. Pengecekan Kekebalan Model 5, 6, dan 7

Pengecekan kekebalan model 5, 6, dan 7 dalam penelitian ini akan mengugnakan System Generalized Method of Moment (Sys-GMM). Model-model tersebut dapat dinyatakan kebal apabila memiliki kesesuaian hasil dengan metode Sys-GMM. Dalam metode Sys-GMM, lag *dependent* variabel akan digunakan sebagai instrumen. Penelitian ini menggunakan 2 lag *dependent*, sebagaimana mengacu pada Roodman (2009) bahwa setidaknya instrumen yang digunakan menggunakan 2 lag *dependent* variabel.

Sebagai contoh, model 5a apabila diestimasi dengan metode Sys-GMM, bentuknya adalah:

$$\mathrm{UNEM_{it}} = \delta \mathrm{L1.UNEM_i} + \delta \mathrm{L2.UNEM_i} + \beta_1 \mathrm{YUNDER_{it}} + \beta_2 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \mathrm{u_{it}}$$

dimana, $\mathrm{UNEM_{it}}$ adalah tingkat pengangguran di provinsi i pada waktu t, sedangkan YUNDER adalah indeks defisit pendidikan tingkat Provinsi, dan HMM merupakan persentase horizontal *mismatch* tingkat Provinsi. Sementara itu, $\delta \mathrm{L1.UNEM_i}$ adalah skalar dari lag-1 sedangkan $\delta \mathrm{L2.UNEM_i}$ adalah skalar dari lag-2 dari UNEM atau tingkat pengangguran tahun sebelumnya. Skalar dalam hal ini maksudnya adalah besaran nilai variabel yang digunakan sebagai instrumen (mengapa menggunakan skalar, karena untuk membedakannya dengan vektor/variabel *independent*/variabel eksplanatori). Kedua lag *dependent* ini akan dipergunakan sebagai instrumen untuk mengantisipasi persoalan endogeneitas. Adapun Xit merupakan variabel-variabel untuk mengontrol heterogenitas (dalam hal ini adalah LogDI, LogFDI, GINI, dan INLFASI), dan uit adalah μi + vit yang mengikuti one-way error component model

Sementara itu, contoh bentuk model 6a dalam metode Sys-GMM adalah sebagaimana berikut:

$$\mathrm{Growth_{it}} = \delta \mathrm{L1.Growth_i} + \delta \mathrm{L2.Growth_i} + \beta_1 \mathrm{YUNDER_{it}} + \beta_3 \mathrm{HMM_{it}} + \beta_j \mathrm{X_{itj}} + \mathrm{u_{it}}$$

dimana, $\mathrm{Growth_{it}}$ adalah pertumbuhan ekonomi Provinsi i pada waktu t yang diukur dengan real GDP growth, sedangkan YUNDER adalah rata-rata defisit

pendidikan, dan HMM adalah persentase horizontal *mismatch*. δL1.Growth$_i$ dan δL2.Growth$_i$ adalah skalar dari lag-1 dan lag2 dari Growth atau tingkat pertumbuhan ekonomi tahun sebelumnya. Lag 1 dan 2 ini akan dipergunakan sebagai instrumen untuk mengantisipasi persoalan endogeneitas. Dalam model pertumbuhan ekonomi ini, akan ditambahkan dengan Lag 1 dan 2 dari masing-masing variabel *independent* sebagai tambahan instrumen. Adapun Xit merupakan variabel-variabel kontrol (dalam hal ini adalah LogDI, LogFDI, dan GINI), dan uit adalah μi + vit yang mengikuti one-way error component model.

Untuk memberikan keyakinan yang lebih kuat terkait dengan kekebalan model 6, penelitian ini juga mengestimasi Growth yang diukur dengan real GDP per-capita menggunakan metode Sys-GMM. Jika Growth diukur dengan real GDP per-capita, maka pengukuran Growth tersebut dilakukan dengan metode produksi. Hal ini karena data real GDP per-capita yang dirilis oleh BPS didapatkan dengan menggunakan metode produksi dari sektor lapangan usaha. Sementara itu, data real Growth GDP didapatkan dengan metode pengeluaran.

Terakhir, dalam mengecek kekebalan model 7, penelitian ini mengukur produktivitas tenaga kerja dengan menggunakan pertumbuhan GDP per-tenaga kerja. Metode pengukuran produktivitas seperti ini mengacu pada definisi dari Bank Dunia yang menyebutkan bahwa GDP per person employed (GDP per-tenaga kerja) dapat mengukur produktivitas tenaga kerja-output per unit dari input tenaga kerja. Adapun contoh bentuk dari model 7a dengan analisis sys-GMM adalah:

$$\mathrm{GPEREMPLOY_{it}} = \delta\mathrm{L1.GPEREMPLOY_i} + \delta\mathrm{L2.GPEREMPLOY_i} + \beta_1\mathrm{YUNDER_{it}} + \beta_2\mathrm{HMM_{it}} + \beta_j\mathrm{X_{itj}} + \mathrm{u_{it}}$$

dimana, GPEREMPLOY$_{it}$ adalah produktivitas tenaga kerja di Provinsi i pada waktu t yang dikur dengan pertumbuhan GDP per-tenaga kerja, sedangkan YUNDER adalah rata-rata defisit pendidikan, dan HMM adalah persentase rata-rata horizontal *mismatch* tingkat Provinsi. δL1.LP$_i$ dan δL2.LP$_i$ adalah skalar dari lag-1 dan lag-2 dari produktivitas tenaga kerja tahun sebelumnya. Lag *dependent* ini akan dipergunakan sebagai instrumen untuk mengantisipasi persoalan endogeneitas. Adapun Xit merupakan variabel-variabel kontrol (dalam hal ini

adalah LogDI, LogFDI, INFLASI, dan TPT), dan uit adalah μi + vit yang mengikuti one-way error component model

3.5. Analisis Data

Terdapat 8 metode analisis data yang digunakan dalam penelitian ini. Perbedaan metode analisis data disesuaikan dengan karakteristik data variabel terikat yang dipergunakan untuk setiap model penelitian. Metode-metode analisis data yang akan digunakan antara lain : Logit, Probit model, Heckit model, OLS, Regresi HLR yang berbasis maximum likelihood, analisis regresi data panel, analisis instrumental variabel, dan system generalized method of moment (Sys GMM). Kedelapan metode tersebut masing-masing akan digunakan untuk mengestimasi model atau untuk melakukan pengecekan kekebalan model.

3.5.1. Analisis Regresi Logistik (Logit Model)

Analisis regresi logistik dalam penelitian ini digunakan untuk mengestimasi model yang variabel terikatnya merupakan data kategorikal yaitu model 1, 2, dan 3. Terdapat beberapa bentuk regresi logistik yang banyak digunakan peneliti diantarannya yaitu logit model dan probit model. Keduanya merupakan pengembangan dari linear probability model (LPM). Dalam konteks ini, LPM memiliki sejumlah kekurangan. Beberapa kekurangan diantarannya yakni bahwa nilai probilitas yang dihasilkan dari model LPM adalah antara 0 dan 1, tetapi tidak ada jaminan estimasinya akan berada dalam batas tersebut. Hal ini karena LPM merupakan bagian dari OLS yang tidak memperhitungkan batasan estimasi probabilitas harus ada pada batas 0 dan 1. Selain itu, dalam LPM, error term LPM mengandung heteroskedaktisitas serta cenderung akan sulit terdistribusi normal karena variabel bebas hanya bernilai 0 dan 1.

Berdasarkan keterbatasan-keterbatasan pada LPM tersebut, Gujarati (2015) menyarankan penggunaan metode logit atau probit. Hosmer et al. (2013) dan Gujarati (2015) menjelaskan bahwa logit model digunakan untuk menduga hubungan antara pengaruh variabel eksplanatori terhadap variabel terikat ketika variabel terikat tersebut bersifat kategorikal. Secara spesifik, variabel terikat yang ditelaah dalam logit model adalah probabilitas (Ghozali & Ratmono, 2017). Sehingga, model logit dalam penelitian ini menelaah seberapa besar probabilitas

seseorang mengalami overeducated atau tidak dan mengalami undereducated atau tidak berdasarkan nilai dari variabel-variabel bebas. Menurut Ghozali & Ratmono (2017), salah satu keunggulan dari logit model atau yang juga disebut sebagai regresi logistik adalah bahwa tidak memerlukan asumsi distribusi normal multivariat. Dalam penelitian ini, logit model akan digunakan untuk mengestimasi model 1, 2, dan 3.

Meskipun model Logit bersifat linear, tetapi tidak dapat diestimasi dengan menggunakan OLS (Gujarati & Porter, 2013). Hal ini karena jika menggunakan OLS, maka nilai Logit akan diestimasi dengan formula Li = ln (1/0) apabila individu mengalami *overeducation*, dan Li = ln(0/1) apabila tidak mengalami *overeducation*. Formula-formula tersebut tentu tidak akan menghasilkan nilai apa-apa. Oleh karena itu, Logit model dalam penelitian ini akan diestimasi dengan metode maksimum likelihood (ML). Hal ini mengacu pada Gujarati (2015), bahwa metode yang paling populer untuk digunakan dalam mengestimasi model Logit adalah ML.

Dikarenakan data dalam penelitian ini adalah data survei yang memiliki fitur variabel pembobot (weight), maka metode Logit yang akan digunakan juga akan menggunakan format data survei. Dengan menggunakan format data survei ini, maka jumlah populasi yang diwakili oleh sampel akan dapat diketahui. Dalam format survei tersebut, maka standar error yang dihasilkan dapat dilinearisasi (linearized standar error). Selain dapat dilinearisasi, standar error yang dihasilkan juga dapat diestimasi dengan menggunakan metode bootstrap atau jacknife. Akan tetapi, dikarenakan model Logit ini bersifat linear, maka penelitian ini akan menggunakan linearized standar error.

3.5.1.1. Uji Kualitas Data

Dalam model Logit, karena pendekatan estimasinya bukan menggunakan OLS, maka tidak diperlukan uji asumsi klasik BLUE. Bahkan, jika mengacu pada Ghozali & Ratmono (2017) analisis regresi logistik atau Logit model ini umumnya digunakan jika asumsi distribusi normalitas multivariat tidak terpenuhi. Artinya, dalam model Logit, tidak diperlukan adanya uji normalitas, heteroskedaktisitas, multikoliniearitas, dan autokorelasi. Berdasarkan hal tersebut,

pengujian kualitas data dalam model Logit pada penelitian ini hanya akan berfokus pada kualitas model.

3.5.1.2. Uji Kualitas Model dan Pengujian Hipotesis

Hipotesis yang terbangun dari model 1, 2, dan 3 bersifat simultan. Oleh karena itu, pengujian hipotesis dalam penelitian ini akan dilakukan dengan menggunakan ukuran pengaruh simultan (uji F). Di samping melakukan pengujian hipotesis tersebut, penelitian ini juga akan melakukan pengujian kualitas model dengan menggunakan ukuran *goodness of fit* (GOF).

Dalam pengujian hipotesis simultan, metode OLS akan menghasilkan nilai R^2 sebagai besaran pengaruh simultan. Akan tetapi, jika mengacu pada Gujarati (2015), menggunakan R^2 pada variabel terikat dengan nilai kategorikal (logit model) akan menjadi bias atau tidak bermakna. Hal ini karena R^2 berasal dari metode berbasis OLS, sedangkan Logit dalam penelitian ini diestimasi dengan ML. Atas dasar itu, penelitian ini tidak menggunakan R^2 sebagai ukuran pengaruh simultan pada model 1, 2, dan 3.

Ukuran pengaruh simultan pada model Logit lazimnya menggunakan McFadden R^2 yang merupakan nilai R^2 dari estimasi ML. Fungsi dari nilai McFadden R^2 sama dengan R^2 pada metode OLS, yaitu untuk menjelaskan *goodness of fit* dari garis regresi. Nilai McFadden R^2 ini berkisar antara 0 sampai dengan 1. Interpretasinya adalah, semakin besar nilai McFadden R^2, maka kemampuan variabel-variabel bebas dalam menginterpretasikan variabel terikat menjadi semakin kuat. Selain nilai McFadden R^2, regresi logistik juga akan menghasilkan LR Chi-Square (likelihood ratio Chi^2) atau LR statistik. Probabilitas dari nilai LR Chi^2 inilah yang dapat digunakan untuk menguji hipotesis simultan pada model Logit. Apabila nilai probabilitasnya lebih rendah dari 0.05, maka variabel-variabel bebas dapat menjelaskan variabel terikat. Dengan kata lain, perubahan varian yang terjadi pada variabel terikat adalah karena terjadinya perubahan pada variabel bebasnya.

Mengacu pada Gujarati (2015), bahwa uji LR Chi-Square adalah untuk menguji pengaruh simultan seperti halnya uji F. Dalam format data survei dengan pembobot, nilai McFadden R^2 serta LR Chi^2 tidak dapat dihasilkan. Sebagai gantinya, regresi logistik atau Logit model dengan menggunakan format data

survei berpembobot akan menghasilkan nilai F statistik (F-hitung) dan probabilitasnya. F-statistik inilah yang akan digunakan untuk menguji hiptesis simultan model 1, 2, dan 3 dalam penelitian ini. Hipotesis nol untuk F statistik ini adalah bahwa tidak ada satupun variabel bebas yang signifikan terhadap variabel bebas. Oleh karena itu, jika nilai probabilitas dari F statistik ini lebih rendah dari 0.05, maka hipotesis nol harus ditolak dengan menerima hipotesis alternatif. Adapun hipotesis alternatifnya berarti variabel-variabel bebas berpengaruh secara simultan terhadap variabel bebas. Dengan kata lain, jika hipotesis nol ditolak, maka variabel-variabel bebas dalam model Logit merupakan determinan penting bagi variabel terikat.

Selanjutnya, terkait dengan ukuran GOF dalam model Logit, ukuran statistik yang biasanya digunakan adalah uji Hosmer and Lemeshow's (HL). Mengacu pada Ghozali & Ratmono (2017), uji HL dilakukan untuk menelaah apakah model Logit bersifat fit. Hipotesis nol untuk uji HL ini adalah bahwa tidak terdapat perbedaan antara model dengan data sehingga model memiliki *goodness of fit*. Sehingga, jika statistik HL lebih besar dari 0.05, maka hipotesis nol diterima, yang berarti model mampu memprediksi nilai observasinya (Ghozali & Ratmono, 2017).

Namun kembali lagi, dikarenakan format data yang digunakan untuk menguji model 1, 2, dan 3 ini adalah data survei dengan pembobot, maka uji HL tidak dapat dilakukan. Sekalipun jika menggunakan model Logit dengan format data sampel tanpa pembobot, nilai HL yang dihasilkan tidak akan presisi karena terlalu banyaknya jumlah sampel (Yu et al., 2017). Sebagai pengganti uji HL, penelitian ini akan melakukan pengujian GOF pada data survei dengan pendekatan uji F-adjusted mean residual sebagaimana mengacu pada Archer et al. (2007). Menurut Archer et al. (2007), uji GOF yang dinilai lebih presisi untuk menguji model regresi logistik pada data survey adalah dengan pendekatan tes statistik F-adjusted mean residual.

Hasil estimasi regresi logistik pada data survei akan menghasilkan F statistik dan setelahnya, akan menghasilkan F-adjusted mean residual. F statistik merupakan overall fit bahwa jika probabilitas F statistik lebih besar dari 0.05, maka tidak ada satupun variabel bebas yang mempengaruhi variabel terikat.

Sebaliknya, jika probabilitas F statistik lebih rendah dari 0.05, maka berarti variabel-variabel bebas berpengaruh secara simultan terhadap variabel terikat.

Adapun F-adjusted mean residual adalah untuk menentukan *goodness of fit* model. Semakin rendah F-adjusted mean residual menunjukkan model semakin memiliki *goodness of fit* yang baik. Jika probabilitas F-adjusted mean residual yang dihasilkan lebih besar dari 0.05, maka model tidak memiliki *goodness of fit* yang memadai. Sebaliknya, jika probabilitas F-adjusted mean residual tersebut lebih besar dari 0.05, maka model memiliki *goodness of fit* yang baik.

3.5.2. Probit Model

Probit (probability unit) model dalam penelitian ini akan digunakan untuk melakukan pengecekan kekebalan model 1, 2, dan 3 tentang determinan vertikal dan horizontal *mismatch* pendidikan. Pada prisipnya, Probit model ini relatif mirip dengan metode Logit, akan tetapi error term dalam model Probit ini menggunakan distribusi normal, sehingga bersifat standar normal cumulative distribution function (CDF). Maksud dari CDF di sini karena CDF merupakan fungsi distribusi kumulatif atau tabel Z (Gujarati, 2015).

Dikarenakan distribusinya adalah standar normal CDF, maka asumsi normalitas dalam metode Probit diperlukan. Selain itu, model Probit juga memerlukan asumsi homoskedaktisitas. Output estimasi dari model Probit ini adalah koefisien. Probabilitas akan semakin tinggi apabila koefisien yang dihasilkan positif. Jika koefisiennya negatif, maka menunjukkan probabilitas yang lebih rendah.

3.5.2.1. Uji Kualitas Data

Uji kualitas data untuk metode Probit dalam penelitian ini menggunakan 2 pengujian yaitu uji normalitas dan heteroskedaktisitas. Pengujian normalitas dalam model Probit ini dilakukan secara grafis dengan menggunakan histogram. Dalam histogram tersebut, axis Y yang digunakan adalah Kernel Densit, sedangkan axis X nya adalah linear prediction (prediksi linear) dari data. Metode grafis seperti ini setara dengan uji normalitas skewness and kurtosis (SK test). Tetapi, karena banyaknya jumlah sampel, maka uji SK ini tidak dapat dilakukan.

Uji kualitas data yang kedua adalah uji heteroskedaktisitas. Uji ini juga dilakukan secara grafis dengan menggunakan scatterplot. Axis Y yang digunakan

adalah residual kuadrat dari data sedangkan axis X nya adalah linear predicition. Jika data distribusinya acak (tidak berpola), maka data dapat dinyatakan tidak mengalami heteroskedaktisitas.

3.5.2.2. Uji Kualitas Model

Kualitas model dalam analisis Probit dapat ditentukan dari nilai Pseudo R^2. McFadden (2021) menjelaskan bahwa jika nilai Pseudo R^2 suatu model Probit berada pada kisaran 0.2 hingga 0.4, maka model dapat dinyatakan memiliki GOF yang memadai (good model fit). Sementara itu, jika nilainya lebih dari 0.4, maka menunjukkan excellent good fit. Selain itu, hipotesis simultan model yang menentukan GOF dalam model Probit ini dapat menggunakan uji HL (hosmer and Lemeshow). Akan tetapi, dalam format data survei berpembobot, statistik Pseudo R^2 dan uji HL tersebut tidak dapat dihasilkan.

Sebagai ganti statistik Pseudo R^2, maka akan menggunakan nilai F-statistik. Jika probabilitas F-statistik kurang dari 0.05, maka hipotesis simultan dapat diterima. Artinya, variabel-variabel ekspalatori yang dimodelkan dinyatakan berkaitan atau berpengaruh dengan variabel *dependent*-nya. Sementara itu, untuk uji GOF-nya, sebagai ganti uji HL, akan digunakan uji F-adjusted mean residual dari Archer et al. (2007). Jika probabiltias F-adjusted lebih dari 0.05, maka model memiliki GOF yang memadai.

3.5.3. Heckit Model (Heckman Selection Model)

Metode Heckit akan digunakan untuk mengestimasi model 4a, 4b, dan 4c. Metode ini digunakan karena mengestimasi pengaruh *mismatch* pendidikan terhadap pendapatan memiliki kemungkinan bias seleksi. Bias seleksi dalam konteks *mismatch* terjadi karena tidak seluruh individu yang berpendidikan telah bekerja sehingga dapat memiliki pendapatan. Besar juga terjadi kemungkinan bahwa informasi mengenai pendapatan hanya diungkapkan oleh individu yang bekerja di sektor gaji (waged sector).

Puhani (2000) mencontohkan bias seleksi dalam mengestimasi pengaruh pendidikan terhadap tingkat gaji. Bias seleksi terjadi ketika terdapat sampel yang memiliki tingkat pendidikan yang lebih lama tetapi belum bekerja karena alasan tertentu. Orang-orang dengan pendidikan lebih tinggi tentu mengharapkan upah yang lebih tinggi sehingga mereka relatif akan menunggu pekerjaan dengan gaji

yang sesuai. Sementara itu, orang-orang dengan tingkat pendidikan lebih rendah, akan mendapatkan tawaran gaji yang lebih rendah, sehingga lowongan pekerjaan untuk mereka relatif lebih luas. Akibatnya, orang-orang dengan pendidikan yang lebih rendah justru tidak banyak yang menganggur dibandingkan dengan yang pendidikannya lebih tinggi.

Persoalan bias seleksi seperti ini dapat mengaburkan hasil penelitian. Dimana, seolah-olah, pendidikan tidak mempengaruhi tingkat pengangguran dan pendapatan. Mengestimasi pengaruh *mismatch* vertikal dan horizontal terhadap pendapatan juga cenderung berpotensi mengalami bias seleksi. Ketika akan menelaah tentang pengaruh *mismatch* pendidikan terhadap gaji, maka orang yang dalam data sampel dinyatakan menganggur (tidak mendapat gaji), tidak akan disertakan ke dalam sampel. Padahal, orang yang menganggur tersebut secara teoritis bisa saja dikarenakan penawaran gaji yang diterimanya rendah, sehingga memutuskan tidak mengambil pekerjaan yang ditawarkan. Terjadinya bias seleksi juga mengakibatkan hasil estimasi berbasis regresi OLS menjadi kurang akurat (karena orang yang berpendidikan tinggi masih dalam proses mencari kerja).

Mayoritas penelitian yang menelaah pengaruh *mismatch* terhadap pendapatan menggunakan metode Heckit untuk mengantisipasi bias seleksi karena pendapatan hanya diungkapkan oleh mereka yang telah bekerja. Beberapa penelitian diantarannya yakni Romero & Jiménez (2017) dan Caroleo & Pastore (2018) menggunakan metode Heckit. Menurut Caroleo & Pastore (2018), dalam menganalisis *mismatch* pendidikan, OLS tidak dapat mengontrol kemungkinan adanya perbedaan aspek yang tidak terobservasi (*unobserved heterogeneity*) antara *mismatch* dengan sampel yang tidak sedang bekerja (menganggur). Sampel yang tidak sedang bekerja tersebut mungkin juga dapat mengalami *mismatch* jika memasuki kerja.

Dalam penelitian ini, masalah bias seleksi tersebut juga terjadi karena tidak seluruh sampel telah berstatus bekerja. Mengacu pada data Sakernas 2022, total sampel yang tersedia adalah 752,688 dengan total populasi sebanyak 209,420,383 orang yang merupakan penduduk usia kerja (15 tahun ke atas). Dari jumlah tersebut, hanya sebanyak 135.2 juta yang statusnya adalah penduduk bekerja. Rinciannya adalah, sebanyak 125 juta bekerja minimal 1 jam per-minggu,

1.4 juta melakukan kegiatan untuk memperoleh penghasilan, 5.3 juta membantu kegiatan usaha/pekerjaan orang lain, dan 2.5 juta bekerja namun sedang tidak aktif.

Dari jumlah-jumlah tersebut, jelas terlihat bahwa terdapat cukup banyak penduduk yang statusnya bukan sebagai penduduk bekerja. Baik itu karena memang sedang menganggur, ataupun karena bukan angkatan kerja (misalnya seperti ibu rumah tangga atau masih sedang bersekolah). Maka dari itu, pengaruh *mismatch* vertikal dan horizontal dalam penelitian ini cenderung sulit apabila diestimasi oleh OLS. Dengan demikian, penelitian ini juga akan menggunakan metode Heckman selection model (Heckit Model) dalam mengantisipasi bias seleksi.

Metode Heckit dikembangkan oleh Heckman (1979). Heckit model merupakan analisis regresi yang mengestimasi sampel terseleksi (censored sample). Dalam Heckit model, bias seleksi diperlakukan sebagai ommited variabel. Dimana, menurut Gujarati (2015), Heckit model mengestimasi 2 persamaan. Persamaan pertama berkaitan dengan persamaan seleksi (selection equation), sedangkan persamaan kedua adalah persamaan hasil estimasi regresinya. Persamaan pertama diestimasi untuk menentukan terdapat bukti empiris bahwa variabel-variabel eksplanatori mempengaruhi kemungkinan seseorang bekerja. Sedangkan persamaan kedua adalah persamaan untuk menentukan bagaimana pengaruh *mismatch* vertikal dan horizontal pendidikan terhadap pendapatan setelah dilakukan seleksi atas sampel.

Persamaan pertama yang berkaitan dengan persamaan seleksi yang berkaitan kemungkinan bagi seorang individu bekerja. Variabel *dependent* untuk persamaan seleksi ini adalah bekerja yang nilainya kategorikal (0 = tidak bekerja, dan 1 = bekerja). Sementara itu, variabel eksplanatori yang digunakan untuk menentukan apakah seseorang itu bekerja ataukah tidak adalah gender perempuan, pedesaan, pendidikan, status masih sekolah, dan migrasi. Persamaan pertama ini akan diestimasi dengan menggunakan metode Probit.

Dalam metode Heckit, terdapat 2 pendekatan yang dapat digunakan yaitu pendekatan maximum likelihood (ML) dan two step consistent estimator. Penelitian ini memilih untuk menggunakan pendekatan ML karena dapat

menggunakan cluster robust standar error tingkat provinsi, sedangkan dalam pendekatan two step tidak dapat dilakukan. Selain itu, penggunaan pendekatan ML dalam metode Heckit ini juga karena penelitian ini menggunakan pembobot dalam frekuensi sampelnya. Penggunaan pembobot dalam Heckit model hanya dapat dilakukan apabila menggunakan estimator ML.

Selanjutnya, untuk model persamaan kedua (main equation) dalam metode Heckit, penelitian ini memposisikan LnIncome sebagai variabel *dependent* yang dipengaruhi oleh *mismatch* pendidikan sebagaimana yang dirumuskan dalam model 4a, 4b, dan 4c. Penggunaan metode Heckit seperti ini, apabila mengacu pada Cutillo & Pietro (2006), dapat mengantisipasi bias seleksi sekaligus bias endogenitas. Dimana, potensi masalah bias endogenitas tersebut adalah jika vertikal *mismatch* dan horizontal berkorelasi dengan error term.

3.5.3.1. Uji Kualitas Data

Meskipun penelitian ini menggunakan pendekatan ML dalam metode Heckit, tetapi tetap terdapat 2 persamaan yang dihasilkan. Persamaan pertama adalah tentang persamaan seleksi yang diestimasi dengan metode Probit. Dikarenakan metode Probit digunakan, maka asumsi normalitas dan hetroskedaktisitas diperlukan untuk menilai kualitas distribusi data dan model yang dibangun. Dalam penelitian ini, uji normalitas persamaan seleksi dilakukan secara grafis dengan membuat histogram distribusi prediksi linear sebagai axis X dan nilai kernel density distribusi normal sebagai variabel Y-nya. Apabila distribusi data prediksi linear tersebut berada di dalam garis lonceng (garis regresi), maka data dapat dinyatakan berdistribusi normal.

Adapun untuk menguji heteroskedaktisitasnya, penelitian ini menggunakan pendekatan grafis dengan scatter plot distribusi probabilitas data model seleksi sebagai axis X dan residual^2 sebagai axis Y-nya. Jika distribusinya berpola (tidak acak), maka terdapat masalah heteroskedaktisitas dalam persamaan seleksi-nya. Namun, apabila asumsi normalitas dan heteroskedaktisitas dalam persamaan seleksi terlanggar sekalipun, bukan menjadi perhatian utama dalam penelitian ini. Sebab, perhatian utamanya adalah untuk menelaah kualitas data persamaan utama (main equation) Heckit model. Persamaan utama ini adalah terkait dengan pengaruh *mismatch* pendidikan terhadap pendapatan.

Suatu persamaan utama model Heckit dinyatakan berkualitas apabila mencapai asumsi normalitas. Hal ini karena model Heckit itu sendiri mengasumsikan distribusi standar error yang normal. Oleh karenanya, persamaan utama dalam Heckit model ini hanya akan diuji normalitasnya. Metode pengujian normalitasnya dilakukan secara grafis dengan menggunakan histrogram nilai prediksi liniear hasil estimasi Heckit model.

3.5.3.2. Uji Kualitas Model dan Pengujian Hipotesis

Uji kualitas model persamaan seleksi dilakukan dengan meninjau nilai statistik Pseudo R^2 yang dihasilkan. Jika angkanya lebih dari 0.2, maka model persamaan seleksi dapat dinyatakan memiliki fit yang memadai. Sementara itu, untuk menguji model fit dari persamaan utama Heckit model (atau juga disebut Heckman correction), penelitian ini menggunakan uji Wald Chi^2. Semakin tinggi nilai Wald Chi^2 , maka kemampuan prediksi model (prediction power) semakin tinggi. Kemudian, jika nilai probabilitas dari Wald Chi^2 tersebut kurang dari 0.05, maka model dapat dinyatakan fit (GOF memadai).

Adapun untuk menguji hipotesis parsial, penelitian ini menggunakan uji t dengan meninjau nilai signifikansi dari setiap koefisien regresi yang dihasilkan. Batas nilai alpha yang digunakan adalah 0.05, sehingga jika nilai signifikansi (probabilitas) dari koefisien kurang dari 0.05, maka dapat dinyatakan bahwa variabel eksplanatori memiliki pengaruh yang signifikan terhadap variabel *dependent*.

3.5.4. Ordinary Least Square (OLS) Estimator

Estimator OLS dalam penelitian ini akan digunakan untuk menganalisis model 4d. Hal ini karena model 4d jumlah sampelnya relatif yang paling sedikit, sehingga kemungkinan data dapat berdistribusi normal dan terbebas dari masalah heteroskedaktisitas. Selain itu, OLS ini juga digunakan karena model 4d sudah tidak lagi mengalami masalah bias seleksi. Sebab, sampel yang digunakan untuk model 4b sudah mengekslusikan sampel dengan status tidak bekerja. Selain itu, variabel *dependent* yang akan diuji dalam model 4d ini adalah logaritma natural dari pendapatan per-jam (bukan lagi pendapatan per-bulan seperti dalam model 4a hingga 4c). Oleh karenanya, heterogenitas pendapatan yang disebabkan oleh jumlah jam kerja sudah tidak lagi terjadi.

Estimator OLS adalah estimator yang dapat menghasilkan koefisien yang *best linear unbias estimator* (BLUE). Artinya, koefisien yang dihasilkan oleh OLS akan menjadi yang terbaik, paling akurat, dan paling tidak bias. Mengacu pada Daniels & Minot (2020), maksud dari 'best' adalah varians error term nya paling rendah. Maksud dari linear adalah bahwa *dependent* variabel adalah fungsi linear dari *independent* variabel, sedangkan 'unbias' maksudnya adalah koefisien yang dihasilkan tidak akan lebih besar atau lebih kecil daripada koefisien yang sesungguhnya.

Meskipun estimator OLS merupakan yang BLUE, tapi cenderung ketat atas asumsi. Kondisi yang diperlukan untuk mencapai hasil OLS yang BLUE jika mengacu kepada Daniels & Minot (2020) adalah bahwa pengukuran *independent* variabel tanpa error, persamaan regresi disusun dengan benar (tidak terdapat omitted variabel), tidak terjadi multikoliniearitas, varians error konstan (homoscedaktis), error tidak berkorelasi satu sama lain (autokorelasi), dan tidak tidak terdapat endogenitas. Asumsi-asumsi inilah yang disebut sebagai asumsi klasik dalam OLS. Satu lagi asumsi yang perlu terjadi pada OLS, meskipun bukan syarat BLUE adalah bahwa error term harus berdistribusi normal (normalitas).

3.5.4.1. Uji Kualitas Data

Uji kualitas data dalam analisis OLS difokuskan pada pengujian asumsi klasik. Asumsi klasik yang akan digunakan adalah normalitas, multikolinearitas, dan heteroskedaktisitas. Uji autokorelasi tidak diperlukan karena data bersifat cross sectional. Autokorelasi atau yang juga sering disebt sebagai korelasi serial hanya terjadi pada data time series atau yang mengandung time series. Sementara itu, penelitian ini mengasumsikan bahwa model 4d tidak mengandung masalah endogenitas karena variabel-variabel yang diuji bukanlah variabel-variabel indikator ekonomi makro yang relatif sulit dibedakan mana variabel *independent* dan *dependent*nya.

Dalam menguji normalitas, penelitian ini akan menggunakan histogram dengan axis Y adalah normal k density sedagkan axis X nya adalah residual dari hasil regresi OLS. Jika data residual berada di dalam garis normal density, maka error term data dapat dijustifikasi berdistribusi normal. Penggunaan grafis dalam menguji normalitas dalam penelitian ini disebabkan banyaknya jumlah sampel,

sehingga metode statistik seperti Jarque Berra, Shapiro Wilk, maupun yang lainnya menjadi tidak efektif.

Untuk mendeteksi ada tidaknya multikolinearitas, penelitian ini mengkorelasikan seluruh variabel *independent* yang akan digunakan dalam model 4d. Jika terdapat variabel *independent* yang berkorelasi satu sama lain sebesar 0.8, maka salah satu variabel *independent* tersebut tidak akan disertakan ke dalam model regresi. Hal ini mengacu pada Ghozali (2018), bahwa batas toleransi korelasi antar variabel *independent* untuk suatu model regresi adalah 0.8.

Untuk uji heteroskedaktisitas, penelitian ini menggunakan scatterplot dengan axis Y adalah residual dan axis X nya adalah fitted value atau linear prediction dari hasil analisis regresi. Jika data fitted value tersebut tersebar secara konstans sehingga terlihat acak (tidak berpola), maka tidak terjadi heteroskedaktisitas. Dalam konteks ini, heteroskedaktisitas merupakan kondisi dimana varians (sebaran) dari error term tidak konstan. Sederhananya adalah, jika sebaran kesalahan (error) dari setiap sampel observasi tidak sama, maka telah terjadi masalah heteroskedaktisitas.

3.5.4.2. Uji Kualitas Model dan Uji Hipotesis

Kualitas model regresi OLS dalam penelitian ini diuji dengan menggunakan uji F. Uji F ini juga dapat digunakan untuk menguji hipotesis simultan. Apabila nilai probabilitas nilai statistik F lebih rendah dari 0.05, maka pengaruh simultan dapat diterima. Artinya, variabel-variabel *independent* berpengaruh secara simultan terhadap variabel *independent*. Dengan kata lain, model memilik GOF yang memadai.

Selain uji F, kualitas model OLS juga dapat ditinjau dari nilai R^2. Nilai R2 tersebut menunjukkan seberapa besar pengaruh simultan variabel-variabel *independent* terhadap variabel *dependent*. Semakin tinggi R^2, maka pengaruhnya akan semakin tinggi. Adapun untuk uji hipotesis parsial, akan menggunakan uji t dengan meninjau nilai probabilitas (signifikansi) dari koefisien yang dihasilkan. Batas alpha yang ditentukan adalah 0.05 (5%), sehingga jika nilai sig lebih rendah dari 0.05, maka hipotesis parsial dapat diterima.

3.5.5. Heteroscedactic Linear Regression (HLR)

Penelitian ini berupaya untuk menggunakan estimator OLS dalam menguji kekebalan model 4. OLS merupakan estimator terbaik atau yang BLUE apabila asumsi-asumsinya dapat terpenuhi. Asumsi sebagaimana dimaksud yakni normalitas, heteroskedaktisitas, autokorelasi, dan multikolinearitas. Dikarenakan salah satu variabel yang digunakan dalam model 4 adalah pendapatan dengan jumlah sampel yang banyak, maka model berpotensi mengalami masalah heteroskedaktisitas. Ketika asumsi heteroskedaktisitas OLS terlanggar, maka penelitian ini akan menggunakan HLR (heteroscedactif linear regression). Metode HL ini menggunakan pendekatan maximum likelihood sehingga tetap memerlukan asumsi distribusi normal.

3.5.5.1. Uji Kualitas Data

Uji kualitas data dalam HLR yang akan digunakan dalam penelitian ini hanyalah uji normalitas. Pengujian normalitas dilakukan secara grafis, dengan menggunakan line diagram. Axis atau sumbu Y menggunakan linear prediction, sedangkan sumbu X nya adalah inverse normal. Jika data mengikuti garis regresi, maka data dapat dinyatakan berdistribusi normal.

3.5.5.2. Uji Kulitas Model dan Uji Hipotesis

Model HLR menggunakan pendekatan ML, oleh karena itu, uji kualitas modelnya dapat menggunakan statistik Wald Chi^2. Tingginya statistik Wald Chi^2 tersebut menunjukkan semakin tinggi juga kemampuan prediksi model. Jika probabilitas statistik Wald Chi^2 ini lebih rendah dari 0.05, maka model dapat dinyatakan memiliki GOF yang memadai. Selain itu, dapat dijustifikasi bahwa variabel-variabel eksplanatori yang digunakan berpengaruh secara simultan terhadap variabel *dependent*.

Adapun untuk uji parsialnya, akan menggunakan uji t. Hipotesis nol yang diuji bahwa tidak terdapat perbedaan dalam koefisien variabel X terhadap Y. Oleh karenanya, jika probabilitas dari koefisien variabel eksplanatori lebih rendah dari 0.05, maka hipotesis nol harus ditolak. Artinya, terdapat pengaruh yang signifikan dari variabel eksplanatori terhadap variabel *dependent* secara parsial.

3.5.6. Analisis Regresi Data Panel (OLS, FE, dan RE)

Analisis regresi data panel dalam penelitian ini digunakan untuk menganalisa model yang menggunakan data panel yaitu model 5, 6, dan 7. Dikarenakan menggunakan data panel, maka akan terdapat 3 alternatif estimator yang dapat digunakan untuk mengestimasinya. Baltagi (2005) menyebutkan, estimator pertama yakni *ordinary least square* (OLS). Estimator kedua yaitu f*ixed effect* (FE) atau yang juga dikenal sebagai *least square dummy variable* (LSDV). Sedangkan alternatif estimator ketiga adalah *random effect* (RE) atau yang berbasis *generalized least square* (GLS).

Jika model terbaik yang terpilih adalah OLS, maka data ditafsirkan bukan sebagai data panel, tetapi data *cross sectonal* biasa. Oleh karenanya, OLS ini tidak mengantisipasi masalah keragaman waktu dalam data panel. Sementara itu, jika FE yang terpilih, maka nilai *intercept* (konstanta) diasumsikan bersifat tetap sepanjang waktu (*time invariant*). Dengan demikian, diperlukan untuk memasukan *dummy differential constanta* untuk model FE. Sementara itu, jika estimator RE yang terpilih, maka *error* yang terjadi dalam konteks individu (dalam hal ini adalah Provinsi) diasumsikan terjadi secara random (acak).

Estimator mana yang dinilai terbaik akan ditentukan melalui 3 pengujian. Pertama, uji Chow untuk menentukan model yang terbaik antara OLS dengan FE. Kedua, uji Hausman untuk menentukan model terbaik antara FE dengan RE. Ketiga, uji Breuch Pagan LM untuk menentukan model terbaik antara OLS dengan RE. Hipotesis nol (*null hypothesis*) untuk uji Chow adalah bahwa OLS lebih baik daripada FE. Sehingga, jika nilai signifikansi (probabilitas) dari koefisien Chow test lebih besar dari 0.05, maka hipotesis nol diterima.

Sebaliknya, jika nilai signifikansi dari koefisien Chow test tersebut lebih rendah dari 0.05, maka hipotesis alternatif diterima, yang berarti FE lebih baik daripada OLS. Untuk uji Hausman, hipotesis nol (null hypothesis) nya adalah bahwa RE lebih baik daripada FE. Dengan demikian, jika nilai signifikasi dari koefisien Hausman test ini lebih besar dari 0.05, maka hipotesis nol diterima. Sebaliknya, jika signifikansi koefisien Hausman test ini lebih rendah dari 0.05, maka hipotesis alternatif diterima, berarti FE lebih baik daripada RE.

3.5.6.1. Uji Kualitas Data Analisis Regresi Data Panel

Uji kualitas data dalam analisis regresi data panel dimaksudkan untuk memeriksa kualitas data dan estimasi yang dihasilkan. Dikarenakan salah satu model yang dihasilkan dari analisis regresi data panel ini adalah OLS, maka diperlukan uji asumsi klasik untuk memasikan estimator bersifat BLUE (best linear unbias estimator). Jika estimator OLS ini memenuhi asumsi BLUE, maka estimator OLS tersebut merupakan yang terbaik, dan harus dipilih. Dalam hal ini, asumsi BLUE pada OLS akan tercapai apabila data memenuhi asumsi normalitas, multikolinearitas, autokorelasi, dan heteroskedaktisitas. Sementara itu, jika model yang terpilih adalah FE atau RE, maka tidak seluruh uji tersebut diperlukan.

Dalam data panel, asumsi normalitas cenderung sulit didapatkan. Hal ini karena normalitas data lazimnya dilakukan untuk mengetahui apakah distribusi data pada data cross-sectional cenderung proporsional dan tidak melenceng jauh dari standar deviasinya. Meski demikian, penelitian ini tetap melakukan pengujian normalitas data untuk model 5, 6, dan 7 pada data panel dengan menggunakan uji skewness dan kurtosis (sktest) dalam bentuk Joint Test (JT). Namun, uji ini hanya tersedia untuk model yang dihasilkan oleh OLS dan RE, karena untuk FE, normalitas tidak diperlukan.

Uji normalitas dengan uji skewness dan kurtosis ini akan menghasilkan 2 JT yaitu JT e dan JT u. JT e adalah joint test untuk normalitas dalam erorr term (overall error term), sedangkan JT u adalah joint test untuk normalitas dalam time invariant individual (individual error term). Hipotesis nol untuk uji ini adalah bahwa data berdistribusi normal. Oleh karena itu, jika probabilitas JT lebih rendah dari 0.000, maka data tidak berdistribusi normal karena hipotesis nol ditolak.

Selanjutnya, uji heteroskedaktisitas pada analisis data panel dalam penelitian ini menggunakan uji M Wald dengan hipotesis nol berupa data bersifat homoskedaktis. Artinya, jika probabilitas M Wald ini lebih rendah dari 0.000, maka telah terjadi masalah heteroskedaktisitas dalam model karena hipotesis nol nya ditolak. Tetapi, M Wald test ini hanya dapat dilakukan pada model yang diestimasi dengan FE.

Jika model yang terbaik adalah RE, maka uji heteroskedaktisitasnya akan menggunakan metode Breusch Pagan / Cook-Wisberg (BP/CW). Hipotesis nol

untuk uji ini adalah bahwa varians bersifat konstan, sehingga hipotesis alternatifnya adalah bahwa varians bersifat heteroskedaktis. Secara sederhana, heteroskedaktisitas merupakan suatu kondisi dimana varians dari setiap error yang dihasilkan dari setiap observasi variabel bebas mengalami ketidaksamaan. Jika terjadi masalah heteroskedastisitas, maka estimator dalam model regresi menjadi tidak efisien dalam melakukan prediksi.

Adapun masalah autokorelasi (serial correlation) dideteksi dengan menggunakan uji Wooldgride dengan hipotesis nol bahwa data tidak mengalami masalah serial korelasi. Jika probabilitas dari nilai Wooldgride ini lebih rendah dari 0.000, maka dapat dideteksi model mengalami masalah autokorelasi. Uji Wooldgride ini dapat dilakukan pada seluruh estimator (OLS, FE, dan RE). Pengujian autokorelasi ini penting dilakukan untuk menentukan apakah model regresi linear memiliki korelasi antara kesalahan pengganggu pada periode t dengan kesalahan pengganggu pada periode t-1. Jika terjadi korelasi, maka terdapat masalah autokorelasi. Konsekuensi dari permasalahan autokorelasi adalah bahwa varian sampel tidak dapat menggambarkan varian populasinya.

Jika masalah heteroskedaktisitas dan autokorelasi pada analisis regresi data panel ditemukan, maka penelitian ini akan mengestimasi ulang modelnya dengan mengganti distribusi standar errornya dengan robust standar error. Hal ini sebagaimana yang dijelaskan oleh Baltagi (2005) bahwa robust standar error dapat digunakan untuk mengantisipasi masalah heteroskedaktisitas dan korelasi serial (khususnya pada data panel).

Jika mengacu pada Hilbe (2015), salah satu indikator yang dapat menyebabkan model fit adalah bahwa variabel eksplanatori tidak saling berkorelasi satu sama lain. Oleh karena itu, asumsi linearitas atau anti multikoliniearitas harus dipenuhi. Penelitian ini melakukan pengujian multikolinieritas tersebut dengan mengkorelasikan seluruh variabel bebas. Jika terdapat 2 variabel bebas yang berkorelasi satu sama lain sebesar lebih dari 0.7, maka salah satu variabel bebas tersebut tidak akan disertakan ke dalam model. Hal ini mengacu pada Ghozali (2018) yang menyebutkan jika nilai koefisien korelasi antar variabel bebas lebih besar dari 0.7, maka mengindikasikan telah terjadi masalah multikoliniearitas.

3.5.6.2. Uji Kualitas Model dan Pengujian Hipotesis

Uji kualitas model 5, 6, dan 7 dalam penelitian ini ditelaah berdasarkan uji statistik F dan koefisien determinasi R^2. Sehingga, hipotesis untuk model 5, 6, dan 7 dalam penelitian ini akan diuji dengan uji F simultan dan uji t parsial. Uji F akan diestimasi berdasarkan nilai probabilitas dari nilai statistik F model yang dihasilkan. Uji statistik F ini juga merupakan ukuran dari *goodness of fit* (GOF) pada model. Dalam sejumlah literatur, misalnya Ghozali (2018), menyebutkan bahwa uji F adalah uji pengaruh simultan. Sementara itu, Gujarati (2015) menyebut uji F sebagai signifikansi keseluruhan regresi (overall significance regression). Statistik F dapat diestimasi berdasarkan nilai R^2 yang dihasilkan dengan menggunakan rumus sebagaimana berikut:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Menurut Gujarati (2015), uji F dilakukan untuk menentukan apakah slope koefisien secara simultan sama dengan 0. Sehingga, hipotesis nol untuk uji F ini adalah bahwa seluruh variabel bebas yang terdapat pada model tidak memiliki pengaruh terhadap variabel bebas secara bersamaan. Dengan demikian, jika nilai probabilitas dari F-statistik yang dihasilkan lebih rendah dari 0.05, maka hipotesis nol harus ditolak dengan menerima hipotesis alternatif. Hipotesis alternatif tersebut adalah bahwa variabel-variabel bebas berpengaruh secara simultan terhadap variabel terikat.

Uji statistik F dalam data panel hanya dapat dilakukan pada estimator OLS dan FE. Sementara itu, jika estimator RE yang terpilih, maka uji hipotesis simultan atau uji GOF nya menggunakan uji Wald Chi^2. Penafsiran uji ini sama dengan uji F statistik. Selain uji F, penelitian ini juga mempertimbangkan nilai R2 untuk menilai kualitas model yang dibangun. Menurut Gujarati (2015), koefisien determinasi ini juga merupakan ukuran GOF dari garis regresi yang diestimasi untuk memberikan proporsi atau persentase dari variasi total dalam variabel terikat yang dijelaskan oleh semua variabel bebas. Jika mengacu pada Ghozali (2018), koefisien determinasi ini mengukur kemampuan model dalam menerangkan variasi variabel terikat.

Adapun uji t parsial (uji statistik t) akan digunakan untuk menguji pengaruh parsial hipotesis turunan model. Pada model 5, terdapat 3 hipotesis yang akan diuji yaitu mengenai pengaruh parsial dari *overeducation*, *undereducation*, dan horizontal *mismatch* terhadap tingkat pengangguran. Begitupun halnya dengan model 6, terdapat 3 hipotesis parsial yang akan diuji yaitu pengaruh parsial *overeducation*, *undereducation*, dan *mismatch* horizontal terhadap pertumbuhan ekonomi. Adapun model 7, 3 hipotesis parsial yang akan diuji adalah pengaruh *overeducation*, *undereducation*, dan *mismatch* horizontal terhadap produktivitas tenaga kerja.

3.5.7. Analisis Instrumental Variabel (IV)

Dalam penelitian ini, estimator instrumental variabel (IV) digunakan untuk mengantisipasi masalah endogenitas dalam model 5, 6, dan 7. Secara ekonometrik, masalah endogenitas merupakan kondisi dimana variabel eksplanatory berkorelasi dengan error term. Jika kondisi tersebut terjadi, maka koefisien pengaruh variabel eksplanatori terhadap variabel *dependent* menjadi diragukan karena variabel eksplanatorinya cenderung dipengaruhi oleh error term.

Masalah endogenitas berpotensi terjadi ketika menganalisis variabel-variabel ekonomi (Baltagi, 2005). Misalnya, variabel seperti investasi, konsumsi, produksi, permintaan dan penawaran tenaga kerja, hingga pertumbuhan ekonomi cenderung dapat menimbulkan masalah endogeneitas. Dalam penelitian ini, terdapat sejumlah model yang menggunakan variabel-variabel ekonomi dalam skala makro yaitu tingkat pengangguran, pertumbuhan ekonomi, dan produktivitas tenaga kerja. Model 6 dan 7 misalnya, menelaah tentang pengaruh *mismatch* pendidikan terhadap pertumbuhan ekonomi dan produktivitas tenaga kerja dengan menggunakan data panel. Atas dasar itu, kedua variabel tersebut berpotensi mengalami bias endogenitas.

Analisis ini akan digunakan setelah analisis regresi data panel dengan estimator OLS, FE, dan RE. Mula-mula, penelitian ini akan menganalisis model 5, 6, dan 7 dengan menggunakan metode regresi data panel dengan membandingkan model terbaik yang dihasilkan dari estimator OLS, FE, dan RE. Setelah estimasi dihasilkan, selanjutnya penelitian ini akan menggunakan estimator yang terpilih dengan menambahkan variabel instrumental untuk mengontrol endogenitas.

Misalnya, jika yang terpilih adalah FE, maka metode IV akan mengikuti estimator tersebut sehingga menjadi IV FE.

Selah estimasi yang dihasilkan dari estimator IV dihasilkan, maka selanjutnya akan dilakukan uji Hausman untuk membandingkan estimator mana yang lebih baik antara analisis data panel biasa (OLS, FE, atau RE) dengan estimator IV. Jika estimator IV yang terpilih, maka model awal dinilai mengandung masalah endogenitas sehingga perlu diantisipasi dengan IV.

3.5.7.1. Uji Kualitas Data

Pengujian kualitas data dalam analisis IV dilakukan dengan mendeteksi masalah heteroskedaktisitas dan serial korelasi. Metode uji yang digunakan untuk uji heteroskedaktisitas bergantung pada jenis estimator IV yang digunakan. Jika estimator yang digunakan adalah IV FE, maka uji heteroskedaktisitas yang akan digunakan adalah M Wald, sedangkan jika IV RE, maka menggunakan uji Uji Breuch Pagan/Cook-Wisberg. Adapun untuk uji serial korelasi, menggunakan uji Wooldgride.

Apabila terdapat masalah heteroskedaktisitas atau serial korelasi, maka penelitian ini akan menggunakan robust standar error untuk mengatasinya. Robust standar error tersebut dinilai cukup kuat sehingga estimasi yang dihasilkan dapat lebih akurat. Kemudian, untuk memastikan apakah estimator IV lebih baik daripada estimator regresi data panel biasa (OLS, FE, atau RE), maka akan dilakukan uji Hausman Chi^2.

3.5.7.2. Uji Kualitas Model

Uji kualitas model dalam analisis IV dilakukan dengan menelaah nilai overall R^2 atau setara dengan nilai R^2 pada analisis regresi berbasis OLS. Selain itu, ukuran statistik yang digunakan untuk menentukan apakah model bersifat fit serta dapat menguji hipotesis simultan adalah dengan uji Wald Chi^2. Hipotesis nol untuk uji ini adalah bahwa tidak terdapat hubungan antara variabel bebas terhadap variabel terikat. Apabila probabilitas Wald Chi^2 lebih rendah dari 0.05, maka model yang dibangun memiliki model fit (GOF) yang memadai. Dengan kata lain, variabel-variabel eksplanatori dapat mempengaruhi variabel *dependent*.

3.5.8. System Generalized Method of Moments

Analisis system generalized method of moments (Sys-GMM) dalam penelitian ini akan digunakan untuk mengecek kekebalan model 5, 6, dan 7. Tujuan utama penggunaan Sys GMM tersebut adalah untuk mengontrol kemungkinan adanya masalah endogenitas. Meskipun kemungkinan adanya masalah endogenitas tersebut telah diatasi dengan menggunakan estimator IV, tetapi penelitian ini memandang perlu untuk memberikan keyakinan atas kekebalan model dengan metode Sys GMM. Metode ini direkomendasikan oleh Wooldridge (2001) dan Baltagi (2005).

Metode Sys GMM ini diinisiasi oleh Arellano & Bover (1995) dan Blundell & Bond (1998). Penggunaan Sys GMM dalam penelitian ini digunakan untuk mengestimasi ulang model 5, 6, dan 7 karena sejumlah kondisi. Kondisi pertama, adalah karena sys GMM lebih efisien daripada metode first difference (FD) GMM dan instrumental variabel (IV) biasa pada kondisi *limited time series* data, atau i jauh lebih besar daripada t (Blundell & Bond, 1998). Model 5, 6 dan 7 dalam penelitian ini memiliki data *time series* sebanyak 11 unit (2012 sampai dengan 2022), sedangkan data cross sectional sebanyak 33 unit yang merupakan Provinsi di Indonesia.

Kondisi kedua, digunakannya Sys GMM dalam pengecekan kekebalan model pada penelitian ini adalah untuk mengontrol *time-unvarying regional specific effect* atau efek spesifik regional (dalam hal ini adalah provinsi) yang tidak berubah terhadap waktu, menghindari masalah asumsi heteroskedaktisitas dan serial korelasi, serta dapat meningkatkan presisi dan mereduksi bias sampel terbatas (Blundell et al., 2000). Dalam mengimplementasikan metode Sys GMM tersebut, penelitian ini menggunakan lag 2 dari variabel terikat sebagai instrumen. Selain itu, penelitian ini juga menggunakan sejumlah lag *independent* variabel sebagai instrumen (penjelasan lebih lengkapnya tersedia pada setiap pengujian model pada BAB IV).

3.5.8.1. Uji Kualitas Data

Penelitian ini melakukan uji kualitas data pada metode Sys GMM melalui uji validitas dan konsistensi instrumen. Penelitian ini menggunakan uji Sargan untuk menentukan validitas instrumen, serta uji AR (Arrelano Bond) untuk

menguji konsistensi instrumen. Hipotesis nol untuk uji Sargan adalah bahwa instrumen bersifat valid dalam residu dan overeidentifying restrinction. Atas dasar itu, jika p-value dari Sargan coefficient lebih besar dari 0.05, maka hipotesis nol tidak dapat ditolak, artinya instrumen Sys GMM dinyatakan valid. Sebaliknya, jika p-value lebih rendah dari 0.05, maka hipotesis nol harus ditolak, dan hipotesis alternatifnya menunjukkan bahwa instrumen Sys GMM tidak valid.

Adapun hipotesis nol untuk uji AR adalah bahwa instrumen system GMM bersifat konsisten. Oleh karenanya, jika nilai p-value AR (2) (orde kedua) lebih besar dari 0.05, maka instrumen Sys GMM dinyatakan konsisten. Selain kedua uji tersebut, uji kualitas data yang juga penting dilakukan dalam analisis Sys GMM adalah mendeteksi multikolinieritas. Penelitian ini mendeteksi multikolinieritas dengan mengkorelasikan variabel-variabel eksplanatori pada model 5, 6 dan 7.

3.5.8.2. Uji Kualitas Model dan Pengujian Hipotesis

Uji kualitas model dan pengujian hipotesis untuk model 5 hingga 7 dalam metode Sys GMM ini menggunakan Wald Chi^2 test. Wald test ini setara dengan uji F pada metode regresi liniear. Uji Wald ini adalah untuk menelaah apakah terdapat hubungan antara variabel-variabel bebas terhadap variabel terikat secara simultan. Hipotesis nol untuk uji ini adalah bahwa tidak terdapat hubungan antara variabel bebas terhadap variabel terikat. Oleh karena itu, jika nilai probabilitas (p-value) dari Wald Chi^2 test lebih tinggi dari 0.05, maka hipotesis nol dapat ditolak.