Financial Econometric Cheat Sheet: Linear Regression

I Putu Sukma Hendrawan, M.S.M.

February 11, 2025

Tipe Data

- **Menurut skala pengukuran**
  - Nominal: kualitatif, tanpa urutan yang berarti. Contoh: 1=Laki-laki, 0=Perempuan
  - Ordinal: kualitatif, dengan urutan yang berarti. Contoh: 1=Sarjana, 2=Magister, 3=Doktor
  - Interval: kuantitatif, tanpa nol yang absolut, penjumlahan dan pengurangan dimungkinkan namun perkalian dan pembagian tidak dimungkinkan. Contoh: tahun, tanggal
  - Rasio: kuantitatif, dengan nol yang absolut. Contoh: pendapatan, yield, harga.
- **Menurut derajat pengendalian atas lingkungan data**
  - Experimental: data dikumpulkan melalui eksperimentasi pada unit analisis tertentu.
  - Observasional: data retrospektif. Contoh: data pasar modal, data rilisan BPS
- **Menurut cara pengumpulan data**
  - Runut waktu atau time-series: Data dikumpulkan untuk interval waktu yang reguler/tetap. Contoh: data harga saham penutupan harian Bank Mandiri pada Bulan Januari 2024.
  - Cross-sectional: Snapshot suatu grup individual pada satu titik waktu tertentu. Contoh: data pembayaran dividen oleh emiten di Bursa Efek Indonesia pada tahun 2023.
  - Panel atau longitudinal: Kombinasi antara data time-series dan cross-section. Contoh: data pembayaran dividen oleh emiten di Bursa Efek Indonesia pada tahun 2015-2023.
- **Menurut cara kalkulasi**
  - Continuous: Dapat berupa nilai berapapun dan tidak terbatas pada presisi desimal tertentu. Contoh: data yield to maturity untuk surat utang tertentu adalah 7.428975...%. Merupakan hasil pengukuran (measurement).
  - Discrete: Berupa nilai integer tertentu. Contoh: jumlah komisaris independen pada suatu perusahaan. Merupakah hasil penghitungan (count).

Pemodelan Ekonometri

- Spesifikasi masalah - formulasi fenomena yang menjadi perhatian
- Seleksi metodologi - pengumpulan data, pemilihan model ekonomi, dan pendekatan statistik yang relevan. Beberapa pertanyaan dapat diajukan:
  - Variabel apa yang terlibat, variabel mana yang endogen dan eksogen?
  - Apakah terdapat teori dan model ekonomi yang dapat menjelaskan fenomena?
  - Bagaimana cara menganalisisnya, apa data dan teknik statistik yang relevan?
  - Apa hipotesis yang tepat untuk dapat menjawab pertanyaan penelitian?
- Pemodelan ekonometri - Formulasi dan estimasi model ekonometri. Pemodelan ekonometri tidak sama dengan model ekonomi. Contohnya, model ekonomi untuk asset pricing adalah bahwa nilai aset (V) ditentukan oleh imbal hasil instrumen bebas risiko dan profil risiko dari aset $V = f(return, risk)$. Model ekonomi biasanya memuat definisi konseptual yang mesti perlu diperjelas ke dalam definisi operasional. Risiko ddidefinisikan operasional menjadi market risk premium dan size premium $V_{i,t} =$ $\beta_1.Rf_t + \beta_1.\text{Market Premium}_t + \beta_1.\text{Size Premium}_t$.
- Analisis - Pada tahapan ini dilakukan analisis mengenai validitas model ekonometri yang telah disusun. Beberapa pertanyaan yang dapat diajukan:
  - Apakah notasi aljabar pada hipotesis sesuai/terdukung oleh parameter estimasi?
  - Apakah asumsi-asumsi sudah terpenuhi?
- Aplikasi - Penggunaan model ekonometri sesuai tujuannya, misalnya untuk prediksi.

Korelasi dan Kovarians

- Apabila terdapat hubungan antara 2 variabel, jika salah satu variabel menyimpang dari rataannya, maka variabel lainnya juga akan menunjukkan penyimpangan, baik dengan arah yang sama atau arah sebaliknya.
- Dengan arah yang sama: positive covariance. Dengan arah berlawanan: negative covariance.
- $Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- Kelemahan kovarians: bukan ukuran terstandar (standardized measure) sehingga sensitif terhadap skala pengukuran.
- Korelasi: standardized covariance
  - Koefisien korelasi = Pearson Product-Moment Correlation Coefficient = R
  - $R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
  - $-1 \leq R \leq +1$, semakin mendekati -1 maka semakin linear negatif; semakin mendekati +1 maka semakin linear positif; semakin mendekati 0 maka hubungan linear semakin lemah
  - Korelasi adalah gubungan 2 arah: korelasi x dan y sama dengan korelasi y dan x
  - Korelasi tidak terpengaruh satuan (unit-less)
  - Korelasi sangat sensitif terhadap pencilan (outliers)
  - Korelasi bukan hubungan sebab akibat karena korelasi terbatas pada 2 variabel padahal masih dimungkinkan adanya pengaruh variabel lain. Selain itu, hubungan sebab akibat membutuhkan arah hubungan yang tidak terdapat pada korelasi.
- Menguji hipotesis terkait signifikansi korefisien korelasi.
  - Hipotesis nol adalah koefisien korelasi tidak berbeda secara signifikan dari nol atau $H_0 : \rho = 0$.
  - Hipotesis alternatif adalah koefisien korelasi berbeda secara signifikan dari nol atau $H_a : \rho \neq 0$ (two-tail test)
  - Dapatkan t_{calc} dengan formula $t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$
  - Bandingkan t_{calc} dengan daerah kritis $t_{(\alpha/2, n-2)}$ pada tabel distribusi t. Sebagai contoh dengan tingkat signifikansi 0.05; jumlah observasi (n) = 10 maka daerah kritis yang dicari adalah $t_{(\frac{0.05}{2}, 10-2)} = t_{(0.025, 8)}$
  - Apabila $t_{\text{calc}} >$ daerah kritis maka hipotesis nol dapat ditolak atau koefisien korelasi adalah signifikan secara statistk (statistically significant).

Regresi Linear Sederhana

Estimasi Model

- Regresi: Mengukur derajat pengaruh variabelexplanatory/independen/predictor/regressor terhadap variabel response/dependen/outcome/regressand.
- Population regression function (PRF): $y = \beta_0 + \beta_1 x_i$.
- Parameter populasi β_0 and β_1 sangat mungkin tidak diketahui sehingga sample regression function (SRF): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ atau $\hat{y} = b_0 + b_1 x_i$

- Sample regression line i.e., line of best fit: estimasi persamaan garis untuk dataset tertentu yang disajikan pada sebuah scatter plot
- Parameter dari sampel dihitung menggunakan formula sebagai berikut:
  - $b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
  - $b_0 = \frac{(\sum_{i=1}^{n} y_i)-(b_1 \sum_{i=1}^{n} x_i)}{n}$
- Ordinary Least Square (OLS): meminimumkan jumlah kuadrat dari residual (sum of squared residuals/SSR). $\min \sum_{i=1}^{n} e_i^2$ dengan $e_i = y_i - \hat{y_i}$
- y yang merupakan variabel dependen yang diobservasi dengan $\hat{y}$ yang merupakan hasil estimasi menggunakan persamaan regresi.

Pengujian Hipotesis

- Pengujian hipotesis didesain untuk tujuan pengambilan keputusan inferensial tentang parameter populasi melalui sampel yang tersedia, apakah terdapat bukti yang signifikan secara statistik untuk menolak hipotesis tertentu (biasanya hipotesis nol H_0).
- Elemen pengujian hipotesis:
  - **Hipotesis nol atau null hypothesis (H_0)** - hipotesis yang akan diuji
  - **Hipotesis alternatif atau alternative hypothesis (H_a)** - hipotesis yang tidak dapat ditolak apabila hipotesis nol ditolak.
  - **Test statistics** - variabel acak yang distribusi probabilitasnya diketahui pada H_0.
  - **Critical value** - nilai yang akan dibandingkan dengan test statistics untuk pengambilan keputusan apakah H_0 dapat ditolak atau tidak dapat ditolak. Nilai ini merupakan titik pisah batas area ditolak/tidak ditolaknya H_0
  - **Significance level** α - probabilitas penolakan H_0 ketika H_0 mestinya tidak ditolak i.e., type I error. Biasanya: 0.01, 0.05, 0.1.
  - **p-value** - tingkat signifikansi tertinggi di mana H_0 tidak dapat ditolak
  - Rule: apabila p-value lebih kecil dari α maka terdapat bukti statistick unutk menolak H_0.
- Pengujian hipotesis atas koefisien regresi dilakukan sebagai berikut:
  - Pada sisi kiri sample regression function $\hat{y} = b_0 + b_1 x_i$ adalah $\hat{y}$ yang merupakan nilai hasil prediksi atau estimasi menggunakan regresi. Jika ditulis tanpa "hat" maka perlu ditambahkan error atau disturbance term $y = b_0 + b_1 x_i + e$
  - Error atau disturbance term merupakan variabel random yang berdistribusi normal dengan variance σ^2 atau $\epsilon \sim N(0, \sigma^2)$. Karena ϵ adalah error term dari populasi yang tidak diketahui maka yang akan digunakan S^2 sebagai estimator dari ϵ^2.
  - $S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}{n-2}$
  - Standard error dari parameter pada sample regression equation adalah
    s.e.$(b_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ dan
    s.e.$(b_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
  - Menggunakan standard error tersebut dapat dihitung t_{calc} dengan formula: $t_0 = \frac{b_0}{\text{s.e.}(b_0)}$ dan $t_1 = \frac{b_1}{\text{s.e.}(b_1)}$
  - Bandingkan t_{calc} dengan daerah kritis $t_{(\alpha/2, n-2)}$ pada tabel distribusi t. Sebagai contoh dengan tingkat signifikansi 0.05; jumlah observasi (n) = 10 maka daerah kritis yang dicari adalah $t_{(\frac{0.05}{2}, 10-2)} = t_{(0.025,8)}$
  - Hipotesis nol adalah parameter tidak berbeda secara signifikan dari nol atau $H_0 : \beta = 0$.
  - Hipotesis alternatif adalah parameter berbeda secara signifikan dari nol atau $H_a : \beta \neq 0$
  - Apabila $t_{\text{calc}} >$ daerah kritis maka hipotesis nol dapat ditolak atau parameter adalah signifikan secara statistk (statistically significant).

Error Measurement dan Goodness of Fit

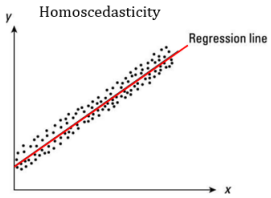
- Total variance dari y adalah jumlag daru total variance yang dapat dijelaskan oleh x i.e., oleh persamaan regresi dan variance yang belum terjelaskan.
- Pernyataan tersebut dapat diubah menjadi pernyataan statistik: sum of squares total (SST) = sum of squares due to regression (SSR) + sum of squares due to error (SSE)
- Penjelasan terstandar: SST = Sum of Squares Total, SSR = Sum of Squares Residuals (SSR), SSE = Sum of Squares Error
- $SST = \sum_{i=i}^{n}(y_i - \bar{y}_i)^2$ i.e., jarak antara nilai observasian y dengan nilai rata-rata dari $\bar{y}$
- $SSR = \sum_{i=i}^{n}(\hat{y}_i - \bar{y}_i)^2$ i.e., jarak antara nilai estimasian $\hat{y}$ dengan nilai rata-rata dari $\bar{y}$
- $SSE = \sum_{i=i}^{n}(y_i - \hat{y}_i)^2$ i.e., jarak antara nilai observasian y dengan nilai estimasian dari $\hat{y}$
- Koefisien determinasi R^2 menyatakan proporsi variasi pda variabel dependen yang dapat dijelaskan oleh variasi pada variabel independen. Koefisien determinasi merupakan ukuran kelaikan suai (goodness of fit). $0 \leq R^2 \leq 1$, semakin kecil nilai R^2 semakin kecil explanatory power dari variabel independen.
- $R^2 = 1 - \frac{SSR}{SST}$
- R^2 relatif sensitif terhadap jumlah variabel independen. Untuk memitigasi bias atas keandalan model akibat penambahan variabel irelevan ke dalam model, digunakan adjusted R-squared.
- Adj.$R^2 = 1 - \frac{n-1}{n-k-1}.R^2$

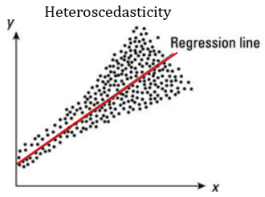
Asumsi model ekonometri

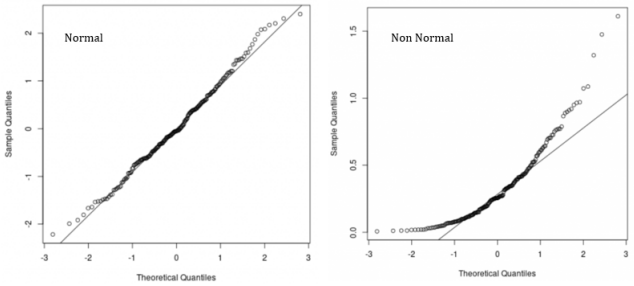
- **Parameters <u>linearity</u>**
  - Variabel dependen y merupakan fungsi linear dari parameter β.
  - Deteksi grafis: residual plot.
  - Deteksi formal: Likelihood-Ratio (LR) test untuk membandingkan goodness of fit antar model regresi.
  - Dampak bila asumsi tidak terpenuhi: parameter/koefisien regresi dan standard error tidak dapat diandalkan (not <u>linear</u> in BLUE).
  - Penanganan: trial and error untuk memperoleh model terbaik.
- **Constant error variance atau <u>homoscedasticity</u>**
  - Variabilitas dari residual adalah konstan $\text{Var}(e|x_1, ..., x_k) = \sigma^2$.
  - Deteksi grafis: pada figur heteroscedasticity terlihat bahwa selisih antara observasian y dengan estimasian $\hat{y}$ berubah seiring perubahan x.
  - Deteksi formal: Breusch-Pagan test dengan hipotesis nol H_0 : homoscedasticity.
  - Dampak bila asumsi tidak terpenuhi: standard error tidak dapat diandalkan atau tidak efisien (not <u>best</u> in BLUE).
  - Penanganan: robust standard error sehingga pengambilan keputusan pada pengujian hipotesis lebih konservatif, weighted least square, transformasi logaritmik.
- **Independent error terms atau <u>no autocorrelation</u> (hanya untuk data runut waktu atau time series)**
  - Tidak terdapat hubungan antara residuals pada observasi x yang satu dengan lainnya. $\text{Corr}(e_t, e_s|x_1, ..., x_k) = \text{Corr}(e_t, e_s) = 0, \forall t \neq s$.
  - Deteksi grafis: correlogram untuk mengetahui korelasi antara e_t dan lagged-nya e_{t-1}
  - Deteksi formal: Durbin-Watson, Breusch-Godfrey dengan hipotesis nol H_0 : no autocorrelation.
  - Dampak bila asumsi tidak terpenuhi: standard error tidak dapat diandalkan atau tidak efisien (not <u>best</u> in BLUE).
  - Penanganan: Generalized Least Square/Cochrane-Orcutt.
- **<u>No perfect multicollinearity</u> (hanya untuk regresi linear berganda)**

- Tidak terdapat hubungan linear antar variabel independen.
- Deteksi menggunakan Variance Inflation Factor (VIF). Apabila VIF < 5 maka tidak terdapat multicollinearity.
- Deteksi dengan correlation matrix antar variabel independen. Koefisien korelasi $|R| \geq 0.80$ menunjukkan indikasi multicollinearity.
- Dampak bila asumsi tidak terpenuhi: persamaan regresi tidak dapat diselesaikan karena jumlah penyelesain tidak terbatas (infinite solutions), parameter dan standard error variabel yang saling berelasi tidak dapat diandalkan.
- Penanganan: menghapus variabel yang saling multikolinear

- **Weak Exogeneity**
  - Tidak terdapat bias akibat tidak dimasukkannya variabel tertentu ke dalam model (omitted variable bias)
  - Weak exogeneity: Tidak terdapat hubungan antara variabel independen dengan residuals $\text{Corr}(x_i, e) = 0, \forall j = 1, ..., k$
  - Deteksi: dapat melalui intuisi dan pemahaman atas konstruk/teori yang diuji serta uji korelasi.
  - Penanganan: penggunaan instrumental variables untuk dapat mengetahui korelasi yang tidak terlihat (unobserved correlation)
- **Normally Distributed Residuals**
  - Residual berdistribusi bebas, normal, dan identik (normally indeppendently and identically distributed (NIID) $e \sim N(0, \sigma_e^2)$.
  - Deteksi grafis: histogram and Q-Q plot. Pada contoh figur Q-Q plot dapat dilihat bahwa residuals berdistribusi normal terlihat dari scatter plot yang berkumpul pada sekitar hypothetical line. Sementara pada figur non-normal terlihat bahwa terdapat scatter plot yang relatif jauh dari hypothetical line.
  - Deteksi formal: Shapiro-Wilk, Komolgorov-Smirnov, Anderson-Darling.
  - Penanganan: Central Limit Theorem







Best Linear Unbiased Estimate (BLUE)

- Apabila asumsi parameters linearity, no perfect multicollinearity, dan exogeneity terpenuhi maka OLS dapat disebut unbiased dan consistent
- Apabila asumsi parameters linearity, no perfect multicollinearity, exogeneity,no autocorrelation, dan homoscedasticity terpenuhi maka OLS dapat disebut Best Linear Unbiased Estimator (BLUE) i.e., efficient
- Apabila asumsi parameters linearity, no perfect multicollinearity, exogeneity, no autocorrelation, homoscedasticity, dan normal errors terpenuhi maka pengujian hipotesis dapat diandalkan.