Revision Notes for ETC2410 – Sample

3. Linear Regression Model

Lecture #3 – The Linear Regression Model

- Simple LRM. **$E(y_i|x_i) = \beta_0 + \beta_1 x_i + u_i$**. (#1) **CONDITIONAL DISTRIBUTION**- model Y with a mean and variance for each level of the predictor X; (#2) **LINEARITY**- that **$E(y_i|x_i) = \beta_0 + \beta_1 x_i$** for each observation i. (#3) **ZERO CONDITIONAL MEAN**- E(u|x) = 0. Deviations occur, but on average = 0
  - **MARGINAL EFFECT**: **β_1**; **1-UNIT INCREASE** in X → E(Y) rises **by β_1**, all other factors constant. Can take the DERIVATIVE of y with respect to x: **$\frac{\partial y}{\partial x} = \beta_1$**. Also, **$\beta_1$ = Cov / Var**.
    - From the linearity assumption, follows that: **$E(y_i|x_i+1) - E(y_i|x_i) = \beta_1$**.
  - **INTERCEPT**: **β_0**; EXPECTED VALUE of the DV (Y) when all predictor variables (X) = **ZERO**. $\hat{\beta}_0$ = the predicted value of y_i when x_i = 0. This is because- **$E(y|x=0) = \beta_0 + \beta_1(0)$**.
  - **RESIDUALS**: **u_i**. captures **UNEXPLAINED VARIATION**; factors affecting the DV NOT CAPTURED by the model's predictors (xi). The residual is: **$\hat{u} = \hat{Y} - Y = y_i - E(y_i \mid x_i)$**.
- OLS estimator. **OLS ESTIMATOR**: ESTIMATES **REGRESSION COEFFICIENTS** (**β_0** and **β_1**) by **MINIMIZING** the SUM OF SQUARED DIFFERENCES b/w **OBSERVED** and **PREDICTED** values.
  - **Aim of OLS estimator**: use **OBSERVED DATA** to estimate the UNKNOWNS β_0 and β_1. Control the fit of the model. There is sampling variability, as these are random variables.
  - **Method of the OLS**: (#1) calculate PREDICTION ERRORS- **$\hat{u}_i = y_i - b_0 - b_1 x_i$**. (#2) substitute into the **SSR Eq**. **$SSR(b_0, b_1) = \sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_1)^2$**. (#3) To **MINIMISE** SSR, set the derivative = 0 with respect to each parameter (β_0 and β_1). This gives TWO equations to solve for TWO unknowns.
    - Formula for b1: **$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}$**. Formula for b0: **$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$**.

Properties of data that hold during OLS. (**$y_i = \beta_0 + \beta_1 x_i + u_i$, i = 1, …, n** by the method of OLS):

- (P1): **$\sum_{i=1}^{n} \hat{u}_i = 0$**. Residuals sum to zero. On avg, model is right- balance in middle of the data.
- (P2): **$\sum_{i=1}^{n} x_i \hat{u}_i = 0$**. Errors are NOT **SYSTEMATICALLY RELATED** to predictor. Errors are EVENLY SPREAD across the x-axis, model captured all linear information in the predictor.
  - For P2: the vector UHAT is **ORTHOGONAL** to columns of X- **x'u** = **0**.
- (P3): **$\bar{y} = \overline{y^{\wedge}}$**, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{y^{\wedge}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$. Avg of actual values = avg of the predicted.

- R^2. **R-SQUARED** (**COEFFICIENT OF DETERMINATION**): measure of **GOODNESS OF FIT**. PROPORTION of **SAMPLE VARIATION** in y explained by x. **R^2 = SSE / SST = 1 – SSR/SST**.
  - **SST**: SST → total sample variation in y (total variation of observed data around its mean). SST does not consider a model. **SST** = **SSE** + **SSR**. DECOMPOSED into SSR and SSE:
    - **SSE**: part of SST explained by IVs in model. **$\hat{\beta}_0 + \hat{\beta}_1 x$ → SSE**, explaining Var(YHAT). Measures variation in **PREDICTED VALUES** around the mean. : **$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$**.
    - **SSR**: the part of SST NOT explained by the model. **UHAT → SSR**, accounting for variation not explained by predictors. **$\hat{u}: \sum_{i=1}^{n} \hat{u}_i^2$**.

- Multiple LRM. The conditional mean of y depends on "k" explanatory variables. For the general MLR, it follows that: **$E(y_i \mid x_{i1} + 1, x_{i2}, \ldots, x_{ik}) - E(y_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \beta_1$**.
  - Coefficient interpretation: b1 measures average Δ in Y in response to 1-unit Δ in X, **HOLDING VALUES OF ALL OTHER PREDICTORS CONSTANT**. **EXPLICITLY** account.

- <u>MLR in matrix notation</u>. (#1) n observations = n equations with the unknown coefficients. (#2) Formulate in matrix algebra – stack equations for each n. (#3) write the model as **$y = X\beta + u$**.
  - Vectors- (#1) y & u = **n x 1** (stack observations from 1 to n). (#2) coefficient V β (**k+1**) x **1**.
  - X is a regressor matrix, with (#1) a **column of 1's** for b0, (#2) **k+1** columns and **n** rows.
- <u>OLS in Matrix Form</u>. **$\hat{\beta} = (X'X)^{-1}X'y$** is the OLS estimator of β in matrix notation.
  - Define a <u>VECTOR</u> of **OLS estimates** for β. (#1) vector of OLS **FITTED VALUES** is $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. (#2) vector of OLS **RESIDUALS** is $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. (#3) So, the **SSR** equals $\sum_{i=1}^{n} \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.
  - X'X <u>MUST BE</u> an **INVERTIBLE MATRIX** → columns of X must be **LINEARLY INDEPENDENT**.
    - **LINEAR INDEPENDENCE**: no columns in the matrix can be built by adding or scaling others (linear combination). Can't recreate one column by mixing others.
      - <u>Interpreting linear dependence</u>: a regressor is not adding any <u>NEW</u> info.