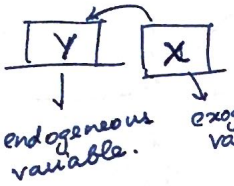
Linear Regression

Statistics and Econometrics.

Linear Regression - Fundamentals

- Linear Regression estimator $\hat{Y}$ is makes an approximate estimate of $E(Y|X)$ given that Y is the dependent and X is the independent variable.
- Population Regression function and Sample Regression function.
- Linear Regression is linear in parameter. Not necessarily in variables.
- Stochastic and deterministic terms (regressors) in the model.
- Minimization of squares of residuals and the OLS criterion. Normal equations.
- Properties of OLS Regression. Goodness of fit measure.
- Classical Linear Regression Model (CLRM) Assumptions. Constant mean for errors, Homoscedasticity, No autocorrelation or serial correlation, etc. Problems of endogenity. Omitted variable bias, Reverse causality.
- Unbiasedness and Variance properties of linear regression (referred to as BLUE properties).
- Endogeneity leads to OLS estimators being biased. Not autocorrelation nor heteroscedasticity.
- Hypothesis testing in regression models, Interval estimation, Q-statistic, Standard errors, Confidence interval, Life expectancy - GDP example.
- Linear regression model with 2 independent variables. Derivations for estimators, coefficients, R^2, and unbiasedness and variance of estimators.

Linear Regression - basics



Influence of changes in x that leads to changes in y $\longrightarrow$ Regression

Multiplier effect. in macroeconomics.

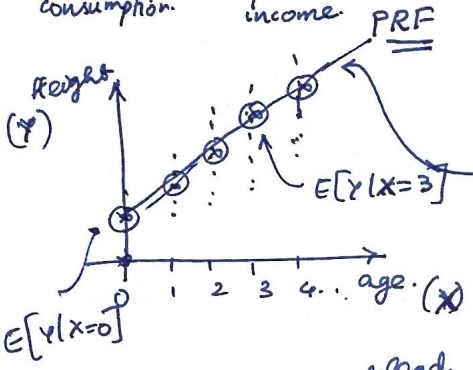
$C = f(I)$

consumption. disposable income.

$C = a + bI$

$\frac{1}{1-b}$ $\rightarrow$ multiplier effect

If disposable inc. increases by 1 unit,



$E[Y|X=0]$

locus of $E[Y|X=a]$ $\forall a \in X$ is the population regression function.

All these can be modelled using linear regression.

$E[Y|X] = \alpha + \beta X$ (line)

$E[Y|X] = \alpha + \beta \log X$

$E[Y|X] = \alpha + \beta X + \delta X^3$ } non linear

The regression function is linear in parameter.

i.e $E[Y|X] = \alpha^2 + \sqrt{\beta} X$ is NOT ALLOWED.

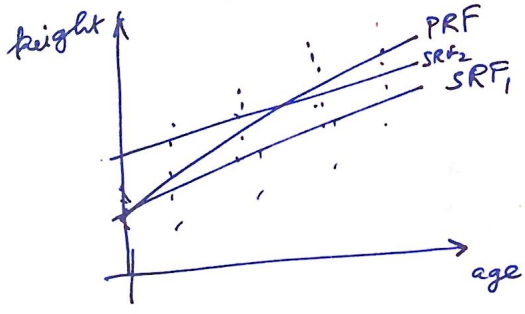
$\text{logit } x = \frac{e^x}{1+e^x}$

Logit regression: $\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ } $\rightarrow$ Non linear regression. logit probit etc.

Assume the simplest form to model.

$E[Y|X] = \alpha + \beta X$.

Now each one takes a different sample and does the same expt $\rightarrow$ obtains the SRF sample regression function



$\hat{Y} = \hat{\alpha} + \hat{\beta} X$

$\hat{Y}$ predicted $\hat{Y}$ $\rightarrow$ It is an estimate of $E[Y|X]$

Need not be the same as $E[Y|X]$

$\hat{Y} \approx E[Y|X]$

$\hat{y} = \hat{\alpha} + \hat{\beta} X$

$\rightarrow$ treating this in a deterministic format. Because X influences $\hat{Y}$ is known & fixed.

But there could be other factors too that influence Y (height) not just X (age).

Stochastic and deterministic terms in regression model

The actual weight can be explained by using stochastic terms to model height.

$$y_i = \alpha + \beta_i x_i + u_i$$

u_i ← disturbance term / randomness term. ⇓ Stochastic term.

Because of inherent randomness in human behavior.

$$Y_i = \alpha + \beta_i x + \delta \Sigma_i + u_i$$

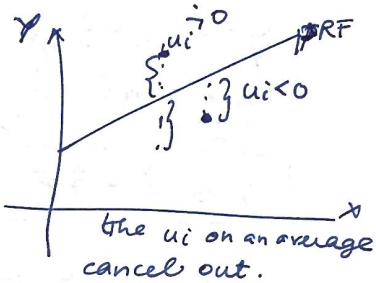
u_i → Captures all the remaining effects that are not captured by x_i

$\delta \Sigma_i$ ↓ we introduce this because,

$Y_i = \alpha + \beta X_i$ ← deterministic regressor

$+ u_i$ ← stochastic regressor.

→ PRF.

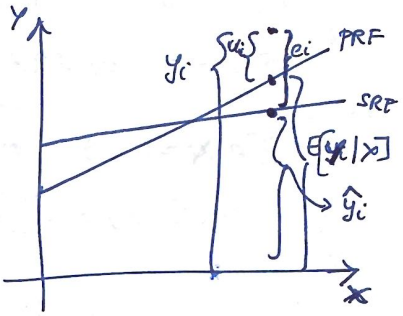


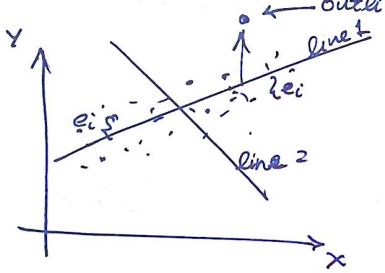
The u_i on an average cancel out.

SRF: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i + e_i$ ← residuals / error terms. "residuals"

e_i's and u_i's will not be the same.

You can't reduce u_i's but you can reduce the e_i's → inherent randomness in the system





← outliers / influencial points.

regressions → means ⇓ Always influenced by outliers. Hence outliers significantly affect regression.

Why line 1 & not line 2 ?

→ minimize errors

→ min e_i^2 →

$$\min \Sigma (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

e_i^1 → +ve & −ve cancel out

e_i^3 → same problem.

The $\hat{\alpha}$ and $\hat{\beta}$ obtained by minimizing are called ordinary least square estimates.

Linear : Use OLS → why ? Gauss Markov.

Non linear : Use MLE

: Use MoM

$$\min \Sigma (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\alpha}^2} = 0 \Rightarrow 2 \Sigma (y_i - \hat{\alpha} - \hat{\beta} x_i)(-1) = 0$$

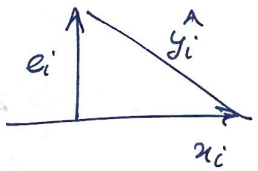
$$\frac{\partial \Sigma e_i^2}{\partial \beta^2} = 0 \Rightarrow 2 \Sigma (y_i - \hat{\alpha} - \hat{\beta} x_i)(-x_i) = 0$$

Note - You talk of distribution only for u_i, not for e_i's

Normal Equations

$$\sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$
$$\sum (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0$$
$$\Rightarrow \boxed{\sum e_i = 0 \quad \sum e_i x_i = 0}$$

e_i's are always perpendicular to x_i's.



$\sum e_i = 0$ }

$$\sum y_i - n\hat{\alpha} - \hat{\beta} \sum x_i = 0$$
$$\frac{\sum y_i}{n} = \frac{n\hat{\alpha}}{n} + \hat{\beta} \frac{\sum x_i}{n}$$
$$\boxed{\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}}$$ → line passing through $\bar{x}$ & $\bar{y}$

∴ Any line passing through $\bar{x}$ and $\bar{y}$ will have $\boxed{\sum e_i = 0}$ → $\bar{e} = 0$

$\sum e_i = 0 \Rightarrow$ SRF always passes through $\bar{x}$ and $\bar{y}$ (sample means ‖ not population means.)

$\sum e_i x_i = 0 \Rightarrow \text{cov}(e, x) = 0$ (sample covariance.)

$$\text{cov}(e, x) = \frac{\sum (x_i - \bar{x})(e_i - \bar{e})}{n} = 0 \quad \text{But } \bar{e} = 0,$$

$$= \frac{\sum (x_i - \bar{x}) e_i}{n} = \frac{\sum x_i e_i - \bar{x} \sum e_i}{n} = 0$$

Properties of OLS Regression

① $\sum e_i = 0 \Rightarrow \bar{y} = \hat{\bar{y}}$ ⇒ SRF always passes through $\bar{x}, \bar{y}$.

$$y = \hat{\alpha} + \hat{\beta} x_i + e_i$$
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$
$$\frac{\sum y}{n} = \frac{n\hat{\alpha}}{n} + \hat{\beta} \frac{\sum x_i}{n} + \frac{\sum e_i}{n}$$
$$\frac{\sum \hat{y}_i}{n} = \frac{n\hat{\alpha}}{n} + \hat{\beta} \frac{\sum x_i}{n}$$

② $\sum x_i e_i = 0 \Rightarrow$ x and e are not correlated.

$\hat{y}$ is a linear function of x.

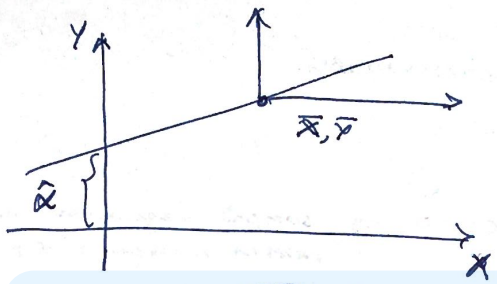
$\boxed{\text{Cov}(x, e) = 0} \Rightarrow \text{Cov}(\hat{y}, e) = 0$

$$\text{cov}(\hat{y}, e) = \text{cov}(\hat{\alpha} + \hat{\beta} x, e) = \text{Cov}(\hat{\alpha}, e) + \hat{\beta}\, \text{Cov}(x, e)$$

$\text{Cov}(\hat{\alpha}, e) = 0$, $\text{Cov}(x, e) = 0$

$$\boxed{\text{Cov}(\hat{y}, e) = 0}$$

③



This shift of origin leads to disappearance of $\hat{\alpha}$

$\hat{\alpha} = 0$ but slope $\hat{\beta}$ remains the same.

$Y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$

$\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}$

$$Y_i - \bar{Y} = \hat{\beta}(x_i - \bar{x}) + e_i$$

↓ shifting ~~from~~ of origin to $(\bar{X}, \bar{Y})$

SRF in deviation form.

$$y_i = \hat{\beta} x_i + e_i$$

$$Y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$

SRF in the raw form

subtract averages. Make $\hat{\alpha} = 0$.

$$\min \Sigma e_i^2 = \Sigma (y_i - \hat{\beta} x_i)^2$$

$$\frac{\partial \Sigma e_i^2}{\partial \beta} = 2\Sigma (y_i - \hat{\beta} x_i)(-x_i) = 0 \Rightarrow \Sigma y_i x_i = \hat{\beta} \Sigma x_i^2$$

$$\Rightarrow \boxed{\hat{\beta} = \frac{\Sigma y_i x_i}{\Sigma x_i^2}}$$

formula for the slope.

$\Sigma e_i = 0 \quad \Sigma Y_i = n\hat{\alpha} + \hat{\beta} \Sigma X_i \Rightarrow \frac{\Sigma Y_i - \beta \Sigma x_i}{n} = \hat{\alpha}$

$\Sigma e_i x_i = 0 \quad \Sigma Y_i x_i = \hat{\alpha} \Sigma x_i + \hat{\beta} \Sigma x_i^2$

$$\Sigma Y_i x_i = \frac{\Sigma Y_i - \beta \Sigma x_i}{n} \Sigma x_i + \hat{\beta} \Sigma x_i^2$$

$$n \Sigma Y_i x_i = \Sigma Y_i \Sigma x_i + \hat{\beta}\left(n \Sigma x_i^2 - \Sigma x_i \right)$$

$$\hat{\beta} = \frac{n \Sigma Y_i X_i - \Sigma Y_i \Sigma X_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$\hat{\beta} = \frac{\Sigma y_i x_i / n}{\Sigma x_i^2 / n} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

④ Goodness-of-fit

R^2 is a factor only in OLS method. Not for MLE etc.

$y_i = \hat{\beta} x_i + e_i$

$\Sigma y_i^2 = \Sigma (\hat{\beta} x_i + e_i)^2$

~~$\Sigma (\hat{\beta} \Sigma y_i)^2 = \hat{\beta}^2 \Sigma x_i$~~

$$\Sigma y_i^2 = \hat{\beta}^2 \Sigma x_i^2 + \Sigma e_i^2 + 2\hat{\beta} \Sigma x_i e_i \quad (\Sigma x_i e_i \to 0)$$

TSS total — ESS explained — RSS residual

$$R^2 = \frac{ESS}{TSS} \times 100 = \%$$

$R^2 = 0.95 \leftarrow 95\%$ variation in Y can be explained by X.

CLRM Assumptions

PRF $\Rightarrow Y_i = \beta_1 + \beta_2 X_i + u_i$ ← random variable

SRF $\Rightarrow Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

$\min \sum e_i^2 \quad \frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = 0 \Rightarrow \sum e_i = 0 \Rightarrow$ sample regression line passes through $\bar{x}$ & $\bar{y}$

$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = 0 \Rightarrow \sum e_i X_i = 0 \Rightarrow \text{cov}(X, e) = 0$

$\text{cov}(\hat{Y}, e) = 0$

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$

$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i$

$y_i = \hat{\beta}_2 x_i + e_i$

$\sum y_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum e_i^2$

TSS ESS RSS

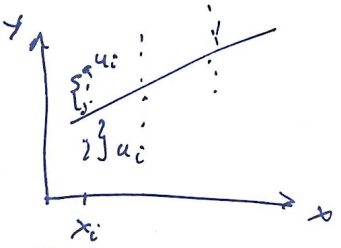
Numerical properties of OLS.

u_i ← random variable → Y_i → $\hat{\beta}_2$

Statistical Properties of OLS

CLRM assumptions

CLRM - classical Linear Regression Model assumptions

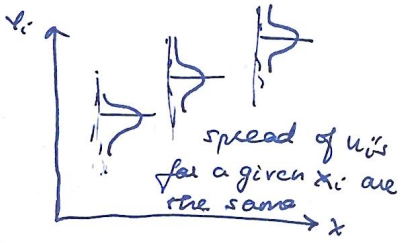


① $E[u_i | X_i] = 0$

The average of u_i's for a given X_i is zero. ⇒ consequence of randomness of u_i's

Otherwise → systematic over/under estimation.

$E[u_i | X_i] = 0 \Rightarrow E(u_i) = 0$



② $E[u_i^2 | X_i] = \sigma^2$ (Homoscedasticity)

Heteroscedasticity problem:

↳ variance of u_i changes with X_i

↳ common with cross section data.

Homoscedasticity $\Rightarrow E[u_i^2 | X_i] = \sigma^2 \Rightarrow E[E(u_i^2 | X_i)] = E[u_i^2] = \sigma^2$

Heteroscedasticity $\Rightarrow E[u_i^2 | X_i] = \sigma_i^2$

$\text{Var}[u_i] = E[u_i^2] - (E[u_i])^2$

$= \sigma^2 - 0 = \sigma^2$

③ $E[u_i u_j | X_i X_j] = 0$ ← mostly untrue for time series data. Eg: GDP → always correlated with previous years data (↳ doesn't change overnight ? {persistence to shock ?})

$\text{Cov}(u_i, u_j) = E[u_i u_j] - E[u_i] E[u_j]$

$E[u_i u_j]$ → 0 (no correlation); $E[u_i]$ → 0; $E[u_j]$ → 0

$\text{Cov}(u_i, u_j) = 0$

$E[u_t, u_{t-1}] = 0$ ← a very strong assumption to make in time series data.

$E[u_t, u_{t-4}] = 0$ ← The data is not correlated for 4 lags

lag: difference between t & t−4

Individual outcome is not correlated with another individuals' outcome.

$E[u_i u_j | x_i x_j] = 0 \Rightarrow$ no auto correlation / serial correlation

Cluster level correlation ⇒ stratified sampling outcomes to a particular data are correlated ⇒ school data

(A) $E[u_i x_i] = 0$ $\quad cov(u_i, x_i) = 0$

Return to education estimation

$wage_i = \alpha + \beta x_i + u_i$ (School)

x_i → schooling.

ability is also a factor for wage

↓

can't be measured & can't be included as a factor

therefore goes into u_i

since ability also determines the number of years of schooling one has attended ⇒ Problem of endogeneity

- Problem of endogeneity ⇒
  - omitted variable bias
  - reverse causality
  - simultaneity.
- Problem of feedback:

x_i causes y_i But also,

$y_i \longrightarrow x_i$

Unbiasedness and Variance Properties of Linear Regression

$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$ $\quad \boxed{w_i = \frac{x_i}{\sum x_i^2}}$ $\quad \sum w_i = \sum \frac{x_i}{\sum x_i^2} = \frac{\sum x_i}{\sum x_i^2} = \frac{\sum (x_i - \bar{x})}{\sum x_i^2}$

$\sum w_i = \frac{n\bar{x} - n\bar{x}}{\sum x_i^2} = 0$

$\therefore \boxed{\sum w_i = 0}$

$\sum w_i x_i = \sum \frac{x_i x_i}{\sum x_i^2} = \frac{\sum x_i^2}{\sum x_i^2} \Rightarrow \boxed{\sum w_i x_i = 1}$

$\sum w_i^2 = \sum \left(\frac{x_i}{\sum x_i^2}\right)^2 = \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{1}{\sum x_i^2} \therefore \boxed{\sum w_i^2 = \frac{1}{\sum x_i^2}}$

$\sum w_i x_i = \sum w_i (X_i - \bar{x}) = \sum w_i X_i - \sum w_i \bar{x} = \sum w_i X_i$

$\therefore \sum w_i x_i = \sum w_i X_i = 1.$

Note

$w_i = \frac{x_i}{\sum x_i^2}$

$\sum w_i = 0$

$\sum w_i x_i = 1$

$\sum w_i^2 = \frac{1}{\sum x_i^2}$

$\sum w_i X_i = 1$

Unbiased if: $E[\hat{\beta}_1] = \beta_1 \quad E[\hat{\beta}_2] = \beta_2$

Why use β_2 & not $\hat{\beta}_2$ here?

$E[\hat{\beta}_2] = E\left[\frac{\sum x_i y_i}{\sum x_i^2}\right] = E\left[\sum \frac{x_i}{\sum x_i^2} y_i\right] = E[\sum w_i y_i]$

$\sum w_i y_i = \sum w_i (Y_i - \bar{Y}) = \sum w_i Y_i - \sum w_i \bar{Y}$ ($\sum w_i = 0$)

Why use PRF & not SRF?

Because expectations are for populations and not for samples.

$\bar{Y}$ - sample average

$E[Y]$ - population average

$E[\hat{\beta}_2] = E[\sum w_i Y_i] = E[\sum w_i (\beta_1 + \beta_2 X_i + u_i)]$

At the same time this is WRONG!!

$\hat{\beta}_2 = \sum w_i Y_i = \sum w_i [\beta_1 + \beta_2 X_i + u_i]$ ← WRONG

$$E[\hat{\beta}_2] = E[\beta_1 \textstyle\sum w_i^{\,0} + \beta_2 \sum w_i x_i^{\,1} + \sum w_i u_i]$$

$\sum w_i x_i = \sum w_i x_i = 1$

$$E[\hat{\beta}_2] = E[\beta_2 + \sum w_i u_i]$$

$u_i \rightarrow$ random variable

$Y_i \rightarrow$ function of u_i. Hence R.V.

$x_i \rightarrow$ are fixed. Not RV.

$w_i \rightarrow$ function of x_i. Not RV.

$$E[\hat{\beta}_2] = E[\beta_2] + E[\sum w_i u_i]$$

$$E[\hat{\beta}_2] = \beta_2 + \sum w_i \, E[u_i]^{\,0}$$

Because of CLRM assumption.

Expectation operator does NOT act on these

$$E[\beta_1] = E[\bar{Y} - \hat{\beta}_2 \bar{X}].$$

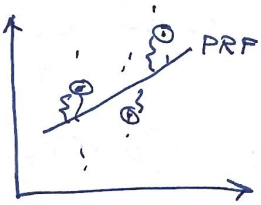
$$E[\hat{\beta}_1] = E[(\beta_1 + \beta_2 \bar{X} + \bar{u}) - \hat{\beta}_2 \bar{X}].$$

$$E[\hat{\beta}_1] = E[\beta_1 - (\hat{\beta}_2 - \beta_2)\bar{X} + \bar{u}]$$

$$E[\hat{\beta}_1] = \beta_1 - \bar{X} E[\hat{\beta}_2 - \beta_2]^{\,0} + E[\bar{u}].$$

$\because E[\hat{\beta}_2] = \beta_2$

$\bar{u} \neq 0$



population $E[u_i] = 0$

sample: $\bar{u} \neq 0$.

OLS min $\Rightarrow \bar{e} = 0$ you take points correspo to the sample above.

$$E[\bar{u}] = E\left[\frac{\sum u_i}{n}\right] = \frac{\sum E[u_i]^{\,0}}{n}$$

CLRM.

$\therefore E[\bar{u}] = 0$

$\therefore$ $E[\hat{\beta}_1] = \beta_1$ | $E[\hat{\beta}_2] = \beta_2$ — OLS estimators are unbiased.

OLS estimators are still unbiased even if you have autocorrelation or heteroscedasticity.

OLS estimator doesn't work if $E[u_i] \neq 0$

$\therefore$ Endogeneity problem $\Rightarrow$ OLS is biased.

$\because \hat{\beta}_2 = \sum w_i Y_i$

$$\text{var}[\hat{\beta}_2] = E[\hat{\beta}_2 - \beta_2]^2 = E[\sum w_i (\beta_1 + \beta_2 x_i + u_i) - \beta_2]^2$$

$$= E[\beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i u_i - \beta_2]^2$$

$$= E[\beta_1 \textstyle\sum w_i^{\,0} + \beta_2 \sum w_i x_i^{\,1} + \sum w_i u_i - \beta_2]^2$$

$$= E[\beta_2 + \sum w_i u_i - \beta_2]^2 = E\left[\sum_{i=1}^{n} w_i u_i\right]^2$$

$$= E[w_1 u_1 + \cdots + w_n u_n]^2 = E[w_1^2 u_1^2 + \ldots + w_n^2 u_n^2 + 2 w_1 w_2 u_1 u_2 + \ldots]$$

$$= E[\sum w_i^2 u_i^2 + \sum_{i \neq j} \sum w_i w_j u_i u_j].$$

$$= \sum w_i^2 E[u_i^2] + \sum_{i \neq j} \sum w_i w_j E[u_i u_j].$$

$E[u_i^2] \rightarrow \sigma^2$

$E[u_i u_j] \rightarrow 0$ (CLRM assumption)

$$\text{var}[\hat{\beta}_2] = \sum w_i^2 \sigma^2$$

$$\therefore \text{var}[\hat{\beta}_2] = \frac{\sigma^2}{\sum x_i^2}$$

$$\text{Var}[\hat{\beta}_1] = E[\hat{\beta}_1 - \beta_1]^2 = E[\bar{Y} - \hat{\beta}_2 \bar{X} - \beta_1]^2$$

$$= E[(\beta_1 + \beta_2 \bar{X} + \bar{u}) - \hat{\beta}_2 \bar{X} - \beta_1]^2$$

$$= E[\beta_1 - (\hat{\beta}_2 - \beta_2)\bar{X} + \bar{u} - \beta_1]^2$$

$$= E[(\hat{\beta}_2 - \beta_2)^2 \bar{X}^2 + \bar{u}^2 - 2(\hat{\beta}_2 - \beta_2)\bar{X}\,\bar{u}]$$

$$= \bar{X}^2 \text{Var}[\hat{\beta}_2] + E[\bar{u}^2] - 2\bar{X} E[(\hat{\beta}_2 - \beta_2)\bar{u}]$$

$\text{Var}[u] = \sigma^2$

$\text{Var}[\bar{u}] = \frac{\sigma^2}{n}$

$$= \frac{\bar{X}^2 \sigma^2}{\sum x_i^2} + E[\bar{u}^2] - 2\bar{X} E[(\hat{\beta}_2 - \beta_2)\bar{u}]$$

$$E[(\hat{\beta}_2 - \beta_2)\bar{u}] = E[(\hat{\beta}_2 - \beta_2)\frac{\sum u_i}{n}] = \frac{1}{n} E[\sum w_i u_i \sum u_i]$$

$$= \frac{1}{n} E\left[\sum w_i u_i^2 + \sum_{i \neq j}\sum (w_i + w_j) u_i u_j\right]$$

$$= \frac{1}{n} \sum w_i \underbrace{E[u_i^2]}_{\sigma^2} + \sum_{i \neq j}\sum (w_i + w_j) E[u_i u_j] \quad = 0$$

$$= \frac{1}{n} \sum w_i \sigma^2 = 0 \text{ ॥}$$

$$E[\bar{u}^2] = E\left[\frac{\sum u_i}{n}\right]^2 = \frac{1}{n^2} E[\sum u_i^2]$$

$$= \frac{1}{n^2} E\left[\sum u_i^2 + \sum_{i \neq j}\sum u_i u_j\right]$$

$$= \frac{1}{n^2}\left[\sum \sigma^2 + 0\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

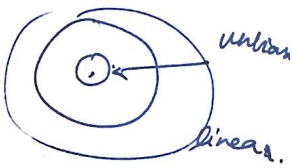
$$\therefore \text{Var}[\hat{\beta}_1] = \frac{\bar{X}^2 \sigma^2}{\sum x_i^2} + \frac{\sigma^2}{n} = \sigma^2\left[\frac{1}{n} + \frac{(\sum X_i)^2}{n^2 \sum x_i^2}\right]$$

Hypothesis testing in regression models

$E[u_i|x_i] = 0$
$E[u_i^2|x_i] = \sigma^2$
$E[u_i u_j | x_i x_j] = 0$
$E[u_i x_i] = 0$

$E[\hat{\beta}_2] = \beta_2$ } unbiasedness.
$E[\hat{\beta}_1] = \beta_1$
$Var[\hat{\beta}_2] = \frac{\sigma^2}{\sum x_i^2}$
$Var[\hat{\beta}_1] = \sigma^2 \left[\frac{1}{n} - \frac{\bar{x}}{\sum x_i^2}\right]$

OLS is the best estimator only among linear unbiased estimators.



statistical questions : (1) Estimation ? → PE
(2) Inference ? → Interval + Hyp Testing.

Interval Estimation / Hypothesis Testing.

$u_i \sim N(0, \sigma^2)$ → assumption.

$Y_i = \beta_1 + \beta_2 x_i + u_i$

$Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$ why ? because you treat the x_i's as fixed.

$w_i = \frac{x_i}{\sum x_i^2}$ $\quad \hat{\beta}_2 = \sum w_i u_i$

$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum x_i^2})$

You make assumptions only on unknowns viz, u_i's.

$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ $\quad \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}}{\sum x_i^2}\right)\right)$

β_2 here.

Pivot → should be a function of the parameter that you are trying to estimate ; but the distribution should be independent of the parameters.

$Q = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0, 1)$.

σ^2 → var of u. & u is not observed.

u is not observed in the sample. Therefore you can't find 's' i.e the sample variance of 'u'

Thus, you instead use the counterpart of u in the sample, which is, 'e'.

$\hat{\sigma}^2 = \frac{\sum (e_i - \bar{e})^2}{n-1}$ But $\bar{e} = 0$ OLS property.

$\therefore \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-1}$

n rand. var. You get e_i by doing : $e_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x$

To get e_i you have to estimate $\hat{\beta}_1$ & $\hat{\beta}_2$.

$\hat{\beta}_1$ & $\hat{\beta}_2$ $\Rightarrow$ two parameters (coeffs) are estimated in regression.

$\therefore$ You lose 2 degrees of freedom.

$$\therefore \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

$$\therefore \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2/\sum x_i^2}} \sim N(0,1)$$

$$\frac{Z}{\sqrt{u/m}} \sim t(m).$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

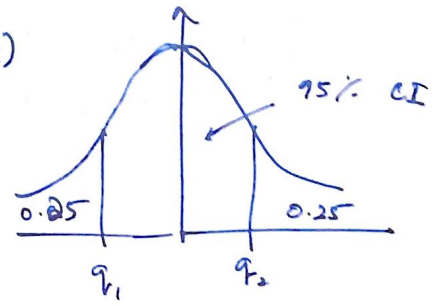
$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\therefore \frac{\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2/\sum x_i^2}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2(n-2)}}} \sim t(n-2) \Rightarrow \boxed{\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2/\sum x_i^2}} \sim t(n-2)} \quad \leftarrow Q_2$$

mean stder. called SE because it is an estimated value.

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}_2\left[\frac{1}{n} + \frac{\bar{X}}{\sum x_i^2}\right]}} \sim t(n-2) \quad \leftarrow Q_1$$

$SE(\hat{\beta}_1)$



$$P[q_1 < Q < q_2] = 0.95$$

$$P\left[q_1 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < q_2\right] = 0.95$$

$$P[q_1 SE(\hat{\beta}_1) < \hat{\beta}_1 - \beta_1 < q_2 SE(\hat{\beta}_1)] = 0.95$$

Symm. dist.

$$P[-q\, SE(\hat{\beta}_1) < \hat{\beta}_1 - \beta_1 < q\, SE(\hat{\beta}_1)] = 0.95$$

$$\boxed{CI = \hat{\beta}_1 \pm q.SE(\hat{\beta}_1)}$$

$$\boxed{CI = \hat{\beta}_2 \pm q\, SE(\hat{\beta}_2)}$$

'q' value comes from the t table at (n-2) dof.

If u_i takes binary values $\{0, 1\}$ the normality assumption cannot be made.

- You go for different functional forms ——— (How?)
- Sometimes changing the functional forms to log often makes a skewed distribution → normal.

Hypothesis Testing.

- Question : Does X have any effect on Y ?.

On slope — If not the corresponding β should be zero!.

$H_0: \beta_2 = 0 \Rightarrow$ X has NO EFFECT on Y.

$H_1: \beta_2 \neq 0$

$Y_i = f(L, K)$. Assume CRS. Cobb-Douglas func. form

$Y_i = \beta_1 L^{\alpha} K^{\beta}$ $\quad \alpha + \beta = 1$

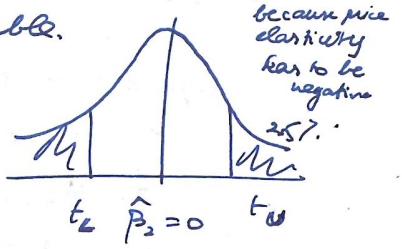
Demand Elasticity .4

Corporates want to know this

$H_1 = \beta_2 < 0$ because price elasticity has to be negative

$t_{crit} \leftarrow$ 5% of 1% → got from the table.

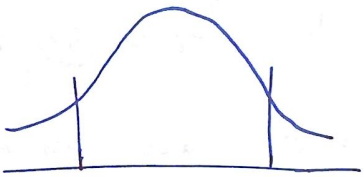
$$t_{calc} = \frac{\hat{\beta}_2 - \beta_2^{\overset{H_0}{=}0}}{SE(\hat{\beta}_2)} \overset{a}{=} \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}.$$



If t_{calc} falls within the acceptance region, accept H_0.

On Intercept

$H_0: \beta_1 = 0$



$Y_i = \beta_1 + \beta_2 X_i + u_i$

$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$

$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

If $\beta_1 = 0 \Rightarrow$ Test for $\bar{Y}$.

$R^2 = \frac{ESS}{TSS}$ OLS property. $\quad \sum y_i^2 = \hat{\beta}_1^2 \sum x_i^2 + \sum e_i^2$.

$TSS = ESS + RSS$

↘ all positive number

If you don't have the intercept, → this formula is violated.

t_{crit}

$$t_{calc} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}.$$

β_1 β_2.

Say. $Y_i = 4 + 0.9X + e_i$

⇒ Check for significance of the parameter.

⇒ Statistical significance matters ~~more~~. more than the values of parameters.

⇒ Although the eqn suggest $\Delta x \Rightarrow \Delta y$.
Doesn't mean anything ~~unless~~ unless we know the statistical significance.

Say. $y_i = 4 + 0.17x_i + e_i$

⇒ Doesn't mean that x_i has a small influence on y_i

⇒ It depends rather on the significance (again)
This is because the values are governed by the ~~number~~ units · in which y_i & x_i are measured

Life Expectancy. GDP

Says. $y_i = 4 + 0.009x_i + e_i$

✗ 1$ ↑ GDP ⇒ Life expectancy increases by 0.009 years

✓ 1000$ ↑ GDP ⇒ Life expectancy increases by 9 years

Inc GDP per capita.

Says, $y_i = (4.0) + 0.009x_i + e_i$

doesn't matter

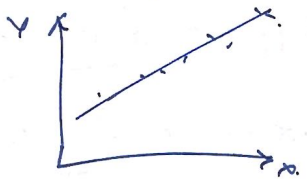
• There are very few cases where intercept actually matters. However you can't eliminate the intercept from regression.

If you take the GDP of the entire population of all countries & regress on Life Expectancy

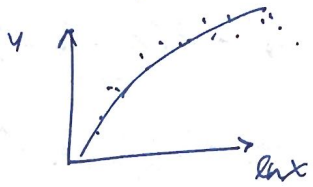
— Is it the PRF or SRF data.

— Ans ⇒ SRF because,

$Y = \alpha + \beta X + u$



$Y = \alpha + \beta \ln X + u$



$Y \rightarrow$ life expectancy in years

$x \rightarrow \ln$(GDP per capita).

Why nonlinear? → At lower levels of GDP per capita, adding income increases life exp significantly

$Y = \alpha + \beta \ln x + u$. → however at higher levels of GDP, a saturation is reached.

$\frac{\partial Y}{\partial x} = \beta \frac{1}{x}$

$\frac{\partial Y}{\frac{\partial x}{x}(100)} = \frac{\beta}{100}$ ← absolute change; relative change.

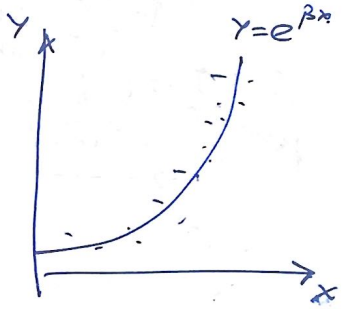
1% Δ in x leads to $\beta/100$ units change in Y.

$\ln Y = \alpha + \beta x + u$

$\frac{1}{Y}\frac{\partial Y}{\partial x} = \beta$

$\frac{\partial Y / Y \times 100}{\partial x} = \beta \times 100$

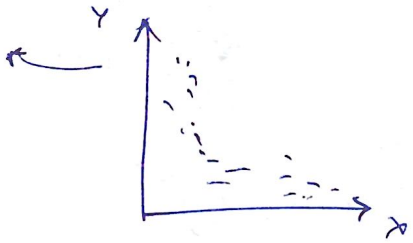
1 unit Δ in x leads $\beta \times 100$ % change in Y.



$\ln Y = \alpha + \beta \ln x + u$

$\frac{1}{Y}\frac{\partial Y}{\partial x} = \beta \frac{1}{x}$

$\frac{\partial Y / Y \times 100}{\partial x / x \times 100} = \beta$



$\ln y = \beta \ln x$.

$e^{\ln y} = e^{\beta \cdot \ln x}$

$y = e^{\beta} + e^{\ln x}$

$y = e^{\beta} + x$?

You can compare R^2 between these $\begin{cases} Y = \alpha + \beta x + u \\ Y = \alpha + \beta \ln x + u. \end{cases}$

$Y = \alpha + \beta x + \delta z + \gamma(zx) + u$ ← interaction effect of z & x on Y.

$\frac{\partial Y}{\partial x} = \beta + \gamma(z)$ ← nonlinear effect of x on Y.

effect depends on the value of z.

Linear Regression with 2 independent variables

PRF $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i$

SRF $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$

$$\Sigma e_i^2 = \Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_1} = 0 \Rightarrow 2\Sigma \underbrace{(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})}_{e_i}(-1) = 0 \Rightarrow \boxed{\Sigma e_i = 0}$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_2} = 0 \Rightarrow 2\Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{2i}) = 0 \Rightarrow \boxed{\Sigma e_i X_{2i} = 0}$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_3} = 0 \Rightarrow 2\Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{3i}) = 0 \Rightarrow \boxed{\Sigma e_i X_{3i} = 0}$$

3 normal equations →

$$\Sigma e_i = 0$$
$$\Sigma e_i X_{2i} = 0$$
$$\Sigma e_i X_{3i} = 0$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$

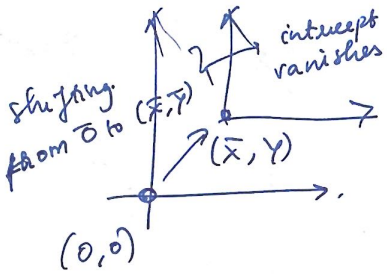
3 dim → Regression plane

k dim ⇒ Regression Hyperplane.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$
$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$

(−)

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_{2i} - \bar{X}_2) + \hat{\beta}_3 (X_{3i} - \bar{X}_3) + e_i$$



$$\boxed{y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + e_i}$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_2} = 0 \Rightarrow \Sigma y_i x_{2i} = \hat{\beta}_2 \Sigma x_{2i}^2 + \hat{\beta}_3 \Sigma x_{2i} x_{3i}$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_3} = 0 \Rightarrow \Sigma y_i x_{3i} = \hat{\beta}_2 \Sigma x_{2i} x_{3i} + \hat{\beta}_3 \Sigma x_{3i}^2$$

$$\Sigma e_i^2 = \Sigma (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})^2$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_2} = 2\Sigma (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})(-x_{2i}) = 0$$

$$\Sigma (x_{2i} y_i - \hat{\beta}_2 x_{2i}^2 - \hat{\beta}_3 x_{2i} x_{3i}) = 0$$

$$\Rightarrow \Sigma x_{2i} y_i = \hat{\beta}_2 \Sigma x_{2i}^2 + \Sigma \hat{\beta}_3 x_{2i} x_{3i}$$

$$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_3} = 0 \Rightarrow 2\Sigma (x_{3i} y_i$$

$$\sum y_i x_{2i} = \hat{\beta}_2 \sum x_{2i}^2 + \hat{\beta}_3 \sum x_{2i} x_{3i}$$

$$\sum y_i x_{3i} = \hat{\beta}_2 \sum x_{2i} x_{3i} + \hat{\beta}_3 \sum x_{3i}^2$$

$\hat{\beta}_2 \rightarrow a$, $\hat{\beta}_3 \rightarrow b$

$p = a \cdot (\alpha) + b\,(\beta)$

$q = a(\alpha) + b(1)$

$$\hat{\beta}_2 = \frac{\begin{vmatrix} \sum y_i x_{2i} & \sum x_{2i} x_{3i} \\ \sum y_i x_{3i} & \sum x_{3i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{2i}^2 & \sum x_{2i} x_{3i} \\ \sum x_{2i} x_{3i} & \sum x_{3i}^2 \end{vmatrix}} = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum x_{2i} x_{3i} \sum y_i x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum x_{2i} \sum x_{3i} \cdot \sum y_i x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_2 = \frac{\frac{\sum y_i x_{2i}}{\sum x_{2i}^2} \sum x_{3i}^2 - \sum x_{2i} x_{3i} \left(\frac{\sum y_i x_{3i}}{\sum x_{3i}^2}\right)}{\frac{\sum x_{2i}^2 \sum x_{3i}^2}{\sum x_{2i}^2} - \frac{\sum x_{2i} x_{3i} \sum x_{2i} x_{3i}}{\sum x_{3i}^2}}$$

$$\hat{\beta}_2 = \frac{b_{12} - b_{13} b_{32}}{1 - b_{32} b_{23}}$$

$b_{12} \rightarrow$ simple regression of y on x_2.

$y = \alpha + \beta x + \mu$

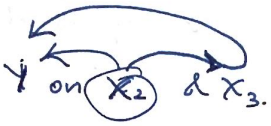
$$\rho_{xy} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}} \quad \begin{matrix} \leftarrow \text{cov} \\ \leftarrow \text{stdev.} \end{matrix} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \frac{\hat{\beta}}{\rho} = \frac{\frac{\sum x_i y_i}{\sum x_i^2}}{\frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}} = \frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}}$$

$$\hat{\beta} = \frac{\rho \sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}} = \frac{\rho s_y}{s_x}$$

$$\hat{\beta}_2 = \frac{b_{12} - b_{13} b_{32}}{1 - b_{32} b_{23}} = b_{12} \text{ if } \rho_{x_2 x_3} = 0$$

If X_2 & X_3 are not correlated. then 2 var regression & a 3 var regression will give the same coefficients.



2 var reg ⇒ both effects taken into account

3 var reg ⇒ removing the effect of X_3 on X_2.
i.e the $b_{13} b_{32}$ term.

assuming if all the effects are true.

Then $b_{12} > \hat{\beta}_2$ ⇒ you overestimate by performing simple regression.

$$\hat{\beta}_2 = \frac{b_{12} - b_{13} b_{32}}{1 - b_{32} b_{23}} = \frac{\rho_{12} \frac{s_1}{s_2} - \rho_{13} \frac{s_1}{s_3} \rho_{32} \frac{s_3}{s_2}}{1 - \rho_{32} \rho_{23} \frac{s_3}{s_2} \cdot \frac{s_2}{s_3}}$$

$$\hat{\beta}_2 = \frac{\rho_{12} \frac{s_1}{s_2} - \rho_{13} \rho_{32} \frac{s_1}{s_2}}{1 - \rho_{32} \rho_{23}}$$

$$\hat{\beta}_2 = \frac{s_1}{s_2} \left[\frac{\rho_{12} - \rho_{13} \rho_{32}}{1 - \rho_{32} \rho_{23}} \right] = \left(\frac{\rho_{12} - \rho_{13} \rho_{23}}{1 - \rho_{23}^2} \right) \frac{s_1}{s_2}$$

If corr 2,3 is high $(1 - \rho_{23}^2)$ is very small, it just inflates the effect much more

Regress Y on X_3
→ take residuals → e_{13i}

Reg Y on X_3
→ take residuals → e_{23i}

Reg e_{13i} on e_{23i}

$$\text{slope} = \frac{\sum e_{13i} e_{23i}}{\sum e_{23i}^2} = \frac{\sum (y_i - b_{13} x_{3i})(e_{23i})}{\sum e_{23i}^2}$$

$$\text{slope} = \frac{\sum y_i e_{23i} - \sum b_{13} x_{3i} e_{23i}}{\sum e_{23i}^2}$$

$\sum b_{13} x_{3i} e_{23i} \to 0$ because of the normal equation.

$$\text{slope} = \frac{\sum y_i e_{23i}}{\sum e_{23i}^2} = \frac{\sum y_i (x_{2i} - b_{23} x_{3i})}{\sum (x_{2i} - b_{23} x_{3i})^2}$$

$$= \frac{\sum y_i x_{2i} - b_{23} \sum y_i x_{3i}}{\sum x_{2i}^2 + b_{23}^2 \sum x_{3i}^2 - 2 b_{23} \sum x_{2i} x_{3i}}$$

$$= \frac{\sum y_i x_{2i} - \frac{\sum x_{2i} x_{3i}}{\sum x_{3i}^2} \sum y_i x_{3i}}{\sum x_{2i}^2 + \left(\frac{\sum x_{2i} x_{3i}}{\sum x_{3i}^2}\right)^2 \sum x_{3i}^2 - 2 \frac{\sum x_{2i} x_{3i} \sum x_{2i} x_{3i}}{\sum x_{3i}^2}}$$

$$= \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum x_{2i} x_{3i} \sum y_i x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 + \left(\sum x_{2i} x_{3i}\right)^2 - 2\left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_{12.3} = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum x_{2i} x_{3i} \sum y_i x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

neg of 1 on 2 keeping 3 const

$$=$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \qquad \text{Var}(\hat{\beta}_1) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}}{\sum x_i^2}\right]$$

$$\text{Var}\left(E[\hat{\beta}_{12.3}]\right) = E\left[\frac{\sum e_{13i} e_{23i}}{\sum e_{23i}^2}\right] = E\left[\sum (y_i - b_{13}\right.$$

effect of x_2 on y keeping x_3 const → normal equations

$$= E\left[\frac{\sum (y_i - b_{13} x_{3i}) (e_{23i})}{\sum e_{23i}^2}\right]$$ ($\sum x_{3i} e_{23i} = 0$)

you can change one of the $y_i \to Y_i$ or $x_i \to X_i$

$$= E\left[\frac{\sum Y_i e_{23i}}{\sum e_{23i}^2}\right] = E\left[\frac{\sum (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i) e_{23i}}{\sum e_{23i}^2}\right]$$

$$= E\left[\frac{\beta_1 \sum e_{23i} + \beta_2 \sum X_{2i} e_{23i} + \beta_3 \sum X_{3i} e_{23i} + \sum u_i e_{23i}}{\sum e_{23i}^2}\right]$$

($\sum e_{23i} = 0$, $\sum X_{3i} e_{23i} = 0$)

$$E\left[\hat{\beta}_{12.3}\right] = E\left[\frac{\beta_1 \cdot 0 + \beta_2 \sum x_{2i} e_{23i} + \beta_3 \cdot 0 + \sum u_i e_{23i}}{\sum e_{23i}^2}\right]$$

$$E\left[\hat{\beta}_{12.3}\right] = E\left[\beta_2 \frac{\sum x_{2i} e_{23i}}{\sum e_{23i}^2} + \frac{\sum u_i e_{23i}}{\sum e_{23i}^2}\right]$$

$$E\left[\hat{\beta}_{12.3}\right] = \beta_2 \frac{\sum x_{2i} e_{23i}}{\sum e_{23i}^2} + \frac{\sum e_{23i} E[u_i]}{\sum e_{23i}}$$ → 0 by CLRM.

$$E\left[\hat{\beta}_{12.3}\right] = \beta_2 \frac{\sum x_{2i} e_{23i}}{\sum e_{23i}^2}$$

Yon ×

$$\frac{\sum y_i e_i}{\sum e_i^2} = \frac{\sum (\hat{y}_i + e_i) e_i}{\sum e_i^2} = \frac{\sum \hat{y}_i e_i + \sum e_i^2}{\sum e_i^2}$$ ($\sum \hat{y}_i e_i \to 0$)

$$E\left[\hat{\beta}_{12.3}\right] = \beta_2 \frac{\sum (\hat{x}_{2i} + e_{23i}) e_{23i}}{\sum e_{23i}^2}$$

$$E\left[\hat{\beta}_{12.3}\right] = \beta_2 \left(\frac{\sum \hat{x}_{2i} e_{23i}}{\sum e_{23i}^2} \to 0 + \frac{\sum e_{23i}^2}{\sum e_{23i}^2}\right)$$

$$E\left[\hat{\beta}_{12.3}\right] = \beta_2 \cdot 1 \Rightarrow \boxed{E\left[\hat{\beta}_{12.3}\right] = \beta_2}$$

unbiasedness

$$\text{Var}\left[\hat{\beta}_{12.3}\right] = E\left[\frac{\sum e_{13} e_{23i}}{\sum e_{23i}^2} - \beta_2\right]^2$$

$$= E\left[\frac{\sum (y_i - b_{13} x_{3i}) e_{23i}}{\sum e_{23i}^2} - \beta_2\right]^2$$

$$= E\left[\frac{\sum y_i e_{23i}}{\sum e_{23i}^2} - \beta_2\right]^2$$

$$= E\left[\frac{\sum (\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i) e_{23i}}{\sum e_{23i}^2} - \beta_2\right]^2$$

$$= E\left[\frac{\beta_1 \sum e_{23i} + \beta_2 \sum x_{2i} e_{23i} + \beta_3 \sum x_{3i} e_{23i} + \sum u_i e_{23i}}{\sum e_{23i}^2} - \beta_2\right]$$

($\sum e_{23i} \to 0$, $\sum x_{3i} e_{23i} \to 0$)

$$\text{var}\left[\hat{\beta}_{12.3}\right] = E\left[\beta_2 + \frac{\sum x_{2i} e_{23i}}{\sum e_{23i}^2} + \frac{\sum u_i e_{23i}}{\sum e_{23i}^2} - \beta_2\right]^2$$

$$\text{var}\left[\hat{\beta}_{12.3}\right] = E\left[\frac{\sum u_i e_{23i}}{\sum e_{23i}^2}\right]^2 = \frac{1}{\left(\sum e_{23i}^2\right)^2} E\left[\sum u_i e_{23i}\right]^2$$

$$= \frac{1}{\left(\sum e_{23i}^2\right)^2} E\left[\sum u_i^2 e_{23i}^2 + \sum_{i \neq j}\sum u_i u_j e_{23i} e_{23j}\right]$$

$$= \frac{1}{\left(\sum e_{23i}^2\right)^2}\left[\sum e_{23i}^2 E\left[u_i^2\right] + \sum_{i \neq j}\sum e_{23i} e_{23j} E\left[u_i u_j\right]\right]$$

($E[u_i^2] \to \sigma^2$, $E[u_i u_j] \to 0$)

$$= \frac{\sigma^2 \sum e_{23i}^2}{\left(\sum e_{23i}^2\right)^2} = \frac{\sigma^2}{\sum e_{23i}^2} = \frac{\sigma^2}{(1 - R_{23}^2)\sum x_{2i}^2}$$

$e \uparrow$, $R^2 = e^2$ in 2 var reg. $\Rightarrow R^2 \uparrow$.

$Y \to X_2, X_3$ $1 - R^2$ is very low $\Rightarrow$ stderr is very high

X_2 & X_3 are coll $\Rightarrow$ t stat is very less.

$\therefore$ you would wrongly conclude that..

X_2 & X_3 have no effect on Y.

But R^2 still is high $\to$ problem of MULTICOLLINEARITY.

k var regression : var $\Rightarrow$ $\dfrac{\sigma^2}{\left(1 - R^2_{2.345\cdots k}\right)\sum x_{2j}^2}$.

Multicollinearity

- not a problem of the population but of the sample.
- SE is high $\Rightarrow$ β is low.
- Remedial measure ?

Y on K & L — Large firms high K & high L

K & L are correlated.

You end up concluding that K & L has no effect on Y.

Reg. $\frac{Y}{L}$ on $\frac{K}{L}$. But forecast will be correct because R^2 is high.

BUILD YOUR FUNDAMENTALS!