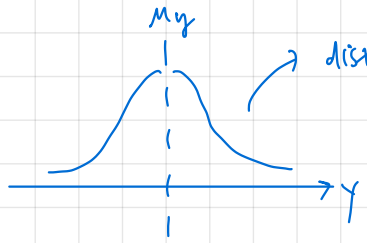
# Bivariate linear regression

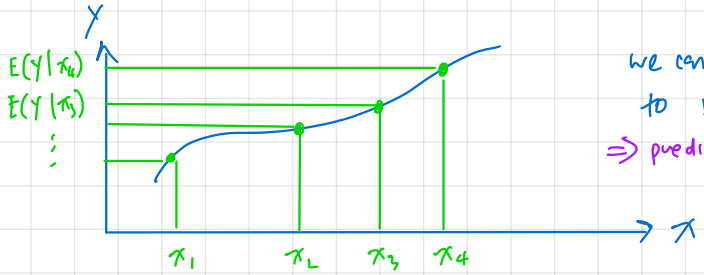
## ① Linear regression model

1) introducing another variable ū iterated expectation

↳ suppose we believe some RV $Y$ is related to some RV $X$.
ie. for a given value of $x$, there is some distribution of $y$.




distribution of $Y$ $\Rightarrow$ $E(Y) = E(E(Y|X=x))$




we can use some function $\hat{y} = f(x)$ to predict $E(Y|X=x)$
$\Rightarrow$ predicts expectation of $Y$ ie. r/s that holds on average

2) using a linear model

↳ suppose we think there is some linear population r/s between $Y$ and $X$

$$E(Y|X) = \beta_0 + \beta_1 X \Rightarrow E(Y_i|X_i) = \beta_0 + \beta_1 X_i \text{ for some observation } i$$

↳ $X$ is not the sole determinant of $Y$, so for each $X, Y$ observation we can abstract out these effects as some random error $u$

$$Y_i = \beta_1 X_i + \beta_0 + u_i$$

coefficients

dependent variable

independent variable

population error
↳ incorporates factors responsible for difference between prediction & actual value

## ② ordinary least squares estimation

↳ given some model $f(x)$, we can find the values of coefficients by solving simultaneously the closed form solutions describing the minimisation of sum of squared difference between model & observations

↳ intuition for using SSE: square penalises large deviations and punishes all deviations, positive or negative

## estimating $\beta_0$ and $\beta_1$ using OLS estimation

1. we posit that $Y_i = \beta_1 X_i + \beta_0 + u_i$

1.1 so we can define estimators $\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0$, $\hat{u}_i = Y_i - \hat{Y}_i$

1.2 To perform OLS estimation, we minimise $(Y - \hat{Y})^2$ on aggregate by summing.
So our objective function $Q = \sum (Y_i - \hat{Y})^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)^2$

2. we can use partial differentiation to derive the relevant equations.

$$Q = \sum (Y - \hat{Y})^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)^2$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = 0$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\sum Y_i X_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum Y_i X_i - \hat{\beta}_0 \sum X_i}{\sum X_i^2}$$

$$= \frac{\sum Y_i X_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum X_i}{\sum X_i^2} = \frac{\sum Y_i X_i - \bar{Y} \sum X_i}{\sum X_i^2 - \bar{X} \sum X_i}$$

re-express as $\bar{X}$

$$= \frac{\sum Y_i X_i - \bar{Y}(n\bar{X})}{\sum X_i^2 - \bar{X}(n\bar{X})} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2}$$

$$\hat{\beta}_1 = \frac{cov(X,Y)}{V(X)} = \frac{S_{xy}}{S_x^2}$$

$$= \frac{\frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$$

$$\frac{1}{n} \sum \hat{u}_i = 0$$

$$\frac{1}{n} \sum \hat{Y}_i = \bar{Y}$$

$$\sum \hat{u}_i X_i = 0$$

$$cov(\hat{u} X) = 0$$

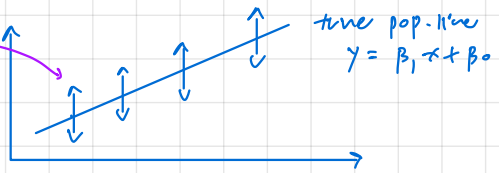
$$TSS = SSR + ESS$$

## ③ Assumptions & sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

### 1) Assumptions

1. given $X_i$, the conditional distribution of the population error term has a mean of 0. → more important, $\rho_{X_i, u_i} = 0$ for $\hat{\beta}_1$

$$E(u_i \mid X_i = x) = 0$$ → errors centered @ line




↳ $\rho_{X_i, u_i} = 0$, but not necessarily converse

2. $(X_i, Y_i)$ are independently and identically distributed
3. large outliers are rare

### 2) mean and variance of $\hat{\beta}_1$

$$E(\hat{\beta}_1) = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{1}{n} \frac{Var((X_i - \bar{X})u_i)}{Var(X_i)^2}$$

$\frac{1}{n-2}\sum_i (X_i - \bar{X})^2 \hat{u}_i^2$

↳ $\frac{1}{n}\sum_i (X_i - \bar{X})^2$

scaling affects here

$n \to \infty$, $\hat{\beta}_1 \to N\left(\beta_1, \frac{1}{n} \frac{Var((X - \bar{X})u)}{Var(X)^2}\right)$

$n \geq 100$ normally sufficient.

↳ Because $Var(\hat{\beta}_1) \propto \frac{1}{n}$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1$ is a consistent estimator.

$$\lim_{n \to \infty} \hat{\beta} = \beta_1$$

↳ the larger $Var(X)$, the smaller $Var(\hat{\beta}_1)$
⇒ intuition: longer horizontal, better line

# proof that $\hat{\beta}_1$ is unbiased given OLS assumptions

1. From partial derivative : $\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

1.1 Also, $Y_i = \beta_0 + \beta_1 X_i + u_i$, $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$ $\rightarrow Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$

1.2 so $\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(\beta_1 (X_i - \bar{X}) + (u_i - \bar{u}))}{\sum (X_i - \bar{X})^2}$

1.3 so $\hat{\beta}_1 - \beta_1 = \frac{\sum (X_i - \bar{X})(u_i - \bar{u})}{\sum (X_i - \bar{X})^2}$

1.4 $\hat{\beta}_1 - \beta_1 = \frac{\sum (X_i - \bar{X}) u_i - \sum (X_i - \bar{X}) \bar{u}}{\sum (X_i - \bar{X})^2}$ $\sum (X_i - \bar{X}) \bar{u} = \sum X_i - n\bar{X} = 0$

$= \frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2}$

2. Theorem : $E(Y) = E(E(Y|X))$

2.1 $E(\hat{\beta} - \beta_1) = E(E(\hat{\beta} - \beta_1 \mid X_1, X_2 \dots X_n))$

↓ estimation conditioned on observations

$= E\left(E\left(\frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2} \mid X_1 \dots\right)\right)$ → conditioned on $X_i$ is itself

$= E\left(\frac{1}{\sum (X_i - \bar{X})^2} E\left(\sum (X_i - \bar{X}) u_i \mid X_1 \dots X_n\right)\right)$

$E(u_i \mid X_1 \dots X_n)$ is a constant ⇒ take out

$f(u_i \mid X_1 \dots X_n)$ not constant ↓ linear

$E\left(\frac{1}{\sum (X_i - \bar{X})^2} E\left(\sum (X_i - \bar{X}) \mid X_1 \dots\right) k\right)$ conditioned on $X$ is itself

$= E\left(\frac{\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \cdot k\right)$ ($\sum (X_i - \bar{X}) = 0$)

$E(u_i \mid X_1 \dots X_n) = k X_i$

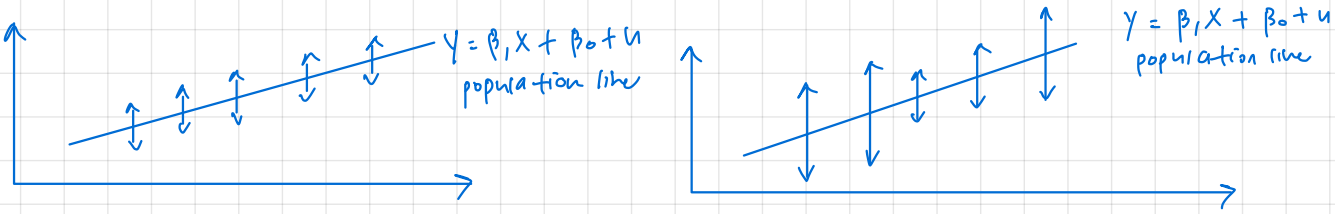
$k \frac{\sum (X_i - \bar{X}) X_i - (X_i - \bar{X})\bar{X} + (X_i - \bar{X})\bar{X}}{\sum (X_i - \bar{X})^2} = k + \frac{\sum (X_i - \bar{X})\bar{X}}{\sum (X_i - \bar{X})^2}$

$= k$

↳ constant bias

3) homo- and heteroskedasticity

↳ constant $Var(u_i \mid X = x)$ — non-constant




↳ whether homo- or heteroskedastic, need to consider underlying relationships and behaviours being modelled (eg. income vs. years of education ⇒ older likely to be more spread)

↳ implications on computation of variance

regardless: $Var(\hat{\beta}_1) = \frac{1}{n} \frac{Var((X_i - \bar{X})u_i)}{Var(X_i)^2} \Rightarrow SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)}$

↳ robust to heteroskadicity

if homoskedastic, then $Var(\hat{\beta}_1)$ simplifies:

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{\sigma_X^2} \cdot \frac{1}{n} \to \frac{1}{n-2}\sum \hat{u}^2$$

↳ $\frac{1}{n}\sum (X_i - \bar{X})^2$

→ if homoskedastic, estimates will converge as $n \to \infty$

## ④ Hypothesis testing

↳ using OLS estimators to estimate $\beta_1$ and $\beta_0$ — coefficients that describe a relationship, we can test if given the data, their magnitudes are statistically different from 0 (ie. no relationship) or some other value

1. $H_0 : \beta_1 = k$, $H_1 : \beta_1 \neq k$
   ↳ under $H_0$, since $n \geq 100$, $\hat{\beta} \sim N(\beta_1, SE(\hat{\beta}_1))$ approximately
2. compute $b_1$ and se
3. compute t-statistic — $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim$ approx normal
4. compute p-value
5. If two tailed, compute confidence interval

$\hat{y}_i = \hat{b}_1 x_i + \hat{b}_0$ $\quad \hat{u}_i = y_i - \hat{y}_i$

$\hat{b}_1 = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$

$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$

$\hat{se}(\hat{\beta}_1) = \sqrt{\frac{var(c(x_i - \bar{x})\hat{u}_i)}{var(x_i)^2}}$ — scaling affects here ⇒ take out, $k^2$ on var, $k$ on $se(\hat{\beta}_1)$

$= \sqrt{\frac{\frac{1}{n-2}\sum (x_i - \bar{x})^2 \hat{u}_i^2}{[\frac{1}{n}\sum (x_i - \bar{x})]^2}}$

$m = \pm t_{n-1, \alpha/2} \cdot se$

## (5) measures of fit

1) $R^2$ value

↳ measures fraction of variance of $Y$ that is explained by $X$. unitless. $0 \le R^2 \le 1$

↳ in the case of a single regressor, $R^2$ is the square of the sample correlation between $X$ & $Y$

⇒ perfect correlation ($|R| = 1$) means all points lie exactly on the line

$$Y_i = \hat{Y}_i + \hat{u}_i$$

$$R^2 = \frac{\text{explained sum of squares}}{\text{total sum of squares}} = \frac{\frac{1}{n-1}\sum_i (\hat{Y}_i - \bar{\hat{Y}})^2}{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2}$$

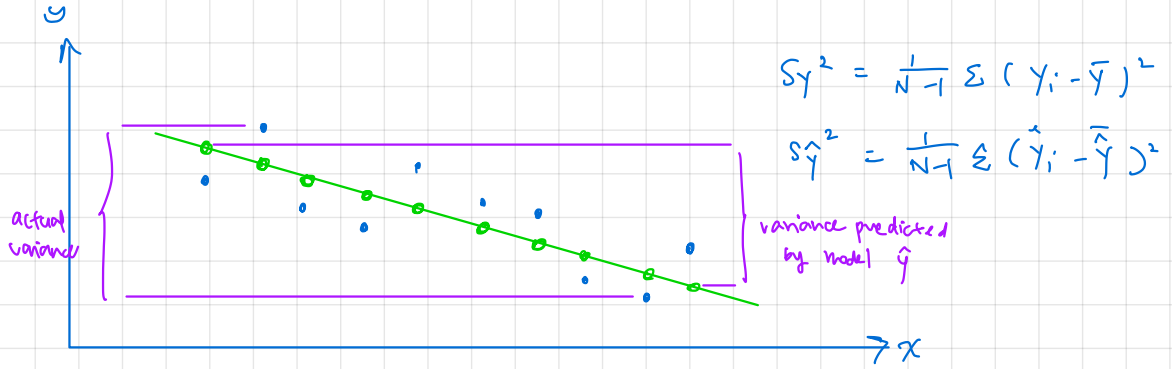
($\hat{Y}_i$: predicted variance given $X_i$; $Y_i$: actual observed variance.)

$$= 1 - \frac{\text{sum of squared residuals}}{\text{total sum of squares}} = \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2}$$

(sum of squared residuals: fraction of variance not explained by regression)

↳ explained sum of squares + sum of squared residuals = total sum of squares

↳ intuition: there are two sources of variation in $Y$; $X$ and some other unmodelled factors. If $R^2 = 0$, then the regression of $Y$ on $X$ explains none of the variation of $Y$ and $ESS = 0$. If $R^2 = 1$, the regression explains all variation of $Y$ and $ESS = TSS$




$$S_Y^2 = \frac{1}{N-1}\sum (Y_i - \bar{Y})^2$$

$$S_{\hat{Y}}^2 = \frac{1}{N-1}\sum (\hat{Y}_i - \bar{\hat{Y}})^2$$

2) Standard error of regression (SER)

↳ measure of the spread of observations about regression line wrt $Y$ (in units of $Y$)

⇒ sample std. deviation of OLS residuals

(both can normalise w.r.t. $\bar{Y}$)

$$SER = \sqrt{\frac{1}{n-2}\sum_i (\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_i \hat{u}_i^2}$$

$\bar{\hat{u}} = \frac{1}{n}\sum \hat{u}_i = 0$ by definition of OLS estimation

$$RMSE = \sqrt{\frac{1}{n}\sum_i \hat{u}_i^2}$$

## ⑥ interpretation of regression line

1) slope, intercept, predictions
   - slope: $\Delta y$ for every unit change in $x$
   - intercept: $\hat{y}$ when $x = 0$
   - predictions → predicting outside of modelled range is not a good idea as it is uninformed

2) changing units

$Y \to aY \Rightarrow \hat{\beta}_1 \to a\hat{\beta}_1$, $\hat{\beta}_0 \to a\hat{\beta}_0$

$X \to bX \Rightarrow \hat{\beta}_1 \to \frac{\hat{\beta}_1}{b}$, $\hat{\beta}_0$ unchanged

3) regression when $X$ is a binary variable

↳ $\beta_1$ is not so much a slope, but a translation brought about by presence of $X$

$$Y_i = \beta_1 X_i + \beta_0 + u_i$$

- $X_i = 0$ → $Y_i = \beta_0 + u_i$
- $X_i = 1$ → $Y_i = \beta_1 + \beta_0 + u_i$

$\hat{Y}$ in this case gives the mean value when $X$ is either value.

Report results

$$\hat{y} = \underset{(se_{b_0})}{b_0} - \underset{(se_{b_1})}{b_1 X}, \quad R^2 = \_\_, \quad SER = \_\_$$

## ⑦ prediction using OLS

$(X, Y)$ are randomly drawn from the same population as $(X_i, Y_i)$

## ⑧ omitted variable bias

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(u_i - \bar{u})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} = \beta_1 + \frac{s_{xu}}{s_x^2}$$

$$= \beta_1 + \frac{s_{xu}}{s_x \cdot s_x} \cdot \frac{s_u}{s_u}$$

or

$$= \beta_1 + \frac{s_{xu}}{s_x s_u} \cdot \frac{s_u}{s_x}$$

$$\boxed{E(\hat{\beta}_1) = \beta_1 + \rho_{xu} \cdot \frac{\sigma_u}{\sigma_x}}$$

intuition: bias arises from the violation of $E(u_i | X_i) = 0$ is fundamentally because $u_i$ and $X_i$ are correlated

$E(u_i | X_i) = 0 \longrightarrow \rho_{u,x} = 0$, but converse untrue.

$E(u_i | X_i) = k \longrightarrow E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = \beta_0 + k$

### intuition behind $\rho_{xu}$ and omitted variable bias

1. Suppose we have an omitted variable $z_i$ and coefficient $\beta_z$.

$$y_i = \beta_0 + \beta_1 x_i + (\beta_z z_i + \varepsilon_i)$$

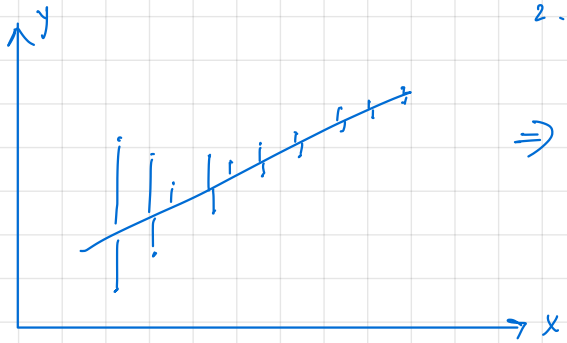
($\varepsilon_i$: true error; $\beta_z z_i$: omitted variable)

1.1 we have erroneously modelled it as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, so any error due to a change in $z$ is encapsulated in our $\hat{u}_i$ when we fit the model.

2. Suppose $\beta_z$ is negative. Suppose also that $x_i$ and $z_i$ are positively correlated. Suppose also that $\beta_1$ is positive

2.1 So as $x_i$ increases, $z_i$ increases, and because $\beta_1 > 0$ but $\beta_z < 0$ the correlation between

1. suppose $\rho_{xu} < 0$. As $x$ increases, $u$ decreases.
2. Then $E(\hat{\beta}_1) < \beta_1$ - ie. flatter or more negative than it really is




⇒ intuition:

$\rho_{xu}$ arises because

$\rho$

Notes

1. unit change → take note of whether it is applied to $Y$ or $X$ → $\hat{\beta}_1$ compensate
   ↳ scale

# Multiple regression

## ① omitted variable bias

↳ in a basic bivariate model, we assume that $y$ is determined by a single variable $X$. But there might be other variables that also determine $y$ that are correlated with $x$, muddling our study of the relationship between $y$ and $X$.

**When does it occur?**

1. the omitted variable is correlated with the included regressor(s)
2. the omitted variable is a determinant of the dependent variable

**Why do omitted variables cause bias?**

1. Let us simplify the expression of $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i - (x_i - \bar{x})\bar{y}}{\sum (x_i - \bar{x}) x_i - (x_i - \bar{x})\bar{x}}$$

2. Suppose the true model is $Y = \beta_0 + \beta_1 X + \underbrace{\beta_2 Z}_{\text{omitted variable}} + \varepsilon$

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i)}{\sum (X_i - \bar{X}) X_i}$$

$$= \beta_1 + \beta_2 \frac{\sum (X_i - \bar{X}) Z_i}{\sum (X_i - \bar{X}) X_i} + \frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X}) X_i}$$

$$= \beta_1 + \beta_2 \frac{\sum (X_i - \bar{X})(Z_i - \bar{Z})}{\sum (X_i - \bar{X})^2} + \frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2}$$

$$\downarrow$$

$$\frac{S_{XZ}}{S_X^2} = \frac{S_{XZ} \cdot S_Z}{S_X \cdot S_X \cdot S_Z} = \frac{S_{XZ}}{S_X S_Z} \cdot \frac{S_Z}{S_X}$$

3. Taking expectation, the omitted variable causes bias in the estimation of $\beta_1$.

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \rho_{XZ} \frac{\sigma_Z}{\sigma_X} + E\left( \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} E(u_i \mid X_i) \right)$$

## How do we observe omitted variable bias in our model ?

1. Let us simplify the expression of $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i - (x_i - \bar{x})\bar{y}}{\sum (x_i - \bar{x})^2}$$

2. Suppose the true model is $Y = \beta_0 + \beta_1 X + \underbrace{\beta_2 Z}_{\text{omitted variable}} + \varepsilon$

2.1 Then $\hat{\beta} = \frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + u_i)}{\sum (x_i - \bar{x})^2}$

$$= \beta_1 + \frac{\sum (x_i - \bar{x})(\beta_2 z_i + u_i)}{\sum (x_i - \bar{x})^2}$$

2.2 If we consider that in the model we do not consider $Z_i$, thus the $\beta_2 z_i + u_i$ term are represented as a single estimator of error, $\tilde{u} = \beta_2 z_i + u_i$

so $\hat{\beta} = \beta_1 + \frac{\sum (x_i - \bar{x})\tilde{u}_i}{\sum (x_i - \bar{x})^2} = \beta_1 + \frac{\sum (x_i - \bar{x})(\tilde{u}_i - \bar{\tilde{u}}_i)}{\sum (x_i - \bar{x})^2}$

3. We can express this in terms of correlation between $X_i$ and $\tilde{u}_i$

3.1 $\beta_1 + \frac{S_{x\tilde{u}}}{S_x^2} = \beta_1 + \frac{S_{x\tilde{u}}}{S_x^2} \cdot \frac{S_{\tilde{u}}}{S_{\tilde{u}}}$

$$= \beta_1 + \frac{S_{x\tilde{u}}}{S_{\tilde{u}} S_x} \cdot \frac{S_{\tilde{u}}}{S_x}$$

$$= \beta_1 + r_{x\tilde{u}} \cdot \frac{S_{\tilde{u}}}{S_x}$$

3.2 Taking expectation, we observe that the correlation between $X$ and $\tilde{u}$ reflects the omitted variable bias.

$$E(\hat{\beta}_1) = \beta_1 + \rho_{x\tilde{u}} \cdot \frac{\sigma_{\tilde{u}}}{\sigma_x}$$

## ② multiple regression model

↳ functional form: we posit that $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_K X_K + \varepsilon_i$

**interpretation**

↳ the model describes the relationship between $Y$ and regressors $X_1 \dots X_K$ on average.

i.e. $E(Y \mid X_1, X_2 \dots X_K) = \beta_0 + \beta_1 X_1 \cdots \beta_K X_K$

↳ the coefficient $\beta_j$ is the effect on $Y$ of a unit change in $X_j$, holding all other regressors constant

**OLS estimator**

↳ just as in the bivariate case, we try to minimize the total sum of squared errors.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \dots + \hat{\beta}_K X_K$$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

$$S = \sum_{i=1}^{n} \hat{u}_i^2$$

$$\left.\begin{array}{l} \frac{\partial S}{\partial \beta_0} = 0 \\ \frac{\partial S}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial S}{\partial \beta_K} = 0 \end{array}\right\} \Rightarrow \text{solve in linear algebra and calculus to get closed form solutions}$$

## ③ concepts of multicollinearity

1) perfect multicollinearity

↳ regressors are said to exhibit perfect multicollinearity when one of the regressors is a perfect linear function of other regressors. eg. $X_1 = kX_2$

2) imperfect collinearity

↳ independent concept from perfect collinearity

↳ regressors are said to exhibit imperfect multicollinearity when they are highly (but not perfectly) correlated

## ④ Sampling distributions in multivariable regression

### 1) Least-square assumptions

**the standard three**

1. $E(u_i \mid X_{1i}, X_{2i} \dots X_{ki}) = 0$
2. $(X_{1i}, X_{2i} \dots X_{ki})_j$, $j \in (1, \dots n)$ are identically & independently distributed
3. large outliers are unlikely

**no perfect collinearity** → if present, likely a logical error in dataset / choice of variables

↳ when perfect multicollinearity is present, then it is as if you have >1 instance of a given variable. eg. $X_2 = X_1 \times 100$ (percentages)

↳ in trying to estimate $\beta_j$, we intuitively hold all other X constant. Given perfect multicollinearity, it is as if trying to simultaneously hold constant & vary a variable

↳ mathematically, it leads to a divide by zero error in the OLS estimator

### 2) estimator distributions

in general, if LSA assumptions hold, sample is large

↳ $\hat{\beta}_j$ is unbiased & consistent: $E(\hat{\beta}_j) = \beta_j$, $Var(\hat{\beta}_j) \propto \frac{1}{n}$

↳ $\hat{\beta}_j$ are well approximated by multivariate normal distributions

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}) \text{ approximately.}$$

↳ because we don't know $\sigma^2_{\hat{\beta}_j}$, we estimate with $SE(\hat{\beta}_j)$

$$\hat{\beta}_j \sim t_{n-1}(\beta_j, SE(\hat{\beta}_j))$$

**imperfect multicollinearity**

↳ when two or more regressors are highly correlated, then sampling variance of these variables become very large.

↳ intuition: high correlation, move together. In OLS estimation, in eping all other variables constant, those regressors have a very small independent spread

$$\Rightarrow Var(\hat{\beta}_j) \propto \frac{1}{Var(X_j)} \Rightarrow \text{high sampling variance}$$

↳ $\hat{\beta}_j$ will still be unbiased and consistent, though

## ⑤ Binary multivariate regression

1) dummy variable trap
↳ if there are G binary (ie. mutually exclusive) variables, if there is an intercept in the regression, and all G regressors are included, it will fail because of perfect collinearity

intuition
↳ imagine that $\beta_0$ is coefficient of a variable $X_0 = 1$.
↳ G binary variables are mutually exclusive $\Rightarrow$ $X_j = 1 - X_1 - X_2 \dots$ (if any is 1, then 0)
$\Rightarrow$ perfect collinearity between $X_0$ and $X_j$

$\rightarrow$ solution: exclude $\beta_0$, or regress only G-1 variables

2) interpretation of coefficients
1. Suppose K categories, inclusion of intercept & removal of 1 regressor

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 \dots \hat{\beta}_{K-1} X_{K-1i} + \hat{u}_i$$

↳ $E(\hat{y} \mid X_1 = 0, X_2 = 0 \dots) = \beta_0 \Rightarrow \hat{\beta}_0$ estimates expected value of y given $K^{th}$ category
↳ $E(\hat{y} \mid X_1 = 0, \dots X_j = 1, \dots) = \beta_0 + \beta_j \Rightarrow \hat{\beta}_j$ estimates expected difference in y between $j^{th}$ & $K^{th}$ (base) categories

2. Suppose exclusion of intercept and K regressors

$$\hat{Y}_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \dots \hat{\beta}_K X_{Ki} + \hat{u}_i$$

↳ $E(\hat{y}_i \mid X_1 = 0 \dots X_j = 1) = \hat{\beta}_j \Rightarrow \hat{\beta}_j$ estimates expected value of y given one of these categories

## ⑥ Hypothesis testing

### 1) single variable tests

same as bivariate case: $t_{n-1} = \dfrac{\hat{b_j} - b_0}{se(\hat{b_j})}$

### 2) joint hypothesis testing

↳ joint hypothesis: a hypothesis that imposes two or more restrictions on the regression coefficients

$H_0$: $\beta_j = k_j \wedge \underbrace{\beta_m = k_m}_{\text{1 restriction}}, \ldots$ for a total of $q$ restrictions

$H_1$: at least 1 of $q$ restrictions do not hold

**why do $q$ $t$ tests not work?**

↳ if we did $q$ $t$-tests, we fail to model a distribution that considers simultaneous restrictions $\Longrightarrow$ akin to binomial distribution rather than modelling from a single joint distribution

**The F statistic**

↳ we need some statistic with a known distribution that models the joint probability distribution of $k$ random variables, that could be uncorrelated with one another

$\Rightarrow$ we can use the F-statistic

1. The F statistic models the ratio of two $\chi^2$ variables.

$$F \sim \frac{\frac{\chi^2_m}{m}}{\frac{\chi^2_n}{n}}$$

2. we can cleverly aggregate our $\hat{\beta}_j$ distributions as a single $\chi^2$ random variable.

2.1 $\hat{\beta}_j \longrightarrow N(\beta_j, Var(\hat{\beta}_j))$ as $n \to \infty$

2.2 if we square them: $\hat{\beta}_j^2 \sim \chi^2_1$ approximately

2.3 if we sum them and divide by number of restrictions $q$: $\frac{\chi^2_q}{q}$

2.4 by CLT: $\frac{\chi^2_m}{m} \to 1$ as $m \to \infty$. so $\frac{\chi^2}{q} = \frac{\frac{\chi^2_q}{q}}{\frac{\chi^2_m}{m}}$, $m \to \infty$

$\Rightarrow$ F statistic

## the F-test

1. $H_0$, $H_1$, $q$ joint hypotheses
2. calculate $q$ $t$-statistics.
3. $f = \frac{\sum_{1}^{q} t_i^2}{q} \Big/ \frac{\chi^2_n}{n}$ - normally we assume $n$ is large $\approx \infty$
4. calculate p-value

note that even if $q=1$, the f-test is valid. But by convention, t test for single variable, else F test.

3) single restrictions involving multiple coefficients eg. theory suggests $\beta_1 = \beta_2$

↳ not a joint hypothesis test w multiple restrictions.

↳ we can do an F test (abstracted by package)

↳ we can also transform the problem into a single variable test by forming desired r/s as a coefficient eg. $H_0: \beta_1 = \beta_2$

1. suppose we believe $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
2. then we can rewrite as $Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + \varepsilon_i$

$= \beta_0 + \gamma X_{1i} + \beta_2 V_i$, $V = X_{1i} + X_{2i}$

3. then we can regress and test that $H_0: \gamma = \beta_1 - \beta_2 = 0$ using a t-test.

## ⑦ measures of fit

**standard error of regression**

↳ estimates standard deviation of error term $u_i$ ⇒ spread of $Y$ about $\hat{Y}$

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}, \quad s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1}$$

$\bar{\hat{u}} = \frac{1}{n}\sum \hat{u}_i = 0$ ; $k$ = no. of regressors

**$R^2$**

↳ estimates fraction of sample variance of $Y_i$ predicted by regressors.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

⇒ but $R^2$ never decreases when you increase no. of regressors, so it inflates how well model is fitting the data

↳ intuition = OLS estimator minimises TSS.

ie. $\min \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} \dots)^2$

any time you add a new $\hat{\beta}_j$, so long as $b_j \neq 0$, by definition TSS will fall, since minimisation.

⇒ we need a metric that punishes model complexity

**Adjusted $\bar{R}^2$**

↳ $R^2$ that deflates as more regressors are included

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}$$

$k$ = no. of regressors

↳ things to note about $\bar{R}^2$ :

1. $\bar{R}^2 < R^2$, because $\frac{n-1}{n-k-1} > 1$
2. Adding a regressor has 2 effects on $\bar{R}^2$ → SSR ↓ due to complexity, ↑ $\bar{R}^2$; → $\frac{n-1}{n-k-1}$ ↓, ↓ $\bar{R}^2$
3. $\bar{R}^2$ can be negative. occurs when regressors reduce SSR by such a small amount that it fails to offset $\frac{n-1}{n-k-1}$

⇒ rule of thumb: negative, fails to explain any variation

**RMSE**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

## Notes on $R^2$

1. what $R^2$ tells you
   - ↳ how much of variability is explained by model
   - ↳ high $R^2$ → regressors are good are predicting $Y_i$

2. what $R^2$ does not tell you
   - ↳ does not say anything about $\rho_{ux}$ and omitted variable bias
   - ↳ does not say anything about causality → RCTs and hypothesis tests
   - ↳ does not say anything about statistical significance → hypothesis tests

## ⑧ model design

↳ when estimating a causal effect, we want to include variables of interest and control variables (to avoid omitted variable bias)

↳ all we need is that estimators of coefficients of variables of interest are unbiased and consistent. It does not matter if estimators of coefficients for control variables are biased

### relaxing LSA #1

↳ to study variables of interest (i.e. $\hat{\beta}_j$ unbiased, consistent), we simply need to ensure that $u_i$ is uncorrelated with them.

↳ suppose we have $X_1, X_2 \dots X_K$ variables of interest & $W_1, W_2 \dots W_j$ control variables.

⇒ to ensure $\hat{\beta}_{X1} \dots \hat{\beta}_{XK}$ unbiased, we simply need $E(u_i \mid X_1 \dots X_K, W_1 \dots W_j)$

$= E(u_i \mid W_1 \dots W_j)$

## A general approach

1. Specify a base specification (set of variables)
   - ↳ should contain the variables of interest & a set of control variables suggested by judgement or theory

2. Specify a range of plausible alternative specifications
   - ↳ other possible control variables you can use to observe if your coefficients change when you include them (they shouldn't, at least within CI across diff. specifications)
   - ↳ can also be alternative proxies for control variables
     - e.g. % on income support } SES
     - % on food support }

3. run regressions
   - ↳ pay attention to whether estimates of coefficients of interest, fitting metrics, numerically similar across the different specifications
     - within CI of previous regressions
     - yes → estimates are generally reliable
     - no → probably suffers from omitted variable bias

# Non-linear regression

## ① non-linear models

↳ in general, we can model the <u>average</u> relationship between variables as some generic function and use different methods to estimate them

$$Y = f(\{x_{ij}\}) + u_i$$

$$\mathbb{E}(Y|\{x_{ij}\}) = f(\{x_{ij}\})$$

### effect on Y of a change in X

↳ in linear specifications, coefficients are intuitively interpretable as $\frac{\Delta Y}{\Delta X_i}$

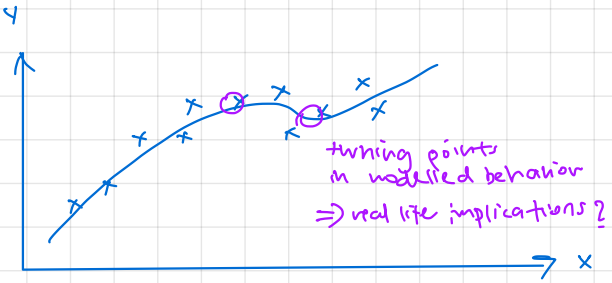
↳ in non-linear specifications, the interpretation depends on functional form

↳ more generally, when we fit non-linear models, we should consider the <u>expected</u> change in Y <u>given</u> some change in $X_i$ from one value to another

$$\Delta\hat{y} = \hat{f}(x_1 + \Delta x_1, x_2, x_3 \dots) - \hat{f}(x_1, x_2 \dots)$$

⇒ our fitted function estimates expected <u>change</u> at each value of independent variable

↳ shape, gradients & predictions at ranges of X are informative about behaviour trying to study, individual coefficients less so




### Standard errors of estimated effects

↳ in the linear case, $\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_i x_i \dots$

$$\Rightarrow \Delta\hat{y} = \hat{\beta}_i(\Delta x_i) \Rightarrow SE(\Delta\hat{y}) : SE(\hat{\beta}_i)$$

↳ in the non-linear case, it depends on the functional form.

eg. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \Rightarrow \Delta\hat{y} = \hat{\beta}_1(\Delta x) + \hat{\beta}_2((x+\Delta x)^2 - x^2))$

$$\Rightarrow SE(\Delta\hat{y}) = SE(\underbrace{\Delta x(\hat{\beta}_1) + ((x+\Delta x)^2 - x^2)(\hat{\beta}_2)}_{\text{depends on functional form}})$$

## ② A general approach to fitting non-linear models

1. use economic theory to identify a shape between variables ⇒ choose functional form(s)
   eg. gradient, turning points, plateaus, etc.

2. estimate parameters using OLS or some other techniques
   ↳ note that in many cases we do a non-linear transformation on the input variables, and expanding them allows us to use multi-variable OLS estimation on them

3. Evaluate the model
   - 3.1 F test : non-linear vs. linear model
   - 3.2 T-test : determine degree of non-linearity
   - 3.3 graph it : see how well it fits the data → residuals, data about line

4. compare statistically significant models
   ↳ when using $\bar{R}^2$ to compare models, make sure the basis is same eg. $\ln(Y) - \ln(X)$ } ok
   vs. $\ln(Y) - X$

   $\ln(Y) - X$ } comparing
   vs $Y - X$ } diff things!
   ⇒ $\bar{R}^2$'s capture of variation not comparable

## ③ Polynomial models

1) functional form and OLS estimation

$$Y_j = \beta_0 + \beta_{ij1} X_{ij} + \beta_{ij2} X_{ij}^2 \dots \beta_{ijr} X_{ij}^{(r)} + u_j$$

(r) to the $r^{th}$ degree

↳ observe that polynomial regression is identical to multiple regression, except that instead of independent variables, the variables are powers
⇒ we can run OLS estimation as in multiple regression

⇒ interpret as we do for non-linear models

2) evaluating and choosing degree

1. based on theory, pick a maximum value of r and fit a model
2. test for non-linearity : if all $X_i^r$, $r>1$ terms are needed using F-test
3. use a t-test to test $X^r$ if single variable, else F-test to test $X_i^r$.
   if you do not reject it, remove power r terms and run the regression $\bar{w}$ r-1
   ↳ remove to avoid imperfect multicollinearity, since statistically insignificant terms will move closely w others to fit noise
4. compare all statistically significant models w $\bar{R}^2$

## ④ logarithmic models

### 1) properties of logarithms

$$\ln\left(\frac{a}{b}\right) = \ln a - \ln b$$

$$\ln(ab) = \ln a + \ln b$$

$$\ln(a^b) = b \ln a$$

### logarithms & percentages

$$\ln(x+\Delta x) - \ln(x) = \ln \frac{x+\Delta x}{x}$$

$$= \ln\left(1 + \frac{\Delta x}{x}\right)$$

$$\approx \frac{\Delta x}{x} \quad \text{when } \Delta x \text{ is small}$$

### 2) three log models

| Case | Population Regression Model | Interpretation of $\beta_1$ |
|---|---|---|
| I. linear-log | $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ | A 1% change in $X$ is associated with a change in $Y$ of $0.01\beta_1$ |
| II. log-linear | $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$ | A change in $X$ by one unit ($\Delta X = 1$) is associated with a $100\beta_1$% change in $Y$ |
| III. log-log | $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ | A 1% change in $X$ is associated with a $\beta_1$% change in $Y$ |

### linear-log model

$$Y_j = \beta_0 + \sum \beta_i \ln(X_{ij}) + u_j$$

$$Y_i + \Delta Y_i = \beta_0 + \beta_1 \ln(X_i + \Delta X)$$

$$\Delta Y_i = \beta_1 [\ln(X_i + \Delta X) - \ln(X)] \approx \beta_1 \left(\frac{\Delta X_i}{X_i}\right) \approx \beta (0.01) k$$

⇒ A k% change in X predict an average change in Y by $\beta_1 \times k$%

### log-linear model

$$\ln(Y_j) = \beta_0 + \sum \beta_i X_{ij} + u_j$$

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X)$$

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X \approx \frac{\Delta Y}{Y}$$

⇒ A unit change in X is associated with a $\beta_1$% change in Y on average

### log-log model

$$\ln(Y_j) = \beta_0 + \sum \beta_i \ln(X_{ij}) + u_j$$

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$$

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 [\ln(X + \Delta X) - \ln(X)]$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$$

⇒ a k% change in X is associated with an average $\beta_1 \cdot k$% change in Y

### predicted values of Y when Y is in log

⇒ taking the exponent is biased

e.g. $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$

$$\Leftrightarrow Y_i = e^{\beta_0 + \beta_1 X_i} \cdot e^{u_i}$$

scaled bias

$$E(Y_i | X_i) = E(e^{\beta_0 + \beta_1 X_i} | X_i) \cdot E(e^{u_i} | X_i)$$

⇒ even if $E(u_i | X_i) = 0$, $E(e^{u_i} | X_i) \neq 1$ in general

3) fitting logarithmic models

↳ observe that as in the polynomial case, we are simply doing a transformation of the variables, and they effectively take the form of multiple regression ⇒ OLS estimation

## (5) functions that are non-linear in their parameters

↳ so far, we have expressed non-linearity by transforming the input variables, but the estimated parameters $\beta_i$ are still linear. we like them because they are conveniently decomposable into a linear function and $\beta_i$ can be estimated by OLS

⇒ how about non-linear parameters?

### general functions that are non-linear in parameters

$$y_i = f(x_1, x_2 \dots, \underbrace{\beta_0, \beta_1 \dots}_{\text{parameters also inside}}) + u_i$$

### non-linear OLS estimation

$$Q = \sum_i [y_i - f(x_1, x_2 \dots, b_0, \dots b_1)]^2 \Rightarrow \text{minimize}$$

↳ in general, if there is no close form solution, then we estimate by numerical solving (guessing until minimum) ⇒ like gradient descent

↳ if LSAs hold, the estimators of $\hat{\beta}_i$ are consistent and normally distributed with a large dataset, so we can use t-tests to test their significance

---

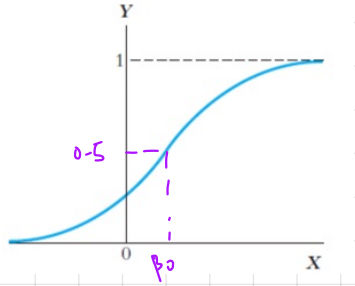
### logistic regression

$$y_j = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i x_{ij})}} + u_i$$

$\beta_0$: shifts laterally

↳ properties:

1. $y \to 1$ as $x \to \infty$
   $y \to 0$ as $x \to -\infty$
2. gradient steepest at $x = 0$
3. as $x \to \infty$ or $-\infty$, $\frac{dy}{dx} \to 0$



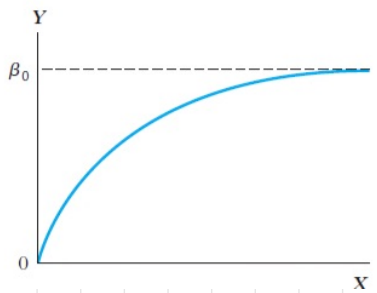

### negative exponential regression

$$y_i = \beta_0 [1 - e^{-\beta_1 (x_i - \beta_2)}]$$

two parameters per regressor

↳ properties:

1. $y \to \beta_0$ as $x \to \infty$
2. $\frac{dy}{dx} > 0$, $\frac{d^2y}{dx^2} < 0$
3. undefined for negative values of $x$



## ⑥ Interaction between variables

↳ suppose variables modify each others' effects on $Y$ at different values
eg. at high student-teacher ratio, effects of SES more pronounced

⇒ how can we model this? we can introduce interaction terms and estimate their coefficients between variables we think will interact

↳ is there an interaction?
⇒ we do a t-test on the interaction terms' coefficients)

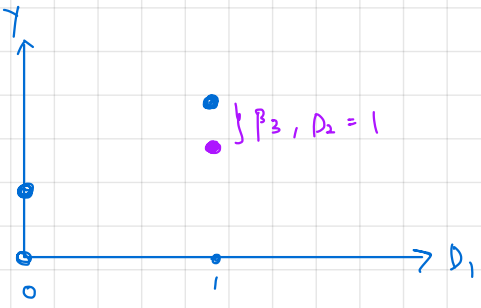
### binary-binary

$$Y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + u_i$$

$$E(Y_i \mid D_1 = 1, D_2 = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$$

$$E(Y_i \mid D_1 = 0, D_2 = d_2) = \beta_0 + \beta_2 d_2$$

Expected difference when $D_1 = 1$ vs. $D_1 = 0$ ⇒ $\beta_1 + \beta_3 d_2$




⇒ effect on $Y$ of change in $D_1$ now depends on $D_2$
⇒ $\beta_3$ is the difference in effect on $Y$ of $D_1$ when $D_2 = 1$ vs. $D_2 = 0$

---

### continuous-continuous

↳ effectively finer granularity of binary-binary case

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$\beta_3 X_1 X_2$: extra effect from changing both $X_1$ or $X_2$

$\beta_1$: ↳ effect of $X_1$ holding $X_2$ constant

$$\frac{dy}{dX_1} = \beta_1 + \beta_3 X_2$$

⇒ $\beta_3 \cdot X_2$ : the expected difference in effect of a unit change in $X_1$ given $X_2 = x_2$

## Continuous – binary

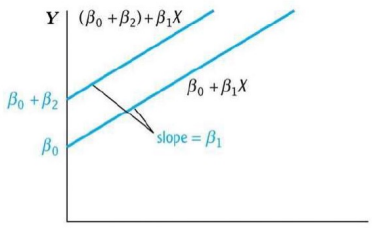
↳ 1. test if they have diff. intercept ie. $\beta_2 = 0$

↳ inclusion of binary term allows for different offset

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i$$

$D_i = 0$: $y_i = \beta_0 + \beta_1 x_i + u_i$

$D_i = 1$: $y_i = (\beta_0 + \beta_2) + \beta_1 x_i + u_i$



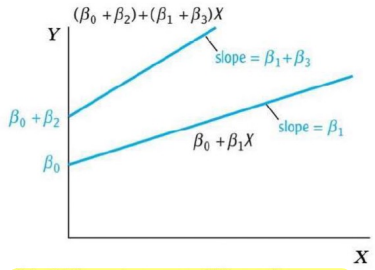

(a) Different intercepts, same slope

↳ inclusion of continuous – binary interaction allows the effect of change in X to be different (ie. different gradient) ↳ 2. test if they have diff gradient ie. $\beta_3 = 0$ (t-test)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + u_i$$

if $D_i = 1$: $y_i = \beta_0 + (\beta_1 + \beta_3) x_i + \beta_2$

if $D_i = 0$: $y_i = \beta_0 + \beta_1 x_i$




(b) Different intercepts, different slopes

⇒ $\beta_3$ - difference in effect of unit change in X, for observations where $D_i = 1$ vs. $D_i = 0$

---

Note:

1. percentages in decimal. But if % change is large, log approximation will be bad
2.

# Assessing regression

## ① Validity of a study

**internal validity** — for a study to be internally valid:

1. the estimator of the causal effect of interest should be unbiased and consistent
2. hypothesis tests have the correct significance level; standard errors are correctly estimated (hetero vs. homoskedasticity)

**external validity** — A study is externally valid if its inferences can be generalised to other populations & settings

## ② Threats to internal validity

**omitted variable bias** $E(u_i|x_i) \neq 0$

1. the omitted variable is correlated with the included regressor(s)
2. the omitted variable is a determinant of the dependent variable

⇒ solutions: include omitted variables, use panel data, instrumental variables or RCT

**incorrect functional form**

↳ bias can arise if the wrong functional form is used. we will observe trends in the residuals. solution: fit better models

**measurement error**

**errors in $x_i$**

↳ in general, $\hat{\beta}_i$ will be biased and inconsistent

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (x_i - \tilde{x}_i) + u_i]$$

$\tilde{X}_i$: observed data; $v_i$

⇒ check corr. between $\tilde{x}_i$ and $v_i$ $= \dfrac{cov(\tilde{x}_i, v_i)}{\sigma_{\tilde{x}_i}, \sigma_{u_i}}$

1. $cov(\tilde{x}_i, v_i) = 0$ iff $E(\quad | \quad) = 0$, then $\hat{\beta}_1$ unbiased & consistent
2. $cov(\tilde{x}_i, v_i) = cov(\tilde{x}_i, \beta x_i - \beta \tilde{x}_i + u_i)$

$= cov(\tilde{x}_i, \beta x_i) - cov(\tilde{x}_i, \beta \tilde{x}_i) + cov(\tilde{x}_i, u_i)$

$= \beta (cov(\tilde{x}_i, \beta x_i) - \sigma_{\tilde{x}_i}^2)$ — $cov(\tilde{x}_i, u_i) = 0$, since $E(u_i|x_i) = 0$

in general, $\neq 0$, so $E(v_i|x_i) \neq 0$

**errors in $y_i$**

↳ in general, no bias to $\hat{\beta}_i$, but $Var(\hat{\beta}_i)$ will be larger

$$Y_i = \beta_0 + \beta_1 x_i + u_i + w_i$$

⇒ so long as measurement error on $Y_i$ is random, $E(w_i|x_i) = 0$, so $E(v_i|x_i) = E(w_i|x_i) + E(u_i|x_i) = 0$

**sample selection bias** — in some cases, SRS is thwarted because the dependent variable is only observed for a restricted sample eg. education on earnings — we only study the sub-group that is employed → we fail to study effect of education earnings had that group worked.

**simultaneous causality bias**

↳ if $x \to y$ but $y \to x$, then $\hat{\beta}_1$ will be picking up the simultaneous effects
⇒ OLS estimator of $\hat{\beta}_1$ will be biased and inconsistent

Suppose: $Y_i = \beta_0 + \beta_1 X_i + u_i$ ⇒ we are fitting this

$X_i = \gamma_0 + \gamma_1 Y_i + v_i$

$\gamma_1 > 0$ : $u_i \uparrow, Y_i \uparrow, X_i \uparrow$ ; $\rho_{u_i, X_i} > 0$

$\gamma_1 < 0$ : $u_i \uparrow, Y_i \uparrow, X_i \downarrow$ ; $\rho_{u_i, X_i} < 0$

$E(u_i | X_i) \neq 0$

more formally

$cov(X_i, u_i) = 0$ iff. $E(u_i | X_i) = 0$

1. $cov(X_i, u_i) = cov(\gamma_0 + \gamma_1 Y_i + v_i, u_i)$

$= cov(\gamma_0, u_i) + cov(\gamma_1 Y_i, u_i) + cov(v_i, u_i)$ (first two = 0)

$= cov(\gamma_1(\beta_0 + \beta_1 X_i + u_i), u_i)$

$= cov(\gamma_1 \beta_0, u_i) + cov(\gamma \beta_1 X_i, u_i) + cov(\gamma_1 u_i, u_i)$ (first = 0)

$= \gamma_1 \beta_1 cov(X_i, u_i) + \gamma_1 \sigma_u^2$ ← $cov(u_i, u_i) = var(u_i)$

$(1 - \gamma_1 \beta_1) cov(X_i, u_i) = \gamma_1 \sigma_u^2 \Rightarrow cov(X_i, u_i) = \frac{\gamma_1 \sigma_u^2}{1 - \gamma_1 \beta_1} \neq 0$

⇒ $E(u_i | X_i) \neq 0$

↳ solution: RCT

③ threats to external validity

↳ whether one can generalize results found in a study to other populations & settings depend on how similar the populations & settings are
eg. results from canada → can apply to SG?

↳ practical significance, beyond statistical significance!

Notes

1. comment — statistical, practical significance, bias { OVB; simultaneous causality; errors in measurement

# Regression with panel data

## ① panel data and formulation

**panel data** data for $n$ different entities observed at $T$ time periods.

notation as $\left( \{X_{jit}\}_{j \in [1,v]}, Y_{it} \right)$.

$v$ variables of entity $i$ at time $t$

↳ a balanced panel has all observations — variables are observed for each entity and each time period.

### Regression and OVB

↳ suppose we want to study the causal relationship between $Y$ and $\{X_j\}_{j \in [1,v]}$, and we have a set of panel data.

↳ simplistically, we can model as $Y_{it} = \beta_0 + \beta_1 X_{it} + \dots + \varepsilon$

↳ but it is conceivable to think that there is omitted variable bias for unobserved variables. The dimensions of time and entity mean that we could split these variables.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \dots + A_1 Z_i + \Gamma W_t + \varepsilon$$

- $X_{it}$: variable of interest changing in space and time
- $Z_i$: omitted variable that only changes across entities (e.g. gender)
- $W_t$: omitted variable that only changes across time (e.g. GST rate)

### entity and time fixed effects

↳ variables in panel data inherently vary across entities and time. In studying whether some set of $X$ affects $Y$, we can simply abstract out the effects of "irrelevant" effects specific to each entity and time period as **fixed effects**

$$Y_{it} = \alpha_i + \gamma_t + \beta_1 X_{it} + \dots + \varepsilon$$

- $\alpha_i$: entity specific baseline
- $\gamma_t$: time specific baseline

## ② fixed effects regression

### T=2, Take difference

$$y_{i1} = \alpha_i + \lambda_1 + \beta_1 X_{i1} + \varepsilon_{i1}$$

$$y_{i2} = \alpha_i + \lambda_2 + \beta_1 X_{i2} + \varepsilon_{i2}$$

$$y_{i2} - y_{i1} = \Delta y_{it}$$

$$\Delta y_{it} = \beta_0 + \beta_1 \Delta X_{it} + \Delta \varepsilon_{it}$$

$$(\lambda_2 - \lambda_1) + \beta_1 (X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

### using binary variables

- model has $T \times n$ different intercepts $(\lambda_t + \alpha_i)$ for each entity-time period pair.
- to avoid perfect collinearity, we can model intercepts as $(n-1)$ binary variables, one for each entity and another $(T-1)$ binary variables, one for each period.

$$y_{it} = \alpha_i + \lambda_t + \beta_1 X_{it} + \varepsilon_{it}$$

$$\gamma_i = \alpha_i - \alpha_1$$

$$\gamma_t = \lambda_t - \lambda_1$$

$$\beta_0 = (\alpha_1 + \lambda_1)$$

$$y_{it} = \beta_0 + \beta_1 X_{it} + \ldots + \gamma_2 D_{2i} + \ldots + \gamma_n D_{ni} + \delta_2 T_{2t} + \ldots + \delta_T T_{Tt}$$

$\gamma_n$: offset for entity $n$; $D_{ni}$: is $i$ the $n^{th}$ entity

$\delta_T$: offset for time $T$; $T_{Tt}$: is $t$ the $T^{th}$ period

### de-meaning

$$y_{it} = \alpha_i + \gamma_t + \beta_1 X_{it} + \varepsilon_{it}$$

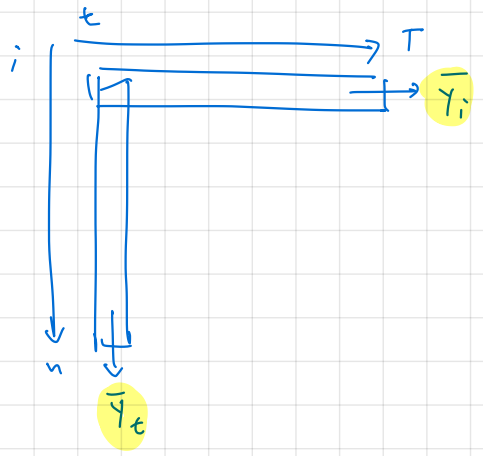
$$\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t$$

$$= \cancel{\alpha_i + \lambda_t} + \beta_1 X_{it} + \varepsilon_{it}$$

$$- (\cancel{\alpha_i} + \beta_1 \bar{X}_i + \bar{\varepsilon}_i)$$ time effect lost, assuming avg. = 0

$$- (\cancel{\lambda_t} + \beta_1 \bar{X}_t + \bar{\varepsilon}_t)$$ entity effect lost, assuming avg. = 0

$$= \beta_1 (\underbrace{X_{it} - \bar{X}_i - \bar{X}_t}_{\tilde{X}_{it}}) + (\underbrace{\varepsilon_{it} - \bar{\varepsilon}_i - \bar{\varepsilon}_t}_{\tilde{\varepsilon}_{it}})$$




$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0 \qquad \frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$$

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_t = \lambda_t \qquad \frac{1}{T}\sum_{t=1}^{T} \lambda_t = 0$$

estimated directly by mean if no other regressors

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} = \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it} + \cancel{\frac{1}{T}\sum_{t=1}^{T} \alpha_i}^{\alpha_i} + \cancel{\frac{1}{T}\sum_{t=1}^{T} \lambda_t}^{0}$$

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it} = \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_{it} + \frac{1}{n}\sum_{i=1}^{n} u_i + \cancel{\frac{1}{n}\sum_{i=1}^{n} \alpha_i}^{0} + \cancel{\frac{1}{n}\sum_{i=1}^{n} \lambda_t}^{\lambda_t}$$

## ③ estimation and estimator distribution

### OLS estimation and estimator distributions

↳ we formulate the model as one of the three forms in fixed effects regression and use OLS estimation to estimate $\hat{\beta}_j$.

↳ Here, we describe the fixed effects estimator $\hat{\beta}_j$ is the OLS estimator obtained using demeaning.

↳ recall that for cross sectional data:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_j x_j + \cdots))^2$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = 0 \quad , \quad \frac{\partial Q}{\partial \hat{\beta}_j} = 0$$

$$\Rightarrow \hat{\beta}_j = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

$\hat{\beta}_j^{\text{before after}} = \sum_{i=1}^{n} \frac{(x_{i2} - x_{i1})(y_{i2} - y_{i1})}{(x_{i2} - x_{i1})^2}$

replace $x_{ij}$ by $\tilde{x}_{ijt} = x_{ijt} - \bar{x}_{ij} - \bar{x}_{jt}$

$y_i$ by $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t$

$$\Rightarrow \hat{\beta}_j = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ijt} \tilde{y}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ijt}^2}$$

$$Q_x = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{it}^2 \qquad n_i = \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{u}_{it}$$

$\sqrt{nT}(\hat{\beta}_1 - \beta_1) \sim N(0, \frac{Var(n)}{Q_x^2})$ approximately as $n \to \infty$

$$E(\hat{\beta}_j) = \beta_1 + \frac{\frac{1}{n} \cdot \frac{1}{T} \cdot \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ijt} E(u_{it} | x_{ijt})}{\frac{1}{n} \cdot \frac{1}{T} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{ijt}^2}$$

$$\lim_{n \to \infty} Var(\hat{\beta}_j) = \frac{1}{nT} \frac{\sigma_{n_i}^2}{Q_x^2}$$

$$Var(n_i) = var\left(\frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{u}_{it}\right) = \frac{1}{T} var(\tilde{x}_{i1} \tilde{u}_{i1} + \tilde{x}_{i2} \tilde{u}_{i2} \cdots)$$

$$= \frac{1}{T} \left( var(\tilde{x}_{i1} \tilde{u}_{i1}) + Var(\tilde{x}_{i2} \tilde{u}_{i2}) \cdots + 2\, cov(\tilde{x}_{i1} \tilde{u}_{i1}, \tilde{x}_{i2} \tilde{u}_{i2}) \cdots \right)$$

## Serial correlation

The most common form of serial correlation is first-order serial correlation, in which the error in time $t$ is correlated to the previous $(t-1)$ period error.

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1$$

$\rho > 1$ : positive autocorrelation → error terms tend to have same sign in related periods

$\rho < 1$ : negative autocorrelation → error terms tend to have different sign in related periods

## Heteroskedasticity robust vs. clustered SE

↳ recall that the heteroskedasticity robust estimator of variance is as:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} \qquad \widehat{Var}(\hat{\beta}_1) = \frac{1}{n} \sum \frac{Var((X_i - \bar{X}) u_i)}{Var(X_i)^2}$$

↳ above, we have derived that in fixed effects regression:

$$\lim_{n\to\infty} Var(\hat{\beta}_j) = \frac{1}{nT} \frac{\sigma_{u_i}^2}{Q_x^2}$$

$$Q_x = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2$$

$$Var(u_i) = Var\left( \frac{1}{T} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{u}_{it} \right) = \frac{1}{T} Var\left( \tilde{X}_{i1} \tilde{u}_{i1} + \tilde{X}_{i2} \tilde{u}_{i2} \dots \right)$$

$$= \frac{1}{T} \left( Var(\tilde{X}_{i1} \tilde{u}_{i1}) + Var(\tilde{X}_{i2} \tilde{u}_{i2}) \dots + 2\,cov(\tilde{X}_{i1} \tilde{u}_{i1}, \tilde{X}_{i2} \tilde{u}_{i2}) \dots \right)$$

⇒ heteroskedasticity robust estimator does not take into account of correlations in time. So we need one that does.

## Clustered SE

$$SE_{clustered}(\hat{\beta}) = \sqrt{\frac{1}{nT} \frac{\hat{\sigma}_{\hat{u}}^2}{\hat{Q}_x^2}}$$

$$\hat{\sigma}_{\hat{u}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( \underbrace{\frac{1}{T} \sum_{t=1}^{T} \tilde{X}_{it} \hat{u}_{it}}_{\hat{w}_i} - \left( \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \right) \right)$$

works out to 0 assuming $u_i$ and $X_i$ uncorr.

clustering by state

⇒ i.e. for one $i$, avg. across all its time, because entity likely dependent across time, but independent across state

## assumptions of fixed effects regression

1. $E(u_{it} \mid X_{i1}, X_{i2} \dots) = 0$ for each entity, $u_{it}$ has mean 0 given a specific entity effect and the history of $x_j$ for that state. i.e. no OVB

2. $(X_{i1}, X_{i2} \dots X_{iT})$ are i.i.d. entity tuples are independently distributed. i.e. collection of observations for a given entity has no bearing on that of another.

3. $(X_{it}, Y_{it})$ have finite $4^{th}$ moments. no extreme outliers

4. No perfect multicolinearity

5. $corr(u_{it}, u_{is} \mid X_{ijt}, X_{ijs}, \alpha_i) = 0$ for $t \neq s$ given some variable $x_j$, the error terms are uncorrelated over time within a state. i.e. no serial correlation
   - ↳ if holds, can use heteroskedasticity robust SE
   - ↳ also can if just 2 time periods
   - ↳ else use clustered SE

# Regression with limited dependent variable

## ① binary dependent variables

1) **limited dependent variables** dependent variables with limited range . eg. binary

2) regression with binary variables

↳ what does it mean to fit a line to a dependent variable that can only take on only two values, 0 and 1? ⇒ conditional probability!

↳ interpret the regression as modelling the probability that the dependent variable = 1.

model $f(\{X_j\})$ predicts $E(Y \mid \{X_j\})$

$$E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\})(1) + P(Y=0 \mid \{X_j\})(0)$$

---

## ② linear probability model

↳ use of a linear model to predict probability

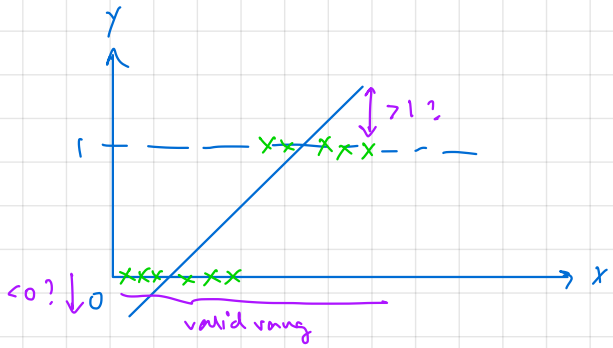
**functional form** $E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\}) = \beta_0 + \sum_j \beta_j X_j$

**estimation** OLS

**estimator distributions** same as that of OLS

**inference** same as that of OLS

**shortcomings**




1. probabilities cannot exceed 1 or be negative.
2. effects of regressor(s) on probability likely to be non-linear, have saturation, etc.

---

## ③ probit & logit regression

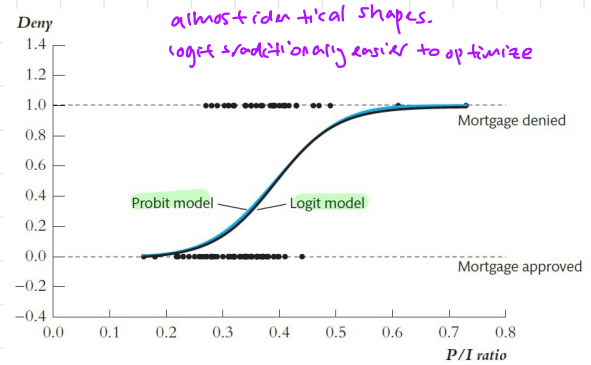
$z = \beta_0 + \sum_j \beta_j X_j$

$P(Y=1 \mid \{X_j\})$

**probit model** $\Phi(z)$, $\Phi$ is cumulative standard normal

**logit model** $\frac{1}{1+e^{-z}}$, logistic function

almost identical shapes.
logit is traditionally easier to optimize



## ④ estimation and inference of probit & logit regression

### i) non-linear least squares

↳ in probit & logit regression, the coefficients are non-linear with respect to the prediction, so ordinary least squares no longer works. But the idea of taking derivates to find a closed form solution still applies

$$Q = \sum_{i=1}^{n} \left[ y_i - f\left( \hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right) \right]^2 \quad L^2 \text{ loss}$$

$$\frac{\partial Q}{\partial \hat{\beta}_j} = 0 \implies \text{solve}$$

1. estimators are consistent
2. normally distributed in large samples

} but, MLE estimators have smaller variance and are thus more efficient

### ii) maximum likelihood estimation

idea: assuming our data was collected SRS, iid, then can we estimate $\beta_j$ by finding the coefficients that maximise the probability of observing this dataset?

1. probability of observing $Y_i \mid \{X_{ij}\}$: $p_i = f(\{\beta_j\}, \{X_{ij}\})$ (logit or probit functions)
2. we can model it as $P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$
3. assuming iid, $P(Y_1 = y_1, Y_2 = y_2 \dots) = \prod_{i=1}^{n} P_i(Y_i = y_i)$
4. convert product to log for convenience

$$P(Y_1 = y_1, Y_2 = y_2 \dots) = \sum_{i=1}^{n} (y_i)\ln(p_i) + (1-y_i)\ln(1-p_i)$$

(can split for convenience of expression)

5. optimise by gradient ascent / descent (if negative log, which STATA does)

### iii) MLE estimator distributions

↳ $\hat{\beta}_j^{MLE}$ is unbiased & consistent, and normally distributed in large samples

↳ $\hat{\beta}_j^{MLE}$ achieves minimum asymptotic variance among unbiased estimators, and is thus the most efficient estimator in binary dependent regression

⇒ note that $SE(\hat{\beta}_j^{MLE})$ also needs to be robustly adjusted for heteroskedasticity

⇒ inference is identical to OLS for practical purposes

## ⑤ measures of fit

**$R^2$** $R^2$ is a poor measure of fit for the LPM. This also true for the probit and logit models. Ultimately, the data is inherently non-linear and limited; $R^2$, which measures deviation from a continuous line is thus not a good metric.

**Accuracy** Does not give a good indicator of quality of prediction ⟹ whether $p = 51\%$ or $90\%$ is the same.

**Pseudo-$R^2$** compares the likelihood of the estimated model ie. $P(\{b_j\} \mid \{x_{ij}\})$ to the likelihood when none of the X's are included as regressors — that is, probability $p_i$ is fixed

$$\text{pseudo-}R^2 = 1 - \frac{\ln\left(L^{fmax}_{probit}\right)}{\ln\left(L^{fmax}_{bernoulli}\right)}$$

conversion to log likelihood

( analogous to $R^2 = 1 - \frac{ESS}{TSS}$,

$1 - \frac{p'}{\bar{p}}$ → $p'$: explained by model; $\bar{p}$: baseline p of occurance

$$L^{fmax}_{probit} = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \quad p_i = \sigma\left(\hat{b}_0 + \sum b_j x_{ij}\right)$$

$$L^{fmax}_{bernoulli} = \prod_{i=1}^{n} \hat{p}^{y_i}(1-\hat{p})^{1-y_i}$$

($\hat{p}$: fixed)

## ⑥ comparing LPM, logit & probit models

↳ all three models are just approximations to the unknown population regression function $E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\})$

1. The LPM is easiest to use and interpret, but cannot capture the non-linear nature of the true function

2. The probit and logit models are identical for all intents and purposes, but as with most non-linear models, interpretation is trickier

3. In some datasets, all three models may be very similar

Notes

1. linear vs. probit — talk about saturation / tails, nonsensical predictions, qualitative difference in predictions across a range

# Instrumental variables regression

## ① Instrumental variable regression

### 1) linear regression, endo- and exogeneity

↳ in linear regression, there are three big threats to internal validity:
- → OVB
- → simultaneous causality
- → errors in variable

(endogeneous variables) $\text{corr}(X_{ij}, u_i) \neq 0 \Rightarrow E(u_i | X_{ij}) \neq 0$

(exogeneous variable) $\text{corr}(X_{ij}, u_i) = 0$

$E(u_i | X_{ij}) \neq 0 \Leftarrow$ $X_j$ is correlated with $u$

↳ multiple regression & panel data regression help w/ OVB ⇒ but how do we fix other biases?

### 2) idea behind IV regression

↳ corr. between $X_{ij}$ and $u_i$ is problematic because the estimator of $\hat{\beta}_j$ becomes biased, and intuitively so because minimising $L^2$ loss wouldn't be "representative"

↳ but, if we could replace $X_{ij}$ with the "component" of $X_{ij}$ not corr. w/ $u_i$, that would work! After all, it is still $X_{ij}$ changing, just that the reason it is changing is entirely unrelated to $u_i$ ⇒ so $E(u_i | X_{ij}) = 0$

⇒ how do we isolate that "component" of $X_{ij}$? ⇒ use other variables unrelated to $u_i$ to predict $X_{ij}$, then regress $Y$ on $\hat{X}_j$!

those other variables are precisely IVs

### 3) instrumental variables

1. instrument relevance: $\text{corr}(Z_i, X_{ij}) \neq 0 \Rightarrow Z_i$ predicts $X_{ij}$
2. instrument exogeneity: $\text{corr}(Z_i, u_i) = 0 \Rightarrow Z_i$ does not introduce bias

---

## ② Two stage least-squares estimator

### 1) model formulation

↳ idea: to find the unbiased effect of $X$ on $Y$, we induce unbiased changes in $X$ through an exogeneous variable $Z$. $Z$ might also predict $Y$, though. So to estimate $\hat{\beta}_j$, we simply need to account for the effects of $Z$ on $Y$.

$$x_i = \pi_0 + \pi_1 z_i + v_i \iff z_i = -\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1} x_i + V_i$$

$$y_i = \gamma_0 + \gamma_1 z_i + w_i$$

$z_i$ predicts both $x_i$ and $y_i$ w/o OVB.

$$E(v_i | z_i) = 0$$
$$E(w_i | z_i) = 0$$

$$y_i = \underbrace{\left(\gamma_0 - \gamma_1 \frac{\pi_0}{\pi_1}\right)}_{\beta_0} + \underbrace{\frac{\gamma_1}{\pi_1}}_{\beta_1} x_i + \underbrace{\left(w_i - \frac{1}{\pi_1} v_i\right)}_{u_i} \Rightarrow E(u_i | x_i) = 0 !$$

$$\beta_1 = \frac{\gamma_1}{\pi_1}$$

OLS estimation of $\pi_0, \pi_1, \gamma_1, \gamma_0$.

$$\gamma_1 = \frac{cov(y_i, z_i)}{var(z_i)}$$
$$\pi_1 = \frac{cov(x_i, z_i)}{var(z_i)}$$

$$\Rightarrow \boxed{\beta_1 = \frac{\gamma_1}{\pi_1} = \frac{cov(y_i, z_i)}{cov(x_i, z_i)}}$$

2) estimator distribution

$$\hat{\beta}_j^{TSLS} = \frac{s_{yz}}{s_{xz}} = \frac{\frac{1}{n-1}\sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n-1}\sum (x_i - \bar{x})(z_i - \bar{z})}$$

$$\hat{\beta}_1^{TSLS} \sim N\left(\beta_1, \frac{Var((z_i - \mu_z) u_i)}{n \cdot cov(x_i, z_i)^2}\right)$$

$\hat{\beta}_j^{TSLS}$ is unbiased if $z$ exogeneous

$$s_{zy} = \frac{1}{n-1}\sum (z_i - \bar{z})(y_i - \bar{y})$$
$$= \frac{1}{n-1}\sum (z_i - \bar{z})(\beta_1 (x_i - \bar{x}) + (u_i - \bar{u}))$$
$$= \beta_1 s_{zx} + \sum (z_i - \bar{z})(u_i - \bar{u})$$

$$\hat{\beta}_1^{TSLS} = \frac{s_{zy}}{s_{zx}} = \beta_1 + \frac{\frac{1}{n-1}\sum (z_i - \bar{z}) u_i}{\frac{1}{n-1}\sum (z_i - \bar{z})(x_i - \bar{x})}$$

$\frac{1}{n-1}\sum (z_i - \bar{z}) u_i \to s_{zu} \Rightarrow E(s_{zu}) = cov(z, u) = 0$ if $z$ is exogenous

$\frac{1}{n-1}\sum (z_i - \bar{z})(x_i - \bar{x}) \to s_{xz}$

$\hat{\beta}_j^{TSLS}$ approximately normal w/ known variance in large samples

1. $$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\frac{1}{n-1}\sum (z_i - \bar{z}) u_i}{\frac{1}{n-1}\sum (z_i - \bar{z})(x_i - \bar{x})} \iff \frac{\frac{1}{n}\sum (z_i - \bar{z}) u_i}{\frac{1}{n}\sum (z_i - \bar{z})(x_i - \bar{x})}$$

($s_{zu}$ → $cov(z,u)$; $s_{zx}$ → $cov(z,x)$)

2. since $\frac{1}{n}\sum (z_i - \bar{z}) u_i$ is the mean of $(z_i - \bar{z}) u_i$, and because $E(u_i | z_i) = 0$ (exogeneity assumption) we know

$$E\left(\frac{1}{n}\sum (z_i - \bar{z}) u_i\right) = 0, \quad Var\left(\frac{1}{n}\sum (z_i - \bar{z}) u_i\right) = \frac{Var((z_i - \bar{z}) u_i)}{n}$$

3. Also, rearranging, we have $(\hat{\beta}_1 - \beta_1)\, cov(z, x) = \frac{1}{n}\sum (z_i - \mu_z) u_i$ as $n \to \infty$.

4. By CLT, as $n \to \infty$, $\frac{1}{n}\sum (z_i - \bar{z}) u_i \sim N\left(0, \frac{Var((z_i - \bar{z}) u_i)}{n}\right)$ approximately.

5. Since $cov(z, x)$ is a constant, as $n \to \infty$, $\hat{\beta}_1 \sim N\left(\beta_1, \frac{Var((z_i - \mu_z) u_i)}{n \cdot cov(x, z)^2}\right)$ (scaling the variance)

3) estimation and inference in TSLS

one shot "two stage"

two stage

$\hat{\beta}_1^{TSLS} = \frac{S_{yz}}{S_{xz}}$

$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_i$

$\hat{SE}(\hat{\beta}_1^{TSLS})$, robust is ok to use directly

$\hat{se}$ must be heteroskedasticity robust, stin!

unbiased but standard errors will be wrong since they do not account for 2 stage

preferred

4) assumptions of IV regression $Y_i = \beta_0 + \underbrace{\sum_{i=1}^{k} \beta_k X_{ki}}_{\text{endogenous}} + \underbrace{\sum_{j=k+1}^{r} \beta_j W_{ji}}_{\text{exogenous}} + u_i$

1. $E(u_i \mid W_{1i} \dots W_{ji}) = 0$
2. $(Y_i, X_i \dots W_i \dots)$ is i.i.d.
3. no extreme outliers
4. IV $Z_1, \dots Z_k$ are valid

---

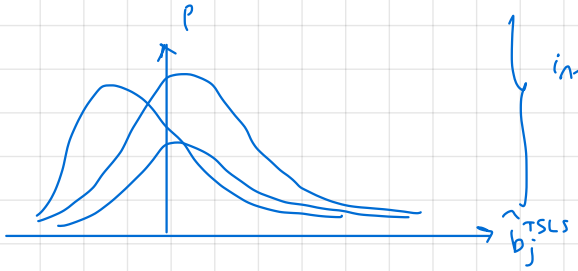
## (3) Checking instrument strength

1) weak (irrelevant) instruments

↳ a way to think about instrument relevance w.r.t. $\hat{\beta}_j$ is like sample size — higher corr., larger variation in X explained and triggered

↳ Statistical inference using TSLS relies on $\hat{\beta}_j^{TSLS}$ being approx. normal in large samples. But if instruments are weak, distribution is non-normal, even in large samples, and distributed more like ratio of two normal variables

⇒ not consistent, even if $n \to \infty$




intuition: $\hat{\beta}_j^{TSLS} = \frac{cov(Z,Y)}{cov(Z,X)}$ ← small

very sensitive

2) checking instrument relevance

idea: regress endogeneous $X$ on $Z_1 \dots Z_m$, controlled (OVB) with $W_1 \dots W_n$. Then do an F test that coefficients on $Z_1 \dots Z_m$ are zero. If rejection probability is low, instruments are weak

rule of thumb: first stage F test, f stat < 10 ⇒ weak instrument set

↳ bias of TSLS in weak instrument: $E(\hat{\beta}_j^{TSLS}) - \beta_j \approx \frac{E(\hat{\beta}_j^{OLS}) - \beta_j}{E(F) - 1}$

↳ if $E(F) = 10$, then $\frac{E(\hat{\beta}_j^{TSLS}) - \beta_j}{E(\hat{\beta}_j^{OLS}) - \beta_j} \approx \frac{1}{10-1}$ ⇒ that is, bias of TSLS relative to bias of (bad, since simultaneous) OLS is about 10%, which is generally acceptable

⇒ if $f > 10$, relative bias is $< 10\%$, vice versa

3) dealing with weak instruments

↳ if multiple instruments, some probably weaker than others — drop weaker, retest until satisfactory subset

---

④ checking instrument validity

under, exactly, over identified: no. of IVs less than, equal to, more than endogeneous variables $X_j$

checking instrument exogeneity: J test of overidentified instruments

↳ idea: exogeneity of IVs mean that they are uncorrelated with $u_i$. It logically follows that they should be statistically uncorrelated in our estimate of $u_i$, $\hat{u}_i = Y_i - (\hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X \dots + \hat{\beta}_{k+1}^{TSLS} W \dots)$

↑ true $X$, rather than predicted $\hat{X}$. Take note! Not $\hat{X}$.

1. estimate $Y_i = \beta_0 + \beta_1 X_1 \dots + \beta_{k+1} W_{k+1} \dots$ ⇒ get $\hat{u}_i = Y_i - \hat{Y}_i$
2. regress $\hat{u}_i$ on all supposedly exogeneous variables ⇒ $\hat{u}_i = \delta_1 z_1 \dots + \delta_{m+1} W_{m+1} \dots$
3. compute F stat. that $\delta_1 \dots \delta_{m+1}$, all coefficients on IV, are 0.
4. compute J stat., that is $J = mf$ (m: no. of IV)
5. compute p-value. $J \sim \chi^2(m-k \text{ dof})$ (m − k: no. of IV − no. of X endogeneous) → reject ⇒ not exogeneous

**how the J test works** we can derive a relationship using the above assumptions and equations that $J \sim \chi^2(m-k \text{ dof}) \Rightarrow$ reject using p-value.

**interpreting the J test** → what it does say
- hypothesis about instrument exogeneity (all)

what it does not say
- which instruments are the non-exogenous ones → use theory / intuition to decide removal

---

## ⑤ choosing instrumental variables

1. economic theory
2. random phenomena → cause variation in one, but not other (eg. birthdate)

Notes

1. Recall $\beta_1 = \frac{\gamma_1}{\pi_1} = \frac{cov(y_i, z_i)}{cov(x_i, z_i)} = \frac{cov(z_i, \overbrace{\beta_0 + \beta_1 x_i + \beta_2 w_i + u_i}^{\text{"true" functional form}})}{cov(x_i, z_i)}$

$= \frac{\beta_1 cov(z_i, x_i)}{cov(x_i, z_i)}$ ideally since all other $cov(z_i, -) = 0$

⇒ but if $z_i$ corr. to $w_i$, then

$\hat{\beta}_1 = \frac{\beta_1 cov(z_i, x_i) + \beta_2 cov(z_i, w_i)}{cov(x_i, z_i)}$ biased & inconsistent!

# Time series regression (forecasting)

## ① Fundamental concepts

### 1) Fundamental operators

**The jth lag** $Y_{t-j}$ is the $j$th lag. We use the lag operator $L$ to denote:

$$LY_t = Y_{t-1} \iff L^j Y_t = Y_{t-j}$$

**Differences** $\Delta Y_t = Y_t - Y_{t-1}$. $\Delta = (1-L)$

↳ $\Delta$ can be iteratively applied. $\Delta^2 Y_t = (1-L)^2 Y_t$

$$= (1 - 2L + L^2) Y_t$$

$$= Y_t - 2Y_{t-1} - Y_{t-2}$$

↳ note that $\Delta^j Y_t \neq Y_t - Y_j$ ⇒ substitute to $L$, then solve

### 2) autocovariance and autocorrelation

**$j^{th}$ autocovariance** the covariance between $Y_t$ and $Y_{t-j}$.

**$j^{th}$ autocorrelation** the correlation between $Y_t$ and $Y_{t-j}$.

$$\rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{Var}(Y_t)\,\text{Var}(Y_{t-j})}}$$

### 3) types of autocorrelation

**Geometric decay**

$\rho(Y_t, Y_{t-k}) \approx c^k$ for some $c < 1$

**slow decay** power law

$\rho(Y_t, Y_{t-k}) \approx k^d$ for some $d < 0$

---

## ② stationarity of processes

**Stochastic processes**

↳ traditionally defined as a collection of random variables indexed by some set

$\{X_t\}_{t \in T}$ ⇒ ie. an indexed set of RVs

↳ time series processes can be thought of as stochastic processes

**stationarity** a mathematical property that a stochastic process has when all the random variables of that process are identically distributed.

⇒ for any set $X_t \dots X_{t+n}$ in a stochastic process, all $X_t \dots X_{t+n}$ have the same probability distribution

⇒ the joint distribution of $Y_t, Y_{t-1} \dots Y_{t-K}$ is identical to $Y_{t-h}, Y_{t-h-1} \dots Y_{t-h-K}$ for any $h$. All corr. joint distributions are identical & time displacement invariant

the joint distributions $(Y_{s+1}, Y_{s+2} \dots Y_{s+T})$ do not depend on $s$, regardless of $T$ (though may be different for different $T$)

**covariance stationarity** a stochastic process exhibits covariance stationarity if its mean, variance and autocovariance do not depend on $t$ and variables $Y_t$ have finite variance.

$cov(X_t, X_{t+j})$ is independent of $t$, at most dependent on $j$

$E(Y_{t+j}) = \mu_j$ is independent of $t$, at most dependent on $j$

⇒ for EC3304, we set:

$$cov(Y_t, Y_{t+j}) = \sigma^2 \ \forall j \iff Var(Y_t) = \sigma^2 \ (j=0)$$

$$E(Y_t) = E(Y_{t+j}) = \mu \ \forall j$$

**stationarity up to order m**

↳ strict stationarity is hard in practice

↳ often we simply require that stationarity holds up to a certain point in history

↳ that is, $\forall h \in [0, H]$, $\{Y_{t+i}\}_{i \in [h, T+h]}$, all $Y_{t+i}$ are identically distributed or have the same mean, variance & $j^{th}$ autocovariance in the case of weak (autocovariance) stationarity

## ③ ergodicity

**definition** a stochastic process is said to be ergodic if its statistical properties can be deduced from a single, sufficiently long random sample of the process

↳ suppose autocovariance stationary process $\{Y_{t+h}\}_{h\in[0,T]}$ has mean $\mu$.

↳ then $E(Y_t) = E(Y_{t+1}) = E(Y_{t+i}) = \mu$

↳ then if $\frac{1}{T}\int_0^T \{Y_{t+h}\}_{h\in[0,T]}\, dt \to \mu$ as $T \to \infty$, we say it is ergodic for the 1st moment (mean)

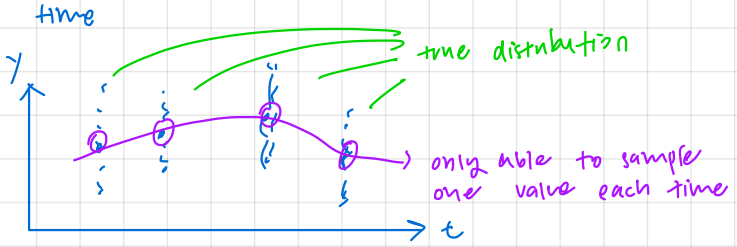
observe that this is only possible if faraway variables are essentially independent

⇒ if the statistical property is ergodic, then the long-horizon forecast converges to the unconditional value of property

---

## ④ why are stationarity & ergodicity important?

↳ with stochastic processes of time, it is by definition that there is no way to collect multiple data points for each variable, since that entails going back in time




↳ stationarity ⇒ $Y_t$ is identically distributed at all $t$

↳ ergodicity ⇒ w enough time steps, we can estimate the stationary properties!

⇒ can be simple like mean → then $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_{t,n} = \frac{1}{T}\sum_{i=1}^{T} Y_t$

⇒ can also be a more complicated recurrence relationship like

$Y_t = \beta_0 + \beta_1 Y_{t-1}$, and $\beta_0$ and $\beta_1$ can be estimated w enough time steps of data

---

## ⑤ Autoregression on stationary & ergodic time series

**autoregression** An autoregression is a regression model in which $Y_t$ is regressed against its own lagged values

↳ the number of lagged values is called the order of the regression

## autoregressive distributed lag model

↳ suppose there is some stable (stationary) relationship between $Y_t$ and its previous $p$ values, and also some other $x$ and their previous $k$ values

↳ that is, the population distribution of $Y_t$ follows

$$Y_t = \beta_0 + \beta_1 Y_{t-1} \dots + \beta_p Y_{t-p} + \delta_1 X_{t-1} \dots \delta_k X_{t-k} + u_t$$

($u_t$: some error term)

if $E(u_t \mid Y_{t-1}, Y_{t-2} \dots Y_{t-p}, X_{t-1}, \dots X_{t-k}) = 0$, then we can estimate $\hat{Y}_t$ w OLS regression such that $E(\hat{Y}_t) = E(Y_t)$.

---

ergodicity → we set up tuples of $Y_t, Y_{t-1} \dots Y_{t-k}$, and use OLS to estimate $\beta_0, \beta_1 \dots \delta_1 \dots$

stationarity → no variables dependent on time

## least squares assumptions

↳ analogous to those of OLS on cross sectional / panel data

1. $E(u_t \mid Y_{t-1}, Y_{t-2} \dots X_{t-1}, X_{t-2} \dots) = 0$. That is $E(u_t \mid \text{regressors}) = 0$
2. All $Y_t, X_{1t}, X_{2t} \dots$ have a stationary distribution. (No dependence on time needs to be modelled)
3. $(Y_t, X_{1t}, X_{2t} \dots)$ and $(Y_{t-j}, X_{1,t-j}, X_{2,t-j} \dots)$ are independent as $j$ get large. That is, all regressors are ergotic.
4. No perfect multicollinearity or outliers.

## ⑥ forecasting errors

### 1) forecast error, MSFE & RMSFE

$$FE = Y_{t+1} - \hat{Y}_{t+1}$$

$$MSFE = E\left((Y_{t+1} - \hat{Y}_{t+1})^2\right)$$

$$RMSFE = \sqrt{MSFE}$$

### 2) oracle forecast

suppose $\hat{\beta}_0, \hat{\beta}_1 \dots$ are estimated perfectly, with zero variance.

then $\hat{Y}_t = \beta_0 Y_{t+1} + \beta_1 Y_{t-2} \dots + u_t$.

Then $E(\hat{Y}_t) = 0$, $Var(\hat{Y}_t) = Var(u_t)$.

⇒ This case is known as the oracle forecast, and this variance and MSFE = $Var(u_t)$ for an oracle model.

1. observe $Var(X) = E(X^2) - E(X)^2$

2. $MSFE = E(Y_T - \hat{Y}_T)^2$

$= E(u_T^2) + E\left(((\beta_0 - \hat{\beta}_0) + (\hat{\beta}_1 - \beta_1) Y_{T-1} \dots)^2\right)$

$= E(u_T^2)$ for oracle forecast

$= Var(u_t)$ since $E(u_t \mid Y_{t-1} \dots) = 0$

but in the general case, is bias variance trade off.

↑ $MSE = bias^2 + Var + \sigma_u^2$

observe that this is $Var(\hat{\beta}_0 + \dots) = Var(\hat{Y}_T)$

for the same reason that mean = 0, $E(X)^2 = 0$ if unbiased

### 3) estimation of MSFE

$$MSFE = E(u_T^2) - E\left(((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) Y_{T-1} \dots)^2\right)$$

$E(u_T^2)$: variance of random error

$E\left(((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) Y_{T-1} \dots)^2\right)$: accumulated squared error from coefficients (variance of $\hat{Y}$)

recall that ergodicity only maintains for long time horizon! How long? ⇒ could be biased

**1. approximation by SER** if no. of predictors small compared to sample size, contribution of estimation error will be small

$$\hat{MSFE} \simeq E(u_T^2) = \sigma_u^2 \simeq SER$$

**2. estimation by final prediction error**

↳ assuming errors are homoskedastic

$$\hat{MSFE} = \frac{T+p+1}{T} s_{\hat{u}}^2 = \frac{T+p+1}{T-p-1} \frac{SSR}{T}$$

3. estimate by pseudo OOS

↳ the SER method ignores estimation error
↳ the FPE method requires homoskedasticity

} & formulation also require stationarity, since functional form of $\hat{Y}$ above based on stationary

↳ pseudo OOS simulates "real time"

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, $P$, for which you will generate pseudo out-of-sample forecasts; for example, $P$ might be 10% or 20% of the sample size. Let $s = T - P$.
2. Estimate the forecasting regression using the estimation sample—that is, using observations $t = 1, \ldots, s$.
3. Compute the forecast for the first period beyond this shortened sample, $s + 1$; call this $\tilde{Y}_{s+1|s}$.
4. Compute the forecast error, $\tilde{u}_{s+1} = Y_{s+1} - \tilde{Y}_{s+1|s}$.
5. Repeat steps 2 through 4 for the remaining periods, $s = T - P + 1$ to $T - 1$ (reestimate the regression for each period). The pseudo out-of-sample forecasts are $\tilde{Y}_{s+1|s}, s = T - P, \ldots, T - 1$, and the pseudo out-of-sample forecast errors are $\tilde{u}_{s+1}, s = T - P, \ldots, T - 1$.

$$\widetilde{MSFE}_{POOS} = \frac{1}{P}\sum_{s=T-P+1}^{T} \tilde{u}_{s.}^2$$

P samples for OOS testing, take avg. estimated errors

---

⑦ forecast uncertainty

forecast intervals: forecast intervals are like a confidence interval except that it pertains to a forecast.

↳ A key difference, though, is that CI are justified by CLN, so we can use the appropriate CI.

↳ in time series forecasting, we need to estimate the distribution of $u_t$ or make assumptions about it in order to compute forecast intervals.

⇒ in practice, it is convenient to assume $u_t$ is normally distributed and heteroskedastic. Under assumptions of stationarity, forecast error is the sum of $u_T$ and estimation error of coefficients.

$$E(Y_T - \hat{Y}_T) = E(u_T) + E((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)Y_T \cdots)$$

assuming ergodicity, second term is approx. normal by CLT, then if $u_t$ is normal, forecast error is also normal

⇒ in which case we formulate $\hat{y}_T \pm z_\alpha \widehat{RMSE}$ ← normally distributed

$\widehat{RMSE}$ PPE or $\widehat{RMSE}$ OOS

fun charts forecast distributions are frequently portrayed graphically as a fan chart


---

## ⑧ Determining the order of autoregression

↳ how do we decide on autoregressors? unlike multiple regression on panel data, where the objective was unbiased estimates, here we simply care about MSFE

↳ how do we decide the trade-off between exploiting autocorrelation vs. estimating too many and introducing more error into our MSFE?

F-stat similar to multivariate regression

Bayesian information criterion

$$BIC(p) = \ln\frac{SSR(p)}{T} + (p+1)\frac{\ln T}{T}$$ } intuition: SSR ↓, BIC ↓; p ↑, BIC ↑ (penalty)

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2$$ sum of squared residuals

↳ Fit $p^{th}$ model and compute BIC(p). Pick model w lowest BIC.

⇒ $\hat{p}$ which minimizes BIC is a consistent estimator of true lag length, ie. length of lag in underlying population model

⇒ minimising BIC maximises the posterior probability of the model being the true model

Akaike information criterion

↳ but in forecasting, "true model" is irrelevant — all that matters is MSE

$$AIC(p) = \ln\frac{SSR(p)}{T} + (p+1)\frac{2}{T}$$

SSR ↓, AIC ↓; p ↑, AIC ↑; T ↑, AIC ↓

↳ penalty for more p smaller than BIC

⇒ AIC estimates KL divergence between model distribution of $\hat{Y}_i$ and data distribution $Y_i$ ⇒ how much info lost using the $p^{th}$ model

⇒ not consistent like BIC, but efficient — if true model not considered, picks best approximation w probability 1

**Data for comparing different models**

↳ make sure same amt. of data for each
↳ i.e. ∴ necessarily decreases no. of samples

---

## (9) Non-stationarity (trends w time)

1) trends

**definition** a persistent, long-term movement of a variable overtime

**deterministic trends** $Y_t$ is a non-random function of time. That is, we can include $t$ as a regressor in the model, and the relationship between $Y_t$ and $t$ is stable

$$\text{eg. } Y_t = \beta_0 + \alpha_1 t + \alpha_2 t^2 + \beta_1 Y_{t-1} \cdots$$

**stochastic trends** $Y_t$ is a random function of time. That is, the way $Y_t$ fluctuates is a function of time. i.e. $\text{Var}(Y_t) = Y(t)$

⇒ when a trend is present, by definition, $(Y_t, Y_{t-1} \ldots)$ cannot be stationary. That is, the distribution of $Y_t$ must be different at each time step.

**random walk model**

special case of AR(1) where $\beta_0 = 0$, $\beta_1 = 1$

$$Y_t = Y_{t-1} + u_t, \quad u_t \text{ is i.i.d.}$$

1. $E(Y_t) = Y_{t-1} + E(u_t)$, since $Y_{t-1}$ is realised

2. $Y_t = Y_{t-2} + u_{t-1} + u_t \ldots$

$$= Y_{t-k} + \sum_{i=1}^{k} u_{t-i} = Y_0 + \sum_{i=1}^{t} u_{t-i}$$

$\text{Var}(Y_t) = 0 + t\,\text{Var}(u_t)$ if homoskedastic

3. $\text{cov}(Y_t, Y_{t-k}) = \text{cov}\left(Y_0 + \sum_{i=1}^{t-} u_i, Y_0 + \sum_{i=1}^{t-k} u_i\right)$

**random walk with drift**

$Y_t = \beta_0 + Y_{t-1} + u_t$, $u_t$ is i.i.d.

1. $E(Y_t) = \beta_0 + Y_{t-1} + E(u_t)$
2. $\text{Var}(Y_t)$ same $= t\,\text{Var}(u_t)$
3. $\text{cov}(Y_t, Y_{t-k}) = (t-k)\,\text{Var}(u_t)$, since $\text{Var}(\beta_0) = 0$

Since $u_i$ is i.i.d, $u_i$ in different periods are independent.

so $cov(u_i, u_j)$, if $j = 0$

Then $cov\left(\sum_{i=1}^{t} u_i, \sum_{i=1}^{k} u_{t-i}\right)$ resolves to $\sum_{i=1}^{t-k} var(u_i) = (t-k)\, var(u_i)$

Since $Var(Y_t)$ and $Var(Y_t, Y_{t-k})$ depend on $t$, stochastic trend!

2) testing for non-stationarity of $Y_T$

Stationarity of $Y_T$

1. if $Y_t$ follows an AR(p) model, $Y_t$ is stationary if its roots $z$ are all greater than 1 in absolute value.

1.1 $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} \dots + u_t$, $u_t$ is i.i.d.

$= \beta_0 + \beta_1 (\beta_0 + \beta_1 Y_{t-2} \dots) + \beta_2 (\beta_0 + \beta_1 Y_{t-3}) \dots + u_t$

1.2 $Y_t$ = some function of $u_t$.

2. recall that a stochastic process is autocovariance stationary if $E(Y_T)$, $cov(Y_T, Y_{T-k})$ and $Var(Y_T) = cov(Y_T, Y_{T-0})$ is constant.

2.1 $cov(Y_T, Y_{T-k})$ = some function of $u_T$.

2.2 $Var(Y_T)$ = some function of $u_T$

2.3 $E(Y_T)$ = some function of $u_T$.

2.4 For 2.1, 2.2, 2.3 to be constant, roots $z_1 \dots z_p$ must be greater than 1.

$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} \dots + u_t$, $u_t$ is i.i.d.

$= \beta_0 + \beta_1 (\beta_0 + \beta_1 Y_{t-2} \dots) + \beta_2 (\beta_0 + \beta_1 Y_{t-3}) \dots + u_t$

roots are the $z$ that satisfy $1 - \beta_1 z - \beta_2 z^2 \dots - \beta_p z^p = 0$

$p$ roots. if all $|z| > 1$, stationary.

$\Leftrightarrow$ sum of coefficients $\beta_1$ to $\beta_p = 1$ if not stationary

## Dickey Fuller test in the AR(1) model

↳ main idea: not stationary $\rightarrow \sum_{i=1}^{p} \beta_i = 1 \Rightarrow H_0$: not stationary, try to reject

$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$

$H_0: \beta_1 - 1 = 0$

$H_1: \beta_1 - 1 < 0$

fit $Y_t - Y_{t-1} = \beta_0 + \beta_1 Y_{t-1} - Y_{t-1} + u_t$

$\Leftrightarrow \Delta Y_t = \beta_0 + (\beta_1 - 1) Y_{t-1} + u_t$

test $\frac{\beta}{SE(\beta)}$, where $\beta$ is DF distributed against critical value

## Extending the dickey fuller test to AR(p)

1. For $p = 2$, observe that:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$$

$$= \beta_0 + (\beta_1 + \beta_2) Y_{t-1} - \beta_2 Y_{t-1} + \beta_2 Y_{t-2}$$

$$= \beta_0 + (\beta_1 + \beta_2) Y_{t-1} - \beta_2 \Delta Y_{t-1}$$

2. For $p = 3$, observe that:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}$$

$$= \beta_0 + (\beta_1 + \beta_2 + \beta_3) Y_{t-1} - (\beta_2 + \beta_3)(Y_{t-1} - Y_{t-2}) - \beta_3 (Y_{t-2} - Y_{t-3})$$

$$= \beta_0 + (\beta_1 + \beta_2 + \beta_3) Y_{t-1} - (\beta_2 + \beta_3) \Delta Y_{t-1} - \beta_3 \Delta Y_{t-2}$$

3. By induction, we can show that if we regress

$$\Delta Y_t = \beta_0 + \beta Y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta Y_{t-i}, \quad \beta = \sum_{i=1}^{p-1} \beta_i - 1$$

3.1 This is the augmented Dickey Fuller test.

$H_0: \beta = \sum_{i=1}^{p} \beta_i - 1 = 0 \rightarrow$ non-stationary

$H_1: \beta < 0 \rightarrow$ not-non-stationary

test against ADF statistic distribution by $\frac{\beta}{SE(\beta)}$ vs. $ADF_\alpha$

4. We choose p by the same ways.

3) testing for stochastic & deterministic non-stationarity

↳ we extend the augmented DF test by including function of t

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta Y_{t-i}, \quad \delta = \sum_{i=1}^{p-1} \beta_i - 1$$

(include time terms)

$H_0: \delta = \sum_{i=1}^{p} \beta_i - 1 = 0 \rightarrow$ non-stationary

$H_1: \delta < 0 \rightarrow$ not-non-stationary

} test against ADF statistic distribution by $\frac{\delta}{SE(\delta)}$ vs. $ADF_\tau$

---

(10) problems caused by stochastic trends

1) bias & non-normality of t-statistic

↳ t-stat of OLS estimates of $\beta_i$ in $AR(p)$ models will be non-normal and biased toward 0, even in large samples ⇒ invalid hypothesis testing

2) spurious regressions

↳ stochastic trends can lead to two time series eg. $Y_t$, $X_t$ to appear related when they are not, if both are stochastic

---

(11) Dealing with non-stationarity

De-trending the data

↳ observe that in random walk model, stochasticity arises from influence of previous step ⇒ de-trend by differencing

$$Y_t = \beta_0 + Y_{t-1} + u_t$$

$$\Delta Y_t = \beta_0 + u_t \Rightarrow \text{stationary about } \beta_0$$