W2 PROBABILITY & STATISTICS REFRESHER

- Reference: Appendices B-1, B-2, B-3, B-4 of the textbook
  - Random variables (discrete, continuous) and their probability distribution
  - Mean, variance, standard deviation
  - Properties of expectation
  - Covariance and correlation
  - Joint and conditional distributions
  - Conditional expectation function as the fundamental target of modelling

Random variables

- Economic and financial variables are by nature random. We do not know what their values will be until we observe them
- A random variable is a rule that assigns a numerical outcome to an event in each possible state of the world
- For e.g. the first wage offer that a BCom graduate receives in the job market is a random variable. The value of ASX200 index tomorrow is a random variable. Other examples are…
- A discrete random variable has a finite number of distinct outcomes. For e.g. rolling a die is a random variable with 6 distinct outcomes
- A continuous random variable can take a continuum of values within some interval. For e.g. rainfall in Melbourne in May can be any number in the range from 0.00 to 200.00mm.
- While the outcomes are uncertain, they are not haphazard. The rule assigns each outcome to an event according to a probability

A random variable and its probability distribution

- A discrete random variable is fully described by

its possible values	$x_1, x_2, \ldots, x_m$
probability corresponding to each value	$p_1, p_2, \ldots, p_m$

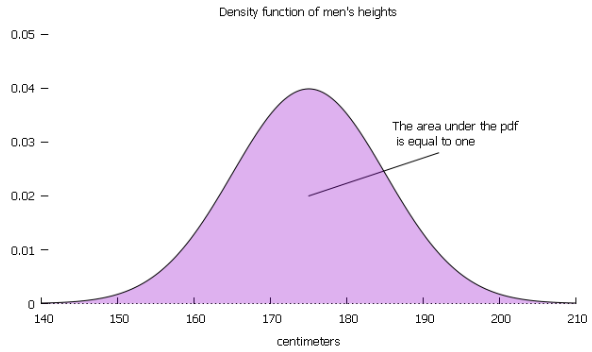
with the interpretation that $P(X = x_1) = p_1$, $P(X = x_2) = p_2$, ..., $P(X = x_m) = p_m$.

- The probability density function (pdf) for a discrete random variable X is a function f with $f(x_i) = p_i$, $i = 1, 2, \ldots, m$ and $f(x) = 0$ for all other x
- Probabilities of all possible outcomes of a random variable must sum to 1

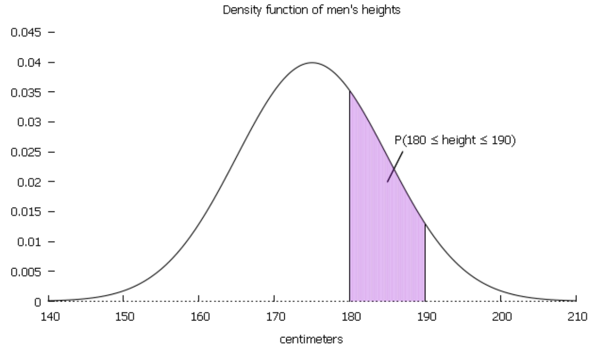
$$\sum_{i=1}^{m} p_i = p_1 + p_2 + \cdots + p_m = 1$$

- Examples:
  1. Rolling a die
  2. Sex of a baby who is not yet born. Is it a random variable?
  3. The starting wage offer to a BCom graduate

- The probability density function (pdf) for a continuous random variable X is a function f such that $P(a \leq X \leq b)$ is the area under the pdf between a and b
- The total area under the pdf is equal to 1
- Example: Distribution of men's height
- The area under the pdf is equal to 1



- The probability of that the height of a randomly selected man lies in a certain interval is the area under the pdf over that interval



Features of probability distributions:

1. Measures of Central Tendency (textbook ref: B-3)

- Expected value or mean of a discrete random variable is given by

$$E(X) = p_1x_1 + p_2x_2 + \cdots + p_mx_m = \sum_{i=1}^{m} p_ix_i$$

and for a continuous random variable is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- Intuitively, expected value is the long-run average if we observe X many, many, many times
- It is convention to use the Greek letter μ to denote expected value:

$$\boxed{\mu_X = E(X)}$$

- Another measure of central tendency is the median of X, which is the "middle-most" outcome of X, i.e. x_{med} such that $P(X \leq x_{med}) = 0.5$
- Finally, there is the mode which is the most likely value, i.e. the outcome with the highest probability. It is not a widely used measure of central tendency

2. Measures of dispersion (textbook ref: B-3)

- Variance of a random variable:

$$\sigma_X^2 = Var(X) = E(X - \mu_X)^2$$

- Variance is a measure of spread of the distribution of X around its mean
- If X is an action with different possible outcomes, then Var(X) gives an indication of riskiness of that action
- Standard deviation is the square root of the variance. In finance, sd is called the volatility in X

$$\sigma_X = sd(X) = \sqrt{E(X - \mu_X)^2}$$

- The advantage of sd over var is that it has the same units as X

Properties of the Expected Value (textbook ref: B-30

1. For any constant c, $E(c) = c$
2. For any constants a and b,
$$E(aX + b) = aE(X) + b$$
3. Expected value is a **linear** operator, meaning that expected value of sum of several variables is the sum of their expected values:
$$E(X + Y + Z) = E(X) + E(Y) + E(Z)$$
   - The above three properties imply that for any constants a, b, c and d and random variables X, Y and Z,
$$E(a + bX + cY + dZ) = a + bE(X) + cE(Y) + dE(Z)$$

Question (Answer: B)

We have put one quarter of our savings in a term deposit (a risk free investment) with annual return of 2% and invested the other three quarters in an investment fund with expected annual return of 3% and variance of 4.

The expected value of the annual return of our portfolio is

A. $\frac{2+3}{2} = 1.5\%$

B. $\frac{1}{4} \times 2 + \frac{3}{4} \times 3 = 2.75\%$

C. $\frac{3}{4} \times 3 = 2.25\%$

D. $\frac{1}{4} \times 2 + \frac{3}{4} \times 3 \times \sqrt{4} = 5\%$

E. $\frac{1}{4} \times 0 + \frac{3}{4} \times 4 = 3\%$

- It is important to have in mind that E is a linear operator, so it "goes through" sums of random variables, but it does not go through non-linear transformations of random variables. For example:

$$E(X^2) \neq (E(X))^2$$

$$E(\log X) \neq \log(E(X))$$

$$E(XY) \neq E(X)E(Y) \text{ unless } X \text{ and } Y \text{ are statistically independent}$$

- Using properties of expectations, we can now show that

$$Var(X) = E(X^2) - \mu^2$$

$$\begin{aligned} Var(X) &= E(X - \mu)^2 = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

Properties of the Variance (textbook ref: B-3)

1. For any constant *c*, *Var*(*c*) = 0
2. For any constants *a* and *b*,

$$Var(aX+b)=a^2Var(X)$$

3. There is a third property related to the variance of linear combinations of random variables that is very important and we will see later after we introduce the covariance

Question (Answer: A)

We have put one quarter of our savings in a term deposit (a risk free investment) with annual return of 2% and invested the other three quarters in an investment fund with expected annual return of 3% and variance of 4.

The variance of the annual return of our portfolio is

A. $\left(\frac{3}{4}\right)^2\times 4=2.25$

B. $\frac{1}{4}\times 0+\frac{3}{4}\times 4=3$

C. $\frac{1}{4}\times 2+\frac{3}{4}\times 4=3.25$

D. $\left(\frac{1}{4}\right)^2\times 2+\left(\frac{3}{4}\right)^2\times 3=1.8125$

E. $\left(\frac{1}{4}\right)^2\times 2+\left(\frac{3}{4}\right)^2\times 4=2.375$

Properties of the Conditional Expectation

- Conditional expectation of Y given X is generally a function of X
- Property 1: Conditional on X, any function of X is no longer a random variable and can be treated as a known constant, and then the usual properties of expectations apply. For example, if X, Y and Z are random variables and a, b and c are constants, then

$$E(XY\mid X)=XE(Y\mid X)$$

or

$$E((a+bX+cXY+X^2Z)\mid X)=a+bX+cXE(Y\mid X)+X^2E(Z\mid X)$$

- Property 2: If E(Y | X) = c where c is a constant that does not depend on X , then E(Y) = c . This is intuitive: if no matter what *X* happens to be, we always expect *Y* to be *c* , then the expected value of *Y* must be *c* regardless of *X*, i.e. the unconditional expectation of *Y* must be *c*.

Important features of joint probability distribution of two random variables: Measures of Association (textbook ref: B-4)

- Statistical dependence tells us that knowing the outcome of one variable is informative about probability distribution of another
- To analyse the nature dependence, we can look at the joint probability distribution of random variables
- This is too complicated when random variables have many possible outcomes (e.g. per capita income and life span, or returns on Telstra and BHP stocks)
- We simplify the question to: when X is above its mean, is Y more likely to be below or above its mean?
- This corresponds to the popular notion of X and Y being "positively or negatively correlated"

Covariance

- Question: "when X is above its mean, is Y more likely to be below or above its mean?"
- We can answer this by looking at the sign of the covariant between X and Y defined as:

$$Cov(X,Y)=E((X-\mu_X)(Y-\mu_Y))=E(XY)-\mu_X\mu_Y$$

- If X and Y are independent Cov(X,Y) = 0
- For any constants a_1, b_1, a_2 and b_2

$$Cov(a_1X+b_1,a_2Y+b_2)=a_1a_2Cov(X,Y)$$

Correlation

- Only the sign of covariance is informative. Its magnitude changes when we scale variables $Cov(aX,bY)=a\ b\ Cov(X,Y)$
- A better and unit free measure of association is correlation which is defined as: $Corr(X,Y)=\dfrac{Cov(X,Y)}{sd(X)sd(Y)}$
- Correlation is always between -1 and +1, and its magnitude, as well as its sign, is meaningful
- Correlation does not change if we change the units of measurement $Corr(a_1X+b_1,a_2Y+b_2)=Corr(X,Y)$

Variance of sums of variables: Diversification (textbook ref: B-4)

- One of the important principles of risk management is "Don't put all your eggs in one basket"
- The scientific basis of this is that:

$$Var(aX+bY)=a^2Var(X)+b^2Var(Y)+2abCov(X,Y)$$

- Example. You have the choice of buying shares of company A with mean return of 10 percent and standard deviation of 5%, or shares of company B with mean return of 10% and sd of 10%. Which would you prefer?
- Obviously A is less risky, and you prefer A to B
- Now consider a portfolio of investing 0.8 of your capital in company A and the rest in B, where, as before, A has mean return of 10% and sd of 10%. What are the return and the risk of this position with the added assumption that the returns to A and B are independent
- Denoting the portfolio return by Z, we have
- We can see that this portfolio has the same expected return as A, and is safer than A

$$Z=0.8A+0.2B$$
$$E(Z)=E(0.8A+0.2B)=$$
$$Var(Z)=Var(0.8A+0.2B)=$$

Diversification in econometrics - Averaging (lect 2 - 1:18)

- Suppose we are interested in starting salaries of BCom graduates. This is a random variable with many possibilities and a probability distribution
- Let's denote this random variable by *Y*. We also denote its population mean and variance by μ and σ^2. We are interested in estimating μ, which is the expected wage of a BCom graduate
- Suppose we choose one BCom graduate at random and denote his/her starting salary by Y_1. Certainly Y_1 is also a random variable with the same possible outcomes and probabilities as *Y*. Therefore, $E(Y_1)=\mu$. So it is OK to take the value of Y_1 as an estimate of μ, and the variance of this estimator is σ^2.
- But if we take 2 independent observations and use their average as our estimator of μ, we have:

$$\begin{aligned}E(\tfrac{1}{2}(Y_1+Y_2)) &= \tfrac{1}{2}(\mu+\mu)=\mu\\ Var(\tfrac{1}{2}(Y_1+Y_2)) &= \tfrac{1}{4}Var(Y_1)+\tfrac{1}{4}Var(Y_2)\\ &= \tfrac{1}{4}(\sigma^2+\sigma^2)=\sigma^2/2\end{aligned}$$

- Now consider a sample of n independent observations of starting salaries of BCom graduate $\{Y_1, Y_2,\ldots, Y_n\}$
- Y_1 to Y_n are i.i.d. (independent and identically distributed) with mean μ and variance σ^2
- Their average is a portfolio of that gives each of these *n* random variables the same weight of 1/n. So

$$E\left(\bar{Y}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu.$$

$$\begin{aligned}Var\left(\bar{Y}\right)=Var\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right] &= \frac{1}{n^2}Var\left[\sum_{i=1}^{n}Y_i\right]\\ &= \frac{1}{n^2}\sum_{i=1}^{n}Var\left(Y_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}.\end{aligned}$$

- The samples average has the same expected value as Y but a lot less risk. In this way, we use the scientific concept of diversification in econometrics to find better estimators!

Key concepts and their importance

- Business and economic variables can be thought of as random variables whose outcomes are determined by their probability distribution
  - A measure of central tendency of the distribution of a random variable is its expected value
  - Important measures of dispersion of a random variable are variance and standards deviation. These are used as measures of risk
- Covariance and correlation are measures of linear statistical dependence between two random variables
  - Correlation is unit free and measures the strength and direction of association
  - Statistical dependence or association does not imply causality
  - Two random variables that have non-zero covariance or correlation are statistically dependent, meaning that knowing the outcome of one of the two random variables gives us useful information about the other
  - Averaging is a form of diversification and it reduces risk

Estimators and their Unbiasedness

- *Estimator is a function of sample data*
- An estimator is a formula that combines sample information and produces an estimate for parameters of interest
- Example: Sample average is an estimator for the population mean
- Example: $\widehat{\beta} = (\mathbf{X'X})^{-1}\,\mathbf{X'y}$ is an estimator for the parameter vector β in the multiple regression model. (Population beta is a fixed vector)
- Since estimators are functions of sample observations, they… change as the sample changes unlike the fixed population beta
- While we do not know the values of population parameters, we can use the power of mathematics to investigate if the expected value of the estimator is indeed the parameter of interest
- Definition: An estimator is an unbiased estimator of a parameter of interest if its expected value is the parameter of interest

The Expected Value of the OLS Estimator

- Under the following assumption, $E(\widehat{\beta}\,) = \beta$

Multiple Regression Model

MLR.1 Linear in Parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$	**E.1 Linear in Parameters** $\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}}\ \underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\mathbf{u}}$
MLR.2 Random Sampling We have a random sample of n observations	
MLR.4 Zero Conditional Mean $E(u \mid x_1, x_2, \ldots, x_k) = 0$	**E.3 Zero Conditional Mean** $E(\mathbf{u} \mid \mathbf{X}) = \underset{(n\times 1)}{\mathbf{0}}$
MLR.3 No Perfect Collinearity None of $x_1, x_2, \ldots, x_k$ is a constant and there are no *exact linear* relationships among them	**E.2 No Perfect Collinearity** $\mathbf{X}$ has rank $k+1$

- We use an important property of conditional expectations to prove that the OLS estimator is unbiased: if E(z | w) = 0 E(g (w) z) = 0 for any function *g.* For example E(wz) = 0, E(w²z) = 0, etc.
- Now let's show $E(\widehat{\beta}\,) = \beta$
- The assumption of no perfect collinearity (E.2) is immediately required because…

S1: Using the population model (Assumption E.1), substitute for **y** in the estimator's formula and simplify

$$\begin{aligned}\widehat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})\\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

S2: Take expectations

$$\begin{aligned}E(\widehat{\beta}) &= E(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})\\ &= \beta + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})\\ &= \beta\end{aligned}$$

Using Assumption E.3, since E(**u** | **X**) = 0 ⇒ E((**X'X**)$^{-1}$**X'u**) = **0**

- Note: all assumptions were needed and were used in this proof

- Are these assumptions too strong?
- Linearity in parameters is not too strong, and does not exclude non-linear relationships between **y** and **x** (more on this later)
- **Random sample** is not too strong for cross-sectional data if participation in the sample is not voluntary. Randomness is obviously not correct for time series data
- **Perfect multicollinearity is quite unlikely unless we have done something silly** like using income in dollars and income in cents in the list of independent variables, or we have fallen into the "dummy variable trap" (more on this later)
- **Zero conditional mean is not a problem for predictive analytics**, because for best prediction, we always want our best estimate of E(y | x$_1$, x$_2$,…., x$_k$) for a set of observed predictors
- **Zero conditional mean can be a problem for prescriptive analytics (causal analysis)** when we want to establish the causal effect of one of the **x** variables, say **x$_1$** on **y** controlling for an attribute that we cannot measure. E.g. Causal effect of education on wage keeping ability constant:

$$\begin{aligned}\text{We want to estimate } \beta_1 \text{ in:}\quad wage &= \beta_0 + \beta_1 educ + \beta_2 ability + u\\ \text{We run the regression:}\quad wage &= \hat{\alpha}_0 + \hat{\alpha}_1 educ + \hat{v}\end{aligned}$$

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2 \frac{\partial ability}{\partial educ} \neq \beta_1$$

- $\widehat{\alpha}_1$ Is a biased estimator of β_1
- This is referred to as "omitted variable bias"
- One solution is to add a measurable proxy for ability, e.g. IQ
- **Zero conditional mean** can be quite restrictive in time series analysis as well
- It implies that the error term in any time period ***t*** is uncorrelated with each of the regressors, in *all time periods*, past, present and future
- Assumption **MLR.4** is violated when the regression model contains a lag of the dependent variable as a regressor. E.g. we want to predict this quarter's GDP using GDP outcomes for the past 4 quarters
- In this case, the regression parameters are biased estimators

The Variance of the OLS Estimator

- Coming up with unbiased estimators is not too hard. But how the estimates that produce are dispersed around the mean determines how precise they are
- To study the variance of $\widehat{\beta}$, we need to learn about variances of a vector of random variables **(var-cov matrix)**

- We introduce an extra assumption:

MLR.1 Linear in Parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$	**E.1 Linear in Parameters** $\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}}\ \underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\mathbf{u}}$
MLR.2 Random Sampling We have a random sample of n observations	
MLR.4 Zero Conditional Mean $E(u \mid x_1, x_2, \ldots, x_k) = 0$	**E.3 Zero Conditional Mean** $E(\mathbf{u} \mid \mathbf{X}) = \underset{(n\times 1)}{\mathbf{0}}$
MLR.3 No Perfect Collinearity None of $x_1, x_2, \ldots, x_k$ is a constant and there are no *exact linear* relationships among them	**E.2 No Perfect Collinearity** $\mathbf{X}$ has rank $k+1$
MLR.5 Homoskedasticity $E(u^2 \mid x_1, x_2, \ldots, x_k) = \sigma^2$	**E.4 Homo+Randomness** $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$

- Under these assumptions, the variance-covariance matrix of the OLS estimator conditional on **X** is because (in lecture notes L4 S25)

$$Var(\widehat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- We can immediately see that given **X** the OLS estimator will be more precise (i.e. its variance will be smaller) the smaller σ² is
- It can also be seen (not as obvious) that as we add observations to the sample, the variance of the estimator decreases, which also makes sense
- Gauss-Markov Theorem: Under Assumptions E.1 to E.4 (or MLR.1 to MLR.5) $\widehat{\beta}$ is the best linear unbiased estimator (B.L.U.E) of β
- This means that there is no other estimator that can be written as a linear combination of elements of **y** that will be unbiased and will have a lower variance than $\widehat{\beta}$
- This is the reason that everybody loves the OLS estimator

Estimating the Error Variance

- We can calculate $\widehat{\beta}$, and we showed that but we cannot compute this because we do not know σ²:

$$Var(\widehat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- An unbiased estimator of σ² is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k-1}$$

- The square root of $\hat{\sigma}^2$, $\hat{\sigma}$, is reported by views in regression output under the name **standard error of the regression**, and it is a measure of how good the fit is (the smaller, the better)
- Why do we divide SSR by n - k - 1 instead of *n*?
- In order to get an unbiased estimator of σ²:

$$E\left(\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k-1}\right) = \sigma^2$$

  If we divide by *n* the expected value of the estimator will be slightly different from the true parameter (proof of unbiasedness of $\hat{\sigma}^2$ is not required)
- Of course if the sample size *n* is large, this bias is very small
- Think about dimensions: **ŷ** is in the column space of **X**, so it is in a subspace with dimension k + 1
- **û** is orthogonal to column space of **X**, so it is in a subspace with dimension n - (k + 1) = n - k - 1. So even though there are *n* coordinates in **û**, only n - k - 1 of those are free (it has n - k - 1 **degrees of freedom**)

- We use the result on the t distribution to test the null hypothesis about a single β_j
- Most routinely we use it to test if controlling for all other x, x_j has no partial effect on y:

$$H_0 : \beta_j = 0$$

for which we use the **t statistic** (or **t ratio**)

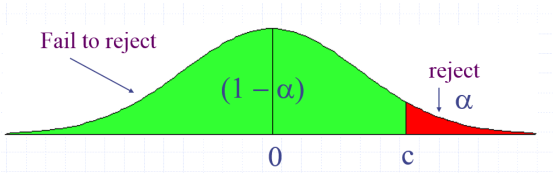
$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

which is computed automatically by most statistical packages for each estimated coefficient

- But to conduct the test we need an alternative hypothesis

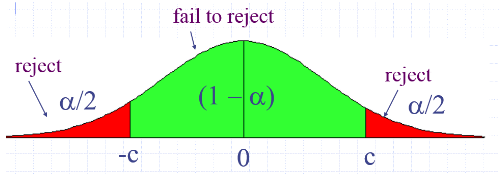
- The alternative hypothesis can be one-sided, such as $H_1 : \beta_j < 0$ or $H_1 : \beta_j > 0$, or it can be two-sided $H_1 : \beta_j \neq 0$
- The alternative determines what kind of evidence is considered legitimate against the null. For example, for $H_1 : \beta_j < 0$ we only are excited to reject the null if we find evidence that it is negative. We won't be interested in any evidence that β_j is positive, and we won't reject the null if we found such evidence
- Example: The effect of missing lectures on final performance, the alternative hypothesis is that missing lectures has negative effect on your final, even after controlling for how smart you are. We are not interested in any evidence that missing lectures improves final performance
- With two sided alternatives, we take any evidence that β_j may not be zero, whether positive or negative, as legitimate

- We also need to specify α the size or the significance level of the test, which is the probability that we wrongly reject the null when it is true (Type I error)
- Using the t_{n-k-1} distribution, the significance level and the type of alternative, we determine the critical value that defines the rejection region
- For $H_1 : \beta_j > 0$



- For $H_1 : \beta_j < 0$

- For $H_1 : \beta_j \neq 0$



- If t_{calc} (the value of the t-statistic in our sample) falls in the critical region, we reject the null
- When we **reject** the null, we say that x_j **is** statistically significant at the α% level
- When we **fail to reject** the null, we say that x_j **is not** statistically significant at the α =% level

- We can use the t test to test the null hypothesis:

$$H_0 : \beta_j = r$$

where ***r*** is a constant, not necessarily zero. Under the assumptions of CLM, we know that:

$$\frac{\hat{\beta}_j - r}{se(\hat{\beta}_j)} \sim t_{n-k-1} \text{ if } H_0 \text{ is true}$$

So we test the null using this t statistic

- Percentiles of t distribution with various degrees of freedom are given in Table G.2 on page 833 (5th edition) of the textbook

Confidence Intervals

- Another way to use classical statistical testing is to construct a confidence interval using the same critical value as was used for a two-sided test
- A (1 - α)% confidence interval is defined as

$$\hat{\beta}_j \pm c \times se(\hat{\beta}_j)$$

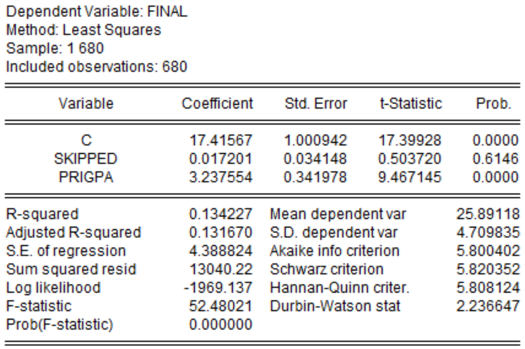
where **c** is the $(1-\frac{\alpha}{2})$ percentile of a t_{n-k-1} distribution

- The interpretation of a (1 - α)% confidence interval is that the interval will cover the true parameter with probability (1 - α)
- If the confidence interval does not contain zero, we can deduce that x_j **is** statistically significant at the α% level

p-value of a test

- An alternative to classical approach to hypothesis testing is to ask "based on the evidence in the sample, what is the value that for all significance levels less than that value the null would not be rejected, and for all significance levels bigger than that, the null would be rejected?"
- To find this, need to compute the t statistic under the null, then look up what percentile it is in the t_{n-k-1} distribution
- This value is called the p-value
- Most statistical packages report the p-value for the null of $\beta_j = 0$, assuming a two-sided test. Obviously, if you want this for a one-sided alternative, just divide the two-sided p-value by 2

Example: Effect of missing classes on the final score



```
Dependent Variable: FINAL
Method: Least Squares
Sample: 1 680
Included observations: 680

      Variable        Coefficient   Std. Error   t-Statistic    Prob.

          C            17.41567     1.000942     17.39928      0.0000
       SKIPPED         0.017201     0.034148     0.503720      0.6146
        PRIGPA         3.237554     0.341978     9.467145      0.0000

R-squared              0.134227   Mean dependent var       25.89118
Adjusted R-squared     0.131670   S.D. dependent var       4.709835
S.E. of regression     4.388824   Akaike info criterion    5.800402
Sum squared resid      13040.22   Schwarz criterion        5.820352
Log likelihood        -1969.137   Hannan-Quinn criter.     5.808124
F-statistic            52.48021   Durbin-Watson stat       2.236647
Prob(F-statistic)      0.000000
```

Testing multiple linear restrictions: The *F* test

- Sometimes we want to test multiple restrictions. For example in the regression model β

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

we are always interested in the overall significance of the model by testing

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

or we may be interested in testing

$$H_0 : \beta_3 = \beta_4 = 0$$

or even a more exotic hypothesis such as:

$$H_0 : \beta_1 = -\beta_2, \ \ \beta_3 = \beta_4 = 0$$

- The first null involves _4_ restrictions, the second involves ___ restriction and the third involves ___ restrictions
- The alternative can only be that at least one of these restrictions is not true
- The test statistic involves estimating two equations, one without restrictions (the unrestricted model) and one with the restrictions imposed (the restricted model), and seeing how much their sum of squared residuals differ
- This is particularly easy for testing exclusion restrictions like the first two nulls on the previous slide

Example: for

$$H_0 : \beta_3 = \beta_4 = 0 \text{ (note: 2 restrictions)}$$

the alternative is:

$$H_1 : at\ least\ one\ of\ \beta_3\ or\ \beta_4\ is\ not\ zero$$

the unrestricted model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \qquad \text{(UR)}$$

and the restricted model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad \text{(R)}$$

Models involving logarithms: log-level

- This is not satisfactory because it predicts that regardless of what your wage currently is, an extra year of schooling will add $42.06 to your wage
- It is more realistic to assume that it adds a *constant percentage* to your wage, not a constant dollar amount
- How can we incorporate this in the model?

> Logarithm of y on x: $\widehat{\log(y)} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2$
> $\widehat{\beta_1} \times 100$: the percentage change in predicted y as x_1 increases by 1 unit, keeping x_2 constant

- In our example, we use natural **logarithm of wage as the dependent variable**

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

- Holding ***IQ*** and ***u*** fixed

$$\Delta \log(wage) = \beta_1 \Delta educ$$

so

$$\beta_1 = \frac{\Delta \log(wage)}{\Delta educ}$$

- Useful result from calculus

$$100 \cdot \Delta \log(wage) \approx \%\Delta wage$$

- This leads to a simple interpretation of β_1 :

$100\beta_1 \approx \%\Delta wage$ when $\Delta educ = 1$ holding IQ constant

- If we do not multiply by 100, we have the decimal version (the proportionate change)
- In this example, $100\beta_1$ is often called the ***return to education*** (just like an investment). This measure is free of units of measurement of wage (currency, price level)

- Let's revisit the wage equation

$$\begin{aligned} \widehat{\log(wage)} &= 5.66 + 0.039\ educ + 0.006\ IQ \\ n &= 935,\ R^2 = .130 \end{aligned}$$

- These results tell us that…
- Warning: This *R*-squared is not directly comparable to the *R*-squared when ***wage*** is the dependent variable. We can only compare *R*-squared of two models if they have the same dependent variable. The total variation (SSTs) in ***wage$_i$*** and **log(*wage$_i$*)** are completely different

Models involving logarithms: level-log

- We can use logarithmic transformation of ***x*** as well
- y on log of x: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\log(x_1) + \widehat{\beta_2}x_2$
  $\widehat{\beta_1}/100$: the change in predicted y as x_1 increases by 1%, keeping x_2 constant

- Example: The context: determining the effect of cigarette smoking during pregnancy on health of babies. Data: birth weight in kg, family income in $s, mother's education in years, number of cigarettes smoked per week by the mother during pregnancy

$$\widehat{bwght} = 3.22 + 0.050\log(finc) + 0.001educ - 0.013cigs$$
$$n = 1387,\ R^2 = 0.03$$

- The coefficient of log(finc): Consider newborn babies whose mothers have the same level of education and the same smoking habits. Every percentage increase in family income increases the predicted birth weight by 0.0005kg = 0.5g

Models involving logarithms: log-log

- **log of *y*** on **log of *x***: $\widehat{\log(y)} = \widehat{\beta_0} + \widehat{\beta_1}\log(x_1) + \widehat{\beta_2}x_2$
  $\widehat{\beta_1}$: the percentage change in predicted y as x_1 increases by 1%, keeping x_2 constant. $\widehat{\beta_1}$ in this case is also called the estimated elasticity of y with respect to x_1 all else constant.

- Example 6.7 in textbook: Predicting CEO salaries based on sales, market value of the firm (mktval) and years that CEO has been in his/her current position (tenure):

$$\begin{aligned} \widehat{\log(salary)} &= 4.50 + 0.16\log(sales) + 0.11\log(mktval) + 0.01tenure \\ n &= 177,\ R^2 = .318 \end{aligned}$$

The coefficient of log(*sales*): In firms with the exact same market valuation with CEOs who have the same level of experience, a 1 percent increase in sales increases the predicted CEO salary by 0.16%

Considerations for using levels or logarithms (see 6-2a)

1. A variable must have a strictly positive range to be a candidate for logarithmic transformation
2. Thinking about the problem: does it make sense that a unit change in ***x*** leads to a constant change in the magnitude of ***y*** or a constant % change in ***y***?
3. Looking at the scatter plot, if there is only one ***x***
4. Explanatory variables that are measured in years, such as years of education, experience or age, are not logged
5. Variables that are already in percentages (such as interest rate or tax rate) are not logged. A unit change in these variables already is a 1 percent change
6. If a variable is positively skewed (like income or wealth), taking logarithms makes it distribution less skewed

Other non-linear models: Quadratic terms

- We can have x^2 as well as x in a multiple regression model:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x + \widehat{\beta_2}x^2$$

- In this model

$$\frac{\partial \hat{y}}{\partial x} = \widehat{\beta_1} + 2\widehat{\beta_2}x$$

that is, the change in predicted ***y*** as ***x*** increases depends on ***x***

- Here, the coefficients of ***x*** and ***x²*** on their own do not have meaningful interpretations, because…
- $\frac{\partial \hat{y}}{\partial x} = 0$ at $x = -\frac{\widehat{\beta_1}}{2\widehat{\beta_2}}$. At this level of x the predicted y is at its maximum if $\widehat{\beta_2} < 0$, and it is at its minimum if $\widehat{\beta_2} > 0$

Examples of the quadratic model

- Sleep and age: Predicting how long women sleep from their age and education level. Data: age, years of education, and minutes slept in a week recorded by women who participated in a survey

$$\widehat{sleep} = 4428.07 - 49.30age + 0.58age^2 - 13.92educ$$
$$n = 305,\ R^2 = 0.03$$

- Keeping education constant, the predicted sleep reaches its minimum at the age $\frac{49.3}{2 \times 0.58} = 42.5$
- House price and distance to the nearest train station: Data: price (000$s), area (m²), number of bedrooms and distance from the train station (km) for 120 houses sold in a suburb of Melbourne in a certain month:

$$\widehat{price} = -29.08 + 1.22area + 63.76beds + 1169.71train - 687.64train^2$$
$$n = 120,\ R^2 = 0.54$$

- The ideal distance from the train station is $\frac{1169.71}{2 \times 687.64} = 0.85$km because…

Considerations for using a linear or quadratic model

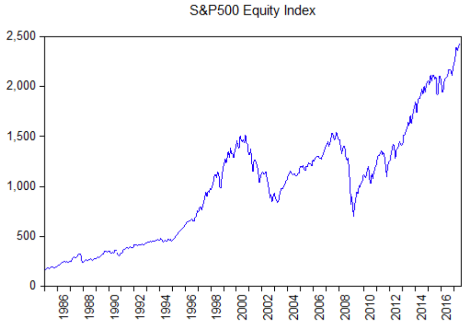
1. Thinking about the problem: is a unit increase in ***x*** likely to lead to a constant change in ***y*** for all values of ***x***, or is it likely to lead to a change that is increasing or decreasing in ***x***?
2. Is there an optimal or peak level of ***x*** for ***y***? Example: wage and age, house price and distance to train station
3. If there is only one ***x***, looking at scatter plot can give us insights
4. In multiple regression, there are tests that we can use to check the specification of the functional form (RESET test, to be covered later)
5. When in doubt, we can add the quadratic term and check its statistical significance, or see if it improves the adjusted R²

Transformation of persistent time series data

- A number of economic and financial series, such as interest rates, foreign exchange rates, price series of an asset tend to be highly persistent
- This means that the past heavily affects the future (but not vice versa)
- A time series can be subject to different types of persistence (deterministic or stochastic)
- A common feature of persistence is lack of mean-reversion. This is evident by visual inspection of a line chart of the time series

Empirical example

- E.g. Below is displayed the Standard and Poors Composite Price Index from January 1985 to July 2017 (monthly observations)



Transformation of persistent time series data (cont.)

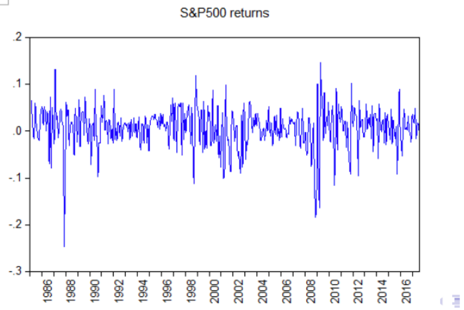
- In such cases the researcher transforms the time series by (log) differencing over the preceding period
- The transformed series is then easier to handle and has more attractive statistical properties
- More precisely, assume that the S&P price index at time t is denoted by P_t
- The said log differencing is expressed as:

$$\log(P_t) - \log(P_{t-1}) = \log(\frac{P_t}{P_{t-1}})$$

$$= \log(1 + \frac{P_t - P_{t-1}}{P_{t-1}}) = \log(1 + r_t) \approx r_t,$$

where 100 x r_t denotes the $\%\Delta P_t$ (for small r_t)

- By differencing the logarithmic transformation to our S&P500 price series, we obtain the S&P500 returns (i.e. log-returns)



Model Selection Criteria

- ***Parsimony*** is very important in predictive analytics (which includes forecasting). You may have heard about the KISS principle. If not, google it!
- We want models that have predictive power, but are as parsimonious as possible
- We cannot use R^2 to select models, because R^2 always increases as we make the model bigger, even when we add irrelevant and insignificant predictors
- One can use t-stats and drop insignificant predictors, but when there are many predictors, and several of them are insignificant, the model that we end up with depends on which predictor we drop first

- Model selection criteria are designed to help us with selecting among competing models
- All model selection criteria balance the (lack of) fit of the model (given by its sum of squared residuals) with the size of the model (given by the number of parameters)
- These criteria can be used when modelling time series data as well
- There are many model selection criteria, differing on the penalty that they place on the lack of parsimony

1. Adjusted R2 (also known as R2)

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

2. Akaike Information Criteria (AIC)

$$AIC = c_1 + \ln(SSR) + 2k/n$$

3. Hannan-Quinn Criterion (HQ)

$$HQ = c_2 + \ln(SSR) + 2k\ln(\ln(n))/n$$

4. Schwarz or Bayesian Information Criterion (SIC or BIC)

$$BIC = c_3 + \ln(SSR) + k\ln(n)/n$$

- c_1, c_2 and c_3 are const constants that do not depend on the fit or number of parameters, so play no important role. ln is the natural logarithm. Also, all models are assumed to include an intercept

- BIC gives the largest penalty to lack of parsimony, i.e. if we use BIC to select among models, the model we end up with would be the same or smaller than the model that we end up with if we used any of the other criteria
- The order of the penalties that each criterion places on parsimony relative to fit is (for n >16)

$$P(BIC) > P(HQ) > P(AIC) > P(\bar{R}^2)$$

- Remember that with BIC, HQ or AIC, we choose the model with the smallest value for the criterion, whereas with $\bar{R}^2$, we choose the model with the largest $\bar{R}^2$
- Different software may report different values for the same criterion. That is because some include c_1, c_2 and c_3 and some don't. The outcome of the model selection exercise does not depend on these constants, so regardless of the software, the final results should be the same.

- Example: Making an app to predict body fat using height (H), weight (W) and abdomen circumference (A)

Predictors	R^2	$\bar{R}^2$	*AIC*	*HQ*	*SC*
W	0.376	0.373	6.474	6.485	6.502
A	0.662	0.661	5.860	5.872	5.888
H	0.008	0.004	6.937	6.949	6.965
W, A	0.719	0.716	**5.685**	**5.702**	**5.727**
W, H	0.461	0.457	6.334	6.351	6.376
A, H	0.688	0.686	5.788	5.805	5.830
W, A, H	0.721	**0.718**	**5.685**	5.707	5.741

- For different class of models, experts use different criteria
- My favourite is HQ (because of my research on multivariate time series, and because Ted Hannah was a great Australian statistician), although not all software report HQ

Confidence intervals for the Conditional Mean versus Prediction Intervals

- Remember that the population model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \text{ for all } i \qquad (1)$$

with the CLM assumptions implying that

$$E(y_i \mid x_{i1}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \text{ for all } i \qquad (2)$$

- Our estimated regression model provides

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}, \text{ for all } i \qquad (3)$$

- Comparing (3) and (2), we see that $\hat{y}_i$ gives us the best estimate of the conditional expectation of y_i given $x_{i1}, \ldots, x_{ik}$
- Also, since u_i is not predictable given $x_{i1}, \ldots, x_{ik}$, $\hat{y}_i$ is also our best prediction for y_i

Solution 1: Robust Standard Errors

- Since OLS estimator is still unbiased, we may be happy to live with the OLS even if it is not BLUE. But the real practical problem is that t- and F-statistics based on OLS standard errors are unusable
  - ▸ Recall the derivation of $Var(\widehat{\beta} \mid \mathbf{X})$:

$$Var(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} [\mathbf{X}' Var(\mathbf{u} \mid \mathbf{X}) \mathbf{X}] (\mathbf{X}'\mathbf{X})^{-1}$$

  - ▸ With homoskedasticity,

$$Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n \Rightarrow Var(\widehat{\beta} \mid \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

  - ▸ With HTSK

$$Var(\mathbf{u} \mid \mathbf{X}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

  - ▸ Therefore, with HTSK

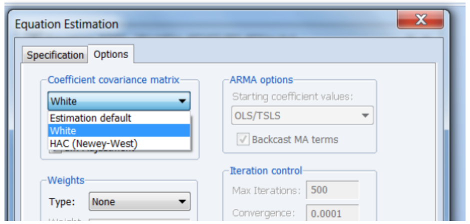
$$Var(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \mathbf{X} \right] (\mathbf{X}'\mathbf{X})^{-1}$$

  - ▸ Amazingly, White proved that:

$$\widehat{Var}(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' \begin{pmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_n^2 \end{pmatrix} \mathbf{X} \right] (\mathbf{X}'\mathbf{X})^{-1}$$

    is a reliable estimator for $Var(\widehat{\beta} \mid \mathbf{X})$ in large samples

- The square root of the diagonal elements of this matrix are called White standard errors or robust standard errors, which most statistical packages compute. These are reliable for inference
- Back to the example. The option of robust standard errors is under the Options tab of the equation window:



Equation Estimation

Specification | Options

Coefficient covariance matrix: White
- Estimation default
- White
- HAC (Newey-West)

ARMA options
Starting coefficient values: OLS/TSLS
☑ Backcast MA terms

Weights
Type: None

Iteration control
Max Iterations: 500
Convergence: 0.0001

- With this option, we get:

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.2042	19.7337	-0.0610	0.9513
INC	0.8248	0.1039	7.9408	0.0000
AGE	-1.3218	1.1055	-1.1956	0.2320
AGE^2	0.0256	0.0141	1.8066	0.0710

R-squared	0.1229	Mean dependent var	13.5950
F-statistic	93.9855	Durbin-Watson stat	1.9576
Prob(F-statistic)	0.0000	Wald F-statistic	40.1225
Prob(Wald F-statistic)	0.0000		

- Compared with the original regression results:

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.2042	15.2807	-0.0788	0.9372
INC	0.8248	0.0603	13.6790	0.0000
AGE	-1.3218	0.7675	-1.7222	0.0852
AGE^2	0.0256	0.0090	2.8406	0.0045

R-squared	0.1229	Mean dependent var	13.5950
Adjusted R-squared	0.1216	S.D. dependent var	47.5906
S.E. of regression	44.6045	Akaike info criterion	10.4355
F-statistic	93.9855	Durbin-Watson stat	1.9576
Prob(F-statistic)	0.0000		

Solution 2: Transform the Model

a. ***Logarithmic transformation of y may do the trick:*** If the population model has log(y) as the dependent variable but we have used y, this kind of mis-specification can show up as heteroskedastic errors. So, if log-transformation is admissible (i.e. if y is positive), moving to a log model may solve the problem, and the OLS estimator on the log-transformed model will then be BLUE and standard errors will be useful. Of course when we consider transforming y, we should think if a log-level or a log-log model makes better sense

b. ***Weighted least squares:*** When there is good reason to believe that variance of each error is proportional to a known function of a single independent variable, then we can transform the model in a way to eliminate HTSK and then use OLS on the transformed model. This estimator is the weighted least squares (WLS) estimator, which we derive on the next slide.

Weighted Least Squares

- Suppose the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \text{ for } i = 1, \ldots, n \qquad (1)$$

  satisfies the assumptions needed for unbiasedness of OLS, and we have

$$Var(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 h_i$$

  where h_i is a known function of one of x's, or a function of a variable z as long as $E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}, z_i) = 0$. For example $h_i = x_{i1}$, or $h_i = x_{i1}^2$, or $h_i = 1/z_i$.
- Multiplying both sides of equation (1) by $w_i = \frac{1}{\sqrt{h_i}}$ eliminates HTSK because:

$$\frac{1}{\sqrt{h_i}} y_i = \beta_0 \cdot \underline{\frac{1}{\sqrt{h_i}}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{1}{\sqrt{h_i}} u_i$$

- The transformed (or "weighted") model:

$$(w_i y_i) = \beta_0 w_i + \beta_1 (w_i x_{i1}) + \beta_2 (w_i x_{i2}) + \cdots + \beta_k (w_i x_{ik}) + (w_i u_i) \quad (2)$$

  satisfies all assumptions of the Gauss-Markov theorem, so the OLS estimator of its parameters is BLUE.
- More importantly, equation (2) has the same parameters as equation (1). So, OLS on the weighted model will produce BLUE of β_0 to β_k and we can test any hypotheses on these parameters based on the weighted model.
- Note that the transformed model does not have a constant term, and β_0 is the coefficient of w_i in the transformed model
- This estimator is called the weighted least squares (WLS) estimator of β

- In the financial wealth example, the auxiliary regression suggests that the variance changes with income. Since income is positive for all observations (why is this important?), we hypothesise that Var (u_i | inc_i, age_i) = $\sigma^2 inc_i$
- We create $w_i = \frac{1}{\sqrt{inc_i}}$ [Eviews command: `series w=1/@sqrt(inc)`] and we run the weighted regression

Dependent Variable: W*NETTFA
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
W	-3.7769	11.4808	-0.3290	0.7422
W*INC	0.7938	0.0627	12.6587	0.0000
W*AGE	-0.8890	0.5756	-1.5444	0.1227
W*(AGE^2)	0.0174	0.0067	2.5864	0.0098

R-squared	0.0810	Mean dependent var	2.1807

- The standard errors are now reliable for inference and for forming confidence intervals