ECO375 Midterm Review

Matt Tudball

University of Toronto St. George

October 17, 2017

ECO375 Midterm Review

Welcome back!

Today's coverage:

- Review of statistical terminology
- Least squares estimation (matrix form included)
- MLR assumptions and properties of the least squares estimator
- Inference (hypothesis testing)
- Selected topics: Heteroscedasticity, non-linear functional forms, linear probability model, etc.

Do not exclusively study from these slides. I do not cover everything that may appear on the midterm.

The Statistical Problem

- **Data Generating Process:** The data generating process is the true underlying mechanism that is generating the data we observe in our sample. We often assume that it takes a particular functional form. For example,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
  Given $x_{i1}, x_{i2}, ..., x_{ik}$ this is the (unobservable) process which is generating the data y_i.
- **Parameters:** The above example is a **parametric model**, meaning that it can be characterised by a set of parameter values, $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$.
- The **statistical problem** is to use data coming out of the data generating process in order to make inferences about the true underlying parameters; in our example $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$.

Least Squares Estimation

- **Estimator:** An estimator is a statistic (i.e. a function of the data) that is used to infer the value of an unknown parameter in a statistical model.
- **Residual:** A residual is the difference between an observed value y_i and the fitted (or predicted) value $\hat{y}_i$ implied by an estimator, often denoted $\hat{u}_i = y_i - \hat{y}_i$.
- Given the parametric model assumed in the previous slide, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$ and some data drawn from it, $\{y_i, x_{i1}, x_{i2}, ..., x_{ik}\}_{i=1}^n$, a natural way of calculating our estimators for the true parameters $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$ is by a process called least squares.
- The intuition behind least squares is that we want to choose estimators $\{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k\}$ to minimise the sum of square residuals, which can be thought of as the 'error' in our prediction.

Least Squares Estimation

- We can write the sum of square residuals in the following form, which we call our **objective function**,

$$\min_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

- Naturally to find $\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\}$ which minimise this objective function, we need to take the first derivative with respect to each of the estimators, set each derivative equal to 0 and then solve the resulting system of $k+1$ equations in $k+1$ unknowns.

$$0 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i2} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$\vdots$$

$$0 = \sum_{i=1}^{n} x_{ik} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

Least Squares Estimation

- We can simplify the system of equations in the previous slide considerably by recognising that,
  $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_k x_{ik} = y_i - \hat{y}_i = \hat{u}_i$.
- Let's rewrite it,
  $0 = \sum_{i=1}^{n} \hat{u}_i$
  $0 = \sum_{i=1}^{n} x_{i1} \hat{u}_i$
  $0 = \sum_{i=1}^{n} x_{i2} \hat{u}_i$
  $\vdots$
  $0 = \sum_{i=1}^{n} x_{ik} \hat{u}_i$
- This implies that the sample mean of $\hat{u}_i$ is always equal to 0 and that the sample covariance between $\hat{u}_i$ and each of the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ is also 0.
- This does **not** imply anything about the properties of the true population error term u_i. $\hat{u}_i$ only has these properties due to the way in which the least squares objective function is solved.

- It is very difficult to explicitly solve this system of $k+1$ unknowns. If $k=1$, however, then we are running a **simple regression** and it is relatively straight-forward to solve for $\{\hat{\beta}_0, \hat{\beta}_1\}$.
- Let's look at the FOCs,
  $0 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \right)$
  $0 = \sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \right)$
- Solving these 2 equations in 2 unknowns, we obtain the following closed-form expressions,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$$

- The least squares prediction for observation i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$. The least squares residual for observation i is $\hat{u}_i = y_i - \hat{y}_i$.

Partialling Out

- What if we are interested in an expression for a particular estimator $\hat{\beta}_1$ when $k > 1$?
- We might have some intuition that if we could "transform" $\hat{\beta}_1$ into a simple regression, then we could find a closed-form expression as in the previous slide. This is the idea behind partialling out.
- Steps:
  1. Regress x_1 on x_2, x_3, ..., x_k and calculate the residual $\hat{r}_1$. We can think of this residual as capturing the unique variation in x_1 that is not explained by x_2, x_3, ..., x_k.
     $x_{i1} = \hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + ... + \hat{\alpha}_k x_{ik} + \hat{r}_{i1}$
     $\hat{r}_1 = x_{i1} - \hat{x}_{i1}$
  2. Regress y on $\hat{r}_1$.
     $y_i = \hat{\lambda}_0 + \hat{\lambda}_1 \hat{r}_{i1} + \hat{\epsilon}_i$
  3. The resulting slope estimate $\hat{\lambda}_1$ is always equal to $\hat{\beta}_1$.

$$\hat{\beta}_1 = \hat{\lambda}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{r}_{i1} - 0)}{\sum_{i=1}^{n}(\hat{r}_{i1} - 0)^2} = \frac{\sum_{i=1}^{n} y_i \hat{r}_{i1}}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

Matrix Form of MLR

- Another approach is to "vectorise" our data and coefficients and then solve for the vector of coefficients rather than each coefficient individually. By doing this we can take advantage of some of the nice properties of matrix algebra.
- Note that,

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times(k+1)}{\boldsymbol{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \underset{(k+1)\times 1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\underset{n\times 1}{\boldsymbol{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

- We therefore assume that our statistical model is $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$

Matrix Form of MLR

- Our objective function can also be written in matrix form as the inner product of the residual vector $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\beta$,

$$
\begin{aligned}
\min_{\hat{\beta}} \boldsymbol{u}^T \boldsymbol{u} &= \min_{\hat{\beta}} (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) \\
&= \min_{\hat{\beta}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T \hat{\beta} \right)^2 \\
&= \min_{\hat{\beta}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)^2
\end{aligned}
$$

where $\boldsymbol{x}_i$ is the $(k+1) \times 1$ vector of explanatory variables for observation i.

- You can see that this matrix formulation of the objective function is equivalent to the objective function we wrote previously.

Matrix Form of MLR

- Let's take first order conditions of this objective function,

$$\min_{\hat{\beta}} \; (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

- $\frac{\partial}{\partial \hat{\beta}}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = 0 \Rightarrow \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = 0.$
- We can now solve for $\hat{\beta}$,

$$\begin{aligned}
\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) &= 0 \\
\left(\boldsymbol{X}^T\boldsymbol{X}\right)\hat{\beta} &= \boldsymbol{X}^T\boldsymbol{y} \\
\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\hat{\beta} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
\boldsymbol{I}_{(k+1)}\hat{\beta} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
\hat{\beta} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

- Note that we must assume that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible in order to solve for $\hat{\beta}$. This is only true if $\boldsymbol{X}$ has full column rank equal to $k+1$.

MLR Assumptions

- Notice that, so far, we haven't made any assumptions about the underlying population from which $\{y_i, x_{i1}, x_{i2}, ..., x_{ik}\}_{i=1}^{n}$ is being drawn. We would calculate these same MLR estimators even if the true data generating process were not linear!
- However, the MLR estimators may not have "good properties" in the absence of certain assumptions over the population.
- What properties do we want our estimators to have?
- **Unbiasedness:** $\mathbb{E}(\hat{\beta}|\boldsymbol{X}) = \beta$
- **Consistency:** $\hat{\beta}$ converges to β as n goes to infinity.
- **Efficiency:** $\hat{\beta}$ has minimum variance (i.e. most precision) within the class of linear estimators.
- **Known Sampling Distribution:** $\hat{\beta}$ has a known finite sample distribution (this will allow us to do inference in finite sample).
- The next slide will cover the assumptions needed for our MLR estimators to achieve these properties.

MLR Assumptions

MLR.1 The model is linear in parameters:
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i = \boldsymbol{x}_i^T \beta + u_i$

MLR.2 We have a random sample $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i)\}_{i=1}^n$ of size n following the model in MLR.1.

MLR.3 None of the independent variables is constant and there is no perfect multicollinearity (i.e. no exact linear relationship among the independent variables). This is the same as $\boldsymbol{X}$ having full column rank equal to $k+1$.

MLR.4 The error u_i has an expected value of zero conditional on all x_i
$\mathbb{E}(u_i | x_{i1}, x_{i2}, ..., x_{ik}) = 0$ for $i = 1, ..., n$.

MLR.5 The error u_i homoscedastic, i.e. it has the same variance for all x_i
$\text{Var}(u_i | x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$

MLR.6 The error u_i is independent of the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ and is normally distributed with mean 0 and variance σ^2
$u_i \sim \mathcal{N}(0, \sigma^2)$

Unbiasedness

- Why is the MLR estimator unbiased under these assumptions?
- First note,

$$\begin{aligned}\hat{\beta} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\beta+\boldsymbol{u}\right) \\ &= \beta+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}\end{aligned}$$

- Then,

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}|\boldsymbol{X}\right) &= \mathbb{E}\left(\beta+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}|\boldsymbol{X}\right) \\ &= \beta+\mathbb{E}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}|\boldsymbol{X}\right) \\ &= \beta+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathbb{E}\left(\boldsymbol{u}|\boldsymbol{X}\right) \\ &= \beta \text{ by MLR.4}\end{aligned}$$

Omitted Variable Bias

- What if we omit a relevant variable? Will $\hat{\beta}$ still be unbiased?
- Let's assume that the actual population relationship between y and a set of x's is,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
  satisfying the assumptions MLR.1 to MLR.4
- Denote the estimator for β_1 and β_k by running this correctly specified regression as $\hat{\beta}_1$ and $\hat{\beta}_k$ respectively.
- But suppose that we accidentally specify the the model,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{k-1} x_{ik-1} + u_i$
  whereby we omit a relevant variable x_k.
- Denote the estimator for β_1 obtained in this misspecified model by $\tilde{\beta}_1$.

Omitted Variable Bias

- By the "partialling out" procedure we know that $\tilde{\beta}_1$ can be written as,

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \tilde{r}_{i1} y_i}{\sum_{i=1}^{n} \tilde{r}_{i1}^2}$$

  where $\tilde{r}_{i1}$ are the sample residuals from the regression,
  $x_{i1} = \alpha_0 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + ... + \alpha_{k-1} x_{ik-1} + r_i$
- It can be shown that,

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \frac{\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik}}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1 \end{aligned}$$

  where $\tilde{\delta}_1$ is the "partialling out" coefficient for x_1 from the regression of x_k on x_1, x_2, ..., x_{k-1}.
- The bias in $\tilde{\beta}_1$ is
  $\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1 - \beta_1 | \boldsymbol{X}) = \text{E}(\hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1 - \beta_1 | \boldsymbol{X}) = \beta_1 + \beta_k \tilde{\delta}_1 - \beta_1 = \beta_k \tilde{\delta}_1$.

Known Sampling Distribution

- Why does the OLS estimator have a known sampling distribution under these assumptions?
- Note that MLR.6 assumes that $\boldsymbol{u}$ is normally distributed, $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.
- Since $\hat{\beta} = \beta + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}$ this means that,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \text{Var}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}\right)\right)$$

- Note that,

$$\begin{aligned}
\text{Var}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{u}\right) &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{u})\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)^T \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{u})\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
\end{aligned}$$

Known Sampling Distribution

- Note that the j^{th} diagonal element of $\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ describes the variance of the estimator $\hat{\beta}_j$. We will call the square root of this diagonal element the **standard error** of $\hat{\beta}_j$.
- Note that by MLR.6,

$$\frac{\hat{\beta}_j-\beta_j}{se\left(\hat{\beta}_j\right)} \sim \mathcal{N}\left(0,1\right)$$

- However, this expression is a function of the true variance of $\boldsymbol{u}$ denoted by σ^2, which we cannot observe.
- We must therefore use an estimator for σ^2 (not derived here) which is usually taken to be $\widehat{\sigma}^2 = \frac{1}{n-k}\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = \frac{1}{n-k}\sum_{i=1}^n \hat{u}_i^2$
- We denote the estimator for the standard error of $\hat{\beta}_j$ using $\widehat{\sigma}^2$ as $\widehat{se}(\hat{\beta}_j)$.
- Note that again by MLR.6,

$$\frac{\hat{\beta}_j-\beta_j}{\widehat{se}\left(\hat{\beta}_j\right)} \sim t_{n-k-1}$$

which is the t-distribution on $n-k-1$ degrees of freedom.

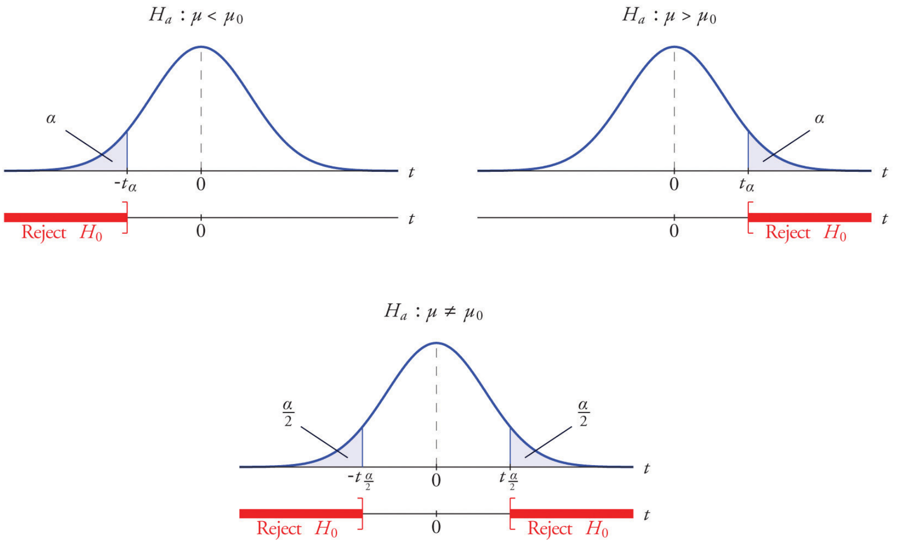
Hypothesis Testing: t-test

- It is very useful to know that the **t-statistic** $t = \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \sim t_{n-k-1}$ since the t-distribution, given a **null hypothesis** value of β_j, does not depend on any of the unobserved true parameter values.
- This means that we can start to ask questions about the likelihood of observing a given estimate $\hat{\beta}_j$ given a null hypothesis value of β_j. This is the main intuition behind hypothesis testing.
- The simple hypothesis test is related to the value of a single parameter value,
- **Two-Sided Hypothesis Test:** $\text{H}_0 : \beta_j = a$ and $\text{H}_1 : \beta_j \neq 0$ where a is some fixed constant, H_0 is the null hypothesis and H_1 is the **alternative hypothesis**.
- **One-Sided Hypothesis Test:** $\text{H}_0 : \beta_j > a$ and $\text{H}_1 : \beta_j < a$ or $\text{H}_0 < a$ and $\text{H}_1 : \beta_j > a$.

Hypothesis Testing: t-test

- How do we actually test these null and alternative hypotheses?
- We need to specify one more parameter, called the **significance level** and usually denoted by α. α represents the probability under the null of rejecting the null hypothesis when it is actually true.
- From the significance level α we can obtain a **critical value** T which is a value on the t-distribution such that, if we observed a t-statistic more extreme than this, we would reject the null hypothesis.
- For a two-sided test, we obtain a critical value $T_{\alpha/2}$ such that if $|t| > |T_{\alpha/2}|$ then we would reject the null hypothesis, where $|\cdot|$ denotes absolute value.
- For a one-sided test $H_0 : \beta_j > a$ ($H_0 : \beta_j < a$), we obtain a critical value T_α such that if $t > T_\alpha$ ($t < T_\alpha$) we would reject the null hypothesis.

Hypothesis Testing: t-test



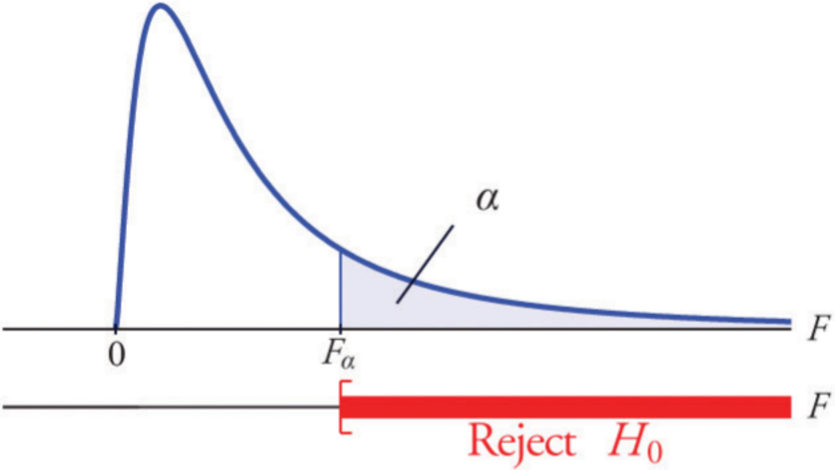
Hypothesis Testing: F-test

- What if we want to test hypotheses over multiple estimates? For example, $\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
- For hypotheses of this form we will use an **F-test**. The **F-statistic** takes the form,

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)} \sim F_{q,n-k-1}$$

  - SSR_r denotes the sum of squared residuals of the **restricted model** in which we assume the null hypothesis is true. R^2_r is the R^2 from this restricted model.
  - SSR_{ur} denotes the sum of squared residuals of the **unrestricted model** in which we make no assumptions over the coefficients in the null hypothesis. R^2_{ur} is the R^2 from this unrestricted model.
  - q is the number of restrictions being tested. In the example above, it is 3.
  - $n-k-1$ is the degrees of freedom of the unrestricted model.

Hypothesis Testing: F-test



Heteroscedasticity

- All of our hypothesis tests so far have assumed **homoscedasticity** of the error term $\boldsymbol{u}$. This says that,

$$\text{Var}(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- This assumption may be violated in practice. In particular, when the error term is **heteroscedastic**, we assume that for each observation i, $\text{Var}(\boldsymbol{u}_i|\boldsymbol{x}_i) = \sigma_i^2$. Then,

$$\text{Var}(\boldsymbol{u}|\boldsymbol{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Heteroscedasticity

- What problems does heteroscedasticity cause?
- It doesn't affect the unbiasedness of the OLS estimator since MLR.4 still holds.
- It will, however, mean that our t-statistics do not follow a t-distribution and our F-statistics do not follow an F-distribution. Therefore, we cannot make inferences using the methods developed in the previous slides.
- What can we do?
- One solution is to use **White (robust) standard errors**. For the simple regression model $y_i = \beta_0 + \beta_1 x_i + u$ under heteroscedasticity, the variance of the estimator is $\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$
- A natural way to find an estimator for the variance is to find an estimator for σ_i^2. What is the simplest estimator? $\hat{u}_i^2$ of course! This looks like $\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$

Heteroscedasticity: GLS and FGLS

- Suppose u_i is the true error term with variance σ_i^2. We can think that there exists a transformation p_i such that $u_i^* = p_i u_i$ has constant variance σ^2. This is known as **Generalised Least Squares (GLS)**.
- A popular choice of transformation is $p_i = 1/\sqrt{h_i}$ where $\sigma_i^2 = \mathbb{E}(u_i^2) = h_i \sigma^2$.
- Then if we transform the data in our OLS regression by $p_i = 1/\sqrt{h_i}$ we have that $\mathbb{E}(u_i^{*2}) = \mathbb{E}(u_i^2/h_i) = \mathbb{E}(u_i^2)/h_i = \sigma^2 h_i/h_i = \sigma^2$.
- The problem with this approach is that we need to know the functional form of h_i.
- A more tractable approach is to model heteroscedasticity as a function of $\boldsymbol{x}_i$. This is called **Feasible Generalised Least Squares (FGLS)**. Here are the steps,
  1. Estimate the original model by OLS and recover $\hat{u}_i$.
  2. Estimate the model $ln\left(\hat{u}_i\right) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_k x_{ik} + e$
  3. Set $\hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + ... + \hat{\alpha}_{ik})$
  4. Transform the model using weights $1/\sqrt{\hat{h}_i}$ and estimate by OLS.

Selected Topics: Dummy Independent Variable

- Suppose that x_i is binary (i.e. it only takes on values 0 or 1). This is called a **dummy variable**. How would we interpret a linear regression $y_i = \beta_0 + \beta_1 x_i + u_i$?
- Note that $\mathbb{E}(y_i|x_i = 1) = \beta_0 + \beta_1$ and $\mathbb{E}(y_i|x_i = 0) = \beta_0$.
- Thus $\beta_1 = \mathbb{E}(y_i|x_i = 1) - \mathbb{E}(y_i|x_i = 0)$ and $\beta_0 = \mathbb{E}(y_i|x_i = 0)$.
- Interpretation of this regression is therefore very simple. Suppose that x_i is completing a university degree and y_i is hourly wages.
- We can say that completing a university degree corresponds to a β_1 increase in hourly wages on average.
- We can also say that β_0 is average hourly wages among those without a university degree.

Selected Topics: Linear Probability Model

- Suppose now that y_i is binary and $y_i = \beta_0 + \beta_1 x_i + u_i$. This is called the **Linear Probability Model**.
- Note that $\mathbb{E}(y_i|x_i) = \mathbb{P}(y_i = 1|x_i) = \beta_0 + \beta_1 x_i$.
- There is nothing restricting the predicted probability to be between 0 and 1.
- Note also that, since y_i is a Bernoulli random variable, $\text{Var}(y_i|x_i) = \text{Var}(u_i|x_i) = \mathbb{P}(y_i = 1|x_i)\left(1 - \mathbb{P}(y_i = 1|x_i)\right) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)$.
- Therefore the errors in the LPM necessarily heteroscedastic.
- We can interpret the coefficient as $\beta_1 = \frac{\partial \mathbb{P}(y_i=1|x_i)}{\partial x_i}$, so it is the change in the likelihood of observing $y_i = 1$ with respect to x_i.

Selected Topics: Log Regressions

- Often when working with real data we will come across non-linear relationships which violate MLR.1. However, it is occasionally possible to transform the data into a linear relationship.
- **Log-Linear Regression:** Suppose it looks like $y_i = e^{\beta_0 + \beta_1 x_i + u_i}$. Then if we apply the natural log to both sides we obtain $\ln(y_i) = \beta_0 + \beta_1 x_i + u_i$. We interpret the coefficient β_1 by saying "a 1 unit increase in x is associated with a $100\beta_1$% change in y".
- **Log-Log Regression:** A log-log model is useful when we think the relationship is a constant elasticity model $y_i = e^{\beta_0} x_i^{\beta_1} e^{u_i}$. Then if we apply the natural log to both sides we obtain $\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + u_i$. We interpret the coefficient β_1 by saying "a 1% increase in x is associated with a β_1% change in y"
- **Linear-Log Regression:** Suppose it looks like $e^{y_i} = e^{\beta_0} x_i^{\beta_1} e^{u_i}$. Then if we apply the natural log to both sides we obtain $y_i = \beta_0 + \beta_1 \ln(x_i) + u_i$. We interpret the coefficient β_1 by saying "a 1% increase in x is associated with a $\beta_1/100$ unit change in y"