NOVA IMS INFORMATION MANAGEMENT SCHOOL

The Linear Regression Model.

Statistics for Data Science

Bruno Damásio
bdamasio@novaims.unl.pt
@bmpdamasio

2025/2026

Table of contents

Introduction

What is Regression Analysis?

Regression analysis is a discipline that "aims to give empirical content to relations". Specifically, it allows to uncover a relationship between two or more variables.

Cross-Sectional Data

- Sample of individuals, households, firms, cities, states, countries, etc. taken at a given point in time.
- An important feature of cross-sectional data: they are obtained by random sampling from the underlying population.
- For example, suppose that y_i is the i-th observation of the dependent variable and x_i is the i-th observation of the explanatory variable
- Random sampling means that $\{(y_i, x_i)\}$ is an independent sequence.
- This implies that for $i \neq j$

$$\text{Cov}(y_i, y_j) = \text{Cov}(x_i, x_j) = \text{Cov}(y_i, x_j) = 0$$

Cross-Sectional Data

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Figure 1: Cross-sectional data on growth rates and country characteristics

Causality and Ceteris Paribus

Definition of causal effect of x on y:

> "How does variable y change if variable x is changed but all other relevant factors are held constant?"

- Most questions are ceteris paribus questions
- It is important to define which causal effect one is interested in
- It is useful to describe how an experiment would have to be designed to infer the causal effect in question

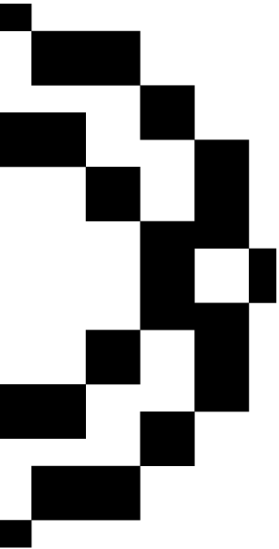
Causality and Ceteris Paribus

Measuring the return to education

- *If a person is chosen from the population and given another year of education, by how much will his or her wage increase?*
- Implicit assumption: all other factors that influence wages such as experience, family background, intelligence etc. are held fixed

Experiment:

- Choose a group of people; randomly assign different amounts of education to them (infeasible!); compare wage outcomes
- Problem without random assignment: amount of education is related to other factors that influence wages (e.g. intelligence)



The Linear Regression Model

The SLRM

Assume that we have have two (observable) variables y and x. And that we are interested in:

- explaining y in terms of x
- studying how y varies with changes in x

In writing down a model that will "explain y in terms of x", we must confront three issues.

1. how do we allow for other factors to affect y?
2. what is the functional relationship between y and x?
3. how can we be sure we are capturing a ceteris paribus relationship between y and x?

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

Equation (1), which is assumed to hold in the population of interest, defines the simple linear regression model.

The SLRM

Table 1: Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

The error term

- The variable u is called the error term, random disturbance ("disturbs an otherwise stable relationship").
- Represents unobserved factors other than x that affect y.
- A simple regression treats all factors affecting y other than x as being unobserved (Why?).
- The disturbance arises for several reasons (Discuss).

The SLRM

- Equation (1) also addresses the issue of the functional relationship between y and x.
- If the other factors in u are held fixed, and assuming $\Delta u = 0$, then x has a linear effect on y:

$$\Delta y = \beta_1 \Delta x$$

- This means that β_1 is the slope parameter in the relationship between y and x, holding the other factors in u fixed; it is of primary interest in applied economics.
- The intercept parameter β_0, sometimes called the constant term, also has its uses, although it is rarely central to an analysis.
- The linearity of (1) implies that a one-unit change in x has the same effect on y, regardless of the initial value of x (This is unrealistic for many economic applications).

The SLRM

- Example: A simple wage function

$$wage = \beta_0 + \beta_1 educ + u$$

(Discuss)

Conditional mean assumption

- When is there a causal interpretation?
- Conditional mean independence assumption

$$\mathrm{E}[u \mid x] = 0$$

  Unobserved factors do not contain information on explanatory variables.
- The conditional mean independence assumption in the wage function is unlikely to hold because individuals with more education will also be more intelligent on average.

Orthogonality conditions

- Now that we have discussed the basic ingredients of the simple regression model, we will address the important issue of how to estimate β_0 and β_1
- We have

$$\mathrm{E}[u \mid x] = 0 \Rightarrow \begin{cases} \mathrm{E}[xu] & = 0 \\ \mathrm{E}[u] & = 0 \end{cases} \Leftrightarrow \begin{cases} \mathrm{E}[x(y - \beta_0 - \beta_1 x)] & = 0 \\ \mathrm{E}[y - \beta_0 - \beta_1 x] & = 0 \end{cases}$$

Method of moments approach

- Since there are two unknown parameters to estimate, we might hope that previeos equations can be used to obtain good estimators for the parameters β_0 and β_1
- Assuming a random sample of n individuals and using the method of the moments we have:

$$n^{-1} \sum x_i \left(y - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{2}$$

and

$$n^{-1} \sum \left(y - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{3}$$

The OLS estimator in the SLRM

Using the basic properties of the summation operator we have:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

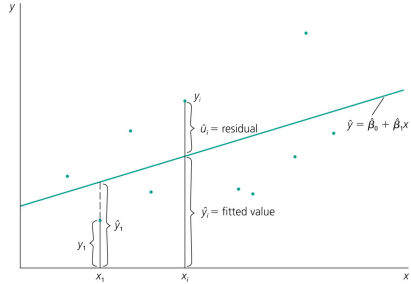
and

$$\hat{\beta}_1 = \frac{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5}$$

or, the sample covariance between x and y divided by the sample variance of x.

The OLS estimator in the SLRM

- The estimates given in 4 and 5 are called the ordinary least squares (OLS) estimates of β_0 and β_1.
- To justify this name, for any $\hat{\beta}_0$ and $\hat{\beta}_1$, define a fitted value for y when $x = x_i$: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The residual for observation i is can be defined as $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Fit as good as possible a regression line through the data points:



Minimizing the SSR

Put otherwise, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the sum of squared residuals,

$$\text{SSR}(\hat{\beta}_0, \hat{\beta}_1) \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{6}$$

The foc for the minimum are found by setting the partial derivatives of (6) equal to zero:

$$\begin{aligned} \frac{\delta}{\delta \hat{\beta}_0} \text{SSR}(\hat{\beta}_0, \hat{\beta}_1) &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \frac{\delta}{\delta \hat{\beta}_1} \text{SSR}(\hat{\beta}_0, \hat{\beta}_1) &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i \end{aligned}$$

Normal Equations

And these yield so called normal equations:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{7}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{8}$$

Solutions

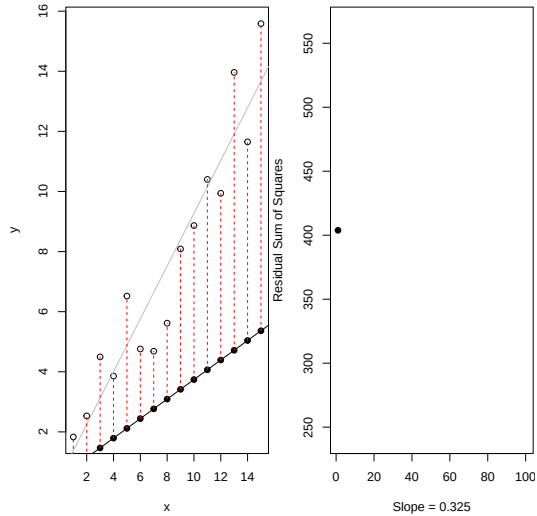
Again, we have:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9}$$

and

$$\hat{\beta}_1 = \frac{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{x})}{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{10}$$

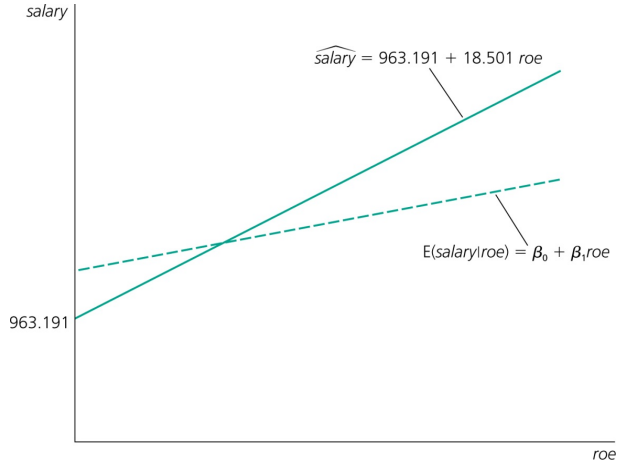
Why OLS?



Example in R

```
data(ceosal1, package='wooldridge')

# OLS regression
lm( salary ~ roe, data=ceosal1 )

##
## Call:
## lm(formula = salary ~ roe, data = ceosal1)
##
## Coefficients:
## (Intercept)          roe
##       963.2         18.5
```

PRF vs Fitted Function



Example in R

```
data(ceosal1, package='wooldridge')
attach(ceosal1)

cov(roe,salary)
var(roe)
mean(salary)
mean(roe)
( b1hat <- cov(roe,salary)/var(roe) )
( b0hat <- mean(salary) - b1hat*mean(roe) )
detach(ceosal1)
```

Goodness of fit: The R-square

- It is often useful to compute a number that summarizes how well the OLS regression line fits the data.
- The R-square (coefficient of determination):

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

- the ratio of the explained variation compared to the total variation;
- It is interpreted as the fraction of the sample variation in y that is explained by x.
- A value of R^2 that is nearly equal to zero indicates a poor fit of the OLS line: very little of the variation in the y_i is captured by the variation in the $\hat{y}_i$.

Goodness of fit: The R-square

```
data(ceosal1, package='wooldridge')
CEOregres <- lm( salary ~ roe, data=ceosal1 )
# Calculate predicted values & residuals:
sal.hat <- fitted(CEOregres)
u.hat <- resid(CEOregres)
```

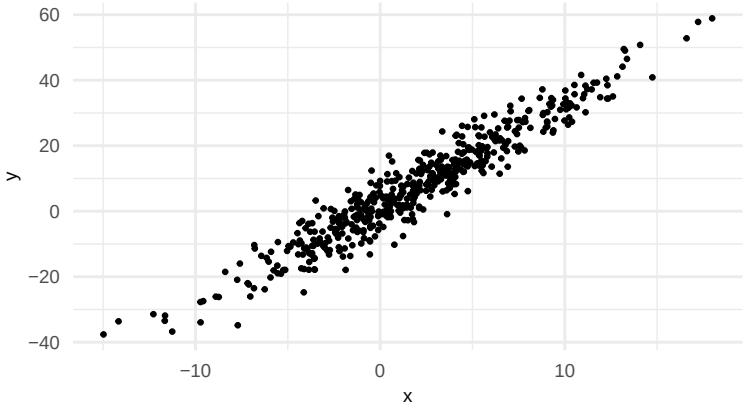
Goodness of fit: The R-square

```
# Calculate R^2 in three different ways:
sal <- ceosal1$salary
var(sal.hat) / var(sal)

## [1] 0.01318862

1 - var(u.hat) / var(sal)

## [1] 0.01318862

cor(sal, sal.hat)^2

## [1] 0.01318862
```

Functional Form

- Linear relationships are not nearly general enough for all applications.
- Fortunately, it is rather easy to incorporate many nonlinearities into simple regression analysis by appropriately defining the dependent and independent variables
- Logarithmic transformation is one of the most applied transformation for economic variables.

Level-Level

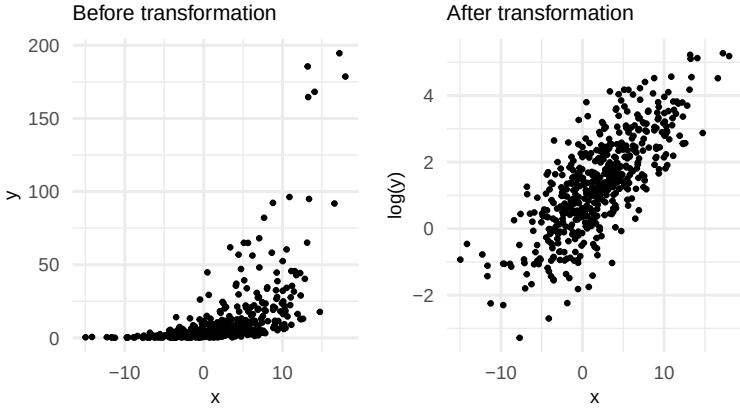


In this plot, we have a linear relationship between x and y. No transformation is needed!

Level-Level interpretation

The interpretation in the level-level specification is: "If x increases one unit, it is expected that the y increases $\hat{\beta}_1$ units."

Log-Level

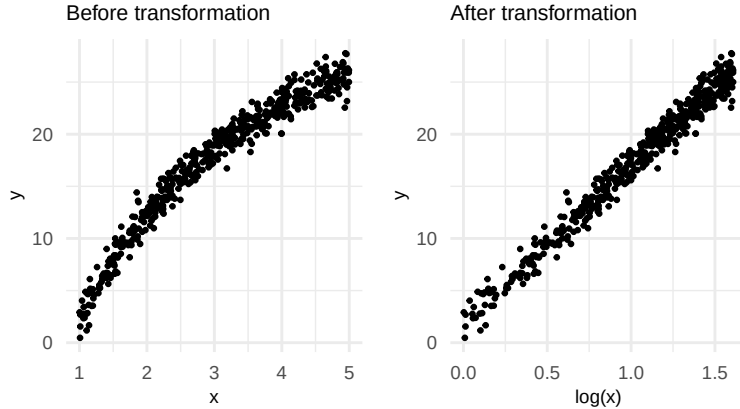


Before transformation, we clearly have a non-linear relationship, similar to an exponential function. After transformation, we have a linear relationship.

Log-Level interpretation

The interpretation in the log-level specification is: "If x increases one unit, it is expected that the y increases $\hat{\beta}_1 \times 100$ %."

Level-Log

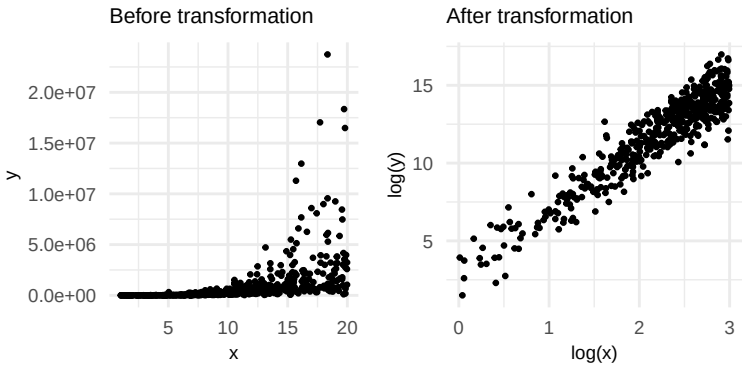


This time, before transformation, we have a non-linear relationship, similar to an logarithm function. After transformation, we have a linear relationship.

Level-Log interpretation

The interpretation in the level-log specification is: "If x increases 1%, it is expected that the y increases $\hat{\beta}_1/100$ units ."

Log-Log



Finally, in a log-log specification, we have a linear relationship, after transforming both variables.

Log-Log interpretation

The interpretation in the log-log specification is: "If x increases 1%, it is expected that the y increases $\hat{\beta}_1$ % ."

Units of Measurement and Functional Form

Table 2: Summary of Functional Forms Involving Logarithms

Model	Dep. Variable	Regressor	Interpretation of β_1
Level-level	y	x	$\Delta \mathrm{E}[y \mid x_i] = \beta_1 \Delta x$
Log-level	$\log(y)$	x	$\%\Delta \mathrm{E}[y \mid x_i] \approx 100\beta_1 \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta \mathrm{E}[y \mid x_i] \approx \beta_1 \%\Delta x$
Level-Log	y	$\log(x)$	$\Delta \mathrm{E}[y \mid x_i] \approx (\beta_1/100)\%\Delta x$

(Board)

Remark: If the parameter are estimated (which is always the case) we need to add the term "estimated". For example, "The estimated change in $\mathrm{E}[y_i \mid x_i]$ ...".

The Multiple Linear Regression Model

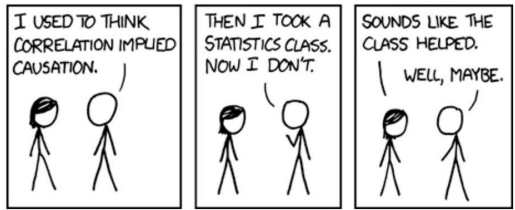
Ceteris Paribus or Holding Other Factors Fixed

- The primary drawback in using simple regression is that it is very difficult to draw ceteris paribus conclusions about how x affects y
- The key assumption, $\mathrm{E}[u \mid \mathbf{x}]$ —that all other factors affecting y are uncorrelated with x is often unrealistic
- Multiple regression analysis is more amenable to ceteris paribus analysis because it allows us to explicitly control for many other factors that simultaneously affect the dependent variable.

Ceteris Paribus or Holding Other Factors Fixed

Ceteris Paribus: "other (relevant) factors being equal", plays an important role in causal analysis.

Ceteris Paribus or Holding Other Factors Fixed



Ceteris Paribus or Holding Other Factors Fixed

- Suppose that wages depend on education and labor force experience. Your goal is to measure the "return to education".
- If your analysis involves only wages and education you may not uncover the ceteris paribus effect of education on wages.
- Consider the following data:

monthly wages (Euros)	years of experience	years of education
1500	6	9
1500	0	15
1600	1	15
2000	8	12
2500	10	12

Ceteris Paribus or Holding Other Factors Fixed

- Previous example shows that the focus

$$\frac{d\mathrm{E}\left[wages \mid education\right]}{deducation}$$

  is incorrect.

- Rather, we need to analyze

$$\frac{\partial\mathrm{E}\left[wages \mid education, experience\right]}{\partial education}$$

Ceteris Paribus or Holding Other Factors Fixed

- In economics and other social sciences you have non experimental data, so in principle, it is difficult to estimate the ceteris paribus effects.
- However, we will see that econometric methods can simulate a ceteris paribus experiment.
- We will be able to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.
- We will see that the procedure is quite simple: if the focus is on the relation between x_{i1} and y_i and x_{i2}, x_{i3} are the controlled variables, just add them in a regression model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

(just simple as that!) In practice the problem is to find the appropriate controlled variables.

Introduction

- The general linear regression with k explanatory variables is just an extension of the simple regression as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i \tag{11}$$

- Still need to make a zero conditional mean assumption, so now assume that $\mathrm{E}[u \mid \mathbf{X}] = 0$
- At a minimum, all factors in the unobserved error term be uncorrelated with the explanatory variables
- β_0 is still the intercept, $\beta_0, \cdots, \beta_k$ denote slope parameters.

Introduction

- Because
$$\frac{\partial}{\partial x_j} \mathsf{E}\left[y \mid \mathbf{X}\right] \approx \beta_j$$
- β_j represents the marginal effect of x_j
- indicates the amount y is expected to change as x_j changes by one unit and other variables are kept constant (ceteris paribus).

Interpreting the OLS Regression Equation

- In the fitted model:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$$
- The estimates $\hat{\beta}_j$ have partial effect, or ceteris paribus, interpretations
- we have:
$$\Delta\hat{y}_i = \hat{\beta}_1 \Delta x_{i1} + \cdots + \hat{\beta}_k \Delta x_{ik}$$
- so we can obtain the predicted change in y given the changes in x_j.
- In particular, when $x_2, \cdots, x_k$ are held fixed, so that $\Delta x_j = 0$, for $j = 2, \cdots, k$, we have:
$$\Delta\hat{y}_i = \hat{\beta}_1 \Delta x_{i1}$$

OLS Fitted Values and Residuals

The fitted model is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \tag{12}$$

The residuals for observation i is:

$$\hat{u}_i = y_i - \hat{y}_i \tag{13}$$

Again, the following properties hold:

1. $\bar{y} = \bar{\hat{y}}$
2. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k$
3. The sample covariance between each x_j, for $j = 1, \cdots, k$ and $\hat{u}_i$ is zero. Hence, the sample covariance between $\hat{y}_i$ and $\hat{u}_i$.

The MLRM: Assumptions

MLR1: Linear in Parameters

The model in the population can be written as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i \\ &= \mathbf{x}_i' \boldsymbol{\beta} + u_i \end{aligned} \tag{14}$$

MLR2: Random Sampling

We have a random sample of n observations,

$$\{(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i)\, i = 1, \cdots, n\}$$

following the population model in Assumption MLR1.

The MLRM: Assumptions

MLR3: No Perfect Collinearity

In the sample there are no exact linear relationships among the independent variables.

Put otherwise, the matrix $\mathbf{X}$ has rank $k + 1$.

MLR4: Zero Conditional Mean

Conditional on the entire matrix $\mathbf{X}$, each error u_i has zero mean: $\mathrm{E}[u_i \mid \mathbf{X}] = 0$

The MLRM: Assumptions

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables: $\text{Var}[u \mid \mathbf{X}] = \sigma^2$.

In matrix form:

$$\text{Var}(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$$

where $\mathbf{I}_n$ denotes the n-dimensional identity matrix.

MLR6: Normality

We have:

$$u_i \mid \mathbf{X} \sim N\left(0, \sigma^2\right)$$

Finite Sample Properties

Theorem 1

Unbiasedness of OLS

Under assumptions MLR1-4 the OLS estimators are unbiased estimators of the population parameters. We have:

$$E\left[\hat{\beta} \mid \mathbf{X}\right] = \beta$$

Variances of OLS estimators

Theorem 2

Variances of OLS estimators

Under assumptions MLR1-MLR5 the variance-covariance matrix of the OLS estimators is:

$$Var\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

The Gauss-Markov Theorem

Theorem 3

Gauss-Markov Theorem

Under assumptions MLR1 - MLR5, the OLS estimators are the **b***est* **l***inear* **u***nbiased* **e***stimators (BLUE) of the regression coefficients, i.e*

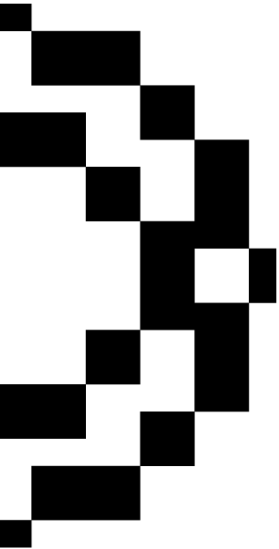
$$Var\left(\hat{\beta}_j \mid \mathbf{X}\right) \leq Var\left(\widetilde{\beta}_j \mid \mathbf{X}\right)$$

where $\widetilde{\beta}_j$ is any unbiased estimator linear in y.

In the matrix sense

$$\text{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) \leq \text{Var}\left(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}\right)$$

means that $\text{Var}\left(\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right) - \text{Var}\left(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}\right)$ is a psd matrix.



Inference: Introduction

Statistical Inference under Normality

Statistical inference in the regression model:

- Hypothesis tests about population parameters
- Construction of confidence intervals

Sampling distributions of the OLS estimators:

- The OLS estimators are random variables
- We already know their expected values and their variances
- However, for hypothesis tests we need to know their distribution
- In order to derive their distribution we need additional assumptions
- Assumption about distribution of errors: normal distribution

Statistical Inference under Normality

Assumption MLR6: Normality

We have:

$$u_i \mid \mathbf{X} \sim N(0, \sigma^2)$$

Equivalently,

$$\mathbf{u} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Discussion of the normality assumption

- The error term is the sum of "many" different unobserved factors
- Sums of independent factors are normally distributed (CLT)
- Problems:
  - How many different factors? Number large enough?
  - Possibly very heterogeneous distributions of individual factors
  - How independent are the different factors?
- The normality of the error term is an empirical question
- At least the error distribution should be "close" to normal
- In many cases, normality is questionable or impossible by definition
- In some cases, normality can be achieved through transformations of the dependent variable
- Under normality, OLS is BUE (see MLE)
- For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size

Terminology

Remark

- Assumptions MLR1-5 $\Longrightarrow$ Gauss-Markov Assumptions
- Assumptions MLR1-6 $\Longrightarrow$ Classical Assumptions or Classical Linear Model Assumptions (CLM Assumptions)

The t Test

Theorem 4

Normal sampling distribution of the OLS

Under Assumptions MLR1-6 we have:

$$\hat{\beta}_j \mid \mathbf{X} \sim N\left(\beta_j, Var\left(\hat{\beta}_j \mid \mathbf{X}\right)\right).$$

Therefore,

$$z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j \mid \mathbf{X}\right)}} \sim N(0,1)$$

Statistical Inference under Normality

Theorem 5

t-distribution for standardized estimators

Under Assumptions MLR1-6 we have:

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right)}} \sim t_{(n-k-1)}$$

The t Test: general setup

- Suppose that we have a hypothesis about the β_j

$$H_0 : \beta_j = a$$

where a is a specific value, say $a = 0$, and that this hypothesis is tested against the alternative hypothesis

$$H_1 : \beta_j \neq a.$$

- Two possibilities:
  - If $|t_{obs}| > t_{\alpha/2} \Longrightarrow$ reject H_0 at the $\alpha \times 100\%$ level
  - If $|t_{obs}| \leq t_{\alpha/2} \Longrightarrow$ do not reject H_0

where

$$t_{obs} = \frac{\hat{\beta}_j - a}{\text{se}\left(\hat{\beta}_j\right)}$$

and $t_{\alpha/2} : \text{P}\left(t > t_{\alpha/2}\right) = \alpha/2$

The t Test: general setup

- Under the null hypothesis we have

$$t = \frac{\hat{\beta}_j - a}{\text{se}\left(\hat{\beta}_j\right)} \sim t_{(n-k-1)}$$

- If we observe $|t_{obs}| > t_{\alpha/2}$ and the H_0 is true, then a low-probability event has occurred.
- We take $|t_{obs}| > t_{\alpha/2}$ as an evidence against the null and the decision should be to reject H_0.

Other cases:

- $H_1 : \beta_j > a$: if $t_{obs} > t_\alpha \implies$ reject the null at the $\alpha \times 100\%$ level; otherwise do not reject.
- $H_1 : \beta_j < a$: if $t_{obs} < -t_\alpha \implies$ reject the null at the $\alpha \times 100\%$ level; otherwise do not reject.

The t Test: general setup

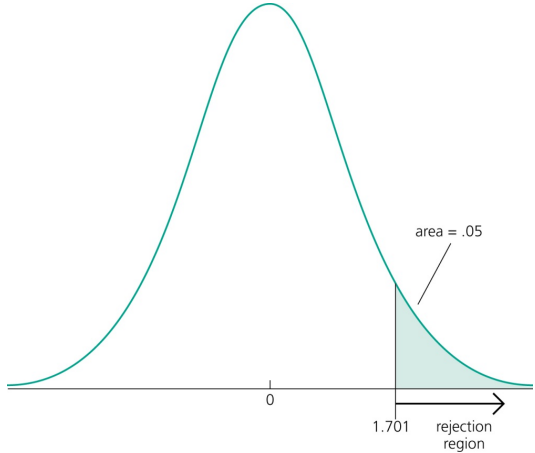


Figure 2: 5% rejection rule for the alternative $H_1 : \beta_j > 0$ with 28 *df*

The t Test: general setup

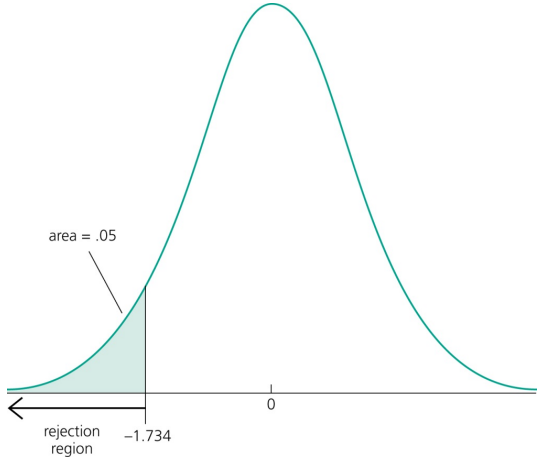


Figure 3: 5% rejection rule for the alternative $H_1 : \beta_j < 0$ with 18 *df*

The t Test: general setup

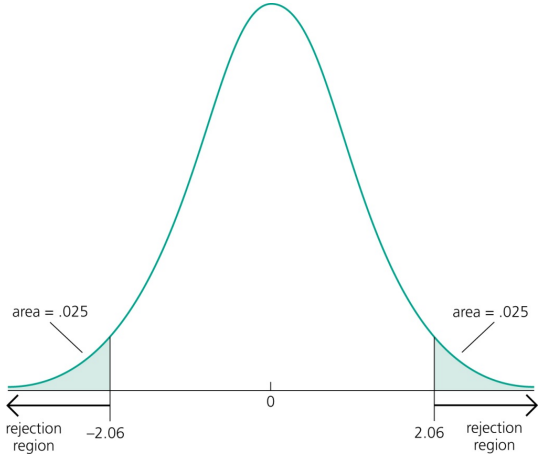


Figure 4: 5% rejection rule for the alternative $H_1 : \beta_j \neq 0$ with 25 *df*

The t Test: the p-value

- p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
- Calculate p-value for $H_0 : \beta_j = a$:
  - $H_1 : \beta_j \neq a \Longrightarrow p\text{-value} = 2\mathrm{P}\left(t > |t_{obs}| \mid H_0 \text{ is true}\right).$
  - $H_1 : \beta_j > a \Longrightarrow p\text{-value} = \mathrm{P}\left(t > t_{obs} \mid H_0 \text{ is true}\right).$
  - $H_1 : \beta_j < a \Longrightarrow p\text{-value} = \mathrm{P}\left(t < t_{obs} \mid H_0 \text{ is true}\right).$
- For example, a p-value $= 0.02$ shows little evidence supporting $H : 0$
- At the 5% level you should reject the null.
- Rejection rule:

$$
\begin{aligned}
p\text{-value} > \alpha \quad &\Rightarrow \quad \text{do not reject } H_0 \text{ at the } \alpha \times 100\% \text{ level} \\
p\text{-value} \leq \alpha \quad &\Rightarrow \quad \text{reject } H_0 \text{ at the } \alpha \times 100\% \text{ level}
\end{aligned}
$$

The t Test: the p-value

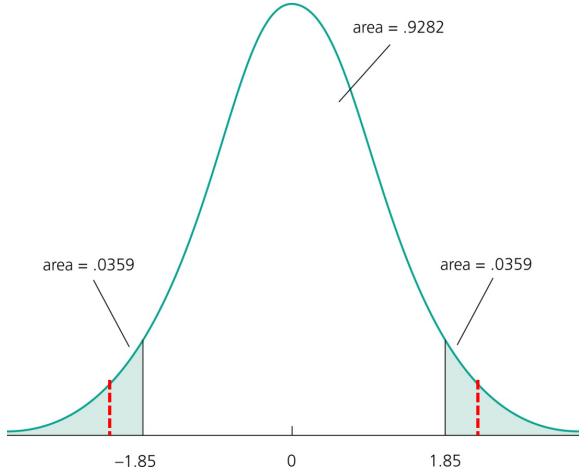


Figure 5: p-value against a two-sided alternative, when $t_{obs} = 1.85$ and $df = 40$.

The t Test: standard case

- Very often we want to test whether there is any relation between x_j and on the expected value of y without imposing a sign on the partial effect *a priori*
- Thus, in most applications, our primary interest lies in testing the null hypothesis

$$H_0 : \beta_j = 0$$

- β_j measures the partial effect of x_j on the expected value of y, after controlling for all other independent variables
- $\beta_j = 0$ means that, once the other variables have been accounted for, x_j has no effect on the expected value of y.

The t Test: standard case

- For example, in model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

- The null $H_0 : \beta_3 = 0$ means that, once education and tenure have been accounted for, the number of years in the workforce (exper) has no effect on hourly wage. If it is true, it implies that a person's work history prior to the current employment does not affect wage.
- If $\beta_3 > 0$, then prior work experience contributes to productivity, and hence to wage.

The t Test: standard case

```
data(wage1, package='wooldridge')
# OLS regression:
summary( lm(log(wage) ~ educ+exper+tenure, data=wage1) )


##
## Call:
## lm(formula = log(wage) ~ educ + exper + tenure, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05802 -0.29645 -0.03265  0.28788  1.42809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.284360   0.104190   2.729  0.00656 **
## educ        0.092029   0.007330  12.555  < 2e-16 ***
## exper       0.004121   0.001723   2.391  0.01714 *
## tenure      0.022067   0.003094   7.133 3.29e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4409 on 522 degrees of freedom
## Multiple R-squared:  0.316,Adjusted R-squared:  0.3121
## F-statistic: 80.39 on 3 and 522 DF,  p-value: < 2.2e-16
```

The t Test: standard case

```
# CV for alpha=5% and 1% using the t distribution
#with 522 d.f.:
alpha <- c(0.05, 0.01)
qt(1-alpha, 522)

## [1] 1.647778 2.333513

# Critical values for alpha=5% and 1% using the
#normal approximation:
qnorm(1-alpha)

## [1] 1.644854 2.326348
```

The t Test: further comments

Correct wording in reporting the outcome of a test involving $H_0 : \beta_j = a$ vs $H_1 : \beta_j \neq a$

- When the null is rejected we say that $\hat{\beta}_j$ (not β_j) is significantly different from a at $\alpha \times 100\%$ level.
- When the null isn't rejected we say that we say that $\hat{\beta}_j$ (not β_j) is not significantly different from a at $\alpha \times 100\%$ level.

Correct wording in reporting the outcome of a test involving $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

- When the null is rejected we say that $\hat{\beta}_j$ (not β_j) is significantly different from zero at $\alpha \times 100\%$ level, or the regressor (on β_j) is relevant to explain $\mathrm{E}[y \mid \mathbf{X}]$.
- When the null isn't rejected we say that $\hat{\beta}_j$ (not β_j) is not significantly different from zero at $\alpha \times 100\%$ level, or the regressor (on β_j) is not relevant to explain $\mathrm{E}[y \mid \mathbf{X}]$.

The t Test: further comments

More Remarks:

- Rejection of the null is not proof that the null is false. Why?
- Acceptance of the null is not proof that the null is true. Why? We prefer to use the language *we fail to reject* H_0 *at the* $\alpha \times 100\%$ *level* rather than H_0 *is accepted at the* $\alpha \times 100\%$ *level.*
- In a test of type $H_0 : \beta_j = 0$, if se$(\hat{\beta}_j)$ is large ($\hat{\beta}_j$ is an imprecise estimator) is more difficult to reject the null. The sample contains little information about the true value of β_j.
- In a test of type $H_0 : \beta_j = 0$ vs $H1 : \beta_j < 0$ (or $H1 : \beta_j > 0$), if se$(\hat{\beta}_j)$, p-values must be divided by two.

Statistical Versus Economic Significance

- The statistical significance of a variable is determined by the size of $t_{obs} = \hat{\beta}_j / \text{se}(\hat{\beta}_j)$
- the economic significance of a variable is related to the size and sign of $\hat{\beta}_j$.
- Suppose that in a business activity we have:

$$\widehat{\log(wage_i)} = 0.1 + \underset{(0.001)}{0.01}\, female + \cdots \quad n = 600$$

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. We have

$$\begin{aligned} t_{obs} &= 10, \\ p\text{-value} &\approx 0 \end{aligned}$$

Discuss statistical versus economic significance.

Testing a linear combination of parameters

- Suppose we want to test a single hypothesis involving more than one of the β_j
- To illustrate the general approach, we will consider a simple model to compare the returns to education at junior colleges and four-year colleges; for simplicity, we refer to the latter as "universities"
- Consider the model

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

where:

- jc = number of years attending a two-year college;
- $univ$ = number of years at a four-year college;
- $exper$ = months in the workforce.

Testing a linear combination of parameters

- The hypothesis of interest is whether one year at a junior college is worth one year at a university.
- Under H_0, another year at a junior college and another year at a university lead to the same ceteris paribus percentage increase in wage. As follows:

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 < \beta_2$$

- We cannot simply use the individual t statistics for $\hat{\beta}_1$ and $\hat{\beta}_2$ to test H_0.
- However, for constructing a t statistic for doing the test, we rewrite the null and alternative

$$H_0 : \beta_1 - \beta_2 = 0 \text{ vs. } H_1 : \beta_1 - \beta_2 < 0$$

under the null we have

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{se}\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$$

Testing a linear combination of parameters

```
data(twoyear, package='wooldridge')
reg <- lm(lwage~jc+univ+exper, data=twoyear)
summary(reg)
```

Testing a linear combination of parameters

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper, data = twoyear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10362 -0.28132  0.00551  0.28518  1.78167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.4723256  0.0210602  69.910   <2e-16 ***
## jc          0.0666967  0.0068288   9.767   <2e-16 ***
## univ        0.0768762  0.0023087  33.298   <2e-16 ***
## exper       0.0049442  0.0001575  31.397   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4301 on 6759 degrees of freedom
## Multiple R-squared:  0.2224,Adjusted R-squared:  0.2221
## F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 2.2e-16
```

Testing a linear combination of parameters

Problem:

$$\begin{aligned} \text{se}\left(\hat{\beta}_1 - \hat{\beta}_2\right) &= \sqrt{\widehat{\text{Var}}\left(\hat{\beta}_1\right) + \widehat{\text{Var}}\left(\hat{\beta}_2\right) - 2 \times \widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)} \\ &\neq \text{se}\left(\hat{\beta}_1\right) - \text{se}\left(\hat{\beta}_2\right)! \end{aligned}$$

Testing a linear combination of parameters

Calculate $\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)$

- The variance-covariance matrix of the estimates is:

$$\begin{pmatrix} \widehat{\text{Var}}\left(\hat{\beta}_0\right) & \widehat{\text{cov}}\left(\hat{\beta}_0, \hat{\beta}_1\right) & \cdots & \widehat{\text{cov}}\left(\hat{\beta}_0, \hat{\beta}_k\right) \\ . & \widehat{\text{Var}}\left(\hat{\beta}_1\right) & . & \widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_k\right) \\ . & . & \ddots & \vdots \\ . & . & . & \widehat{\text{Var}}\left(\hat{\beta}_k\right) \end{pmatrix}$$

- Just need to pick up element $(2, 3)$
  Remark: in **R**:`vcov()`

Testing a linear combination of parameters

```
vcov(reg)

##               (Intercept)            jc          univ         exper
## (Intercept)  4.435337e-04 -1.741432e-05 -1.573472e-05 -3.104756e-06
## jc          -1.741432e-05  4.663243e-05  1.927929e-06 -1.718296e-08
## univ        -1.573472e-05  1.927929e-06  5.330230e-06  3.933491e-08
## exper       -3.104756e-06 -1.718296e-08  3.933491e-08  2.479792e-08

sqrt(vcov(reg)[2,2]+vcov(reg)[3,3]-2*vcov(reg)[3,2])

## [1] 0.006935907
```

Testing a linear combination of parameters

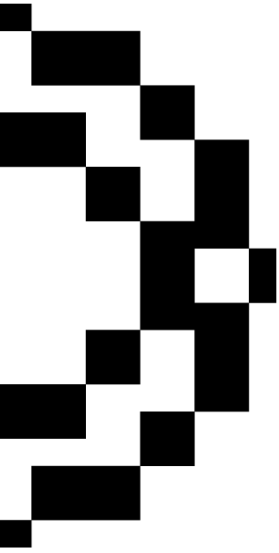
- Calculate t_obs as

$$\begin{aligned} t_{obs} &= \frac{0.0667 - 0.0769}{\sqrt{0.0068^2 + 0.0023^2 - 2 \times 1.927929 \times 10^{-6}}} \\ &= \frac{-0.01018}{0.006936} \\ &= -1.468 \end{aligned}$$

- The rejection region is:

$$W = \{t : t_{obs} < -1.645\}$$

- The p-value is about 0.070, we do not reject the null at the 5% level. There is no evidence against $\beta_1 = \beta_2$ at the 5% level



The F Test

Testing multiple linear restrictions: The F-test

- Let us consider model (*unrestricted model*):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- So far, we have only covered hypotheses involving a single restriction.
- Frequently, we wish to test multiple hypotheses about the underlying slope parameters $\beta_1, \cdots, \beta_k$
- We begin with the leading case of testing whether a set of independent variables has no partial effect on a dependent variable.
- For notational simplicity, assume that it is the last q variables in the list of independent variables
- We have a joint null hypothesis about β

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0 \tag{15}$$

- Which puts q exclusion restrictions on the model.

Testing multiple linear restrictions: The F-test

- When we impose the restrictions under H_0, we are left with the *restricted model*:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

- Thus the restricted model (UR) is a special case of the unrestricted one (R).
- After estimating both models (UR and R), what we need to decide is whether the increase in the SSR (or in the R^2) in going from the unrestricted model to the restricted model is large enough to warrant rejection of the null (15).
- As with all testing, the answer depends on the significance level of the test

Testing multiple linear restrictions: The F-test

- Formally we have

$$\begin{aligned} F &= \frac{\left(\text{SSR}_R - \text{SSR}_{UR}\right)/q}{\text{SSR}_{UR}/\left(n-k-1\right)} \\ &= \frac{\left(R^2_{UR} - R^2_R\right)/q}{\left(1-R^2_{UR}\right)/\left(n-k-1\right)} \sim F_{(q,n-k-1)} \end{aligned}$$

where:

- $q = df_{H_0} - df_{H_1} = df_R - df_{UR}$ (number of restrictions under the null)
- $n - k - 1 = df_{UR}$ (df of the unrestricted model)

Testing multiple linear restrictions: The F-test

- Let F_{obs} be the observed test statistics.
- We have:
  - reject H_0 if $F_{obs} > F_\alpha$ (or p-value $< \alpha$)
  - do not reject H_0 if $F_{obs} \leq F_\alpha$ (or p-value $\geq \alpha$)
- The reasoning is as follow. Under the null hypothesis we have

$$F \sim F_{(k,n-k-1)}$$

- If we observe $F > F_\alpha$ and the H_0 is true, then a low-probability event has occurred.

Remark: the p-value can be calculated as

$$\mathrm{P}\left(F > F_{obs} \mid H_0 \text{ is true}\right)$$

Testing multiple linear restrictions: The F-test

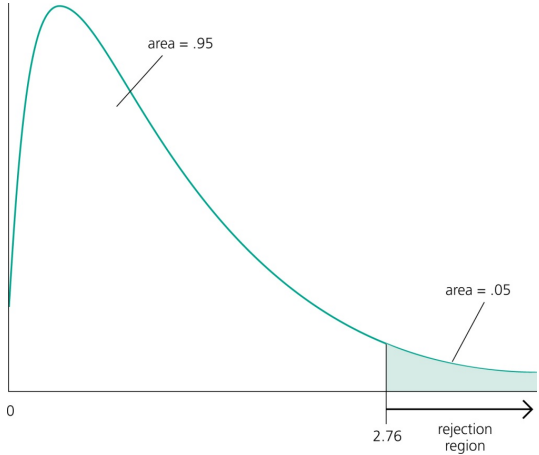


Figure 6: 5% critical value and rejection region in an $F_{3,60}$ distribution.

Testing multiple linear restrictions: The F-test

```
data(mlb1, package='wooldridge')
# Unrestricted OLS regression:
res.ur <- lm(log(salary) ~ years+gamesyr+bavg+hrunsyr+rbisyr, data=mlb1)
# Restricted OLS regression:
res.r <- lm(log(salary) ~ years+gamesyr, data=mlb1)
# R2:
( r2.ur <- summary(res.ur)$r.squared )
```

```
## [1] 0.6278028
```

```
( r2.r <- summary(res.r)$r.squared )
```

```
## [1] 0.5970716
```

```
# F statistic:
( F <- (r2.ur-r2.r) / (1-r2.ur) * 347/3 )
```

```
## [1] 9.550254
```

```
# p value = 1-cdf of the appropriate F distribution:
1-pf(F, 3,347)
```

```
## [1] 4.473708e-06
```

Testing multiple linear restrictions: The F-test

```
data(mlb1, package='wooldridge')

# Unrestricted OLS regression:
res.ur <-
  lm(log(salary) ~ years+gamesyr+bavg+hrunsyr+rbisyr, data=mlb1)
# Load package "car" (which has to be installed on the computer)
library(car)
```

```
## Loading required package:  carData
```

Testing multiple linear restrictions: The F-test

```
# F test
myH0 <- c("bavg","hrunsyr","rbisyr")
linearHypothesis(res.ur, myH0)


##
## Linear hypothesis test:
## bavg = 0
## hrunsyr = 0
## rbisyr = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    350 198.31
## 2    347 183.19  3    15.125 9.5503 4.474e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing multiple linear restrictions: The F-test

- When the null is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- We are testing of overall significance of a regression
- The restricted model is

$$y = \beta_0 + u$$

- The F-statistic becomes

$$\begin{aligned} F &= \frac{\text{SSE}/k}{\text{SSR}/(n-k-1)} \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{(k,n-k-1)} \end{aligned}$$

The t-test vs the F-test

- Are sequential successive t-tests and F-tests equivalent?
- What happens if t-test and F-test suggest different conclusions?

Heteroskedasticity

Introduction

Consider the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error, u, was constant

$$\text{Var} = (u \mid \mathbf{X}) = \sigma^2$$

- If this is not true, that is if the variance of u is different for different values x, then the errors are heteroskedastic

$$\text{Var} = (u \mid \mathbf{X}) = \sigma_i^2$$

Objectives

Answering the following questions

1. What are the consequences of Heteroskedasticity?
2. How to (formally) detect the presence?
3. How to solve the problem?

Consequences

1. The Gauss-Markov Theorem does not hold (OLS is not BLUE)
2. The estimators of the variances $\text{Var}(\hat{\beta}_j \mid \mathbf{X})$ are biased
3. Since the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and t statistics.
4. The usual OLS t statistics do not have t distributions
5. F statistics are no longer F distributed (and the LM statistic no longer has an asymptotic chi-square distribution).
6. In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity

Consequences

However:

1. OLS estimators are still unbiased.
2. OLS estimators are still consistent

Testing for Heteroskedasticity

Consider the following model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Objective, to test:

$$H_0 : Var(u \mid \mathbf{X}) = \sigma^2 \text{ vs } H_1 : Var(u \mid \mathbf{X}) = \sigma_i^2$$

Testing for Heteroskedasticity

Intuition:

- We have $\text{Var}[u \mid \mathbf{X}] = \text{E}\left[u^2 \mid \mathbf{X}\right]$ because $\text{E}[u \mid \mathbf{X}] = 0$
- we want to test whether u^2 is related (in expected value) to one or more of the explanatory variables.
- A simple approach is to assume a linear function:

$$u^2 = \alpha_0 + \alpha_0 x_1 + \cdots + \alpha_k x_k + \textit{error}$$

- And then, using the F-stat or the LM-stat, test:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Testing for Heteroskedasticity

Three approaches:

- Breusch-Pagan Test
- White (full) Test
- White (special) Test

The Breusch-Pagan Test

1. Estimate the model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ by OLS and obtain the residuals $\hat{u}$;
2. Run the regression $\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k + error$ and keep the associated R^2, say R_u^2;
3. Test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ against a bilateral alternative, Using the F-Stat or the LM, as follows:

$$LM = n \times R_u^2 \overset{a}{\sim} \chi^2_{(k)}$$

The Breusch-Pagan Test

```r
data(hprice1, package='wooldridge')

# Estimate model
reg <- lm(price~lotsize+sqrft+bdrms, data=hprice1)
reg
```

```
##
## Call:
## lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1)
##
## Coefficients:
## (Intercept)      lotsize        sqrft        bdrms
##  -21.770308     0.002068     0.122778    13.852522
```

The Breusch-Pagan Test

```r
# Automatic BP test
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 14.092, df = 3, p-value = 0.002782
```

The Breusch-Pagan Test

```r
# Manual regression of squared residuals
summary(lm( resid(reg)^2 ~ lotsize+sqrft+bdrms, data=hprice1))
```

```
##
## Call:
## lm(formula = resid(reg)^2 ~ lotsize + sqrft + bdrms, data = hprice1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9044  -2212  -1256    -97  42582
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.523e+03  3.259e+03  -1.694  0.09390 .
## lotsize      2.015e-01  7.101e-02   2.838  0.00569 **
## sqrft        1.691e+00  1.464e+00   1.155  0.25128
## bdrms        1.042e+03  9.964e+02   1.046  0.29877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6617 on 84 degrees of freedom
## Multiple R-squared:  0.1601,Adjusted R-squared:  0.1301
## F-statistic: 5.339 on 3 and 84 DF,  p-value: 0.002048
```

The White Test

- The White (1980) auxiliary regression considers also the squares of the independent variables and all cross products.
- Consider, without any lost of generality, $k = 3$:

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

The auxiliary regression is:

$$\begin{aligned} \hat{u}^2 &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \\ &+ \alpha_4 x_1^2 + \alpha_5 x_2^2 + \alpha_6 x_3^2 \\ &+ \alpha_7 x_1 x_2 + \alpha_8 x_1 x_3 + \alpha_9 x_2 x_3 + \textit{error} \end{aligned}$$

The White Test

```
# Manual regression of squared residuals
summary(lm( resid(reg)^2 ~ lotsize+sqrft+bdrms+I(lotsize^2)+
              I(sqrft^2)+I(bdrms^2)+I(bdrms*sqrft)+
              I(bdrms*lotsize)+I(lotsize*sqrft),
              data=hprice1))
```

The White Test

```
##
## Call:
## lm(formula = resid(reg)^2 ~ lotsize + sqrft + bdrms + I(lotsize^2) +
##     I(sqrft^2) + I(bdrms^2) + I(bdrms * sqrft) + I(bdrms * lotsize) +
##     I(lotsize * sqrft), data = hprice1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10837  -2310  -1169    880  40657
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.563e+04  1.137e+04   1.374  0.17325
## lotsize             -1.860e+00  6.371e-01  -2.919  0.00459 **
## sqrft               -2.674e+00  8.662e+00  -0.309  0.75838
## bdrms               -1.983e+03  5.438e+03  -0.365  0.71640
## I(lotsize^2)        -4.978e-07  4.631e-06  -0.107  0.91467
## I(sqrft^2)           3.523e-04  1.840e-03   0.191  0.84864
## I(bdrms^2)           2.898e+02  7.588e+02   0.382  0.70362
## I(bdrms * sqrft)    -1.021e+00  1.667e+00  -0.612  0.54210
## I(bdrms * lotsize)   3.146e-01  2.521e-01   1.248  0.21572
## I(lotsize * sqrft)   4.568e-04  2.769e-04   1.650  0.10303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5884 on 78 degrees of freedom
## Multiple R-squared:  0.3833,Adjusted R-squared:  0.3122
## F-statistic: 5.387 on 9 and 78 DF,  p-value: 1.013e-05
```

The White Test

```
# Compute LMobs
whitefull=lm( resid(reg)^2 ~ lotsize+sqrft+bdrms+I(lotsize^2)+
              I(sqrft^2)+I(bdrms^2)+I(bdrms*sqrft)+
              I(bdrms*lotsize)+I(lotsize*sqrft),
              data=hprice1)
summary(whitefull)
```

The White Test

```
##
## Call:
## lm(formula = resid(reg)^2 ~ lotsize + sqrft + bdrms + I(lotsize^2) +
##     I(sqrft^2) + I(bdrms^2) + I(bdrms * sqrft) + I(bdrms * lotsize) +
##     I(lotsize * sqrft), data = hprice1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10837  -2310  -1169    880  40657
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.563e+04  1.137e+04   1.374  0.17325
## lotsize             -1.860e+00  6.371e-01  -2.919  0.00459 **
## sqrft               -2.674e+00  8.662e+00  -0.309  0.75838
## bdrms               -1.983e+03  5.438e+03  -0.365  0.71640
## I(lotsize^2)        -4.978e-07  4.631e-06  -0.107  0.91467
## I(sqrft^2)           3.523e-04  1.840e-03   0.191  0.84864
## I(bdrms^2)           2.898e+02  7.588e+02   0.382  0.70362
## I(bdrms * sqrft)    -1.021e+00  1.667e+00  -0.612  0.54210
## I(bdrms * lotsize)   3.146e-01  2.521e-01   1.248  0.21572
## I(lotsize * sqrft)   4.568e-04  2.769e-04   1.650  0.10303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5884 on 78 degrees of freedom
## Multiple R-squared:  0.3833,Adjusted R-squared:  0.3122
## F-statistic: 5.387 on 9 and 78 DF,  p-value: 1.013e-05
```

The White Test

```
# Compute LMobs
length(reg$res)*summary(whitefull)$r.squared

## [1] 33.73166

#or
nobs(reg)*summary(whitefull)$r.squared

## [1] 33.73166
```

The White Test

```
# Alternative
bptest(reg, ~ lotsize+sqrft+bdrms+I(lotsize^2)+
              I(sqrft^2)+I(bdrms^2)+I(bdrms*sqrft)+
              I(bdrms*lotsize)+I(lotsize*sqrft),
              data=hprice1)
```

The White Test

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 33.732, df = 9, p-value = 9.953e-05
```

The White Test

- With only 3 independent variables the auxiliary has 9 independent variables.
- With 6 the auxiliary regression involves 27 regressors
- This abundance of regressors is a weakness in the pure form of the White
- it uses many degrees of freedom for models with just a moderate number of independent variables.
- The special White test considers the following auxiliary regression

$$\hat{u}^2 = \alpha_0 + \alpha_1 \hat{y} + \alpha_2 \hat{y}^2 + \textit{error}$$

- We can preserve the spirit of the White test while conserving on degrees of freedom by using the OLS fitted values in a test for heteroskedasticity

The White Test

```
bptest(reg, ~ fitted(reg) + I(fitted(reg)^2) )

##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 16.268, df = 2, p-value = 0.0002933
```

Robust Estimation

To account for heteroskedasticity in the model, we must resort to a robust variance-covariance matrix to perform statistical inference.

Robust Estimation

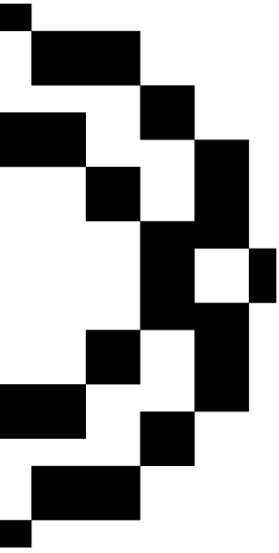
```r
data(hprice1, package="wooldridge")
# load packages (which need to be installed!)
library(lmtest); library(car)
# Estimate model
reg <- lm(price~lotsize+sqrft+bdrms, data=hprice1)
```

Robust Estimation

```
# Usual SE:
coeftest(reg)

##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -2.1770e+01  2.9475e+01 -0.7386  0.462208
## lotsize      2.0677e-03  6.4213e-04  3.2201  0.001823 **
## sqrft        1.2278e-01  1.3237e-02  9.2751 1.658e-14 ***
## bdrms        1.3853e+01  9.0101e+00  1.5374  0.127945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Robust Estimation

```
# Refined White heteroscedasticity-robust SE:
coeftest(reg, vcov=hccm)

##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -21.7703081  41.0326943 -0.5306 0.597124
## lotsize       0.0020677   0.0071485  0.2893 0.773101
## sqrft         0.1227782   0.0407325  3.0143 0.003406 **
## bdrms        13.8525217  11.5617901  1.1981 0.234236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Asymptotic Properties of the OLS

OLS Asymptotics

So far we covered what are called finite sample, small sample, or exact properties of the OLS estimators in the population model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{16}$$

- Properties of OLS that hold for any sample/sample size
  - Expected values/unbiasedness under MLR.1-MLR.4
  - Variance formulas under MLR.1-MLR.5
  - Gauss-Markov Theorem under MLR.1-MLR.5
  - Exact sampling distributions/tests under MLR.1-MLR.6
- Properties of OLS that hold in large samples:
  - Consistency under MLR.1-MLR.4
  - Asymptotic normality/tests under MLR.1-MLR.5

Asymptotic normality of OLS

- In practice, the normality assumption MLR.6 is often questionable
- If MLR.6 does not hold, the results of t- or F-tests may be wrong
- Fortunately, F- and t-tests still work if the sample size is large enough
- Also, OLS estimates are normal in large samples even without MLR.6

Asymptotic normality of OLS

Theorem 4

Asymptotic normality of OLS

Under assumptions MLR1-5, we have:

$$z = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \overset{a}{\sim} N(0,1)$$

Asymptotic normality of OLS

Practical Consequences

- In large samples, the t-distribution is close to the $N(0,1)$ distribution
- As a consequence, t-tests are valid in large samples without MLR.6
- The same is true for confidence intervals and F-tests
- Important: MLR.1 - MLR.5 are still necessary, esp. homoskedasticity

RESET Test

The Ramsey's RESET test

- Ramsey (1969) proposed a general functional form misspecification test, Regression Specification Error Test (RESET), which has proven to be useful.
- The idea behind RESET is fairly simple. If the original model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- then no nonlinear functions of the independent variables should be significant when added to this equation
- Although this often detects functional form problems, it has the drawback of using up many dof if there are many explanatory variables in the original model
- Further, certain kinds of neglected nonlinearities will not be picked up by adding quadratic terms.

The Ramsey's RESET test

RESET adds polynomials in the OLS fitted values to detect general kinds of functional form misspecification, as following

1. estimate the model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ and obtain the fitted values $\hat{y}_i$
2. estimate augmented model $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + u_i$
3. test the null

$$H_0 : \gamma_1 = \gamma_2 = 0$$

using an F-test with 2 and $n - k - 3$ dof (or an LM with 2 dof).

The Ramsey's RESET test

```
data(hprice1, package='wooldridge')
# original linear regression
orig <- lm(price ~ lotsize+sqrft+bdrms, data=hprice1)
# regression for RESET test
RESETreg <- lm(price ~ lotsize+sqrft+bdrms+I(fitted(orig)^2)+
                                        I(fitted(orig)^3), data=hprice1)
RESETreg


##
## Call:
## lm(formula = price ~ lotsize + sqrft + bdrms + I(fitted(orig)^2) +
##     I(fitted(orig)^3), data = hprice1)
##
## Coefficients:
##       (Intercept)            lotsize              sqrft              bdrms
##         1.661e+02          1.537e-04          1.760e-02          2.175e+00
## I(fitted(orig)^2)  I(fitted(orig)^3)
##         3.534e-04          1.546e-06
```

The Ramsey's RESET test

```
# RESET test. H0: all coeffs including "fitted" are=0
library(car)
linearHypothesis(RESETreg, matchCoefs(RESETreg,"fitted"))


##
## Linear hypothesis test:
## I(fitted(orig)^2) = 0
## I(fitted(orig)^3) = 0
##
## Model 1: restricted model
## Model 2: price ~ lotsize + sqrft + bdrms + I(fitted(orig)^2) + I(fitted(orig)^3)
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     84 300724
## 2     82 269984  2     30740 4.6682 0.01202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Ramsey's RESET test

```
# RESET test
library(lmtest)
resettest(orig)


##
##  RESET test
##
## data:  orig
## RESET = 4.6682, df1 = 2, df2 = 82, p-value = 0.01202
```

Multiple Regression Analysis with Qualitative Information

Dummy variables

- A dummy variable is a variable that takes on the value 1 or 0
- Examples: male (= 1 if are male, 0 otherwise), south (= 1 if in the south, 0 otherwise), etc.
- Dummy variables are also called binary variables, for obvious reasons
- Remark: Usual practice is to denote the dummy variable by the name of one of the categories. For example, instead of using gender one can define the variable e.g. as female, which equals 1 if the gender is female and 0 if male.

A Dummy Independent Variable

- Dummy variables can be incorporated into a regression model as any other variables.
- Consider the simple regression model

$$y = \beta_0 + \beta_1 D + \beta_2 x + u$$

- This can be interpreted as an intercept shift
- If $D = 0$, then $y = \beta_0 + \beta_2 x + u$
- If $D = 1$, then $y = (\beta_0 + \beta_1) + \beta_2 x + u$
- β_1 indicates the difference wrt the base group ($D = 0$)
- Formally we have

$$\beta_1 = \mathrm{E}[y \mid \mathbf{x}, D = 1] - \mathrm{E}[y \mid \mathbf{x}, D = 0] \tag{17}$$

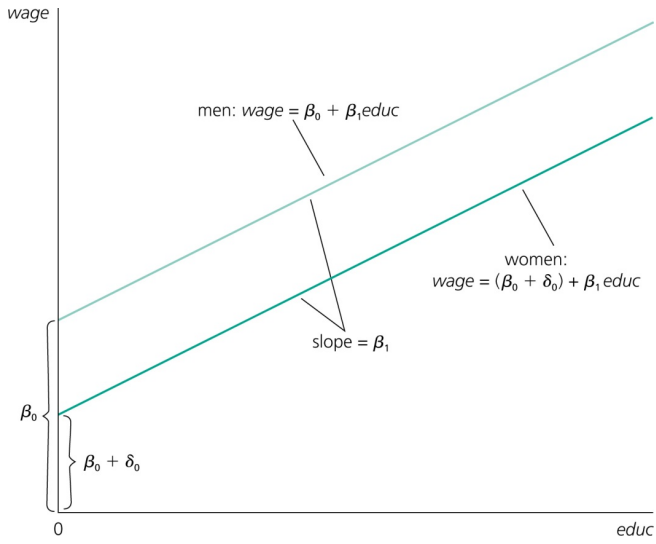
Dummy variables

Consider the model:

$$wage = \beta_0 + \delta_0 fem + \beta_1 educ + u$$

Provide an interpretation on δ_0.

Dummy variables



Dummy variable trap

- This model cannot be estimated (perfect collinearity)

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

- When using dummy variables, one category always has to be omitted:

$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u$$

or

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- Alternatively, one could omit the intercept:

$$wage = \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

- Disadvantages:
  1. More difficult to test for differences between the parameters
  2. R-squared formula only valid if regression contains intercept

Dummy variables

```
data(wage1, package='wooldridge')
lm(wage ~ female+educ+exper+tenure, data=wage1)


##
## Call:
## lm(formula = wage ~ female + educ + exper + tenure, data = wage1)
##
## Coefficients:
## (Intercept)       female         educ        exper       tenure
##     -1.5679      -1.8109       0.5715       0.0254       0.1410
```

Dummy variables

- It can easily be tested whether difference in means is significant

$$H_0 : \delta_0 = 0 \text{ vs } H_1 : \delta_0 \neq 0$$

- The alternative that there is discrimination against women is

$$H_1 : \delta_0 < 0$$

Dummy variables

- A common specification has the dependent variable appearing in logs
- The coefficient on the dummy has a percentage interpretation. In the model:

$$\log(sal) = \beta_0 + \delta_0 fem + \beta_1 educ + u$$

  $\delta_0 \times 100$ is the percentage difference in a woman's wage in relation to that of a man.
- Recall that the log approximation may be inaccurate if the percentage difference is big.
- A more accurate approximation is

$$100\left(e^{\delta_0} - 1\right)$$

Dummy variables

```
data(wage1, package='wooldridge')
lm(log(wage) ~ female+educ+exper+I(exper^2)+tenure+I(tenure^2), data=wage1)


##
## Call:
## lm(formula = log(wage) ~ female + educ + exper + I(exper^2) +
##     tenure + I(tenure^2), data = wage1)
##
## Coefficients:
## (Intercept)       female         educ        exper   I(exper^2)       tenure
##   0.4166910   -0.2965110    0.0801966    0.0294324   -0.0005827    0.0317139
## I(tenure^2)
##  -0.0005852
```

Dummy variables

- We can use several dummy independent variables in the same equation.
- For example, we could add the dummy variable *married* to the wage equation, as follows:

$$sal = \beta_0 + \delta_0 fem + \delta_1 married + \beta_1 educ + u$$

- Again, the introduction of these dummies changes the model intercept:

	Male	Female
Married	$\beta_0 + \delta_1$	$\beta_0 + \delta_0 + \delta_1$
Single	β_0	$\beta_0 + \delta_0$

Dummy variables

```
data(wage1, package='wooldridge')

lm(wage~female+married+educ,
   data=wage1)


##
## Call:
## lm(formula = wage ~ female + married + educ, data = wage1)
##
## Coefficients:
## (Intercept)       female      married         educ
##    -0.04082     -2.08699      1.18153      0.49495
```

Dummy variables for multiple categories

- We can use dummy variables to control for something with multiple categories
- Suppose everyone in your data is either a HS dropout, HS grad only, or college grad
- To compare HS and college grads to HS dropouts, include 2 dummy variables
- hsgrad $= 1$ if HS grad only, 0 otherwise; and colgrad $= 1$ if college grad, 0 otherwise

Dummy variables for multiple categories

- Any categorical variable can be turned into a set of dummy variables
- Because the base group is represented by the intercept, if there are g categories there should be $g - 1$ dummy variables (why?)
- If there are a lot of categories, it may make sense to group some together
- Example: top 10 ranking, $11 - 25$, etc.

Dummy variables for multiple categories

- In the wage example if *married* and *female* are included we have the following possibilities

female	*married*	characterization
1	0	single woman
1	1	married woman
0	1	married man
0	0	single man

- Including only *female* and *married* will allows to estimate the gender gap and the marriage premium
- However, a major limitation of this model is that the the marriage premium is assumed to be the same for men and women.

Dummy variables for multiple categories

```
data(wage1, package='wooldridge')
lm(wage~female+married+educ,
   data=wage1)

##
## Call:
## lm(formula = wage ~ female + married + educ, data = wage1)
##
## Coefficients:
## (Intercept)       female      married         educ
##    -0.04082     -2.08699      1.18153      0.49495
```

Dummy variables for multiple categories

- Let us estimate a model that allows for wage differences among four groups: married men, married women, single men, and single women.
- To do this, we must select a base group;
- we choose *single men*.
- Then, we must define dummy variables for each of the remaining groups.
- Call these *marrmale*, *marrfem*, and *singfem*.

Dummy variables for multiple categories

```
data(wage1, package='wooldridge')
mm=as.integer(wage1$married==1 & wage1$female==0)
mf=as.integer(wage1$married==1 & wage1$female==1)
sm=as.integer(wage1$married==0 & wage1$female==0)
sf=as.integer(wage1$married==0 & wage1$female==1)
lwage=log(wage1$wage)
educ=wage1$educ
lm(lwage~mm+mf+sf+educ)


##
## Call:
## lm(formula = lwage ~ mm + mf + sf + educ)
##
## Coefficients:
## (Intercept)           mm           mf           sf         educ
##     0.58003      0.40026     -0.07285     -0.10316      0.07498
```