BerkeleyHaas
Haas School of Business
University of California Berkeley

Big Data and Better Decisions
Spring 2018

Module 2: Review of Regression Analysis

Contents

List of Figures

1. INTRODUCTION

Regression analysis is often used to model and make predictions about real-world systems. For example, rudimentary weather forecast was based on linear regressions of, say, the amount of rainfall in millimeter on several regressors such as the date and month, rainfall in the previous time period, temperature, humidity, and other such variables. The model then provided predictions of future rainfall given the present values of the regressors.

However, regression analysis in the context of impact evaluations primarily a tool for statistical inference. In fact, statistical research in social science fields such as economics, epidemiology and psychology has extensively relied on regression analysis as a key tool to evaluate hypothesis or research questions. Without a good understanding of statistical inference (hypothesis testing) and application of regression analysis, it will be challenging to conduct impact evaluations. We will assume you already have knowledge of basic statistical and econometric methods. If you need a refresher in these topics, we recommend searching online for various free resources.

In this module, we will demonstrate the use of regression analysis to infer causal effects mainly using the `lm` command in R. Although correlation does not imply causation in general, regression analysis is a tool to test whether the observed association between the outcome (left hand side variable) and the treatment or intervention of interest (a right hand side variable) is statistically significant. Various regression analysis methods and sets of covariates (right hand side variables) are used to ensure that unbiased and precise estimate of this association as possible. However, only a good study design (*e.g.,* a randomized control trial), careful data collection, and support by a plausible biological or economic theory can help us infer causality. Therefore, we must always remember that regression analysis remains a statistical tool and does not replace a good study design and its implementation.

We'll also review basic concepts about statistical inference. In impact evaluation, we always work with a sample from a population of interest (e.g. Oportunidades' beneficiaries). This sample is just one of a massive number of possible samples we could have been taken from this same population. Sampling introduces some uncertainty into our impact estimations because there is always a chance that the estimated effect is due to our specific choice of sample. Therefore, we need an approach that takes estimates computed from a single sample and make them applicable to the entire population.

We also consider a set of techniques to evaluate the validity of assumptions behind the standard regression model. Specifically, we cover techniques to evaluate the role of (over-)influential data, to check the normality of residuals, to assess the presence of heteroscedasticity and multicollinearity, and other ways to evaluate model specification.

At the end of this module we expect you to be able to:

- ✓ Use linear regression as a way to approximate conditional expectation functions.
- ✓ Link our regression models to the standard textbook regression approach based on general linear models.

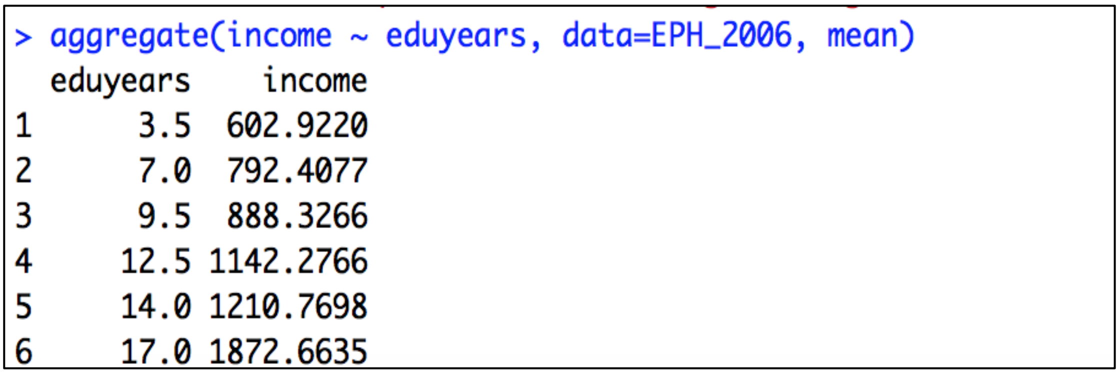
- ✓ Have an understanding of most relevant concepts related to statistical inference and hypothesis testing.
- ✓ Evaluate the assumptions behind linear regression models using a set of specification tests.

2. SET-UP REGRESSION MODEL

We begin by loading the dataset that we will use through this module and performing some basic operations using the commands that we learned in Module 1.2. Let's state a hypothesis we are interested in testing: higher levels of education lead to higher income. Now, perform the following steps in RStudio.

- Load the dataset EPH_2006.csv using the read.csv function. You may notice that the number of variables in this dataset are few but the number of observations is close to 50,000. Indeed, we will introduce even larger datasets in future modules so that you have confidence in using them. This sample includes only employed individuals with positive incomes, dropping some troublesome outliers.
- Use the `aggregate` function to list the descriptive statistics of income by different education level categories. The output should look like Figure 1. You find that individuals with more years of education have higher mean incomes. However, the range (minimum-maximum) at each education level overlaps with the range for some other education levels

```
# mean income
aggregate(income ~ eduyears, data=EPH_2006, mean)
# min
aggregate(income ~ eduyears, data=EPH_2006, min)
# max
aggregate(income ~ eduyears, data=EPH_2006, max)
```



```
> aggregate(income ~ eduyears, data=EPH_2006, mean)
  eduyears    income
1      3.5  602.9220
2      7.0  792.4077
3      9.5  888.3266
4     12.5 1142.2766
5     14.0 1210.7698
6     17.0 1872.6635
```

Figure 1. aggregate function: cross-tabulation of mean income and education levels

- ✓ Use the `group_by` and `mutate` functions from the `dplyr` package to create a new variable called `meanincome` which is equal to the mean of income for each education level.

- ✓ Load the `dplyr` package for this purpose –

  `require(dplyr)`

- ✓ Mean Income:

```
EPH_2006 <- EPH_2006 %>% group_by(eduyears) %>% mutate(meanincome
= mean(income, na.rm = TRUE))
```

  - Rename this variable "Conditional Mean Income" using the following command
```
colnames(EPH_2006)[which(colnames(EPH_2006) == 'meanincome')] <-
'Conditional_Mean_Income'
```
  - Generate another variable called `uniqrecord` which flags a unique record for each level of education. Use the `group_indices` function from dplyr for this.
```
 EPH_2006$uniqrecord <- EPH_2006 %>% group_indices(eduyears)
```
  - Remember that you can use the `? help command` to check the syntax, usage and purpose of the function
```
?group_by                  ?mutate                  ?group_indices
```

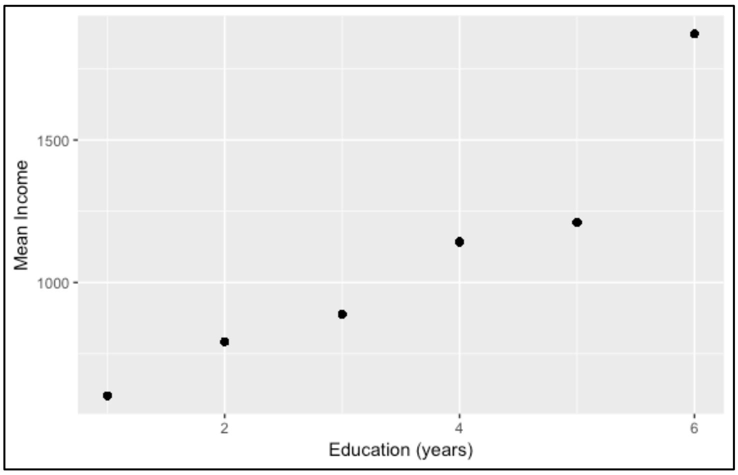
- ✓ Plot mean income levels by education level. Use the `ggplot` function from the ggplot2 package (remember to load it using `require(ggplot2)`) for this.

```
P       <-       EPH_2006       %>%       ggplot(aes(uniqrecord,
   Conditional_Mean_Income))
p + geom_point() +
  xlab("Education (years)") +
  ylab("Mean Income")
```

- ✓ Now, visualize a more informative graph with income and mean income:

```
p <- EPH_2006 %>% ggplot(aes(uniqrecord))
p + geom_point(aes(y=income, color="Income(pesos)")) +
  geom_point(aes(y=Conditional_Mean_Income,           color="Mean
Income")) +
  ylim(0,2000)    +    geom_smooth(aes(y=Conditional_Mean_Income,
color="Fitted Values"), method='lm') +
  xlab("Education (years)") + ylab("Income")
```

We are restricting the sample to incomes less than 2000 so that the fitted line has an identifiable linear slope. You should also see how we have used the `ggplot` function as an illustration of R's powerful graphical capabilities. The graphs produced by each of these commands are shown in Figure 2. The fitted line is the best linear approximation of the relationship between mean income and education years.



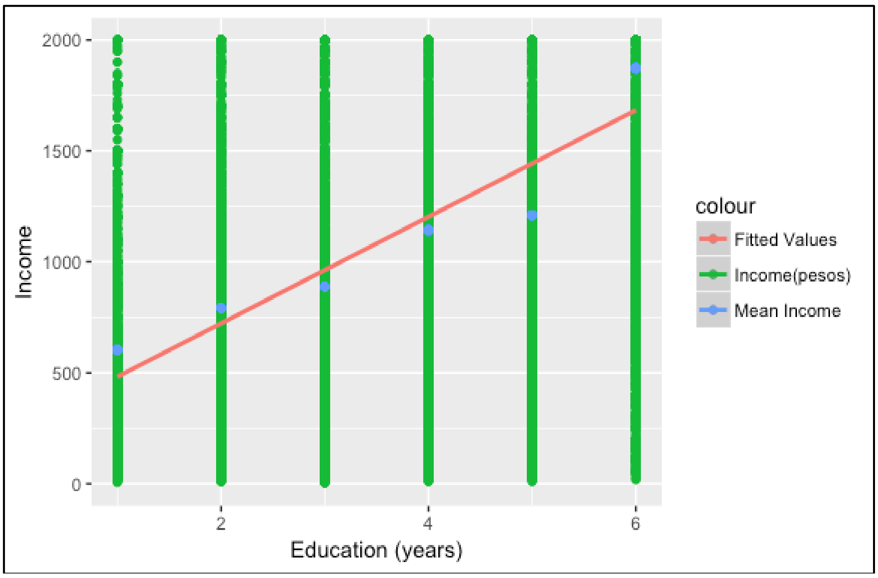


Figure 2. Plots of income, mean income and education levels

- ✓ Define a functional form for the relationship between income and eduyears as:

$$income_i = \beta_0 + \beta_1 eduyears_i + \varepsilon_i$$

Where *i* represent an individual, β_0 is the intercept (or the modeled value of *mean* income when years of education is zero), β_1 is the slope of the line representing a unit change in *mean* income per unit change in years of education, and ε_i represents the deviation in the *actual* income and predicted *mean* income for a given years of education for an individual *i*

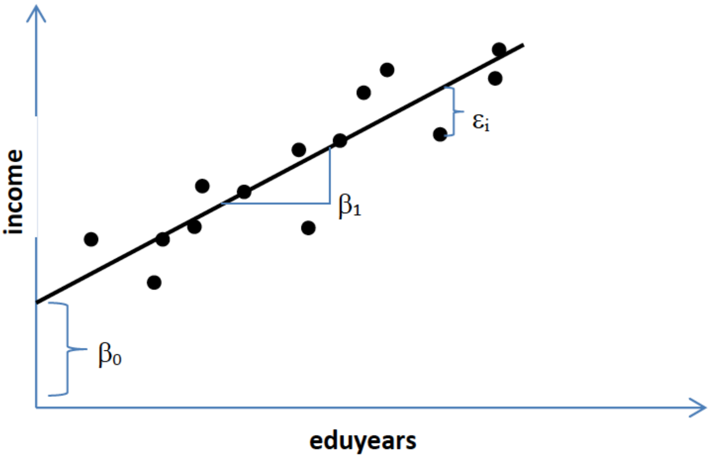
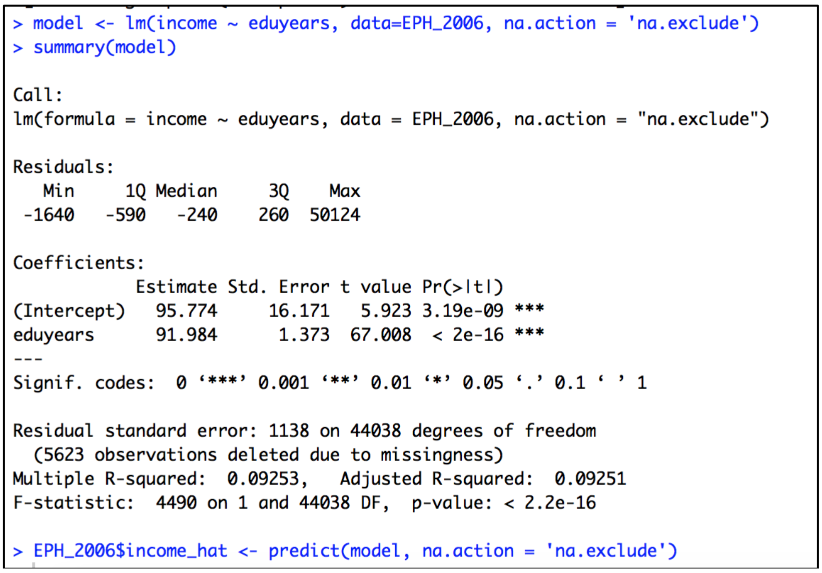


Figure 3. Graphical representation of the fitted regression line

Now run the following regression command and predict mean income; Figure 4 presents the outputs from R.

- `model <- lm(income ~ eduyears, data=EPH_2006, na.action = 'na.exclude')`
- `summary(model)`
- `EPH_2006$income_hat <- predict(model, na.action = 'na.exclude')`



```
> model <- lm(income ~ eduyears, data=EPH_2006, na.action = 'na.exclude')
> summary(model)

Call:
lm(formula = income ~ eduyears, data = EPH_2006, na.action = "na.exclude")

Residuals:
   Min     1Q Median     3Q    Max 
 -1640   -590   -240    260  50124 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   95.774     16.171   5.923 3.19e-09 ***
eduyears      91.984      1.373  67.008  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1138 on 44038 degrees of freedom
  (5623 observations deleted due to missingness)
Multiple R-squared:  0.09253,	Adjusted R-squared:  0.09251 
F-statistic:  4490 on 1 and 44038 DF,  p-value: < 2.2e-16

> EPH_2006$income_hat <- predict(model, na.action = 'na.exclude')
```

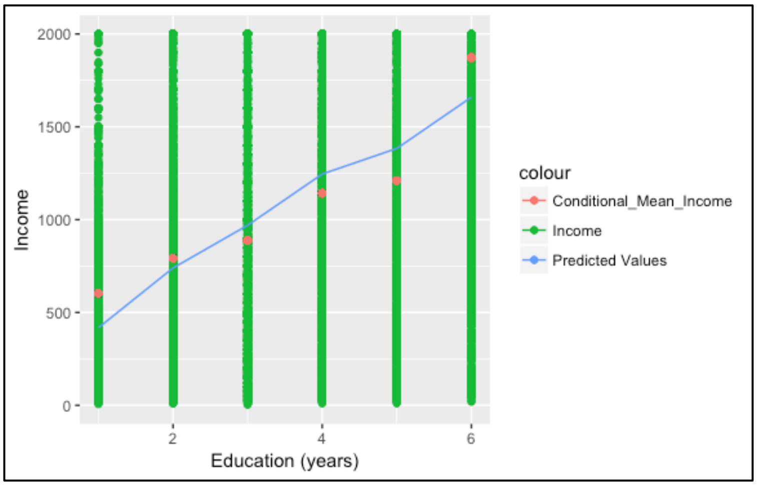
Figure 4. Regress command: relationship between income and education.

✓ Now plot the scatter plot of income and years of education as before, but instead of fitting a line (`method = ‘lm’` command), draw a line using the predicted income values (`income_hat`).

- Do this using the following code:

```
p <- EPH_2006 %>% ggplot(aes(uniqrecord))
p + geom_point(aes(y=income, color="Income")) +
  geom_point(aes(y=Conditional_Mean_Income,
   color="Conditional_Mean_Income")) +
  geom_line(data=EPH_2006[!is.na(EPH_2006$income_hat),],aes(y=income_hat,
   color="Predicted Values")) +
 ylim(0,2000) + xlab("Education (years)") + ylab("Income")
```

Figure 5. Plot of predicted income levels using regression model and education (identical to Fig 2)



Exercise 2.1: From the regression output in Figure 4, state the relationship between income and years of education – is it increasing or decreasing, and at what rate? Also assess whether this association is statistically significant. Comment on whether you would conclude that more education causes increase in the income levels of individuals.

3. STATISTICAL INFERENCE

The goal of this section is to refresh some basic concepts used in statistical inference or hypothesis testing. The need for statistical inference arises because we cannot observe the entire universe or target population. Instead, we have to make predictions about the target population on the basis of a smaller sample from that population, and we have to select a finite number of measurements within that sample. Since we sample only a part of the target population, there is sampling error in all of our estimates. After all, there is no guarantee that a measurement on a sample is precisely the same as that measurement would be over the entire population. For example, if the average age in a sample is 52.3 years, how confident are you that the average age in the entire population is precisely 52.3 years and not, say, 51 years or 48 years?

Theoretically, each sample from a population could yield a different estimate. For example, if we randomly sampled 1000 individuals from a population of 5 million people, then there are a tremendous number of ways of selecting samples consisting of different sets of 1000 individuals. Not all of these samples (indeed, perhaps none of the samples) will have average ages equal to the average age of the population; indeed, some might have sample means that are very different from the population mean. Therefore, the sample mean is best characterized as a distribution of values, where the distribution represents our uncertainty about the population mean. In the next few sections we introduce notation and interpretations to express and understand this uncertainty.

3.1 Key Concepts

As discussed above, we estimate a sample mean only as an approximation of the population mean. Let's denote sample mean as $\bar{Y}$ and the population mean as μ. However, the mean itself tells us nothing about the variance of our estimate. For example, a population with mean age of 53 may comprise 1,000 53-year-olds, but it may instead comprise 500 ten-year-olds and 500 96-year-olds. sample second statistic could give us a sense of the variability of the sample.

Keep in mind that measurement uncertainty is due to sampling, or our inability to know and measure population averages, whereas the variability in the population is a characteristic of the population. We use a statistic called the *variance* to describe the spread of a measurement across the population. The population variance is denoted as σ^2 and the sample variance is denoted as S^2. The formulas that we use to calculate or estimate the population and sample mean and variance are as follows:

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$$

$$\bar{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

$$S^2 = \frac{\sum_{k=1}^{n} (Y_k - \bar{Y})^2}{n-1}$$

where Y_i is the measurement of interest for individual i, N is total number of individuals in the population, k represents an individual in a sample, and n is the sample size.

Sample Mean Theorem: In a random sample from a target population, the sample mean (Ȳ) has an *expectation* equal to the population mean (μ). Expectation is a probabilistic concept which can be interpreted as stating which value of the given parameter should be "expected" given certain information about the population's distribution of the measurement in question; a more comprehensive explanation can be found in any basic statistics and probability textbook.

The Law of Large Numbers: In a random sample from a target population, as the sample size increases, the sample mean converges in probability to the population mean. In other words, no matter the underlying distribution of the measurement of interest, if we have a large sample size, then we can reasonably approximate the population mean by the sample mean.

Central Limit Theorem: The distribution of the sample mean of a variable estimated from a random samples drawn repeatedly from a population distribution with a well-defined expected value for the mean (μ) and variance (σ^2) converges to a normal distribution with mean μ and standard deviation (or standard error) $\sigma/\sqrt{n}$. The central limit theorem plays an extremely important role in statistical inference. No matter how a measurement of interest is distributed across the population, the distribution of its mean can be reasonably approximated by the normal distribution when the sample size is large. This result enables us to conduct hypothesis testing on sample means whether or not we know the underlying distribution of the measurement in question, which we will further discuss below.

Exercise 2.2: Copy and run the following command text in R. We will be asking R to draw different distributions using functions from the ggplot2 and stats packages. You should use the help command (?) for these commands and the distribution functions to better understand them. Although these commands are not necessary for effective policy evaluation, they provide good practice of the R skills we have been building throughout this course.

Play with the options and parameter values to see how the output changes:

```
 *NORMAL DISTRIBUTION
normal <- ggplot(data.frame(x = c(80, 120)), aes(x)) +
  stat_function(fun = dnorm, args = list(mean = 100, sd = 10), colour =
"red") + ylab("") +
  scale_y_continuous(breaks = NULL) +
  ggtitle("Normal density with mean=100 and s.d=5")
normal
 *STANDARD NORMAL DISTRIBUTION
 std_normal <-  ggplot(data.frame(x = c(-4, 4)), aes(x)) +
   stat_function(fun = dnorm, args = list(mean = 0, sd = 1), colour =
 "navy") + ylab("") +
   scale_y_continuous(breaks = NULL) +
   ggtitle("Standard normal density")
 std_normal

 *CHI-SQUARE DISTRIBUTION
 chisq <- ggplot(data.frame(x = c(0, 100)), aes(x)) +
   stat_function(fun = dchisq, args = list(df = 20), colour = "black") +
 ylab("") +
   scale_y_continuous(breaks = NULL) +
   ggtitle("Chi-sq distribution")
 chisq

 *STUDENTS T DISTRIBUTION
 students_t <-  ggplot(data.frame(x = c(-4, 4)), aes(x)) +
   stat_function(fun = dt, args = list(df = 8), colour = "green") +
 ylab("") +
  scale_y_continuous(breaks = NULL) +
  ggtitle("Student's t density")
 students_t

 *F-DISTRIBUTION
 f <-  ggplot(data.frame(x = c(0, 5)), aes(x)) +
     stat_function(fun = df, args = list(df1 = 5, df2 = 10), aes(colour =
 "F(5,10)")) +
     stat_function(fun = df, args = list(df1 = 25, df2 = 50), aes(colour =
 "F(25,50)")) +
     stat_function(fun = df, args = list(df1 = 50, df2 = 100), aes(colour
 = "F(50,100)")) +
     ylab("F-density") +
     scale_y_continuous(breaks = NULL) +
     ggtitle("F-distribution density")
 f
```

3.2 Fun with Dice

In this section, we are going to demonstrate that taking multiple sample means from random samples of a uniformly distributed variable produces a normal distribution. This will solidify your understanding of the Central Limit Theorem and the Law of Large Numbers. Since we are doing this in R, we also hope to learn a new command or two.

To illustrate this idea, consider a simple experiment: the roll of a die. Any time we roll a die, the outcome is one of the possible six values $Yi \in \{1,2,3,4,5,6\}$. Suppose we roll the die a million times.

How many times do you expect number 1 to show up? It is "expected" to show up about 10^9 / 6 times. What is the "expected" population mean? It would be simply (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5.

Let's create a hypothetical population of 50,000 dice rolls in R. Each of the 6 numbers has equal probability of being rolled, so that the expected distribution is uniform and discrete. We create this hypothetical population and display the distribution graphically as follows.

```
set.seed(3254)// When we set the seed, R produces random numbers
in the same way so that results are replicable. It is always a
good idea to set a seed.

dice_rolls_100 <- as.integer(runif(100, 1, 7))
// check the ?runif command to know how to create uniform
distribution of numbers. Now the data will be populated with 100
dice roll results having values between 1 and 6.

summary(dice_rolls_100) // obtain the summary statistics for y.
Observe that the mean is 3.68; not 3.5 as we would expect.
```

```
dice_rolls_1000 <- as.integer(runif(1000, 1, 7))

summary(dice_rolls_1000)
// obtain the summary statistics for y. Observe that the mean
is closer to 3.5 than the previous sample
```

```
dice_rolls_50000 <- as.integer(runif(50000, 1, 7))

summary(dice_rolls_50000)// obtain the summary statistics for y.
Observe that the mean is almost exactly 3.5.
```

The above code demonstrates the law of large numbers. As the sample size increased, the sample mean converges to the population mean. You can see that the data is uniformly distributed with each of the 6 values having equal probability by plotting a histogram of the observations. The result is show in Figure 6.

```
hist(dice_rolls_50000, freq = FALSE, col = 'blue',
main='Probability Distribution', xlab='Dice Value', ylab='Probability')
```

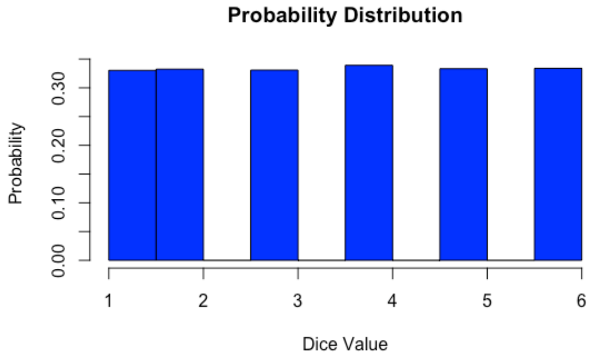


Figure 6. Uniform distribution of dice rolls

Now we assume that the data with 50000 dice rolls is our population. To demonstrate the Central Limit Theorem, we will randomly select 200 rolls from this population and take the mean of that sample. We will repeat the process 400 times, drawing 400 random samples. Plot and summarize the distribution of the sample mean as follows:

```
# Sampling from our population and finding the mean

sample(dice_rolls_50000, 200)
mean(sample(dice_rolls_50000, 200)

# Repeating this 400 times and plotting the distribution of
sample means

sampleMeans <- as.data.frame(replicate(400,
mean(sample(dice_rolls_50000, 200))))

colnames(sampleMeans) <- c('means')

ggplot(sampleMeans, aes(means)) +
geom_histogram(breaks=seq(1,6,0.05), fill="orange",
color="blue") +
  xlim(1,6)


svmat meanY, names( col ) // convert the matrix to a variable
in the dataset so that we can perform operations on it.

sum meanY, d

histogram   meanY,   discrete   frac   xlabel(1(1)6)    ///
graphregion(fcolor(white)) w(0.08) xline(3.5) normal // plot the
distribution of sample mean of Y
```

Figure 7 reports this histogram. We find that the distribution of die rolls is normal (symmetric with thin tails), with a mean close to the population mean of 3.5.

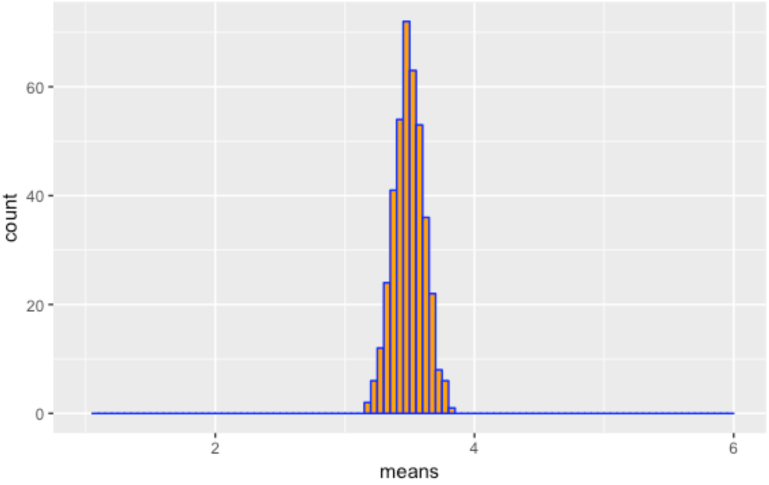


Figure 7. Normal distribution of the sample mean of die rolls

The normal distribution of the mean is also observed for other summary statistics, such as proportion of individuals of a certain categorical type. For example, Figure 8 presents the distribution for the proportion of people with a high-school diploma in the sample. The distribution represents the uncertainty in the sample proportion and this uncertainty reduces as the sample size increases.

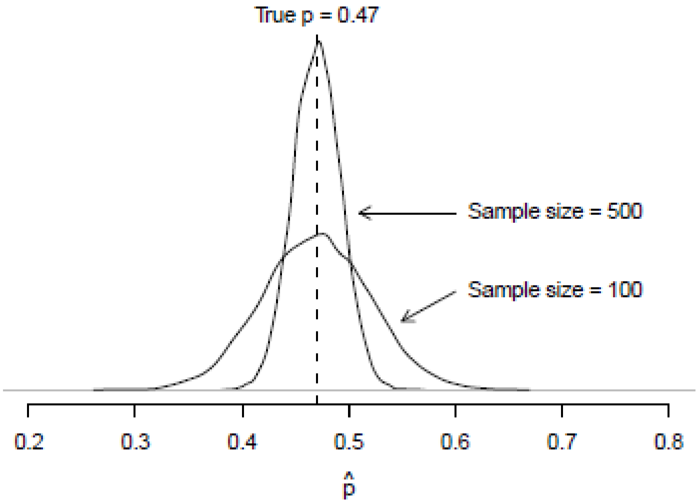


Figure 8. Normal Distribution for the Sample Proportion (Source: Ben Arnold)

3.3 Hypothesis Testing

Hypothesis testing is a fundamental part of any statistical analysis. When we conduct a hypothesis test, we specify a testable null hypothesis and alternative hypothesis, using our data to test whether it provides evidence against the null hypothesis (and therefore in favor of the alternative hypothesis). Below are the key steps in hypothesis testing.

3.3.1 Step 1: Specify the null and alternative hypotheses

For example, the null hypothesis in context of an impact evaluation can be:

H0: The intervention has no effect on the income levels of the household
H1: The intervention changes the income of the household.

The null hypothesis is either rejected or not rejected based on statistical significance tests. Please note: while we sometimes state that "we reject the null hypothesis" with a certain amount of certainty, this never constitutes an acceptance of the alternative hypothesis.

We can also have a null hypothesis like "H0: The mean income in the population is equal to 1200 pesos" and alternative hypothesis like "H1: the mean income in the population is not equal to 1200 pesos."

Sometimes we are interested in checking whether a given statistic is different from zero (statistically significant). In this situation the null hypothesis is, for example, "H0: the rate of increase in income per additional year of education is 0" and the alternative hypothesis is "H0: the rate of increase in income per additional year of education is different from (or greater than) 0".

3.3.2 Step 2: Decide the statistical significance levels or confidence on your inference

No matter large a sample, there always remains a chance that any statistic estimated from the sample is not the population statistic, so that the inference based on such a sample statistic may not accurately reflect the population.

There are two types of errors we consider in hypothesis testing.

- ✓ Type I error: Rejecting the null hypothesis when it is true
- ✓ Type II error: Failing to reject the null hypothesis when it is false

A type I error can be thought of as a "false positive," while a Type II error can be thought of as a "false negative." We will often deal with Type II errors when we discuss the sample design for impact evaluations. For typical hypothesis testing presented here, we mainly worry about the Type I error. We define α, usually known as the statistical significance level, as the probability of committing a Type I error.

How much error are you willing to accept in your statistical inference? Each researcher must determine a level of alpha that is appropriate to his study. There is a tradeoff. The sample size required to make the risk of a Type I error very small is often too high to be affordable for most projects. The values traditionally used in empirical research are 1%, 5% and 10% Type I error rates, but there is no theoretical foundation for assuming these values. In this course, we will typically assume that 5% is an appropriate level of Type 1 error.

3.3.3 Step 3: Calculate the test statistic and its distribution

We are going to discuss variants of the student t-test, which is one of the oldest and most powerful statistical tests to assess whether the difference in two means, or a mean and a constant, is statistically significant (different than 0). In fact, if you are able to design a randomized control trial and achieve perfect balance of covariates at the baseline and perfect compliance with your study design, then a t-test (or its variants) is all that you would need to estimate the effect of the treatment. The general formulae for a t-test are,

✓ **Compare a distribution of sample mean with a fixed value**

$$\mathrm{t} = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

where $\bar{Y}$ is the sample mean, μ_0 is the constant (which can be 0), S is the sample standard deviation, and n is the sample size.

✓ **Compare a distribution of sample mean with another mean distribution**

$$\mathrm{t} = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{{S_Y}^2}{n_Y} + \frac{{S_X}^2}{n_X}}}$$

Where $\bar{Y}$ and $\bar{X}$ are the sample means we wish to compare, S_Y is the sample standard deviation and n_Y is the sample size for the distribution of variable Y, and S_X is the sample standard deviation and n_X is the sample size for the distribution of variable X.

As the numerators in the above formulas suggest, when the two means or the mean and the fixed number are similar or equal the t-statistic is small. Therefore, *under the null hypothesis we expect the t-statistic to be zero*. However, because of = uncertainty in sample means (standard errors), the t-statistic also has an underlying distribution. The standard deviation of t-statistics depends on the degrees of freedom for the t-statistics.

The degrees of freedom are estimated as follows.

✓ Comparing a sample mean distribution with a fixed value: n – 1

✓ Comparing a sample mean distribution with another mean distribution where both have the same variance and same sample size (n): 2n – 2

✓ Comparing a sample mean distribution with another mean distribution where both have the same variance but different sample size (n_1 and n_2): $n_1 + n_2 – 2$

✓ Comparing a sample mean distribution with another mean distribution where both have the different variance and sample size (n_1 and n_2): the formula is complicated.

The properties of a t-distribution are not covered in this module, but you can read more about t-distributions elsewhere if you are interested. However, note that:

- ✓ t-distribution are different for different degrees of freedom
- ✓ As sample size (and thus degrees of freedom) increases, the distribution approaches a standard normal distribution.

3.3.4 Step 4: Compare the t-statistic with the reference distribution

Consider Figure 9, where the t-distribution with sample size of 100 or degrees of freedom of (100 – 2 =) 98 is presented. If the calculated t-statistic is too far away from 0, in the "red zone", then that the researcher can report with high confidence that the difference in the two quantities tested is not zero, and thus that the two quantities have different means. The "red zone" is based on the t-distribution and the Type I error rate (or α) that we set in Step 2. In Figure 9, the Type I error rate or statistical significance level is set at 5% total or 2.5% on either sides. The critical values of t for the given distribution are ±1.98. If the estimated t-statistic is >1.98 or < –1.98 then we state that the null hypothesis is rejected with 95 percent confidence.

For example, when testing whether the mean income is 1200 pesos with a sample size of 100, we obtain a t-statistic of 0.45, implying that the sample does not provide evidence supporting the rejection of the null hypothesis. On other hand, if the t-statistic had been 3.6, then we would reject the null hypothesis that the sample mean is 1200 pesos at a statistical significance level of α=5%.

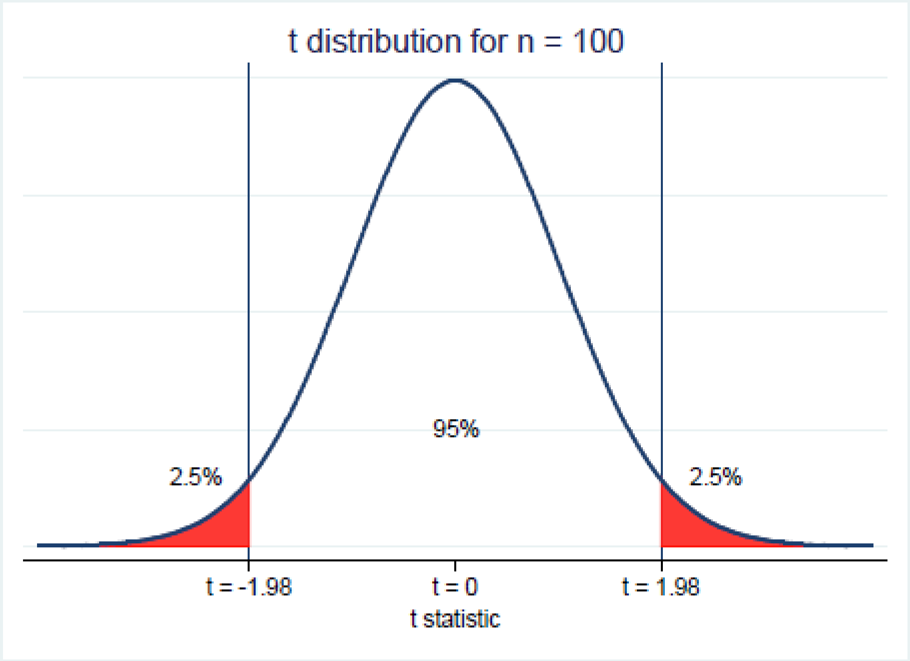


Figure 9. Example of t Distribution

The critical t-value is different when we make one-sided or one-tailed comparisons. If the alternative hypothesis explicitly states that differences in the sample means or a sample mean and a fixed value are *higher* or *lower* than zero, then we are only concerned about the estimated t-statistic being too large in a single direction, either positive or negative. The critical t-value in that one direction is slightly lower than the t-value in the two-tailed comparison, making it slightly 'easier' to reject the null hypothesis.

Another related concept in statistical inference is the p-value. A critical t-value is determined on the basis of degrees of freedom (or sample size) and the pre-set Type I error or statistical significance level, α. We reject or fail to reject the null hypothesis at this significance level. We could also ask at what level we could have to set α in order that we would reject the null hypothesis with confidence level α given our sample. *The **p-value** is the probability of obtaining a test statistic at least as large as the one that was actually observed, assuming that the null hypothesis is true.* In simpler words, it is the probability of finding a difference different than 0 (the null) purely due to random sampling error when in fact the null is true.

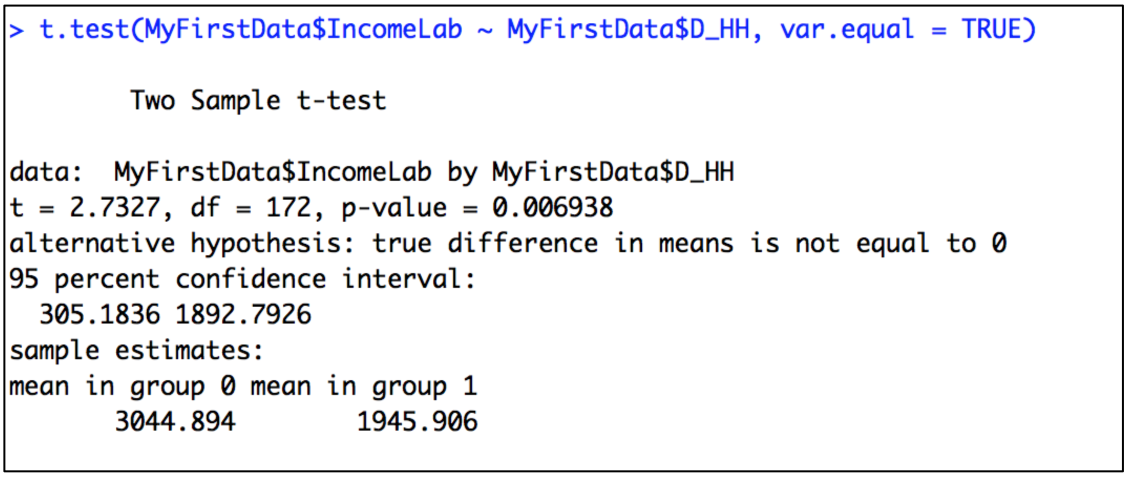
3.4 Example of t-test

We repeat the example from Module 1 for the t test example as a demonstration. The t-statistic is 2.7327. The critical t-value for a distribution with degrees of freedom of (174 – 2 =) 172 for a two sided test at significance level of 5% can be displayed by the following R command:

```
qt(0.025, 172)
```

You will find that the critical t-value is 1.974. Since the t-statistic is larger than the critical value, we reject the null hypothesis with 95 percent confidence.

We also find that the p-value is 0.0069 for a two sided t-test, implying that, if the null hypothesis were true, the difference in mean income between the two groups would occur due to random sampling error about 0.69% of the time.



```
> t.test(MyFirstData$IncomeLab ~ MyFirstData$D_HH, var.equal = TRUE)

        Two Sample t-test

data:  MyFirstData$IncomeLab by MyFirstData$D_HH
t = 2.7327, df = 172, p-value = 0.006938
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  305.1836 1892.7926
sample estimates:
mean in group 0 mean in group 1
       3044.894        1945.906
```

Figure 10. Output of t.test command from Module 1

Exercise 2.3: Load EPH_2006.csv. Run a t-test to assess if the income is different by sex of the person in R. Follows the 4 steps described above and interpret the findings.

4. BASICS OF REGRESSION ANALYSIS

Regression analysis is a tool to study associations or relationships between variables of interest. Under some circumstances, regression analysis can be used to infer causal relationships between covariates and a variable of interest. However, it is important to note that the underlying theory and study design help us to determine whether the observed association is causal; regression analysis alone will only tell us whether the association is statistically significant or not. For example, simple regression analysis suggests that individuals with higher levels of education also have higher incomes, but this does not imply that an individual's obtaining more education will thereby increase their income. In order to establish these kinds of causal relationships, we have to argue from established theories and convincingly reject any alternative explanations of the regression findings. For example, in order to show that providing people with more education increases their income, researchers would need to show that alternative factors, like family wealth or cultural mores, are not causing both higher education *and* higher income, yielding a spurious positive correlation between them. There are several free online courses for you to understand the basics of regression.

4.1 The Impact Evaluation Perspective

Most of the modern impact evaluation literature is based on the ***potential outcomes framework*** (Rubin 1974, Holland 1986) as illustrated below.

Consider the following regression:

$$HH\_Income_i = \beta_0 + \beta_1 Oportunidades_i + \varepsilon_i$$

where we hypothesize that the income for household i (HH_Income_i) is changed by the participation of the household in Oportunidades program ($Oportunidades_i$) and moderated by unmeasured factors summarized in ε_i. We assume that the relationship between the dependent and the independent variables is linear. We can define the *expected* outcome for the households who participated in Oportunidades and those who didn't as follows.

Did Not Participate: $E(HH\_Income_i\,|Oportunidades_i = 0) = \beta_0 + E(\varepsilon_i\,|\,Oportunidades_i = 0)$

Participated: $E(HH\_Income_i\,|Oportunidades_i = 1) = \beta_0 + \beta_1 + E(\varepsilon_i\,|\,Oportunidades_i = 1)$

The impact of Oportunidades is the difference in the conditional expected means for the household income as follows:

$$Impact = [\beta_0 + \beta_1 + E(\varepsilon_i\,|\,Oportunidades_i = 1)] - [\beta_0 + E(\varepsilon_i\,|\,Oportunidades_i = 0)]$$

$$\boldsymbol{Impact = \beta_1 + [E(\varepsilon_i|\,Oportunidades_i = 1) - E(\varepsilon_i|\,Oportunidades_i = 0)]}$$

Notice that the difference between program participants and non-participants consists of the true parameter of interest β_1 and the term which contains non-observable differences between treated and non-treated individuals, $[\boldsymbol{E}(\boldsymbol{\varepsilon_i}|\ \boldsymbol{Oportunidades_i} = \mathbf{1}) - \boldsymbol{E}(\boldsymbol{\varepsilon_i}|\ \boldsymbol{Oportunidades_i} = \mathbf{0})]$. We usually call this term ***selection bias***.

One of the goals of good impact evaluation design and analysis is to mitigate the role of this selection bias. In Randomized Control Trials (RCTs) or experimental evaluations with sufficiently large number of participants in both groups, the non-observable and non-measured factors are expected to be distributed similarly such that the selection bias is zero. In this design, regression analysis is expected to estimate true impacts of the program β_1. In observational or quasi- experimental studies, we use a combination of study design and analysis techniques to rule out or reduce the effect of selection bias as discussed over the next few lectures.

Recall from the previous module' exercises that the t-test and regression analysis both gave the same t-statistic and inference about the impact of a program in a simple setting. However, as the designs get more complex and we have to account for non-compliance and imbalance in the baseline groups (as we will discuss later), we will have to rely on regression analysis instead of simpler group mean tests.

4.2 Inference in Regression Analysis

The logic of hypothesis testing or statistical inference is same as that discussed in section 3.3. in this module. In evaluating the impact of Oportunidades above, we are interested to know if β_1 is statistically different from 0 (no impact) or not.

Therefore, we specify the hypothesis as:

H0: $\beta_1 = 0$
H0: $\beta_1 \neq 0$

The t-statistic is calculated as,

$$t = \frac{\beta_1 - 0}{SE_{\beta_1}}$$

The regression model outputs the t-statistic by default and also reports the p-value as the probability that the coefficient is different from zero (the null is rejected) due to sampling error. We also know the critical t-values for the given degrees of freedom or the critical z-value on basis of standard normal distribution, so we can compare whether the t-statistic is further away from the critical t-value. When we reject the null hypothesis, we say that the coefficient β_1 is statistically significant at the given α level.

4.3 Assumptions of Regressions

As you may remember from previous coursework, the estimated coefficients only approximate the true population parameters under certain assumptions. For the estimates from a regression analysis to be plausible, we should check whether the data conform to these assumptions. Wooldridge (2009) provides the following set of assumptions.

- ✓ Linearity in parameters: $E(Y|X_k) = \beta_0 + \sum_k \beta_k X_k + \varepsilon_i$. This means that the expected mean of the outcome variable (Y) is a linear combination of the regression coefficients and the k predictor variables (X_k) and the error term (ε_i). The variables can be transformed to log, power terms or any other transformation but the model remains additive and linear in terms of coefficients.

- ✓ Random sampling or independence of error: This assumption states that the errors of the outcome variables are uncorrelated with each other Intuitively, it means that each observation of Y is independently draws from the population from other outcomes.

  In practice, observations from the same group or place share common characteristics. For instance, households from the same village share same infrastructure. We also select the samples in multiple stages to ensure that the field work is feasible and within budget. Therefore, the assumption of independent Y is hardly valid in most cases. One econometric tool that we use to mitigate this 'clustered' correlation between the error terms of our sample is clustered standard errors; these can be implemented using the `lm.cluster` command from the `multiwayvcov` package. (See the documentation for more information).

- ✓ No perfect collinearity: No two variables can be identical, and no variable can be a linear combination of any other set of variables.

- ✓ Homoscedasticity. This means that different outcomes observations have the same error variance; the distribution of the individual-characteristic error terms assigned to each individual must have equal variability. In practice, this assumption is often violated and we say that we have heteroskedastic error; clustering standard errors (see above) partially accounts for this problem

- ✓ Zero conditional mean for the error: $E(\varepsilon_i|X_k) = 0$. This means that on average we expect the error term to have no effect on the conditional / expected outcome or that the factors included in the error term would on average cancel out themselves.

- ✓ Normality: $\varepsilon_i \sim N(0, \sigma^2)$. The error term is distributed normally with mean of 0 (assumption 5) and variance equal to that of the outcome (Y).

We will demonstrate a few tests to see if above assumptions hold or are violated in a case study next.

4.4 Example of Regression Analysis

Now let's use real data from Oportunidades dataset called *DataFinal.dta* for to demonstrate the concepts presented so far as follows.

- ✓ Open the data:

```
DataFinal <- read.csv("DataFinal.csv")
```

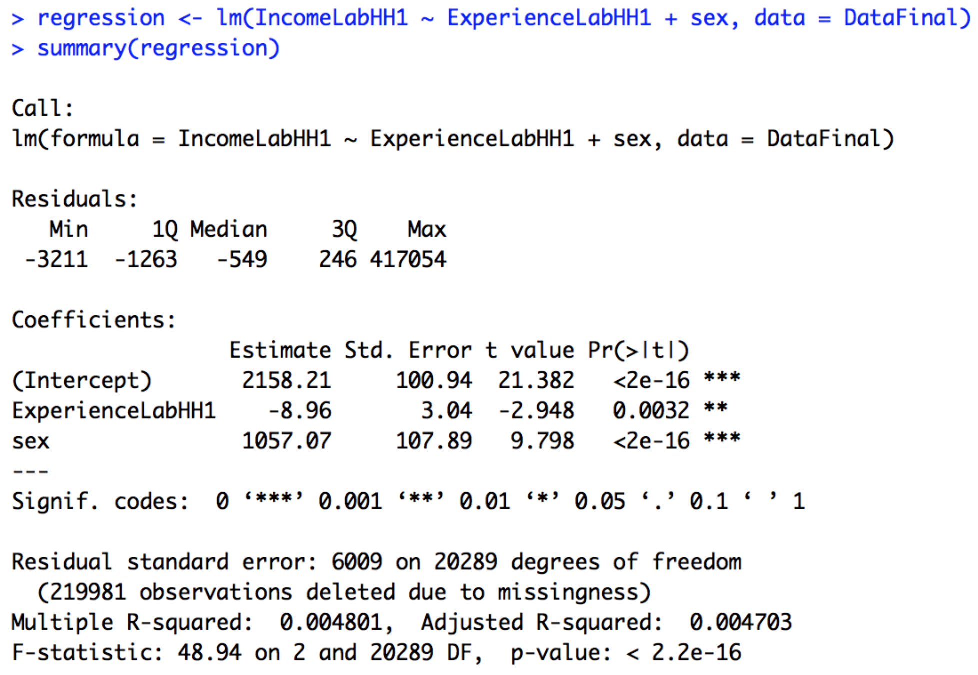
✓ Use the ‘summary’ function to see the variables in the dataset. For sake of demonstration, let’s change the experience of the household head from months to years as follows,

```
summary(DataFinal)
DataFinal$ExperienceLabHH1 <-
DataFinal$ExperienceLabHH1/12
summary(DataFinal$ExperienceLabHH1)
```

✓ Let’s assume that we expect the income of the household head to be determined by the years of experience and the person’s sex. We can specify a regression model to analyze this association as:

```
regression <- lm(IncomeLabHH1 ~ ExperienceLabHH1 + sex, data =
DataFinal)
summary(regression)
```

Figure 12 presents the output from R:



```
> regression <- lm(IncomeLabHH1 ~ ExperienceLabHH1 + sex, data = DataFinal)
> summary(regression)

Call:
lm(formula = IncomeLabHH1 ~ ExperienceLabHH1 + sex, data = DataFinal)

Residuals:
   Min     1Q Median     3Q    Max 
 -3211  -1263   -549    246 417054 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2158.21     100.94  21.382   <2e-16 ***
ExperienceLabHH1    -8.96       3.04  -2.948   0.0032 ** 
sex               1057.07     107.89   9.798   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6009 on 20289 degrees of freedom
  (219981 observations deleted due to missingness)
Multiple R-squared:  0.004801,	Adjusted R-squared:  0.004703 
F-statistic: 48.94 on 2 and 20289 DF,  p-value: < 2.2e-16
```

Figure 12. Case study application: output of regression model

✓ The coefficient associated with years of experience (ExperienceLabHH1) is negative (-8.96), indicating that more years of experience are associated with lower wage income. This seems to go against our intuition; we will return to interpreting this coefficient at a later time.

The sex variable has a positive coefficient (1.057), which implies that men are associated with higher wages than women. The exact interpretation of the estimated coefficients depends on their variable type. In this case, the dependent variable is continuous, so the coefficients should be interpreted using ratio units, the unit of the covariate over the unit of the dependent variable. We find that an additional year of experience (continuous variable) is associated with a 8.96 peso reduction in wage, while men are associated with wages 1,057 pesos higher than those of women.

The F-statistic is statistically significant at the 5% level (indeed, it's significant at the 0.1% level), indicating that the likelihood that model has no explanatory power is very small. If we look at the model coefficients associated with sex and the proxy of experience, we note that both are also statistically significant (p == 0.009) at a 95% confidence level. In other words, both experience and gender are important factors in predicting wage income. The R^2 of 0.0048 indicates that about 0.5% of the variance of wages can be explained by the model. ***Note:*** in prediction or forecasting applications of regression models, the R^2 is important because we want the model to largely explain the variation in the dependent variable, but low R^2 statistics are less of a concern in statistical inference, which attempts to understand key components of the variation in the dependent variable rather than fully explain it.

4.5 Diagnosing Regression Assumptions

In Section 4.3, we listed the key assumptions in regression analysis. Below, we evaluate whether these assumptions hold true for a case study application using *DataFinal.csv*

4.5.1 Detecting influential data

Some observations have a very strong or disproportionate effect on the regression estimates. It is important to ensure that these observations are not wholly responsible for the statistical and economic significance of our coefficients; their importance suggests the possibility of either erroneous data collection or model misidentification. There are three main ways we can categorize the "unusual" observations:

- ✓ Outliers: In linear regression, an outlier is an observation with a large residual. In other words, it is an observation whose dependent variable value is highly unusual given its covariate values. An outlier may be an indication of a sampling peculiarity or may indicate a data entry error or other problem.

- ✓ Leverage: An observation with an extreme value on a predictor variable is called a point with high 'leverage. Leverage is a measure of how far a measurement deviates from the relevant covariate's mean. These leverage points can have an outsized effect on the estimate of regression coefficients.

- ✓ Influence: An observation is said to be influential if removing the observation substantially changes the estimated coefficients. Influence can be thought of as the product of leverage and "outlier-ness."

There is no fully-generalizable method to detect high-influence observations, but you can combine a few statistics and graphical tools to identify such observations. However, whether these observations should be removed from your analysis (especially for a small dataset) is a decision that must be made carefully.

4.5.2 Normality of error term / residual

Now consider the normality of residuals assumption. This assumption important implications for standard methods of inference (or estimation of the standard errors) as well as the validity of the hypothesis testing. However, a violation of this assumption does not necessarily bias the coefficient. So long as residuals are independently and identically distributed (i.i.d.) and the linearity and multicollinearity conditions hold, least square regression coefficients are unbiased (though the standard errors reported by R may be incorrect).

We recommend checking the normality of the residuals by conducting the following test: Run a regression model, estimate the residuals by subtracting predicted outcome values from the observed values, and then plot their distribution and compare it with a normal distribution:

```
regression_model_resid <- resid(regression_model)
# subsetting to values < 10000 to remove outliers
regression_model_resid       <-       subset(regression_model_resid,
regression_model_resid < 10000)
plot(density(regression_model_resid),   main   =   "Kernel   Density
estimate", xlab = "Residuals", ylab = "Density", col="red")
curve(dnorm(x,mean(regression_model_resid),sd(regression_model_resid
)), add = TRUE, col="green")
legend("topright", inset=0.2, legend = c("Kernel density estimate",
"Normal Density"), col = c("red","green"), lty=1:2, cex=0.7)
```

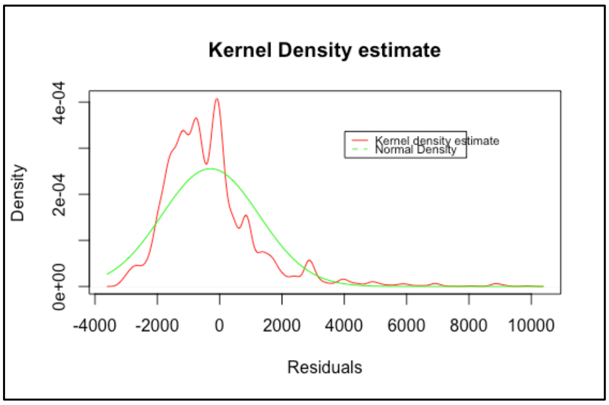


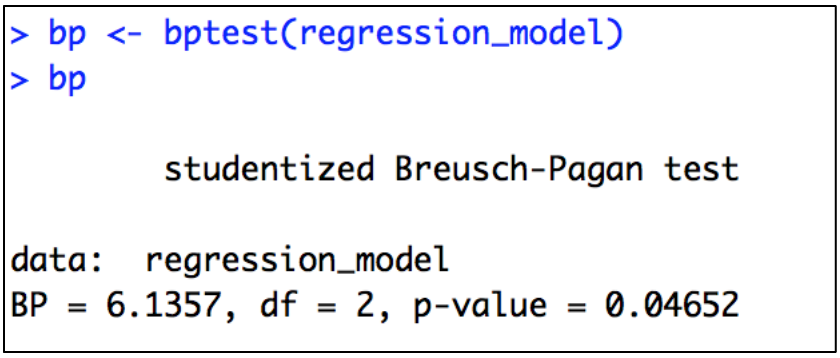
Figure 13. Comparison of distribution of residuals with normal distribution

We find that the overlap with the normal distribution is poor in Figure 13 (the residuals appear to be right-skewed), and thus infer that the normality assumption is violated.

4.5.3 Evaluating homoscedasticity

There are graphical and non-graphical methods to detect heteroscedasticity. Non-graphical methods are based on the computation of statistical tests such as the Breusch-Pagan test. The `bptest` function is present in the `lmtest` package. The null hypothesis in this tests is that residuals have constant variance (i.e. are homoscedastic). For example, run the following commands:

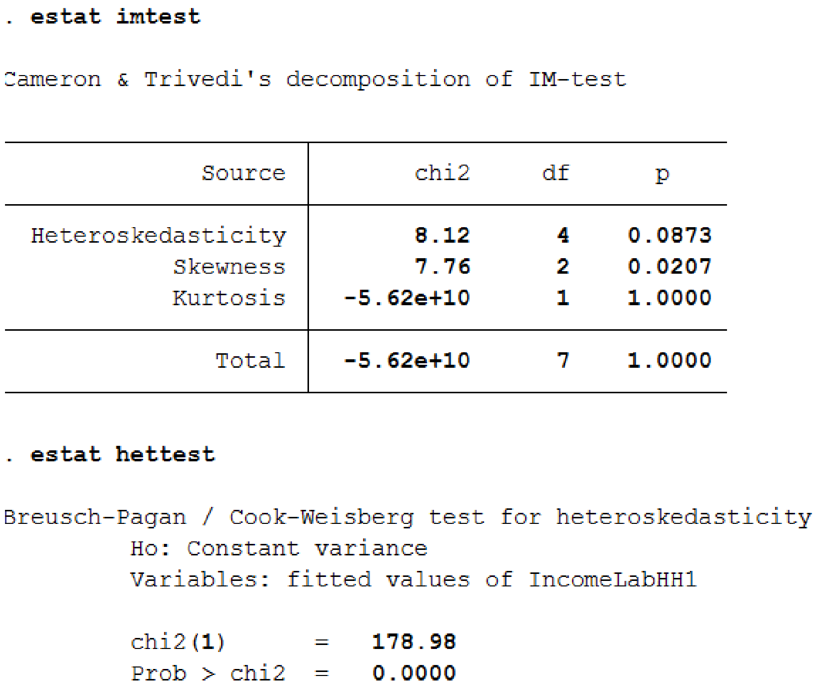
```
bp <- bptest(regression_model)
bp
```



```
> bp <- bptest(regression_model)
> bp

        studentized Breusch-Pagan test

data:  regression_model
BP = 6.1357, df = 2, p-value = 0.04652
```



```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       8.12      4    0.0873
            Skewness |       7.76      2    0.0207
            Kurtosis |  -5.62e+10      1    1.0000
---------------------+-----------------------------
               Total |  -5.62e+10      7    1.0000
---------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of IncomeLabHH1

         chi2(1)      =   178.98
         Prob > chi2  =   0.0000
```

Figure 14. Checking for heteroscedasticity

In this example, the p-values are <0.05, so we can reject the null hypothesis of homoscedasticity. A common practice is to supplement the information provided by these diagnostic tests with graphical information displaying the severity of the problem. A simple method to detect heteroscedasticity is to make a scatter plot of residuals on the predicted values. For this we can use the plot function with the model, which gives the first plot as the Residuals v Fitted plot:

```
plot(regression_model)
```

Figure 15 displays the first output plot of above command. As we move to the right, the dispersion of the residuals increases, indicating the presence of heteroscedasticity.

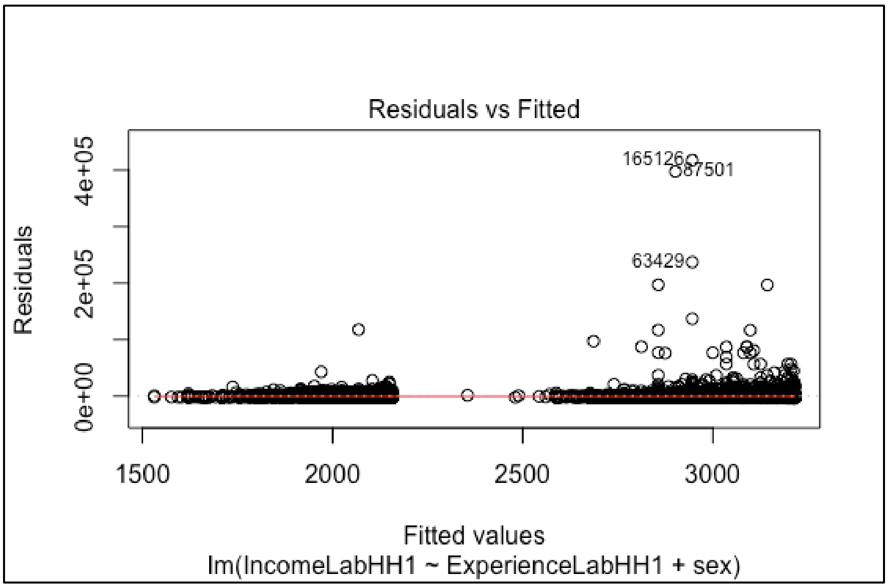
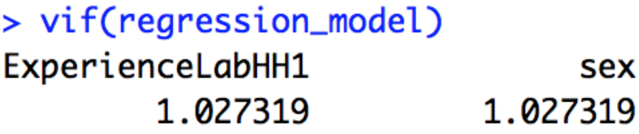


Figure 15. Graphical Display of Heteroscedastic Residuals

The presence of heteroscedasticity does not bias the coefficient estimates, but we need to adjust the estimated standard errors of those coefficients. In fact, we recommend that researchers *always* report standard errors robust to heteroscedasticity. R enables heteroskedastic-robust standard errors using the `lm.cluster` command from the `multiwayvcov` package. As discussed previously, most real life samples are multistage (for example, select a state, then a region, then villages, and finally households) and the respondents within a cluster may share certain characteristics so that they cannot be assumed independent of each other.

4.5.4 Verifying multicollinearity

Multicollinearity occurs when one covariate is a linear combination of some subset of the other covariates. While correlation between the covariates is often expected, such 'perfect' correlation makes it impossible to estimate unique regression coefficients that minimize least-squared error. In R, this assumption can be verified by using the `vif` function from the `car` package, which provides the Factor Variance Inflation (VIF). VIF values above about 10 indicate the presence of multicollinearity. Figure 16 shows that we find that the VIF is only slightly higher than 1 in our model, implying that multicollinearity is not present.



```
> vif(regression_model)
ExperienceLabHH1              sex
        1.027319         1.027319
```

Figure 16. Test for multicollinearity

4.5.5 Verifying linearity

Linearity means that the relationship between the dependent variable and independent variables is linear. While in most cases we justify this assumption theoretically, we can also conduct a variety of empirical tests of linearity.

4.5.6 Model specification

A regression model can suffer from inclusion of irrelevant variables (over-specification of the model) or omitted variables (under-specification model). Specification problems have important implications for the estimation of regression coefficients. While including irrelevant variables does not bias the estimated coefficients, their inclusion can inflate the variance of standard errors. Omitting relevant variables, on the other hand, leads to biased estimates.

In experimental or quasi-experimental impact evaluation designs with a control group, the covariates of interest are indicator variables for participation in each treatment. Other variables either improve the precision of the estimation, adjust for any baseline imbalance between the two groups, or (in quasi-experimental impact evaluations) control for the selection process of the study participants. In experimental designs, we recommend that researchers take care to not include too many covariates in their baseline models, erring on side of inclusion of fewer covariates in the interest of minimizing spurious enlargement of standard errors. The problem of over- or under- specification is more of a concern when we are doing a cross-sectional analysis and inferring causality without a control group, or when the analysis is only quasi-experimental.

5 BIBLIOGRAPHY/FURTHER READINGS

1. Wooldridge, Jeffrey M. (2009). *Introductory Econometrics: A Modern Approach, 4th Edition*