Social Sciences Intro to Statistics

Week 8.2 Bivariate Regression, Hypothesis Testing

Week 8: Learning goal - Apply understanding of bivariate regression to do hypothesis testing for continuous variables.

Introduction

Lecture overview:

- Hypothesis testing
- Regression with continuous variables
- Hypothesis testing about B1
- Factor Variables

Load packages:

```
library(tidyverse)
library(ggplot2)
library(haven)

load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/els/output_

# ELS data frames
els <- df_els_stu_allobs_fac
```

Hypothesis testing about β_1

Taking what we learned last time about bivariate regression, let's find out how we can conduct hypothesis testing. We are going to test hypotheses using β_1 as the point estimate $\hat{\beta}_1$, which you calculate from R

```
mod1 <- lm(formula = bytxmstd ~ bytxrstd, data = els)

summary(mod1)
#>
#> Call:
#> lm(formula = bytxmstd ~ bytxrstd, data = els)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -26.703  -4.434  -0.071   4.144  39.084
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 7.592331   0.213320   35.59   <2e-16 ***
#> bytxrstd    0.850036   0.004182  203.26   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.715 on 16195 degrees of freedom
#> Multiple R-squared:  0.7184, Adjusted R-squared:  0.7184
#> F-statistic: 4.131e+04 on 1 and 16195 DF,  p-value: < 2.2e-16
```

Regression with continuous variables

Research question

When posed in a correlational way: - What is the relationship between reading test score (X) and math test score (Y)?

When posed in a causal effects way: - What is the effect of reading test score (X) on math test score (Y)?

- Population Linear Regression Model
    - $Y_i = \beta_0 + \beta_1 X_i + u_i$
    - where:
        - Y_i: math test score for student i

    * X_i: reading test score for student i
    * β_1: population regression coefficient, is the average change in the value of Y associated with a one-unit increase in X

- OLS Prediction Line
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - $\hat{Y}_i = 7.59 + 0.85 \times X_i$

Looking above, it seems like the interpretation of β_1 – "the average change in the value of Y associated with a one-unit increase in X" – *is* the answer to our primary research question

- that's because it is!

The fundamental goal of causal inference research is to make statements about β_1

- in causal inference research we specify our research question using the form "What is the effect of X on Y;
- and β_1 represents the relationship between (a one-unit change in) X and Y

But β_1 is a population parameter. We usually don't know it. For two reasons:

1. We usually have data from a single random sample, not the entire population
2. If we are trying to estimate causal relationships then β_1 represents the *causal effect* of a one-unit increase in X on the value of Y, not the correlational/associational relationship between X and Y

- formally, β_1 is the relationship between X and Y if values of X were randomly assigned (i.e., an experiment)
- We usually don't have experimental data, so we use regression (or other methods) as an attempt to "recreate" experimental conditions

Let's put aside the causal/experimental concern for now and focus on the first problem: we want to make statements about β_1 but β_1 is a population parameter and we only have data from a single random sample. so what do we do?:

- calculate OLS estimate $\hat{\beta}_1$
- Use $\hat{\beta}_1$ to test hypotheses about β_1

Hypothesis testing about β_1

We always test the same hypothesis about β_1

- $H_0 : \beta_1 = 0$

  - Means that the slope of the relationship between X and Y is `0`; that is, there is no relationship X and Y

- $H_a : \beta_1 \neq 0$
  - there is a relationship X and Y)

Why this hypothesis?

- In causal research, the research question is "What is the effect of X on Y"
- If we cannot reject $H_0 : \beta_1 = 0$, then this answers our research question

Overview

Recall that we followed these five steps when testing hypotheses about μ_Y and when testing hypotheses about whether the population means of two groups are equal to one another (e.g., $\mu_{treatment} = \mu_{control}$):

1. Assumptions
2. Specify null and alternative hypotheses
3. Test statistic
4. P-value
5. Conclusion

When testing hypotheses about β_1, we follow these same five steps!

1. Assumptions
2. Specify null and alternative hypotheses
   - $H_0 : \beta_1 = 0$
   - $H_a : \beta_1 \neq 0$
3. Test statistic
   - calculate test statistic under the assumption that $H_0 : \beta_1 = 0$ is true
   - Draw the sampling distribution of $\hat{\beta}_1$ centered at $\beta_1 = 0$
   - Plot your point estimate $\hat{\beta}_1$ from your single random sample
   - test statistic t calculates the distance between $H_0 : \beta_1 = 0$ and $\hat{\beta}_1$ in terms of standard errors, so that we can assign probabilities to this distance
4. P-value
   - if the probability (p-value) of observing a $\hat{\beta}_1$ as far away from $H_0 : \beta_1 = 0$ as the one we observed is small, then we reason it is unlikely that H_0 is true, and then we reject H_0
5. Conclusion

Assumptions

For now, we'll state the following assumptions as necessary to test hypotheses about β_1:

- Draw random sample
- sample size is large enough to assume that sampling distribution of $\hat{\beta}_1$ is normally distributed (central limit theorem)

Testing hypotheses about β_1 requires more assumptions; we'll introduce these later

Note: in our example, relationship between reading test score (X) and math test score (Y) for students, we our pretending that our sample is a random sample from the population of all students.

Specify hypotheses

RQ: What is the relationship between reading test score (X) on math test score (Y)?

- Null hypothesis, H_0
  - $H_0 : \beta_1 = 0$
  - in words: there is no relationship between reading test score (X) and math test score (Y)?
- $H_a : \beta_1 \neq 0$
  - in words: there is a relationship between reading test score (X) and math test score (Y)?

Good to set alpha level (rejection region) at the same time we specify null and alternative hypotheses

- let's choose α of `.05`

Note: We almost always test two-sided hypotheses about regression coefficients

- Why? Because we can be wrong about the direction of β_1!
- Some policies can cause more harm than good! In fact, it is quite common to find policies that affect the outcome in the opposite direction than is intended!

Test statistic and p-value

After calculating the OLS estimate $\hat{\beta}_1$, we can calculate a test statistic, t, that will provide evidence necessary to make a decision about rejecting H_0 or not

Recall the general formula for (any) test statistic

- $test\_statistic = \frac{\text{(sample estimate)}-\text{(value hypothesized by } H_0)}{\text{(sample standard error)}}$
- When testing hypothesis $H_0 : \beta_1 = 0$ about a regression coefficient:
  - "sample estimate" is: $\hat{\beta}_1$
  - "value hypothesized by H_0" is: $\beta_{1,H_0} = 0$
  - "sample standard error" is: $SE(\hat{\beta}_1)$, the sample standard error of the regression coefficient, $\hat{\beta}_1$

Test statistic for testing hypothesis about β_1

- $t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$

Calculating t for our RQ: relationship between reading test score (X) and math test score (Y)

- $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.8500}{0.0042} = 203.26$

Based on output from below regression model

- $\hat{\beta}_1$: 0.85
- $SE(\hat{\beta}_1)$: 0.0042
- t: 203.2601
- p-value associated with t: 0

```
mod1 <- lm(bytxmstd ~ bytxrstd, data = els)

summary(mod1)
#>
#> Call:
#> lm(formula = bytxmstd ~ bytxrstd, data = els)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -26.703  -4.434  -0.071   4.144  39.084
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
```

```
#> (Intercept) 7.592331   0.213320   35.59   <2e-16 ***
#> bytxrstd    0.850036   0.004182  203.26   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.715 on 16195 degrees of freedom
#> Multiple R-squared:  0.7184, Adjusted R-squared:  0.7184
#> F-statistic: 4.131e+04 on 1 and 16195 DF,  p-value: < 2.2e-16

# printing output from the element named coefficients
summary(mod1)$coefficients
#>              Estimate  Std. Error   t value      Pr(>|t|)
#> (Intercept) 7.5923310 0.213319614  35.59134 3.345643e-267
#> bytxrstd    0.8500363 0.004182012 203.26012  0.000000e+00
```

Conceptual understanding of test statistic

Conceptually, test statistic for $H_0: \beta_1 = 0$ is the same as test statistic about the value of a single population mean, μ_Y, and is as follows:

We calculate the test statistic under the assumption that $H_0: \beta_1 = 0$ is true

We draw the hypothetical sampling distribution of $\hat{\beta}_1$ centered at $H_0: \beta_1 = 0$

- Imagine we take a random sample from the population and calculate $\hat{\beta}_1$; then do that 1,000 times, 10,000 times
- Each observation in the sampling distribution is an estimate $\hat{\beta}_1$ from a single random sample
- Drawing from the central limit theorem, the sampling distribution is normally distributed
- $SE(\hat{\beta}_1)$, the sample standard error of $\hat{\beta}_1$ is an estimate of how far away, on average, the value of $\hat{\beta}_1$ from a single random sample is from the value of the expected value $E(\hat{\beta}_1)$, which is the mean value of $\hat{\beta}_1$ from an infinite number of random samples
  - recall the "standard error" is also called "standard deviation of the sampling distribution"

We plot our point estimate $\hat{\beta}_1$ from our single random sample on the sampling distribution of $\hat{\beta}_1$ centered at $H_0: \beta_1 = 0$

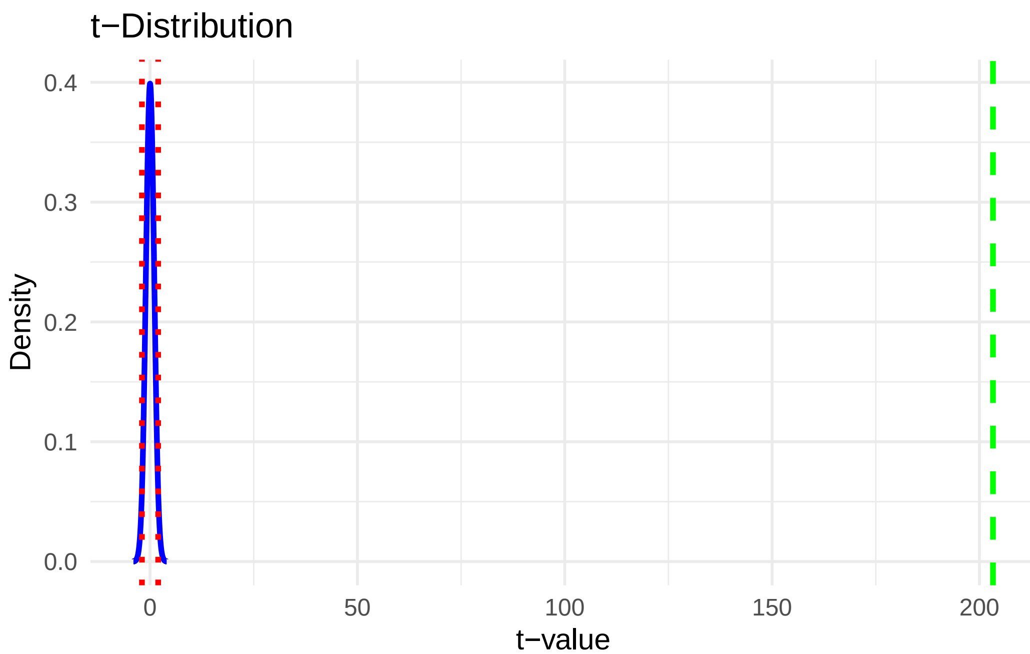
- the test statistic t calculates the distance between $H_0: \beta_1 = 0$ and $\hat{\beta}_1$, and converts this distance in terms of standard errors, $SE(\hat{\beta}_1)$

- Because the sampling distribution is normally distributed, we know that approximately 68% of observations fall within one standard error of the mean, 95% of observations fall within two standard errors of the mean, 99% of observations fall within three standard errors of the mean, etc.
  - the value of t for our regression was: 203.2601!!!
- if the probability (p-value) of observing a $\hat{\beta}_1$ as far away from $H_0 : \beta_1 = 0$ as the one we observed is small, then we reason it is unlikely that H_0 is true, and then we reject H_0

Below, we plot the sampling distribution associated with $H_0 : \beta_1 = 0$.

```r
mod1 <- lm(bytxmstd ~ bytxrstd, data = els)
summary(mod1)
#>
#> Call:
#> lm(formula = bytxmstd ~ bytxrstd, data = els)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -26.703  -4.434  -0.071   4.144  39.084
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 7.592331   0.213320   35.59   <2e-16 ***
#> bytxrstd    0.850036   0.004182  203.26   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.715 on 16195 degrees of freedom
#> Multiple R-squared:  0.7184, Adjusted R-squared:  0.7184
#> F-statistic: 4.131e+04 on 1 and 16195 DF,  p-value: < 2.2e-16

#plot_t_distribution(beta_y = 'bytxmstd', beta_x = 'bytxrstd', data_df = els) #this plot_t_di

beta1 <- coef(summary(mod1))["bytxrstd", "Estimate"]       # beta1 coefficient
se_beta1 <- coef(summary(mod1))["bytxrstd", "Std. Error"]  # standard error of beta1
t_value <- beta1 / se_beta1                                # t-value

# Degrees of freedom for the t-distribution
df <- df.residual(mod1)

# Create a sequence of t-values
t_values <- seq(-4, 4, length.out = 1000)
```

```
# Calculate the density of the t-distribution
density <- dt(t_values, df)

# Create a data frame for plotting
t_dist_df <- data.frame(t_values = t_values, density = density)

# Plot the t-distribution
ggplot(t_dist_df, aes(x = t_values, y = density)) +
  geom_line(color = "blue", size = 1) +
  geom_vline(xintercept = t_value, color = "green", linetype = "dashed", size = 1) +
  geom_vline(xintercept = c(qt(0.975, df), qt(0.025, df)), color = "red", linetype = "dotted"
  labs(title = "t-Distribution", x = "t-value", y = "Density") +
  theme_minimal()
```



p-value

Above, we chose an alpha level of $\alpha = 0.05$; so if our observed p-value is less than .05, then we reject $H_0 : \beta_1 = 0$

- from above, our t-statistic of 203.2601 is associated with a p-value of 0
- Decision:
    - we reject $H_0 : \beta_1 = 0$

- we accept $H_a : \beta_1 \neq 0$

And because two-sided alternative hypotheses are at least as conservative as one-sided alternative hypotheses, we can also conclude that $\beta_1 > 0$

- that is, we can conclude there is a positive relationship between reading test score (X) and math test score (Y)

Our estimate $\hat{\beta}_1 = 0.85$ can be interpreted as follows:

- we estimate that a test score increase reading test score is associated with a 0.85 a test score increase in math test score for students.

Understanding $SE(\hat{\beta}_1)$

Anytime we talk about hypothesis testing, we are using estimates from one random sample to make statements about population parameters

- But our estimates differ from population parameters due to random sampling

Standard error (SE) tells us how far away (on average) an estimate is likely to be from population parameter

- The lower our SE, the closer we are to the population parameter!

When is SE($\hat{\beta}_1$) likely to be low?

- When standard error of the regression (SER) is also low (i.e., our predictions are good!)
- When sample size is big [estimates become more precise as sample size increases]
- When the variance of X is high

Factor Variables

This section briefly introduces a class of variables called "factor" variables; When running regression in R with a categorical X variable (e.g., marital status), the X variable must be factor variable

- For a more thorough introduction, see the lecture Attributes and Class from the course EDUC 260A: Introduction to programming and data management

Object class

Every object in R has a **class**

- Class is an **attribute** of an object
- Object class controls how functions work and defines the rules for how objects can be treated by object oriented programming language
  - E.g., which functions you can apply to object of a particular class
  - E.g., what the function does to one object class, what it does to another object class

Because **class** is an **attribute**, **class_** is additional "meta data" we put on top of the "just the data" part of an object

- the variable has additional attributes (metadata); "class" is one of these attributes
- The "class" of `df_mba$region` is `haven_labelled` (more on this later)
- You can use the `class()` function to identify object class.
- When I encounter a new object I often investigate object by applying `typeof()`, `class()`, and `attributes()` functions

Why is object class important?

- Functions care about object **class**, not object **type**
- Specific functions usually work with only particular **classes** of objects
- "Date" functions usually only work on objects with a date class
- "String" functions usually only work on objects with a character class
- Functions that do mathematical computation usually work on objects with a numeric class

`labelled` object class

Variable labels are labels attached to a specific variable (e.g., marital status) **Value labels** [in Stata] are labels attached to specific values of a variable, e.g.:

- Var value `1` attached to value label "married", `2`="single", `3`="divorced"

`labelled` is object class for importing vars with **value labels** from SAS/SPSS/Stata

- `labelled` object class created by `haven` package
- Characteristics of variables in R data frame with `class==labelled`:
  - Data `type` can be numeric(double) or character

`factor` object class

Factors are an object *class* used to display categorical data (e.g., marital status)

- A factor is an **augmented vector** built by attaching a **levels** attribute to an (atomic) integer vectors
- Usually, we would prefer a categorical variable (e.g., race, school type) to be a factor variable rather than a character variable
- when running regression in *R*, categorical variables must be factor class variables