PSTAT 126 - Regression Analysis – Fall 2017

Lab 2 Handout

**Note**: These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

Goals for this Lab

* Learn how to test a hypothesis for the regression slope using the t-test in the R summary
* Learn how to repeat the hypothesis test using the ANOVA method

Lab Exercise #1

We are continuing to look at the relationship between **bmi** and **diastolic** blood pressure in the **pima** dataset. All of the R commands, output, and answers to the questions are provided for this example.

1. Open and the **pima**  dataset from the **faraway**  package.

> data(pima,package="faraway")

1. Screen the data for zero values, remove them by creating a new dataset with the SUBSET function, and then confirm that subjects with zero values have been removed..

> summary(pima$bmi)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 27.30 32.00 31.99 36.60 67.10

> summary(pima$diastolic)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 62.00 72.00 69.11 80.00 122.00

> newpima<-subset(pima, bmi > 0 & diastolic>0)

> summary(newpima$bmi)

Min. 1st Qu. Median Mean 3rd Qu. Max.

18.20 27.50 32.40 32.47 36.60 67.10

> summary(newpima$diastolic)

Min. 1st Qu. Median Mean 3rd Qu. Max.

24.00 64.00 72.00 72.37 80.00 122.00

1. Perform a hypothesis test for the population regression coefficient using the t-test method.
   1. State the Null and Alternative Hypotheses

H0: 1 = 0; H1: 1≠0

* 1. Use the summary of the linear model to test these hypotheses

> fit1<-lm(diastolic~bmi,data=newpima)

> summary(fit1))

Call:

lm(formula = diastolic ~ bmi)

Residuals:

Min 1Q Median 3Q Max

-54.081 -7.628 -0.331 7.262 54.868

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 55.48694 2.11810 26.197 < 2e-16 \*\*\*

bmi 0.51989 0.06382 8.147 1.63e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.86 on 727 degrees of freedom

Multiple R-squared: 0.08365, Adjusted R-squared: 0.08239

F-statistic: 66.37 on 1 and 727 DF, p-value: 1.63e-15

* 1. What value of *t* did you obtain for the regression slope that predicts diastolic from BMI?

t=8.147

* 1. What is the p-value for the slope? Is it less than 0.05?

p = 0.00000000000000163, it is definitely less than 0.05

* 1. Draw a statistical conclusion regarding the null hypothesis above and state why

We reject the null hypothesis because p < 0.05.

1. We will now repeat the hypothesis test using the ANOVA method. The null and alternative hypotheses are the same.
   1. Generate the ANOVA output in R

> anova(fit1)

Analysis of Variance Table

Response: diastolic

Df Sum Sq Mean Sq F value Pr(>F)

bmi 1 9328 9327.6 66.369 1.63e-15 \*\*\*

Residuals 727 102174 140.5

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* 1. What is the value of SS Total?

SSTotal = SSRegression + SSResidual = 9328 + 102174 = 111502

* 1. What is the value of dfTotal?

dfTotal = dfRegression + dfResidual = 1 + 727 = 728

* 1. Complete the ANOVA source table using the R output.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **SS** | **df** | **MS** | **F** | **p** |
| Regression | 9328 | 1 | 9327.6 | 66.369 | 0.00000000000000163 |
| Error | 102174 | 727 | 140.5 | --- | --- |
| Total | 111502 | 728 | --- |  |  |

* 1. Draw a statistical conclusion and state why

We reject the null hypothesis because p < 0.05

* 1. Are we getting the same answer from the t-test and the ANOVA?

Yes.

R commands for Exercise #1

data(pima,package='faraway')

summary(pima$bmi)

summary(pima$diastolic)

newpima<-subset(pima, bmi > 0 & diastolic>0)

summary(newpima$bmi)

summary(newpima$diastolic)

fit1<-lm(diastolic~bmi,data=newpima)

summary(fit1)

anova(fit1)

Lab Exercise #2

Repeat all parts of Exercise 1 for the relationship between education (**educ**) and weekly wages (**wage**) in the **uswages** dataset from the **faraway** package. Note that zero values are acceptable for both wage and education, so there is no need to subset the data.

1. Open and the **uswages**  dataset from the **faraway**  package.
2. Perform a hypothesis test for the population regression coefficient using the t-test method.
   1. State the Null and Alternative Hypotheses
   2. Use the summary of the linear model to test these hypotheses
   3. What value of *t* did you obtain for the regression slope that predicts **wage** from **educ**?
   4. What is the p-value for the slope? Is it less than 0.05?
   5. Draw a statistical conclusion regarding the null hypothesis above and state why
3. Repeat the hypothesis test using the ANOVA method. The null and alternative hypotheses are the same.
   1. Generate the ANOVA output in R
   2. What is the value of SS Total?
   3. What is the value of dfTotal?
   4. Complete the ANOVA source table using the R output.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **SS** | **df** | **MS** | **F** | **p** |
| Regression |  |  |  |  |  |
| Error |  |  |  |  |  |
| Total |  |  |  |  |  |

* 1. Draw a statistical conclusion and state why
  2. Are we getting the same answer from the t-test and the ANOVA?

1. For extra practice, create a scatterplot for **educa** and **wage**, with appropriate labels.

R commands for Exercise #2

data(uswages,package="faraway")

fit2<-lm(wage~educ,data=uswages)

summary(fit2)

anova(fit2)