

R Assignment Module 13: Inference for Regression

Instructions

This is your final R assignment! As usual, use an R script file for your work. Paste your code with R output and write in your analysis in a word file. Save to a pdf and post on Canvas by midnight on Sunday, July 19.

Load the data “faithful” from R using the `data()` command. There are two variables, waiting and eruptions. Waiting is the waiting time to an eruption and eruptions is the duration of the eruption that followed the waiting time.

If you use the command `attach(faithful)` you don’t need to type `-faithful$variable-` name when you work through this, but you can only type `-variable-` instead.

1a. Use the `plot` function to create a scatterplot with waiting as the x or explanatory variable and eruptions as the y or dependent variable. The entries for your plot are: `plot(x variable name, y variable name, main=“Plot Title”, xlab=“x label”, ylab=“y label”, pch=20)`. The last entry is the plotting character. “pch=20” will give you small dots. If you search on Google you can find lists of the numbers that coincide with other characters like triangles, etc. Describe the relationship between waiting and eruptions.

1b. Run the regression using the `lm()` command. It will be easier if you define an object that you can call up later. I usually use `fit` as the object: `fit<-lm(eruptions~waiting)`. The `lm` command reads regress eruptions on waiting, eruptions~waiting. Type in `summary(fit)` to see the regression results.

1c. Now let’s add the regression line to the plot. To do this, use the same plot function you wrote for 1a but add below it add the following lines:

```
abline(lm(eruptions~waiting), col=“purple”)
```

```
abline(h=mean(eruptions), col=“green”)
```

The first is the regression line and the second is a line that is the sample mean of eruptions. If you didn’t use the regression, the mean would be our primary summary measure. Note how much we can reduce the variation in the data by using the regression model.

1d. Now interpret the regression in evaluating the hypothesis test below. In your write up, I want you to discuss significance based on the output you saw in `summary(fit)` along with goodness of fit summarized in the reported R-squared. What is the p-value? What does the p-value mean?

$$H_0 : \beta_{\text{waiting}} = 0$$

$$H_a : \beta_{\text{waiting}} \neq 0$$

1e. Use the information from the regression output to construct a 95% confidence interval for the population slope coefficient, β_{waiting} . To find t^* you can look up the value in the t-table using 270 degrees of freedom or use the `qt` function in R to find the t^* such that $P(T > t^*) = 0.025$.

1f. Generate a 95% confidence interval for the conditional mean $E(Y|x = 76) = E(\text{eruptions}|\text{waiting} = 76) = \mu_y(76)$. What this interval is trying to capture is the point in the y-space on the *population* regression line. The confidence interval uses the estimate from the regression evaluated at the median of waiting, 76, and subtracts the margin of error for the lower bound and adds the margin of error for the upper bound. The confidence interval procedure will capture the true $\mu_y(76)$ 95% of the time with repeated samples of size $n=272$. With R this is pretty simple. In fact here are the commands:

```
newdata <- data.frame(waiting=76)
```

```
ci.int<-predict(fit, newdata, interval="confidence")
```

```
ci.int
```

2a-f. Repeat steps a-f above but use the R data set mtcars to explore how mpg (miles per gallon) relates to wt (car weight). For part (f) use the median of wt as the value for conditional mean 95% confidence interval. You may also use 270 degrees of freedom.