Lab 9 - Regression

Statistics 411: March 7, 2018

1. Consider exercise 29 on page 205 in `ex0729` in the Sleuth3 data set.

   (a) Plot the data using a scatterplot.

   (b) Fit a SLR.

   (c) Report the regression equation using proper notation (e.g., $\hat{\mu}\{Y|X\} = ...$).

   (d) Use the `abline()` function to add the fit line to the scatterplot.

   (e) Diagnose the fit of the line using `diagANOVA()` from the course website, `source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")`

   (f) Test whether there is a linear relationship (i.e., correlation) between Distance to the polling station and the Overestimate at a significance level of $\alpha = 1\%$ by building a 99% CI for β_1. Report on all 6 steps of the hypothesis test including a decision and conclusion in terms of the problem.

2. Consider the case study of pH versus time in section 7.1.2 in `case0702` Sleuth3 data set.

   (a) Plot the data using a scatterplot.

   (b) Fit a SLR.

   (c) Use the `abline()` function to add the fit line to the scatterplot.

   (d) Diagnose the fit of the line using `diagANOVA()` from the course website, `source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")`

   (e) You will see that there is a problem with model fit. Which assumption is violated? What approach does the book use to mitigate the violated assumption?

   (f) Use the book's approach to address the violated assumption then re-perform steps (a)-(d).

   (g) Test whether there is an association between pH and Time at a significance level of $\alpha = 1\%$. Report on all 6 steps of the hypothesis test including a decision and conclusion in terms of the problem.

3. The number of insurance claims per month per 100 employees was monitored by an employer over several years as a function of average monthly temperature (in degrees F).

   (a) Get the data from the course web page and save into an object named `claims`, by executing executing
   `claims =`
   `read.csv("http://www.math.montana.edu/parker/courses/STAT411/Lab9_Claims.CSV")`

   (b) Log-transform (using a natural log) the claims rates, then fit a SLR model to the log-transformed rates as a function of average monthly temperature.

   (c) Plot the log-transformed rates vs temperature using a scatterplot.

   (d) Use the `abline()` function to add the fit line from (b) to the scatterplot.

   (e) Diagnose the fit of the line.

   (f) The regression equation is

   $$\hat{\mu}\{\ln[\text{claims}]|T\} = \hat{\beta}_0 + \hat{\beta}_1 T$$

   where T is average monthly temperature. Anti-log both sides of this equation to get the best fit exponential to these data:

   $$\widehat{median}\{\text{claims}|T\} = \hat{C}e^{\hat{r}\times T}.$$

   In other words, what are $\hat{C}$ and $\hat{r}$ in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$?

   (g) Give a 95% CI for the true value of C (C is interpreted as the claims rate when $T = 0$ degrees Fahrenheit).