Arsen Akopian
Sagar Parikh
Section 04645

1/22/2014 Class Notes

In this lecture, we went over basic administrative issues, homework 1, data summarizing methods, and regression methods and uses. We used the pickup.csv and pickup.r in analyzing pickup truck prices off of craigslist. This was our introduction to quantitative data analysis.

1. Introduction
   i. Basic Administrative Comments
      a. Students from different sections are allowed to work in groups for homework assignments
      b. Math calculations can be done by hand or on a computer
      c. Calculus will be used several times throughout the semester
      d. Students don't need to memorize r codes or scripts
      e. Scribers are allowed to use laptops
      f. R scripts don't have to be turned in with the homework unless specified
   ii. Helpful Websites
      a. Rseek.org
      b. Stackoverflow.com

2. Homework 1
   i. Biggest Pitfalls:
      a. Verify data frame correctness in import frame
      b. Use the factor command to distinguish categorical variables
      c. Calculus Optimization: Take derivate of the function to find the minimum

3. Data Summarizing
   i. Use of groupmean vs. lm model
      a. The lm function returns baseline offset values while groupmean returns direct group means.

         Direct group means = $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_n$ for n groups
         Baseline offset form = $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_n$ for n groups

         $\hat{\theta}_2 = \hat{\beta}_1 + \hat{\beta}_2$
         $\hat{\theta}_3 = \hat{\beta}_1 + \hat{\beta}_3$

         The lm function returns $\hat{\beta}$ values, and the groupmean function returns $\hat{\theta}$ values

   ii. Standard Deviation for N data points

Arsen Akopian
Sagar Parikh
Section 04645

a. $s_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}$

b. Use N-1 for convention; some call it an unbiased estimator, but this is an unconvincing argument.

c. Variance is listed in units squared; square root is taken to get to units that make sense.

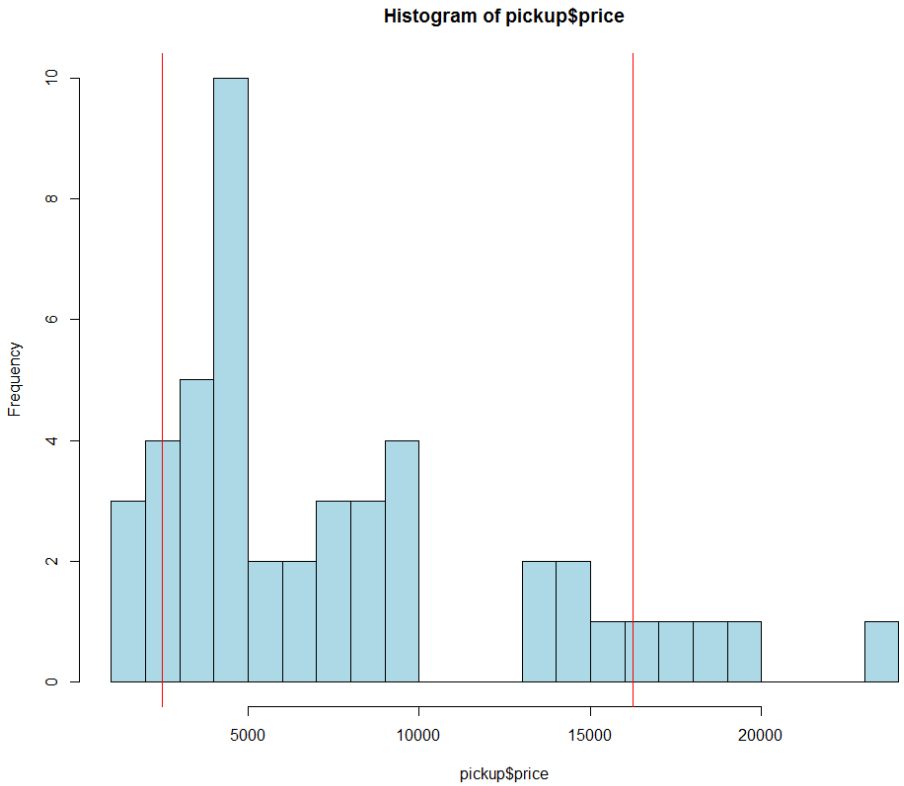
iii. Coverage Intervals

a. Coverage interval is the interval that covers a certain percentage of the data.

b. A coverage interval of 80% is an interval that covers 80% of the points by excluding the top and bottom 10%

c. Consider this the data set analog to the trimmed mean

d. R code: qdata(c(0.1,0.9), price, data=pickup) – This function gives the middle 80% of prices for the data in pickup.csv



4. Regression Methods

i. Least Squares Lines

a. Least squares lines try to best fit the data so that data analysis and prediction can be done

b. General Form: $\hat{y} = B_0 + B_1 x$; B_0 is the intercept while $B_1$is the slope for x.

c. $y = B_0 + B_1 x_1 + e_i \rightarrow e_i$ is not necessarily equal to 0

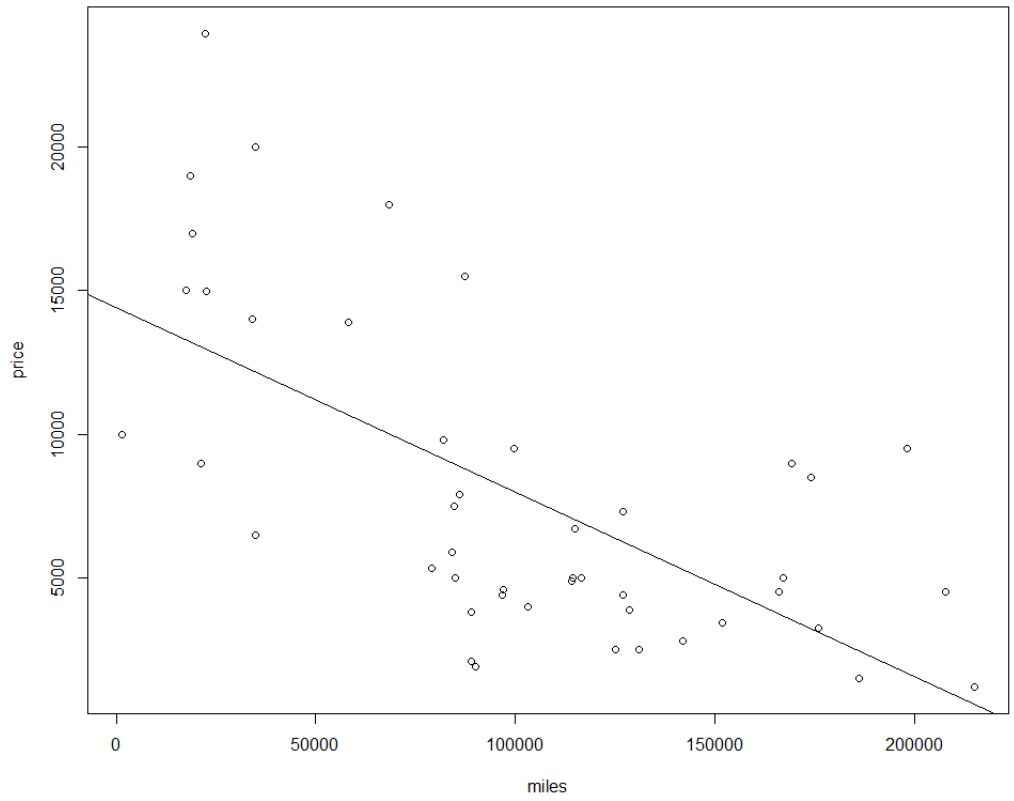
d. The idea for linear regression is to make $\sum {e_i}^2$ as small as possible

Arsen Akopian
Sagar Parikh
Section 04645

e. The command abline(lm(price~miles, data=pickup)) will show the least squares regression line.

f. To get the formula for the line, use coef(lm(price~miles, data=pickup)) to get where miles is the slope

```
coef(model1)
 (Intercept)        miles
1.441938e+04 -6.429944e-02
```



ii. Regression Models

a. Plug-in Prediction: Regression models can be used to forecast a Y value using a given X. Simply substitute the X value into the $\hat{y} = B_0 + B_1 x$ equation.

b. Summarizing trends using slope: By looking at the slope, or $\frac{dy}{dx}$, and multiplying it by Δx, you can interpret the data using relevant units. For example, while the price of the pickup trucks will not change for every additional mile driven, an interval of 10,000 would allow for more meaningful analysis.

c. Taking the 'x-ness' out of y (Statistical Adjustment): By graphing e_i , you can see the residuals for the data (e_i). This accounts for the trend in the data and shows how y relates to $\hat{y}$.

Arsen Akopian
Sagar Parikh
Section 04645

d. Reduce Uncertainty: Looking at the spread of the residuals will reduce uncertainty as the standard deviation of the residuals is typically smaller than the standard deviation of the data set.