Air Quality

Brandon Bevan

May 7, 2018

Based from the Coursera course "Reproducible Research" by Johns Hopkins University

The goal of this document is to provide an example of "literate statistical programming" by "weaving" together English text, R Code, and graphics provided by ggplot and R's builtin plotting capabilities.

"Literate statistical programming" with R Markdown files allows for "reproducible" research through the ability of the critic to

1. Download the markdown file
2. Re-run the analyses in R
3. Regenerate the HTML (or pdf)

In this document, we provide a regression analysis of air quality data.

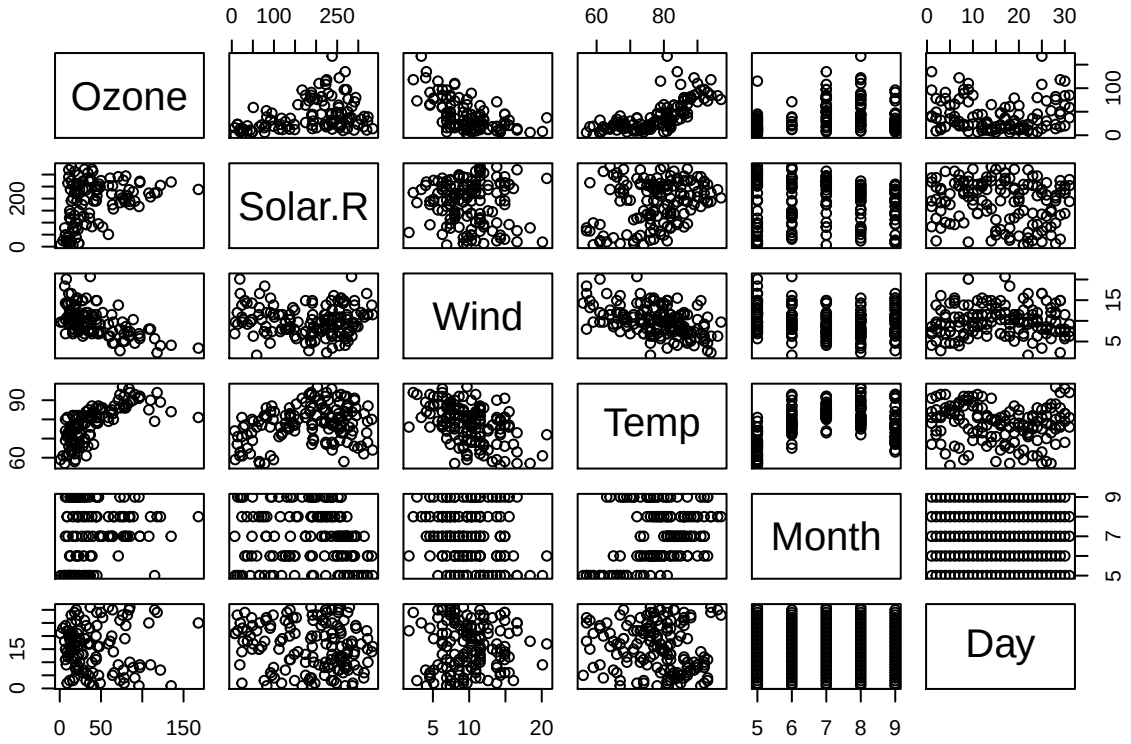
```
library(datasets)
data(airquality)
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

As can be seen, the variables within the data set are Ozone levels, Solar Radiation levels, Wind, Temperature, Month, and Day measurements.

Here is a plot of each pair of variables against one another.

```
pairs(airquality)
```



We will test a regression model of Ozone versus Solar Radiation.

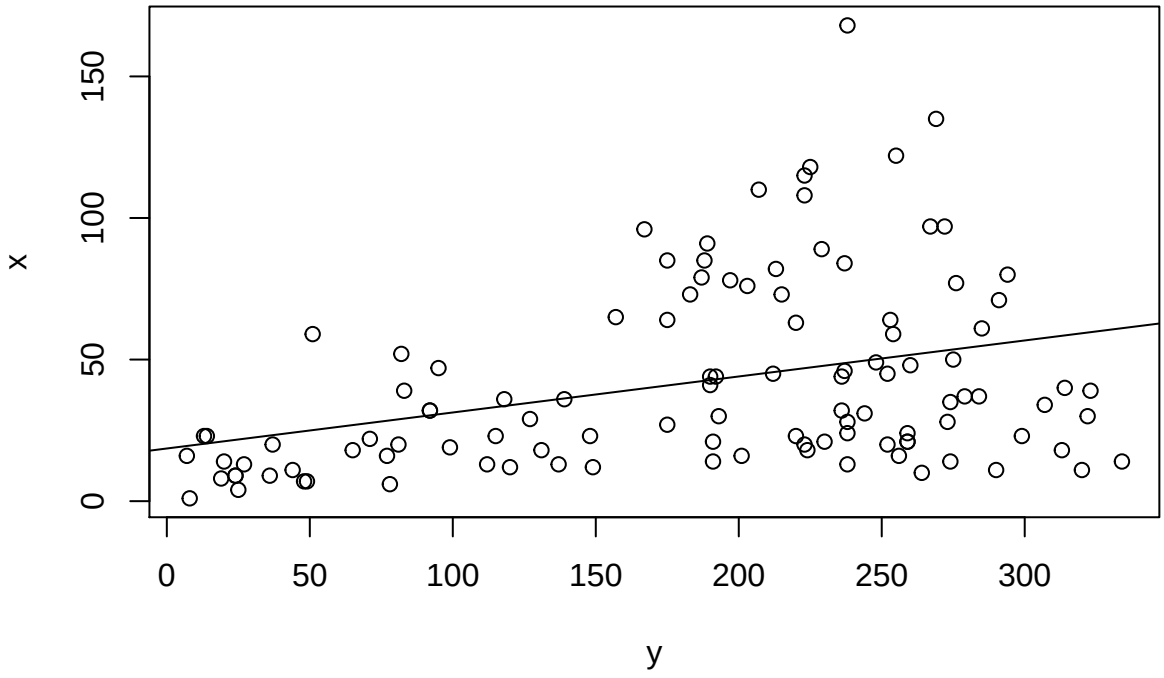
```
library(stats)
fit <- lm(Ozone ~ Solar.R, airquality)
summary(fit)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.292 -21.361  -8.864  16.373 119.136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.59873    6.74790   2.756 0.006856 **
## Solar.R      0.12717    0.03278   3.880 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.33 on 109 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.1213, Adjusted R-squared:  0.1133
## F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

Next, we plot the regression line.

```r
y <- airquality$Solar.R
x <- airquality$Ozone

plot(y,x)
abline(fit)
```



In conclusion, I have realized that the gap in my understanding is in interpreting measures of significance for regression lines. I am also unsure how to interpret the Mean Squared Error, whether it should be close to zero or not. Therefore, the next step in my education is learning Hypothesis Testing for regression analysis.