

# Lab 1: data.gov

*James Barbour*

*January 16, 2017*

## Aviation Accidents and Fatalities, 1975-2014

This data set is provided by the NTSB, found [here](#). Initially, it was poorly formatted and unusable, but since it is a small dataset, I was able to quickly manually clean up the data.

The data set contains the number of aviation accidents (All) from 1975 to 2014. For each year, the number of fatal accidents (Fatal), the total number of fatalities (Total), number of fatalities aboard flights (Aboard), and flight hours logged that year (Flight.Hours).

I decided to run a multiple linear regression on the dataset to predict the number of accidents from the year and flight hours. While flight hours alone isn't a reliable predictor of accidents, the year and flight hours together predict the number of total accidents fairly reliably with an adjusted  $R^2$  value of 0.89. This is likely due to an increase in safety standards and technology combined with the increase in the average size of commercial airplanes over the years.

```
datagov <- read.csv("table10_2014.csv")
summary(datagov)
```

```
##      Year      All      Fatal      Total
## Min.   :1975   Min.   :1221   Min.   :222.0   Min.   : 391.0
## 1st Qu.:1984   1st Qu.:1694   1st Qu.:323.0   1st Qu.: 572.0
## Median :1994   Median :2056   Median :404.0   Median : 734.0
## Mean   :1994   Mean   :2336   Mean   :428.4   Mean   : 799.9
## 3rd Qu.:2004   3rd Qu.:2878   3rd Qu.:521.5   3rd Qu.:1004.5
## Max.   :2014   Max.   :4216   Max.   :719.0   Max.   :1556.0
##      Aboard      Flight.Hours
## Min.   : 386.0   Min.   :18103000
## 1st Qu.: 558.5   1st Qu.:23891000
## Median : 727.0   Median :25998000
## Mean   : 781.5   Mean   :26752641
## 3rd Qu.: 983.0   3rd Qu.:28736000
## Max.   :1398.0   Max.   :38641000
```

```
head(datagov)
```

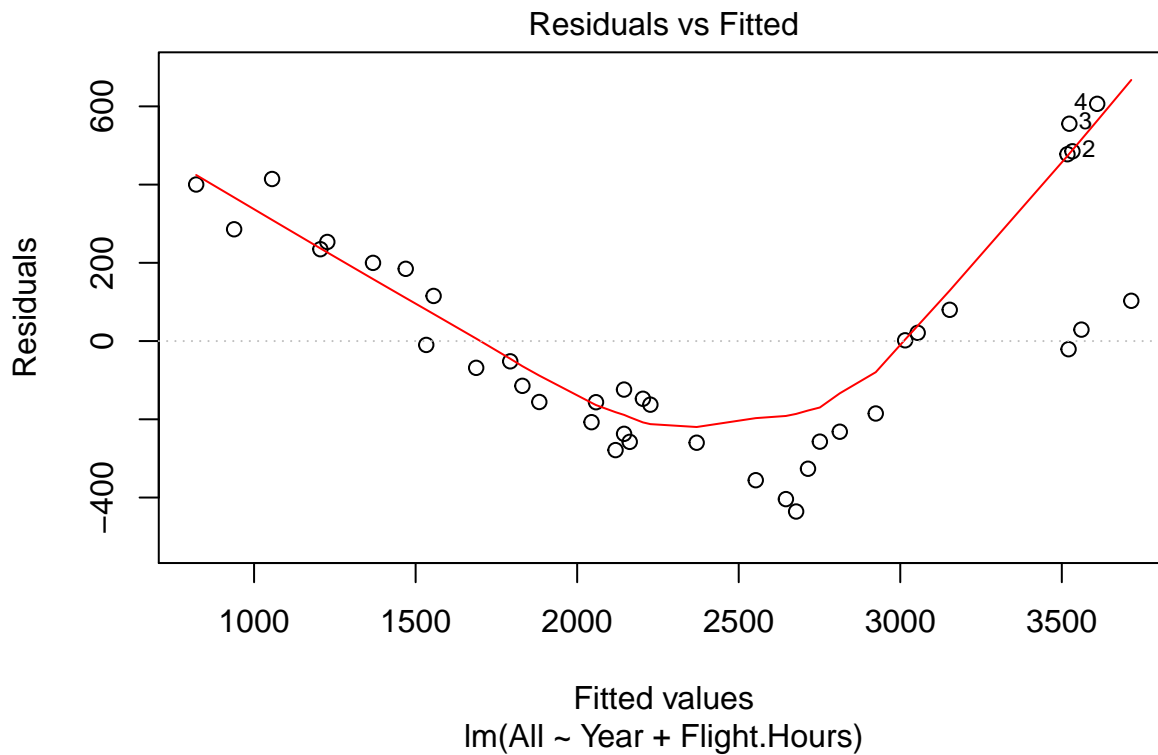
```
##   Year All Fatal Total Aboard Flight.Hours
## 1 1975 3995   633  1252   1231    28799000
## 2 1976 4018   658  1216   1203    30476000
## 3 1977 4079   661  1276   1265    31578000
## 4 1978 4216   719  1556   1398    34887000
## 5 1979 3818   631  1221   1203    38641000
## 6 1980 3590   618  1239   1230    36402000
```

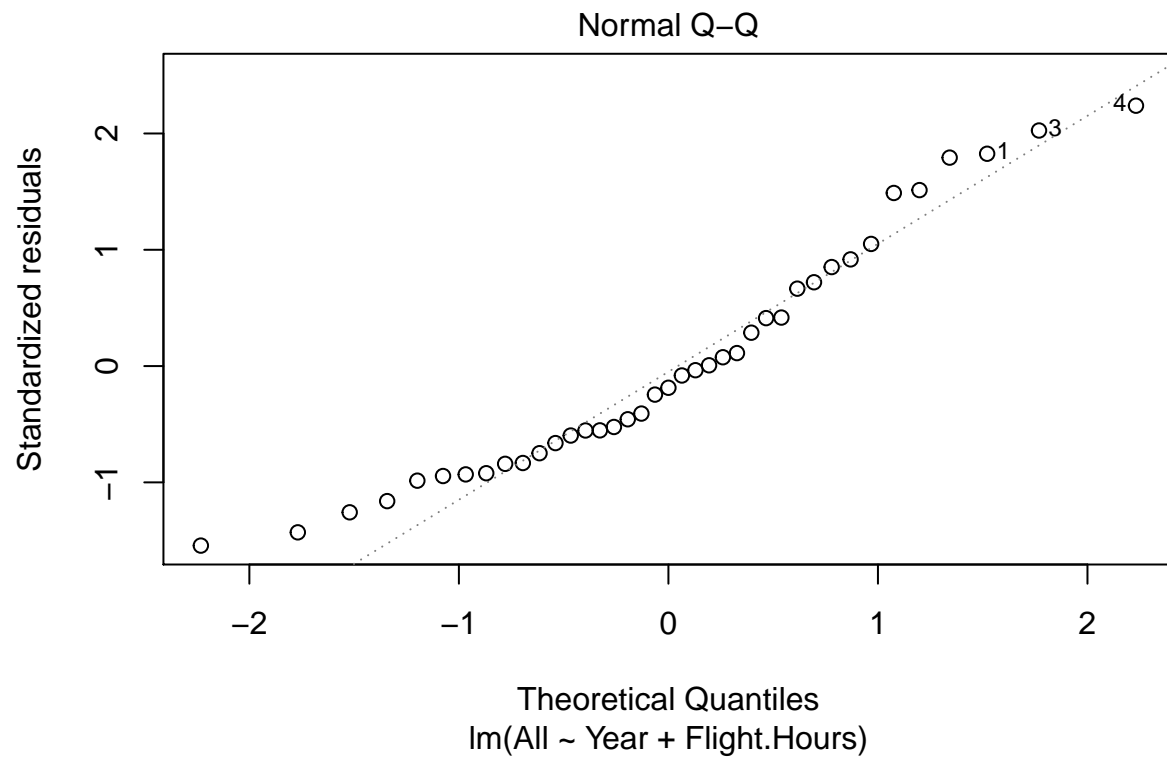
```
hours.lm = lm(All ~ Year + Flight.Hours, data = datagov)
summary(hours.lm)
```

```
##
## Call:
## lm(formula = All ~ Year + Flight.Hours, data = datagov)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -435.72 -219.56  -51.83  192.26  606.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.153e+05  1.473e+04   7.833 2.74e-09 ***
## Year        -5.726e+01   7.184e+00  -7.970 1.83e-09 ***
## Flight.Hours  4.335e-05   1.805e-05   2.402  0.0216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.9 on 36 degrees of freedom
## Multiple R-squared:  0.8988, Adjusted R-squared:  0.8932
## F-statistic: 159.9 on 2 and 36 DF,  p-value: < 2.2e-16
```

```
plot(hours.lm, which = 1:2)
```





### Total accidents per year

Points sized by number of fatalities, alpha is flight hours per year

```
ggplot(data = datagov) + geom_point(mapping = aes(x = Year, y = All, size = Total, alpha = Flight.Hours))
```

