# STATS 700 Homework 1

*Cameron Hollingshead*

*September 19, 2017*

## 1 A First Impression of the Data

Welcome to my solution to Homework 1! For this small series of code, we shall use a dataset on airline safety, from the Aviation Safety Network. For 56 different airlines, we have information on the following variables: available seat kilometers per week, incidents (1985-99 and 2000-14), fatal accidents (1985-99 and 2000-14), and fatalities (1985-99 and 2000-14).

We will examine some basic attributes of this data set, and draw some inferences where appropriate. First, the dimensionality of the data:

```
dim(air)
```

```
## [1] 56  7
```

56 observations, 7 variables. That matches how we described the data above. Phew.

Next, let us produce some generic summary statistics of the different variables.

```
summary(air)
```

```
##  avail_seat_km_per_week incidents_85_99  fatal_accidents_85_99
##  Min.   :2.594e+08      Min.   : 0.000   Min.   : 0.000
##  1st Qu.:4.740e+08      1st Qu.: 2.000   1st Qu.: 0.000
##  Median :8.029e+08      Median : 4.000   Median : 1.000
##  Mean   :1.385e+09      Mean   : 7.179   Mean   : 2.179
##  3rd Qu.:1.847e+09      3rd Qu.: 8.000   3rd Qu.: 3.000
##  Max.   :7.139e+09      Max.   :76.000   Max.   :14.000
##  fatalities_85_99 incidents_00_14  fatal_accidents_00_14 fatalities_00_14
##  Min.   :  0.0    Min.   : 0.000   Min.   :0.0000        Min.   :  0.00
##  1st Qu.:  0.0    1st Qu.: 1.000   1st Qu.:0.0000        1st Qu.:  0.00
##  Median : 48.5    Median : 3.000   Median :0.0000        Median :  0.00
##  Mean   :112.4    Mean   : 4.125   Mean   :0.6607        Mean   : 55.52
##  3rd Qu.:184.2    3rd Qu.: 5.250   3rd Qu.:1.0000        3rd Qu.: 83.25
##  Max.   :535.0    Max.   :24.000   Max.   :3.0000        Max.   :537.00
```

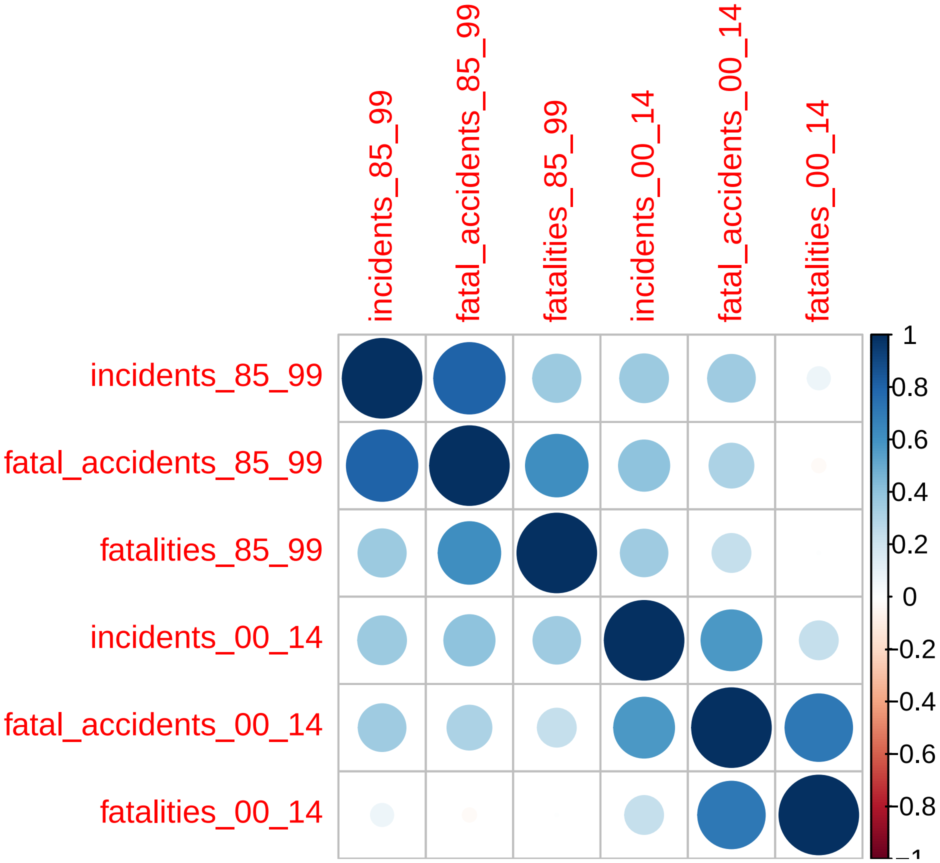
As we might hope, the minimum and first quartile values for incidents, fatal accidents, and fatalities are either 0 or close to 0. As we move closer to the upper range for those variables, we can see greater disparity between the third quartile and maximum values. We may consider applying a Poisson model on predicting incidents, fatal accidents, and fatalities in future data.

Another interesting attribute we can see is that, while the maximum number of fatal accidents for an airline dropped from 14 to 3, the maximum number of fatalities for an airline slightly increased from 535 to 537. My untested intuition tells me that number of passengers per plane have increased, between these 15-year periods.

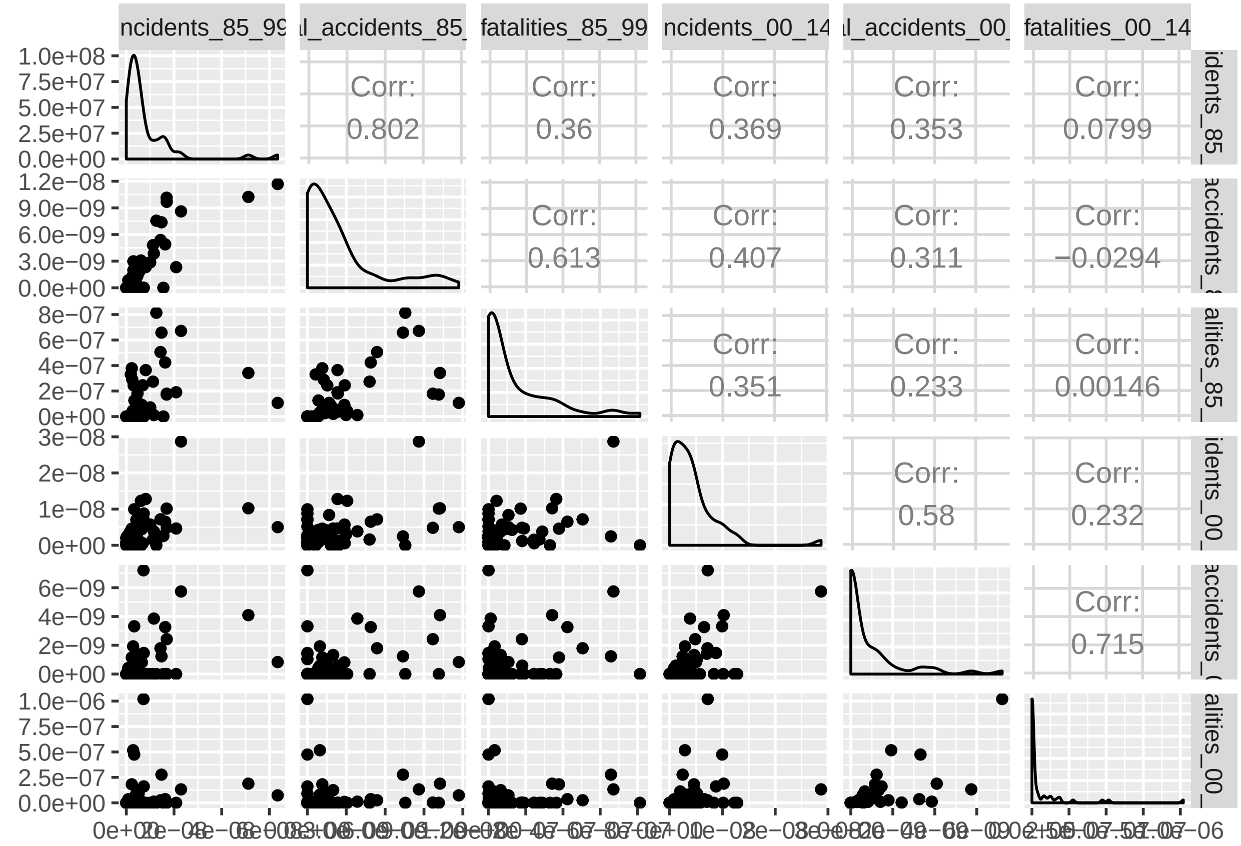
## 2 A More Detailed Picture

Visualizations of the data may give greater insight as to how we may set up our probability model. Before we show such visualizations, we shall reshape the data so that incidents, fatal accidents, and fatalities of each time period are weighted by available seat km per week. Here are scatter plots and correlation plots for the six different variables:

```
air.reshape = air[,2:7] / air[,1]
corrplot(cor(air.reshape))
```



```
ggpairs(air.reshape)
```

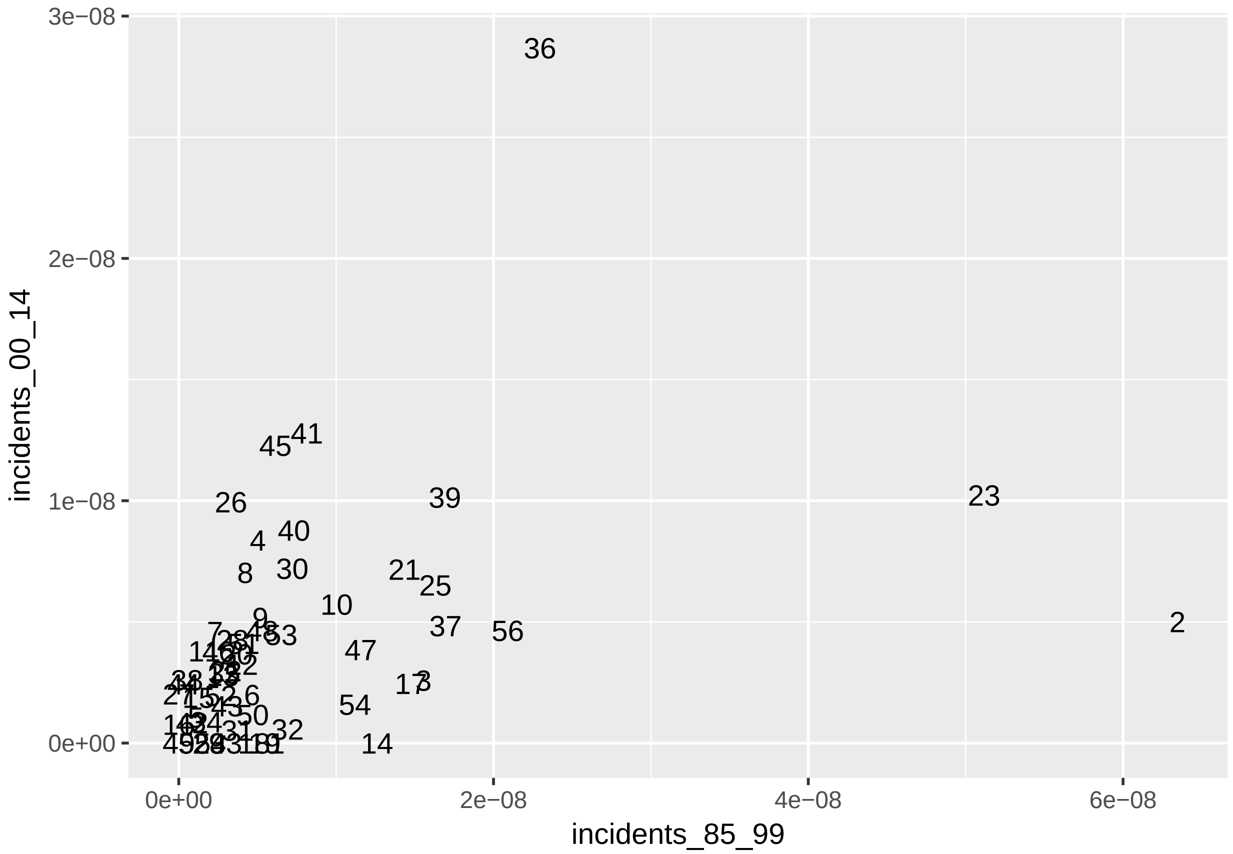


One attribute of these plots and corresponding correlation coefficients is that number of fatalities in a particular time period only correlates strongly with number of fatal accidents in the same time period. I would be shocked, if they were not correlated. More importantly, it seems clearer that number of fatalities will not be as useful of a variable, perhaps unless we can connect it with an airline's available seat km per week.

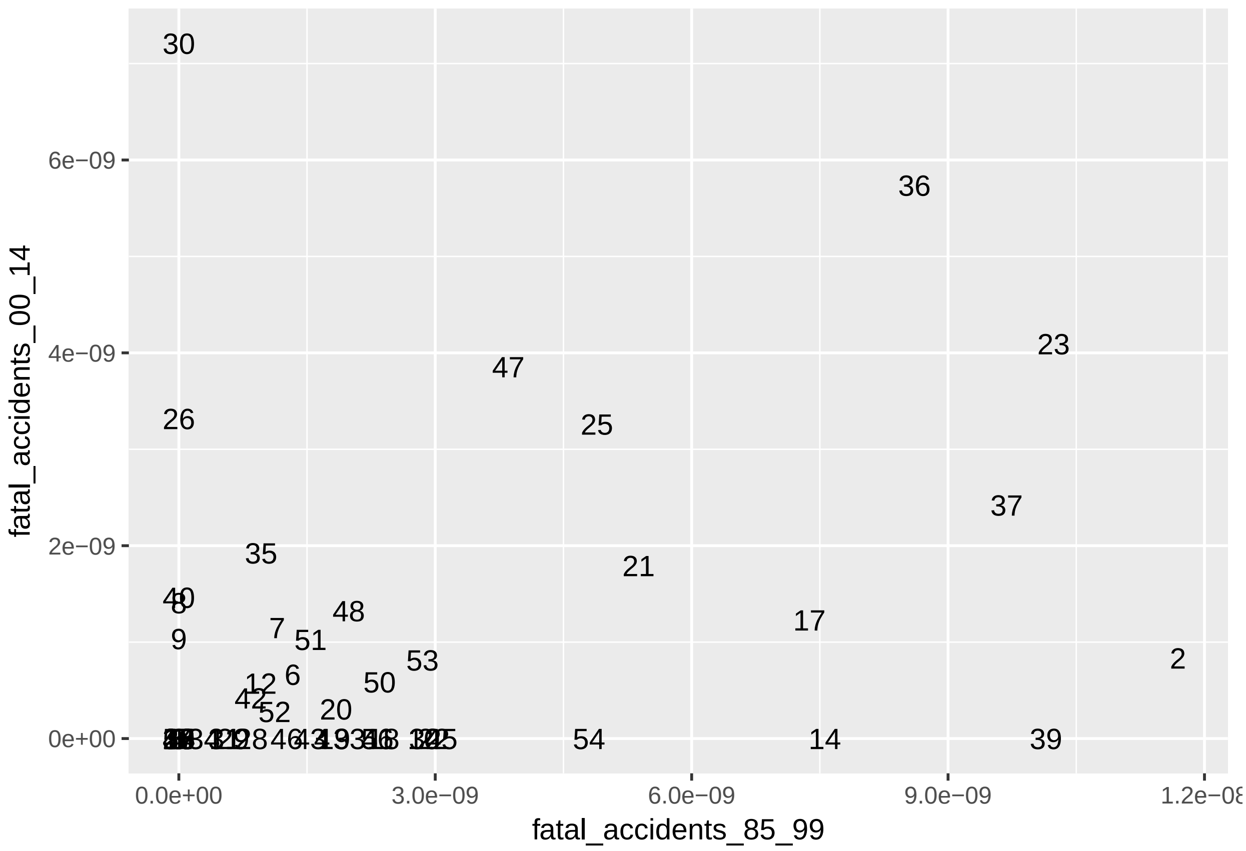
Ultimately, we may want to use this data, in order to see if number of incidents / fatal accidents for a particular airline can be used to further predict those same numbers in a future period. However, as you can observe on the plots between incidents and fatal accidents of different time periods, correlation is not that strong.

On the new scale, let us see if there are any airlines that are relatively high in rate of incidents and fatal accidents:

```
ggplot(data = air.reshape, aes(incidents_85_99, incidents_00_14)) + geom_text(aes(label=row(air.reshape
```



```
ggplot(data = air.reshape, aes(fatal_accidents_85_99, fatal_accidents_00_14)) + geom_text(aes(label=row
```



Corresponding the index numbers to specific airlines, Pakistan International, Ethiopian Airlines, and Aeroflot, a Russian airline, seem to be especially high for all four parameters. What some may infer from that is that these airlines are the most consistently dangerous, among the 56 airlines in the dataset. Let us avoid drawing such conclusions, so soon after looking at the data.

# 3 A Statistical Model Might Be Nice

Sometimes, simple linear regression can suggest where one should focus, with respect to some dataset and for what variables we wish to predict. Since fatal accidents and fatalities are highly correlated, we will only use incidents and fatal accidents of the earlier time period to predict incidents and fatal accidents on the later period. We will also put all variables on a log scale, to mitigate the variance size and better observe relationships.

```
model = glm(log(incidents_00_14 + 1) + log(fatal_accidents_00_14 + 1) ~ log(incidents_85_99 + 1) + log(
summary(model)
```

```
##
## Call:
## glm(formula = log(incidents_00_14 + 1) + log(fatal_accidents_00_14 +
##     1) ~ log(incidents_85_99 + 1) + log(fatal_accidents_85_99 +
##     1), family = poisson, data = air.reshape)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.262e-04  -5.474e-05  -1.691e-05   1.696e-05   1.986e-04
```

```
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.948e+01  2.635e+03  -0.007    0.994
## log(incidents_85_99 + 1)      5.474e+06  2.027e+11   0.000    1.000
## log(fatal_accidents_85_99 + 1)  1.011e+08  8.644e+11   0.000    1.000
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.9152e-07  on 55  degrees of freedom
## Residual deviance: 2.3927e-07  on 53  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 20
```

Uh, oh. Judging from those type I errors, I get the sense that incidents and fatal accidents from an earlier time are not reliable predictors of incidents and fatal accidents for the future. That is, if an airline had a reputation (in past 15 years) for having relatively high number of incidents and fatal accidents, that does not even usually mean they will have a similar reputation in the next 15 years.

What does that say about Pakistan International, Ethiopian Airlines, and Aeroflot? To be continued… possibly.

# 4 How I Feel About the Data

It took me a while to find a data set that interested me, in regards to predicting uncertain information. I want to work with something like this, but most likely, not this particular data. This data seems to pertain strictly to the question, "Does an airline's past record on accidents imply a similar future record?" That seems too trivial for my taste. On top of that, the fatality variable seems altogether useless, even misleading as to how "dangerous" an airline is.

With that said, the general question, "Which airlines will encounter the most accidents?," would interest me, and I would like to find a data set that can better answer that question. The data used in this R script seems to be an anti-thesis to that. We shall see if I can find a more suitable dataset.

If I were to use this data anyway, other problems I may encounter is there only being two time periods for prediction, with not enough information to break the longer periods into smaller ones. Also, I would figure that 'available seat km per week' is liable to great change from one period to another. Averaging over the entire 30 years neglects, say, airlines that have grown or declined, with respect to the rest of the market.

As you can tell, I don't like this data. Rather than making due with what I am given, I will almost certainly burn my bridges and start anew elsewhere.