Should travelers avoid flying airlines that have had crashes in the past?

An analysis by Mark Jin for

Almond FinTech

1. Hypothesis Testing

We formulate the hypotheses as such:

- **H0**

  There's no positive relationship in airline crash rates between 1985-1999 and 2000-2014.

- **H1**

  There is a positive relationship in airline crash rates between 1985-1999 and 2000-2014

Data Treatment

Add column for regional subsidiary

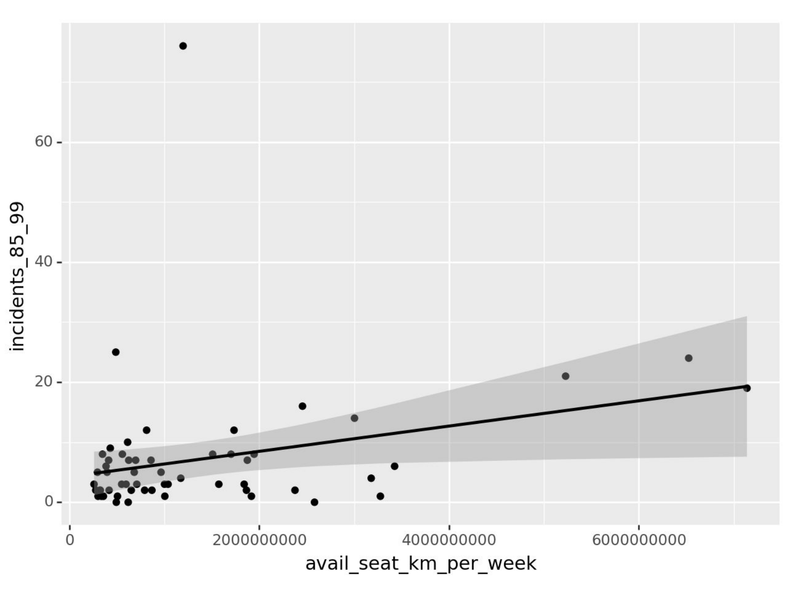
Remove dirty variables

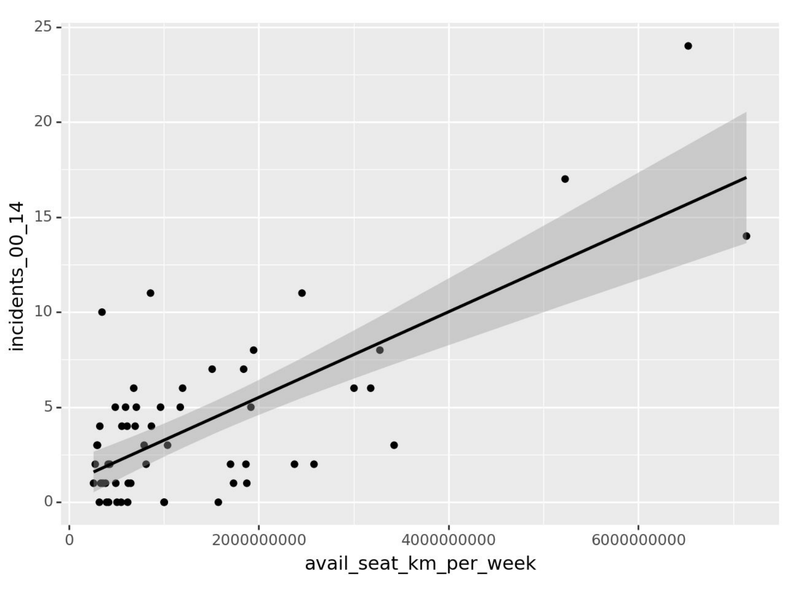
- Fatal accidents and fatalities values are sparse and contains many outliers. Additionally, customers don’t not want to fly with airlines with a high probability of crashing, whether the crashes involved fatalities is secondary

Calculate incident rate

- Calculate incidents per 1 billion available seat kilometer (ASK)

Is ASK Appropriate for Calculating Incident Rate?





On lower ASKs, incidents per ASK is lower in ‘00–’14 flights, evidence of safer flights.

On higher ASKs, incidents per ASK are about equal in both categories. But later versions of planes also likely have more seats than earlier versions of planes, thus increasing incidents per ASK despite having fewer incidents per flight.

How to Best Calculate Incident Rate?

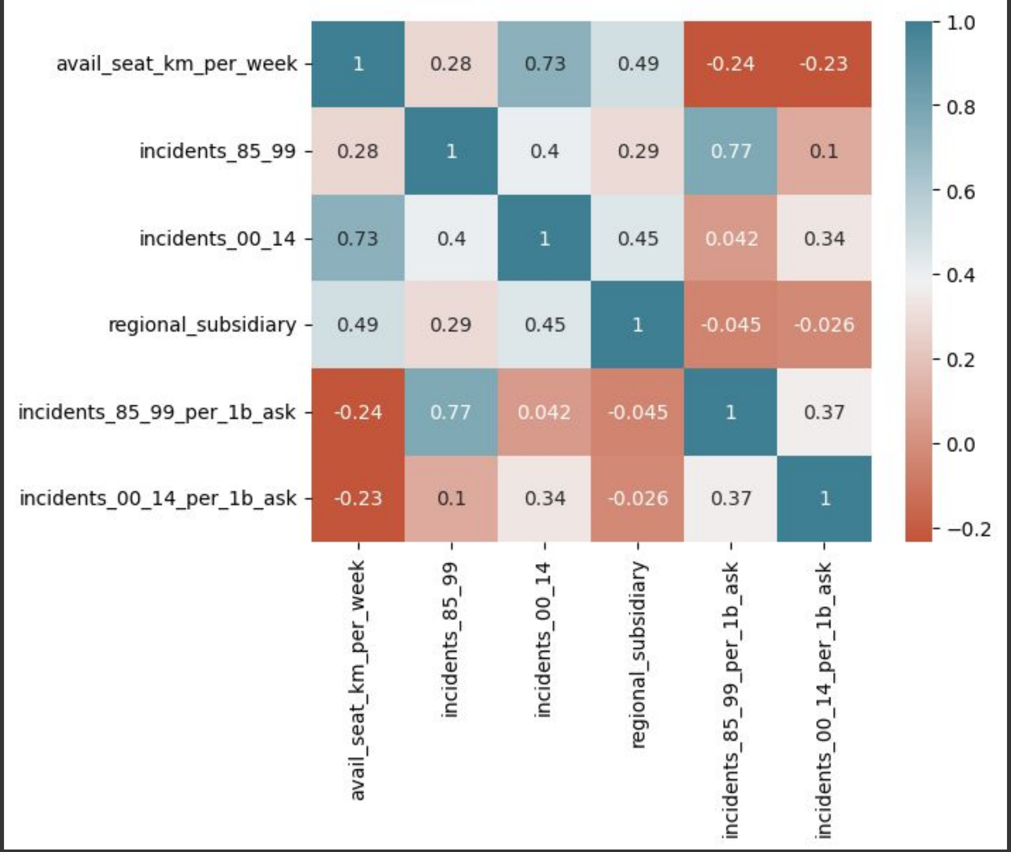
➔ **Ideal**

Calculate number of incidents per flight (takeoff, cruise, landing == 1 flight). Crashes rarely happen during the cruise phase. Assuming the number of seats on a plane doesn't significantly affect a plane's crash rate.

➔ **Reality**

ASK is the only denominator we can use in this dataset to obtain incident rates, so we must make do with what we have

Correlation Matrix



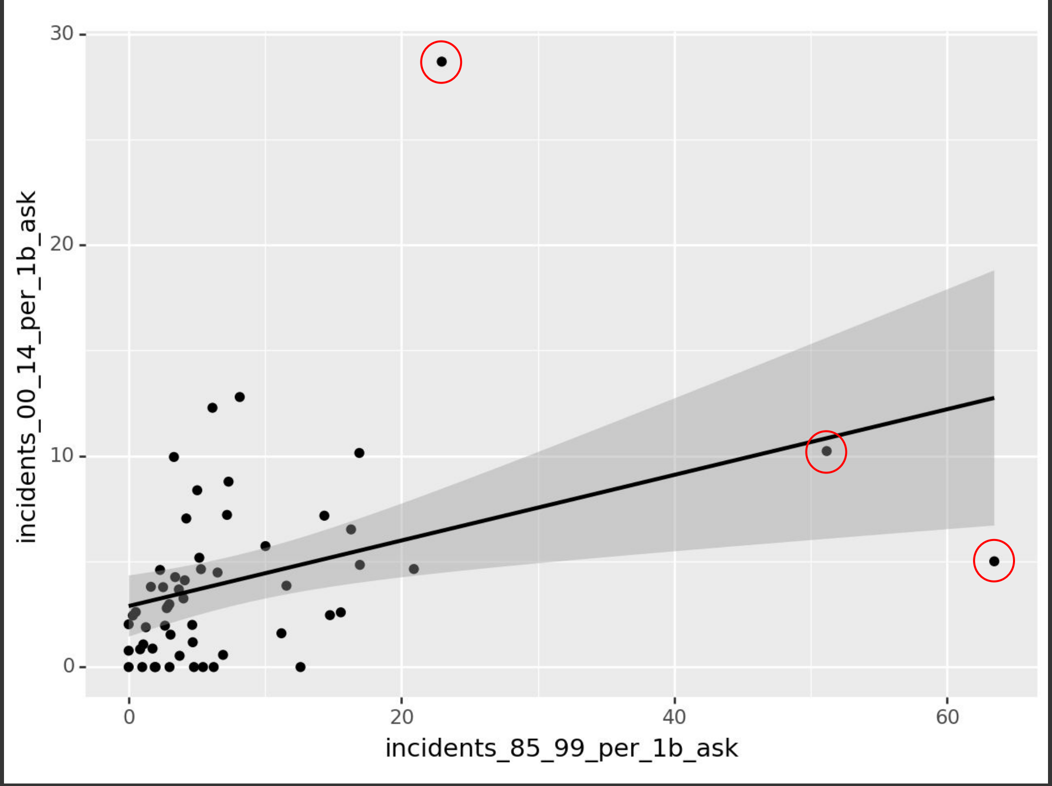
	avail_seat_km_per_week	incidents_85_99	incidents_00_14	regional_subsidiary	incidents_85_99_per_1b_ask	incidents_00_14_per_1b_ask
avail_seat_km_per_week	1	0.28	0.73	0.49	-0.24	-0.23
incidents_85_99	0.28	1	0.4	0.29	0.77	0.1
incidents_00_14	0.73	0.4	1	0.45	0.042	0.34
regional_subsidiary	0.49	0.29	0.45	1	-0.045	-0.026
incidents_85_99_per_1b_ask	-0.24	0.77	0.042	-0.045	1	0.37
incidents_00_14_per_1b_ask	-0.23	0.1	0.34	-0.026	0.37	1

Insights

The correlations between an airline having regional_subsidiaries and incident rates are near 0.

There is a small correlation (0.37) between 85_99_incident rate and 00_14_incident rate

Relationship between Incident Rate in ‘85-’99 and ‘00-’14

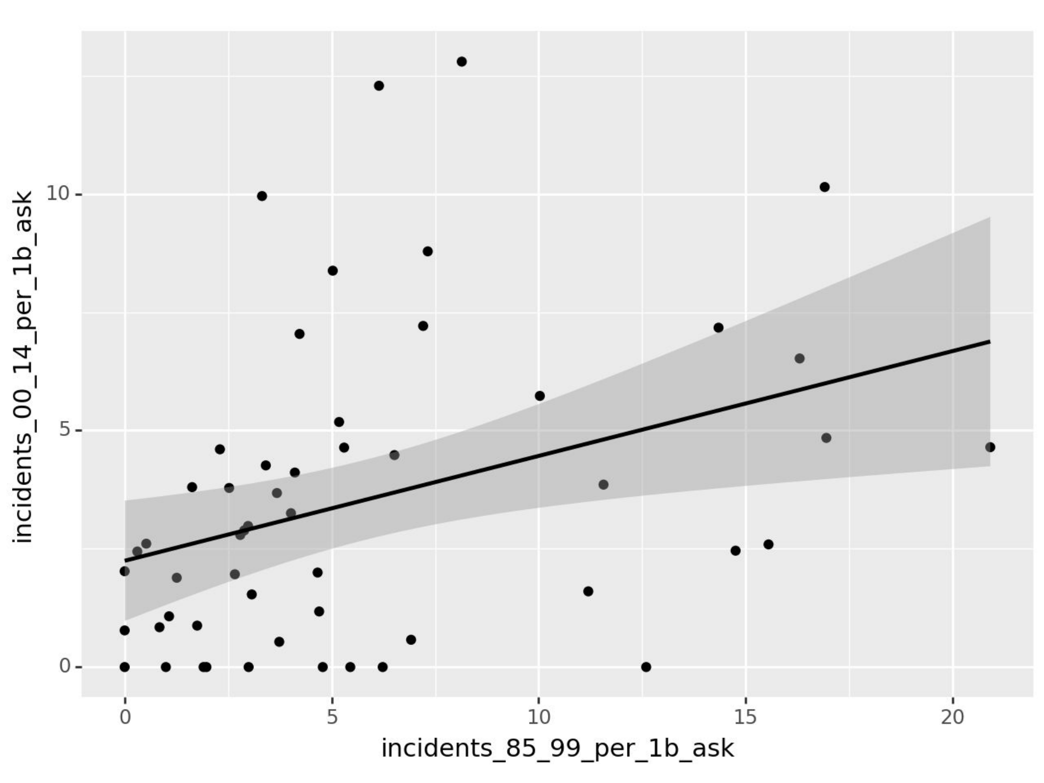


Insights

Aeroflot, Ethiopian Airlines, and Pakistan International are outliers here that may be disproportionately affecting our dataset.

Remove them to observe changes in regression line fit.

Relationship between Incident Rate in ‘85-’99 and ‘00-’14 (no outliers)



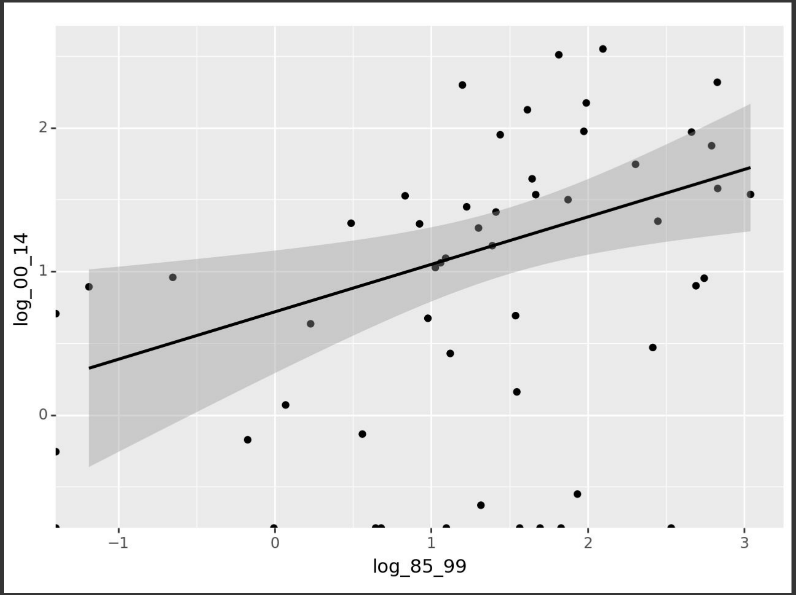
Insights

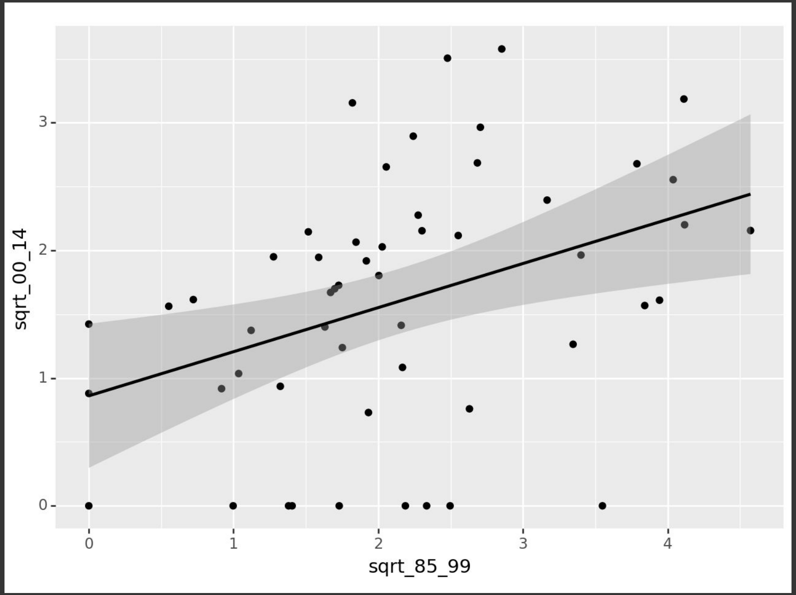
Line poorly fits data

Heteroscedastic data is not appropriate for a linear regression model.

Attempt to treat data to be more homoscedastic.

Treating incident rates for heteroscedasticity





Calculating the log and square root of the incident rates. Square roots of the incident rates seems to be more homoscedastic, but neither treats low r-squared values.

This means that attempts at a linear regression will very poorly explain the relationship.

Conclusion

Insufficient evidence to form a meaningful relationship between the incidents in 85-99 and 00-14.

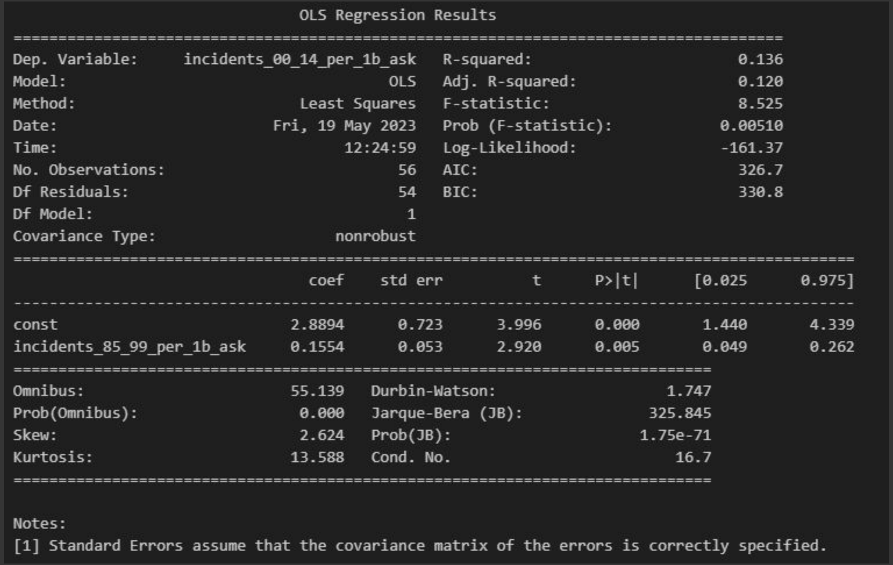
Cannot reject H0, i.e. I cannot suggest a relationship in airline crash rates between 1985-1999 and 2000-2014.

Incident rates are very very low. Even if there were to be a positive correlation between airline incidents in 85-99 and 00-14, it is still very safe to use such airlines

Ultimately, we don't have enough evidence to suggest travelers avoid flying airlines that have had crashes in the past, except for Aeroflot, Ethiopian Airlines, Pakistan International

— Regression Results

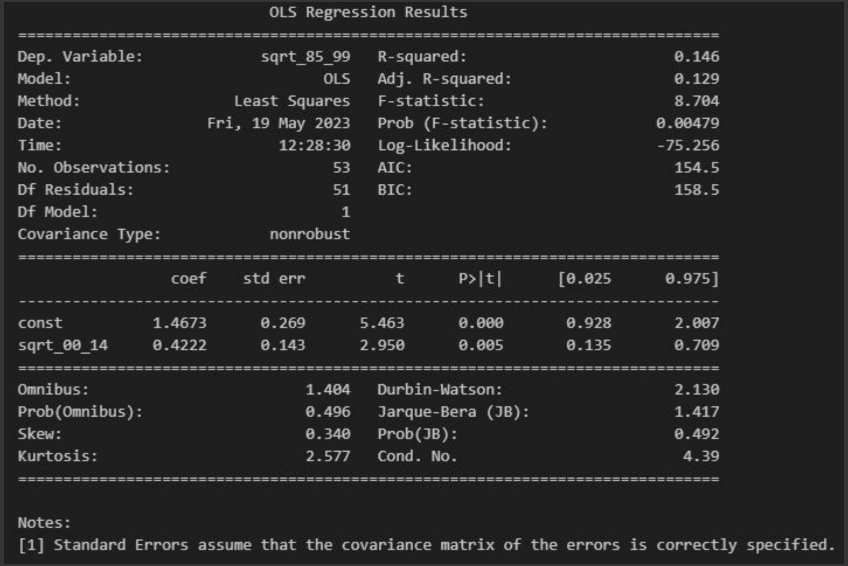
LM for incidents_00_14_per_1b_ask and incidents_00_14_per_1b_ask



```
                                 OLS Regression Results
======================================================================================
Dep. Variable:     incidents_00_14_per_1b_ask   R-squared:                       0.136
Model:                                    OLS   Adj. R-squared:                  0.120
Method:                         Least Squares   F-statistic:                     8.525
Date:                        Fri, 19 May 2023   Prob (F-statistic):            0.00510
Time:                                12:24:59   Log-Likelihood:                -161.37
No. Observations:                          56   AIC:                             326.7
Df Residuals:                              54   BIC:                             330.8
Df Model:                                   1
Covariance Type:                    nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                        2.8894      0.723      3.996      0.000       1.440       4.339
incidents_85_99_per_1b_ask   0.1554      0.053      2.920      0.005       0.049       0.262
==============================================================================
Omnibus:                       55.139   Durbin-Watson:                   1.747
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              325.845
Skew:                           2.624   Prob(JB):                     1.75e-71
Kurtosis:                      13.588   Cond. No.                         16.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

LM for sqrt_85_99 and sqrt_00_14



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             sqrt_85_99   R-squared:                       0.146
Model:                            OLS   Adj. R-squared:                  0.129
Method:                 Least Squares   F-statistic:                     8.704
Date:                Fri, 19 May 2023   Prob (F-statistic):            0.00479
Time:                        12:28:30   Log-Likelihood:                -75.256
No. Observations:                  53   AIC:                             154.5
Df Residuals:                      51   BIC:                             158.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.4673      0.269      5.463      0.000       0.928       2.007
sqrt_00_14     0.4222      0.143      2.950      0.005       0.135       0.709
==============================================================================
Omnibus:                        1.404   Durbin-Watson:                   2.130
Prob(Omnibus):                  0.496   Jarque-Bera (JB):                1.417
Skew:                           0.340   Prob(JB):                        0.492
Kurtosis:                       2.577   Cond. No.                         4.39
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

A p-value of 0.005 for these models indicate that there is a linear relationship between the 2 variables. However, the R-squared value being a very low 0.13 and 0.14 means that these models explain explains a very little amount of the variation in the dependent variable. With the R-squared values being so poor, there is a lack of confidence to extrapolate the findings the larger, actual population.