A Statistical Examination of Fatalities in Aircraft Crashes

Kaja Coraor
University of North Carolina at Chapel Hill
coraor@live.unc.edu

I. Introduction

In the past century there have been thousands of airplane crashes. This paper examines the proportion of fatalities among the passengers and crew on board these aircrafts. It uses a variety of statistical techniques to analyze a number of possible explanatory variables including season, time of day, and number of people on board the plane. The SAS commands along with their outputs and explanations are included in the SAS Code section of the paper. A full list of the SAS commands is available in the appendix.

II. Data

The data contains information on airplane crashes around the world between 1908 and 2009. The data can be found at https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq. Variable descriptions were obtained from http://www.planecrashinfo.com/database.htm and can also be found below. There are 5268 observations in the dataset.

Note that the dataset was restructured to add columns for month and year (based on the "date" in the original dataset) along with hemisphere and season (based on "date" and "location") from original dataset. Hemisphere and season are approximate values should not be interpreted as exactly descriptive of the crash. The restructured dataset also contains a column for "proportion of fatalities among people on board" and is an exact representation of the crash based on the "aboard" and "fatalities" values in the original dataset. The restructured dataset contains 18 variables.

Response Variable: proportion of fatalities among people on board (ProportionFatalities)

Variables in dataset:

1. Date (date of accident - mm/dd/yyyy)
2. Month (month of accident - mm - January = 1, December = 12)
3. Year (year of accident - yyyy - 1908 to 2009)
4. Time (local time when/where accident occured - 24 hour format)
5. TimeInMinutes (number of minutes after 12:00AM local time that the accident occurred)
6. Location (location of crash)
7. Hemisphere (hemisphere of crash - North or South)
8. Season (season during crash - Fall/Winter/Spring/Summer)
9. Winter (1 if Winter, 0 if a different season)
10. Spring (1 if Spring, 0 if a different season)
11. Summer (1 if Summer, 0 if a different season)
8. Operator (airline or operator of aircraft)
9. Flight Number (flight number assigned by aircraft operator)
10. Route (complete or partial route flown prior to accident)
11. Type (aircraft type)
12. Registration (ICAO registration of aircraft)
13. cn/ln (Construction or serial number / line or fuselage number)
14. Aboard (total aboard - crew and passengers)
15. Fatalities (total fatalities aboard - crew and passengers)
16. Proportion of Fatalities Among People on Board (Fatalities/Aboard)
17. Ground (total killed on the ground)
18. Summary (brief description of accident and cause if known)

III. SAS Code

```
>proc import out= plane DATAFILE="/home/coraor0/Stor 455 Project/added_
>               columns_Airplane_Crashes_and_Fatalities_Since_1908.xlsx"
>   DBMS=xlsx REPLACE;  SHEET="data";  GETNAMES=YES;
>run;
```

The "input" procedure was used to read in the data (an xlsx file) and store it as a dataset called "plane." The data was examined to ensure it was imported correctly.

```
>data nomissing;
>   SET plane;
>   IF (Month = . or Year = . or TimeInMinutes = . Winter = . or Spring =
>       . or Summer = . or Aboard = . or Fatalities = . or
>       ProportionFatalities = . or Ground = .) THEN delete;
>run;
```

A new dataset called "nomissing" was created in order to exclude all records from the "plane" dataset that contained missing values. Note that only records with missing values in specific columns are omitted; since only numerical variables will be used in the regressions there is no need to drop observations that contain missing values for string variables such as "summary" or "route."

```
>title Scatter Plot Matrix';
>proc sgscatter data=nomissing;
>   label TimeInMinutes='Time';
>   matrix Month Year TimeInMinutes Winter Spring Summer Aboard Fatalities
>          Ground ProportionFatalities / transparency=0.8
>   markerattrs=graphdata3(symbol=circlefilled);
>run;
```

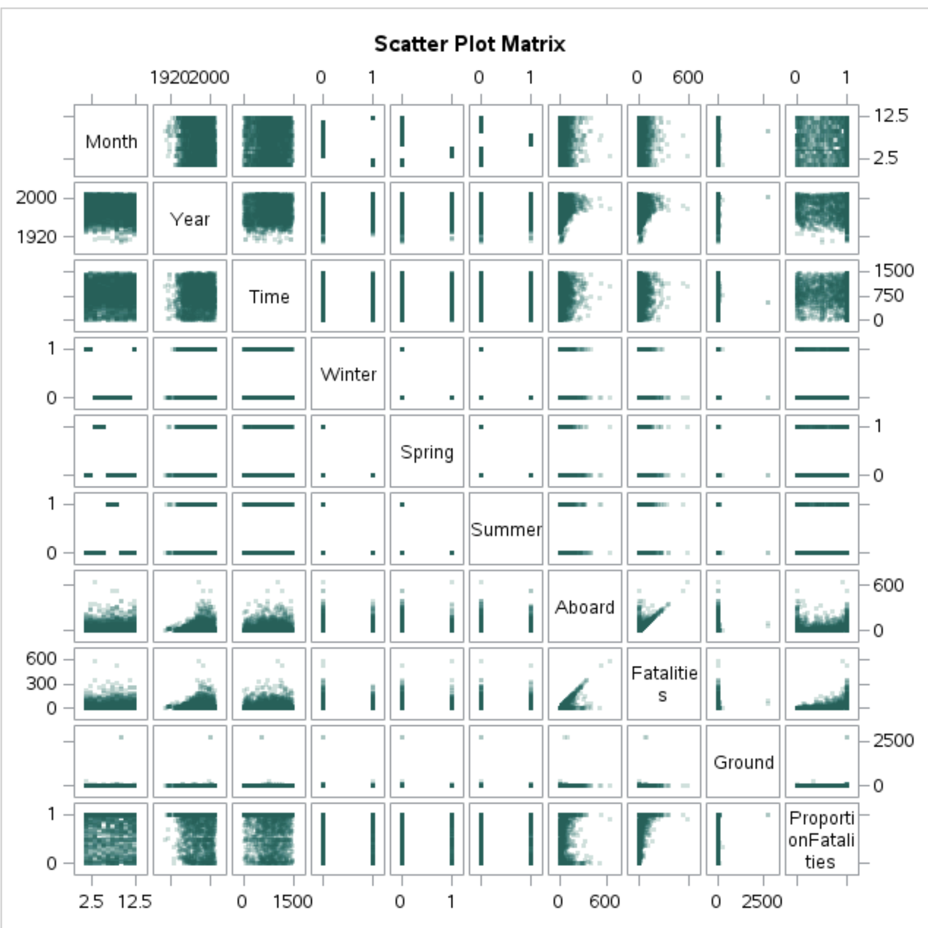

Figure 1: Scatter Plots

In order to perform some diagnostics on the data, a scatterplot was made using SAS's "Scatter Plot Matrix" Snippet. The results are shown above. It appears that "aboard" and "fatalities" have a fairly linear relationship, something that will need to be revisited later in the analysis. "ProportionFatalities" and "fatalities" also seem to have a positive relationship, though more analysis is necessary to determine whether or not the relationship is linear. "Aboard" and "fatalities" both seem to have a positive relationship with "year," which may be explained simply by the increase in the number of observations available in the later years (due to an increase in the number of commercial flights). Aside from all of these possible relationships evident in the scatter plot, it should be noted that the Ground variable seems to have some outliers. This can be examined more closely with a scatter plot of the Ground observations.

```
>proc gplot data=nomissing;
>  plot ProportionFatalities* Ground ;
>run;
```

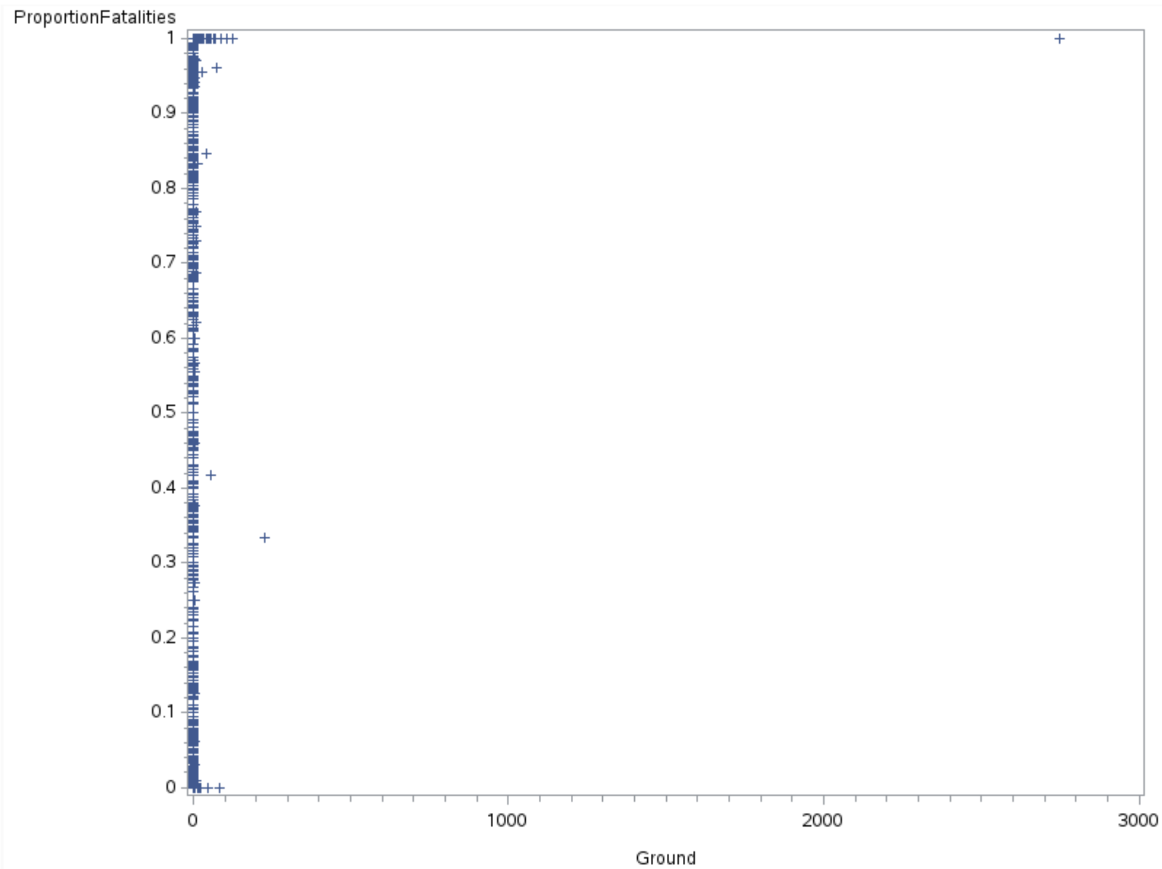

Figure 2: Ground scatter plot

It seems that there is one observation around 2800 whereas the rest are close to 0. To get some more information about this observation, we can print all observations whose Ground value is above some threshold.

```
>proc print data=nomissing;
>  var Date Location Operator Route Type Aboard Fatalities Ground;
>  where Ground > 1000;
>run;
```

Obs	Date	Location	Operator	Route	Type	Aboard	Fatalities	Ground
2604	09/11/2001	New York City, New York	American Airlines	Boston - Los Angeles	Boeing 767-223ER	92	92	2750
2605	09/11/2001	New York City, New York	United Air Lines	Boston - Los Angeles	Boeing B-767-222	65	65	2750

Figure 3: Observations with >1000 ground deaths

After examining the data, it is clear that the outliers in Ground are not due to typos or mistakes but rather are representative of two of the planes that crashed in the terrorist attack on the Twin Towers in September of 2001. Despite being true values, these observations may skew the results. We will perform some tests later in the analysis to determine whether or not these outliers (and other potential observations) are influential and should be excluded from the analysis.

```
>proc means data=nomissing;
> var ProportionFatalities Month Year TimeInMinutes Winter Spring Summer
>     Aboard Fatalities Ground;
>run;
```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
ProportionFatalities	ProportionFatalities	3026	0.8270078	0.3048384	0	1.0000000
Month	Month	3026	6.6140119	3.5446230	1.0000000	12.0000000
Year	Year	3026	1977.42	20.4836109	1908.00	2009.00
TimeInMinutes	TimeInMinutes	3026	797.5905486	361.2650352	1.0000000	1439.00
Winter	Winter	3026	0.2650364	0.4414255	0	1.0000000
Spring	Spring	3026	0.2303371	0.4211182	0	1.0000000
Summer	Summer	3026	0.2478519	0.4318368	0	1.0000000
Aboard	Aboard	3026	34.0974884	51.7955237	1.0000000	644.0000000
Fatalities	Fatalities	3026	24.7914739	40.5837301	0	583.0000000
Ground	Ground	3026	2.6024455	71.0173224	0	2750.00

Figure 4: MEANS Procedure for all variables

The above statement provides some descriptive statistics for each of the numerical variables in the dataset. This information can provide some insights into potential outliers. For example, since the TimeInMinutes variable represents the number of minutes after 12:00AM at which the accident occurred, the value can be no larger than 1440 (the total number of minutes in a day). Therefore if the MEANS procedure shows a maximum for TimeInMinutes larger than 1440 this would suggest that there are some outliers in the dataset that may need to be deleted. Similarly if the maximum for Fatalities were higher than the maximum for Aboard this would imply a potential issue as Fatalities is only measured among number of people on board the plane (Aboard). However based on the results shown above it seems like there are no noticeable issues with the dataset.

```
>proc univariate data=newdata2 alpha=.05;
> var ProportionFatalities;
> histogram / endpoints = 0 to 1.0 by 0.1;
>run;
```

The "univariate" procedure was used to obtain more detailed information about the response variable (proportion of fatalities among people on board – ProportionFatalities). The results are shown on the following page.

The UNIVARIATE Procedure
Variable: ProportionFatalities (ProportionFatalities)

Moments			
N	3026	Sum Weights	3026
Mean	0.82700776	Sum Observations	2502.52547
Std Deviation	0.30483842	Variance	0.09292646
Skewness	-1.6198269	Kurtosis	1.1764133
Uncorrected SS	2350.71052	Corrected SS	281.102543
Coeff Variation	36.8604061	Std Error Mean	0.0055416

Basic Statistical Measures			
Location		Variability	
Mean	0.827008	Std Deviation	0.30484
Median	1.000000	Variance	0.09293
Mode	1.000000	Range	1.00000
		Interquartile Range	0.22222

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	149.2363	Pr > |t|	<.0001
Sign	M	1488	Pr >= |M|	<.0001
Signed Rank	S	2214888	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1.0000000
99%	1.0000000
95%	1.0000000
90%	1.0000000
75% Q3	1.0000000
50% Median	1.0000000
25% Q1	0.7777778
10%	0.2500000
5%	0.0714286
1%	0.0000000
0% Min	0.0000000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	3006	1	3020
0	2978	1	3022
0	2959	1	3023
0	2948	1	3025
0	2939	1	3026

Figure 5: UNIVARIATE Procedure

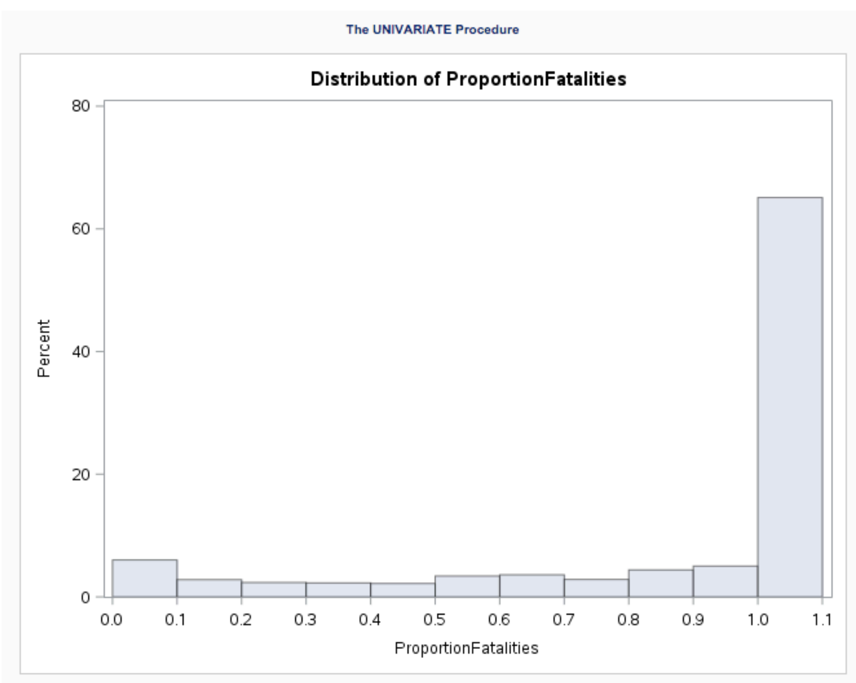

Figure 6: UNIVARIATE Procedure Histogram

It may seem strange that the tallest bin in the histogram is the 1.0-1.1 bar, but it is important to note that a bin contains all values greater than *or equal to* the leftmost value and less than the rightmost value. Therefore all observations represented by the 1.0-1.1 bar in the histogram have ProportionFatalities equal to 1.0 (i.e. all persons on board that airplane died in the crash).

```
>proc corr data=newdata2;
> var ProportionFatalities Month Year TimeInMinutes Winter Spring
>     Summer Fatalities Ground;
>run;
```

Pearson Correlation Coefficients, N = 3026 Prob > |r| under H0: Rho=0										
	ProportionFatalities	**Month**	**Year**	**TimeInMinutes**	**Winter**	**Spring**	**Summer**	**Aboard**	**Fatalities**	**Ground**
ProportionFatalities ProportionFatalities	1.00000	-0.00597 0.7426	-0.03107 0.0874	-0.01646 0.3654	-0.00482 0.7909	0.02767 0.1281	-0.04540 0.0125	-0.21588 <.0001	0.19948 <.0001	0.01328 0.4654
Month Month	-0.00597 0.7426	1.00000	-0.03212 0.0773	0.02249 0.2162	-0.23397 <.0001	-0.41745 <.0001	0.07397 <.0001	0.03892 0.0323	0.03167 0.0816	0.01716 0.3452
Year Year	-0.03107 0.0874	-0.03212 0.0773	1.00000	0.03785 0.0374	-0.01568 0.3884	0.00193 0.9154	0.04410 0.0153	0.06124 0.0008	0.01984 0.2753	0.03118 0.0864
TimeInMinutes TimeInMinutes	-0.01646 0.3654	0.02249 0.2162	0.03785 0.0374	1.00000	0.01338 0.4620	0.00205 0.9104	-0.01875 0.3026	0.02567 0.1580	0.01341 0.4609	-0.01812 0.3189
Winter Winter	-0.00482 0.7909	-0.23397 <.0001	-0.01568 0.3884	0.01338 0.4620	1.00000	-0.32851 <.0001	-0.34472 <.0001	-0.00438 0.8095	-0.02577 0.1563	-0.01400 0.4416
Spring Spring	0.02767 0.1281	-0.41745 <.0001	0.00193 0.9154	0.00205 0.9104	-0.32851 <.0001	1.00000	-0.31403 <.0001	-0.03774 0.0379	-0.02050 0.2597	-0.01469 0.4192
Summer Summer	-0.04540 0.0125	0.07397 <.0001	0.04410 0.0153	-0.01875 0.3026	-0.34472 <.0001	-0.31403 <.0001	1.00000	0.03698 0.0420	0.01753 0.3350	-0.01504 0.4084
Aboard Aboard	-0.21588 <.0001	0.03892 0.0323	0.06124 0.0008	0.02567 0.1580	-0.00438 0.8095	-0.03774 0.0379	0.03698 0.0420	1.00000	0.76371 <.0001	0.02173 0.2320
Fatalities Fatalities	0.19948 <.0001	0.03167 0.0816	0.01984 0.2753	0.01341 0.4609	-0.02577 0.1563	-0.02050 0.2597	0.01753 0.3350	0.76371 <.0001	1.00000	0.03466 0.0566
Ground Ground	0.01328 0.4654	0.01716 0.3452	0.03118 0.0864	-0.01812 0.3189	-0.01400 0.4416	-0.01469 0.4192	-0.01504 0.4084	0.02173 0.2320	0.03466 0.0566	1.00000

Figure 7: Pearson Correlation Coefficients

The "corr" procedure provides the Pearson correlation coefficients for each of the variables specified in the SAS statement. At the 99% confidence level there are a few variables that have a significant linear relationship. Winter, Spring, and Summer all correlate with Month as well as with each other. This makes sense intuitively because Winter, Spring, and Summer are indicator variables for season and since each season has designated months (depending on the hemisphere) the month should be correlated with season. There are two other significant correlations in the table above. The first is Fatalities and ProportionFatalities, which have a positive correlation coefficient of .19948. This follows from the idea that each plane has a limit on the number of people aboard; the higher the number of fatalities, the higher one would expect the proportion of fatalities to be. Fatalities is also strongly correlated with Aboard, having a correlation coefficient of .76371. This is logical for a similar intuitive reason to why Fatalities is correlated with ProportionFatalities; if a plane crash has a high number of fatalities then the number of people aboard the plane must have also been high. Based on these correlations and the logical argument, it seems like Fatalities should be excluded from the model as it introduces a large amount of redundancy to the model.

Before delving further into the analysis we need to identify potential outliers and determine whether or not they are influential. The hat matrix diagonals can be used to identify any outliers, and there are a number of tests that can be used to determine which are influential and should be deleted for the model. The DFFITS and Hat Matrix Diagonals tests are described on the next page.

1. Hat Matrix Diagonals – if the value is greater than $2 * p/n$ then the cases are influential
   - Assuming we use a model with 8 explanatory variables (Month, Year, TimeInMinutes, Winter, Spring, Summer, Aboard, Ground), hat matrix value must be greater than $2 * 9/3206$ or 0.00595.
2. DFFITS - if the magnitude of the DFITTS value is larger than $2\sqrt{p/n}$ then the cases are influential
   - Assuming we use a model with 8 explanatory variables (Month, Year, TimeInMinutes, Winter, Spring, Summer, Aboard, Ground), the absolute value of the DFFITS value must be greater than $2\sqrt{9/3026}$ or 0.1091. Note that Fatalities was excluded from the model; this will be explained later in the analysis.

We could also examine the Cook's Distances, R-Student values, and the DFBETAS in order to determine which outliers should be discarded, but in this case DFFITS and the hat matrix diagonals should be sufficient.

```
>proc reg data=newdata;
> model ProportionFatalities = Month Year TimeInMinutes Winter Spring
>       Summer Aboard Ground;
> output out=outdata r=residual h=hat rstudent=rstudent dffits=dffits;
>run;
>proc print data=outdata;
> var ProportionFatalities Month Year TimeInMinutes Winter Spring
> Summer Aboard Ground residual hat rstudent dffits;
> where hat > 2*9/3026 or dffits > 2*sqrt(9/3026);
>run;
```

Obs	ProportionFatalities	Month	Year	TimeInMinutes	Winter	Spring	Summer	Aboard	Ground	residual	hat	rstudent	dffits
17	1	12	1923	150	1	0	0	52	0	0.17972	0.00662	0.60574	0.04943
1147	0.205240175	11	1970	1025	0	0	0	229	0	-0.38829	0.00618	-1.30871	-0.10319
1244	0.585227273	12	1972	1422	1	0	0	176	0	-0.05588	0.00653	-0.18831	-0.01527
1304	0.003355705	2	1974	1290	1	0	0	298	0	-0.48570	0.01109	-1.64133	-0.17378
1307	1	3	1974	701	0	1	0	346	0	0.55215	0.01397	1.86886	0.22247
1342	1	12	1974	1335	1	0	0	191	0	0.37746	0.00682	1.27257	0.10544
1357	0.46969697	4	1975	990	0	1	0	330	0	0.00462	0.01265	0.01563	0.00177
1434	0.905279503	3	1977	1027	0	1	0	644	0	0.83543	0.04832	2.88053	0.64907
1476	1	1	1978	1215	1	0	0	213	0	0.40443	0.00655	1.36337	0.11073
1512	0.698473282	11	1978	1410	0	0	0	262	0	0.15180	0.00854	0.51210	0.04752
1520	0.837209302	12	1978	39	1	0	0	129	0	0.12620	0.00603	0.42522	0.03311
1521	0.052910053	12	1978	1095	1	0	0	189	0	-0.57330	0.00629	-1.93302	-0.15377
1539	1	5	1979	904	0	1	0	271	2	0.46093	0.00862	1.55563	0.14509
1567	1	11	1979	769	0	0	0	257	0	0.44157	0.00745	1.48935	0.12905
1594	1	8	1980	1148	0	0	1	301	0	0.53177	0.01020	1.79638	0.18236
1602	0.006872852	12	1980	1392	1	0	0	291	0	-0.48807	0.01193	-1.65003	-0.18127
1631	1	12	1981	533	1	0	0	180	0	0.35821	0.00610	1.20720	0.09459
1638	0.009433962	1	1982	1176	1	0	0	212	0	-0.58676	0.00641	-1.97857	-0.15897
1653	0	6	1982	1244	0	0	1	257	0	-0.52226	0.00784	-1.76210	-0.15668
1664	0.126903553	9	1982	720	0	0	0	394	0	-0.25923	0.01744	-0.87856	-0.11705
1694	1	9	1983	1106	0	0	0	269	0	0.46053	0.00836	1.55406	0.14271
1747	0	2	1985	615	1	0	0	274	0	-0.52252	0.00909	-1.76410	-0.16898
1757	1	6	1985	435	0	0	1	329	0	0.56171	0.01234	1.89970	0.21239
1762	0.992366412	8	1985	1136	0	0	1	524	0	0.80540	0.03078	2.75143	0.49035
1777	1	12	1985	405	1	0	0	256	0	0.45352	0.00980	1.53150	0.15235
1804	0.041666667	9	1986	360	0	0	0	384	1	-0.35933	0.01714	-1.21777	-0.16081
1882	1	7	1988	655	0	0	1	290	0	0.51547	0.00931	1.74048	0.16874
1912	1	12	1988	1143	1	0	0	259	11	0.46380	0.00960	1.56606	0.15416
1921	0.025280899	2	1989	129	1	0	0	356	0	-0.39758	0.01605	-1.34673	-0.17199
1941	0.375838926	7	1989	960	0	0	1	298	0	-0.09567	0.00978	-0.32295	-0.03210
2011	0.566371681	10	1990	555	0	0	0	226	0	-0.03025	0.00611	-0.10194	-0.00799
2017	0.04040404	12	1990	825	1	0	0	198	0	-0.57396	0.00660	-1.93553	-0.15772
2040	1	5	1991	1397	0	1	0	223	0	0.40813	0.00688	1.37610	0.11457
2047	1	7	1991	520	0	0	1	261	0	0.47858	0.00780	1.61455	0.14312
2097	0	7	1992	1061	0	0	1	292	0	-0.47741	0.00952	-1.61200	-0.15801
2116	0.164705882	12	1992	473	1	0	0	340	0	-0.27392	0.01515	-0.92730	-0.11502
2129	0.007575758	4	1993	70	0	1	0	264	0	-0.54418	0.00979	-1.83797	-0.18277
2171	0.974169742	4	1994	1216	0	1	0	271	0	0.44167	0.00902	1.49087	0.14220
2210	0.003412969	12	1994	690	1	0	0	293	0	-0.49183	0.01161	-1.66250	-0.18019
2217	0.012552301	12	1994	1020	1	0	0	239	0	-0.54758	0.00853	-1.84829	-0.17148
2275	0.975609756	12	1995	1298	1	0	0	164	0	0.32398	0.00598	1.09175	0.08466
2298	0.010909091	6	1996	727	0	0	1	275	0	-0.48990	0.00852	-1.65340	-0.15331
2304	1	7	1996	1231	0	0	1	230	0	0.44721	0.00645	1.50762	0.12151
2324	1	11	1996	1120	0	0	0	349	0	0.56446	0.01382	1.91041	0.22611
2362	0.901574803	8	1997	102	0	0	1	254	0	0.36916	0.00866	1.24574	0.11641
2371	1	9	1997	814	0	0	0	234	0	0.41740	0.00649	1.40710	0.11375
2384	0.002544529	12	1997	1390	1	0	0	393	0	-0.36010	0.01967	-1.22194	-0.17308
2424	1	9	1998	1290	0	0	0	229	0	0.41561	0.00678	1.40127	0.11581
2473	0.001934236	7	1999	685	0	0	1	517	0	-0.19446	0.02993	-0.66323	-0.11650
2479	0.00952381	8	1999	1125	0	0	1	315	0	-0.43665	0.01120	-1.47552	-0.15707
2492	1	10	1999	412	0	0	0	217	0	0.39301	0.00616	1.32459	0.10429
2504	0.050955414	12	1999	580	1	0	0	314	2	-0.41781	0.01325	-1.41328	-0.16377
2598	0	8	2001	1126	0	0	1	304	0	-0.45949	0.01052	-1.55223	-0.16006
2604	1	9	2001	527	0	0	0	92	2750	0.03375	0.49511	0.15954	0.15799
2605	1	9	2001	543	0	0	0	65	2750	-0.00003	0.49511	-0.00012	-0.00012
2617	1	11	2001	556	0	0	0	260	5	0.44848	0.00812	1.51319	0.13695
2656	1	5	2002	929	0	1	0	225	0	0.40918	0.00647	1.37934	0.11133
2699	1	2	2003	1050	1	0	0	275	0	0.48704	0.00943	1.64450	0.16047
2830	0	8	2005	964	0	0	1	309	0	-0.45367	0.01080	-1.53277	-0.16015
2971	0.14953271	6	2008	1245	0	0	1	214	0	-0.42034	0.00618	-1.41677	-0.11174
3025	1	6	2009	15	0	0	1	228	0	0.43699	0.00811	1.47439	0.13331

Figure 8: Outliers & Influential Cases

The table on the previous page lists all observations where either the Hat Matrix Diagonal is above the threshold (indicating an outlier) or the DFFITS magnitude is greater than its threshold (indicating an influential case).

```
>data newdata2;
> set outdata;
> if hat > 2*9/3026 and dffits > 2*sqrt(9/3026) then delete;
> keep ProportionFatalities Month Year TimeInMinutes Winter Spring Summer
> Aboard Ground;
>run;
```

The statement above drops all observations that are marked as influential outliers by the Hat Matrix Diagonals test and the DFFITS test.

```
>proc reg data=newdata2;
> model ProportionFatalities = Month Year TimeInMinutes Winter Spring
>       Summer Aboard Ground;
> output out=temp student=r;
> plot ProportionFatalities*(Month Year TimeInMinutes Winter Spring Summer
>      Aboard Ground);
> plot student.*(Month Year TimeInMinutes Winter Spring Summer Aboard
>      Ground p.);
> plot student.*nqq.;
>run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: ProportionFatalities ProportionFatalities

Number of Observations Read	3001
Number of Observations Used	3001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	21.02596	2.62824	30.32	<.0001
Error	2992	259.38185	0.08669		
Corrected Total	3000	280.40781			

Root MSE	0.29443	R-Square	0.0750
Dependent Mean	0.82564	Adj R-Sq	0.0725
Coeff Var	35.66130		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.49597	0.52039	2.87	0.0041
Month	Month	1	0.00003056	0.00199	0.02	0.9877
Year	Year	1	-0.00029996	0.00026299	-1.14	0.2541
TimeInMinutes	TimeInMinutes	1	-0.00001018	0.00001491	-0.68	0.4949
Winter	Winter	1	-0.01313	0.01767	-0.74	0.4575
Spring	Spring	1	-0.00383	0.01964	-0.19	0.8454
Summer	Summer	1	-0.03203	0.01626	-1.97	0.0490
Aboard	Aboard	1	-0.00178	0.00011782	-15.15	<.0001
Ground	Ground	1	0.00006985	0.00010617	0.66	0.5106

Figure 9-1: REG Procedure for diagnostics

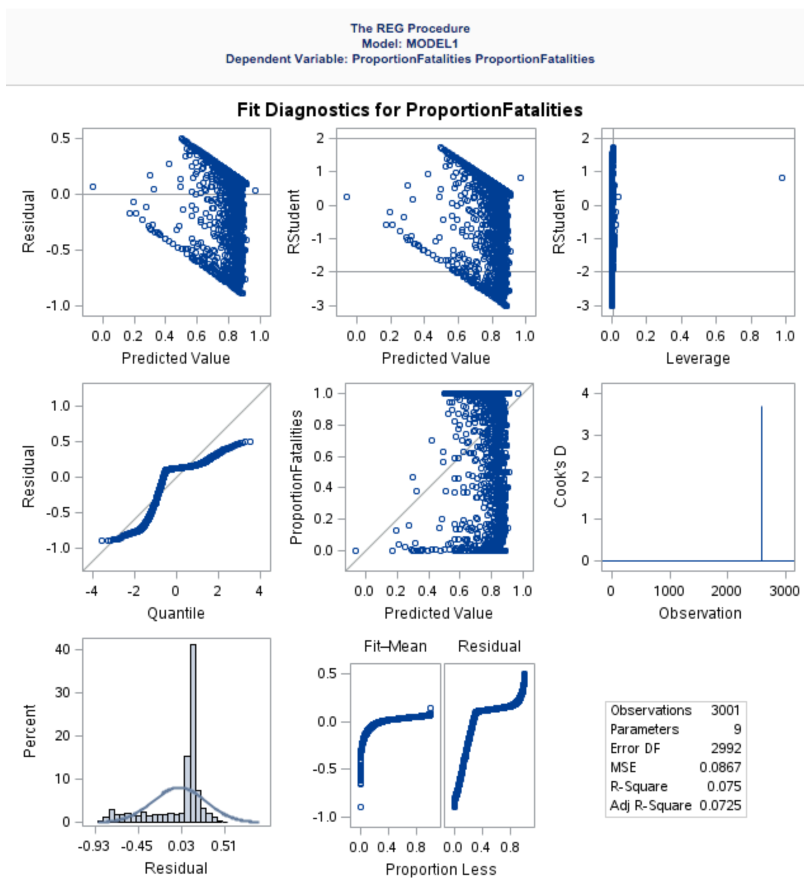

Figure 9-2: REG Procedure Graphs

The statement above performs a regression on the new dataset (without the outliers) and excluding the Fatalities variable, with the results displayed above. Unfortunately the R^2 value is quite low at 0.0750, which means the model only explains about 7.5% of the variable in the response data. Ideally R^2 value would be much closer to 100%. A low R^2 value does not, however, mean that the model is meaningless or unusable. There can still be statistically significant predictors in the model, but a low R^2 value does mean that predictions of the response variable will not be very precise. The p-value for the F test is less than 0.0001 which means the model is significant (despite the low R^2 value). The only explanatory variables with p-values less than 0.05 are Aboard and Summer. Note that the residual plots in Figure 9-2 do not represent a random Gaussian distribution around zero. This suggests that the error terms are not normal and therefore a linear model is not necessarily the best model for this system. However since we have not studied nonlinear regressions, we will continue with the linear regression.

```
>proc reg data=newdata2;
> model ProportionFatalities = Month Year TimeInMinutes Winter Spring
>       Summer Aboard Ground;
> test1: test Winter, Spring, Summer;
>run;
```

The REG Procedure
Model: MODEL1

Test test1 Results for Dependent Variable ProportionFatalities				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	0.15029	1.73	0.1579
Denominator	2992	0.08669		

Figure 10: F Test for model without seasons

The statement above performs an F test to determine whether or not Winter/Spring/Summer should be included in the model. Essentially this statement compares two models, one which includes all variables (Month, Year, TimeInMinutes, Winter, Spring, Summer, Aboard, Ground) with the model that excludes Winter, Spring, and Summer. Because the second model is nested inside the first model, this comparison can easily be achieved by performing an F test in which the hypotheses are as follows:

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$
H_1: β_4 and β_5 and β_6 are not all 0 (i.e. at least one is nonzero)

In this case, β_4 represents the regression coefficient of the Winter indicator variable, β_5 represents that of the Spring indicator variable, and β_6 represents the coefficient for the Summer variable. The p-value for the F test (shown in Figure 10 above) is 0.1579 which is not significant at the 95% confidence level. This suggests that Winter, Spring, and Summer are all not useful for predicting ProportionFatalities in a linear model therefore they can be excluded from the model.

```
>proc reg data=newdata2;
> model ProportionFatalities = Month Year TimeInMinutes Winter Spring
>       Summer Aboard Ground / VIF TOL;
>run;
```

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1.52772	0.51964	2.94	0.0033	.	0
Month	Month	1	0.00031307	0.00152	0.21	0.8367	0.99632	1.00370
Year	Year	1	-0.00032319	0.00026280	-1.23	0.2189	0.99470	1.00533
TimeInMinutes	TimeInMinutes	1	-0.00000977	0.00001491	-0.66	0.5123	0.99748	1.00253
Aboard	Aboard	1	-0.00179	0.00011775	-15.24	<.0001	0.99579	1.00423
Ground	Ground	1	0.00007429	0.00010616	0.70	0.4841	0.99899	1.00101

Figure 11: Variance Inflation Analysis

The statement above checks for multicolinearity in the model. A variance inflation (VIF) value greater than 10 would suggest that there is excessive multicolinearity and some of the variables should be removed from the model. Since none of the variables have a VIF value over 10 there does not seem to be an issue of multicolinearity.

```
>data transformations;
> set newdata2;
> _id_ = _n_;
> month_year = Month*Year;
> month_time = Month*TimeInMinutes;
> month_aboard = Month*Aboard;
> month_ground = Month*Ground;
> year_time = Year*TimeInMinutes;
> year_aboard = Year*Aboard;
> year_ground = Year*Ground;
> time_aboard = TimeInMinutes*Aboard;
> time_ground = TimeInMinutes*Ground;
> aboard_ground = Aboard*Ground;
> aboard2 = Aboard*Aboard;
> ground2 = Ground*Ground;
> log_aboard = log(Aboard+1);
> log_ground = log(Ground+1);
>run;
```

In order to find a model with better fit, some interaction terms need to be explored. The above statement creates a number of interaction terms in a new dataset called "transformations."

```
>proc reg data=transformations;
> Stepwise: model ProportionFatalities= Month Year TimeInMinutes
>                 Aboard Ground month_year month_time month_aboard
>                 month_ground year_time year_aboard year_ground
>                 time_aboard time_ground aboard_ground aboard2 ground2
>                 log_aboard log_ground / selection=stepwise;
>run;
>quit;
```

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	year_aboard			1	0.0727	0.0727	60.6185	235.29	<.0001
2	log_aboard			2	0.0033	0.0761	51.6521	10.79	0.0010
3	aboard2			3	0.0084	0.0845	25.9963	27.45	<.0001
4	Year		Year	4	0.0021	0.0866	21.0555	6.90	0.0086
5	time_aboard			5	0.0008	0.0873	20.5321	2.51	0.1131

Figure 12: Stepwise algorithm with interaction terms

The statement above performs the stepwise selection algorithm to determine which variables to include in the dataset. The algorithm proceeded through 5 steps before arriving at the results printed above. The results show that year, log(aboard), and aboard^2 are statistically significant along with the interaction terms between year and aboard as well as time and aboard.

```
>proc reg data=transformations;
> Forward: model ProportionFatalities = Month Year TimeInMinutes Aboard
>                Ground year_aboard time_aboard aboard2 log_aboard /
>                selection=FORWARD vif tol slentry=0.1;
> Backward: model ProportionFatalities = Month Year TimeInMinutes Aboard
>                 Ground year_aboard time_aboard aboard2 log_aboard /
>                 selection=B vif tol slstay=0.1;
> Stepwise: model ProportionFatalities = Month Year TimeInMinutes Aboard
>                 Ground year_aboard time_aboard aboard2 log_aboard /
>                 selection=stepwise vif tol slentry=0.1 slstay=0.1;
> rsquare: model ProportionFatalities = Month Year TimeInMinutes Aboard
>                Ground year_aboard time_aboard aboard2 log_aboard /
>                selection=rsquare vif tol;
> adjrsq: model ProportionFatalities = Month Year TimeInMinutes Aboard
>               Ground year_aboard time_aboard aboard2 log_aboard /
>               selection=adjrsq vif tol;
> cp: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground
>           year_aboard time_aboard aboard2 log_aboard / selection=cp vif
>           tol;
>run;
>quit;
```

There are a number of different models that could be adequate for the dataset, and there is no way to determine which is definitively the "best" model. However there are methods to compare various models. The selection algorithms above provide multiple models, which can later be compared using these various methods. The forward and stepwise algorithms both provide the same model involving: Year, year_aboard, aboard^2, and log(aboard). This model has an R^2 value of 0.0866 and a root MSE of 0.29239. The backward elimination method includes the untransformed aboard variable rather than the year_aboard interaction term. The rsquare selection method provides a model with 9 variables and an R^2 value of 0.08880. While this value is slightly better than the R^2 value of the previous model, this does not necessarily mean that the model is better. In fact, introducing additional variables to a model often increases the R^2 value without actually improving the fit of the model. Instead it may lead to "overfitting" where the model may fit the given dataset very well but will not generalize to other datasets. The adjrsq method also included a number of variables and is likely subject to overfitting. The cp method provided the same model as the forward selection algorithm and stepwise algorithm with an additional time_aboard term. I chose to use the model from the forward and stepwise selection algorithms as they seem the most logical to me. However it is worth mentioning that there is no way to determine the "best" model so at this point any of the aforementioned models could be chosen.

```
>proc reg data=transformations;
> model ProportionFatalities = Year year_aboard aboard2 log_aboard;
> output out=temp student=r;
> plot ProportionFatalities*(Year year_aboard aboard2 log_aboard);
> plot student.*(Year year_aboard aboard2 log_aboard p.);
> plot student.*nqq.;
>run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: ProportionFatalities ProportionFatalities

Number of Observations Read	3001
Number of Observations Used	3001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	24.27361	6.06840	70.98	<.0001
Error	2996	256.13420	0.08549		
Corrected Total	3000	280.40781			

Root MSE	0.29239	R-Square	0.0866
Dependent Mean	0.82564	Adj R-Sq	0.0853
Coeff Var	35.41368		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.41199	0.53974	4.47	<.0001
Year	Year	1	-0.00070984	0.00027015	-2.63	0.0086
year_aboard		1	6.401562E-7	2.461956E-7	2.60	0.0094
aboard2		1	-0.00000733	0.00000133	-5.50	<.0001
log_aboard		1	-0.07157	0.01105	-6.48	<.0001

Figure 13: REG Procedure on new model

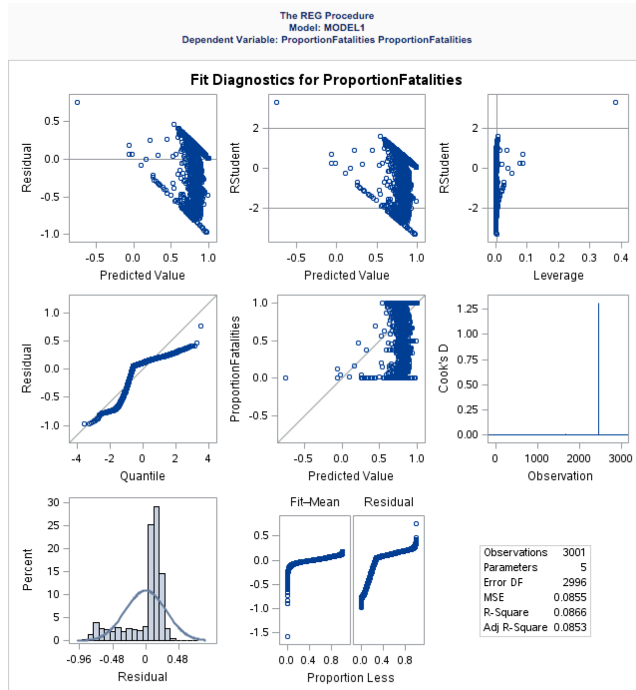

Figure 14: Diagnostic Plots

This statement performs a regression on the new model with Year, aboard^2, and the interaction terms between year and aboard as well as log and aboard. The R^2 value for this model is 0.0866 with a root MSE of 0.29239. The plot on the quantile plot (left column middle row in Figure 14) shows a slightly better approximation to a normal line than we had seen previously, which suggests that the new model has a slightly more normal error term, although it still does not look perfectly linear.

```
>proc reg data=transformations;
> model ProportionFatalities = Year year_aboard aboard2 log_aboard
>       / vif tol;
>run;
>quit;
```

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	2.41199	0.53974	4.47	<.0001	.	0
Year	Year	1	-0.00070984	0.00027015	-2.63	0.0086	0.92758	1.07807
year_aboard		1	6.401562E-7	2.461956E-7	2.60	0.0094	0.05689	17.57728
aboard2		1	-0.00000733	0.00000133	-5.50	<.0001	0.12792	7.81755
log_aboard		1	-0.07157	0.01105	-6.48	<.0001	0.16704	5.98641

Figure 15: VIF TOL Analysis for new model

Now that we have a new model we have to check for multicolinearity again. This time there does seem to be an issue of colinearity in the model. The VIF value for the interaction term between year and aboard is 17.57728. Because this value is larger than 10 it suggests that the variable should be excluded from the model.

```
>proc reg data=transformations;
> model ProportionFatalities = Year aboard2 log_aboard;
>run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: ProportionFatalities ProportionFatalities

Number of Observations Read	3001
Number of Observations Used	3001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	23.69560	7.89853	92.21	<.0001
Error	2997	256.71221	0.08566		
Corrected Total	3000	280.40781			

Root MSE	0.29267	**R-Square**	0.0845
Dependent Mean	0.82564	**Adj R-Sq**	0.0836
Coeff Var	35.44770		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.10073	0.52680	3.99	<.0001
Year	Year	1	-0.00057258	0.00026520	-2.16	0.0309
aboard2		1	-0.00000419	5.589718E-7	-7.49	<.0001
log_aboard		1	-0.04636	0.00530	-8.74	<.0001

Figure 16: Regression for final model

The above shows the regression output for the final model that excludes the interaction term between year and abroad, as this term was shown to introduce multicolinearity to the model.

IV. Summary

After using various analytical techniques, it seems that the majority of the variables in the dataset are not significant and should be excluded from the model. That leaves the following for the model:

$$ProportionFatalities = 2.10073 - 0.00057258 * Year - 0.00000419 * Aboard^2 - 0.04636 * \log(Aboard)$$

As was noted earlier, the R^2 value is small for this model, which makes any predictions made using the model very imprecise. Despite the impreciseness, it is still possible to make predictions using this model. For example, if we predict the expected proportion of fatalities on a plane that crashes in 2064 with 143 people aboard, we would expect this proportion to be

$$ProportionFatalities = 2.10073 - 0.00057258 * 2064 - 0.00000419 * 143^2 - 0.04636 * \log(143)$$

This yields a proportion of 0.603166. Therefore we would expect that approximately 60.3% of people on board the plane would suffer fatalities. In other words, if you happened to be on that plane with the 142 other people, you would have a 39.7% chance of surviving the crash. However as was previously noted, this value is not precise because of the low R^2 value, so this prediction has a very wide confidence interval.

Examining the coefficients more closely, we see that year has a negative coefficient, which suggests that plane crashes in later years are less deadly (i.e. they have lower proportions of fatalities) than crashes in earlier years. This can be loosely interpreted as "provided your plane crashes today, you are less likely to die in that crash than you would have been if you were in that same crash many years ago" although this is a very loose interpretation as there are many other factors that affect your likelihood in surviving a plane crash. The second coefficient (for the $aboard^2$ term) suggests that number of people on board the plane and proportion of fatalities in the crash are negatively quadratically related. That is, as the number of people on board the plane increases, the proportion of fatalities in the crash will decrease quadratically. The last coefficient also deals with number of people on board the plane but suggests that proportion of fatalities in the crash decreases with the natural log of the number of people on board.

Overall, this equation does not seem like a very accurate model to predict the proportion of fatalities in an airplane crash. One of the most important things to note is that the errors did not seem to be random as we would expect with our model (this can be seen in the many residual plots), therefore our assumptions for the model failed.

In order to improve the model we would likely need to obtain a better dataset that contains more information on the plane and the crash. For example, if we could find out the year that each plane was built or the amount of experience the pilot had or even the manufacturer of the engine, that may provide more insights into the proportion of fatalities in a crash and could further improve the model.

V. Appendix

```
/*
  The data contains information on airplane crashes around the world
  between 1908 and 2009. The data was obtained from
  https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq

  Variable descriptions can be found at http://www.planecrashinfo.com/database.htm
  or copied below.

  Number of records: 5268
  Number of variables in original dataset: 13

  The dataset was restructured to add columns for month and year
  (based on the "date" in the original dataset) along with hemisphere
  and season (based on "date" and "location") from original dataset.
  Hemisphere and season are approximate values should not be interpreted
  as exactly descriptive of the crash. The restructured dataset
  also contains a column for "proportion of fatalities among people
  on board" and is an exact representation of the crash based on the
  "aboard" and "fatalities" values in the original dataset.

  Number of variables in final dataset: 18

  Reponse Variable: proportion of fatalities among people on board

  *-----------------------------------------------------------------------------------------------*
  | Variable Information:                                                                         |
  | 1. Date (date of accident - mm/dd/yyyy)                                                       |
  | 2. Month (month of accident - mm - January = 1, December = 12)                                |
  | 3. Year (year of accident - yyyy - 1908 to 2009)                                              |
  | 4. Time (local time when/where accident occured - 24 hour format)                             |
  | 5. Location (location of crash)                                                               |
  | 6. Hemisphere (hemisphere of crash - North or South)                                          |
  | 7. Season (season during crash - Fall/Winter/Spring/Summer)                                   |
  | 8. Operator (airline or operator of aircraft)                                                 |
  | 9. Flight Number (flight number assigned by aircraft operator)                                |
  |10. Route (complete or partial route flown prior to accident)                                  |
  |11. Type (aircraft type)                                                                       |
  |12. Registration (ICAO registration of aircraft)                                               |
  |13. cn/ln (Construction or serial number / line or fuselage number                             |
  |14. Aboard (total aboard - crew and passengers)                                                |
  |15. Fatalities (total fatalities aboard - crew and passengers)                                 |
  |16. Proportion of Fatalities Among People on Board (Fatalities/Aboard)                         |
  |17. Ground (total killed on the ground)                                                        |
  |18. Summary (brief description of accident and cause if known)                                 |
  *-----------------------------------------------------------------------------------------------*

*/
```

```
/* Read in data */
 PROC IMPORT OUT= plane DATAFILE= "/home/coraor0/Stor 455
Project/added_columns_Airplane_Crashes_and_Fatalities_Since_1908.xlsx"
         DBMS=xlsx REPLACE;
    SHEET="data";
    GETNAMES=YES;
 RUN;

 /* Print data */
 proc print data=plane;
 run;

 /* Drop observations with missing data in numerical variables */
 DATA nomissing;
   SET plane;
   IF Month = . or Year = . or TimeInMinutes = . or Winter = . or Spring = . or Summer = . or Aboard = . or
Fatalities = . or ProportionFatalities = . or Ground = . or TimeInMinutes > 1440 then delete;
 RUN;

 /* Print data with no missing values */
 proc print data=nomissing;
 run;

 /*--Scatter Plot Matrix--*/
title 'Scatter Plot Matrix';
 proc sgscatter data=nomissing;
   label TimeInMinutes='Time';
   matrix Month Year TimeInMinutes Winter Spring Summer Aboard Fatalities Ground ProportionFatalities /
     transparency=0.8 markerattrs=graphdata3(symbol=circlefilled);
 run;

/* Scatter plot for Ground */
 proc gplot data=nomissing;
  plot ProportionFatalities* Ground ;
 run;

 /* print observations with outliers in Ground */
 proc print data=nomissing;
 var Date Location Operator Route Type Aboard Fatalities Ground;
  where Ground > 1000;
 run;

 /* Get summary descriptive statistics for each variable */
 proc means data=nomissing;
  var ProportionFatalities Month Year TimeInMinutes Winter Spring Summer Aboard Fatalities Ground;
 run;

  /* Drop any outliers in TimeInMinutes */
 DATA newdata;
   SET nomissing;
   IF TimeInMinutes > 1440 then delete;
 RUN;

  /* Scatter plots for time */
 proc gplot data=newdata;
```

```
 plot ProportionFatalities* TimeInMinutes ;
run;

/* Histogram for response variable */
proc univariate data=newdata alpha=.05;
var ProportionFatalities;
histogram / endpoints = 0 to 1.0 by 0.1;
run;

 /* Correlation matrix */
proc corr data=newdata;
var ProportionFatalities Month Year TimeInMinutes Winter Spring Summer Aboard Fatalities Ground;
run;

/* Identify potential outliers and influential cases */
proc reg data=newdata;
 model ProportionFatalities = Month Year TimeInMinutes Winter Spring Summer Aboard Ground;
 output out=outdata r=residual h=hat rstudent=rstudent dffits=dffits;
run;
proc print data=outdata;
	var ProportionFatalities Month Year TimeInMinutes Winter Spring Summer Aboard Ground residual
hat rstudent dffits;
	where hat > 2*9/3026 or dffits > 2*sqrt(9/3026);
run;
data newdata2;
  set outdata;
  if hat > 2*9/3026 and dffits > 2*sqrt(9/3026) then delete;
  keep ProportionFatalities Month Year TimeInMinutes Winter Spring Summer Aboard Ground;
run;

/* Diagnostics */
proc reg data=newdata2;
 model ProportionFatalities = Month Year TimeInMinutes Winter Spring Summer Aboard Ground;
 output out=temp student=r;
 plot ProportionFatalities*(Month Year TimeInMinutes Winter Spring Summer Aboard Ground);
 plot student.*(Month Year TimeInMinutes Winter Spring Summer Aboard Ground p.);
 plot student.*nqq.;
run;

/* Test Winter/Spring/Summer */
proc reg data=newdata2;
 model ProportionFatalities = Month Year TimeInMinutes Winter Spring Summer Aboard Ground;
 test1: test Winter, Spring, Summer;
run;

 /* check for multicolinearity */
proc reg data=newdata2;
 model ProportionFatalities = Month Year TimeInMinutes Aboard Ground
 / VIF TOL;
run;

/* Transformations */
data transformations;
 set newdata2;
 _id_ = _n_;
```

```
 month_year = Month*Year;
 month_time = Month*TimeInMinutes;
 month_aboard = Month*Aboard;
 month_ground = Month*Ground;
 year_time = Year*TimeInMinutes;
 year_aboard = Year*Aboard;
 year_ground = Year*Ground;
 time_aboard = TimeInMinutes*Aboard;
 time_ground = TimeInMinutes*Ground;
 aboard_ground = Aboard*Ground;
 aboard2 = Aboard*Aboard;
 ground2 = Ground*Ground;
 log_aboard = log(Aboard+1);
 log_ground = log(Ground+1);
run;
proc reg data=transformations;
Stepwise: model ProportionFatalities= Month Year TimeInMinutes Aboard Ground
 month_year month_time month_aboard month_ground year_time year_aboard
 year_ground time_aboard time_ground aboard_ground aboard2 ground2
 log_aboard log_ground / selection=stepwise;
run;
quit;

/* Model Selection */
proc reg data=transformations;
Forward: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard time_aboard
aboard2 log_aboard
                                / selection=FORWARD vif tol slentry=0.1;
Backward: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard
time_aboard aboard2 log_aboard
                                / selection=B vif tol slstay=0.1;
Stepwise: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard time_aboard
aboard2 log_aboard
                                / selection=stepwise vif tol slentry=0.1 slstay=0.1;
rsquare: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard time_aboard
aboard2 log_aboard
                                / selection=rsquare vif tol;
adjrsq: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard time_aboard
aboard2 log_aboard
                                / selection=adjrsq vif tol;
cp: model ProportionFatalities = Month Year TimeInMinutes Aboard Ground year_aboard time_aboard
aboard2 log_aboard
                                / selection=cp vif tol;
run;
quit;

/* Diagnostics */
proc reg data=transformations;
 model ProportionFatalities = Year year_aboard aboard2 log_aboard;
 output out=temp student=r;
 plot ProportionFatalities*(Year year_aboard aboard2 log_aboard);
 plot student.*(Year year_aboard aboard2 log_aboard p.);
 plot student.*nqq.;
run;
```

```
/* Added variable plots */
proc reg data = transformations;
 model ProportionFatalities = Year year_aboard aboard2 log_aboard / partial;
run;

/* Check for multicolinearity */
proc reg data=transformations;
 model ProportionFatalities = Year year_aboard aboard2 log_aboard
  / vif tol;
run;
quit;

/* Final Model */
proc reg data=transformations;
 model ProportionFatalities = Year aboard2 log_aboard;
run;
```