
ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: “Big” Data Analytics
Final Report: Commercial Airlines Boarding Time Linear Regression

Chris Fu
ISyE 2028
Dr. Yao Xie
Bonus Project
November 29, 2015

One of the issues major airlines face are flight delays and flight cancellations. This causes a chain reaction as other connecting flights are delayed or cancelled. This is very problematic not only for the passengers but for the airlines as well. According to the Federal Aviation Administration, it is estimated that flight delays cost airlines \$22 billion yearly. I am working in conjunction with Aerospace Engineering Professor Dr. John-Paul Clarke and Industrial Engineering Professor Dr. David Goldsman. The question at hand is based on the time history of the boarding pass scans, how is the probability of late departure times affected. The main focus of this paper is to determine whether there is a linear relationship between:

1. Number of people on each flight and the amount of time it takes from first passenger to board the plane to the last passenger that boards the plane
2. Capacity of each flight and the amount of time it takes from first passenger to last passenger
3. Number of people on each flight and amount of time it takes from first passenger to pushback (moving the plane from passenger terminal to a runway or taxiway)
4. Capacity of each flight and the amount of time it takes from first passenger to pushback.

The data we received is from a major US airline company. It looks at a single day's worth of passengers who boarded each plane at differing times. For each entry there is:

- The time of when the passenger scanned their boarding pass and boarded the plane
- The flight number of the plane
- The type of aircraft
- The seat the individual was assigned to
- The time of the plane's pushback (when the plane leaves the passenger terminal to the runway or taxiway)

The data has quite well over 10,000 entries and to perform statistical analysis by hand would not be the most efficient way. I used Python programming that I learned over the past summer in my CS 2316 class to extract/filter out useful information. I was able to create a spreadsheet that consists of:

- The flight number of the plane
- Amount of people in each flight (1)
- The capacity that an aircraft can hold (2)
- The time that the first passenger of each flight boarded on the plane
- The time that the last passenger of each flight boarded on the plane
- The time which the plane pushback (OutTime)
- Time difference between the last passenger and first passenger that boarded the flight (3)
- Time difference between pushback and the first passenger (4)

With the extracting of useful information complete, I began the analysis process. From the guidance of Dr. Clarke and Dr. Goldsman, I used Linear Regression between:

- (4) on (1)
- (3) on (1)
- (4) on (2)
- (3) on (2)

Before we continue with analysis, here are some terms we need to understand:

- R^2 : Coefficient of determination. A Statistic that will give some information about the goodness of fit of a model. How good is the fit?
- Multiple R: Correlation coefficient between Y variable and X variable
- ANOVA: Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures. This is very important as it helps us in understanding about the regression
- Significance F: significance of the f-test. The lower the value, the better the regression analysis
- P-value: the lower the value, the better. Ex. Intercept is 95% correct if p-value is 5%

After looking at the linear regression model, there were some data points (outliers) that distracted my results. These data points were events where a particular plane remained at the passenger terminal for over 24 hours until pushback. This made it difficult to determine whether a different regression analysis should be used given the characteristics of the model. I then removed the outliers and created the linear regression model once more. With the outliers removed, the R^2 improved drastically from 0.00066 in Diagram (1) to 0.1512 in Diagram (5). With R^2 , this means that we are able to explain 15% of variability in our Y variable (Difference between pushback and first passenger) from our X variable (amount of people in flight). The significance F turned out to be 1.53×10^{-30} . This is good as the linear regression model is good for these two variables. The p-value for the intercept is .1239 and for the slope is 1.53×10^{-30} . While the p value for the intercept is rather high, the slope is pretty good. With all the data given, it is safe to assume that the (amount of people in flight) and the (difference between pushback and first passenger) have a linear relationship.

More Analysis:

Amount of people in flight VS Difference between pushback and first passenger

R^2	Significance F	P-value for intercept	P-value for slope
0.151269334	1.53442E-30	0.123934161	1.53442E-30

Capacity of flight VS Difference between pushback and first passenger

R^2	Significance F	P-value for intercept	P-value for slope
0.137848672	8.87692E-28	0.037560398	8.87692E-28

Amount of people in flight VS Difference between last passenger and first passenger

R^2	Significance F	P-value for intercept	P-value for slope
0.005271811	0.039195071	2.19199E-12	0.039195071

Capacity of flight VS Difference between last passenger and first passenger

R^2	Significance F	P-value for intercept	P-value for slope
0.001635369	0.251179442	3.97304E-07	0.251179442

From all four relationships, we see that our R^2 are all under .20. Having a low R^2 does not mean it is bad. For this project, it was actually expected that the coefficient of determination was going to be low. Any field that

attempts to predict human behavior will bound to have a R^2 that is lower than 50%. This is simply because humans are more difficult to predict than physical/mechanical processes.

Despite the low R^2 , we can draw important conclusions from the significance F. We see that for the first two relationships, the significance F was extremely low. The latter two relationships end up with values still under 1.

Conclusion

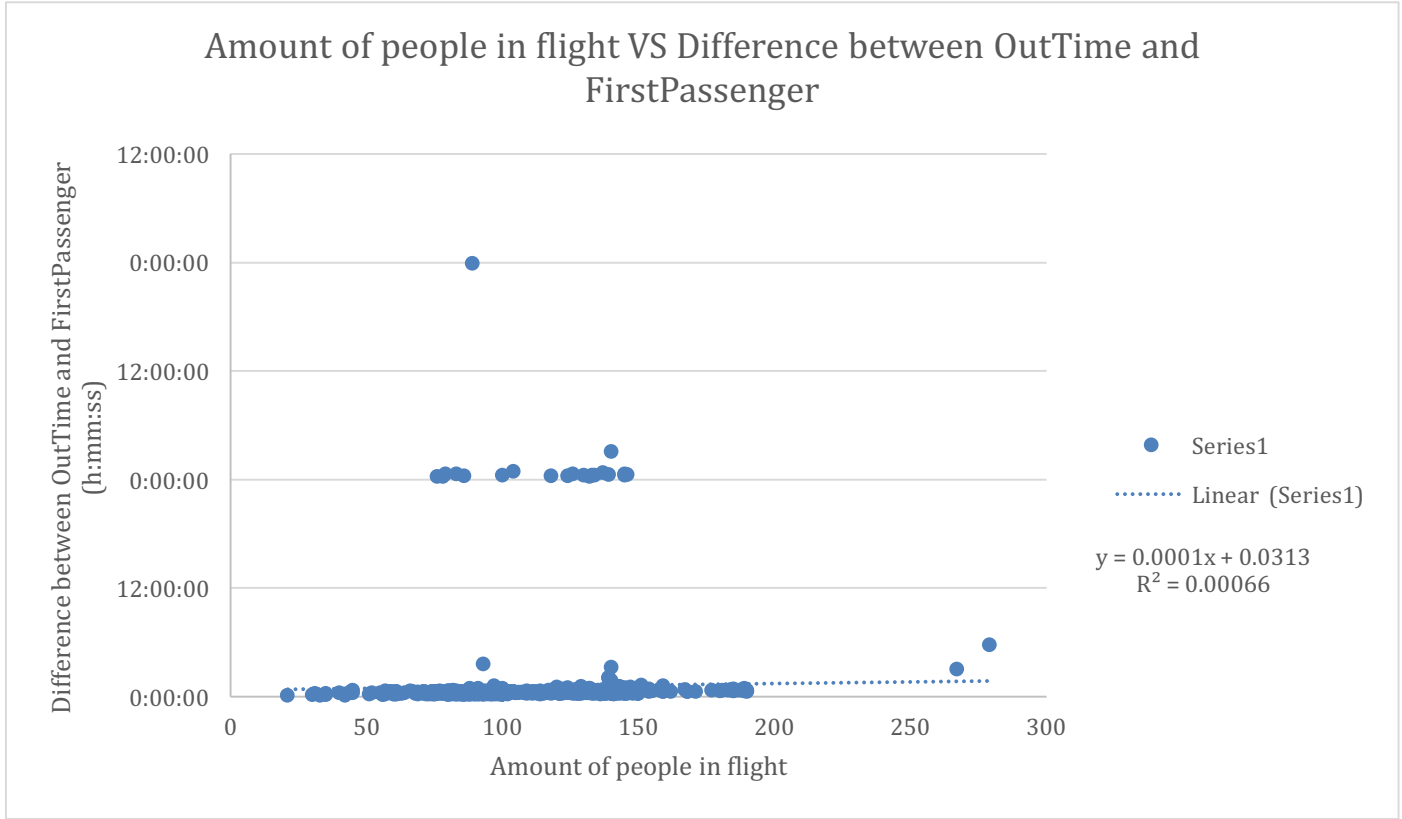
The four relationships we analyzed on first glance at the data appears to be related in a linear fashion. In regards to the 'Capacity of flight VS Difference between pushback and first passenger' and the 'Capacity of flight VS Difference between last passenger and first passenger', determining whether it is related linearly would need to be analyzed further. The graphs of those two relationships are rather awkward and do not follow a normal linear graph. As for 'Amount of people in flight VS Difference between pushback and first passenger' and 'Amount of people in flight VS Difference between last passenger and first passenger', by looking at the graph and the data, I am fairly certain that these two relationships follow a linear trend. That as the amount of people in a flight increases, the 'difference between pushback and first passenger' and 'difference between last passenger and first passenger' will increase as well.

For future analysis, other regression models may be worthwhile to examine in the future. Multiple Linear Regression of:

- 'Difference between last passenger and first passenger' on 'Amount of people on flight' AND 'Capacity of flight'
- 'Difference between pushback and first passenger' on 'Amount of people on flight' AND 'Capacity of flight'

will also be worth looking at to see if the two independent variables (explanatory variables) affect the dependent variable (response variable).

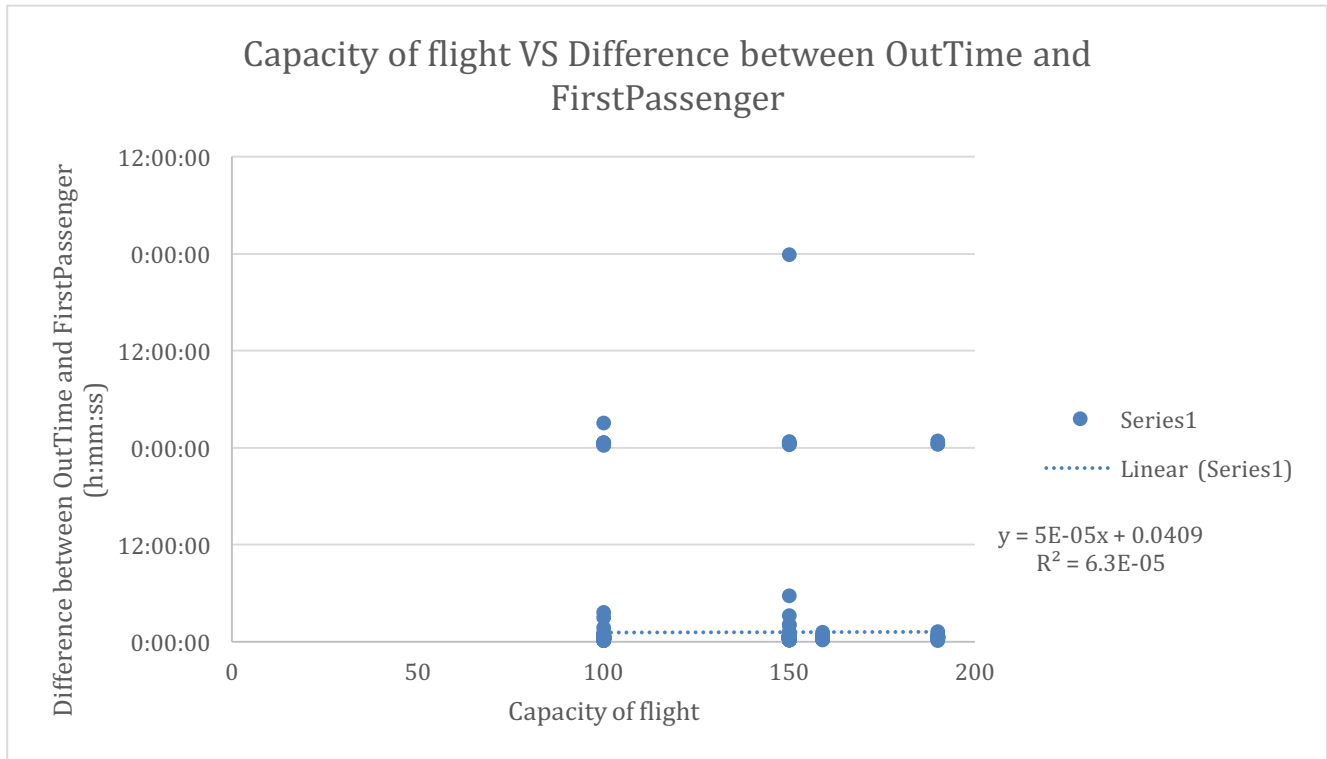
Appendix



Amount of people in flight VS Difference between pushback and first passenger
Diagram (1)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.025660132							
R Square	0.000658442							
Adjusted R Square	-0.000551414							
Standard Error	0.169306491							
Observations	828							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.015600234	0.015600234	0.54423176	0.460893738			
Residual	826	23.67703234	0.028664688					
Total	827	23.69263257						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.031306336	0.022978271	1.362432207	0.173432891	-0.013796336	0.076409008	-0.013796336	0.076409008
X Variable 1	0.000141417	0.000191694	0.737720652	0.460893738	-0.000234848	0.000517681	-0.000234848	0.000517681

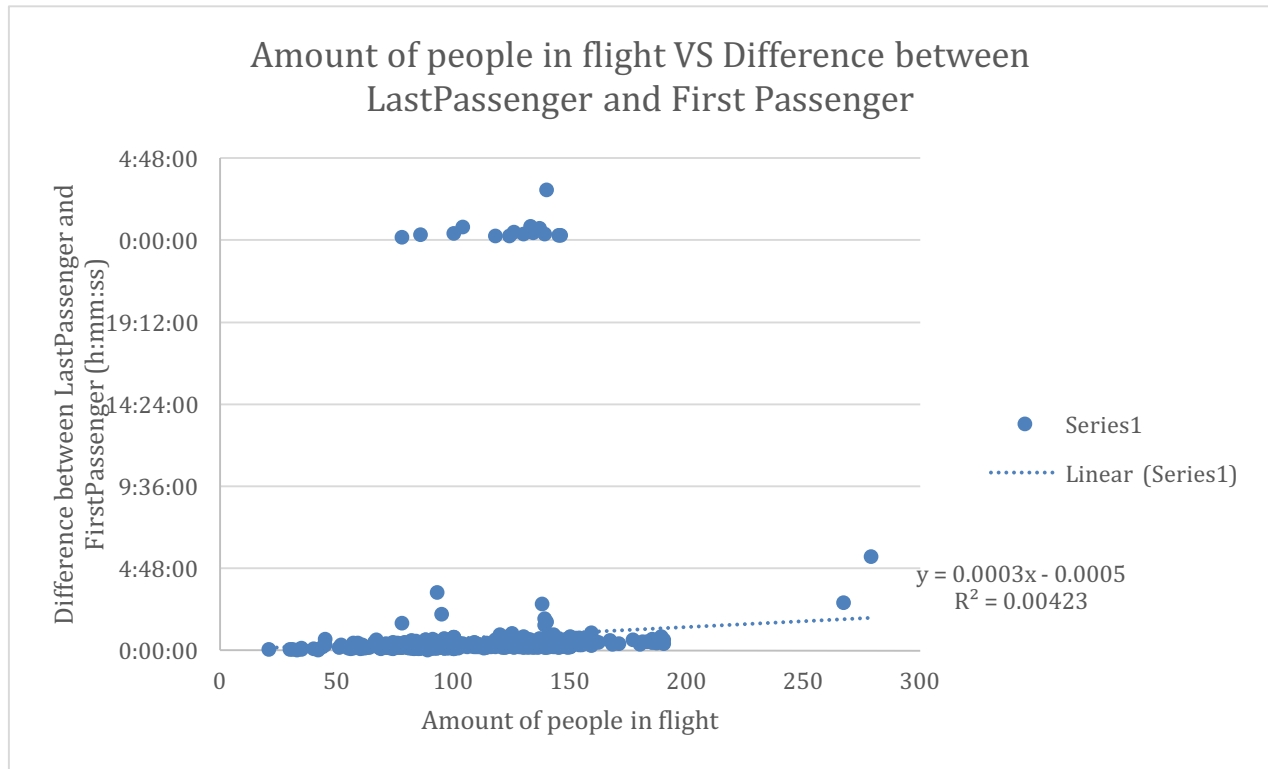
Corresponding Linear Regression summary output



**Capacity of flight VS Difference between pushback and first passenger
Diagram (2)**

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.065039559							
R Square	0.004230144							
Adjusted R Square	0.003024612							
Standard Error	0.135179488							
Observations	828							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.064120638	0.064120638	3.508942445	0.061391359			
Residual	826	15.09390592	0.018273494					
Total	827	15.15802656						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-0.000547203	0.018346555	-0.029825908	0.976213101	-0.036558557	0.035464152	-0.036558557	0.035464152
X Variable 1	0.000286704	0.000153054	1.873217138	0.061391359	-1.37172E-05	0.000587125	-1.37172E-05	0.000587125

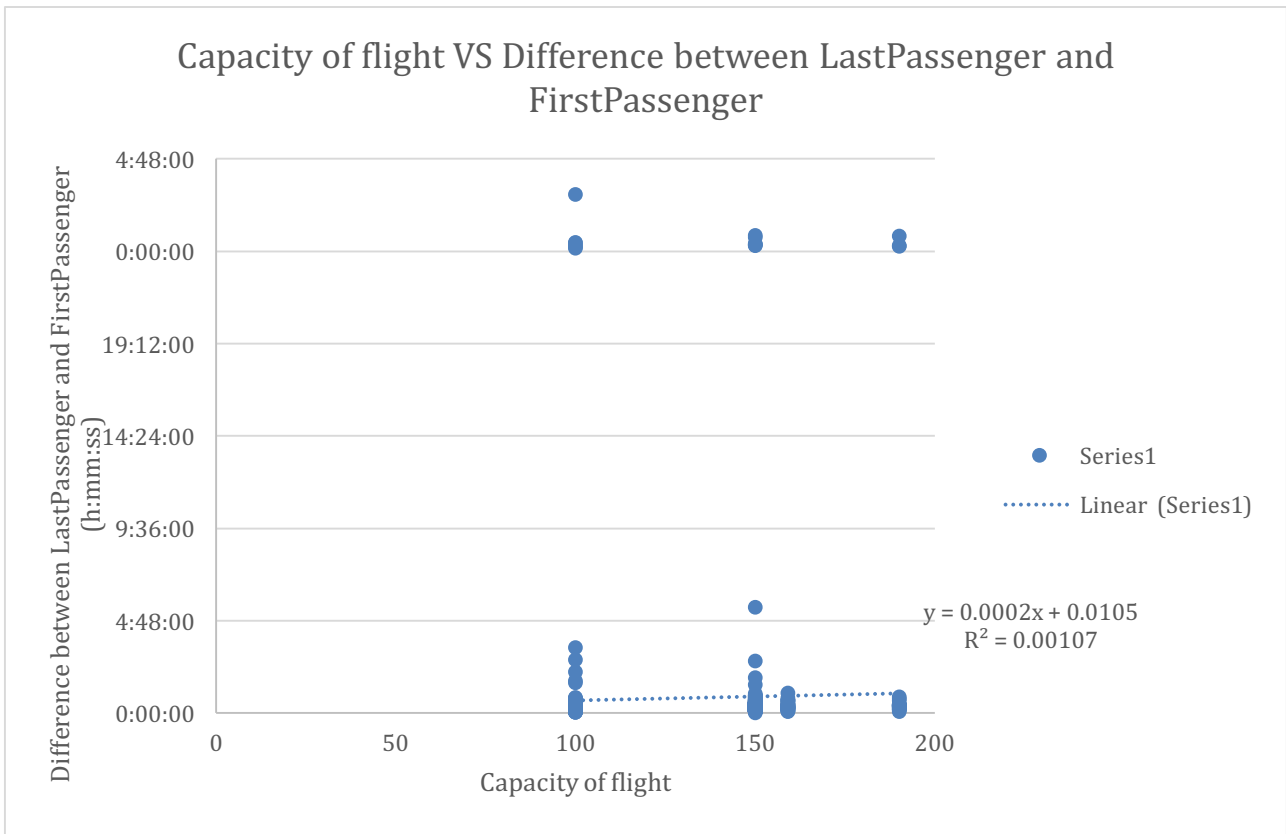
Corresponding Linear Regression summary output



Amount of people in flight VS Difference between last passenger and first passenger
Diagram (3)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.007961471							
R Square	6.3385E-05							
Adjusted R Square	-0.001147192							
Standard Error	0.169356891							
Observations	828							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.001501758	0.001501758	0.052359346	0.819064182			
Residual	826	23.69113081	0.028681756					
Total	827	23.69263257						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.04094879	0.030054409	1.362488596	0.173415113	-0.01804321	0.09994079	-0.01804321	0.09994079
X Variable 1	5.10363E-05	0.00022304	0.228821646	0.819064182	-0.000386755	0.000488828	-0.000386755	0.000488828

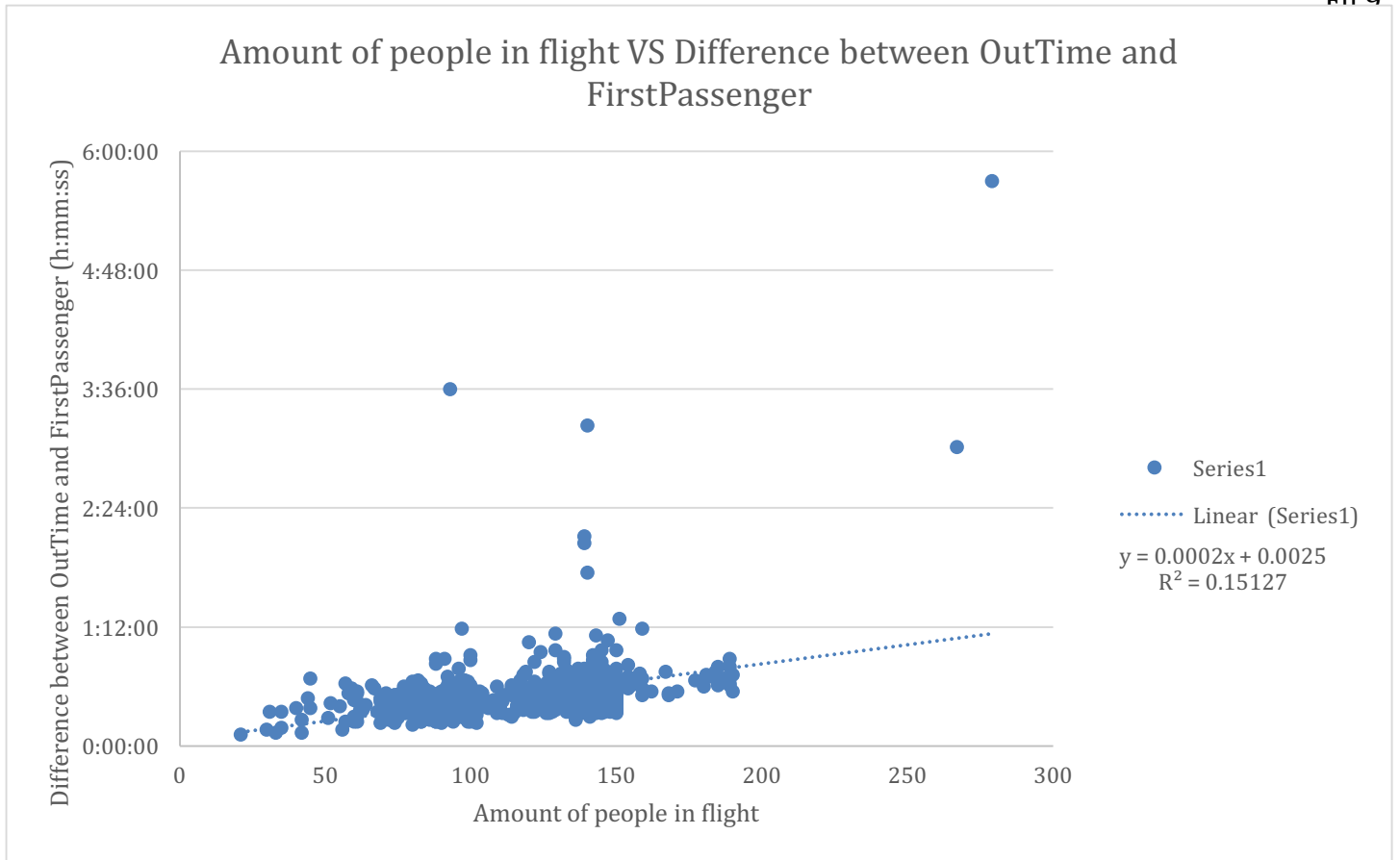
Corresponding Linear Regression summary output



Capacity of flight VS Difference between last passenger and first passenger
Diagram (4)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.03268425							
R Square	0.00106826							
Adjusted R Square	-0.0001411							
Standard Error	0.135393936							
Observations	828							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.016192716	0.016192716	0.88332653	0.347566943			
Residual	826	15.14183384	0.018331518					
Total	827	15.15802656						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.01052921	0.024027276	0.438219028	0.661342095	-0.036632492	0.057690911	-0.036632492	0.057690911
X Variable 1	0.000167587	0.000178311	0.939854526	0.347566943	-0.00018241	0.000517583	-0.00018241	0.000517583

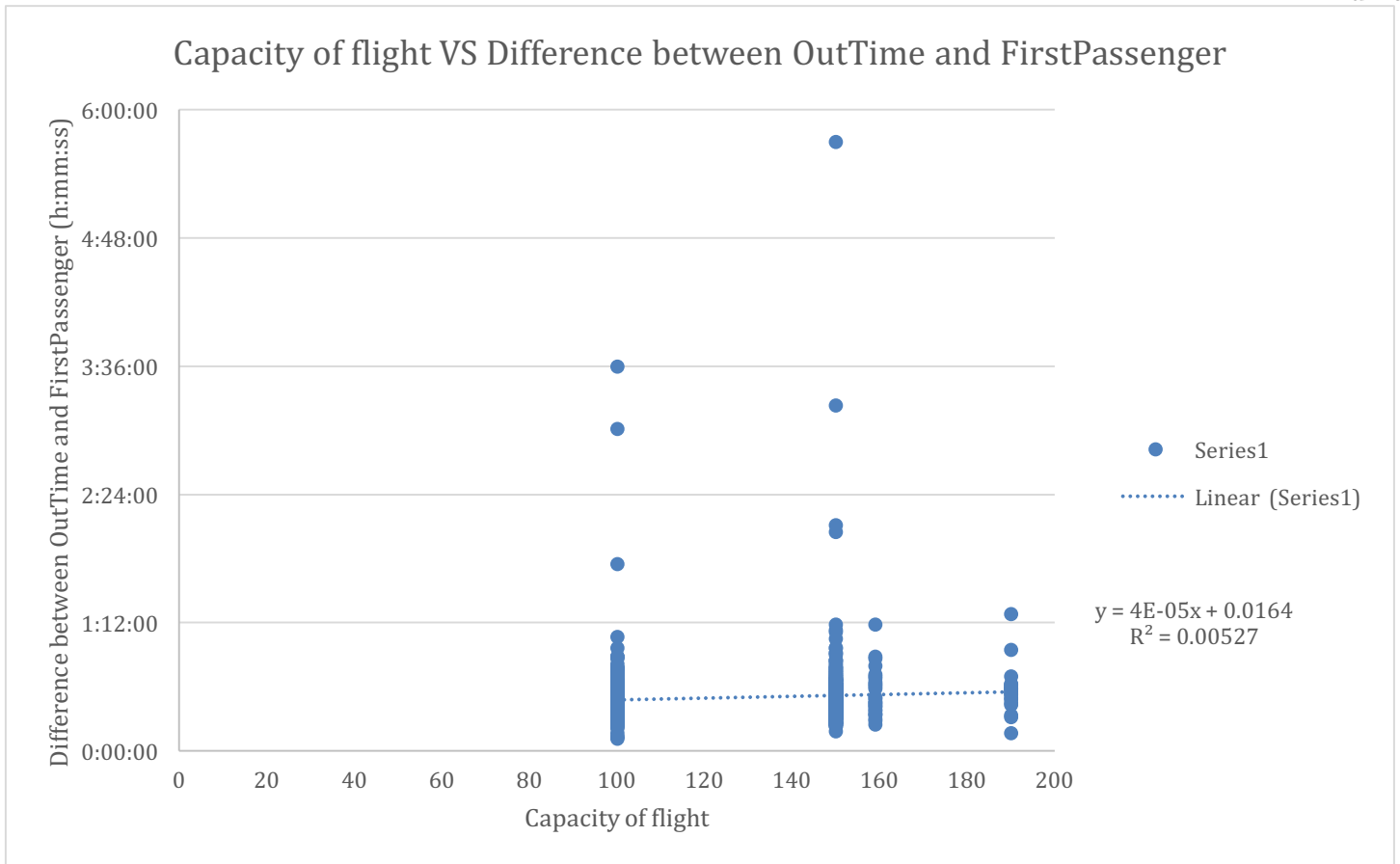
Corresponding Linear Regression summary output



Amount of people in flight VS Difference between pushback and first passenger
Removed outliers
 Diagram (5)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.388933585							
R Square	0.151269334							
Adjusted R Square	0.15021501							
Standard Error	0.011701519							
Observations	807							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.019645422	0.019645422	143.4752137	1.53442E-30			
Residual	805	0.110225063	0.000136926					
Total	806	0.129870485						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.002466784	0.001601726	1.540078623	0.123934161	-0.000677268	0.005610836	-0.000677268	0.005610836
X Variable 1	0.000160024	1.33597E-05	11.97811395	1.53442E-30	0.0001338	0.000186248	0.0001338	0.000186248

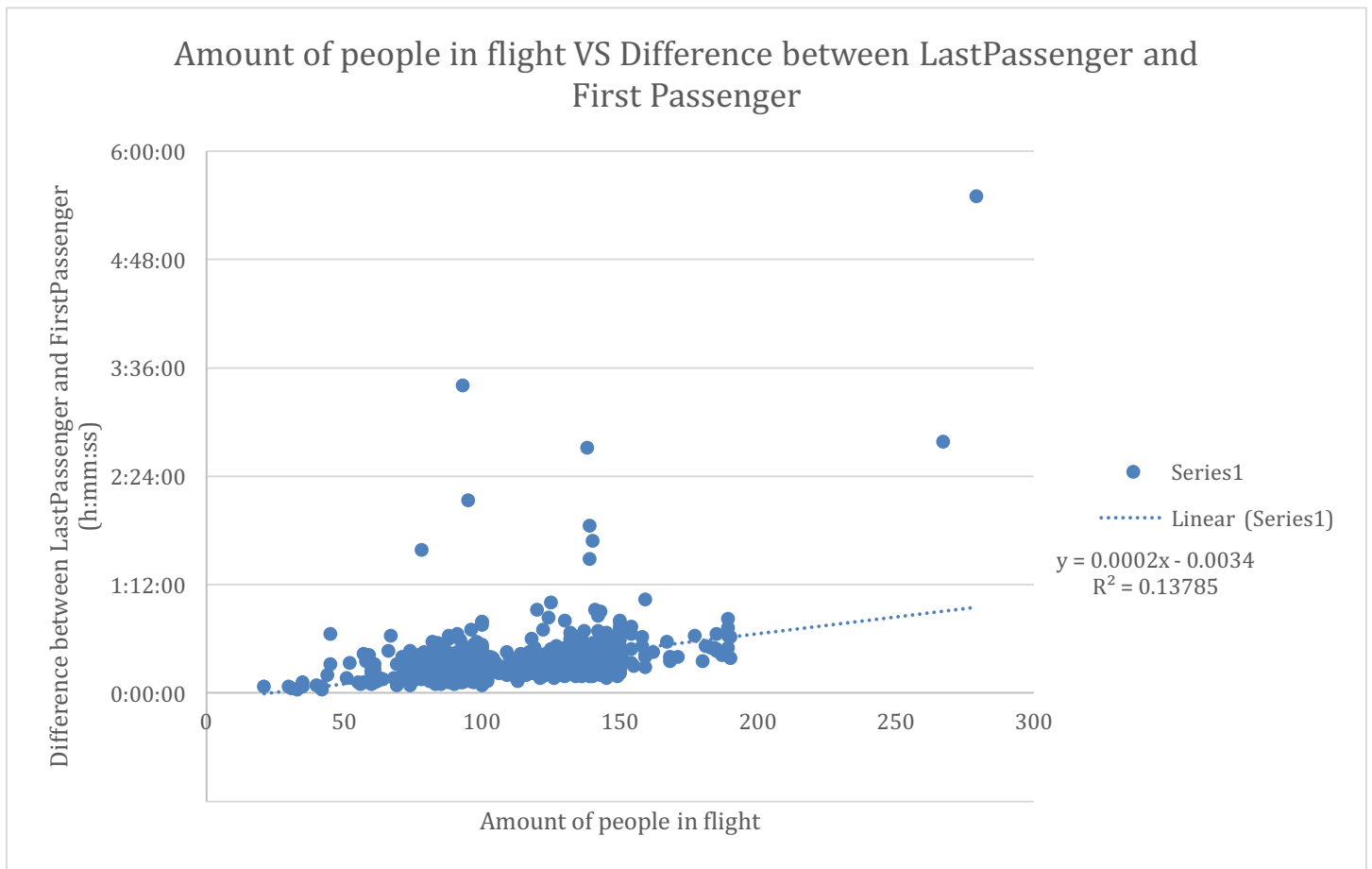
Corresponding Linear Regression summary output



Capacity of flight VS Difference between pushback and first passenger
Removed outliers
Diagram (6)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.371279776							
R Square	0.137848672							
Adjusted R Square	0.136777677							
Standard Error	0.011860474							
Observations	807							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.018105857	0.018105857	128.7107928	8.87692E-28			
Residual	805	0.113240036	0.000140671					
Total	806	0.131345892						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-0.003381822	0.001623484	-2.083064592	0.037560398	-0.006568583	-0.00019506	-0.006568583	-0.00019506
X Variable 1	0.000153626	1.35412E-05	11.34507791	8.87692E-28	0.000127046	0.000180206	0.000127046	0.000180206

Corresponding Linear Regression summary output

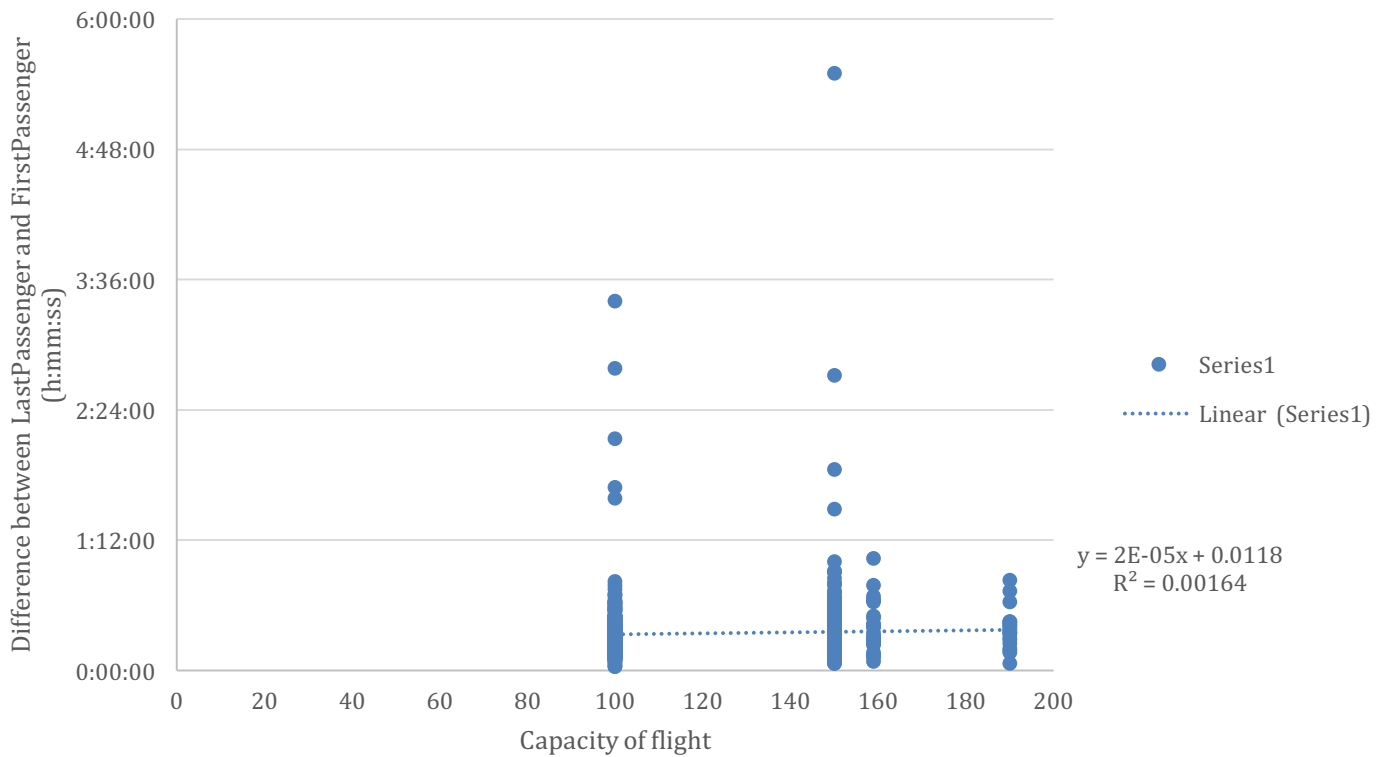


Amount of people in flight VS Difference between last passenger and first passenger
Removed outliers
 Diagram (7)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.072607238							
R Square	0.005271811							
Adjusted R Square	0.004036124							
Standard Error	0.012668042							
Observations	807							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.000684653	0.000684653	4.266298982	0.039195071			
Residual	805	0.129185832	0.000160479					
Total	806	0.129870485						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.016360083	0.002293621	7.132862243	2.19199E-12	0.011857899	0.020862267	0.011857899	0.020862267
X Variable 1	3.51662E-05	1.70255E-05	2.065502114	0.039195071	1.74659E-06	6.85857E-05	1.74659E-06	6.85857E-05

Corresponding Linear Regression summary output

Capacity of flight VS Difference between LastPassenger and FirstPassenger



Capacity of flight VS Difference between last passenger and first passenger
Removed outliers
 Diagram (8)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.040439697							
R Square	0.001635369							
Adjusted R Square	0.000395165							
Standard Error	0.012763063							
Observations	807							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	0.000214799	0.000214799	1.31862855	0.251179442			
Residual	805	0.131131093	0.000162896					
Total	806	0.131345892						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.011814332	0.002310825	5.11260333	3.97304E-07	0.007278378	0.01635029	0.007278378	0.016350286
X Variable 1	1.96973E-05	1.71532E-05	1.148315527	0.251179442	-1.3973E-05	5.3368E-05	-1.3973E-05	5.33675E-05

Corresponding Linear Regression summary output

References

<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>

<https://www.youtube.com/watch?v=Cltt47Ah3Q4>

<https://www.youtube.com/watch?v=zYm6koNvAts>