Final Paper

Cecilia Strombeck

November 2014

GA Data Science

**Do the existence of baggage fees delay flights?**

**Problem Statement & Hypothesis**

The aim of this project was to determine whether there is a statistically significant difference in actual flight time performance between airlines with versus without baggage fees.

The hypothesis is that airlines that charge a fee for checking baggage will have longer flight times.

The reasoning being that when airlines charge passengers for checking bags, passengers are more likely to bring their luggage on board as a carry-on, which leads to added difficulties in finding space in the overhead compartments, prolonging boarding times and delaying flights. [[1]](#footnote--1) For anyone who travels on domestic commercial airlines, they will likely have experienced the inconvenience of having to find space for their carry-on items in the packed over-head bins and the time spent by you, fellow passengers and flight attendant staff on puzzling in bags in the limited space.[[2]](#footnote-0) In recent years, the airline industry has introduced and raised baggage fees as a way to operate with a profit, which as a result has led more passengers to bring carry-ons instead of checking baggage.

**Data**

***Flight Data***

Data are retrieved from the Bureau of Transportation Statistics (BTS) and are called ‘On-Time Performance’. Monthly data sets are available for download in csv files, with the most recently available data are for August 2015.[[3]](#footnote-1) I used data for the 12-month-period June 2014-May 2015. JetBlue introduced baggage fees in June 2014, and in order to have as many observations as possible of flights without baggage fees through the entire time period, I chose a 12-month-period before this change was introduced.

Filename(s): On\_Time\_On\_Time\_Performance\_*Year*\_*Month* eg*.*

On\_Time\_On\_Time\_Performance\_2015\_8

Data Dictionaries: <http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time>

Readme file (see folder)

http://www.transtats.bts.gov/glossary.asp

***Airline names***

The ‘On-Time Performance’ datasets from BTS only included *UniqueCarrier*, *AirlineID* and *Carrier*, without the full airline name, but BTS also had a separate lookup table that I downloaded in order to get the full airline names. The variable added was *Description* and it was merged using *AirlineID*.

Filename: L\_AIRLINE\_ID.csv

Source: http://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID=236&DB\_Short\_Name=On-Time

***Airport coordinates***  
Data with longitude and latitude information for each airport was obtained from a datahub page. The longitude and latitude variables were added for both the airport of origin and destination airport as *lon\_origin*, *lat\_origin*, *lon\_dest,* and *lat\_dest.* These were added for the calculation of target time, explained below.

Filename: global\_airports.csv

Source & Data Dictionary: http://datahub.io/dataset/global-airports.

***Bag Fees***

This information was manually entered into a csv spreadsheet for the airlines of interest using aggregated information provided on various websites.[[4]](#footnote-2) When I came across discrepancies across sources, I went directly to the airline’s website to verify the fee. In this process I discovered that JetBlue started charging for the first checked bag in June 2015.[[5]](#footnote-3)

**Data Exploration & Processing**

The full data set of the 12 monthly files combined contained more than 6 million observations.

***Times***

I investigated the meaning of the different variables for departure, arrival and duration of flight. I discovered that there was no variable for the scheduled departure time, but only the actual departure time (*DepTime*), as measured at the time of departure from the origin airport gate. That being said, I discovered that this time could be inferred by adding the variable *DepDelay* to *DepTime*, as *DepDelay* is described in the glossary as “The difference between the schedule departure time and the actual departure time from the origin airport gate.” For example, if the departure time (*DepTime*) was 9:55AM, and *DepDelay* was 5, this means that the scheduled departure time was 9:50AM. Note that there was also a variable (*DepDelayMinutes*) that only accounted for positive values of delays, wheras *DepDelay* could be negative in the cases when the flight left the gate before the scheduled departure time. See Sensitivities below.

The Actual Travel Time was calculated as the difference between the scheduled departure and the arrival at the gate at the destination airport, although I ran a sensitivity calculating it differently (see below).

Target Time is calculated based on the distance between the two airports and the distance traveled in the east-west direction (or reverse) to account for the jet stream, plus an additional 43.2 minutes to account for taxiing at each airport. The following formula (adopted from FiveThirtyEight) was used:

.117\*distance + .517\*(lon\_origin-lon\_dest) + 43.2

In the BTS data they count delays based on any arrival or departure times that deviate from the scheduled times. So if Airline A and B fly the same route on the same time, but if Airline B had scheduled the route to take 10 minutes longer, that could lead the Airline A being counted as being ‘on-time,’ whereas Airline B could be ‘delayed.’ This analysis is performed independent from the scheduled arrival time, in order to not favor airlines that might ‘pad’ their schedules.

I calculated delays based on the difference between Actual Travel Time and Target Time. This time difference ranged from -53 minutes, meaning that a flight was 53 minutes faster than the calculated Target Time for that route. On the opposite side of the spectrum, one flight from SFO-JFK was +1770, in other words 30 hours behind target. Note that I removed observations where the flight was cancelled or diverted.

***Correct airlines***

The data set only includes the 14 U.S. air carriers that have at least 1 percent of total domestic scheduled-service passenger revenues. I do not consider that to be a notable limitation of the analysis.

In some cases multiple air carriers (from the same airline alliance) may share the same aircraft but name it differently. Based on some brief research, it appears that this practice has largely been discontinued in the US. Regardless, for the main purpose of this analysis, the only two airlines that do not have checked baggage fees this practice does not exist since they are not part of any airline alliances.

A major obstacle was the fact that there are some regional air carriers that operate flights, even though, for all relevant purposes for the passenger it is actually recognized as one of the familiar airlines. For example, you could be taking a Delta flight (that is even painted in the Delta colors etc), but in the data it would appear as an Envoy Air flight. [[6]](#footnote-4) Fortunately, FiveThirtyEight had research how to map these flights to the ‘familiar airlines’ and I adopted their methodology.

***Limiting data set***

To simplify the analysis, I only looked at flights on the 20 most common routes*.* This made it less computationally intensive, since the full data set included 6 million observations.

As mentioned above, I also removed observations for flights that were cancelled or diverted.

**Modeling**

Null hypothesis: There is no relationship between the existence of baggage fees and an airline’s actual flight time performance.

I used a multiple linear regression model for this analysis, which was run in two iterations, first without the baggage fee dummy and then with.

The outcome was a continuous variable of the difference between the target time and the actual travel time.

The dependent variables were dummies for airport of origin, destination airport, airlines and the existence of a baggage fee for the 1st checked bag.

The baggage fees were mostly in the range of $25-$35 dollars, and therefore I did not think it would add much value to treat it as a continuous variable. Note that Spirit also charges for carry-on bags, which therefore may not give passengers as strong incentive to bring a bag on board instead of checking in.

***Sensitivities***

The calculation of actual travel time, is taken as the difference between scheduled departure time and the actual arrival time.

1) An alternative approach I took was to calculate travel time using the actual departure time, if the departure happens before the scheduled departure time. The calculation of this is done by using *DepDelayMinutes* instead of *DepDelay* in the calculation of actual travel time. This alternative approach would make sense if you assume that passengers do not get the same amount of utility from the minutes arrived early as the utility lost from the minutes arrived late. If someone is late they may miss scheduled appointments, whereas if you arrive early you were prevented from scheduling anything during this time since it was unanticipated (for better or worse).

2) The second alternative approach I took was to use *DepDelayMinute* as the dependent variable.

**Possible Next Steps**

***Interpret variables differently***

In the current regression model, the outcome is a continuous variable as the difference between the target time and the actual travel time. First of all, this could be set up such that it did not allow for negative values, in other words, only counting flights as being *on target* or being slower than target. Secondly, this could be changed to being a binary variable by setting a threshold where a flight is considered delayed or not, which could be set at a value greater than 0. For example, if the time difference was greater than 10 minutes it would be counted as a delay, whereas otherwise it would not. Under this alteration the linear regression model would have to be changed to a logistic regression model.

Currently, the baggage fee is set as a dummy variable for whether or not the airline has a fee for the first checked bag for passengers traveling in economy class. Alternatively, this could be changed to being a continuous variable using the price of the first checked bag or even accounting for the price of the second checked bag, either in the same variable or in a separate continuous variable altogether.

***Limit scope of flights***

Limit the routes based on nature of travel. If it were possible to infer what types of travel passenger made on specific routes/travel times this could be associated with more/less bags being brought on board. For example, passengers traveling for vacation may be more likely to bring sizable carry-on items than business travelers doing a same-day-return trip BOS-NY for a meeting.

Checking if results were different if adding load factor (the % seats filled on-board) and filtering results on that. For example, one could only look at flights that had a load factor >95%. It would make sense that loading the over-head compartment would take longer time when the flight is close-to or fully loaded.

***Adding Additional Features***

The load factor could be added as an explanatory variable. The use of other features could also be explores to obtain a higher explanatory power of the model.

**Conclusion & Key Takeaways**

Based on the approach I took, in running monthly as well as a full-year multiple linear regression, I cannot reject the null hypothesis that there is no relationship between the existence of baggage fees and flight delays.

This would have been valuable information to know since it could serve as an added reason for passengers to elect travel with airlines that do *not* charge baggage fees.

1. http://usatoday30.usatoday.com/travel/flights/2008-09-15-carry-on-bags-fees\_N.htm [↑](#footnote-ref--1)
2. http://www.wsj.com/articles/what-are-the-odds-your-carry-on-wont-fit-1435167020 [↑](#footnote-ref-0)
3. As of November 9, 2015. [↑](#footnote-ref-1)
4. <http://www.tripadvisor.com/AirlineFees>

   <http://www.airfarewatchdog.com/blog/3801089/airline-baggage-fees-chart/>

   <http://www.seatguru.com/airlines/Spirit_Airlines/baggage.php> [↑](#footnote-ref-2)
5. See <http://www.nbcnews.com/news/us-news/jetblue-starts-charging-25-first-checked-bag-n384446> [↑](#footnote-ref-3)
6. The largest three regional carries are Envoy Air, ExpressJet Airlines and SkyWest Airlines. [↑](#footnote-ref-4)