# NYCFlights: Arrival Delay Logictic Model

*Abhishek Chaturvedi*

*04/30/2017*

# Logsitic and Inverse Logistic Transformation

- Write an R function for the logistic function. The function should accept a `numeric` vector with values `[-Inf,Inf]` and produce a numeric vector in the the range `[0,1]`.
- Plot the logistic function from `[-10,10]`
- Write a R function for the inverse logistic function. The function should accept a `numeric` vector with values `[0,1]` and prodcuce a numeric vector in the range `[-Inf,Inf]`
- Plot the Inverse Logistic function from `[0,1]`

**Hint:** For plotting curves see `?graphics::curve` or `?ggplot2::stat_function`

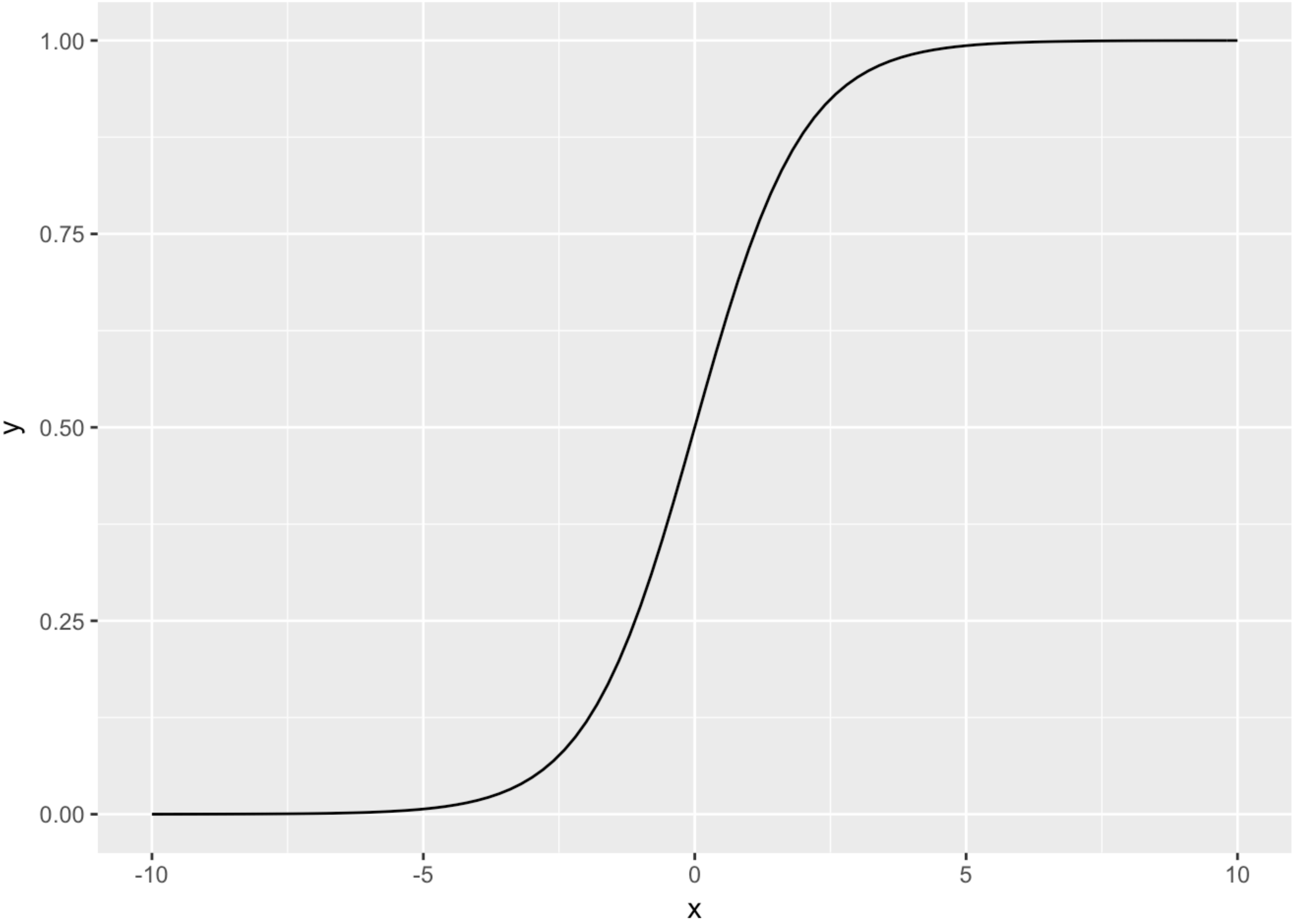
```
library("ggplot2")

myLogistic <- function(x) {
  return(1/(1+exp(-x)))
}

x1 <- c(-10:10)
myLogistic(x1)
```

```
##  [1] 4.539787e-05 1.233946e-04 3.353501e-04 9.110512e-04 2.472623e-03
##  [6] 6.692851e-03 1.798621e-02 4.742587e-02 1.192029e-01 2.689414e-01
## [11] 5.000000e-01 7.310586e-01 8.807971e-01 9.525741e-01 9.820138e-01
## [16] 9.933071e-01 9.975274e-01 9.990889e-01 9.996646e-01 9.998766e-01
## [21] 9.999546e-01
```

```
ggplot(data.frame(x = x1) , aes(x)) +
  stat_function(fun = myLogistic, geom = "line")
```

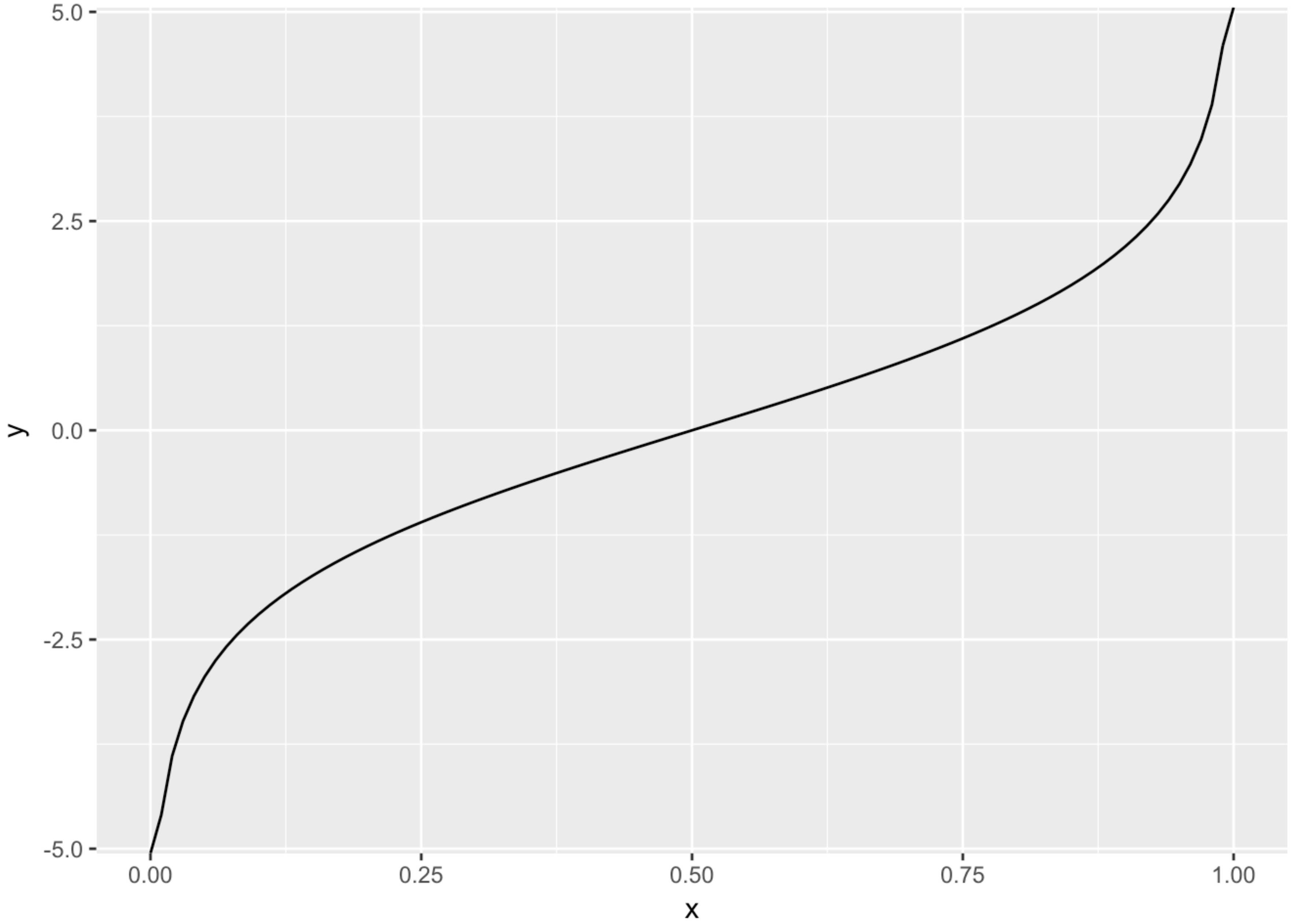


```
myInverseLogistic <- function(x){
  return(log(x/(1-x)))
}

x2 <- c(0,0.05,0.1,0.2,0.3,0.6,0.8,0.9,0.95,1)
myInverseLogistic(x2)
```

```
##  [1]       -Inf -2.9444390 -2.1972246 -1.3862944 -0.8472979  0.4054651
##  [7]  1.3862944  2.1972246  2.9444390        Inf
```

```
ggplot(data.frame(x = x2) , aes(x)) +
  stat_function(fun = myInverseLogistic, geom = "line")
```



# NYCFlights Model

Using the rectangular data that you created from the earlier assignment and following theexample from the text and class, create a model for arr_delay >= 22 minutes. Describe/Explain each of the steps and show all work.

KNIT YOUR DOCUMENT AS *HTML* AND SUBMIT IT AND THE `Rmd` file to your repository.

```r
library('data.table')

flightsData <- fread("data/flights.csv")
planesData <- fread("data/planes.csv")
airportsData <- fread("data/airports.csv")
weatherData <- fread("data/weather.csv")

flightsPlanesData <- merge(flightsData,planesData, by.x="tailnum", by.y="tailnum" , all.x = TRUE, suffixes = c(".flights", ".planes"))

flightsPlanesOriginData <- merge(flightsPlanesData,airportsData, by.x = "origin", by.y = "faa", all.x = TRUE, suffixes = c(".flights", ".origin"))

flightsPlanesOriginDestinationData <- merge(flightsPlanesOriginData,airportsData, by.x = "dest", by.y = "faa", all.x = TRUE, suffixes = c(".origin", ".dest"))

flightsPlanesOriginDestinationWeatherData <- merge(flightsPlanesOriginDestinationData,weatherData, by.x = c("year.flights", "month", "day", "origin", "time_hour"), by.y = c("year", "month", "day", "origin", "time_hour"), all.x = TRUE, suffixes = c(".flights", ".weather"), allow.cartesian=FALSE)

flights_data_joined <- flightsPlanesOriginDestinationWeatherData[sample(.N,50000)]

# my old model consists of arr_delay,humid,dep_time,sched_dep_time,sched_arr_time,dep_delay,origin

y <- "arr_delay"
xs <- c('humid','dep_time', 'sched_dep_time','sched_arr_time','dep_delay','origin')

yx <- flights_data_joined[,c(y,xs), with=FALSE]
model1 <- lm(arr_delay ~ ., data=yx)
summary(model1)
```

```
##
## Call:
## lm(formula = arr_delay ~ ., data = yx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.119 -10.359  -1.735   8.358 132.335
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.079390   0.528221 -15.295  < 2e-16 ***
## humid           0.105557   0.005004  21.094  < 2e-16 ***
## dep_time        0.013881   0.005893   2.355  0.01851 *
## sched_dep_time  0.008887   0.005901   1.506  0.13206
## sched_arr_time -0.021751   0.001169 -18.613  < 2e-16 ***
## dep_delay       1.007193   0.010404  96.813  < 2e-16 ***
## originJFK       0.721309   0.234173   3.080  0.00207 **
## originLGA       1.838268   0.227083   8.095 5.92e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.64 on 31370 degrees of freedom
##   (18622 observations deleted due to missingness)
## Multiple R-squared:  0.8118, Adjusted R-squared:  0.8118
## F-statistic: 1.933e+04 on 7 and 31370 DF,  p-value: < 2.2e-16
```

```
# modifying data to create a new column 'gt22' for flights more than 22 hours delay
yx <- within(yx, gt22 <- arr_delay>=22)

model2 <- glm(formula = gt22 ~ . - arr_delay, family=binomial, data=yx)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = gt22 ~ . - arr_delay, family = binomial, data = yx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1992  -0.3127  -0.2391  -0.1837   3.1684
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.439e+00  1.402e-01 -31.656  < 2e-16 ***
## humid           1.863e-02  1.256e-03  14.832  < 2e-16 ***
## dep_time        3.500e-05  1.494e-03   0.023    0.981
## sched_dep_time  2.863e-04  1.497e-03   0.191    0.848
## sched_arr_time -6.545e-05  3.020e-04  -0.217    0.828
## dep_delay       1.106e-01  3.127e-03  35.378  < 2e-16 ***
## originJFK       6.084e-02  6.113e-02   0.995    0.320
## originLGA       3.141e-01  5.923e-02   5.302 1.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26337  on 31377  degrees of freedom
## Residual deviance: 12438  on 31370  degrees of freedom
##   (18622 observations deleted due to missingness)
## AIC: 12454
##
## Number of Fisher Scoring iterations: 7
```

```
# naive model to compare with above model

naiveModel <- glm(formula = gt22 ~ 1, family=binomial, data=yx)

summary(naiveModel)
```

```
## 
## Call:
## glm(formula = gt22 ~ 1, family = binomial, data = yx)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6602  -0.6602  -0.6602  -0.6602   1.8058  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.41258    0.01143  -123.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 48071  on 48598  degrees of freedom
## Residual deviance: 48071  on 48598  degrees of freedom
##   (1401 observations deleted due to missingness)
## AIC: 48073
## 
## Number of Fisher Scoring iterations: 4
```

# Question:

Is this a good model?

AIC has improved compare to the naive model which suggests its a good model

# PART B:

Your model should be good at explaining tardiness. Now, assume that your job is to predict arrival delays a month in advance. You can no longer use all the features in your model. Retrain your model using only features that will be *known* only a month in advance of the departure time. Show all steps as above.

# Answer:

## my old model consists of

arr_delay,humid,dep_time,sched_dep_time,sched_arr_time,dep_delay,origin

out of this humid, dep_delay will not be available so the new model will be

```
y <- "arr_delay"
xs <- c('dep_time', 'sched_dep_time','sched_arr_time','origin')

yx <- flights_data_joined[,c(y,xs), with=FALSE]

model3 <- lm(arr_delay ~ . , data=yx)

summary(model3)
```

```
##
## Call:
## lm(formula = arr_delay ~ ., data = yx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -90.66  -22.22   -9.12    7.98 1148.98
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.192e+01  6.980e-01 -17.081  < 2e-16 ***
## dep_time        7.144e-02  1.383e-03  51.676  < 2e-16 ***
## sched_dep_time -4.905e-02  1.443e-03 -33.984  < 2e-16 ***
## sched_arr_time -6.672e-03  6.503e-04 -10.261  < 2e-16 ***
## originJFK      -3.309e+00  4.746e-01  -6.973 3.15e-12 ***
## originLGA      -1.447e+00  4.828e-01  -2.997  0.00273 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.27 on 48593 degrees of freedom
##   (1401 observations deleted due to missingness)
## Multiple R-squared:  0.08125,    Adjusted R-squared:  0.08116
## F-statistic: 859.5 on 5 and 48593 DF,  p-value: < 2.2e-16
```

```
# lets add carrier

y <- "arr_delay"
xs <- c('dep_time', 'sched_dep_time','sched_arr_time','origin','carrier')

yx <- flights_data_joined[,c(y,xs), with=FALSE]
model4 <- lm(arr_delay ~ . , data=yx)

summary(model4)
```

```
##
## Call:
## lm(formula = arr_delay ~ ., data = yx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -82.25  -22.30   -8.93    8.24 1144.43
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.452e+01  1.195e+00 -12.158  < 2e-16 ***
## dep_time        7.031e-02  1.378e-03  51.033  < 2e-16 ***
## sched_dep_time -5.005e-02  1.442e-03 -34.702  < 2e-16 ***
## sched_arr_time -4.658e-03  6.584e-04  -7.075 1.51e-12 ***
## originJFK      -1.995e+00  6.461e-01  -3.088 0.002013 **
## originLGA      -3.376e-01  5.956e-01  -0.567 0.570796
## carrierAA      -3.778e+00  1.058e+00  -3.571 0.000355 ***
## carrierAS      -1.272e+01  4.417e+00  -2.879 0.003988 **
## carrierB6       4.767e+00  9.633e-01   4.948 7.52e-07 ***
## carrierDL      -2.747e+00  9.914e-01  -2.771 0.005583 **
## carrierEV       8.803e+00  1.079e+00   8.157 3.51e-16 ***
## carrierF9       1.398e+01  4.344e+00   3.219 0.001285 **
## carrierFL       1.348e+01  2.180e+00   6.184 6.32e-10 ***
## carrierHA      -2.840e-01  7.034e+00  -0.040 0.967794
## carrierMQ       5.123e+00  1.129e+00   4.539 5.66e-06 ***
## carrierOO      -2.697e+00  1.759e+01  -0.153 0.878143
## carrierUA      -2.751e+00  1.054e+00  -2.611 0.009031 **
## carrierUS       7.186e-02  1.199e+00   0.060 0.952200
## carrierVX       5.418e-01  1.789e+00   0.303 0.762028
## carrierWN       5.176e+00  1.377e+00   3.758 0.000171 ***
## carrierYV      -1.292e+00  4.963e+00  -0.260 0.794616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.02 on 48578 degrees of freedom
##   (1401 observations deleted due to missingness)
## Multiple R-squared:  0.09192,    Adjusted R-squared:  0.09155
## F-statistic: 245.9 on 20 and 48578 DF,  p-value: < 2.2e-16
```