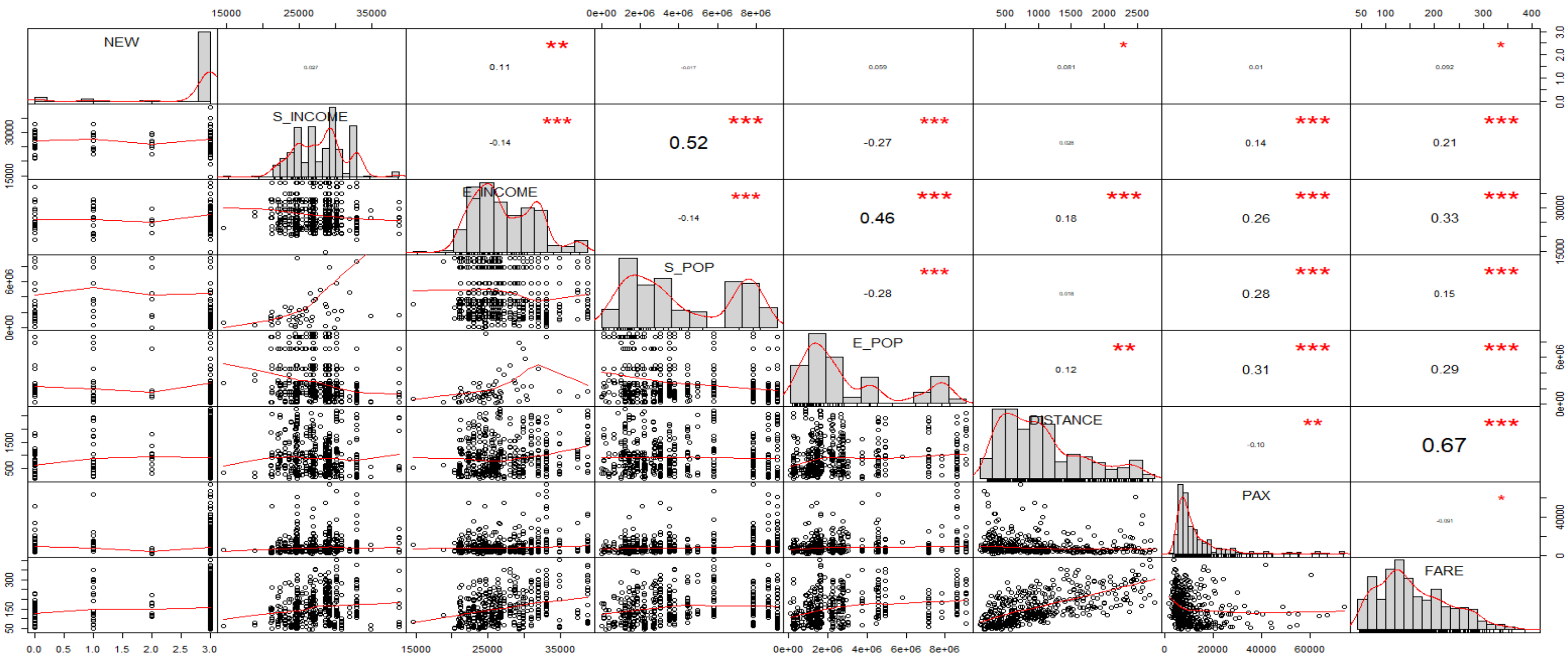
Predicting Airfare
Linear Regression Models

S_CODE:	Starting airport's code
S_CITY:	Starting city
E_CODE:	Ending airport's code
E_CITY:	Ending city
COUPON:	Average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
NEW:	Number of new carriers entering that route between Q3-96 and Q2-97
VACATION:	Whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
SW:	Whether Southwest Airlines serves that route (Yes) or not (No)
HI:	Herfindel Index – measure of market concentration (refer to BMGT 681)
S_INCOME:	Starting city's average personal income
E_INCOME:	Ending city's average personal income
S_POP:	Starting city's population
E_POP:	Ending city's population
SLOT:	Whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
GATE:	Whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
DISTANCE	Distance between two endpoint airports in miles
PAX:	Number of passengers on that route during period of data collection
FARE:	Average fare on that route

a.

For the correlation table I used the function cor() and just shared the part of that below that only shares the correlation of **FARE** with other numerical predictors. After that I created a correlation chart for the scatter plots. Looking at scatter plots and correlation chart it can be said that “**DISTANCE”** is best single predictor for Fare.

```
                FARE
NEW        0.09172969
S_INCOME   0.20913485
E_INCOME   0.32609229
S_POP      0.14509708
E_POP      0.28504299
DISTANCE   0.67001599
PAX       -0.09070541
FARE       1.00000000
```



b. .

To identify the single best categorical predictor for the Fare we used the given function and then calculated the mean difference. The following table shares the results of the mean differences. As the largest mean difference signifies the best categorical predictor thus from the results it can be interpreted that “**SW”** is the single best categorical predictor for the Fare.

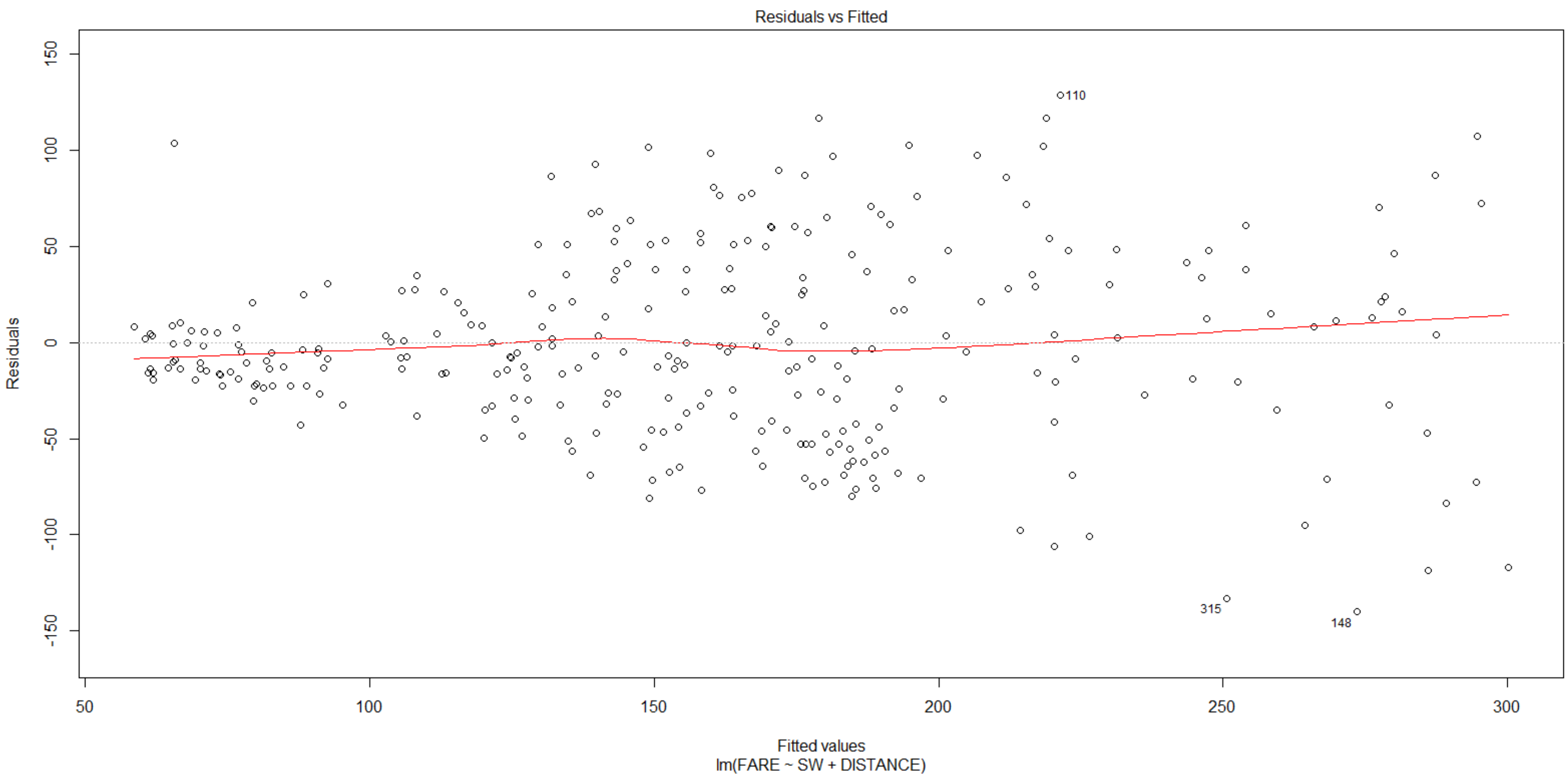
```
VarRows meanRows
1     Vac 47.57162
2      SW 89.80052
3    Slot 35.23372
4    Gate 40.03308
```

c. **.**

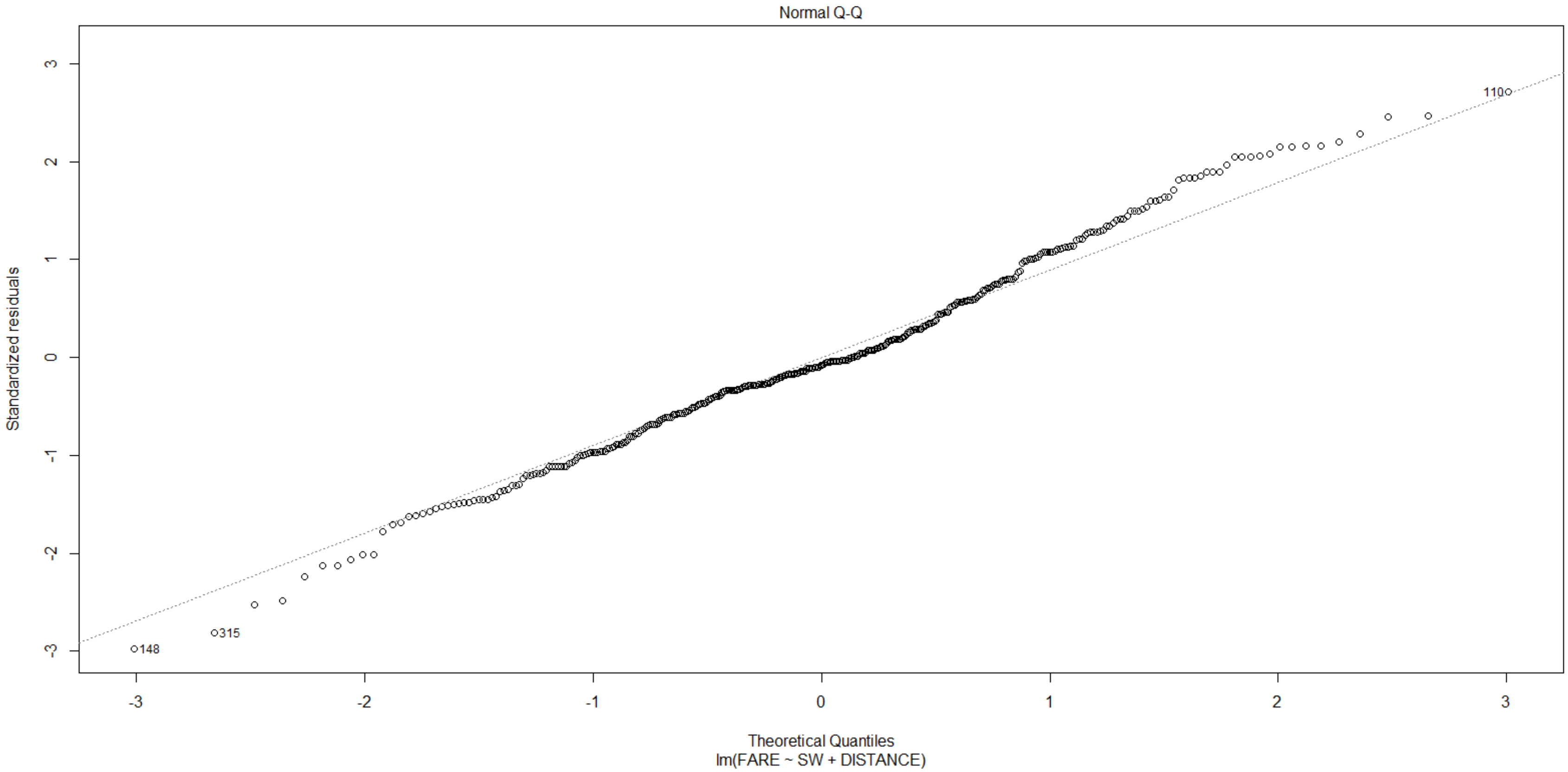
Summary of the model

```
Call:
lm(formula = FARE ~ SW + DISTANCE, data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-140.191  -28.939   -3.925   28.178  128.526 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 112.427592   4.922178   22.84   <2e-16 ***
SWYes       -63.109943   5.452649  -11.57   <2e-16 ***
DISTANCE      0.067923   0.003804   17.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.5 on 379 degrees of freedom
Multiple R-squared:  0.5887,	Adjusted R-squared:  0.5865 
F-statistic: 271.3 on 2 and 379 DF,  p-value: < 2.2e-16
```

The next two plots are for plot of residual vs Predicted values



Looking at the above plot it seems to violate constant variance assumption as the shape of the scatter plot looks like to follow a funnel pattern. This suggest residual variance lower for smaller values of x and the variance increases as the value of x increases.



From the above QQ-Plot it can be seen that generally the points lie on a straight line with a little deviations and there are few outliers as well. But it can also be noted that at the end the point deviate a little more from the line.

BIC	**4054.431**
R-Squared	**0.5887181**
MSE	**2361.868**

d. .

```
Call:
lm(formula = FARE ~ SW + DISTANCE + HI + S_INCOME + E_INCOME + 
    S_POP + E_POP + PAX + VACATION + SLOT + GATE, data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-102.023  -21.742   -1.107   19.377  103.178 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.215e+01  2.588e+01   0.856  0.39263    
SWYes       -4.016e+01  4.685e+00  -8.572 2.80e-16 ***
DISTANCE     7.702e-02  3.135e-03  24.565  < 2e-16 ***
HI           8.465e-03  1.187e-03   7.131 5.26e-12 ***
S_INCOME     1.340e-03  6.324e-04   2.119  0.03474 *  
E_INCOME     7.151e-04  4.759e-04   1.502  0.13385    
S_POP        2.607e-06  8.168e-07   3.192  0.00154 ** 
E_POP        4.686e-06  9.680e-07   4.840 1.91e-06 ***
PAX         -7.814e-04  1.653e-04  -4.727 3.25e-06 ***
VACATIONYes -3.826e+01  4.547e+00  -8.416 8.65e-16 ***
SLOTFree    -1.345e+01  4.738e+00  -2.839  0.00477 ** 
GATEFree    -2.129e+01  4.922e+00  -4.325 1.96e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.55 on 370 degrees of freedom
Multiple R-squared:  0.7876,	Adjusted R-squared:  0.7812 
F-statistic: 124.7 on 11 and 370 DF,  p-value: < 2.2e-16
```

BIC	**3855.601**
R-Squared	**0.7875503**
MSE	**1386.321**

e. .

Results of Model developed in Part(c).

BIC	**4054.431**
R-Squared	**0.5887181**
MSE	**2361.868**

Results of Model developed in Part(d).

BIC	**3855.601**
R-Squared	**0.7875503**
MSE	**1386.321**

As we know smaller BIC and MSE values and larger R-Squared value suggest a better model. Thus, looking at the above results we can conclude that model developed in Part(d) is better as compared to model developed in Part(c).