Harshaanth Thiyagaraja Kumar

Linear Regression - Predicting Airfares on New Routes

Several new airports have opened in major cities, opening the market for new routes (a route refers to a pair of airports), and Southwest has not announced whether it will cover routes to/from these cities. In order to price flights on these routes, a major airline collected information on 638 air routes in the United States. Some factors are known about these new routes: the distance traveled, demographics of the city where the new airport is located, and whether this city is a vacation destiny. Other factors are yet unknown (e.g., the number of passengers that will travel this route). A major unknown factor is whether Southwest or another discount airline will travel on these new routes. Southwest's strategy (point-to-point routes covering only major cities, use of secondary airports, standardized fleet, low fares) has been very different from the model followed by the older and bigger airlines (hub-and-spoke model extending to even smaller cities, presence in primary airports, variety in fleet, pursuit of high-end business travelers). The presence of discount airlines is therefore believed to reduce the fares greatly.

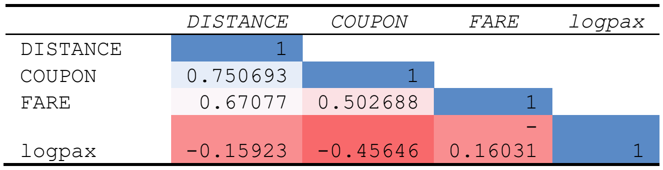
The file Airfares.xls contains real data that were collected for the third quarter of 1996. They consist of the following predictors and response:

Use the dataset in assignment 1(after data cleaning)

Take a log transformation of pax (use log10)

Don't delete any outliers you observe.

1. Create a correlation table between response and all **numerical** predictors. Use Excel *conditional formatting - color scale* to highlight the extreme values in the table. Report the correlation table. Based on the correlation table, what seems to be the best single predictor of FARE?

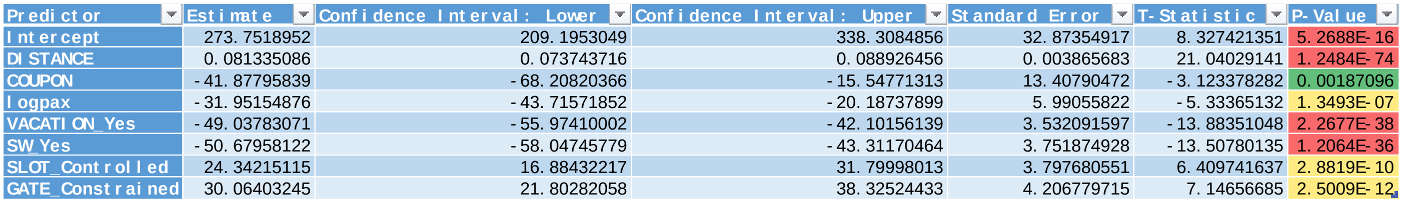


	DISTANCE	*COUPON*	*FARE*	*logpax*
DISTANCE	1			
COUPON	0.750693	1		
FARE	0.67077	0.502688	1	
logpax	-0.15923	-0.45646	-0.16031	1

distance

2. Transform categorical variables into dummies,
use "no" and "free" as baseline dummy variables, run a linear regression
Report this model's

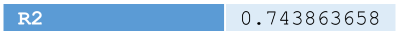
(1) *Regression equation* (write down the equation and paste the table on LinReg_output sheet),



Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	273.7518952	209.1953049	338.3084856	32.87354917	8.327421351	5.2688E-16
DISTANCE	0.081335086	0.073743716	0.088926456	0.003865683	21.04029141	1.2484E-74
COUPON	-41.87795839	-68.20820366	-15.54771313	13.40790472	-3.123378282	0.00187096
logpax	-31.95154876	-43.71571852	-20.18737899	5.99055822	-5.33365132	1.3493E-07
VACATION_Yes	-49.03783071	-55.97410002	-42.10156139	3.532091597	-13.88351048	2.2677E-38
SW_Yes	-50.67958122	-58.04745779	-43.31170464	3.751874928	-13.50780135	1.2064E-36
SLOT_Controlled	24.34215115	16.88432217	31.79998013	3.797680551	6.409741637	2.8819E-10
GATE_Constrained	30.06403245	21.80282058	38.32524433	4.206779715	7.14656685	2.5009E-12

Fare=273.75+0.08distance-41.88coupon-31.95logpax-49.04vacation_yes-50.68SW_yes+24.34SLOT_controlled+30.06GATE_constrained

(2) What is the R2? Explain it.



R2	0.743863658

74.39% of the variation in FARE can be explained by the regression equation.

(3) Interpret the coefficients (all of them).

If the distance increases by 1 mile, the fare would increase by $0.08 holding other variables constant.

If the average number of stops increases by 1, the fare would decrease by $41.88 holding other variables constant.

If logpax increases by 1, the fare would decrease by $31.95 holding other variables constant.

The fare of a vacation route is $49.04 lower than the fare of a non-vacation route holding other variables constant.

If southwest serves the route, the fare would be $50.68 lower than a route where southwest doesn’t serve holding other variables constant.

If either endpoint airport is slot controlled, the fare would be $24.34 higher holding other variables constant.

If either endpoint airport has gate constraints, the fare would be $30.06 higher holding other variables constant.

(4) What’s the predicted fare if the flight is non-stop, it is not a vacation route, southwest will not serve the route, the end airport is not slot controlled, there has no gate constraints, and the distance is 1000 miles, number of passengers is 5,000.

Fare=273.75+0.08distance-41.88coupon-31.95logpax-49.04vacation_yes-50.68SW_yes+24.34SLOT_controlled+30.06GATE_constrained

=273.75+0.08*1000-41.88*1-31.95*log(5000+1)-49.04*0 -50.68*0 +24.34*0 +30.06*0

=193.6

(5) Any suggestions regarding the firm's pricing strategy.

Price higher on routes with longer distance, fewer stops and fewer passengers, price lower on routes served by southwest and vacation routes, and price higher on routes flying busy airports

Submission checklist

1. Word or PDF file with answers to questions, including snapshots of required reports and charts.
2. Excel Workbooks.