Lab #2

By

Olivia Rancour (orr2)
Rishi Reddy (rc81)
Nikhil Preeth Birra (nb35)
Vir Benipal (bs31)

Professor Prince

To ,

The Manager,
The Butler Trucking Company,
California,USA

From,

University of Akron,
Akron,Ohio
USA.

Date:3/10/2015

Subject: To estimate the total daily travel time for the drivers.

Executive Summary

The purpose of this report is to show the results of our multiple linear regression model. Butler Trucking Company is an independent trucking company is southern California. The managers believe that the total daily travel in hours (Times) are closely related to: the number of miles traveled in making daily deliveries (Miles), the number of gallons of gasoline consumed (GasolineConsumption), the number of deliveries on a driving assignment (Deliveries), and whether or not the assignment requires the driver to travel on a congested highway (Highway).

Null and Alternative Hypotheses

The manager believes that the total daily time travelled in hours belongs to miles, gas consumption, delivers, highway. So first we set our null hypothesis:

Ho:$\beta 1 = \square 2 = \square 3 = \square 3 = \square 4 = 0$ and our alternative hypothesis:

Ha:at least one of the alternative variables not equal to zero.

Correlation matrix showing dependency

Considering the variables time, deliveries, gasoline consumption,and miles we get the below correlation table (table 1). From the correlation table, the value of +1 indicates that two variables are perfectly related in a positive linear sense. In the correlation matrix we can see that miles and GasolineConsumption is the same i.e miles≃GasolineConsumption.That is one of the variables can be ignored for further analysis.

	Miles	GasolineConsumption	Deliveries	Time
Miles	1.0000	0.9571	0.0258	0.6938
GasolineConsumption	0.9571	1.0000	0.0316	0.6587
Deliveries	0.0258	0.0316	1.0000	0.5973
Time	0.6938	0.6587	0.5973	1.0000

Table 1:- Correlation matrix table

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.262415	0.181471	-1.45	0.1492
Miles	0.0735648	0.00678	10.85	<.0001*
GasolineConsumption	-0.070927	0.072549	-0.98	0.3290
Deliveries	0.6741337	0.02363	28.53	<.0001*
Highway[1-0]	0.9941993	0.076811	12.94	<.0001*

Table 2:- Fit model parameter estimates

To know which of the variables to ignore we have further analyzed using the fit model (table 2).The parameter estimates can be obtained from the fit model, in the parameter estimates we have to consider the column Prob>|t| . We can see that gasoline consumption alone is greater than 0.0001 so we reject Gasoline consumption.

Accepted independent variable	**Unaccepted independent variable**
Miles($\beta 1$)	**Gasoline Consumption($\beta 4$)**
Ho:$\beta 1 = 0$,Ha:$\beta 1 \neq 0$	Ho:$\beta 4 = 0$,Ha:$\beta 4 \neq 0$
Rejected Ho since p<0.05,we can conclude that Mile are significant predictors for time.	Did not rejected Ho since p=0.3290 and >0.05,we can conclude that Gasoline Consumption is not significant predictors for time
Delivers($\beta 2$)	
Ho:$\beta 2 = 0$,Ha:$\beta 2 \neq 0$	
Rejected Ho since p<0.05,we can conclude that delivers are significant predictors for time.	
Highway($\beta 3$)	
Ho:$\beta 3 = 0$,Ha:$\beta 3 \neq 0$	
Rejected Ho since p<0.05,we can conclude that HIghway is a significant predictor for time.	

Interpretations of Coefficients (The betas)

We began testing each variable we set up a model equation that includes each independent variable.

times = -.330229 + $\beta 1$(0.0672203) + $\beta 2$(0.6735158) + $\beta 3$(0.9980033)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.330229	0.167678	-1.97	0.0498*	.
Miles	0.0672203	0.001961	34.27	<.0001*	1.0006683
Deliveries	0.6735158	0.02362	28.51	<.0001*	1.0035528
Highway[1-0]	0.9980033	0.076707	13.01	<.0001*	1.002885

For every mile,delivery and highway, the average time increase is by 0.067 hours,0.67 hours and 0.99 hours respectively.
The equation is:

- Times = -.330229 + 0.0672203 (miles) + 0.6735158 (Deliveries) + 0.9980033 (Highway)
- When highway=0, the equation is
    - times = -.330229 + 0.0672203 (miles) + 0.6735158 (Deliveries) + 0.9980033 (0)
    - so times = -.330229 + 0.0672203 (miles) + 0.6735158 (Deliveries)
- When highway=1, the equation is
    - times = -.330229 + 0.0672203 (miles) + 0.6735158 (Deliveries) + 0.9980033 (1)
    - so times = 0.6677743 + 0.0672203 (miles) + 0.6735158 (Deliveries)

The difference in the y-intercepts of these two lines is 0.9980033, which is the value of regression coefficient for time.

This intercepts tells us that on average, it takes 0.9980033 times more to travel on highway 1 (congested) than highway 0 (Non congested).

Coefficients of determination

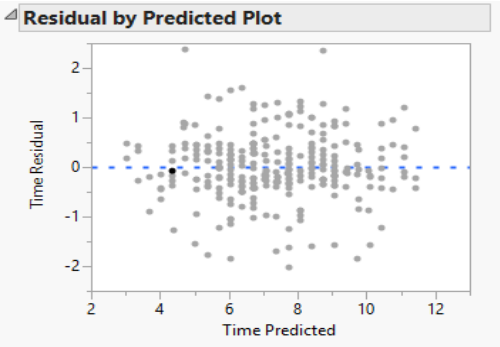
The correlation between the observed time and the predictor time is 88.3%. So higher the value of the $\square^2$the more useful the model.

- Considering the below table the individual values of p for the parameters miles, deliveries, and highways are less than 0.05, which indicates the parameters are significant.
- While checking for the multicollinearity no parameter have VIF value greater than 10, so we can conclude there is no multicollinearity.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.330229	0.167678	-1.97	0.0498*	.
Miles	0.0672203	0.001961	34.27	<.0001*	1.0006683
Deliveries	0.6735158	0.02362	28.51	<.0001*	1.0035528
Highway[1-0]	0.9980033	0.076707	13.01	<.0001*	1.002885

Residual Plots



Check for Homoscedasticity:

Miles, deliveries, and highways are good indicators of time due to the scattered appearance of the residuals. The Variance range as few outliers in the graph, so we fail the residual test.

Now we remove the two outliers from the above graph and interpret the coefficients and residual graph test again. The value of the R^2 becomes 0.892545.

Summary of Fit

RSquare	0.892545
RSquare Adj	0.891449
Root Mean Square Error	0.63564
Mean of Response	7.271812
Observations (or Sum Wgts)	298

Times = -.363746 + $\beta 1$(0.0672565) + $\beta 2$(0.6729487) + $\beta 3$(1.0309854)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.363746	0.161447	-2.25	0.0250*	.
Miles	0.0672565	0.001884	35.70	<.0001*	1.0004507
Deliveries	0.6729487	0.022704	29.64	<.0001*	1.003378
Highway[1-0]	1.0309854	0.073793	13.97	<.0001*	1.0029285

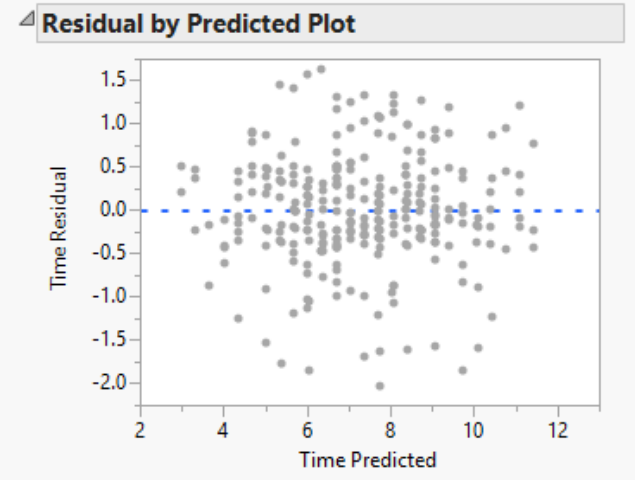
For every mile, delivery and highway the average time increase is by 0.067 hours,0.67 hours and 1.03 hours.

The equation is:

- Times = -.363746 + 0.0672565 (miles) + 0.6729487 (Deliveries) + 1.0309854 (Highway)
- When highway=0, the equation is
    - times = -.363746 + 0.0672565 (miles) + 0.6729487 (Deliveries) + 1.0309854 (0)
    - so times = -.363746 + 0.0672565 (miles) + 0.6729487 (Deliveries)
- When highway=1, the equation is
    - times = -.363746 + 0.0672565 (miles) + 0.6729487 (Deliveries) + 1.0309854 (1)
    - so times = 0.6672394 + 0.0672565 (miles) + 0.6729487 (Deliveries)

The difference in the y-intercepts of these two lines is 1.0309854, which is the value of regression coefficient for time.

This intercepts tells us that on average, it takes 1.0309854 times more to travel on highway 1 (congested) than highway 0 (Non congested).



Check for Homoscedasticity:

Miles, deliveries, and highways are good indicators of time due to the scattered appearance of the residuals. The Variance range as few outliers in the graph, so we pass the residual test.

Findings

In summary, there seems to be a positive relationship between miles, deliveries and highways with total travel time. Gasoline consumption has a p value >0.05 which makes it ineffective and can be removed. We have also removed the outliers in the residual plot and calculated an effective VIF equation giving us accurate results. We hope this information helps you in understanding how total daily times are affected by different variables.