

Regression Analysis Report with Annual Motor Sales/ SAS Studio

Didem Bulut Aykurt

MIS540-1 – Introduction to Business Intelligence

Colorado State University-Global Campus

Dr. Alin Tomoiaga

October 23, 2022

Introduction

In this case, I worked on the Annual Motor Sales dataset to apply a linear regression model by dependent variables are the location of Kansas City(KC), Chicago, Houston, Houston, Oklahoma City(OKC), Omaha, and Little Rock with continues variable “month” using SAS Studio. The dataset describes concerns about small engine products sold from six locations. The data sources are from CSUGlobal University. Twenty-four intakes represent aggregated data and six locations dependent variable with one continuous variable, “month.”

Linear regression is a statistical method to model the relationship between a dependent variable and one or more independent variables. It can help you to estimate how the dependent variable changes as the independent variables change and to test if there is a significant difference among the groups of the independent variable(s).

There are different types of linear regression, depending on the number and nature of the independent and dependent variables. Some common types are:

Simple linear regression: This is used when you have one independent variable and one dependent variable.

Multiple linear regression: This is used when you have more than one independent variable and one dependent variable.

Logistic regression: This is used when you have one or more independent variables and a binary dependent variable.

To perform a linear regression, we need to check some assumptions, such as **linearity, normality, homogeneity of variance, and independence of observations**. If the premises are met, we can calculate the regression coefficients, which are the values that determine the

shape and position of the line. We can also calculate the R-squared, which measures how well the line fits the data, and the p-value, which measures how likely the difference among the groups is due to chance.

Logistic regression does not require all the same assumptions as linear regression. Logistic regression does not assume linearity, normality, or homogeneity of variance, but it does take independence of observations. It also assumes that the response variable is binary, that there is no multicollinearity among the explanatory variables, and that there is a large sample size.

The linear regression equation measures the relationship between two variables ranging from -1 to +1. The result of -1 means a robust negative correlation, +1 is a solid positive correlation, and 0 means no relation between the two variables. The equation is below.

$$Y=a+bX$$

X: the independent variable

Y: the dependent variable

b: the slope of the line

a: the intercept when $x=0$ the value of the Y

The result of the linear regression model in SAS Studio.

The first result table has statistical details. We can say ANOVA table. That table has significant statistical points like R-square, P-value, F-value, Sum of squares, and RMSE.

The f value and P value in regression analysis on the analysis of variance table show that the model explains variation in the response, reconciling the p-value with the f statistic for

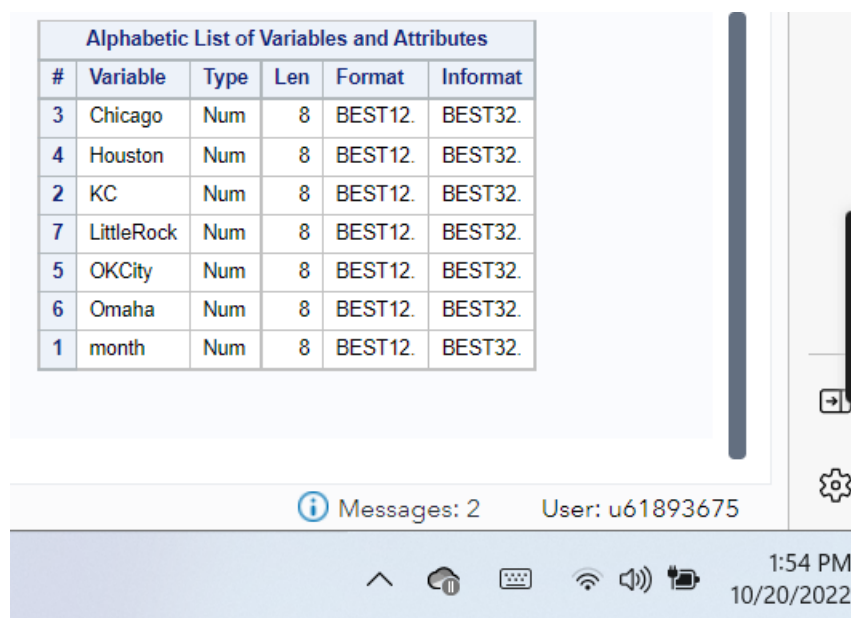
the model to result in a significant variant level. The p-value is minimal, which tells the model is doing a good job explaining much of the target's variability.

R^2 in linear regression: That allows us to measure how well the linear regression explains or fits data. R-square between 0 to 1 and the highest number is a better result, higher or equal to 0.7.

The p-value and t-value on the parameter estimate table tell us if a group of variables is jointly significant or another means when determining which rejects or fails to reject the null hypothesis. If the p-value is lower than the alpha is 0.05, the dataset is statistically substantial or rejects the null hypothesis.

Look at each result to see how our data fit a linear, normal distribution, and hypothesis result.

Figure 1: *The result of the SAS Studio description of the Annual Motor Sales dataset is as follows:*



Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
3	Chicago	Num	8	BEST12.	BEST32.
4	Houston	Num	8	BEST12.	BEST32.
2	KC	Num	8	BEST12.	BEST32.
7	LittleRock	Num	8	BEST12.	BEST32.
5	OKCity	Num	8	BEST12.	BEST32.
6	Omaha	Num	8	BEST12.	BEST32.
1	month	Num	8	BEST12.	BEST32.

Business Question and Hypothesis

1) Is there a difference in the monthly products sold for the period in KC?

- Null hypothesis: No statistically significant difference exists in KC's

monthly products sold for 24 months.

- Alternative hypothesis: There is a significant difference between KC's

monthly products sold in 24 months.

10/19/22, 6:24 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: KC

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6129.58696	6129.58696	1.73	0.2023
Error	22	78066	3548.46574		
Corrected Total	23	84196			

Root MSE	59.56900	R-Square	0.0728
Dependent Mean	3015.41667	Adj R-Sq	0.0307
Coeff Var	1.97548		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	3044.27536	25.09943	121.29	<.0001
month	1	-2.30870	1.75659	-1.31	0.2023

KC result of $Pr > F$ is 0.2; the p-value is higher than 0.05, which means there is not enough evidence to reject the null hypothesis and that the predictor variable does not significantly affect the response variable after controlling for other variables in the model. The linear regression model can't support the KC dataset. KC's f-value result with df table for alpha is 0.05 than df of model 1, and df for error is 22, that result is 4.3009. The result of KC's f statistic 1.73 is lower than the critical value of 4.3009. $Pr > t$ value of 0.001 is lower than 0.05, so we can reject

the null hypothesis, but the model can't support the dataset. That is why we can't say reject or fail to reject the null hypothesis. R-square is 0.07, which means the dataset fits 7%; it's a weak result to accept the statistical result. Thus, we might apply a different model or add more variables to get a better result.

Our Y intercept is KC product sold of result 3044.27536, and the slope of the line is - 2.30870. Let's predict sales next three months.

For the month of 25 predict product sales in KC= $3044.27536 + (-2.30870 * 25) = 2,986.55786$

For the month of 26 predict product sales in KC= $3044.27536 + (-2.30870 * 26) = 2,984.24916$

For the month of 27 predicted product sales in KC= $3044.27536 + (-2.30870 * 27) = 2,981.94046$

2) Is there a difference between monthly products sold for 24 months in Chicago?

- Null hypothesis: There is no difference between the monthly products sold for the period in Chicago.
- Alternative hypothesis: There is a significant difference between the monthly products sold in the period for Chicago.

10/19/22, 6:26 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: Chicago

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7364401	7364401	1230.68	<.0001
Error	22	131648	5984.01367		
Corrected Total	23	7496049			

Root MSE	77.35641	R-Square	0.9824
Dependent Mean	3142.70833	Adj R-Sq	0.9816
Coeff Var	2.46146		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2142.40942	32.59416	65.73	<.0001
month	1	80.02391	2.28112	35.08	<.0001

The Chicago result of Pr>F is 0.0001 is smaller than 0.05 alpha, which means the model explains significantly for each detail. The linear regression model supports the dataset.

Chicago's f-value result with df table for alpha is 0.05 than df of model 1, and df for error is 22, which is 4.3009. The result of Chicago's f statistic 1230.68 is higher than the critical value of 4.3009. Pr>t value of 0.001 is lower than 0.05, which means rejecting the null hypothesis. As a result, the parameter estimate is statistically significant. R-square is 0.98, which means the model fits 98% dataset.

The case Y intercept is Chicago product sold of result 2142.40942, and the slope of the line is 80.02391. Let's predict sales next three months.

For the month of 25 predicted product sales in Chicago= $2142.40942 + 80.02391 * 25 = 4,143.007171$.

For the month of 26, predicted product sales in Chicago= $2142.40942+80.02391*26=$
4,223.03108

For the month of 27, predicted product sales in Chicago= $2142.40942+80.02391*27=$
4303.05499

3) Is there a difference in the monthly products sold for a period in Houston?

- Null hypothesis: There is no difference in Houston's monthly products sold in 24 months.
- Alternative hypothesis: There is a statistically significant difference in monthly products sold for the period in Houston.

10/19/22, 7:15 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: Houston

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9527.04348	9527.04348	0.08	0.7856
Error	22	2763119	125596		
Corrected Total	23	2772646			

Root MSE	354.39569	R-Square	0.0034
Dependent Mean	4497.91667	Adj R-Sq	-0.0419
Coeff Var	7.87911		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	4461.93841	149.32480	29.88	<.0001
month	1	2.87826	10.45056	0.28	0.7856

Houston's result of Pr>F is 0.7856 is higher than 0.05, meaning the model can't explain each detail. The linear regression model can't support the Houston dataset. Houston's f-value

result with df table for alpha is 0.05 than df of model 1, and df for error is 22, that result is 4.3009. The result of Houston's f statistic of 0.08 is lower than the critical value of 4.3009. Pr>t value of 0.001 is lower than 0.05, so we can reject the null hypothesis, but the model can't support the dataset. That is why we can't say reject or fail to reject the null hypothesis because R-square is 0.003, which means the dataset fits 0.3%; the model doesn't provide the dataset. Thus, we might apply a different model or add more variables to get a better result.

Our Y intercept is Houston product sold of result 4461.93841, and the slope of the line is 2.87826. Let's predict sales next three months.

For the month of 25 predicted product sales in Houston= $4461.93841 + 2.87826 * 25 = 4,533.89491$

For the month of 26, predicted product sales in Houston= $4461.93841 + 2.87826 * 26 = 4,536.77317$

For the month of 27, predicted product sales in Houston= $4461.93841 + 2.87826 * 27 = 4,539.65143$

4) Is there a difference between monthly products sold for 24 months in OKCity?

- Null hypothesis: There is no difference between the monthly products sold for the last 24 months in OKCity.
- Alternative hypothesis: There is a significant difference between the monthly products sold in 24 months for OKCity.

10/19/22, 7:16 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: OKCity

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	274947587	274947587	1567.70	<.0001
Error	22	3858412	175382		
Corrected Total	23	278805999			

Root MSE	418.78680	R-Square	0.9862
Dependent Mean	9117.70833	Adj R-Sq	0.9855
Coeff Var	4.59311		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	15230	176.45602	86.31	<.0001
month	1	-488.96304	12.34935	-39.59	<.0001

OKCity result of Pr>F is 0.0001 is smaller than 0.05, which means the model explains significantly for each detail—the linear regression model support dataset. OKCity's f-value result with df table for alpha is 0.05 than df of model 1, and df for error is 22. That result is 4.3009. The result of OKCity's f statistic 1567.70 is higher than the critical value of 4.3009. Pr>t value of 0.001 is lower than 0.05, which means rejecting the null hypothesis. As a result, the parameter estimate is statistically significant. R-square is 0.98, as the model fits 98% dataset.

The Y intercept is OKCity's product sold of result 15230, and the slope of the line is -488.96304. Let's predict the sale next three months.

For the month of 25 predicted product sales in OKCity= $15230 - 488.96304 * 25 = 3,005.924$

For the month of 26, predicted product sales in OKCity= $15230 - 488.96304 * 26 = 2,516.96096$

For the month of 27, predicted product sales in OKCity= $15230 - 488.96304 * 27 = 2,027.99792$

5) Is there differences between the monthly products sold in the period in Omaha?

- Null hypothesis: No difference between monthly products sold in Omaha.
- Alternative hypothesis: there is a difference between monthly products sold in Omaha.

10/19/22, 7:17 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: Omaha

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1294.26087	1294.26087	6.20	0.0208
Error	22	4589.07246	208.59420		
Corrected Total	23	5883.33333			

Root MSE	14.44279	R-Square	0.2200
Dependent Mean	5021.66667	Adj R-Sq	0.1845
Coeff Var	0.28761		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	5034.92754	6.08548	827.37	<.0001
month	1	-1.06087	0.42589	-2.49	0.0208

Omaha's result of $Pr > F$ is 0.02 is smaller than 0.05. Omaha's f-value result with df table for alpha is 0.05 than df of model 1, and df for error is 22, which is 4.3009. The result of Omaha's f statistic of 6.20 is higher than the critical value of 4.3009, meaning the model does

not support data. P>t value of 0.001 is lower than 0.05, so we can reject the null hypothesis, but the model can't support the dataset. That is why we can't say reject or fail to reject the null hypothesis because R-square is 0.22, meaning the dataset fits 22%, and the model doesn't provide the dataset. Thus, the linear regression model can't support the Omaha dataset. We might apply a different model or add more variables to get a better result.

Our Y intercept is Omaha product sold of result 5034.92754, and the slope of the line is - 1.06087. Let's predict the sale next three months.

For the month of 25 predicted product sales in Omaha= $5034.92754 - 1.06087 * 25 = 5,008.40579$

For the month of 26, predicted product sales in Omaha= $5034.92754 - 1.06087 * 26 = 5,007.34492$

For the month of 27, predicted product sales in Omaha= $5034.92754 - 1.06087 * 27 = 5,006.28405$

6) Is there a statistically different in the monthly product sold for the period in LittleRock?

- Null hypothesis: there is no difference between the monthly product sold for a period in LittleRock.
- Alternative hypothesis: there is a difference between the monthly product sold in a period for LittleRock.

10/20/22, 1:32 PM

Results: Linear Regression

Model: MODEL1
Dependent Variable: LittleRock

Number of Observations Read	48
Number of Observations Used	24
Number of Observations with Missing Values	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	302.69565	302.69565	0.02	0.8929
Error	22	358881	16313		
Corrected Total	23	359183			

Root MSE	127.72140	R-Square	0.0008
Dependent Mean	4779.16667	Adj R-Sq	-0.0446
Coeff Var	2.67246		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	4785.57971	53.81547	88.93	<.0001
month	1	-0.51304	3.76630	-0.14	0.8929

Little Rock's result of $Pr > F$ is 0.8929 is higher than 0.05, meaning the model does not explain each detail. The linear regression model does not support the Little Rock dataset. Little Rock's f-value result with df table for alpha is 0.05 than df of model 1, and df for error is 22, which is 4.3009. The result of Little Rock's f statistic of 0.02 is lower than the critical value of 4.3009. $Pr > t$ value of 0.001 is lower than 0.05, so we can reject the null hypothesis, but the model can't support the dataset. We can't say leave or fail to reject the null hypothesis because the R-square is 0.0008, which means the dataset fits 0.08%; the model doesn't provide the dataset. Thus, we might apply a different model or add more variables for better results.

Our Y intercept is Little Rock product sold of result 4785.57971, and the slope of the line is -0.51304. Let's predict sales next three months.

For the month of 25, predicted product sales in Little Rock= $4785.57971 - 0.51304 * 25 =$
4772.75271

For the month of 26, predicted product sales in Little Rock= $4785.57971 - 0.51304 * 26 =$
4772.24067

For the month of 27, predicted product sales in Little Rock= $4785.57971 - 0.51304 * 27 =$
4771.72763

Reference

Sharda, R., Delen, D., & Turban, E. (2022). Business intelligence, analytics, and data science: A managerial perspective (4th ed.). Pearson.

Elliott, A. C., & Woodward, W. A. (2023). SAS Essentials: Mastering SAS for data analytics (3rd ed.). John Wiley & Sons.

Annual Motor Sales dataset from CSUGlobal University.

Taboga, M. (n.d.). R squared of a linear regression. <https://statlect.com/fundamentals-of-statistics/R-squared-of-a-linear-regression>

Zach (2019). How to Read the F-distribution Table. <https://www.statology.org/how-to-read-the-f-distribution-table/>