Hypothesis Testing and Regression Analysis in STATA

Introduction

Hypothesis is a statement or claim asserted about a certain phenomenon. Hypothesis testing is therefore, the process of substantiating the claim/ hypothesis. In this exercise, we are using the auto data which is a data set containing prices and other attributes of different types of cars. The data comes pre-installed with STATA and various statistical procedures were used to perform hypothesis tests. All tests were conducted at 95% confidence level.

Question 1.

In order to test the hypothesis that the average price of a car is $7000, an independent sample t-test was used. The following are the hypothesis that were formulated:

H_0: the average price of a car is $ 7000 that is $\mu = \$\,7000$.

H_1: the average price of a car is different from $ 7000 that is, is $\mu \neq \$\,7000$.

The following are the results of the independent sample test.

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   price |      74    6165.257    342.8719    2949.496    5481.914      6848.6
------------------------------------------------------------------------------
    mean = mean(price)                                            t =  -2.4346
Ho: mean = 7000                                  degrees of freedom =       73

   Ha: mean < 7000              Ha: mean != 7000              Ha: mean > 7000
 Pr(T < t) = 0.0087         Pr(|T| > |t|) = 0.0174          Pr(T > t) = 0.9913
```

From the table above, the p-value for the t-test (0.0174, 0.0087) is less than the alpha value (0.05). We therefore reject the null hypothesis and conclude that the average price of a car is less than $ 7000.

Question 2

A two independent sample t-test was used to test the hypothesis that foreign cars are more expensive than domestic cars. The following hypothesis were formulated:

H_0: there is no significance mean difference in the price of foreign and domestic cars that is

$\mu_F = \mu_D$

H_1: there is a significance mean difference in prices of foreign/imported cars and domestic cars that is; $\mu_F \neq \mu_D$

The following are results of the two independent sample t-test.

```
. ttest price, by(foreign)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Domestic	52	6072.423	429.4911	3097.104	5210.184	6934.662
Foreign	22	6384.682	558.9942	2621.915	5222.19	7547.174
combined	74	6165.257	342.8719	2949.496	5481.914	6848.6
diff		-312.2587	754.4488		-1816.225	1191.708

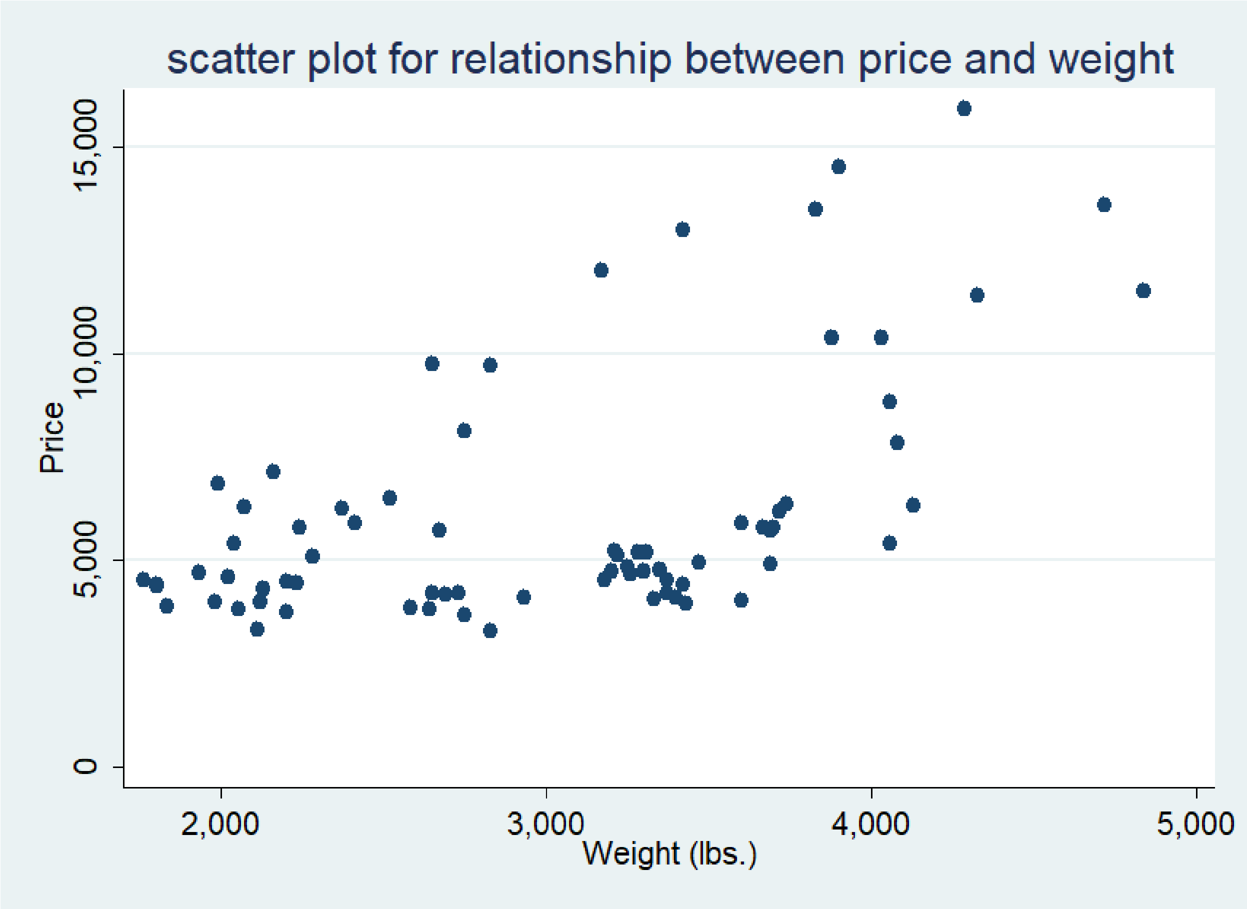
```
    diff = mean(Domestic) - mean(Foreign)                     t =  -0.4139
Ho: diff = 0                                 degrees of freedom =       72

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3401         Pr(|T| > |t|) = 0.6802          Pr(T > t) = 0.6599

.
```

From the results above, the p-values for testing the H_0 against H_1 (0.3401, 0.6802, 0.6599) are all greater than the alpha value (0.05). We therefore fail to reject H_0 and conclude that there is no statistically significant mean difference in the prices of imported/foreign and domestic cars.

Question 3.

Our aim is to investigate the relationship between the variables price ad weight. A scatter plot is the best visualization tool to depict this relationship. Thereafter we are going to perform a correlation analysis to determine the strength of the relationship if it exist.

As can be observed from the scatter plot below, there exist a positive liner relationship between price and weight variables.



Correlation analysis was done to determine the strength of the relstionship and its summarised iin the table table below.

```
. correlate price weight
(obs=74)

             |    price   weight
-------------+------------------
       price |   1.0000
      weight |   0.5386   1.0000

.
```

The correlation coefficient between price and weight is 0.5386 which implies that the two variables are strong and positively correlated.

Question 4

A regression analysis was conducted to determine the factors that are important to consider when purchasing a car. Significant factors were selected and were interpreted at 0.05 significance level. All the numeric variables were used as predictors and price as the response variable. The regression analysis is summarized below

```
. regress price mpg rep78 weight length displacement headroom trunk gear_ratio
```

Source	SS	df	MS
Model	285190003	8	35648750.3
Residual	291606956	60	4860115.94
Total	576796959	68	8482308.22

Number of obs =	69
F(8, 60) =	7.33
Prob > F =	0.0000
R-squared =	0.4944
Adj R-squared =	0.4270
Root MSE =	2204.6

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-111.6099	80.12397	-1.39	0.169	-271.8817	48.66192
rep78	880.5384	307.3861	2.86	0.006	265.6746	1495.402
weight	3.828068	1.545112	2.48	0.016	.7373826	6.918752
length	-105.7757	42.35789	-2.50	0.015	-190.5041	-21.04735
displacement	15.65373	9.084363	1.72	0.090	-2.517706	33.82516
headroom	-716	421.7548	-1.70	0.095	-1559.635	127.6353
trunk	72.08116	104.9038	0.69	0.495	-137.7576	281.9199
gear_ratio	1674.071	1074.051	1.56	0.124	-474.3504	3822.493
_cons	6856.602	6885.375	1.00	0.323	-6916.199	20629.4

From ANOVA table in the above table, $F_0 > F_\alpha$ implying that our model is statistically significant at 5% significance level. Only repair record of the car, weight and length of the car are statistically significant in the model as their p-values are less than 0.05 which is the level of significance. Furthermore, R squared value is 0.4944 implying that only 49.44% of the variation in response is accounted for by the predictors in the above table. The following is the summary of the adjusted regression model with all the significant predictors.

```
. regress price rep78 weight length

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  3,    65) =   16.16
       Model |  246375736     3  82125245.5           Prob > F      =  0.0000
    Residual |  330421222    65  5083403.42           R-squared     =  0.4271
-------------+------------------------------           Adj R-squared =  0.4007
       Total |  576796959    68  8482308.22           Root MSE      =  2254.6

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rep78 |   844.9462   302.0363     2.80   0.007      241.738    1448.154
      weight |   5.252098   1.103427     4.76   0.000     3.048401    7.455794
      length |  -103.6016   37.78457    -2.74   0.008    -179.0626   -28.14063
       _cons |   6850.952   4312.738     1.59   0.117    -1762.181    15464.08
------------------------------------------------------------------------------
```

The model is still significant and the R squared value is 0.4271 which means that only 42.71% of the variations in price are accounted for by the predictors in the model. Following is the interpretation of the regression coefficients in relation to the response variable (price).

Intercept

The coefficient for the y-intercept is $6850.952 which represent the mean value for price when all the predictor values are zero.

Repair record (rep78)

The coefficient for this factor is 844.9462 which means a that holding all the other factors constant, there is a $844.9462 increase in price of the car. unit increase in rep79

Weight

The coefficient for this variable is 5.252 implying that, holding all the other factors constant, there is a $5.252 increase in price of the car, for every unit increase in the weight of the car.

Length

Coefficient of this variable is -103.6016, implying that for every one unit increase in length, there is a corresponding $103.60 decrease in the price of the car after adjusting for both weight and rep78. The corresponding regression equation is therefore;

$$price = 6850.952 + 844.9462 * rep78 + 5.252 * weight - 103.60 * length$$