NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Week 3 Assignment - Computational: Statistical Inference in Linear Regression
MSDS 410

This assignment has two parts, the first is intended to be sure that you understand the mechanics of hypothesis testing and the information provided from a typical regression analysis. The second part asks you to begin to apply statistical inference using regression models with the AMES data.

In this assignment we will review model output from R and perform hypothesis specifications and computations related to statistical inference for linear regression. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to four places, i.e. X.xxxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results.

Any computations that involve “the log function”, denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

PART 1: MECHANICS AND COMPUTATIONS (30 points)

Model 1: Let's consider the following R output for a regression model which we will refer to as Model 1. (Note 1: In the ANOVA table, I have added 2 rows – (1) Model DF and Model SS - which is the sum of the rows corresponding to all the 4 variables (2) Total DF and Total SS - which is the sum of all the rows;

Note 2: The F test corresponding to the Model denotes the overall significance test. In R output, you will see that at the bottom of the Coefficients table)

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1974.53	1974.53	209.8340	< 0.0001
X2	1	118.8642568	118.8642568	12.6339	0.0007
X3	1	32.47012585	32.47012585	3.4512	0.0676
X4	1	0.435606985	0.435606985	0.0463	0.8303
Residuals	67	630.36	9.41		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 4 rows)	4	2126	531.50		<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	11.3303	1.9941	5.68	<.0001
X1	2.186	0.4104		<.0001
X2	8.2743	2.3391	3.54	0.0007
X3	0.49182	0.2647	1.86	0.0676
X4	-0.49356	2.2943	-0.22	0.8303

Residual standard error: 3.06730 on 67 degrees of freedom	
Multiple R-sqaured: 0.7713, Adjusted R-squared: 0.7577	
F-statistic: on 4 and 67 DF, p-value < 0.0001	

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
4	5	0.7713	166.2129	168.9481	X1 X2 X3 X4

(1) (3 points) How many observations are in the sample data?

Residuals df = n – k - 1

67 = n – 4 – 1

n = 67 + 4 + 1 = 72

We have 72 observations

(2) (3 points) Write out the null and alternate hypotheses for the t-test for Beta1.

In OLS regression the statistical inference for the individual regression coefficients can be performed by using a t-test. If there is a significant linear relationship between the independent variable X and the dependent variable Y, the beta1 will not equal zero. The null hypothesis states that the slope (beta 1) is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

- H0: Beta1 = 0; this is a null hypothesis; If it is true, x1 varaible is <u>not</u> a significant predictor to predict a response Y varaible
- Ha: Beta1 ≠ 0; this is an alternative hypothesis. If it is true, x1 varaible <u>is</u> a significant predictor to predict a response Y varaible

(3) (3 points) Compute the t- statistic for Beta1. Conduct the hypothesis test and interpret the result.

t-statistic = Beta1 / SE
where b1 is the slope of the sample regression line, and SE is the standard error of the slope.
t-statistic = 2.186 / 0.4104 = 5.3265;

We can reject H0 based on the value of t-statistic and the given significance level. This decision can be made by using the p-value of the t-statistic or by using the critical value for the significance level. With df = 71, t-statistic = 5.3265, and significance level, we get p-value = 0.00001. When the p-value is low, null must go. We reject null hypthesis (beta1 = 0) and conclude that x1 varaible is a significant predictor to predict a response Y varaible.

(4) (3 points) Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.

R-Squared = SS Model / SS Total = 2126 / 2756.37 = 0.7713

(5) (3 points) Compute the Adjusted R-Squared value for Model 1. Discuss why Adjusted R-squared and the R-squared values are different.

Adjusted R-Squared = [R-Squared] – ((1 – [R-Squared]) * [# of predictors] / [residuals df])
= 0.7713 – ((1 – 0.7713) * 4 /67) = 0.7713 – 0.01365 = 0.7577

(6) (3 points) Write out the null and alternate hypotheses for the Overall F-test.

Consider regression model $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$
The overall F-test for a regression effect is a joint hypothesis test that at least one of the predictor varaibles has a non-zero coefficient.

H0: $B_1 = B_2 = B_3 = B_4 = 0$; there is no relationship between predictors and a response Y varaible
Ha: there is at least 1 inequality

(7) (3 points) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.

F-statistics = MS regression / MS residual = 531.5 / 9.41 = 56.4825

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there is at least one predictor correlates to response variable Y.

Model 2: Now let's consider the following R output for an alternate regression model which we will refer to as Model 2.

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1928.27000	1928.27000	218.8890	<.0001
X2	1	136.92075	136.92075	15.5426	0.0002
X3	1	40.75872	40.75872	4.6267	0.0352
X4	1	0.16736	0.16736	0.0190	0.8908
X5	1	54.77667	54.77667	6.2180	0.0152
X6	1	22.86647	22.86647	2.5957	0.112
Residuals	65	572.60910	8.80937		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 6 rows)	6	2183.75946	363.96	41.3200	<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	14.3902	2.89157	4.98	<.0001
X1	1.97132	0.43653	4.52	<.0001
X2	9.13895	2.30071	3.97	0.0002
X3	0.56485	0.26266	2.15	0.0352
X4	0.33371	2.42131	0.14	0.8908
X5	1.90698	0.76459	2.49	0.0152
X6	-1.0433	0.64759	-1.61	0.112
Residual standard error: 2.968 on 65 degrees of freedom				
Multiple R-sqaured: 0.7923, Adjusted R-squared: 0.7731				
F-statistic: 41.32 on 6 and 65 DF, p-value < 0.0001				

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
6	7	0.7923	163.2947	166.7792	X1 X2 X3 X4 X5 X6

(8) (3 points) Now let's consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

Model 1 is nested in Model 2 because Model 2 has more predictors than Model 1. We can also say that Model 1 is reduced (subset) model and Model 2 is full model. We can use a F-test for nested models to decide wether or not to include an additional predictor variable in the final model. Here are the models in equation forms:

- Reduced Model: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$
- Full Model: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6$

(9) (3 points) Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.

Ho: $B_5 = B_6 = 0$
Ha: there is at least 1 inequality
If p-value is small, I'll be reducing error significantly with full model. Full model will be significant.

(10) (3 points) Compute the F-statistic for a nested F-test using Model 1 and Model 2. Conduct the hypothesis test and interpret the results.

F- statistics = (([SS residuals of reduced model] - [SS residuals of full model]) / 2) / [mean sq. residuals of full model] = ((630.36 – 572.6091) / 2) / 8.8094 = 3.2778

So, F- statistics is small (p-value is big), do not reject null hypothesis and conclude that the reduced model is a good enough; do not need full model.

PART II: APPLICATION (20 points)

For this part of the assignment, you are to use the AMES Housing Data you worked with during Modeling Assignment #1.

Model 3:

(11) Based on your EDA from Modeling Assignment #1, focus on 10 of the continuous quantitative variables that you though/think might be good explanatory variables for SALESPRICE. Is there a way to logically group those variables into 2 or more sets of explanatory variables? For example, some variables might be strictly about size while others might be about quality. Separate the 10 explanatory variables into at least 2 sets of variables. Describe why you created this separation. A set must contain at least 2 variables.

Yes, there is a way to logically group those varaibles into 2 or more sets of explanatory variables. I created this separation because I can see if those variables can be correlated to each other. We don't want to see the predictors correlate to each other.

- Size: TotalFloorSF, TotalBsmtSF, LotArea, FullBath, HalfBath
- Quality: OverallQual, YearRemodel, QualityIndex, HouseAge, price_sqft

(11)Pick one of the sets of explanatory variables. Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Call this Model 3. Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + TotalBsmtSF + LotArea +
    FullBath + HalfBath, data = model3)

Residuals:
    Min      1Q  Median      3Q     Max
-171406  -19840    -355   17259  251773

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)  -39790.366   2908.577  -13.68 < 0.0000000000000002 ***
TotalFloorSF     68.165      2.696   25.28 < 0.0000000000000002 ***
TotalBsmtSF      76.142      2.293   33.21 < 0.0000000000000002 ***
LotArea           0.622      0.106    5.89         0.0000000045 ***
FullBath      17104.928   2017.607    8.48 < 0.0000000000000002 ***
HalfBath      16609.542   1992.234    8.34 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35500 on 1996 degrees of freedom
Multiple R-squared:  0.766,     Adjusted R-squared:  0.765
F-statistic: 1.3e+03 on 5 and 1996 DF,  p-value: <0.0000000000000002
```

a) all model coefficients individually

t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608

H0: $B_1 = 0$; TotalFloorSF is not a significant predictor

Ha: $B_1 \neq 0$; TotalFloorSF is a significant predictor

t-value = 68.165 / 2.696 = 25.2838

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the TotalFloorSF is significantly helping us in predicting the sales price of homes.

H0: $B_2 = 0$; TotalBsmtSF is not a significant predictor

Ha: $B_2 \neq 0$; TotalBsmtSF is a significant predictor

t-value = 76.142 / 2.293 = 33.2063

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the TotalBsmtSF is significantly helping us in predicting the sales price of homes.

H0: $B_3 = 0$; LotArea is not a significant predictor

Ha: $B_3 \neq 0$; LotArea is a significant predictor

t-value = 0.622 / 0.106 = 5.8679

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the LotArea is significantly helping us in predicting the sales price of homes.

H0: $B_4 = 0$; FullBath is not a significant predictor
Ha: $B_4 \neq 0$; FullBath is a significant predictor

t-value = 17104.928 / 2017.60 = 8.4779
t-value > t-critical and we can also see that R generated p-value is very small.
With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the FullBath is significantly helping us in predicting the sales price of homes.

H0: $B_5 = 0$; HalfBath is not a significant predictor
Ha: $B_5 \neq 0$; HalfBath is a significant predictor

t-value = 16609.542 / 1992.234 = 8.3371
t-value > t-critical and we can also see that R generated p-value is very small.
With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the HalfBath is significantly helping us in predicting the sales price of homes.

b) the Omnibus Overall F-test

H0: $B_1 = B_2 = B_3 = B_4 = B_5 = 0$; there is no relationship between predictors and a response Y varaible
Ha: there is at least 1 inequality

F-statistics = MS regression / MS residual = 16482571449167 /1263504604= 1304.5
We can also see the same results from above R generated summary of linear regression statitics.
F-Statistic: 1305; p-value is very small
When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

Model 4:

(13) Pick the other set (or one of the other sets) of explanatory variables. Add this set of variables to those in Model 3. In other words, Model 3 should be nested within Model 4. . Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + TotalBsmtSF + LotArea +
    FullBath + HalfBath + OverallQual + YearRemodel + QualityIndex +
    HouseAge + price_sqft, data = model4)

Residuals:
   Min     1Q Median     3Q    Max
-85404  -5704   -195   4959 142813

Coefficients:
                 Estimate   Std. Error t value                Pr(>|t|)
(Intercept)  -219202.8847   39069.0281   -5.61       0.000000022974 ***
TotalFloorSF     117.7896       1.3097   89.94 < 0.0000000000000002 ***
TotalBsmtSF        7.5323       1.1202    6.72       0.000000000023 ***
LotArea            0.0510       0.0416    1.22              0.22072
FullBath       -1691.4467     911.2484   -1.86              0.06357 .
HalfBath        3248.7943     861.0846    3.77              0.00017 ***
OverallQual     5520.2704     514.3396   10.73 < 0.0000000000000002 ***
YearRemodel        3.7097      19.8958    0.19              0.85211
QualityIndex    -506.1400      58.3788   -8.67 < 0.0000000000000002 ***
HouseAge         118.2276      19.2966    6.13       0.000000001077 ***
price_sqft      1552.3664      19.5331   79.47 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13600 on 1991 degrees of freedom
Multiple R-squared:  0.966,     Adjusted R-squared:  0.966
F-statistic: 5.66e+03 on 10 and 1991 DF,  p-value: <0.0000000000000002
```

a) all model coefficients individually

t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608

H0: $B_1 = 0$; TotalFloorSF is not a significant predictor

Ha: $B_1 \neq 0$; TotalFloorSF is a significant predictor

t-value = 117.7896 / 1.3097 = 8963

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the TotalFloorSF is significantly helping us in predicting the sales price of homes.

H0: $B_2 = 0$; TotalBsmtSF is not a significant predictor

Ha: $B_2 \neq 0$; TotalBsmtSF is a significant predictor

t-value = 7.5323 / 1.1202 = 6.7241

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the TotalBsmtSF is significantly helping us in predicting the sales price of homes.

H0: $B_3 = 0$; LotArea is not a significant predictor

Ha: $B_3 \neq 0$; LotArea is a significant predictor

t-value = 0.0510 / 0.0416 = 1.2260

t-value < t-critical and we can also see that R generated p-value is larger than alpha 2.5%

With 95% confidence (two-tail test), we do not reject the null hypothesis and conclude that the LotArea is not significantly helping us in predicting the sales price of homes. If I already have 9 other predictors, I can drop this predictor.

H0: $B_4 = 0$; FullBath is not a significant predictor
Ha: $B_4 \neq 0$; FullBath is a significant predictor

t-value = -1691.4467 / 911.2484 = -1.8562

t-value < t-critical (in absolute form) and we can also see that R generated p-value is larger than alpha 2.5%

With 95% confidence (two-tail test), we do not reject the null hypothesis and conclude that the FullBath is not significantly helping us in predicting the sales price of homes. We can also drop this predictor.

H0: $B_5 = 0$; HalfBath is not a significant predictor
Ha: $B_5 \neq 0$; HalfBath is a significant predictor

t-value = 3248.7943 / 861.084 = 3.7729

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the HalfBath is significantly helping us in predicting the sales price of homes.

H0: $B_6 = 0$; OverallQual is not a significant predictor
Ha: $B_6 \neq 0$; OverallQual is a significant predictor

t-value = 5520.2704 / 514.3396 = 10.7327

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the OverallQual is significantly helping us in predicting the sales price of homes.

H0: $B_7 = 0$; YearRemodel is not a significant predictor
Ha: $B_7 \neq 0$; YearRemodel is a significant predictor

t-value = 3.7097 / 19.8958 = 0.1865

t-value < t-critical and we can also see that R generated p-value is larger than alpha 2.5%

With 95% confidence (two-tail test), we do not reject the null hypothesis and conclude that the YearRemodel is not significantly helping us in predicting the sales price of homes. We can also drop this predictor.

H0: $B_8 = 0$; QualityIndex is not a significant predictor
Ha: $B_8 \neq 0$; QualityIndex is a significan predictor

t-value = -506.1400 / 58.3788 = -8.6700

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the QualityIndex is significantly helping us in predicting the sales price of homes.

H0: $B_9 = 0$; HouseAge is not a significant predictor

Ha: $B_9 \neq 0$; HouseAge is a significant predictor

t-value = 118.2276 / 19.2966 = 6.1269

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the HouseAge is significantly helping us in predicting the sales price of homes.

H0: $B_{10} = 0$; price_sqft is not a significant predictor

Ha: $B_{10} \neq 0$; price_sqft is a significant predictor

t-value = 1552.3664 / 19.533 = 79.4740

t-value > t-critical and we can also see that R generated very small p-value.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the price_sqft is significantly helping us in predicting the sales price of homes.

b) the Omnibus Overall F-test

Ho: $B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = B_7 = B_8 = B_9 = B_{10} = 0$

Ha: there is at least 1 inequality

F-statistics = MS regression / MS residual = 1039894769060 / 183726991 = 5660

We can also see the same results from above R generated summary of linear regression statitics.

F-Statistic: 5660; p-value is very small

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

Nested Model:

(14) Write out the null and alternate hypotheses for a nested F-test using Model 3 and Model 4, to determine if the Model 4 variables, as a set, are useful for predicting SALEPRICE or not. Compute the F-statistic for this nested F-test and interpret the results.

```
Analysis of Variance Table

Model 1: SalePrice ~ TotalFloorSF + TotalBsmtSF + LotArea + FullBath +
    HalfBath
Model 2: SalePrice ~ TotalFloorSF + TotalBsmtSF + LotArea + FullBath +
    HalfBath + OverallQual + YearRemodel + QualityIndex + HouseAge +
    price_sqft
  Res.Df           RSS Df     Sum of Sq    F              Pr(>F)
1   1996 2521955190051
2   1991  365800439375  5 2156154750676 2347 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Ho: $B_5 = B_6 = B_7 = B_8 = B_9 = B_{10} = 0$

- Ha: there is at least 1 inequality
  F- statistics = (([SS residuals of reduced model] - [SS residuals of full model]) / 2) / [mean sq. residuals of full model] = ((16482571449167 - 1039894769060)/2)/183726991 = 42026. Since the individual ANOVA summaries showed very big numbers, it might have concatenated the digits because of not enough space. I will go with F value of above ANOVA summary of both models which is 2347.

  According to F- statistics 2347 and very small p-value, we reject null hypothesis and conclude that the reduced model is not a good enough; full model is significant.

Assignment Document:

Results should be presented and discussed in the numerical order of the questions given. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document MUST be submitted in pdf format. Please use the naming convention: CompAssign2_YourLastName.pdf.