Scott Jue
1/22/2023
MSDS 410

Modeling Assignment #3: Statistical Inference in Linear Regression

PART 1:

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?

There are 72 observations in the sample data.

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

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

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

t = 2.186 / 0.4101
t = 5.3265

The degrees of freedom for a multiple linear regression model is the number of observations minus the number of parameters. Therefore, using the degrees of freedom of 67 and a 0.05 significance level, the critical t-static value is 1.996. We can see that the computed t-statistic of 5.3265 is greater than the critical t-value. Therefore, we can reject the null hypothesis that $\beta_1 = 0$. We can interpret this result as having X1, meaning that $\beta_1 \neq 0$, included in the model provides significant information for predicting the response variable (Y).

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

R-Squared = (2756.37-630.36) / 2756.37
R-Squared = 0.7713

The R-Squared value can be interpreted as approximately 77.13% of the variability in the response variable (Y) can be accounted for by the explanatory variances X1, X2, X3, and X4 in Model 1.

(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 = $1 - \left[\frac{(1-0.7713)(72-1)}{72-4-1}\right]$

Adjust R-Squared = 0.7577

When explanatory variables are added to the model, R-Squared value will increase even if the variable does not significantly improve the model. Adjusted R-squared value increases only when the explanatory variable is significant and helps with prediction of the response variable. Furthermore, ff the addition of an explanatory variable does not improve the model, then adjusted R-Squared will penalize the addition of this variable and the value will decrease.

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

Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

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

$$F = \left[\frac{\frac{(2756.37 - 630.36)}{4}}{\left(\frac{630.36}{72-4-1}\right)} \right]$$

F = 56.4926

Using the degrees of freedom of 4 for the numerator and 67 for the denominator and a 0.05 significance level, the critical F-static value is 2.5087. As we can see, the computed F-statistic value of 56.4926 is much greater than the critical F value of 2.5087. Therefore, we can reject the null hypothesis and conclude that the explanatory variables included in the model significantly help to predict the response variable.

Model 2

Now let's consider the following R output for an alternative 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 includes variables X1, X2, X3 and X4, while Model 2 consists of variables X1, X2, X3, X4, X5, and X6. As we can see Model 2 contains the same variables as Model 1 which are X1, X2, X3, and X4, but adds the additional variables X5 and X6. Therefore, Model 2 nests Model 1.

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

Null hypothesis → H_0: $\beta_5 = \beta_6 = 0$
Alternative hypotheses → H_A: $\beta_5 \neq 0$ or $\beta_6 \neq 0$

(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.

$$F = \left[\frac{\left(\frac{2183.75946 - 2126}{2} \right)}{\left(\frac{572.6091}{(72-6-1)} \right)} \right]$$

F = 3.2783

For the F-test, the critical F-value at the 0.05 significance level is 3.1381. Therefore, since the computed F-value of 3.273 is greater than the critical value, we can reject the null hypothesis and conclude that the addition of the X5 and X6 explanatory variable to the model significantly helps to predict the response variable.

PART II: APPLICATION

Introduction

For this section, we will use the Ames, IA housing dataset to build two multiple linear regression models that predict the sale price of a typical home in Ames, IA between 2006 and 2010. From our previous exploratory data analysis, we will use a subset of the original data that only includes observations with residential zones, single-family homes, normal sale conditions, total square feet less than 4,000, all public utilities, and at least a 1 car garage. Additionally, any observations that have null values have been removed prior to model fitting since they lack completeness. The first model (Model 3) will be a reduced model using only one type of variable. The second model (Model 4) expands on the first model and adds additional explanatory variables. We will then perform a series of hypotheses tests to evaluate the usefulness the explanatory variables. Lastly, we will also compare the two models to see if the full model is a better fit for the dataset than the reduced model.

Model 3

(11) The 10 continuous explanatory variables I have chosen have been split into 3 sets. The first set which contains 5 variables are related to interior size. The second set of variables which consist of 2 variables are related to the size of the lot. The third set of variables are related to size of the porch. By separating the variables into different groups by type we can learn what type of variables are better for predicting to response variable SalePrice. Since there are many variables within a specific type, we can build an initial multiple regression model using one type of variable, then nest that model inside another model that adds additional types of variables to see how much the model improves. From there, we can determine which types of variables are significant in helping us predict the response variable. The table below summarize these 10 variables and the category type.

Variable	Category Type
FirstFlrSF	Interior
SecondFlrSF	Interior
TotalBsmtSF	Interior
GrLivArea	Interior
GarageArea	Interior
LotArea	Lot
LotFrontage	Lot
ScreenPorch	Porch
EnclosedPorch	Porch
OpenPorchSF	Porch

(12) For Model 3, I will use the set of explanatory variables that are related to the interior of the house. Below is the equation for the linear model using these interior variables which results in a R-squared value of 0.7797.

Fitted Model 3:

Y_hat = -49318.311 + 101.1 *FirstFlrSF + 101.902*SecondFlrSF + 63.727*TotalBsmtSF – 22.308*GrLivArea + 95.318*GarageArea

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -49318.311   3368.759 -14.640  < 2e-16 ***
FirstFlrSF     101.100     18.746   5.393 8.01e-08 ***
SecondFlrSF    101.902     18.337   5.557 3.23e-08 ***
TotalBsmtSF     63.727      3.532  18.045  < 2e-16 ***
GrLivArea      -22.308     18.169  -1.228     0.22
GarageArea      95.318      5.811  16.404  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34950 on 1533 degrees of freedom
Multiple R-squared:  0.7797,    Adjusted R-squared:  0.779
F-statistic:  1085 on 5 and 1533 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq   F value Pr(>F)
FirstFlrSF      1 3.5847e+12 3.5847e+12 2934.8128 <2e-16 ***
SecondFlrSF     1 2.1290e+12 2.1290e+12 1743.0385 <2e-16 ***
TotalBsmtSF     1 5.8418e+11 5.8418e+11  478.2723 <2e-16 ***
GrLivArea       1 2.0670e+09 2.0670e+09    1.6923 0.1935
GarageArea      1 3.2869e+11 3.2869e+11  269.0978 <2e-16 ***
Residuals    1533 1.8725e+12 1.2214e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the below hypothesis tests of the individual model coefficients, we will use a critical t-value of 1.9615 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,533.

***Intercept*:**

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = -14.64

Since the absolute value of t is greater than the critical t-value of 1.9615, then we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. However, the intercept value cannot be interpreted outside of this dataset since the value is negative. That is if all other variables are equal to 0, then the predicted SalePrice is negative which is not only outside the range of the dataset but it is also not feasible for SalePrice of a home to be negative.

FirstFlrSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 5.393

The t-value of 5.393 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, FirstFlrSF, included in the model provides significant information for predicting the response variable (Y).

SecondFlrSF:

Null hypothesis → H_0: $\beta_2 = 0$
Alternative hypotheses → H_A: $\beta_2 \neq 0$

t = 5.557

The t-value of 5.557 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_2 = 0$ and conclude that $\beta_2 \neq 0$. We can interpret this result as having the explanatory variable, SecondFlrSF, included in the model provides significant information for predicting the response variable (Y).

TotalBsmtSF:

Null hypothesis → H_0: $\beta_3 = 0$
Alternative hypotheses → H_A: $\beta_3 \neq 0$

t = 18.045

The t-value of 18.045 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_3 = 0$ and conclude that $\beta_3 \neq 0$. We can interpret this result as having the explanatory variable, TotalBsmtSF, included in the model provides significant information for predicting the response variable (Y).

GrLivArea:

Null hypothesis → H_0: $\beta_4 = 0$
Alternative hypotheses → H_A: $\beta_4 \neq 0$

t = -1.228

The absolute t-value of -1.228 is less than the critical t-value of 1.9615; therefore, we fail reject the null hypothesis and conclude that $\beta_4 = 0$. We can interpret this result as having the explanatory variable, GrLivArea, included in the linear model does not provides significant information in predicting response variable (Y).

GarageArea:

Null hypothesis → H_0: $\beta_5 = 0$
Alternative hypotheses → H_A: $\beta_5 \neq 0$

t = 16.404

The t-value of 16.404 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_5 = 0$ and conclude that $\beta_5 \neq 0$. We can interpret this result as having the explanatory variable, BsmtUnfSF, included in the model provides significant information for predicting the response variable (Y).

Next, we will perform the Omnibus Overall F-test for Model 3. For this hypothesis test of the overall model, we will use a critical F-value of 2.2199 which is the value associated with a

numerator degrees of freedom of 5, for the 5 explanatory variables, and a denominator degrees of freedom of 1,909, for the residuals, at the 0.05 significance level.

Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

$$F = \left[\frac{\frac{(8.501071e+12 - 1.872457e+12)}{5}}{\left(\frac{1.872457e+12}{1539-5-1}\right)} \right]$$

F = 1085.3827

The F-statistic of 1085.3827 is greater than the critical F-value of 2.2199. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Model 4

(13) Model 4 will use the same set of explanatory variables as Model 3, but we will also add the explanatory variables that are related to the lot size. Therefore, we will add the LotArea and LotFrontage variables to Model 4. Below is the equation for this linear regression model using the interior and lot related variables which results in a R-squared value of 0.7861.

Fitted Model 4:

Y_hat = -57390 + 88.12**FirstFlrSF + 96.65*SecondFlrSF + 65.61*TotalBsmtSF – 18.78*GrLivArea + 90.92*GarageArea + 0.7253*LotArea + 164.4*LotFrontage

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.739e+04  3.906e+03 -14.691  < 2e-16 ***
FirstFlrSF   8.812e+01  1.864e+01   4.728 2.48e-06 ***
SecondFlrSF  9.665e+01  1.812e+01   5.334 1.10e-07 ***
TotalBsmtSF  6.561e+01  3.497e+00  18.766  < 2e-16 ***
GrLivArea   -1.878e+01  1.795e+01  -1.047  0.29547
GarageArea   9.092e+01  5.823e+00  15.612  < 2e-16 ***
LotArea      7.253e-01  1.389e-01   5.223 2.00e-07 ***
LotFrontage  1.644e+02  5.298e+01   3.103  0.00195 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34460 on 1531 degrees of freedom
Multiple R-squared:  0.7861,    Adjusted R-squared:  0.7851
F-statistic: 803.7 on 7 and 1531 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq   F value    Pr(>F)
FirstFlrSF     1 3.5847e+12 3.5847e+12 3017.8694 < 2.2e-16 ***
SecondFlrSF    1 2.1290e+12 2.1290e+12 1792.3673 < 2.2e-16 ***
TotalBsmtSF    1 5.8418e+11 5.8418e+11  491.8076 < 2.2e-16 ***
GrLivArea      1 2.0670e+09 2.0670e+09    1.7402  0.187315
GarageArea     1 3.2869e+11 3.2869e+11  276.7134 < 2.2e-16 ***
LotArea        1 4.2474e+10 4.2474e+10   35.7584 2.775e-09 ***
LotFrontage    1 1.1434e+10 1.1434e+10    9.6262  0.001953 **
Residuals   1531 1.8185e+12 1.1878e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For hypothesis testing of all model coefficients individually, we will use a critical t-value of 1.9615 which is the value associated with a two tailed test at the 0.05 significance level and with degrees of freedom of 1,531.

***Intercept*:**

Null hypothesis → H_0: $\beta_0 = 0$
Alternative hypotheses → H_A: $\beta_0 \neq 0$

t = -14.691

Since the absolute value of t is greater than the critical t-value of 1.9615, then we can reject the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_0 \neq 0$. However, the intercept value cannot be interpreted outside of this dataset since the value is negative. That is if all other variables are equal to 0, then the predicted SalePrice is negative which is not only outside the range of the dataset, but it is also not feasible for SalePrice of a home to be negative.

FirstFlrSF:

Null hypothesis → H_0: $\beta_1 = 0$
Alternative hypotheses → H_A: $\beta_1 \neq 0$

t = 4.728

The t-value of 4.728 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_1 = 0$ and conclude that $\beta_1 \neq 0$. We can interpret this result as having the explanatory variable, FirstFlrSF, included in the model provides significant information for predicting the response variable (Y).

SecondFlrSF:

Null hypothesis → H_0: $\beta_2 = 0$

Alternative hypotheses → H_A: $\beta_2 \neq 0$

t = 5.334

The t-value of 5.334 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_2 = 0$ and conclude that $\beta_2 \neq 0$. We can interpret this result as having the

explanatory variable, SecondFlrSF, included in the model provides significant information for predicting the response variable (Y).

TotalBsmtSF:

Null hypothesis → H_0: $\beta_3 = 0$

Alternative hypotheses → H_A: $\beta_3 \neq 0$

t = 18.766

The t-value of 18.766 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_3 = 0$ and conclude that $\beta_3 \neq 0$. We can interpret this result as having the explanatory variable, TotalBsmtSF, included in the model provides significant information for predicting the response variable (Y).

GrLivArea:

Null hypothesis → H_0: $\beta_4 = 0$

Alternative hypotheses → H_A: $\beta_4 \neq 0$

t = -1.047

The absolute t-value of -1.047 is less than the critical t-value of 1.9615; therefore, we fail reject the null hypothesis and conclude that $\beta_4 = 0$. We can interpret this result as having the explanatory variable, GrLivArea, included in the linear model does not provides significant information in predicting response variable (Y).

GarageArea:

Null hypothesis → H_0: $\beta_5 = 0$

Alternative hypotheses → H_A: $\beta_5 \neq 0$

t = 15.612

The t-value of 15.612 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_5 = 0$ and conclude that $\beta_5 \neq 0$. We can interpret this result as having the explanatory variable, BsmtUnfSF, included in the model provides significant information for predicting the response variable (Y).

LotArea:

Null hypothesis → H_0: $\beta_6 = 0$

Alternative hypotheses → H_A: $\beta_6 \neq 0$

t = 5.223

The t-value of 5.223 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_6 = 0$ and conclude that $\beta_6 \neq 0$. We can interpret this result as having the

explanatory variable, LotArea, included in the model provides significant information for predicting the response variable (Y).

LotFrontage:

Null hypothesis → H_0: $\beta_7 = 0$

Alternative hypotheses → H_A: $\beta_7 \neq 0$

t = 3.103

The t-value of 3.103 is greater than the critical t-value of 1.9615; therefore, we can reject the null hypothesis that $\beta_7 = 0$ and conclude that $\beta_7 \neq 0$. We can interpret this result as having the explanatory variable, LotFrontage, included in the model provides significant information for predicting the response variable (Y).

Next, we will perform the Omnibus Overall F-test for Model 4. For this hypothesis test of the overall model, we will use a critical F-value of 2.0155 which is the value associated with a numerator degrees of freedom of 7, for the 7 explanatory variables, and a denominator degrees of freedom of 1,531, for the residuals, at the 0.05 significance level.

Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

$$F = \left[\frac{\frac{(8.501071e+12 - 1.818549e+12)}{7}}{\left(\frac{1.818549e+12}{1539-7-1}\right)} \right]$$

F = 803.6975

The F-statistic of 803.6975 is greater than the critical F-value of 2.0155. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Nested Model

(14) Next, we will use the nested F-test using Model 3 and Model 4 to determine if the additionay lot-related variables are useful for predicting SalePrice. The critical F-value for at the 0.05 significance level is 3.0016. Below are the null and alternate hypotheses for the nested F-test.

Null hypothesis → H_0: $\beta_6 = \beta_7 = 0$
Alternative hypotheses → H_A: $\beta_6 \neq 0$ or $\beta_7 \neq 0$

$$F = \left[\frac{\left(\frac{6.682522e+12 - 6.628614e+12}{2}\right)}{\left(\frac{1.818549e+12}{(1539-7-1)}\right)} \right]$$

F = 22.6923

As we can see the computed F-value of 22.6923 is greater than the critical F-value of 3.0016. Therefore, we can reject the null hypothesis and conclude that the set of variables related to lot size adds significant information to the linear model for predicting the SalePrice.

Conclusion

From this analysis we were able to fit 2 regression models, Model 3 and Model 4, that predict the sale price of a typical home in Ames, Iowa between 2006 and 2010. Model 3 is a reduced model fitted only using interior size variables as the explanatory variables. This model had a R-Squared value of 0.7797 and adjusted R-Squared value of 0.779. Model 4 is an expanded model that nests Model 3 and also includes the set of variables related to the lot size. This model had a R-Squared value of 0.7861 and adjusted R-Squared value of 0.7851. With these variables added to the model, Model 4 accounts for approximately 79% of the variability in the response variable SalePrice Furthermore, a nested F-test was performed to confirm that the addition of the lot variables provided significant useful information for predicting SalePrice. As such, Model 4 is a better fitting model compared to Model 3.