Robin Singh



|  |
| --- |
| Modeling Assignment #2: Building Linear Regression Models  *MSDS 410* |

In this assignment we will begin building regression models to predict home sale price (SALEPRICE) using the variables in the AMES data. This assignment walks you through a number of modeling experiences, these are delineated below as tasks. Each task should be completed and written about separately. This is not necessarily the sequence of steps of what you should do in a modeling setting, but it is intended to give you perspective on modeling.

**Purpose:** In this modeling assignment we will begin building linear regression models to predict the home sale price. As such the response variable is: SALEPRICE (Y). We will begin by fitting specific models and looking at diagnostic and model fit information. Models will progressively become more involved and complex over the span of this assignment.

**Data:** The data for this assignment is the Ames, Iowa housing data set. This data is posted in Canvas.

**Explanatory Variables:** All continuous variables in the AMES Housing data set

**Assignment Tasks:**

1. Define the Sample Population – Exploratory Data Analysis

In Modeling Assignment #1, you were exposed to the idea of a Sample Population and the traditions of Exploratory Data Analysis. Every time you start a modeling endeavor, these two tasks need to be completed and formalized

As it says above, we are building regression models for the response variable SalePrice(Y). In order to do this, you need to know the Sample Population. Without this, it is not possible to infer results from the sample to the larger population, it makes the notion of hypothesis testing for population parameter values irrelevant, and it makes the process of determining outliers highly problematic. Frankly, it throws your whole purpose for modeling into chaos.

Defining the Sample Population is actually a very powerful tool for you as the modeler. It gives you license to define what aspects of the data are legitimate for you to work with. You don’t have to model ALL of the data you are given in one model. You can break the data up into parts and model them separately. Why would you want to do this? Well, are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence? Are there certain kinds of properties that are not like the others? Could one be a derelict property such that it is not like the others? Could one be a mansion such that it is not like the other properties in the data set? You get to define this! In doing this, often many records with extreme scores are eliminated from modeling consideration. Define your Target Population and hence the Sample using ‘drop conditions’. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population. If you want to use your conditions from Modeling Assignment #1, that is fine. If you feel you need to make changes, now is the time to do so.

Once the Sample Population is clearly well defined, and you’ve selected only those records and fit the Sample Population definition, you can then continue to perform a detailed Exploratory Data Analysis (EDA). Usually, this is broken up into two parts. The first is data preparation (or data cleaning). Here, you concern yourself with any remaining missing values, extreme scores, and outliers.

* 1. Are there variables with missing values? Should values for these variables be imputed or “fixed”? You can impute values for the missing data points by using a mean or median for the variable. Or, maybe use a decision tree, other contextual information, or models. For variables with large numbers of missing values, you may want to simply eliminate that variable from the dataset. One option is to not do anything. In R, the default way that missing values are handled is to remove the record with a missing value from the computation, if the variable with the missing value is included in the function. Always keep this fact in mind.
  2. Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing. Note: This may be something you do in conjunction with the EDA as you find extreme values.

Then, you can turn your attention to understanding the data more deeply. You were exposed to the EDA ideas and traditions in Module 1. For a full blown modeling project, you would want to exam all of the variables in your data set. Some suggestions for things that you could do are:

* 1. Obtain histograms for each continuous variable
  2. Obtain summary statistics, such as: Means, standard deviations, minimum, maximum, median for all continuous variables
  3. Are the explanatory variables correlated to the response variable?
  4. Are the explanatory variables correlated amongst themselves?
  5. Obtain scatterplots of explanatory variables with the response variable.
  6. Do you want to create new variables to make the analysis more easily interpretable? For example, you might want to create a variable like PRICE per SQR FOOT. This could be a more meaningful response variable than total home sale price. I’m sure with a little bit of google searching you can find other variables that you would want to compute and potentially use. This is totally voluntary on your part. Not at all required. Do this if you have the interest or think such variables might be of value.

The amount of preparation and EDA is totally up to you. From prior experience, up to 90% of ones time modeling data is spent on data cleaning and preparation issues, depending on the type of data one is working with. Just remember: Garbage in 🡪 Garbage out! For this assignment, you want to be sure you have a dataset that you are comfortable working with for the remainder of this assignment. All of the statistics and graphs you produce are for you. They will be helpful as you go through the following steps, but you do not need to report anything here. There is nothing that needs to be written for task (0).

*PART A: Simple Linear Regression Models*

1. Let Y = sale price be the dependent or response variable. Select “the best” continuous explanatory variable from the AMES data set to predict Y. What criteria did you use to select this variable? Fit a simple linear regression model using X to predict Y. Call this Model 1. You should:
   1. Make a scatterplot of Y and X and overlay the regression line on the cloud of data.

Chart, scatter chart

Description automatically generated

* 1. Report the model in equation form and interpret each coefficient of the model in the context of this problem.

A screenshot of a computer

Description automatically generated with low confidence

Y=14913.4+110.9\*B1

Slope of the linear model is 110.9, and 14913.4 is the intercept.

The intercept is what the Prediction would equal if TotalFloorSF was 0.

So, the TotalFloorSF multiplied by 110.9 + the intercept will give us an estimated calculation of the SalePrice variable Y.

* 1. Report and interpret the R-squared value in the context of this problem.

Multiple R-squared: 0.5622, Adjusted R-squared: 0.562

This is how well the TotalFloorSF variable will predict the SalePrice variable.

* 1. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

Graphical user interface, text, application, email

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Coefficient Model:

Null: B=0

Alt: B != 0

T-stat: 110.851/2.054=53.968

n-p-2=2270-1-2=2267

99% confidence interval t-stat with alpha=0.01: t(0.01/2,267)= 2.576

53.968>2.326

We can reject the null hypothesis that B is not equal to zero.

Omnibus Model:

SS Regression = 6,206,700,000,000

SS Error = 4,832,700,000,000

SS Total = 6,206,700,000,000+4,832,700,000,000= 1.1039\*(10^13)

F-stat = (SSR/p)/(SSE/(n-p-1)=6,206,700,000,000/(4,832,700,000,000/2268)=2912.8221

Pr>F=2.2e-16, which is <0.0001

We can still reject the null hypothesis that the B1 slope is equal to 0, meaning that the B variable does have a significant amount to do with the model and the SalePrice variable.

* 1. The validity of the hypothesis tests is dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. To assess this, use the model from part a) to calculate predicted values for each record. Then use the predicted values to compute residuals. Yes, many of the packages automatically give you the predicted and residuals, but you should know how to code and compute these values. Next standardize the residuals but subtracting off the mean and dividing by the standard deviation for each residual (i.e. you will have to obtain those summary statistics first).

Graphical user interface, text, application

Description automatically generated

Check on the underlying assumptions by plotting:

* + - Histogram of the standardized residuals

Chart, histogram

Description automatically generated

* + - Scatterplot of standardized residuals (Y) by predicted values (X)

Chart, scatter chart

Description automatically generated

Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or influential points?

The data of the standardized residuals look to be quite normal according to the histogram of it, as there is almost no skew.

There is, however, a fair bit of heteroskedasticity in the residual data, as the errors seem to fan out over time, as seen in the scatterplot of the residuals compared to the prediction values. The point as to where the heteroskedasticity seems to increase is at around the $250,000 point.

Independence seems to be implied on this dataset, as it is cross-sectional data.

1. Let Y = sale price be the dependent or response variable. Use the OVERALL QUALITY variable as the explanatory variable (X) to predict Y. Fit a simple linear regression model using X to predict Y. Call this Model 2. You should:

Graphical user interface

Description automatically generated

* 1. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.

Chart

Description automatically generated

* 1. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Is there anything different about the interpretation of coefficients here as opposed to those of Model 1? Can you say a 1 unit change in X is the same across all possible values of X?

Y=-83302+42778\*B1

The intercept is -83,302 and slope=42,778 for this model.

In this linear model, since I removed the 1 and 2 OverallQual ratings due to lack of data, the model will begin at 3. If the OverallQual rating was below 2, the price of the house would be negative, which does not make sense in this case, and is hence just a placeholder for that case. I cannot say that a 1 unit change is the same across all variables as many of the prices do overlap with the nearby ratings.

* 1. Report and interpret the R-squared value in the context of this problem.

Text

Description automatically generated

The r^2 value is 0.6271, which is better than the R^2 value of the last model.

This is how well the X variable predicts the Y variable.

* 1. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

Text

Description automatically generated

Text

Description automatically generated

Coefficient model:

42778/692.7=61.7554

99% confidence interval t-stat with alpha=0.01: t(0.01/2,267)= 2.576

61.7554>2.326

We can reject the null hypothesis that B is not equal to zero.

Omnibus Model:

SS Regression=6.9225e+12

SS Error= 4.1169e+12

F-stat=(6.9225e+12/1)/ (4.1169e+12/2268)=3813.6049

Pr(>F) is < 2.2e+16

We can still reject the null hypothesis that the B1 slope is equal to 0, meaning that the B variable does have a significant amount to do with the model and the SalePrice variable.

* 1. Check on the underlying assumptions. You can do this by hand or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or influential points?

Graphical user interface, application

Description automatically generated

Chart, histogram

Description automatically generated

Diagram

Description automatically generated

The data does have a slight right skew, but this could have been eliminated with better cutoffs for outliers. I had cut off values from the left side and not the right, so it should still be considered normal.

There does appear to be fanning out of the errors in this model down the line, and it starts earlier. However, they do appear to vary quite a bit between themselves, meaning there is still some heteroskedasticity.

1. Of the above 2 models, which one fits better? On what criteria are you assessing the model fit?

Model2 seems to be the models that fits better compared to model 1. This is due to higher R^2 values and a lower standard deviation on model 2. Model 2 also has a lower sum of squared errors when analyzing the residuals.

*PART B: Multiple Linear Regression Models*

1. Fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These two explanatory(X) variables should be: the explanatory variables from Model 1 and Model 2 above. Call this Model 3. You should:
   1. Report Model 3 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression models above?

Graphical user interface, text

Description automatically generated

Y=-95236.76+65.01\*B1+29108.1\*B2

B1 is the variable for TotalFloorSF and B2 is the variable for OverallQual, while -95236.76 is the intercept, or the Y value for if both input variables were zero.

This model is different as it is a multi-linear regression model. Two coefficients are taken into account as opposed to one in each of the previous models.

* 1. Report and interpret the R-squared value in the context of this problem. Does this multiple linear regression model fit better than the simple linear regression models? How do you know? Calculate the difference between R-squared for Model 3 and R-squared for Model 1. How would you interpret this difference?

Text

Description automatically generated

R^2 for this model=0.7564; Adjusted R^2=0.7562

It does fit better than the standard linear regression models, as there adjusted R^2 value, which takes into account the artificial increase in R^2 when adding another variable, is significantly higher as well.

Difference: Model3-Model1=0.7564-0.5622=0.1942

* 1. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

Text

Description automatically generated

Text

Description automatically generated

Coefficient B1:

Null: B1=0

Alt: B1 != 0

T-stat: 65.015/1.874=34.6932

95% confidence interval=1.96

34.6932>1.96

Can reject null hypothesis that B1=0

Coefficeint B2:

T-Stat: 29108.103/684.642=42.5158

95%CI=1.96

42.5158>1.96

Can reject null hypothesis that B2=0.

Null: B1=B2=0

Alt: B1 != B2 != 0

Omnibus:

F-stat=3520 on p=2 and 2267 DF; p-value is <2.2e-16, so can reject null hypothesis that B1=B2=0

* 1. Check on the underlying assumptions. You can do this by hand, or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or points of concern?

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

The data appears to be normal with almost no skew in any direction.

The heteroskedasticity has also significantly decreased with both of the variable combined, with the errors fanning out at a later price vale of arount$320,000. Some outliers exist, as seen in the scatterplot, with the low point before the $400,000 mark, as well as some after it.

* 1. Based on this information, should you want to retain both variables as predictor variables of Y? Discuss why or why not.

Yes, I would want to retain both variables as predictor variables of Y, because of the higher R^2 value gotten from this model, the more normal appearing data, as well as the low heteroskedasticity in this data as well. It seems to be a lot more accurate in its predictions.

1. Select any other continuous variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be your variable of choice plus the explanatory variables from Model 3. Call this Model 4. You should:

A screenshot of a computer

Description automatically generated with low confidence

* 1. Report Model 4 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression models above?

Y=-43851.81+(68.08\*B1)+(22744.83\*B2)+(-463.33\*B3)

-43851.81 is the intercept, 68.08 is the coefficient for B1 or TotalFloorSF, 22744.83 is the coefficient for B2 or OverallQual, and -463.33 is the coefficient for B3 or HouseAge.

This model has three coefficients compared to the two in the last model and 1 in the one before.

* 1. Report and interpret the R-squared value in the context of this problem. Does this multiple linear regression model fit better than the simple linear regression models? How do you know? Calculate the difference between R-squared for Model 4 and R-squared for Model 3. How would you interpret this difference? Does your variable of choice help to improve the model’s explanatory ability?

Multiple R-squared: 0.7826, Adjusted R-squared: 0.7823

This one also appears to fit slightly better than the last model because it has a greater adjusted R^2 value, which takes into account the inflation of the R^2 value that occurs when adding another variable into the model.

Model4-Model3=0.7826-0.7564=0.0262

This is a slight increase in R^2 from the last value, indicating it may better predict the SalePrice value than the previous model.

* 1. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

Text

Description automatically generated

Coefficient B1:

Null: B1=0

Alt: B1 != 0

T-stat: 68.08/1.78=38.2472

95% confidence interval=1.96

38.2472>1.96

Can reject null hypothesis that B1=0

Coefficeint B2:

T-Stat: 22744.83/752.91=30.2092

95%CI=1.96

30.2092>1.96

Can reject null hypothesis that B2=0.

Omnibus:

Null: B1=B2=B3=0

Alt: B1 != B2 != B3 != 0

F-stat=2719 on p=3 and 2266 DF; p-value is <2.2e-16, so can reject null hypothesis that B1=B2=B3=0

* 1. Check on the underlying assumptions. You can do this by hand, or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or points of concern?

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

The normality has remained the same and there are no issues there. The homoskedasticity has also slightly decreased from last time as well, improving the accuracy of the model.

Some outliers still exist, and the errors increase after the $300,000 mark, but not as much as the previous model.

* 1. Based on this information, should you want to retain all three variables as predictor variables of Y? Discuss why or why not.

Yes, I would retain all three variables as a predictor because the R^2 and adjusted R^2 value have both increased, and the homoskedasticity in the model has also increased indicating a more accurate model.

*PART C: Multiple Linear Regression Models on Transformed Response Variable*

1. Refit Model 1, Model 3 and Model 4 using the Natural Log of SALEPRICE as the response variable. This is LOG base e, or LN () on your calculator. You’ll have to find the appropriate function using R. Perform an analysis of goodness-of-fit to compare the Natural Log of SALEPRICE models to the original models. Which transformed model fits the best? Do the transformed models fit better than the original models? You do not need to report all of the output like was done in Parts A and B. Rather, you should construct a table to summarize your findings so that the comparisons can be made easily. What is the best way or statistic to use, to make comparisons between models? You may need more than one table to do this adequately, if you have more than 1 criteria.

|  |  |  |
| --- | --- | --- |
| **Model** | **R^2** | **R^2 Adjusted** |
| Model 1 | 0.5622 | 0.562 |
| Log Model 1 | 0.5631 | 0.5629 |
| Model 3 | 0.7564 | 0.7562 |
| Log Model 3 | 0.7726 | 0.7724 |
| Model 4 | 0.7826 | 0.7823 |
| Log Model 4 | 0.8187 | 0.8185 |

|  |  |
| --- | --- |
| Model | F-Stat |
| Model 1 | 2913 |
| Log Model 1 | 2923 |
| Model 3 | 2267 |
| Log Model 3 | 3852 |
| Model 4 | 2719 |
| Log Model 4 | 3411 |

As you can see in the charts above, the R^2 and adjusted R^2 values are higher in the models with the regression done on the log of SalePrice compared to the ones without. This means that the data is better fitting on the log(SalePrice) models than the ones without. The F-stat is also higher in the models with regression done on the log(SalePrice) compared to the ones without. The higher F-Stat value indicates the X variables become more significant on the Y variables.

1. How is the interpretation of the LN(SalePrice) models different from the SalePrice models? Discuss if the improvement of model fit justifies the use of the Log(SALEPRRICE) response variable, relative to interpretation and explanation to a non-technical audience, like your manager or other executives.

LN(SalePrice) causes a dataset do become more normalized and reduces the skew or exponentiality of it.

LN(SalePrice) offers a layer of standardization of the output variables to a certain range. Hence, it allows the model do be fit to a certain range as well and potentially work better. It allows the dataset to become more normalized and can then cause us to get smaller errors in our model.

*PART D: Multiple Linear Regression and Influential Points*

1. Use Model 4 for this part. Even after you have cleaned your data, still you may have unusually large residuals, which you can see from the residual plots. These are called ‘influential’ points. Sometimes, we find that a small subset of ‘influential’ points exerts a disproportionate influence on the model coefficients. These points can be identified by several statistics such as DFFITS, Cook’s Distance, Leverage, and Influence. Fit Model 4 using a regression function from one of the comprehensive regression packages (like lessR). Obtain output data with these statistics (DFFITS, etc.) for individual records so that you can identify the influential points. Use the threshold value given in the text book (Like that on Page 112 of Chatterjee and Hadi). Then refit the model after removing the influential points. How many influential points did you find & remove? When you refitted the model, did the model improve? The other side of the coin is that if you remove data points due to them being “influential” and not looking like you might want them to look, some would argue that such an action is the modeler biasing the data. Comment on whether or not you find the improvement of model fit justifies the potential for the modeler biasing the result by removing potentially legitimate data points.

If I used the method proposed by the book, none of the entries would be removed. However, if I use another common threshold, such as 4/number of entries, I will remove 122 entries from my data.

When those entries were removed, the R^2 value of my data did improve. It rose to 0.8223. Hence, I do think the removal of the values was beneficial. The p-value, however, did increase as well. This made the model more accurate than the log(SalesPrice) model, but without standardizing.

Graphical user interface, text, application

Description automatically generated

*PART E: Beginning to Think About a Final Model*

1. Use Model 4 to start with for this part. So far, we have fit a few models to predict SALEPRICE(Y). But, there are many other continuous variables in the data set. You could use theory, or your background knowledge, to select variables for inclusion in a multiple regression model. Many modelers do this. It gives a nice place to start the search process. On the technical side, in this assignment, we have been looking at change in R-squared when a new variable has been added to an existing model to isolate the explanatory contribution of that new variable. And, we have been looking at hypothesis tests on the individual coefficients.

Use the concept of Change in R-squared, plus anything else you wish, to put together a reasonable approach to find a good, comprehensive multiple regression model to predict SALEPRICE(Y). Any of the continuous variables can be considered fair game as explanatory variables. This can feel like an overwhelming task. You don’t need to go overboard, or kill yourself, in doing this. We will learn about automated approaches to do this shortly. But, for now, I’d like you to think about how you would do this by hand.

Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset. For this task you need to:

1. Explain your approach

My approach would be to take all of the aspects applied to the most recent model, including eliminating influential factors, and adding more continuous variables to the model in order to get the best prediction. I will also make sure my variables encompass most of the aspects of the property. For that I chose the variables, TotalFloorSF, OverallQual, HouseAge, and LotArea. This would be in addition to the preprocessing conducted before this assignment began.

1. Report the model you determined and interpret the coefficients

I chose the variables, TotalFloorSF, OverallQual, HouseAge, and LotArea. This would be in addition to the preprocessing conducted before this assignment began.

1. Report the coefficient and ANOVA tables.

Table

Description automatically generated

1. Report goodness of fit

The R^2 value is the highest so far at 0.8441, with the adjusted R^2 value at 0.8438. This is a pretty good fit to the data according to the R^2.

1. Check on underlying model assumptions.

Chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated

The data still appears normal, and the homoscedasticity has increases quite a bit compared to the last models, with the fanning out further delayed, and no significant outliers.

*CONCLUSION / REFLECTION*

Please write a conclusion / reflection section that, at minimum, addresses the questions:

* In what ways do variable transformation and outlier deletion impact the modeling process and the results?
* Are these analytical activities a benefit or do they create additional difficulties?
* Can you trust statistical hypothesis test results in regression?
* What do you consider to be next steps in the modeling process?

This assignment was one that has allowed me to learn quite a bit about preprocessing of data and its impact on the models you will create. The transformation and outlier deletion impacts the modeling process quite significantly. It allows for the model to be quite more accurate, as shown by the increase in R^2 and adjusted R^2 values. This may be controversial, as it may get rid of some legitimate entries in the data, but this may allow pro predictions on statistically normal data to be more accurate. These analytical activities are a benefit, as they allow for better performing models on normal data and could allow for better studies and predictions based on these models with that accuracy. I believe you can trust statistical hypothesis tests in regression as they show how impactful the data may really be to the model, and how a lot of factors go into the dependent variable in the models. The next steps in the modelling process should be working with alternative regressions and statistical inferences, as well as streamlining the models to even better their performance.

**Assignment Document:**

Results should be presented and discussed in an organized manner, preferably listed by task number and letter. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment2\_LastName.pdf.