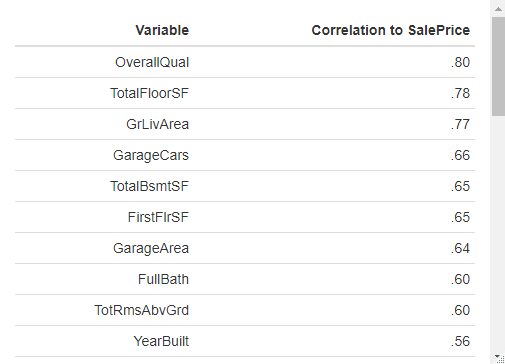
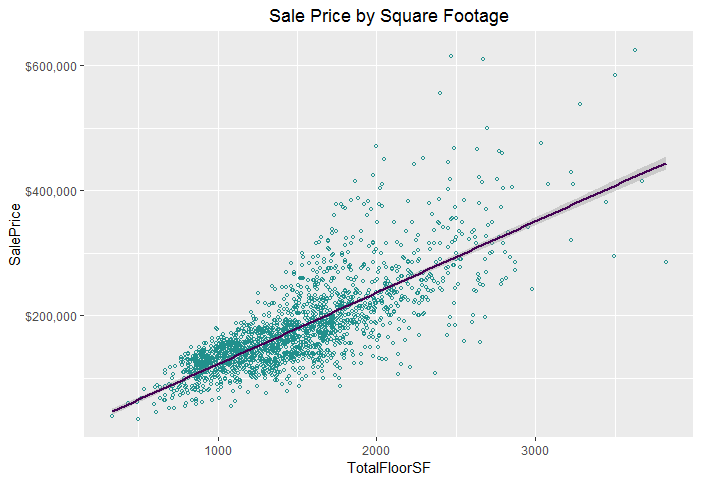


|  |
| --- |
| Modeling Assignment #2: Building Linear Regression Models  *MSDS 410* |

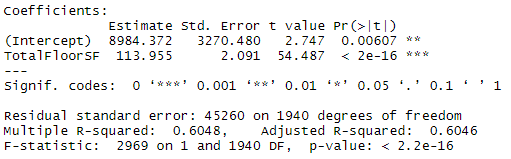
*PART A: Simple Linear Regression Models*

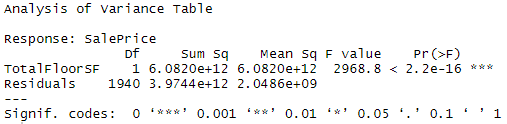
1. I have chosen TotalFloorSF as my explanatory variable to predict SalePrice. The following table shows the 10 non-factor variables with the highest Pearson correlation to SalePrice. TotalFloorSF has the highest correlation after OverallQual (which I have ignored due to it being used in section 2 below.)



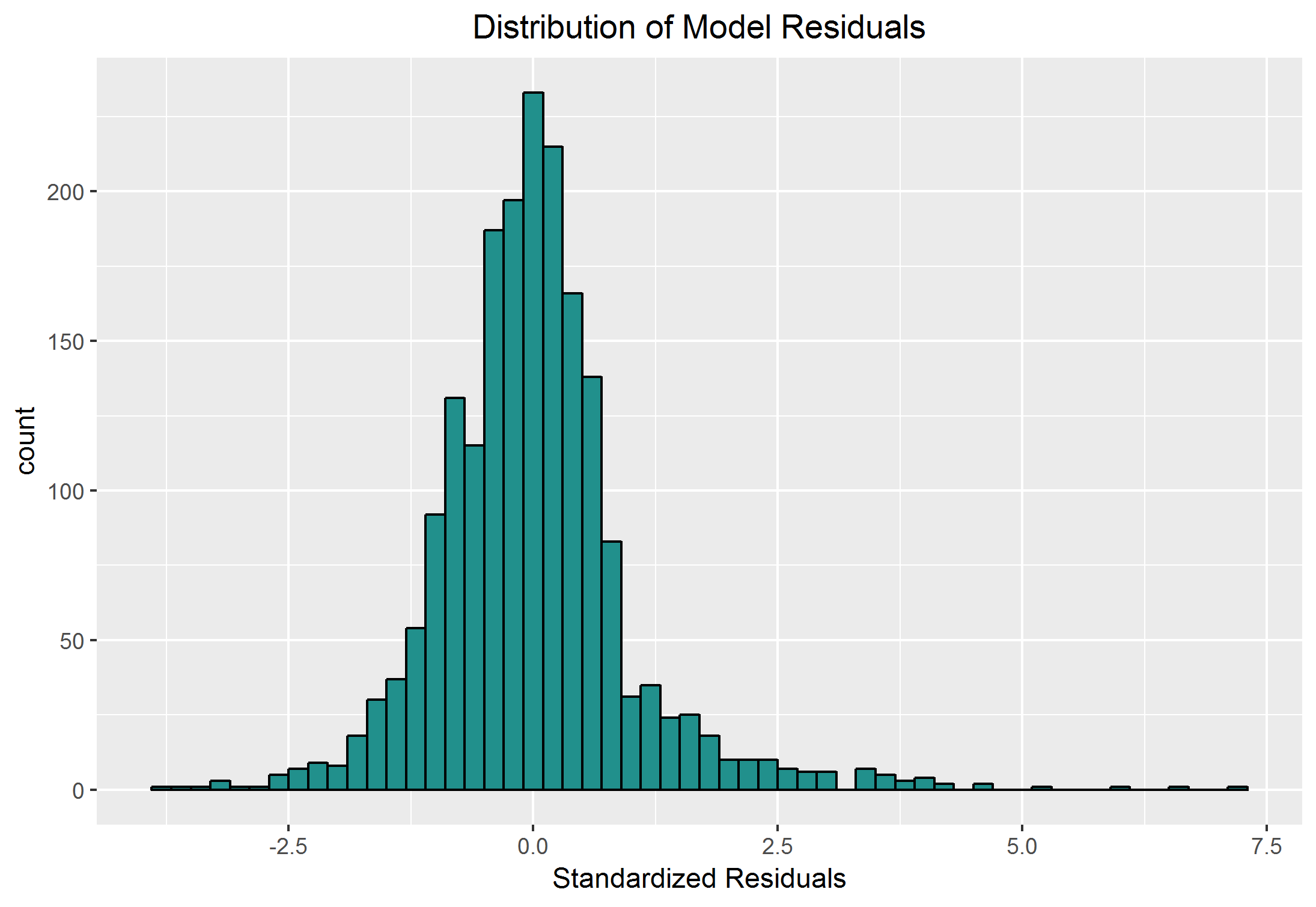


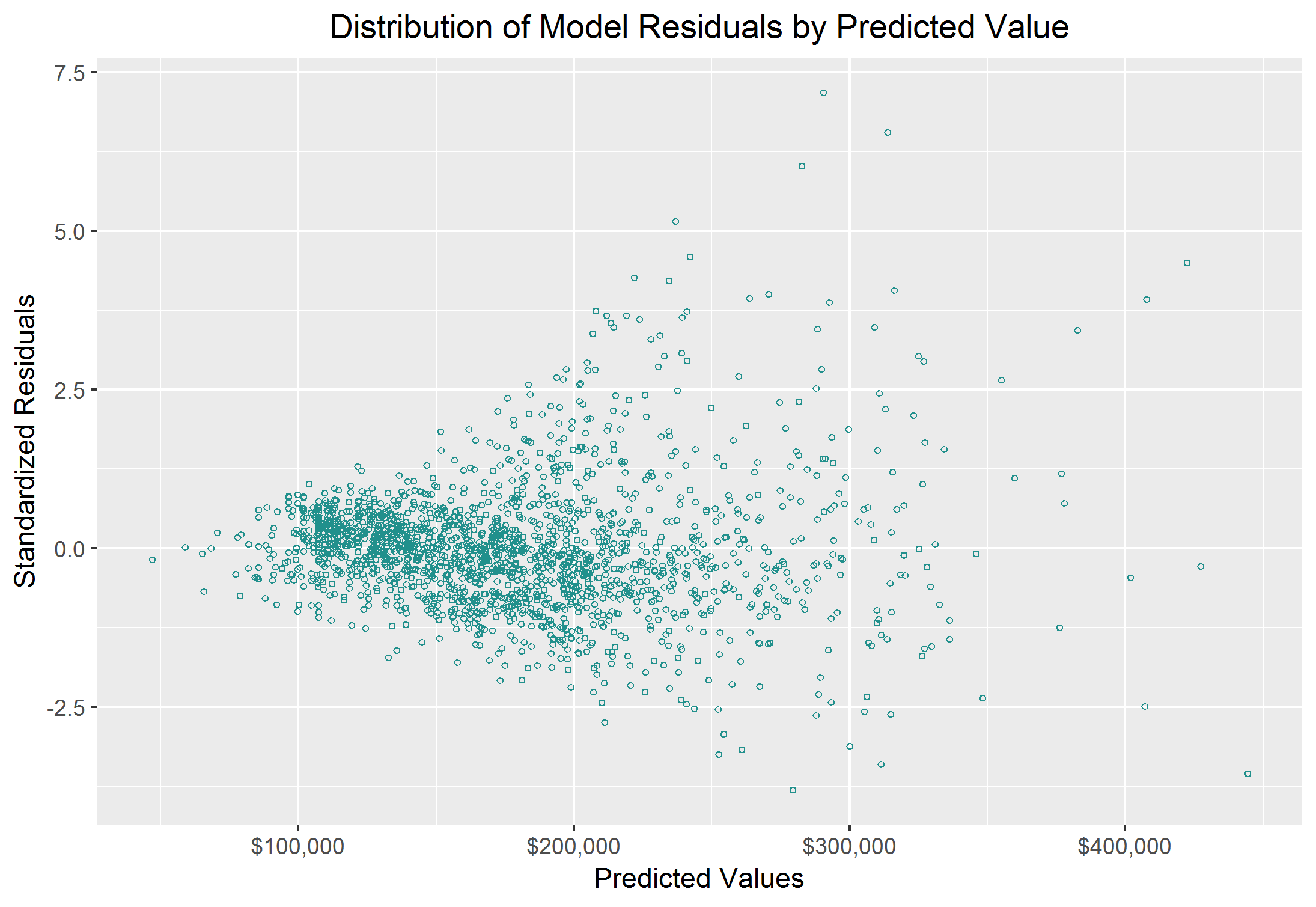
* 1. Y = 113.955\*TotalFloorSF + 8984.372
     + 8984.372: Our intercept is the value of a home that has 0 total square feet. This is outside of the realistic parameters of our model and could not be used to estimate an empty plot of land. It simply serves as the y-intercept of the line of best fit.
     + 113.955\*TotalFloorSF: For every square foot increase in size of the home, the predicted price increases by $113.96.
  2. R-Squared: 0.605
     + This model accounts for 60.5% of the variance in SalePrice.





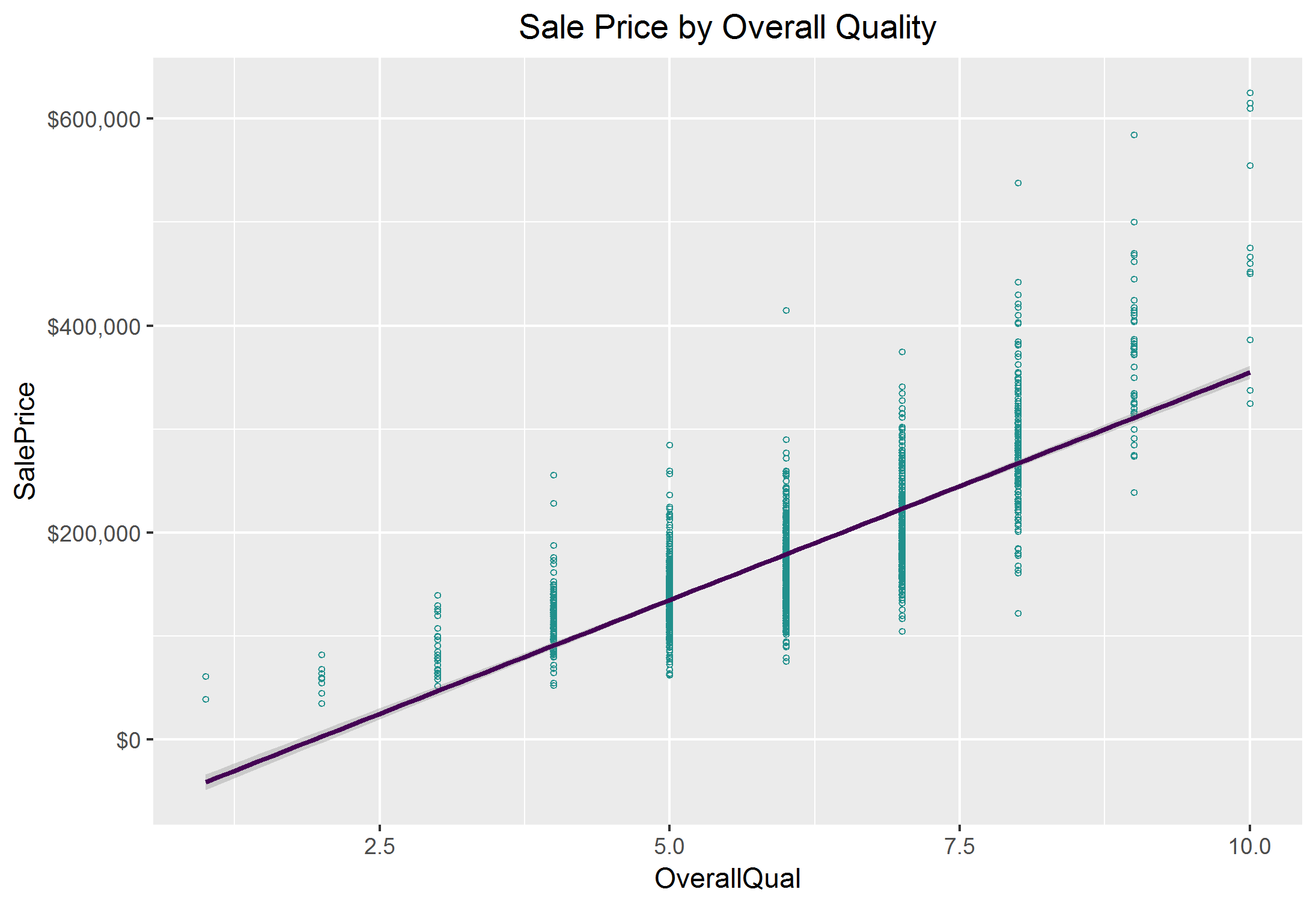
* Intercept Hypothesis Test:
  + H0: β0 = 0
  + HA: β0 does not equal 0
  + Reject the null hypothesis at the 0.001 level
* TotalFloorSF Hypothesis Test:
  + H0: β1 = 0
  + HA: β1 does not equal 0
  + Reject the null hypothesis at the 0 level
* Omnibus Hypothesis Test:
  + H0: All betas included in model (in our case only one) = 0
  + HA: At least one beta does not equal 0
  + We reject the null hypothesis and conclude that our model significantly explains the variance of SalePrice.
  1. The validity of the hypothesis tests are 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). Check on the underlying assumptions by plotting:



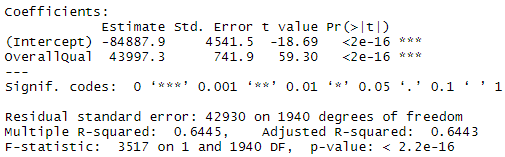


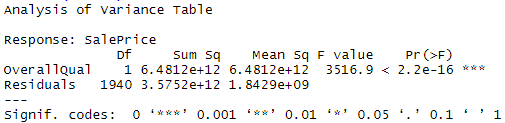
In the scatter plot graph, there is a clear fan shape formed by the residuals, which suggests a violation of the assumption of homoscedasticity. There appear to be some extreme outliers, which can be observed in the far right tail of the histogram. While these outliers could strongly influence our model, even without those points there is still a general fan shape of the standardized residuals as the predicted sale price increases. The issue that this causes is that as our predicted price increases, the observed and expected error of our model increases. The higher predictions are less accurate than the lower predictions.

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:

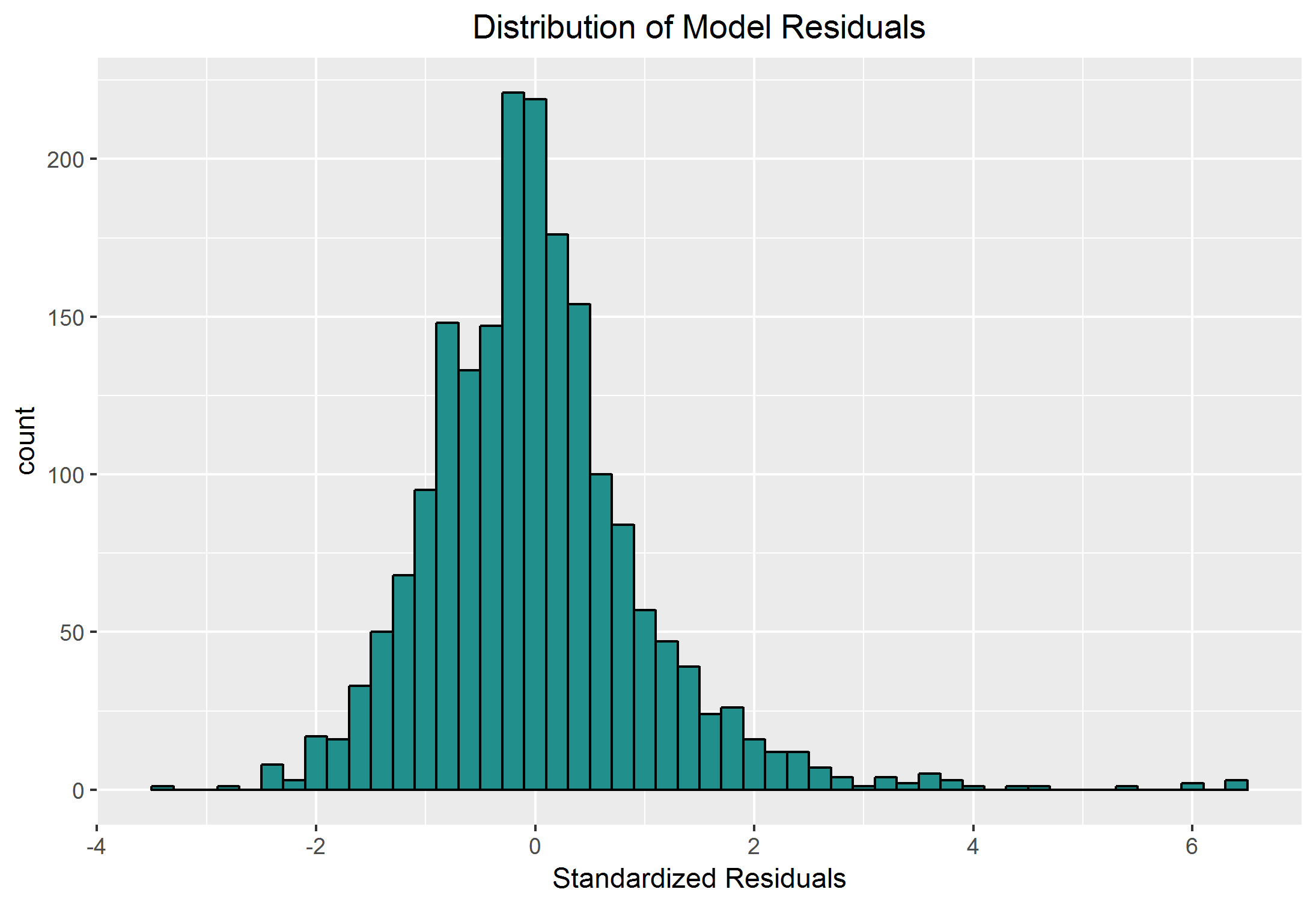


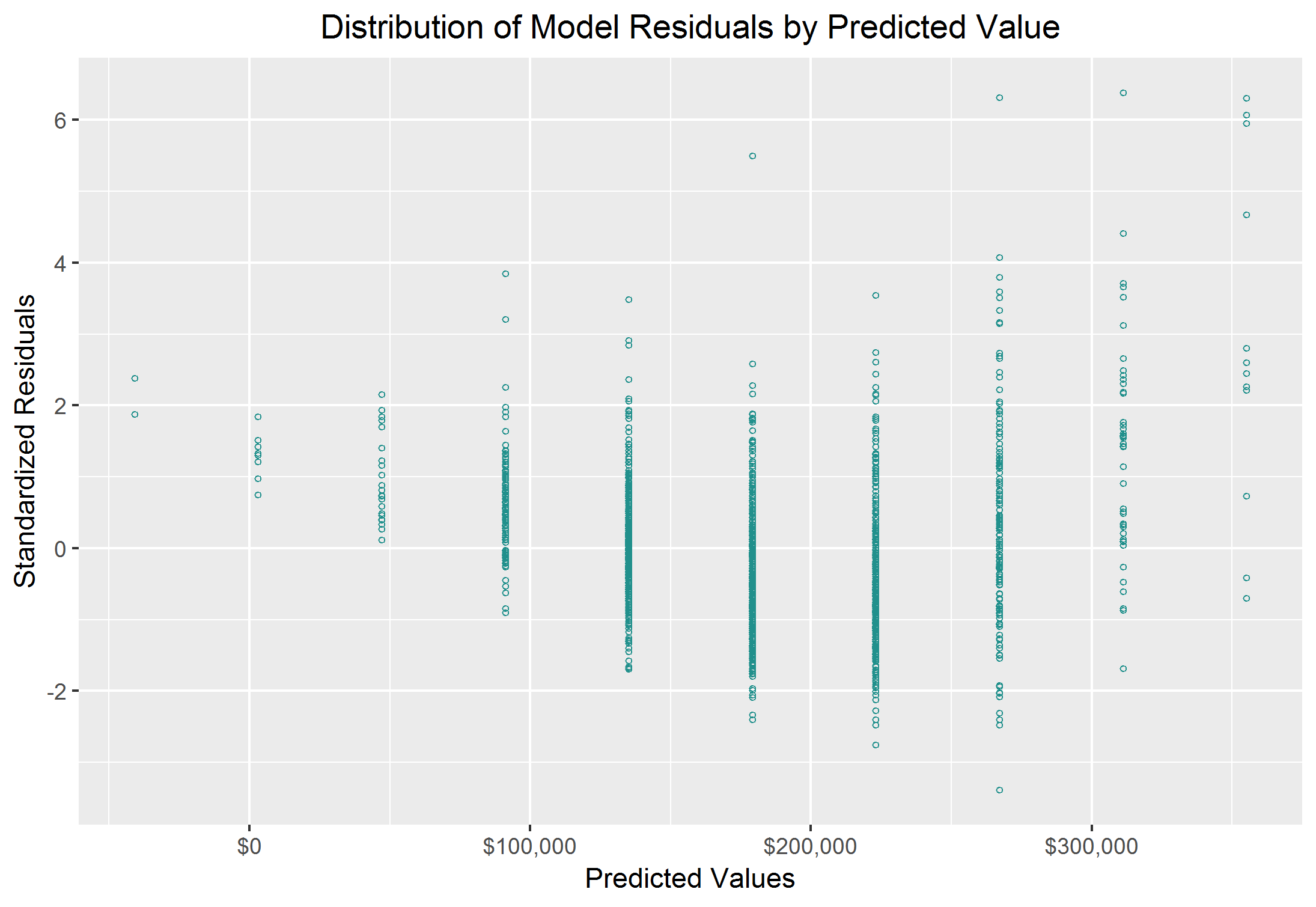
* 1. Y = 43997.254\*OverallQual - 84887.949
     + -84887.949: Our intercept is the value of a home that was rated a 0 in Overall Quality. This is outside of the realistic parameters as the lowest quality rating observed is 1. It simply serves as the y-intercept of the line of best fit.
     + 43997.254\*OverallQual: For every one point increase in Overall Quality, the predicted price increases by $43,997.25.
     + OverallQual is an ordinal variable, unlike TotalSqFt form model 1, which was totally continuous. Each level of Overall Quality has sale prices at many different values. For prediction purposes this poses a challenge because this model will assume that for a given Quality score each home will have the exact same sale price.
  2. R-Squared: 0.644
     + This model accounts for 64.4% of the variance in SalePrice.





* Intercept Hypothesis Test:
  + H0: β0 = 0
  + HA: β0 does not equal 0
  + Reject the null hypothesis at the 0 level
* OverallQual Hypothesis Test:
  + H0: β1 = 0
  + HA: β1 does not equal 0
  + Reject the null hypothesis at the 0 level
* Omnibus Hypothesis Test:
  + H0: All betas included in model (in our case only one) = 0
  + HA: At least one beta does not equal 0
  + We reject the null hypothesis and conclude that our model significantly explains the variance of SalePrice.



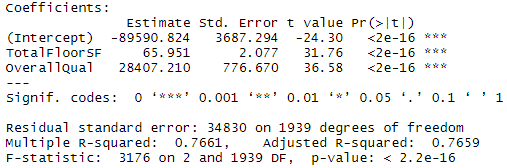


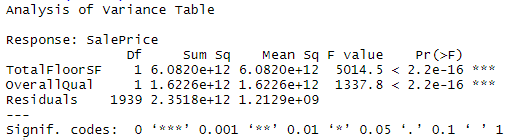
The residuals in the scatter plot seem to show a concave fanning pattern over the predicted sales price variable, where they trend more negative as the price increases until trending back positive over the last three quality levels. This shows a heteroscedastic nature for the residuals of this model as well. This histogram of the standardized residuals is right-tailed, which violates the normality of residuals assumption. There appear to be approximately five outlier results with highly positive standardized residuals and one with very negative standardized residual. These points could strongly affect the model.

1. Model 2 shows a better fit. It has an r-squared value about 4% greater than model 1. The t-statistic on the explanatory variable in model 2 is greater than that of model 1, which suggests a greater statistical significance for that explanatory variable.

*PART B: Multiple Linear Regression Models*

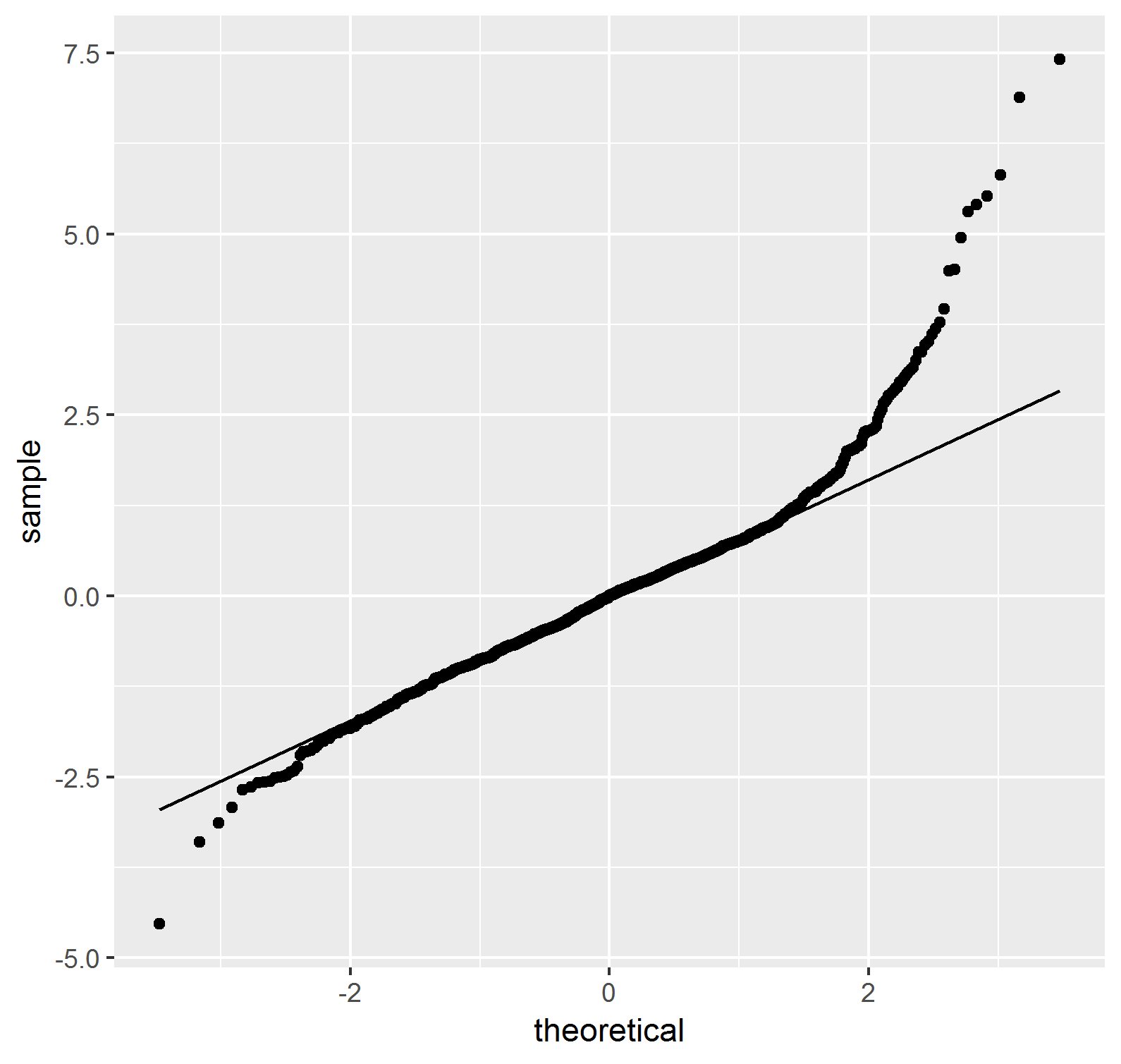
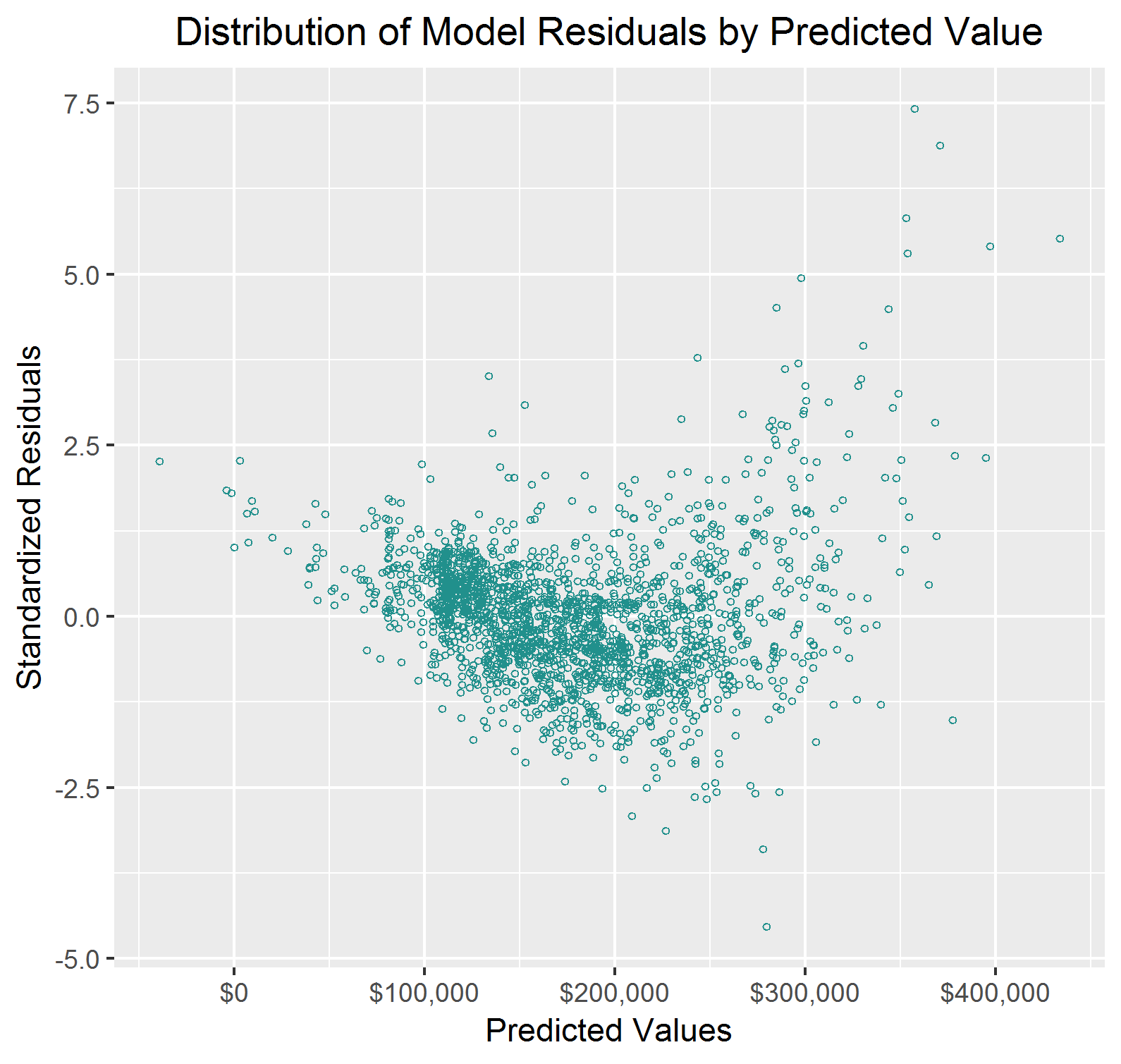
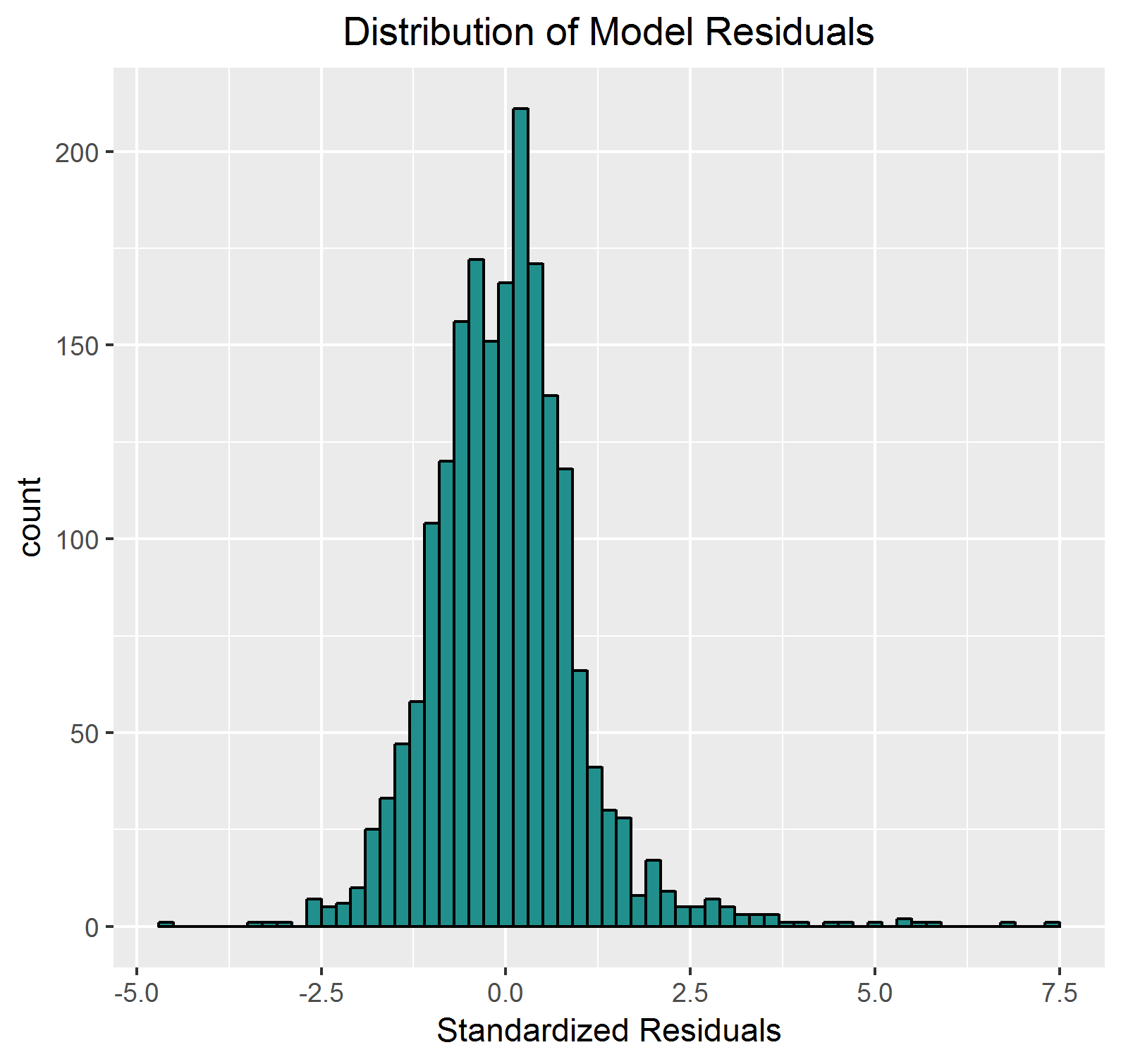
1. Fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These two explanatory(X) variables should be: the explanatory variables from Model 1 and Model 2 above. Call this Model 3. You should:
   1. Y = 65.951\*TotalFloorSF + 28407.21\*OverallQual - 89590.824
      * -89590.824: Our intercept is the value of a home that was rated a 0 in Overall Quality and has zero total square footage. This is outside of the realistic parameters as the lowest quality rating observed is 1 and we do not have any empty plots of land in our sample. It simply serves as the y-intercept of the line of best fit.
      * 28407.21\*OverallQual: For every one point increase in Overall Quality while holding square footage constant, the predicted price increases by $28,407.21.
      * 65.951\*TotalFloorSF: For every increase of one square foot while holding the overall quality of the home constant, the predicted sale price increases by $65.95.
      * In this model both β1 and β2 are lower than they were in their respective simple linear models.
   2. R-Squared: 0.776
      * This model accounts for 76.6% of the variance in SalePrice.
      * This model fits better than either of the simple linear regressions. There is a 0.161 increase in r-squared from model 1 to model 3. That means that our multiple linear regression explains an extra 16% of sale price variance in comparison to the simple linear regression using only total square footage.





Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests. Intercept Hypothesis Test:

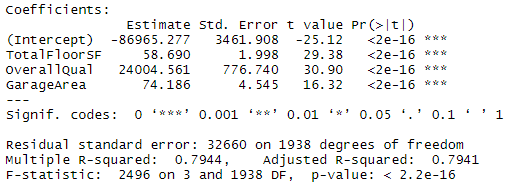
* Intercept Hypothesis Test:
  + H0: β0 = 0
  + HA: β0 does not equal 0
  + Reject the null hypothesis at the 0 level
* TotalFloorSF Hypothesis Test:
  + H0: β1 = 0
  + HA: β1 does not equal 0
  + Reject the null hypothesis at the 0 level
* OverallQual Hypothesis Test:
  + H0: β2 = 0
  + HA: β2 does not equal 0
  + Reject the null hypothesis at the 0 level
* Omnibus Hypothesis Test:
  + H0: β1 = β2 = 0
  + HA: At least one beta does not equal 0
  + We reject the null hypothesis and conclude that at least one βdoes not equal zero.

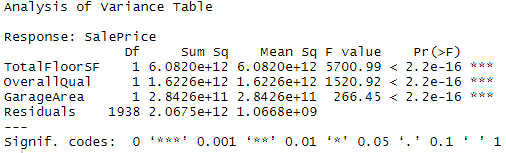


In the scatter plot for this model, it appears that we can still observe the concavity that we saw in the scatter plot for model 2. The extreme outliers in the histogram are also still present, which can also be seen in the Q-Q plot. There are still some extreme outliers affecting the model at the high end of our predicted sale price, however even with those removed it appears our model would violate the assumption of homoscedasticity.

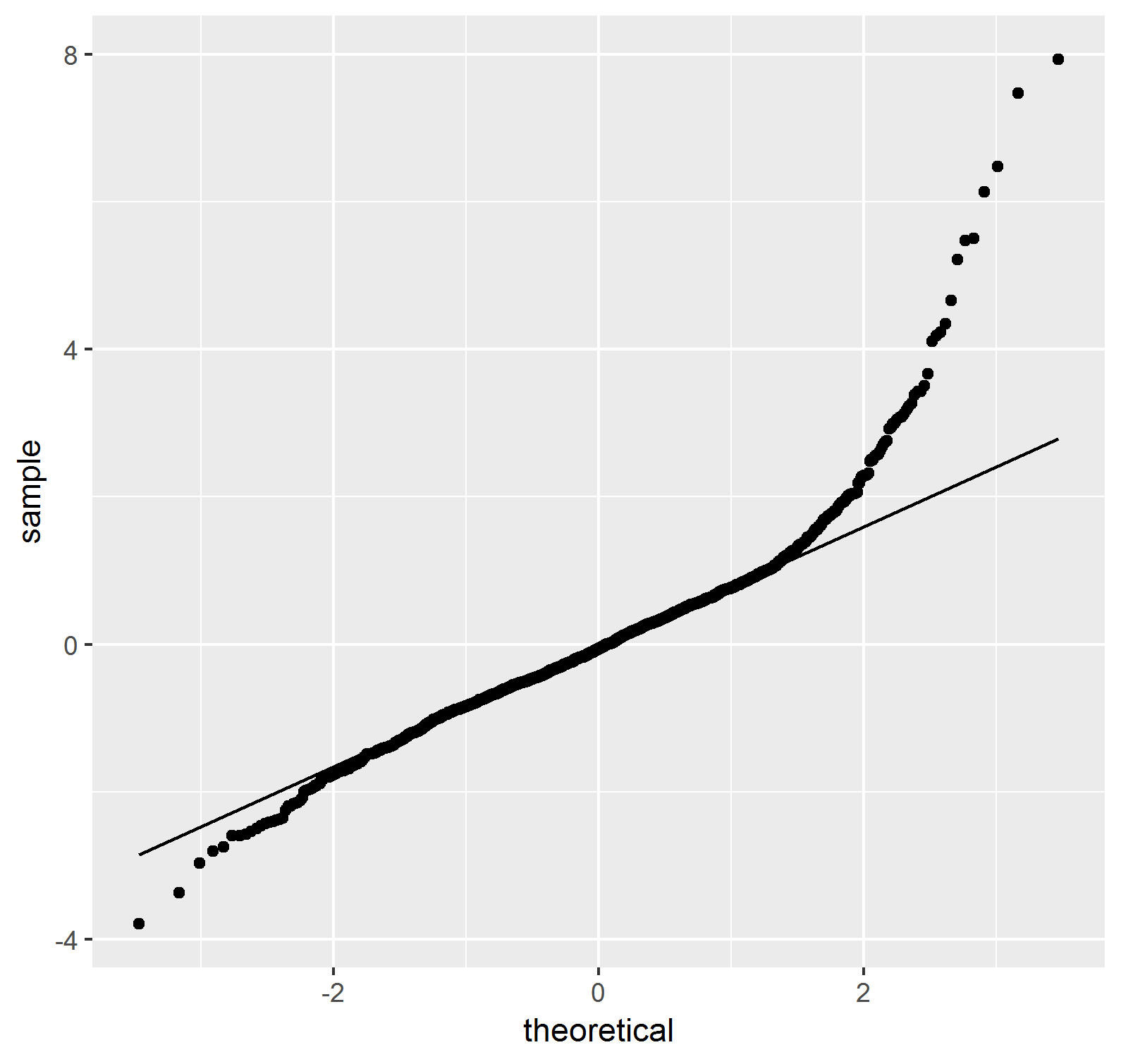
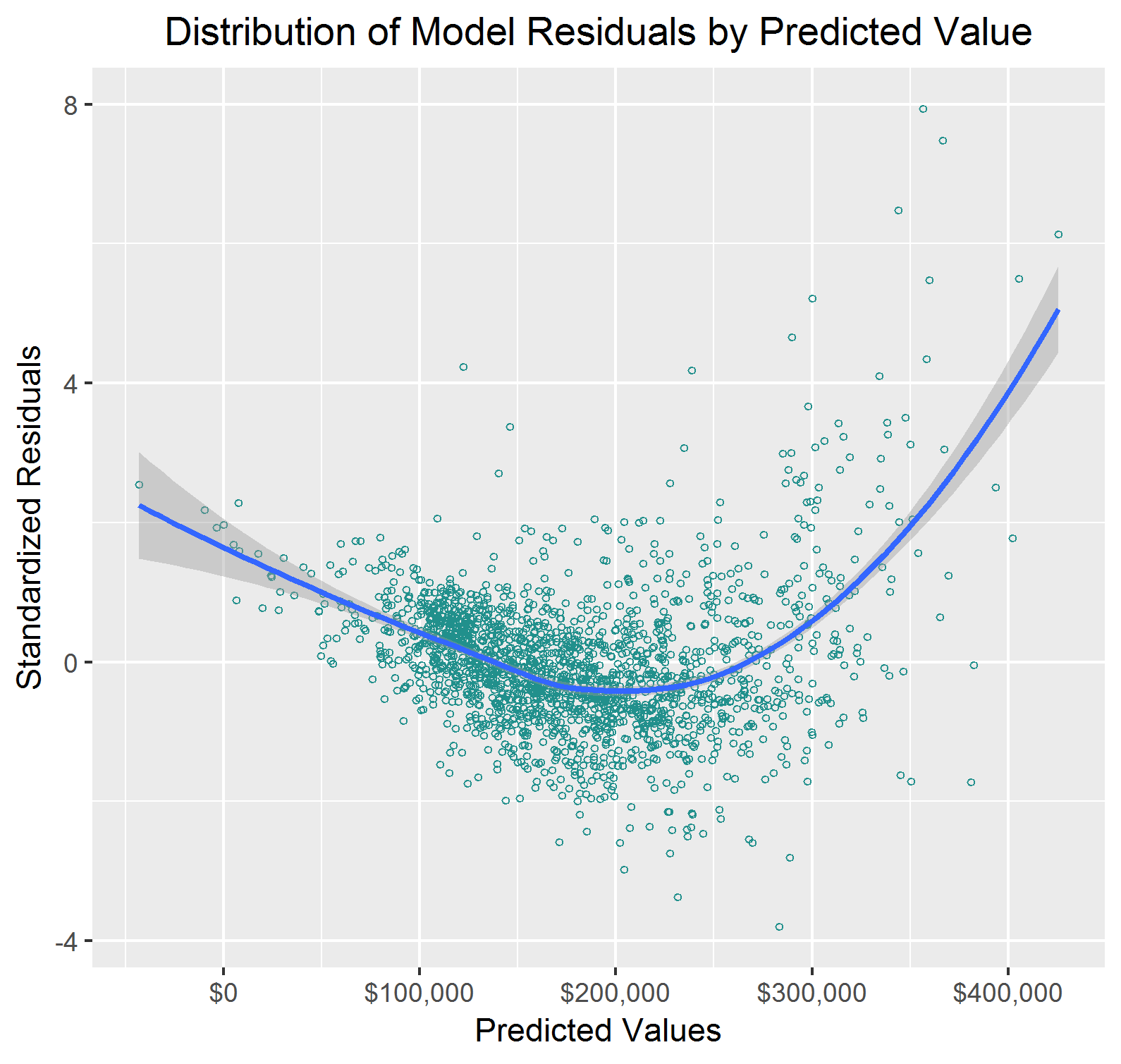
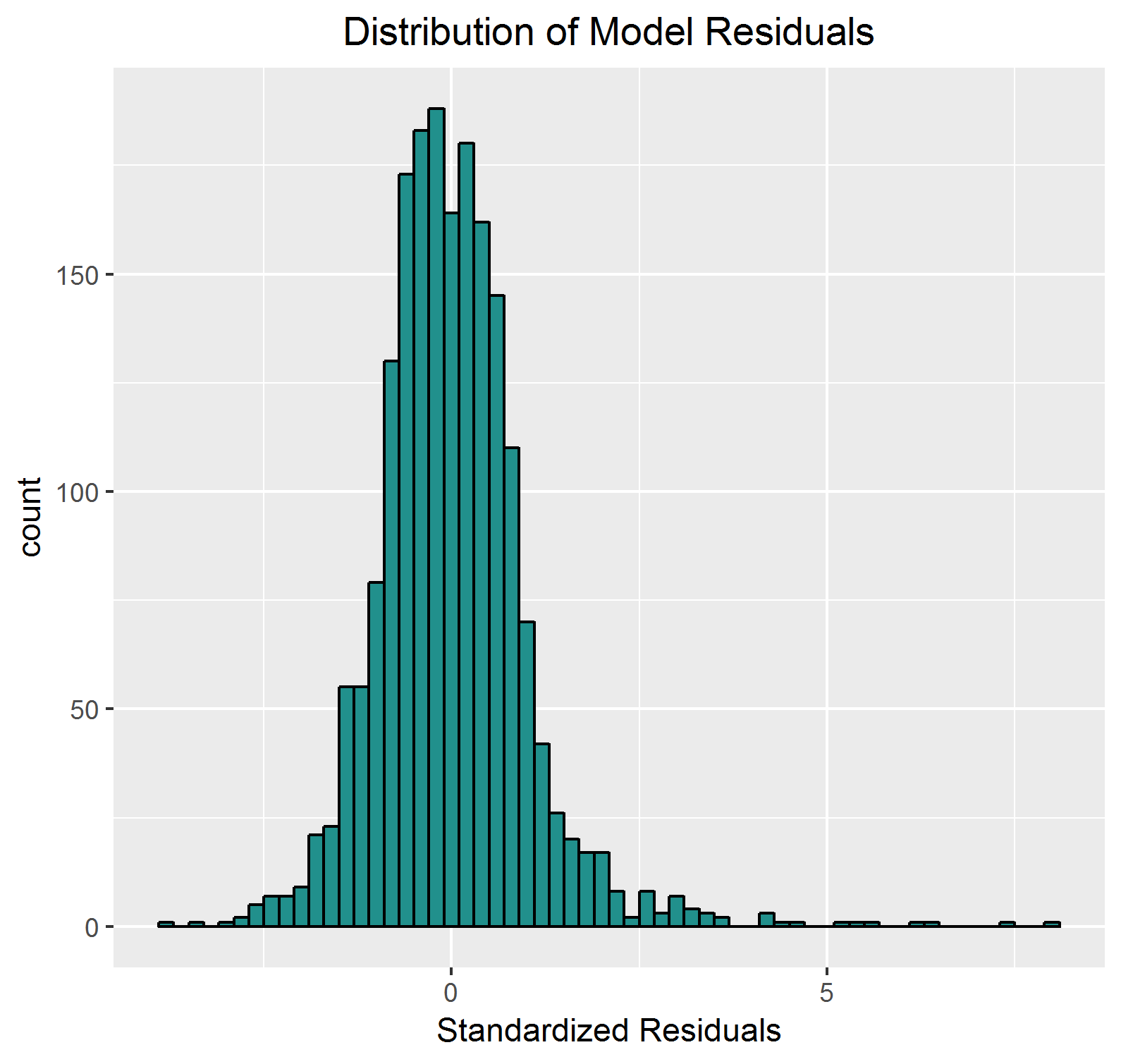
* 1. Based on this information, we should retain both variables as predictor variables of Y. All of the current downfalls of model 3 are also more apparent in the simple linear models. Despite the homoscedasticity violations, the multiple linear regression model is a clear improvement over either simple linear model, which is plainly seen in the large increase in R-squared.

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:
   1. Y = 58.69\*TotalFloorSF + 24004.561\*OverallQual + 74.186\*GarageArea - 86965.277
      * -86965.277: Our intercept is the value of a home that was rated a 0 in Overall Quality and has zero total square footage and no garage. This is outside of the realistic parameters as the lowest quality rating observed is 1 and we do not have any empty plots of land in our sample. It simply serves as the y-intercept of the line of best fit.
      * 74.186\*GarageArea: For every one square foot increase in the size of the garage while holding Overall Quality and house square footage constant, we predict an increase in sale price of $74.18.
      * 24004.561\*OverallQual: For every one point increase in Overall Quality while holding home and garage square footage constant, the predicted price increases by $24,004.56.
      * 58.69\*TotalFloorSF: For every increase of one square foot in home size while holding the overall quality of the home and garage size constant, the predicted sale price increases by $58.69.
      * In this model both β1 and β2 are lower than they were in their respective simple linear models and the previous multiple linear model.
   2. R-Squared: 0.794
      * This model accounts for 79.4% of the variance in SalePrice.
      * This model fits better than either of the simple linear regressions. There is a 0.19 increase in r-squared from model 1 to model 4. That means that our multiple linear regression explains an extra 19% of sale price variance in comparison to the simple linear regression using only total square footage.
      * R-squared increase vs model 3: 0.028. The addition of Garage Area accounts for an additional 2.8% of the variation in Sale Price. This variable provides a modest increase over model 3.





* Intercept Hypothesis Test:
  + H0: β0 = 0
  + HA: β0 does not equal 0
  + Reject the null hypothesis at the 0 level
* TotalFloorSF Hypothesis Test:
  + H0: β1 = 0
  + HA: β1 does not equal 0
  + Reject the null hypothesis at the 0 level
* OverallQual Hypothesis Test:
  + H0: β2 = 0
  + HA: β2 does not equal 0
  + Reject the null hypothesis at the 0 level
* GarageArea Hypothesis Test:
  + H0: β3 = 0
  + HA: β3 does not equal 0
  + Reject the null hypothesis at the 0 level
* Omnibus Hypothesis Test:
  + H0: β1 = β2 = 0
  + HA: At least one beta does not equal 0
  + We reject the null hypothesis and conclude that at least one βdoes not equal zero.

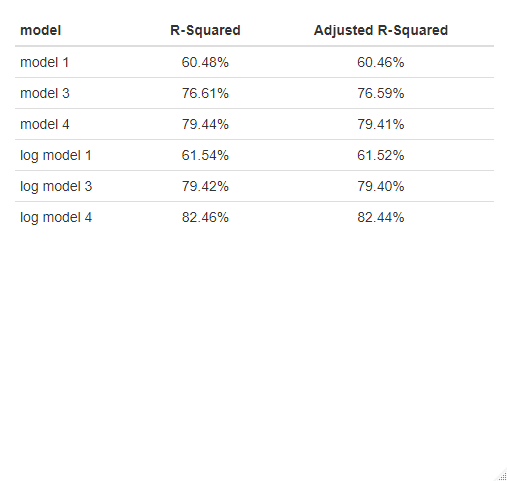


As shown in the LOESS regression line plotted on the scatter plot of standardized residuals, we again see the same trends in the residuals as in previous models. This model violates the assumptions of homoscedasticity in the same manner as previous models. This is also apparent in our QQ plot and histogram. There are outliers that are affecting the model as well.

* 1. I would argue again in favor of keeping all three variables as predictors. All explanatory variables are highly significant and the inclusion of Garage Area results in a 3% increase in adjusted r-squared over model 3.

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

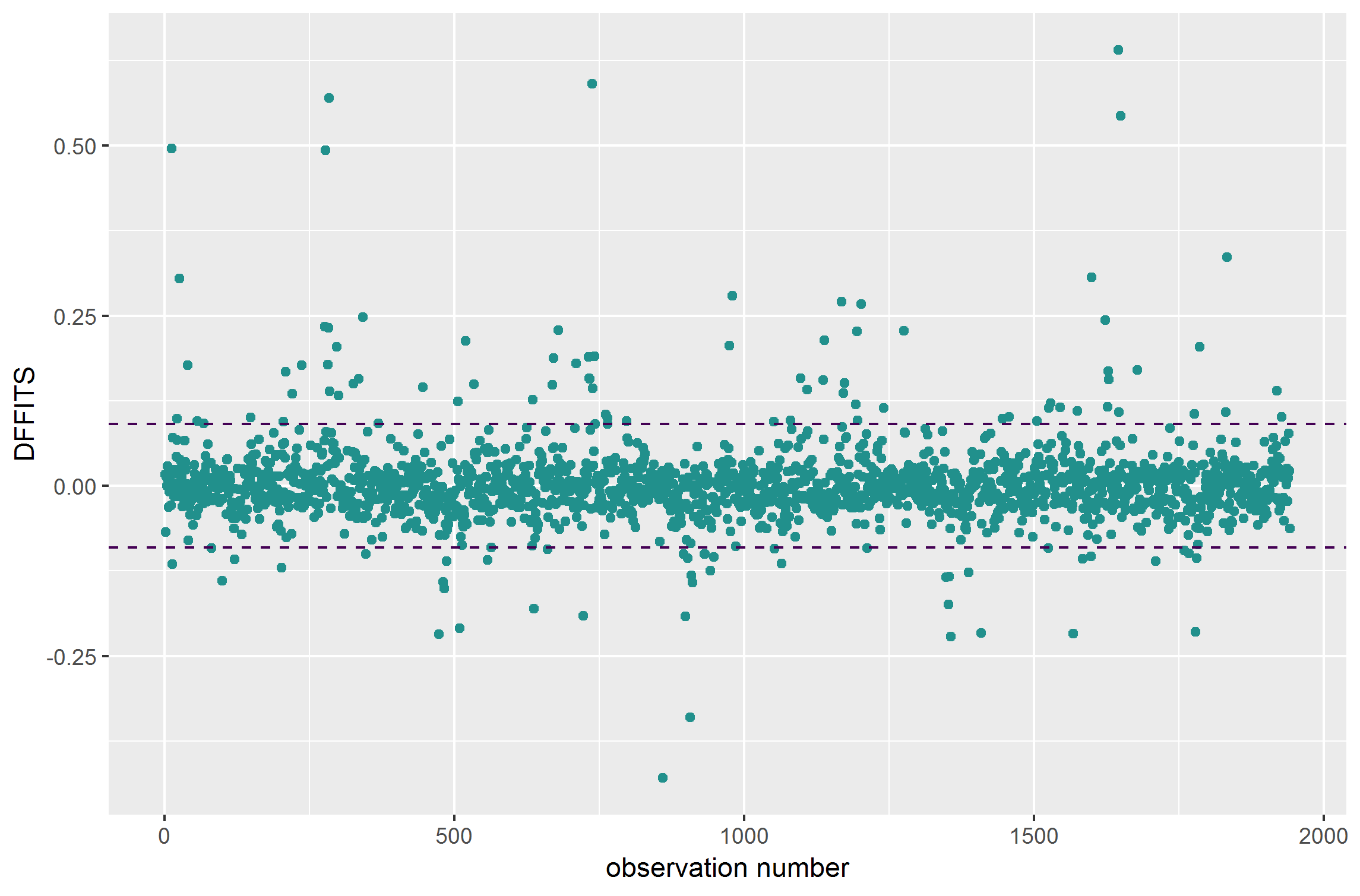
1. The transformed model 4 provides the best fit of all the models we have investigated so far, including showing increases in R-Squared and adjusted R-Squared over the non-transformed models above.



1. The output of our log transformed model can be interpreted as eY = Sale Price. This is different from the non-transformed model because those did not require any further transformation of the output in order to predict the sale price. The increase in predictive power of the transformed model justifies the transformation of the target variable. Any audience with a sufficiently technical background typically does not need to know the finer details in how a prediction is made. They are typically only concerned with how accurate the model ends up being. For any non-technical person who is interested in learning about the inner workings of the model, it is imperative that a data scientist is able to distill the ideas behind the methodology down in a way that is easily understood to a non-technical audience.

*PART D: Multiple Linear Regression and Influential Points*

1. DFFITS threshold: 2\*sqrt((p+1)/(n-p-1)) = 2\*sqrt((3+1)/(1942-3-1)) = 0.09086



As seen in the chart above, we have many influential points based on DFFITS values, 125 total. After the removal of these 125 influential points, model 4 improves substantially. The new R-Squared value is 0.8388, which represents a 4% improvement over the original model 4.

However I do not believe that this improvement justifies the removal of the influential data points. The 125 influential points represents ~6% of the total sample. Removing them simply because their inclusion does not work well with our model feels disingenuous and misleading, especially considering we removed outlier observations before beginning the analysis. Removal after fitting the models post-hoc is simply choosing the data to make the model more convenient to use.

*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
2. Report the model you determined and interpret the coefficients
3. Report the coefficient and ANOVA tables.
4. Report goodness of fit
5. Check on underlying model assumptions.

*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?

**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.