Applied Linear Regression Project Report

By

Charlie Silkin

Analysis 1

Question 1:

The model I chose is a standard OLS model with influential points removed from the data set.

The fitted regression equation of my model is: $\widehat{SalePrice} = 31050.021 + 83.791(GrLivArea)$

My first step was to check the model assumptions for a standard OLS model with SalePrice as the response and GrLivArea as the predictor with no data points removed from the data set. I found that the p-value for the Shapiro-Wilk Normality Test (0.446) was greater than the α = 0.05 significance level. As a result, we fail to reject the null hypothesis that the errors follow a normal distribution and conclude that the normality assumption is not violated. The Q-Q Plot for this model (Figure 1) also supports this conclusion, as most of the points line up along the qqline.

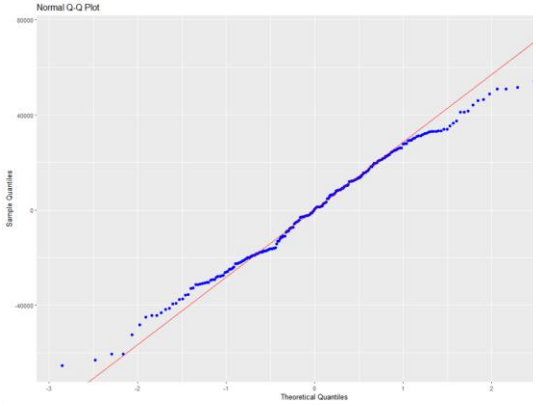

Figure 1

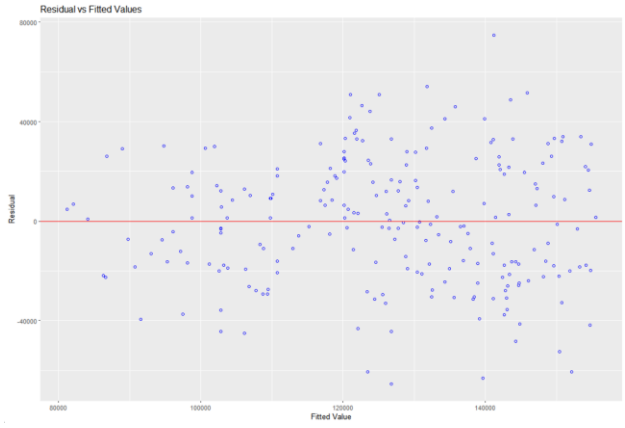

Figure 2

The constant variance assumption, however, is violated for this model. The p-value for the Breusch-Pagan hypothesis test to check this assumption (0.009946) is less than the α = 0.05 significance level, so we reject the null and conclude that the errors have a non-constant variance and that the constant variance assumption has been violated. This makes sense, since the spread of the residuals appears to increase as the fitted value increases, as evidenced by the Residuals vs. Fitted Values plot (Figure 2).

I then checked to see if a transformation of the predictors would result in the constant variance assumption being upheld by creating a BoxCox plot (Figure 3) to check to see if the BoxCox value of λ wasn't equal to 1. Since the 95% confidence interval does contain the value of $\lambda = 1$, combined with the fact that the normality assumption wasn't violated in the original OLS model, I concluded that a variable transformation wouldn't be appropriate for the given parameters.

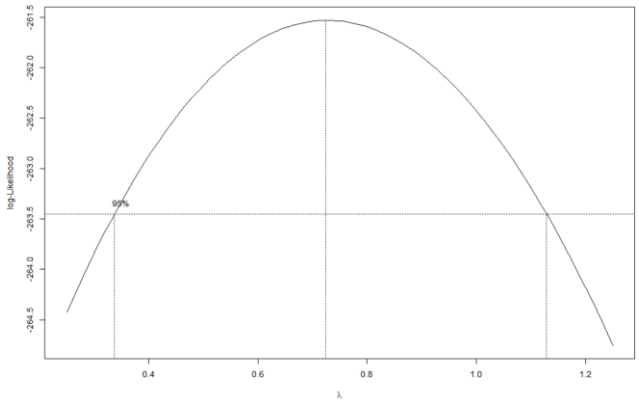

Figure 3

Lastly, I checked to see if there were any observations of high influence in the model – the observations were outliers and high leverage points – and I found 8 such points in the data set: observations number 1, 4, 22, 56, 67, 82, 165, 219, and 220. After removing the points from the data set, I refit the original OLS model and conducted the hypothesis tests for the model assumptions again. The p-values for both the Shapiro-Wilk test (0.0882) and the Breusch-Pagan test (0.06629) were both greater than the $\alpha = 0.05$ significance level, so reject the null hypotheses and conclude that the assumptions weren't violated. As seen in Figure 4, the spread of the residuals appears to remain constant even with an increase in the fitted values.

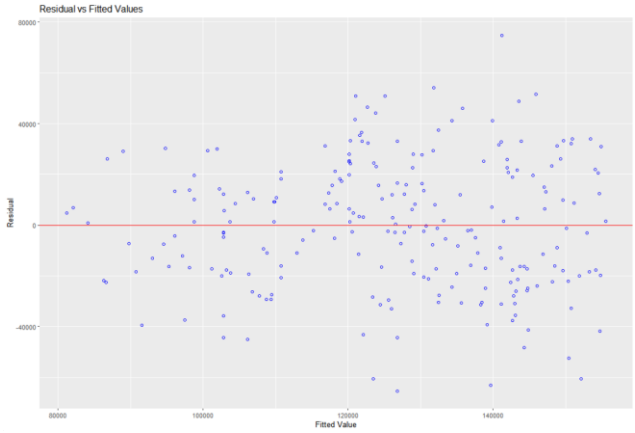

Figure 4

Question 2:

The fitted regression equation of my model is: $\widehat{SalePrice} = 27326.822 + 86.387(GrLivArea)$

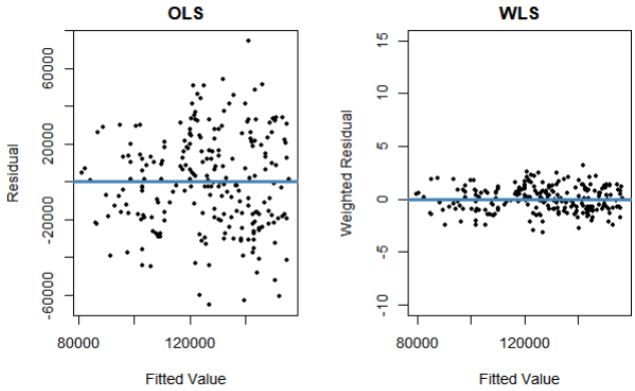

Figure 5

For this model, I chose weights corresponding to $w_i = \frac{1}{x_i}$, since in the original OLS model, the error terms increase as the fitted value increases (i.e., the errors are proportional to the predictor). As seen in Figure 5, the spread of the weighted residuals is a lot smaller compared to the spread of the residuals in the original OLS model.

I checked the model assumptions for this model to ensure its validity. The p-values for both the Shapiro-Wilk test (0.3963) and the Breusch-Pagan test (0.9996) were both greater than the α = 0.05 significance level, so reject the null hypotheses and conclude that the assumptions weren't violated. As seen in the Q-Q plot (Figure 6), most of the points lie along the qqline. In addition, the spread of the residuals is constant even as the fitted value increases (see fitted vs. weighted residuals plot in Figure 5).

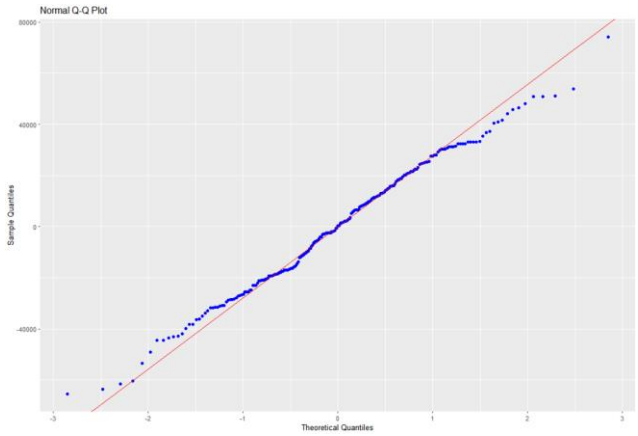

Figure 6

Question 3:

The model I would recommend to the Ames Reality Group would be the Weighted Least Squares model I developed in Question 2 for two reasons: 1) for the WLS model, the necessary adjustment to make to the observations in order for the model assumptions to not be violated is to reduce the weight of outliers, as opposed to removing them entirely for the OLS model in Question 1; and 2) the R^2 for the WLS model – at 0.3972 – is higher than the R^2 for the OLS model (0.3714), meaning that 39.72% of the variation in the sale price can be explained by the living area, as opposed to 37.14% for the OLS model.

To test the overall significance of regression, I performed an F-test. The p-value from the test is equal to $2.2 * 10^{-16}$, so we reject the null hypothesis that the slope is zero and conclude that there is a significant linear relationship between Living Area and Sale Price, as further reinforced by the scatterplot of the response vs. the predictor in Figure 7.

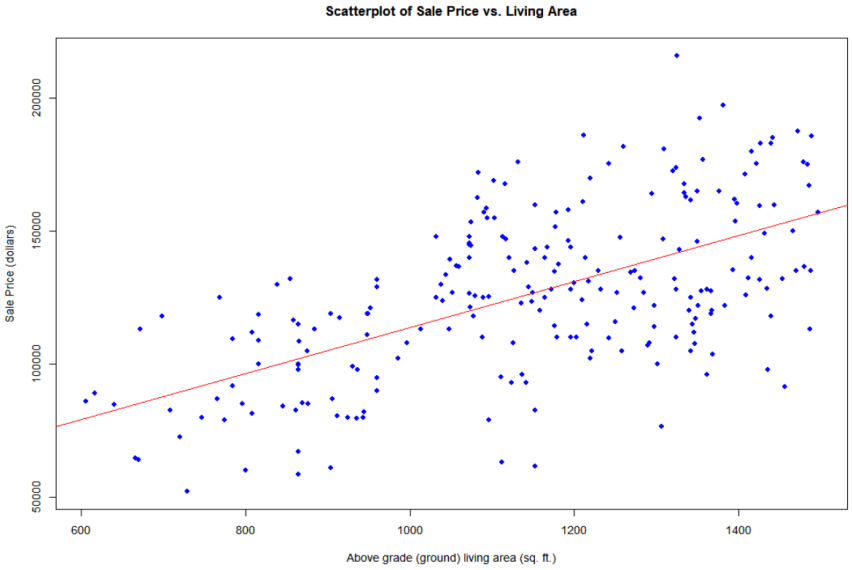

Figure 7

The final fitted regression (as stated earlier) is $\widehat{SalePrice} = 27326.822 + 86.387(GrLivArea)$. The slope of this equation is 86.387, meaning that for each additional square foot of living area, the mean sale price is estimated to increase by $86.39. To better estimate the true value of the slope, I calculated a 95% confidence interval for the slope, resulting in a lower bound of 72.47 and an upper bound of 100.31, meaning we are 95% confident that for an increase of one square foot of living area, the average increase in sale price is between $72.47 and $100.31.

Analysis 2:

Question 1:

Multicollinearity appears to be an issue in this model. As seen in the correlation plot (Figure 8), there is a very strong correlation between the number of cars in the garage and the garage area. In addition, there is also a large positive correlation between the square feet of the basement and the square feet of the first floor as well as a large negative correlation between the square feet of the first floor and the square feet of the second floor. I also checked the variance inflation factors (VIF's) of the predictors to see if they were all less than or equal to 5. I found that the GrLivArea, FirstFlrSf, and SecondFlrSf predictors all have VIF values greater than 5, suggesting multicollinearity exists.

Figure 8

In order to fix the multicollinearity issue, I first tried various combinations of collinear predictors, including $I(FirstFlrSf + SecondFlrSf)$, $I((FirstFlrSf + SecondFlrSf)/GrLivArea)$, and $I(GarageCars/GarageArea)$. However, none of the variable transformations wound up being significant, as all the p-values were greater than the α = 0.05 significance level.

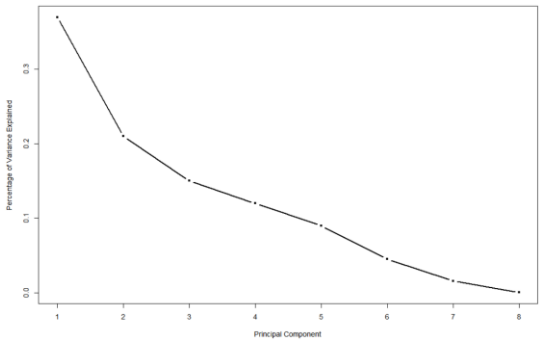

Figure 9

Since the variable transformations didn't result in a significant model, I wound up using PCR with 7 components to fix the multicollinearity issue. To determine the number of components, I used 10-fold cross validation. The results of the CV line up with the scree plot depicted in Figure 9, as there don't appear to be any elbows which result in the change in variance leveling off.

As such, my fitted regression equation for this model is:

$LOG(SalePrice_i) = 11.8141 - 0.125z_{i1} - 0.035z_{i2} + 0.026z_{i3} - 0.032z_{i4} - 0.024z_{i5} + 0.034z_{i6} + 0.038z_{i7}$, with an adjusted R^2 value of 0.6479, meaning that 64.79% of the variation in the LOG of SalePrice can be explained by the seven principal components.

Question 2:

To determine if the six given variables (TotalBsmtSf, GrLivArea, FirstFlrSf, GarageCars, Fireplaces, and ScreenPorch) are sufficient to model LOG(SalePrice), the appropriate statistical test to conduct is a nested model F-test, comparing the linear model with just the six predictors to a linear model with all of the predictors in the data set.

Hypotheses:

H_0: $\beta_7 = \beta_8 = \cdots = \beta_{26} = 0$

H_1: At least one of $\beta_7 \ldots \beta_{26} \neq 0$

P-Value:

p-value = $2.2 * 10^{-16}$

Statistical Decision:

Reject H_0 at the $\alpha = 0.05$ significance level

Conclusion in Context:

At least one of the other β coefficients corresponding to one of the other predictors not given is significant, so the six variables given are not sufficient to model LOG(SalePrice)

Question 3:

The MLR model for LOG(SalePrice) as a function of TotalBsmtSf, FirstFlrSf, GrLivArea, GarageCars, Fireplaces, and ScreenPorch has the following fitted regression equation:

$$\widehat{LOG(SalePrice)} = 10.879 + 0.000289(TotalBsmtSf) + 0.000354(GrLivArea) + 0.0000651(FirstFlrSf) + 0.128(GarageCars) + 0.0625(FirePlaces) + 0.0000724(ScreenPorch)$$

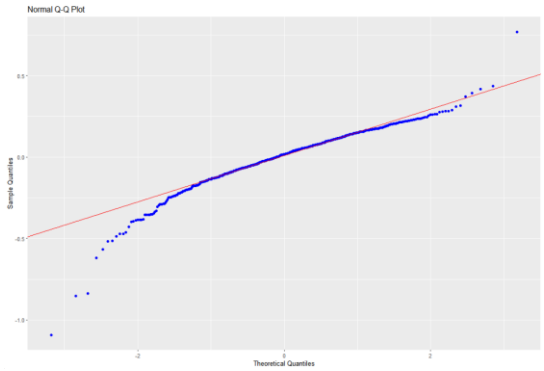

Figure 10

For this MLR model, I chose to estimate the coefficients using Huber's method since the error terms for the OLS and LAD regression models don't follow a normal distribution.

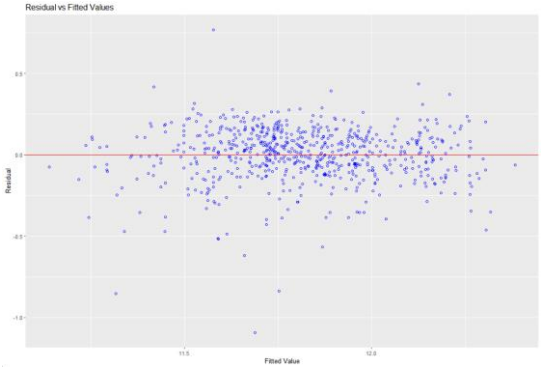

Figure 11

The error terms appear to have a constant variance, as seen in the fitted vs. residuals plot in Figure 11, which is supported by the p-value of Breusch-Pagan test (0.1914) being greater than the α = 0.05 significance level. The p-value for the LAD and Huber's models are the same.

As seen in Figure 10, most of the points on the Q-Q plot for the OLS model don't stay along the qqline, especially along the tails. The results of the Q-Q Plot line up with the results of the Shapiro-Wilk test for the OLS model, which has a p-value of $2.2*10^{-16}$, which is less than the α = 0.05 significance level, so we reject H_0 and conclude that the errors aren't normally distributed. The p-values for the Shapiro-Wilk test for all the other regression methods are also less than the α = 0.05 significance level.

```
Bootstrap bca confidence intervals

                Estimate        2.5 %       97.5 %
(Intercept) 1.087947e+01  1.080050e+01 1.092912e+01
TotalBsmtSf 2.893525e-04  2.472781e-04 3.377739e-04
GrLivArea   3.539829e-04  2.932622e-04 4.160527e-04
FirstFlrSf  6.511164e-05  3.648631e-06 1.285350e-04
GarageCars  1.281807e-01  1.099158e-01 1.465823e-01
Fireplaces  6.247433e-02  4.229214e-02 8.304625e-02
ScreenPorch 7.241030e-05 -1.471257e-04 2.876368e-04
```

Figure 12

Since the error terms aren't normally distributed, the confidence intervals for the slope parameters and the intercept are invalid, so we need a Bootstrap confidence interval (which will estimate the coefficients and create confidence intervals using the residuals of the OLS model) to check the significance (Figure 12).

Based on the results of the bootstrap, the TotalBsmtSf, GrLivArea, FirstFlrSf, GarageCars, and Fireplaces predictors are significant, since 0 is not within the lower and upper bounds of the

confidence intervals. For each slope parameter, for each one unit increase, the LOG of the sale price is expected to increase by the value of the parameter estimate.

Analysis 3:

Question 1:

The recommended model to deploy would be the lasso model, since among the 7 different models, the lasso model with coefficients corresponding to the value of λ has the lowest RMSE.

Model	RMSE
WLS (Analysis 1)	127263.4
Huber's Method (Analysis 2, Question 3)	0.1657047
Backward Selection – AIC	0.09845752
Best Subset Selection	0.1030226
PCR – 5 Principal Components	0.1157739
Ridge Regression	0.09790307
Lasso – Lambda Min	0.09711727

Table 1

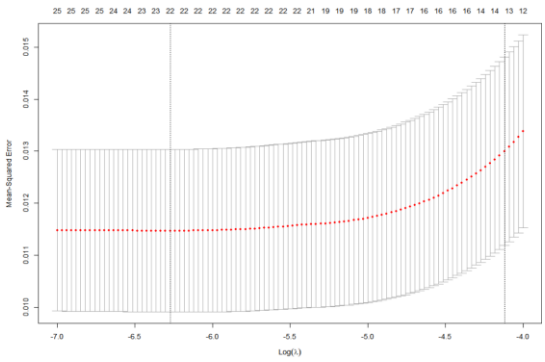

Figure 13

In order to find the λ values, I created a plot with many possible λ values along the natural-log scale and used 10-fold cross-validation to determine lambda.min and lambda.1se (Figure 13). I found lamba.min to be equal to 0.00172309 and lambda.1se to be equal to 0.0116256. In order to choose between the lambda values, I calculated the RMSE for both models on the testing data, and found that the RMSE for the model corresponding to lambda.min (0.097) was less than the RMSE for the model corresponding to lambda.1se (0.1043899).

	s1
(Intercept)	3.078840e+00
LotArea	7.845706e-06
OverallQual	7.510178e-02
OverallCond	4.863434e-02
YearBuilt	3.152281e-03
YearRemodAdd	6.027442e-04
BsmtFinSf1	7.863889e-05
BsmtFinSf2	3.542410e-05
TotalBsmtSf	7.656185e-05
FirstFlrSf	8.703603e-05
SecondFlrSf	.
GrLivArea	2.715194e-04
BsmtFullBath	1.731848e-02
BsmtHalfBath	1.661040e-02
FullBath	.
HalfBath	.
BedroomAbvGr	-1.531815e-02
KitchenAbvGr	-5.324017e-02
TotRmsAbvGrd	1.521604e-02
Fireplaces	2.929021e-02
GarageCars	1.000321e-02
GarageArea	1.474117e-04
WoodDeckSf	3.824206e-05
OpenPorchSf	3.382050e-05
EnclosedPorch	-2.478143e-05
ThreeSsnPorch	.
ScreenPorch	2.047545e-04

Figure 14

As seen in Figure 14, the lasso zeroed-out the SecondFlrSf, FullBath, HalfBath, and ThreeSsnPorch predictors.

Question 2:

Fitted Regression Equation:

$$LOG(\widehat{SalePrice}) = 2.901 + 0.08987(OverallQual) + 0.05641(OverallCond) + 0.003767(YearBuilt) + 0.0001154 * I(BsmtFin1 + BsmtFin2) + 0.0001593(FirstFlrSf) + 0.0003426(GrLivArea) + 0.0000102 * I(LotArea + GarageArea)$$

RMSE = 0.1059366

New Predictors Created:

1) I(BsmtFinSf1 + BsmtFinSf2) → The sum of finished square feet from both Type 1 and Type 2 basements
2) I(LotArea + GarageArea) → The sum of the house's parking areas (the garage and the outside parking lot/driveway)

Process:

Among the seven models in Question 1, I noticed that the model for best subset selection – while not having the highest RMSE – contained the least amount of predictors (8). Using the "which" matrix, the R code computed the RSS values needed to find the lowest AIC and BIC, and highest R^2 values among various subsets of predictors, all three landing on a subset of 8 predictors, which I then further aggregated into the required 7 predictors.