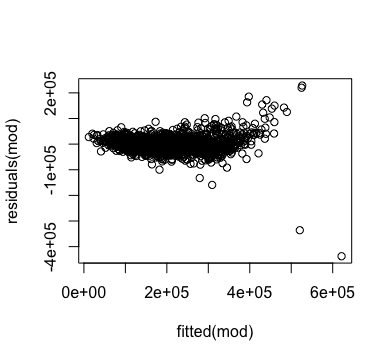
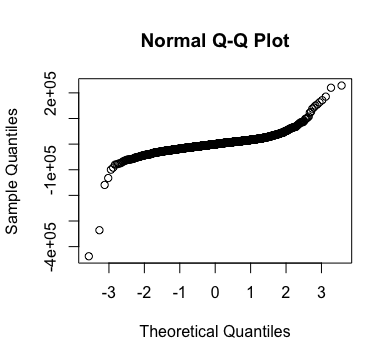
**Question 1: Exploratory Data Analysis**

(a) Begin by doing some exploratory analysis. Use the code below to plot examine the relationships between the explanatory variables and *SalePrice*.

* The relationships between the following explanatory variables and *SalePrice* are linear.
  + Overall.Qual, Year.Built, Bsmt.Qual, BsmtFin.SF, Total.Bsmt.SF, Heating.QC, X1st.Flr.SF, Gr.Liv.Area, Kitchen.Qual, TotRms.AbvGrd, Fireplaces, Firplace.Qu, Garage.Finish, Garage.Cars, Garage.Area, Bath
* Other variables do not exhibit a strong linear relationship with *SalePrice.*
* The residual plot below shows that constant variance is not upheld.



* While not certain given the information provided, it is probably reasonable to assume that the data points are not independent of each other. Housing prices in the same area are highly affected by each other.



* The Normal Q-Q plot illustrates a non-linear curve; there is deviation from normality.

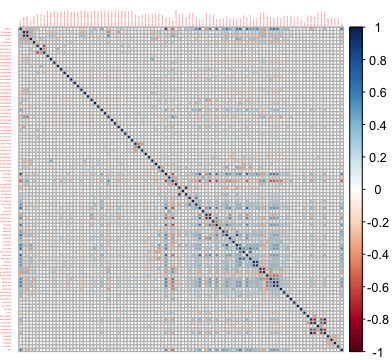
(b)

(i) Based on this plot, what observations can you make about the relationship the predictors and *SalePrice*? How many of predictors correlate strongly with *SalePrice*?

- Most variables do not correlate strongly with *SalePrice*.

- While the color is difficult to tell, about 5~8 predictors correlate strongly with *SalePrice*.

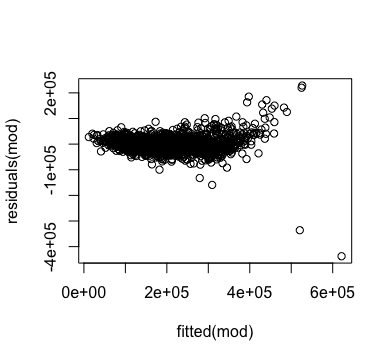
(ii) How many of the Is there any evidence of multicollinearity? If so, why?



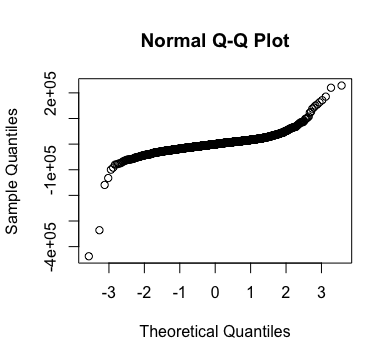
* While it is difficult to tell the exact number of relationships suggesting multicollinearity, there appears to be about 30~40 such relationships going by the number of “dark” spots in one side(triangle) of the plot.
* It makes sense since a lot of the variables are correlated to each other; for instance, the overall quality is probably some factor of the year built, basement quality, kitchen quality, etc.

**Question 2: Fitting the Linear Regression Model.**

(a) Use the *lm()* function in R to fit the full model. Plot the fitted values against the residuals. What observations can you make from this plot in terms of the model assumptions (linear, constant variance, etc.)? How do these observations compare to those you made for the previous question?



* Residuals appear to be centered around 0, so linearity holds.
* Variance appears to increase as *SalePrice* increases.
* Independence cannot be determined with the above.

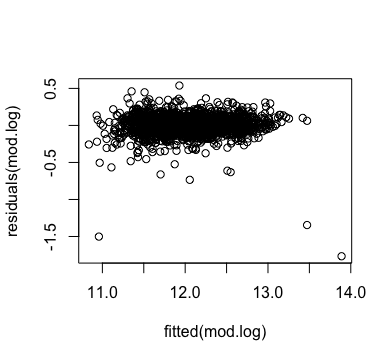


* Based on the Normal Q-Q Plot, the residuals appear to have a departure from normality.

(b) From the plot in part (a), it appears that there is a problem with the constant variance assumption. Perform a Box-Cox transformation of the response variable *SalePrice* using the *boxcox()* function in R. What is the best choice for the transformation parameter λ?

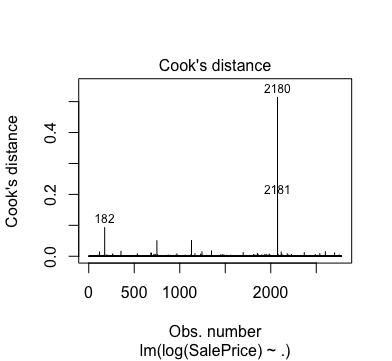
- The value that maximizes the log-likelihood is 0.14.

(c) Fit the model again using a log-transformation of the response variable *SalePrice* and plot the fitted values against the residuals? Are there any outliers?



- There appear to be 3 outliers whose residual values are close to -1.5.

(d) Compute the cook’s distance for each of the observations? Using the rule of thumb that observations should be excluded if their cook’s distance is > 1, identify any points which should be excluded. If you exclude any points, refit the model one final time.



* We can identify three points(182, 2180, 2181) that have Cook’s distances that exceed 1.
* Based on the rule of thumb, these points should be excluded.
* The refined model without outliers returns a summary as follows

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.028e+01 6.811e-02 150.866 < 2e-16 \*\*\*

MS.ZoningRL -2.238e-03 2.216e-02 -0.101 0.919570

MS.ZoningRM -4.575e-02 2.489e-02 -1.838 0.066146 .

Lot.Area 3.587e-06 6.772e-07 5.296 1.28e-07 \*\*\*

Lot.ShapeReg 6.573e-03 5.359e-03 1.227 0.220050

Land.ContourHLS 8.208e-02 1.734e-02 4.734 2.32e-06 \*\*\*

Land.ContourLow 4.888e-02 2.249e-02 2.174 0.029805 \*

Land.ContourLvl 4.344e-02 1.248e-02 3.482 0.000506 \*\*\*

NeighborhoodBlueste 7.423e-03 4.737e-02 0.157 0.875483

NeighborhoodBrDale -3.605e-02 3.836e-02 -0.940 0.347504

NeighborhoodBrkSide -1.142e-02 3.256e-02 -0.351 0.725871

NeighborhoodClearCr 3.317e-02 3.342e-02 0.992 0.321060

NeighborhoodCollgCr -1.879e-02 2.639e-02 -0.712 0.476382

NeighborhoodCrawfor 7.961e-02 2.995e-02 2.658 0.007898 \*\*

NeighborhoodEdwards -8.459e-02 2.834e-02 -2.985 0.002865 \*\*

NeighborhoodGilbert -4.401e-02 2.757e-02 -1.596 0.110542

NeighborhoodGreens 1.122e-01 4.992e-02 2.248 0.024683 \*

NeighborhoodGrnHill 5.841e-01 8.710e-02 6.706 2.44e-11 \*\*\*

NeighborhoodIDOTRR -5.049e-02 3.592e-02 -1.406 0.159987

NeighborhoodMeadowV -1.296e-01 3.600e-02 -3.600 0.000323 \*\*\*

NeighborhoodMitchel -3.566e-02 2.856e-02 -1.248 0.211960

NeighborhoodNAmes -4.204e-02 2.770e-02 -1.517 0.129306

NeighborhoodNoRidge 5.385e-02 3.017e-02 1.785 0.074376 .

NeighborhoodNPkVill -2.895e-02 3.618e-02 -0.800 0.423574

NeighborhoodNridgHt 9.698e-02 2.654e-02 3.654 0.000264 \*\*\*

NeighborhoodNWAmes -6.089e-02 2.833e-02 -2.149 0.031698 \*

NeighborhoodOldTown -6.160e-02 3.293e-02 -1.871 0.061501 .

NeighborhoodSawyer -3.382e-02 2.879e-02 -1.175 0.240131

NeighborhoodSawyerW -4.385e-02 2.761e-02 -1.588 0.112379

NeighborhoodSomerst 4.914e-02 3.140e-02 1.565 0.117724

NeighborhoodStoneBr 1.048e-01 2.975e-02 3.523 0.000433 \*\*\*

NeighborhoodSWISU -7.012e-02 3.447e-02 -2.034 0.042014 \*

NeighborhoodTimber 5.676e-03 2.910e-02 0.195 0.845376

NeighborhoodVeenker -3.204e-02 3.663e-02 -0.875 0.381824

Condition.1Norm 4.636e-02 8.538e-03 5.429 6.16e-08 \*\*\*

Condition.1Rail 2.265e-02 1.612e-02 1.404 0.160286

Condition.1PosN 1.093e-02 1.714e-02 0.638 0.523685

Bldg.Type2fmCon -1.662e-02 1.742e-02 -0.954 0.340161

Bldg.TypeDuplex -5.398e-02 1.493e-02 -3.615 0.000305 \*\*\*

Bldg.TypeTwnhs -1.281e-01 1.826e-02 -7.017 2.87e-12 \*\*\*

Bldg.TypeTwnhsE -6.472e-02 1.246e-02 -5.196 2.18e-07 \*\*\*

House.Style1Story 7.996e-03 1.140e-02 0.701 0.483282

House.StyleSplit 2.483e-02 1.313e-02 1.891 0.058779 .

Overall.Qual 5.814e-02 3.301e-03 17.614 < 2e-16 \*\*\*

Overall.Cond 4.733e-02 2.754e-03 17.183 < 2e-16 \*\*\*

Year.Built -1.584e-03 2.287e-04 -6.925 5.42e-12 \*\*\*

Year.Remod.AddTRUE 1.299e-02 5.572e-03 2.331 0.019829 \*

Roof.StyleGarble 5.008e-03 2.278e-02 0.220 0.826015

Roof.StyleGarbles 4.484e-02 3.575e-02 1.254 0.209875

Roof.StyleHip 6.709e-03 2.328e-02 0.288 0.773223

Mas.Vnr.TypeBrick -1.773e-03 7.437e-03 -0.238 0.811586

Mas.Vnr.TypeStone 9.188e-03 1.042e-02 0.882 0.377951

Mas.Vnr.Area -6.973e-06 2.016e-05 -0.346 0.729461

Exter.Qual 7.380e-03 7.212e-03 1.023 0.306268

Exter.Cond 4.421e-03 6.786e-03 0.652 0.514773

FoundationCBlock 2.080e-02 1.023e-02 2.034 0.042059 \*

FoundationPConc 3.604e-02 1.132e-02 3.184 0.001471 \*\*

FoundationOther 5.313e-02 2.183e-02 2.434 0.015006 \*

Bsmt.Qual 2.538e-02 4.973e-03 5.103 3.57e-07 \*\*\*

Bsmt.Cond -1.146e-02 5.326e-03 -2.152 0.031511 \*

BsmtFin.SF 5.983e-05 7.251e-06 8.252 2.41e-16 \*\*\*

Total.Bsmt.SF 5.162e-05 1.267e-05 4.075 4.73e-05 \*\*\*

Heating.QC 1.517e-02 3.149e-03 4.816 1.55e-06 \*\*\*

Central.AirY 3.092e-02 1.149e-02 2.691 0.007175 \*\*

X1st.Flr.SF 1.924e-05 1.854e-05 1.038 0.299583

Gr.Liv.Area 2.118e-04 1.595e-05 13.281 < 2e-16 \*\*\*

Bedroom.AbvGr 5.413e-03 4.374e-03 1.238 0.215949

Kitchen.Qual 2.632e-02 5.386e-03 4.886 1.09e-06 \*\*\*

TotRms.AbvGrd -5.491e-04 3.035e-03 -0.181 0.856430

Fireplaces 1.709e-02 7.413e-03 2.305 0.021245 \*

Fireplace.Qu 6.054e-03 2.702e-03 2.241 0.025121 \*

Garage.TypeAttchd 2.547e-02 9.313e-03 2.735 0.006277 \*\*

Garage.TypeDetchd 1.561e-02 1.112e-02 1.403 0.160611

Garage.TypeNo 1.272e-02 1.779e-02 0.715 0.474453

Garage.Finish 6.506e-03 4.026e-03 1.616 0.106234

Garage.Cars 3.074e-02 7.736e-03 3.974 7.26e-05 \*\*\*

Garage.Area 4.485e-05 2.605e-05 1.722 0.085186 .

Paved.Drive 2.312e-02 5.474e-03 4.223 2.49e-05 \*\*\*

Wood.Deck.SF 4.731e-05 1.997e-05 2.369 0.017914 \*

Open.Porch.SF 1.345e-05 3.790e-05 0.355 0.722777

Enclosed.Porch 1.678e-04 3.879e-05 4.326 1.58e-05 \*\*\*

X3Ssn.Porch 1.027e-04 8.910e-05 1.152 0.249405

Screen.Porch 2.587e-04 4.199e-05 6.161 8.32e-10 \*\*\*

Pool.Area 6.319e-05 6.578e-05 0.961 0.336808

Fence -3.886e-04 1.989e-03 -0.195 0.845124

Sale.TypeOther 9.114e-03 2.050e-02 0.445 0.656686

Sale.TypeNEW 1.124e-01 5.573e-02 2.017 0.043843 \*

Sale.TypeWD 8.408e-03 1.440e-02 0.584 0.559330

Sale.ConditionOther 9.457e-02 1.694e-02 5.584 2.59e-08 \*\*\*

Sale.ConditionNormal 1.025e-01 1.010e-02 10.155 < 2e-16 \*\*\*

Sale.ConditionPartial 2.654e-02 5.395e-02 0.492 0.622728

Yr.Sold2007 -7.606e-03 6.724e-03 -1.131 0.258126

Yr.Sold2008 1.576e-03 6.944e-03 0.227 0.820498

Yr.Sold2009 -1.383e-02 6.961e-03 -1.987 0.047000 \*

Yr.Sold2010 -2.542e-03 8.244e-03 -0.308 0.757865

Bath 3.286e-02 4.845e-03 6.782 1.45e-11 \*\*\*

---

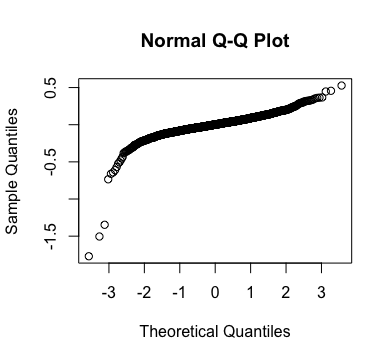
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1158 on 2679 degrees of freedom

Multiple R-squared: 0.9174, Adjusted R-squared: 0.9145

F-statistic: 313.1 on 95 and 2679 DF, p-value: < 2.2e-16

(e) Finally, produce a quantile-quantile plot to check the normality of the residuals and comment on the results. What are the implications of these results for the estimation of and inference on the regression coefficients?



* The Normal Q-Q Plot of the model residuals still shows deviation from normality.
* Therefore, estimation and inference based on regression coefficients of this model may be misleading.

**Question 3: Variable Selection via Stepwise Regression.**

(a) Use the summary function *summary()* to look at the final model in the previous question. Notice that many of the coefficients for the explanatory variables are not statistically significant, an indication that model selection could be useful in simplifying the model. Use the *drop1()* function in R to see which variable can be dropped to most improve the AIC value of the model. Which variable can be removed to produce the biggest drop in AIC?

Df Sum of Sq RSS AIC

<none> 35.916 -11872

MS.Zoning 2 0.1628 36.078 -11863

Lot.Area 1 0.3756 36.291 -11845

Lot.Shape 1 0.0196 35.935 -11872

Land.Contour 3 0.3038 36.219 -11854

Neighborhood 26 4.4297 40.345 -11601

Condition.1 3 0.4649 36.381 -11842

Bldg.Type 4 0.9108 36.826 -11810

House.Style 2 0.0560 35.972 -11871

Overall.Qual 1 4.1704 40.086 -11569

Overall.Cond 1 3.9554 39.871 -11584

Year.Built 1 0.6491 36.565 -11824

Year.Remod.Add 1 0.0719 35.988 -11868

Roof.Style 3 0.0295 35.945 -11875

Mas.Vnr.Type 2 0.0177 35.933 -11874

Mas.Vnr.Area 1 0.0016 35.917 -11873

Exter.Qual 1 0.0139 35.930 -11872

Exter.Cond 1 0.0060 35.922 -11873

Foundation 3 0.1741 36.090 -11864

Bsmt.Qual 1 0.3526 36.268 -11846

Bsmt.Cond 1 0.0611 35.977 -11869

BsmtFin.SF 1 0.9113 36.827 -11804

Total.Bsmt.SF 1 0.2216 36.137 -11856

Heating.QC 1 0.3108 36.226 -11850

Central.Air 1 0.0966 36.012 -11866

X1st.Flr.SF 1 0.0147 35.930 -11872

Gr.Liv.Area 1 2.3633 38.279 -11697

Bedroom.AbvGr 1 0.0209 35.937 -11872

Kitchen.Qual 1 0.3207 36.236 -11849

TotRms.AbvGrd 1 0.0005 35.916 -11874

Fireplaces 1 0.0720 35.988 -11868

Fireplace.Qu 1 0.0669 35.983 -11868

Garage.Type 3 0.1145 36.030 -11869

Garage.Finish 1 0.0341 35.950 -11871

Garage.Cars 1 0.2087 36.124 -11858

Garage.Area 1 0.0399 35.956 -11870

Paved.Drive 1 0.2357 36.151 -11855

Wood.Deck.SF 1 0.0762 35.992 -11868

Open.Porch.SF 1 0.0019 35.918 -11873

Enclosed.Porch 1 0.2535 36.169 -11854

X3Ssn.Porch 1 0.0178 35.933 -11872

Screen.Porch 1 0.4983 36.414 -11835

Pool.Area 1 0.0122 35.928 -11873

Fence 1 0.0004 35.916 -11874

Sale.Type 3 0.0534 35.969 -11873

Sale.Condition 3 1.4110 37.327 -11771

Yr.Sold 4 0.0896 36.005 -11873

Bath 1 0.6187 36.534 -11826

* Dropping Roof.Style produces the biggest drop in AIC.

(b) Use the *stepAIC()* function in R to perform stepwise variable selection with AIC as the criterion to select the best model. How many variables are included in the final model? How much does AIC improve?

* Starting with the full model and using “both” as the direction
* 32 variables were selected (73 coefficients including intercept, dummy variables from categorical variables)
* AIC decreased by 23 from -11872 to -11895.

**Question 4: Variable Selection via LASSO and Elastic Net.**

(a) While stepwise regression performs variable selection and estimation iteratively, LASSO is a method which performs variable selection and estimation simultaneously. Use the *glmnet()* function in R with the full model to estimate the model using LASSO, choosing the best value for λ using cross validation. What is the best value for λ? How many of the 96 model coefficients are nonzero in the fit?

* The best value for lambda is 0.001352236.
* 75 out of 96 model coefficients are nonzero.

(b) Elastic net combines the model selection benefits of LASSO (l1 penalty) with the regularization effects of ridge regression (l2 penalty). Use the *glmnet()* function in R with the full model to estimate the model using the elastic net, choosing the best value for λ using cross validation. What is the best value for λ? How many of the 96 model coefficients are nonzero in the fit? How does the number of nonzero parameter estimate compare to the LASSO fit and how do you explain the difference?

* Best value for lambda is 0.002045834.
* 78 of 96 model coefficients are nonzero.
* The number of nonzero coefficients is larger for the Elastic Net fit vs. the LASSO fit.
* This is because LASSO regression forces some coefficients to zero, while Ridge regression does not. Thereby, because Elastic Net is a combination of LASSO and Ridge regression, it will have more nonzero coefficients than LASSO.

**Question 5: Variable Selection via Group LASSO.**

(a) Use the *gglasso()* function in R with the full model to estimate the model using Group LASSO, choosing the best value for λ using cross validation. What is the best value for λ? How many of the 96 model coefficients are nonzero in the fit?

* The best value for lambda is 0.6508434
* 12 of 96 model coefficients are nonzero in the fit.