Case Study, Linear Regression

Zizhen Song, Beiming Liu, Jason Liu, Christopher Csiszar

Part 1

Task 1

The data ("housing.txt") as provided has many variables which needed cleaning or modifying, even before proceeding to any analysis work. All factor variables ("factor" here meaning categorical and descriptive variables) which contained missing entries were re-formatted to include the null values as a factor level in itself. We then looked at missing values among the continuous variables, and decided on appropriate remedies pertaining to each variable's logical intent:

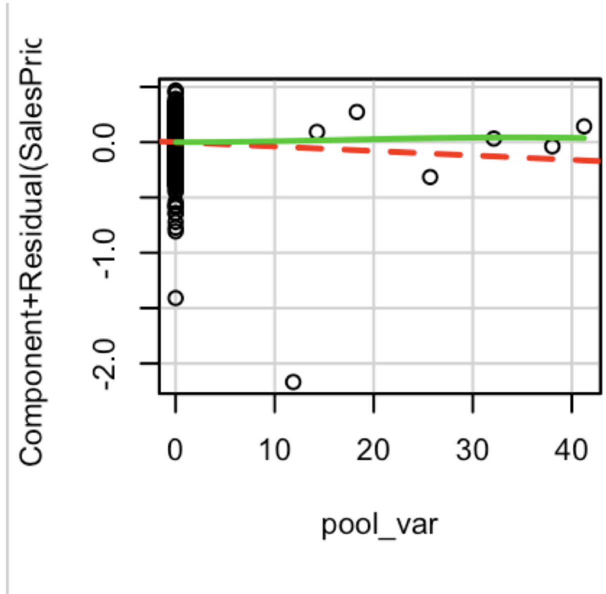
- LotFrontage: missing values were replaced by the mean LotFrontage value
- MasVnrArea: missing values were replaced by 0
- GarageYrBlt: missing values were replaced by the mean GarageYrBlt value

In general, means were chosen twice above due to their conformity when scaling the data in the next step - this was decided after attempting to replace each by 0, min, and max of the variable. We then scaled the entire dataset to allow for any model's coefficients to be compared more easily.

Our initial approach to determine what the most influential variables were when pricing a house in Ames, Iowa was to use a stepwise regression model to compare variables to each other. We ran into a significant problem with this approach however, since all angles of approach led to our data matrix to be singular and thus non-invertible. We eventually abandoned this approach in favor of the LASSO.

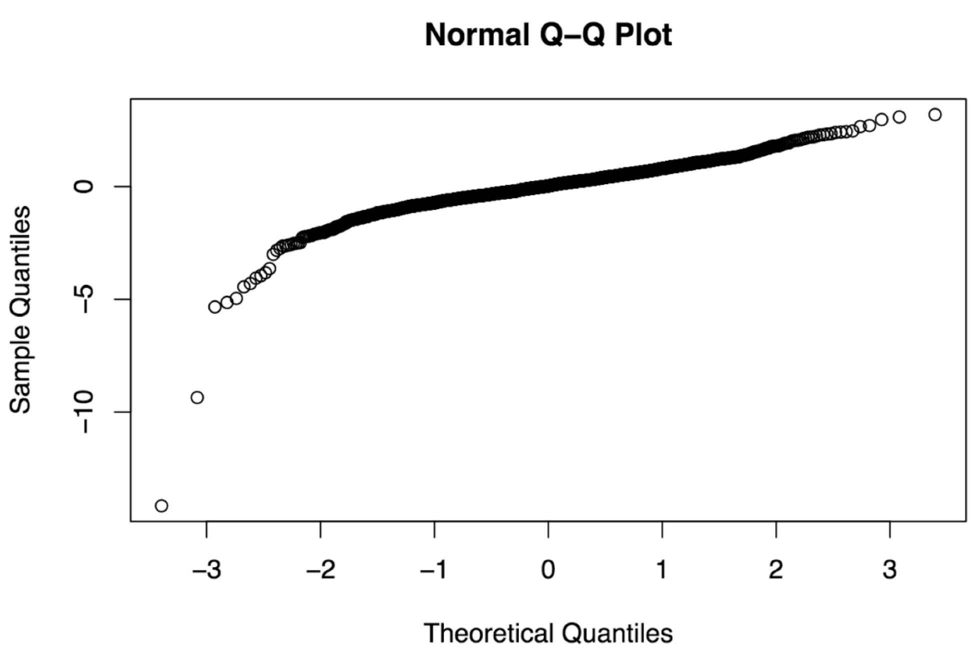
We run LASSO on the data to perform variable selection, and then transformed our response variable (SalesPrice) by taking its log - experience and data observation having told us at this point that might result in a more normal residual distribution. We fit only the selected variables and the response with OLS, and run a Cook's Distance test to determine if any particularly influential or outlier points in the data might be negatively affecting our fit.

We found that of the six influential and outlier data points, five were observations where pools were present as a house factor. Given that only seven observations of pools are recorded in the housing data in total, we suspected that transforming the two pool variables (PoolArea, PoolQC) into a combined single variable (pool_var, which multiplied the pool's surface area by an integer representation of its Quality Grade) would better describe a pool's relation to final house price. While the hunch proved correct and the better metric presented fewer outliers in the data, refitting on OLS showed that the combined pool variable was no longer significant to any reasonable degree. This, combined with insight from the graph below, led us to manually excluded all pool variables from the data at this point.

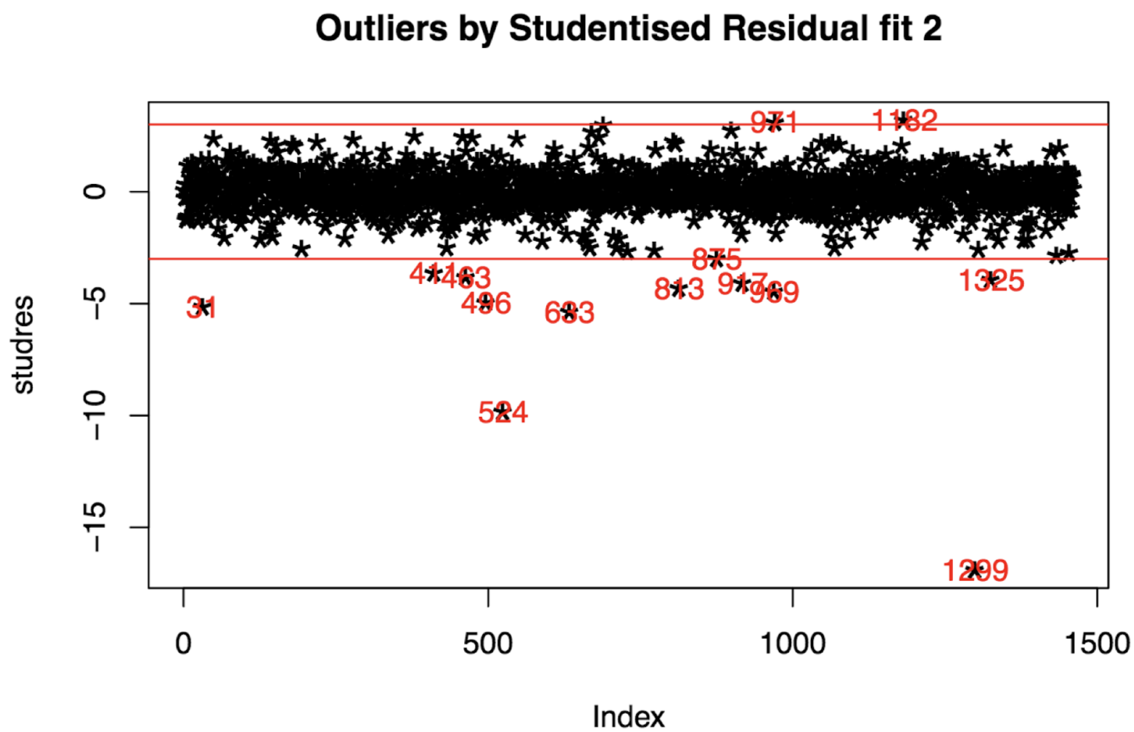


To further refine our OLS model, we exclude variables that are not significant to a 5% level, since LASSO does not consider significance of coefficients when dropping variables. To do so confidently, we quickly run an ANOVA F-test to determine whether dropping these variables is appropriate: our ANOVA F-test value was much greater than 0.05, thus we could not reject the possibility that the variables we wished to eliminate were not affecting price at all, meaning it was OK to drop them from the model henceforth.

Running a QQ plot of standardized residuals shows that their normality assumption does not hold even after taking the log of sales price; in fact a KS test at this point in the analysis shows "D = 0.1017, p-value = 9.319e-06".

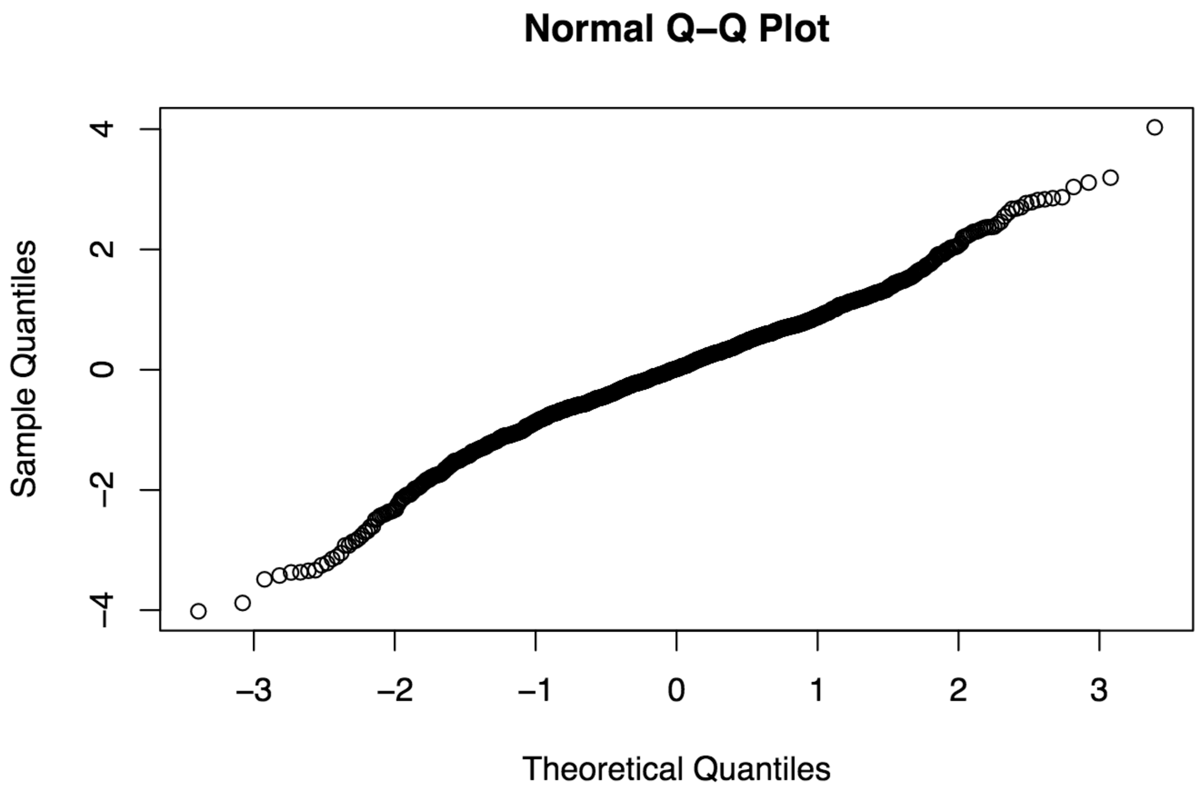


We resorted to deleting outliers from the data using a studentized residual test and the rule of thumb abs(Z) < 3.

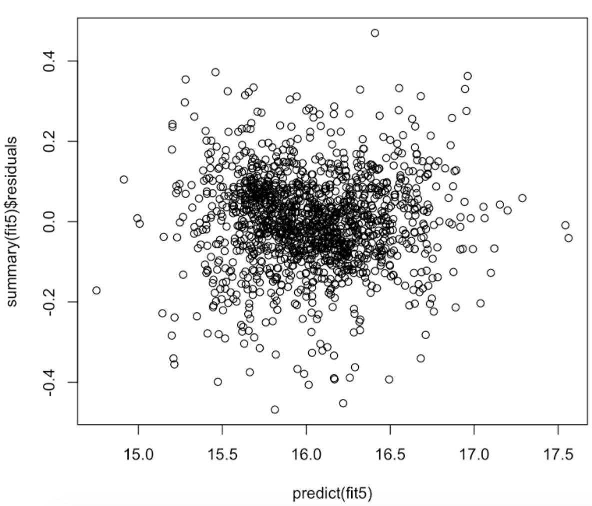


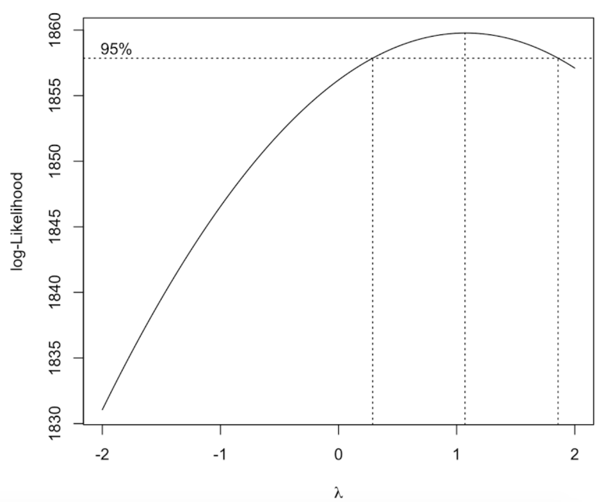
At this point, we maintained all outlier deletions and variable deletions through LASSO and significance / ANOVA, but reverted to unscaled X data to fit our final linear model through OLS. We reverted to unscaled X data because we wanted our final model to be more immediately interpretable, as explanation is the focus of this part of our report. We did however also proceed with another OLS model with scaled data, to observe the differences and effect of parameters out of interest.

Our efforts at variable and data removal were not in vain, as the QQ plot of residuals of our final model looks much more normally distributed.



We also checked whether variance is constant by the plot of residual against predicted value, no obvious pattern can be found. In addition, the boxcox plot shows that lambda = 1 is in 95% confidence interval, which means that we don't need to do other transformations, proved that our assumption to use log(SalePrice) is right. (Since boxcox only works with positive values for the response variable Y, we try to predict a shifted version Y+mu, where mu is greater than the minimum value of Y, in this case we chose mu = 4)





Furthermore, we checked the variance inflation factor (VIF) score of our fitted model, and none of them is greater than 6, which shows that we don’t need to worry about multicollinearity.

```
     LotArea  OverallQual  OverallCond    YearBuilt YearRemodAdd  TotalBsmtSF    GrLivArea
    1.164303     2.946651     1.470653     3.042520     2.060725     1.688632     5.247875
BsmtFullBath     FullBath KitchenAbvGr TotRmsAbvGrd   Fireplaces   GarageCars   WoodDeckSF
    1.219478     2.392535     1.315985     3.639222     1.460649     1.880440     1.175293
 ScreenPorch
    1.073110
```

Last but not Least, we checked our the R-squared p-value of F-statistic, and MSE of our final model, where we have R-squared = 0.9102, p-value of F-statistic < 2.2e-16 and MSE = 0.01356364. Now we can say that our model is very solid.

We observe the following most important factors of house prices in Ames, Iowa, according to our model in decreasing order of importance:

- OverallQual: Rates the overall material and finish of the house
- GarageCars: Size of garage in car capacity
- BsmtFullBath: Basement full bathrooms
- OverallCond: Rates the overall condition of the house

Thinking about the above variables at arm's length for a moment, a sound logical argument can be made for each one directly relating to the sales price of a house. We thus believe these results to be credible.

Task 2

Based on the data we have of Morty home, our prediction for Morty's house's final sales price is $153,197.10 based on the model we developed above. An upper bound of $192,939.50 is returned by the ceiling of a 95% confidence interval around our prediction, which would be our best guess as to an absolute max price reasonably possible.

Upon determining what Morty can do to increase the value of his house, we consulted the following list of relevant properties he can try working on (derived from our model). We sorted the coefficients of our scaled model to compare the effect of each parameter on

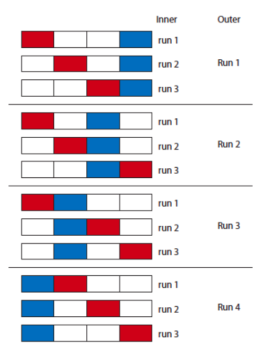
```
  (Intercept)   OverallQual    GarageCars  BsmtFullBath   OverallCond
 2.857096e+00  7.049313e-02  6.499393e-02  5.530747e-02  4.964991e-02
   Fireplaces  TotRmsAbvGrd     YearBuilt      FullBath  YearRemodAdd
 3.888633e-02  5.002907e-03  3.038200e-03  2.110147e-03  8.954240e-04
    GrLivArea   ScreenPorch   TotalBsmtSF    WoodDeckSF       LotArea
 2.637803e-04  2.272170e-04  1.592129e-04  6.126336e-05  2.637637e-06
 KitchenAbvGr
-9.793637e-02
```

Morty already has a full bathroom in his basement, so immediate jobs he can perform (not relating to an update of the entire house, such as upgrading all materials) is to increase the size of his garage to accommodate one more car, build a fireplace, and bring all his bathrooms to above grade (or more so).

Part2

We have gone through four Regression models during the lecture. OLS is unbiased estimator which is perfect for explanatory model. Ridge and lasso are biased estimator but with lower variance of the model, hence better predictive power. Ridge is especially good for eliminate collinearity while keep the whole model, lasso should be used in sparse beta vector where only few predictors is useful. And elastic net regression is in between of lasso and ridge.

For lasso and ridge, fix the hold out set, choose100 lambda value, run 3-fold cross validation on them and get 100 average MSPEs, picking the lambda with minimum (average MSPE), and this is run 1. Repeat the above for four times, each time with different hold out set. Eventually we have four lambdas, use hold out set as testing and rest as training, pick the ultimate lambda with lowest MSPE, construct new model with it.



Blue: hold out set; Red Testing set; White: Training set
While for elastic net regression, what we did is exactly the same as lasso and ridge except there are another for loop to tuning the alpha parameter.

After the calculation, we can see that Elastic net is with the minimum MSPE. The result is not that surprise as we can see from below that beta vector is quite sparse, meaning only few portions of the predictor get left in the model, that makes it suitable to use lasso or elastic net. Another fun fact here is the MSPE for OLS estimator is very high compare to the other threes, that can be explained by whole model's overfitting problem make the variance too large, although have small bias and high r^2, the explanatory power is weak.

```
MSPE
     Ridge      Lasso Elastic.Net       Ols
0.09541001 0.08613225  0.08435617 0.5731535
```

(all numeric scaled)

```
LMABDA.ALPHA
    Ridge      Lasso Elastic.Net.lambda Elastic.Net.alpha
0.8463827 0.01823476         0.08307418               0.2
```

Tuning parameters

```
MSPE
     Ridge      Lasso Elastic.Net        Ols
602143218 543590232   532381189 3617235771
```

(x without and scaling)

```
LMABDA.ALPHA
   Ridge    Lasso Elastic.Net.lambda Elastic.Net.alpha
67238.76 1448.615           6599.621               0.2
```

Tuning parameters

As far as coefficient comparison:

```
> Coefficients
                       Ridge        Lasso Elastic.Net
(Intercept)      -0.111362998 -0.064704670 -0.08507836
Id               -0.005258628  0.000000000  0.00000000
MSSubClass       -0.025661679 -0.076660937 -0.05942992
MSZoningFV        0.036545520  0.000000000  0.00000000
MSZoningRH        0.002245711  0.000000000  0.00000000
MSZoningRL        0.039219645  0.000000000  0.00000000
MSZoningRM       -0.042304479 -0.043289372 -0.04813570
LotFrontage       0.015102853  0.000000000  0.00000000
LotArea           0.033286755  0.032473939  0.03333478
StreetPave        0.192919786  0.000000000  0.04617831
AlleyPave         0.004539867  0.000000000  0.00000000
AlleyNA           0.015246402  0.000000000  0.00000000
LotShapeIR2       0.088422807  0.035779432  0.04542198
LotShapeIR3      -0.199688041 -0.115935280 -0.13571956
LotShapeReg      -0.021616241  0.000000000  0.00000000
LandContourHLS    0.085989121  0.002181435  0.02126830
LandContourLow   -0.004253214  0.000000000  0.00000000
LandContourLvl    0.027054210  0.000000000  0.00000000
UtilitiesNoSeWa  -0.335502738  0.000000000  0.00000000
LotConfigCulDSac  0.089026210  0.095524646  0.09855649
```

We can see that Ridge kept the whole model while Elastic net only kept a portion of them, and lasso's predictors is just a subset of elastic net.

Taking a look at the predictors that still exist in elastic net model.
We can see that some interesting fact in here that neighborhood is a very strong predictor, almost every condition is associate with the price change. Which make sense, Sunnyvale and Daly city have totally different house price for the exact same apartment, that is why real estate is always about "location, location, location".

```
           LotArea          StreetPave         LotShapeIR2      LandContourHLS    LotConfigCulDSac NeighborhoodBrkSide
       0.033334781         0.046178311         0.045421975         0.021268303         0.098556491         0.008661092
NeighborhoodCrawfor NeighborhoodNoRidge NeighborhoodNridgHt NeighborhoodSomerst NeighborhoodStoneBr NeighborhoodVeenker
       0.195070562         0.467568991         0.446022832         0.071167379         0.473709307         0.011041240
    Condition1Norm         OverallQual         OverallCond           YearBuilt        YearRemodAdd        RoofStyleHip
       0.083672604         0.198714138         0.046044530         0.049470243         0.037813952         0.008060825
    RoofMatlWdShngl  Exterior1stBrkFace  Exterior1stCemntBd  Exterior2ndCmentBd  Exterior2ndImStucc          MasVnrArea
       0.797273922         0.136694809         0.089104350         0.005908721         0.108631354         0.039096774
    FoundationPConc      BsmtExposureGd     BsmtFinType1GLQ          BsmtFinSF1         TotalBsmtSF           X1stFlrSF
       0.031200404         0.197203757         0.064601171         0.049840362         0.052250295         0.044052532
         X2ndFlrSF           GrLivArea        BsmtFullBath            FullBath            HalfBath        TotRmsAbvGrd
       0.039178237         0.190944120         0.024526702         0.029625563         0.010313852         0.049378964
     FunctionalTyp          Fireplaces   GarageTypeBuiltIn         GarageYrBlt          GarageCars          GarageArea
       0.091480366         0.040397526         0.044844888         0.006092850         0.067482846         0.028049039
      GarageQualGd          WoodDeckSF         OpenPorchSF         ScreenPorch            PoolArea         SaleTypeNew
       0.027031512         0.024129234         0.002610172         0.014574512         0.020192711         0.153763780
SaleConditionPartial
       0.068127967
```