IS605 Final Exam

Marco Siqueira Campos

1. Instructions:

Your final is due by the end of day on 05/24/2016. You should post your solutions to your GitHub account. You are also expected to make a short presentation during our last meeting (3-5 minutes) or post a recording to the board. This project will show off your ability to understand the elements of the class.

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques> . I want you to do the following.

Pick **one** of the quantitative independent variables from the training data set (train.csv) , and define that variable as X. Pick **SalePrice** as the dependent variable, and define it as Y for the next analysis.

Competition information:

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

1. Probability:

Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the 4th quartile of the X variable, and the small letter "y" is estimated as the 2d quartile of the Y variable. Interpret the meaning of all probabilities.

1. P(X>x | Y>y) b. P(X>x, Y>y) c. P(X<x | Y>y)

Follow steps was made:

a-Choose one quantitative variable: **LotArea.**

b-Find the value of 2nd Quartile (median) for **SalePrice.**

c-Find the value of 4rd Quartile for **LotArea.**

d-Compute the frequency (probability) for combination **SalePrice** and **LotArea.**

e-Compute the probability *‘a’* to *‘c’*.

My X quantitative variable is **LotArea** and the 4thQuartile is the maximum value=215200

P(X>x)=P(LotArea>215200) = 0

P(X<x)=P(LotArea<215200) = 1

My Y variable is **SalePrice** and the 2nd Quartile (median) is = 163000

P(Y>y)=P(SalePrice>163000) = 0.5

P(Y<y)=P(SalePrice<163000) = 0.5

Contingency table

|  |  |  |  |
| --- | --- | --- | --- |
|  | **P(Y** | **P(Y>y)** | **Total** |
| **P(X>x)** | 0 | 0 | 0 |
| **P(X** | 0.5 | 0.5 | 1 |
| **total** | 0.5 | 0.5 | 1 |

Tab. 1

1. P(X>x | Y>y)

P(A|B) = P(A B) / P(B) = 0/0.5 = 0

1. P(X>x, Y>y)

P(AB) = 0

1. P(X<x | Y>y)

P(A|B) = P(A B) / P(B) = 0.5 /0.5 = 1

Interpret the meaning of all probabilities.

1. Is the probability of P(X>x) with reduced sample space to P(Y>y)

In this case is probability of **LotArea**>215200 with only sample space of **SalePrice**>163000.

As P(**LotArea**>215200) = 0, the final probability will be 0.

1. Is the probability that both P(X>x) and P(Y>y) , is the probability that both P(X>x) and P(Y>y), or **LotArea**>215200 and **SalePrice**>163000 happen, is equivalent to P(AB).

As **LotArea**>215200 is 0 the probability will be zero.

1. Is the probability of P(X<x) with reduced sample space to P(Y>y)

In this case is probability of **LotArea<**215200 with only sample space of **SalePrice**>163000.

The final probability will be 1.

Does splitting the training data in this fashion make them independent? In other words, does P(X|Y)=P(X)P(Y))? Check mathematically, and then evaluate by running a Chi Square test for association. You might have to research this.

No, this splitting the train data cannot change the independence of data, for P(X and Y) is only P(X) \* P(Y) when is the variable are independent.

To avoid cell with 0 at chi-squared test I will use for test X = P(X>x), x is the 3rd quartile and Y = P(Y>y), y is the 2nd quartile

Variable Frequency Probability

X = P(X>3Q) = P(LotArea>11600) = 365 365/1460 = 0.25

Y = P(Y>2Q)= P(SalePrice>163000) = 728 728/1460 = 0.4986

P(X|Y) = P(X)P(Y)

For P(X|Y) I counted the frequency in data base 276/728 = 0.379

P(X|Y)=0.379 ≠ P(X).P(Y) = 0.25 \* 0.4986 = 0.12465

Contingency table for Chi-squared test

|  |  |  |  |
| --- | --- | --- | --- |
|  | **P(Y** | **P(Y>y)** | **Total** |
| **P(X>x)** | 89 | 276 | 365 |
| **P(X** | 643 | 452 | 1095 |
| **total** | 732 | 728 | 1460 |

Tab. 2

library(MASS)

X\_GT = c(89, 276)

X\_LT = c(643, 452)

XY = as.data.frame(rbind(X\_GT, X\_LT))

names(XY) = c('Y\_LT', 'Y\_GT')

XY

Y\_LT Y\_GT

X\_GT 89 276

X\_LT 643 452

chisq.test(XY)

Pearson's Chi-squared test with Yates' continuity correction

data: XY

X-squared = 127.74, df = 1, p-value < 2.2e-16

the row and the column variables are statistically significantly associated (p-value < 0.05), there is no independence between X and Y.

1. Descriptive and inferencial Statistics:

Provide univariate descriptive statistics and appropriate plots for both variables. Provide a scatterplot of X and Y. Transform both variables simultaneously using Box-Cox transformations. You might have to research this. Using the transformed variables, run a correlation analysis and interpret. Test the hypothesis that the correlation between these variables is 0 and provide a 99% confidence interval. Discuss the meaning of your analysis.

Descriptive analysis for X, **LotArea**

summary(train$LotArea)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1300 7554 9478 10520 11600 215200

sd(train$LotArea)

[1] 9981.265

par(mfrow=c(2,2))

hist(train$LotArea, col = "blue")

boxplot(train$LotArea, main="Boxplot LotArea")

qqnorm(train$LotArea)

qqline(train$LotArea)

par(mfrow=c(1,1))

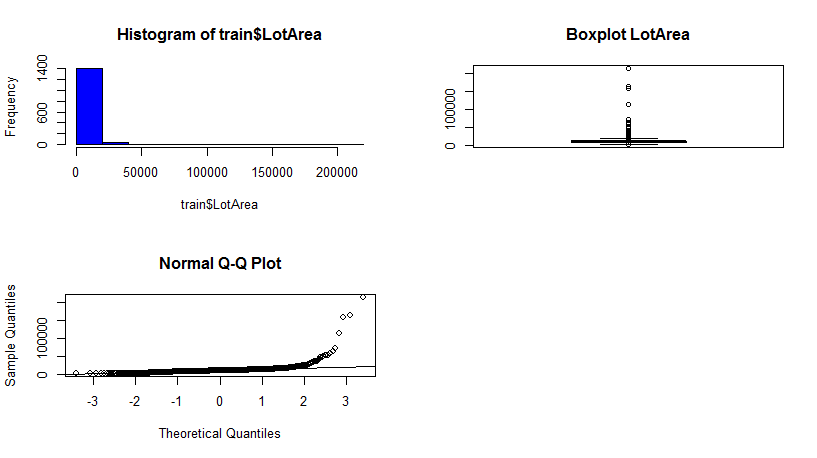


Fig. 1

Descriptive analysis for Y, **SalePrice**

summary(train$SalePrice)

Min. 1st Qu. Median Mean 3rd Qu. Max.

34900 130000 163000 180900 214000 755000

sd(train$SalePrice)

[1] 79442.5

par(mfrow=c(2,2))

hist(train$SalePrice, col = "red")

boxplot(train$SalePrice, main="Boxplot SalePrice")

qqnorm(train$SalePrice)

qqline(train$SalePrice)

par(mfrow=c(1,1))

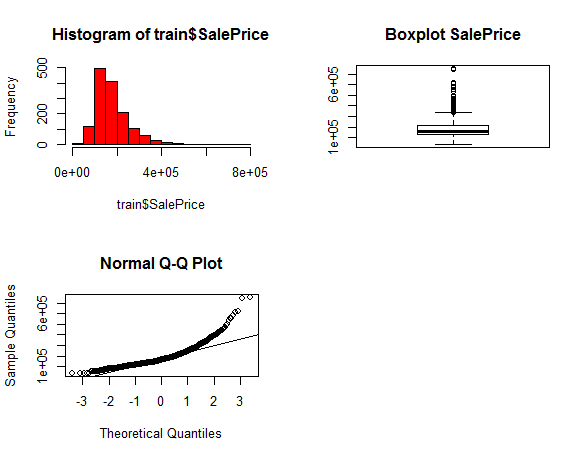


Fig.2

plot(train$LotArea,train$SalePrice, main = "Scatterplot SalePrice by LotArea ", col="red")

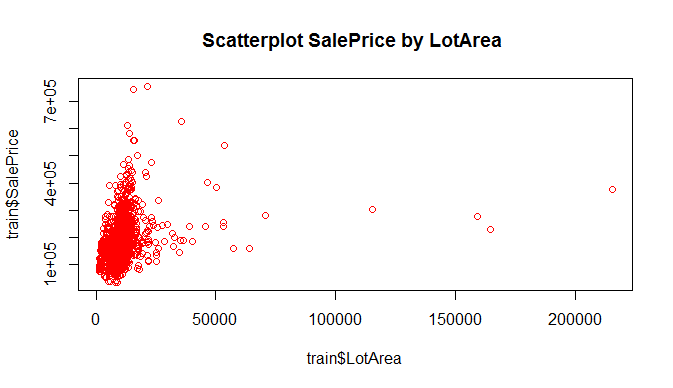


Fig. 3

The two distribution, are asymmetric, right skew, are not normal. The outliers at boxplot is not a real outlier is more characteristic at this kind of distribution.

The **LotArea** have a huge variation, we have a high concentration in beginning and we have some cases with very large area, the 3rd quartile is far from the maximum value.

The scatterplot show a light positive relationship between the two variables.

For Box-Cox transformation our greatest objective is to reduce the non-normality of the residual for X (**LotArea**) and Y (**SalePrice**), because our main interest is in correlation between X and Y, and do not individually do the transformation of X and Y;

We found the lambda parameter of Y that give the better results at the residual.

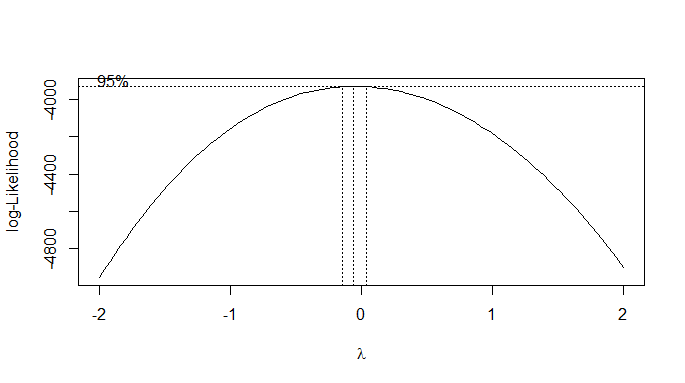


Fig. 4 Box-Cox Lambda Plot

The Box-Cox transformation Z= =

Lambda = -0.06060606

lm( train$SalePrice~ train$LotArea)

bc <- boxcox(train$SalePrice ~ train$LotArea)

lambda <- bc$x[which.max(bc$y)]

mnew <- lm(train$SalePrice^lambda ~ train$LotArea)

op <- par(pty = "s", mfrow = c(1, 2))

qqnorm(m$residuals, main="Normal QQ Plot - after"); qqline(m$residuals)

qqnorm(mnew$residuals, main="Normal QQ Plot - before trans"); qqline(mnew$residuals)

par(op)

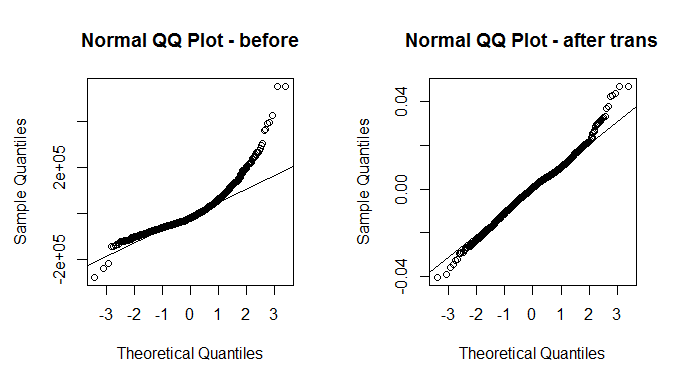


Fig. 4 Residual QQ Plot before and after lambda transformation

Correlation between X and Z (Y transformed).

The correlation between X and Z is: r = -0.2558218

Hypothesis test for:

H0: θ = 0

H1: θ ≠ 0

As p < 0.05, we reject the null hypothesis, H0: θ = 0, and the correlation coefficient is statistically different from zero for significant level of α=0.05.

The confidence interval for 99% is P(-0.3177247 > θ > -0.1917473 )=0,99

cor.test(train$SalePrice^lambda,train$LotArea, method = "pearson", alternative = "two.sided",

+ estimate="rho", conf.level = 0.99)

Pearson's product-moment correlation

data: train$SalePrice^lambda and train$LotArea

t = -10.104, df = 1458, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

99 percent confidence interval:

-0.3177247 -0.1917473

sample estimates:

cor

-0.2558218

The meaning of the correlation analysis is: There is weak linear negative correlation between X and Z (Y transformed variable by Box-Cox). The correlation coefficient is -0.26, and the Confidence Interval with 0.99 is (-0.32, -0.19).

A warning must be made here, in the original correlation, X and Y, the correlation was positive, the transformation altered meaning and practical meaning. The analysis here must be done considering that the correlation is really positive rather than negative.

1. Linear Algebra and Correlation:

Invert your correlation matrix (This is known as the precision matrix and contains variance inflation factors on the diagonal). Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix.

To do the correlation matrix the follow variable was choose: **SalePrice, LotArea, GarageArea** and **MasVnrArea**.

Correlation matrix

m<-train[,c("LotArea","SalePrice","GarageArea","MasVnrArea")]

cor(na.omit(m))

LotArea SalePrice GarageArea MasVnrArea

LotArea 1.0000000 0.2646740 0.1807779 0.1041598

SalePrice 0.2646740 1.0000000 0.6224917 0.4774930

GarageArea 0.1807779 0.6224917 1.0000000 0.3730665

MasVnrArea 0.1041598 0.4774930 0.3730665 1.0000000

Invert the correlation matrix, the precision matrix

mcor<-cor(na.omit(m))

solve(mcor)

LotArea SalePrice GarageArea MasVnrArea

LotArea 1.07670387 -0.2811282 -0.03239281 0.03417217

SalePrice -0.28112823 1.9162968 -0.94284209 -0.53399337

GarageArea -0.03239281 -0.9428421 1.65371565 -0.16337131

MasVnrArea 0.03417217 -0.5339934 -0.16337131 1.31236711

Multiplying the correlation matrix by the precision matrix

mp<-solve(mcor)

mcor%\*%mp

LotArea SalePrice GarageArea MasVnrArea

LotArea 1.000000e+00 -3.469447e-17 3.469447e-18 0.000000e+00

SalePrice -4.510281e-17 1.000000e+00 -6.938894e-17 1.110223e-16

GarageArea -1.214306e-17 -8.326673e-17 1.000000e+00 0.000000e+00

MasVnrArea -1.387779e-17 -1.110223e-16 0.000000e+00 1.000000e+00

Multiplying the precision matrix by correlation matrix

mp%\*%mcor

LotArea SalePrice GarageArea MasVnrArea

LotArea 1.000000e+00 -5.204170e-17 -1.908196e-17 -3.469447e-17

SalePrice -9.020562e-17 1.000000e+00 -4.163336e-16 -1.110223e-16

GarageArea 1.110223e-16 2.498002e-16 1.000000e+00 8.326673e-17

MasVnrArea 2.775558e-17 1.110223e-16 5.551115e-17 1.000000e+0

1. Calculus-Based Probability & Statistics:

Many times, it makes sense to fit a closed form distribution to data. For your non-transformed independent variable, location shift it so that the minimum value is above zero. Then load the MASS package and run fitdistr to fit a density function of your choice. (See <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html> ). Find the optimal value of the parameters for this distribution, and then take 1000 samples from this distribution (e.g., rexp(1000, ) for an exponential). Plot a histogram and compare it with a histogram of your non-transformed original variable.

References:

1- <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/boxcox.html>

2- <http://rcompanion.org/handbook/I_12.html>

3- <http://stackoverflow.com/questions/33999512/how-to-use-the-box-cox-power-transformation-in-r>

4- <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.test.html>

5- <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html>