Assignment #4: Interpreting Regression Models (100 points)

Data: The data for this assignment is the Ames, Iowa housing data set. This data will be made available by your instructor.

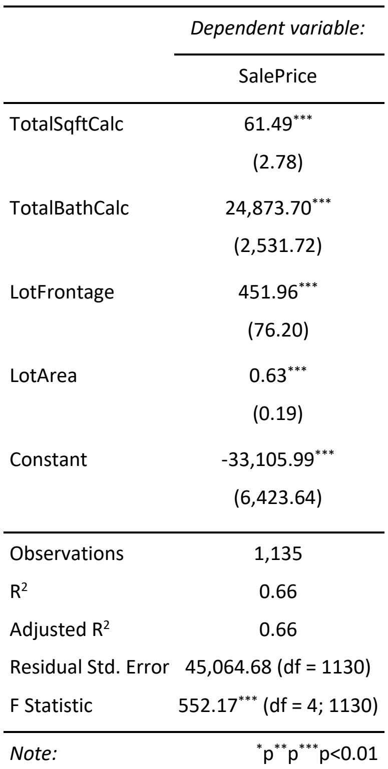
Assignment Instructions:

In this assignment we will learn how to interpret regression models for the home sale price. We will fit specific models, interpret them as regression models and within context, and score the models to produce estimated home prices for hypothetical homes. Use the starter script to help you complete this assignment.

Part 1: Decomposing the Home Value

Fit the MLR model with the response variable SalePrice and the predictor variables TotalSqftCalc, TotalBathCalc, LotFrontage, and LotArea. You should get this fitted model.

Table 1: Model #1



	Dependent variable:
	SalePrice
TotalSqftCalc	61.49^{***}
	(2.78)
TotalBathCalc	$24,873.70^{***}$
	(2,531.72)
LotFrontage	451.96^{***}
	(76.20)
LotArea	0.63^{***}
	(0.19)
Constant	$-33,105.99^{***}$
	(6,423.64)
Observations	1,135
R^2	0.66
Adjusted R^2	0.66
Residual Std. Error	45,064.68 (df = 1130)
F Statistic	552.17^{***} (df = 4; 1130)
Note:	$^{*}p^{**}p^{***}p<0.01$

In R the regression coefficients should show like this.

```
> model.1$coef

  (Intercept) TotalSqftCalc TotalBathCalc   LotFrontage       LotArea

-3.310599e+04  6.148968e+01  2.487370e+04  4.519610e+02  6.263278e-01
```

In textbook regression problems we are taught to interpret the model coefficients as the change in Y for a unit change in X. However, when we are presenting and discussion our models, we do not discuss them in that manner. Typically we want to discuss our models in canonical, or natural or contextual, units.

We will consider the 'typical' house in Ames, IA to have the following values for the model features.

TotalSqftCalc	TotalBathCalc	LotFrontage	LotArea
2100	2.5	75	11000

Let's compute 'model scores' (home price estimates) for this typical house and some variants on this house. The starter script for this assignment will show you how to use matrix multiplication in R to compute these values.

(1) (5 points) Using Model #1 compute the estimated home price for the typical home with (TotalSqftCalc, TotalBathCalc, LotFrontage,   LotArea) = (2100, 2.5, 75, 11000).

198993.3 dollars

(2) (5 points) Add 400 sqft to our typical home. Using Model #1 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage,   LotArea) = (2500, 2.5, 75, 11000).

223589.1 dollars

(3) (5 points) Add 1.5 bathrooms to our typical home. Using Model #1 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage,   LotArea) = (2100, 4.0, 75, 11000).

236303.8 dollars

(4) (5 points) Add 4000 sqft of lot size to our typical home. Using Model #1 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage, LotArea) = (2100, 2.5, 75, 15000).

201498.6 dollars

Part 2: How should we interpret a regression model for log(SalePrice)?

Fit the MLR model with the response variable log(SalePrice) and the predictor variables TotalSqftCalc, TotalBathCalc, LotFrontage, and LotArea. You should get this fitted model.

Table 2: Model #2

	Dependent variable:
	log(SalePrice)
TotalSqftCalc	0.0002***
	(0.0000)
TotalBathCalc	0.17***
	(0.01)
LotFrontage	0.002***
	(0.0003)
LotArea	0.0000***
	(0.0000)
Constant	11.08***
	(0.03)
Observations	1,135
R^2	0.68
Adjusted R^2	0.67

Residual Std. Error	0.19 (df = 1130)
F Statistic	588.77*** (df = 4; 1130)
Note:	*p**p***p<0.01

When we log-transform the response variable Y in a linear model we create (or specify) a log-linear model. While Model #1 is specified with additive effects on the raw scale (SalePrice), Model #2 is specified with additive effects on the log-scale, that means that the effects are multiplicative on the raw scale. The interpretation of this transformation is initially problematic for our new greenhorn students. Let's walk through this model in the same manner that we did for Model #1 to get a better understanding of this transformation. Help and hints are available in the starter script.

(5) (5 points) Jensen's Inequality – f(E[X]) <= E[f(X)] for a convex function f

Switching back and forth between the original scale and the log-scale can cause some confusion in greenhorns. Some of that confusion can be explained by Jensen's Inequality. The confusion arises because we expect quantities to be equal, when in fact they will not be. Compute the exp(mean(log(ames.df$SalePrice))) and note that it is less than mean(exp(log(ames.df$SalePrice))) = mean(ames.df$SalePrice).

exp(mean(log(ames.df$SalePrice))) = 185348.6

mean(exp(log(ames.df$SalePrice))) = 197211.5

mean(ames.df$SalePrice) = 197211.5

(6) (5 points) Using Model #2 compute the estimated home price for the typical home with (TotalSqftCalc, TotalBathCalc, LotFrontage, LotArea) = (2100, 2.5, 75, 11000). Name this value v.0.

v.0 = 12.14898
exp(v.0) = 188901.6 dollars

(7) (5 points) Add 400 sqft to our typical home. Using Model #2 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage, LotArea) = (2500, 2.5, 75, 11000). Name this value v.1.

v.1 = 12.23472

exp(v.1) = 205811.9 dollars

(8) (5 points) Compute the percent increase in the home value from adding 400 sqft to our typical home. Verify your answer by computing exp(v.1) / exp(v.0). Note that the percent increase is the decimal portion, i.e. a factor of 1.23 means a 23% increase.

8.95% increase

Exp(v.1)/exp(v.0) = 1.089519

(9) (5 points) Add 1.5 bathrooms to our typical home. Using Model #2 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage, LotArea) = (2100, 4.0, 75, 11000). Name this value v.2.

v.2 = 12.40779

exp(v.2) = 244700.4 dollars

(10) (5 points) Compute the percent increase in the home value from adding 1.5 bathrooms to our typical home. Verify your answer by computing exp(v.2) / exp(v.0). Note that the percent increase is the decimal portion, i.e. a factor of 1.23 means a 23% increase.

29.54% increase

Exp(v.2)/exp(v.0) = 1.295385

(11) (5 points) Add 4000 sqft of lot size to our typical home. Using Model #2 compute the estimated home price for a home with
(TotalSqftCalc, TotalBathCalc, LotFrontage, LotArea) = (2100, 2.5, 75, 15000). Name this value v.3.

v.3 = 12.158

exp(v.3) = 190612.8 dollars

(12) (5 points) Compute the percent increase in the home value from adding 4000 sqft of lot size to our typical home. Verify your answer by computing exp(v.3) / exp(v.0). Note that the percent increase is the decimal portion, i.e. a factor of 1.23 means a 23% increase.

0.009% increase

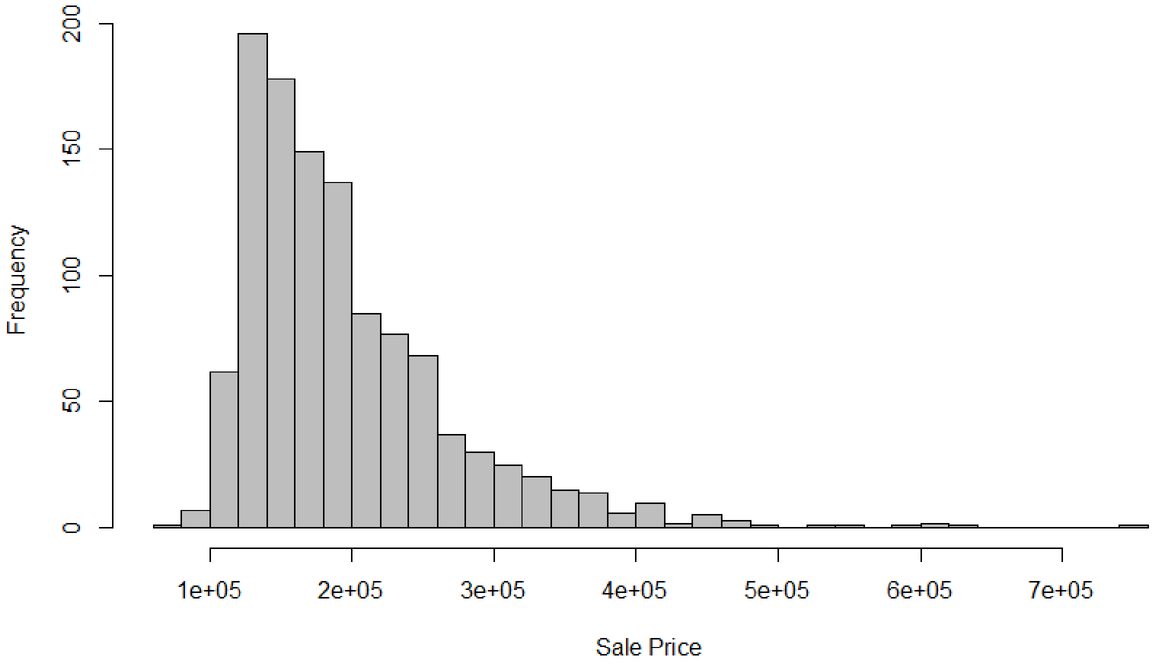
Exp(v.3)/exp(v.0) = 1.009058

Part 3: Why did we transform the response variable?

Frequently a highly skewed response variable is transformed towards symmetry. The two typical transformations are the log transformation and the square root transformation. The square root transformation is used when the response variable contains zeros.

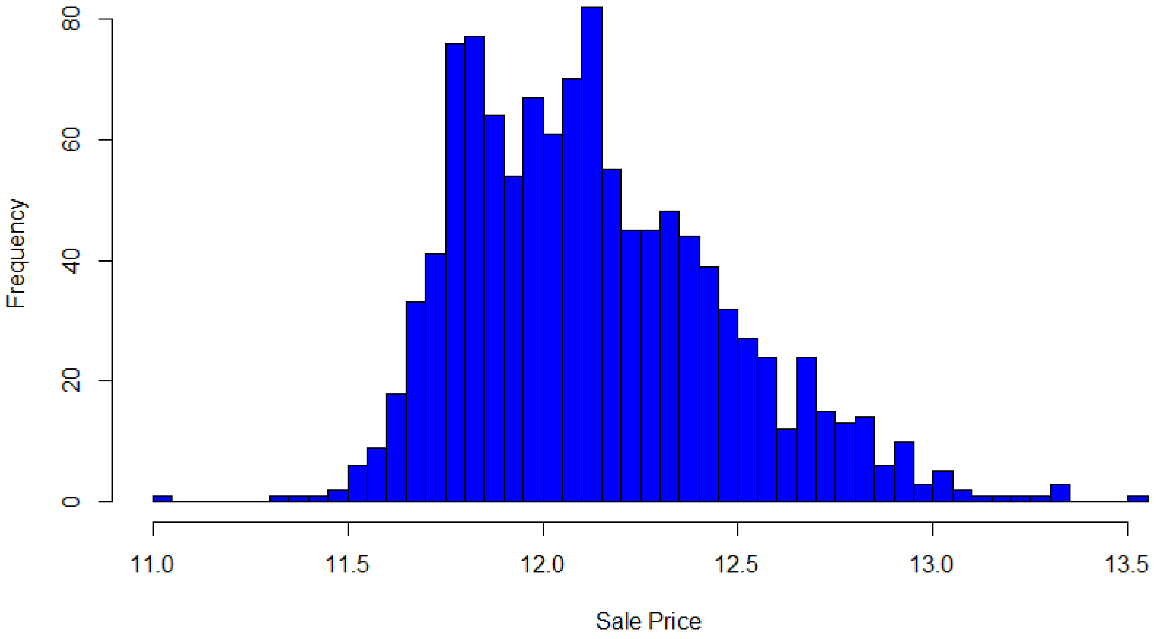
We can compute the skewness of a variable using the skewness() function in the moments package. Install the package in order to complete the questions in this section.

(13) (10 points) Produce a histogram with n=40 bins and compute the skewness for SalePrice. Color the histogram grey. Paste the histogram into the template.



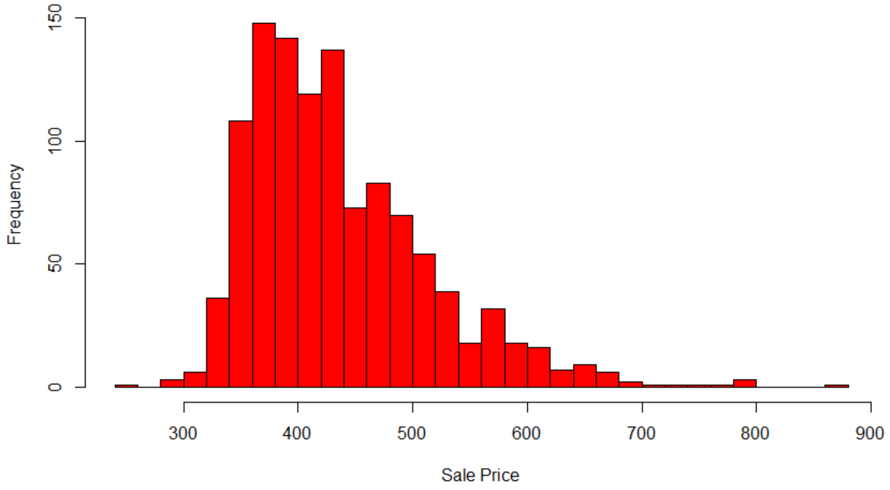
Skewness = 1.928376

(14) (10 points) Produce a histogram with n=40 bins and compute the skewness for log(SalePrice). Color the histogram blue. Paste the histogram into the template.



Skewness = 0.6652866

(15) (10 points) Produce a histogram with n=40 bins and compute the skewness for sqrt(SalePrice). Color the histogram red. Paste the histogram into the template.

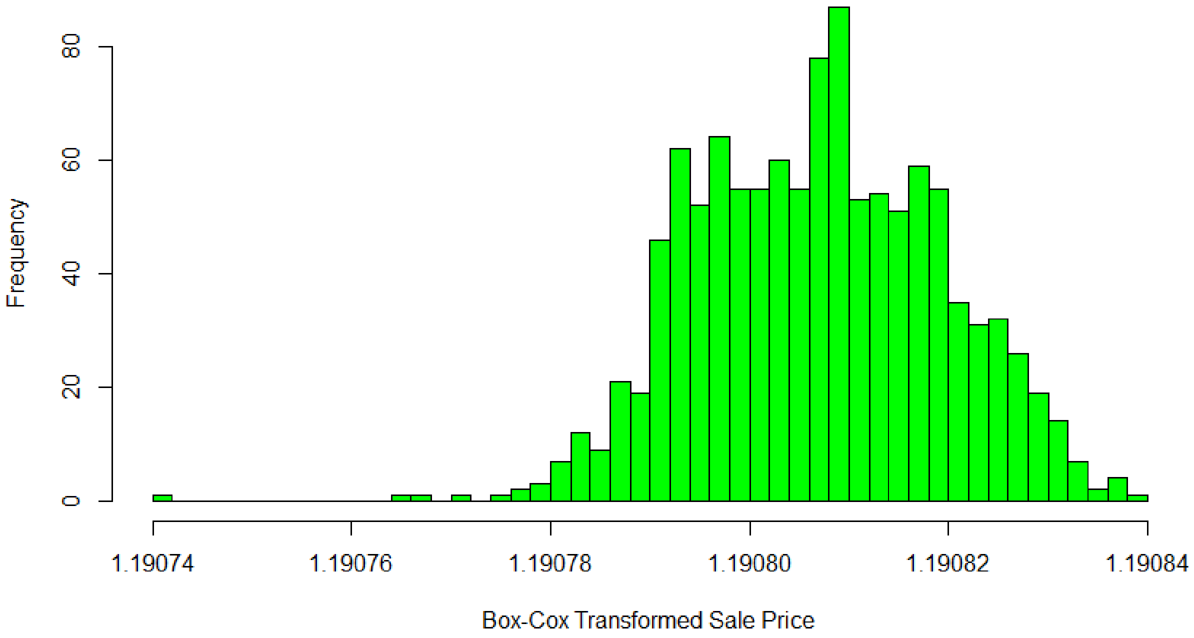


Skewness = 1.212218

Part 4: The Box-Cox Transformation

The Box-Cox transformation is available in the forecast package. Install the forecast package and use the BoxCox.lambda() function to compute the optimal lambda for the Box-Cox transformation and the BoxCox() function to make the optimal Box-Cox transformation of SalePrice.

(16) (10 points) Produce a histogram with n=40 bins and compute the skewness for BoxCox(SalePrice). Color the histogram green. Paste the histogram into the template.



Skewness = -0.1184782