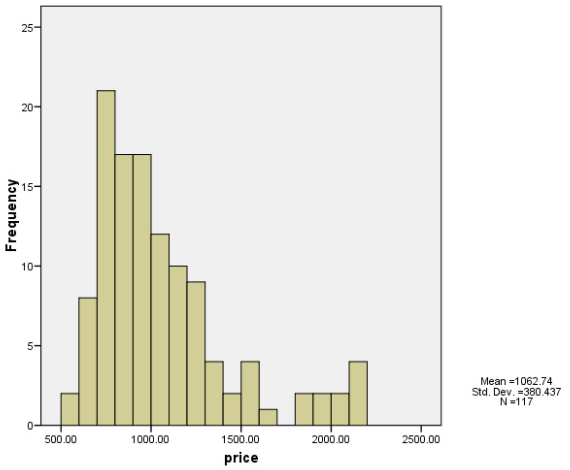
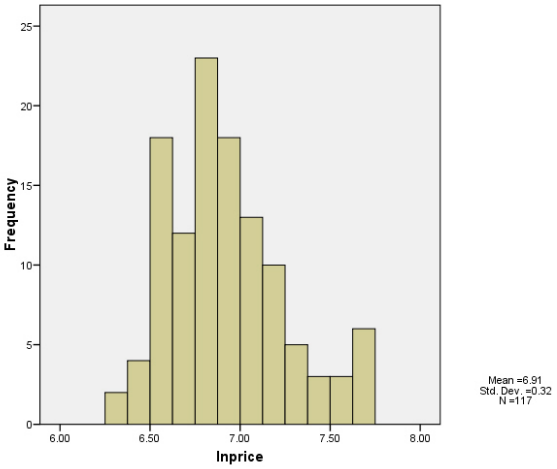
1 Transformations in Multiple Linear Regression

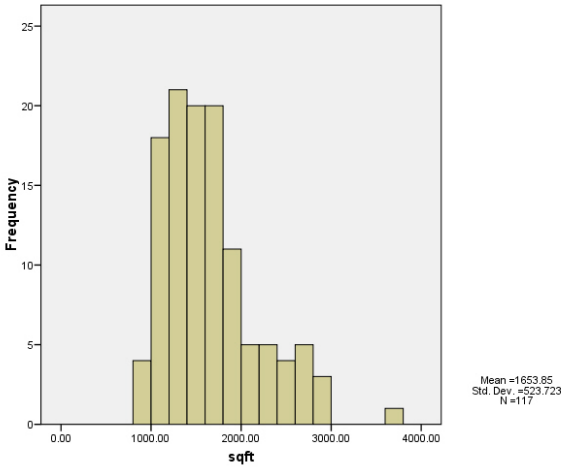
1.1 Logarithmic

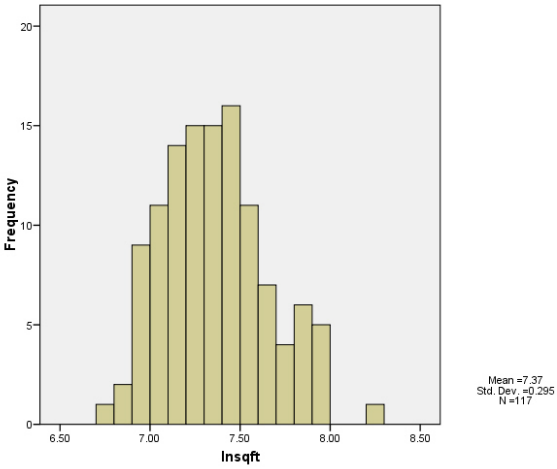
Albuquerque Real Estate Data:
The distribution of the response variable y =price is skewed to the right. Transforming it with the logarithmic function (ln), will result in a more "normal" distribution.





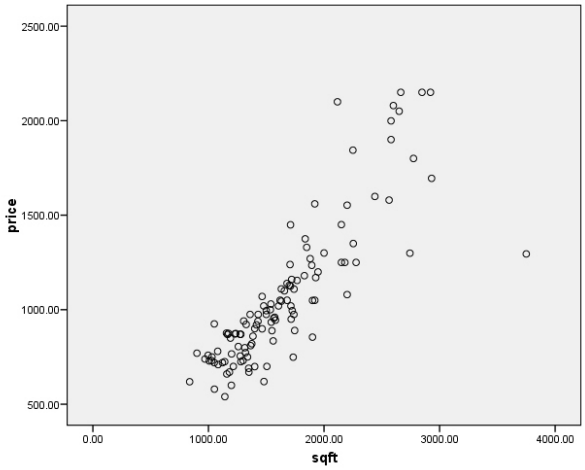
The same observation is true for sqft

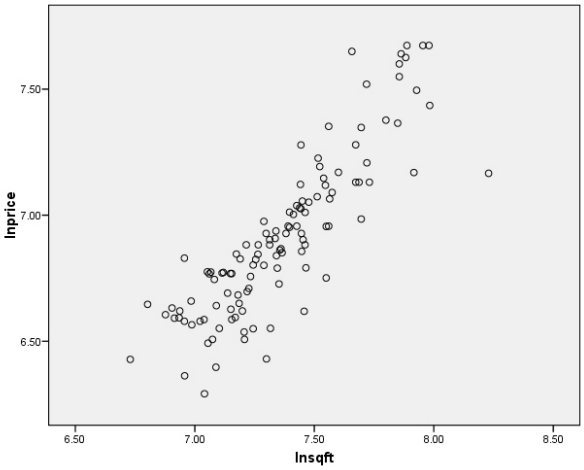




Using the transformed variables in a linear regression model will improve the fit but also the observations in the Residual Analysis.

First compare the **scatterplots**.





Correlation coefficients
$r_{price,sqft} = .845$ and $r_{lnprice,lnsqft} = .863$

Models

$$price = \beta_0 + \beta_1 sqft + e, \quad e \sim \mathcal{N}(0, \sigma), \qquad (1)$$

$$ln(price) = \beta_0 + \beta_1 ln(sqft) + e, \quad e \sim \mathcal{N}(0, \sigma), \qquad (2)$$

this is equivalent to

$$price = e^{\beta_0 + \beta_1 ln(sqft) + e} = e^{\beta_0}\ sqft^{\beta_1}\ e^{e}, \quad e \sim \mathcal{N}(0, \sigma)$$

Regression Analysis
Model (1) R Square = .714
$\hat{\beta}_0 = 47.819, SE(\beta_0) = 62.855$
$\hat{\beta}_1 = .614, SE(\beta_1) = .036$

Model (2) R Square = .745
$\hat{\beta}_0 = .017, SE(\beta_0) = .377$
$\hat{\beta}_1 = .936, SE(\beta_1) = .051$

1.2 Box-Cox Method

Box-Cox Method (paper in 1964) for finding the "best" transformation, to obtain normally distributed sample data.
Consider the following transformations for different values of λ

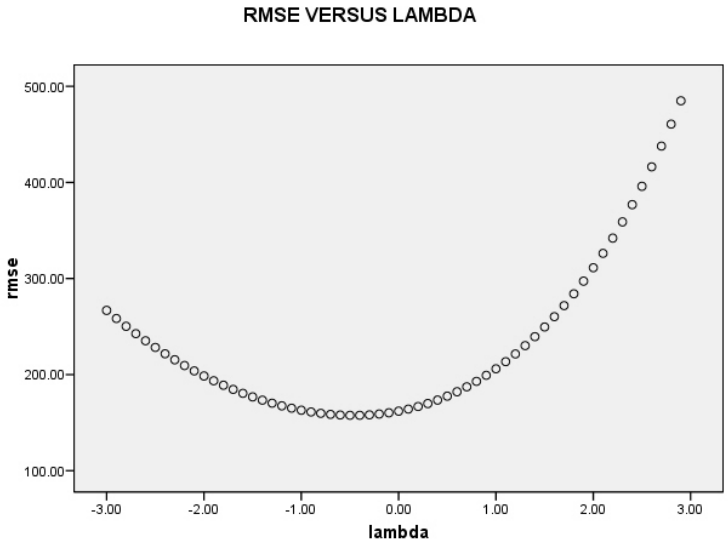
$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(y), & \text{if } \lambda = 0 \end{cases}$$

After the transformation of the response variable, run the regression model and obtain the SSE from the ANOVA table.
Find the λ with the smallest SSE ($\sqrt{MSE}$). Given those results choose the "best" transformation according to the following table.

Best λ	Equation	Name
-2.5 to -1.5	$1/y^2$	inverse square
-1.5 to -0.75	$1/y$	reciprocal
-0.75 to -0.25	$1/\sqrt{y}$	inverse square root
-0.25 to 0.25	$\ln(y)$	natural log
0.25 to 0.75	$\sqrt{y}$	square root
0.75 to 1.5	y	none
1.5 to 2.5	y^2	square

Example: Albuquerque
y =price, x =lnsqft

λ	RMSE
-3.00	266.84
-2.50	228.29
-2.00	198.56
-1.50	176.81
-1.00	162.94
-.50	157.60
.00	161.98
.50	177.58
1.00	206.05
1.50	249.54
2.00	311.24
2.50	395.94
3.00	484.91



According to these results the best lambda is -0.5, therefore the transformation $1/\sqrt{y}$ should be chosen.