BUDT 730
Data, Models and Decisions

Lecture 13

Regression Analysis (5)

Transformation of Variables

Prof. Sujin Kim

Quiz 10: Catalog_Marketing_Reg.xlsx

- Build a linear regression model: AmountSpent = Salary + Gender
  - Gender: 1 if male, 0 if female
  - Write the two regression equations:
    - Equation for male (1)
    - Equation for female (0)

  - Interpret the coefficient of Gender

- Add an interaction term to the model
  - Write the two regression equations:
    - Equation for male (1)
    - Equation for female (0)
  - Is the interaction term useful for explaining AmountSpent? Explain why or why not.

Practice: Catalog_Marketing_Reg.xlsx

- Build a linear regression model: AmountSpent = Salary + Gender
  - Write the regression equation

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.516e+01	4.680e+01	-0.538	0.591	
Salary	2.180e-02	7.357e-04	29.626	<2e-16	***
factor(Gender)1	3.866e+01	4.503e+01	0.859	0.391	

AmountSpent = - 25.16 + 0.02179586 Salary +38.66 (Gender =1)

Gender = 0	AmountSpent = - 25.16 + 0.0218 Salary
Gender = 1	AmountSpent = 13.50+ 0.0218 Salary

  - Interpret the coefficient of Gender
  - On average, AmounSpent by male is $38.66 larger than AmountSpent by female when the salary is the same.

Practice: Catalog_Marketing_Reg.xlsx

- Add an interaction term to the model
  - Write the regression equation

AmountSpent = - 51.91 + 0.0224 Salary +101.74 (Gender =1)-0.0011 Salary*(Gender=1)

Gender = 0	AmountSpent = - 51.91 + 0.0224 Salary
Gender = 1	AmountSpent = 49.83 + 0.0212 Salary

  - Is the interaction term useful for explaining AmountSpent? Explain why or why not

Overall, two models are very similar. Particularly, the coefficient of Salary does now change much with interaction variable. The interaction term does not contribute much on explain AmpuntSpent.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -2.516e+01  4.680e+01  -0.538    0.591    
Salary          2.180e-02  7.357e-04  29.626   <2e-16 ***
factor(Gender)1 3.866e+01  4.503e+01   0.859    0.391    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.2 on 997 degrees of freedom
Multiple R-squared:  0.4898,	Adjusted R-squared:  0.4888 
F-statistic: 478.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -51.909526  58.522788  -0.887    0.375    
Salary                  0.022351   0.001036  21.582   <2e-16 ***
factor(Gender)1       101.738029  94.282615   1.079    0.281    
Salary:factor(Gender)1  -0.001121   0.001472  -0.762    0.447    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.3 on 996 degrees of freedom
Multiple R-squared:  0.4901,	Adjusted R-squared:  0.4886 
F-statistic: 319.1 on 3 and 996 DF,  p-value: < 2.2e-16
```

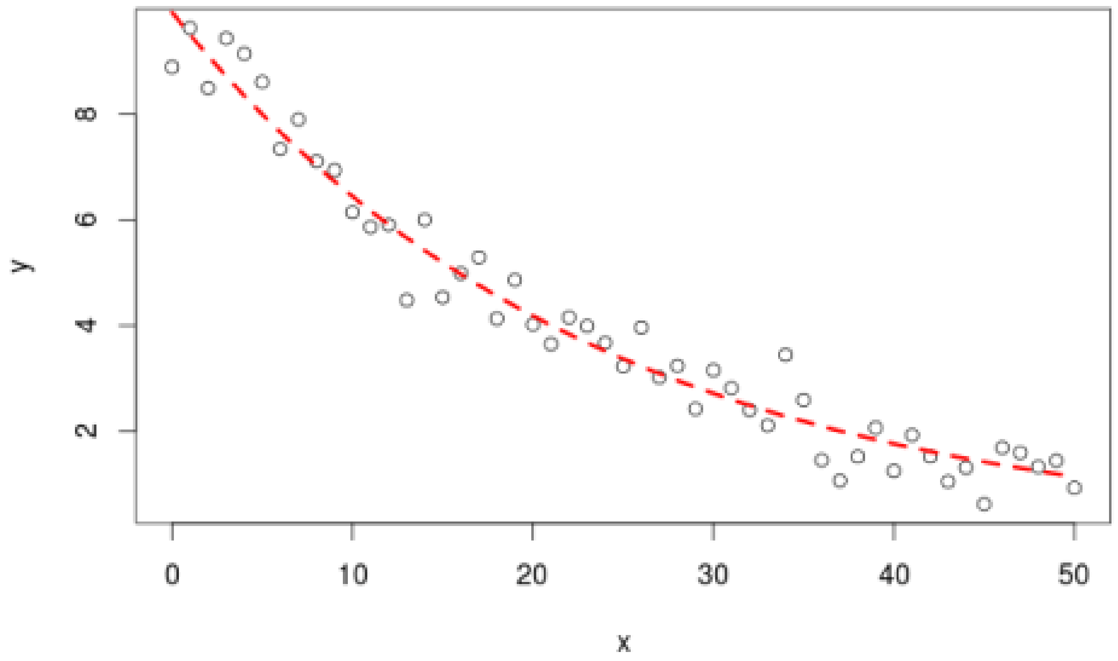
Variable Transformations

- Several types of independent variables can be used in regression equations:
  - Dummy variables
  - Interaction variables
  - Nonlinear transformations
- Dataset:
  - `DetergentSales.xlsx`

Extractor Functions for lm

Function	Description
`summary`	returns summary information about the regression
`plot`	makes diagnostic plots
`coef`	returns the coefficients
`confint`	returns confidence intervals for the coefficients
`vcov`	estimated covariance between parameter estimates
`residuals`	returns the residuals (can be abbreviated `resid`)
`fitted`	returns fitted values, $\widehat{y}_i$
`deviance`	returns RSS
`predict`	performs predictions
`anova`	finds various sums of squares
`AIC`	is used for model selection
`model.matrix`	matrix used to fit model mathematically

Table 11.1: Generic extractor functions for many of R's modeling functions, including `lm`.



Regression Model with Nonlinear Variables

Linear vs. Nonlinear Models

- So far we have focused on linear regression models.
- Consider a simple linear regression model:

$$Y = a + bX$$

- Linear models assume the change in Y associated with a unit increase in X does not depend on the value of X .
  - In other words: the slope is constant for all X

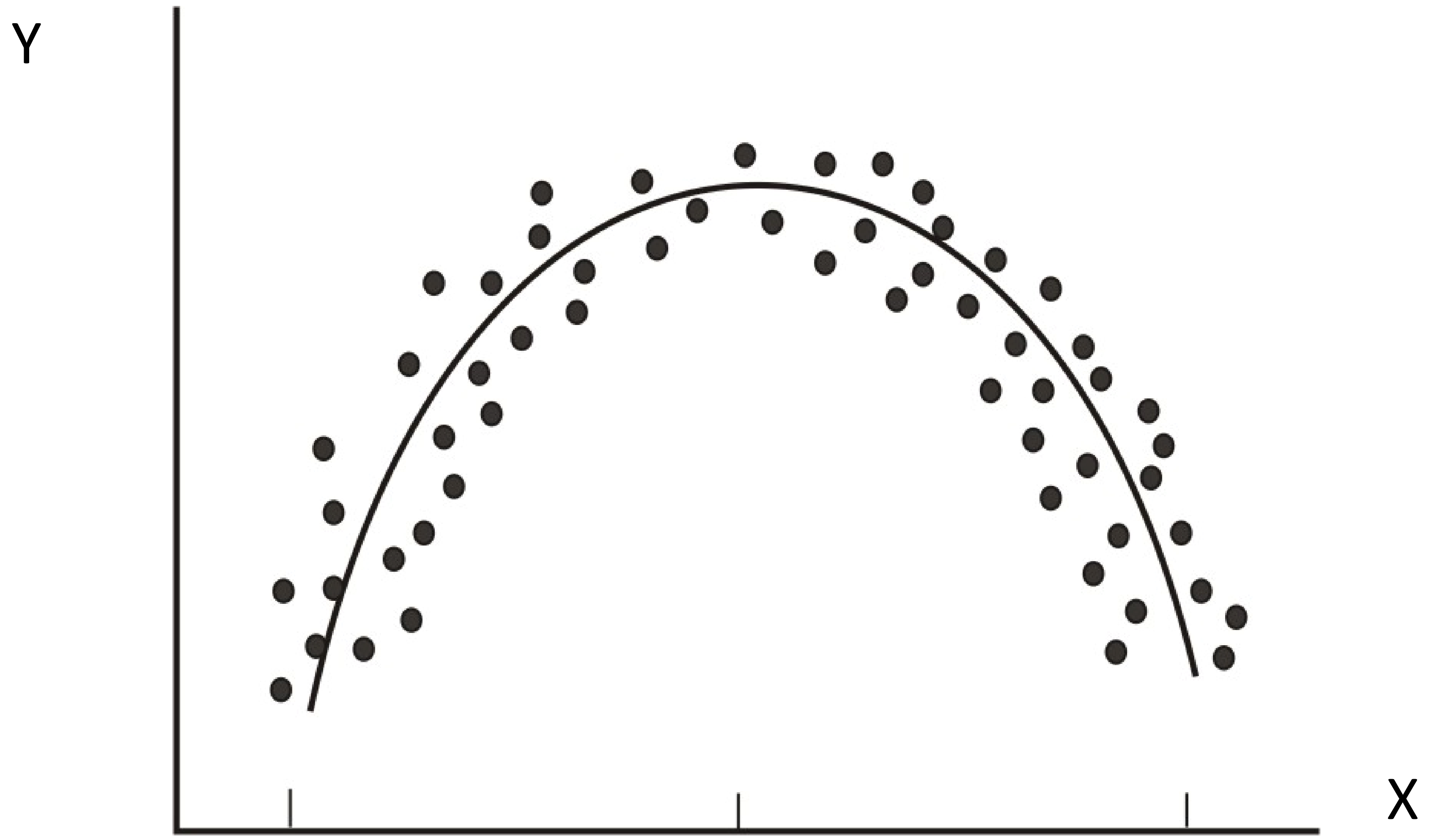
$$\frac{dY}{dX} = b$$

- Now, we model the relation between X and Y by a nonlinear function

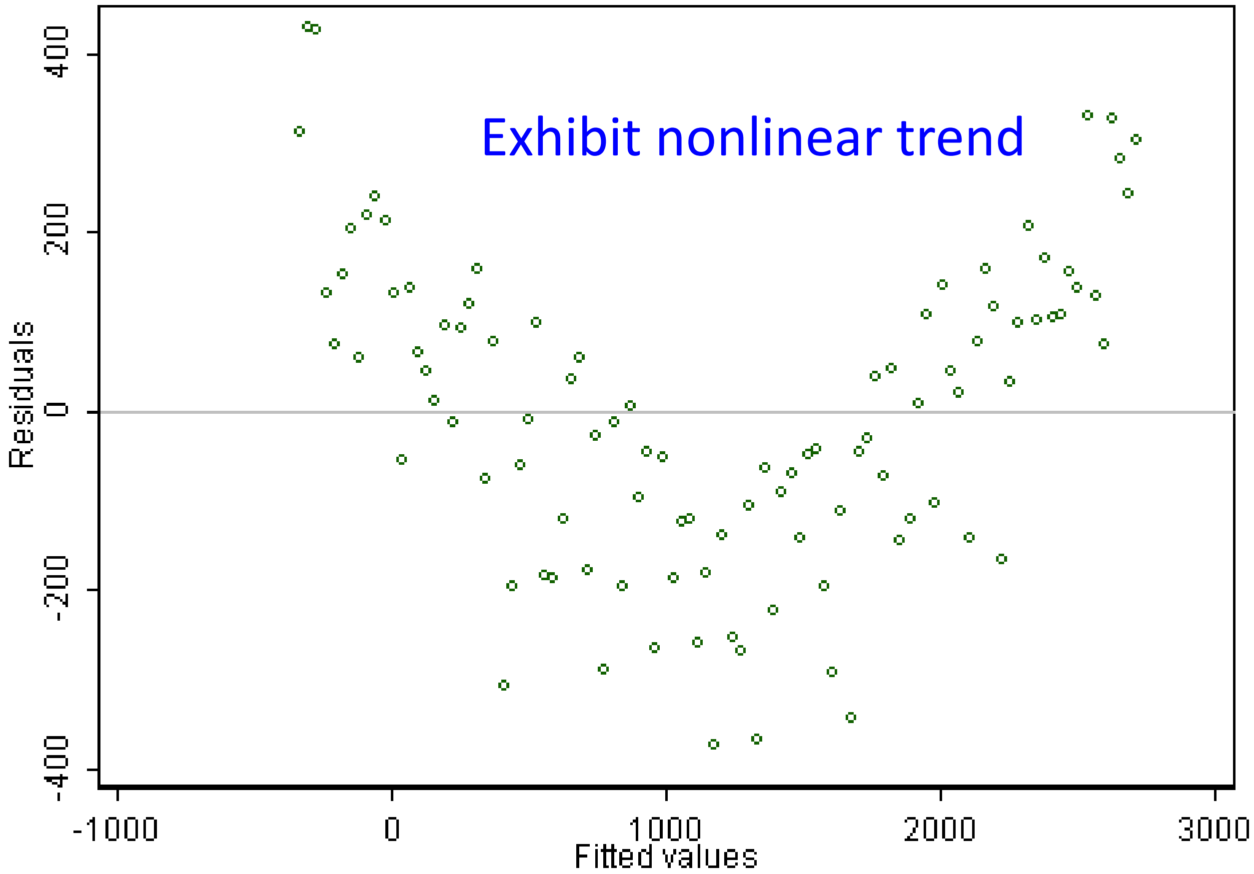
Nonlinear Transformations

- We can transform the dependent and/or the independent variables
- Common nonlinear transformations
  - Natural logarithm, square root, reciprocal, square
- When to use Transformations?
  - Visualize the relationships between variables
    - Scatter plots: Does the relationship look linear?
    - Fitted values vs. residuals: Is there a pattern? If modeled appropriately, residuals should randomly vary around zero.
  - Use domain knowledge

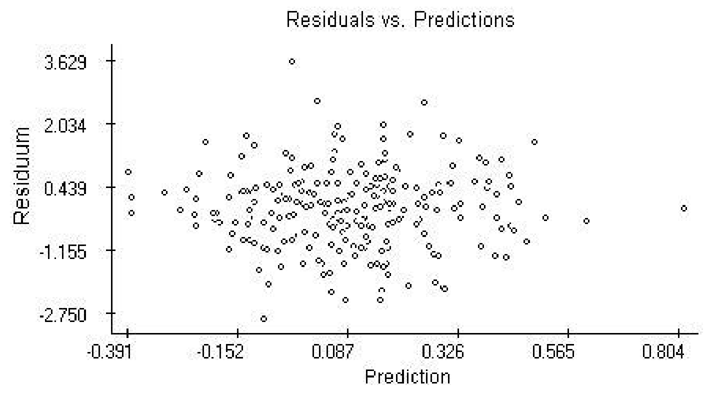
Detecting Nonlinear Relationships: Scatter Plots



Detecting Nonlinear Relationships: Residuals



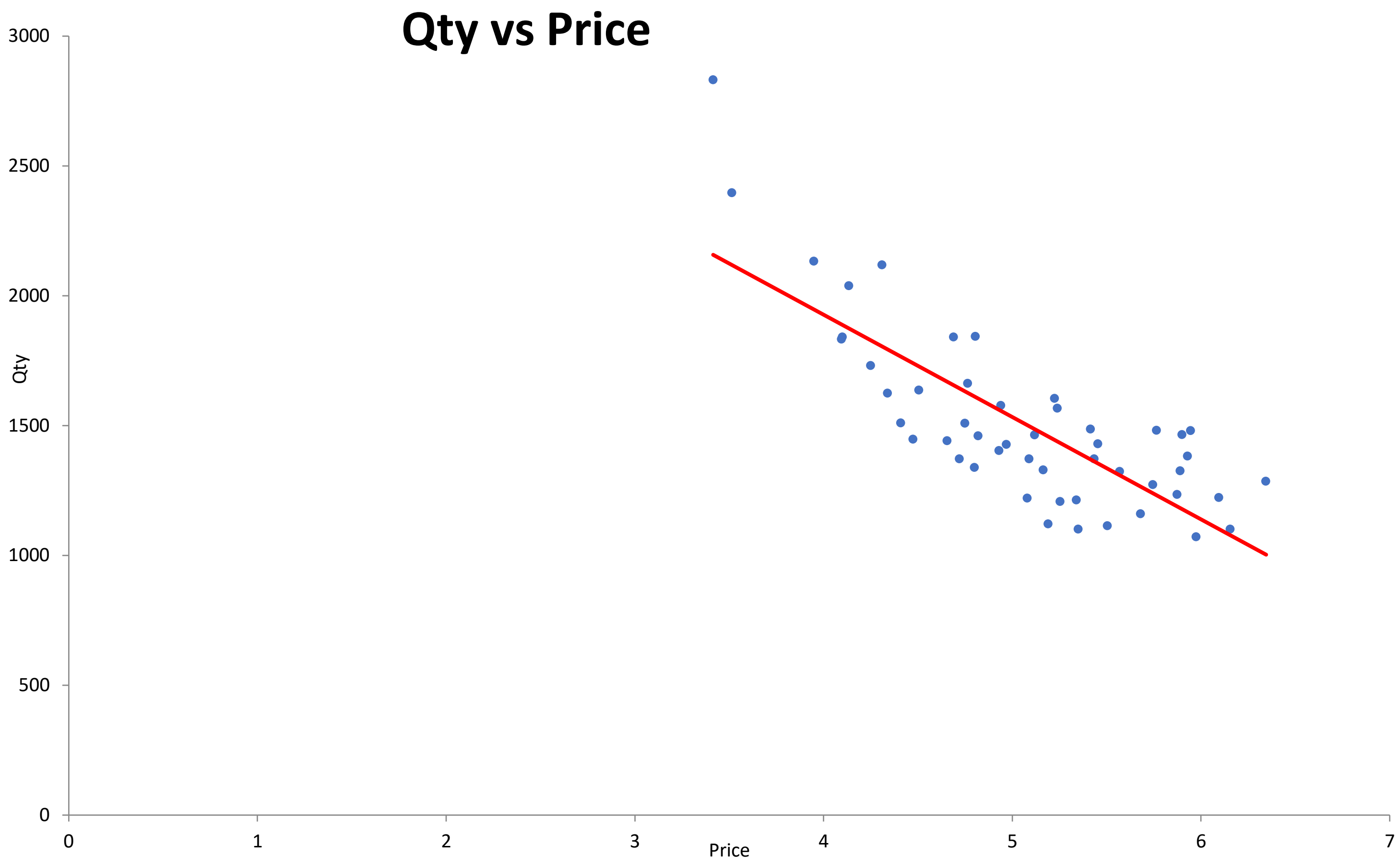
Residual plot without trend



Example: Detergent Sales

- A brand manager at a consumer goods firm is studying the sales of the firm's flagship brand of laundry detergent, Mr. Clean
- Weekly data over a 50-week period are obtained from a particular sales district, including the prevailing retail price for a 5-lb. box of Mr. Clean for that week and the number of boxes sold
- Goal is to construct a regression model that explains and predicts the demand for Mr. Clean as a function of its price
  - Explore linear, quadratic, logarithm, exponential, and log-log models

Mr. Clean Data – DetergentSales.xlsx



Model 1: Simple Linear Regression

```
Call:
lm(formula = Qty ~ Price)

Residuals:
    Min      1Q  Median      3Q     Max 
-337.03 -153.10   -5.54  156.54  674.36 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3503.6      225.3  15.553  < 2e-16 ***
Price         -394.1       44.1  -8.937 8.78e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 214 on 48 degrees of freedom
Multiple R-squared:  0.6246,	Adjusted R-squared:  0.6168 
F-statistic: 79.87 on 1 and 48 DF,  p-value: 8.779e-12
```

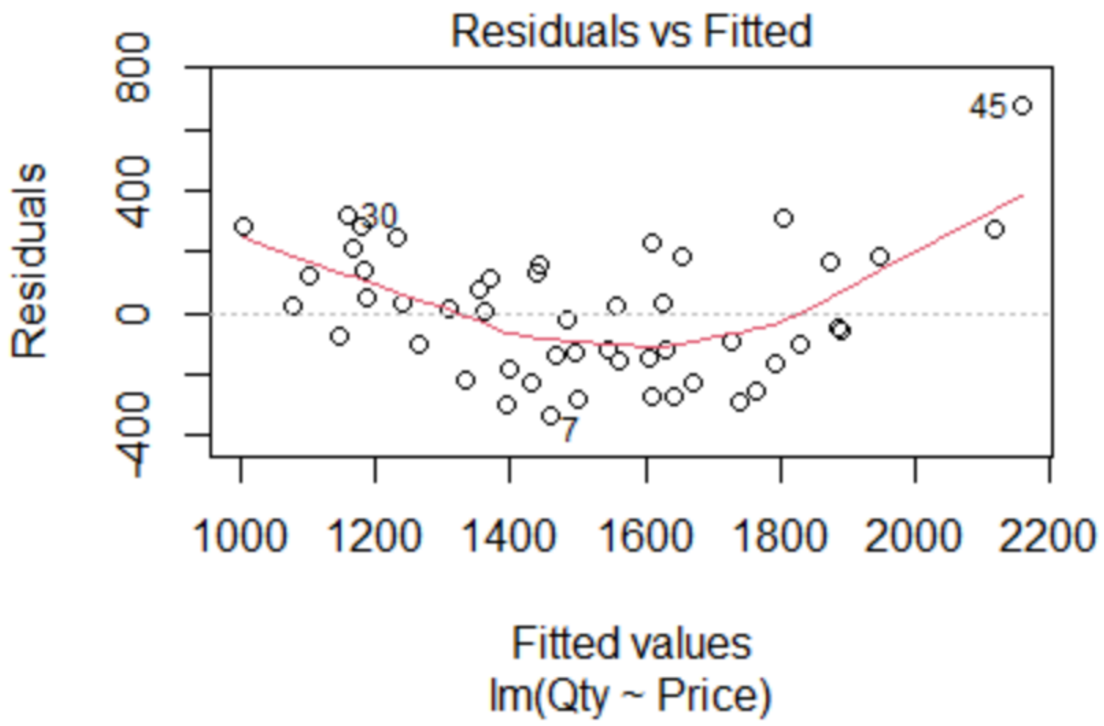
Qty = 3503.6 - 394.1 Price

b = - 394.1 implies that when price increases by $1, demand decreases, on average, by 394 boxes.

Residual Plots

```r
# fitted value vs residual plot
plot(resid(Model1)~fitted(Model1))
abline(h=0)
qqnorm(resid(Model1))
qqline(resid(Model1), col="red")
hist(resid(Model1))

# You can also generate a residual plot using plot()
plot(Model1)
```

Mr. Clean Data: Simple Linear Regression

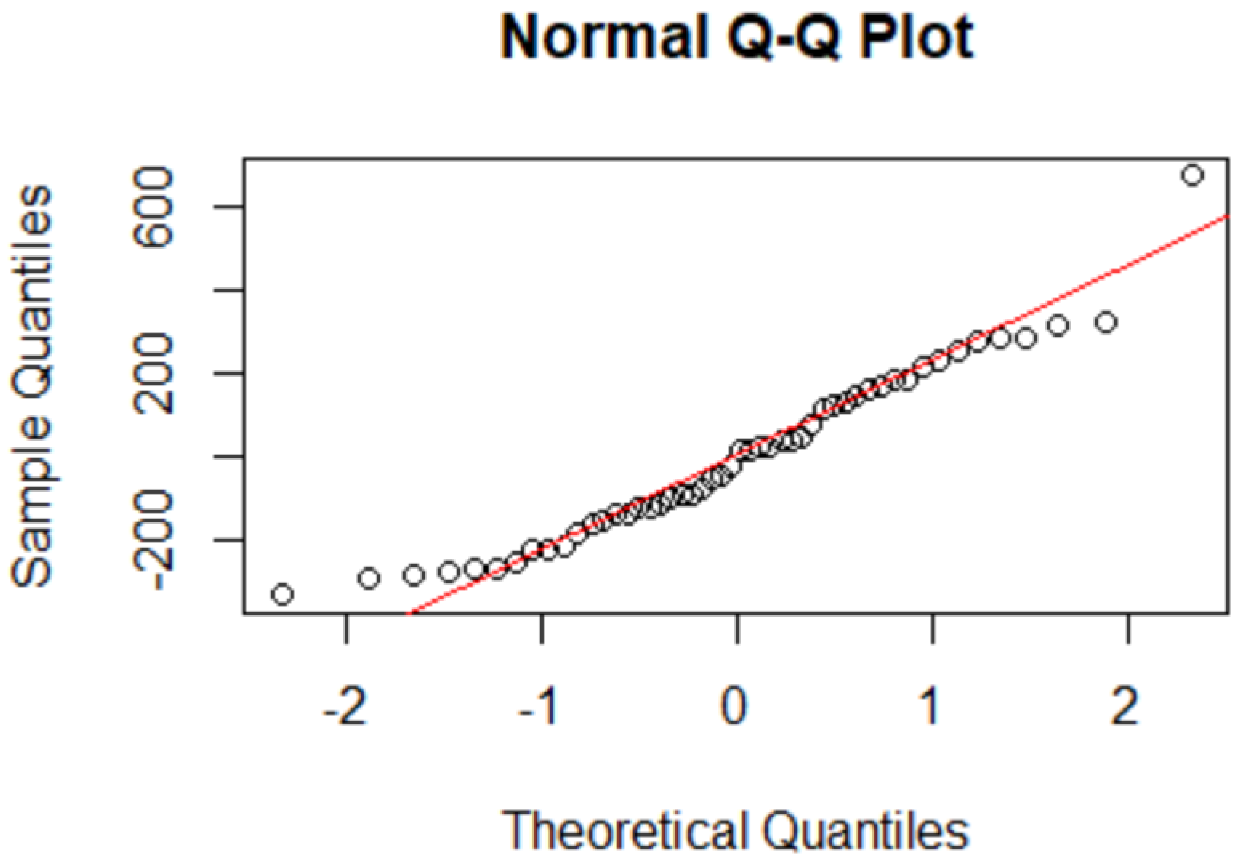


Non-linear pattern in the residuals: a parabolic shape

Residual Analysis: Normality

- You can check the normality by forming a histogram or a Q-Q plot of the residuals.
  - The histogram should be approximately symmetric and bell-shaped, and the points of a Q-Q plot should be close to a 45 degree line.
  - If there is an obvious skewness or some other nonnormal property, this indicates a violation of the normality assumption.

Mr. Clean Data: Simple Linear Regression



Model 2: Quadratic Model

- The quadratic model has the form:

$$Y = a + b_1X + b_2X^2$$

- For Mr. Clean the regression formula is:

$$Qty = a + b_1\,Price + b_2(Price)_2$$

- Interpretation can be tricky - do not have an easy interpretation.
- What happens to Y if we increase X by one unit?
- It depends on the value of X:

$$\frac{dY}{dX} = b_1 + 2b_2X$$

Mr. Clean Data: Quadratic Model

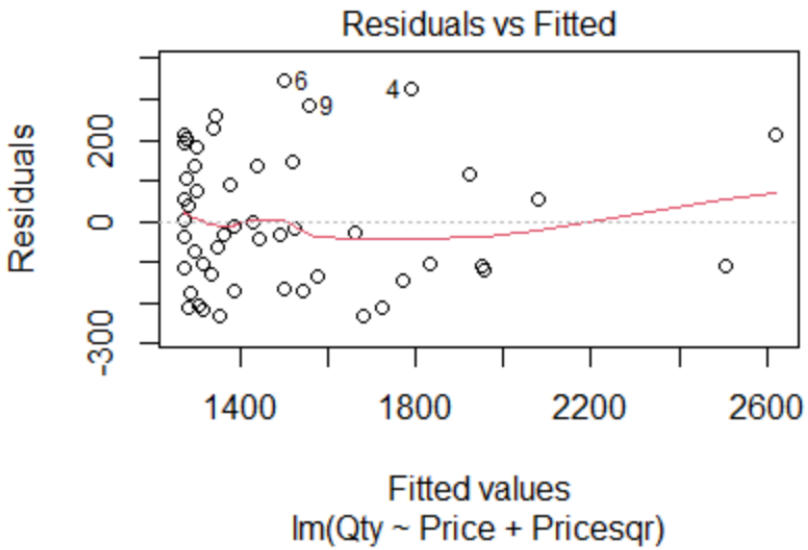
```
Pricesqr<-(Price)^2
Model2<- lm(Qty~Price+Pricesqr)
summary(Model2)
```

```
Call:
lm(formula = Qty ~ Price + Pricesqr)

Residuals:
    Min      1Q  Median      3Q     Max 
-233.46 -127.23  -28.39  129.43  343.45 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9331.81    1011.54   9.225 4.04e-12 ***
Price       -2790.24     411.13  -6.787 1.72e-08 ***
Pricesqr      241.46      41.29   5.848 4.57e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 164.6 on 47 degrees of freedom
Multiple R-squared:  0.7827,	Adjusted R-squared:  0.7735 
F-statistic: 84.66 on 2 and 47 DF,  p-value: 2.629e-16
```

$$Qty = 9331.81 - 2790.24 Price + 241.46\,(Price)^2$$

$-2790.24 + 2 * 241.46\, Price$ is the rate of change of demand with respect to Price

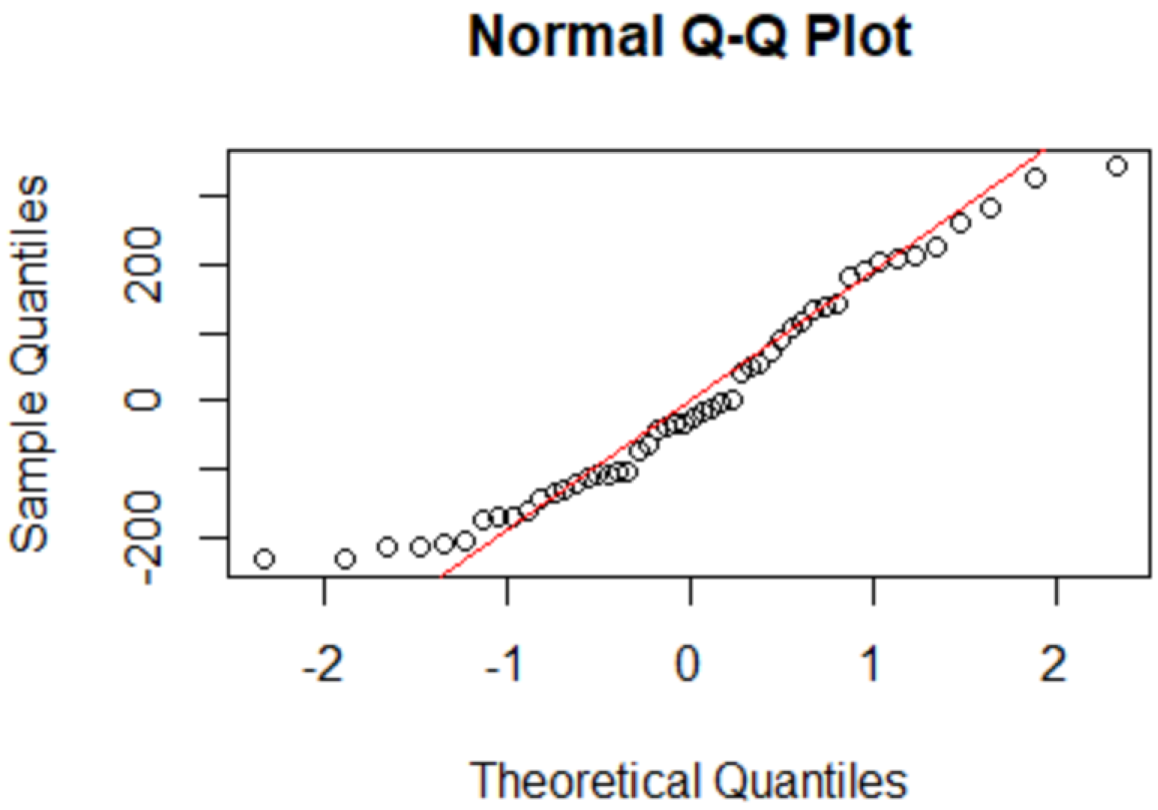
Mr. Clean Data: Quadratic Model

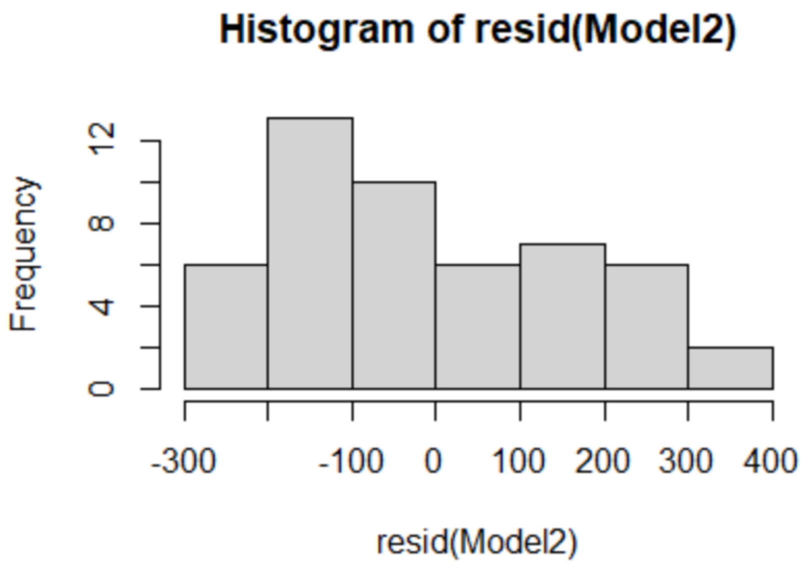


The residuals are skewed to the right

Remarks: Because the quadratic model does not allow for an isolated **interpretation** of the individual predictors, it is in practice often less preferred. This is especially true if the focus is on learning **new insight** about the relationship between *X* and *Y*. If on the other hand, if the goal is purely improved **prediction**, then a quadratic model could be a good choice.

Mr. Clean Data: Quadratic Model





Model 3: Log Model

- Y = *a* + *b* * Log(X)
- More naturally interpretable than quadratic models

$$dY = b\frac{dX}{X}$$

  - dX (infinitesimal change in X) $\approx \Delta X, dY \approx \Delta \mathrm{Y}$
  - The quantity $(\Delta X)/X$ represents a small proportional increase in *X*. Therefore $100 \cdot (\Delta X)/X$ is a small percentage change in *X*.
  - *(b (ΔX)/X)* is the change in Y when X increases by a small proportional amount.
  - On average, Y in increases **approximately** by ***b/100,*** when X increases by **1%** .

  Note: $dY = b\frac{dX}{X} \approx \mathrm{b} * 1\% = \mathrm{b} * 0.01 = \mathrm{b}/100$

Mr. Clean Data: Log Model

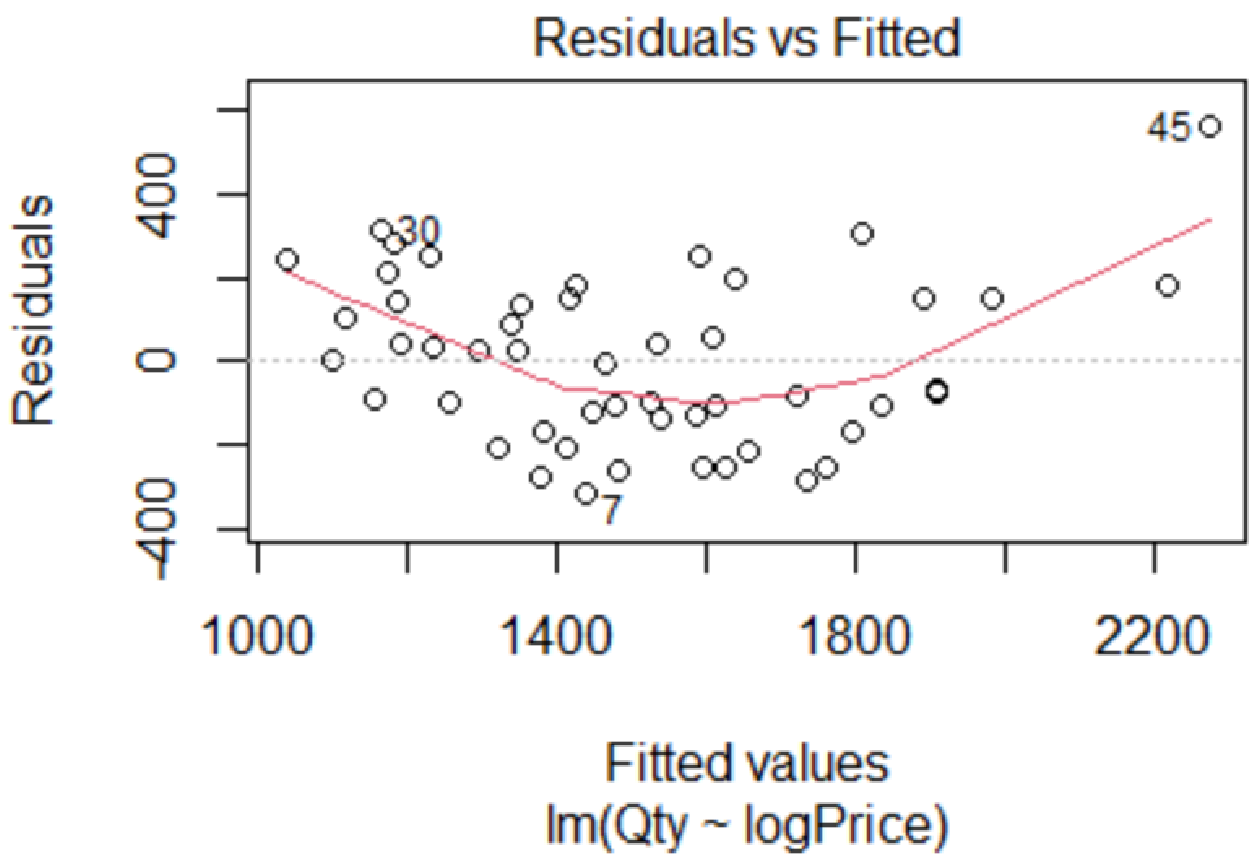
```
logprice<-log(Price)
Model3<- lm(Qty~logprice)
summary(Model3)
```

```
Call:
lm(formula = Qty ~ logPrice)

Residuals:
    Min      1Q  Median      3Q     Max 
-318.36 -134.88   -0.33  146.66  557.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4724.1      321.7   14.68  < 2e-16 ***
logPrice     -1994.7      198.8  -10.03 2.28e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 198.5 on 48 degrees of freedom
Multiple R-squared:  0.6771,	Adjusted R-squared:  0.6704 
F-statistic: 100.7 on 1 and 48 DF,  p-value: 2.277e-13
```

Qty = 4724.1 - 1994.7 Log(Price)

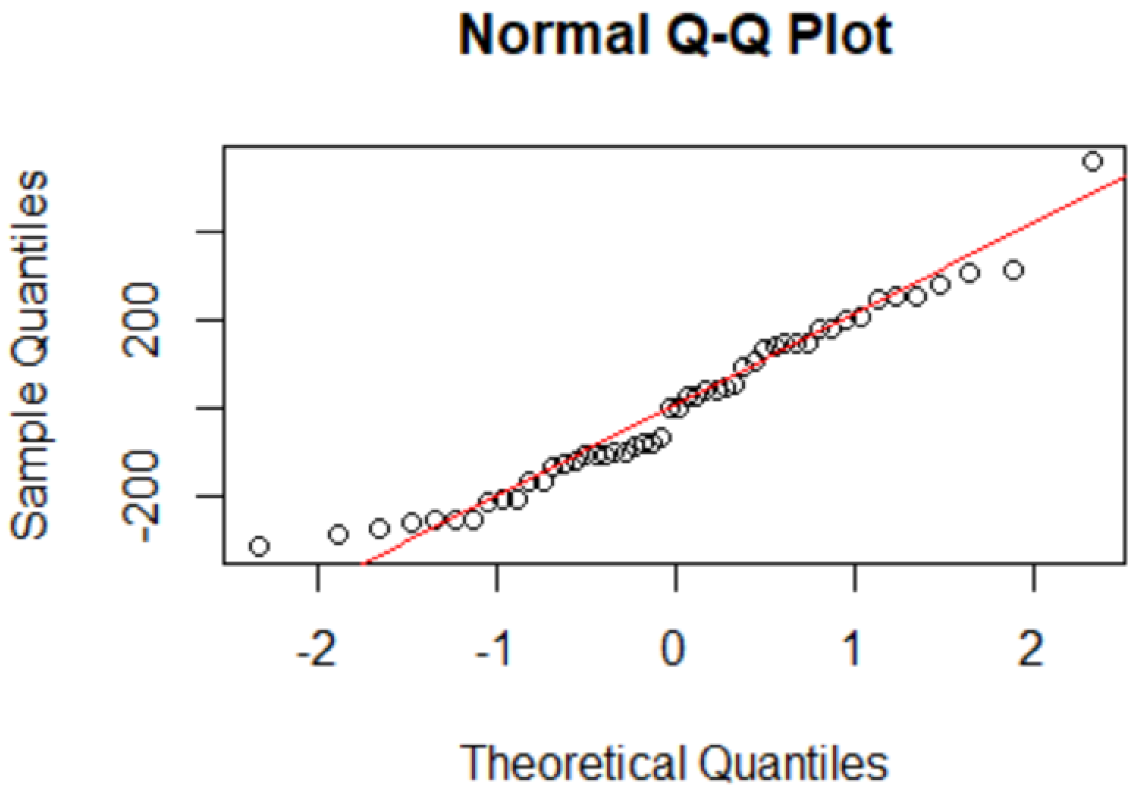
b=-1994 implies that on average the demand decreases approximately by 19 or 20 boxes, when price increases by 1 %.

Mr. Clean Data: Log Model



Much Better!

Mr. Clean Data: Log Model



Model 4: Exponential Model

- *Exponential model is :*

$$Y = c * e^{bx} \qquad \text{(multiplicative model)}$$

- This model implies:

$$Log(Y) \ = \ a \ + \ bX \qquad \text{(additive model)}$$

where $a = Log(c)$

- Note that

$$\frac{dY}{Y} = b\ dX$$

When X increases by **one unit**, the expected percentage change in Y is **approximately b * 100%**

  - Note: 100 *$\frac{dY}{Y}$% $= 100 * b\ dX$ % ≈ b*100%

Mr. Clean Data: Exponential Model

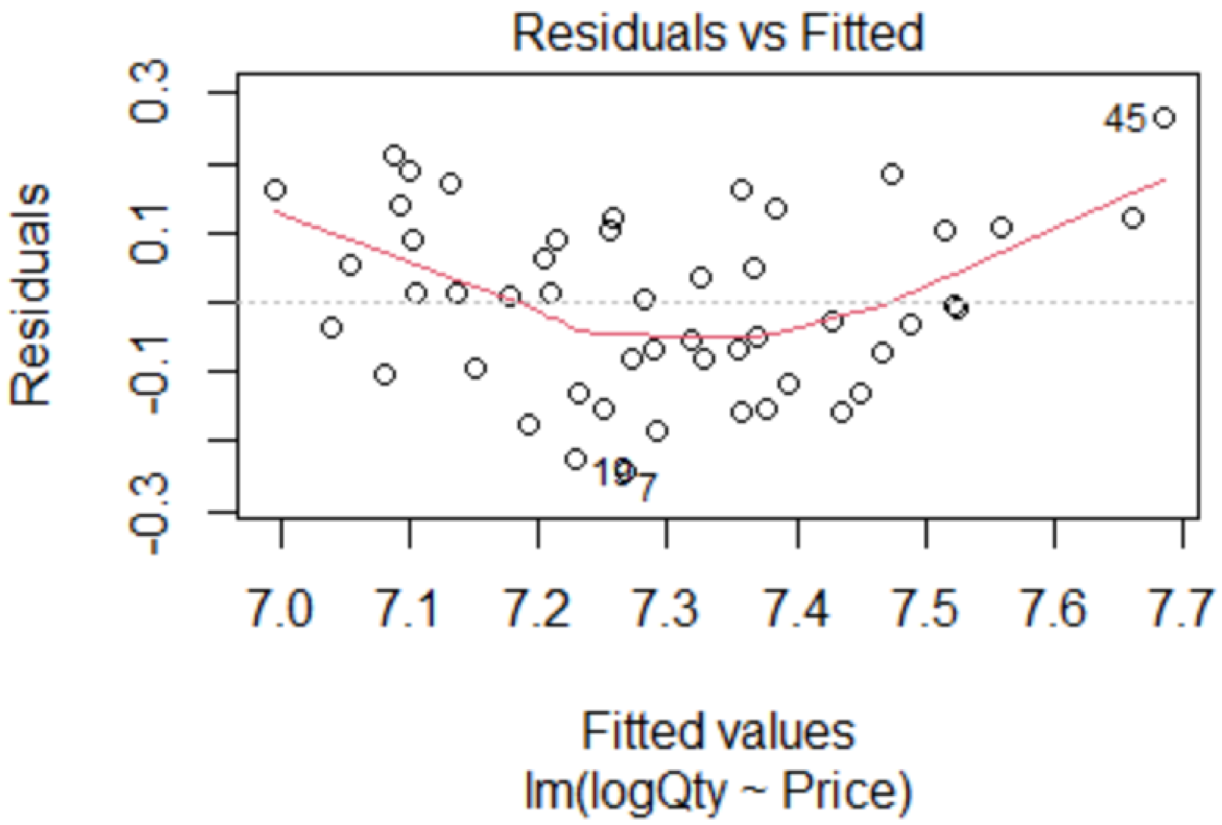
```
logQty<-log(Qty)
Model4<- lm(logQty~Price)
summary(Model4)
```

```
Call:
lm(formula = logQty ~ Price)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.244548 -0.091300  0.000222  0.104257  0.263714 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.49020    0.13343  63.629  < 2e-16 ***
Price       -0.23577    0.02612  -9.026  6.5e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1268 on 48 degrees of freedom
Multiple R-squared:  0.6292,	Adjusted R-squared:  0.6215 
F-statistic: 81.46 on 1 and 48 DF,  p-value: 6.5e-12
```

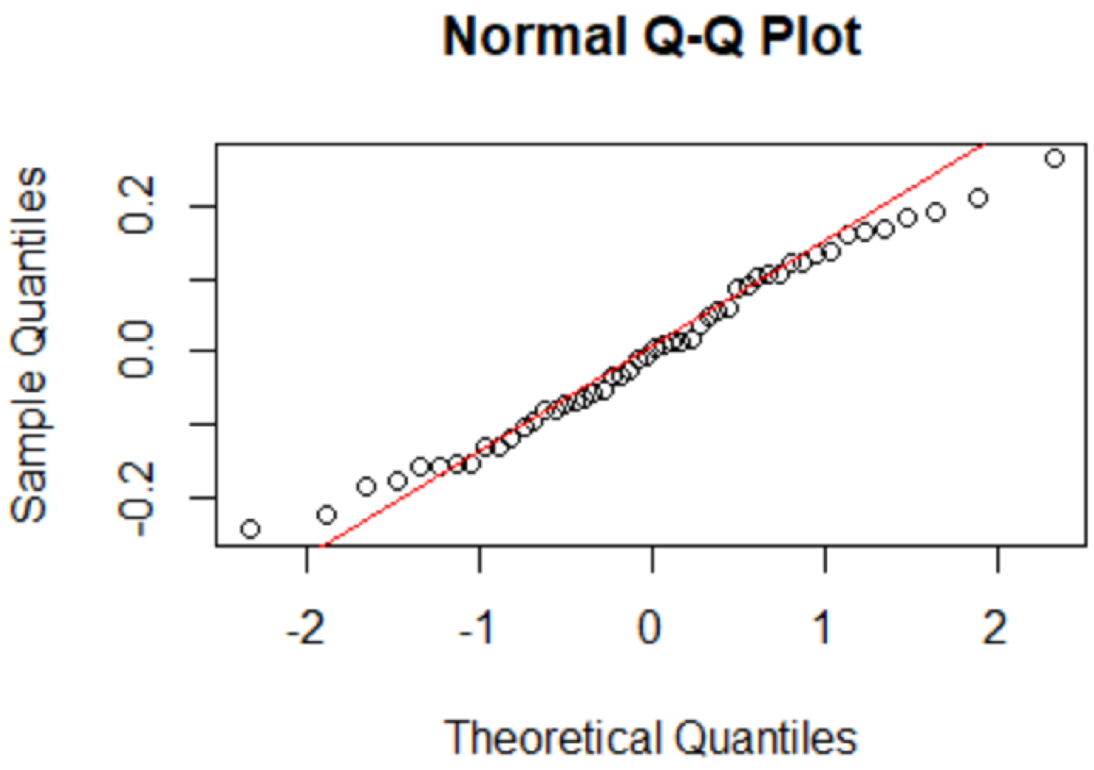
Log(Qty) = 8.49 - 0.23577 Price

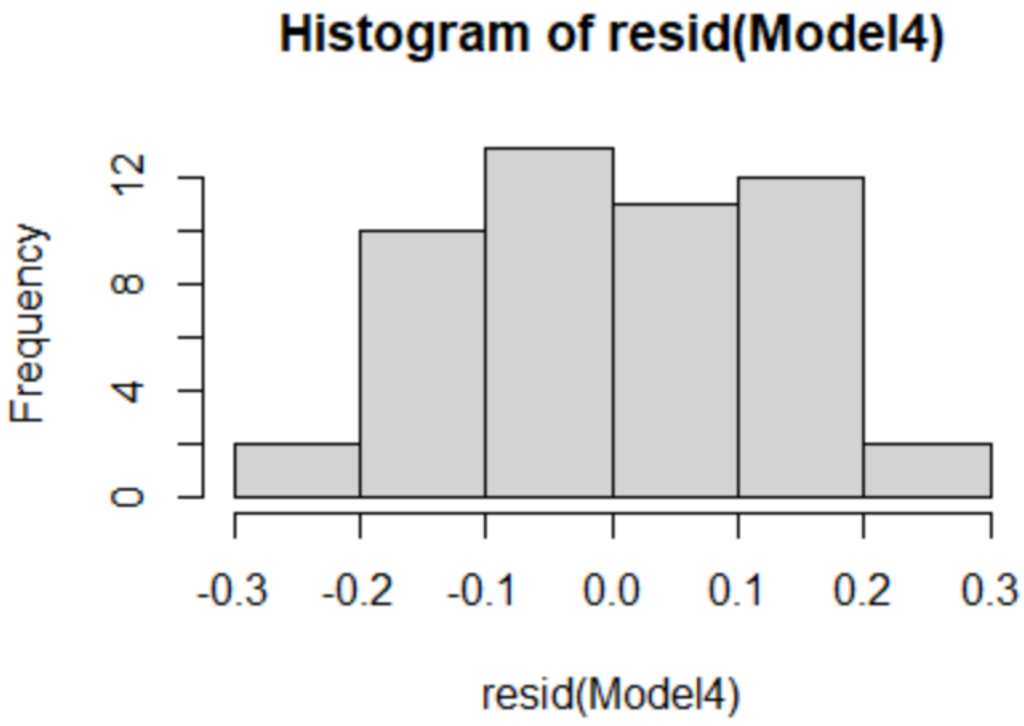
***b* = -0.23577** implies that when price increases by $1, on average demand decreases approximately by 23.58%.

Mr. Clean Data: Exponential Model



Mr. Clean Data: Exponential Model





Model 5: Log-log Model

- The Log Log Model (or Power Model) is applicable when you believe that the price elasticity is constant, that is: when *X* increases by 1% the *Y increases* (on average) by *b*%

$$Log\, Y \;=\; a \;+\; b\, Log\, X$$

- Note that

$$\frac{dY}{Y} = b\,\frac{dX}{X}$$

- *b* is the percentage increase in *Y* when *X* increases by 1%.

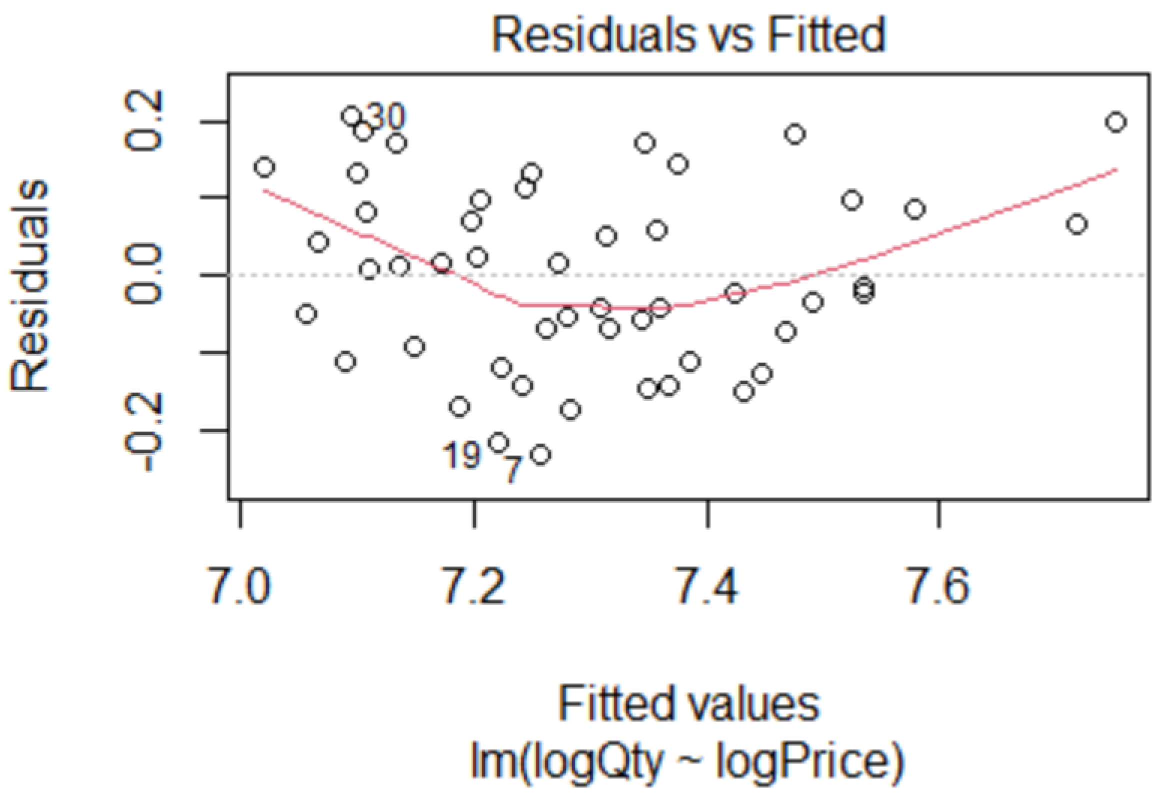
Mr. Clean Data: Log-log Model

```
Call:
lm(formula = logQty ~ logPrice)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.233794 -0.088214 -0.003343  0.093754  0.204465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.2010     0.1943  47.360  < 2e-16 ***
logPrice     -1.1813     0.1201  -9.839  4.3e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1199 on 48 degrees of freedom
Multiple R-squared:  0.6685,	Adjusted R-squared:  0.6616 
F-statistic: 96.81 on 1 and 48 DF,  p-value: 4.296e-13
```

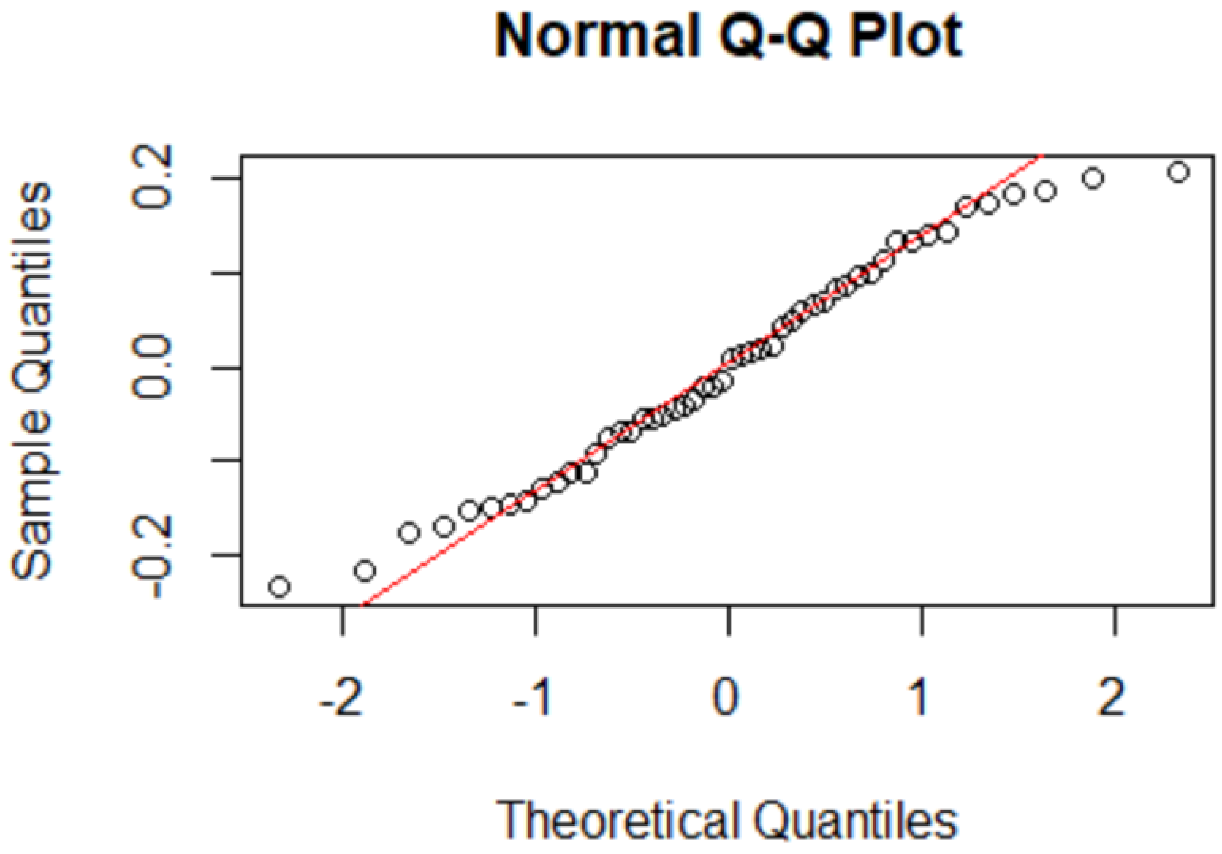
Log(Qty) = 9.2010 - 1.1813 Log(Price)

b = -1.1813 implies that when price increases by 1%, on average demand decreases approximately by *1.18%.*

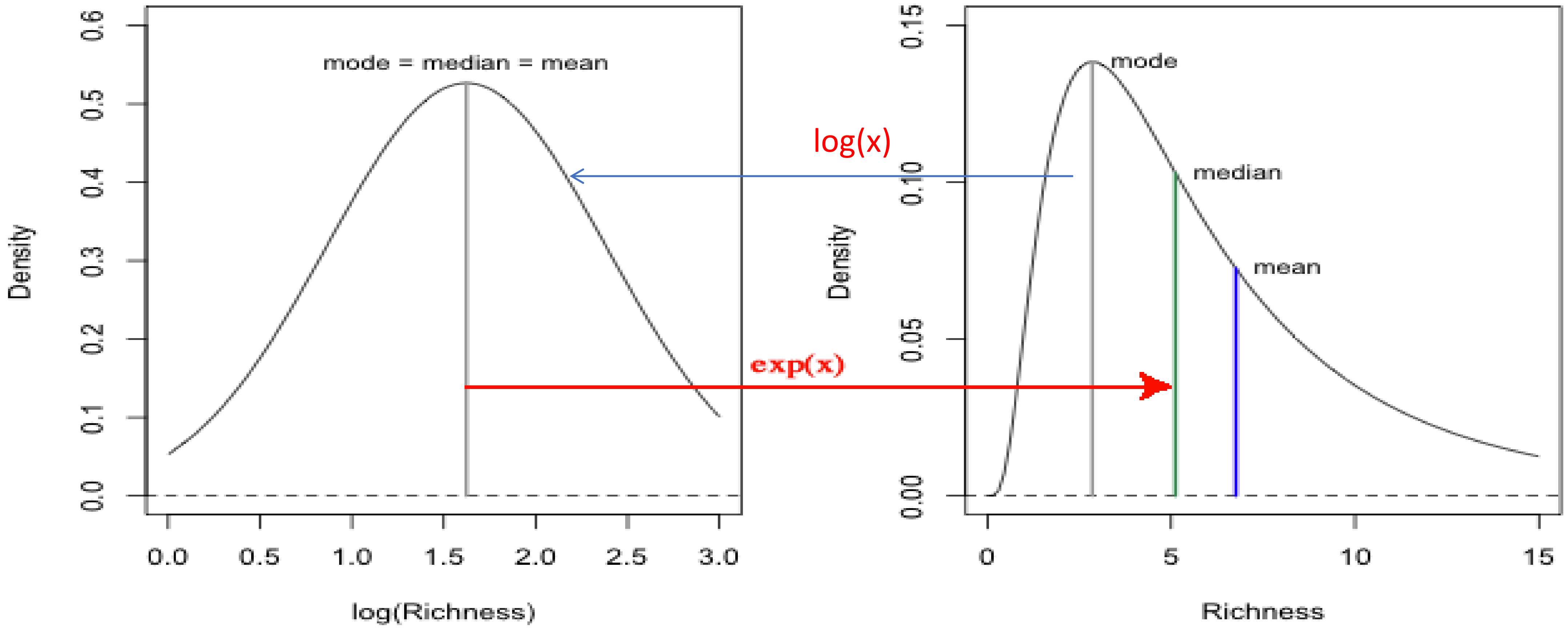
Mr. Clean Data: Log-Log Model



Mr. Clean Data: Log-Log Model

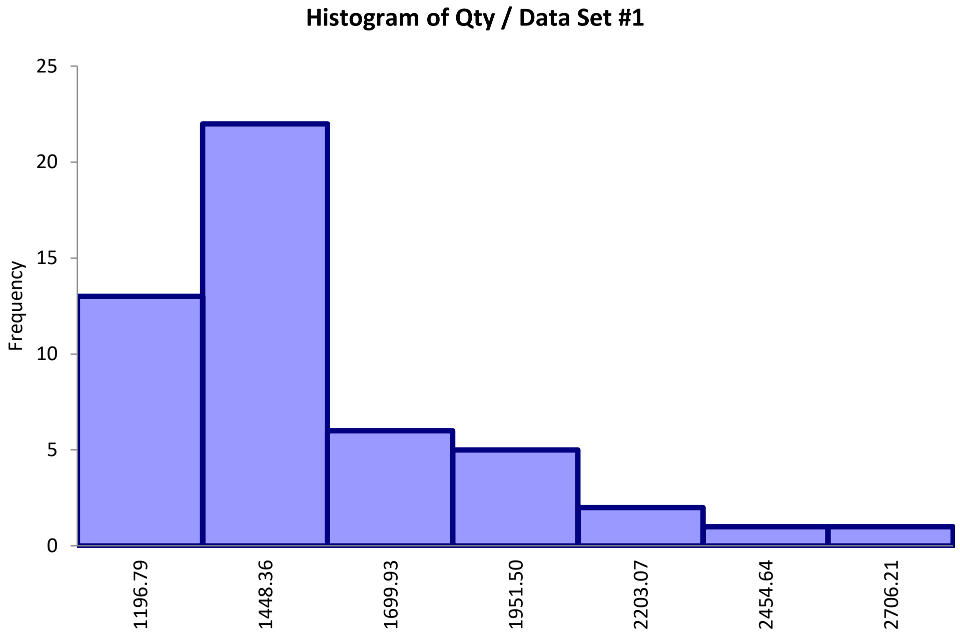


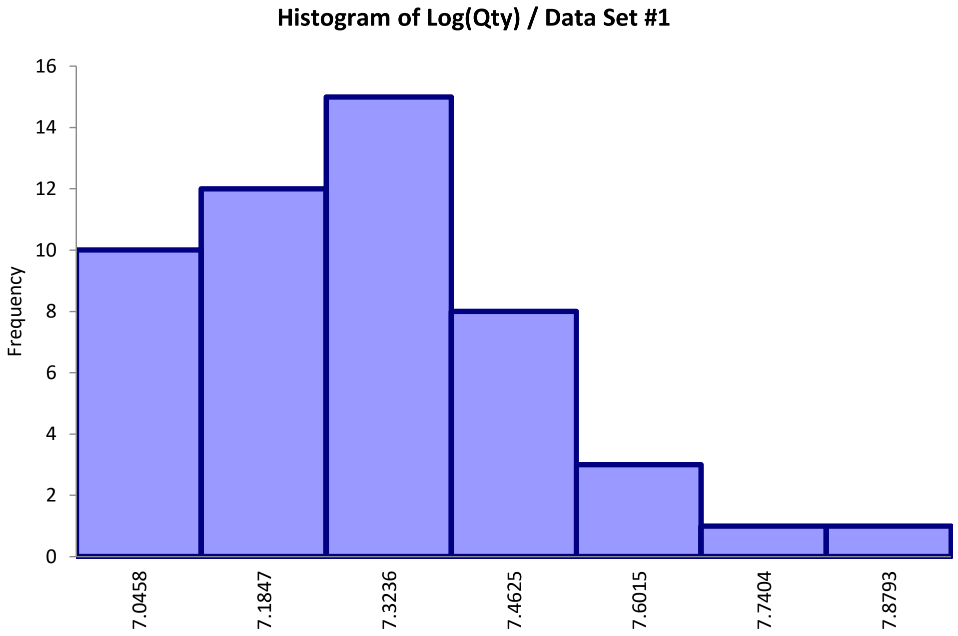
More on Log Transformations



The logarithmic function transforms right-skewed distributions into approximately Normal distributions, which are usually fit better by regression

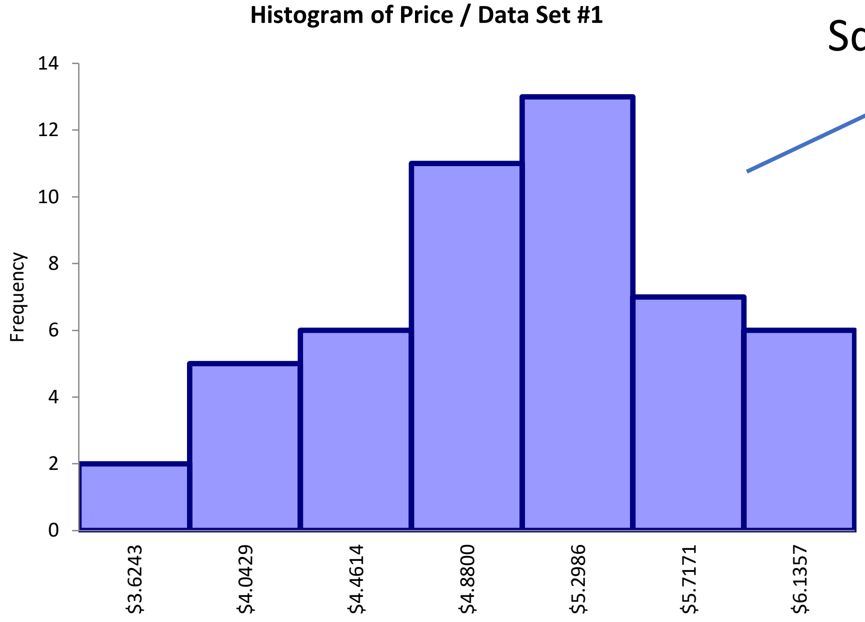
Log Transformation



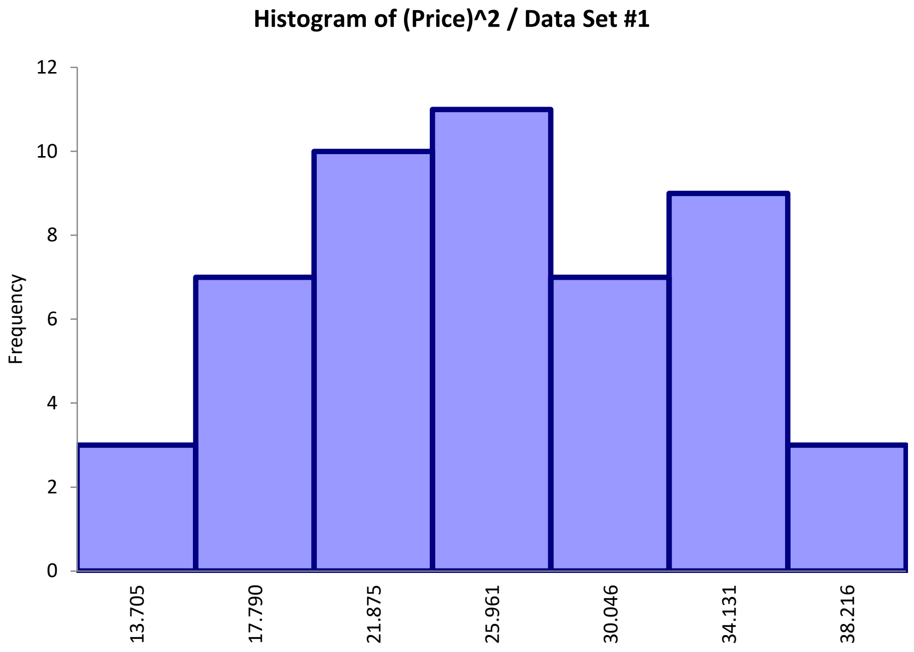


Square transformation

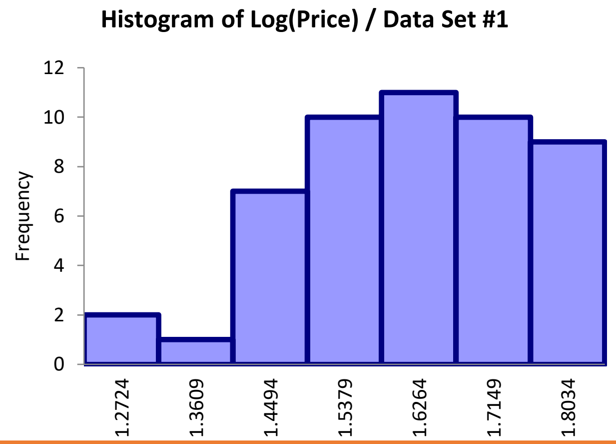
The quadratic model has the highest R^2!



Square



Log



Nonlinear Transformations Summary

Model	Regression Formula	Interpretation of Model Coefficients
Linear	$Y = a + b X$	Increasing X has a constant effect on Y (b)
Quadratic	$Y = a + b_1 X + b_2 X^2$	$b_1 + 2b_2X$ is the rate of change of Y with respect to X
Log	$Y = a + b \text{ Log}(X)$	When X increases by 1%, Y increases (on average) by b / 100
Exponential	$\text{Log}(Y) = a + bX$	When X increases by one unit, the expected percentage change in Y is approximately b * 100%
Log-Log	$\text{Log}(Y) = a + b \text{ Log}(X)$	When X increases by 1%, Y increases (on average) by b%

Practice: Catalog_Marketing_Reg.xlsx

- Build an exponential model: Log(AmountSpent) = Salary + Gender
  - Interpret the coefficient of Salary

- Build a Log-Log model: Log(AmountSpent) = Log(Salary) + Gender
  - Interpret the coefficient of Salary

Next Time…

- Model Validation
- Variable Selection
- Predictive Model