CHAPTER 1

Simple Linear Regression

1.2 Conditions for a Simple Linear Model

1.3 Assessing Conditions

1.4 Transformations

1.5 Outliers and Influential Points

1.2 Conditions for a Simple Linear Regression

There are four conditions (LINE) that comprise the simple linear regression model, which are:

1. **L**inearity: The overall relationship between the response and the other variables has a linear pattern.
2. **I**ndependence: The errors (residuals) are assumed to be independent from one another.
3. **N**ormality: The errors (residuals) follow a normal distribution.
4. **E**qual Variances (heteroscedasticity): The variability in the errors (residuals) is the same for all values of the predictor variables.

The four conditions of the model pretty much tell us what can go wrong with our model, namely:

1. The population regression function is not linear.
2. The error terms are not independent.
3. The error terms are not normally distributed.
4. The error terms do not have equal variance.

We need to identify the following two problems:

- Are there any "outliers"?

1.3 Assessing Conditions (Assumptions)

Two plots are used to assess the conditions.

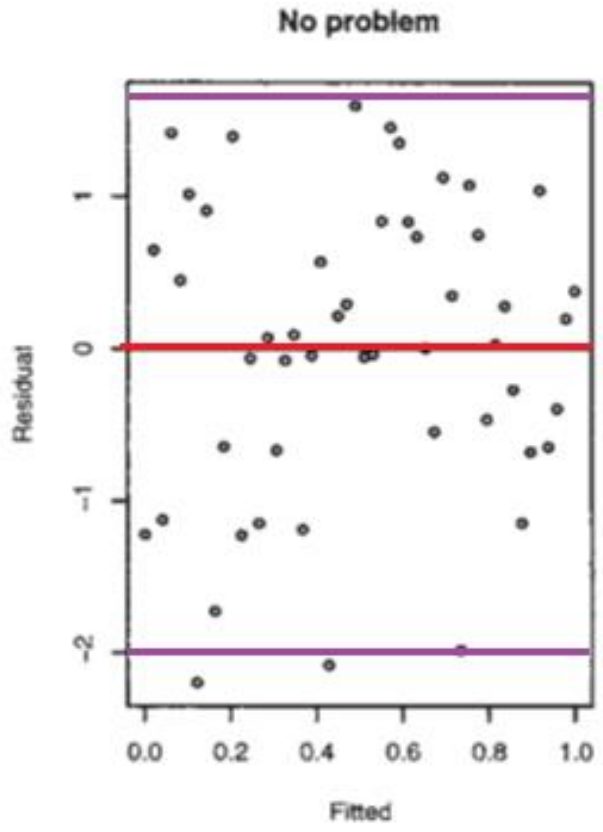
1. Residual versus Fits Plot: A scatterplot of residuals on the y-axis and fitted (predicted) values on the x-axis.

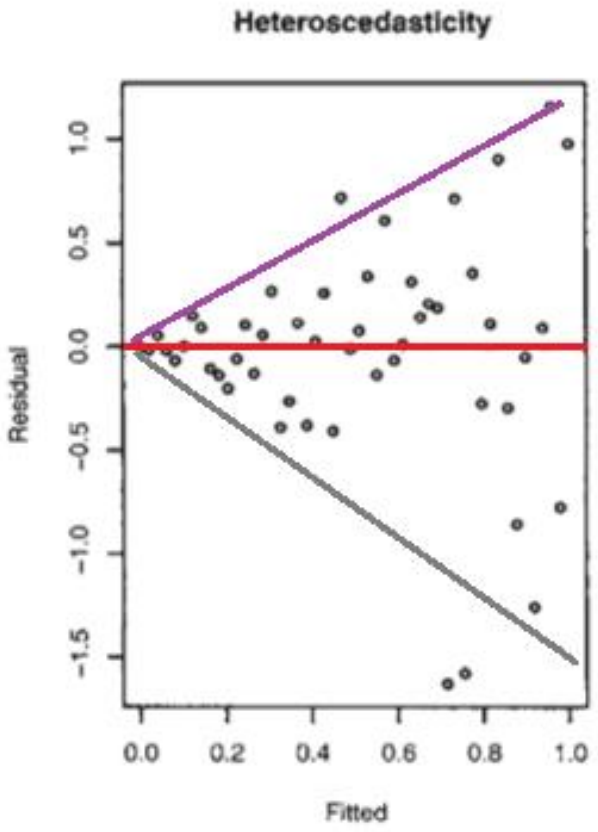
- The plot is used to detect non-linearity, unequal error variances, and outliers.

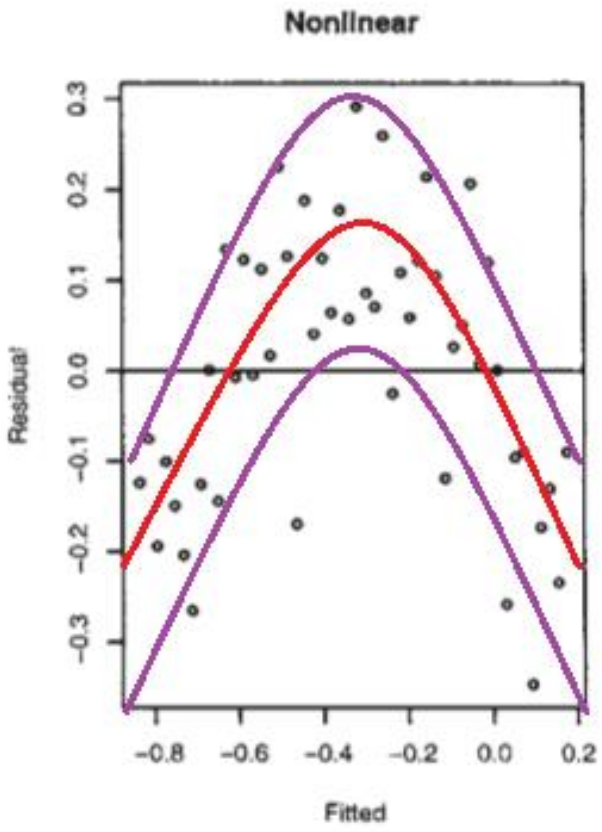
2. Normal Plot (QQ-Plot): is assessing whether or not a data set is approximately normally distributed.

- Histogram of the residuals can be used to check normality assumption.

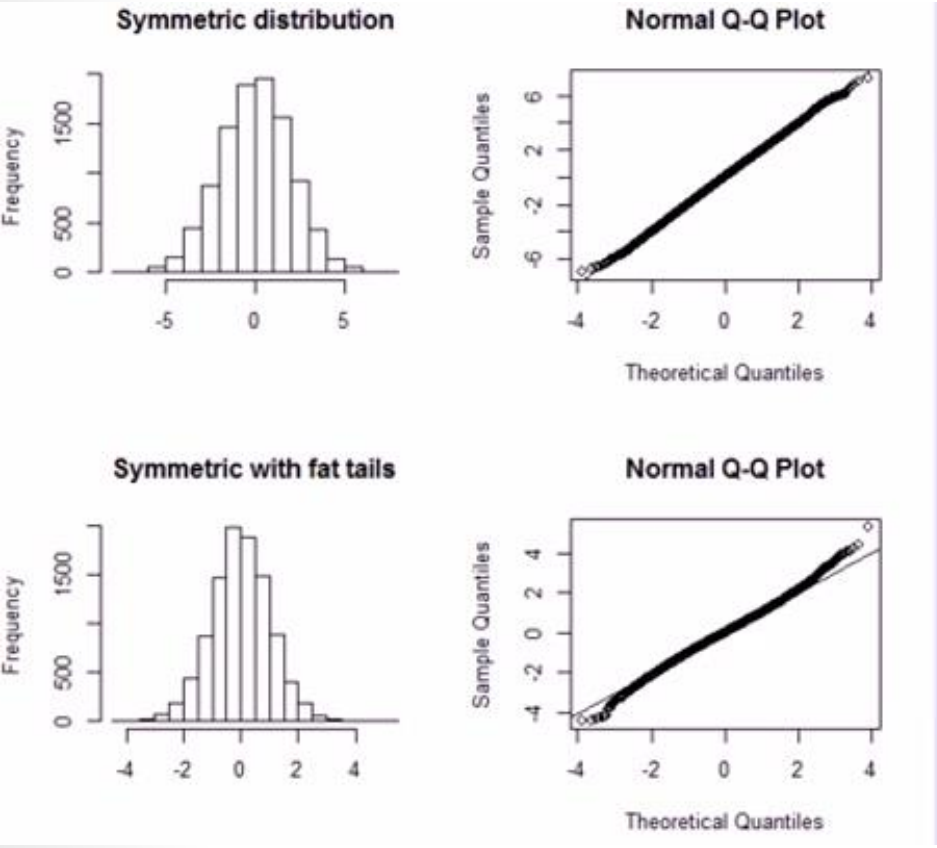
Residual versus Fits Plot

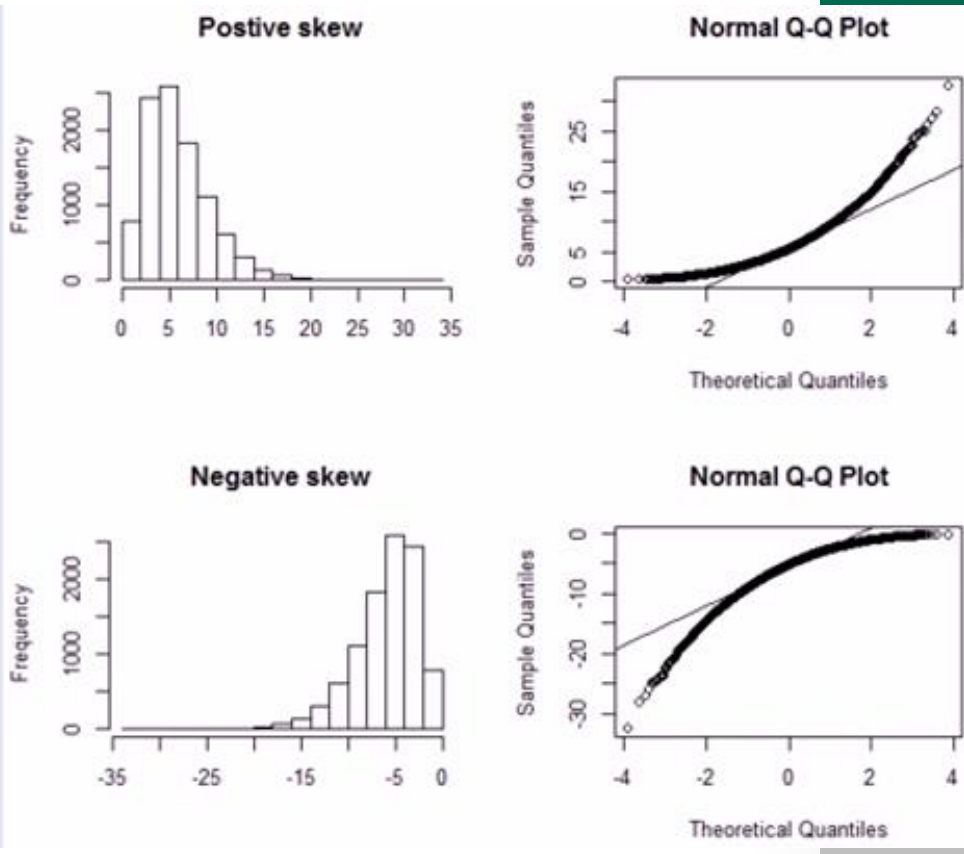






Normal Plot


Symmetric distribution
Frequency
0
500
1500
-5
0
5
Normal Q-Q Plot
Sample Quantiles
-6
-2
2
6
-4
-2
0
2
4
Theoretical Quantiles


Postive skew
Frequency
0
1000
2000
0
5
10
15
20
25
30
35
Normal Q-Q Plot
Sample Quantiles
0
5
15
25
-4
-2
0
2
4
Theoretical Quantiles

Symmetric with fat tails
Frequency
0
500
1500
-4
-2
0
2
4
Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
-4
-2
0
2
4
Theoretical Quantiles

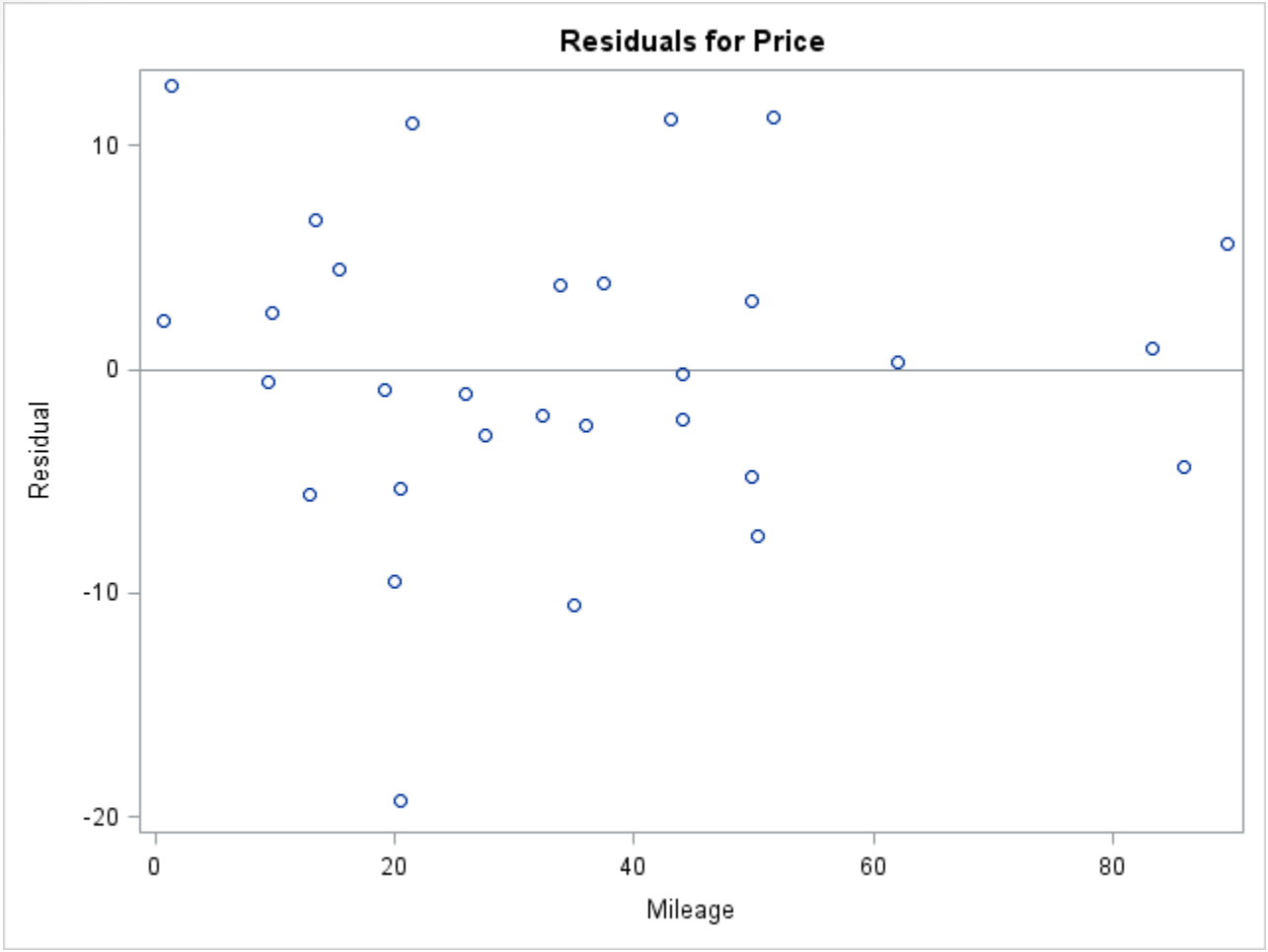
Negative skew
Frequency
0
1000
2000
-35
-25
-15
-5
0
Normal Q-Q Plot
Sample Quantiles
-30
-20
-10
0
-4
-2
0
2
4
Theoretical Quantiles

Example 4: *(Porsche prices)*

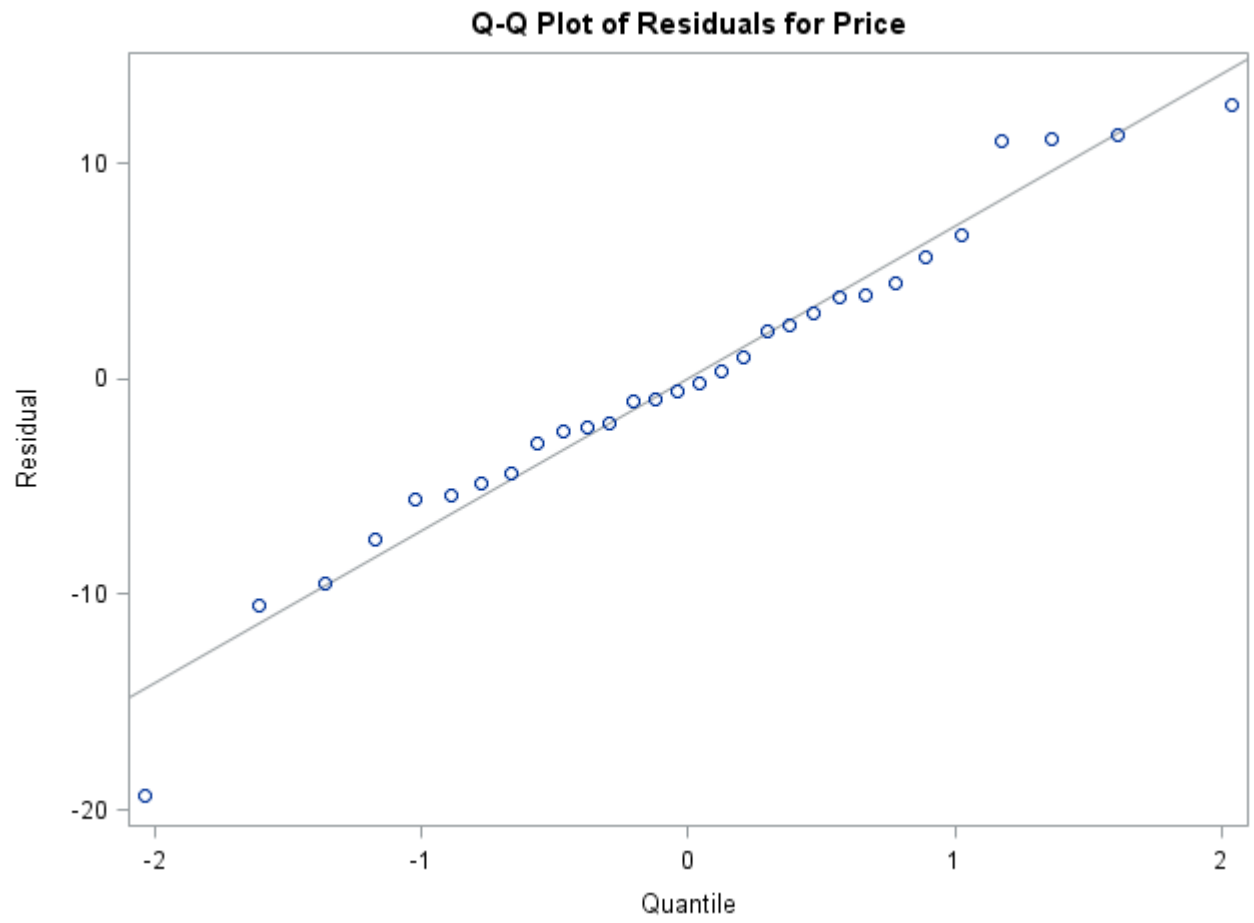
Check the conditions for model to predict Porsche prices based on mileage.

1. **Linearity:** The figure shows that the linearity condition is reasonable. (NO pattern)

2. **Equal (Constant) Variance:** The figure supports the constant variance condition. (No cone shape)



3. Independence: the independence assumption is satisfy. (No pattern according to the observation order)



4. Normality: The qq-plot shows a consistent linear trend that supports the normality condition.

1.4 Transformations:

We have learned about the regression assumptions.

Now, What to do when assumptions are not met?

If one or more of the regression assumptions are not satisfied, then we can consider transformations on one or both of the variables.

- Transform the predictor (x) values only.
- Transform the response (y) values only.
- Transform both the predictor (x) values and response (y) values.

Some Transformations' formulas:

Equation	Name
$1/y^2$	inverse square
$1/y$	reciprocal
$1/\sqrt{y}$	inverse square root
$\ln(y)$	natural log
$\sqrt{y}$	square root
y	none
y^2	square

Assumption 1: Linearity

How to detect a problem?

Plot the scatterplot of x versus y, If we didn't see a linear relationship, then there is a problem with linearity assumption.

Or plot residuals versus fitted values.

If we see any pattern, there is a problem with linearity assumption.

What to do about the problem:

Transform the X values, $X' = f(X)$. Then do the regression using X' instead of X:

$$Y = \beta_0 + \beta_1 . X' + \in$$

Note:

Only use this “solution” if non-linearity is the only problem, not if it also looks like there is non-constant variance or non-normal errors. For these, we will transform Y.

Reason:

The errors are in the vertical direction. Stretching or shrinking the X-axis doesn’t change those, so if they are normal with constant variance, they will stay that way.

Assumption 2: Constant (Equal) variance

How to detect a problem?

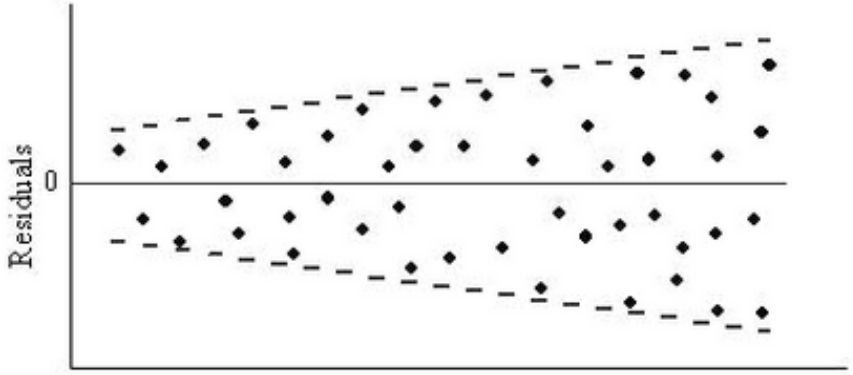
Plot residuals versus fitted values.

If we see increasing or decreasing spread (cone shape), there is a problem with the assumption.

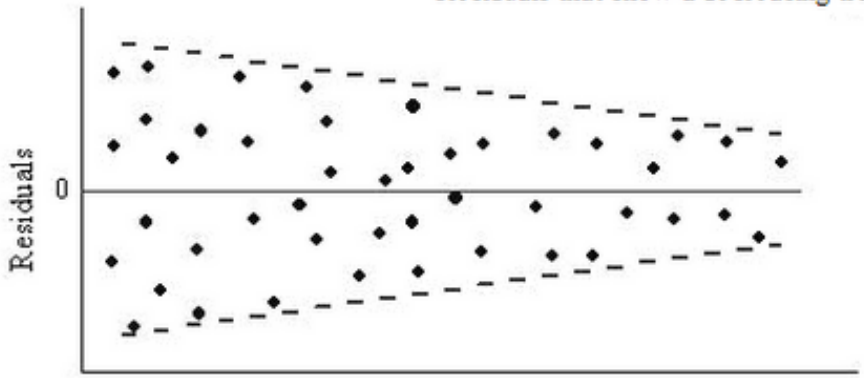
What to do about the problem:

Transform the Y values, or both the X and Y values.

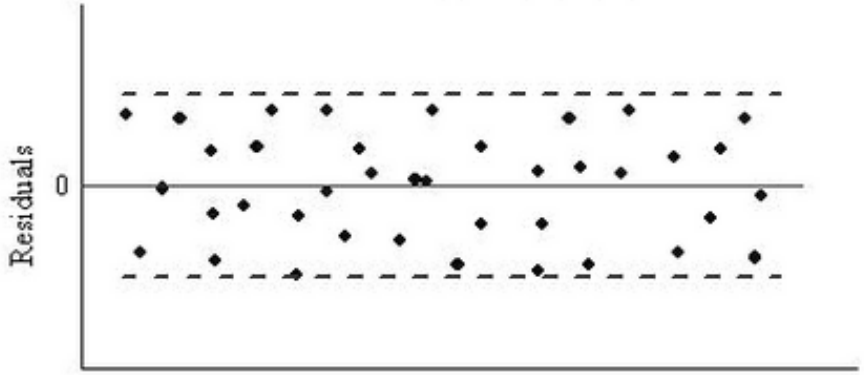
Residuals that show an increasing trend



Residuals that show a decreasing trend



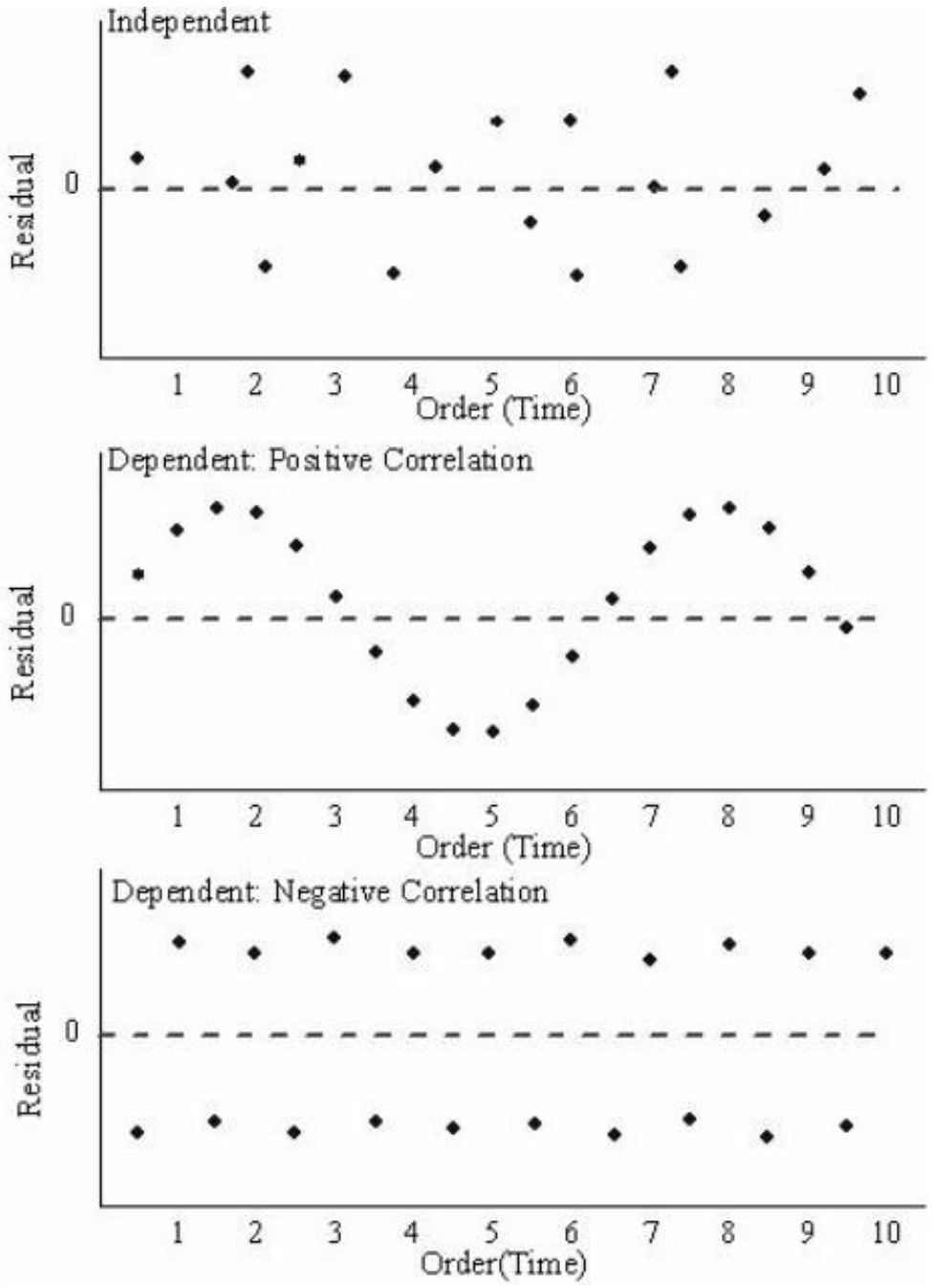
Constant variance



Assumption 3: Independence

1. The main way to check this is to understand how the data were collected. Taking a **random sample** is one way to make sure the observations are independent.

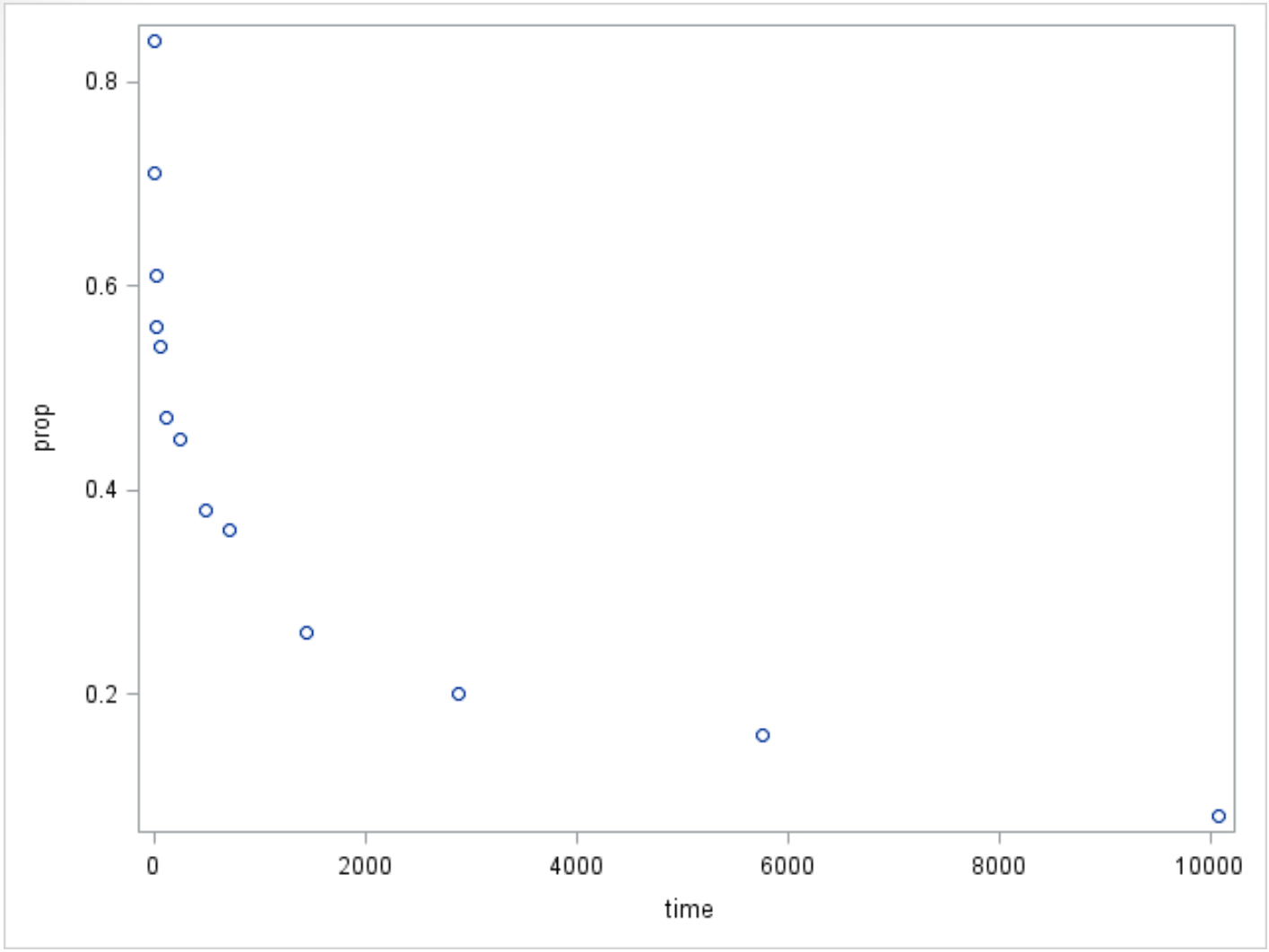
2. If the values were collected over time (or space) it makes sense to plot the residuals versus order collected, and see if there is a trend or cycle. **See page 109 for examples.**


Independent
Residual
0
1 2 3 4 5 6 7 8 9 10
Order (Time)
Dependent: Positive Correlation
Residual
0
1 2 3 4 5 6 7 8 9 10
Order (Time)
Dependent: Negative Correlation
Residual
0
1 2 3 4 5 6 7 8 9 10
Order(Time)

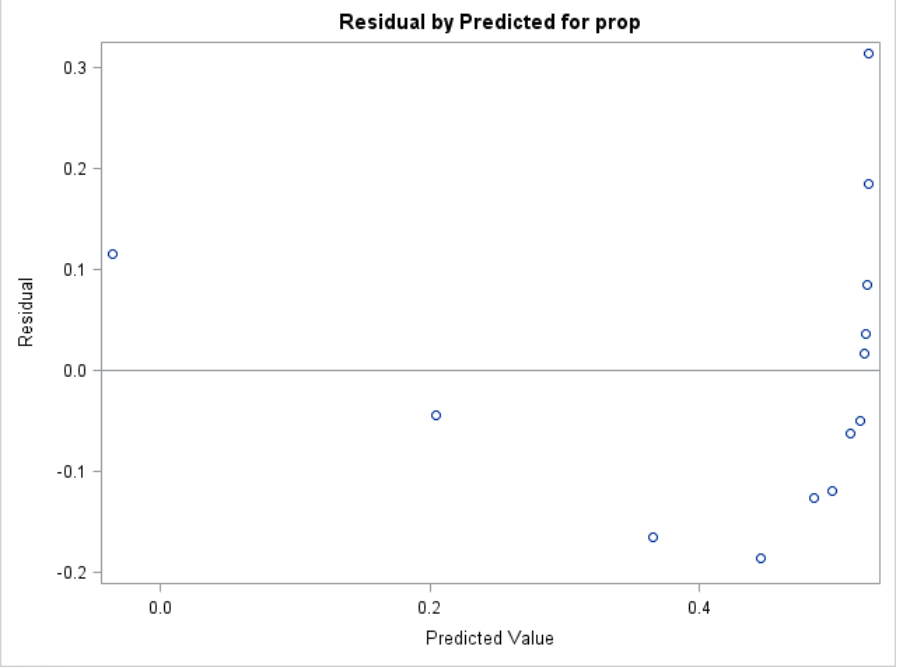
Example 1: *(Capacitor voltage)*

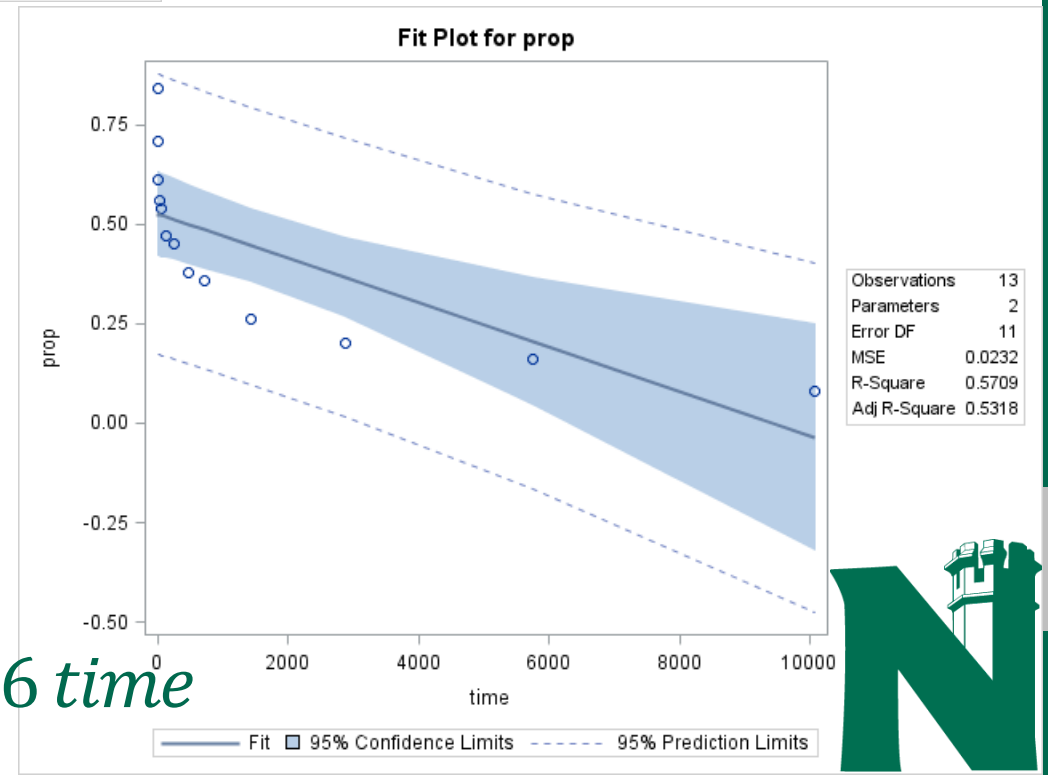
Let's consider the data from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items. The subjects were then asked to recall the items at various times up to a week later. The proportion of items (y = prop) correctly recalled at various times (x = time, in minutes). The data are in the file *wordrecall.csv*.

1. Use scatterplot to determine the relationship between the time and proportion.



There is no linear relationship between the variables.

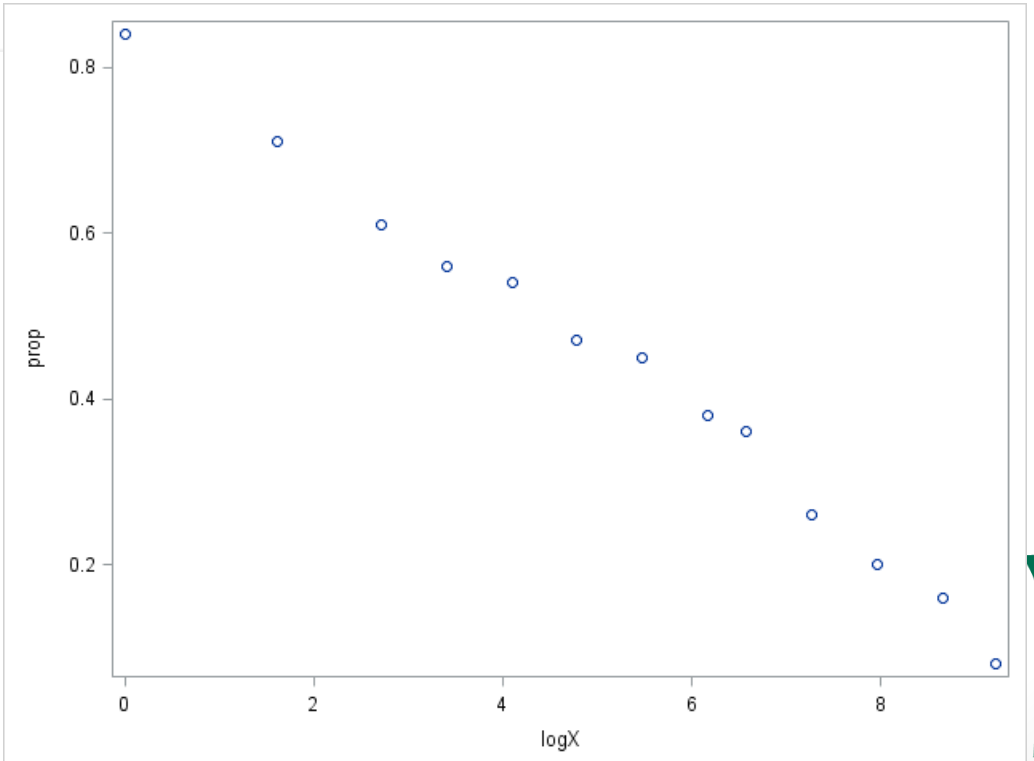




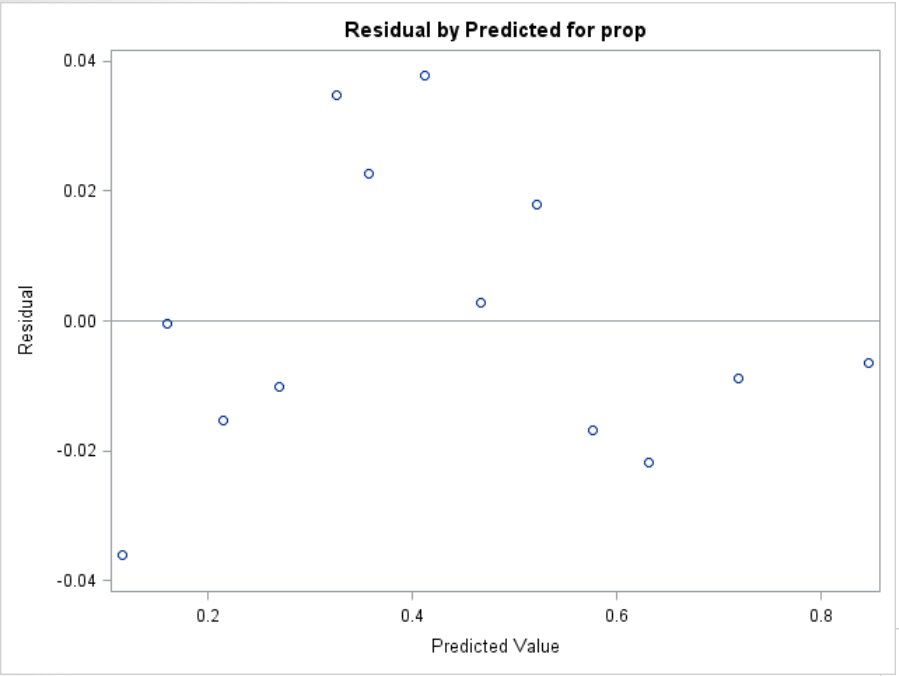
$$\widehat{Prop} = 0.526 - 0.00006\,time$$

2. Transform time using a log transformation and then plot scatterplot of log(time) versus prop.

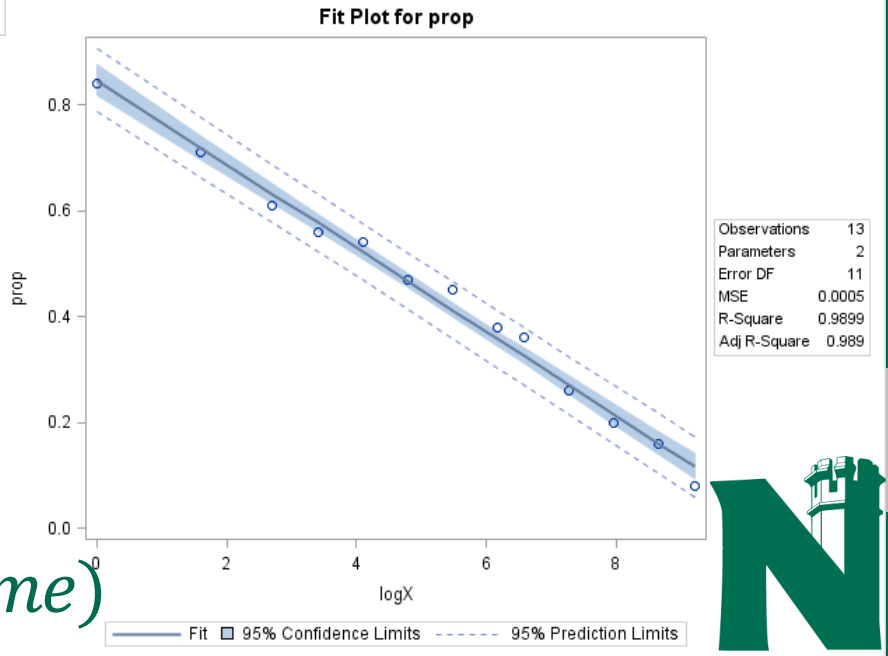
```
data wordrecall_new;
set wordrecall;
expX = exp(time);
recX = 1 / (time);
sqrX = sqrt(time);
logX = log(time);
sqX = time **2;
run;
```



3. Run the regression analysis using log(prop).




```
proc reg data = wordrecall_new plots=all;
model prop = logX;
run;
```



$$\widehat{Prop} = 0.846 - 0.079\log(time)$$

4. Using the original model and transformed model, to predict value of the prop corresponding 5 minutes (time)?

The original model:

$$\widehat{Prop} = 0.526 - 0.00006\ (5) = 0.526$$

The transformation model:

$$\widehat{Prop} = 0.846 - 0.079\ \log(5) = 0.719$$

5. Compare between the residual values of the two models corresponding 5 minutes?

The linear model:

$$residual = 0.526 - 0.71 = -0.184$$

The transformed model:

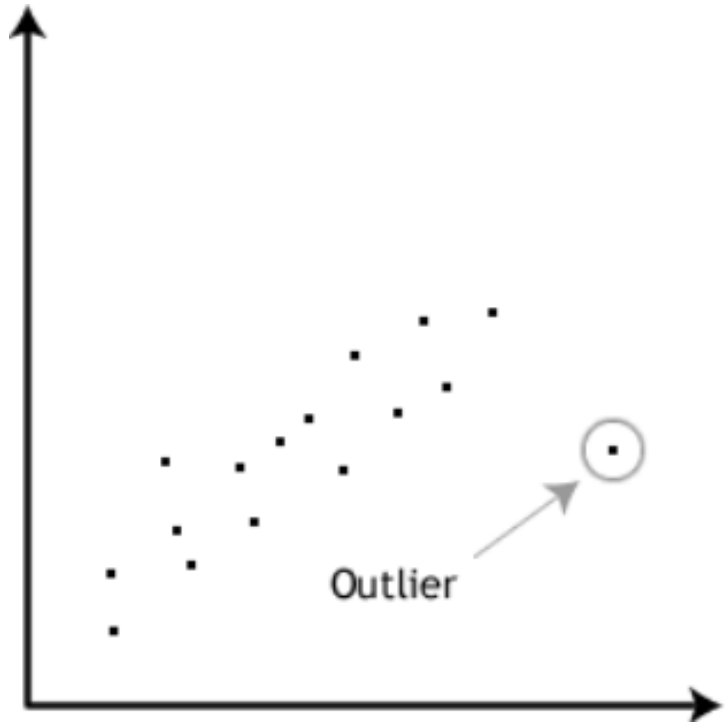
$$residual = 0.719 - 0.71 = -0.009$$

The transformed model fits the data better then the original model.

- We can use the standard error as a measure of the predictions accuracy.

1.5 Outliers and Influential Points:

We classify a data point as an outlier if it stands out away from the pattern of the rest of the data and is not well describe by the model.



The outliers can be:

1. A mistake was made.
   - If it's obvious that a mistake was made in recording the data, it's okay to throw out an outlier and do the analysis without it.
2. The person (or unit) belongs to a different population, and should not be part of the analysis.
   - it's okay to remove the point(s).
3. Sometimes outliers are simply the result of natural variability.
   - In that case, it is NOT okay to discard them. If you do, you will underestimate the variance.

Example 2: *(Olympic long jump)*

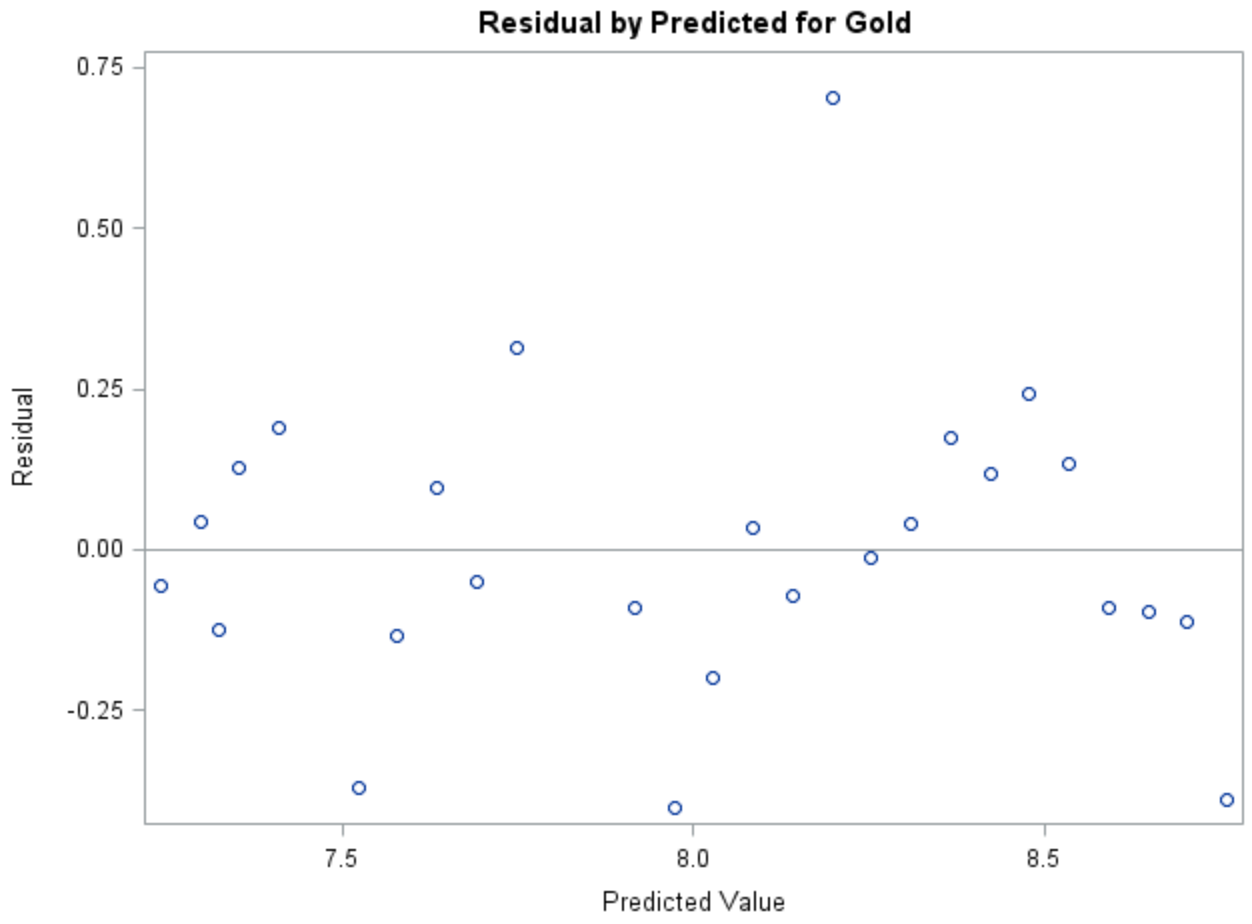
During the 1968 Olympic, Bob Beamon shocked the track and field world by jumping 8.9 meters.

The data are in the file ***LongJumpOlympic.csv***.

1. Use scatterplot to determine the relationship between the year and distance.


Gold
9
8.5
8
7.5
7
1900
1920
1940
1960
1980
2000
2020
Year

2. Using predicted values vs. residual. Can we expect an outlier?



Regression Standard Error:

The standard error of the estimate is a measure of the accuracy of predictions.

For a simple linear regression model, the estimated standard deviation of the error term ($\hat{\sigma}_\epsilon$) based on the least squares fit to a sample of n observations is

$$\hat{\sigma}_\epsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

- SSE = Sum Square of Residuals.
- We can use $\sqrt{MSE}$ to compare between the models to check which fit the data better (less variation).

Example 3: *(Porsche prices)*

For the same dataset in example 1.

Find the regression standard error.

Root MSE	7.17029	R-Square	0.7945
Dependent Mean	50.53667	Adj R-Sq	0.7872
Coeff Var	14.18829		

$$\hat{\sigma}_{\epsilon} = 7.17029$$

Reading Assignment

Read section 1.2-1.6