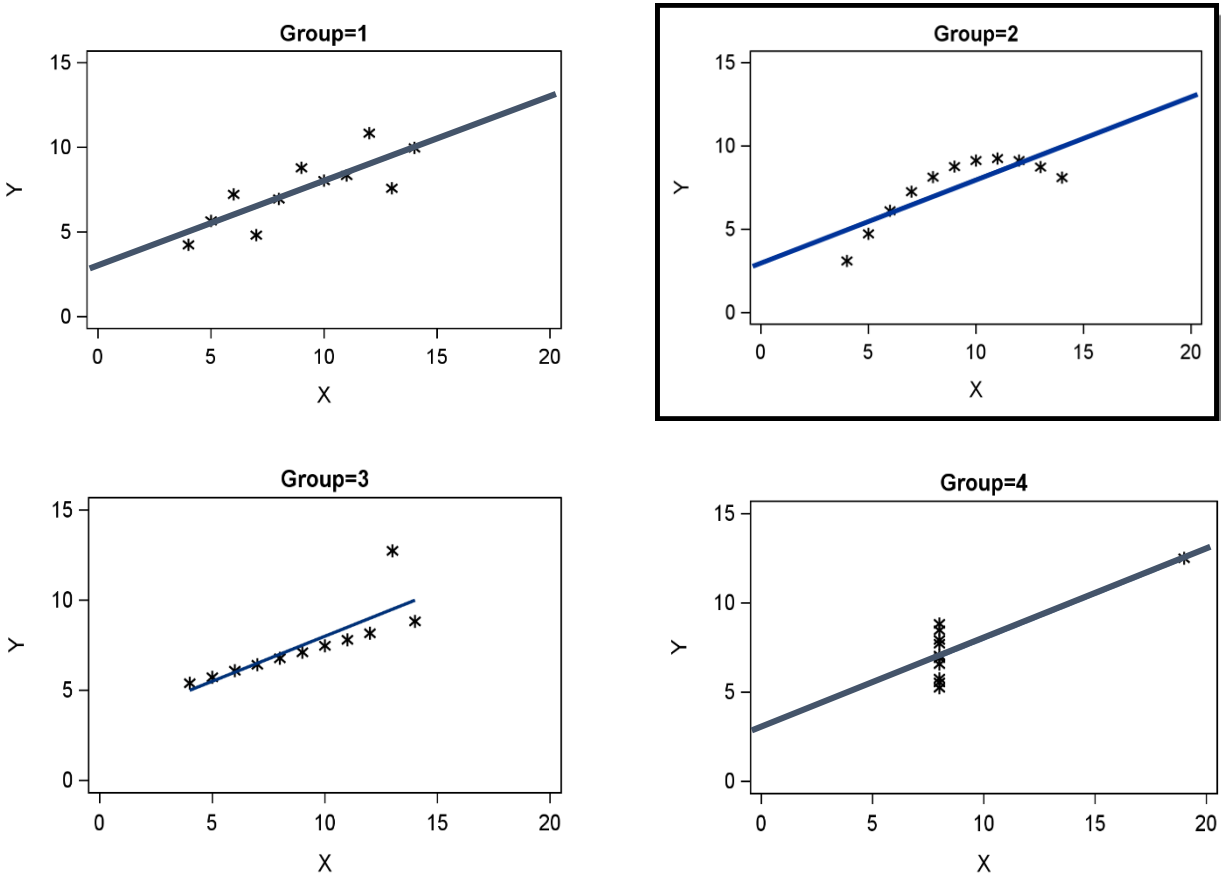
Diagnostics

Class of 2023

Before we start discussing diagnostics, it is ESSENTIAL that you visualize your data!!

Y-hat=3+0.5X

$R^2 = 0.67$



Diagnostics

- Examining Residuals
- Misspecified Model
- Lack of Constant Variance
- Lack of Normality
- Correlated error terms
- Influential points and outliers
- Multicollinearity

- Ames Housing Data complete example

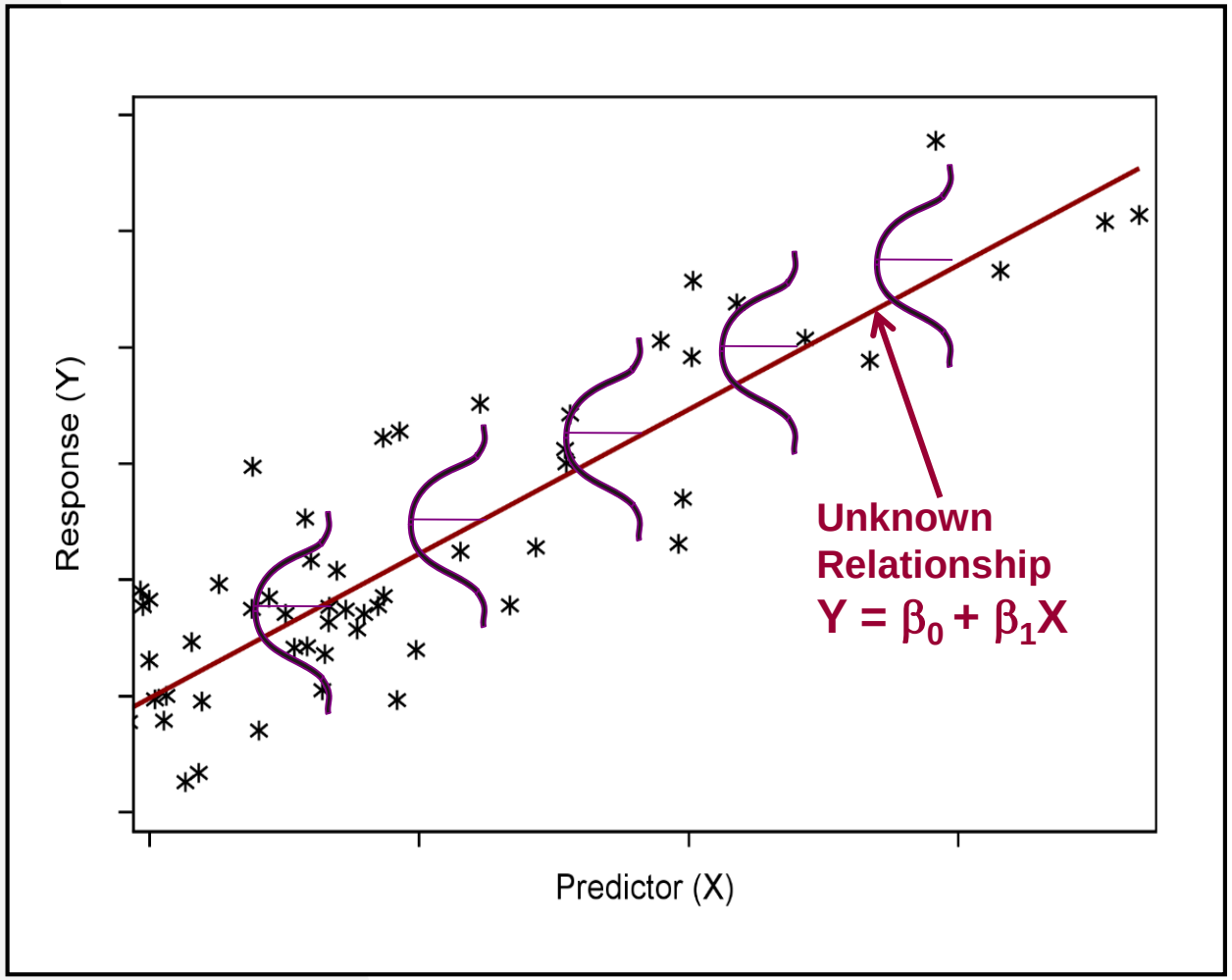
Examining Residuals

Linear Regression Assumptions

- The mean of the Ys is accurately modeled by a **linear** function of the Xs.
- The random error term, ε, is assumed to have a **normal** distribution with a mean of zero.
- The random error term, ε, is assumed to have a **constant variance**, σ^2.
- The errors are **independent**.

- **No perfect collinearity**

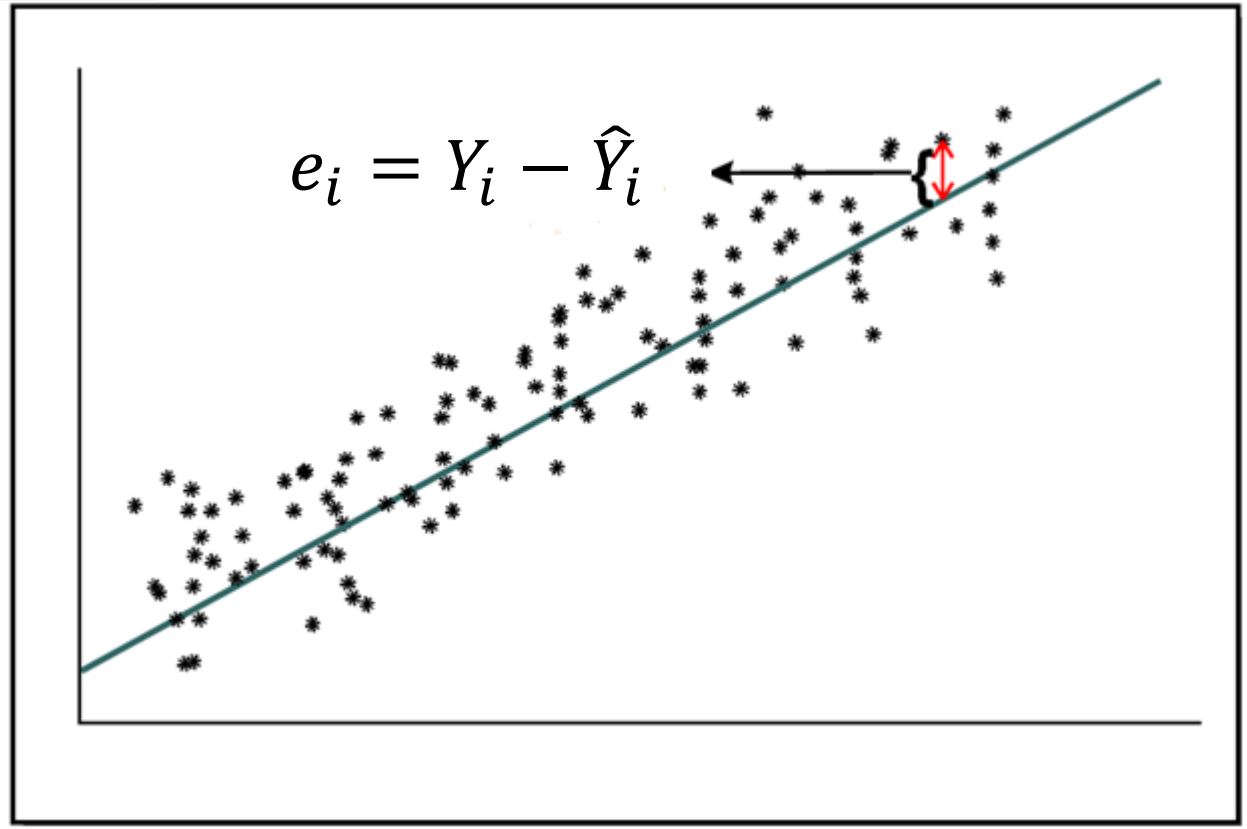
Linear Regression Assumptions



Violation of Model Assumptions

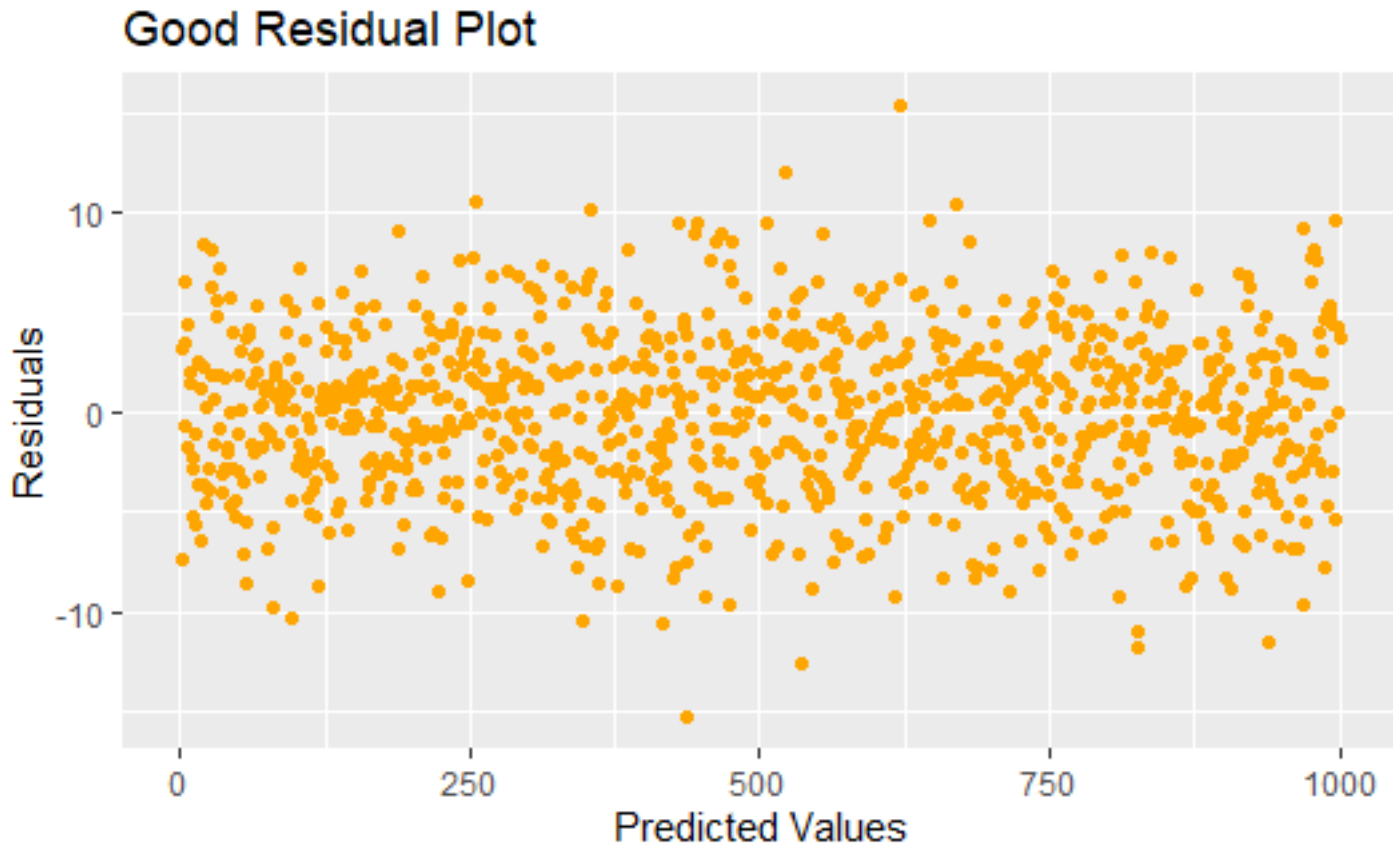
- **Linear in the parameters** – indicates a ***misspecified model***, and therefore the results are not meaningful.
- **Constant Variance** – does not affect the parameter estimates, but the standard errors are compromised.
- **Normality** – does not affect the parameter estimates, but it affects the test results.
- **Independent observations** – does not affect the parameter estimates, but the standard errors are compromised.

Many assumptions are investigated using residuals



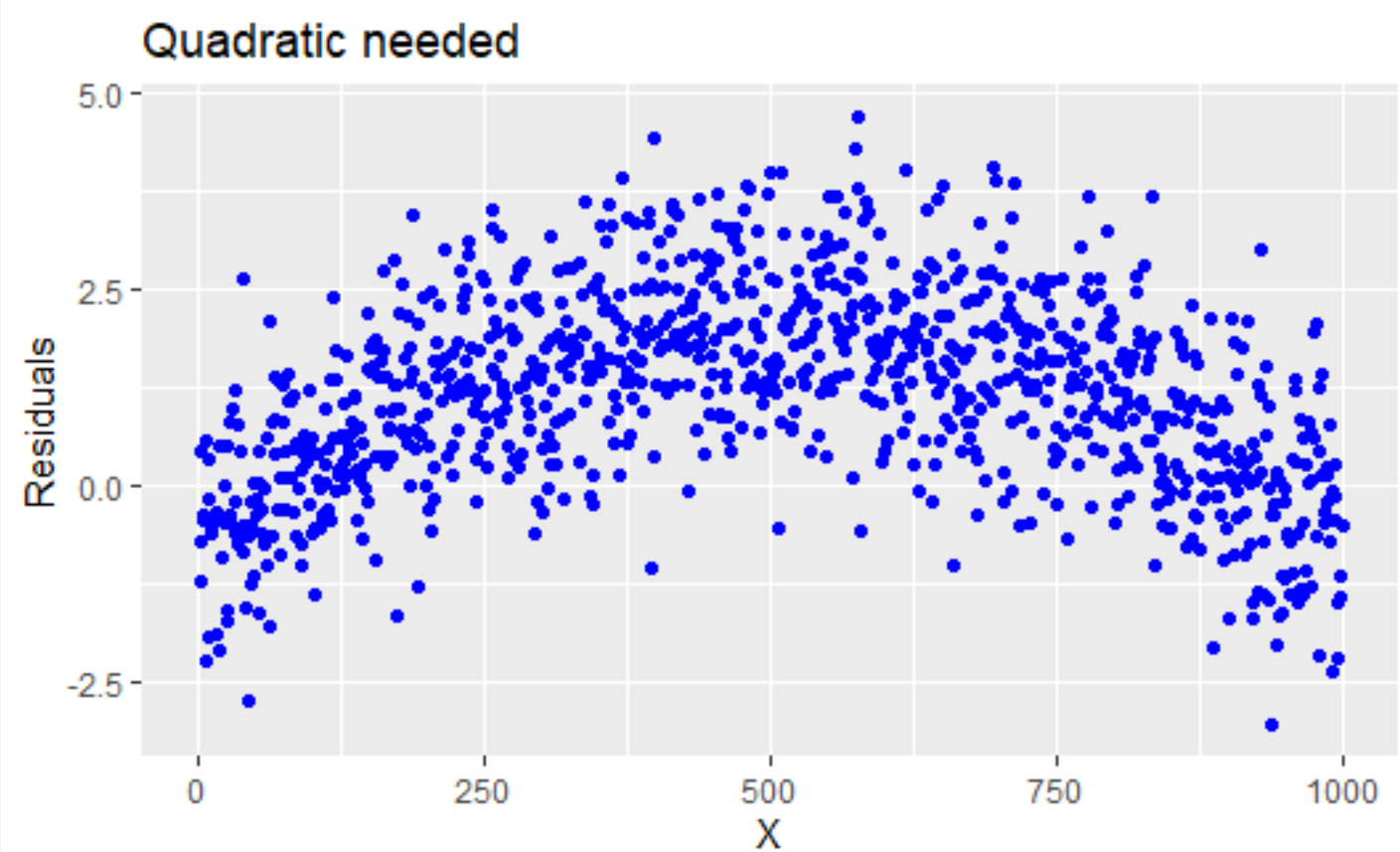
Examining Residual Plots (good residual plot)

- Plot residuals(y-axis) versus each x(x-axis) (or residuals(y-axis) versus predicted value(x-axis))
  - Residuals are randomly scattered about zero reference line.
  - No patterns found.
  - Model form appears to be adequate.



Misspecified Model

Examining Residual Plots-Misspecified Model



- Pattern is detected (for example, curvilinear) in residuals.
- Model form is incorrect.
- Possible remedies, depending on pattern, include polynomial terms, interactions, splines, and so on.

Polynomial Regression Models

- Quadratic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j$$

- Cubic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3 + \varepsilon_j$$

- Polynomial Model with a Cross-Product Term

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j} X_{2j} + \varepsilon_j$$

Model Hierarchy

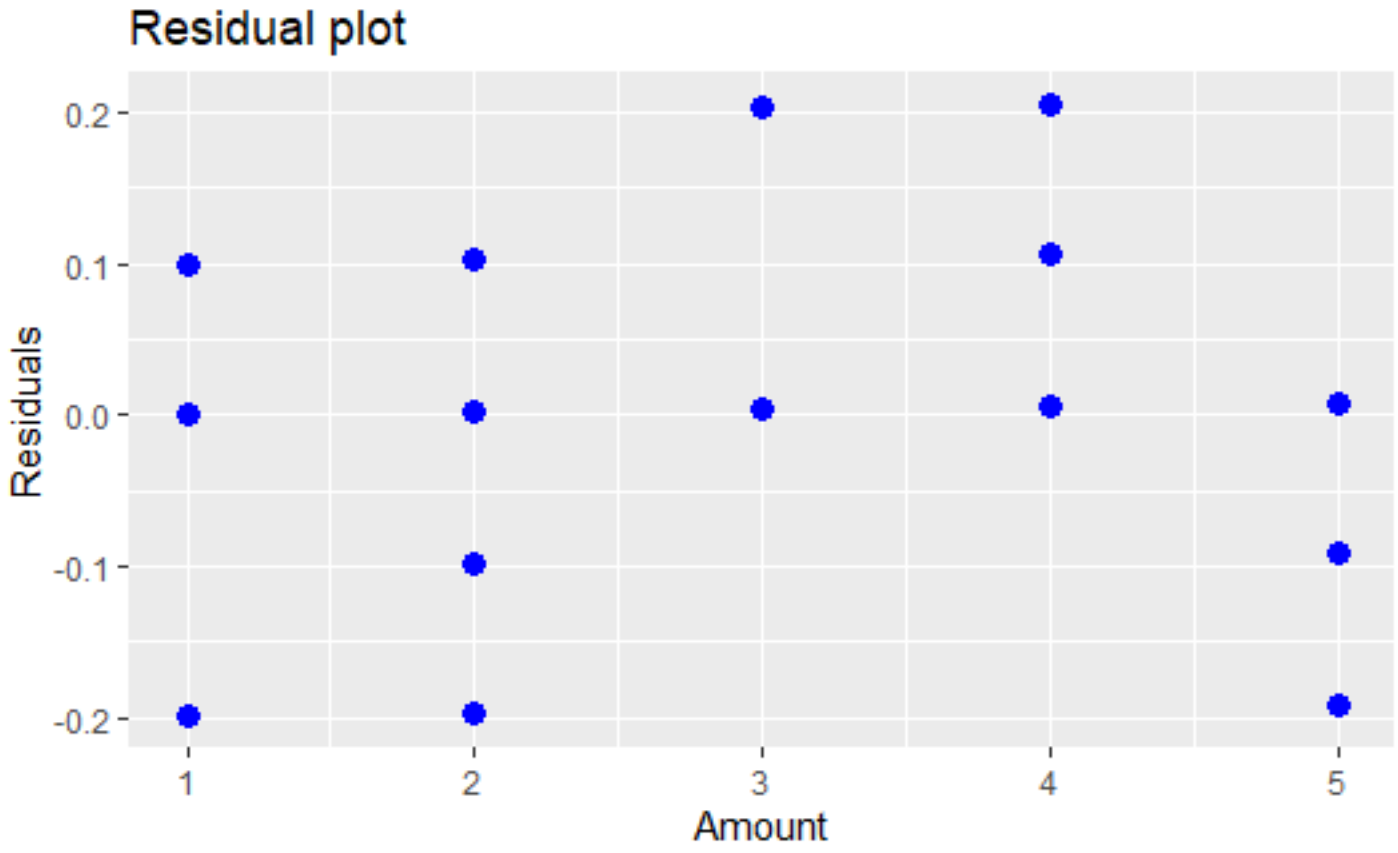
- When adding higher order terms (power terms and/or interactions), you should have ALL lower terms included in the model.
- For example, if you x^3 in the model, you should have x and x^2 also in the model
- If you include an interaction between x_3 and x_4 (x_3x_4) in the model, then x_3 AND x_4 should also be included in the model
- This is referred to as model hierarchy

Example of polynomial regression

A researcher is interested in studying the effect of a chemical additive on paper strength. The response variable is the amount of force required to break the paper (strength) and the explanatory variable is the amount of chemical additive (amount).

Example

```
lm.quad=lm(strength~amount)
 summary(lm.quad)

ggplot(lm.quad,aes(x=amount,y=resid(lm.quad)))+geom_point(
color="blue",size=3)+labs(title="Residual plot", x="Amount",
y="Residuals")
```



Fitting a Quadratic

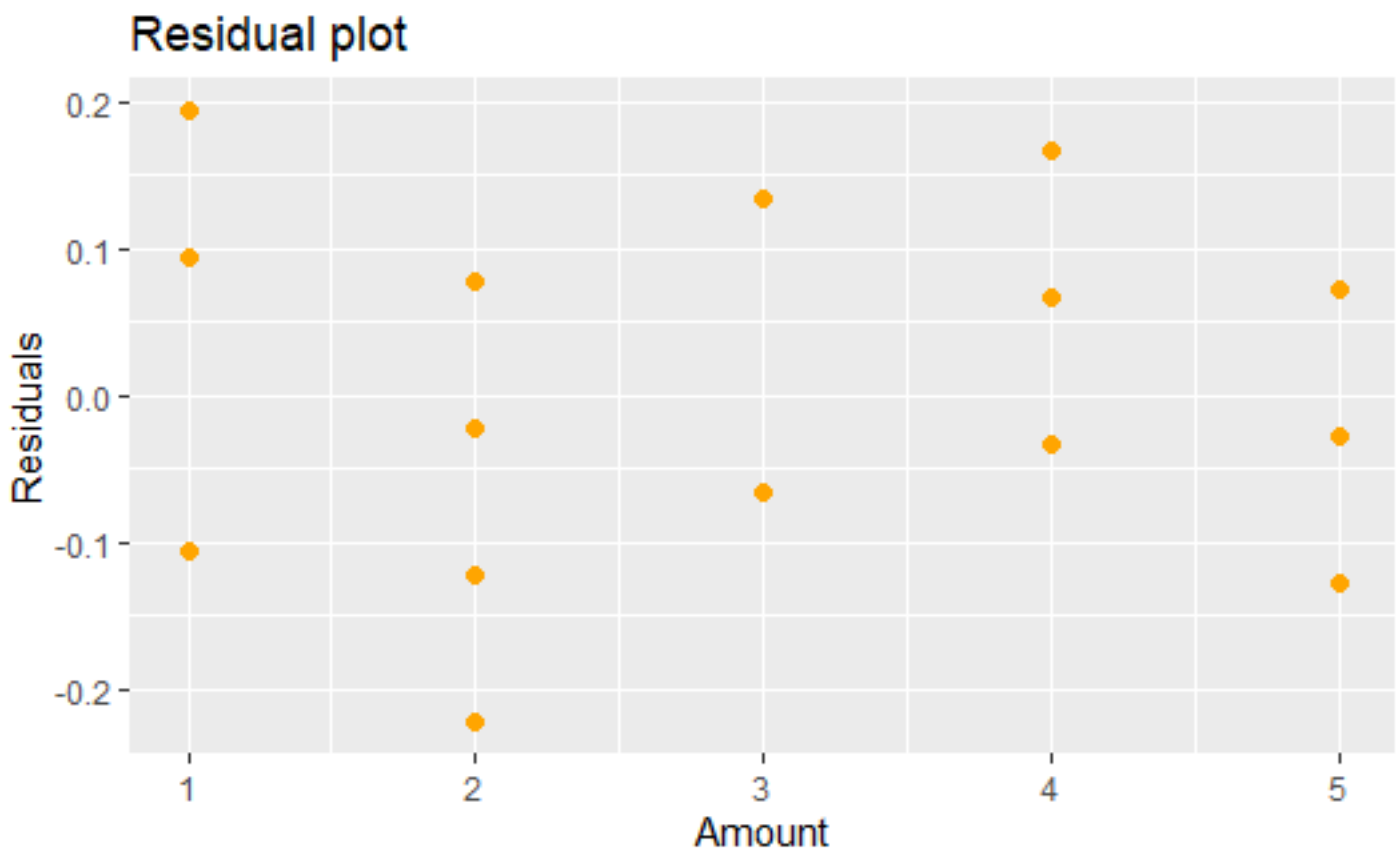
lm.quad=lm(strength~amount + I(amount^2))

summary(lm.quad)

$$\hat{Y}_i = 2.21 + 0.33x_i - 0.04x_i^2$$

```
Call:
lm(formula = strength ~ amount + I(amount^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.22276 -0.06562 -0.02763  0.07602  0.19466

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21334    0.13399  16.519 9.97e-13 ***
amount       0.32928    0.09690   3.398  0.00302 **
I(amount^2) -0.03728    0.01535  -2.428  0.02526 *
```

Residual Plot



Third degree polynomial

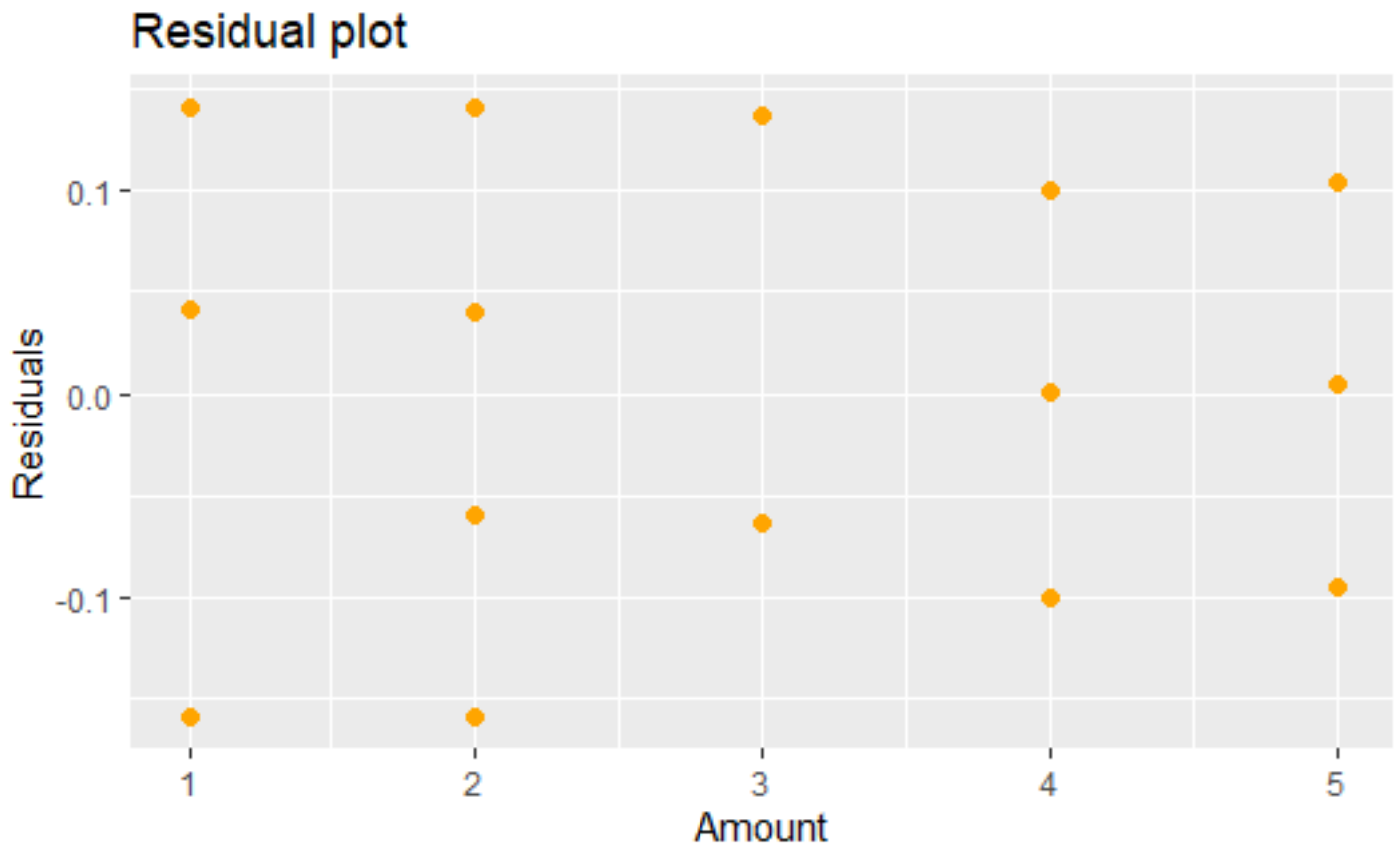
```
lm.quad=lm(strength ~ amount + I(amount^2) + I(amount^3))

summary(lm.quad)
ggplot(lm.quad,aes(x=amount,y=resid(lm.quad)))+geom_point(color="ora
nge",size=2)+labs(title="Residual plot", x="Amount", y="Residuals")
```

Model:

```
Call:
 lm(formula = strength ~ amount + I(amount^2) +
I(amount^3))

 Residuals:
      Min        1Q    Median        3Q       Max
 -0.15941 -0.06360  0.00272  0.08579  0.14142

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept)  2.73280    0.26060  10.487 4.28e-09 ***
 amount      -0.36900    0.32208  -1.146   0.2669
 I(amount^2)  0.22339    0.11651   1.917   0.0712 .
 I(amount^3) -0.02862    0.01270  -2.254   0.0369 *
```

$$\hat{Y}_i = 2.73 - 0.37x_i + 0.22x_i^2 - 0.03x_i^3$$

Residual Plot



When a straight line is inappropriate

Consider the following options:

- Fit a polynomial/more complex regression model.
- Transform the dependent and/or independent variables to obtain linearity.
- Fit a nonlinear regression model, if appropriate.
- Fit a nonparametric regression model (for example, LOESS)

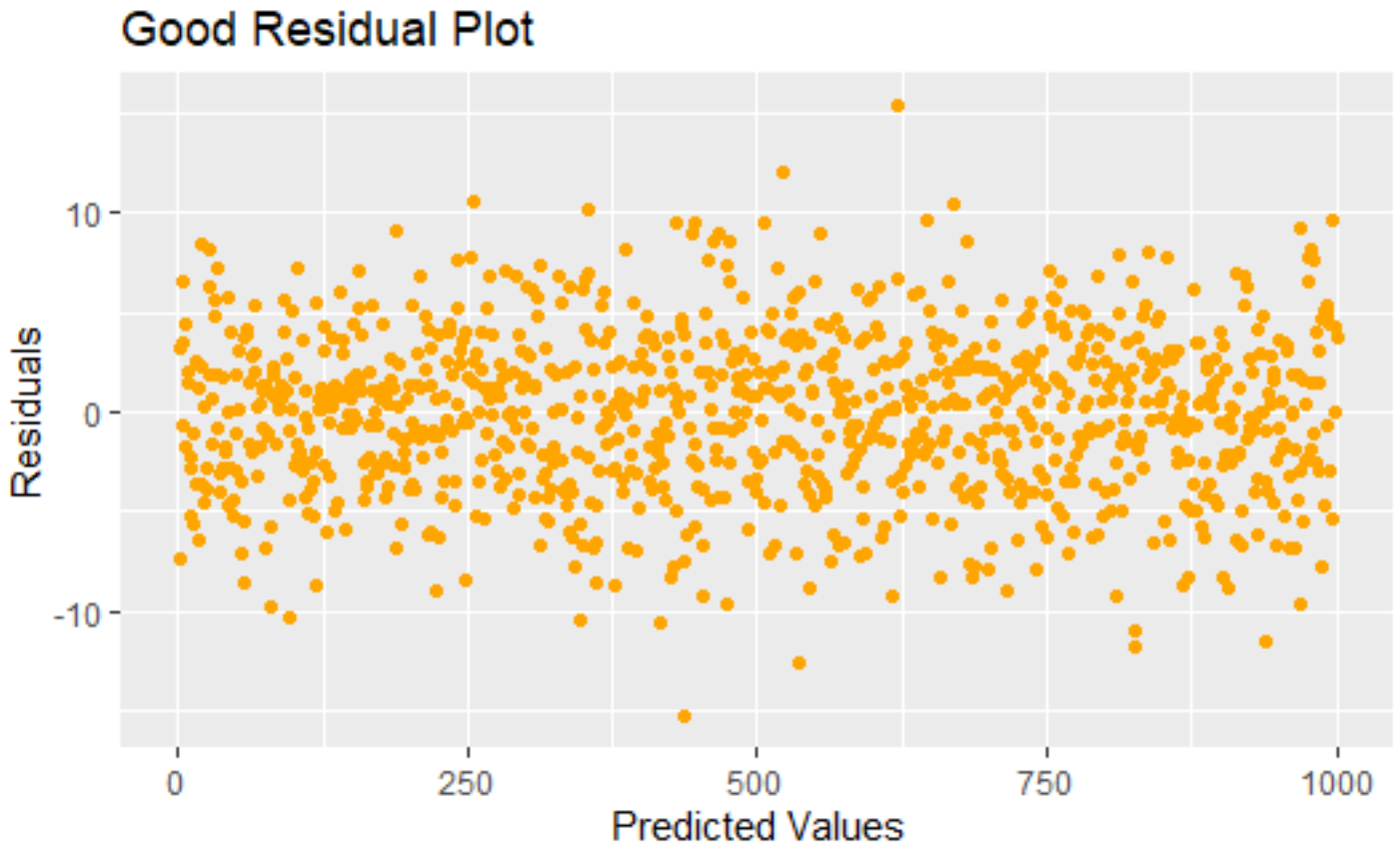
Lack of Constant Variance

Examining Residual Plots-Variance is not constant

- Constant variance assumption is violated.
- Possible remedy is transforming variables to stabilize the variance.
- Procedures that model the non-constant variance can be used.

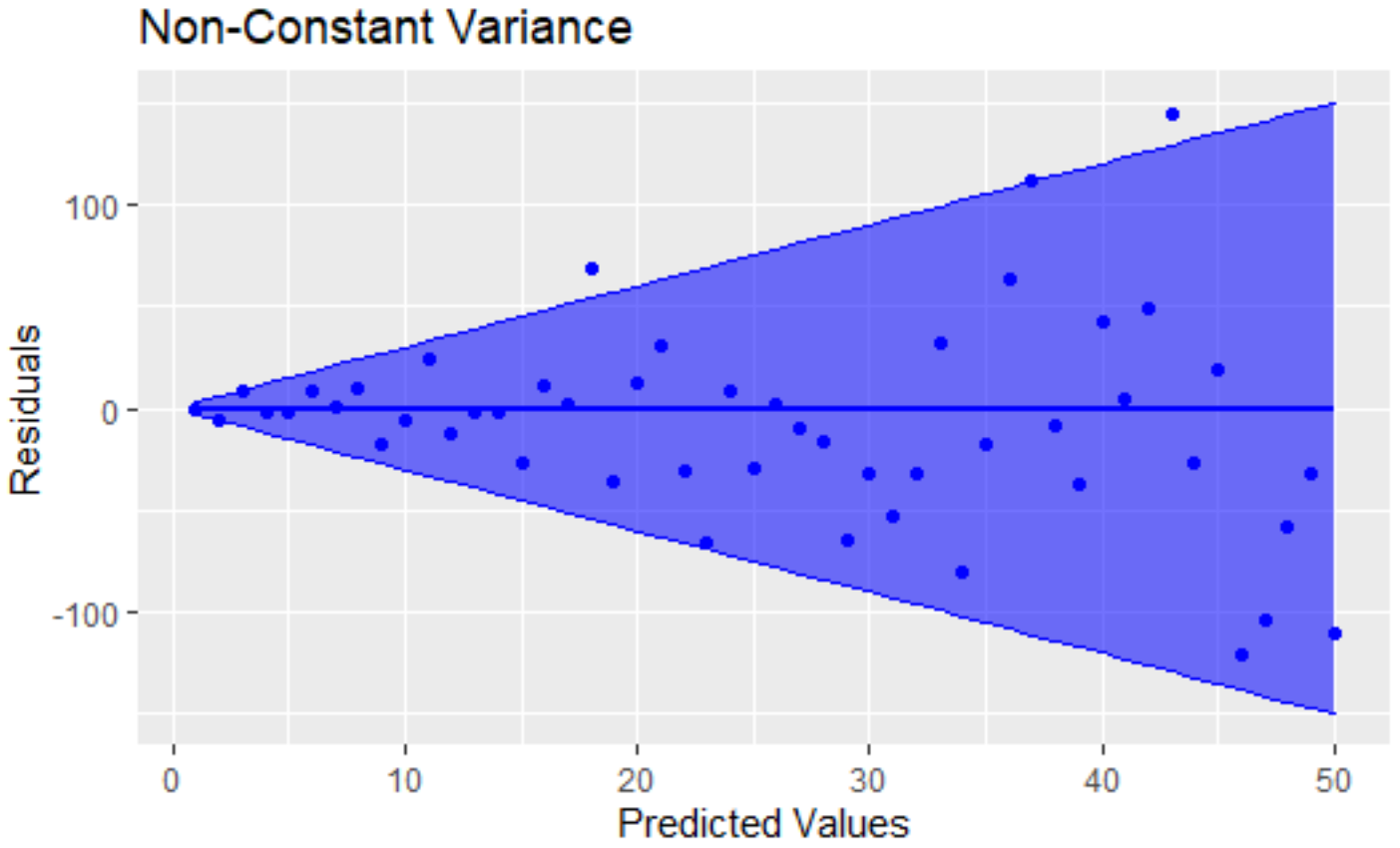
Homoscedasticity

- The random error term, ε, is assumed to have a constant variance, σ^2 (**homoscedasticity)**.
- This is an example of **heteroscedasticity**.
  - Does ***not*** affect the calculation of the parameter estimates.
  - Does affect the standard errors of the parameter estimates.



Homoscedasticity

- The random error term, ε, is assumed to have a constant variance, σ^2 (**homoscedasticity**).
- This is an example of **heteroscedasticity**.
  - Does ***not*** affect the calculation of the parameter estimates.
  - Does affect the standard errors of the parameter estimates.



Heteroscedasticity

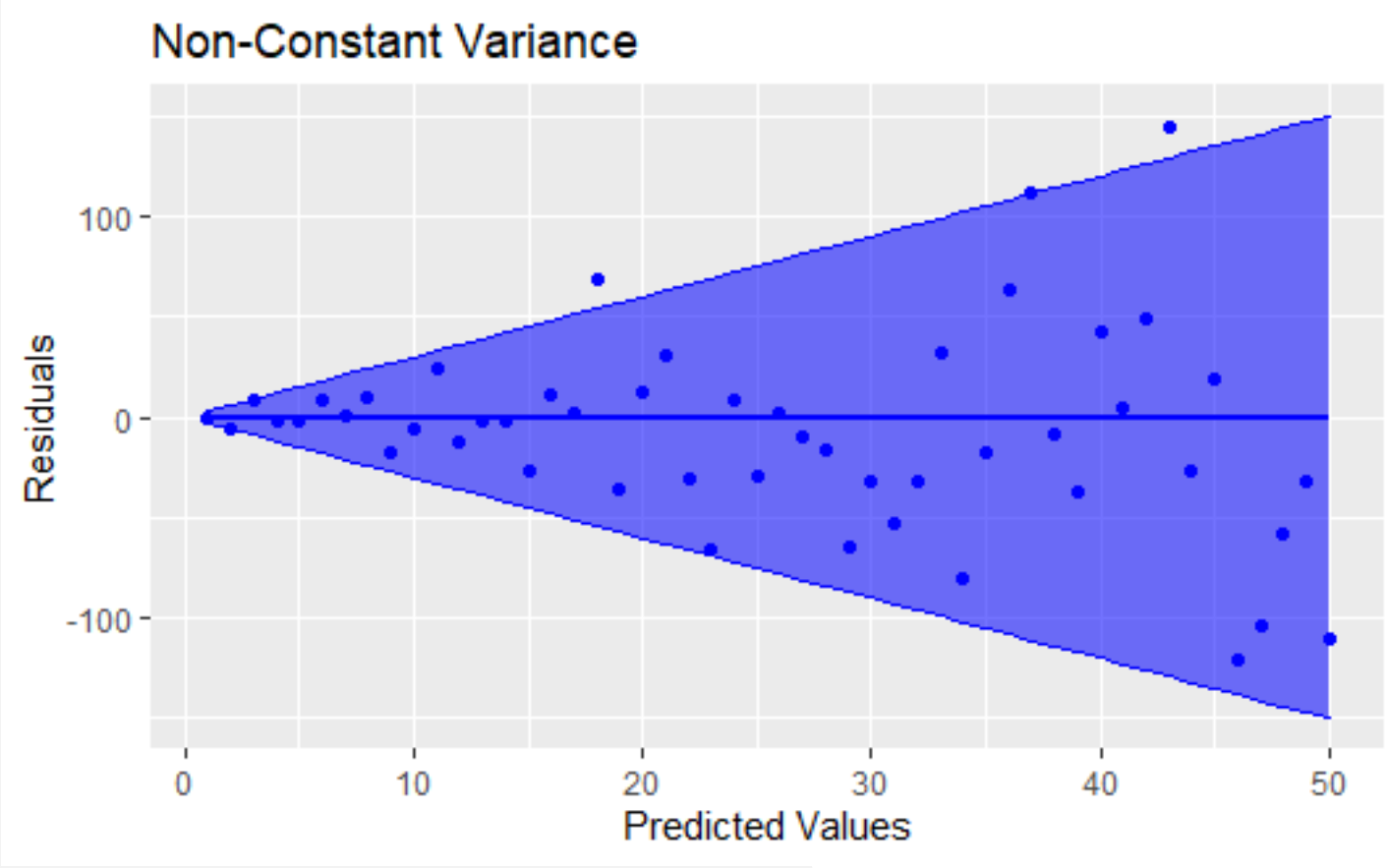
- Any inferences under the traditional assumptions will be incorrect.
- Hypothesis tests and confidence intervals based on the t, F, χ^2 distributions will not be valid.

Detecting Heteroscedasticity

- There are a couple of approaches to detecting heteroscedasticity in a data set.
  1. Plotting residuals and looking for patterns.
  2. Spearman Rank Correlation

     The Spearman correlation uses ranks of the data (still between -1 and 1)

Detecting Heteroscedasticity

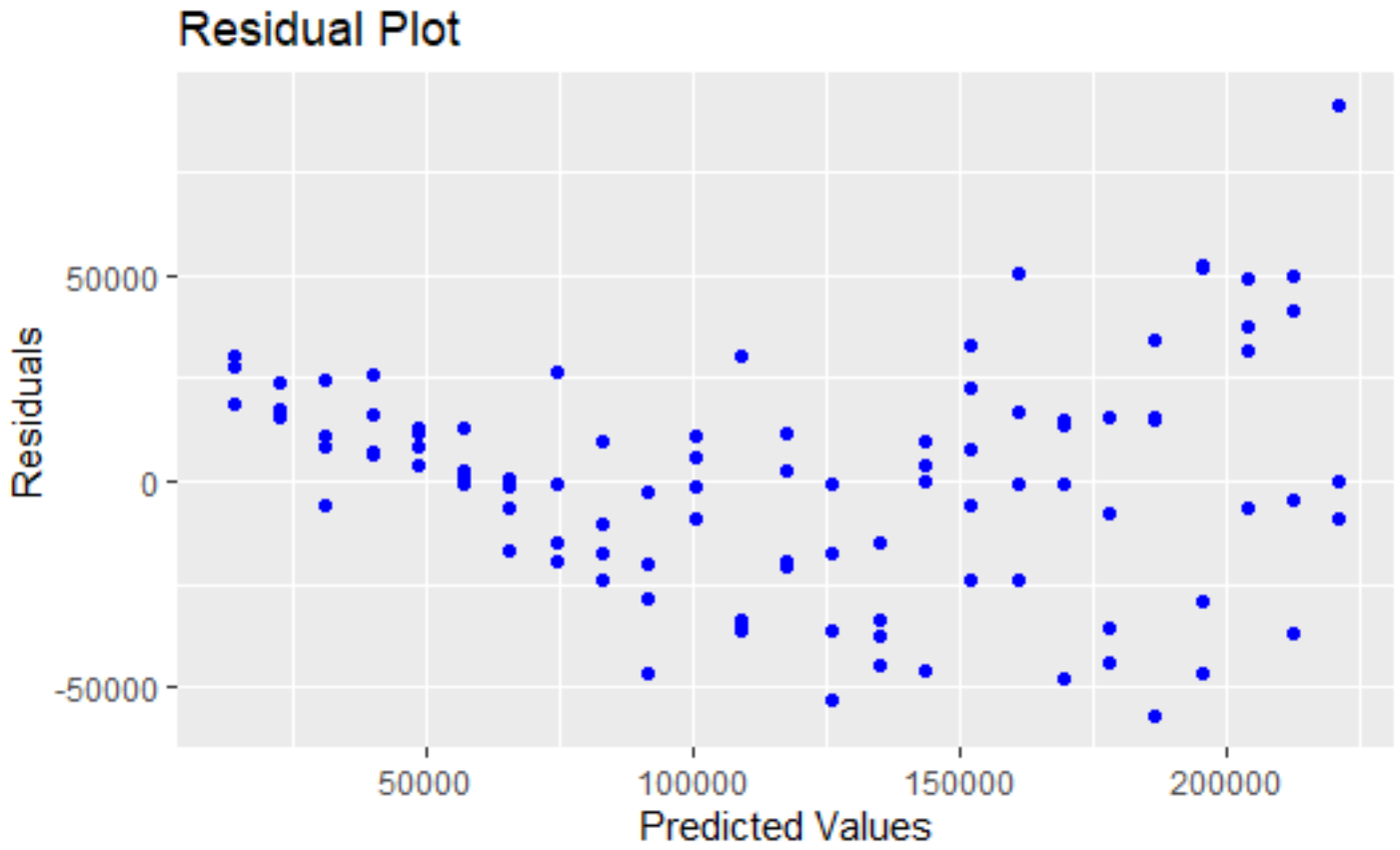


Plotting residuals and looking for patterns.

Spearman Rank Correlation

- If the Spearman rank correlation coefficient between the ***absolute value of the residuals*** and the ***predicted values*** is
  - close to zero, then the variance is potentially homoscedastic
  - positive, then the variance increases as the mean increases
  - negative, then the variance decreases as the mean increases
  - Can perform a test: H_0: variance is homoscedastic
    H_A: variance is heterscedastic
  - If there is a relationship between the absolute value of residuals and predicted value but it is not linear, this test will NOT discover it

Salary data set



lm.var=lm(salary~years)

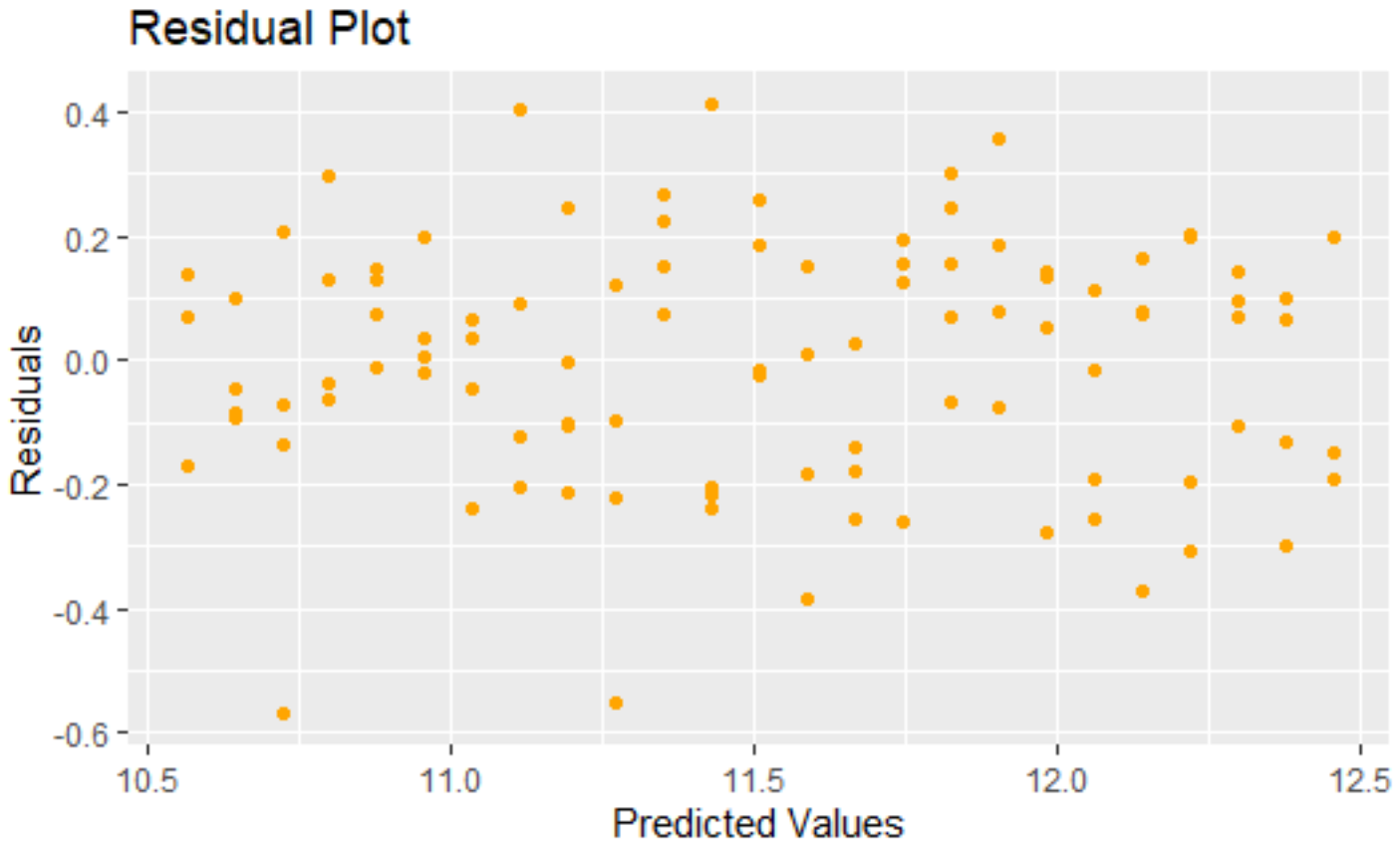
ggplot(lm.var,aes(x=fitted(lm.var),y=resid(lm.var)))+geom_point(color="blue")+labs(title="Residual Plot", x="Predicted Values",y="Residuals")

Spearman rank correlation test

```
cor.test(abs(resid(lm.var)),fitted.values(lm.var),method="spear
man",exact=T)

        Spearman's rank correlation rho

data:  x and y
S = 115122, p-value = 0.001747
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3091986
```

Use variance-stabilizing transformation

```
lm.var=lm(log(salary)~years)

ggplot(lm.var,aes(x=fitted(lm.var),y=resid(lm.var)))+geom_point(color=
"orange")+labs(title="Residual Plot", x="Predicted
Values",y="Residuals")
```



Accounting for Heteroscedasticity

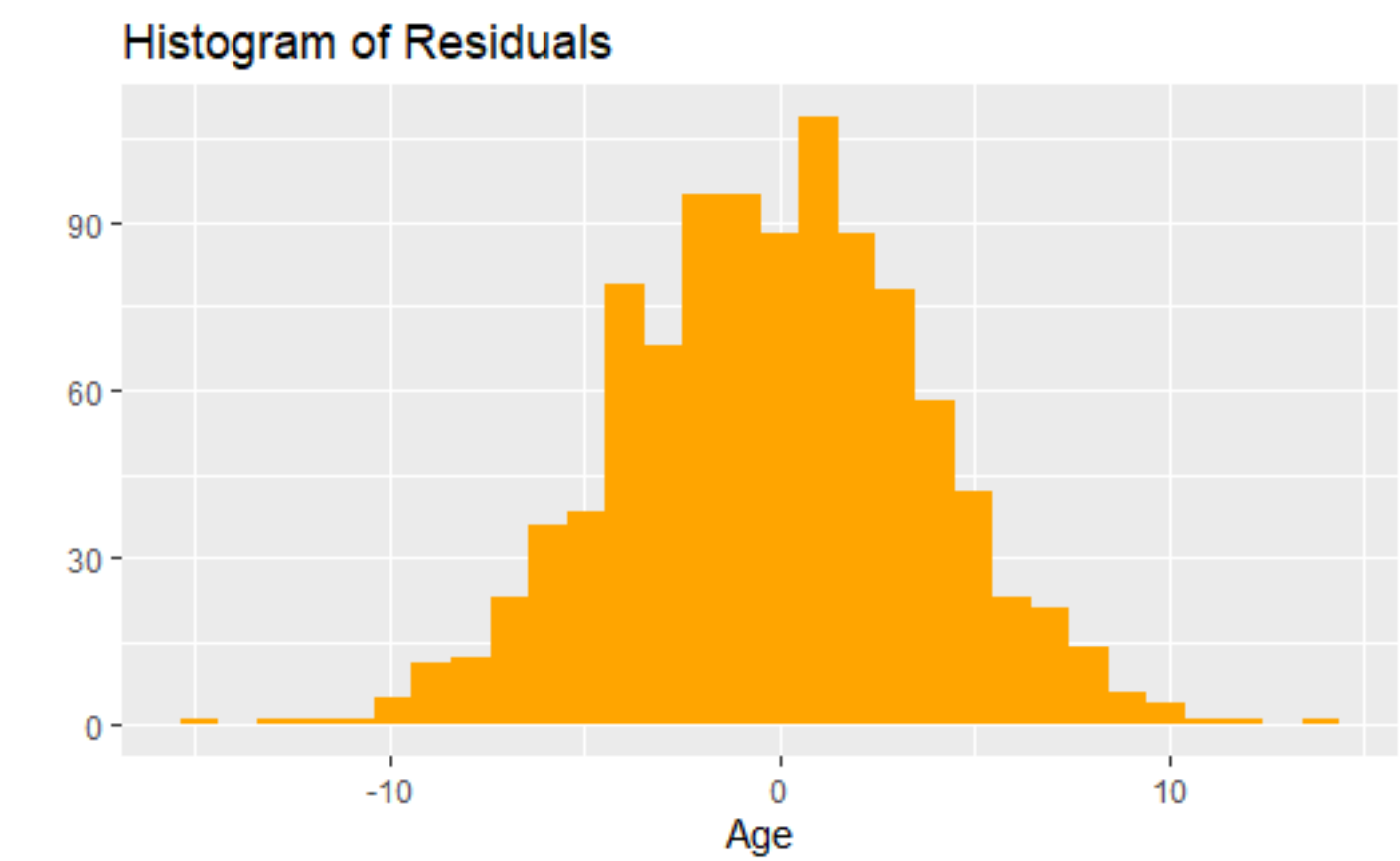
- There are a couple of approaches to account for heteroscedasticity:
  - Use Weighted Least Squares (WLS) or iteratively reweighted least squares (IRLS).
  - Transform data.
  - Use a different distribution (for example, if count data, a Poisson distribution is more appropriate)

Lack of Normality

Detecting Lack of Normality

- Check that the error terms are Normally distributed by examining:
  - Histogram of the residuals
  - Normal probability plot of the residuals (QQ-plot)
  - Formal tests for Normality

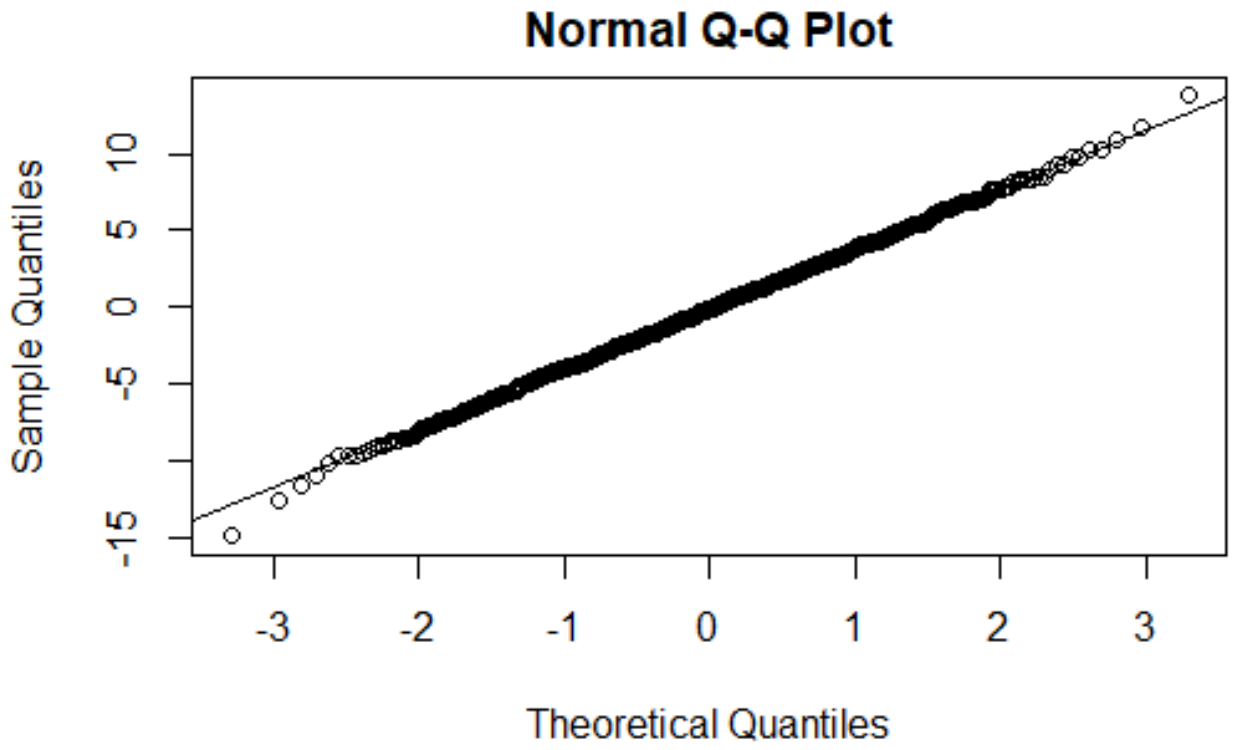
Detecting Lack of Normality

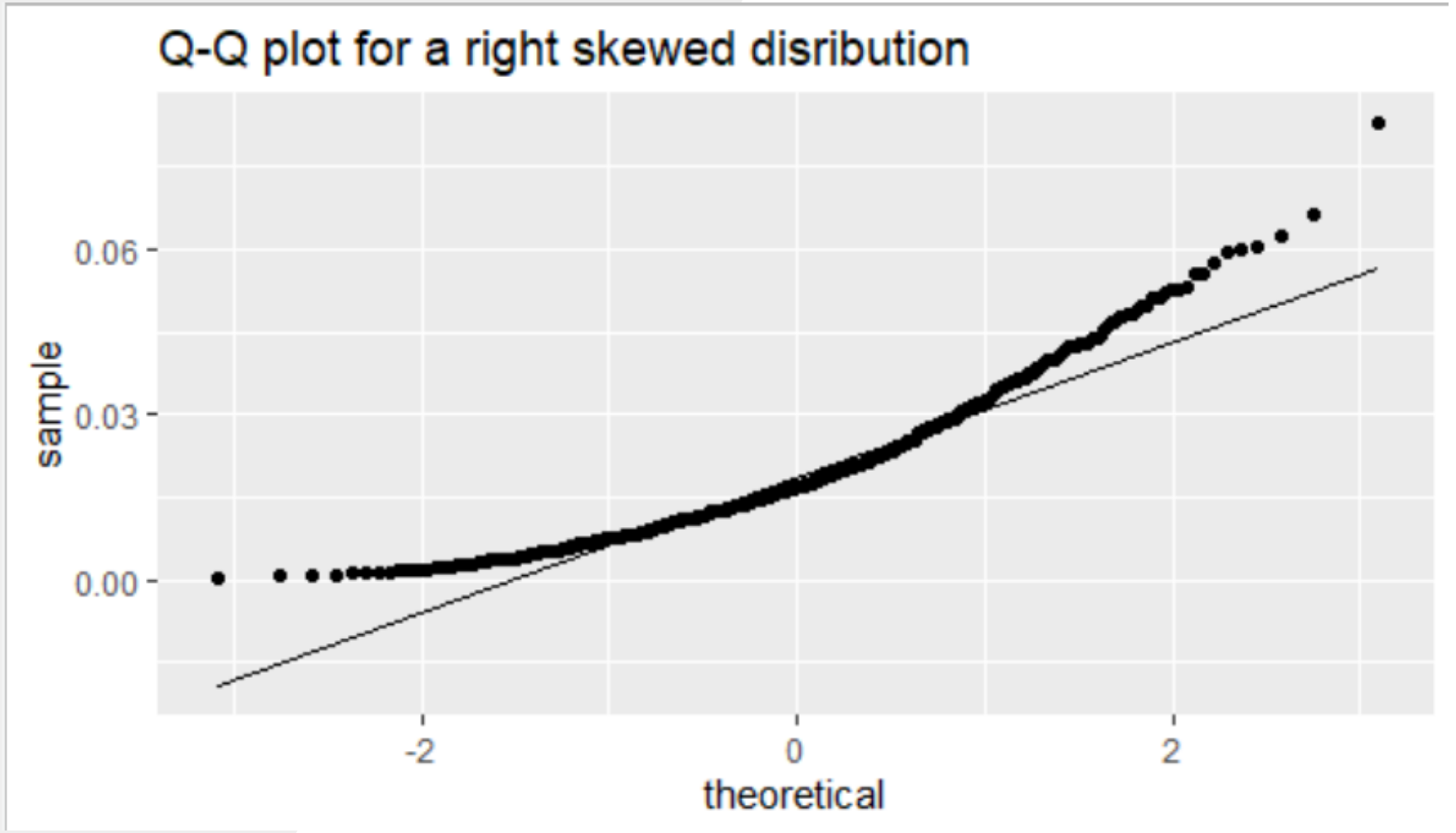


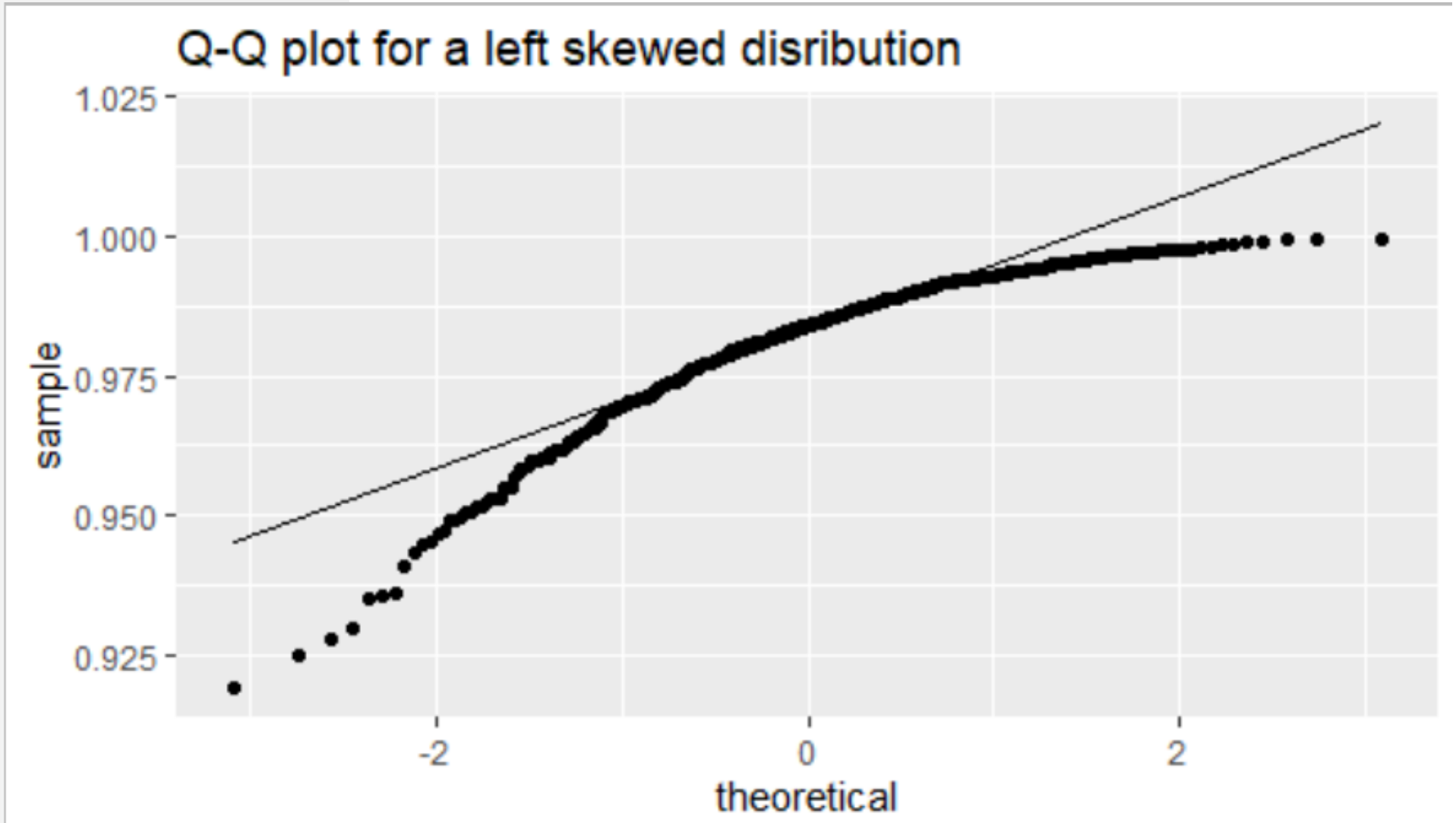
- Check that the error terms are Normally distributed by examining:
  - Histogram of the residuals

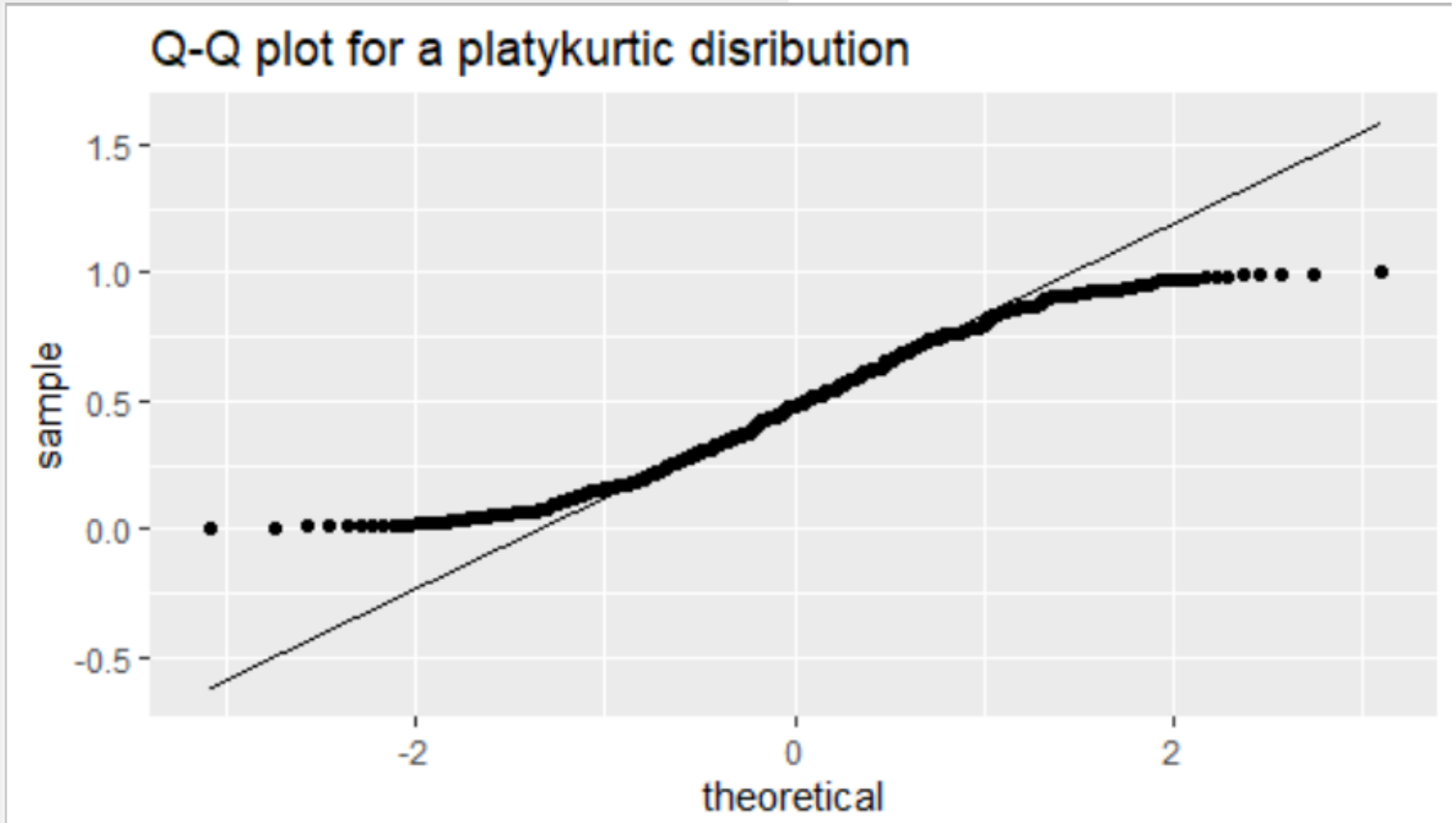
Detecting Lack of Normality

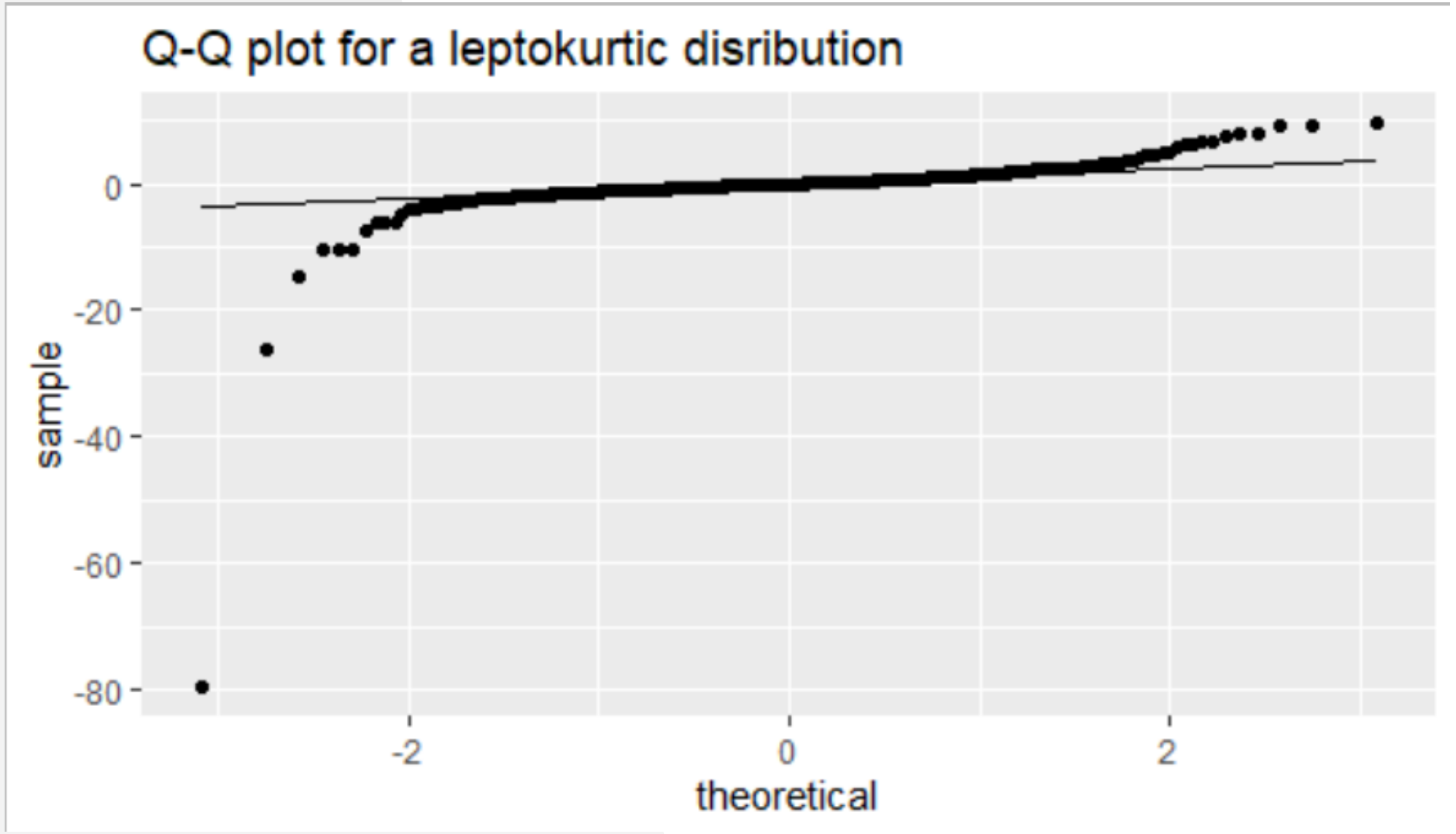
- Check that the error terms are Normally distributed by examining:
  - Histogram of the residuals
  - Normal probability plot of the residuals (QQ-plot)




Q-Q plot for a right skewed disribution
sample
0.06
0.03
0.00
-2
0
2
theoretical


Q-Q plot for a left skewed disribution
1.025
1.000
0.975
0.950
0.925
sample
-2
0
2
theoretical


Q-Q plot for a platykurtic disribution
1.5
1.0
0.5
0.0
-0.5
sample
-2
0
2
theoretical


Q-Q plot for a leptokurtic disribution
0
-20
-40
-60
-80
sample
-2
0
2
theoretical

Detecting Lack of Normality

- Check that the error terms are Normally distributed by examining:
  1. Histogram of the residuals
  2. Normal probability plot of the residuals (QQ-plot)
  3. Formal tests for Normality

H_0: Normality

H_A: Not Normality

Tests for Normality

There are numerous tests for normality. Most have the same null and alternative hypotheses. We will illustrate the following tests:

$$H_0: \text{Normally Distributed}$$
$$H_A: \text{Not Normally Distributed}$$

1. Anderson-Darling is based on the empirical cumulative distribution function of the data and gives more weight to the tails.
2. Shapiro-Wilk test uses the idea of correlation between the sample data and normal scores. The Shapiro-Wilk is better for smaller data sets.

Log(Salary) model

```
> hist(resid(lm.var))
> qqnorm(resid(lm.var))
> qqline(resid(lm.var))
> ad.test(resid(lm.var))
```

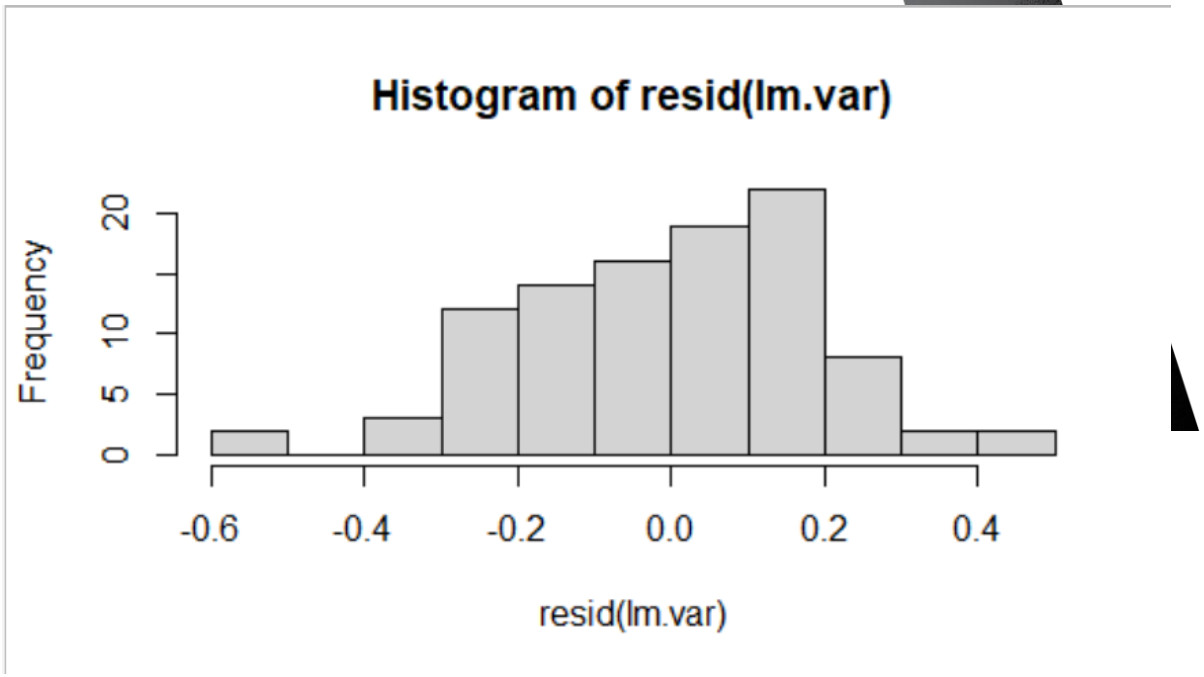
Anderson-Darling normality test

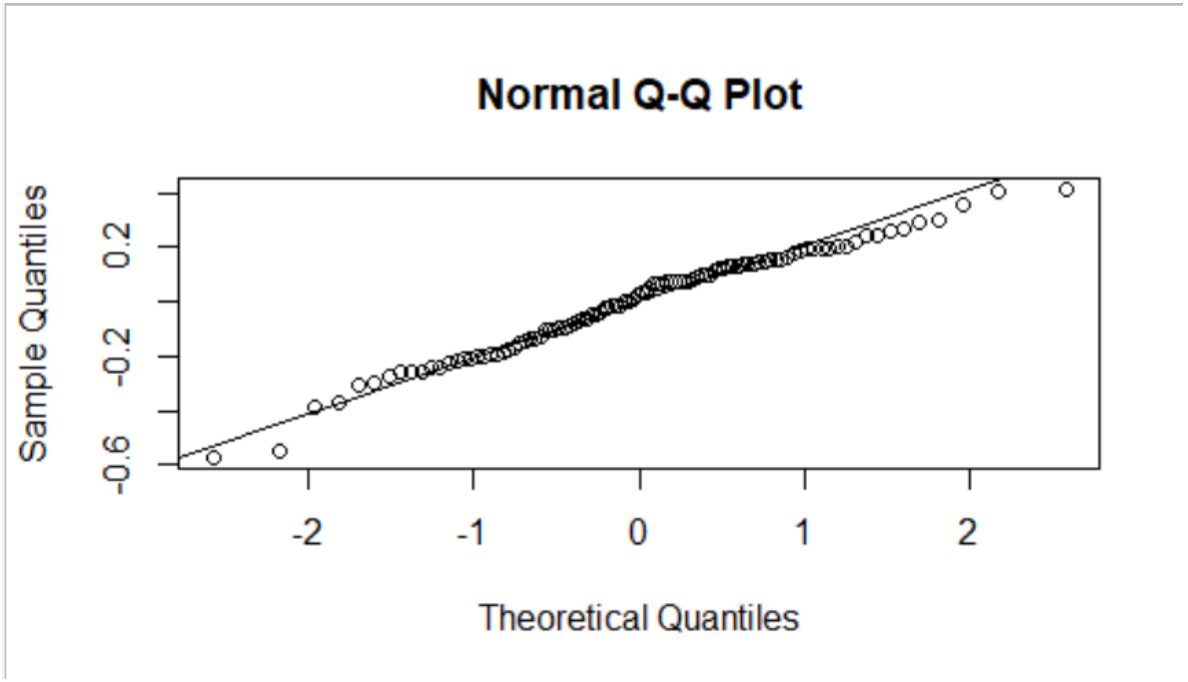
data: resid(lm.var)
A = 0.61387, p-value = 0.1074

```
> shapiro.test(resid(lm.var))
```

Shapiro-Wilk normality test

data: resid(lm.var)
W = 0.98033, p-value = 0.141





Accounting for Lack of Normality

Depends on why the lack of Normality occurred:

- Outliers → Robust Regression
- Nonnormal→ Transformation Needed
  - Can try Box-Cox transformation

Box-Cox transformation

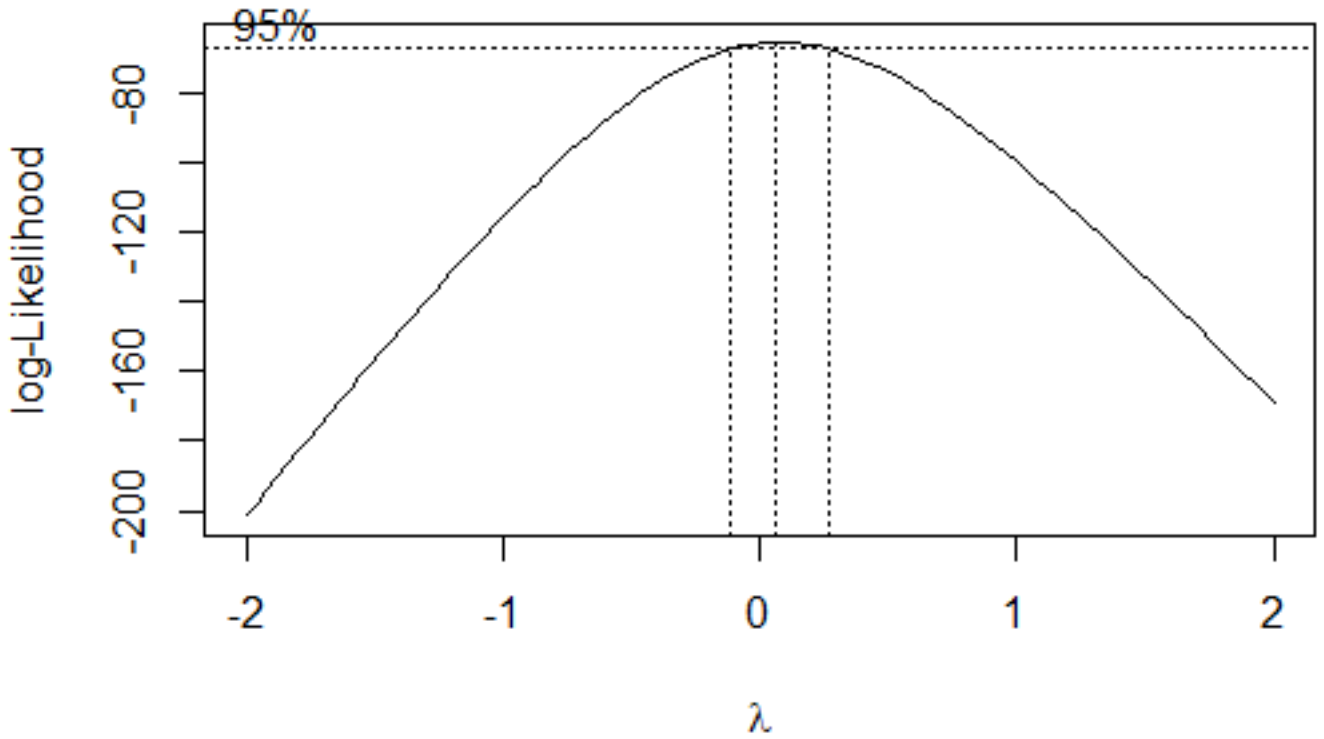
- Box-Cox (1964) developed a method to determine the best (power) transformation to induce normality.
- The Box-Cox transformation has the following form:

$$\begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

  - λ is the power in which the response variable y is raised to (so, if $\lambda = 2$, then we would square y). The exception is when $\lambda=0$ (this the log transformation).

Box-cox on original Salary data set

```
lm.var=lm(salary~years)

boxcox(lm.var)
```



Correlated Error terms

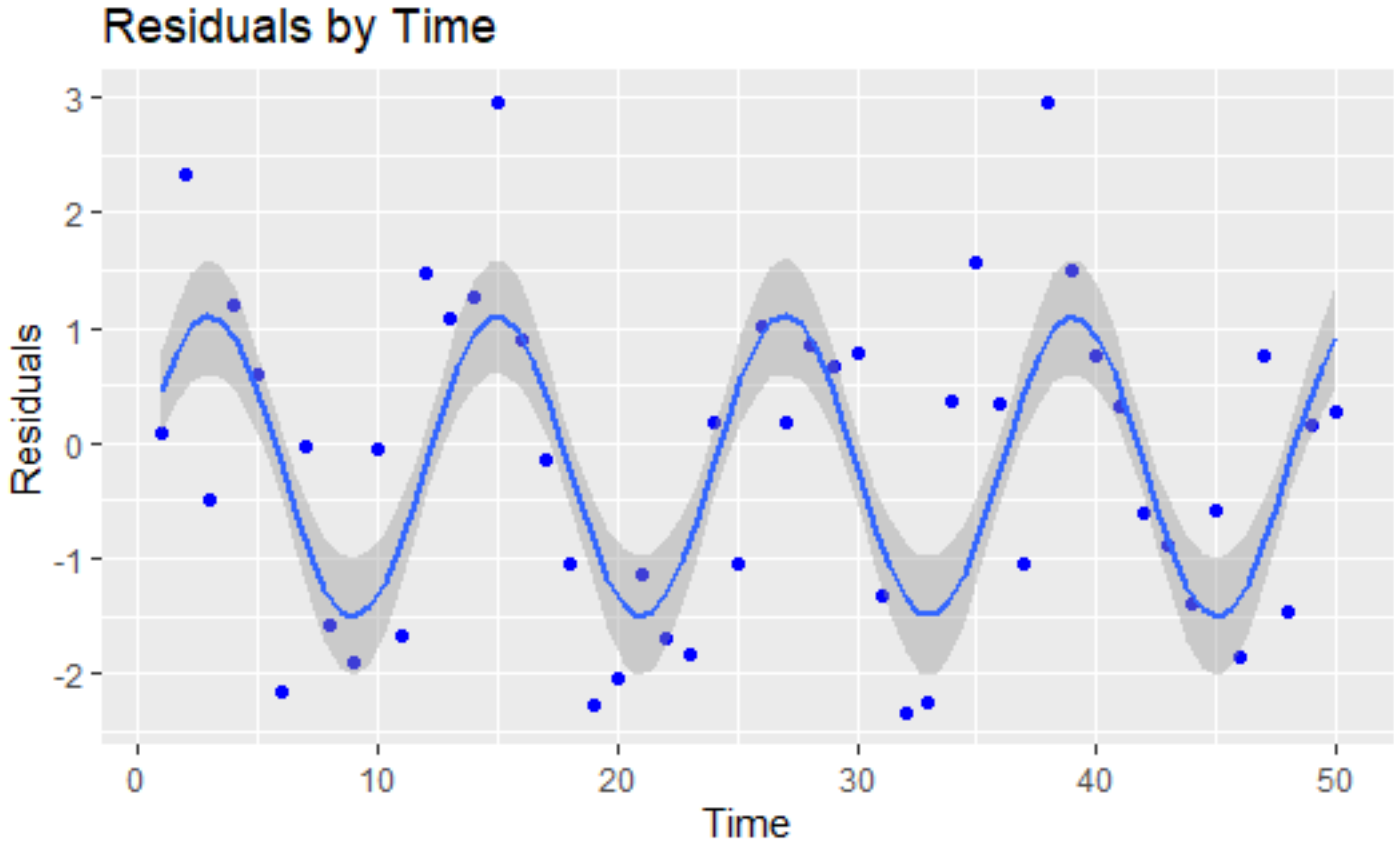
Independence

Know the source of your data:

- Clustered/Grouped data
- Observations connected in some way
- Complex survey designs
- Repeated measures
- Data gathered over time

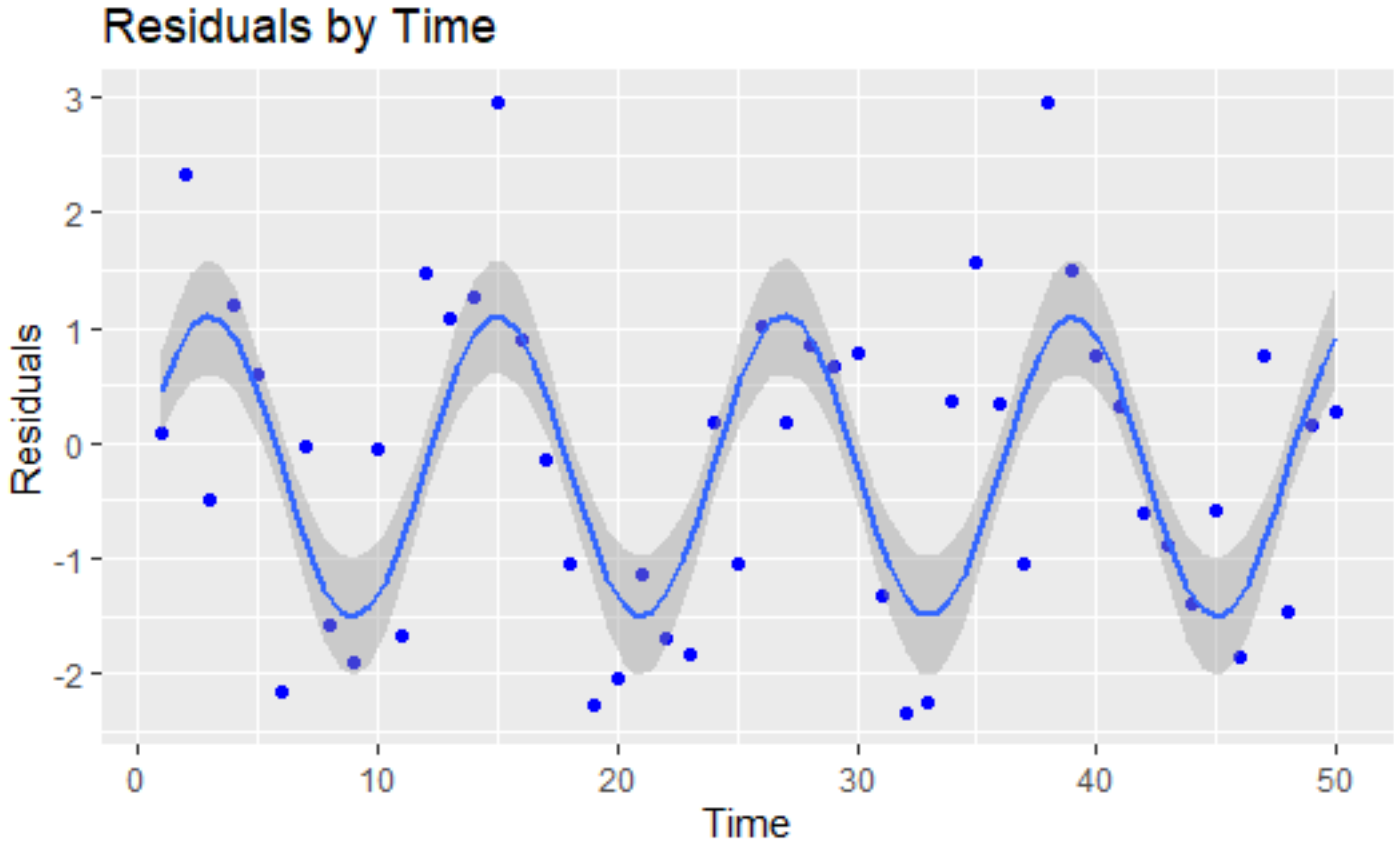
Data dependent on time

- Observations not independent.
- Residuals follow cyclic pattern.
- Collected over time.



Detect non-independence

1. Plots of residuals versus time or other ordering component



Detect non-independence

1. Plots of residuals versus time or other ordering component
2. Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

H_0: No Residual Correlation
H_A: Residual Correlation

Durbin-Watson test

Statistic:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Bounded between 0 and 4. When d=2, we fail to reject H_0 and assume there is not enough evidence supporting autocorrelation. For $d < 2$, possible positive autocorrelation (this is the one usually used). For $d > 2$, there is possible negative autocorrelation.

Google stock data

```
data(google)
x=seq(1,length(google))
lm.model=lm(google~x)
dwtest(lm.model,alternative="greater")
```

```
        Durbin-Watson test

data:  lm.model
DW = 1.842, p-value = 0.0321
alternative hypothesis: true autocorrelation is greater than 0
```

How to handle correlated error terms

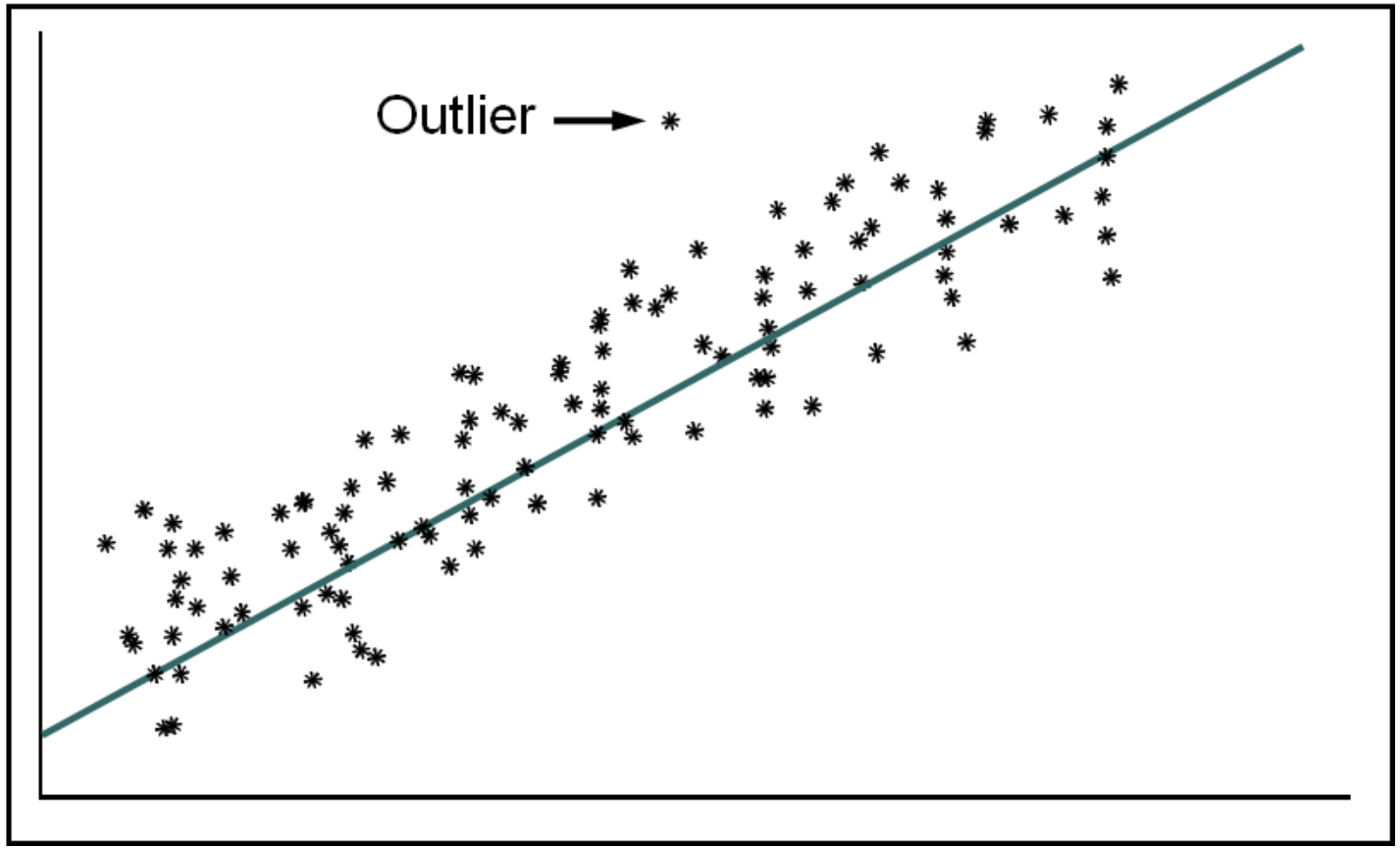
- If correlated due to time, perform time series
- If correlated due to clustered data, perform a hierarchical model
- Longitudinal analysis/panel data

Influential points and outliers

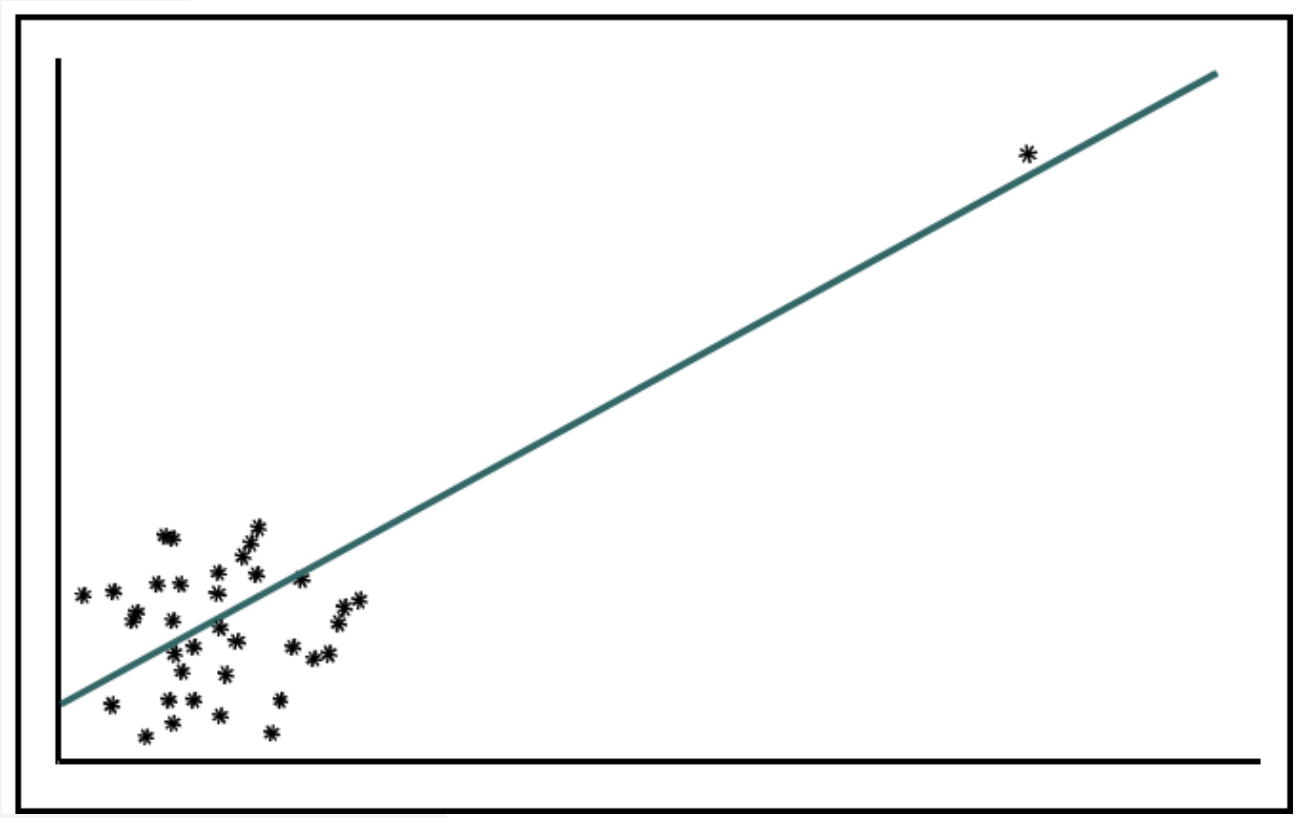
Anomalous observations

- There are two types of anomalous observations that will be discussed:
  1. Outliers – point with a large standardized residual (lie far away from the fitted line in the Y-direction).
  2. Leverage Points – point that falls outside the normal range (far from the mean) in the X-space (possible values of the predictors) and have a large “influence” on the regression line.
- Observations could be one or both of these.

Detecting Outliers

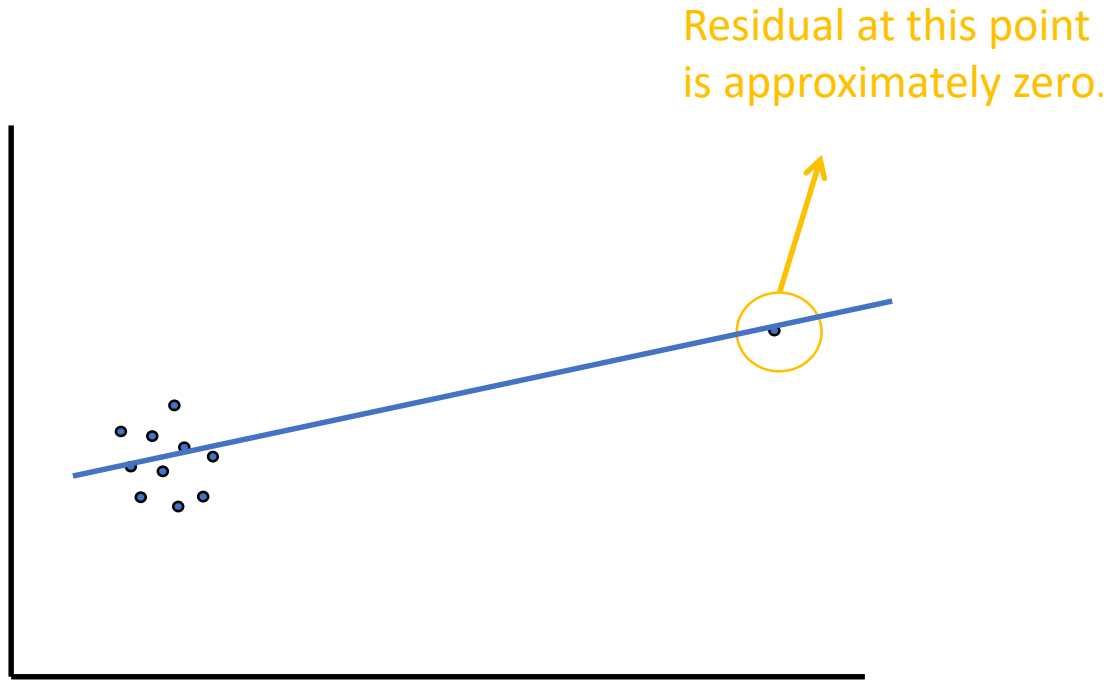


Influential Observations



Residual Analysis

- Don’t only focus efforts on residuals of data.
- Residual analysis only tends to discover outliers instead of leverage points.



Diagnostic Statistics

Statistics that help identify influential observations are the following:

- Internally Studentized residuals (good for detecting outliers)
- Externally Studentized residuals (good for detecting outliers)
- Cook's *D* (good for detecting influential observations)
- DFFITS (good for detecting influential observations)
- DFBETAS (good for detecting influential observations)
- Hat values (good for detecting influential observations)

Studentized Residuals

- Studentized residuals are obtained by dividing the residuals by their standard errors.
- Suggested cutoffs are as follows:
  - $|SR| > 2$ for data sets with a relatively small number of observations
  - $|SR| > 3$ for data sets with a relatively large number of observations

Cook’s D

Cook's distance, also referred to as Cook's D, measures the difference in the regression estimates when the i^{th} observation is left out.
A suggested cutoff is:

$$D_i > \frac{4}{n - p - 1}$$

DFFITS

- $DFFITS_i$ measures the impact that the i^{th} observation has on the predicted value.
- A suggested cutoff for influence is shown below:

$$|\mathbf{DFFITS_i}| > 2\sqrt{\frac{p}{n}}$$

Hat values

- Using matrix notation, the estimate of the parameters is:

$$b = (X'X)^{-1}X'y$$

- Which means the estimated line is:

$$\hat{y} = X(X'X)^{-1}X'y$$

- And the hat values

$$X(X'X)^{-1}X'$$

- A suggested cutoff is:

$$h_{ii} > \frac{2p}{n}$$

DFBETA

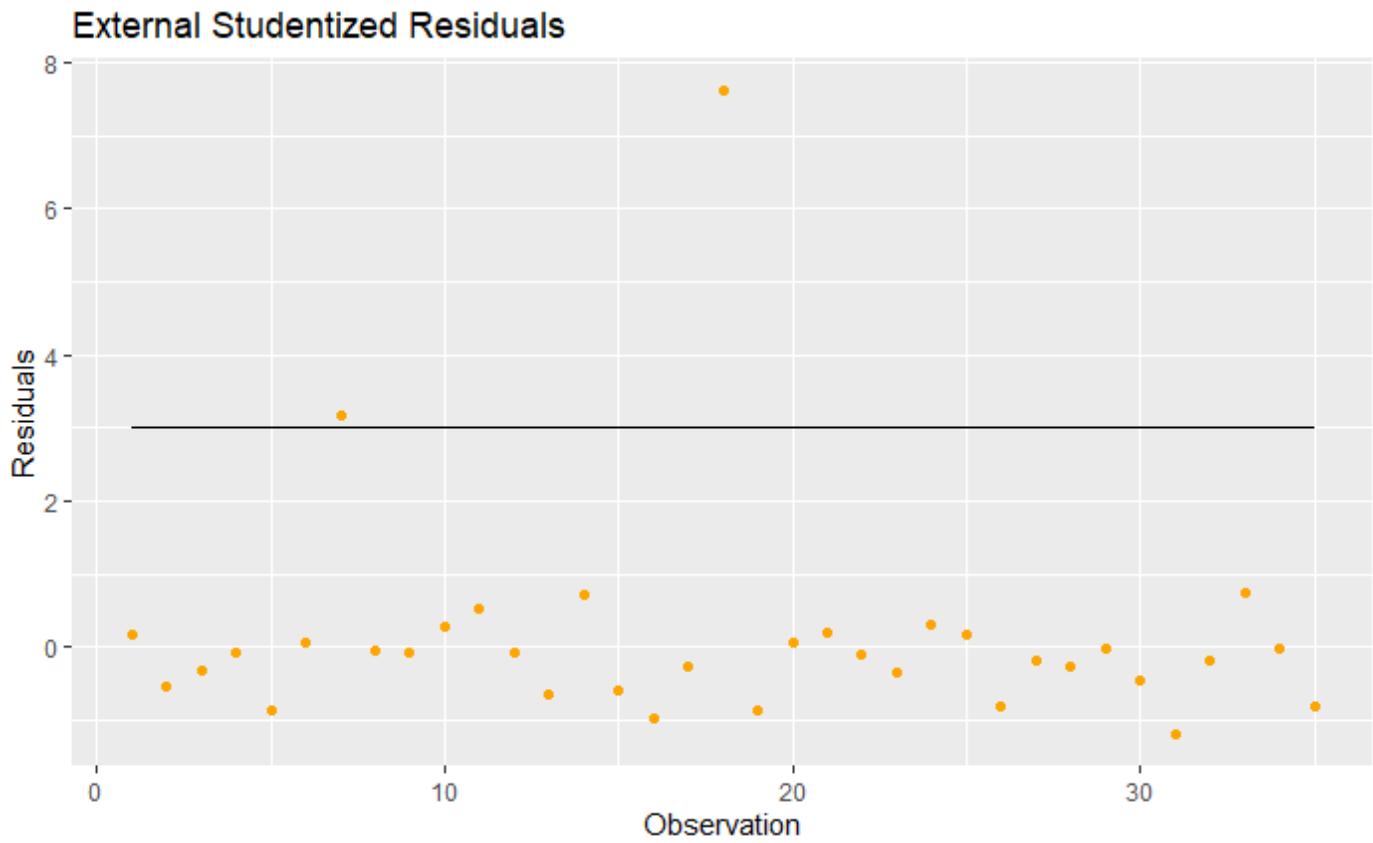
- Measure of change in the j^{th} parameter estimate with deletion of the i^{th} observation
- One DFBETA per parameter per observation
- Helpful in explaining on which parameter coefficient the influence most lies
- A suggested cutoff for influence is shown below:

$$|\mathbf{DFBETA_{ij}}| > 2\sqrt{\frac{1}{n}}$$

Scottish Hill Races

```
url =
'http://www.statsci.org/data/general/hills.txt'
races.table = read.table(url, header=TRUE,
sep='\t')
n.index=seq(1,nrow(races.table))
races.table=cbind(races.table,n.index)
lm.model=lm(Time~Distance+Climb,data=rac
es.table)
```

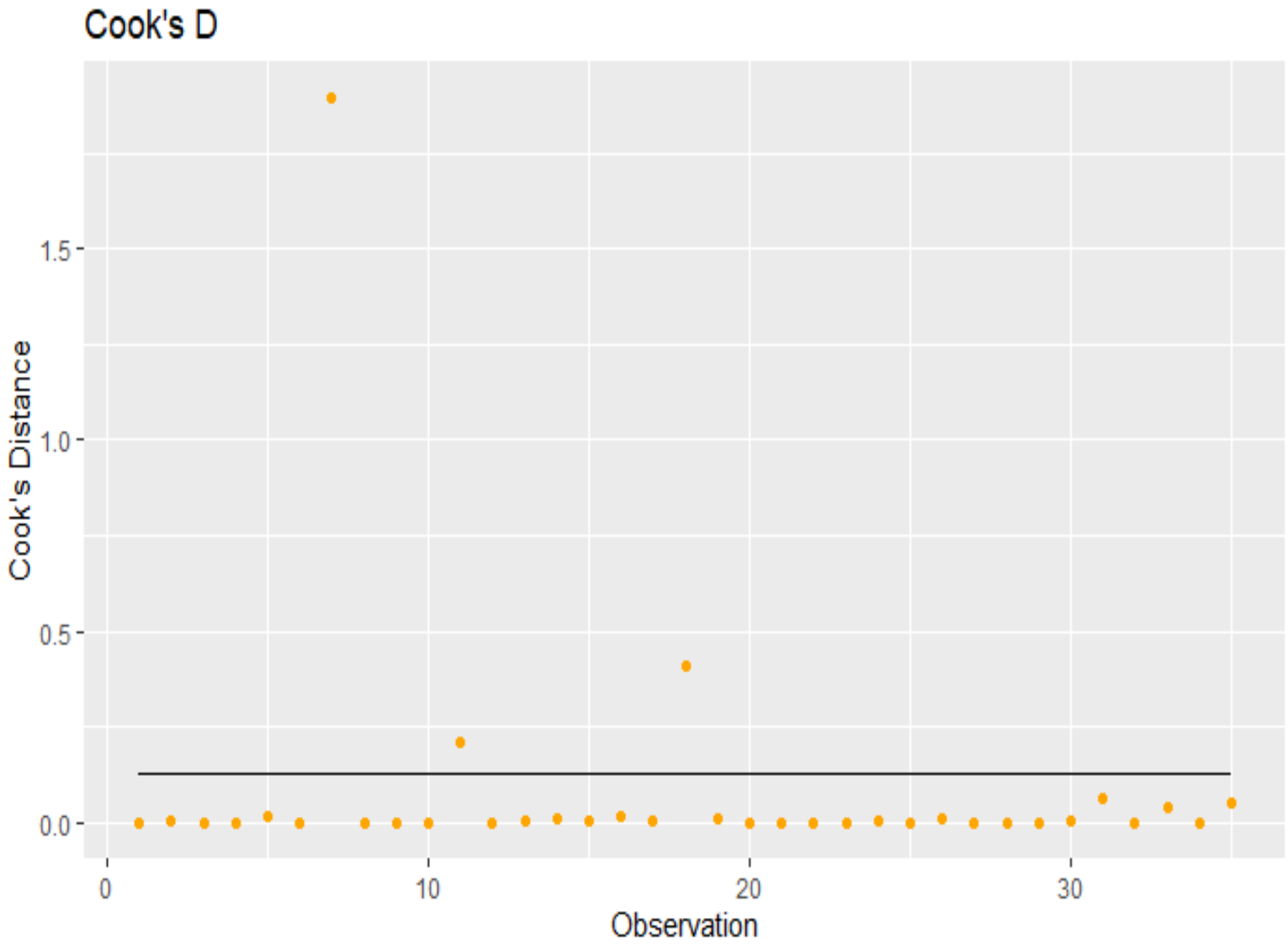
```
ggplot(lm.model,aes(x=n.index,y=rstudent(lm.mod
el)))+geom_point(color="orange")+geom_line(y=-
3)+geom_line(y=3)+labs(title = "External
Studentized
Residuals",x="Observation",y="Residuals")
```



Cook’s D

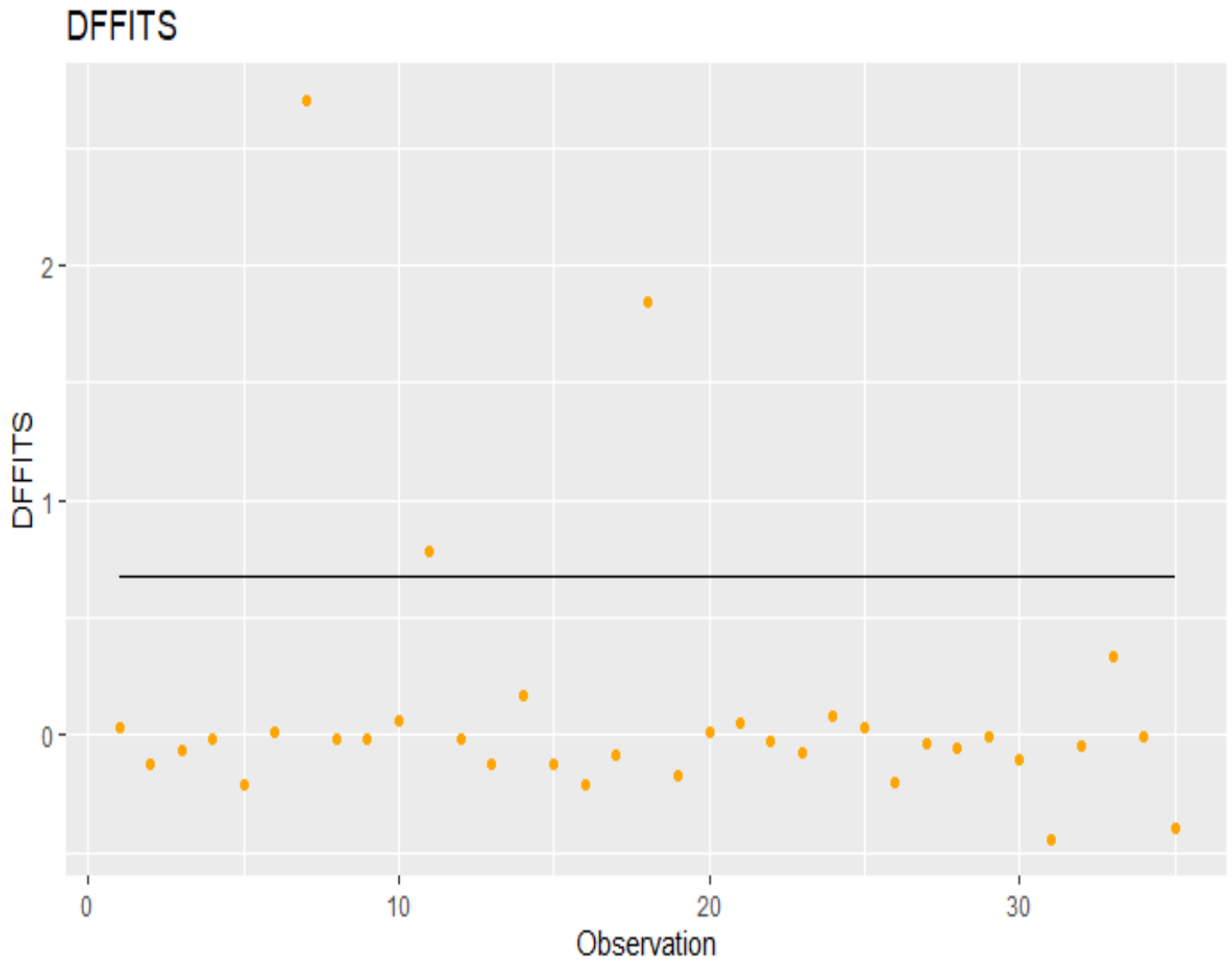
##Cook's D
D.cut=4/(nrow(races.table)-3-1)

ggplot(lm.model,aes(x=n.index,y=cooks.distance(lm.model)))+geom_point(color="orange")+geom_line(y=D.cut)+labs(title = "Cook's D",x="Observation",y="Cook's Distance")



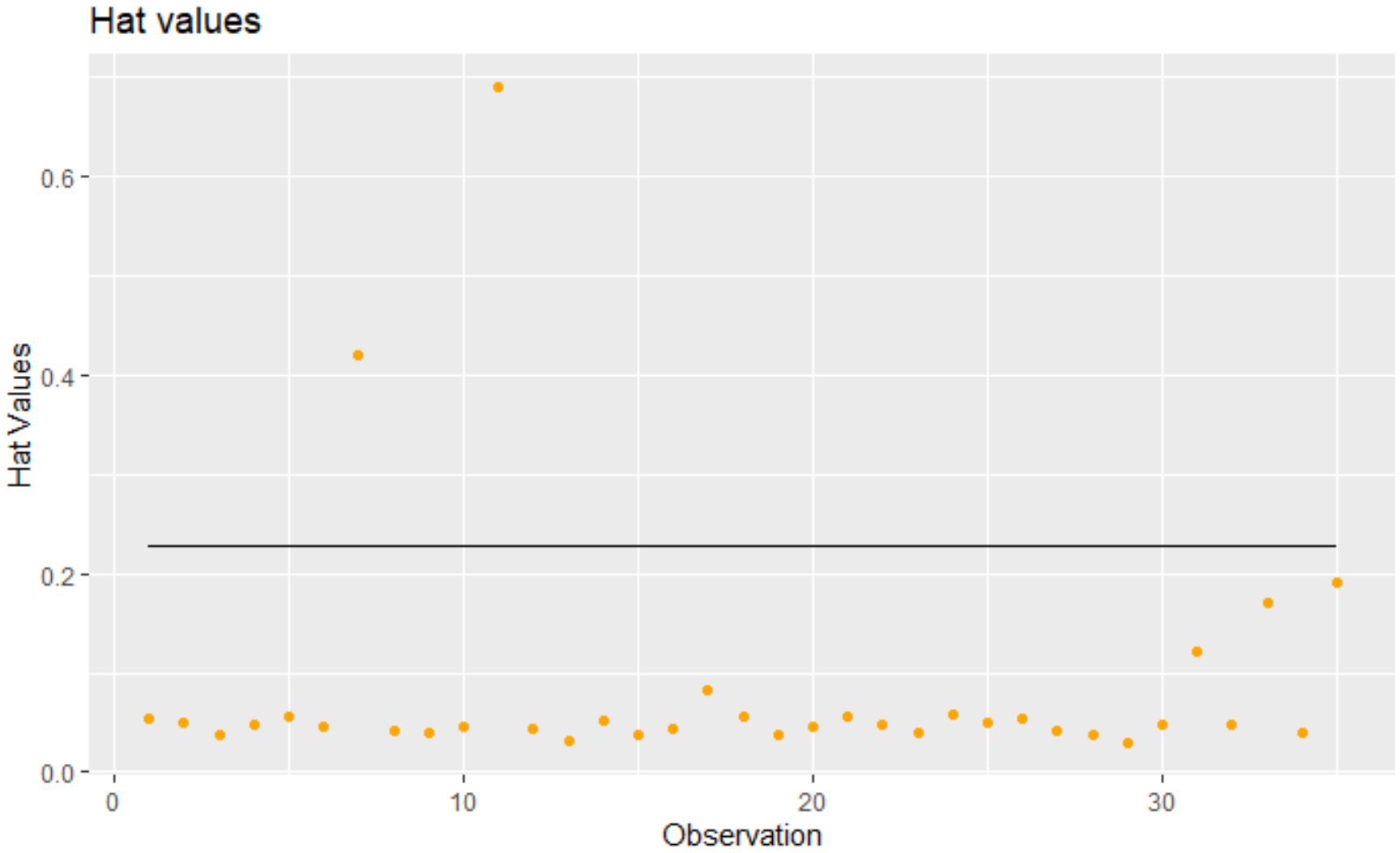
DFFITS

```
df.cut=2*(sqrt((3+1)/nrow(races.table
)))

ggplot(lm.model,aes(x=n.index,y=dffi
ts(lm.model)))+geom_point(color="or
ange")+geom_line(y=df.cut)+geom_li
ne(y=-df.cut)+labs(title =
"DFFITS",x="Observation",y="DFFITS"
)
```

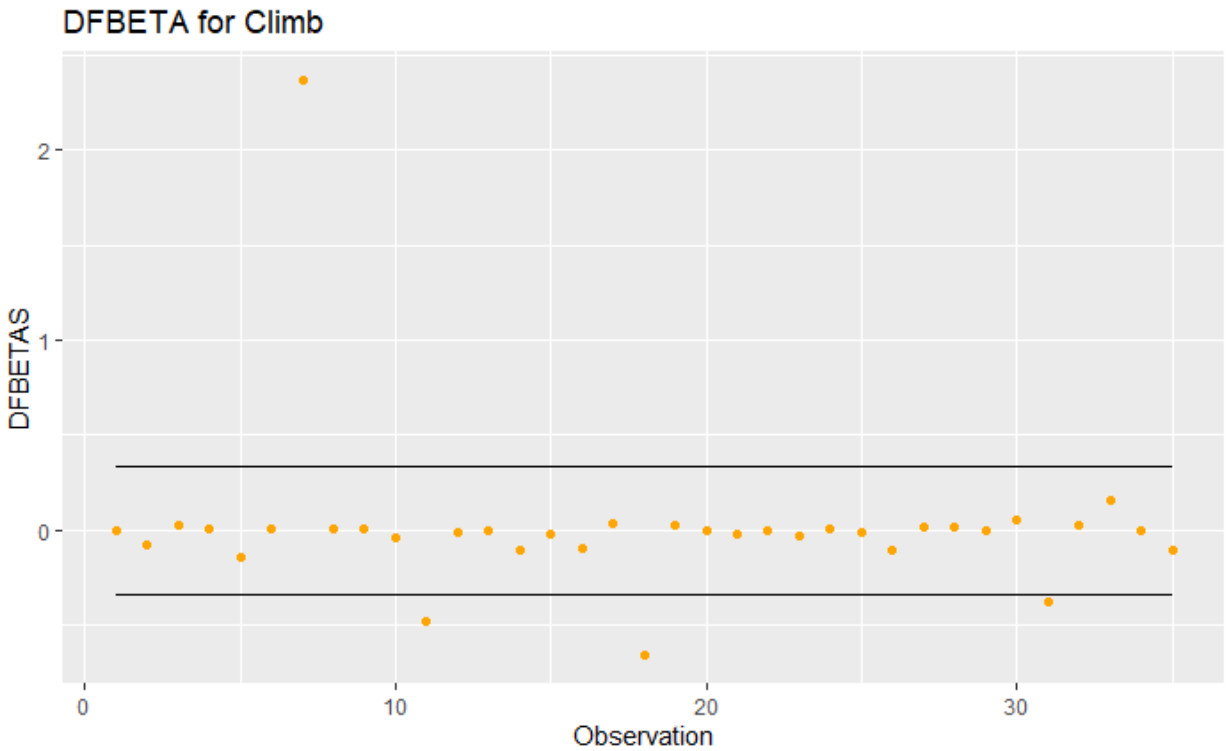


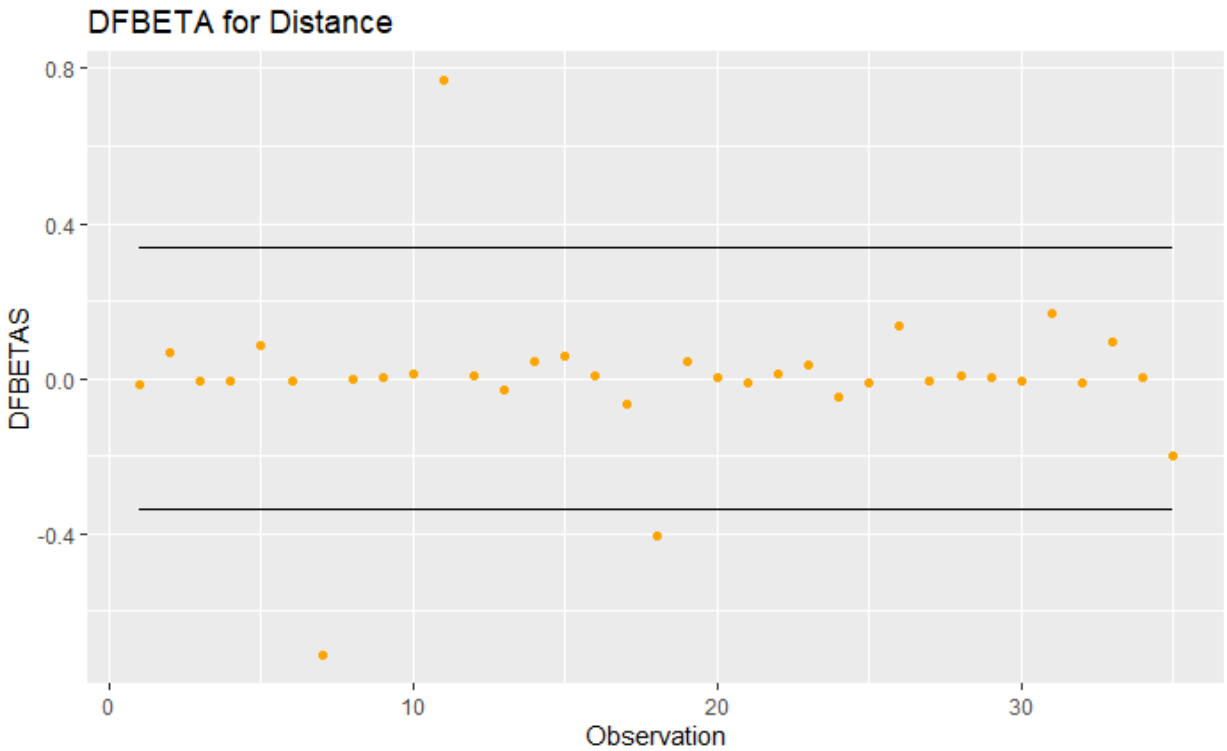
Hat values

```
hat.cut=2*(3+1)/nrow(races.table)

ggplot(lm.model,aes(x=n.index,y=hatvalues(lm.model)))+geom_point(color="orange")+geom_line(y=hat.cut)+labs(title = "Hat values",x="Observation",y="Hat Values")
```



```
db.cut=2/sqrt(nrow(races.table))


ggplot(lm.model,aes(x=n.index,y=dfbetas(lm.model)[,'Climb']))+geom_point(color="orange")+geom_l
ine(y=db.cut)+geom_line(y=-db.cut)+labs(title = "DFBETA for Climb",x="Observation",y="DFBETAS")

ggplot(lm.model,aes(x=n.index,y=dfbetas(lm.model)[,'Distance']))+geom_point(color="orange")+geo
m_line(y=db.cut)+geom_line(y=-db.cut)+labs(title = "DFBETA for
Distance",x="Observation",y="DFBETAS")
```





How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors occurred.
2. If the data is valid, one possible explanation is that the model is not adequate.
    - A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.
    - Nonlinear model
3. Determine the robustness of the inference by running the analysis both with and without the influential observations.
4. Robust Regression (Covered Later in Program)
5. Weighted Least Squares (WLS)

Collinearity

Illustration of Collinearity

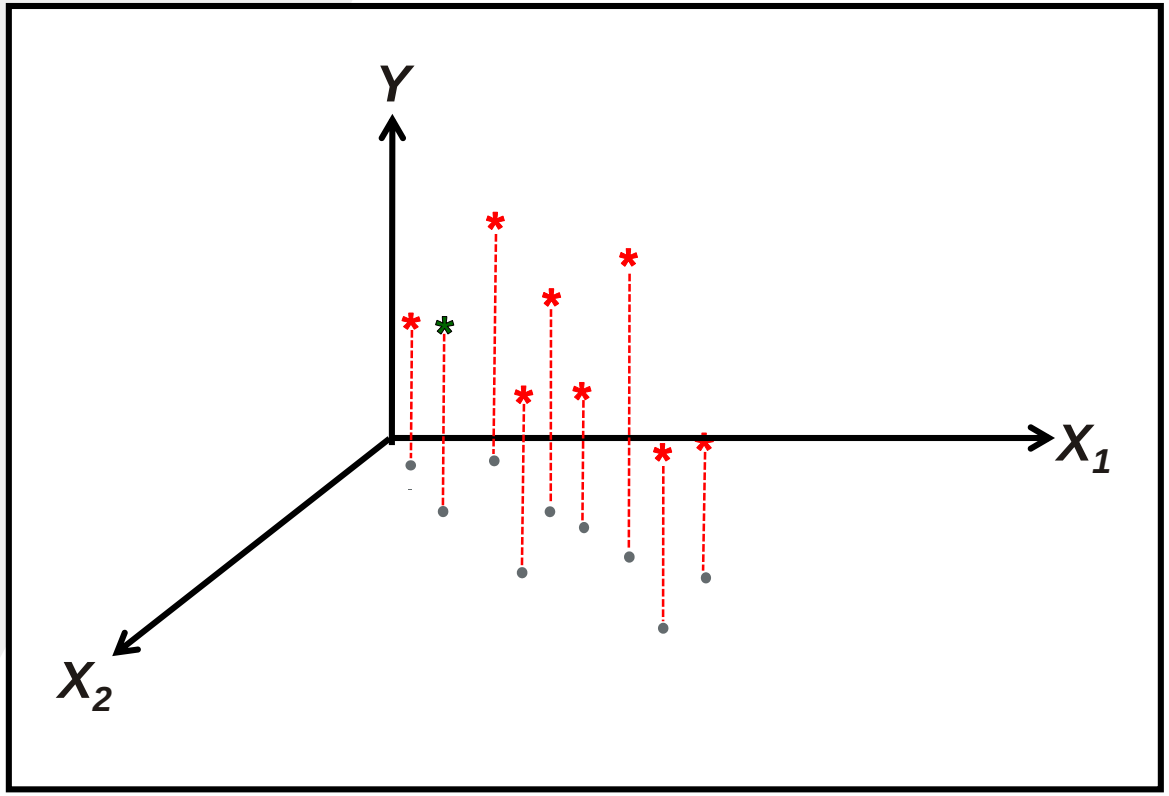


Illustration of Collinearity

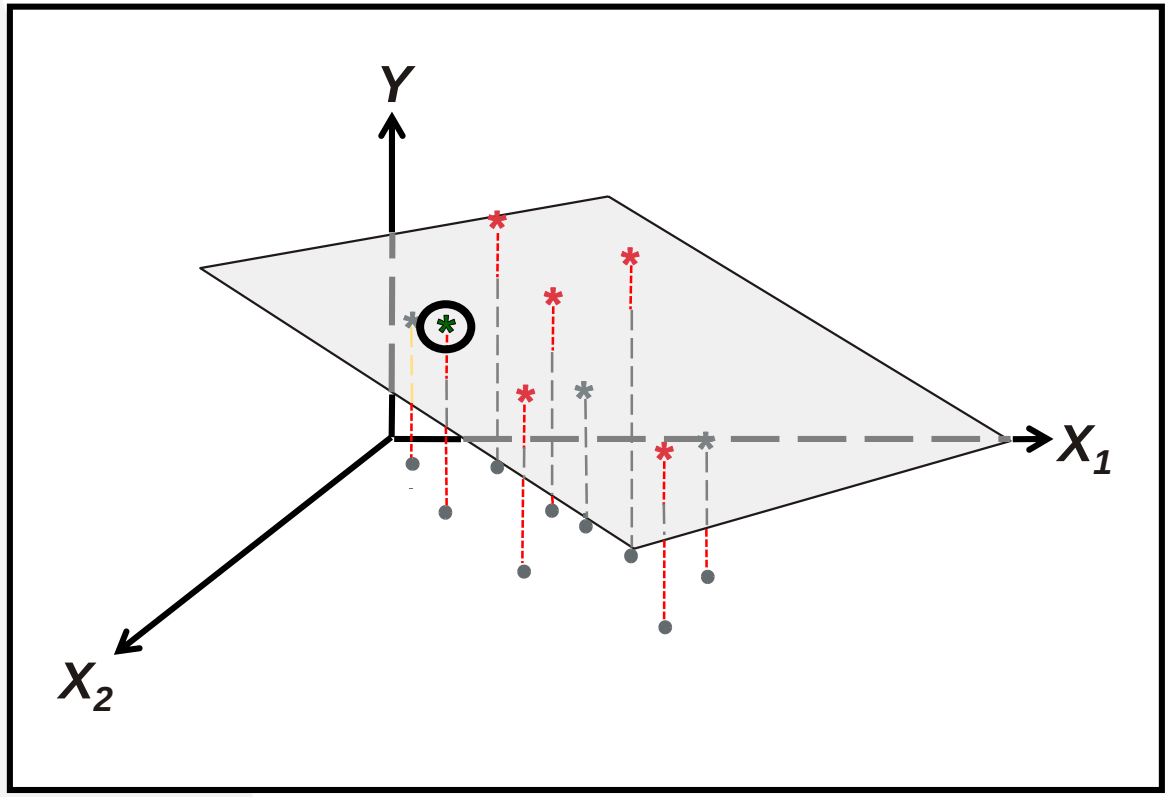
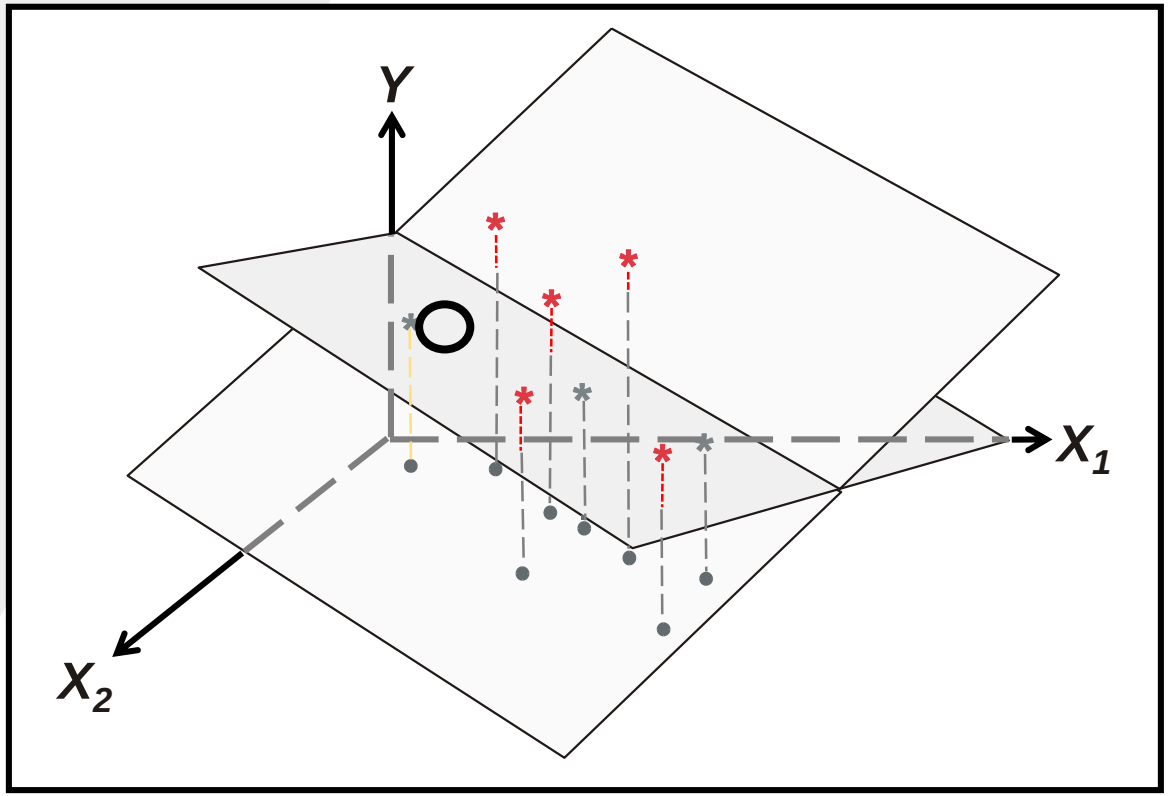


Illustration of Collinearity



Collinearity Diagnostics

- Looking at correlation matrix of predictors
- One of the most commonly used measures is the variance inflation factor (VIF).
- VIF is calculated by

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Values of VIF greater than 10 indicate potential collinearity

MTCARS data set

```
cor(mtcars)
lm.model=lm(mpg~.,data=mtcars)
vif(lm.model)
```

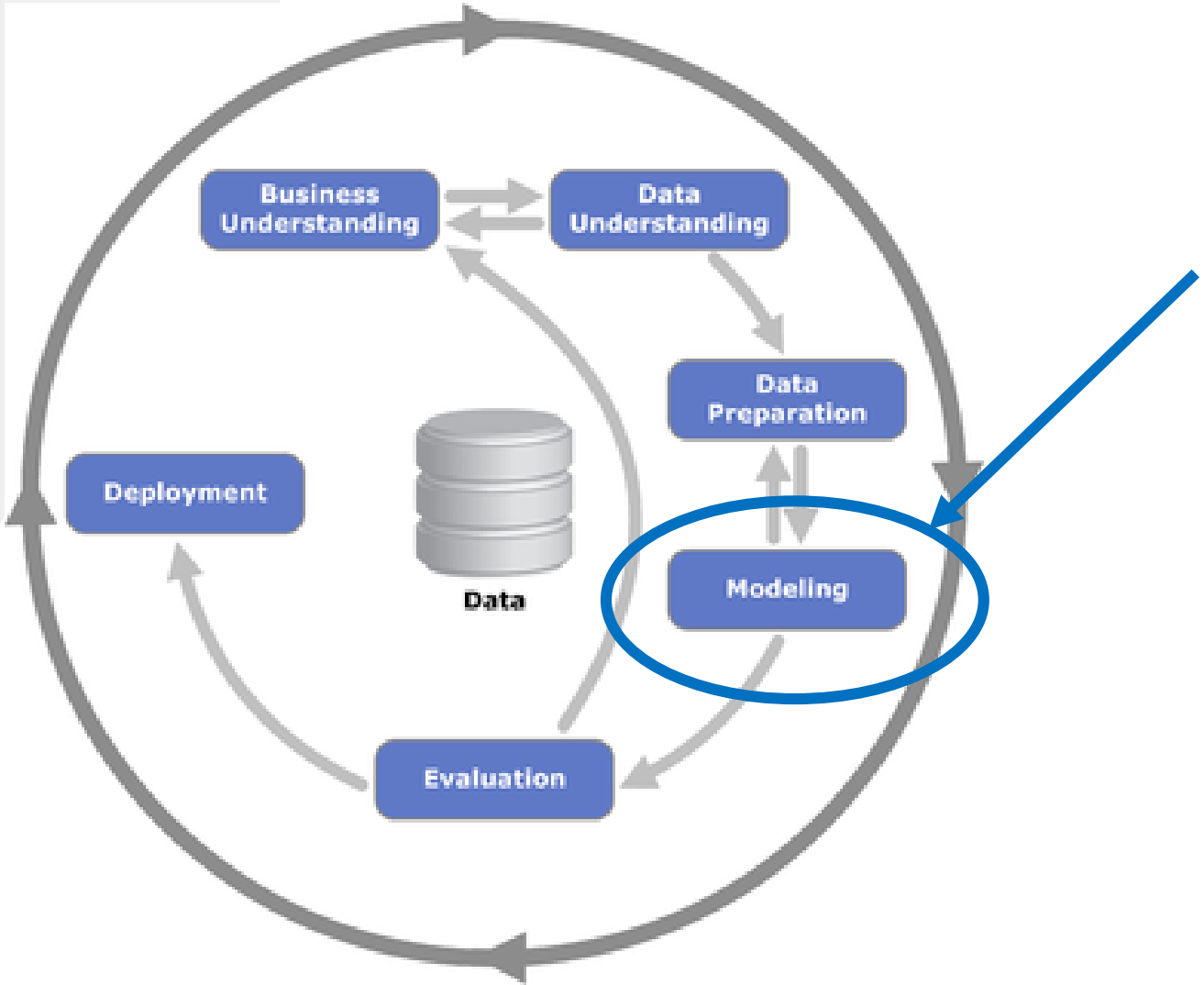
```
     cyl     disp       hp     drat       wt     qsec
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958
      vs       am     gear     carb
 4.965873  4.648487  5.357452  7.908747
```

Dealing with Multicollinearity

- Exclude redundant independent variables.
- Redefine variables.
- Use biased regression techniques (for example, LASSO).
- Center the independent variables in polynomial regression models.

CRISP-DM

Cross-industry standard process for data mining



An Effective Modeling Cycle