Linear Regression

BC2406 UNIT 6

Instructor

Hyeokkoo Eric Kwon

Nanyang Technological University

Objective

- To Develop an Analytics Model that can predict a Continuous Target Variable Y
- To learn Diagnostic Checks on Linear Regression Model
- To Learn How to Detect Multi-Collinearity Problem

Content

- From Association to Prediction
  - Correlation Coefficient (r)
- Linear Regression Model
- Diagnostics Checks
- Complications

mtcars dataset

A standard dataset in R with 32 observations on 11 variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S (0 = V-shaped engine, 1 = straight engine)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

Correlation as a measure of Association between 2 numerical variables

```
cor(mtcars$mpg, mtcars$wt)
## -0.8676594

cor(mtcars$mpg, mtcars$hp)
## -0.7761684

cor(mtcars$mpg, mtcars$qsec)
## 0.418684

cor(mtcars$drat, mtcars$qsec)
## 0.09120476

cor(mtcars$hp, mtcars$cyl)
## 0.8324475
```

- What is the meaning of correlation?

Note: $-1 \leq r \leq 1$

Question

- If r is close to 1 or -1, does this mean X cause Y?
- Poll:
  - Yes, X cause Y:
  - No, X does not cause Y:
  - Still thinking…:

Answer

- If r is close to 1 or -1, does this mean X cause Y?
- Ans: **Not necessarily**.
- Examples:
  - X = Number of ice creams sold, Y = Deaths from Drowning.
  - X = Number of Police Officers Hired, Y = Crime Rate.
  - X = Food Intake (Calories), Y = Weight.

- Correlation ≠ Causation

High Values of r (regardless of sign)

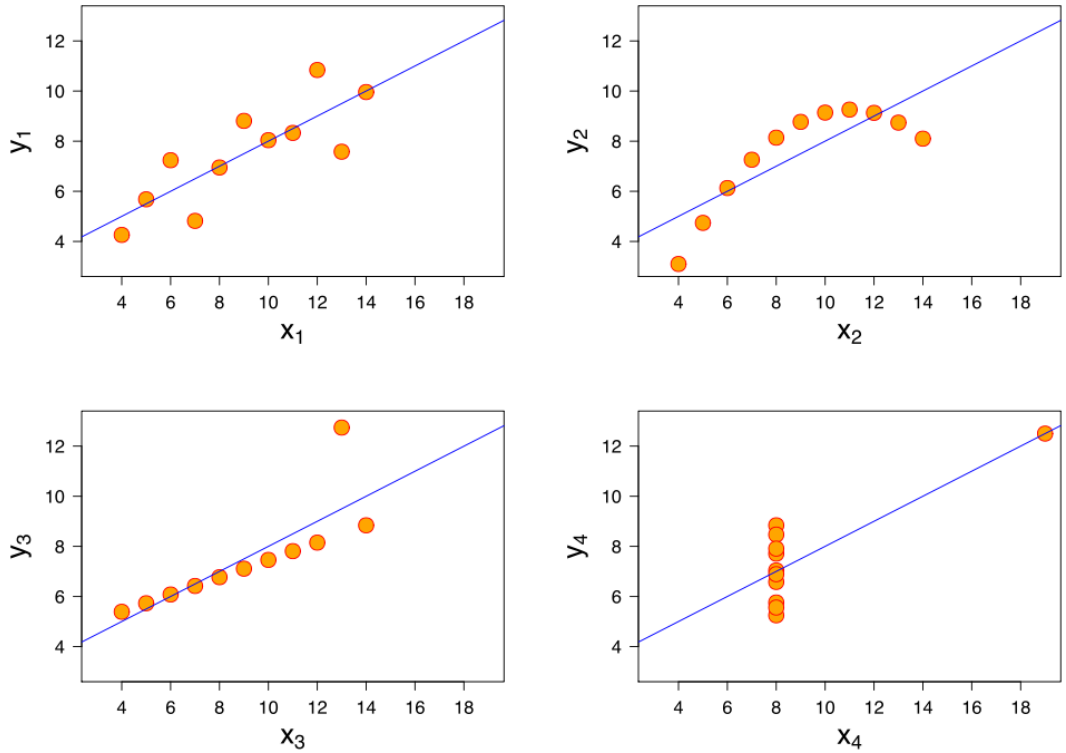
- Suggests a **statistical relationship** that can be exploited for predicting Y using X, even if
  - X does not cause Y
  - i.e. just based on association
- If X really cause Y, then very high confident in predictions of Y. How to "prove" causation?
  - Design of Experiments
  - Clinical Trials
  - Specify the mechanism of action that shows how X cause Y.
- How to prove that Temperature cause Stock Price Fluctuation?
  - Some variables are beyond one's control.
  - Satisfied with strong associations, at least for the time being.

Question

- If r is close to 1 or -1, does this mean X and Y has a linear association?
- Poll:
  - Yes, linear association:
  - No linear association:
  - Still thinking…:

Answer

- If r is close to 1 or -1, does this mean X and Y has a linear association?
- Ans: Maybe.



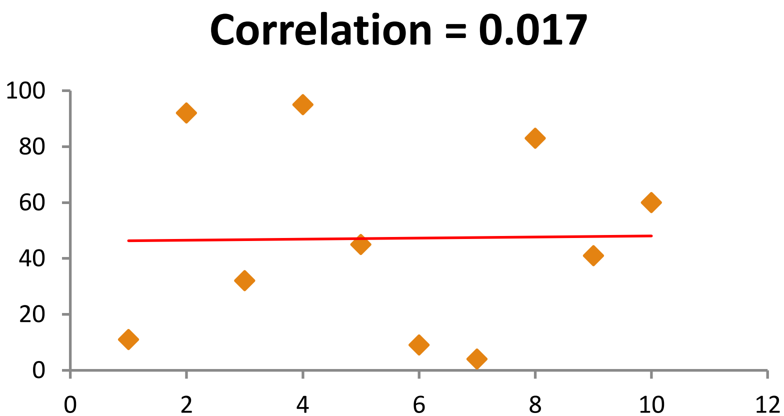
r = 0.816 in each of the charts.

Question

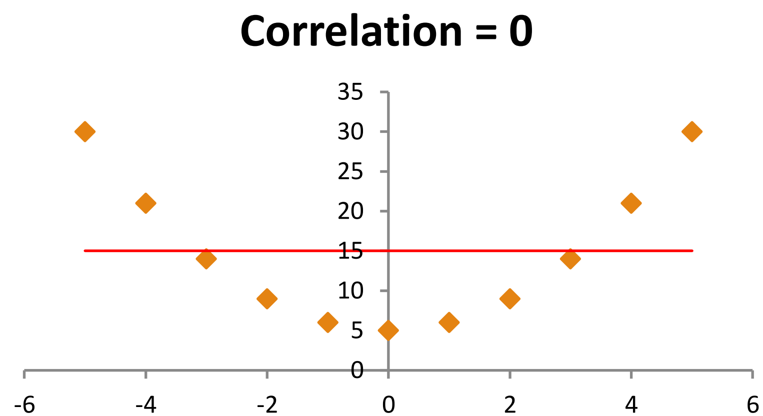
- If r is close to 0, does this mean X and Y has no association?
- Poll:
  - Yes, no association:
  - No, there is association:
  - Still thinking…:

Answer

- If r is close to 0, does this mean X and Y has no association?
- Ans: Maybe.



No association between X and Y.



Clearly a Quadratic association between X and Y.

Conclusions on the Interpretation of r

- If r is close to 1 or -1, then:
  - X is **associated with** Y.
  - **May or may not be linear association** (but regardless, good candidate to be considered for inclusion in analytics model.)
  - Confirm with **Scatterplot** of Y vs X.

- If r is close to 0, then:
  - X definitely **does not have a linear association** with Y.
  - May have **no association** or have **non-linear association**.
  - Confirm with **Scatterplot** of Y vs X.

Correlation is a precise number. What exactly is correlation trying to measure?

Ans: Consistency of the Trend (if any).

Highly Consistent Trend, High |r|:

- Data points all falling close to a straight line.
- Data points all falling close to a rising curve.
- Data points all falling close to a falling curve.

Inconsistent Trend, Low |r|:

- Data Points randomly distributed.
- 50% data points rising trend, 50% data points falling trend.

More than one X can be associated with Y

- Regardless of causal relationship or just association:
  - Y = Weight of a Person
  - X1 = Food Intake (Calories)
  - X2 = Age
  - X3 = Gender
  - X4 = Number of Times to Buffet per month
  - X5 = Metabolic Rate
  - X6 = Amount of Physical Activity per week
  - X7 = Amount of Fresh Fruits consumed
  - X8 = Weight of Mother
  - X9 = Weight of Father
- How do we include/test all of these Xs in "predicting" Y?
  - Correlation is not enough.
  - Use analytics models

Linear Regression Model

Linear Regression Equation

$$\hat{y} = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\text{Straight Line Equation}}$$

The straight line equation is only 50% of the Linear Regression model.

Linear Regression Model

$$\underbrace{y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\hat{y}} + e$$

$\hat{y}$

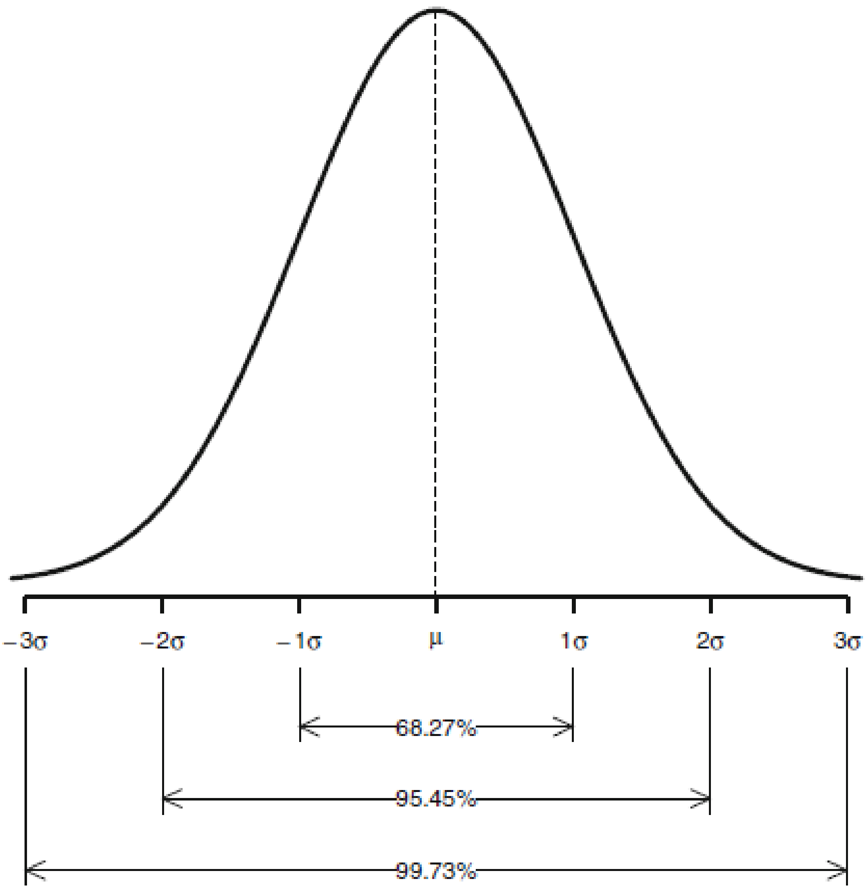
Straight Line Equation

e ~ N(0, σ)

Errors (aka Residuals) follow a Normal Distribution with mean 0 and constant standard deviation.

$$y = \hat{y} + \mathrm{e}$$
$$y - \hat{y} = e$$

Normal Distribution: X ~ N(μ, σ)



μ: Mean controls centre of the bell curve.

σ: Standard Deviation (sigma) controls fatness of the bell curve.

Curve generated by a mathematical function:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Area under the curve = 1

Linear Regression Model Assumptions

$$y = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\hat{y}} + e$$

e ~ N(0, σ)

From the equation above, can you write down the assumptions in words?

Linear Reg Model Assumptions in Words

1. Linear Association between Y and Xs.
2. Errors has a normal distribution with mean 0.
3. Errors are independent of X and has constant standard deviation.

Interpretation of the Linear Regression Line

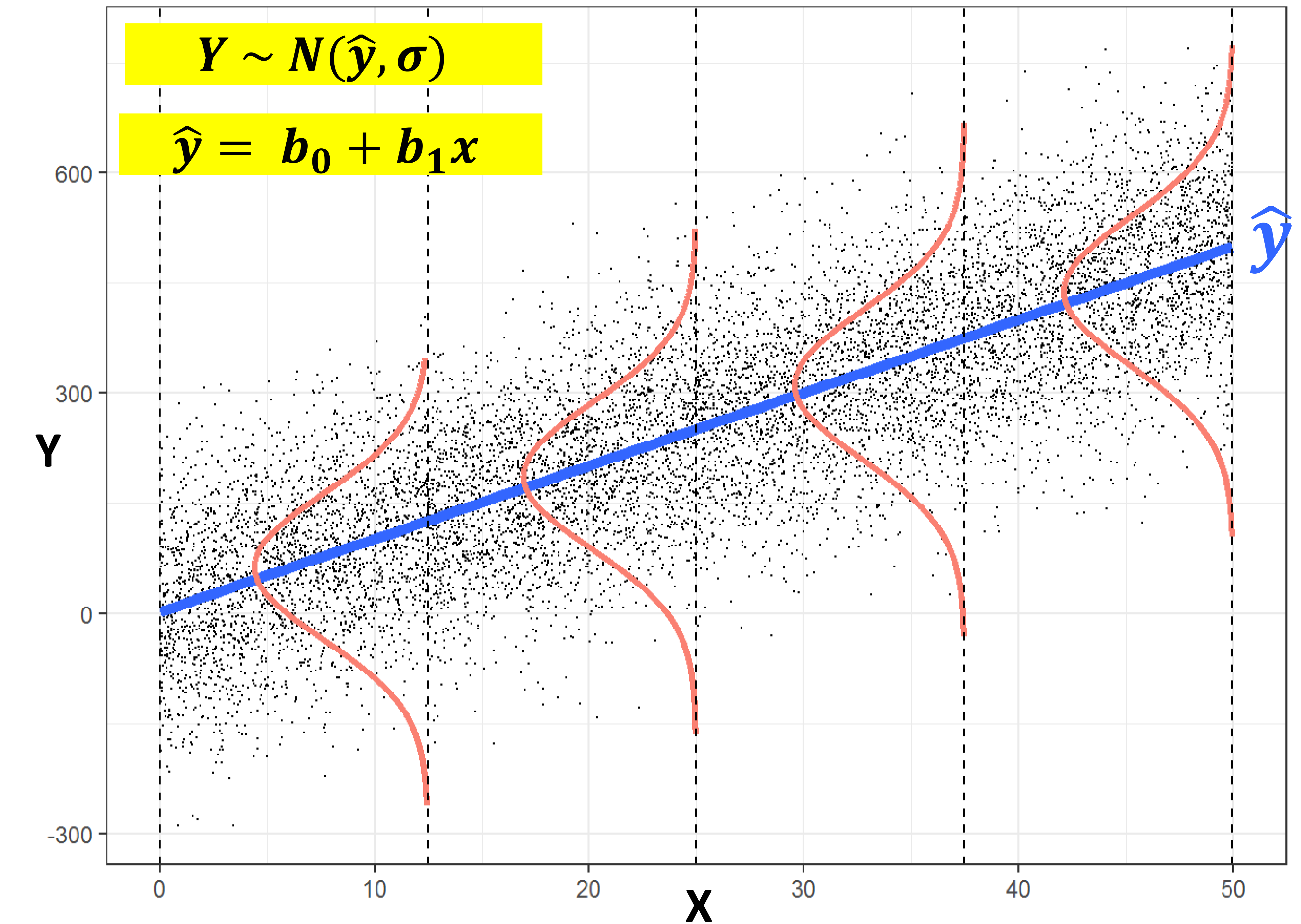
$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m + e$$

$$y = \hat{y} + e\,, \qquad e \sim N(0, \sigma)$$

$$\longrightarrow \quad Y \sim N(\hat{y} + 0\,, \sigma)$$

$$\longrightarrow \quad Y \sim N(\hat{y}, \sigma)$$

The straight line $\hat{y}$, represents the mean value of Y (that has a normal distribution) at each value of Xs.



$Y \sim N(\hat{y}, \sigma)$
$\hat{y} = b_0 + b_1 x$
600
300
0
-300
Y
0
10
20
30
40
50
X
$\hat{y}$

Getting the Regression Equation using R

- Given a dataset, first identify the outcome variable (Y) that you want to predict or estimate.
- Ensure that the Y variable is continuous
- Identify a list of potential X variables that may have an effect on Y.
  - If X is categorical, ensure that X data type is "factor" so that R will auto-generate dummy variables behind-the-scene.
- Use lm() function in Base R to create the linear reg object
- Use summary() to view the results:
  - Model Coefficients are the slope of each X in the model
  - P-value < 5% for statistically significant X variable
  - Adj R Squared for overall goodness of fit of the line to data.
- Do diagnostic checks with plot() function.

Refer to "ADA1-6-1 linreg.R" Rscript

Results of m1

```
> m1 <- lm(mpg ~ wt, data = mtcars)
> summary(m1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,	Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

$$mpg = 37.285 - 5.344 \times wt + e$$

Results of m1

```
> m1 <- lm(mpg ~ wt, data = mtcars)
> summary(m1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 de        of freedom
Multiple R-squared:  0.7528,    Adjuste        squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-va      : 1.294e-10
```

a% risk of concluding that a relationship exists when there is no actual relationship

Default: 5% or 0.05 as cut-off point for p-value

Results of m1

```
> m1 <- lm(mpg ~ wt, data = mtcars)
> summary(m1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Results of m1

```
> m1 <- lm(mpg ~ wt, data = mtcars)
> summary(m1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,	Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

$$mpg = 37.285 - \mathbf{5.344} \times wt + e$$

Results of m1

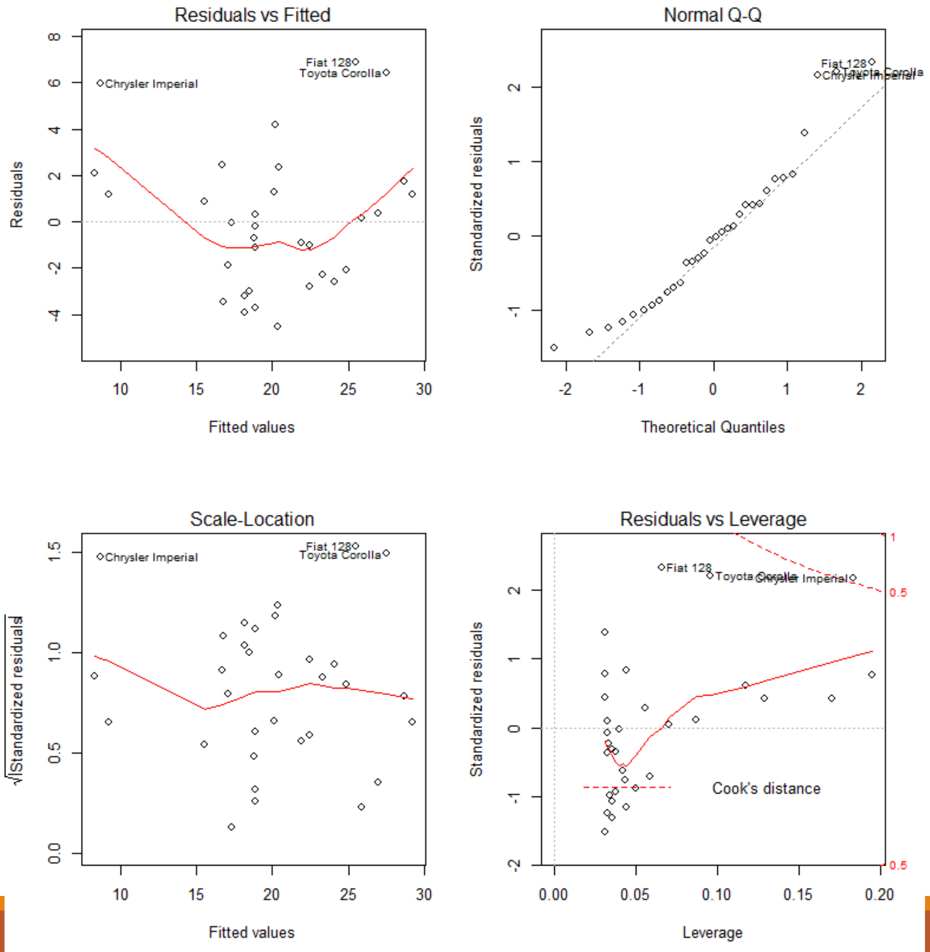
```
> m1 <- lm(mpg ~ wt, data = mtcars)
> summary(m1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,	Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

R-squared represents the Explanation power of the model.

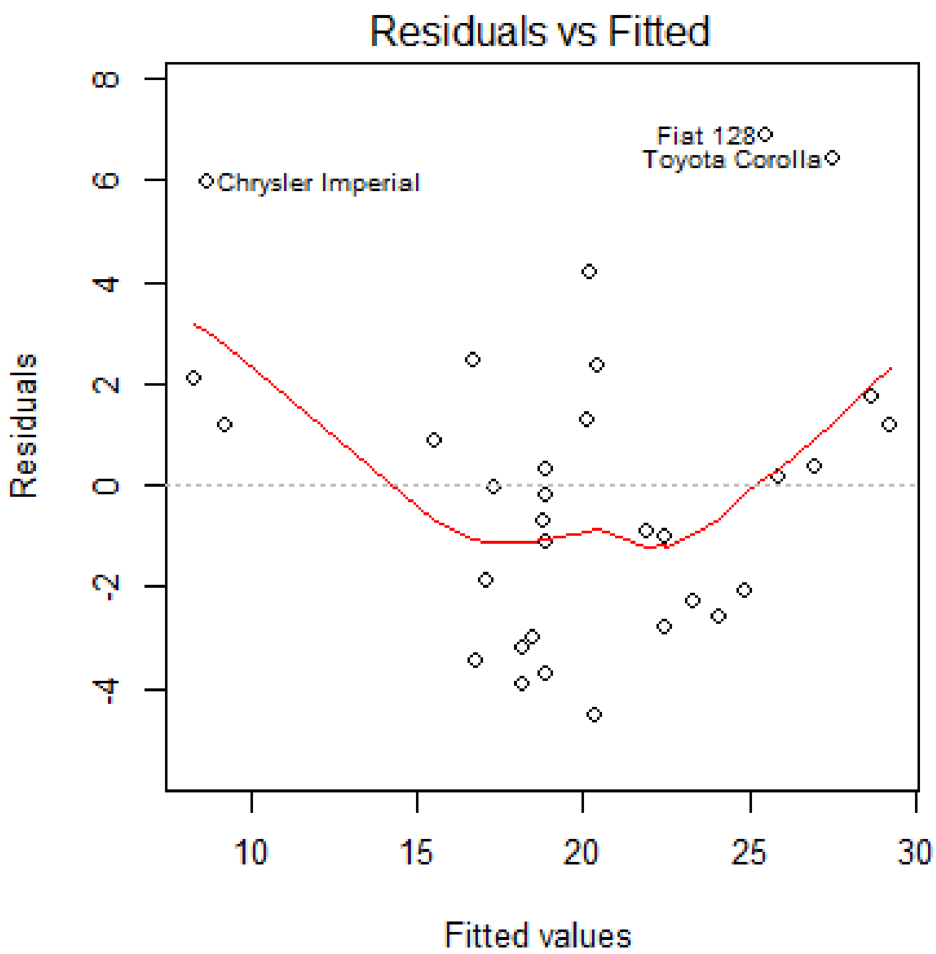
Adjusted R-squared gives a penalty to every additional X variable.

Model Diagnostic Plots

```
> par(mfrow = c(2,2))
> plot(m4)
```



Top Left Chart



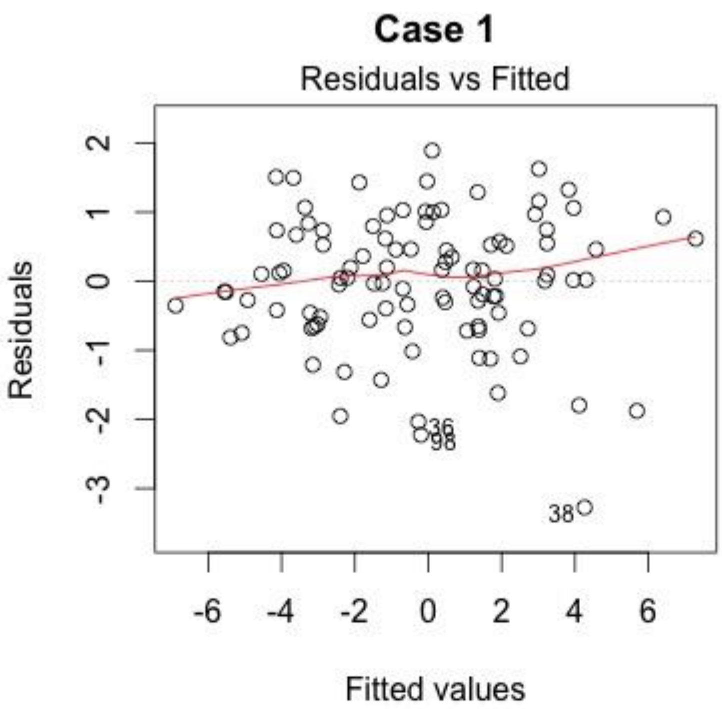
To test

Assumption 1 = **Linear** Association between Y and Xs.

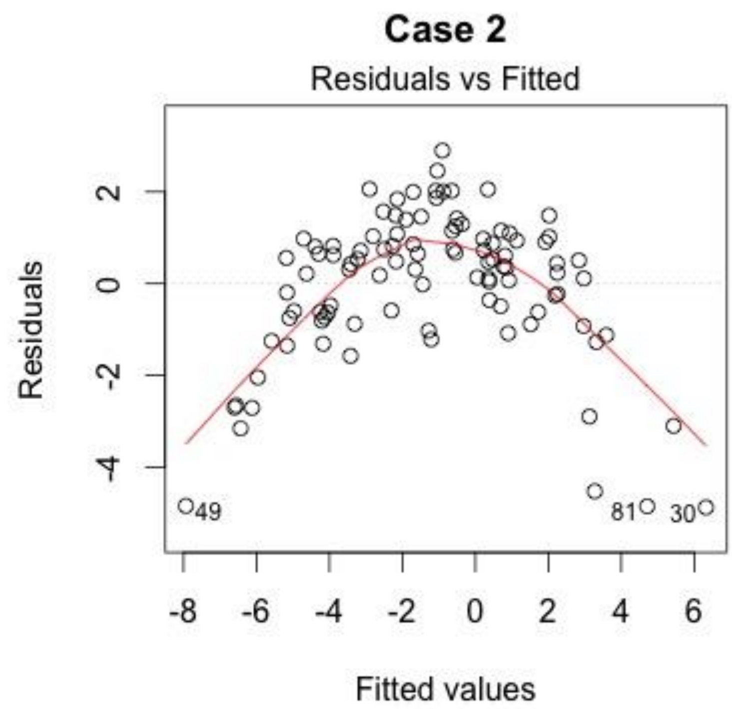
Assumption 2 = Errors has a normal distribution with **mean 0.**

Top Left Chart (Examples)

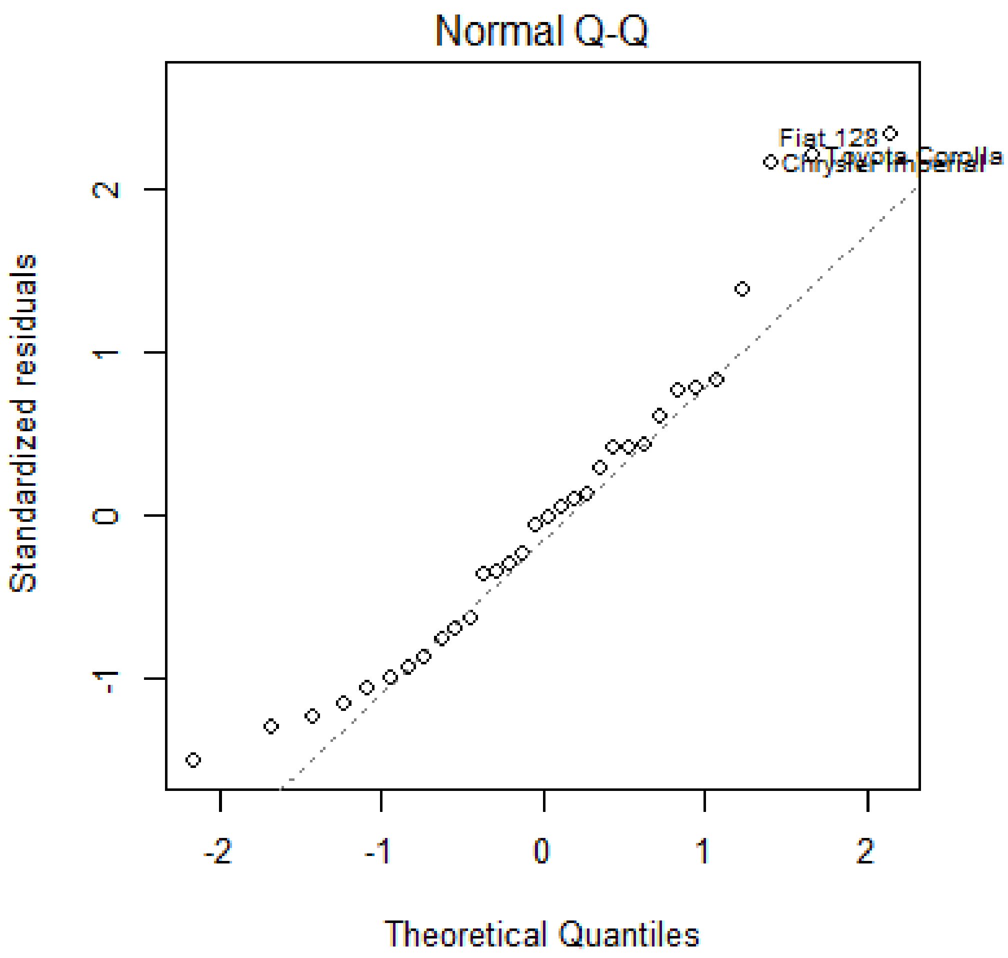
Better Case



Worse Case



Top Right Chart (Q-Q plot)

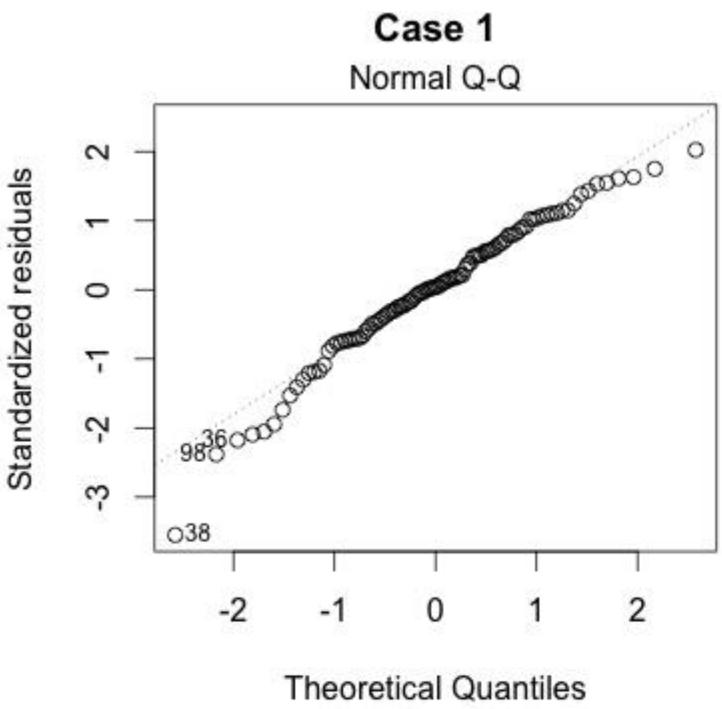


<u>To test</u>

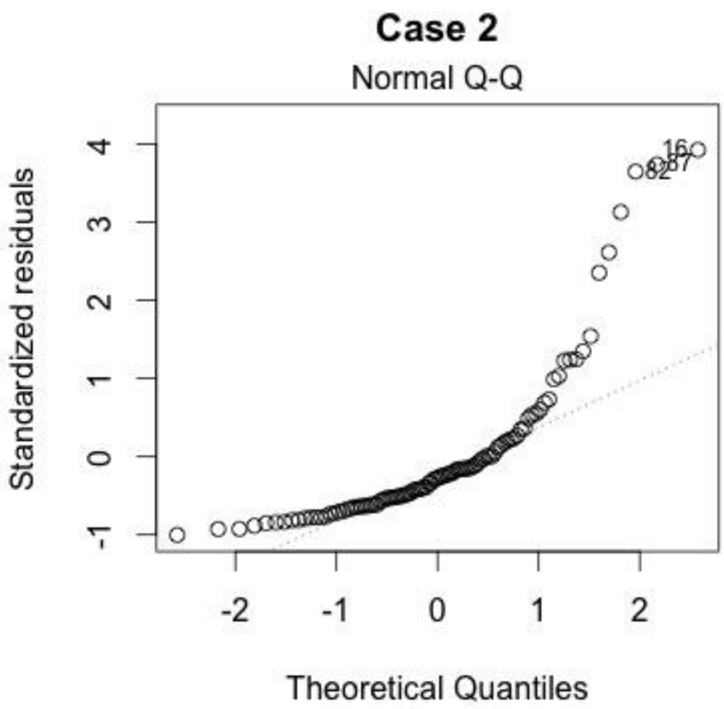
Assumption 2 = Errors has a **normal distribution** with mean 0.

Top Right Chart (Q-Q plot) (Examples)

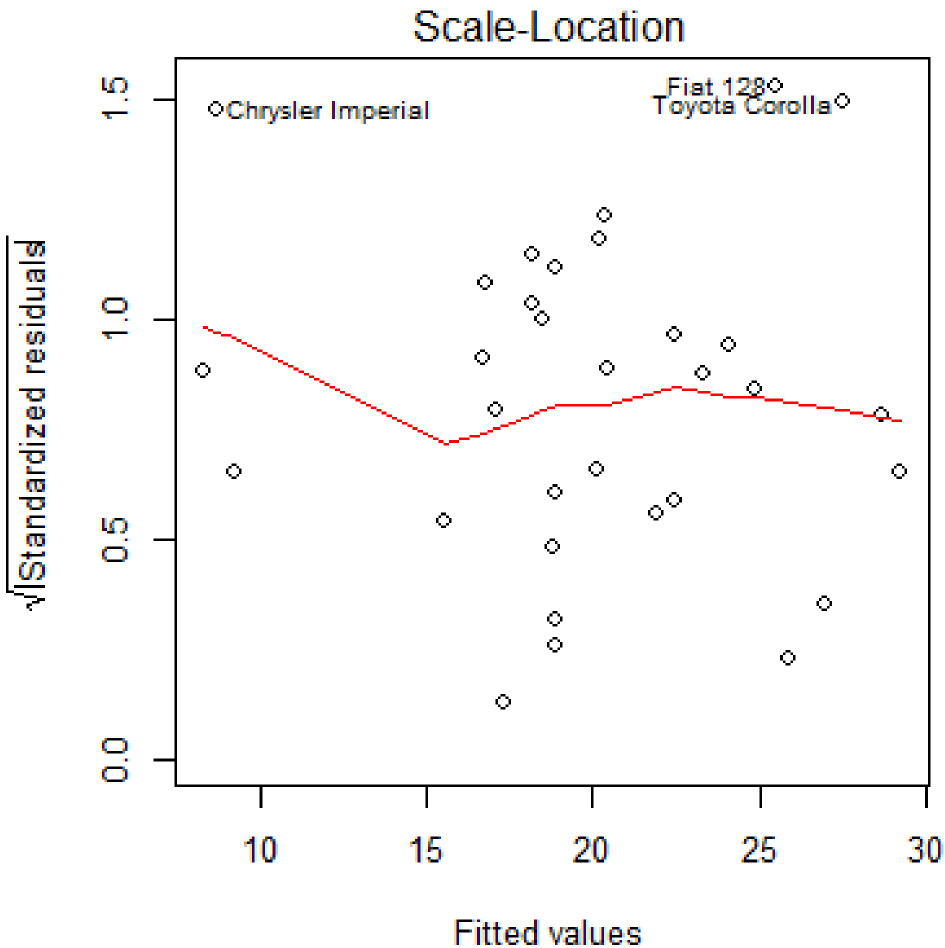
Better Case



Worse Case



Bottom Left Chart

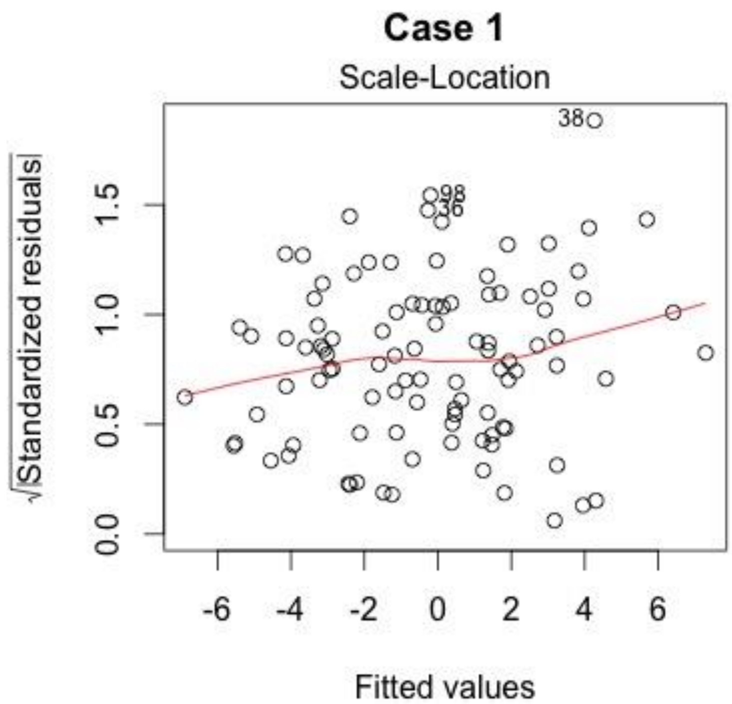


<u>To test</u>

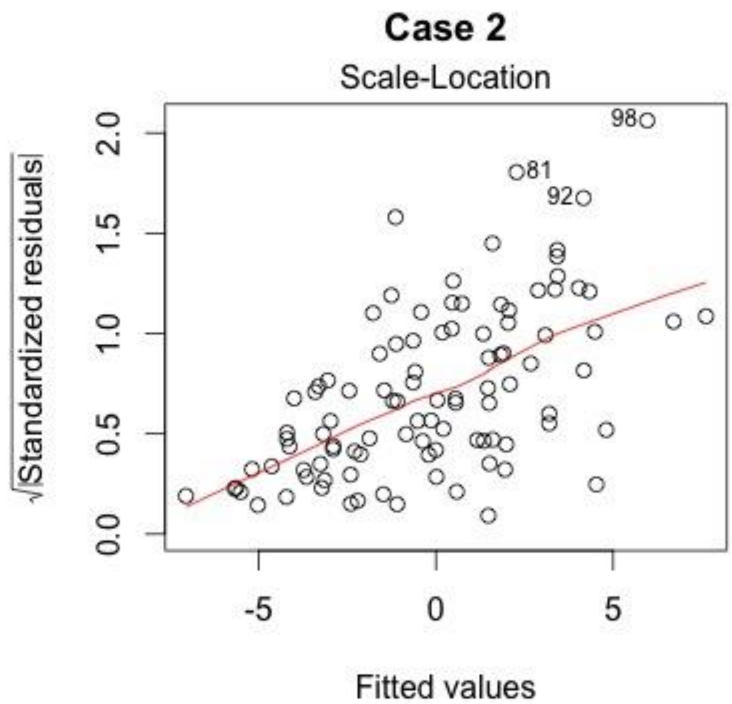
Assumption 3 = Errors are independent of X and has **constant standard deviation**.

Bottom Left Chart (Examples)

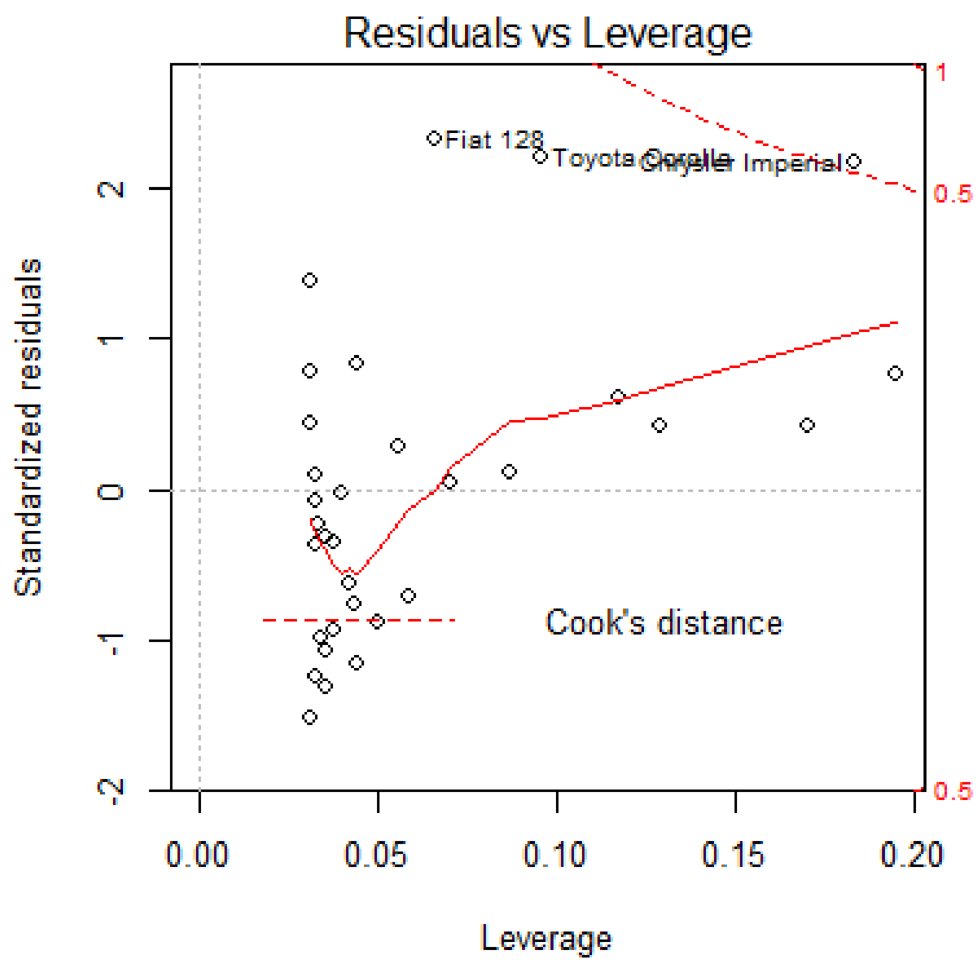
Better Case



Worse Case



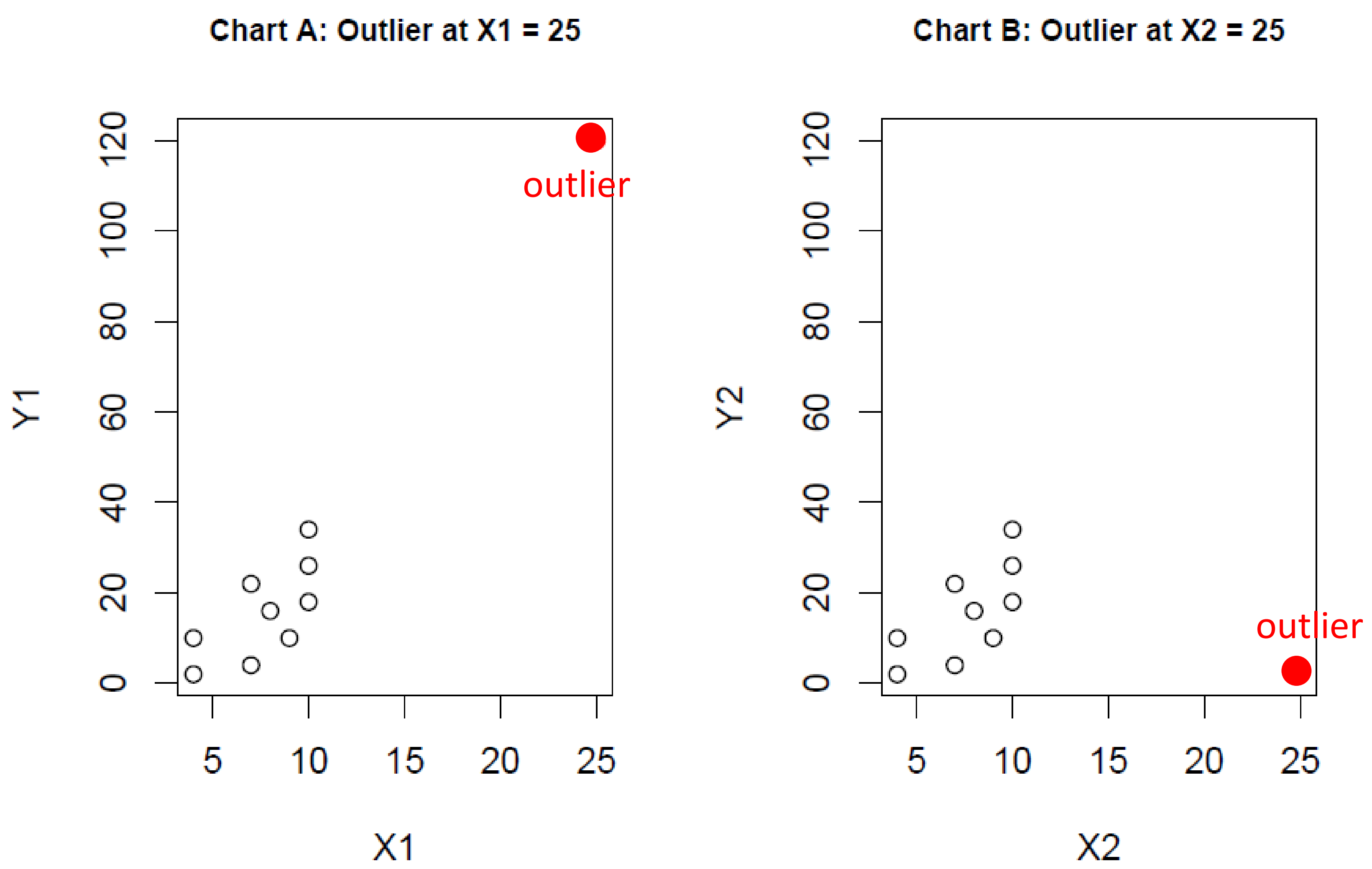
Bottom Right Chart (influential outliers)



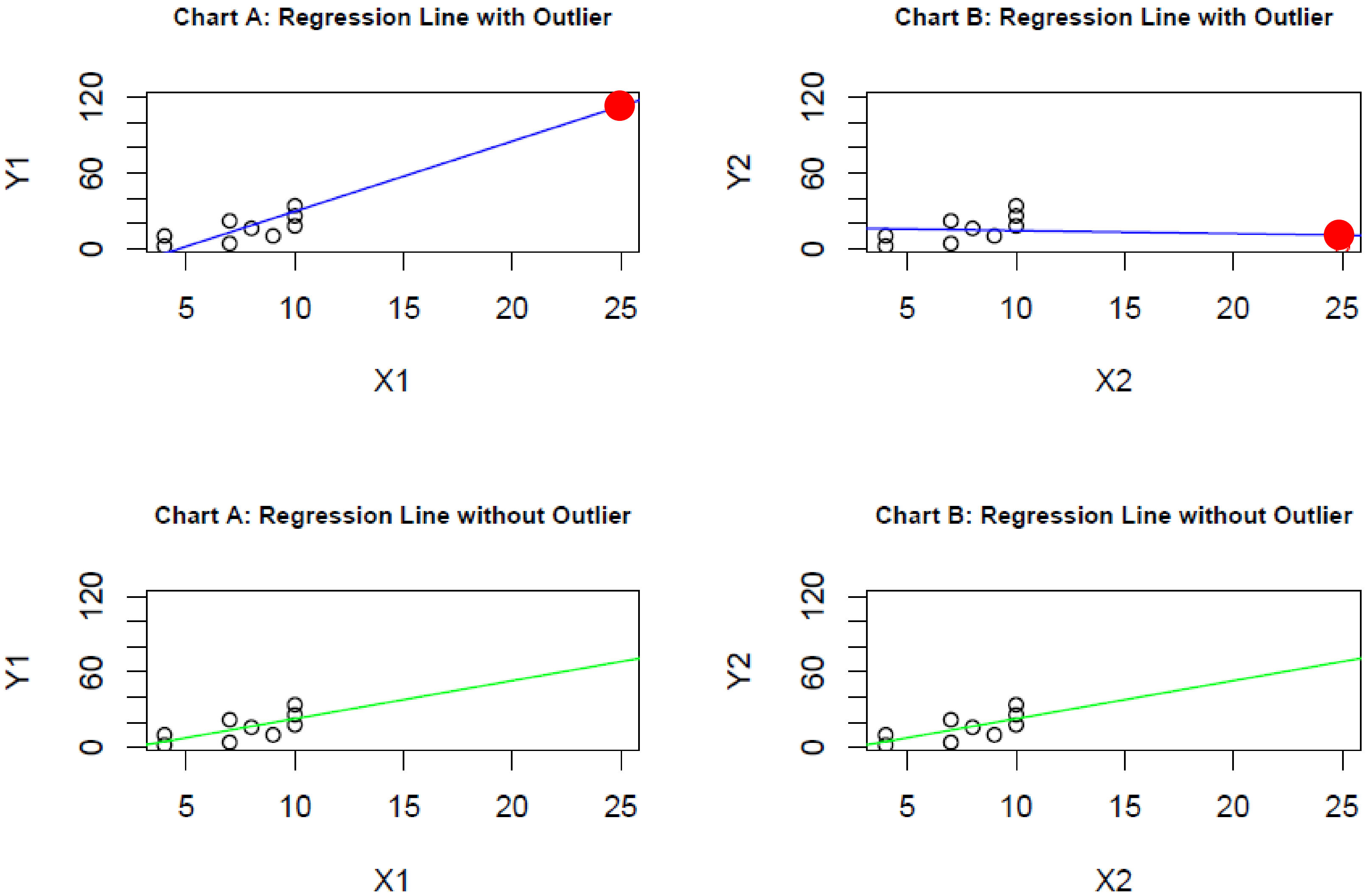
Influential Outliers

- There are two kinds of outliers in any analytics models:
  - Influential
  - Non-influential
- What's the difference?
- Which is more important?

Which chart has influential outlier? A, B or both?



Influence on the model

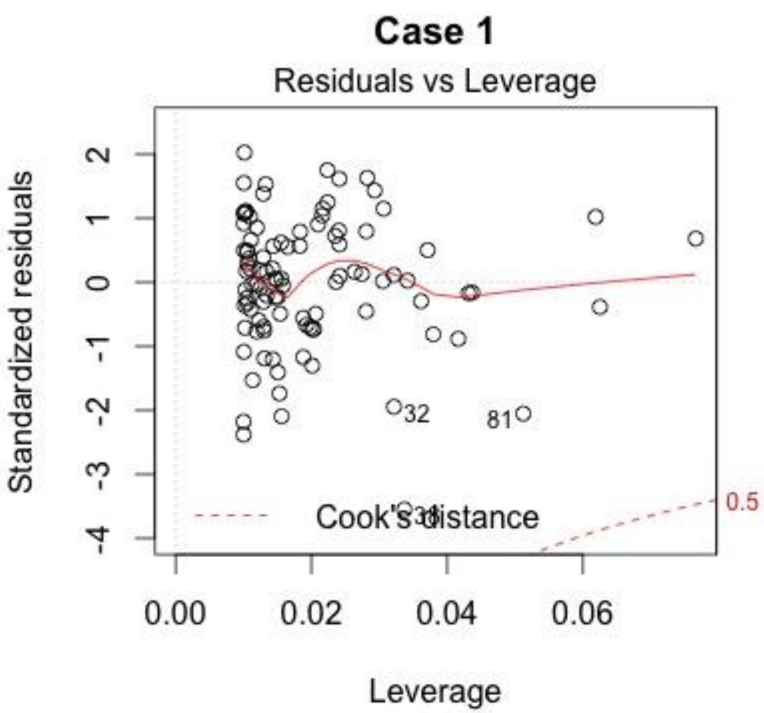


Detecting Influential Outliers

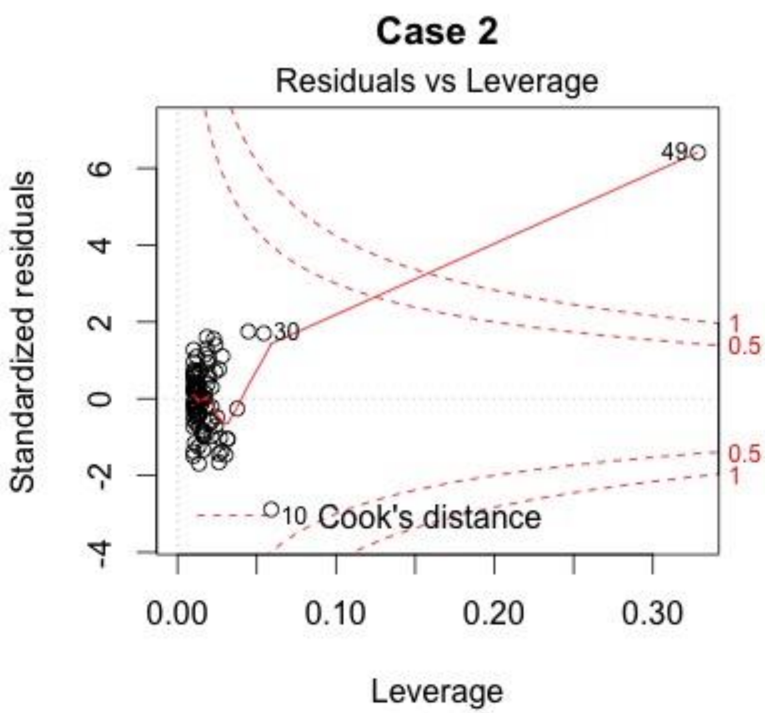
- If model has only one X variable, scatterplot will easily reveal the existence of influential outliers (if any).
- If more than two Xs in the model, scatterplot cannot be done. Use Cook's statistics.
  - Easily presented as a standard model diagnostic plot in R.

Bottom Right Chart (influential outliers)

Better Case



Worse Case



Category X variable in Linear Regression

If X is categorical, which model is correct?

- Y: Salary; X: Occupation Code (1: Clerk, 2: Analyst, 3: Manager)
- Linear Regression Model 1 (without dummy variable):

  Avg.Salary = 1510 + 700 (occ.code)
- Linear Regression Model 2 (with dummy variable):

  Avg.Salary = 1500 + 465 (occ.code == 2) + 1156 (occ.code == 3)

Note: (occ.code == 1) Group is used as a baseline group.

R automatic create dummy variables

- If R recognize a variable as categorical (check that the data type is “factor”),
- Dummy variables will be automatically created.
- If X has k categorical levels, k – 1 dummy variables will be created
- The baseline reference is the smallest categorical level by **alphabetical order**.
  - Baseline reference level can be changed with relevel() function.

How to select which Xs go into the Reg model?

- Expert Opinion +
- Domain knowledge +
- Statistical Opinion
  - P- values of the Xs (less than 5%).
  - Automatic Selection Algorithm
    - Backward Elimination
    - Forward Selection
    - Bidirectional Selection & Elimination
  - Dimension Reduction (Feature Engineering) Techniques
  - Another Model to select variables e.g. CRT.
  - Other methods…

Multicollinearity Detection via Variance Inflation Factor (VIF)

- If there are multicollinear X variables
  - When an X variable can be expressed statistically well as a linear combination of some other X variables
  - It means a lot of information about that X variable is already contained in the other X variables.
- Mathematically, given a Model M, the VIF of the i^{th} X variable, X_i is:

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

where R_i^2 is the R^2 statistic in the linear regression with X_i as the outcome variable (Y) on all the other X variables in the Model M.

VIF – No consensus on cut-off

- Some research papers conclude multicollinearity if VIF > 5 (or equivalently $R_i^2 > 0.8$);
- Others are more strict and conclude multicollinearity if VIF > 10 (or equivalently $R_i^2 > 0.9$);
- For models with dummy variables: If GVIF > 2
- Use **vif()** function from external Rpackage **car**

Demo: Linear Regression on mtcars

Run “ADA1-6-1 linreg.R” Rscript

- Various ways to build a linear regression model
- How to do model diagnostics
- Multicollinearity & VIF
- **caTools package for train vs test set split**
  - How to develop model on trainset
  - How to apply model on testset
  - How to calculate RMSE on both trainset and testset

Summary

- Linear Regression model is not just the straight-line equation.
- Diagnostic checks is a due diligence.
- Complications:
  - Influential Outliers
  - Multicollinearity
  - Categorical X (Make sure R recognize correctly as categorical)