Chapter 7: Linear Regression

March 16

- Lecture
  - Chapter 7 Linear Regression, part 1
- Practical tomorrow
- Homework due next week
- Final Exam to be held during last lecture time slot
  - May 15, 13:45 – 16:30
- Outliers in covariance

Aims for today

- Understand what linear regression is
  - understand linear regression with one predictor
- Understand how we assess the fit of a regression model
  - Least squares and sum of squares
  - *F* and *t* test statistics
  - R^2
- Know how to do regression using **R**
- Understand assumptions of regression and how to evaluate them

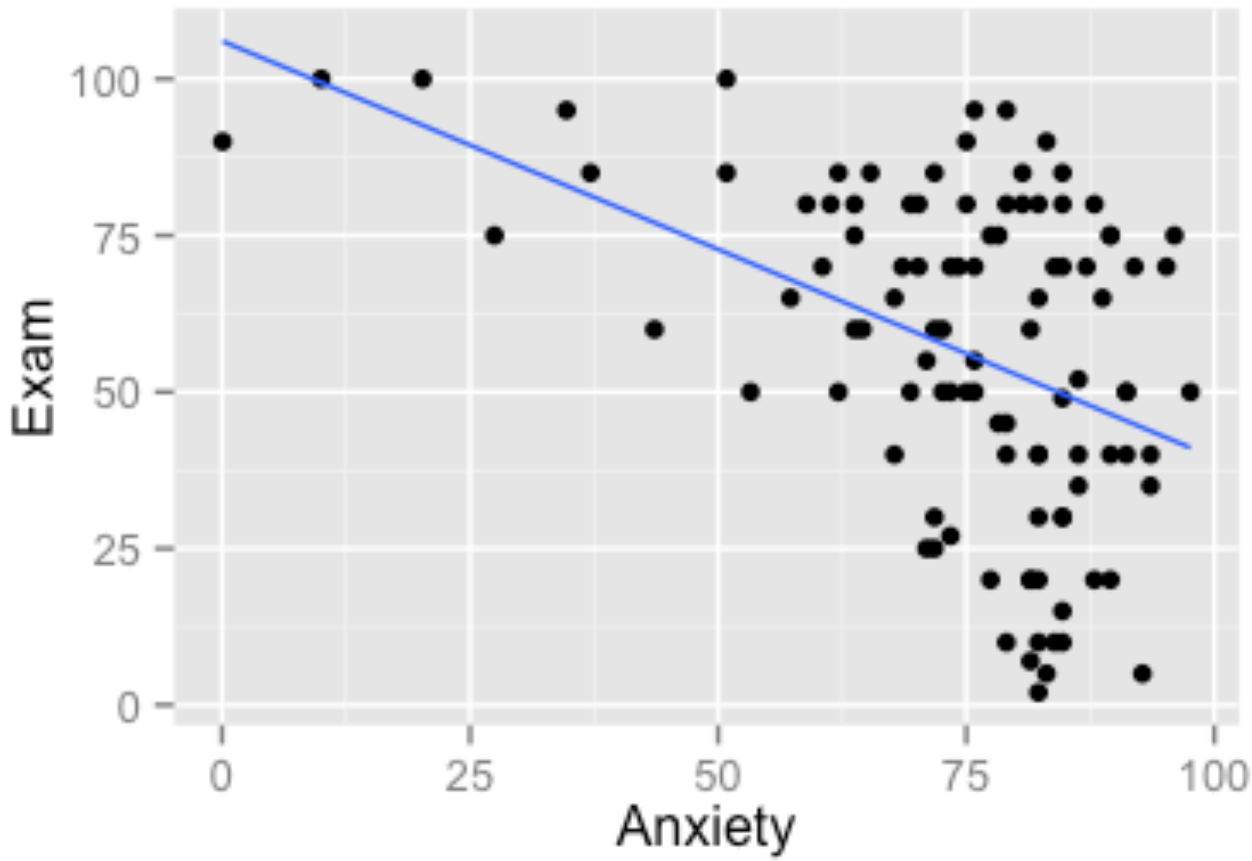
What is Regression?

- A statistical technique that is closely related to correlation
  - with correlation, we were interested in measuring the relationship between existing data points that have values along two variables
- With regression, we <u>go beyond the existing data</u>

What is Regression?

- A way of predicting the value of an outcome variable from the value of one (simple regression) or multiple (multiple regression) predictor variables.
  - It is a hypothetical model of the relationship between variables
  - It is a <u>linear model</u>
    - based on a straight line drawn through the data

What is Regression?

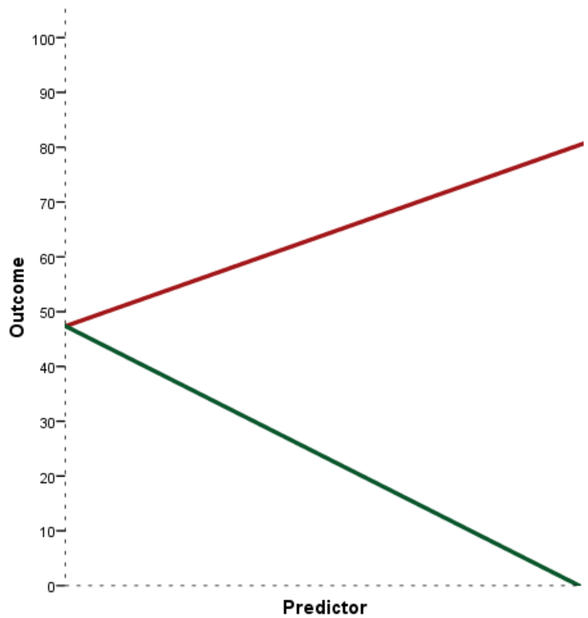


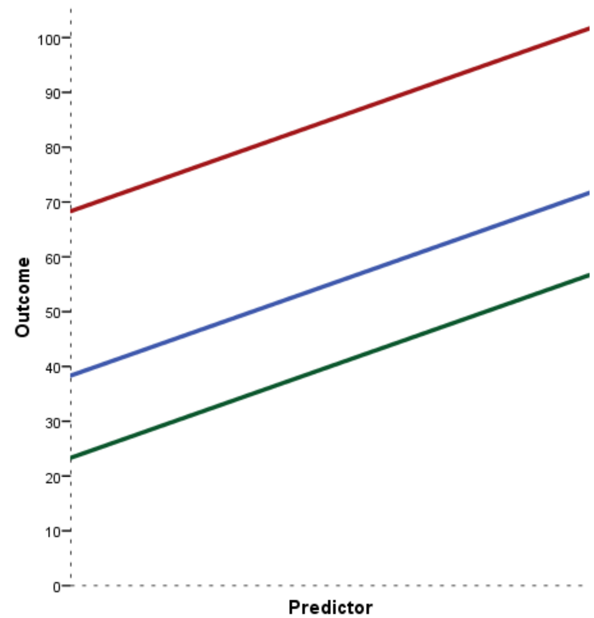
Describing a Straight Line

$$Y_i = b_0 + b_1 X_i$$

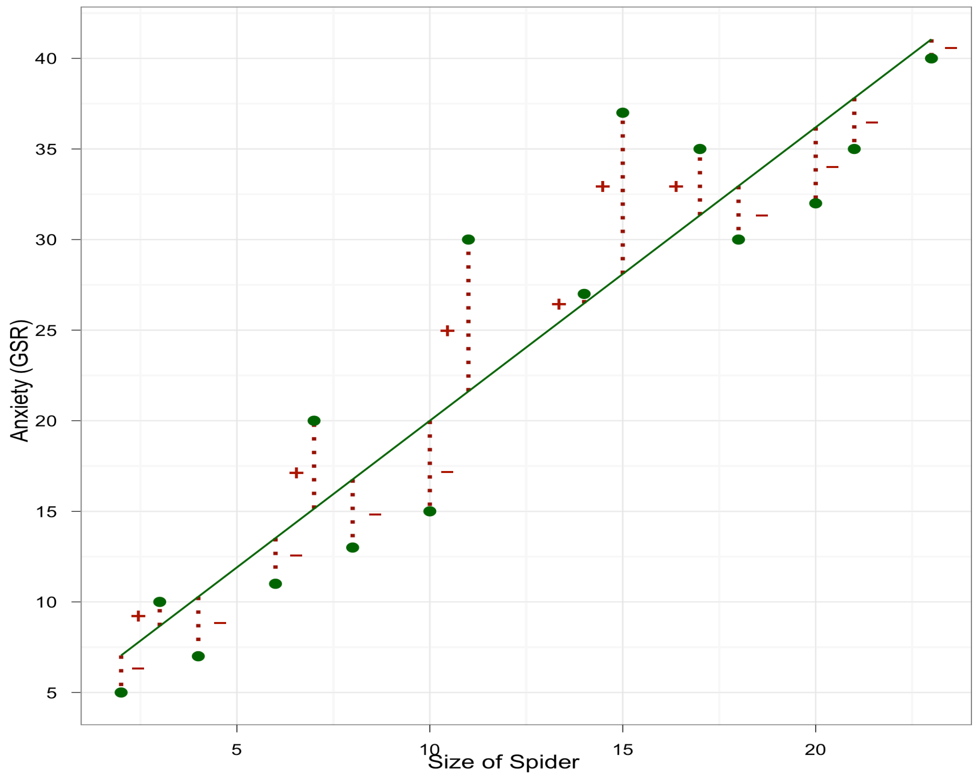
- *"Betas"*
  - b_1
    - Gradient (slope) of the regression line
    - Direction/strength of relationship
  - b_0
    - Intercept (value of Y when X = 0)
    - Point at which the regression line crosses the Y-axis

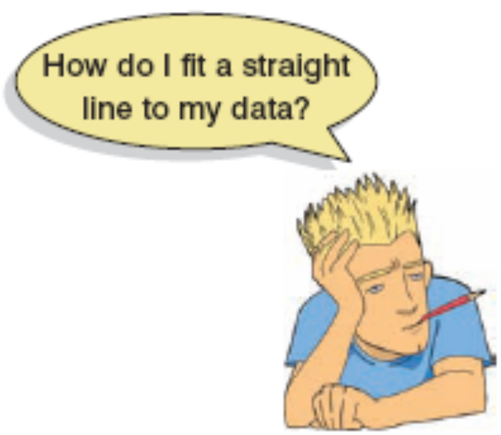
Intercepts and Gradients





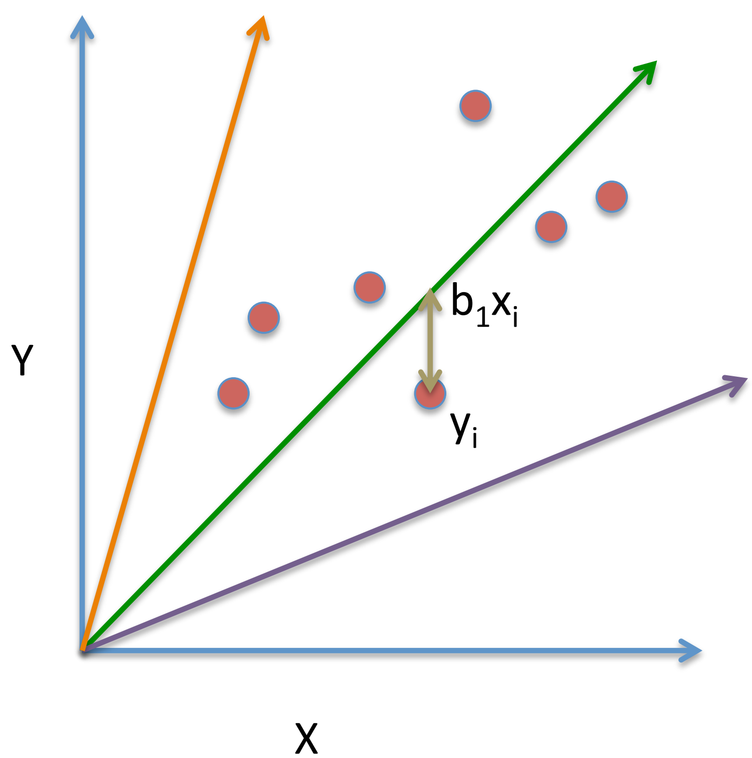
The Method of Least Squares

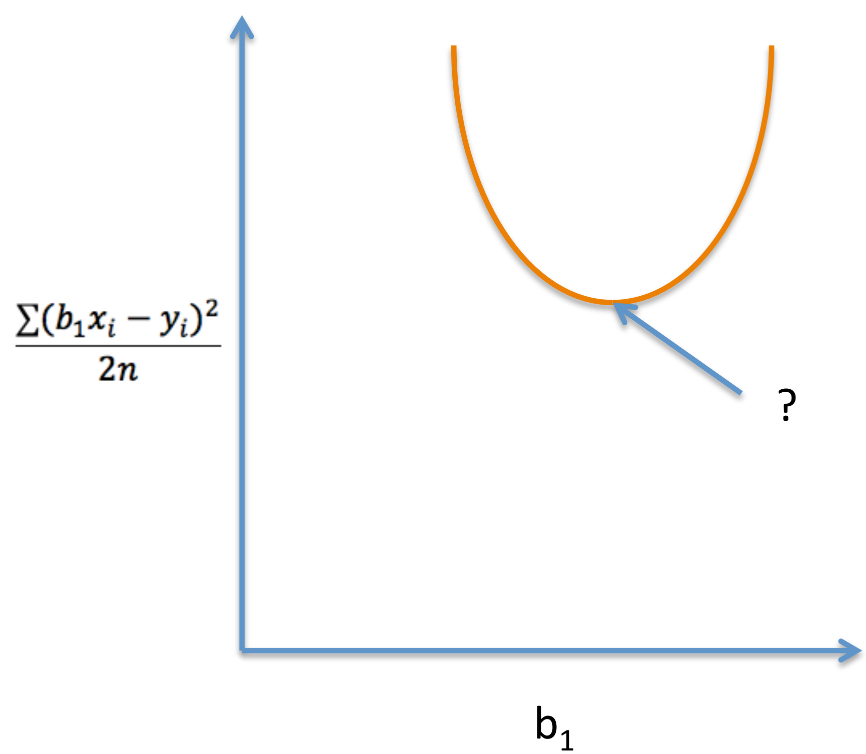




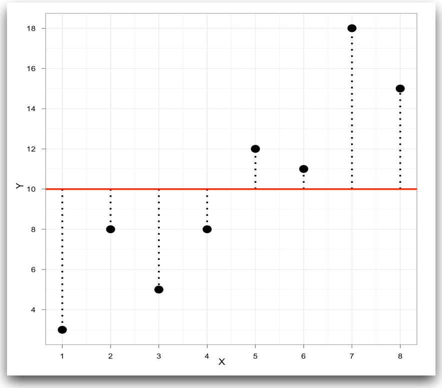
$$\sum(b_0 + b_1x_i - y_i)^2$$

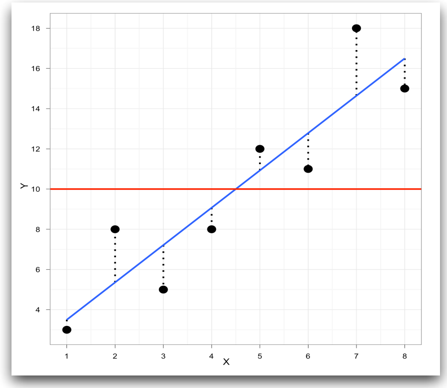
Varying the slope changes residual sum of squares





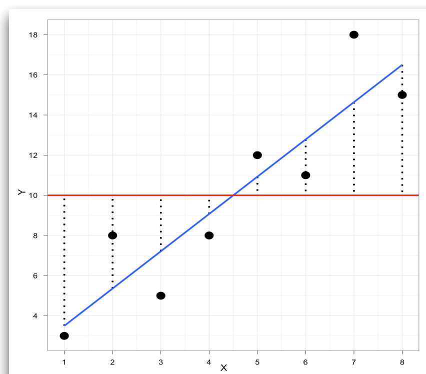
Sums of Squares





SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the regression line



SS_M uses the differences between the mean value of Y and the regression line

Summary

- SS_T
  - Total variability (variability between scores and the mean).
- SS_R
  - Residual/error variability (variability between the regression model and the actual data).
- SS_M
  - Model variability (difference in variability between the model and the mean).

Back to R^2

- $SSm/SSt = R^2$
  - amount of variance explained by the model relative to the amount of variance there was to explain in the first place
  - tells us how good of a fit our regression is
- Which means that the square root is the Pearson correlation coefficient for the data

Two ways of testing the model

- How likely is it that we would see the pattern/ model fit in our sample data if there's no such pattern in the population?
- F statistic
  - Testing the overall model
- t statistic
  - Testing the betas

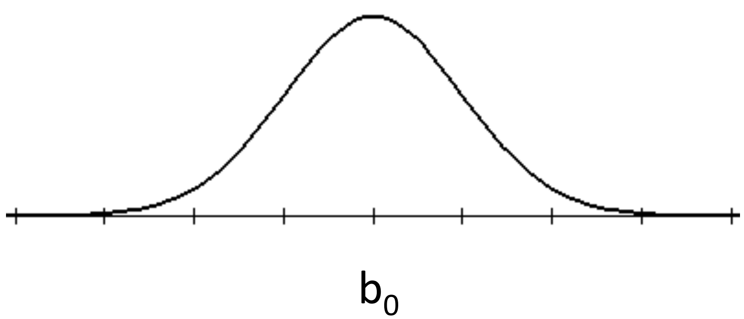
Testing the Overall Model

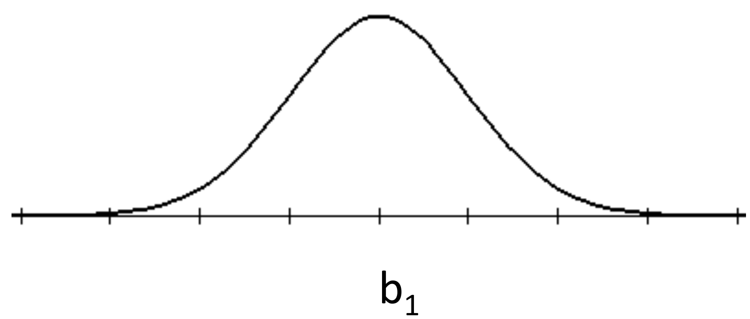
- Mean squared error
  – Sums of squares are total values.
  – They can be expressed as averages.
  – These are called mean squares, MS.
- Explained variance over unexplained variance

$$\boxed{F = \frac{MS_M}{MS_R}}$$

Testing the betas

- null hypothesis: betas = 0
- Standard Errors of the betas
  - standard deviations of the sampling distributions





Testing the betas

- Calculate t test statistic for each beta
  - ratio explained to unexplained variance

$$t = \frac{b_{observed} - b_{expected}}{SE_b} = \frac{b_{observed} - 0}{SE_b}$$

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and downloads) in the week after release
- Predictor variable:
  - The amount (in units of £1000) spent promoting the record before release.

Regression in R

- We run a regression analysis using the *lm()* function – lm stands for 'linear model'. This function takes the general form:

```
newModel<-lm(outcome ~ predictor(s), data = dataFrame)
```

Regression in R

```
albumSales.1 <- lm(sales ~ adverts, data = album1)
```

- or we can specify the columns directly:

  ```
  albumSales.1 <- lm(album1$sales ~
  album1$adverts)
  ```

Output of a Simple Regression

- We have created an object called *albumSales.1* that contains the results of our analysis. We can show the object by executing:

```
summary(albumSales.1)

>Coefficients:
              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) 1.341e+02   7.537e+00   17.799      <2e-16 ***
adverts      9.612e-02   9.632e-03   9.979       <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.99 on 198 degrees of freedom
Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313
F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Using the Model

$$\begin{aligned}\text{Record Sales}_i &= b_0 + b_1\text{Advertising Budget}_i \\ &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right)\end{aligned}$$

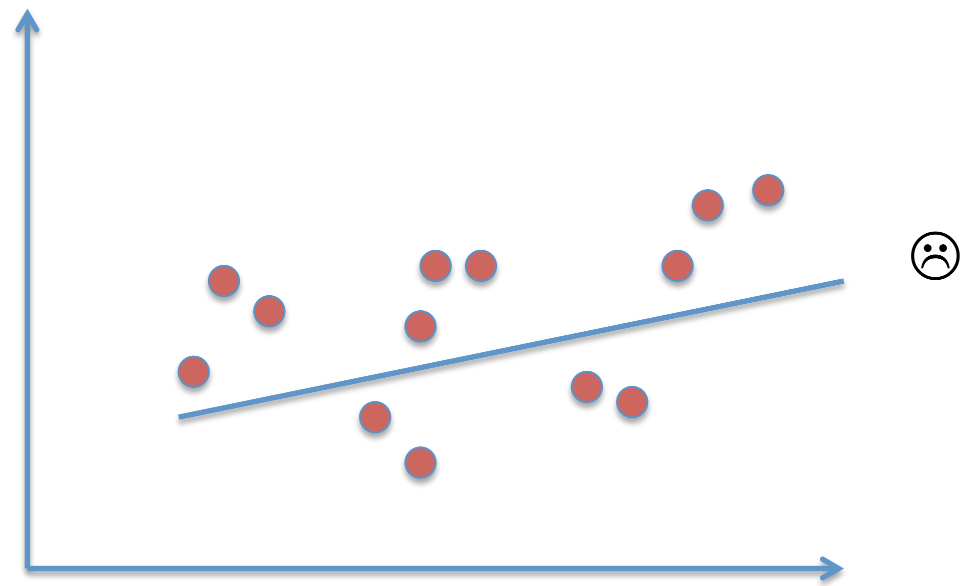
$$\begin{aligned}\text{Record Sales}_i &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right) \\ &= 134.14 + \left(0.09612 \times 100\right) \\ &= 143.75\end{aligned}$$

Assumptions

- Variable types
- Independence
- Linearity
- Assumptions regarding residuals
  - No autocorrelation
  - Homoscedasticity
  - Normally distributed

Assumption of no autocorrelation

- Also called “independent errors”



- durbinWatsonTest(name_of_model)
  - D-W statistic > 2: residuals negatively correlated
  - D-W statistic < 2: residuals positively correlated
  - returns a p-value!

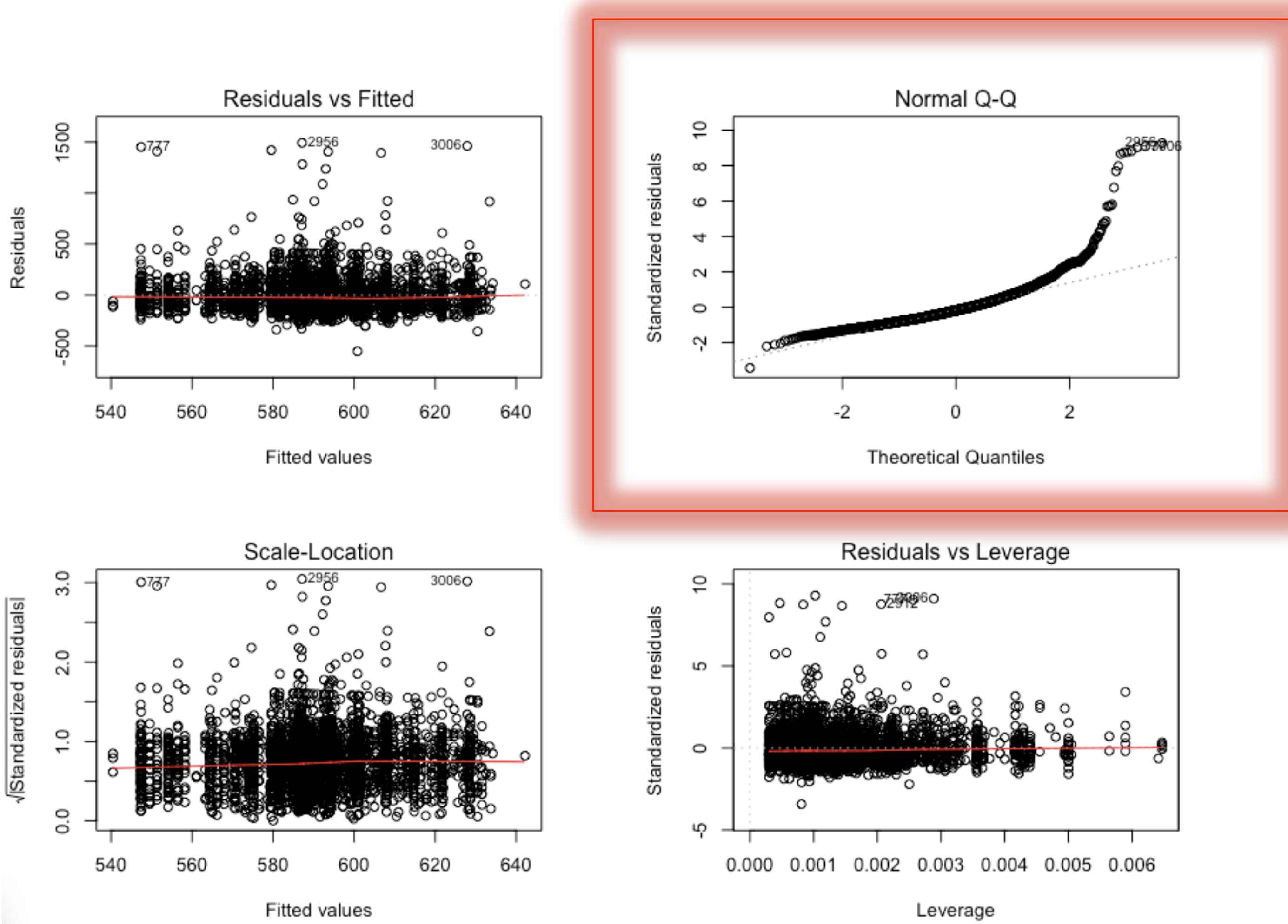
Model Diagnostic Plots in RStudio

```
plot(name_of_model)
```

-or to visualize all 4 graphs at once-

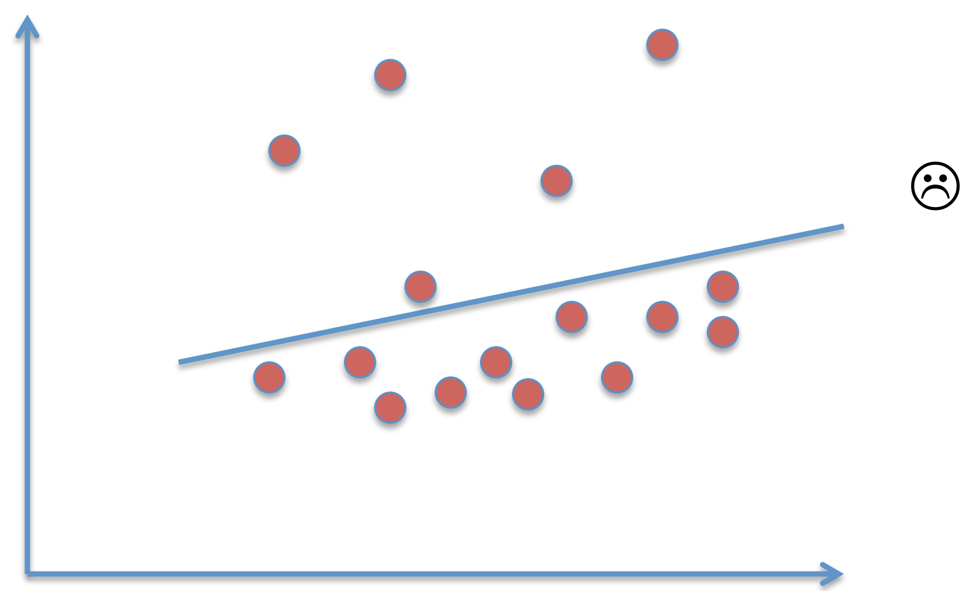
```
par(mfrow = c(2,2))
par(mar = c(4.25,4.25,4.25,4.25))
plot(name_of_model)
```

Model Diagnostic Plots

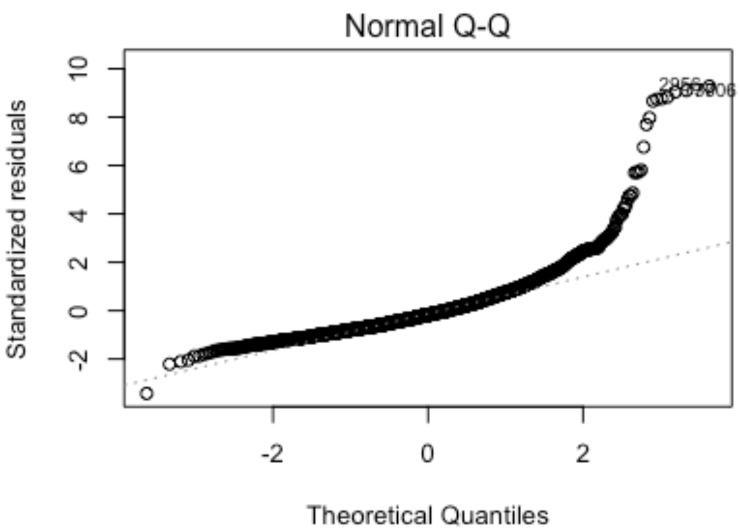


Assumption of Normally Distributed Errors

- Do your residuals form a normal distribution?

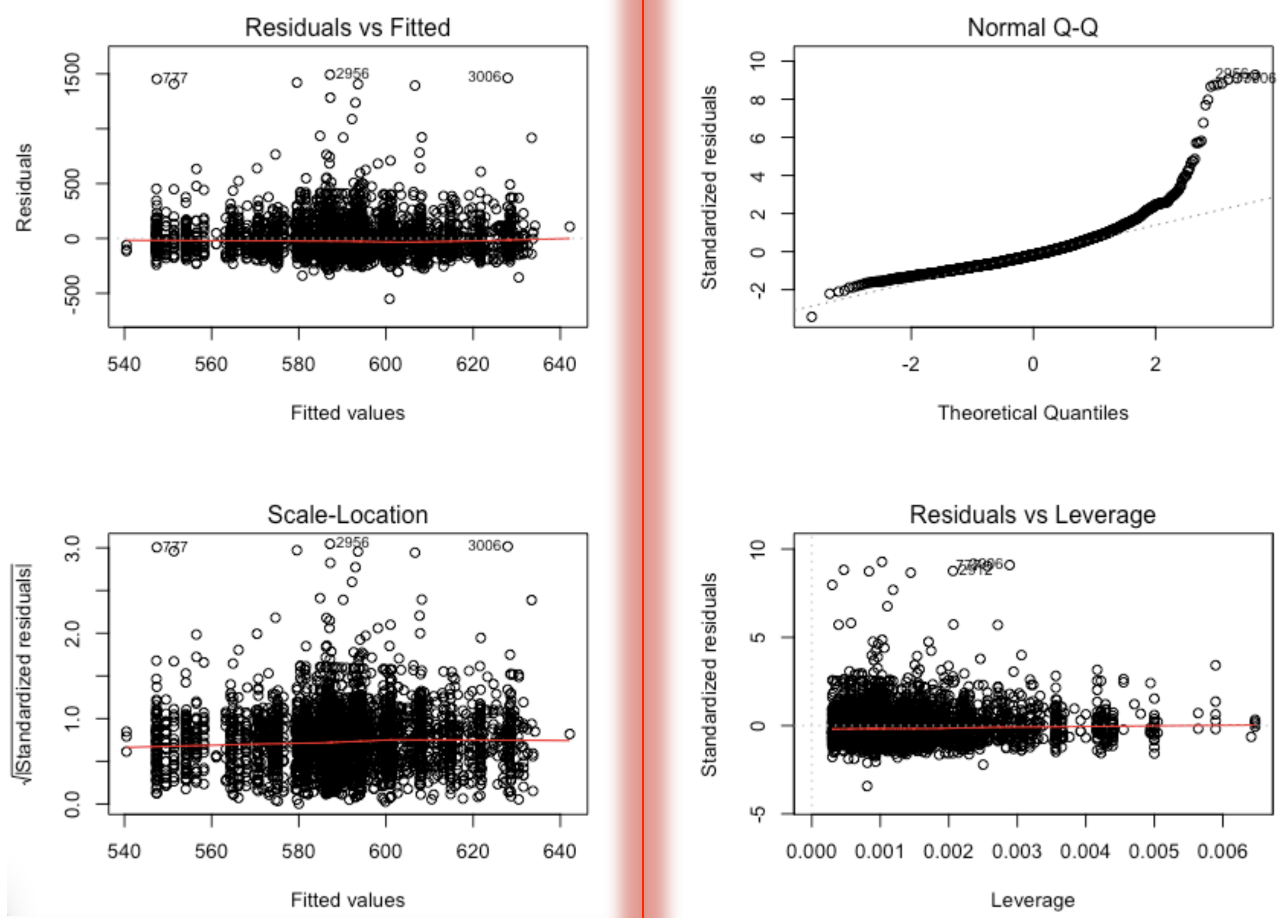


Assumption of Normally Distributed Errors



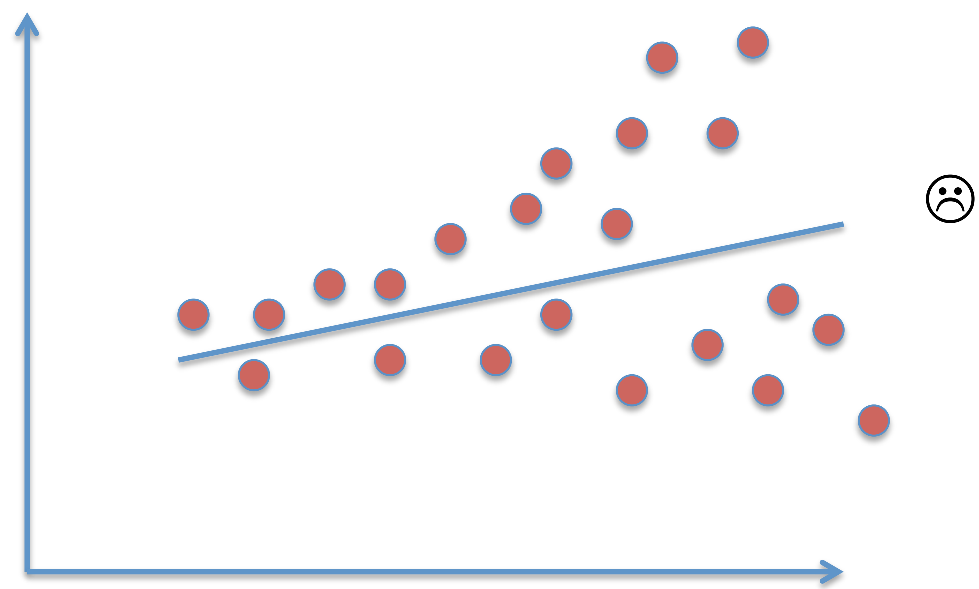
☹

Model Diagnostic Plots

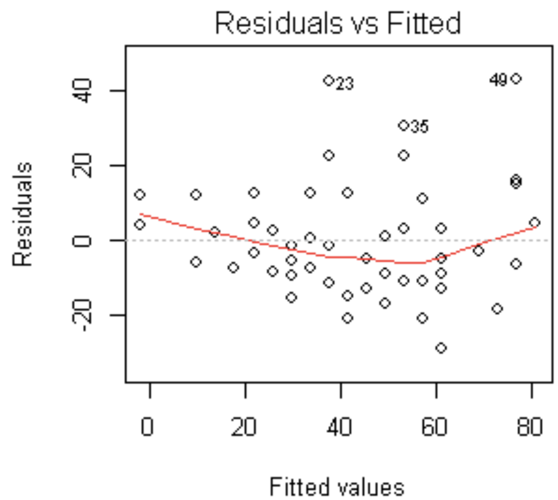


Assumption of Homoscedasticity

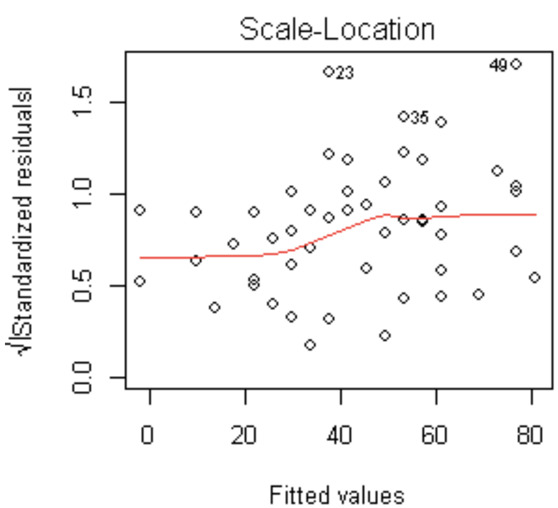
- Are the size of the residuals consistent across the values of your predictor(s)?

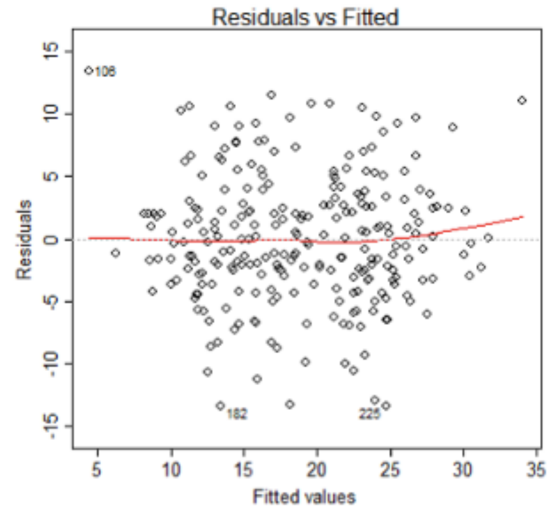


Assumption of Homoscedasticity

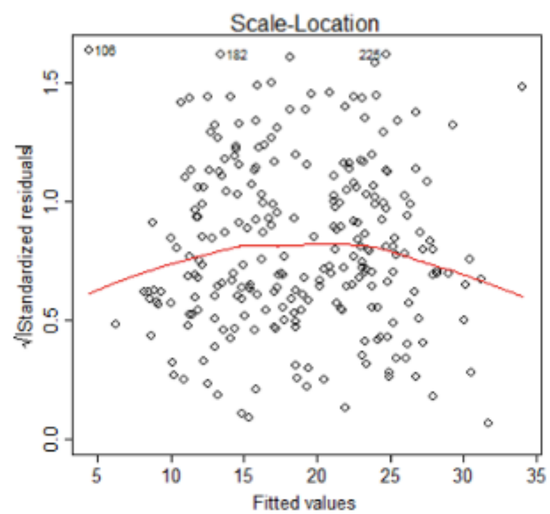


☹

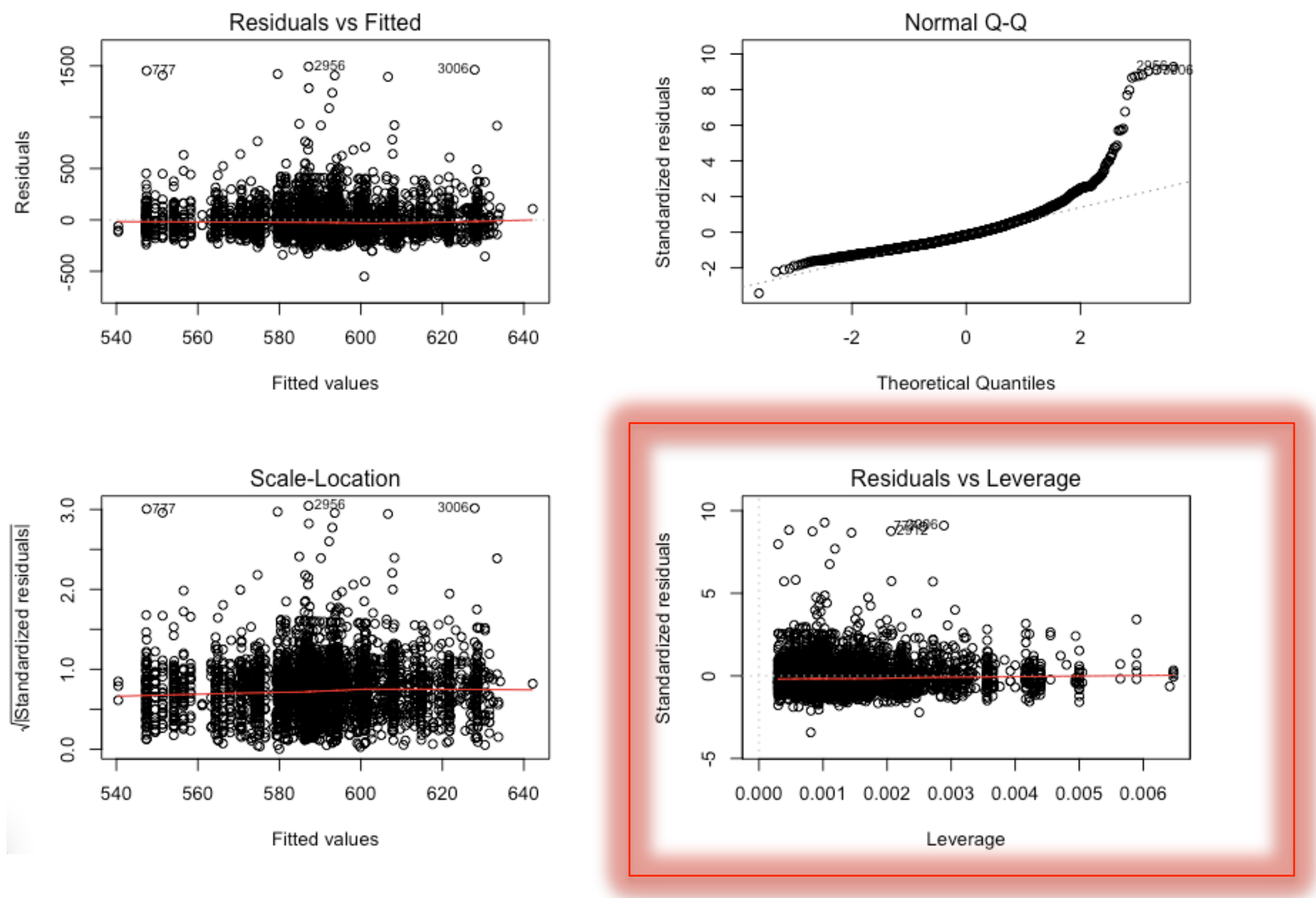




☺



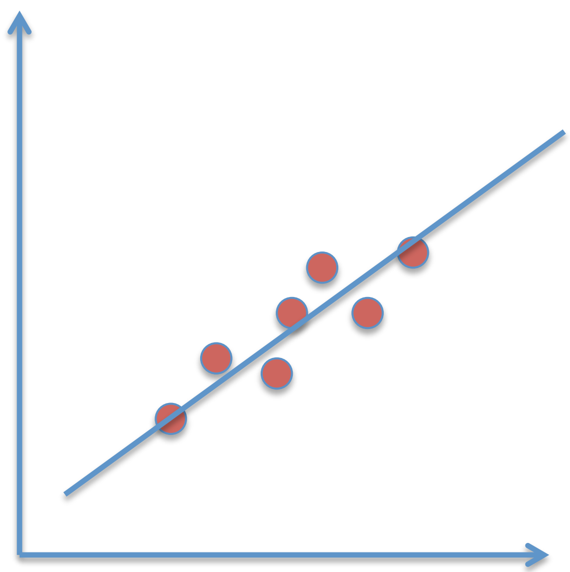
Model Diagnostic Plots

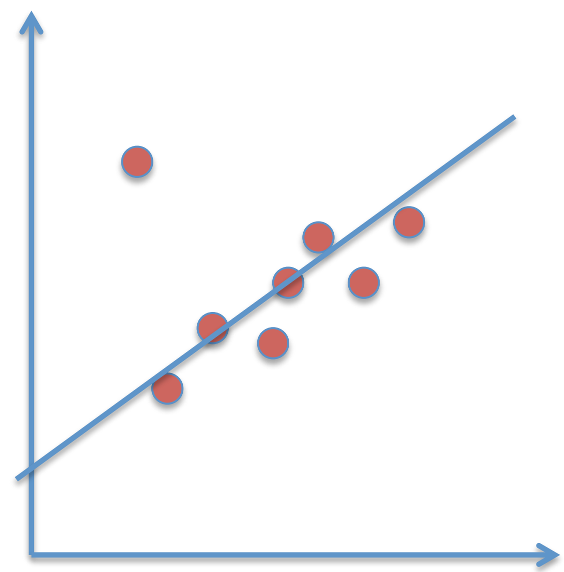


Outliers vs. Influential Points vs. High-Leverage Points

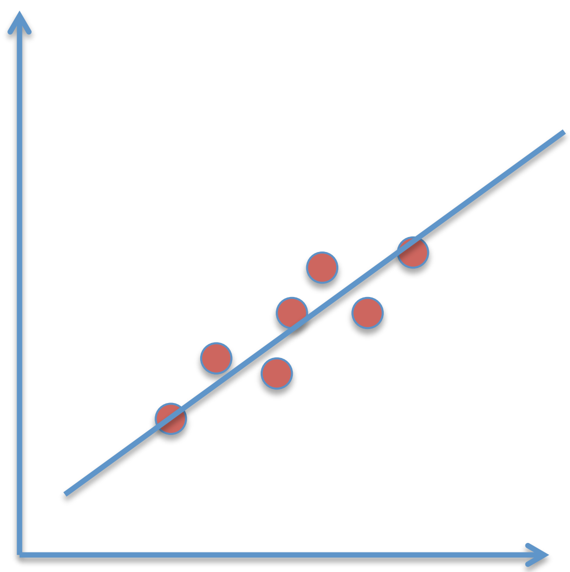
- Outliers
  – extreme points that don't fit the general pattern of the data
- Influential point
  – an outlier that greatly affects the slope of the regression line
- High-leverage point
  – a data point with an extreme x value

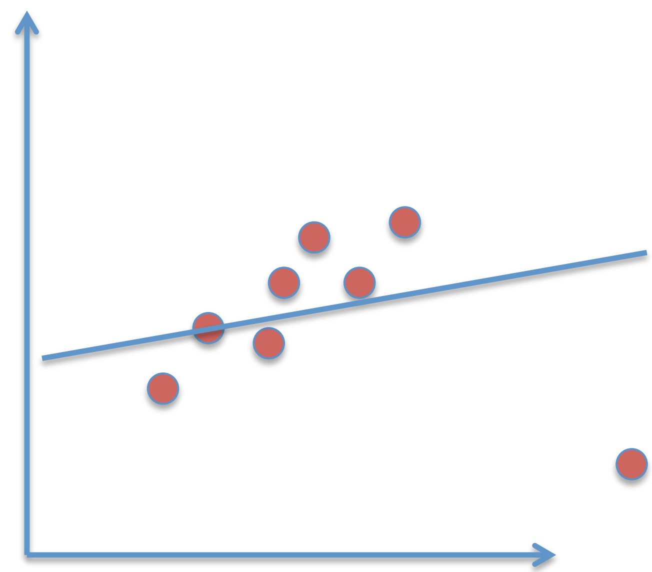
What is this?



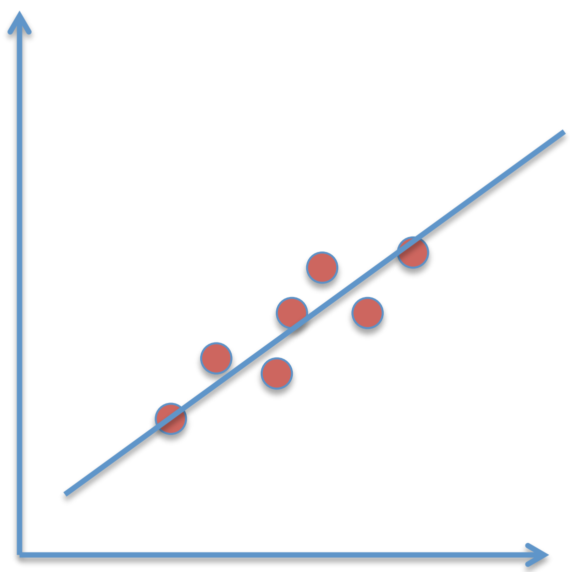


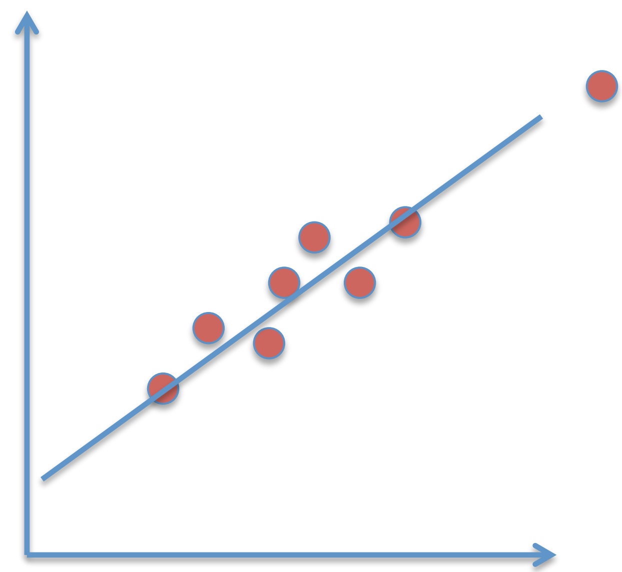
What is this?





What is this?





Detecting Outliers

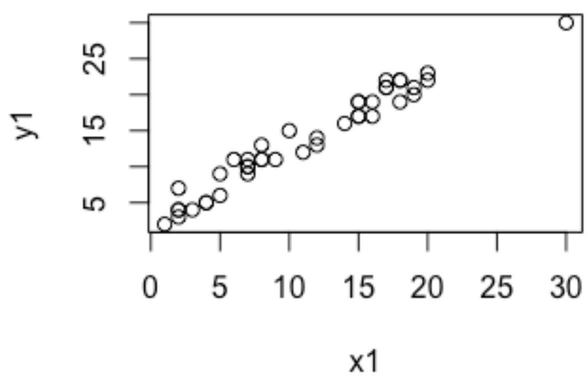
- standardized residuals
  - residuals divided by their standard deviation
    - these are z-scores!
    - remember that 99.9% of data should be between +/- 3.29
  - rstandard(name_of_model)

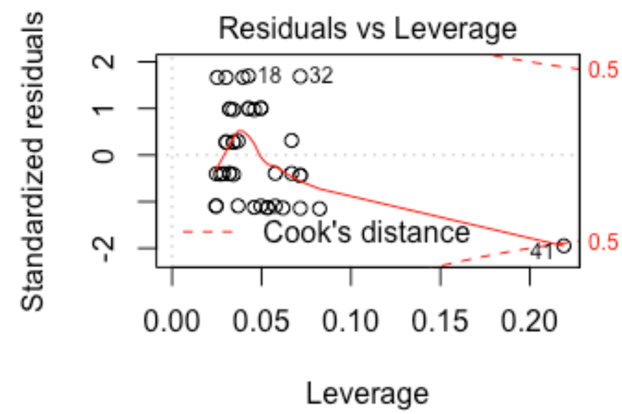
Assessing Influential Cases & Leverage

- Influential cases: Cook's distance
  - values greater than 1 ☹
  - cooks.distance(name_of_model)
- Leverage: Hat values
  - $(k+1)/n$ = ave. hat value for a data set
    - k: number of predictors
    - n: number of participants
  - 2 or 3 times ave. value ☹
  - hatvalues(name_of_model)

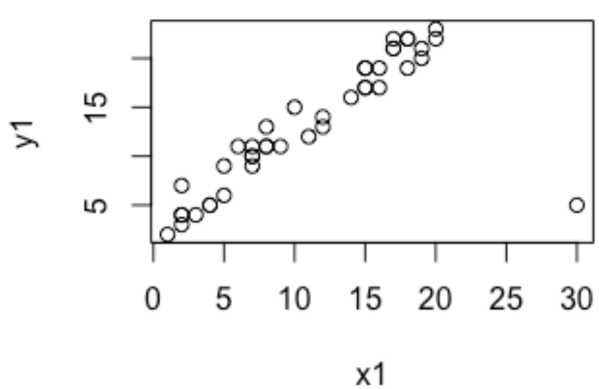
Leverage/Influential Point examples

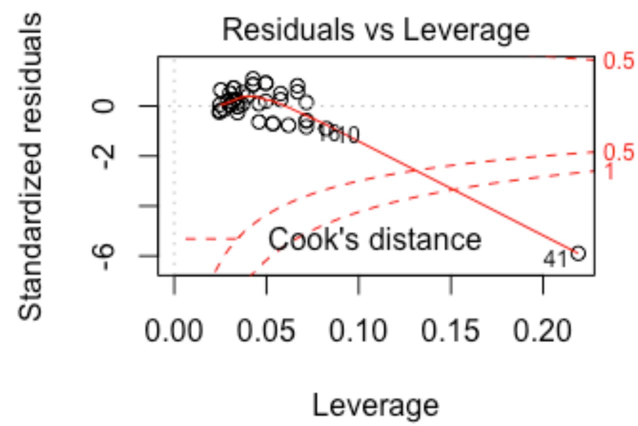
High Leverage Point/No (highly) influential point (and not an outlier)





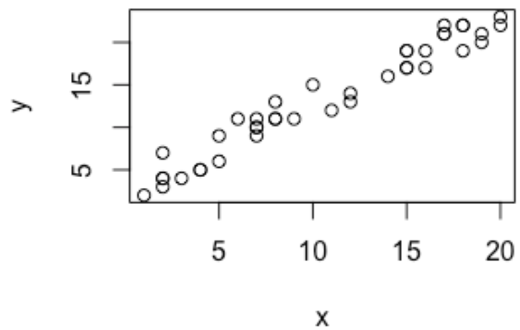
High Leverage Point/Highly influential point (and an outlier)

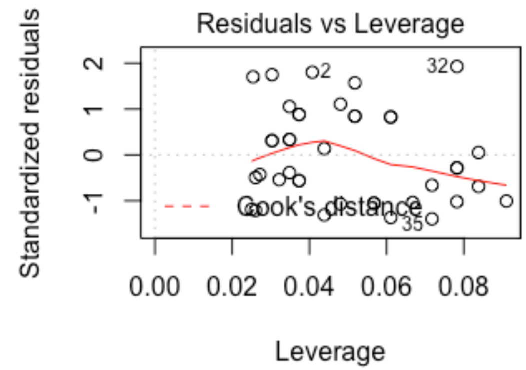




Leverage/Influential Point examples

No high leverage/No influential points





No high leverage/No (highly) influential points (but yes outlier!)