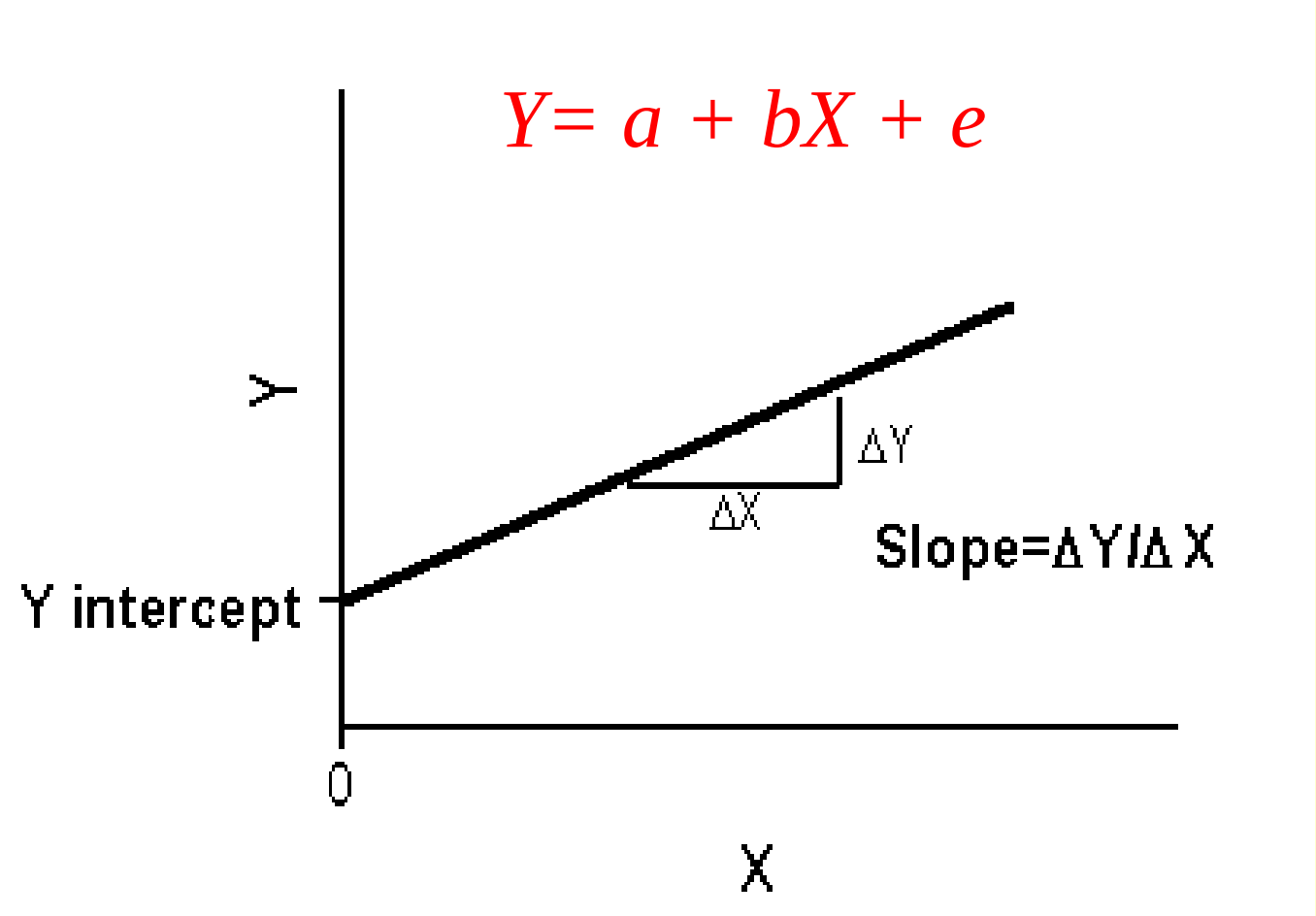
Simple Linear Regression

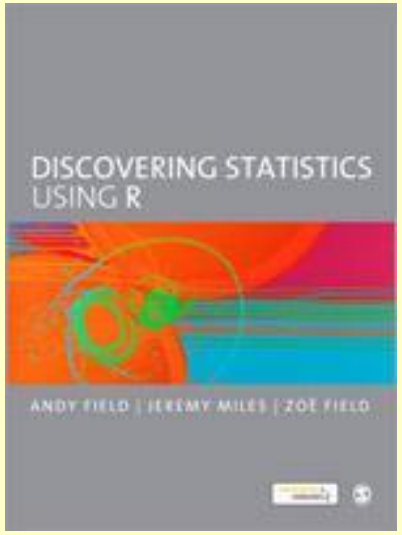


http://www.pelagicos.net/classes_biometry_fa18.htm

Reading - Field: Chapter 7

AIMS

- Understand linear regression with one predictor
- Learn how to assess fit of linear regression
  - Total Sum of Squares
  - Model Sum of Squares
  - Residual Sum of Squares
  - F
  - R^2
- Learn how to do Regression with R
- Interpret regression model results



What is a Regression

The generic term *Regression* refers to methods that allow the prediction of the value of one (dependent) variable from another (independent).

Regression methods are based on various conceptual models of relationship between these two variables.

Examples include:

- Linear / non-linear regression
- Simple / multiple regression

What is Simple Linear Regression

Most simple version of regression:

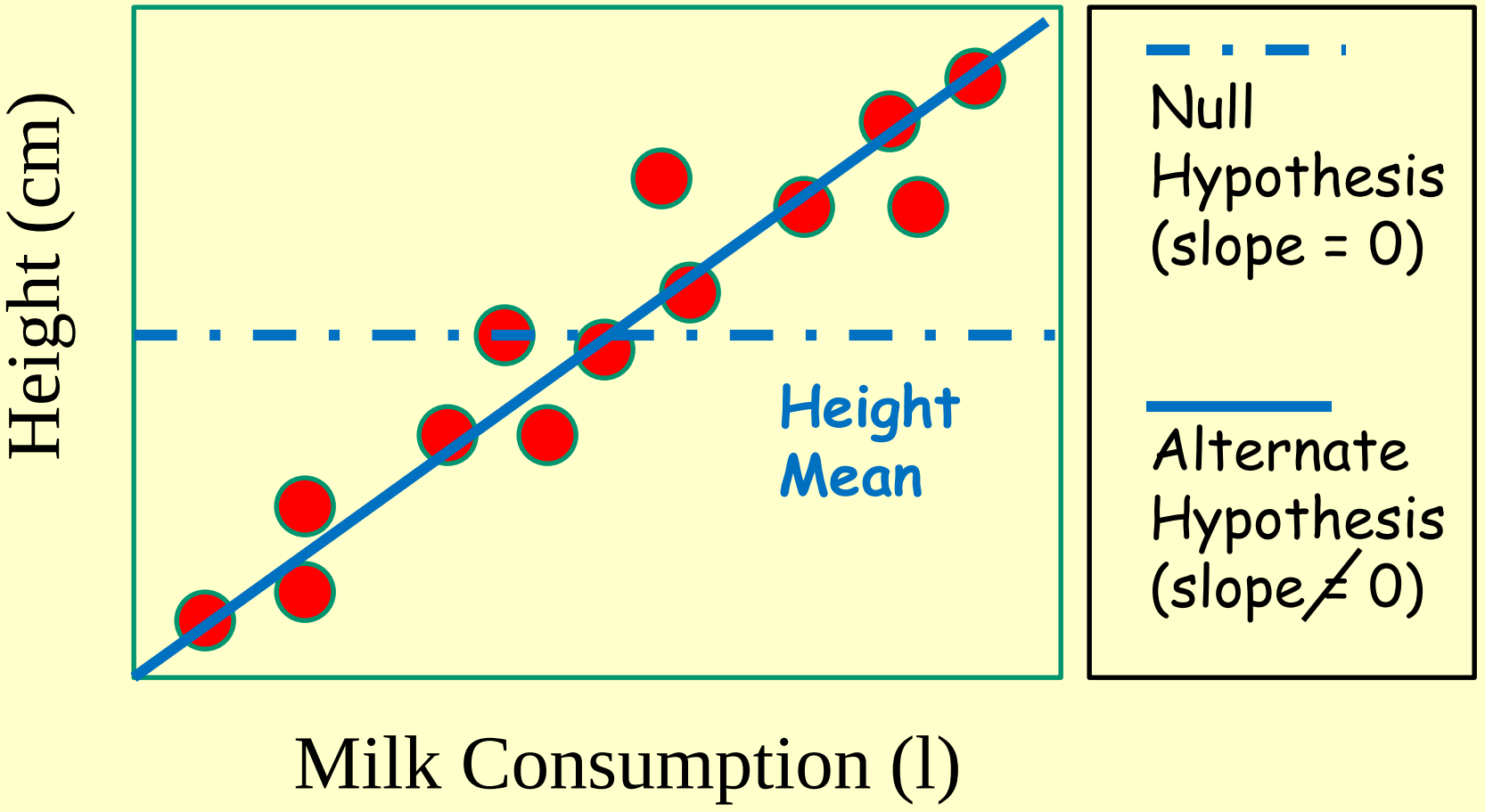
- One independent variable
- Linear relationship

Tests hypothetical model of a linear relationship between two variables:

- Dependent (outcome): Y axis
- Independent (driver): X axis

Simple Linear Regression - How to

Identify the line that best describes relationship between X and Y variables



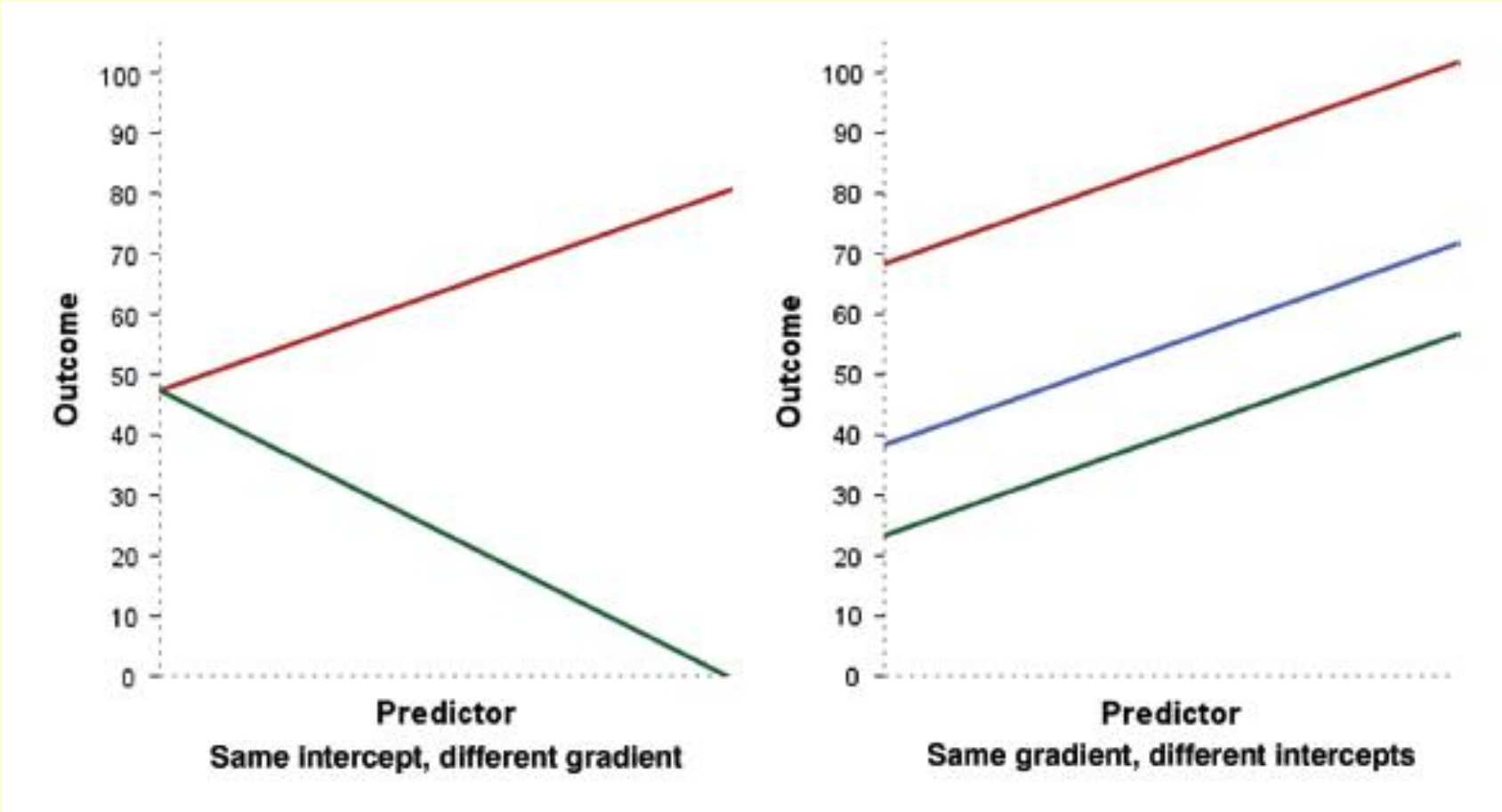
Describing a Straight Line

Model uses linear relationship between X and Y:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- ϵ_i Error (unexplained portion of variation ~ N (μ, σ))
- b_i
  - Regression coefficient for predictor variable
    - Gradient (slope) of the regression line
    - Direction / Magnitude of Relationship
- b_0
  - Intercept (value of Y when X = 0)
  - Point where regression line crosses Y-axis

Intercepts and Gradients



Calculating Slope of Best-fit Line

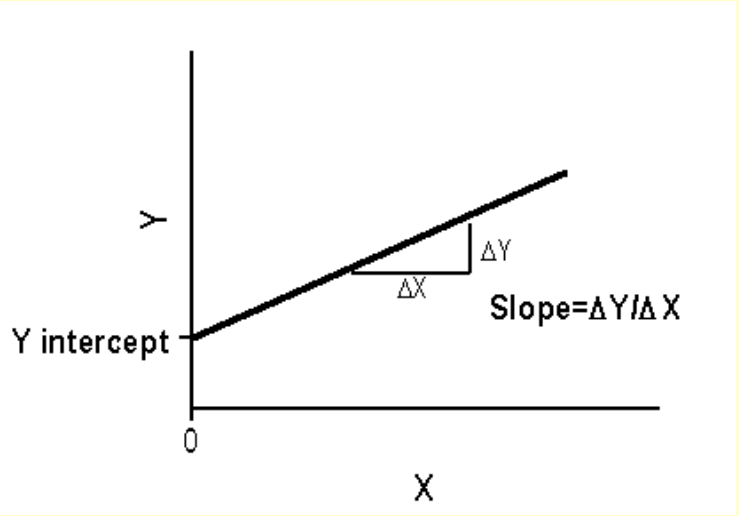
The Regression Coefficient = Slope of Best-fit Line

Covariance between X and Y divided by the variance in X

$$b = \frac{Covariance = \frac{\sum (Xi - \overline{X})(Yi - \overline{Y})}{(n-1)}}{Variance = \frac{\sum (Xi - \overline{X})(Xi - \overline{X})}{(n-1)}}$$

Quantifies best-fit slope of line relating X and Y variables

$$Y = a + bX + e$$



Linear Regression - Assumptions

Linear Regression makes four assumptions:

- (In addition to reliance on "random sampling").

- Variables either interval or ratio measurements.

- Variables normally distributed. No Outliers.

- Linear relationship between the two variables.

How Good is the Fit of the Model

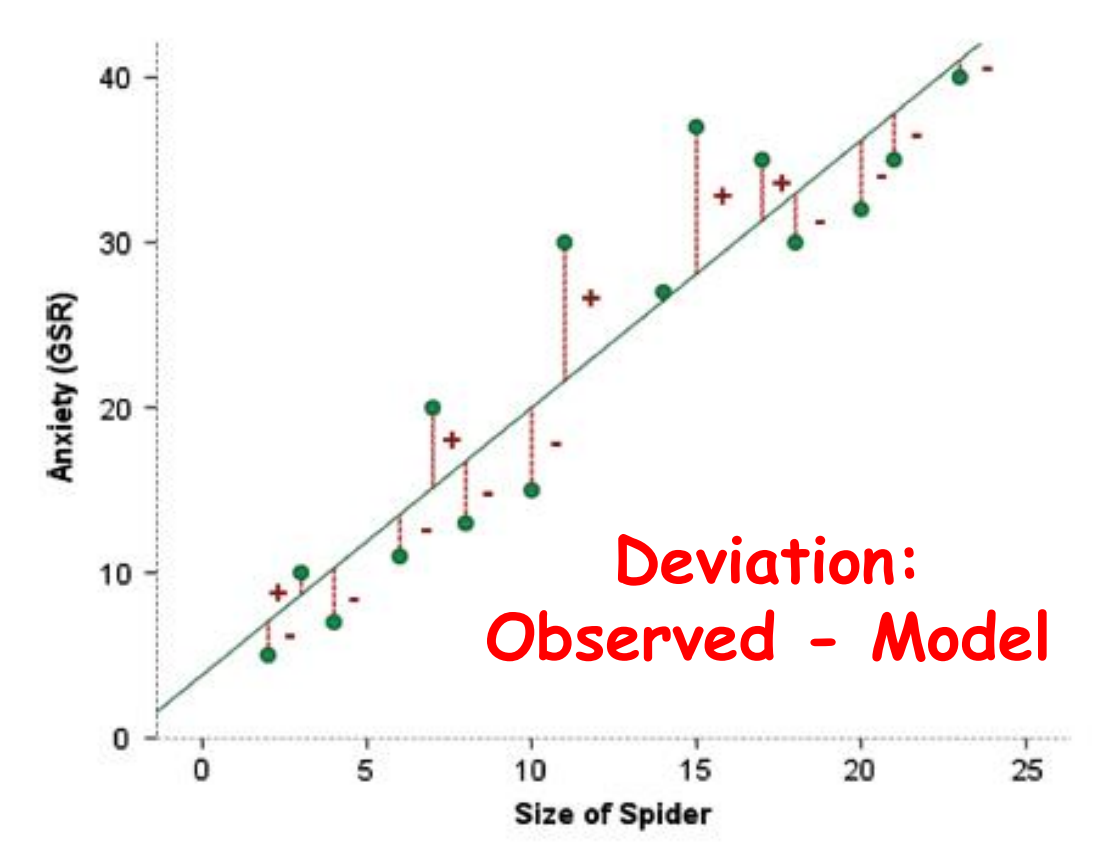
Regression line is based on observations. But, this model might not reflect reality.

- We need a way of testing how well the model fits the observed data: similar to a variance
- Sum of Squares: Sum of the squared deviations (both positive and negative)
- Mean Square: Sum of Squares divided by the degrees of freedom

Measuring Fit

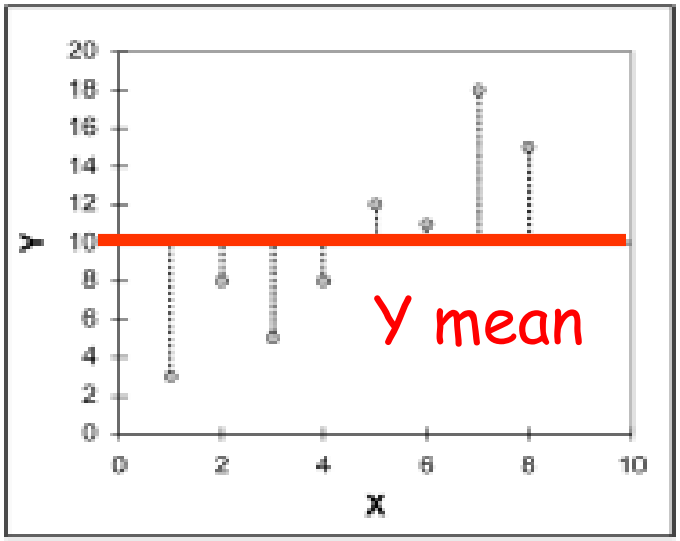
Calculate squared deviations for all data points

Sum of Squares: sum of all squared differences between observed and modeled Y values



Sum of Squares: Σ (Deviations) ^2

Three Different Sum of Squares



df = sample size - 1

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

SS_T - Total SumSquares

Squared difference between observed Y values and their mean calculated from the data

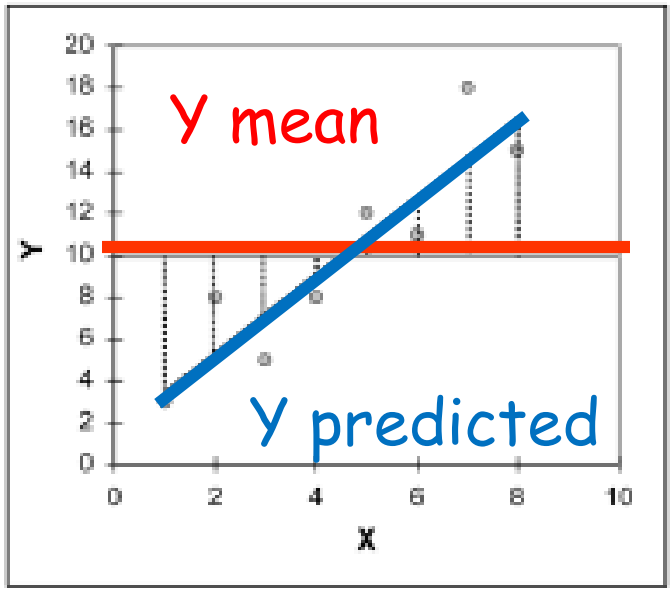
Sum Squares:

$\Sigma (Yi - Y\ mean)\ ^2$

Mean Squared:

SumSquares / df

Three Different Sum of Squares



df = 1 (linear model)

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

SS_M

Model SumSquares

Squared difference between predicted Y values from regression model and mean of Y data

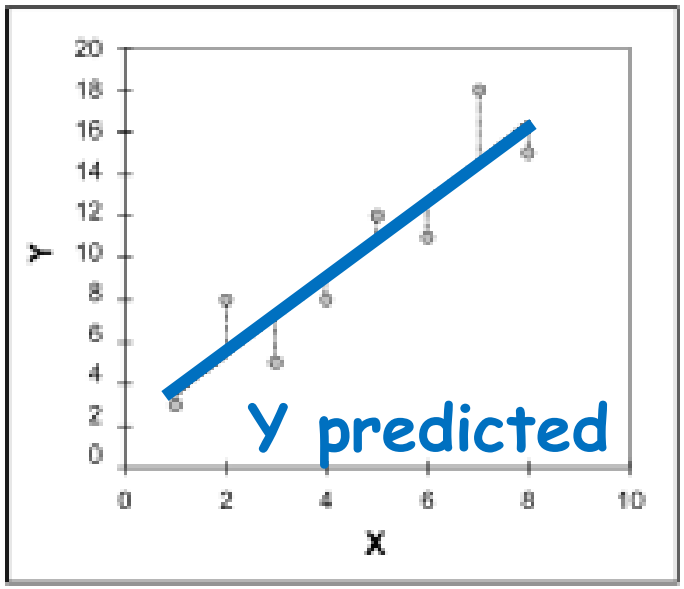
Sum Squares:

Σ (Y predicted – Y mean) ^2

Mean Squared:

SumSquares / 1

Three Different Sum of Squares



df = sample size - 2

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

SS_R

Residual (Error) SumSquares

Squared difference between predicted Y values from regression model and observed Y data

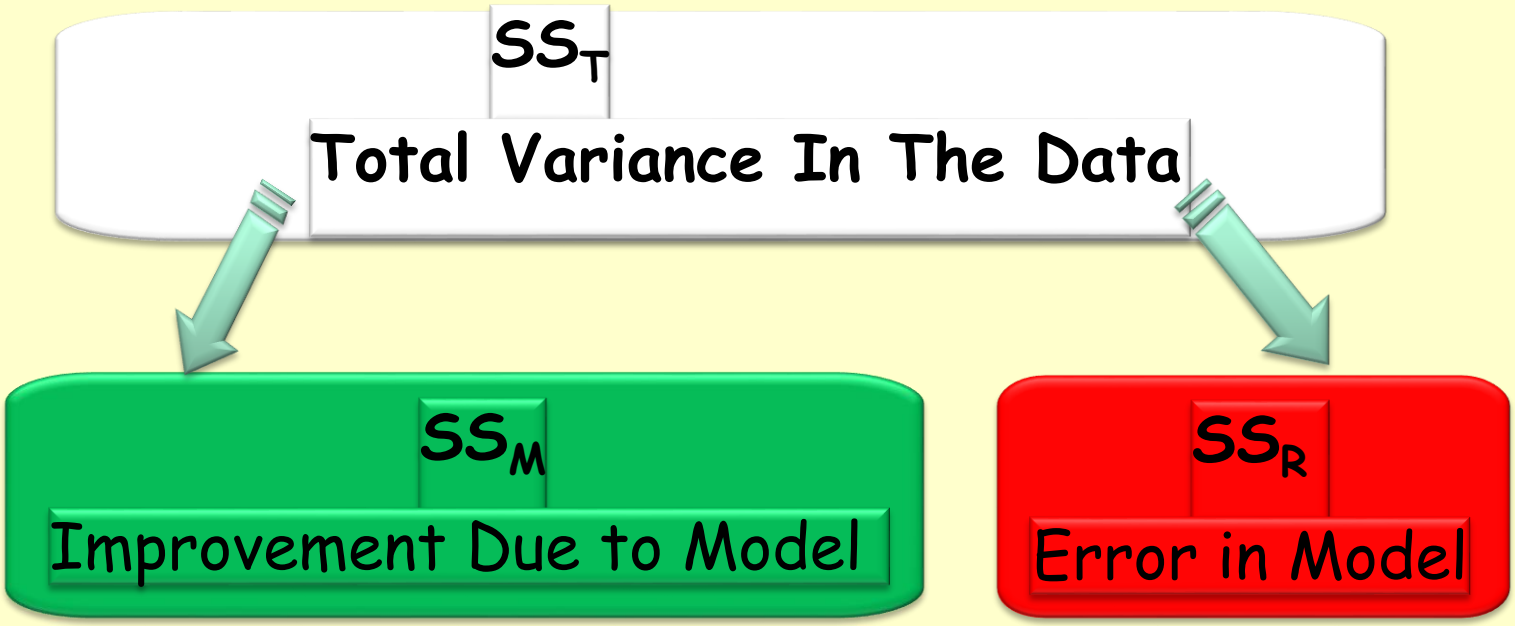
Sum Squares:

Σ (Yi - Y predicted) ^2

Mean Squared:

SumSquares / df

Testing the Model - ANOVA



If model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

Linear Regression -Output

Variable List

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

Sum Squares

Model		Sum of Squares
1	Regression	433687.833
	Residual	862264.167
	Total	1295952.000

SS_M

SS_R

SS_T

Mean Squares

Model	Sum of Squares	df	Mean Square
Regression	433687.833	1	433687.833
Residual	862264.167	198	4354.870
Total	1295952.000	199	

MS_M

MS_R

Testing the Model: R squared

R^2 - Coefficient of Determination

The proportion of total variance accounted for by the regression model

Ranges from
0 (none) to 1 (all)

$$R^2 = \frac{SS_M}{SS_T}$$

Testing the Model: F Test

F statistic: Mean Squares Ratio

- Sums of Squares: sums of squared deviations
- Calculate averages called Mean Squares, MS

F statistic =
ratio of model MS
(regression variance)
divided by residual MS
(error variance)

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

MODEL

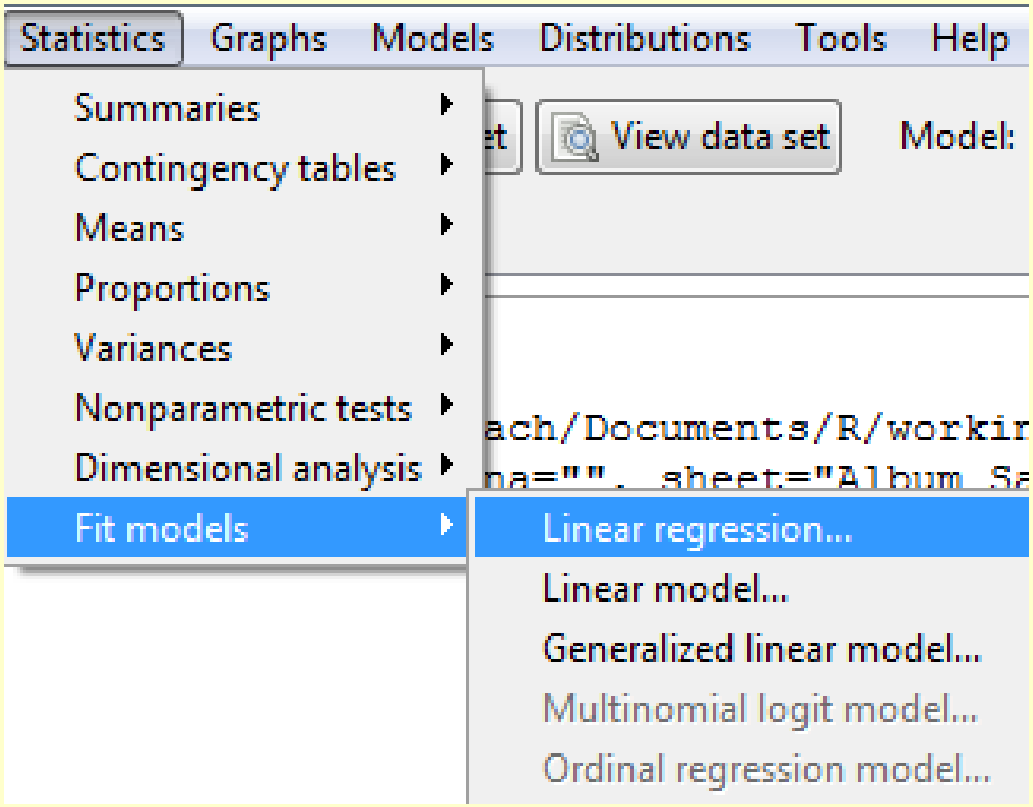
ERROR

NOTE: F ranges from 0 to a very large number
The larger the F value, the stronger model

Linear Regression: An Example

- A record company boss is interested in predicting record sales from advertising.
- Data: AlbumSales.xlsx
  - 200 different album releases
- Outcome variable:
  - Sales in first week after release
- Predictor variable:
  - Amount (£s) spent promoting record before commercial release

Regression in Rcmdr: How to



Statistics /
Fit Models /
Linear regression

- Linear Regression:
  - 1 Response Variable, 1 (or more) Explanatory Variables
- Linear Model:
  - 1 Response Variable, 1 (or more) Explanatory Variables

Regression in Rcmdr: How to

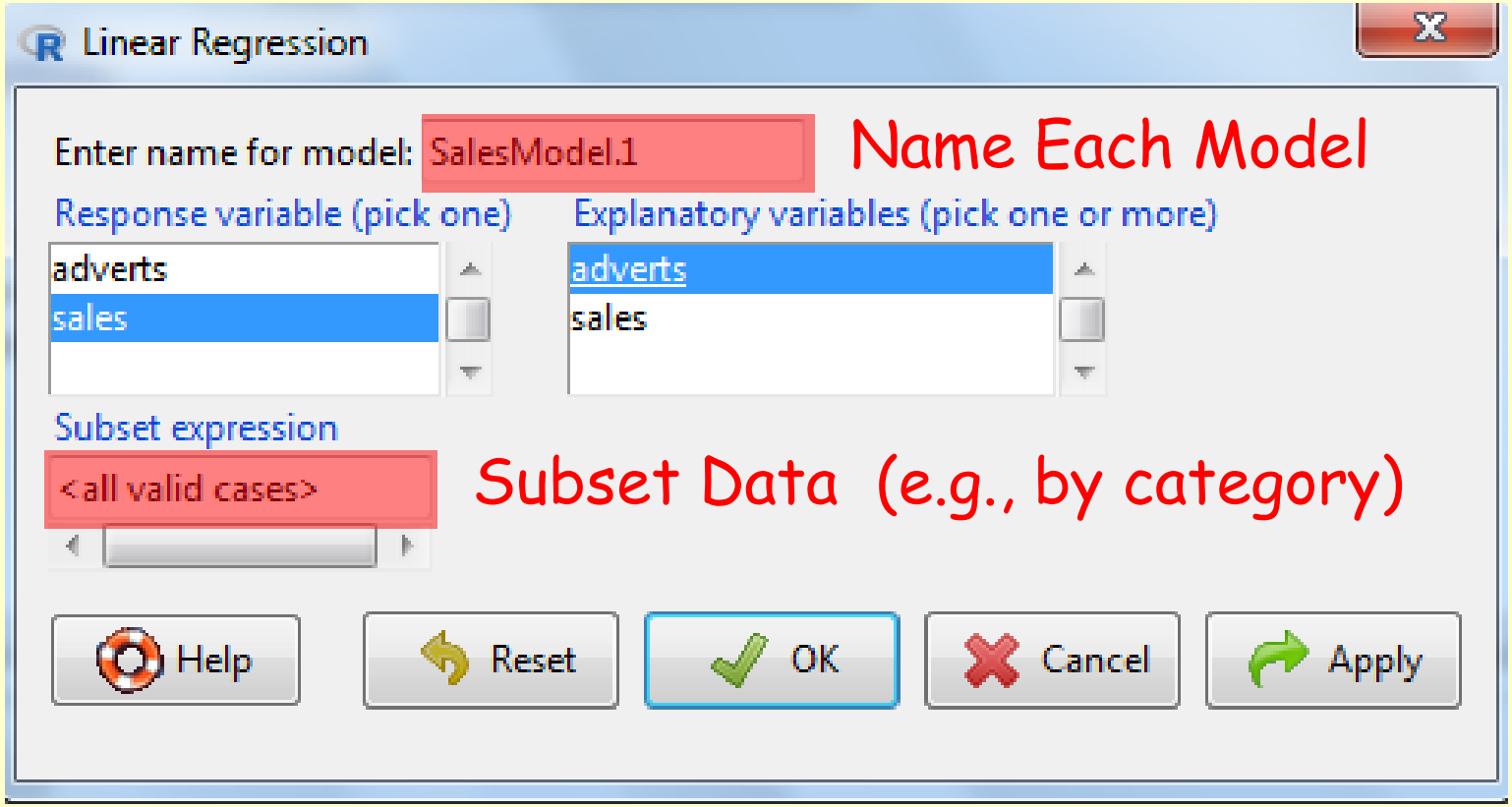
In Rcmdr: Statistics / Fit Models / Linear regression

- We run a regression analysis using the *lm()* function
  NOTE: lm stands for 'linear model'.

- This function takes the general form:
  newModel<-lm(outcome ~ predictor(s),
  data = dataFrame, na.action = an action))

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

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

Linear Regression in Rcmdr: How to



Name the Model – to track the results

Select Dependent / Independent Variables

Linear Regression in Rcmdr: How to

```
> SalesModel.1 <- lm(sales~adverts, data=sales)

> summary(SalesModel.1)
```

Created a new object

- Output:

- Call: lm(formula = sales ~ adverts, data = sales)
  (linear model, sales is a function of adverts)

- Residuals: (Observed – Model)
  Distribution of model residuals

Min	1Q	Median	3Q	Max
-152.949	-43.796	-0.393	37.040	211.866

Regression in Rcmdr: How to

- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	134.139938	7.536575	17.799	<2e-16 ***
adverts	0.096124	0.009632	9.979	<2e-16 ***

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

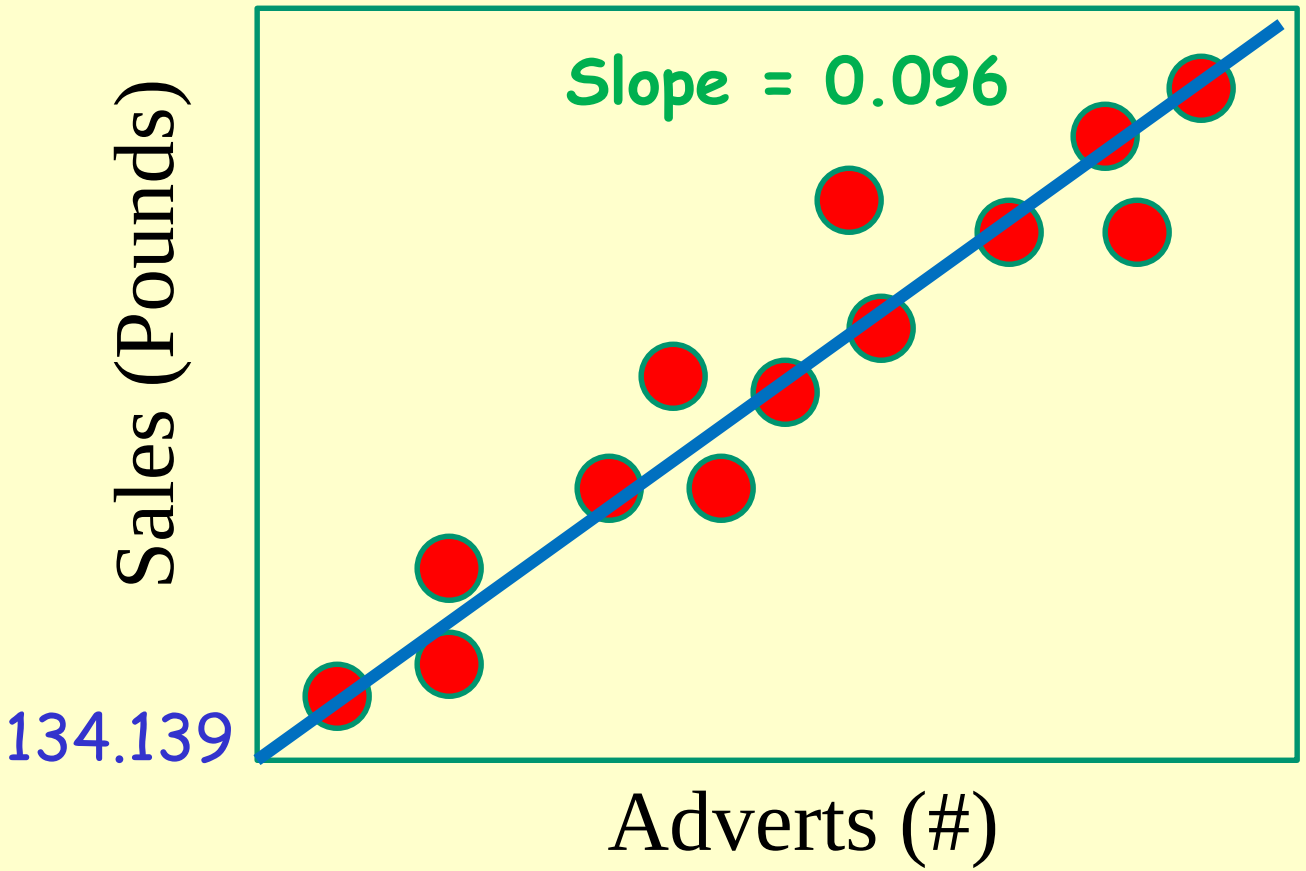
Note: Coefficients used to build linear equation (Y = a + BX)

Sales = 134.139 (+/- 7.53 S.E.)+(0.096 (+/- 0.009) * Adverts)

Regression in Rcmdr: How to

Sales = 134.139 (+/- 7.53 S.E.) + (0.096 (+/- 0.009) * Adverts)

Int. 95% C.I. = 134.139 +/- 14.759 = 148.898 to 119.380
Slope 95% C.I. = 0.096 +/- 0.017 = 0.11364 to 0.07836



Alternate Hypothesis (slope ≠ 0)

Is the intercept significant?

Is the slope significant?

Regression in Rcmdr: How to

- R-squared, F-statistic, p value:

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

Regression in Rcmdr: How to

- Multiple R^2: 0.3346

Proportion of variance in the sample explained by model.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

Equal to squared Pearson correlation coefficient.

- Adjusted R^2: 0.3313

Measure of loss of predictive power (shrinkage in regression).

$$R^2_{adj} = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

Tells us how much variance would be accounted for if model had been derived from the population from which the sample taken. Adjusted by: n and p. (n = sample size) (p = independent variables)

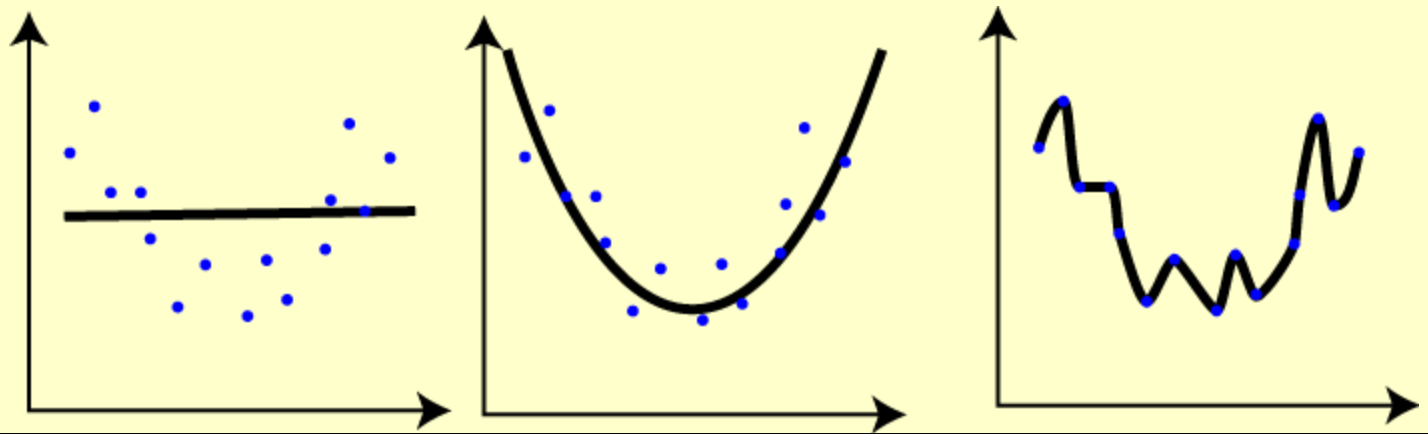
Regression in Rcmdr: How to

- Why use the Adjusted R-squared?

Because R^2 quantifies how well a model fits the data, we could easily pick the model with the larger R^2, the best fit.

Model with more parameters will be able to bend and twist the best-fit line to come closer to the points, and will have a higher R^2. There is no penalty for adding more parameters.

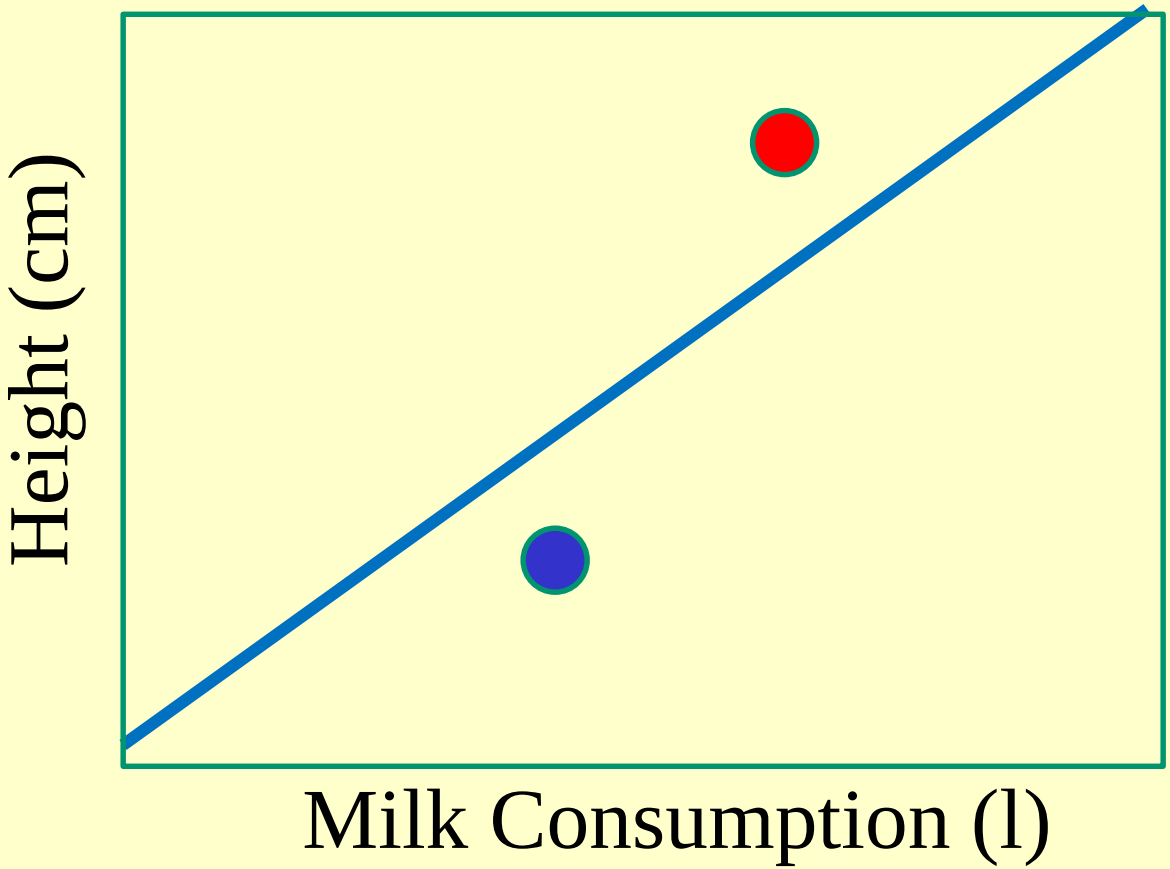
Thus, if you use R^2 as the criteria for picking the best model, you will usually pick the model with the most parameters.



Linear Regression Residuals

Output:

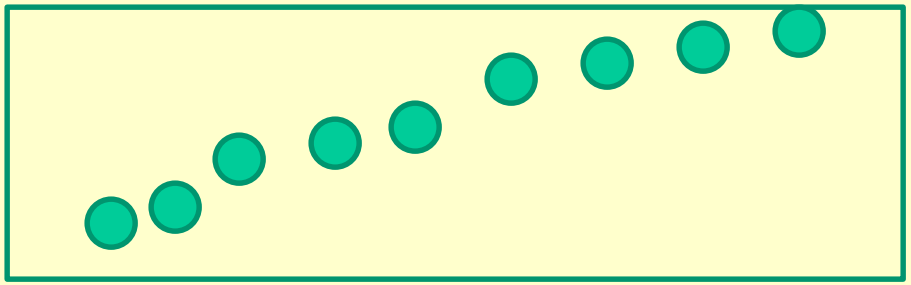
Residuals (look for normality) and perform test

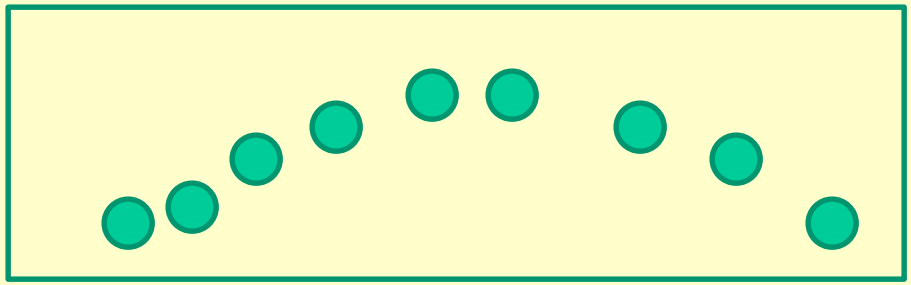


Residual:

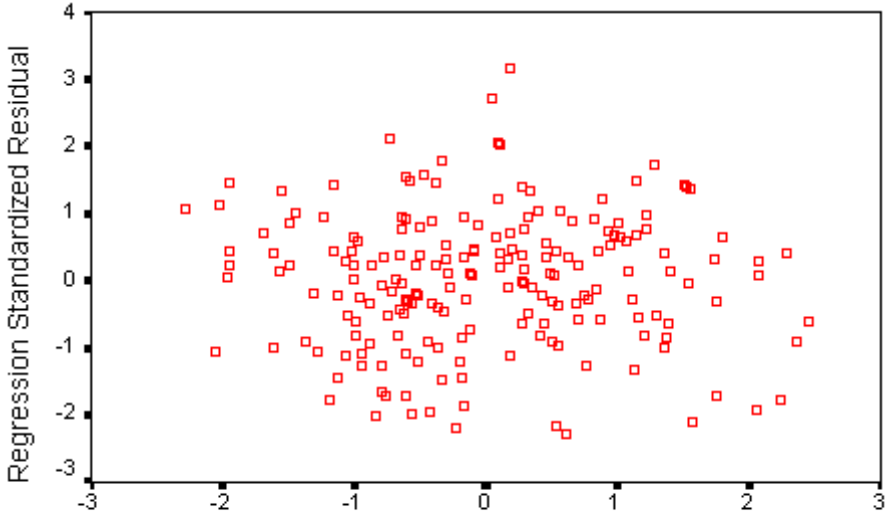
Yobs – Y model

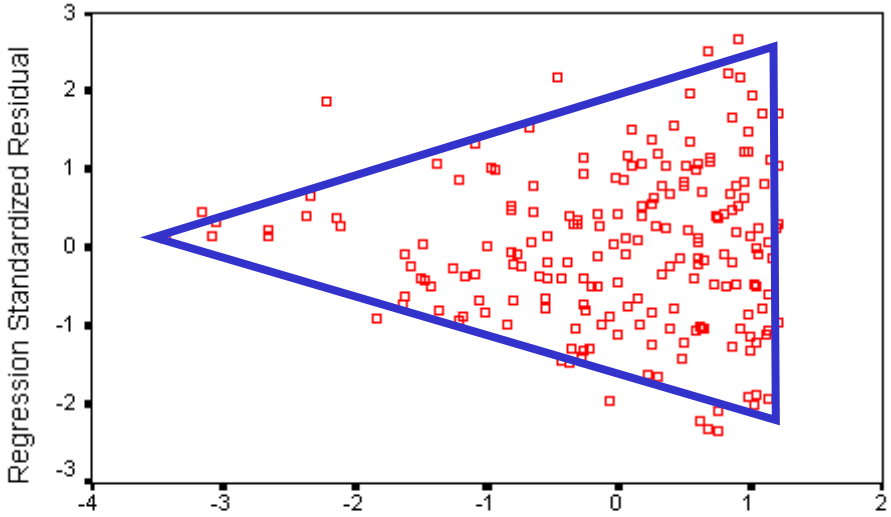
Independence & Homoscedasticity





Errors are not independent: obvious linear or non-linear patterns





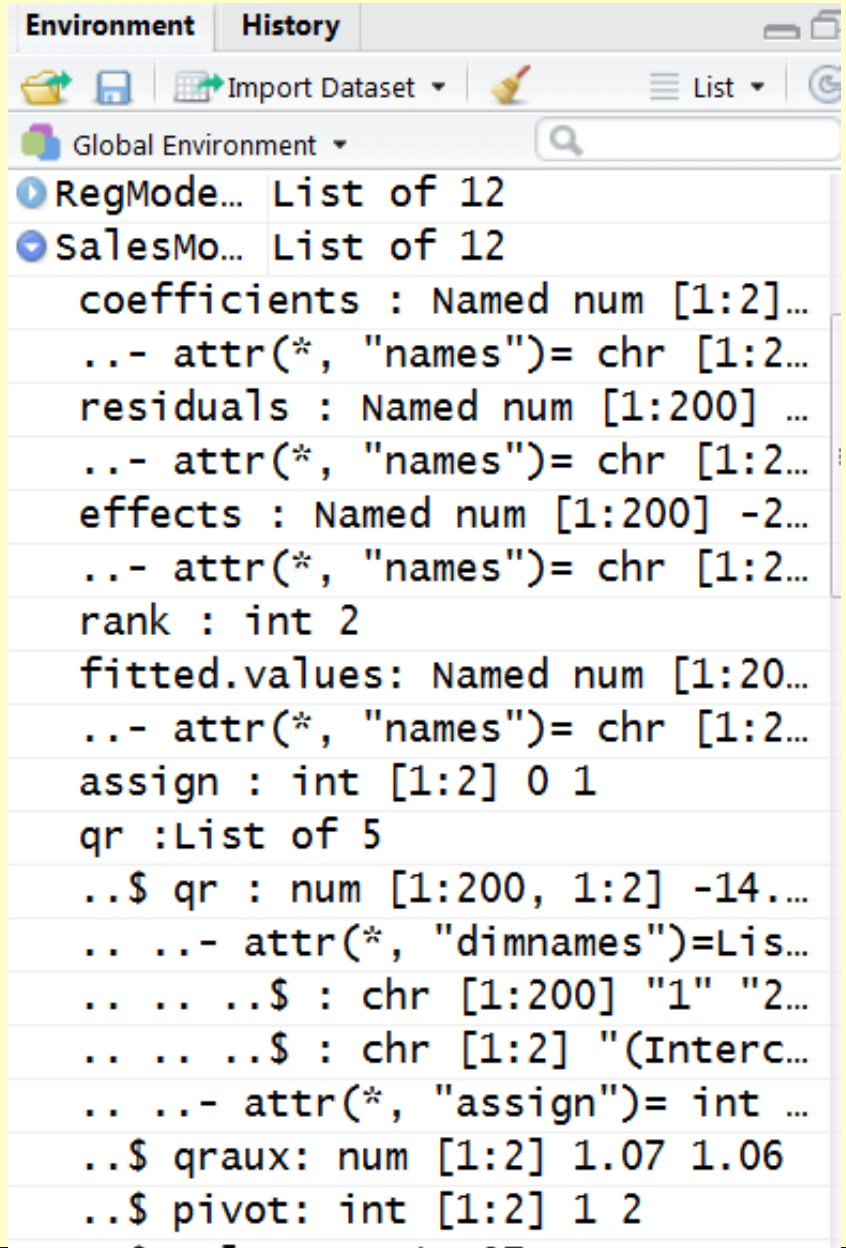
Error cloud: Equal variance with changing predicted value.

Triangular Pattern: Variance increases with predicted value.

Regression in Rcmdr: How to

- Testing for normality, after the regression:
  - Normality of Errors:
  - Homoscedasticity of Errors:
  - Independence of Errors:
  - Extracting more data from the regression output:

SalesModel.1 is an object made up of 12 datasets



Regression in Rcmdr: How to

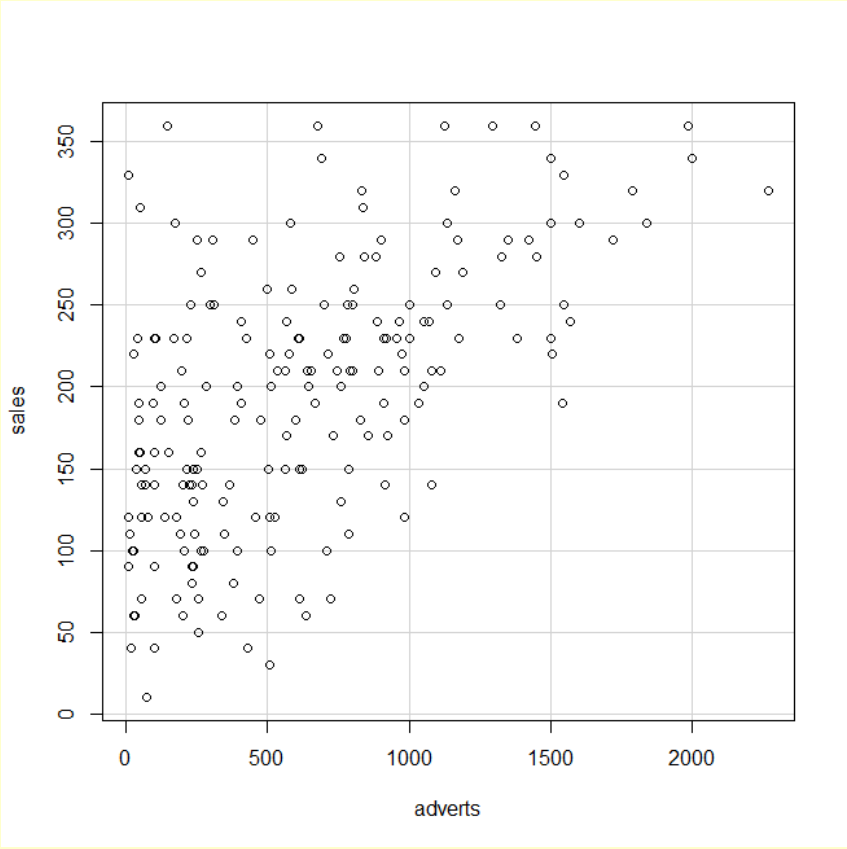
```
> SalesModel.1 <- lm(sales~adverts, data=sales)

> summary(SalesModel.1)    MODEL RESULTS

> anova(RegModel.1)          F TESTS RESULTS
```

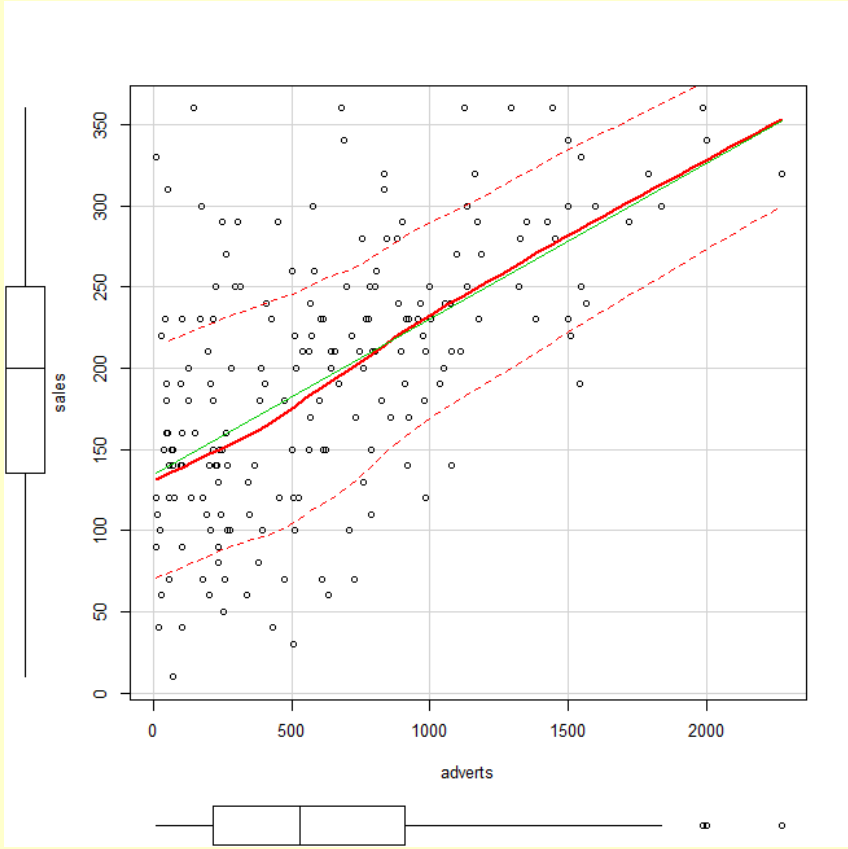
```
Analysis of Variance Table
Response:  sales
           Df  SumSq  MeanSq  Fvalue  Pr(>F)
adverts     1  433688 433688  99.587 <2.2e-16 ***
Residuals 198 862264   4355

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

Regression in Rcmdr: How to

```
> scatterplot(sales~adverts,
reg.line=FALSE, smooth=FALSE,
spread=FALSE, boxplots=FALSE,
span=0.5, ellipse=FALSE,
levels=c(.5, .9), data=sales)
```

```
> scatterplot
(sales~adverts,
data=sales)
```





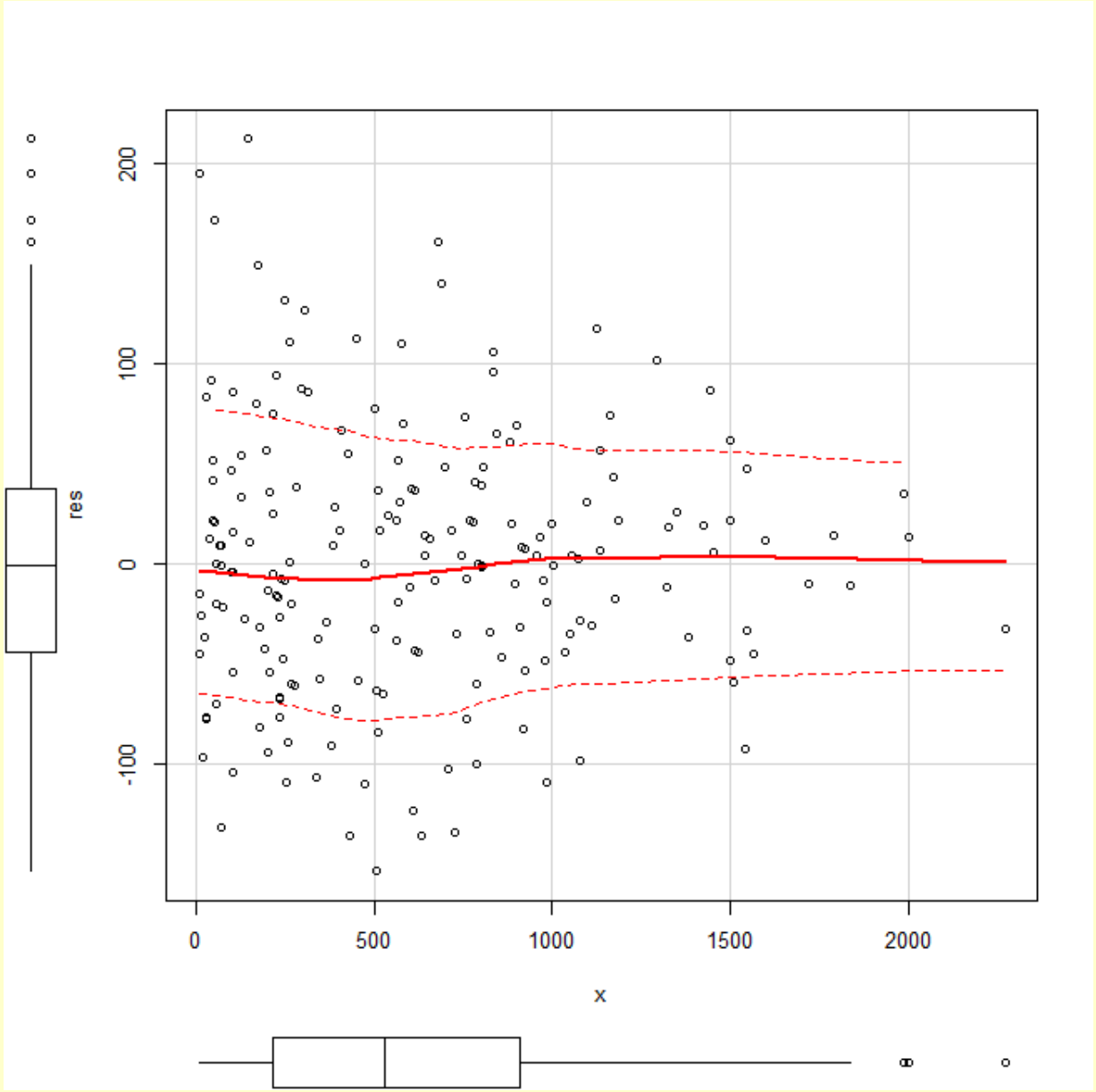
Regression in Rcmdr: How to

```
> res <-residuals(RegModel.1)

> normalityTest (~res, test="shapiro.test")

Shapiro-Wilk normality test

data:  res

W = 0.98995, p-value = 0.1757
```

Regression in Rcmdr: How to

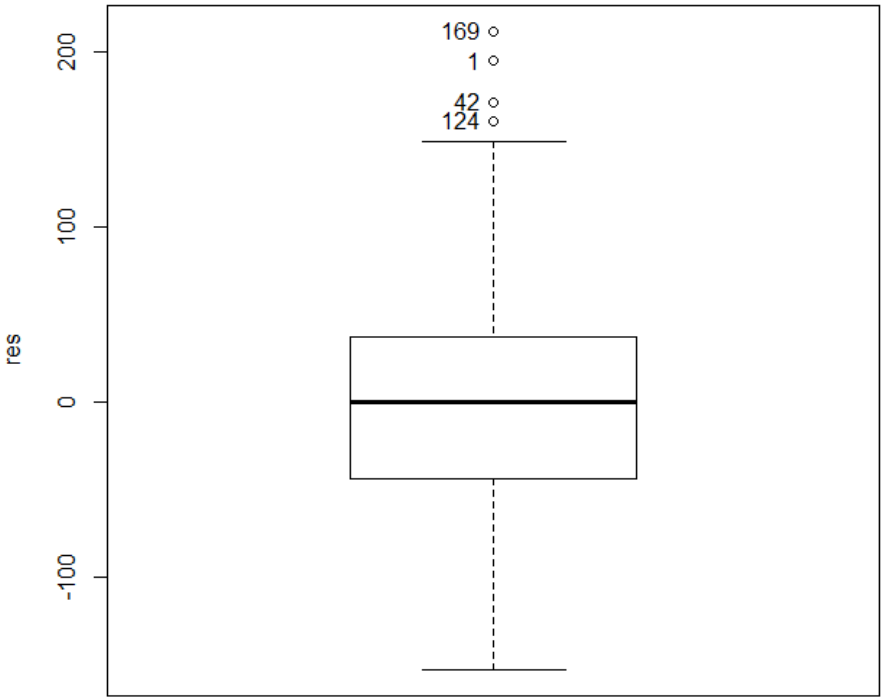
Scatterplot of the residuals (y axis) as a function of the adverts (x axis)

```
> x <- adverts

> y < res

> scatterplot
(res~x,
reg.line=FALSE
```

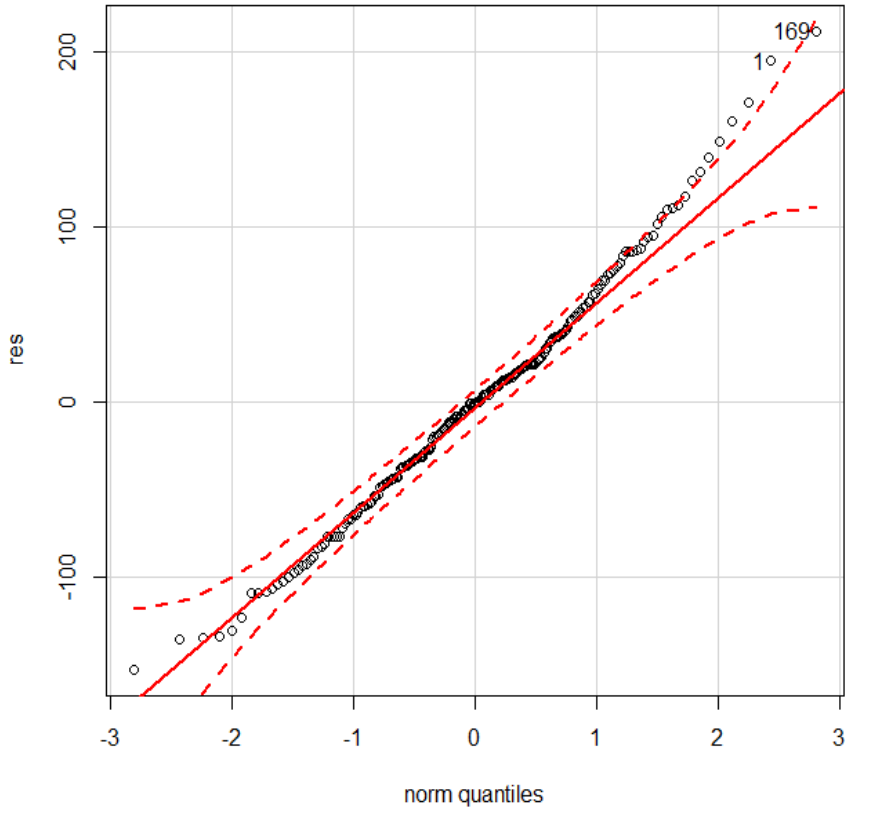


Regression in Rcmdr: How to

```
> Boxplot
( ~ res, id.method="y")
```




```
> qqPlot(res,
dist="norm",
id.method="y", id.n=2,
labels=rownames(sales))
```

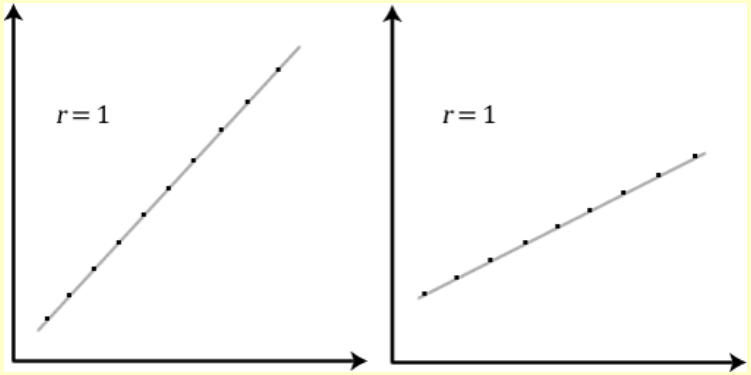


Correlation vs Regression

Main Differences: Pearson correlation is birectional. Regression is not.

Pearson correlation does not measure the slope of best-fit line. Regression does.

For example:
Correlation coefficient of +1 does not mean that for one unit increase in one variable there is one unit increase in the other.



Linear Regression – Summary

1. Define the explanatory variable as the independent variable (predictor), and the response variable as the dependent variable (predicted).

2. Plot the explanatory variable (x) on the x-axis and the response variable (y) on the y-axis, and fit a linear regression model ($y = b0 + b1x$), where $\beta 0$ is the intercept, and $\beta 1$ is the slope.

Note that the point estimates from the observations ($b0$ and $b1$) estimate the population parameters ($\beta 0$ and $\beta 1$), respectively.

3. Define residual (e) as the difference between the observed (y) and predicted ($\hat{y}$) values of the response variable. $Ei = yi - \hat{y}$

Linear Regression – Summary

4. Define the least squares line as the line that minimizes the sum of the squared residuals. Three conditions are necessary for fitting such line: (1) linearity, (2) nearly normal residuals, and (3) constant variability.

5. Define an indicator variable as a binary explanatory variable (with two levels).

6. Interpret the slope as follows:

• when x is numerical: "For each unit increase in x, we would expect y to be lower/higher on average by $|b1|$ units"

• when x is categorical: "The value of the response variable is $|b1|$ units higher/lower for the other level of the explanatory variable, compared to the baseline level."

Linear Regression – Summary

7. The least squares line passes through average of the response and explanatory variables ($\bar{x}$, $\bar{y}$). This allows us to calculate the intercept ($b0$) as follows: $b0 = \bar{y} - b1\,\bar{x}$, where $b1$ is the slope, $\bar{y}$ is the average of the response variable, and $\bar{x}$ is the average of the explanatory variable.

8. Interpret the intercept as:

- "When $x = 0$, we expect y to equal, on average, $b0$." when x is numerical.

- "Expected average value of y is equal to $b0$, the reference level of the explanatory variable." when x is categorical.

9. R^2 quantifies the proportion of the variability in y that is explained by the variability in x.