STATISTICS & ML WITH R

Modeling correlation and regression

2024

M. Chiara Mimmi & Luisa M. Mimmi

WORKSHOP SCHEDULE

- 4 days
  - 1. Intro to R and data analysis
  - 2. Statistical inference & hypothesis testing
  - 3. Modeling correlation and regression
  - 4. Machine Learning; MetaboAnalyst; Power Analysis

- Each day will include:
  - Frontal class (MORNING)
  - Practical training with R about the topics discussed in the morning. (AFTERNOON)

DAY 3 – LECTURE OUTLINE

- Testing and summarizing relationship between 2 variables (**correlation**)
  - Pearson's r analysis (parametric)
    - 2 numerical variables
  - Spearman test (not parametric)
    - 2 numerical variables (non linear relationships)
- Measures of **association**
  - Chi-Square test of independence
    - 2 categorical variables
  - Fisher's Exact Test
    - alternative to the Chi-Square Test of Independence
- Introduction of **regression analysis**
  - Simple linear regression models
  - Multiple Linear Regression models

Summarizing relationships between two variables

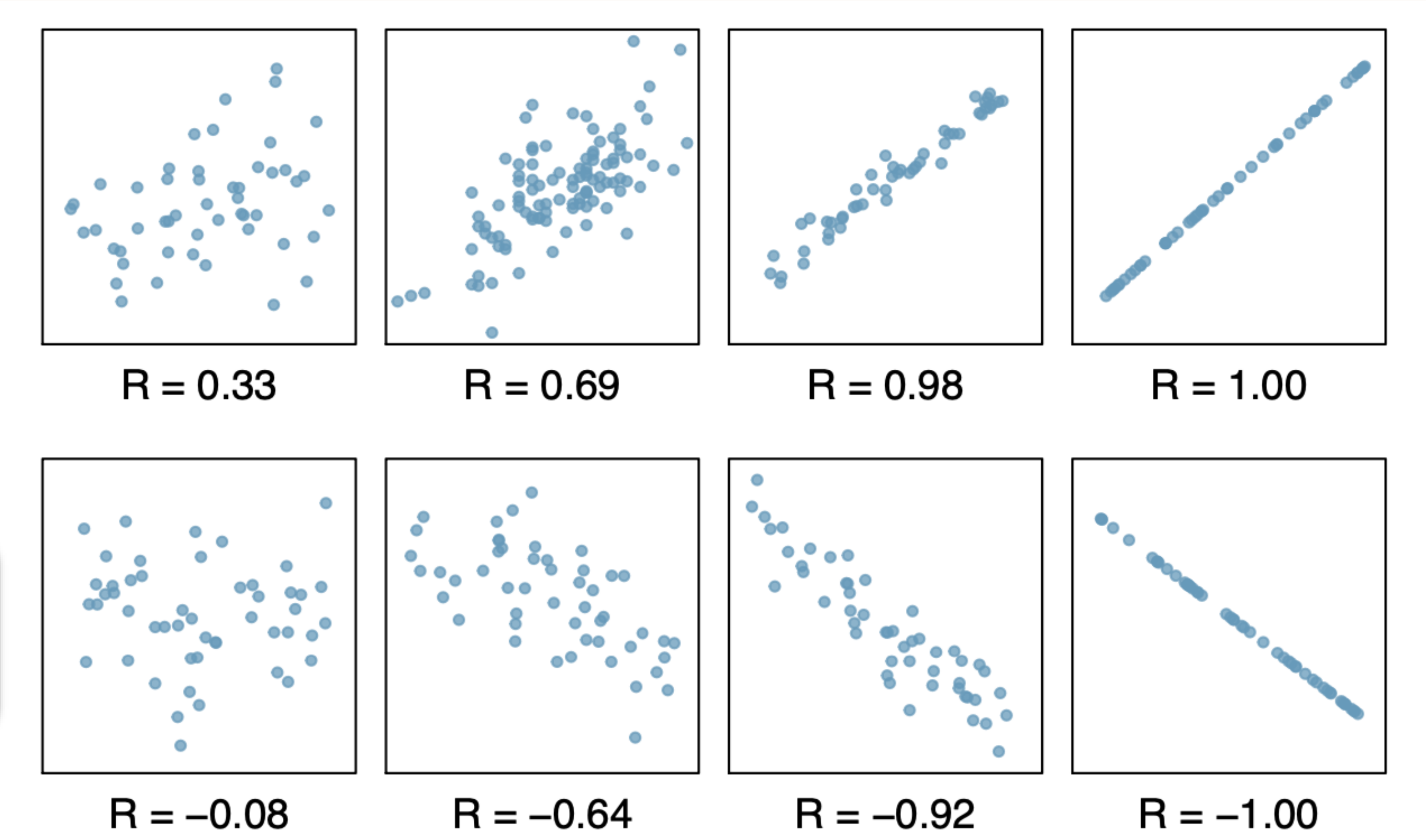
Correlation

Defining correlation

- Correlation is a numerical summary statistic that measures the *strength of a linear relationship* between two variables
  - denoted by **r** (correlation coefficient) which takes values between -1 and 1

positive correlation

negative correlation



Source: Vu, J., & Harrington, D. (2021). *Introductory Statistics for the Life and Biomedical Sciences*. Retrieved from https://www.openintro.org/book/biostat/

Most used measures of correlation

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r (ρ for population)	Linear	Two **quantitative** (**interval** or **ratio**) variables	Normal distribution
Spearman's rs (ρ for population)	Non-linear	Two **ordinal**, **interval** or **ratio** variables	Any distribution
Cramér's V **(Cramér's ϕ)**	Non-linear	Two **nominal** variables	Any distribution
Kendall's τ (tau)	Non-linear	Two **ordinal**, **interval** or **ratio** variables	Any distribution

What is the link between correlation and covariance?

- Covariance is another helpful statistic that **tells whether both variables vary in the same direction** (positive covariance) **or in the opposite direction** (negative covariance)
  - Unlike in correlation, there is no meaning of covariance numerical value **only sign is useful**
  - $Cov\ (X,Y)$ is > 0 --> (X,Y) vary in the same direction
  - $Cov\ (X,Y)$ is < 0 --> (X,Y) vary in the opposite direction
  - $Cov\ (X,Y)$ is ~ 0 --> (X,Y) vary independently from each other

- The general formula for Covariance is:

$$Cov\ (X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Interesting to note that:

$$Cor\ (X,Y) = \frac{Cov\ (X,Y)}{s_x s_y}$$

  - where s_x is the standard deviation of x and s_y is the standard deviation of y
  - dividing Covariance by $s_x s_y$, we obtain Correlation **r** with range [-1, +1]

Correlation between 2 numerical variables

Pearson's correlation (parametric test)

Pearson's correlation

Pearson correlation ($\mathbf{r}$) measures a linear association between 2 CONTINUOUS variables (x and y) or 2 dichotomous variables

- It's also known as a parametric correlation test because it depends to the distribution of the data.
- The Pearson correlation evaluates the linear relationship between two continuous variables.

FORMULA

$$\mathbf{r} = \frac{\sum(\mathrm{x} - \mathrm{m_x})(\mathrm{y} - \mathrm{m_y})}{\sqrt{\sum(\mathrm{x} - \mathrm{m_x})^2 \sum(y - m_y)^2}}$$

WHERE:

x and y are two vectors of length n

m_x and m_y correspond to the means of x and y, respectively.

We can test the statistical significance of the correlation statistic as well.

The p-value (significance level) of the correlation can be determined by calculating

$\mathrm{t\ value} = r\,\frac{\sqrt{\mathrm{n-2}}}{\sqrt{(1-\mathrm{r}^2)}}$ with $d.f. = (n-2)$

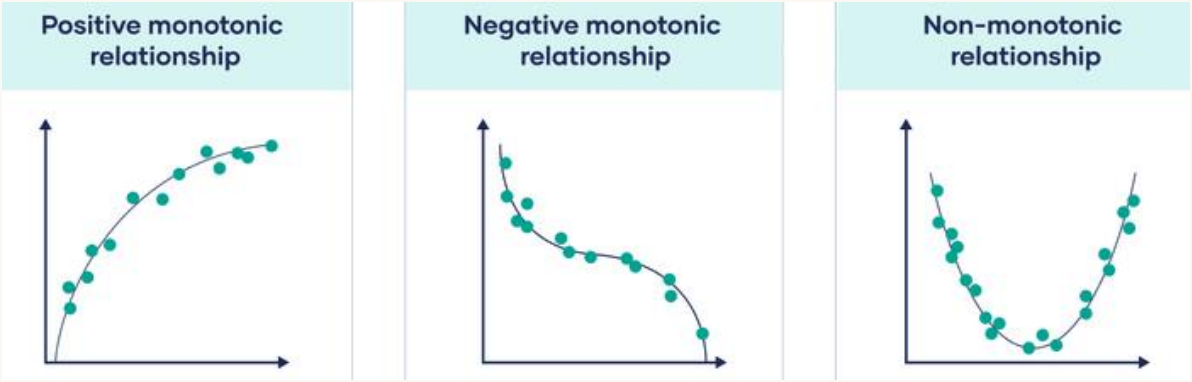
Correlation between 2 numerical variables

Spearman’s correlation (non parametric test)

Spearman's rank order correlation coefficient

Spearman's correlation (r_s or $\rho(rho)$) is a nonparametric alternative to Pearson's correlation, used for

- continuous data with a **non linear**, **monotonic** relationships, or
- **ordinal** data (e.g. Likert scale survey questions*: strongly agree, agree, etc.*)



FORMULA

$$\rho = r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

- where:
  - r_s is Spearman's coefficient of rank correlation.
  - d_i is the difference between the ranks for each (x, y) pair.
  - n is the number of paired observations.

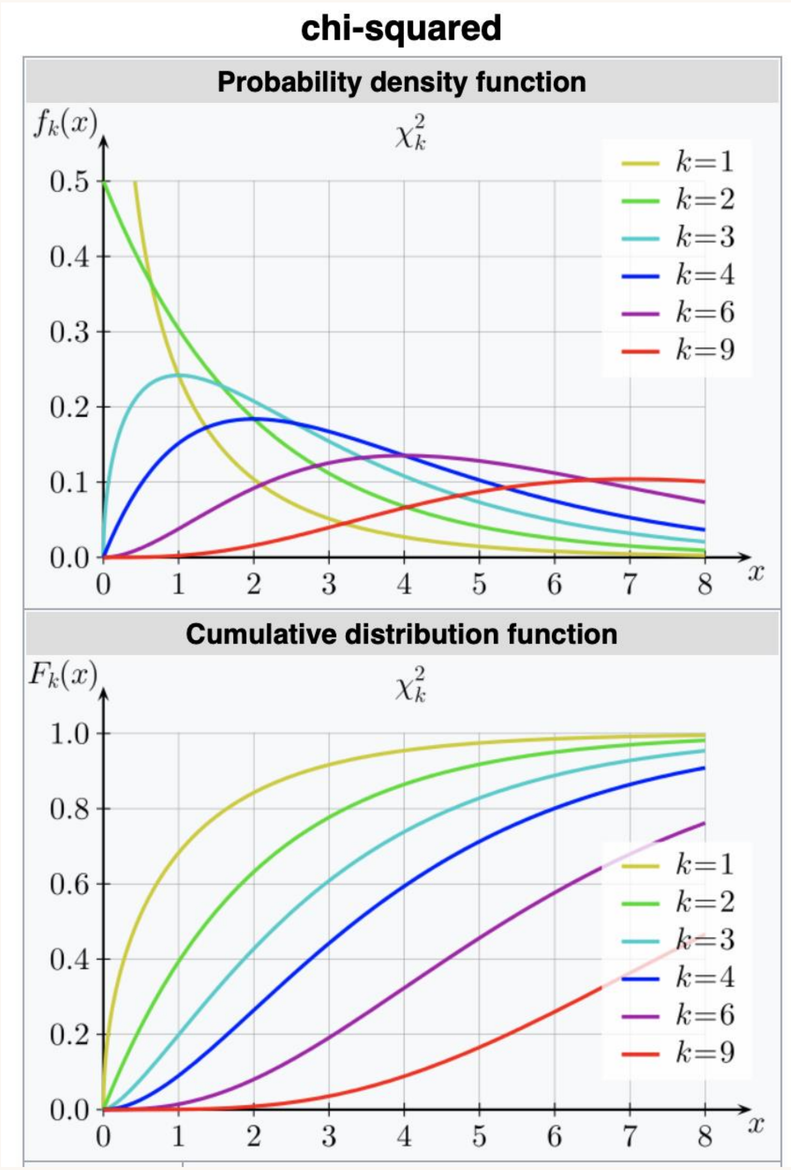
Hypothesis Test: Rank Correlation

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

Chi Squared Distributions

A widely used analytical tool

The chi-squared (χ_k^2) distribution



- The **chi-squared distribution** (χ_k^2) is a family of continuous probability distributions
- It results from the sum of squares of *k* normally distributed random variables, where *k* is the number of degrees of freedom (*df*)
- The mean is equal to the *df* and the variance is equal to 2 x *df*

Notation	$\chi^2(k)$ or χ_k^2
Parameters	$k \in \mathbb{N}^*$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$

Applications of the Chi-Square Test

- Unlike the Normal distribution, very few real-world observations follow a chi-square distribution, but it is used extensively in hypothesis testing (also due to its close relationship with the normal).
  - As k increases, the χ_k^2 distribution looks more and more similar to a normal distribution

- The **Chi-square test** helps to answer the following questions:

1. Independence test
- Are two categorical variables independent of each other?
  - for example, does gender have an impact on whether a person has a Netflix subscription or not?

2. Distribution (or **Goodness of fit) test**
- Are the observed values of two categorical variables equal to the expected values?
  - One question could be, is one of the three video streaming services Netflix, Amazon, and Disney subscribed to above average?

3. Homogeneity test
- Are two or more samples from the same population?
  - One question could be whether the subscription frequencies of the three video streaming services Netflix, Amazon and Disney differ in different age groups.

Correlation between 2 categorical variables

Chi Squared test of independence

A useful tool for categorical variables: contingency tables

- A contingency table summarizes data for 2 categorical variables (each value in the table representing the times a particular combination of outcomes occurs)
- Below we see 2 categorical variables "**gender**" (male, female) and "**has Netflix subscription**" (yes, no)

Frequency			SUM
	Male	Female	
Netflix yes	10	13	**23**
Netflix no	15	14	**29**
SUM	**25**	**27**	

- The row totals (counts across each row) and the column totals (counts across each column) are the **marginal totals**
- Frequencies can also be shown as proportions

Computing the Chi-Square Test of Independence

- E.g. suppose we are testing the independence of the two categorical variables "**gender**" (male, female) and "**has Netflix subscription**" (yes, no)
- The test performs a comparison of these two contingency tables:

Observed Frequency

	Male	Female
Netflix yes	10	13
Netflix no	15	14

Expected Frequency

	Male	Female
Netflix yes	(23x25)/52 = 11.06	(23x27)/52 = 11.94
Netflix no	(29x25)/52 = 13.94	(29x27)/52 = 15.06

IMPORTANT ASSUMPTIONS TO NOTICE:

- The assumption for the **chi-squared (χ^2)** test statistic is that the expected frequencies per cell are > 5
- The **chi-squared (χ^2)** test uses only the categories but NOT rankings

Computing the Chi-Square Test of Independence (computation)

- Let's compute the example for the two variables "**gender**" and "**has Netflix subscription**"

Observed Frequency

	Male	Female
Netflix yes	10	13
Netflix no	15	14

Expected Frequency

	Male	Female
Netflix yes	(23x25)/52 = 11.06	(23x27)/52 = 11.94
Netflix no	(29x25)/52 = 13.94	(29x27)/52 = 15.06

- The **chi-squared (χ^2)** test statistic is calculated via:

$$\chi^2 = \sum_{K=1}^{n} \frac{(O_k - E_k)^2}{E_k} = \frac{(10-11.06)^2}{11.06} + \frac{(13-11.94)^2}{11.94} + \frac{(15-13.94)^2}{13.94} + \frac{(14-15.06)^2}{15.06} = 0.35$$

- where:
  - O_k = observed frequency and
  - $E_k = \textit{Expected frequency} = f(i,j) = \frac{RowSum(i) \times ColumnSum(j)}{N}$
    - calculated for each cell in the contingency table
- The test assumptions are:
  - H_0: (null hypothesis) The two variables are independent.
  - H_1: (alternative hypothesis) The two variables are not independent. (i.e. they are associated)
  - $\text{d.f.} = (n_{rows} - 1)(n_{col} - 1) = 1$

Interpreting the Chi-Square Test of Independence

- The **chi-squared (χ^2)** test statistic calculated value:

  $$\chi^2 = \mathbf{0.35}$$

- BY THE CRITICAL REGION: Looking at the χ^2 distribution, for a significance level of 5% and a *df* of 1, the **critical chi-squared value = 3.841**
  - → Since the calculated chi-squared value=**0.35** is smaller, we **FAIL TO REJECT the null** (*H_0: The two categorical variables are independent*)

- BY THE p VALUE: Also, the **p-value** associated to the $\chi^2 = 0.35$ and

  $\mathrm{d.f.} = (n_{rows}-1)(n_{col}-1) = 1$ is **0.5541**.
  - → Since this p-value is not less than 0.05, we fail to reject the null hypothesis.

- This means we do not have sufficient evidence to say that there is an association between gender and political having a Netflix account!

Chi Squared test (another application)

Goodness of Fit Test for one categorical variable

Chi-Square Goodness of Fit Test

- GOAL: a Chi-Square goodness of fit test is used to **determine whether or not a categorical variable follows a *hypothesized distribution***.
  - With **high** goodness of fit, the values expected based on the model are **close to** the observed values
  - With **low** goodness of fit, the values expected based on the model are **far from** the observed values

- EXAMPLES OF APPLICATION:
  - *Is this sample drawn from a population with 90% right-handed and 10% left-handed people?*
  - *Do offspring have with an equal probability of inheriting all possible genotypic combinations (i.e., unlinked genes)?*

- HYPOTHESIS FORMULATION
  - Null Hypothesis **(H_0):** The population follows the specified distribution.
  - Alternative Hypothesis **(H_a):** The population does not follow the specified distribution.

Chi-Square Goodness of Fit Test (computation)

- FORMULA: The formula is essentially the same as in the independence test

$$\chi^2 = \sum_{K=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

- where O_k = Observed Frequencies and E_k = Expected Frequencies
- … with $df = n - 1$ (number of groups minus 1)

- WHEN SHOULD WE USE IT? (assumptions)
  1. We are testing the distribution of **one categorical variable**
     - if you have a continuous variable, it should be converted to categorical (this is called *data binning*) or a different test can be used (like the Kolmogorov–Smirnov goodness of fit test for continuous variables)
  2. The sample was randomly selected from the population.
  3. There are a minimum of 5 observations expected in each group.

Chi-Square Goodness of Fit Test (example)

- GOAL: examine the appropriateness of hypothesized distribution for a dataset
- CASE: In the FAMuSS study (we'll see later in the lab) volunteers were observed at a university, so we test if their distribution by categorical variable race is the same as (i.e. *representative of*) the general US population?

Race	African.American	Asian	Caucasian	Other	Total
FAMuSS (Observed)	27	55	467	46	595
US Census (Expected)	76.16	5.95	478.38	34.51	595

$$\chi^2 = \sum_{K=1}^{n} \frac{(O_k - E_k)^2}{E_k} = \frac{(27 - 76.16)^2}{76.16} + \frac{(55 - 5.95\ )^2}{5.95} + \frac{(467 - 478.38)^2}{478.38} + \frac{(46 - 34.51)^2}{34.51} == 440.18$$

- where O_k = Observed Frequencies and E_k = Expected Frequencies
- … with $\mathrm{k} = 4,$ and $df = n - 1 = 3$ (number of groups minus 1)

- The χ^2 statistic is extremely large, and the associated p-value < 0.001 →
- We **reject the null hypothesis** ($H_0 =$the sample proportions should equal the population proportions)… in fact, we can see for example the higher Asian representation in sample

Correlation between 2 categorical variables - Fisher's Exact Test

(alternative to the Chi-Square Test of Independence)

Fisher's Exact Test

- Fisher's Exact Test is used to determine whether or not there is a significant association between two categorical variables.

- It is typically used as an alternative to the Chi-Square Test of Independence when one or more of the cell counts in a 2×2 table is less than 5.

- Fisher's Exact Test uses the following null and alternative hypotheses:
  - H_0: (null hypothesis) The two variables are independent.
  - H_1: (alternative hypothesis) The two variables are not independent.

Calculate **effect size** after a Chi-Square Test

3 alternatives to assess “strength” of the association (if any)

Three Ways to Calculate Effect Size for a Chi-Square Test

- So, we have seen 2 commonly used **Chi-Square tests**:
  - **Chi-Square Test for Independence**: Used to determine whether or not there is a significant association between two categorical variables from a single population.
  - **Chi-Square Test for Goodness of Fit**: Used to determine whether or not a categorical variable follows a hypothesized distribution

- For both of these tests, we obtain a p-value that tells us "*if*" an association is found (i.e. we should reject the null hypothesis of the test or not).
- Then, we may wonder about the **effect size** of the test (i.e. "*how strong*" an association is)

- There are 3 ways to measure effect size:
  1. **Phi** (ϕ)
     - for 2 x 2 contingency table
  2. **odds ratio** (OR)
     - for 2 x 2 contingency table
  3. **Cramer's V** (V)
     - for larger tables
       - example in lab

Correlation between… 1 numerical variable and 1 categorical variables

… we have actually met before 😉

Correlation between 1 numerical variable and 1 categorical variables

- Recall that we have already encountered methods for for comparing numerical data across groups in the previous lessons
    1. Using **side-by-side boxplots** for visual comparison of how the distribution of a numerical variable differs by category
    2. Using **One-Way ANOVA** for testing relationships between Numerical and Categorical variables
        - i.e. the extension of the t-test for more than 2 groups

When and why do Regression Analysis?

- Regression the most widely used method of comparison in data analysis. I can be specified as:
  1. Simple Regression Analysis uncovers mean-dependence between 2 variables
  2. Multiple Regression Analysis involves more variables

- Regression is a method used for different purposes:
  1. In CAUSAL ANALYSIS: to uncover the effect of one variable on another variable
  2. In PREDICTIVE ANALYSIS : to assess what to expect of a variable for various values of another variable

…more on this distinction coming up in Lecture #?

Regression Analysis

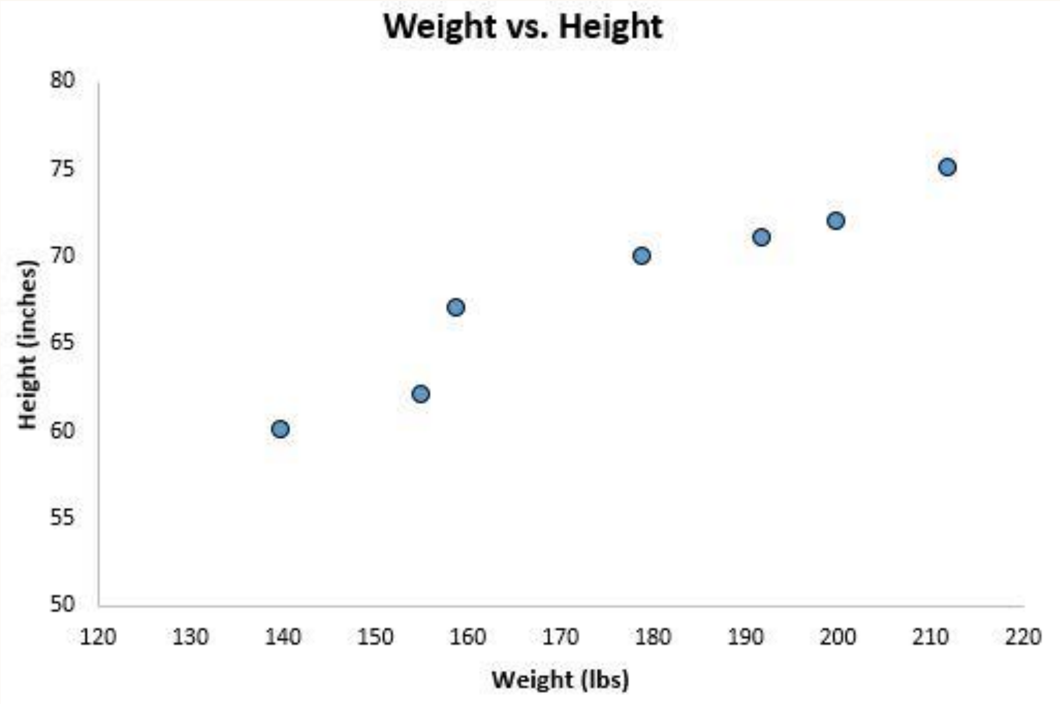
Simple Linear Regression

Regression analysis is a widely used method for prediction and – *given the proper experimental conditions* – for causal explanation

Simple linear regression: example

- Regression models are highly valuable, as they are one of the most common ways to make inferences and predictions
- Linear regression is OK with data that exhibit linear or approximately linear relationships
- **Simple linear regression** is a statistical method you can use to understand the relationship between two variables, x (the predictor variable) and y (the response variable)

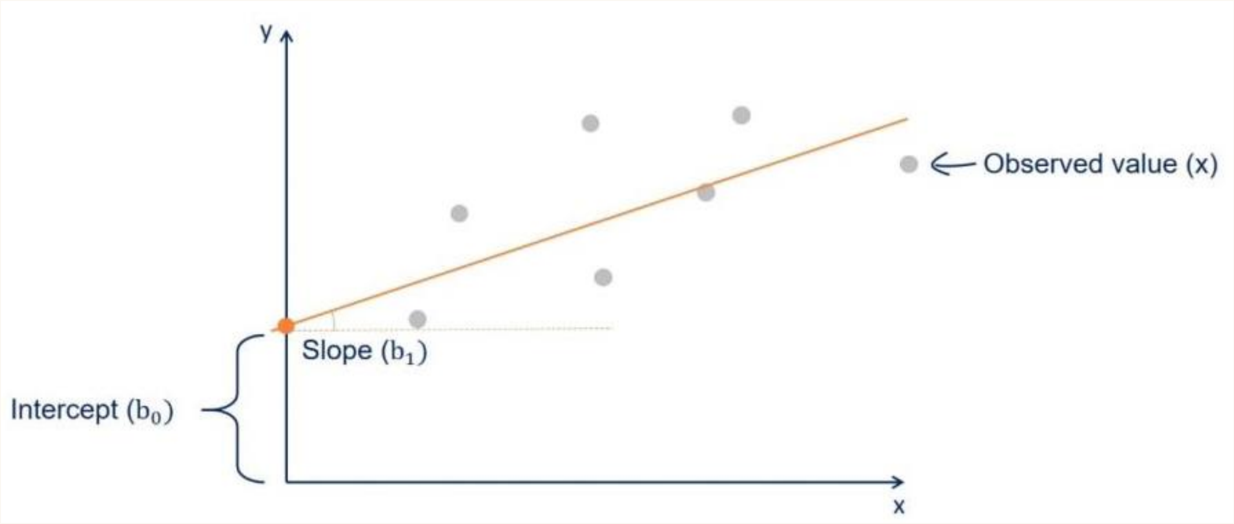
Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75



Functional (linear) relationship and regression

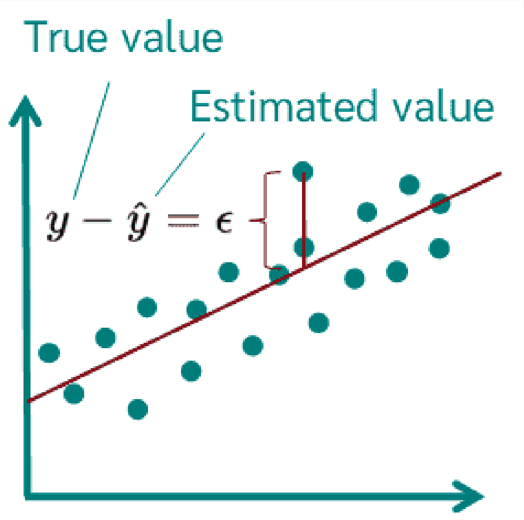
- The correlation coefficient gave us information about the degree to which points (corresponding to x and y pairs) were clustered around a straight line … but nothing about the slope of that line
- **regression analysis**, instead, provides this kind of information:
  - we want to know *exactly how* those 2 variables are related
  - (given we hypothesized a linear relationship) the model has a functional form that provides an **intercept** and a **slope**:

$$y = \boldsymbol{b_0} + \boldsymbol{b_1} x$$



Linear regression (Ordinary Least Square)

- The **OLS regression line** is chosen as to **minimize the difference between estimated values and actual ones**
  - in fact, OLS seeks the minimum *sum of squared distances between each point and the regression line*
  - it is the "best fitting" line given any data of points
- **NOTE**: like with previous **inferential statistics methods**, we are making statements on the population of interest based on some sample data available



Population data of interest	Sample data we have
$y = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = b_0 + b_1 x + e$
y = *true* Y values (dependent/response variable)	$\hat{y}$ = ***estimated*** (or predicted) Y values based on X values
x = *true* X values (independent/explanatory variable)	x = *sample* X values
β_0 = *true* intercept	b_0 = *estimated* intercept
β_1 = *true* slope/coefficient on x	b_1 = *estimated* slope/coefficient on x
ε = *true* residual or unobserved part of y	e = *estimated* residual (error),or unobserved part of Y)

OLS Linear regression interpretation

- The formula for the line of best fit is written as:

$$\hat{y} = b_0 + b_1 x + e$$

  - where $\hat{y}$ is the predicted value of the response variable (height), b_0 is the **y-intercept**, b_1 is the **regression coefficient**, and x is the value of the predictor variable (weight).

- For example, in the case of :

$$\hat{y} = 32.7830 + 0.2001x$$

  - b_0 = **32.7830**. This means **when the predictor variable weight is 0 pounds, the predicted height is 32.7830 inches**.
    - Sometimes the value for b_0 can be useful to know, but not in this specific example
  - b_1 = **0.2001**. This means that for **a one unit increase in the x variable, the y variable is predicted to increase(decrease) by 0.2001 units**. Here, a one pound increase in **weight** is associated with a 0.2001 inch increase in (**expected height**), on average.
    - NOTE: just like with previous hypothesis testing on sample means etc., we are testing the coefficients (b_0 and b_1) for statistical significance under H_0: the coefficient = 0

Assumptions of linear regression

- For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:

1. **Linear relationship**: There exists a linear relationship between the independent variable, x, and the dependent variable, y
2. **Normality**: The residuals of the model are normally distributed.
   - Check normality (OF RESIDUALS) with the known methods (QQplot, Shapiro-Wilk, Kolmogorov Smirnov)
3. **Homoscedasticity**: The residuals have constant variance at every level of x.
4. **Independence**: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
   - This is mostly relevant when working with time series data. Ideally, we don't want there to be a pattern among consecutive residuals.

Diagnostic plotting: residuals

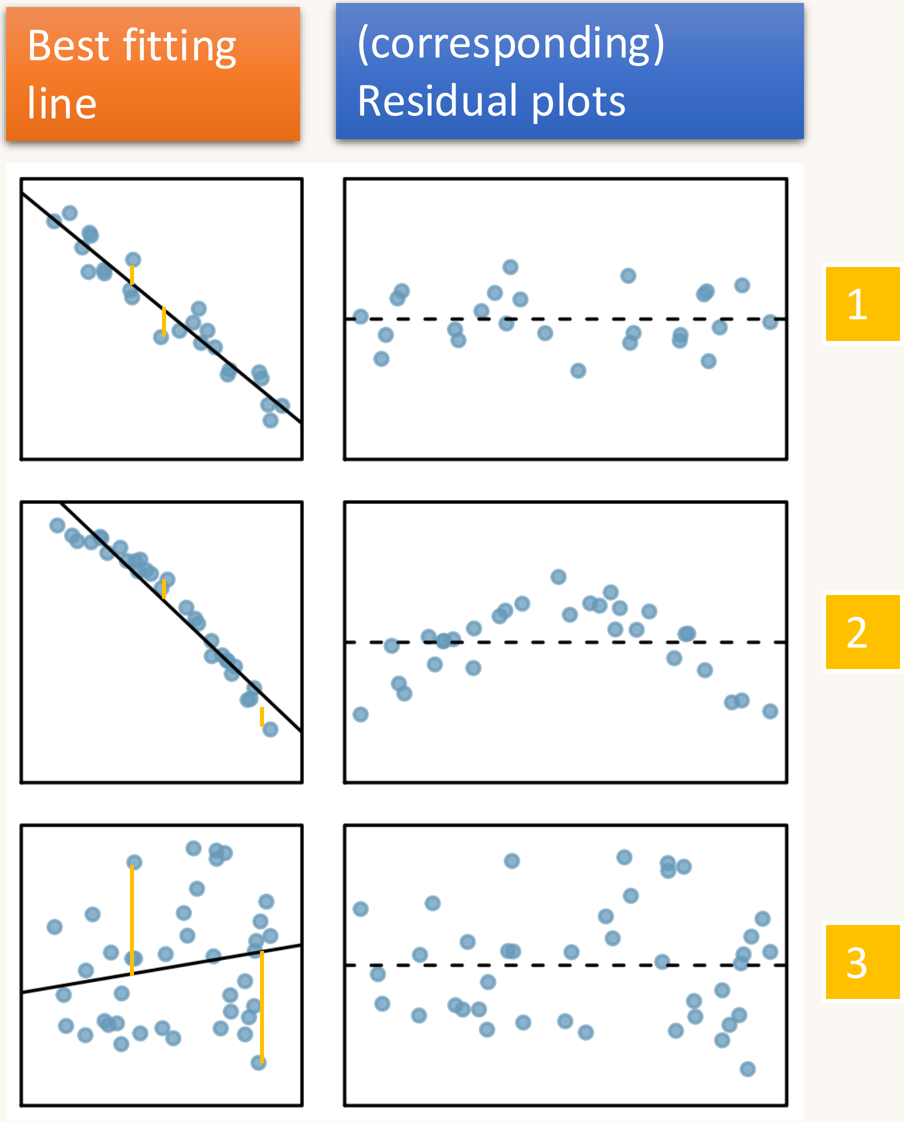
A **residual** is the vertical distance between a data point and the regression line. $y_i - \hat{y}_i$

- y_i: The actual response value for the ith observation
- $\hat{y}_i$: The predicted response value based on the multiple linear regression model

We want to see a residual plot where data shows random scatter above and below the horizontal line

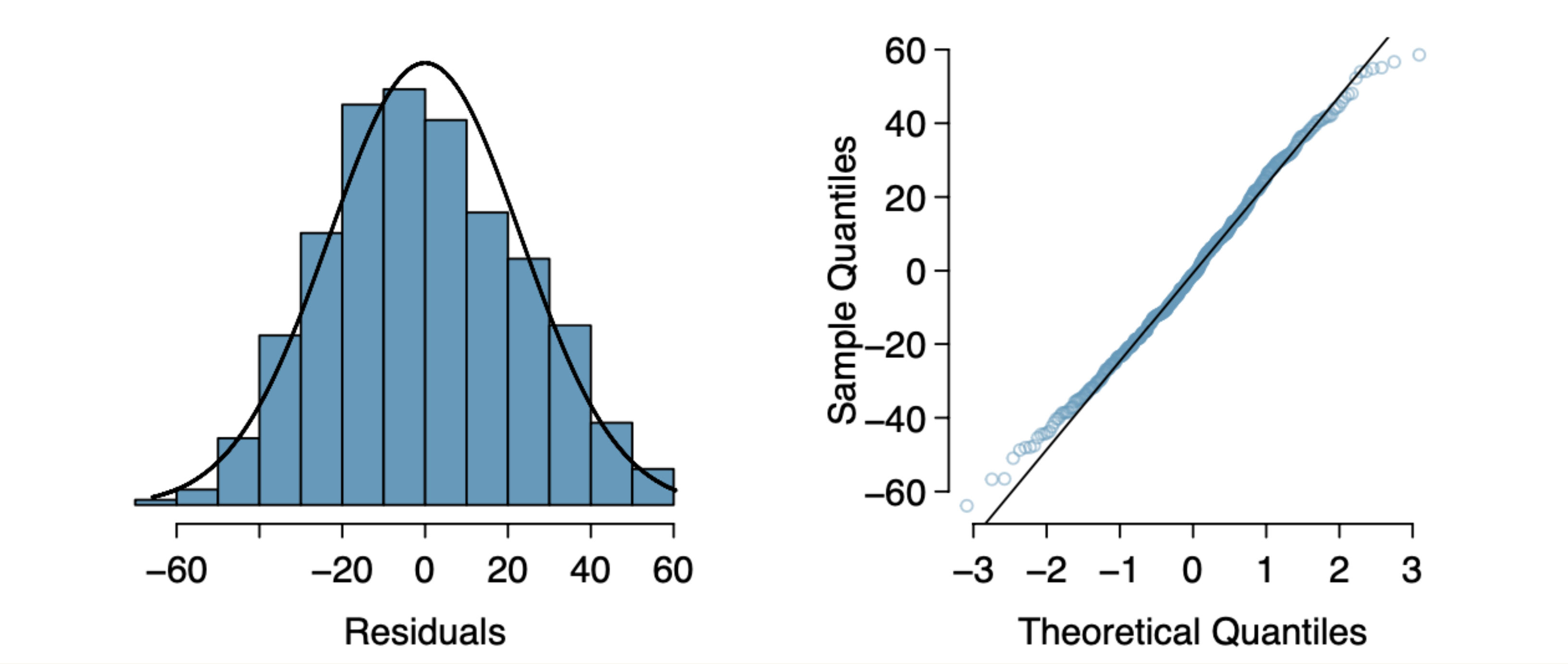
In the example on the right:

- Case 1) linear model is a **particularly good fit!**
- Case 2) the original data cycles below and above the regression line
- Case 3) the variability of the residuals is not constant; the residuals are slightly more variable for larger predicted values.



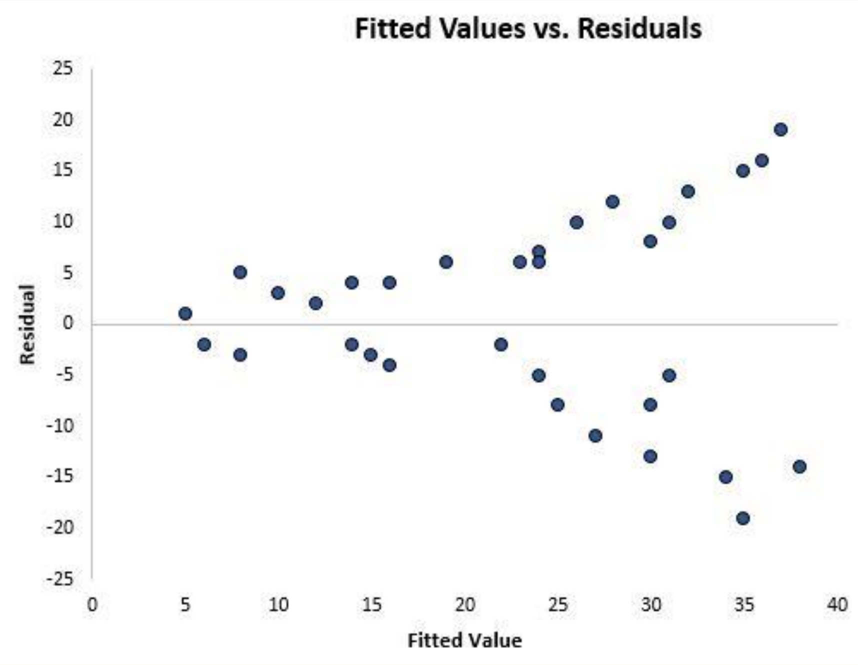
Diagnostic plotting: normality of residuals

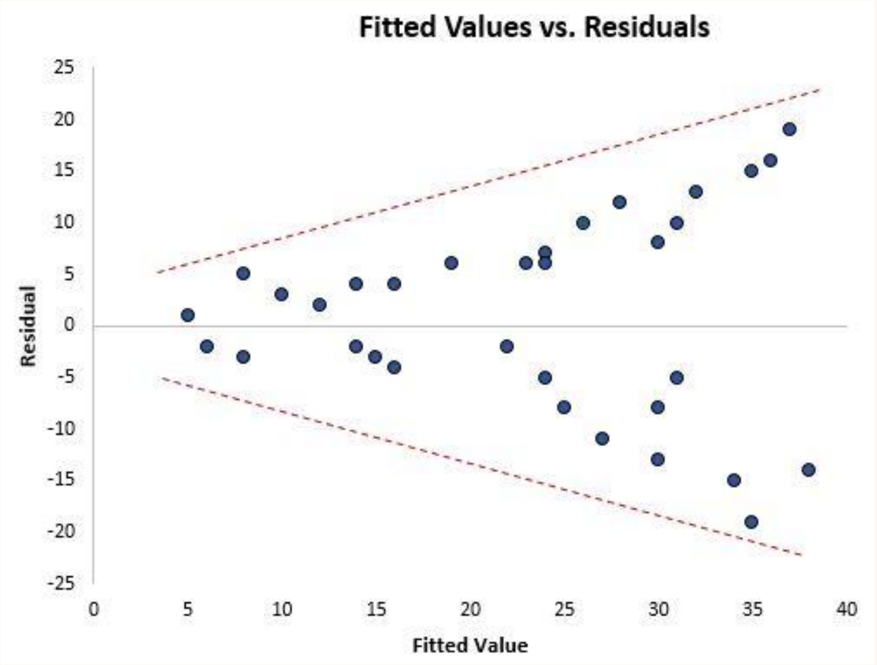
- The residuals of the model are normally distributed.
- Check normality (OF RESIDUALS) with the known methods (QQplot, Shapiro-Wilk, Kolmogorov Smirnov)



Diagnostic plotting: Homoscedasticity

- ASSUMPTION: The residuals (i.e. the error term) have constant variance at every level of x ("homoscedasticity")
- When this is not true, the results of the regression model might be unreliable
- This assumption can be verified by:
  - the "Residual vs. Fitted" plot
  - the Breusch-Pagan Test or the White Test





The Coefficient of Determination or "R Squared" (R^2)

- One way to measure how well the least squares regression line "fits" the data is using the coefficient of determination, denoted as R^2.
- R^2 is the proportion of the variance in the response variable that can be explained by the predictor variable.
- R^2 can range from 0 to 1.
  - A **value close to 0** indicates that data is very spread around the regression line (this doesn't necessarily mean that the model is a bad fit, rather that the data is naturally noisy)
  - A **value close to 1** indicates that the response variable can be perfectly explained without error by the predictor variable.
- For example, an R^2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R^2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable
- **BEWARE OF MISINTERPRETATION:** R^2 measures variability around a regression line… it **doesn't tell if the model is a good fit** or even reasonable !!
  - To assess the performance of linear models, R^2 must be considered along with other measures (e.g. the Residual Standard Error or the significance level of the regression)

Multiple Linear Regression

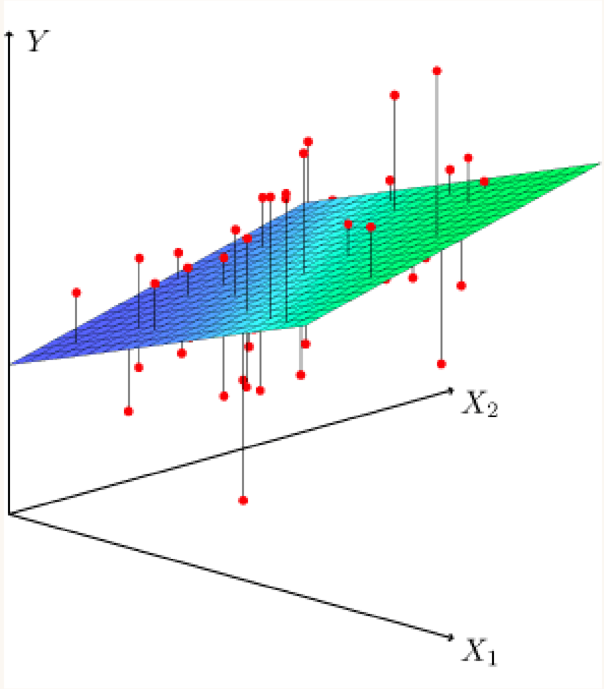
Regression analysis can be used to estimate the linear relationship between a response variable and <u>several</u> predictors

Multiple linear regression: formally

- A multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- where:
  - Y: The response variable
  - X_j: The jth predictor variable
  - β_j: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
  - ε: The error term



- The values for β_0, β_1, β_2, … , β_p are chosen using the least square method, which minimizes the sum of squared residuals (RSS):

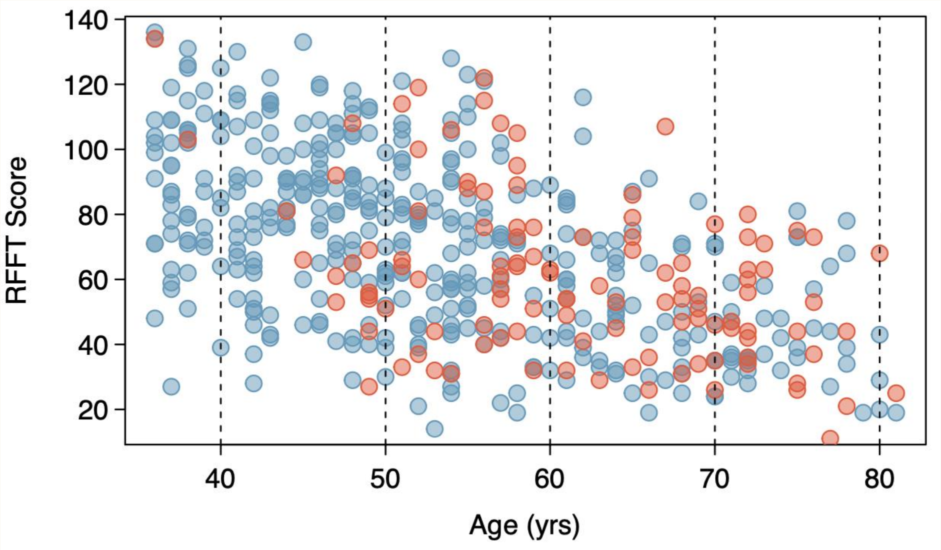
$$\text{RSS} = \sum (y_i - \hat{y}_i)^2$$

- where:
  - Σ: A greek symbol that means sum
  - y_i : The actual response value for the ith observation
  - $\hat{y}_i$: The predicted response value based on the multiple linear regression model

Multiple linear regression: example

- [We'll revisit this in the lab, using the PREVEND dataset]
- STUDY: Statins are a class of drugs widely used to lower cholesterol (which can increase risk for adverse cardiovascular events). However, treatment with a statin might be associated with an increased risk of cognitive decline. Adults of older age are at increased risk for cardiovascular disease, but *also* for cognitive decline
- GOAL: Examine the association of statin use with cognitive ability in an observational cohort, but also accounted for age in the analysis as it could be a potential confounder in this setting
- HYPOTHETICAL MODEL: $RFFT\ score = \beta_0 + \beta_{statin}(STATIN) + \beta_{age}\ (\mathrm{AGE}) + \ \varepsilon$

RFFT score = cognitive function test, with 0 = worst



red dots = Statin users
blue dots = not Statin users

Source: Vu, J., & Harrington, D. (2021). *Introductory Statistics for the Life and Biomedical Sciences*. Retrieved from https://www.openintro.org/book/biostat/

Multiple linear regression: interpreting predictors coefficients

Given our model, we have obtained this prediction equation:

$$E(RFFT) = 137.8822 + 0.8509(STATIN) - 1.2710(\text{AGE})$$

- ESTIMATE for a coefficient b_j is the predicted mean change in $\hat{y}_i$ corresponding to a 1 unit change in x_j, **when the values of all other predictors remain constant.** E.g.:
  - an increase of 1 year of age is associated with a decrease of -1.2710 in RFFT score, **when statin use is the same**
  - **for 2 individuals of the same age**, the RFFT score will be 0.8509 higher for the one taking statins

- [STD. ERROR, T-STATISTIC, P-VALUE]: For each coefficient the model tests the $H_0 : b_j = 0$
  - the association between RFFT score and statin use is not statistically significant, but the association between RFFT score and age is significant

```
Call:
lm(formula = rfft ~ statin + age, data = prevend)

Residuals:
    Min      1Q  Median      3Q     Max
-63.855 -16.860  -1.178  15.730  58.751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.8822     5.1221  26.919   <2e-16 ***
statin        0.8509     2.5957   0.328    0.743
age          -1.2710     0.0943 -13.478   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

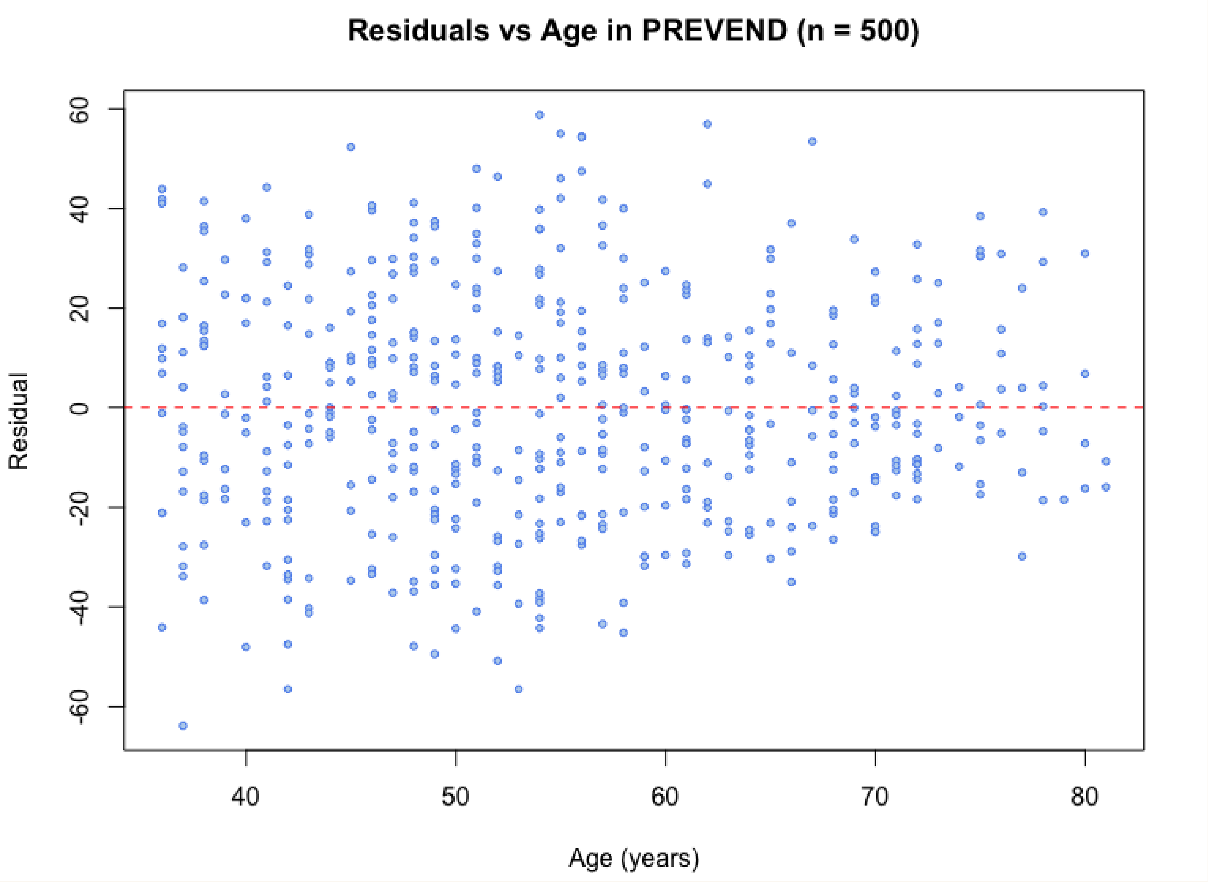
Assumptions for (multiple) linear regression

Similar to those of simple linear regression…

1. **Linearity** : For each predictor variable x_j, change in the predictor is linearly related to change in the response variable y when the value of all other predictors is held constant.
   - It is not possible to make a scatterplot of a response against several simultaneous predictors. Instead, we use a modified residual plot to assess linearity
2. **Normality of residuals**: The residuals are approximately normally distributed.
   - Verified with normal probability plots (Q-Q plots etc.)
3. **Homoscedasticity** (constant variability): The residuals have approximately constant variance at every level of x.
   - Verified by plotting the residual values on the *y*-axis and the predicted values on the *x*-axis
4. **Independent observations**: Each set of observations $(y, x_1, x_2, \dots, x_k)$ is independent
5. (NEW!) **No multicollinearity:** i.e. no situations when there is a strong linear correlation between the independent variables, conditional on the other variables in the model
   - multicollinearity may lead to imprecision or instability of the estimated parameters when a variable changes

Using residuals to check model assumption 1 (on individual predictors)

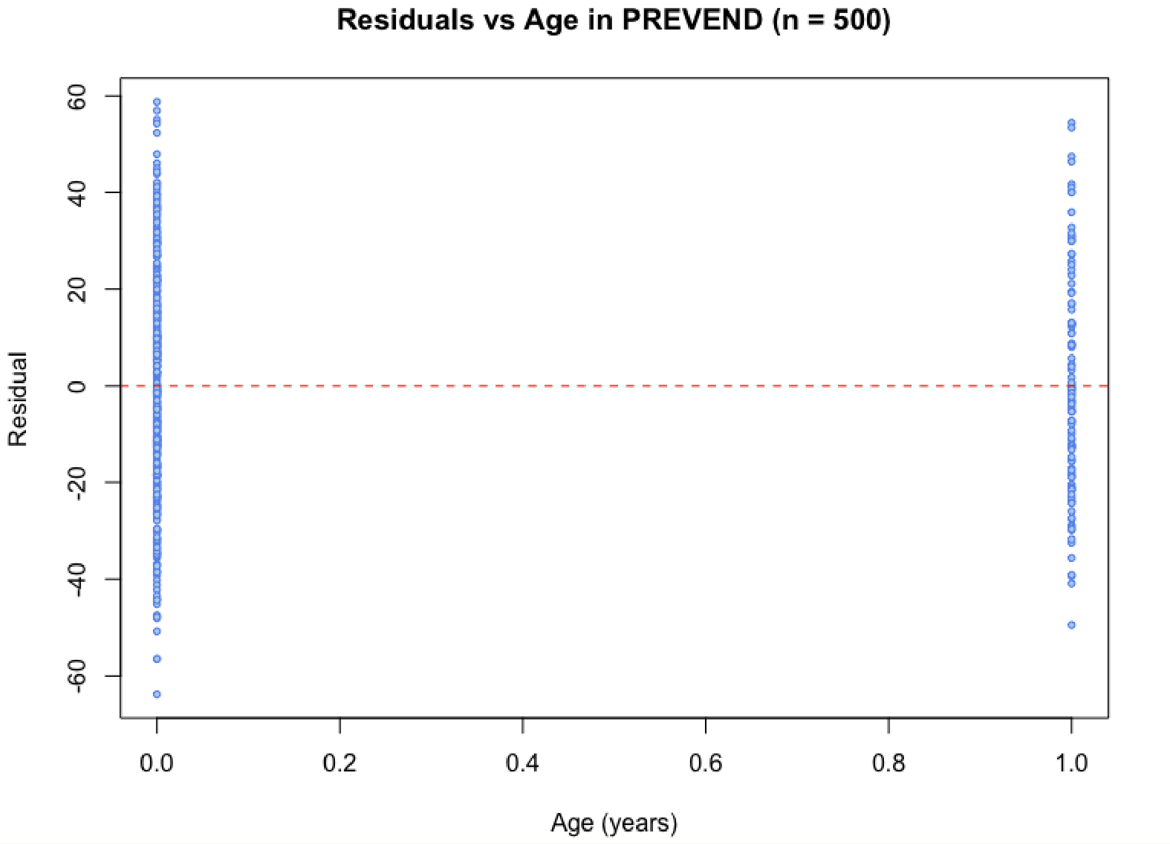
- Assess **linearity** with respect to age using a scatterplot with residual values on the *y*-axis and values of age on the *x*-axis
  - There does not seem to be remaining nonlinearity with respect to age after the model is fit.



Using residuals to check model assumption 1 (on individual predictors)

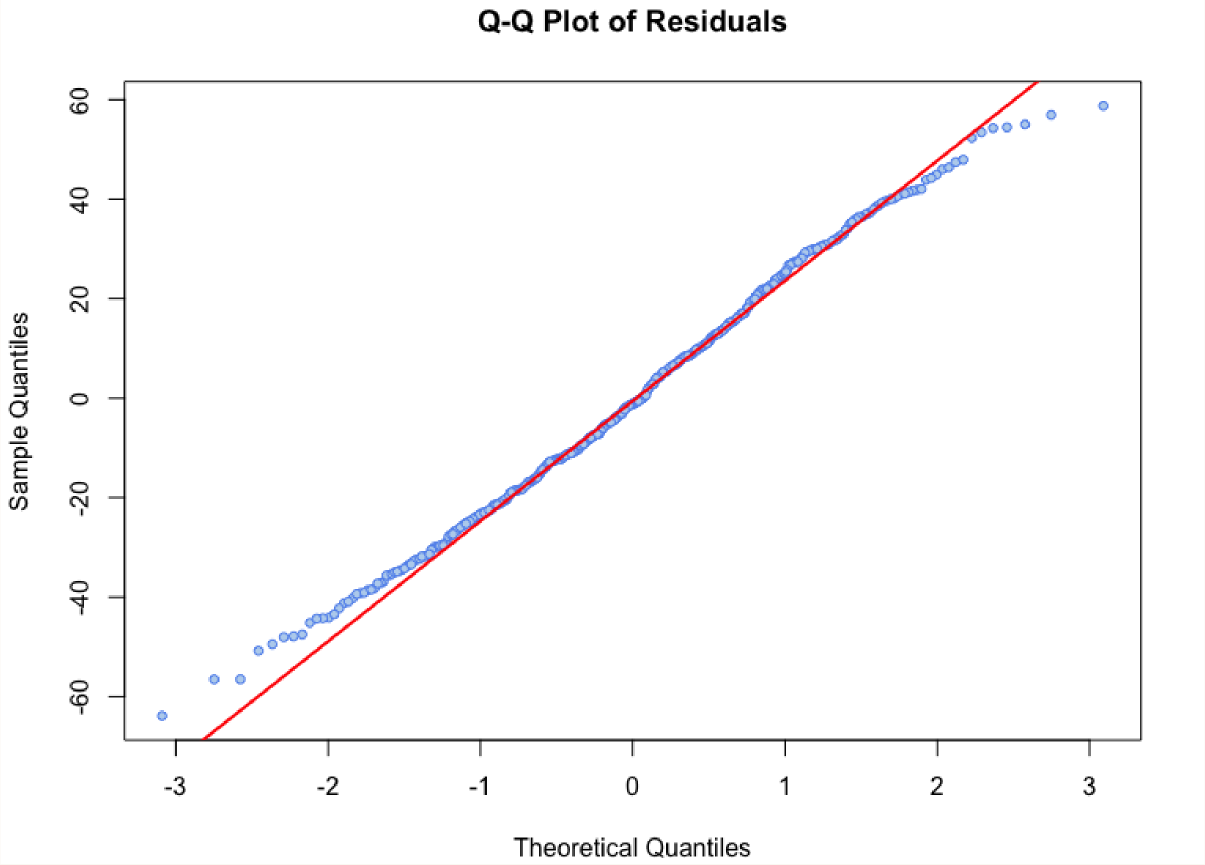
- Assess **linearity** with respect to statin use using a scatterplot with residual values on the *y*-axis and values of age on the *x*-axis
  - It is not necessary to assess linearity with respect to statin use since it is measured as a categorical variable. *A line drawn through two points (that is, the mean of the two groups defined by a binary variable) is necessarily linear*

NOT MEANINGFUL with respect to categorical explanatory variable!



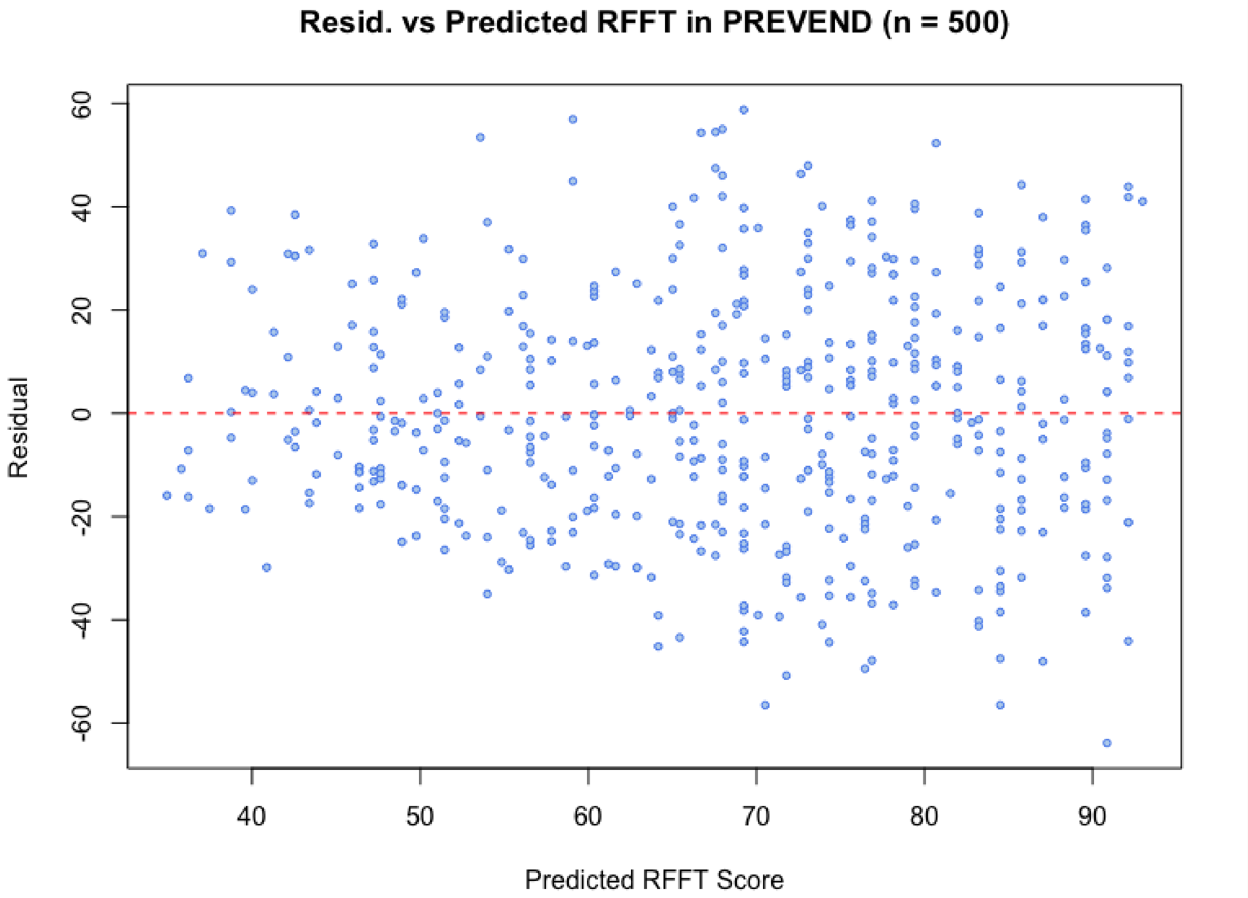
Using residuals to check model assumption 2 (Normality of residuals)

- As in Simple Regression we use Q-Q plots
  - The residuals are reasonably normally distributed, with only slight departures from normality in the tails.



Using residuals to check model assumption 3 (Homoscedasticity)

- As in Simple Regression we plot plot the residual values on the *y*-axis and the predicted values on the *x*-axis
  - It seems reasonable to assume approximately constant variance.



Checking assumption n 6 (no multicollinearity)

- It can be assessed by studying the correlation between each pair of independent variables, or even better, by computing the variance inflation factor (VIF)
  - The **VIF** measures how much the variance of an estimated regression coefficient increases, relative to a situation in which the explanatory variables are strictly independent.
  - A high value of VIF is a sign of multicollinearity (the threshold is generally at 5 or 10)
  - The easiest way to reduce the VIF is to remove some correlated independent variables, or eventually to standardize the data.

Not an issue in our dataset/ model

Variability in the response explained by the model: R^2 and Adj. R^2 in multiple regression

- As in simple regression, R^2 represents the proportion of variability in the response variable explained by the model
  - As variables are added, R^2 always increases
- $Adj.R^2$ incorporates a *penalty* for including predictors that do not contribute much towards explaining observed variation in the response variable
  - $Adj.R^2$ does not have an inherent interpretation, but it is useful while comparing models with different explanatory variables
- *Resid. Std. Err.* (square root of the residual mean squared errors) is the estimated standard deviation of the error of the regression equation and is a good measure of the accuracy of the regression line.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.8822     5.1221  26.919   <2e-16 ***
statin        0.8509     2.5957   0.328    0.743
age          -1.2710     0.0943 -13.478   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.21 on 497 degrees of freedom
Multiple R-squared:  0.2852,    Adjusted R-squared:  0.2823
F-statistic: 99.13 on 2 and 497 DF,  p-value: < 2.2e-16
```

F statistic in multiple regression

- Again, the F-test of overall significance indicates whether this linear regression model provides a **better fit to the data** than a hypothetical model that contains no independent variables (known as the “intercept model”)
  - H_0: (null hypothesis) The intercept model fits the data as well as your model.
  - H_1: (alternative hypothesis) Your model fits the data better than the intercept-only model
- In this case p-value is extremely small, we have sufficient evidence to conclude that this model fits the data better than intercept-only model
- NOTE: in general, if none of the independent variables are statistically significant, the overall F-test is also not statistically significant

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.8822     5.1221  26.919   <2e-16 ***
statin        0.8509     2.5957   0.328    0.743
age          -1.2710     0.0943 -13.478   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.21 on 497 degrees of freedom
Multiple R-squared:  0.2852,	Adjusted R-squared:  0.2823
F-statistic: 99.13 on 2 and 497 DF,  p-value: < 2.2e-16
```

Model performance: ideas for further investigation

- The **PREVEND** data came from a a cross-sectional study, i.e. not from a study in which participants were followed as they aged (i.e., a longitudinal study)
- So, while the model indicates that older patients tend to have lower RFFT scores, **we <u>cannot</u> conclude that RFFT scores decline with age in individuals**
  - only repeated measurements of RFFT taken as (the same) individual participants aged could rule out some explanatory effect of unobserved differences across different age cohorts
- We found that age was a *confounder:* **was it the only one**?
  - Other potential confounders could be education level (also associated to access to health care) and the presence of cardiovascular disease (can lead to vascular dementia and cognitive decline)
  - Residual confounders —frequent in observational studies— can be other variables in a dataset that have not been examined, or variables that were not measured in the study
- A randomized experiment is the best way to eliminate residual confounders, since it ensures that, at least on average, all predictors are not associated with the exposure (i.e. one source of confounding: selection bias).

The details of how a study was designed and how data were collected should always be taken into account when interpreting study results.

Adding a categorical predictor with several levels to the model

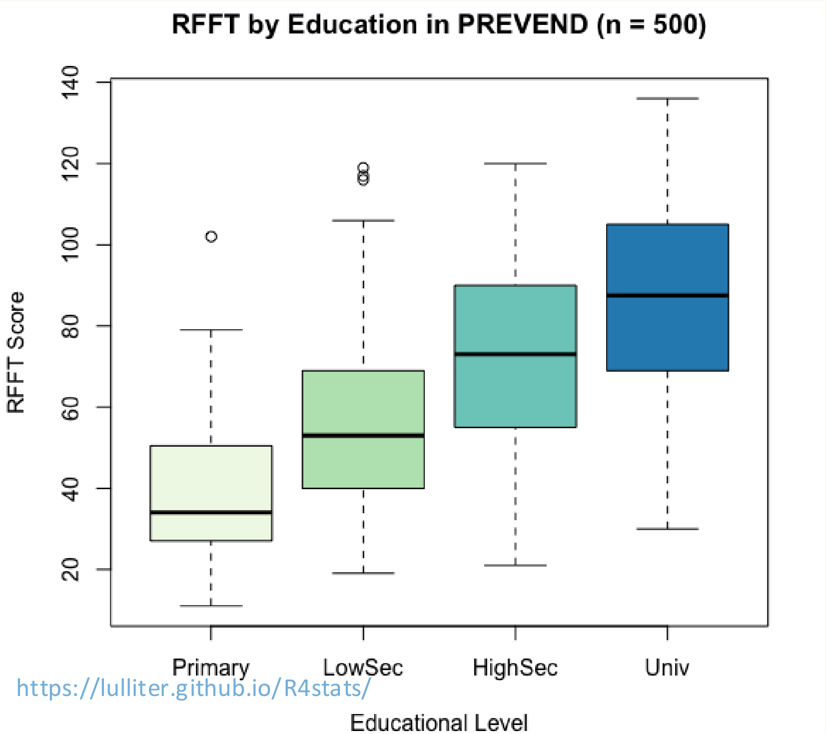
- In a regression model with a categorical variable with more than two levels (e.g. education level), **one of the categories is set as the reference category.** The remaining categories each have an estimated coefficient
  - Each predictor levels can be thought of as binary variables that can take on either 0 or 1

$$E(RFFT) = 40.94 + 14.78(LowerSecond) + 32.13(HigherSecond) + 44.96(Univ)$$

- EXAMPLE: predicted RFFT for individuals in **Lower Secondary Education** level

$$E(RFT) = 40.94 + \mathbf{14.78(1)} + 32.13(\mathbf{0}) + 44.96(\mathbf{0}) = 55.72$$

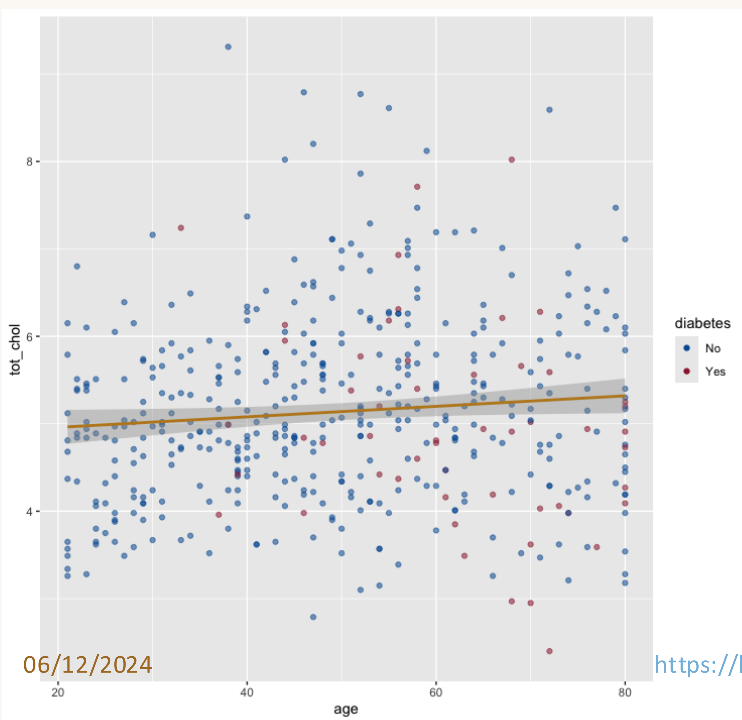
Primary is not in the model because it is the **implicit reference level** (i.e. the intercept value 40.94)



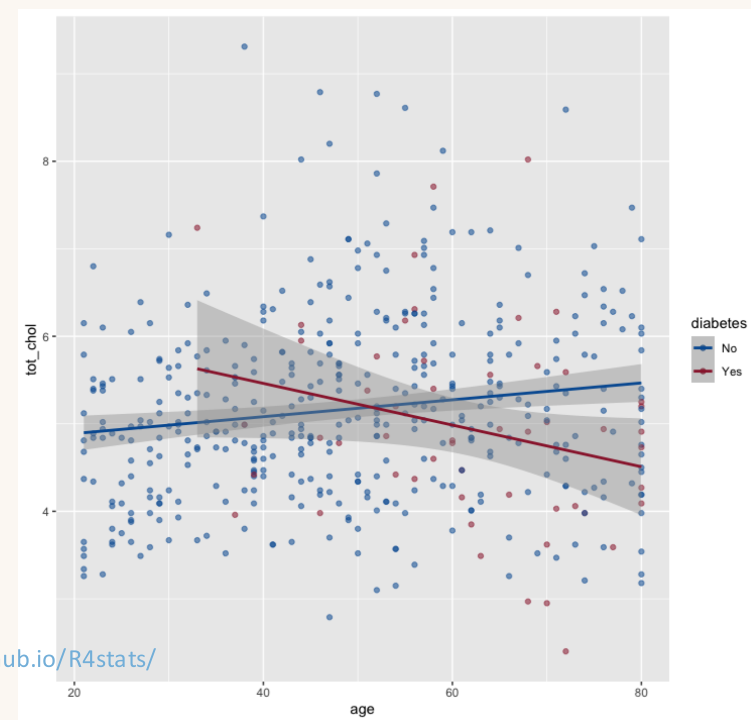
Adding a interaction term to the model specification

- A statistical interaction occurs when the effect of one explanatory variable X_1 on the response Y **depends on the level of another explanatory variable** X_2
- Let's go back to the NHANES dataset and consider a linear model that predicts total cholesterol level (mmol/L) from age (yrs.) and diabetes status.
- Comparing 2 alternative models:
  1. MLR without interaction: $E(\text{TotChol}) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Diabetes})$
  2. MLR with interaction: $E(\text{TotChol}) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Diabetes}) + \beta_2(\text{Diabetes x Age})$
- Model 2 *acknowledges* **the relationship between cholesterol and age depends on diabetes status** (i.e. it "*allows*" the relationship of X_1 with the Y to vary based on the values of X_2)

Linear model on entire sample



Linear model by category (Diabetics or not)



Multiple linear regression: recap

- **Multiple linear regression** is a generalization of simple linear regression to address the relationship between a response variable Y and several predictors X_j, where k is the number of predictors
  - including logical, interval/ratio, or categorical predictors, as well as interaction terms
  - to interprete categorical predictors (>2 levels) one of the category's levels is set as the reference, each remaining level has an estimated coefficient = estimated change relative to the reference

- **Typical applications** of multiple linear regression are:
  1. **PRIMARY PREDICTOR: Estimating an association between a response variable and primary predictor of interest, while adjusting for possible confounding variables**
     - this is the case of the previous example! (Examining the association between statin use and cognitive ability, adjusting by age)
  2. **EXPLANATORY MODELS: Constructing a model that effectively explains the observed variation in the response variable; in other words, to build a predictive model for a response variable**
     - different techniques may be adopted in model selection (i.e. different specifications where we add/subtract explanatory variables)
     - A **parsimonious model (few variables) is usually preferred** over a complex model
     - **R^2** and **Adjusted R^2** can be useful to compare models
     - In particular **Adjusted R^2** helps to balance predictive ability with complexity in a multiple regression model

More advanced topics on REGRESSION…

- This lecture is just an introduction but there is a wide array of topics pertaining to regression analysis…

- Here are some of the many variants and advancements over the linear regression model:
  - LOGISTIC REGRESSION: if the dependent variable is dichotomous (0,1) or nominally scaled
  - POISSON REGRESSION: if the dependent variable is count over a period of time
  - COX PROPORTIONAL HAZARDS REGRESSION: for modeling censored data
  - FUNCTIONAL TRANSFORMATIONS: quadratic, exponential ….
  - GENERALIZED LINEAR MODELS (GLMs): an extension of the linear model where the modelling of error is not Gaussian
  - PANEL REGRESSION MODELS: special regression models that can make use of both the temporal and the inter-individual variation if you have longitudinal data (or time-series cross-sectional or panel data)