BasicR Course

15 – 19 February 2021

DAY 3

Molecular Biotechnology - Master

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

dkfz.

DAY 2 AND 3 – DATA ANALYSIS

Data Import

Wrangling

- Tidy + Manipulating
- Summarizing
- Cleaning

Exploration

- Visualization
- Descriptive Statistics

Day 2

Statistical Inference

- Foundation of inference
- Basic statistical tests
- Linear regression
- Multivariable regression

Day 3

DAY 3

Statistical Inference

- Foundation of inference
- Basic statistical tests
- Linear regression
- Multivariable regression

FOCUS

- **Brief overview of statistical terms**
- **Suitability of statistical tests**
- **Application of statistical tests in R**
- **Interpretation of results**

NOT A STATISTICS COURSE

Course does not cover

- Principles of study design
  - Types of experiments/ studies
  - Reducing bias in study design
- Statistical theory
  - Probability and random variables
  - Central Limit Theorem
  - Hypothesis testing
  - Type 1 and Type 2 error
  - Test statistic and standard error
  - Confidence intervals and p-values

STATISTICAL INFERENCE

– FOUNDATIONS

FOUNDATIONS FOR INFERENCE

Research Question

Do senior citizens (Age>65) on average, have a normal (=120mm Hg) systolic blood pressure?

$SysBP^{>65}$= 120 mm Hg?

2 Hypothesis

Null hypothesis: Senior citizens (Age>65) have **a normal** systolic blood pressure.

H_0 : $SysBP^{>65}$ = 120 mm Hg

Alternate hypothesis: Senior citizens (Age>65) have **an abnormal** systolic blood pressure.

H_A : $SysBP^{>65} \neq$ 120 mm Hg

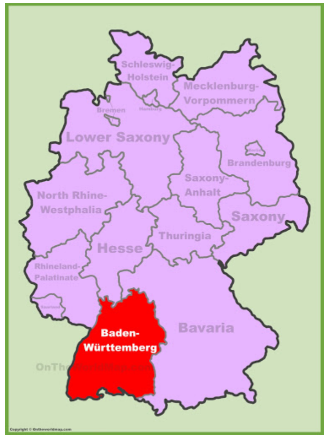
$SysBP^{>65}$: Mean Sys BP for >65 years

BUILDING A STUDY

RESEARCH QUESTION: Do senior citizens (Age>65) on average, have a normal (=120mm Hg) systolic blood pressure?

INFERENTIAL STATISTICS

extend results from our study to the wider population



THE TRUTH

- Collect all 700M people >65 years
- Measure systolic BP for all people
- Calculate the mean (**$SysBP^{>65}$**)
- Compare with 120 mm Hg.
- No need for statistical tests here

STUDY

- Randomly sample 1000 people > 65 years from BW
- Measure systolic BP for all people
- Calculate the mean (**$SysBP^{>65}$**). Let's say 140 mm Hg.

- However, this will only tell you about this specific study
- Even if reduce for SAMPLING BIAS. Still random variability exists

$SysBP^{>65}$: Mean Sys BP for >65 years

ONE SAMPLE T-TEST

WHEN TO USE: Check if the average value of a score/ variable in a population is equal to a set value.

EXAMPLE: Check if mean systolic BP in people > 65 years ($SysBP^{>65}$) is 120 mm Hg.

EFFECT SIZE: Difference between SAMPLE MEAN and set value.

KEY ASSUMPTIONS:

- Random sample of patients from the population
- Systolic BP measurements were independent over the samples
- Systolic BP is a continuous variable
- Systolic BP is normally distributed
- Variances in sample and population are approximately equal

Two tailed test **-** Generally used

$$H_0 : SysBP^{>65} = 120 \text{ mm Hg}$$

$$H_A : SysBP^{>65} \neq 120 \text{ mm Hg}$$

We are looking for **differences in <u>either</u> direction**

One tailed test **-** Rarely used

$$H_0 : SysBP^{>65} > 120 \text{ mm Hg}$$

$$H_A : SysBP^{>65} \leq 120 \text{ mm Hg}$$

We are looking for **differences in <u>one</u> direction**

KEY STATISTICAL TERMS

N: Number of samples (people >65 years)

e.g. in this study. **N** = 1000

SAMPLE MEAN: Mean Systolic BP for >65 years = $\mathbf{SysBP^{>65}}$

e.g. in this study. $\mathbf{SysBP^{>65}}$ = 140 mm Hg

EFFECT SIZE: Difference between SAMPLE MEAN and expected value under $\mathbf{H_0}$. How abnormal is the $SysBP^{>65}$?

e.g. in this study. **EFFECT SIZE =** $\mathbf{SysBP^{>65}}$ -120 mm Hg = 20 mm Hg

STANDARD ERROR (***SE***): Estimate of how variable the SAMPLE MEAN, $\mathbf{SysBP^{>65}}$**.**

- If we perform multiple studies of sampling 1000 people >65 all over the world, and calculate the $\mathbf{SysBP^{>65}}$ for each study. The SE is an estimate of how variable these averages are.
- Smaller SE, if larger **N**.

T-STATISTIC (***t***): Standardized representation of effect size.

- How far away (in terms of standard errors) is the effect size from the Null.

KEY STATISTICAL TERMS

ALPHA THRESHOLD (α): Type 1 Error or False Positive

- *Set by user.* Default is 5%. 1% and 10% are also used.
- But should depend on test! 5% Alpha means that we are comfortable with false positives of 5%.

P-VALUE (*p*): Strength of the evidence against the null hypothesis (against **SysBP**$^{>65}$ = 120 mm Hg).

e.g. in this study. ***p* = 0.03**

- The smaller the p-value the more ***statistically significant*** the finding is.
- We can compare our **P-VALUE** to the selected **ALPHA**.

  p > α Do not reject the null hypothesis (never accept it)

  p < α Reject the null hypothesis

- **Only if Null hypothesis is true (**i.e. in reality **SysBP**$^{>65}$ = 120 mm Hg), and we perform 100 similar studies of sampling 1000 people >65 all over the world, then for only 3 studies we would see **SysBP**$^{>65}$ > 140 mm Hg.

CONFIDENCE INTERVAL (CI): Range of values within which we are reasonably confident that the true effect size lies.

e.g. in this study. 95% CI **= 10-30 mm Hg**

- Threshold e.g. 95% CI depends on selected **ALPHA**
- If we perform 100 studies of sampling 1000 people > 65 all over the world, for 95 studies **the EFFECT SIZE (difference)** will lie within this interval.

REPORTING AND UNDERSTANDING P VALUES

TYPE 1 AND TYPE 2 ERRORS

Type 1 error: FALSE POSITIVE

- Statistical significance does not necessarily mean that the effect is real. If set an **ALPHA of 5%,** this means we are comfortable with false positive rate of 5%.

Type 2 error: FALSE NEGATIVE

- Small studies (small **N**) can show non-significance even when there are real effects. **Lack of power.**

  So, we should not accept the Null hypothesis because we do not get a statistically significant result.

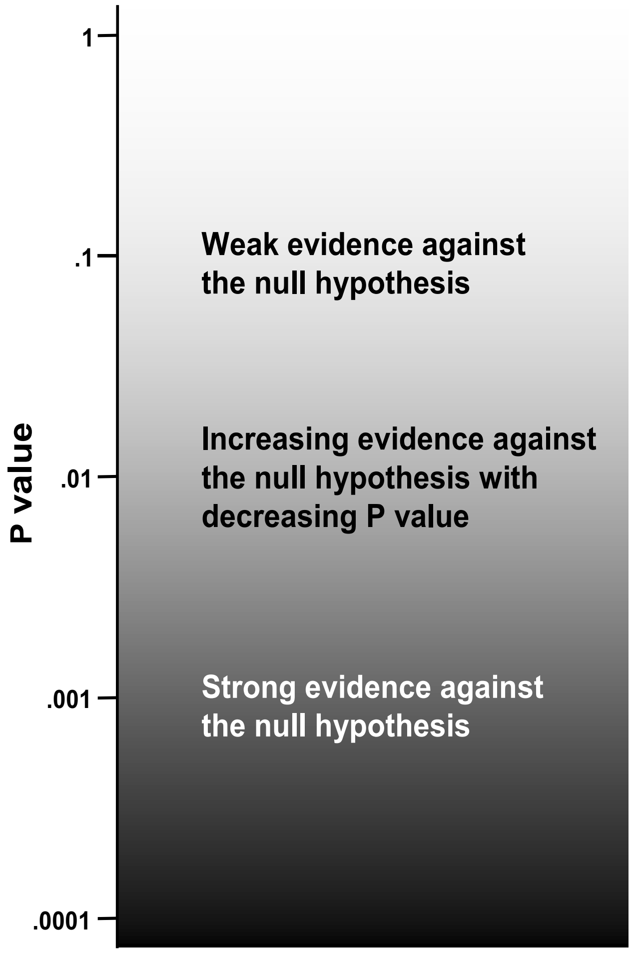
Sterne, J.A. (2001) BMJ

REPORTING AND UNDERSTANDING P VALUES

ALWAYS USE JUDGEMENT

- As the p-value decreases the evidence against the null hypothesis increases. It is a continuous scale.

- P-values, confidence intervals and effect sizes must be considered in combination.
  - Statistical significance (based on p-value) and biological/ clinical significance (based on effect sizes) are not the same thing.
  - Highly powered test (Large **N**) can give statistical significance even if biological effect size is tiny.



Sterne, J.A. (2001) BMJ

STATISTICAL INFERENCE

– STATISTICAL TESTS

ONE SAMPLE T-TEST

WHEN TO USE: Check if the average value of a score/ variable in a population is equal to a set value.

EXAMPLE: Check if mean systolic BP in people > 65 years ($SysBP^{>65}$) is 120 mm Hg.

EFFECT SIZE: Difference between SAMPLE MEAN and set value

KEY ASSUMPTIONS:

- Random sample of patients from the population
- Systolic BP measurements were independent over the samples
- Systolic BP is a continuous variable
- Systolic BP is normally distributed
- Variances in sample and population are approximately equal

***Two tailed test* -** Generally used

$$H_0 : SysBP^{>65} = 120 \text{ mm Hg}$$
$$H_A : SysBP^{>65} \neq 120 \text{ mm Hg}$$

We are looking for **differences in <u>either</u> direction**

***One tailed test* -** Rarely used

$$H_0 : SysBP^{>65} > 120 \text{ mm Hg}$$
$$H_A : SysBP^{>65} \leq 120 \text{ mm Hg}$$

We are looking for **differences in <u>one</u> direction**

TWO SAMPLE T-TEST - UNPAIRED

WHEN TO USE: **Check if the average value of a variable is equal between 2 populations**
Association between a continuous variable and a binary variable (2 groups)

EXAMPLE: Check if mean systolic BP in elderly, > 65 years ($SysBP^{>65}$) and in adults, < 65 years ($SysBP^{<65}$) is equal.

EFFECT SIZE: Difference between SAMPLE MEAN of the 2 populations

KEY ASSUMPTIONS:

- Random sample of patients from both populations
- Systolic BP measurements were independent over the samples within populations, and between populations
- Systolic BP is a continuous variable
- Systolic BP is normally distributed in both populations
- Variances in both samples/ populations are approximately equal.
  In case variance in both populations are unequal, there is an extra step required.

***Two tailed test* - Generally used**

$$H_0 : SysBP^{>65} = SysBP^{<65}$$

$$H_A : SysBP^{>65} \neq SysBP^{<65}$$

We are looking for **differences in <u>either</u> direction**

***One tailed test* - Rarely used**

$$H_0 : SysBP^{>65} > SysBP^{<65}$$

$$H_A : SysBP^{>65} \leq SysBP^{<65}$$

We are looking for **differences in <u>one</u> direction**

TWO SAMPLE T-TEST - PAIRED

WHEN TO USE: **Check if the average value of a variable is equal between paired measurements in a population.**
Association between a continuous variable and 2 matched groups

REPEATED MEASURES: Measure a person at 40 years and then the same person again at 60 years

MATCHED MEASURES: Measure an older twin and younger twin

EXAMPLE: Check if mean systolic BP in elder twin ($SysBP^1$) and in younger twin ($SysBP^2$) is equal.

EFFECT SIZE: Mean of difference between paired measurements

KEY ASSUMPTIONS:

- Random sample of patients from the population.
- Systolic BP measurements were paired over the samples.
- Systolic BP is a continuous variable
- Systolic BP difference between pairs is normally distributed
- Variances of differences in sample and population are approximately equal

Two tailed test - Generally used

$$H_0 : SysBP^1 - SysBP^2 = 0$$
$$H_A : SysBP^1 - SysBP^2 \neq 0$$

We are looking for **differences in <u>either</u> direction**

One tailed test - Rarely used

$$H_0 : SysBP^1 - SysBP^2 > 0$$
$$H_A : SysBP^1 - SysBP^2 \leq 0$$

We are looking for **differences in <u>one</u> direction**

ONE-WAY ANOVA

WHEN TO USE: **Check if the average value of a variable is equal between multiple (>2) populations**
Association between a continuous variable and a categorical variable (>2 groups).

EXAMPLE: Check if mean systolic BP in children < 18 years ($SysBP^{<18}$), in elderly >65 years ($SysBP^{>65}$), and in remaining adults < 65 years ($SysBP^{<65}$) are equal.

EFFECT SIZE: None

KEY ASSUMPTIONS:

- Random sample of patients from all populations
- Systolic BP measurements were independent over the samples within populations, and between populations
- Systolic BP is a continuous variable
- Systolic BP is normally distributed in all populations
- Variances in all samples/ populations are approximately equal.

Only tells you whether there is a difference between the groups -- not which group is different than the rest.

H_0 : SysBP of all age groups are equal. $SysBP^{<18} = SysBP^{<65} = SysBP^{>65}$

H_A : SysBP of at least one age group is different to the others.

CHI-SQUARE TEST OF INDEPENDENCE

WHEN TO USE: Check if the proportion of a categorical score/ outcome is equal between two (or more) groups.
Association between 2 (or more) categorical variables

EXAMPLE: Check if Smoking behavior was associated with AGE GROUP (<65 years vs > 65 years)

H_0: No association between **AGE GROUP** and smoking
H_A: Association between **AGE GROUP** and smoking

OBSERVED	<65	>65	
NO SMOKING	25 (76%)	6 (28%)	31 (57%) (57%)
SMOKING	8 (24%)	15 (72%)	23 (43%)
	33	21	N = 54

EFFECT SIZE: Odds Ratio (76%/28%)

Under H_0

EXPECTED H_0	<65	>65	
NO SMOKING	18.9 (57%)	12.1 (57%)	31 (57%)
SMOKING	14.1 (43%)	8.9 (43%)	23 (43%)
	33	21	N = 54

KEY ASSUMPTIONS:

- Random sample of patients from all populations
- Systolic BP measurements were independent over the samples within groups, and between groups
- Both variables are categorical
- Each particular scenario (i.e. cell) has at least 5 expected cases

DATA DETERMINES THE TEST

GOAL	NORMAL OUTCOME	NON-NORMAL OUTCOME	CATEGORICAL DATA (UNORDERED CATEGORIES)
COMPARE TWO UNPAIRED GROUPS	Unpaired t-test	Mann-Whitney test	Chi square test (Fishers exact test)
COMPARE TWO PAIRED GROUPS	Paired t-test	Wilcoxon test	McNemars test
COMPARE THREE OR MORE UNMATCHED GROUPS	One way ANOVA	Kruskal Wallis test	Chi square test (Fishers exact test)

EXERCISE

Day3_1_StatisticalTests_Exercise.Rmd

STATISTICAL INFERENCE

– LINEAR REGRESSION

WHY DO WE USE?

- Looks at relationship between an outcome and exposure
- Allows one to generate a meaningful effect size per unit change in exposure
- Can consider multiple exposures
- Adjust for confounders and interaction terms
- Allows us to make predictions for ‘new’ obersvations

WHICH REGRESSION TO USE?

OUTCOME	MODEL
Continuous (serum)	Linear
Binary (Dead or alive)	Logistic
Ordinal / ranked (Pain grading 1-3)	Ordinal
Categorical (apples vs. oranges vs. pears)	Multinomial
Count (number of admissions to hospital)	Poisson

LINEAR REGRESSION

FORMALISES RELATIONSHIP

- A more formal description of the relationship between an outcome and one (or more) exposures.

RELIES ON LINEARITY

- Looks for a linear relationship between a predictor and an outcome. Depends on explicitly defining the line which best describes the relationship: the regression line

QUANTIFIES THE ASSOCIATION

- Allows estimation of the value of y (outcome) per unit change in x (exposure)

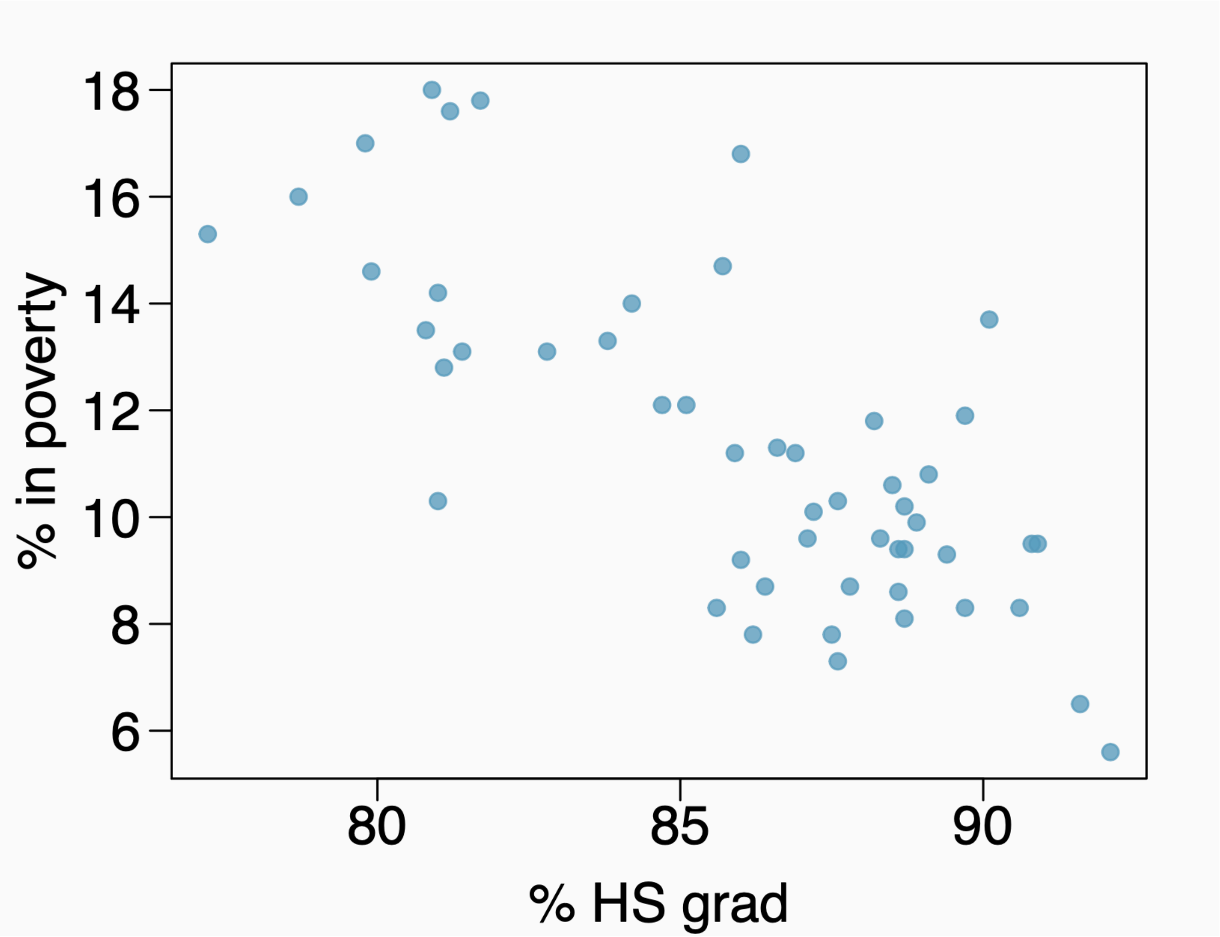
TRANSFORMATION ALLOWED

- Variables can be transformed to show linearity.

ASSUMPTIONS

- Checked using Regression diagnostics.

LINEAR REGRESSION

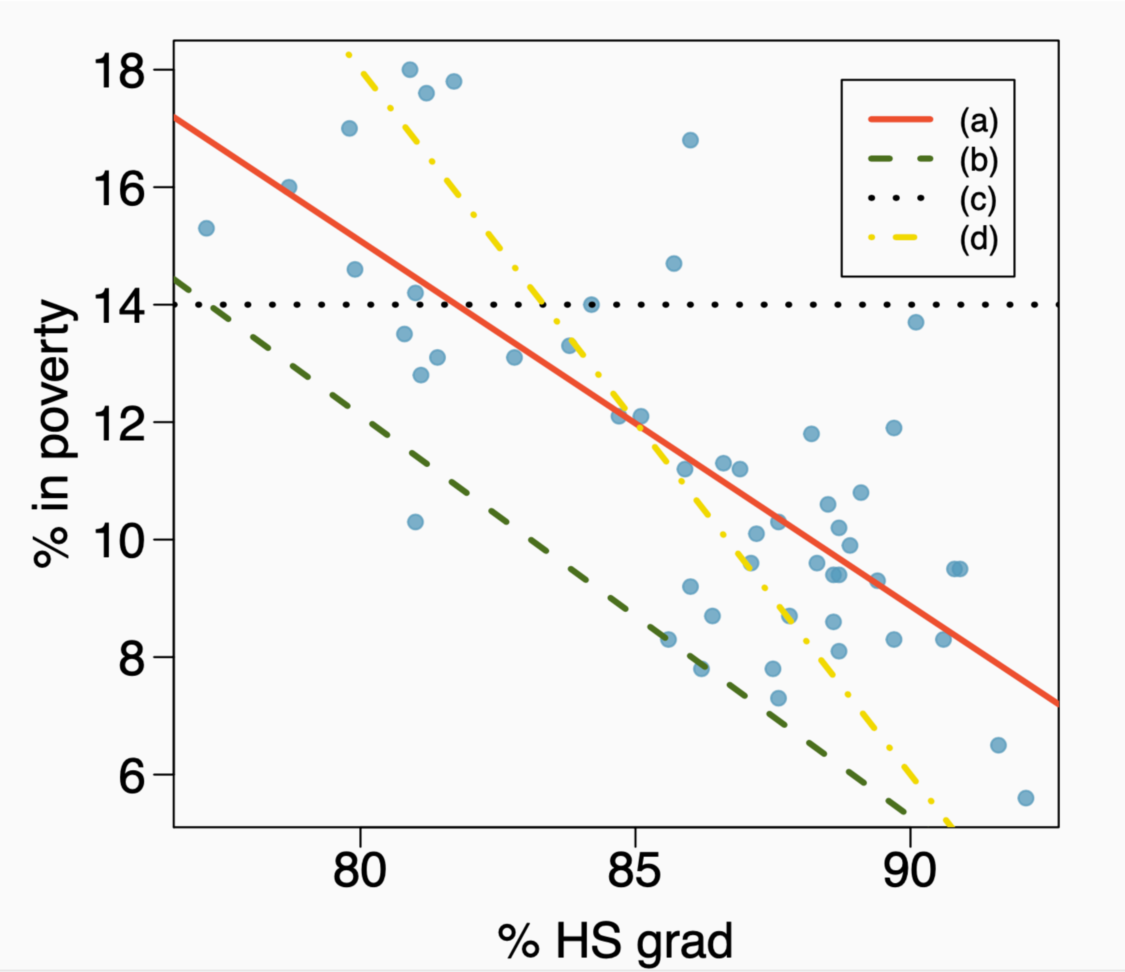


The scatterplot below shows the relationship between high school graduate rate in 50 US states and the % of residents who live below the poverty line.

OUTCOME	Poverty %
EXPOSURE	Graduation %
RELATIONSHIP	Linear Negative Strong

https://www.openintro.org/book/os/

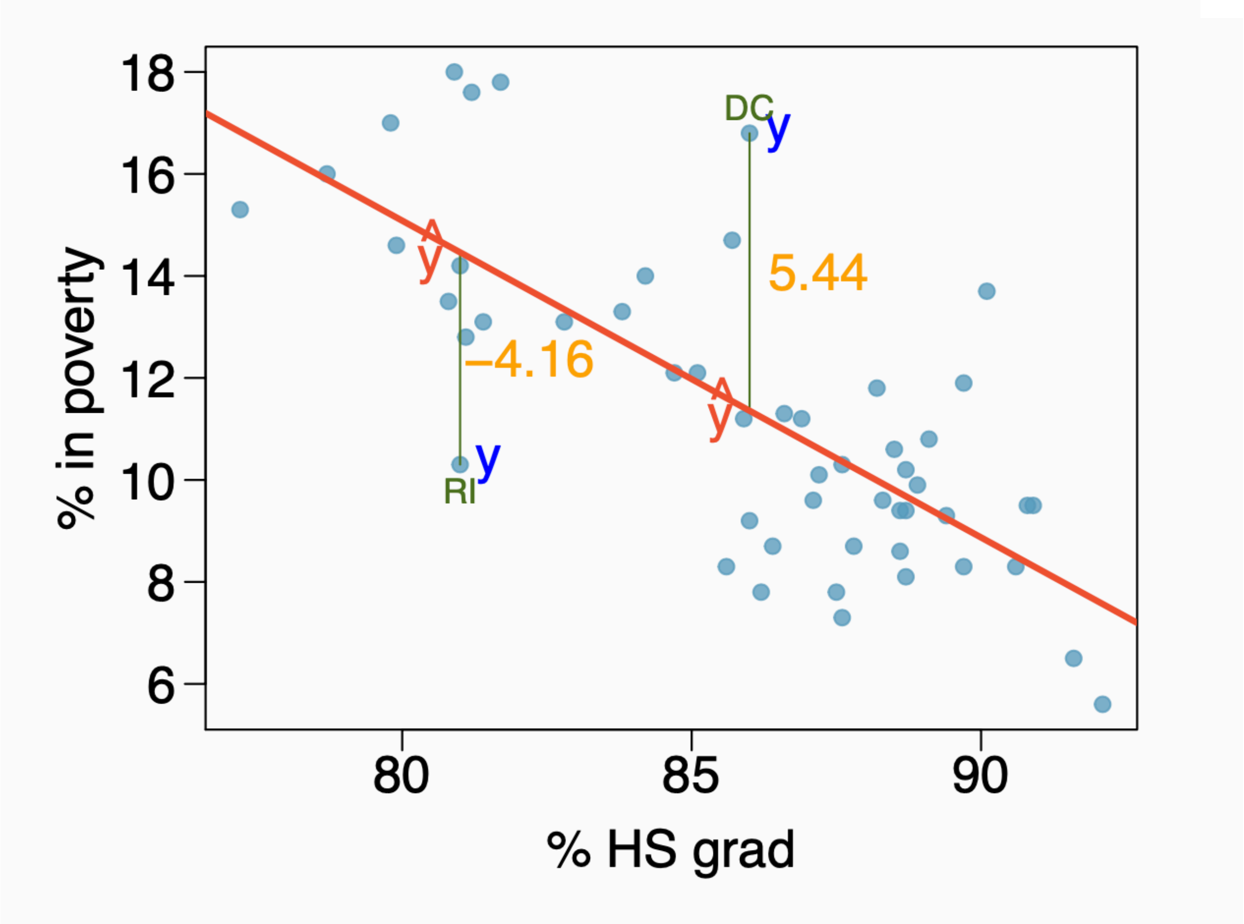
LINE OF BEST FIT



LINE OF BEST FIT? **a)**

https://www.openintro.org/book/os/

RESIDUALS



RESIDUALS: Difference between Data and Line of best Fit

AIM: Minimize these residuals

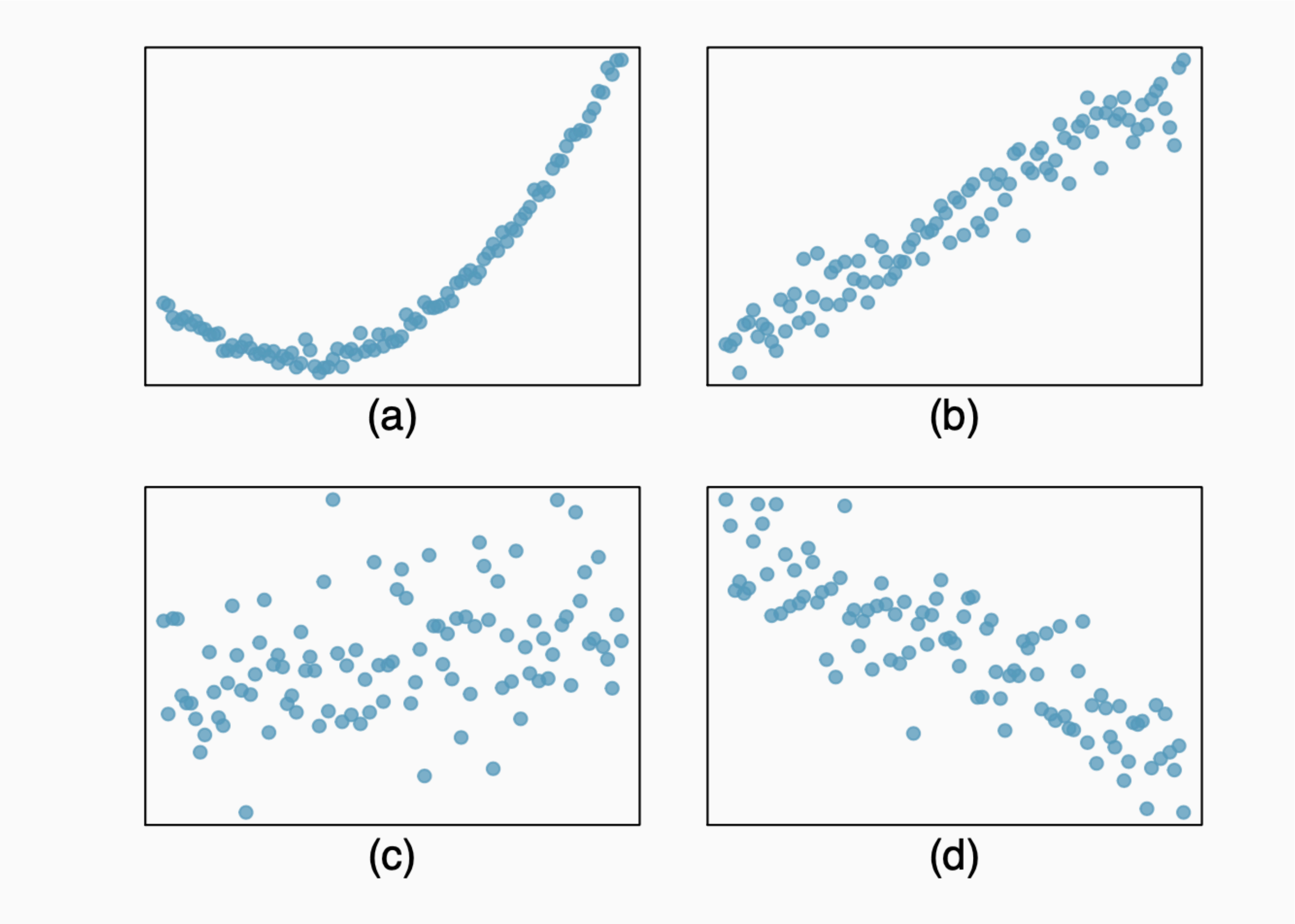
METHOD: Least squares
commonly used
penalizes larger residuals more

CORRELATION: Strength of association (only linear)
-1 to 1
0 No association

https://www.openintro.org/book/os/

dkfz.

CORRELATION

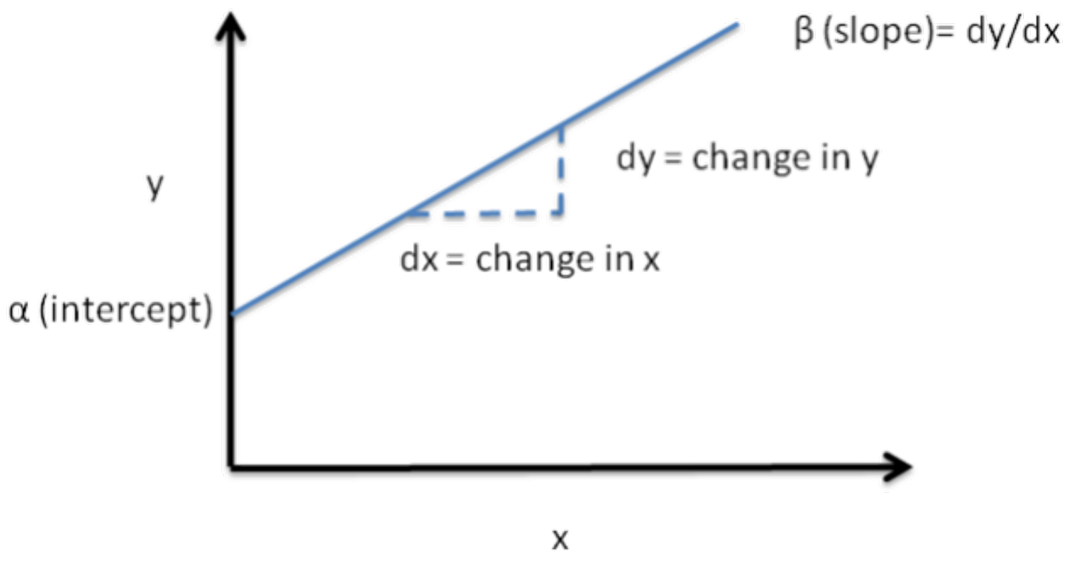


Strongest association? a

Strongest correlation? b

https://www.openintro.org/book/os/

LEAST SQUARES LINE



$$y = \alpha + \beta x + \varepsilon$$

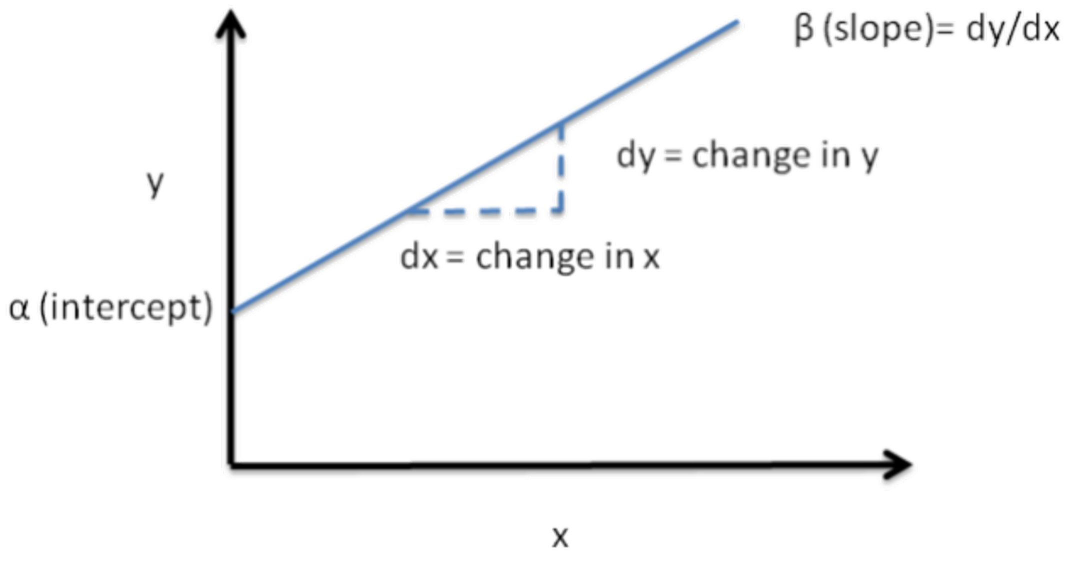
y is a continuous outcome variable
x is a continuous explanatory variable
α is the intercept.
point estimate (one value)
β is the slope or coefficient.
point estimate (one value)
ε is the residuals (error)
ε is normally distributed with mean 0

$$\hat{y} = \alpha + \beta x$$

$\hat{y}$ is the predicted data (line)

CAN BE USED FOR PREDICTIONS

TERMS



Intercept (α)

is the Y value of the regression line when X equals zero. It defines the elevation of the line.

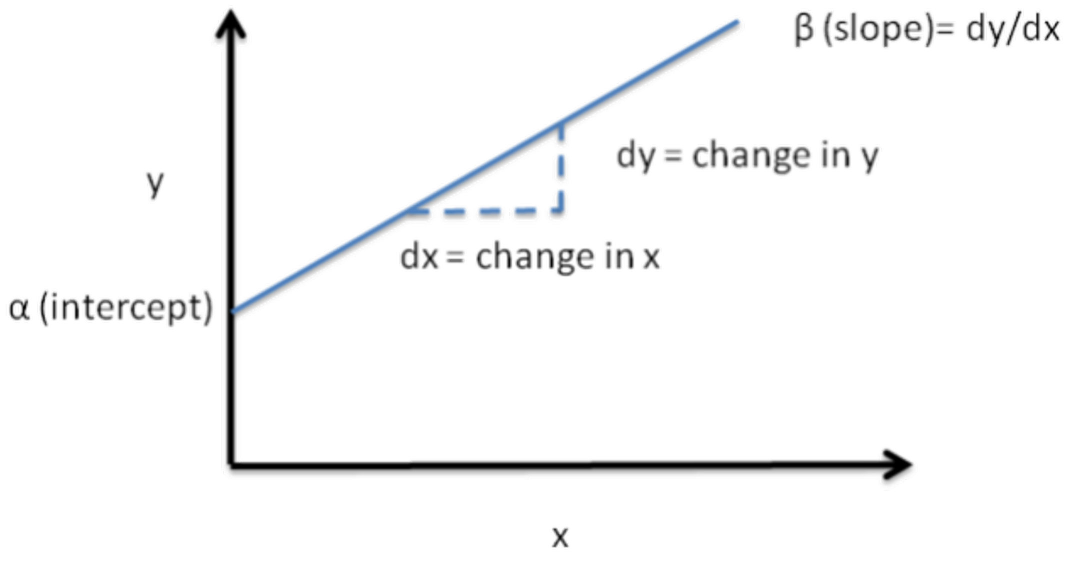
Regression coefficient (β)

It quantifies the slope of the line. It equals the change in outcome (Y) for each unit change in predictor (X). It is expressed in the units of the Y-axis divided by the units of the X-axis. If the slope is positive, Y increases as X increases. If the slope is negative, Y decreases as X increases.

Y and X can be tested to see if they are linearly related

NULL HYPOTHESIS: **No association, β=0**

TERMS



SE AND CONFIDENCE INTERNALS

The standard error values of the slope can be hard to interpret, but their main purpose is to compute the **95% confidence intervals**. You are confident that the real value of the coefficient that you are estimating falls somewhere in this interval 95% of the time.

P VALUE

Probability that this linear relationship is a chance finding.

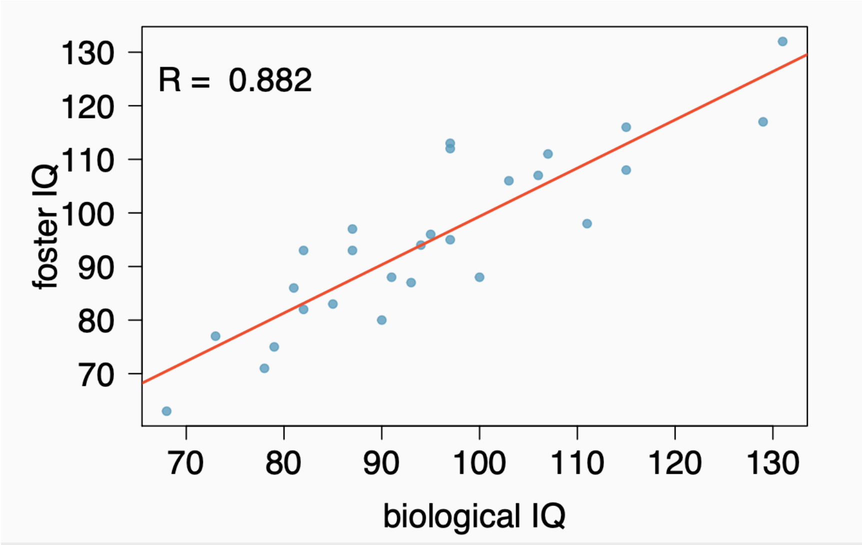
R^2

- is a statistical measure of how well a regression line approximates real data points.

  R=correlation

- How much variation in outcome (%) is explained by exposure?

INTERPRETATION



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Intercept = 9.21
Slope = 0.90

P value < 0.0001
95% CI = [0.71, 1.09]

OUTCOME: IQ of twin raised by foster parents
EXPOSURE: IQ of twin raised by biological parents

RELATIONSHIP

Linear
Positive
Strong. R=0.882

LINEAR REGRESSION

Intercept: 9.21 is the foster IQ if biological IQ is 0
Slope: Increase in 1 (10) biological IQ leads to increase in 0.9 (9) foster IQ
Confidence Interval: Real slope (effect size) is within this interval of [0.71, 1.09]

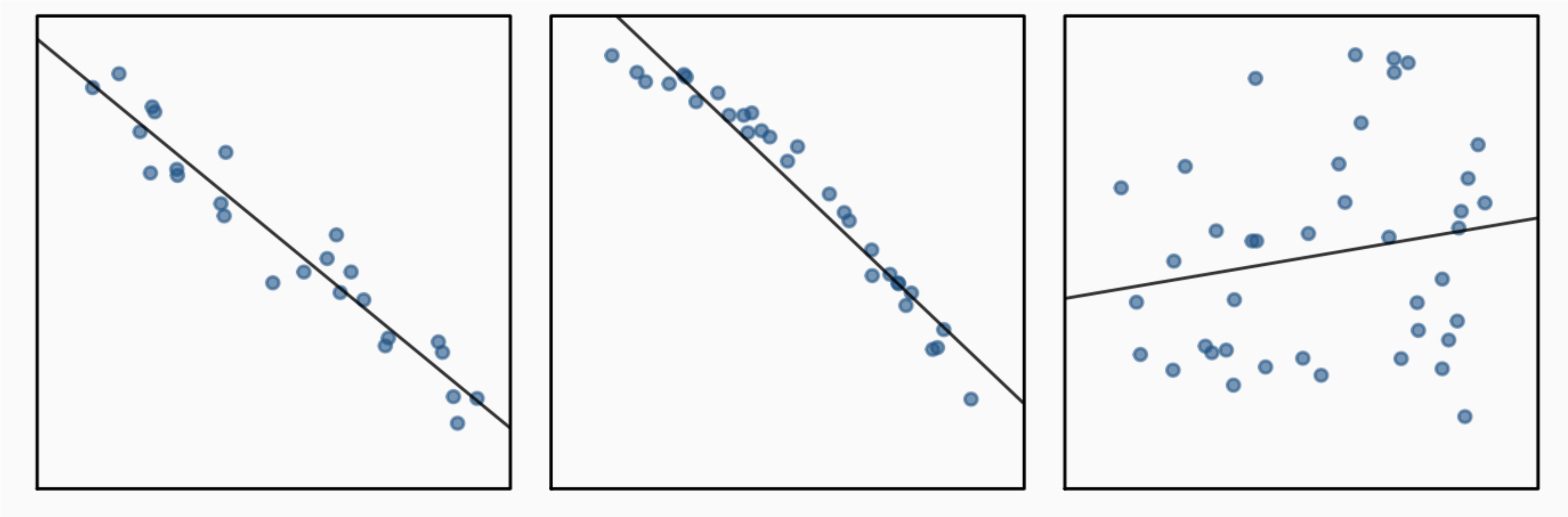
https://www.openintro.org/book/os/

ASSUMPTIONS

1. Linear relationship between exposure and outcome
2. Normality of residuals
3. Constant variability of residuals
4. No extreme outliers

1. LINEAR RELATIONSHIP

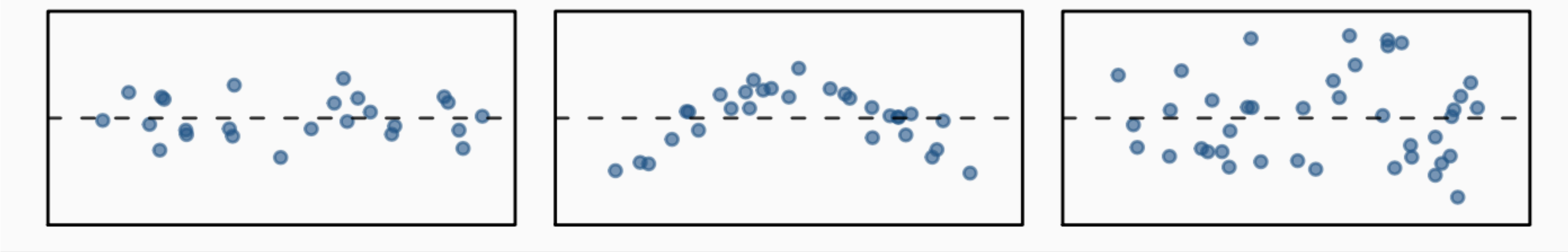
Y



X X X

Y vs X

Residuals



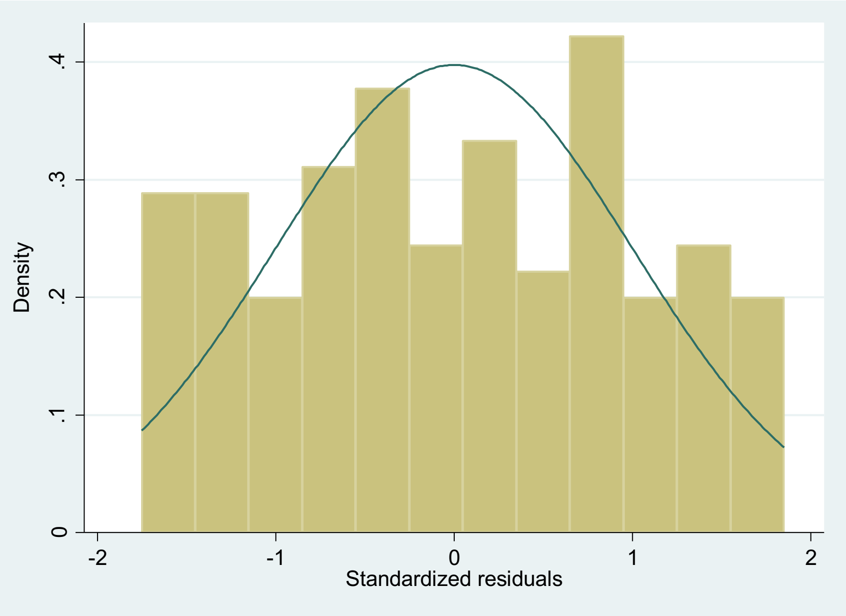
Predicted ($\hat{Y}$) Predicted ($\hat{Y}$) Predicted ($\hat{Y}$)

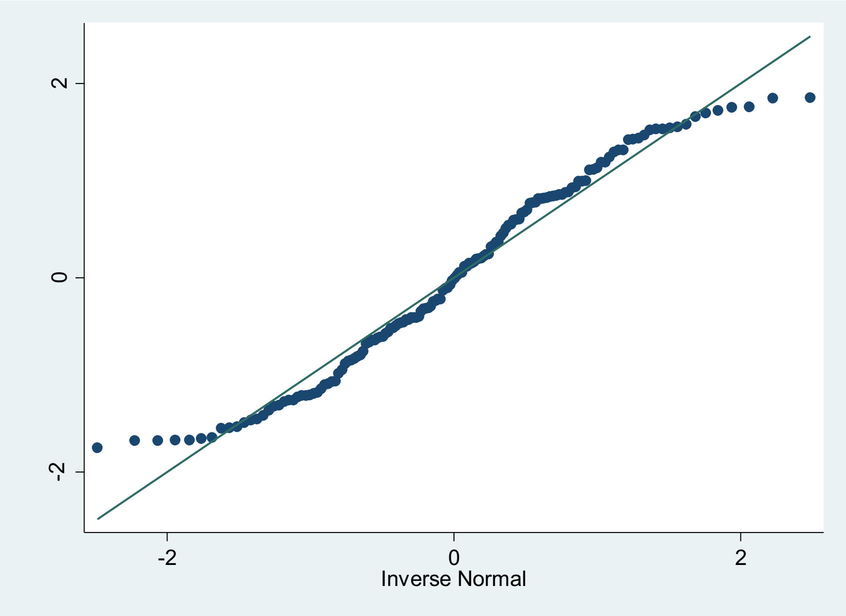
RESIDUAL PLOT

https://www.openintro.org/book/os/

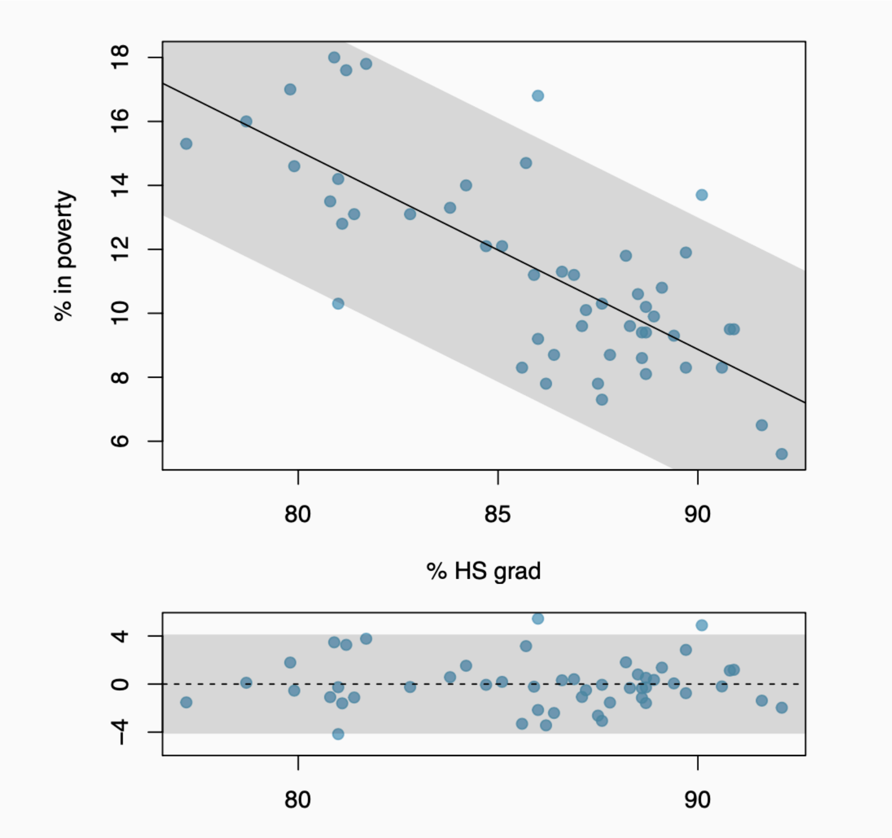
2. NORMALITY OF RESIDUALS

Histogram or QQ plot



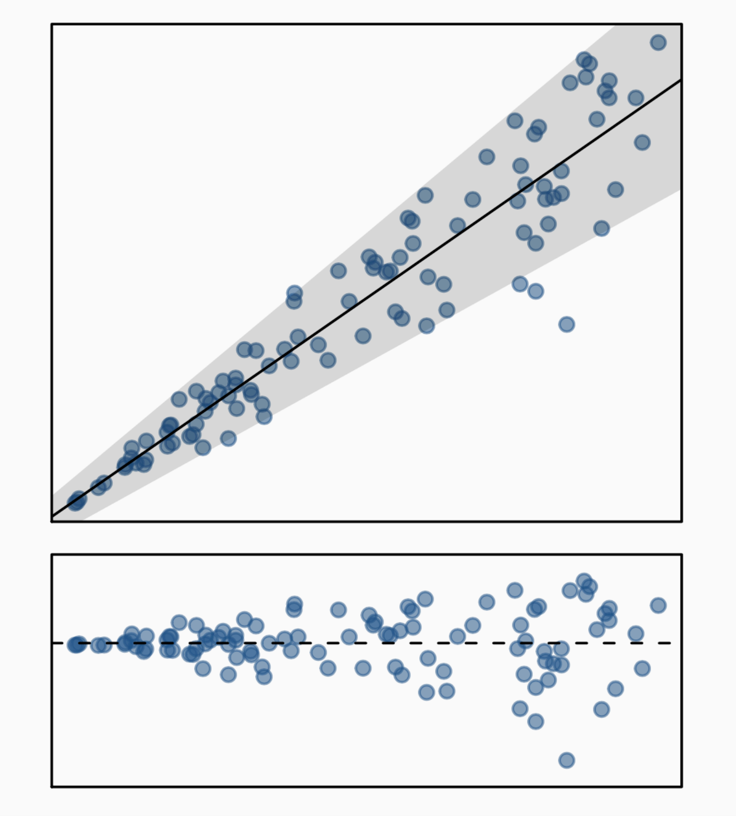


3. CONSTANT VARIABILITY OF RESIDUALS



Y vs X

RESIDUAL PLOT

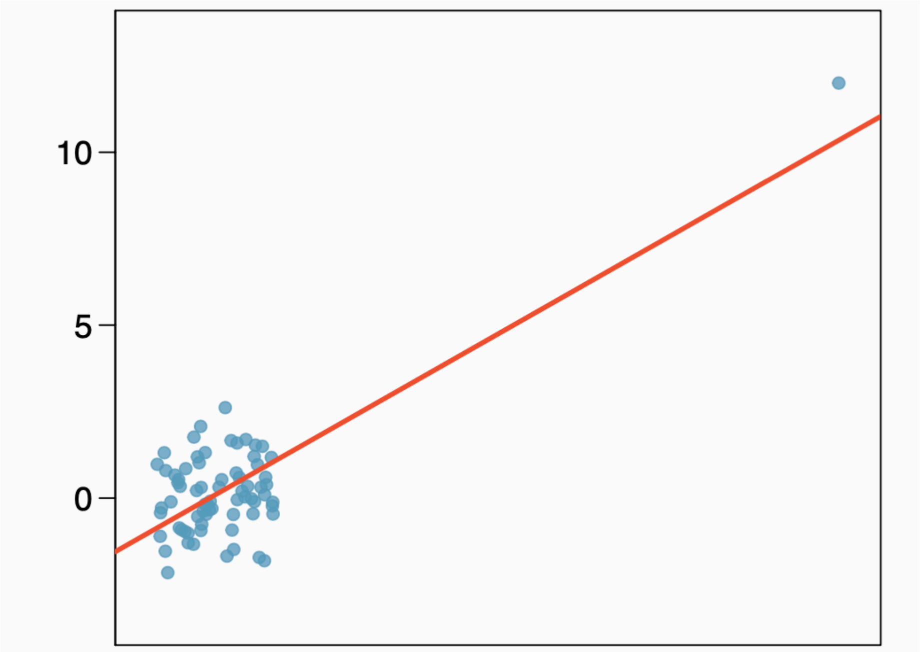


CONSTANT VARIANCE
Homoskedasticity

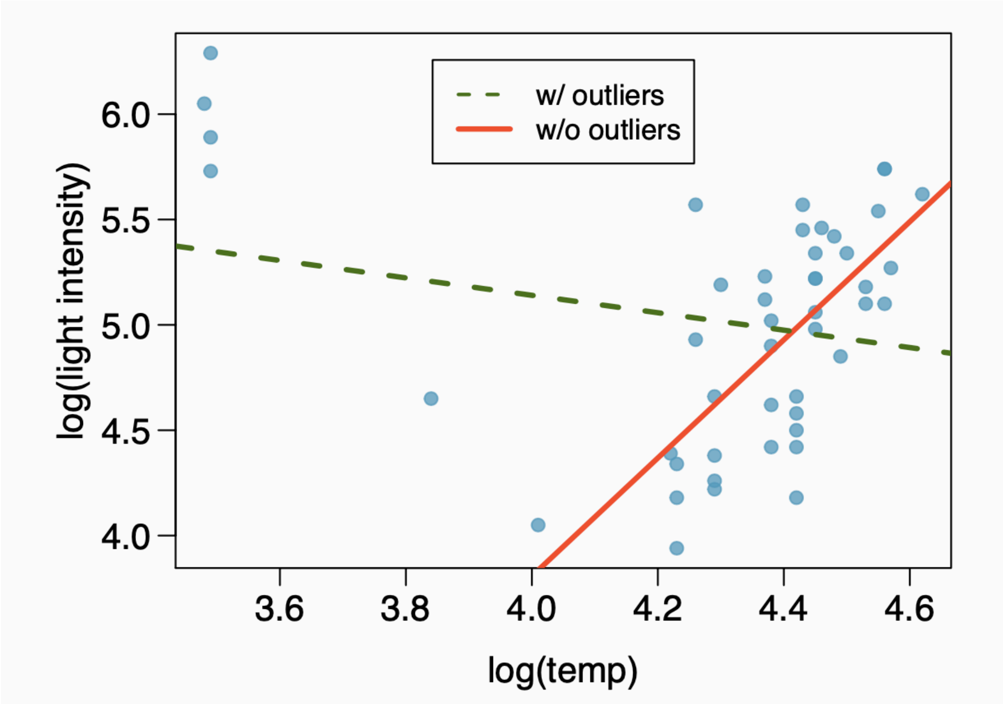
NON-CONSTANT VARIANCE
Heteroskedasticity

https://www.openintro.org/book/os/

4. NO EXTREME OUTLIERS



High leverage but not influential



High leverage and influential

https://www.openintro.org/book/os/

EXERCISE

Day3_2_ LinearRegression_Exercise.Rmd

STATISTICAL INFERENCE

– MULTIVARIABLE REGRESSION

MULTIVARIABLE REGRESSION

SIMPLE LINEAR REGRESSION:

1 Exposure vs. 1 Outcome

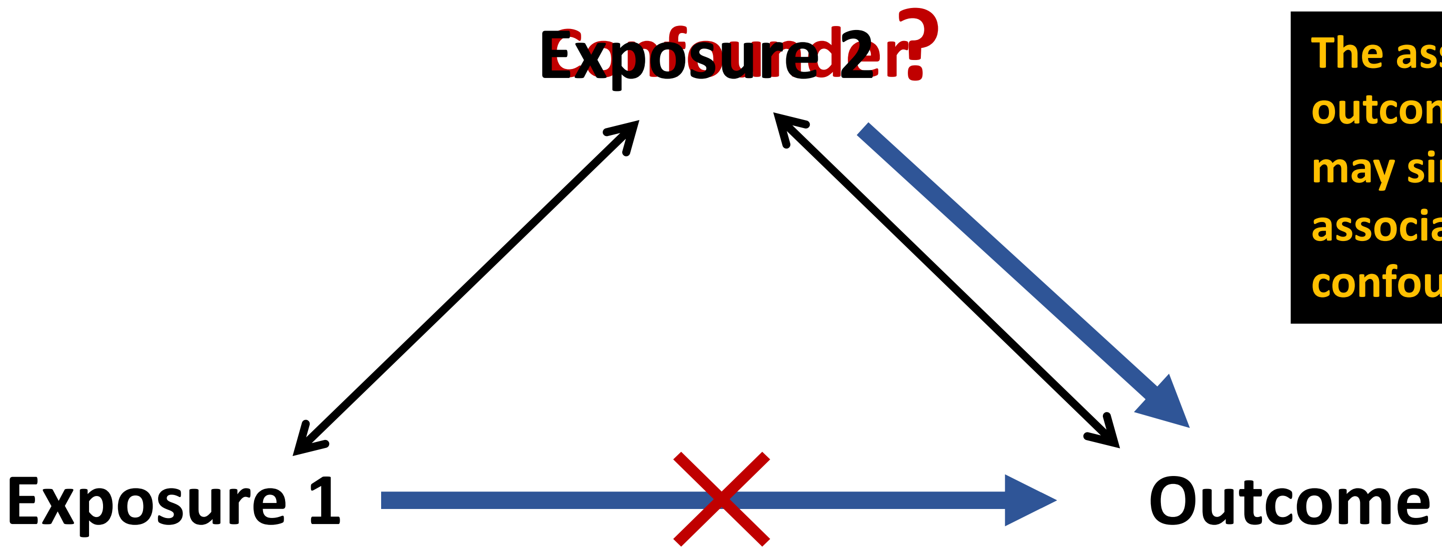
MULTIVARIABLE LINEAR REGRESSION:

2+ Exposures vs. 1 Outcome

Easily done in R, just add another variable with +

CONFOUNDER

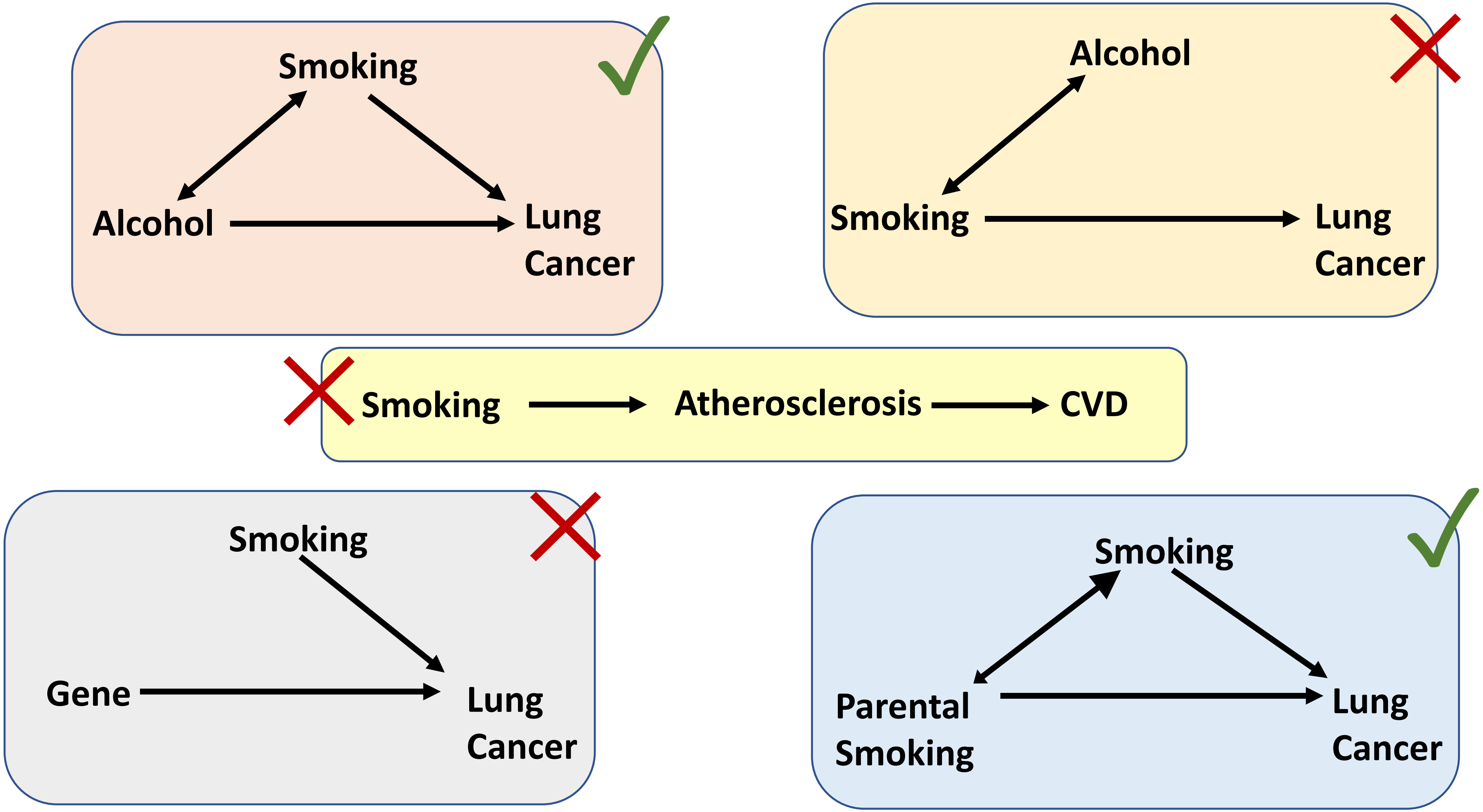
BE CAREFUL!



The association of the outcome with the exposure may simply reflect the association with the confounder

- Any variable that is not the outcome or the exposure is a potential ***confounder***.
- To be a confounder the variable must be **independently *associated with*** exposure and the outcome.
- There is no requirement for the association to exist in the population – it could be a chance feature of your sample!

WHICH OF THERE ARE CONFOUNDER?



FINAL THINGS TO CONSIDER – NOT IN SCOPE

- **COLLINEARITY**: If there is substantial correlation amongst the multiple exposures (predictor and covariates), the model becomes unstable and standard errors become larger.
    - ➢ Calculate variance inflation factors to check for this
- **CATEGORICAL VARIABLES:** If the categorical variables produce small subgroups the model will become unstable.
    - ➢ Combine levels of any covariates with particularly small numbers at some levels
- **MODEL SPECIFICATION:** Do not over-fit the model! With enough covariates can always produce a perfect fit (even with random predictors!). But the model won't be generalizable!
    - ➢ Always omit irrelevant covariates!
- **MULTIPLE TESTING:** Refers to any instance that involves the simultaneous testing of more than one hypothesis. If decisions about the individual hypotheses are based on the unadjusted p-values, then there is typically a large probability that some of the true null hypotheses will be rejected.
    - ➢ Adjust for p-values e.g. Benjamini-Hochberg Procedure decreases the false discovery rate

DAY 2 AND 3 – DATA ANALYSIS

Data Import

Wrangling

- Tidy + Manipulating
- Summarizing
- Cleaning

Exploration

- Visualization
- Descriptive Statistics

Day 2

Statistical Inference

- Foundation of inference
- Basic statistical tests
- Linear regression
- Multivariable regression

Day 3

- **USE THE SUITABLE MODEL**
- **ACCURATE INTERPRETATION IS KEY!**

EXERCISE

Day3_3_ MultivariableRegression_Exercise.Rmd

THANK YOU