Lecture 7 - GLM III Differences between two groups

Caspar J. van Lissa

2023-08-13

Categorical predictors

Binary variables

Examples of binary/dichotomous variables

- Nominal
  - Biological sex (male/female)
  - Student ethnicity (Dutch/foreign student)
  - Has tattoo, has pets (No/Yes)
- Ordinal
  - Performance on exam question (Fail/Pass)
  - Risk of disease (Low/High)

Ways of coding

sex	ethnicity	tattoo	exam	risk
Man	0	1	Fail	1
Woman	1	2	Pass	-1
Man	0	2	Fail	1
Woman	0	1	Pass	-1
Woman	0	1	Pass	1
Man	1	1	Pass	-1

Dummy coding

sexWoman	ethnicity1	tattoo2	examPass
0	0	0	0
1	1	1	1
0	0	1	0
1	0	0	1
1	0	0	1
0	1	0	1

Other ways of coding

- There are other ways to code binary variables
- Outside the scope of this course
- Further reading: https://stats.oarc.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis

Regression with dummy variables

Linear regression

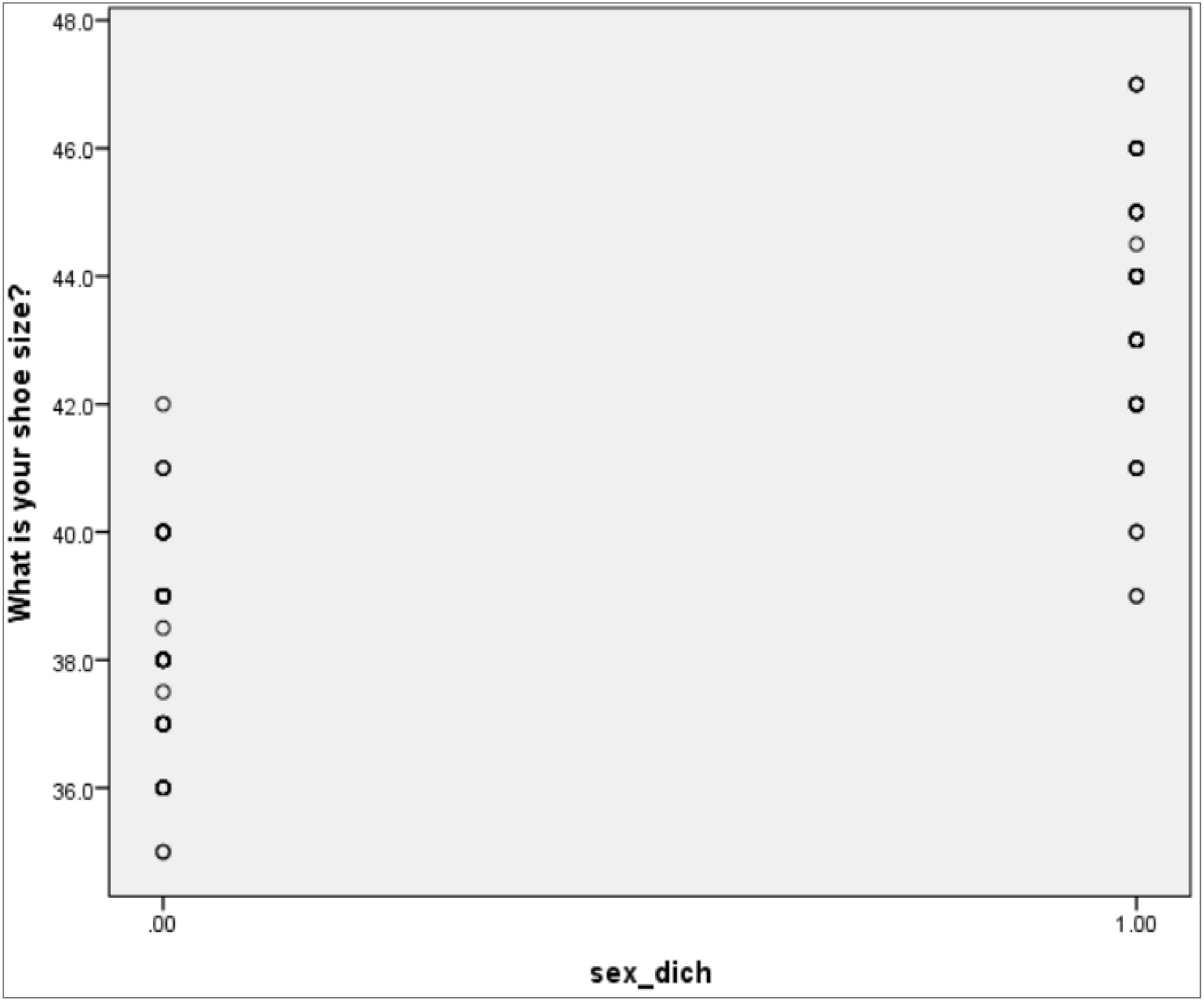
Linear regression is a basic model that can be adapted for various purposes
You've learned $Y_i = a + b * X_i + e_i$

- Where X is a continuous predictor

Today we examine how we can use the same model for *binary predictors*

Scatterplot

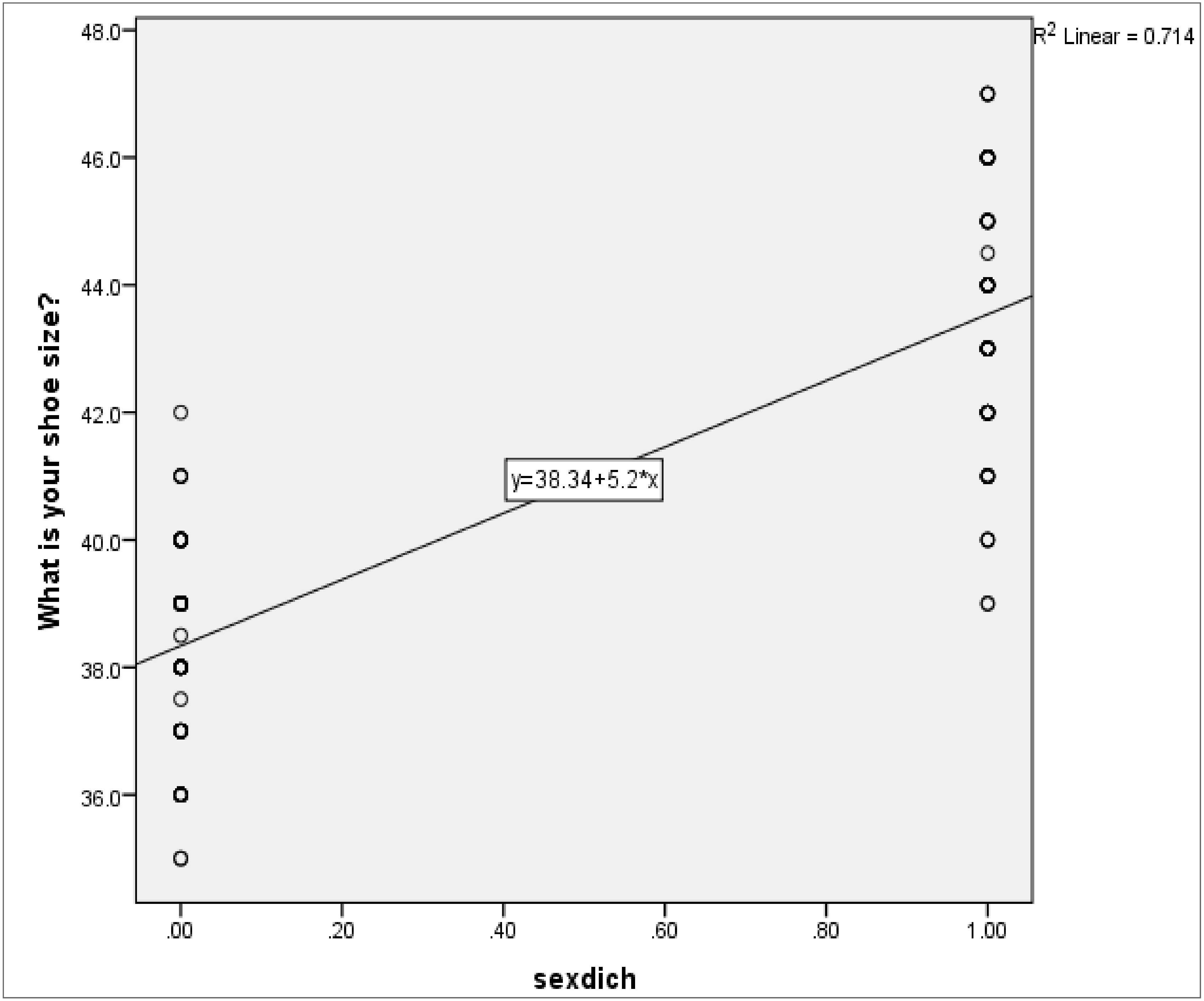
We have previously examined some scatterplots, including this one:



48.0
46.0
44.0
42.0
40.0
38.0
36.0
What is your shoe size?
.00
1.00
sex_dich

Scatterplot

You could imagine drawing a diagonal line through the two categories:



R2 Linear = 0.714
y=38.34+5.2*x
48.0
46.0
44.0
42.0
40.0
38.0
36.0
What is your shoe size?
.00
.20
.40
.60
.80
1.00
sexdich

Coefficients

The formula for a diagonal regression line is:
$Y = a + bX$
a is the intercept

- This is the predicted value **when X equals 0**

b is the slope, how steep the line is

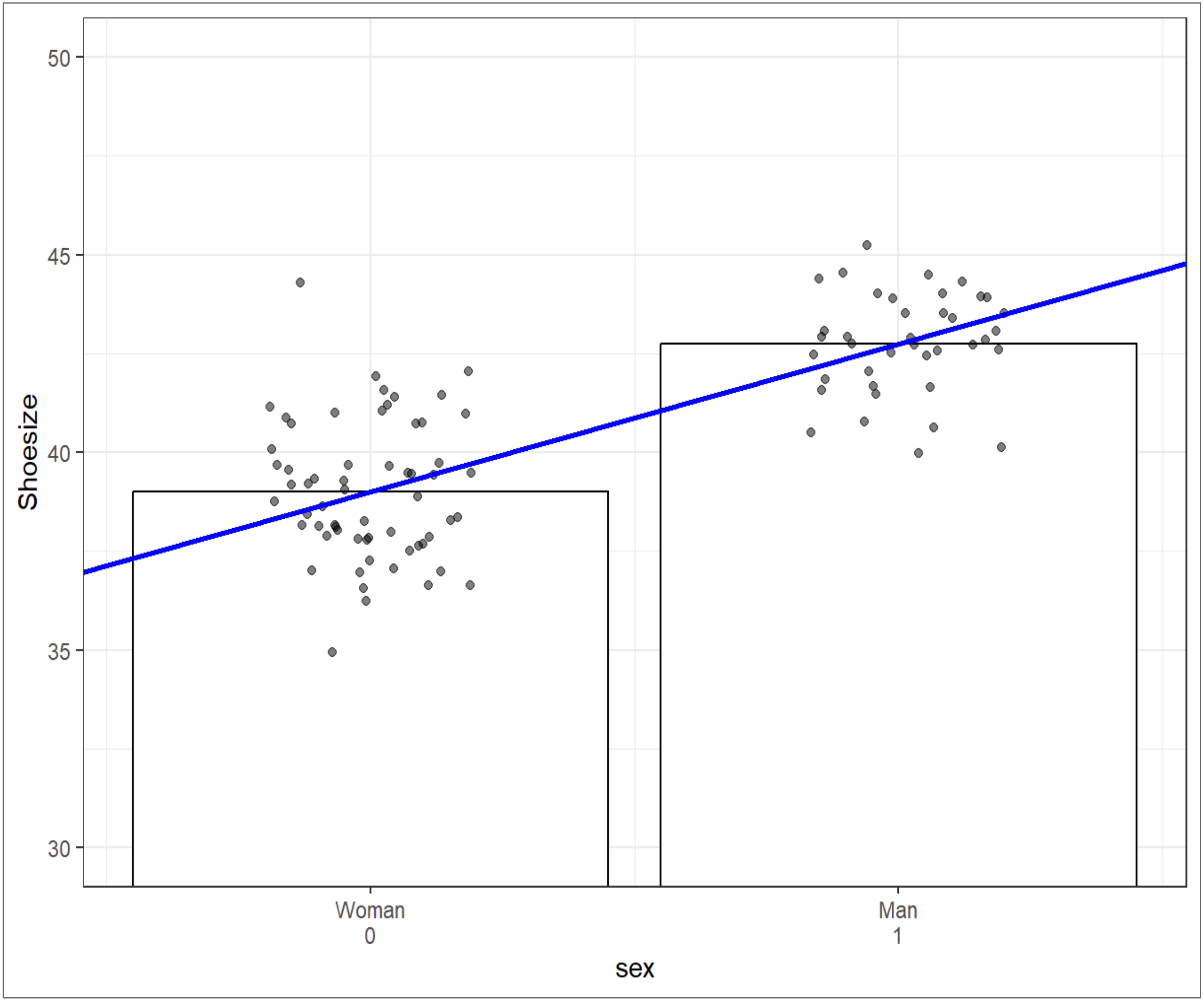
- Y increases by b **when X increases by 1**

Regression with binary predictor

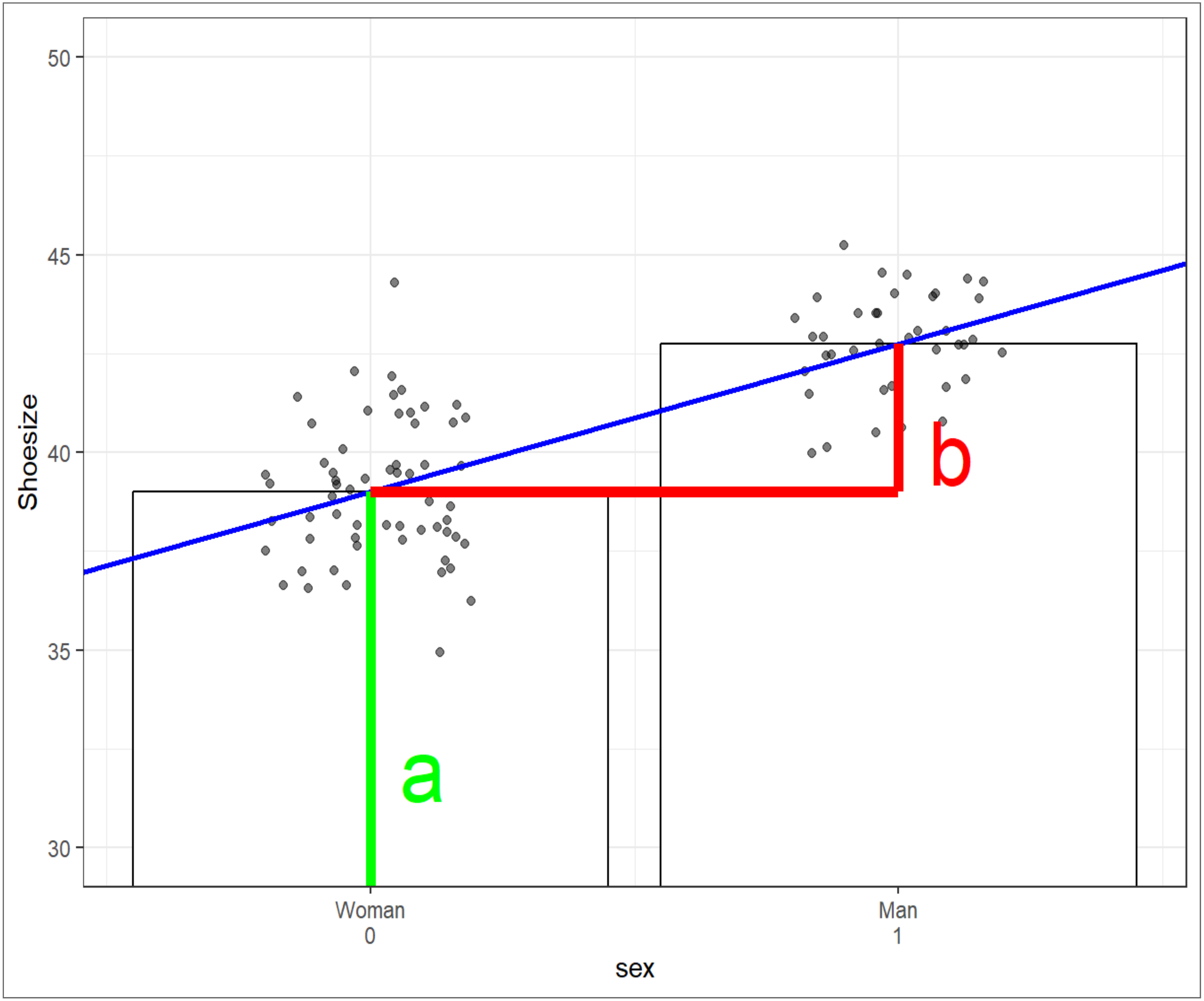
We make clever use of regression to include a binary predictor:

- Assign the value 0 to one of the categories
  - This is the “reference category”
- Assign the value 1 to the other category
- You can enter your data this way, or “recode” existing variables
- Regression will estimate the mean of the reference category and test the difference between the two categories!

Visualization


50
45
40
35
30
Shoesize
Woman
0
Man
1
sex

Visualisation



50
45
40
35
30
Shoesize
a
b
Woman
0
Man
1
sex

Formula

$$\hat{Y}_i = a + b * X_i$$

- $\hat{y}_i$: Individual predicted value of Shoesize
- a: Intercept
- b: Slope
- X_i: Sex (0 = woman, 1 = man)

Fill in the formula

Fill in for women:

- $\hat{Y}_i = a + b * 0 = a$
- So the predicted shoesize for women is the intercept (a)

Fill in for men:

- $\hat{Y}_i = a + b * 1 = a + b$
- So the predicted shoesize for men is the intercept (a) plus the difference between men and women (b)

Regression with dummy

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.845[a]	.714	.713	1.6393

a. Predictors: (Constant), Man

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1182.463	1	1182.463	440.044	.000[b]
	Residual	472.938	176	2.687		
	Total	1655.400	177			

a. Dependent Variable: What is your shoe size?

b. Predictors: (Constant), Man

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38.337	.163		235.033	.000
	Man	5.202	.248	.845	20.977	.000

a. Dependent Variable: What is your shoe size?

Independent samples t-test

Independent Samples t-Test

The independent samples t-test is used to compare the means of two independent groups

- It is equivalent to the t-test of the slope in regression with a binary predictor

Compare: t-test

Compare: Regression

Assumptions

The t-test has the same assumptions as bivariate linear regression, with slight nuances

- Linearity of relationship between X and Y
  - Difference between two groups is linear by definition
- Normality of residuals
  - The outcome is normally distributed in each group
- Homoscedasticity
  - Equality of Variances in both groups (Levene’s test)
- Independence of observations

Levene’s test

By default, SPSS tests the assumption of homoscedasticity using Levene’s test

- This is an F-test, as two sources of variance (i.e., the variances of the two groups) are compared
- $H_0 : S_1^2 = S_2^2$, which is equivalent to $H_0 : S_1^2 - S_2^2 = 0$
- If Levene’s test is *significant*, there is evidence that the two variances are not equal
- SPSS offers a test that allows for unequal variances

But remember: The notion of “assumption checks” risks overfitting data

- Better to make an informed guess (variances equal or not), and stick with it

Demo

M1 =

0

SD1 =

1

M2 =

1

SD2 =

1

Step 1. Hypotheses

The default hypothesis in most software is:

- $H_0 : \mu_1 = \mu 2$, which is equivalent to $H_0 : (\mu_1 - \mu_2) = 0$
- $H_A : (\mu_1 - \mu_2) \neq 0$

But a one-sided test is also possible:

- $H_0 : \mu_1 > \mu 2$, which is equivalent to $H_0 : (\mu_1 - \mu_2) > 0$
- $H_A : (\mu_1 - \mu_2) \leq 0$

Or custom hypothesis

Step 2. Test statistic

Observed group difference minus hypothesized group difference, divided by the appropriate standard error

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_{H_0}}{SE_{\bar{X}_1 - \bar{X}_2}}$$

Standard error:

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{S^2_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})}$$

Where

$$S^2_{pooled} = \frac{(n_1 - 1) * S_1^2 + (n_2 - 1) * S_2^2}{n_1 + n_2 - 2}$$

Step 3. P-value

Use the t-distribution with appropriate degrees of freedom

- $df : n_1 + n_2 - 2$, minus 2 because two parameters are being estimated
  - a and b, or the two means $\bar{X}_1$ and $\bar{X}_2$
- Find p-value in the t-table, online calculator, Excel or SPSS
- Remember: decide whether you assume equal variances or not

Step 4. Draw conclusion

- If $p < \alpha$, the test is significant
- It is very unlikely to observe a group difference at least as large as you observed, if H_0 were true

Effect size

Effect size

The t-test tells us whether the difference between groups is statistically significant

- But is it also *practically* significant?
- Remember statistical power
  - In a large enough sample, even trivial differences between groups become significant
- Effect size measures standardize the difference between the group means
- This makes it interpetable on a meaningful scale (i.e., number of standard deviations)

Visualization

Cohen's D

Cohen's d is an effect size for mean differences, calculated as:

$$\frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}}$$

Difference divided by pooled SD

Interpreting Cohen's d

Larger Cohen's d: bigger difference between the groups
Rule of thumb:

- Small effect size: $d \approx 0.2$
- Medium effect size: $d \approx 0.5$
- Large effect size: $d \approx 0.8$