Correlation and regression in R

Four weeks of PS2010 in one workshop!

Get the slides, code and data here

⬇⬇⬇

bit.ly/44sYP3

https://luke-kendrick.github.io/r-summer25/index.html

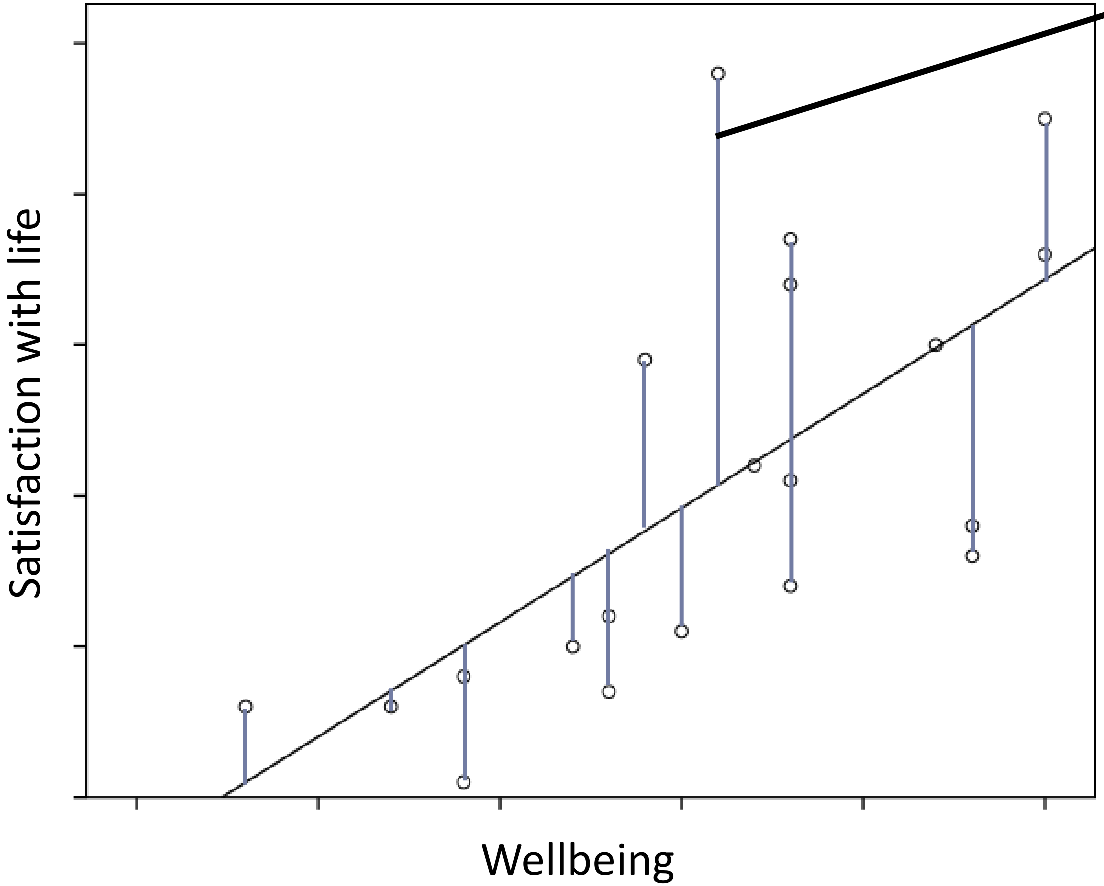
Demo (lecture) dataset: Plan for this workshop…

- The dataset: Lecture_data_R_sat_life.csv (data and R code on Moodle)
- Study with 200 adults, aiming to predict satisfaction with life (SWL)

- Four lectures of content to explore this dataset:

1. Correlations between variables
2. Multiple regression with continuous predictors
3. Multiple regression including binary predictors
4. Assumptions of multiple regression

Understanding "best fit" and variance in correlations



Residual: difference between raw data point and best fit line

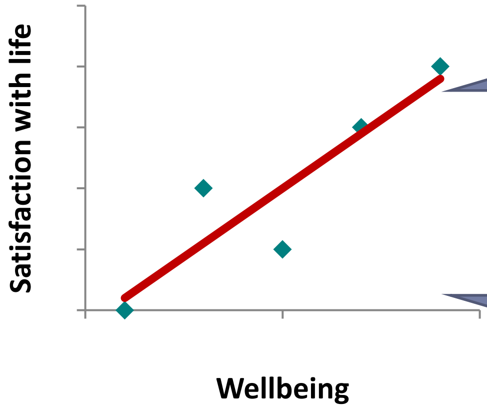
Small residuals indicate a more accurate model

Best fit = residuals are minimised

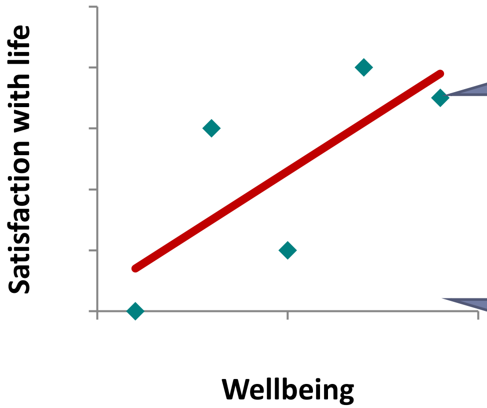
Reducing the random variability (residuals)

Understanding “best fit” and variance in correlations

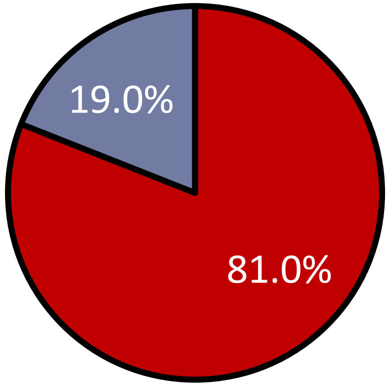
Line of best fit: Aims to reduce the distance between the individual data points and the line that describes the strength of the relationship



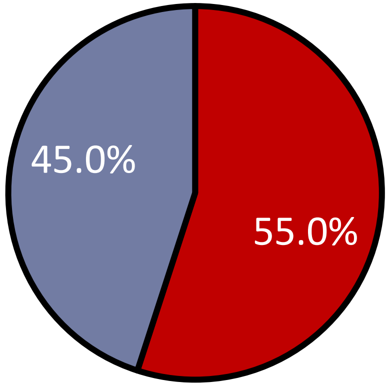
Less spread around the line = less random variance



More spread around the line = more random variance



Model/explained variance:
Variance explained by the line of best fit
Random/error/unexplained variance:
Variance between line of best fit and raw data



Demo dataset: What will I be showing you?

The continuous variables:

- Satisfaction with life (SWL): Questionnaire, scores 5-35, high = more satisfied
- Three wellbeing scales: Questionnaire, high = better wellbeing
  - Psychological (e.g. "Do you feel able to enjoy life?")
  - Physical: (e.g. "Are you happy with your opportunity for exercise/leisure?")
  - Relationships: (e.g. "Are you happy with your friendships and personal relationships?")
- Negative life experiences: number experienced in last 12 months (max. 12)
- Years of education – used as a control variable

The binary/categorical variables:

- Occupational status: 0 = not in employment, 1 = employed
- Relationship status: 0 = single, 1 = in a relationship
- Location of home: 0 = rural, 1 = urban

My dataset: Lecture_data_R_sat_life.csv

Getting R ready for correlation and regression analysis…

- First, set the working directory, and check if needed: `getwd()`
  - Session > Set working directory > Choose directory
- Next, install and load all the packages we need today…

```
install.packages(tidyverse)
install.packages(correlation)
install.packages(gridExtra)
install.packages(ppcor)
install.packages(cocor)
install.packages(car)

library(tidyverse)
library(correlation)
library(gridExtra)
library(ppcor)
library(cocor)
library(car)
```

Why do I need more than just tidyverse?

`correlation` – lets you calculate *r* values
`gridextra` – show multiple scatterplots in a grid
`ppcor` – lets you calculate partial correlations
`cocor` – lets you statistically compare correlations
`car` – needed for testing the statistical assumptions

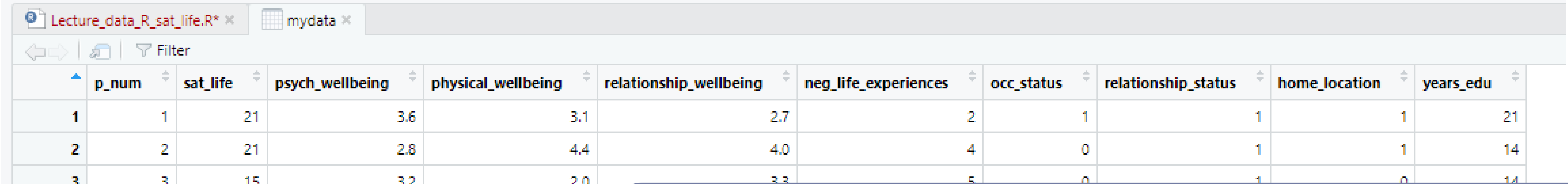
Demo dataset: Getting R ready for analysis…

- Now, open the dataset….

```
mydata <- read_csv("Lecture_data_R_sat_life.csv")
```

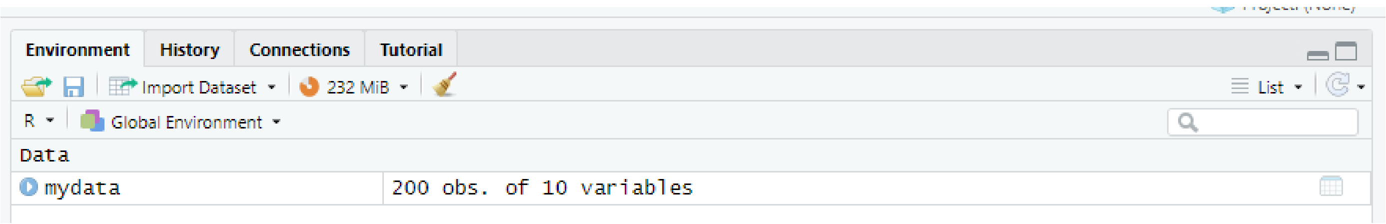
- If you want to look at the data in Excel sheet style…

```
view(mydata)
```



	p_num	sat_life	psych_wellbeing	physical_wellbeing	relationship_wellbeing	neg_life_experiences	occ_status	relationship_status	home_location	years_edu
1	1	21	3.6	3.1	2.7	2	1	1	1	21
2	2	21	2.8	4.4	4.0	4	0	1	1	14
3	3	15	3.2	2.0						
4	4	11	2.7	2.7						
5	5	23	4.2	4.2						
6	6	15	3.4	2.1						
7	7	14	3.3	2.9						
8	8	20	3.2	3.2						
9	9	17	1.4	1.8						
10	10	16	3.3	1.3						
11	11	20	3.1	2.8						
12	12	23	3.4	3.0						

You can also check in the "Environment" tab to see the number of participants (obs.) and number of variables…



Demo dataset: Getting R ready for analysis…

- Next, define variables as continuous (numeric) or binary (factor)
- If you want to double check the names of the variables in your dataset…

```
names(mydata)
```

Check that you have defined all ten variables (ten lines of code)

```
mydata$p_num <- as.numeric(mydata$p_num)
mydata$sat_life <- as.numeric(mydata$sat_life)
mydata$psych_wellbeing <- as.numeric(mydata$psych_wellbeing)
mydata$physical_wellbeing <- as.numeric(mydata$physical_wellbeing)
mydata$relationship_wellbeing <- as.numeric(mydata$relationship_wellbeing)
mydata$neg_life_experiences <- as.numeric(mydata$neg_life_experiences)
mydata$occ_status <- as.factor(mydata$occ_status)
mydata$relationship_status <- as.factor(mydata$relationship_status)
mydata$home_location <- as.factor(mydata$home_location)
mydata$years_edu <- as.numeric(mydata$years_edu)
```

A quick look at the descriptives first…

- Descriptives for the continuous, and frequencies for the categorical

```
summary(mydata)
```

```
     p_num           sat_life      psych_wellbeing physical_wellbeing relationship_wellbeing neg_life_experiences
 Min.   :  1.00   Min.   : 6.00   Min.   :1.400   Min.   :1.000      Min.   :1.200          Min.   :0.000
 1st Qu.: 50.75   1st Qu.:18.00   1st Qu.:2.700   1st Qu.:2.600      1st Qu.:2.500          1st Qu.:2.000
 Median :100.50   Median :20.00   Median :3.250   Median :3.200      Median :3.000          Median :3.000
 Mean   :100.50   Mean   :20.55   Mean   :3.252   Mean   :3.164      Mean   :3.041          Mean   :3.075
 3rd Qu.:150.25   3rd Qu.:23.00   3rd Qu.:3.800   3rd Qu.:3.600      3rd Qu.:3.600          3rd Qu.:4.000
 Max.   :200.00   Max.   :34.00   Max.   :4.900   Max.   :4.900      Max.   :4.900          Max.   :9.000
 occ_status relationship_status home_location   years_edu
 0: 88      0: 92               0: 86         Min.   :12.00
 1:112      1:108               1:114         1st Qu.:14.00
                                              Median :17.00
                                              Mean   :16.44
                                              3rd Qu.:17.00
                                              Max.   :25.00
```

- For example…
  - Mean satisfaction with life is 20.55
  - 92 people are not currently in a relationship, 108 are in a relationship

A quick look at the descriptives first…

- To split this by groups, such as satisfaction with life by occupational status
  - Ask R to give you the descriptives_bygroup
  - Then tell R what variable to group_by(occ_status)
  - Finally, which variable to summarise and which descriptives: mean sd

```
descriptives_bygroup <- mydata %>%
  group_by(occ_status) %>%
  summarise(mean_sat_life = mean(sat_life), sd_sat_life = sd(sat_life))
```

%>% means "now do the next thing"

- Last step is to print the descriptives table into the console

```
print(descriptives_bygroup)
```

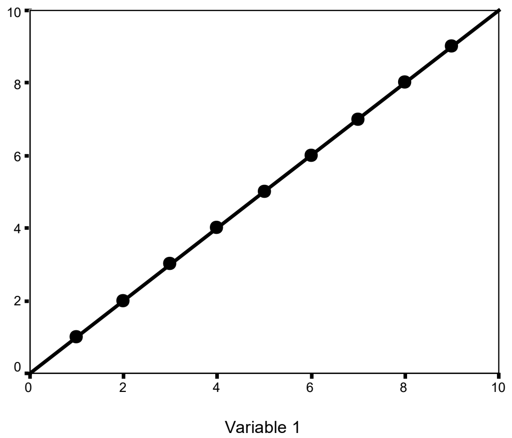
	occ_status <fct>	mean_sat_life <dbl>	sd_sat_life <dbl>
1	0	19.4	4.62
2	1	21.5	4.71

Part 1: Correlations and scatterplots

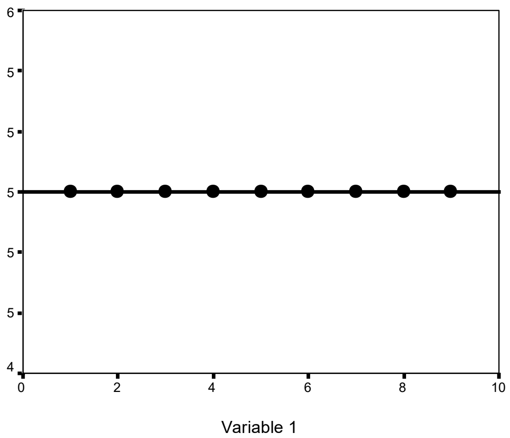
1. Graph correlations
   - How to create a scatterplot with a line of best fit
2. How to run Pearson's correlations
   - Understand and interpret positive and negative correlations
3. Run partial correlations
   - Understand control/confounding variables
4. Statistically compare correlations
   - How to graph two correlations on the same plot

Correlations (Pearson’s r)

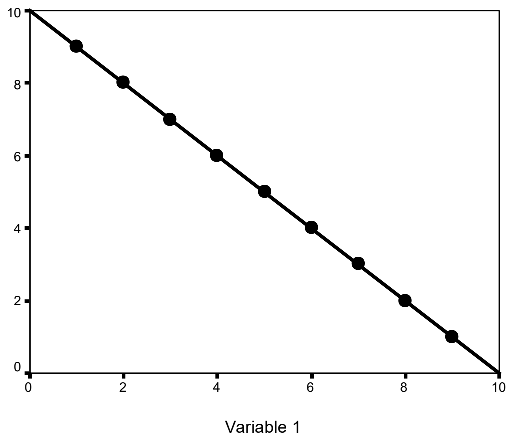
- Is there a *significant* relationship between two variables?
- Research question: Is SWL correlated with wellbeing and negative life events?



Positive correlation
+ive r values



No correlation
r = 0



Negative correlation
-ive r values

- r values range from –1 (perfect negative) to +1 (perfect positive)
- The ‘best fit’ line represents the relationship

Let’s start by making the scatterplots…

Building scatterplots in R…

- Four different scatterplots, with lines of best fit…
  - Plot 1: Psychological wellbeing on x axis
  - Plot 2: Physical wellbeing on x axis
  - Plot 3: Relationship wellbeing on x axis
  - Plot 4: Negative life events on x axis

All four plots have satisfaction with life on the y axis

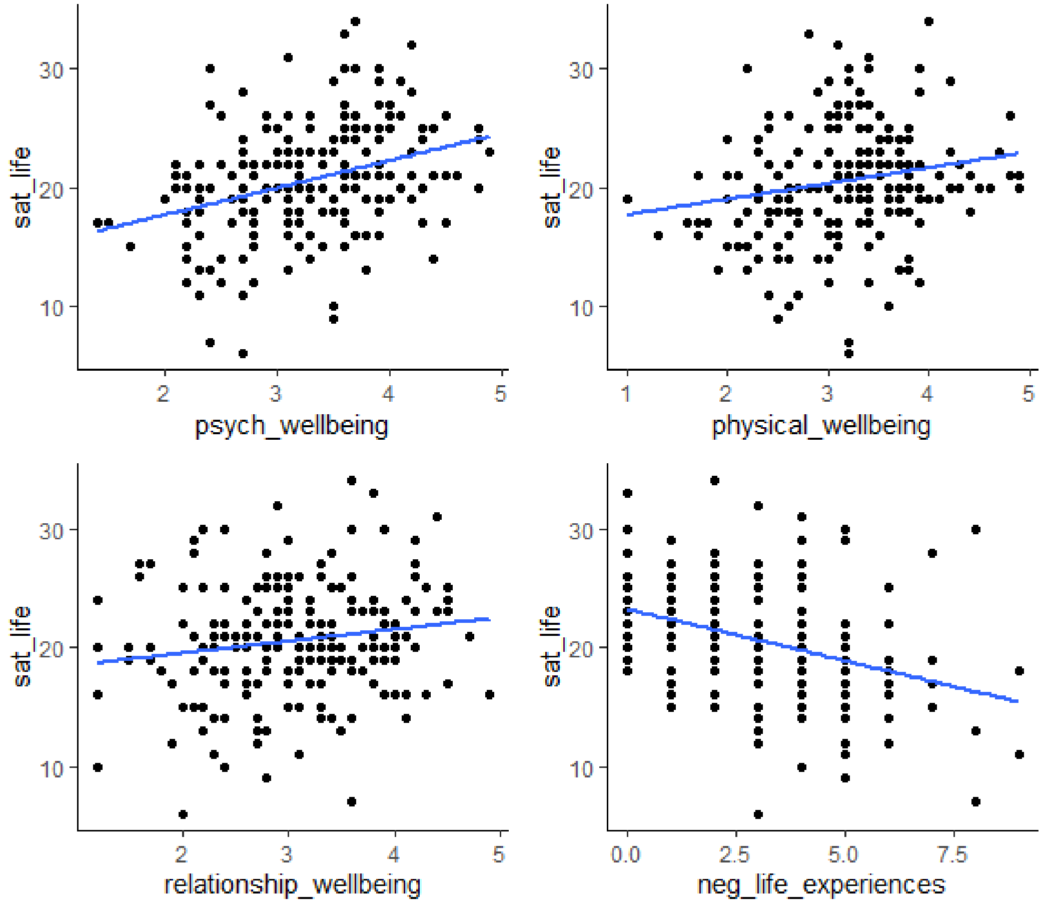
```
plot1 <- ggplot(mydata, aes(x = psych_wellbeing, y = sat_life)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE) +
  theme_classic()
```

Adapt this R code by changing the plot number and the variable on the x axis – you should have four sets of code!

Display the scatterplots in R…

```
grid.arrange(plot1, plot2, plot3, plot4, nrow = 2, ncol = 2)
```

HINT: “nrow = 2, ncol = 2” creates a 2x2 grid to display all four graphs



What if I just want to print one plot?

```
print(plot1)
```

Running Pearson's correlations in R

```
mydata %>%
  dplyr::select(sat_life, psych_wellbeing, physical_wellbeing,
relationship_wellbeing, neg_life_experiences) %>%
  correlation(p_adjust = "none")
```

- Add all the continuous variables you want to correlate
- Tell R to run a correlation (do not adjust the *p* value for multiple tests)
- A quick reminder of APA format for presenting correlations…

$$\boxed{r\ (\mathrm{df}) = .XX,\ p < .XXX}$$

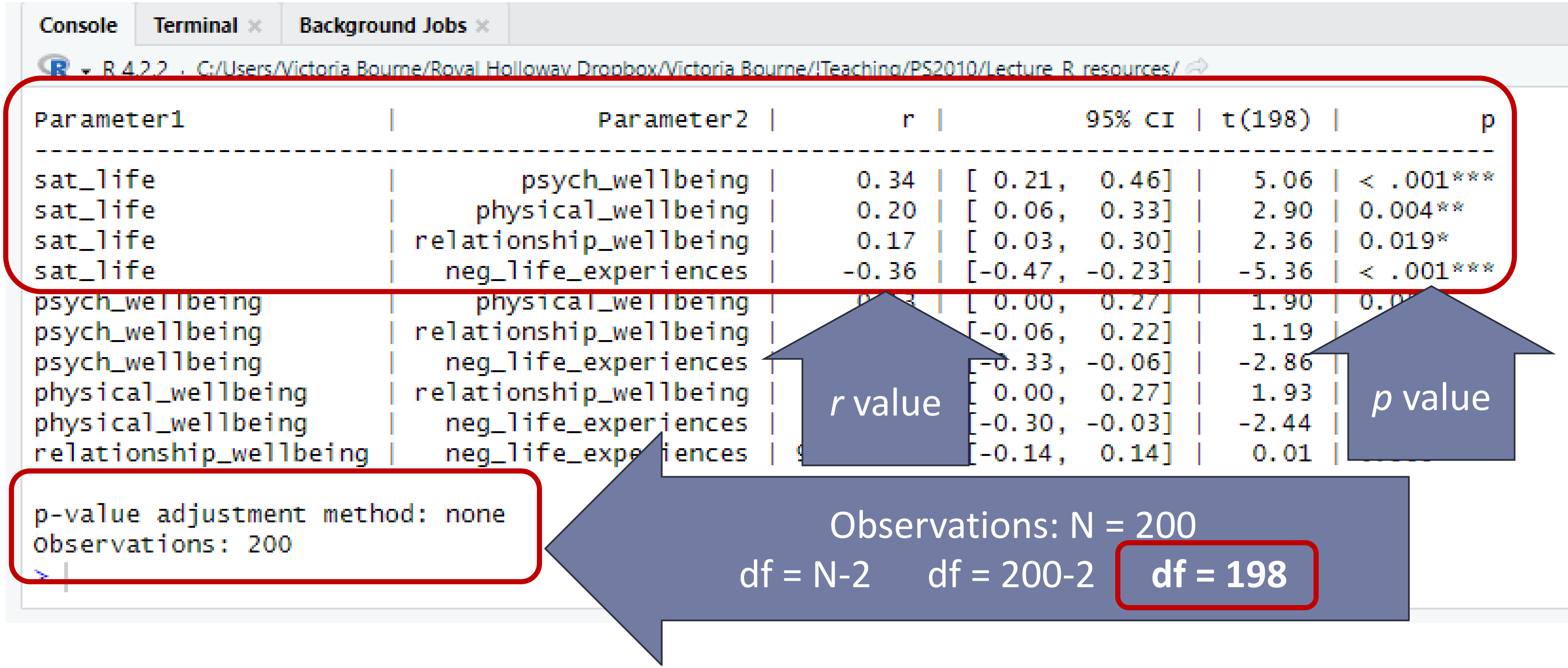
Tells you which statistic you calculated (*r*) lowercase and italicised

Tells you the degrees of freedom (N – 2)

Tells you the statistic calculated value (2 d.p.)

Tells you the significance (*p* value) lowercase and italicised

Interpreting Pearson's correlations in R



Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/Victoria Bourne/Royal Holloway Dropbox/Victoria Bourne/!Teaching/PS2010/Lecture R resources/

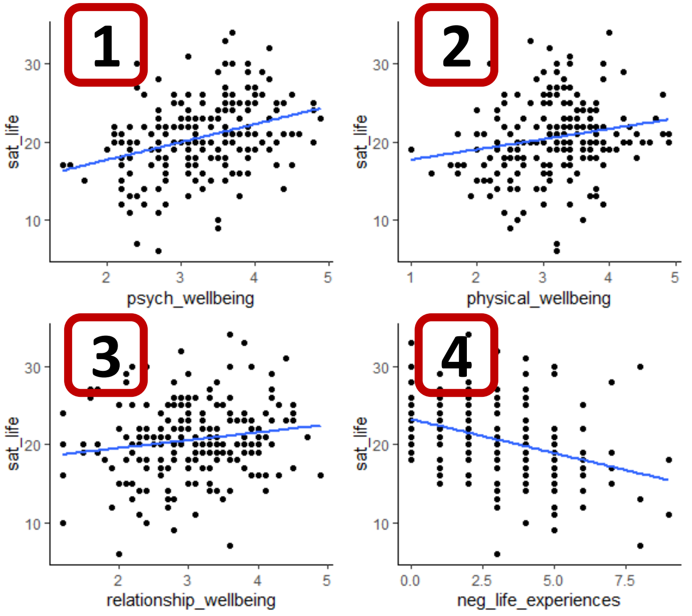
Parameter1	Parameter2	r	95% CI	t(198)	p
sat_life	psych_wellbeing	0.34	[0.21, 0.46]	5.06	< .001***
sat_life	physical_wellbeing	0.20	[0.06, 0.33]	2.90	0.004**
sat_life	relationship_wellbeing	0.17	[0.03, 0.30]	2.36	0.019*
sat_life	neg_life_experiences	-0.36	[-0.47, -0.23]	-5.36	< .001***
psych_wellbeing	physical_wellbeing	[illegible]	[0.00, 0.27]	1.90	[illegible]
psych_wellbeing	relationship_wellbeing	[illegible]	[-0.06, 0.22]	1.19	[illegible]
psych_wellbeing	neg_life_experiences	[illegible]	[-0.33, -0.06]	-2.86	[illegible]
physical_wellbeing	relationship_wellbeing	[illegible]	[0.00, 0.27]	1.93	[illegible]
physical_wellbeing	neg_life_experiences	[illegible]	[-0.30, -0.03]	-2.44	[illegible]
relationship_wellbeing	neg_life_experiences	[illegible]	[-0.14, 0.14]	0.01	[illegible]

```
p-value adjustment method: none
Observations: 200
```

Interpreting Pearson's correlations in R

```
Console    Terminal ×    Background Jobs ×
R 4.2.2 · C:/Users/Victoria Bourne/Royal Holloway Dropbox/Victoria Bourne/!Teaching/PS2010/Lecture R resources/

Parameter1 |                Parameter2 |     r |         95% CI | t(198) |         p
------------------------------------------------------------------------------------
sat_life   |            psych_wellbeing |  0.34 | [ 0.21,  0.46] |   5.06 | < .001***
sat_life   |         physical_wellbeing |  0.20 | [ 0.06,  0.33] |   2.90 | 0.004**
sat_life   |     relationship_wellbeing |  0.17 | [ 0.03,  0.30] |   2.36 | 0.019*
sat_life   |        neg_life_experiences | -0.36 | [-0.47, -0.23] |  -5.36 | < .001***
```

1. SWL and psychological wellbeing: r (198) = .34, p < .001
   - A significant positive relationship
2. SWL and physical wellbeing: r (198) = .20, p = .004
   - A significant positive relationship
3. SWL and relationship wellbeing: r (198) = .17, p = .019
   - A significant positive relationship
4. SWL and negative life events: r (198) = -.36, p < .001
   - A significant negative relationship



Running partial correlations in R

```
pcor.test(mydata$sat_life, mydata$psych_wellbeing,
          mydata$years_edu,
          method = "pearson")
```

- Tell R to run a partial correlation
- Tell it which variables from "my data" you want to correlate
- Tell it which variables from "my data" you want to control for

You need to have a separate piece of R code for each partial correlation!

1. SWL and psychological wellbeing: `mydata$sat_life, mydata$psych_wellbeing,`
2. SWL and physical wellbeing: `mydata$sat_life, mydata$physical_wellbeing,`
3. SWL and relationship wellbeing: `mydata$sat_life, mydata$relationship_wellbeing,`
4. SWL and negative life events: `mydata$sat_life, mydata$neg_life_experiences,`

Interpreting partial correlations in R

- The partial correlation between SWL and psychological wellbeing, controlling for years of education

r (197) = .31, p < .001

```
>
> pcor.test(mydata$sat_life, mydata$psych_wellbeing,
+           mydata$years_edu,
+           method = "pearson")
   estimate      p.value statistic   n gp  Method
1 0.3146133 6.020652e-06  4.652039 200  1 pearson
```

Tells you the r statistic

Present to 2 decimal places

Tells you the p value

Remember: e-06 means move the decimal point 6 places to the left

p = .000006 or p < .001

Tells you the N

df = N – 2 – number of control vars

df = 200 – 2 – 1

df = 197

Comparing correlations in R: Four steps

1

- Define which groups you want to look at separately
- Relationship status: 0 = single, 1 = in a relationship

2

- Run the correlations, specifying which group you are looking at
- A separate piece of R code for each group (single or in a relationship)

3

- Compare two correlations: Is one significantly stronger than the other?
- For this, you need the N and *r* value for each of the correlations

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

Comparing correlations in R: Step One

1

- Define which groups you want to look at separately
- Relationship status: 0 = single, 1 = in a relationship

```
single <- mydata[mydata$relationship_status == "0", ]
relationship <- mydata[mydata$relationship_status == "1", ]
```

- Define by name the group you want to create
  - You will use this name in the R code to select out particular participants
- Tell R which variable from "my data" has the grouping information
- Tell R the value that determines which group someone belongs to
  - A value of "0" means a person belongs in the "single" group
  - A value of "1" means a person belongs in the "in a relationship" group

Comparing correlations in R: Step Two

2

- Run the correlations, specifying which group you are looking at
- A separate piece of R code for each group (single or in a relationship)

```
cor.test(single$sat_life, single$neg_life_experiences,
          method = "pearson")
```

- Tell R which previously defined group you want to look at
- Which two variables do you want to correlate?
  - For now, let's just look at SWL and negative life experiences
- What kind of correlation do you want R to run?

Now repeat this for the other group

```
cor.test(relationship$sat_life, relationship$neg_life_experiences,
          method = "pearson")
```

Comparing correlations in R: Step Two

2

- Run the correlations, specifying which group you are looking at
- A separate piece of R code for each group (single or in a relationship)

Correlation for "single"

```
> cor.test(single$sat_life, single$neg_life_experiences,
+          method = "pearson")

	Pearson's product-moment correlation

data:  single$sat_life and single$neg_life_experiences
t = -5.8042, df = 90, p-value = 9.557e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6565266 -0.3550171
sample estimates:
       cor
-0.5218863
```

r value

Correlation for "relationship"

```
> cor.test(relationship$sat_life, relationship$neg_life_experiences,
+          method = "pearson")

	Pearson's product-moment correlation

data:  relationship$sat_life and relationship$neg_life_experiences
t = -1.8204, df = 106, p-value = 0.07153
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.35151720  0.01537078
sample estimates:
       cor
-0.1741089
```

r value

$r\ (90) = -.52, p < .001$

Do these *r* values differ significantly?

$r\ (106) = -.17, p = .072$

Comparing correlations in R: Step Three

3

- Compare two correlations: Is one significantly stronger than the other?
- For this, you need the N and *r* value for each of the correlations

	***r* value** (remember the –ive!)	**N** (df + 2 = N)
Correlation for single participants	-0.5218863	92
Correlation for participants in a relationship	-0.1741089	108

```
cocor.indep.groups(r value 1, r value 2, N 1, N 2)
```

```
cocor.indep.groups(-0.5218863, -0.1741089, 92, 108)
```

```
fisher1925: Fisher's z (1925)
  z = -2.7972, p-value = 0.0052
  Null hypothesis rejected
```

The two correlations differ significantly (z = -2.80, p = .005)

Comparing correlations in R: Step Four

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

```
plot_cc <- ggplot(mydata, aes(x = neg_life_experiences, y = sat_life, colour = relationship_status)) +
  geom_point(aes(shape = relationship_status)) +
  geom_smooth(aes(linetype = relationship_status), method = "lm", se = FALSE) +
  labs(title = "Negative life experiences vs Satisfaction with life by Relationship status",
       x = "Negative life experiences",
       y = "Satisfaction with life") +
  theme_classic() +
  scale_color_manual(values = c("0" = "grey", "1" = "black ")) +
  scale_linetype_manual(values = c("0" = "solid", "1" = "dashed")) +
  scale_shape_manual(values = c("0" = 16, "1" = 3))
```

Ok – this is a lot of code!
What does it mean?

Comparing correlations in R: Step Four

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

```
plot_cc <- ggplot(mydata, aes(x = neg_life_experiences, y = sat_life, colour = relationship_status)) +
  geom_point(aes(shape = relationship_status)) +
  geom_smooth(aes(linetype = relationship_status), method = "lm", se = FALSE) +
  labs(title = "Negative life experiences vs Satisfaction with life by Relationship status",
       x = "Negative life experiences",
       y = "Satisfaction with life") +
  theme_classic() +
  scale_color_manual(values = c("0" = "grey", "1" = "black ")) +
  scale_linetype_manual(values = c("0" = "solid", "1" = "dashed")) +
  scale_shape_manual(values = c("0" = 16, "1" = 3))
```

What variables are you plotting?

Which continuous variable should be plotted on the x axis (horizontal)?

Which continuous variable should be plotted on the y axis (vertical)?

Which categorical variable defines the separate groups we want to see?

Comparing correlations in R: Step Four

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

```
plot_cc <- ggplot(mydata, aes(x = neg_life_experiences, y = sat_life, colour = relationship_status)) +
  geom_point(aes(shape = relationship_status)) +
  geom_smooth(aes(linetype = relationship_status), method = "lm", se = FALSE) +
  labs(title = "Negative life experiences vs Satisfaction with life by Relationship status",
       x = "Negative life experiences",
       y = "Satisfaction with life") +
  theme_classic() +
  scale_color_manual(values = c("0" = "grey", "1" = "black ")) +
  scale_linetype_manual(values = c("0" = "solid", "1" = "dashed")) +
  scale_shape_manual(values = c("0" = 16, "1" = 3))
```

What titles should be displayed?

What is the title of your graph?

What is the title for your x axis (horizontal)?

What is the title for your y axis (vertical)?

Comparing correlations in R: Step Four

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

```
plot_cc <- ggplot(mydata, aes(x = neg_life_experiences, y = sat_life, colour = relationship_status)) +
  geom_point(aes(shape = relationship_status)) +
  geom_smooth(aes(linetype = relationship_status), method = "lm", se = FALSE) +
  labs(title = "Negative life experiences vs Satisfaction with life by Relationship status",
       x = "Negative life experiences",
       y = "Satisfaction with life") +
  theme_classic() +
  scale_color_manual(values = c("0" = "grey", "1" = "black ")) +
  scale_linetype_manual(values = c("0" = "solid", "1" = "dashed")) +
  scale_shape_manual(values = c("0" = 16, "1" = 3))
```

How are groups visually distinguished?

What colour should each group be? See https://r-charts.com/colors/

What line style should each group have? See https://r-charts.com/base-r/line-types/

What shape should data points be? See https://r-charts.com/base-r/pch-symbols/

Comparing correlations in R: Step Four

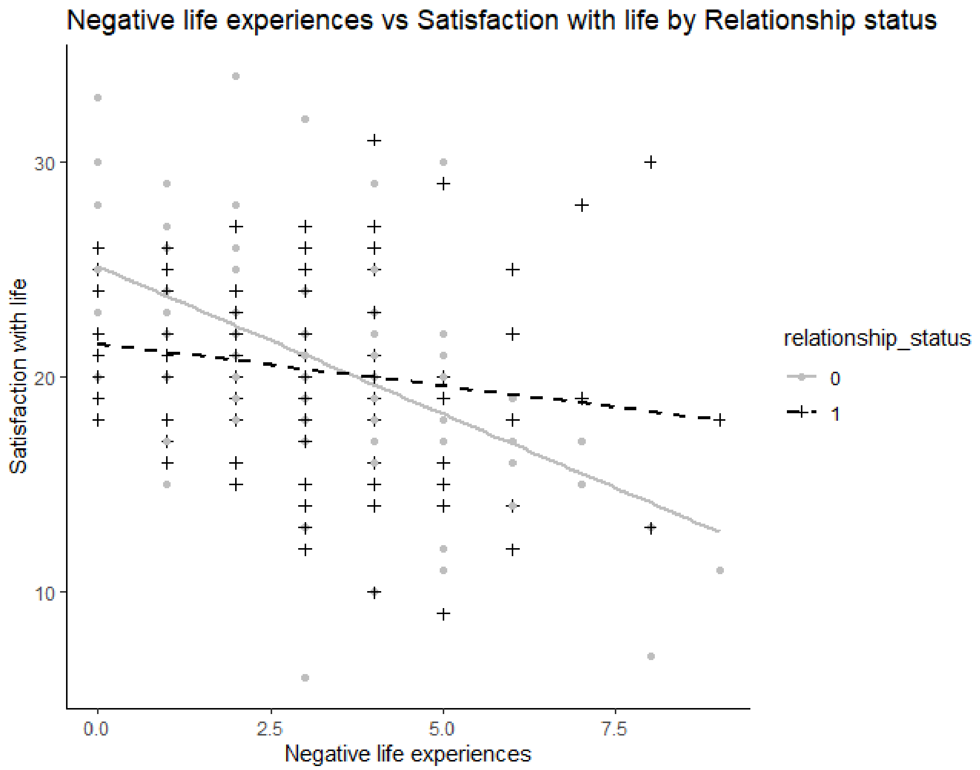
4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

How do I now see this beautiful graph?

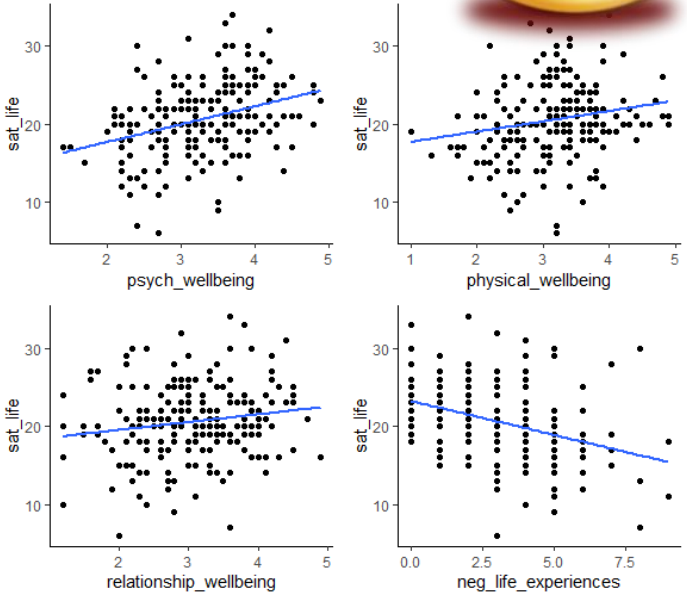
```
print(plot_cc)
```

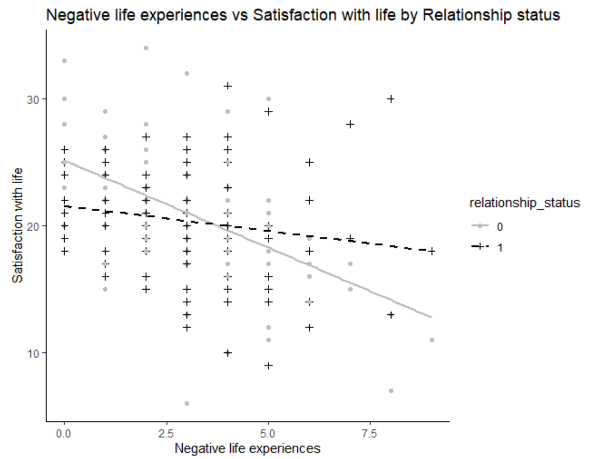
What does all this mean?



What does all this mean?

- SWL is positively correlated with all three measures of wellbeing, and negatively correlated with experiencing negative life events
- These correlations reduce slightly when controlling for years of education
- When looking at the correlation between SWL and experiencing negative life events for people who are single or in a relationship separately
  - For single people: There is a significant negative correlation
  - For those in a relationship: There is no correlation
  - These correlations differ significantly





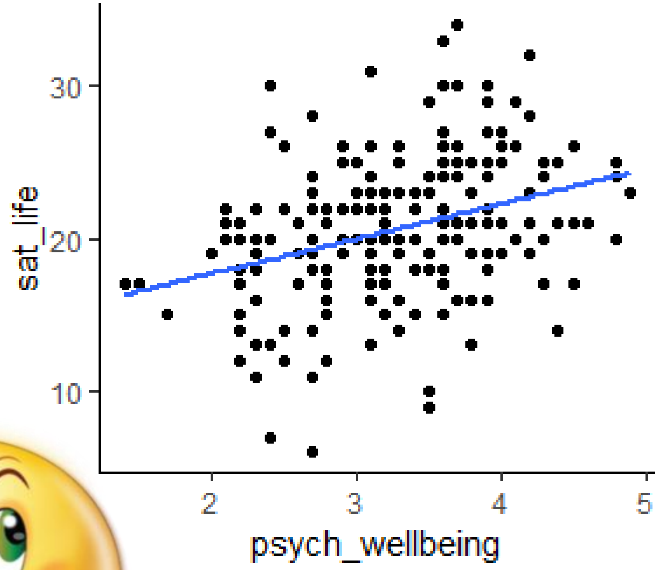
Part 2: Multiple and hierarchical regression

1. Run multiple regression with continuous predictors
2. Run a hierarchical regression
3. Graph significant continuous predictors

We will run <u>two different analyses</u> using continuous variables:
Can we predict satisfaction with life (the outcome variable)…

1. Multiple regression: Using the three wellbeing variables and negative life experiences (continuous predictor variables)?
2. Hierarchical regression: Using the same predictor variables, after controlling for years of education (control variable)?

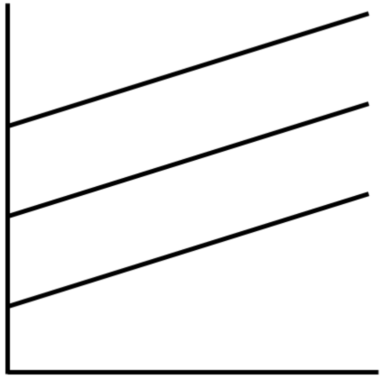
From correlation to regression…



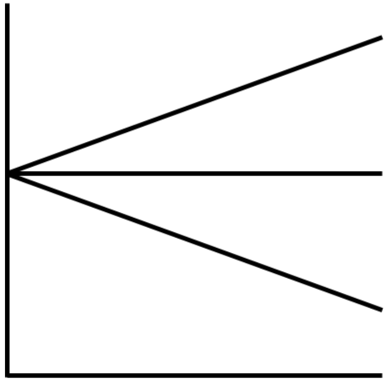
SWL and psychological wellbeing: r (198) = .34, p < .001

- A significant positive relationship

What two pieces of information do I need to describe the line of best fit??



Slope: same
Intercept: different



Slope: different
Intercept: same

- Intercept (β_0)
  - Where the line meets the Y axis
- Slope (β_1) ← Key statistic for interpretation
  - Slope of the regression line
  - For a one unit increase in the predictor variable, what change would you expect to see in the outcome variable?
  - Positive means increasing scores
  - Negative means decreasing scores

Running a multiple regression in R

```
model <- lm(sat_life ~ psych_wellbeing + physical_wellbeing + relationship_wellbeing +
       neg_life_experiences, data = mydata)
summary(model)
```

Build the regression model based on the outcome and predictor variables

- `sat_life ~` (make sure you follow the outcome with a ~)
- `psych_wellbeing + physical_wellbeing + etc.` (have + between each)

Then run the code to show the output in the Console window

- `summary(model)`

Understanding a multiple regression model in two steps…

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · C:/Users/uvjt286/Royal Holloway Dropbox/Victoria Bourne/!Teaching/PS2010/Lecture_R_resources/

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             12.5832     2.2210   5.666 5.20e-08 ***
psych_wellbeing          1.7356     0.4372   3.969 0.000101 ***
physical_wellbeing       0.6581     0.4237   1.553 0.121970    
relationship_wellbeing   0.7933     0.3868   2.051 0.041626 *  
neg_life_experiences    -0.7062     0.1590  -4.442 1.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.227 on 195 degrees of freedom
Multiple R-squared:  0.2311,	Adjusted R-squared:  0.2153 
F-statistic: 14.65 on 4 and 195 DF,  p-value: 1.76e-10
```

Individual predictors: Is each individual predictor significant?

Overall model: Are all the predictors together significant?

Interpreting the overall model (all predictors)

```
Residual standard error: 4.227 on 195 degrees of freedom
Multiple R-squared:  0.2311,    Adjusted R-squared:  0.2153
F-statistic: 14.65 on 4 and 195 DF,  p-value: 1.76e-10
```

Multiple R^2 and Adjusted R^2

How much variance in the outcome variable can the predictors explain?

Report the *adjusted* R^2

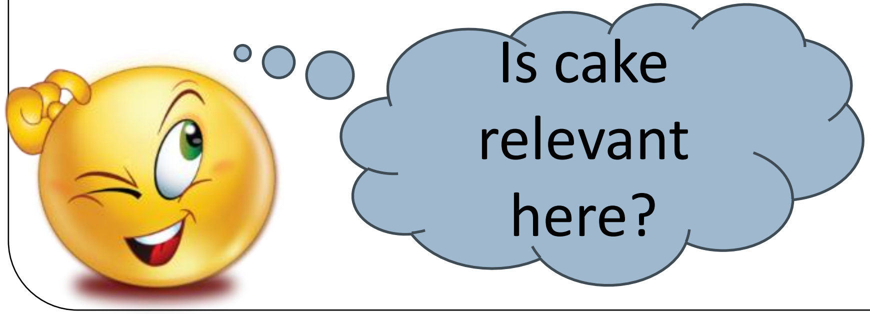
.2153 * 100 = 21.53
21.5% of variance is explained

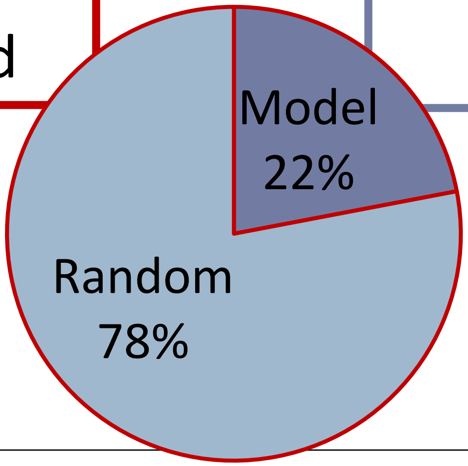
ANOVA

Is the amount of explained variance in the outcome variable significant?

Report in usual APA format

$F\ (4, 195) = 14.65, p < .001$





The overall model, with *all predictors*, is significant ($F\ (4, 195) = 14.65, p < .001$), explaining 21.5% of the variance in SWL

Interpreting the individual predictors

How do I interpret this? What does it mean?

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               12.5832     2.2210   5.666 5.20e-08 ***
psych_wellbeing            1.7356     0.4372   3.969 0.000101 ***
physical_wellbeing         0.6581     0.4237   1.553 0.121970    
relationship_wellbeing     0.7933     0.3868   2.051 0.041626 *  
neg_life_experiences      -0.7062     0.1590  -4.442 1.49e-05 ***
```

Report three statistics for each predictor

β value (slope)

t value (statistic)

p value (sig.)

- Psychological wellbeing: β = 1.74, $t = 3.97$, $p < .001$: Significant predictor (positive β value)
- Physical wellbeing: β = 0.66, $t = 1.55$, $p = .122$: Not significant
- Relationship wellbeing: β = 0.79, $t = 2.05$, $p = .042$: Significant predictor (positive β value)
- Negative life events: β = -0.71, $t = -4.44$, $p < .001$: Significant predictor (negative β value)

Interpreting the individual predictors

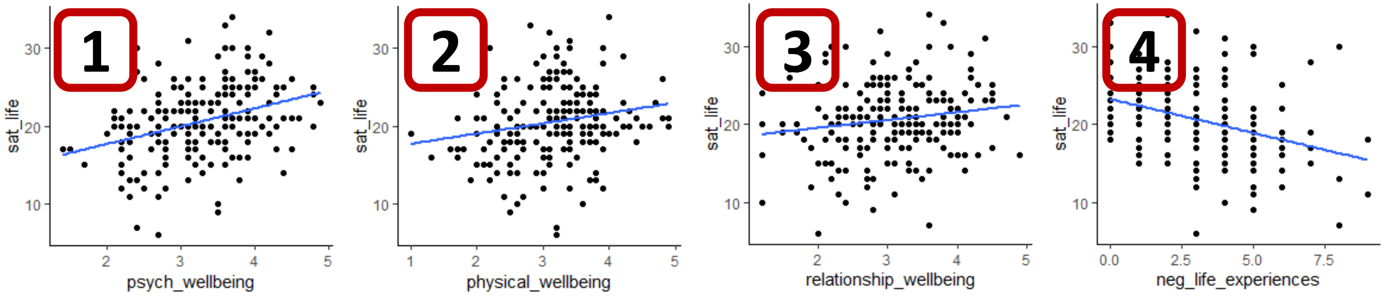
β value is the slope for that predictor

How much does the outcome variable change for a one point increase in this predictor?

+ive or –ive

β values tell you about the *direction*!

1. Psychological wellbeing: β = 1.74, t = 3.97, p < .001
   - A one point increase in WB predicts a **1.74 increase** in SWL
2. Physical wellbeing: β = 0.66, t = 1.55, p = .122
   - Physical WB is not a significant predictor (no interpretation)
3. Relationship wellbeing: β = 0.79, t = 2.05, p = .042
   - A one point increase in WB predicts a **0.79 increase** in SWL
4. Negative life events: β = -0.71, t = -4.44, p < .001
   - A one point increase in NLE predicting a **0.71 decrease** in SWL



Steps to running a hierarchical regression…

1

- Does the control variable explain a significant amount of variance in the model?

2

- Does the final model (control + predictors) explain a significant amount of variance in the model?

3

- Does the final model explain significantly more variance than the control model alone?

4

- Is each individual predictor significant?
- Graph any significant predictors.

1

- Does the control variable explain a significant amount of variance in the model?

```
model1 <- lm(sat_life ~ years_edu, data = mydata)
summary(model1)
```

- Build `model1` – we will use this model name in later code
- Outcome variable is `sat_life ~`
- Control variable is `years_edu,`
- Run the summary code to show the output

Years of education explains a significant amount of the variance in satisfaction with life. NEXT, do the predictor variables explain ***more*** variance?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.3447     2.2528   5.480 1.29e-07 ***
years_edu     0.4991     0.1356   3.681 0.000299 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.628 on 198 degrees of freedom
Multiple R-squared:  0.06406,	Adjusted R-squared:  0.05933 
F-statistic: 13.55 on 1 and 198 DF,  p-value: 0.0002992
```

Individual control variable:

The control variable shows a positive predictive relationship ($\beta = 0.50$, $t = 3.68$, $p < .001$)

Overall control model:

The control model is significant (F (1, 198) = 13.55, $p < .001$), explaining 5.9% of the variance in SWL

2

- Does the final model (control + predictors) explain a significant amount of variance in the model?

```
model2 <- lm(sat_life ~ years_edu + psych_wellbeing + physical_wellbeing +
 relationship_wellbeing + neg_life_experiences, data = mydata)
summary(model2)
```

- Build `model2` – we will use this model name in later code
  - Outcome variable is `sat_life ~`
  - Control and predictors: `years_edu + psych_wellbeing + etc…`
- Run the summary code to show the output: `summary(model2)`

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.2848     2.8672   2.889 0.004298 **
years_edu                 0.2938     0.1260   2.332 0.020735 *
psych_wellbeing           1.6319     0.4346   3.755 0.000229 ***
physical_wellbeing        0.6148     0.4193   1.466 0.144249
relationship_wellbeing    0.7176     0.3839   1.869 0.063109 .
neg_life_experiences     -0.6500     0.1590  -4.087  6.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.18 on 194 degrees of freedom
Multiple R-squared:  0.252,	Adjusted R-squared:  0.2328
F-statistic: 13.07 on 5 and 194 DF,  p-value: 5.663e-11
```

Individual predictor variables:

- We will come back to these in Step 4…

Overall final model:

The final model, including both control and predictor variables, is significant (F (5, 194) = 13.07, p < .001), explaining 23.3% of the variance in SWL

3

- Does the final model explain significantly more variance than the control model alone?

```
r2_control <- summary(model1)$adj.r.squared
r2_full <- summary(model2)$adj.r.squared
r2_change <- r2_full - r2_control
print(r2_change)
```

- `r2_control`: recall `adj.r.squared` for `model1` (control only)
- `r2_full`: recall `adj.r.squared` for `model2` (final model, all variables)
- `r2_change`: calculate the difference `r2_full` minus `r2_control`
- `print`: show the adjusted `r2_change` in the console

```
> r2_change <- r2_full - r2_control
>
> print(r2_change)   # Print the Adj Rsq change
[1] 0.173427
>
```

How much does the variance explained change?

- Adjusted R^2 change = 0.173 or 17.3% increase
- Adjusted R^2 change = 23.3% - 5.9% (from previous slides)

3

- Does the final model explain significantly more variance than the control model alone?

```
anova(model1,model2)
```

- anova: Use an ANOVA to statistically compare the models
  - Compare model1 (only the control variable)
  - With model2 (all variables: control and predictors)

```
Model 1: sat_life ~ years_edu
Model 2: sat_life ~ years_edu + psych_wellbeing + physical_wellbeing +
    relationship_wellbeing + neg_life_experiences
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    198 4241.2
2    194 3389.4  4    851.82 12.189 7.357e-09 ***
---
```

Is the change in Adj R^2 significant?

Yes… (*F* (4, 194) = 12.19, *p* < .001)

After controlling for the variance explained by the control variable of years of education, the predictor variables explain a further 17.3% of the variance in satisfaction with life, which is a significant increase in the variance explained (*F* (4, 194) = 12.19, *p* < .001)

4

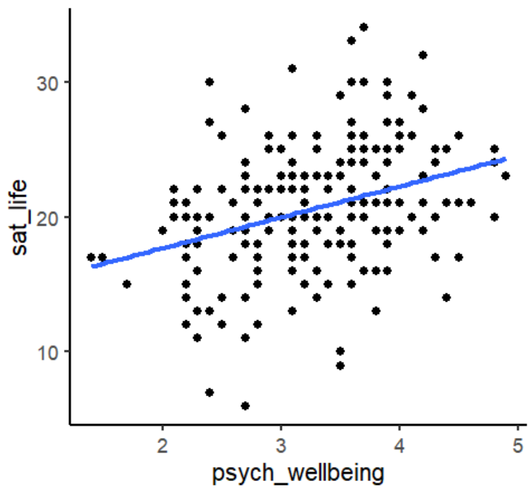
- Is each individual predictor significant?
- Graph any significant predictors.

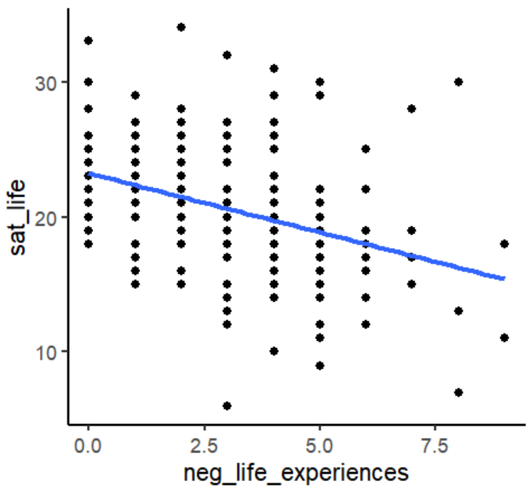
- Go back to the output we created for `model2`

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              8.2848     2.8672   2.889 0.004298 ** 
years_edu                0.2938     0.1260   2.332 0.020735 *  
psych_wellbeing          1.6319     0.4346   3.755 0.000229 ***
physical_wellbeing       0.6148     0.4193   1.466 0.144249    
relationship_wellbeing   0.7176     0.3839   1.869 0.063109 .  
neg_life_experiences    -0.6500     0.1590  -4.087  6.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Individual predictors, after controlling for years of edu:

- Psych WB: ($\beta = 1.63$, $t = 3.76$, $p < .001$)
  - Significant predictor of SWL (positive)
- Physical WB: ($\beta = 0.62$, $t = 1.47$, $p = .144$)
  - Not significant
- Relationship WB: ($\beta = 0.72$, $t = 1.87$, $p = .063$)
  - Not significant
- Negative life events: ($\beta = -0.65$, $t = -4.09$, $p < .001$)
  - Significant predictor of SWL (negative)





Use code from earlier to plot significant predictors

Interpreting a hierarchical regression…

1

- Does the control variable explain a significant amount of variance in the model?

The control model is significant (F (1, 198) = 13.55, p < .001), explaining 5.9% of the variance in SWL and showing a positive predictive relationship (β = 0.50, t = 3.68, p < .001)

2

- Does the final model (control + predictors) explain a significant amount of variance in the model?

The final model, is significant (F (5, 194) = 13.07, p < .001), explaining 23.3% of the variance in SWL

3

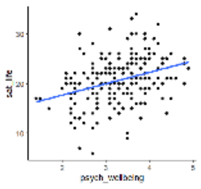
- Does the final model explain significantly more variance than the control model alone?

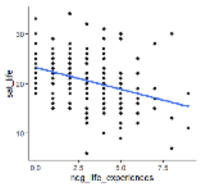
After controlling for the variance explained by Years of Edu, the predictors explain a further 17.3% of the variance in SWL, which is a significant increase (F (4, 194) = 12.19, p < .001)

4

- Is each individual predictor significant? Graph any significant predictors.

SWL is significantly predicted by psych. wellbeing (+ive) and neg. life events (-ive). NOTE: Give full stats for all predictors (inc. NS)

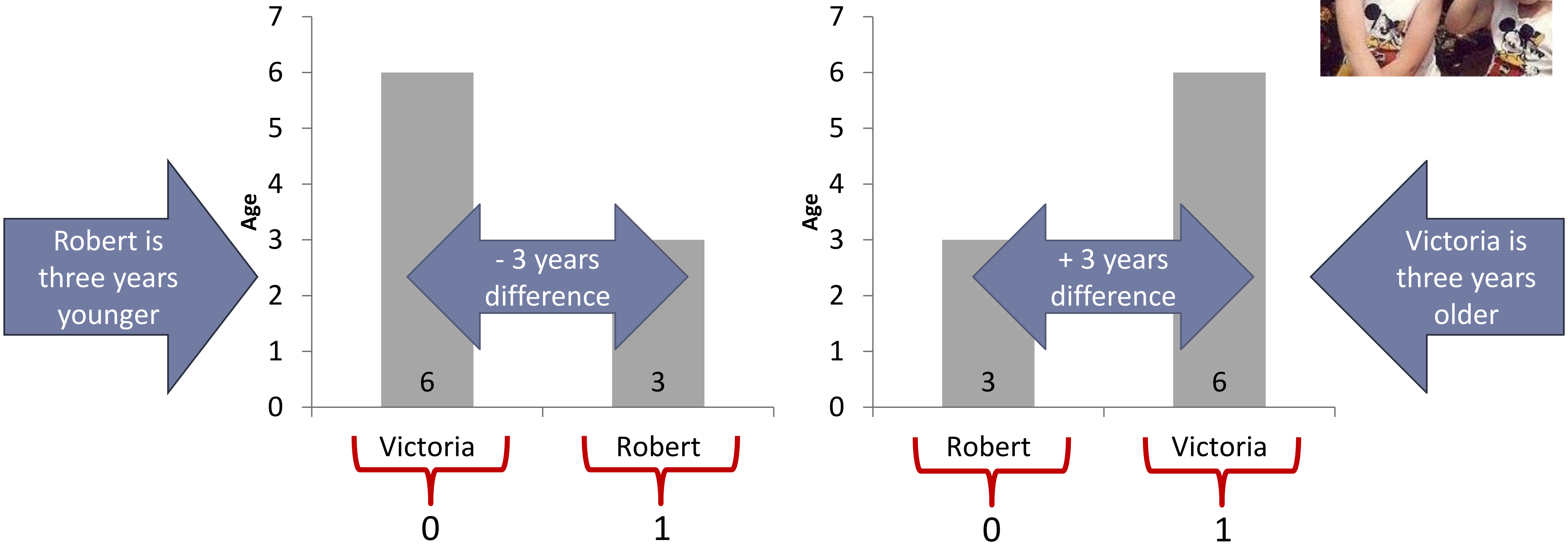




Part 3: Binary and interactive predictors

1. Understand binary predictors and the 0 and 1 coding
2. Interpreting +ive and –ive B values for binary predictors
3. Understand what an interactive predictor is
4. Interpreting significant interactive predictors

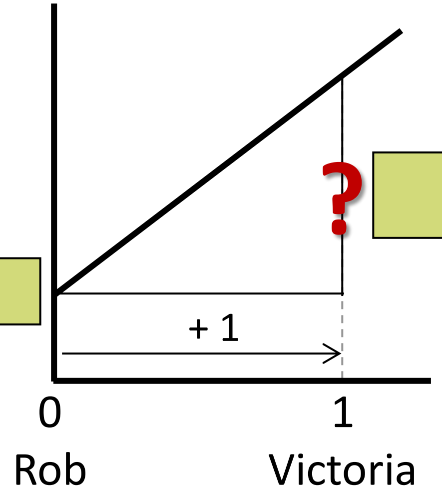
How can we analyse binary variables?



- Difference between two groups is always the same, only the +ive or –ive changes
- Regression tells you if the difference between groups is significant predictor of the outcome
- Direction of the relationship depends on order of coding

Binary predictors and β values

Binary predictors must always be coded as 0 and 1



Intercept (β constant)

Outcome value when the predictor = 0

Slope (β predictor)

Change in outcome for one unit change in predictor (so = 1)

Occupational status:	0 = not in employment	1 = employed
Relationship status:	0 = single	1 = in a relationship
Location of home:	0 = rural	1 = urban

Running a multiple regression in R

Exactly the same as before, but with the extra predictors

```
model <- lm(sat_life ~ psych_wellbeing + physical_wellbeing + relationship_wellbeing +
      neg_life_experiences + occ_status + relationship_status + home_location, data = mydata)
summary(model)
```

Build the regression model based on the outcome and predictor variables

- `sat_life ~` (make sure you follow the outcome with a ~)
- `psych_wellbeing + physical_wellbeing + etc.` (have + between each)
- `summary(model)` show the output in the console

Understanding a multiple regression model in two steps…

- **Overall model:** Are all the predictors together significant?
- **Individual predictors:** Is each individual predictor significant?

Interpreting the overall model (all predictors)

```
Residual standard error: 4.155 on 192 degrees of freedom
Multiple R-squared:  0.2687,    Adjusted R-squared:  0.242
F-statistic: 10.08 on 7 and 192 DF,  p-value: 1.028e-10
```

Multiple R^2 and Adjusted R^2

How much variance in the outcome variable can the predictors explain?

Report the *adjusted R^2*

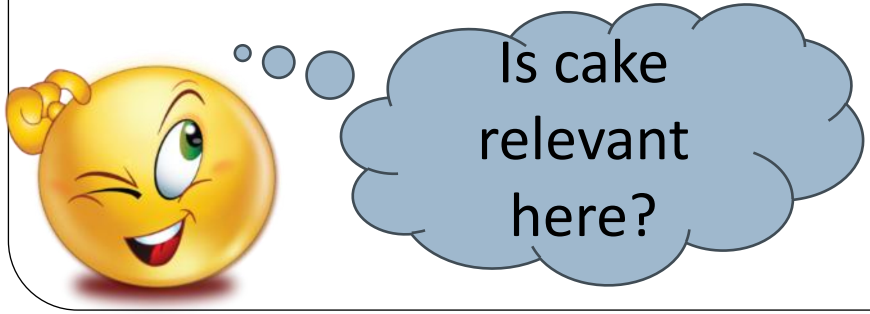
.242 * 100 = 24.2
24.2% of variance is explained

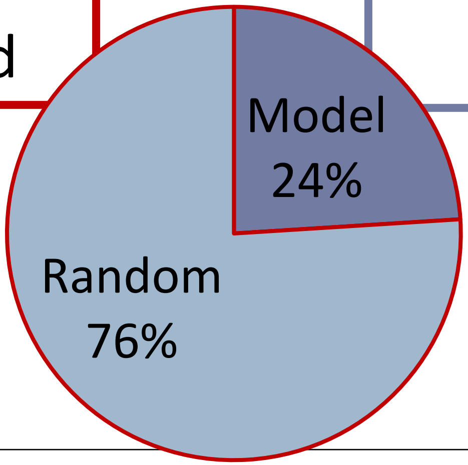
ANOVA

Is the amount of explained variance in the outcome variable significant?

Report in usual APA format

F (7, 192) = 10.08, p < .001





The overall model, with *all predictors*, is significant (F (7, 192) = 10.08, p < .001), explaining 24.2% of the variance in SWL

Interpreting the individual predictors

How do I interpret this? What does it mean?

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             13.4548     2.3414   5.746 3.53e-08 ***
psych_wellbeing          1.6105     0.4411   3.651 0.000336 ***
physical_wellbeing       0.6392     0.4206   1.520 0.130248    
relationship_wellbeing   0.7336     0.3816   1.922 0.056028 .  
neg_life_experiences    -0.7189     0.1590  -4.522 1.07e-05 ***
occ_status1              1.2317     0.6126   2.010 0.045781 *  
relationship_status1    -0.2508     0.6044  -0.415 0.678618    
home_location1          -1.2964     0.6035  -2.148 0.032954 *  
---
```

Report three statistics for each predictor

β value (slope)

t value (statistic)

p value (sig.)

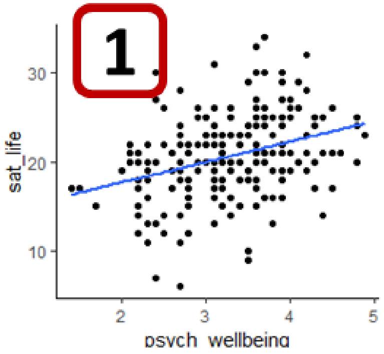
Interpreting the individual continuous predictors

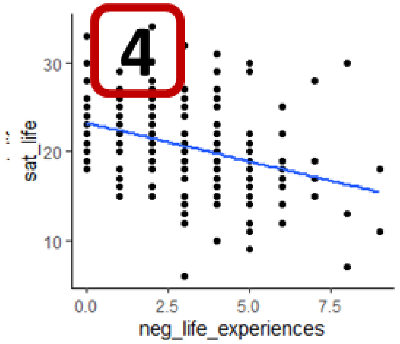
β value is the slope for that predictor

How much does the outcome variable change for a one point increase in this predictor?

+ive or –ive

β values tell you about the *direction*!






```
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              13.4548     2.3414   5.746 3.53e-08 ***
psych_wellbeing           1.6105     0.4411   3.651 0.000336 ***
physical_wellbeing        0.6392     0.4206   1.520 0.130248    
relationship_wellbeing    0.7336     0.3816   1.922 0.056028 .  
neg_life_experiences     -0.7189     0.1590  -4.522 1.07e-05 ***
occ_status1               1.2317     0.6126   2.010 0.045781 *  
relationship_status1     -0.2508     0.6044  -0.415 0.678618    
home_location1           -1.2964     0.6035  -2.148 0.032954 *  
```

1. Psychological wellbeing: β = 1.61, $t = 3.65$, $p < .001$
   - A one point increase in WB predicts a **1.61 increase** in SWL
2. Physical wellbeing: β = 0.64, $t = 1.52$, $p = .130$
   - Physical WB is not a significant predictor (no interpretation)
3. Relationship wellbeing: β = 0.73, $t = 1.92$, $p = .056$
   - Relationship WB is not a significant predictor (no interpretation)
4. Negative life events: β = -0.72, $t = -4.52$, $p < .001$
   - A one point increase in NLE predicting a **0.72 decrease** in SWL

Interpreting the individual binary predictors

β value is the slope for that predictor

How much does the outcome variable change for a one point increase in this predictor?

+ive or –ive

β values tell you about the *direction*!

Occupational status:	0 = not in employ.	1 = employed
Relationship status:	0 = single	1 = in a relationship
Location of home:	0 = rural	1 = urban

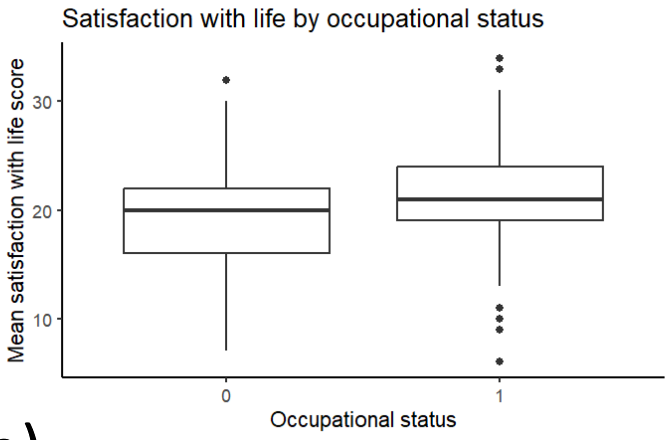
```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              13.4548     2.3414   5.746 3.53e-08 ***
psych_wellbeing           1.6105     0.4411   3.651 0.000336 ***
physical_wellbeing        0.6392     0.4206   1.520 0.130248
relationship_wellbeing    0.7336     0.3816   1.922 0.056028 .
neg_life_experiences     -0.7189     0.1590  -4.522 1.07e-05 ***
occ_status1               1.2317     0.6126   2.010 0.045781 *
relationship_status1     -0.2508     0.6044  -0.415 0.678618
home_location1           -1.2964     0.6035  -2.148 0.032954 *
```

1. Occupational status: β = 1.23, t = 2.01, p = .045
   - A one point increase in occupational status, so being employed, predicts a **1.23 increase** in SWL
2. Relationship status: β = -0.25, t = -0.42, p = .678
   - Relationship status is not a significant predictor (no interpretation)
3. Home location: β = -1.30, t = -2.15, p = .033
   - A one point increase in home location, so living in an urban environment, predicts a **1.30 decrease** in SWL

Graphing significant binary predictors

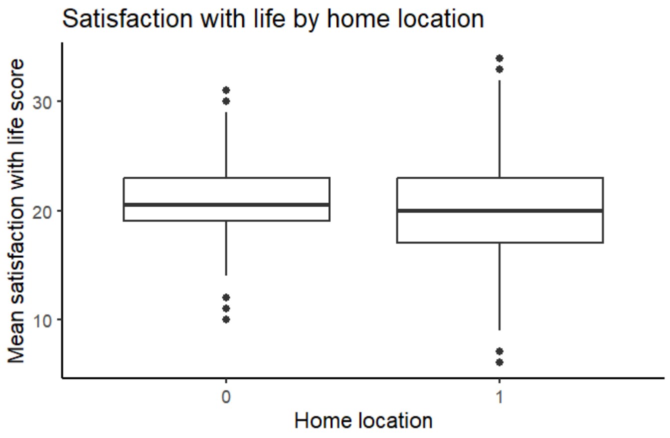
- How to create boxplots…

```
ggplot(mydata, aes(x = occ_status, y = sat_life)) +
  geom_boxplot() +
  labs(title = "Satisfaction with life by occupational status",
       x = "Occupational status",
       y = "Mean satisfaction with life score") +
  theme_classic()
```



- Binary predictor on the X axis (one box for each group)
- Outcome variable on the Y axis

```
ggplot(mydata, aes(x = home_location, y = sat_life)) +
  geom_boxplot() +
  labs(title = "Satisfaction with life by occupational status",
       x = “Home location",
       y = "Mean satisfaction with life score") +
  theme_classic()
```



Descriptive statistics for binary predictors

- If you want to pull out the descriptives for individual groups…

```
descriptives_bygroup <- mydata %>%
  group_by(occ_status) %>%
  summarise(mean_sat_life = mean(sat_life), sd_sat_life = sd(sat_life))
print(descriptives_bygroup)
```

- `descriptives_bygroup <- mydata %>%`
  - Calculate descriptives for individual groups of participants

```
occ_status mean_sat_life sd_sat_life
<fct>              <dbl>       <dbl>
0                   19.4        4.62
1                   21.5        4.71
```

Mean

SD

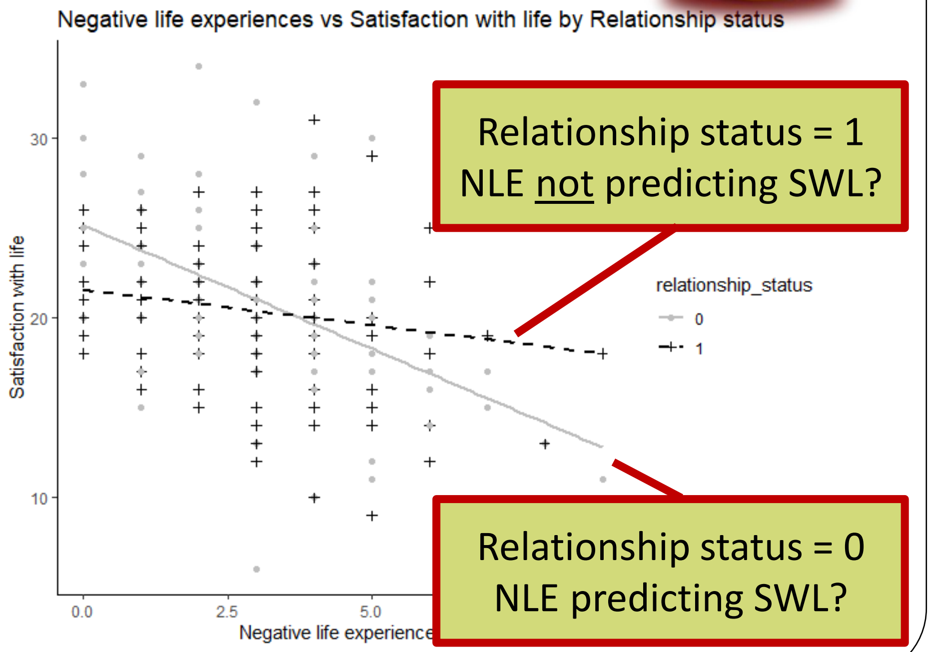
- `group_by(occ_status) %>%`
  - Which categorical/binary variable to group by
- `summarise(mean_sat_life = mean(sat_life), sd_sat_life = sd(sat_life))`
  - The variable to calculate statistics on, and to calculate mean and SD
- Repeat with any other grouping variables you want…

Understanding interactive predictors…

On to a new analysis!!!

- Let's simplify the analysis…
- Outcome variable: Satisfaction with life
- Three predictor variables:
  - Continuous predictor: Negative life experiences (NLE)
  - Binary predictor: Relationship status
  - Interactive predictor: NLE by relationship status

Does this graph look familiar?



Relationship status = 1
NLE <u>not</u> predicting SWL?

Relationship status = 0
NLE predicting SWL?

What do interactive predictors show us?

If we separate out participants by the binary variable, is the predictive relationship between the continuous predictor and the outcome variable different according to groups?

Running the analysis in R

Interactive predictor:
Variable1*Variable2

```
model <- lm(sat_life ~ neg_life_experiences + relationship_status +
                      neg_life_experiences*relationship_status, data = mydata)
summary(model)
```

Build the regression model based on the outcome and predictor variables

- `sat_life ~` (make sure you follow the outcome with a ~)
- `neg_life_experiences + relationship_status + etc.` (have + between each)
  - NOTE: To add an interactive predictor, just add “Variable1*Variable2” into the list of predictors
  - For this analysis: `neg_life_experiences*relationship_status`
- `summary(model)` show the output in the console

Understanding a multiple regression model in two steps…

- **Overall model:** Are all the predictors together significant?
- **Individual predictors:** Is each individual predictor significant?

How do I interpret all this, and what does it mean?

Interpreting the overall model (all predictors)

```
Residual standard error: 4.381 on 196 degrees of freedom
Multiple R-squared:  0.1699,    Adjusted R-squared:  0.1572
F-statistic: 13.37 on 3 and 196 DF,  p-value: 5.646e-08
```

Multiple R^2 and Adjusted R^2

How much variance in the outcome variable can the predictors explain?

Report the *adjusted* R^2

.1572 * 100 = 15.72
15.72% of variance is explained

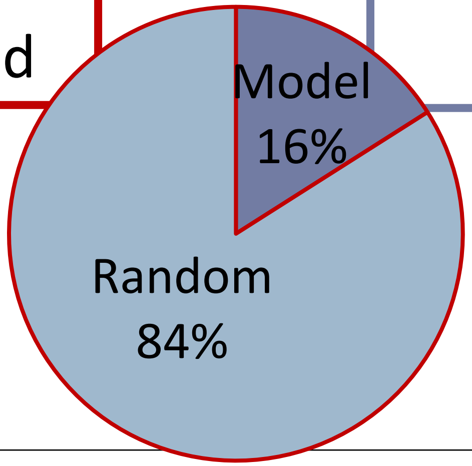
ANOVA

Is the amount of explained variance in the outcome variable significant?

Report in usual APA format

F (3, 196) = 13.37, p < .001

Is cake relevant here?



The overall model, with *all predictors*, is significant (F (3, 196) = 13.37, p < .001), explaining 15.7% of the variance in SWL

Interpreting the individual predictors

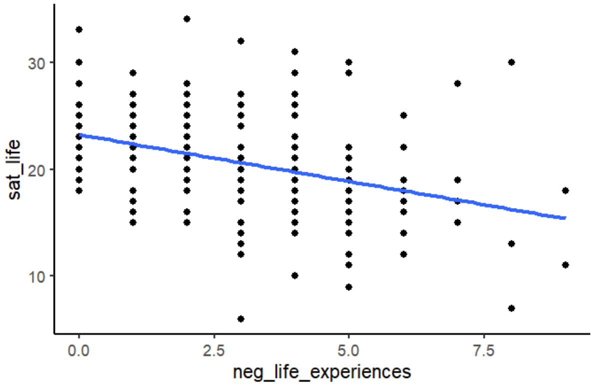
Negative life experiences (continuous predictor)

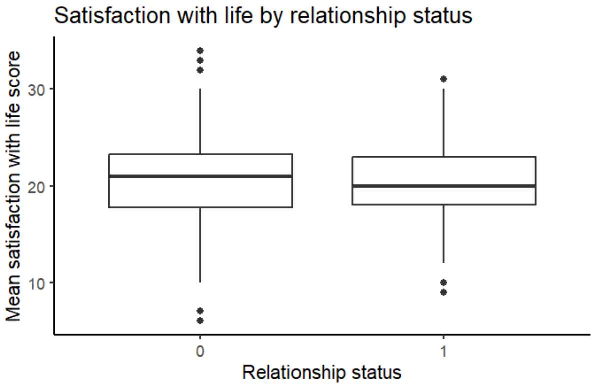
Relationship status (binary predictor)

NLE * relationship status (interactive predictor)

```
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                             25.1320     0.8556  29.374  < 2e-16 ***
neg_life_experiences                    -1.3712     0.2264  -6.057 6.96e-09 ***
relationship_status1                    -3.5785     1.1660  -3.069  0.00245 **
neg_life_experiences:relationship_status1  0.9763     0.3196   3.054  0.00257 **
```

1. Negative life experiences: $\beta = -1.37$, $t = -6.06$, $p < .001$
   - A one point increase in NLE predicts a **<u>1.37 decrease</u>** in SWL
2. Relationship status: $\beta = -3.58$, $t = -3.07$, $p = .003$
   - A one point increase in relationship status, <u>so being in a relationship</u>, predicts a **<u>3.60 decrease</u>** in SWL
3. NLE * relationship status: $\beta = 0.98$, $t = 3.05$, $p = .003$
   - How do we interpret this? Statistically comparing correlations…





Comparing correlations in R (as before!)

If an interactive (binary * continuous) predictor is significant, then we break this down by statistically comparing the correlations – which you already know!

1
- Define which groups you want to look at separately
- Relationship status: 0 = single, 1 = in a relationship

2
- Run the correlations, specifying which group you are looking at
- A separate piece of R code for each group (single or in a relationship)

3
- Compare two correlations: Is one significantly stronger than the other?
- For this, you need the N and *r* value for each of the correlations

4
- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

Comparing correlations: Steps 1 and 2

Go back to Part 1 for a full description of how to run this code and interpret the output

1

- Define which groups you want to look at separately
- Relationship status: 0 = single, 1 = in a relationship

```
single <- mydata[mydata$relationship_status == "0", ]
relationship <- mydata[mydata$relationship_status == "1", ]
```

2

- Run the correlations, specifying which group you are looking at
- A separate piece of R code for each group (single or in a relationship)

```
cor.test(single$sat_life, single$neg_life_experiences,
         method = "pearson")
```

```
cor.test(relationship$sat_life, relationship$neg_life_experiences,
         method = "pearson")
```

Comparing correlations in R: Step Three

Go back to Lecture 15 for a full description of how to run this code and interpret the output

3

- Compare two correlations: Is one significantly stronger than the other?
- For this, you need the N and *r* value for each of the correlations

	***r* value** (remember the –ive!)	**N** (df + 2 = N)
Correlation for single participants	-0.5218863	92
Correlation for participants in a relationship	-0.1741089	108

```
cocor.indep.groups(r value 1, r value 2, N 1, N 2)
```

```
cocor.indep.groups(-0.5218863, -0.1741089, 92, 108)
```

```
fisher1925: Fisher's z (1925)
  z = -2.7972, p-value = 0.0052
  Null hypothesis rejected
```

The two correlations differ significantly (z = -2.80, p = .005)

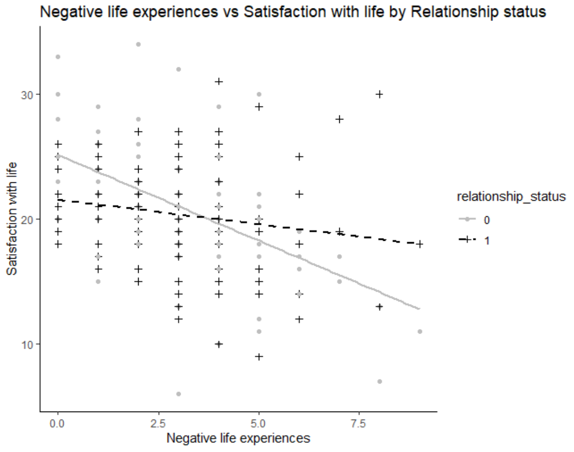
Comparing correlations in R: Step Four

Go back to Lecture 15 for a full description of how to run this code and interpret the output

4

- Plot both correlations on one graph, with a line of best fit for each group
- Use clear formatting to distinguish the two groups

```
plot_cc <- ggplot(mydata, aes(x = neg_life_experiences, y = sat_life, colour = relationship_status)) +
  geom_point(aes(shape = relationship_status)) +
  geom_smooth(aes(linetype = relationship_status), method = "lm", se = FALSE) +
  labs(title = "Negative life experiences vs Satisfaction with life by Relationship status",
       x = "Negative life experiences",
       y = "Satisfaction with life") +
  theme_classic() +
  scale_color_manual(values = c("0" = "grey", "1" = "black ")) +
  scale_linetype_manual(values = c("0" = "solid", "1" = "dashed")) +
  scale_shape_manual(values = c("0" = 16, "1" = 3))
```



Plot the continuous predictor variable should be plotted on the x axis

Plot the outcome variable should be plotted on the y axis

Plot the binary predictor variable as defining the separate groups

Part 4: The assumptions of regression

To understand and evaluate the assumptions of multiple regression:

1. Multicollinearity
2. Distribution of residuals
3. Homoscedasticity
4. Outlier effects

Analysis now (same as earlier):

Outcome: Satisfaction with life
Predictor variables:

- Three wellbeing variables (cont.)
- Negative life experiences (cont.)
- Occupational status (binary)
- Relationship status (binary)
- Home location (binary)

Process for evaluating assumptions

1. Run the multiple regression (don't interpret it yet)
2. Evaluate any evidence of multicollinearity
3. Evaluate the distribution of the residuals
4. Evaluate the homoscedasticity assumption
5. Evaluate and identify participants who are outliers
6. Interpret the multiple regression you ran earlier

Different R code for each assumption.

Let's go through each one in turn…

Running the multiple regression in R

Same analysis as we ran earlier

```
model <- lm(sat_life ~ psych_wellbeing + physical_wellbeing + relationship_wellbeing +
        neg_life_experiences + occ_status + relationship_status + home_location, data = mydata)
summary(model)
```

Build the regression model based on the outcome and predictor variables

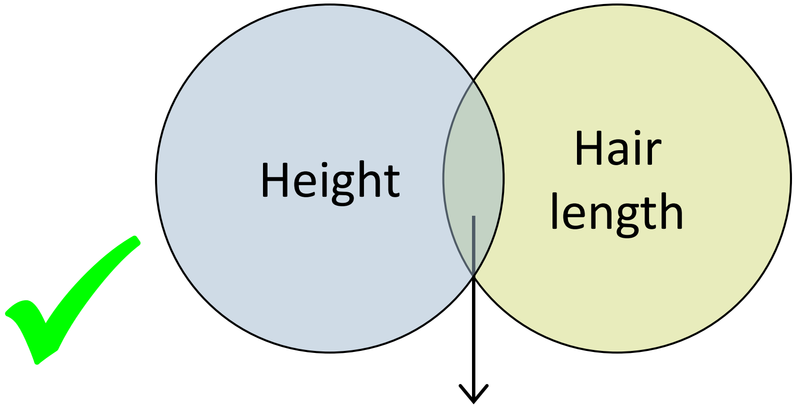
- `sat_life ~` (make sure you follow the outcome with a ~)
- `psych_wellbeing + physical_wellbeing + etc.` (have + between each)
- `summary(model)` show the output in the console

Why do I need to run the analysis now, if we won't look at the results yet?

For parts of the assumptions, R needs the **`model`** to be defined – so we run it now, but look at it later

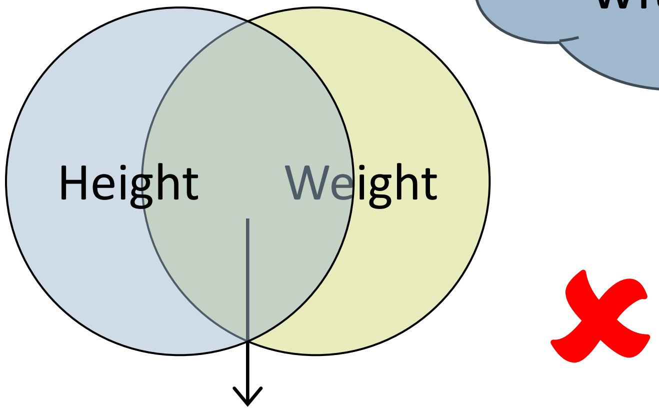
Multiple regression: multicollinearity

- Multicollinearity: when predictor variables are correlated with each other – this is bad!!!
- Say we want to predict a persons age…



Little shared variance

No multicollinearity



Lots of shared variance

Multicollinearity

Multicollinearity: Zero order correlations

- Zero order correlations between <u>predictor</u> variables must be within +/- .9

-1.0	-.9	-.8	-.7	-.6	-.5	-.4	-.3	-.2	-.1	0	+.1	+.2	+.3	+.4	+.5	+.6	+.7	+.8	+.9	+1.0

```
mydata %>%
  dplyr::select(psych_wellbeing, physical_wellbeing, relationship_wellbeing,
                               neg_life_experiences) %>%
  correlation(p_adjust = "none")
```

- Add all predictor variables
- Run a correlation, with no adjustment

All zero order correlations within +/- .9 ✓

```
Parameter1             |             Parameter2 |        r |         95% CI | t(198) |         p
------------------------------------------------------------------------------------------------
psych_wellbeing        |     physical_wellbeing |     0.13 | [ 0.00,  0.27] |   1.90 | 0.059
psych_wellbeing        | relationship_wellbeing |     0.08 | [-0.06,  0.22] |   1.19 | 0.237
psych_wellbeing        |   neg_life_experiences |    -0.20 | [-0.33, -0.06] |  -2.86 | 0.005**
physical_wellbeing     | relationship_wellbeing |     0.14 | [ 0.00,  0.27] |   1.93 | 0.055
physical_wellbeing     |   neg_life_experiences |    -0.17 | [-0.30, -0.03] |  -2.44 | 0.016*
relationship_wellbeing |   neg_life_experiences | 9.63e-04 | [-0.14,  0.14] |   0.01 | 0.989
```

Multicollinearity: Variance inflation factor (VIF)

- How much might the variance explained in a model be artificially inflated by multicollinearity between predictors? High scores are bad!

Values less than 5 are good	5-10: likely multicollinearity	10 + serious multicollinearity

```
vif_values <- vif(model)
print(vif_values)
```

- Calculate the `vif_values`
- Based on the `model` we ran previously
- Then `print` the calculated vif_values into the console

All VIF values are in the acceptable range, with the largest being 1.115 (psychological wellbeing) ✓

```
     psych_wellbeing      physical_wellbeing relationship_wellbeing   neg_life_experiences            occ_status
            1.114775                1.080804               1.032343               1.103654              1.071516
relationship_status           home_location
            1.051545                1.034393
```

Multicollinearity: Tolerance

- Tolerance is calculated as: $1 - R^2$
- If R^2 is the amount of variance explained, tolerance is the amount unexplained
- Low scores are bad! Below 0.2 indicates an issue with multicollinearity

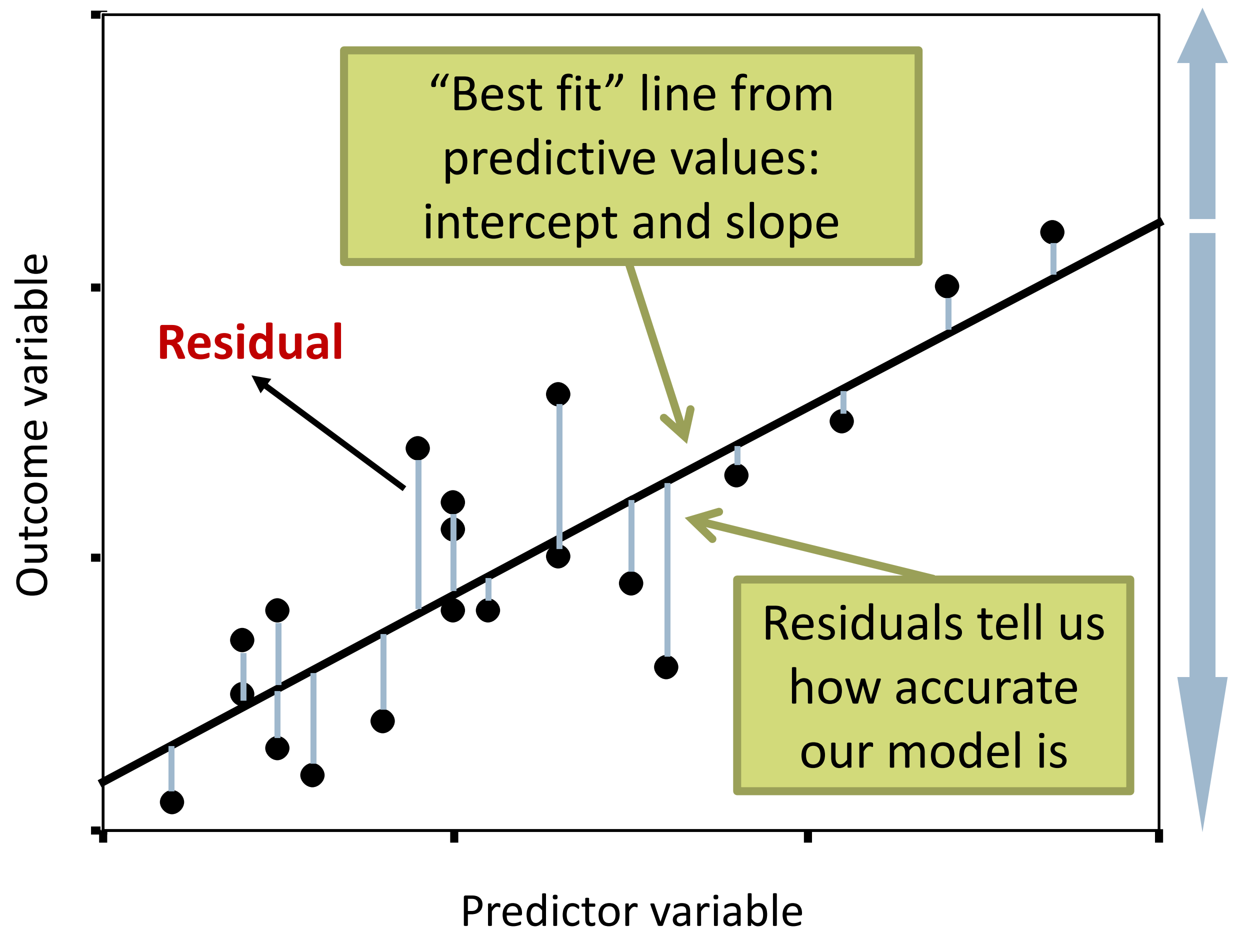
```
tolerance_value <- 1 - summary(model)$r.squared
print(tolerance_value)
```

- Calculate the `tolerance value`
- Subtract the R^2 `calculated in the model` earlier `from 1`
- Then `print` this in the console

```
> tolerance_value <- 1 - summary(model)$r.squared
> print(tolerance_value)
[1] 0.7313278
> |
```

Tolerance is above 0.2, so indicates no problem with multicollinearity

Line of “best” fit… but not perfect! Residuals

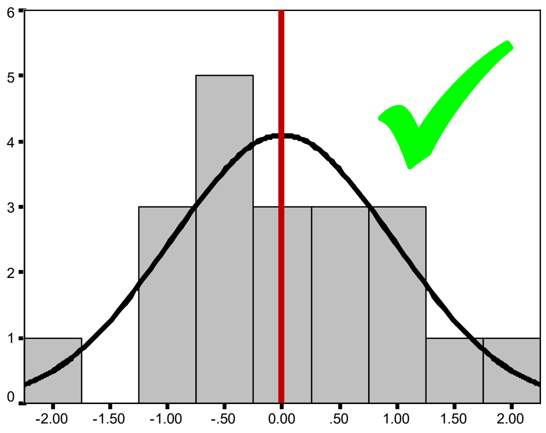


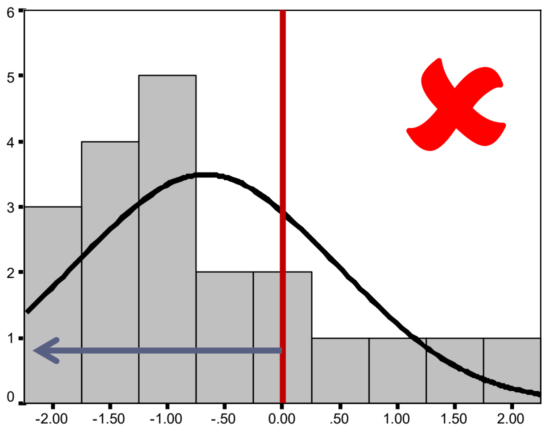
- Positive residual
- Higher score than is predicted by equation
- Underestimate

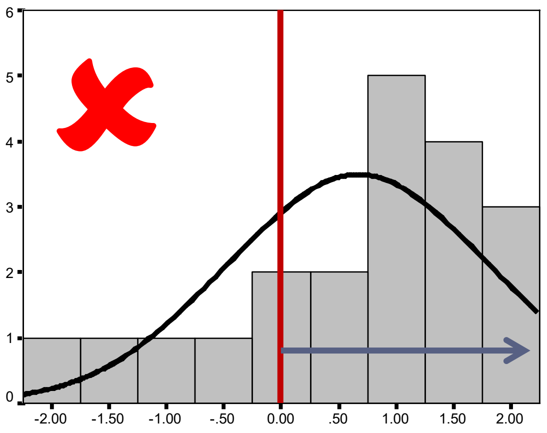
- Negative residual
- Lower score than is predicted by equation
- Overestimate

Assumptions of residuals

- Residuals should be normally distributed
- Histogram: Distribution of residuals for each participant







- Regression model shows no bias in predicting participants scores

- Lots of negative residuals
- **<u>Positive</u>** skew (tail)
- Regression model overestimating

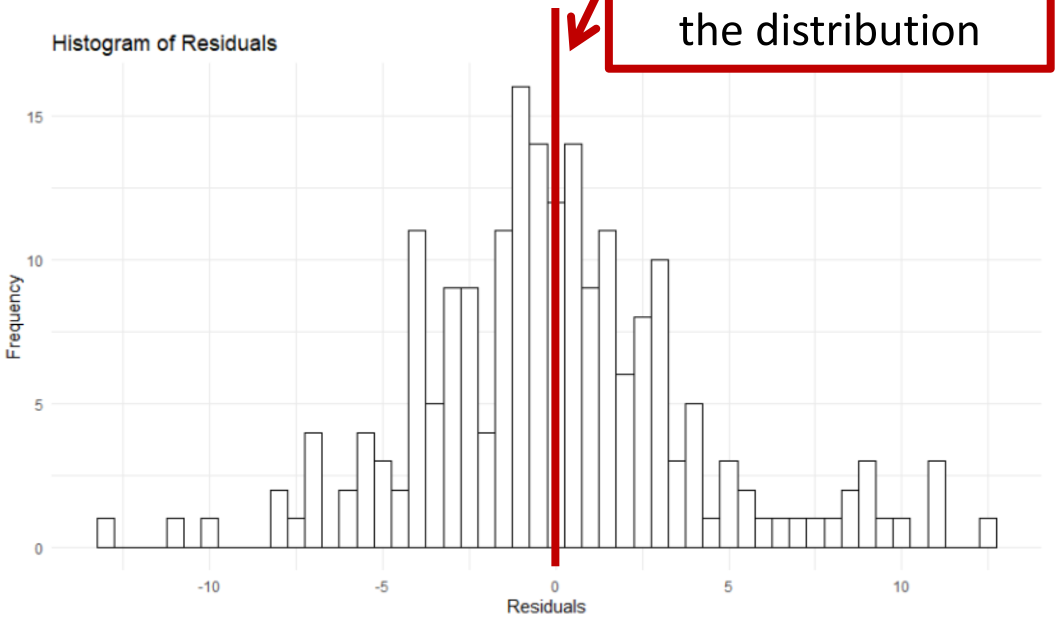
- Lots of positive residuals
- **<u>Negative</u>** skew (tail)
- Regression model underestimating

Distribution of residuals

- Create a histogram with the residual for each participant

```
ggplot(mydata, aes(x = model$residuals)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "white") +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency") +
  theme_minimal()
```

Look at 0 on the x axis as the middle of the distribution

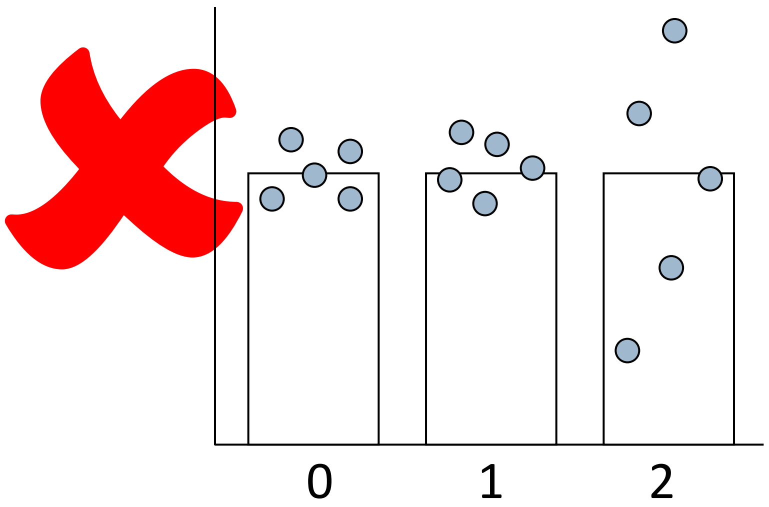


- Plot the model$residuals on the x axis
- Create a geom_histogram
- Add the relevant title

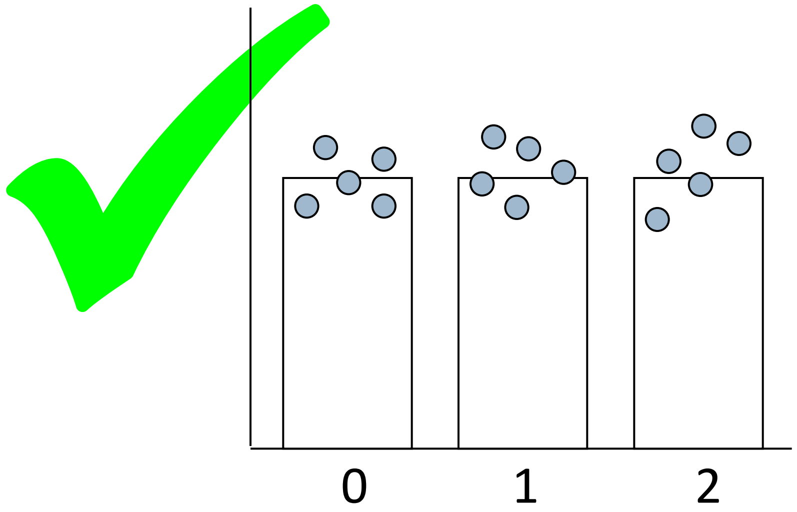
Residuals are roughly normally distributed, so there is no systematic under or over estimation in the model ✓

Assumption of homoscedasticity

- Related to homogeneity of variance in ANOVA



*Hetero*geneity of variance



*Homo*geneity of variance

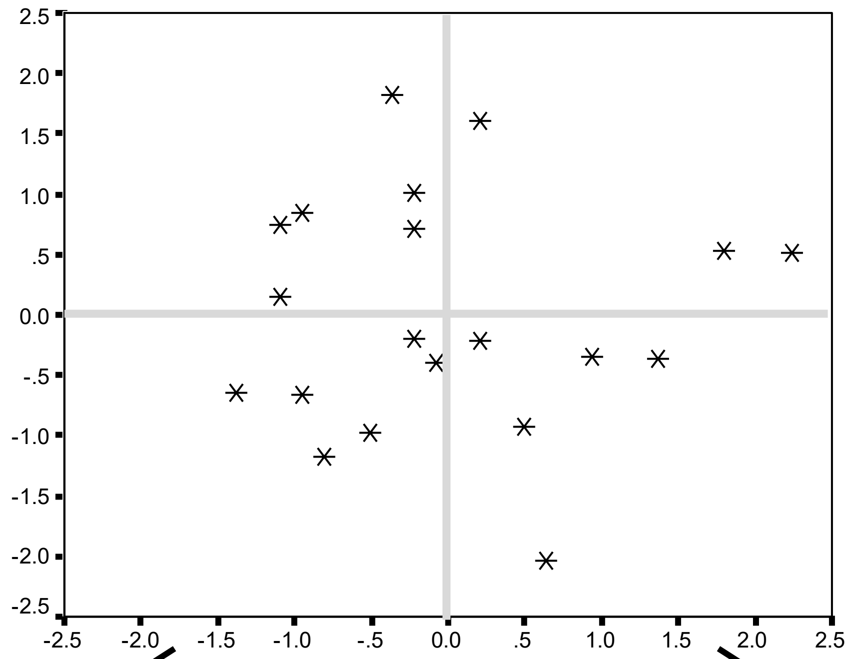
- But, there are no groups in regression…

Assumption of homoscedasticity

Underestimated

Residual (difference between actual and predicted score)

Overestimated



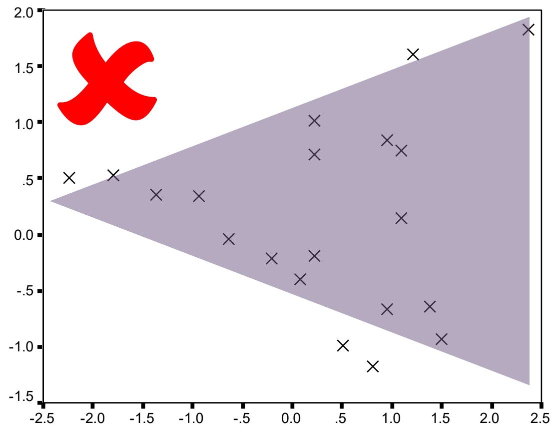
Want points to be randomly distributed across whole graph – no clustering or funnelling: Homoscedasticity ✓

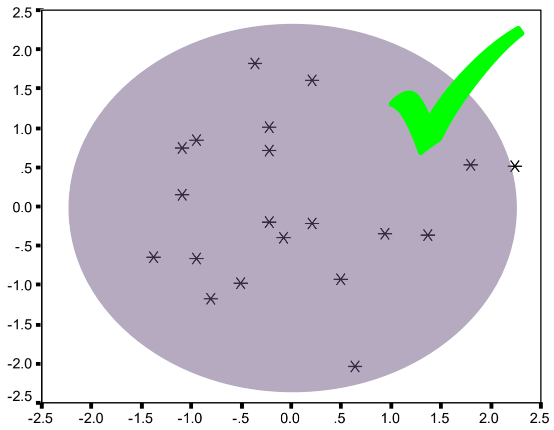
Outcome predicted by the regression equation (fitted value)

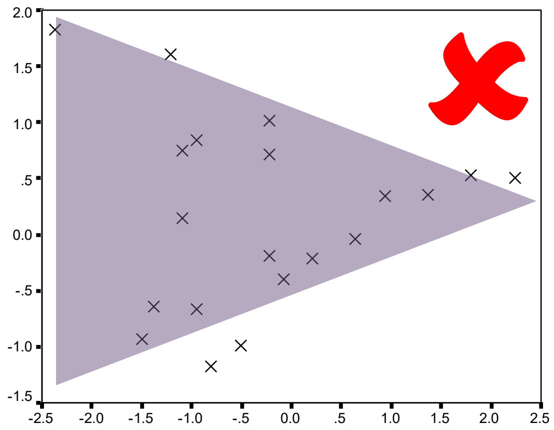
Low predicted values

High predicted values

Assumption of homoscedasticity







Low predicted values – little variance (accurate)

High predicted values – lots of variance

Heteroscedasticity

Variance is similar across the continuum

Homoscedasticity

Low predicted values – lots variance

High predicted values – little variance (accurate)

Heteroscedasticity

Evaluating homoscedasticity

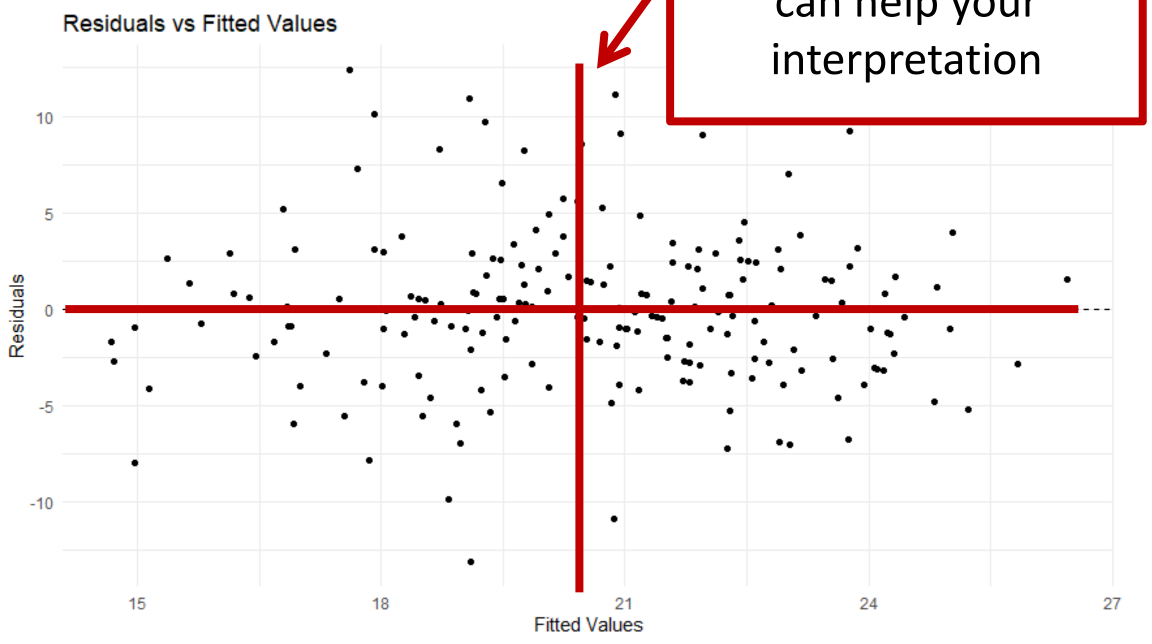
- Create a scatterplot showing predicted (fitted) values against residuals

```
ggplot(mydata, aes(x = model$fitted.values, y = model$residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()
```

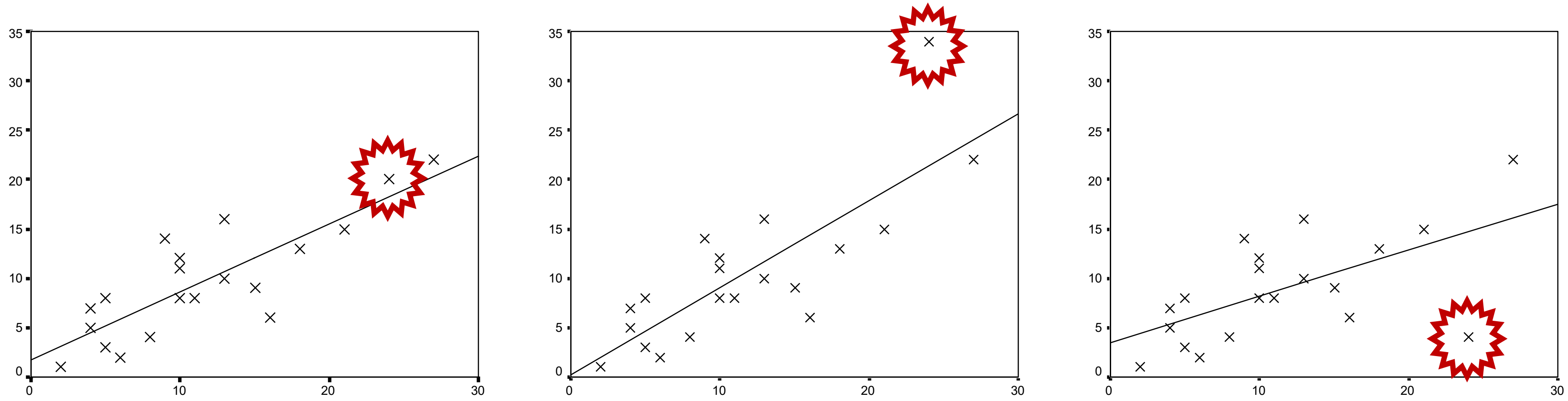
- Fitted (predicted) values on x axis
- Residuals on y axis

The error in the model is evenly distributed across low and high predicted (fitted values), suggesting homoscedasticity in the model ✓

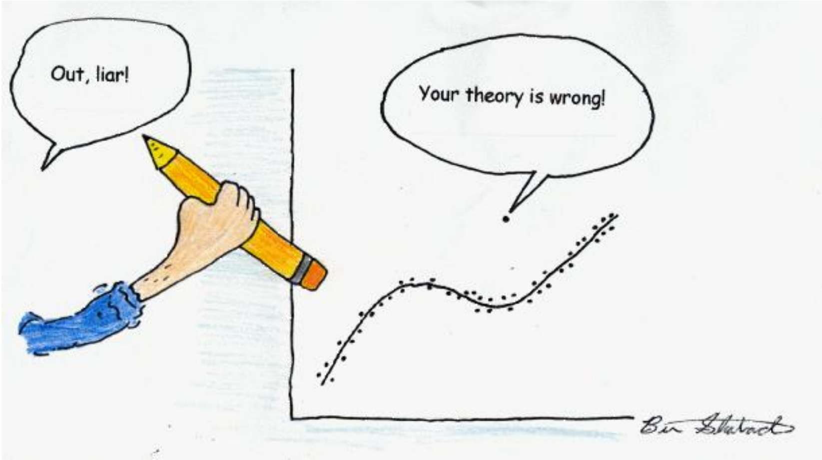
Marking the graph for the four quadrants can help your interpretation



Assessing outlier effects



- Outliers affect the slope and intercept
- What is an outlier?
  - Standardised residual greater than ± 2
- Up to 5% of the sample being outliers is ok!



Identifying outliers

1. Calculate the standardised residual for each participant using the model

```
standardized_residuals <- rstandard(model)
print(standardized_residuals)
```

2. Count (sum) how many participants are "outliers" with std. residuals > 2

```
outliers <- sum(abs(standardized_residuals) > 2)
print(outliers)
```

3. Calculate the percentage of the sample defined as "outliers"

```
percentage_outliers <- (outliers / nrow(mydata)) * 100
print(percentage_outliers)
```

```
> # Determine the number of outliers (absolute value greater than 2)
> outliers <- sum(abs(standardized_residuals) > 2)
> print(outliers)
[1] 15
> # Calculate the percentage of outliers
> percentage_outliers <- (outliers / nrow(mydata)) * 100
> print(percentage_outliers)
[1] 7.5
```

There are 15 participants identified as outliers, making up 7.5% of the sample

Practice dataset…

Practice dataset: What will you be analysing?

The continuous variables:

- Optimism: questionnaire, scores 0 – 10, high scores indicate more optimistic
- Two self-compassion scales: Questionnaire, high = more pos/neg compassion
  - Positive SC (e.g. “I try to be kind towards those things about myself I don’t like”)
  - Negative SC (e.g. “I am hard on myself about my own flaws and weaknesses”)
- Chronological age (in years)
- Reading age – used as a control variable

The binary/categorical variable:

- Extracurricular activities status: 0 = no EC activities, 1 = takes EC activities

Your dataset: WS_data_R_optimism.csv

Practice dataset

Your dataset: WS_data_R_optimism.csv

1. Set and check your working directory
2. Install and load all necessary packages for today:
   - `correlation`
   - `gridextra`
   - `ppcor`
   - `cocor`
   - `car`

   NOTE: You may see errors when you do this – it likely means it is already installed, so just ignore these errors for now
3. Next, define variables as continuous (`numeric`) or binary (`factor`)
   - You can check the variable names using `names(mydata)`
   - All variables should be continuous, other than `extra_curr` - which is binary

ps. there is also code for running descriptives, if you want to try that!

Correlations…

- Is optimism correlated with…
  - Positive self compassion
  - Negative self compassion
  - Age (in years)
- Create scatterplots for the three correlations

- Do the three correlations change if you control for reading age?

- Are the correlations different if you look at children who do and who do not take part in extra curricular activities separately?

Multiple regression…

Run and interpret two analyses…

1. Multiple regression: Is optimism predicted by…
   - Positive self compassion
   - Negative self compassion
   - Age (in years)
   - Create scatterplots for significant predictors

2. Hierarchical regression: After controlling for reading age, is optimism predicted by…
   - Positive self compassion
   - Negative self compassion
   - Age (in years)
   - Create scatterplots for significant predictors

Binary and interactive predictors…

- Is optimism predicted by…
  - Positive self compassion (continuous)
  - Negative self compassion (continuous)
  - Taking part in extra curricular activities (binary)
  - Positive self compassion * extra curricular (interactive predictor)
  - Negative self compassion * extra curricular (interactive predictor)
- Graph and interpret any significant predictors

Evaluating the assumptions...

- Run this analysis...
  - Outcome variable:
    - Optimism
  - Predictor variables:
    - Positive self compassion
    - Negative self compassion
    - Extra curricular activities
- Is this analysis robust and unbiased? →

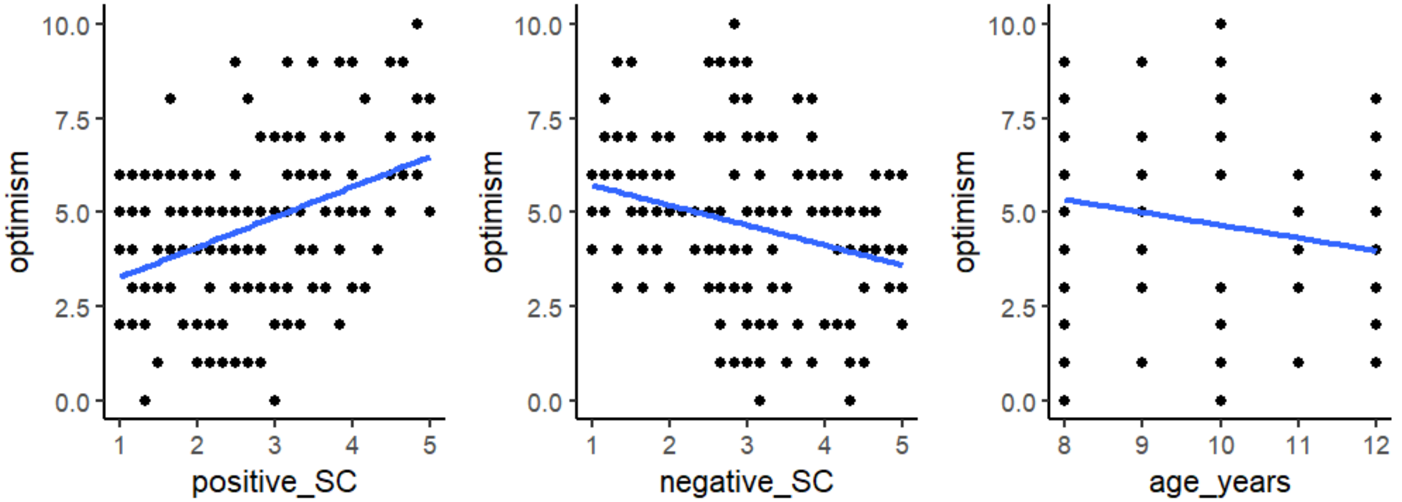
Evaluate the four assumptions:
Multicollinearity
Distribution of residuals
Homoscedasticity
Outliers


IT'S FINISHED.
IT'S DONE.
quickmeme.com

Correlations: Answers in output…

- Is optimism correlated with…
  - Positive self compassion
  - Negative self compassion
  - Age (in years)
- Create scatterplots for the three correlations

```
Parameter1   |  Parameter2 |     r |         95% CI | t(148) |         p
-------------------------------------------------------------------------
optimism     | positive_SC |  0.43 | [ 0.29,  0.55] |   5.73 | < .001***
optimism     | negative_SC | -0.28 | [-0.42, -0.12] |  -3.49 | < .001***
optimism     |   age_years | -0.22 | [-0.37, -0.06] |  -2.78 | 0.006**
positive_SC  | negative_SC | -0.01 | [-0.17,  0.15] |  -0.16 | 0.875
positive_SC  |   age_years | -0.22 | [-0.37, -0.06] |  -2.75 | 0.007**
negative_SC  |   age_years | -0.06 | [-0.22,  0.10] |  -0.73 | 0.464

p-value adjustment method: none
Observations: 150
```



Correlations: Answers in output…

- Do the three correlations change if you control for reading age?

```
> pcor.test(mydata$optimism, mydata$positive_SC,
+           mydata$reading_age,
+           method = "pearson")
   estimate      p.value statistic   n gp  Method
1 0.4187661 1.066678e-07  5.591127 150  1 pearson
>
> pcor.test(mydata$optimism, mydata$negative_SC,
+           mydata$reading_age,
+           method = "pearson")
    estimate     p.value statistic   n gp  Method
1 -0.2925419 0.000294163 -3.709148 150  1 pearson
>
> pcor.test(mydata$optimism, mydata$age_years,
+           mydata$reading_age,
+           method = "pearson")
    estimate    p.value statistic   n gp  Method
1 -0.1948509 0.01725172 -2.408608 150  1 pearson
```

Correlations: Answers in output…

- Are the correlations different if you look at children who do and who do not take part in extra curricular activities separately?

```
>
> cor.test(None$optimism, None$positive_SC,
+          method = "pearson")

        Pearson's product-moment correlation

data:  None$optimism and None$positive_SC
t = 1.7287, df = 71, p-value = 0.08822
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03051012  0.41198645
sample estimates:
      cor
0.2009678

> cor.test(Activities$optimism, Activities$positive_SC,
+          method = "pearson")

        Pearson's product-moment correlation

data:  Activities$optimism and Activities$positive_SC
t = 7.1104, df = 75, p-value = 5.762e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4786046 0.7517012
sample estimates:
      cor
0.6345603
```

```
fisher1925: Fisher's z (1925)
  z = -3.2704, p-value = 0.0011
  Null hypothesis rejected
```

Multiple regression: Answers in output…

1. Multiple regression: Is optimism predicted by…
   - Positive self compassion
   - Negative self compassion
   - Age (in years)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.5631     1.3215   4.966 1.88e-06 ***
positive_SC   0.7315     0.1359   5.382 2.88e-07 ***
negative_SC  -0.5384     0.1358  -3.965 0.000115 ***
age_years    -0.2341     0.1098  -2.132 0.034641 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.827 on 146 degrees of freedom
Multiple R-squared:  0.2768,	Adjusted R-squared:  0.262 
F-statistic: 18.63 on 3 and 146 DF,  p-value: 2.749e-10
```

Hierarchical regression: Answers in output…

2. Hierarchical regression: After controlling for reading age, is optimism predicted by…

- Control variable only:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.75468    0.93594   7.217 2.57e-11 ***
reading_age -0.21653    0.09665  -2.240   0.0266 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.099 on 148 degrees of freedom
Multiple R-squared:  0.0328,	Adjusted R-squared:  0.02626 
F-statistic: 5.019 on 1 and 148 DF,  p-value: 0.02656
```

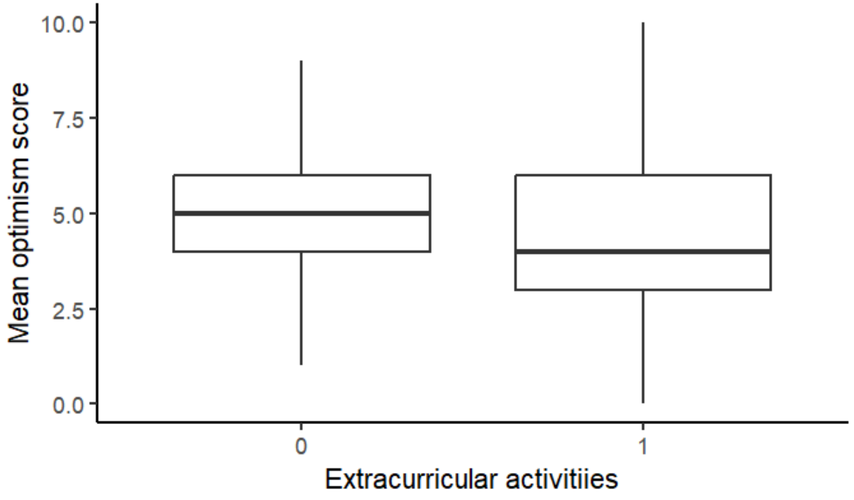
- All predictors and change statistics:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.8769     1.4650   5.377 2.96e-07 ***
reading_age  -0.1696     0.0851  -1.993   0.0482 *  
positive_SC   0.7187     0.1347   5.335 3.60e-07 ***
negative_SC  -0.5533     0.1346  -4.109 6.61e-05 ***
age_years    -0.1957     0.1104  -1.773   0.0783 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.809 on 145 degrees of freedom
Multiple R-squared:  0.2961,	Adjusted R-squared:  0.2767 
F-statistic: 15.25 on 4 and 145 DF,  p-value: 1.975e-10
```

```
> 
> r2_change <- r2_full - r2_control
> 
> print(r2_change)  # Print the Adj Rsq change
[1] 0.2504328
> 
> # Does the model significantly improve?
> 
> anova(model1,model2)
Analysis of Variance Table

Model 1: optimism ~ reading_age
Model 2: optimism ~ reading_age + positive_SC + negative_SC + age_years
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    148 651.79                                  
2    145 474.34  3    177.45 18.081 5.058e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Binary and interactive predictors: Answers in output…

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.56275    0.74243   7.493 6.27e-12 ***
positive_SC             0.32964    0.17263   1.910 0.058178 .  
negative_SC            -0.48178    0.17918  -2.689 0.008016 ** 
extra_curr1            -3.21859    1.08357  -2.970 0.003486 ** 
positive_SC:extra_curr1  0.98208    0.25499   3.851 0.000176 ***
negative_SC:extra_curr1 -0.05604    0.25972  -0.216 0.829480    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.748 on 144 degrees of freedom
Multiple R-squared:  0.3468,	Adjusted R-squared:  0.3241 
F-statistic: 15.29 on 5 and 144 DF,  p-value: 4.849e-12
```

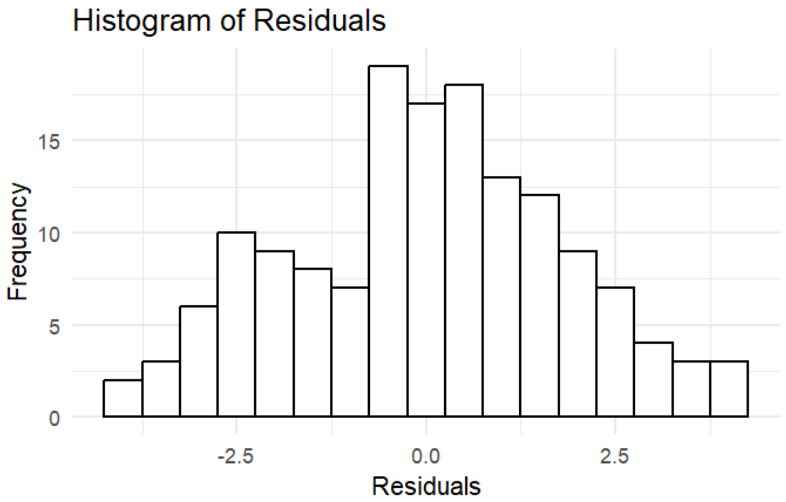



```
>
> cor.test(None$optimism, None$positive_SC,
+          method = "pearson")

	Pearson's product-moment correlation

data:  None$optimism and None$positive_SC
t = 1.7287, df = 71, p-value = 0.08822
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03051012  0.41198645
sample estimates:
      cor 
0.2009678 

> cor.test(Activities$optimism, Activities$positive_SC,
+          method = "pearson")

	Pearson's product-moment correlation

data:  Activities$optimism and Activities$positive_SC
t = 7.1104, df = 75, p-value = 5.762e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4786046 0.7517012
sample estimates:
      cor 
0.6345603 
```

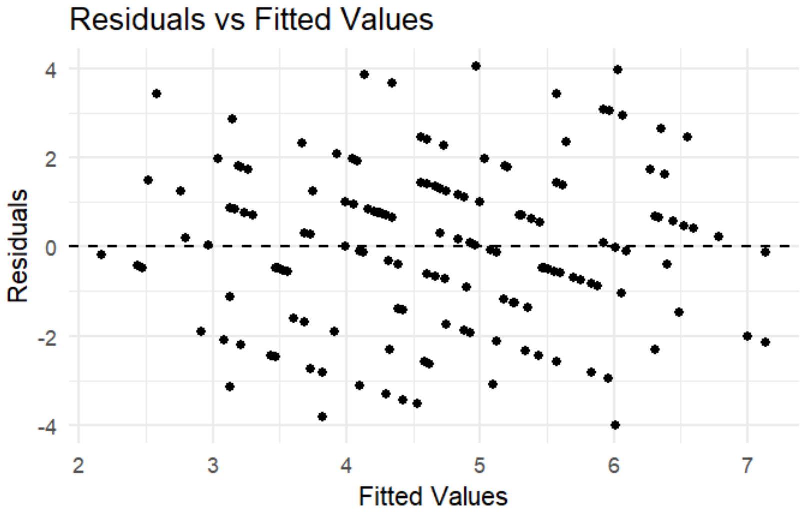
```
fisher1925: Fisher's z (1925)
  z = -3.2704, p-value = 0.0011
  Null hypothesis rejected
```

Evaluating the assumptions: Answers in output…

```
Parameter1   |  Parameter2 |     r |        95% CI | t(148) |     p
--------------------------------------------------------------------
positive_SC | negative_SC | -0.01 | [-0.17, 0.15] |  -0.16 | 0.875

p-value adjustment method: none
Observations: 150
>
> # Multicollinearity, calculate VIF.
>
> vif_values <- vif(model)
> print(vif_values)
positive_SC negative_SC  extra_curr
   1.002759    1.000169    1.002591
>
> # Multicollinearity, calculate tolerance. This is, essentially 1 - the R2
ed).
>
> tolerance_value <- 1 - summary(model)$r.squared
> print(tolerance_value)
[1] 0.7208827
```



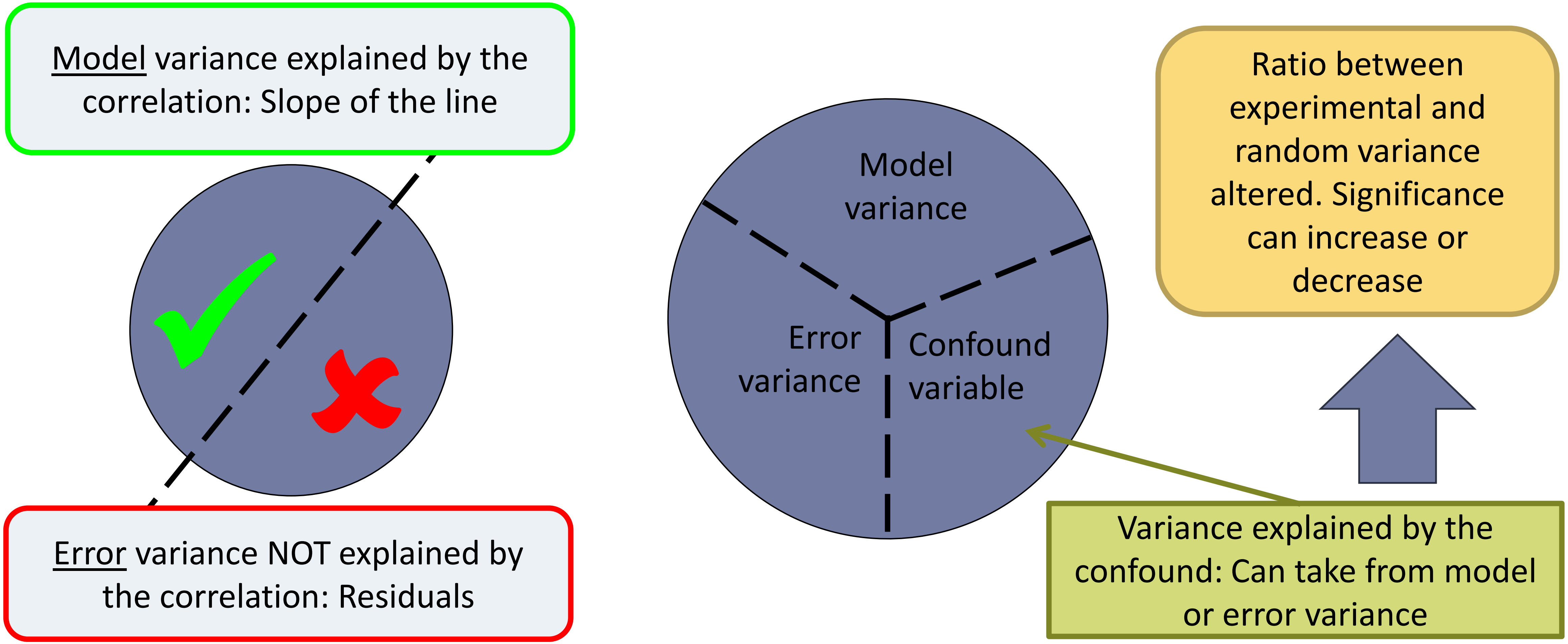

```
> # Determine the number of outliers (absolute value greater than 2)
> outliers <- sum(abs(standardized_residuals) > 2)
> print(outliers)
[1] 6
>
> # Calculate the percentage of outliers
> percentage_outliers <- (outliers / nrow(mydata)) * 100
> print(percentage_outliers)
[1] 4
>
> ##### Yay - regression in R finished!!! #####
```



Conceptual slides from correlation lecture…

Confounding variables

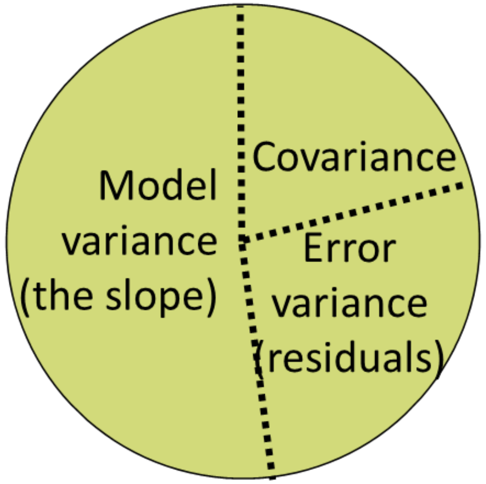
- If we can measure it, we can control for it! Remember ANCOVA...



Comparing zero order and partial correlations

	Zero order correlation (no control variable)	Partial correlation (controlling for years of edu)
SWL and psych. wellbeing	$r\ (198) = .34,\ p < .001$	$r\ (197) = .31,\ p < .001$
SWL and physical wellbeing	$r\ (198) = .20,\ p = .004$	$r\ (197) = .18,\ p = .009$
SWL and relation. wellbeing	$r\ (198) = .17,\ p = .019$	$r\ (197) = .16,\ p = .040$
SWL and negative life events	$r\ (198) = -.36,\ p < .001$	$r\ (197) = -.33,\ p < .001$

Small reduction in all correlations,
but still significant



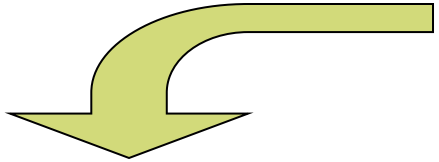
What would this cake look like?

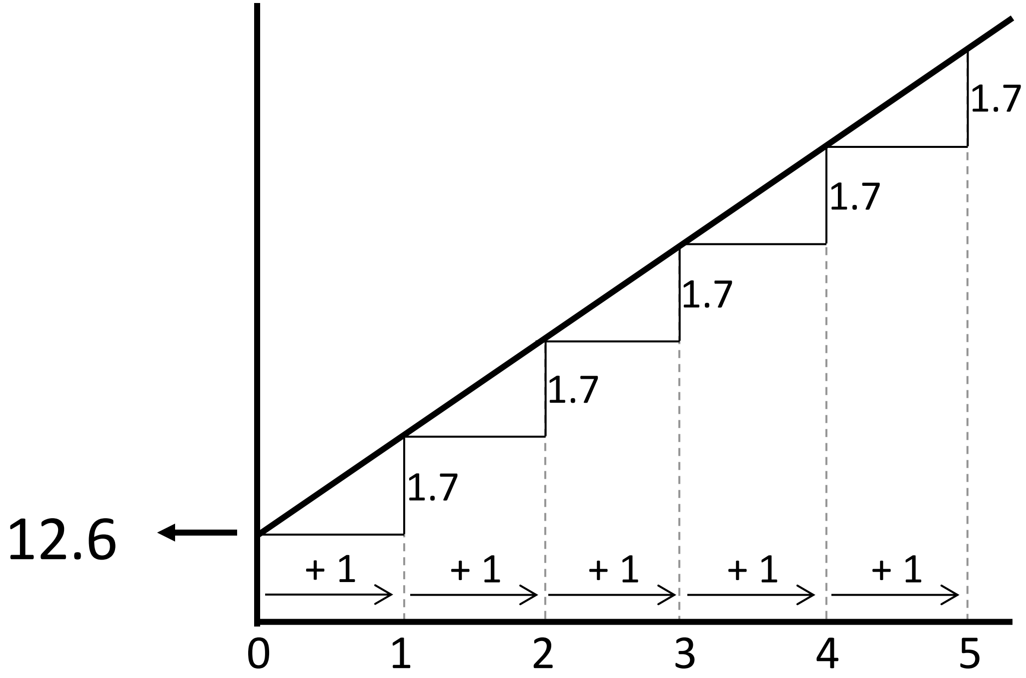
Conceptual slides from first regression lecture…

Regression: the basics

- Move beyond simple correlations:
  - Analyse more than two continuous variables
  - Build predictive models
- Terminology and variables for analysis:
  - Outcome variable: what you are trying to predict
  - Predictor variables: the variables you use to try to predict the outcome variable
  - Control variables: confounding variables that we can measure and control for
- A simple example to start with:
  - Outcome variable: satisfaction with life (SWL)
  - Predictor variable: psychological wellbeing (we will add the other variables later…)
  - If someone has a wellbeing score of 4.0, what will their SWL score be?

Regression as a predictive tool





- Intercept (β_0) = 12.6
  - Satisfaction with life score if psychological wellbeing is 0
  - The baseline/start point
- Slope (β_1) = 1.7
  - The change in satisfaction with life score for a one point increase in psychological wellbeing

Regression as a predictive tool

$12.6 + 1.7 + 1.7 + 1.7 + 1.7 + 1.7 = 21.1$

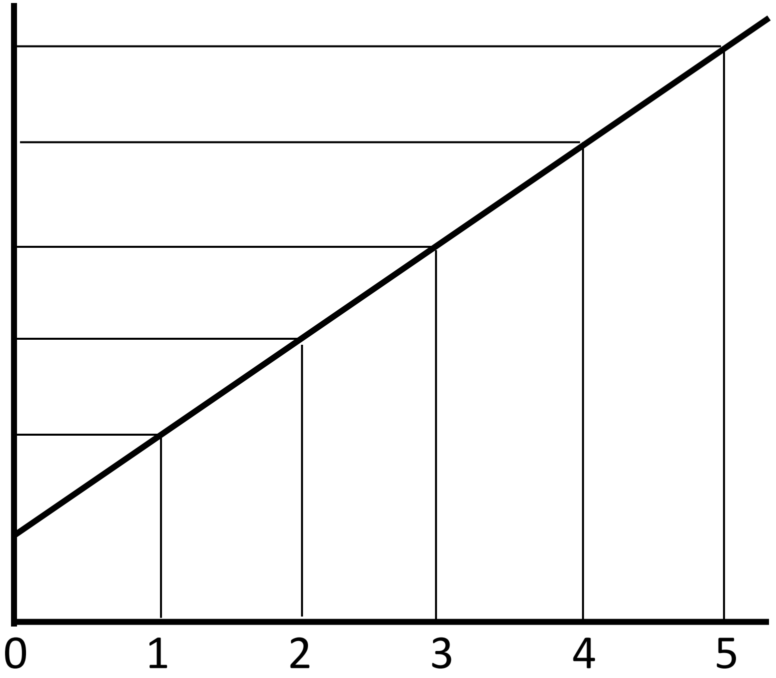
$12.6 + 1.7 + 1.7 + 1.7 + 1.7 = 19.4$

$12.6 + 1.7 + 1.7 + 1.7 = 17.7$

$12.6 + 1.7 + 1.7 = 16.0$

$12.6 + 1.7 = 14.3$

12.6



- Intercept (β_0) = 12.6
  - Satisfaction with life score if psychological wellbeing is 0
  - The baseline/start point
- Slope (β_1) = 1.7
  - The change in satisfaction with life score for a one point increase in psychological wellbeing

Regression as a predictive tool

$$Y = \beta_0 + (\beta_1 * X)$$

The value you want to predict (outcome): satisfaction with life

Intercept: 12.6

Slope: 1.7

The value you know (predictor): psychological wellbeing

$$Y = 12.6 + (1.7 * X)$$

If someone has a wellbeing score of 4.0, what will their satisfaction with life score be?

$$Y = 12.6 + (1.7 * 4.0)$$

$$Y = 12.6 + (6.8)$$

$$Y = 19.4$$

From simple to multiple regression

- Outcome variable: only a single continuous variable
- Predictor variables: can include multiple variables
  - Continuous predictor variables – this week
  - Binary (two group) predictor variables – next week!

What if I want to use more than one predictor variable?

- Outcome variable: satisfaction with life (SWL)
- Predictor variables (all continuous):
  - Psychological wellbeing
  - Physical wellbeing
  - Relationships wellbeing
  - Negative life experiences

Understanding a multiple regression model in two steps…

- **Overall model:** Are all the predictors together significant?
- **Individual predictors:** Is each individual predictor significant?

Multiple regression to predict...

Only include significant predictors!

$$Y = \beta_0 + (\underbrace{\beta_1 * X_1}_{\text{Psychological WB}}) + (\underbrace{\beta_2 * X_2}_{\text{Relationship WB}}) + (\underbrace{\beta_3 * X_3}_{\text{Neg. life events}})$$

Outcome: SWL

Intercept: 12.58

Slope: 1.74

Slope: 0.79

Slope: -0.71

$$Y = 12.58 + (1.74 * X_1) + (0.79 * X_2) + (-0.71 * X_3)$$

Predict SWL for a person with psychological WB of 3.6, relationship WB of 4.1 and NLE of 7

$$Y = 12.58 + (1.74 * 3.6) + (0.79 * 4.1) + (-0.71 * 7)$$

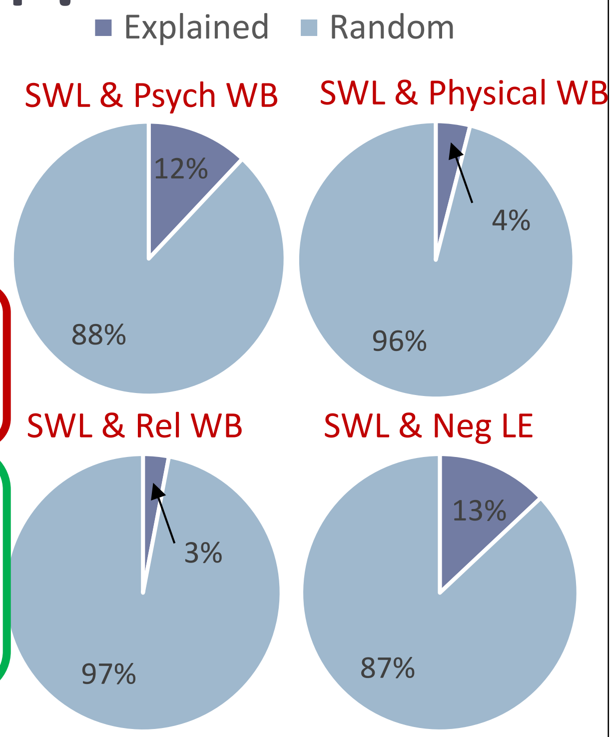
$$Y = 12.58 + (6.264) + (3.239) + (-4.97)$$

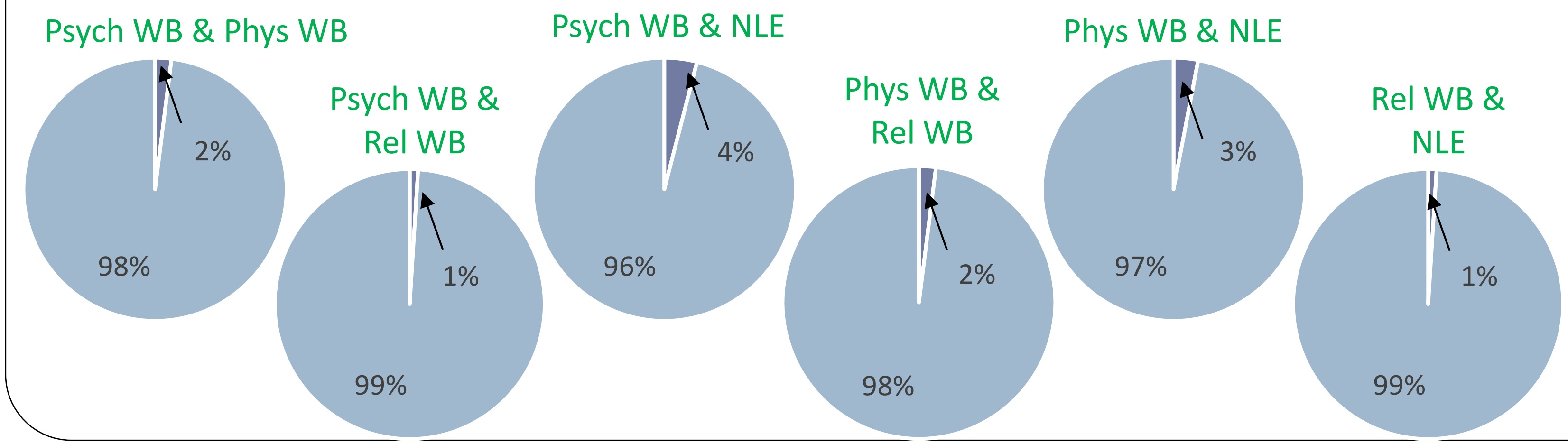
$$Y = 17.113$$

Why can correlations and regression differ?

- Zero order correlations: one slice of cake at a time!

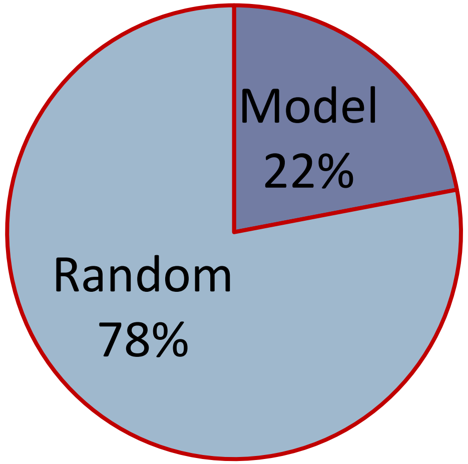
Parameter1	Parameter2	r	95% CI	t(198)	p
sat_life	psych_wellbeing	0.34	[0.21, 0.46]	5.06	< .001***
sat_life	physical_wellbeing	0.20	[0.06, 0.33]	2.90	0.004**
sat_life	relationship_wellbeing	0.17	[0.03, 0.30]	2.36	0.019*
sat_life	neg_life_experiences	-0.36	[-0.47, -0.23]	-5.36	< .001***
psych_wellbeing	physical_wellbeing	0.13	[0.00, 0.27]	1.90	0.059
psych_wellbeing	relationship_wellbeing	0.08	[-0.06, 0.22]	1.19	0.237
psych_wellbeing	neg_life_experiences	-0.20	[-0.33, -0.06]	-2.86	0.005**
physical_wellbeing	relationship_wellbeing	0.14	[0.00, 0.27]	1.93	0.055
physical_wellbeing	neg_life_experiences	-0.17	[-0.30, -0.03]	-2.44	0.016*
relationship_wellbeing	neg_life_experiences	9.63e-04	[-0.14, 0.14]	0.01	0.989

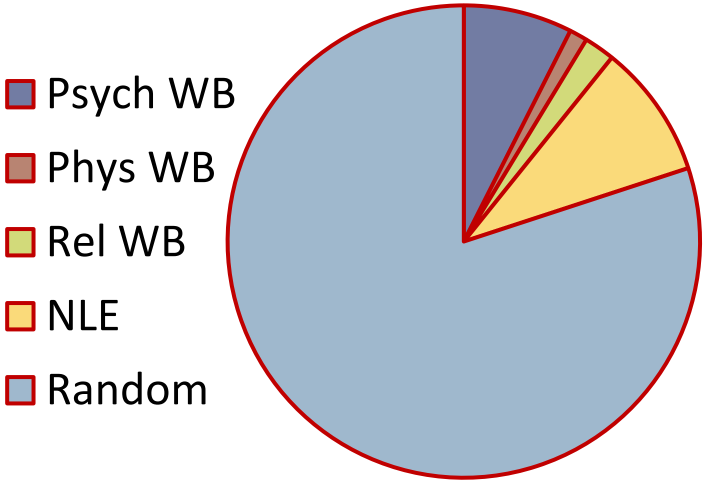




Multiple regression and multiple slices…

- One cake, but with multiple slices: deals with overlapping variance
  - Separate cakes may overestimate the amount of variance explained





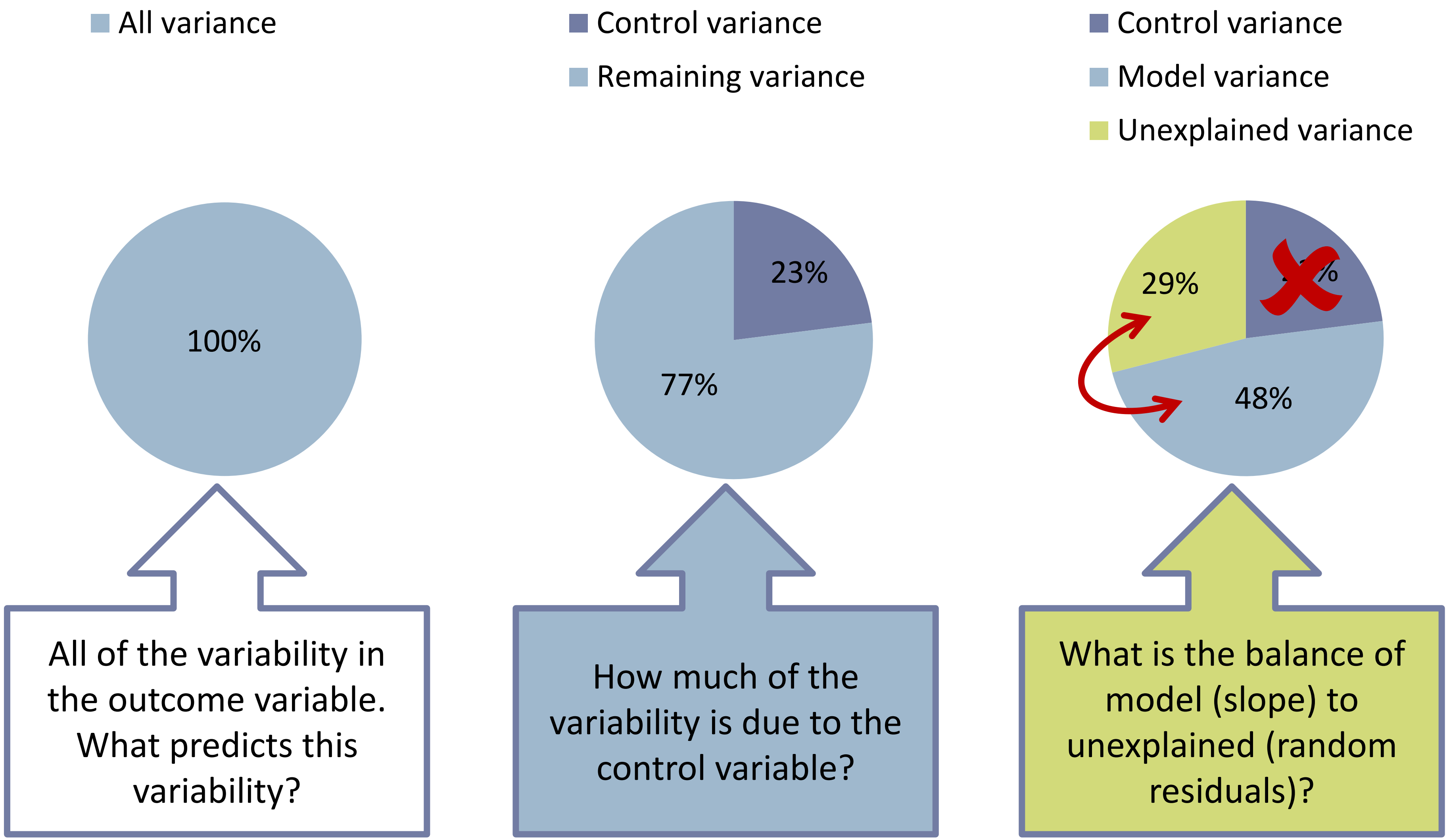
Two slices of cake for the overall model (after accounting for shared variance)

One slice of cake for each predictor variable (after accounting for shared variance)

- Psychological wellbeing:
  - $\beta = 1.74$, $t = 3.97$, $p < .001$
- Physical wellbeing:
  - $\beta = 0.66$, $t = 1.55$, $p = .122$
- Relationship wellbeing:
  - $\beta = 0.79$, $t = 2.05$, $p = .042$
- Negative life events:
  - $\beta = -0.71$, $t = -4.44$, $p < .001$

Adding confounding variables into the model/cake

- Building on partial correlations…
- How can we run a multiple regression, but control for years of education?



Comparing multiple and hierarchical regression

Multiple regression

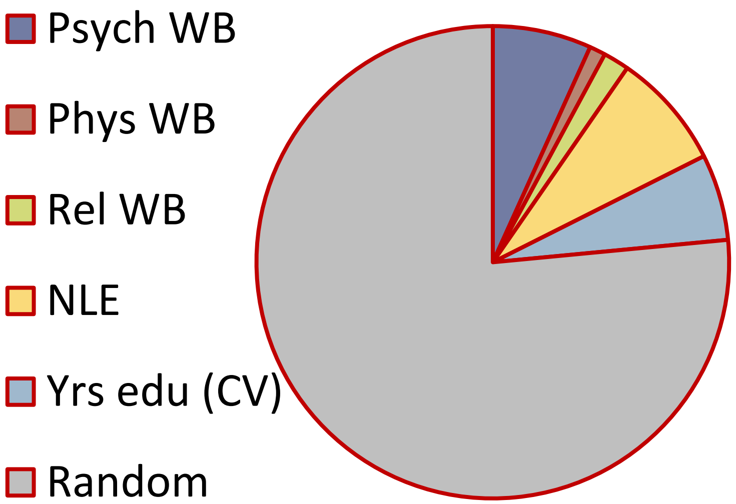
- Psych WB: Sig +ive
- Phys WB: NS
- Rel WB: Sig +ive
- NLE: Sig -ive

Hierarchical regression

- Psych WB: Sig +ive
- Phys WB: NS
- Rel WB: NS
- NLE: Sig -ive

Why do the results change?

Remember the overlapping variance in our cakes?



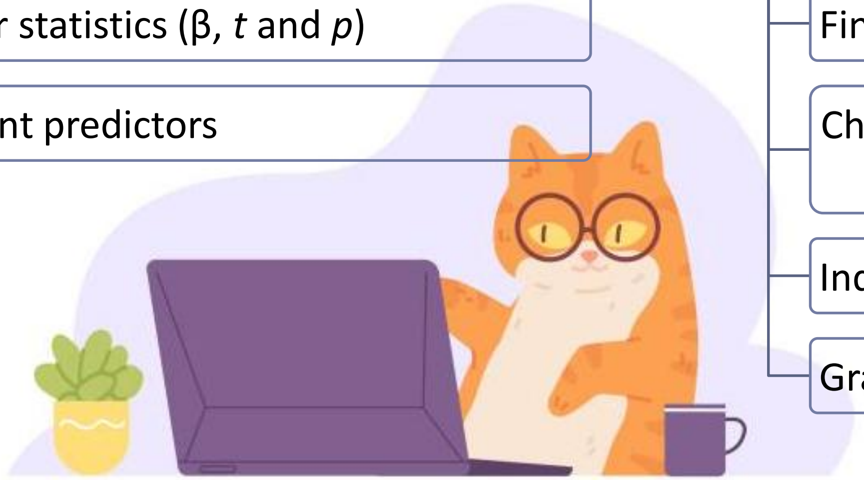
Writing up a regression analysis (Lab report 2!)

Multiple regression

- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met (covered in a later lecture)
- Full model statistics (adj. R2 and ANOVA)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors

Hierarchical regression

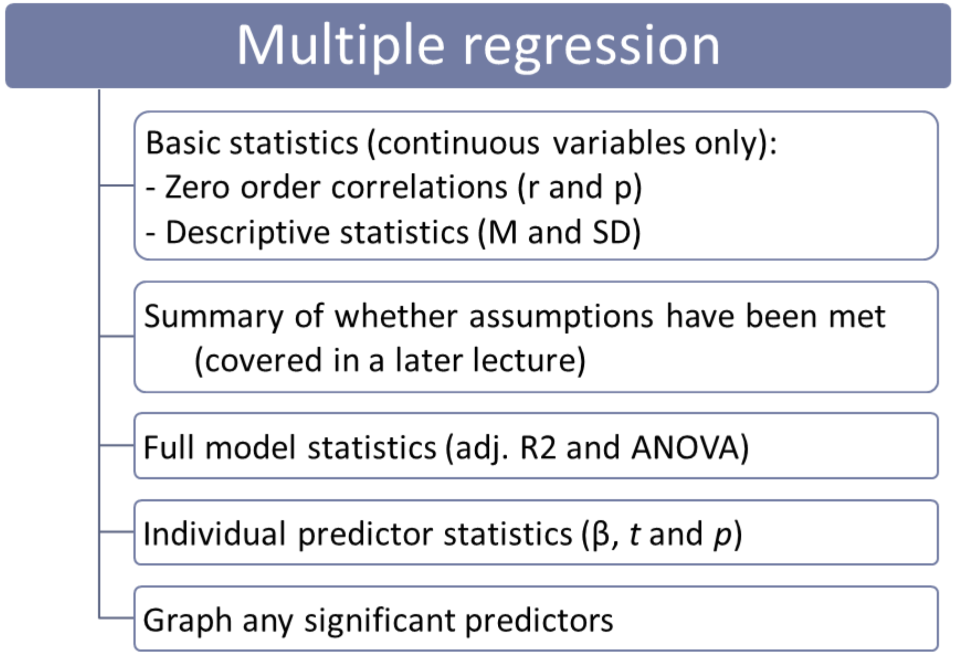
- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met (covered in a later lecture)
- Control model statistics (adj. R2 and ANOVA)
- Final model statistics (adj. R2 and ANOVA)
- Change statistics from model 1 to model 2 (ANOVA and adj. R2 change)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors



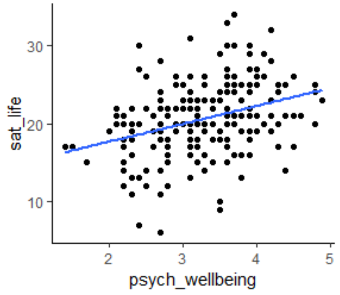
Conceptual slides from second regression lecture…

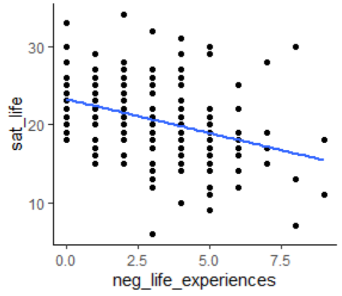
What does this all mean?

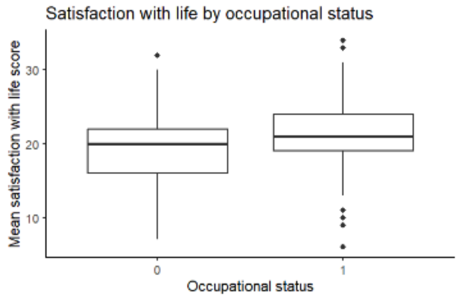
Writing up this multiple regression

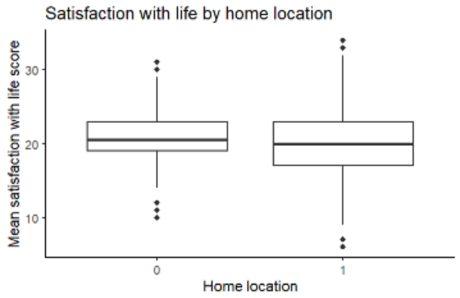


- Present your basic statistics
- Evaluate the assumptions (next lecture)
- The overall model, with *all predictors*, is significant, explaining 24.2% of the variance in SWL
- Looking at individual predictors:
  - Psychological WB predicts higher levels of SWL
  - Negative LE predicts lower levels of SWL
  - Being employed predicts higher levels of SWL
  - Living in an urban env. predicts lower levels of SWL
  - All other predictors were not significant









Remember to present the full statistics for all findings – even if NS!

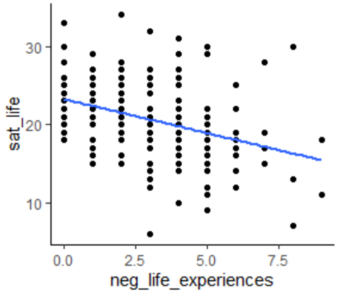
Writing up this multiple regression

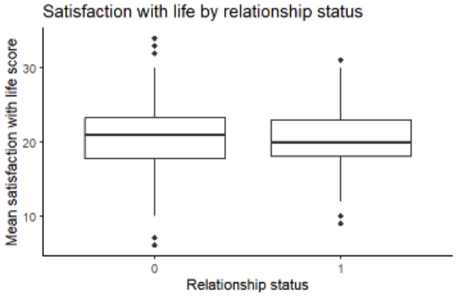
What does this all mean?

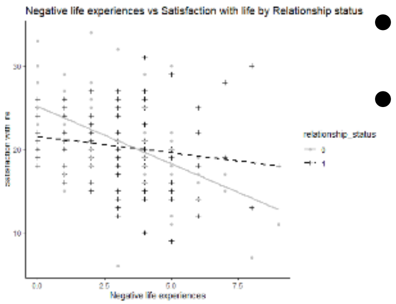
Multiple regression

- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met (covered in a later lecture)
- Full model statistics (adj. R2 and ANOVA)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors

- Present your basic statistics
- Evaluate the assumptions (next lecture)
- The overall model, with *all predictors*, is significant, explaining 15.7% of the variance in SWL
- Looking at individual predictors:
  - Negative LE predicts lower levels of SWL
  - Being in a relationship predicts lower levels of SWL
  - NLE and rel. status interact to significantly predict SWL
    - Single: Significant negative relationship
    - In a rel: No predictive relationship
    - These differ significantly



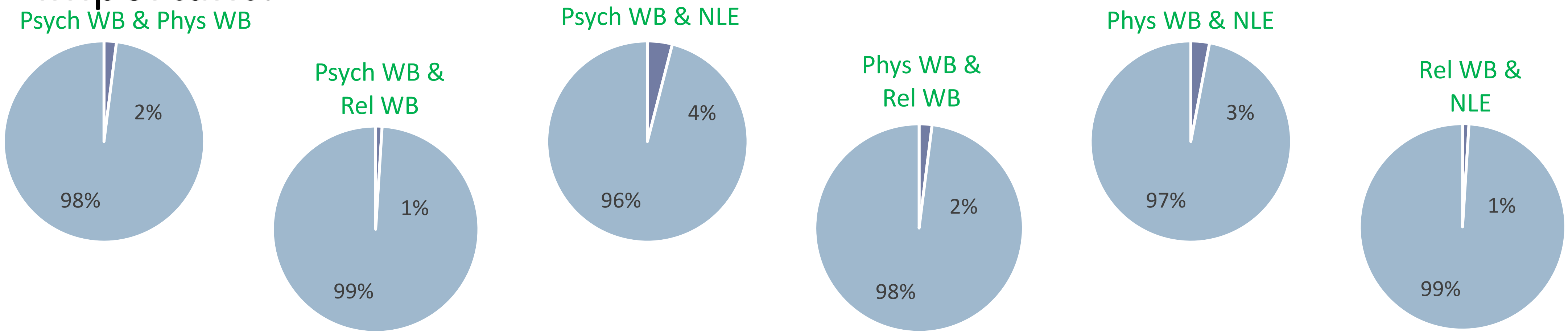




Remember to present the full statistics for all findings – even if NS!

Multicollinearity: Overlapping variance

- If slices of variance overlap too much, how do you know which is important?



Evaluating multicollinearity (three ways)

- 1: Zero order correlations
- 2: Variance inflation factor (VIF)
- 3: Tolerance

Interpreting multicollinearity

Is there any evidence of multicollinearity?

1: Zero order correlations

All within the acceptable range of +/- .9 ☺

2: Variance inflation factor (VIF)

All below 5 ☺

3: Tolerance

Above 0.2 ☺

There is no evidence of multicollinearity in the dataset, therefore each predictor variable makes a unique contribution to the predictive model ✓

Assumption of homoscedasticity

- The amount of variability in the residuals should be similar across all of the scores on the continuum, from low to high scores
  - Is the regression model equally accurate with low scoring predictors and high scoring predictors?

- Homoscedasticity: similar variance of residuals (errors) across the variable continuum (high and low outcomes)
  - Similarly accurate across all scores

- Heteroscedasticity: variance of residuals (errors) differs across the variable continuum (high and low outcomes)
  - More accurate with some scores than others

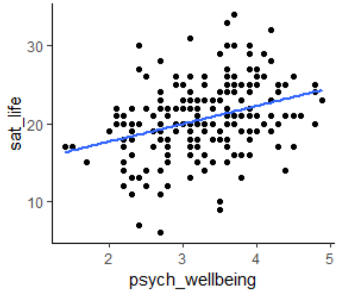
What does this all mean?

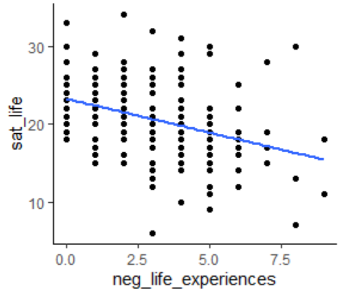
Writing up this multiple regression

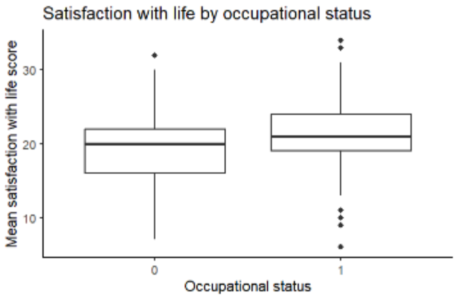
Multiple regression

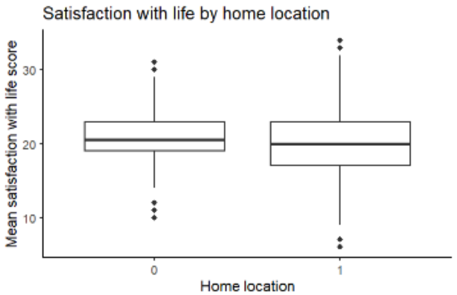
- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met (covered in a later lecture)
- Full model statistics (adj. R2 and ANOVA)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors

- Present your basic statistics
- Assumptions are met, other than too many outliers
- The overall model, with *all predictors*, is significant, explaining 24.2% of the variance in SWL
- Looking at individual predictors:
  - Psychological WB predicts higher levels of SWL
  - Negative LE predicts lower levels of SWL
  - Being employed predicts higher levels of SWL
  - Living in an urban env. predicts lower levels of SWL
  - All other predictors were not significant









Remember to present the full statistics for all findings – even if NS!

What if you violate any assumptions?

Do not attempt this for your lab report!

For PS2010 lab report 2…

- Present the results of your analyses evaluating the assumptions, and draw the appropriate conclusions
- Do not change your analysis plan
- Consider the implications of violating the assumption in your Discussion: Impact on your results?

If you wanted to publish the analysis

- Multicollinearity: Remove a predictor (check which two share the most variance and remove one)
- Too many outliers: Remove them from your dataset and repeat the multiple regression analysis
- Distribution of residuals and homoscedasticity: Often resolved by removing outliers (large residuals)

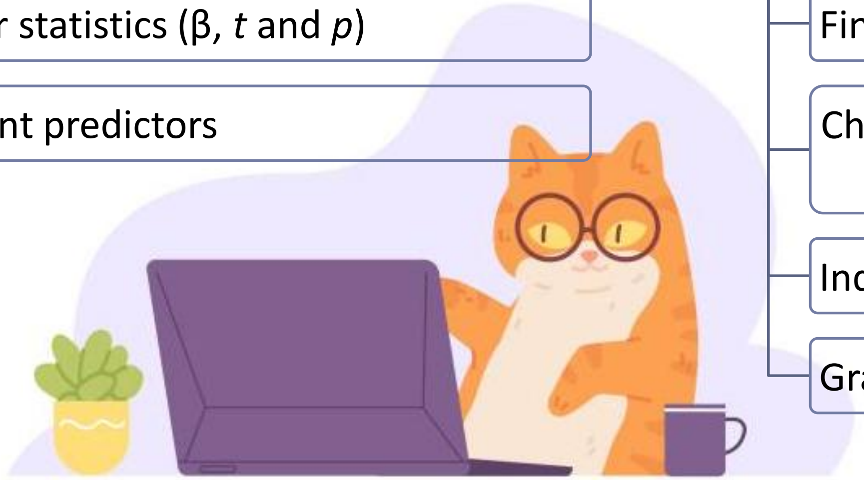
Writing up a regression analysis (Lab report 2!)

Multiple regression

- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met
- Full model statistics (adj. R2 and ANOVA)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors

Hierarchical regression

- Basic statistics (continuous variables only):
  - Zero order correlations (r and p)
  - Descriptive statistics (M and SD)
- Summary of whether assumptions have been met
- Control model statistics (adj. R2 and ANOVA)
- Final model statistics (adj. R2 and ANOVA)
- Change statistics from model 1 to model 2 (ANOVA and adj. R2 change)
- Individual predictor statistics (β, *t* and *p*)
- Graph any significant predictors



Analysis requirements for Lab Report 2

- Compulsory elements of the analysis:
  - Cronbach's alpha to evaluate the reliability of each scale in your designed questionnaire. Report this in the Methods. Covered next week!
  - Run a multiple regression including both continuous and binary predictors
  - Write up the analysis in APA format, using the structure given in this lecture
  - Create the appropriate graphs where necessary
- Optional elements of the analysis:
  - Run a hierarchical regression to include a control variable
- Not necessary in the analysis:
  - Interactive predictors within the regression
  - Factor analysis of your developed questionnaire (next week)