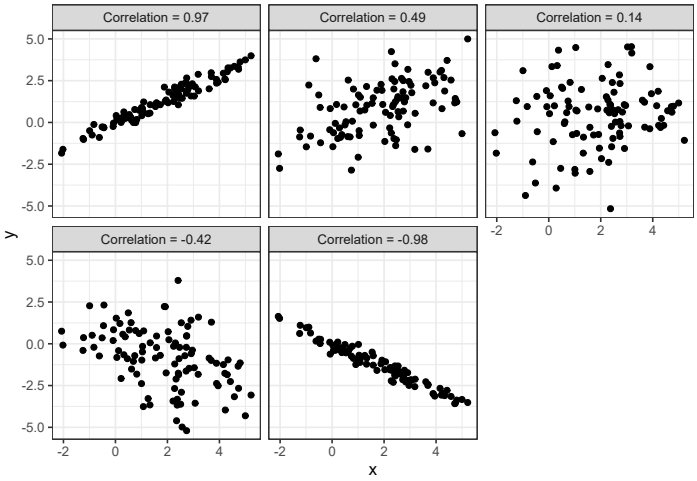
Chapter 5

Introduction to Regression Models

In order to understand the statistical models we will use to analyse longitudinal data a good understanding of regression modelling is necessary. This technique is at the heart of many of the techniques used to understand longitudinal data. We will start with correlations and basic regression models and then cover more advances aspects, such as interactions and non-linear relationships.

5.1 Correlation and Regression

Probably one of the most intuitive way to investigate the relationship between two continuous variables is by using a correlation. This is a summary statistic that tells us how much two variables co-vary. It is standardized so it can take any value between 1 (very strong relationship) and -1 (opposite relationship). A value of 0 indicates that there is no relationship between the variables. To get an intuition about what kind of relationships are represented by this indicator we can look at the following simulated data where we plot the relationship between a set of x and y variables.



We see in these graphs that for positive correlations an increase in x is associated with an increase in y while for negative correlation it's associated with a decrease. We also see that the closer the correlation is to 1 and -1 the clearer the relationship appears while the closer it is to 0 the more scattered the points are.

To calculate the correlation we can use the following formula:

$$r_{xy} = \frac{Cov(x,y)}{SD(x) * SD(y)} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

The covariance of x and y ($Cov(x,y)$) indicates to what degree the values of x change with the value of y. To calculate it we take for each case the deviation of x from the mean ($x_i - \overline{x}$) and multiply it with the deviation of y from its mean ($y_i - \overline{y}$). We add this value for all the cases to get the estimate of the covariance. If x and y vary together this will be a large number, otherwise it will be small. The covariance does not have a scale (can be any number). As such, we standardize the value by dividing it by the standard deviation of x and y. The standard deviation is a measure of variation and is calculated by taking the distance of all the cases from the mean (e.g., $x_i - \overline{x}$), squaring them ($(x_i - \overline{x})^2$), adding them up ($\sum(x_i - \overline{x})^2$) and taking the square root ($\sqrt{\sum(x_i - \overline{x})^2}$).

Fortunately we do not need to calculate the correlation by hand but it is useful to understand how it is calculated. Let's use this on some real data. Looking at the data we prepared previously we can explore the relationship between mental health in wave 1 and mental health in wave 2 using a correlation.

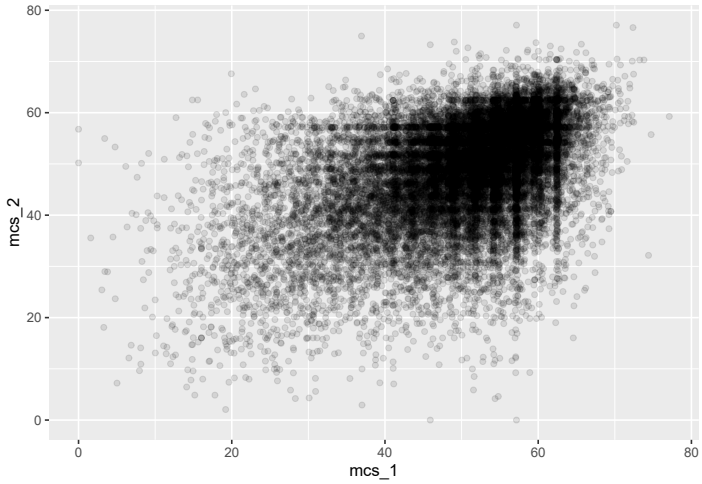
```
cor(usw$mcs_1, usw$mcs_2, use = "complete.obs")
```

```
## [1] 0.5086
```

Notice that we need to use to option: `use = "complete.obs"` *to exclude missing cases.*

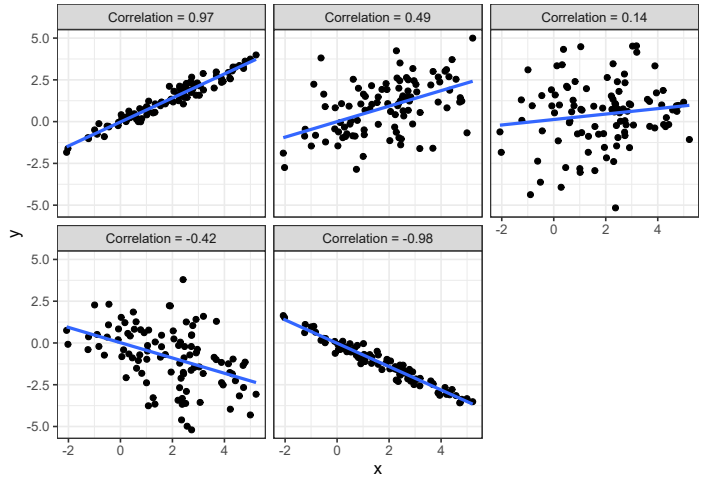
We see that we have a positive, moderate size correlation. This implies that people that have good mental health in wave 1 also have good mental health in wave 2. We can also think of this as "stability", indicating that mental health is fairly stable in time. We can also explore the relationship using a scatter plot.

```
ggplot(usw, aes(mcs_1, mcs_2)) +
  geom_point(alpha = 0.1)
```



We can see a pattern in the data with larger values on "mcs_1" being associated with larger values on "mcs_2". The points are relatively scattered, indicating that while we have a relationship between the two variables they are not identical, indicating some change in time for mental health.

Correlations are a way to look at relationships between two continuous variables. An alternative way is to use a regression model. This tells us how we can represent the relationship between x and y using a line. For example, looking at the simulated data, we can represent the same relationships using a regression line:



A regression also represents the relationship between variables but in a different way. Typically, the regression is represented mathematically using the following formula:

$$y_i = \alpha + \beta * x_i + \sigma_i$$

- y_i is the outcome or the dependent variable. It's the variable we want to explain or predict using out model. This value varies by individual (i)
- α (alpha) is also known as the **intercept** or the expected value when the predictor is 0

- β (beta) is also know as the **slope** and tells us what is the expected change in the dependent variable (y) with a increase of 1 in the independent variable (x)
- σ (sigma) is also know as **the residual**. This represents the unexplained variance of y and is a summary of the distance between the regression line and the observed values of y.

Let's look at an example. Using the simulated data above (where the correlation is 0.97) we can run a regression using the `lm()` command. We give as input the formula. Here *y1* is the dependent variable and so is on the left of the ~ (tilde) symbol while x is the predictor or the independent variable. We also indicate what is the data we are using, in this case this is called "df". We save the object as "m1" and then print it using the `summary()` command.

```
m1 <- lm(y1 ~ x, data = df)

summary(m1)
```

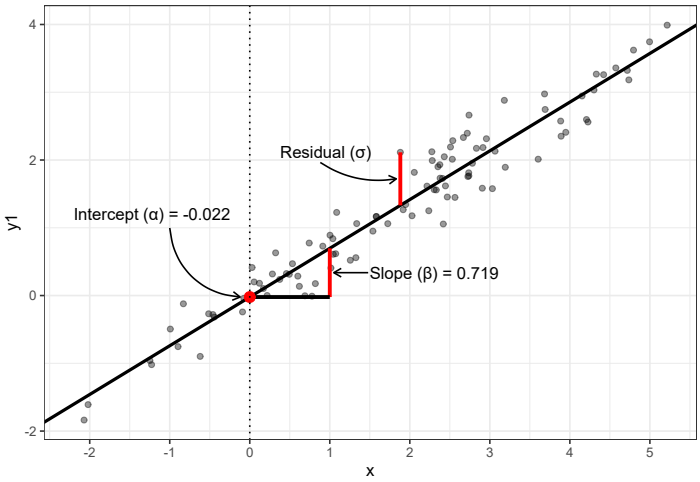
```
##
## Call:
## lm(formula = y1 ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6610 -0.1808 -0.0033  0.1857  0.7829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0217     0.0456   -0.48     0.63
## x             0.7189     0.0182   39.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.303 on 98 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.94
## F-statistic: 1.56e+03 on 1 and 98 DF,  p-value: <2e-16
```

In the output we can see that the intercept is **-0.022**. This tells us that the expected value of *y1* when x is 0. The slope describes the relationship between x and *y1*. Here it indicates that when the value of x increases by 1 then the expected value of *y1* increases by **0.719**. The residual of this relationship is **0.303** and this is the unexplained variance.

We can represent this relationship using the formula introduced above:

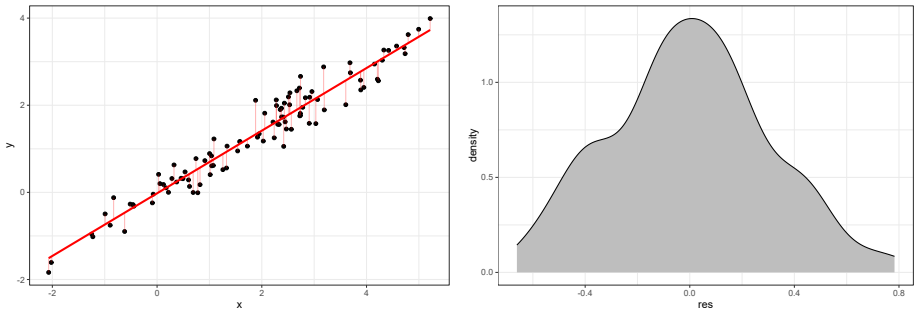
$$y1 = -0.022 + 0.719 * x + 0.303$$

To help us better understand this relationship we can also visualize it using a scatter plot.



Here we can see how the regression line represents the relationship between x and *y1*. We see that the slope indicates the angle of this line. It shows how the value of *y1* increases by **0.719** when x increases by 1. We also see that the intercept is the place where the line intersects the value 0 on the x scale. This indicates that when the value of x is 0 we expect that y will have a value of **-0.022**. Lastly, we see that the residual is an indicator of the distance between the observed scores (the dots in the scatterplot) and the predicted value (the line).

More precisely, the residual is a variable that is created by taking the differences between the observed scored and the predicted values for all the cases. This is assumed to have a normal distribution with a mean of 0. In general, the better our model in explaining the dependent variable the smaller the residual. For the previous regression we can represent the residuals visually using these two graphs:



The graph on the left shows how the residual variable is calculated while the one on the right shows its observed distribution (or density).

Let's apply the regression to our longitudinal data on mental health. Here we explain mental health in wave 2 using mental health in wave 1. This could be a useful way to understand how stable is this variable.

```
m2 <- lm(mcs_2 ~ mcs_1, data = usw)

summary(m2)
```

```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.13  -4.13   1.47   5.28  33.32
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.18288    0.26744    90.4   <2e-16 ***
## mcs_1        0.50635    0.00516    98.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.28 on 27637 degrees of freedom
##   (23355 observations deleted due to missingness)
## Multiple R-squared:  0.259,  Adjusted R-squared:  0.259
## F-statistic: 9.64e+03 on 1 and 27637 DF,  p-value: <2e-16
```

Looking at the results we see that if mental health in wave 1 ("mcs_1") increase by 1 then the expected value of mental health in wave 2 ("mcs_2") increases by **0.506**. This indicates a moderate positive relationship between mental health in wave 1 and mental health in wave 2 (similar to what was indicated by the correlation). Through the intercept we see that for those respondents that had a value of 0 on mental health in wave 1 the expected mental health in wave 2 will be **24.183**.

Because the residual is not standardized it can be hard to interpret. And alternative indicator we can use to understand how well our predictors explain the our is the R-squared. This indicates what proportion of the total variation of the dependent variable is explained by our model. This can range from 0, we explain no variation, to 1, we explain all the variation. In our model the R-squared is **0.259**.

The coefficients interpreted so far describe the relationships in the observed data. Sometimes, we might want inferences about the general population. For example, if we use a sample of the general population, as we are doing when using the Understanding Society data, we might not just want to say something about the respondents in the study but also about what is happening more generally in the population. That is where the standard error, t value and p values come in. The standard error is an estimate of uncertainty and tells us how much we expect the coefficients to vary. We can use it to create the confidence interval which gives us a range in which we expect the coefficient to be in the population.

The p-value, on the other hand, can be used to make a significance test whether the observed score is significantly different from 0 in the population. The null hypothesis of this test is that the observed score is equal to 0 in the population. If we assume a cut-off point of 0.05 we can reject this null hypotheses for p-values bellow that. While this may be useful in some conditions the p-value is a contentious topic in statistics given it's multiple assumptions and misuse. My general recommendation is to focus more on interpreting the main effects from a substantive point of view (are they large or small? are the effects important from a substantive point of view?) and also focus more on presenting the uncertainty of the findings, for example using confidence intervals.

For example, we could calculate the 95% confidence interval for the observed slope by using the formula:

$$\hat{\beta} \pm 1.96 * se(\hat{\beta})$$

Using the observe values from the output

$$0.506 \pm 1.96 * 0.005 = 0.506 \pm 0.0098 = (0.496, 0.516)$$

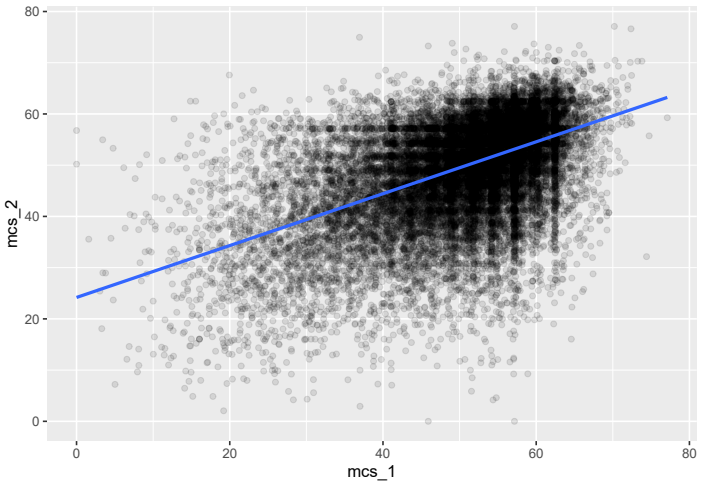
This would indicate that we expect the relationship between mental health in wave 1 and mental health in wave 2 to be somewhere between **0.496** and **0.516** in the population. Again, this comes with some assumptions. For example, because we use the 95% confidence interval it means that 5% the times the population values will be outside of that range.

We can represent the relationship from the regression using the formula introduced above as well:

$$mcs\_2 = 24.183 + 0.506 * mcs\_1 + 8.277$$

We can also visualize it using a graph.

```
ggplot(usw, aes(mcs_1, mcs_2)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", se = F)
```



5.2 Modelling Different Types of Relationships

The regression model is extremely flexible and can be extended in a number of ways. One way to extend is by using multiple predictors. We rarely expect one variable to perfectly explain another. As such, typically we want to include different predictors in our regression models. For example, we might want to expand the previous model explaining mental health in wave 2 by also including age. This has two advantages. Firstly, we can investigate the relationship between age and mental health. Secondly, by including age in the model we may also get a different estimate of the relationship between mental health in wave 1 and mental health in wave 2. This is because, by

including variables in the regression model, we are effectively "controlling", or taking into account, their effects. This is one of the strengths of regression modeling, that make it so popular in the social sciences where we often have observational data and need to control for possible confounders.

To extend the model with age we can simply add it in the formula together with the + symbol:

```
m3 <- lm(mcs_2 ~ mcs_1 + age, data = usw)

summary(m3)
```

```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + age, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.81  -4.09   1.54   5.25  33.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.75917    0.28808    79.0   <2e-16 ***
## mcs_1        0.49992    0.00516    96.8   <2e-16 ***
## age          0.03743    0.00287    13.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.25 on 27636 degrees of freedom
##   (23355 observations deleted due to missingness)
## Multiple R-squared:  0.263,  Adjusted R-squared:  0.263
## F-statistic: 4.94e+03 on 2 and 27636 DF,  p-value: <2e-16
```

Looking at the results we see that there is a positive relationship between age and mental health. When age increases by 1 the expected mental health in wave 2 increases by 0.037. We also notice that the effect of mental health in wave 1 on mental health in wave 2 is slightly smaller (0.5 vs. 0.506). This is because in the new model we are "controlling" for age when calculating this relationship. It appears that some of the relationship between these two variables was explained by age. Finally, we notice that our model is slightly better now with the R-squared being larger than before (0.263 vs. 0.259). It appear that age helps us explain an additional 0.4% of variance compared to the previous model.

We can write up the relationships observed in the data using a formula as before:

$$mcs\_2 = 22.759 + 0.5 * mcs\_1 + 0.037 * age + 8.252$$

When we have multiple predictors the interpretation of the intercept is the expected value of the outcome when all the predictors are 0. This is because when "mcs_1" and "age" become 0, the expected score will be based on the intercept (remember the mean or expected value of the residual is 0):

$$mcs\_2 = 22.759 + 0.5 * 0 + 0.037 * 0 = 22.759$$

So the interpretation would be that we expect a score on mental health in wave 2 of **22.759** for respondents that have a score of 0 for mental health in wave 1 and age 0. While this might be mathematically sound it does not make a lot of sense from a substantive point of view as we do not have respondents of age 0 and cannot predict their values based on our data.

We can make the intercept more useful by recoding the age variable. In chapter 3 we created a centered version of age by subtracting the average from the original variable. This results in the same distribution but with a new average of 0. We can add this variable in the model instead of the original age variable to make the intercept easier to interpret.

```
m3b <- lm(mcs_2 ~ mcs_1 + age_center, data = usw)

summary(m3b)
```

```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + age_center, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.81  -4.09   1.54   5.25  33.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.46753    0.26752    91.5   <2e-16 ***
## mcs_1        0.49992    0.00516    96.8   <2e-16 ***
## age_center   0.03743    0.00287    13.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.25 on 27636 degrees of freedom
##   (23355 observations deleted due to missingness)
## Multiple R-squared:  0.263,  Adjusted R-squared:  0.263
## F-statistic: 4.94e+03 on 2 and 27636 DF,  p-value: <2e-16
```

Notice that most of the coefficients in the regression are the same as before (including the slopes). The only coefficient that changes is the intercept. The new value, **24.468** can be interpreted as the expected mental health in wave 2 for respondents that had 0 on the mental health score in wave 1 and have **average age**. As you can see, the interpretation is much more interesting now. If we wanted we could also center mental health in wave 1 so then the intercept would refer to the expected value for those that had average mental health in wave 1. Remember, that the intercept can be a useful tool to understand your data but you need to be care how you code the predictors.

5.2.1 Categorical Predictors

Another way to expand regression models is by including categorical predictors. So far we only included continuous ones. Let's see how the interpretation of the coefficients is different in this context.

Let's see how education impacts mental health. We have coded education in different ways but let's use here the degree variable we created. Remember that this was coded as a factor if people had a degree or not:

```
count(usw, degree)
```

```
## # A tibble: 3 x 2
##   degree        n
##   <fct>     <int>
## 1 Degree    16491
## 2 No degree 34411
## 3 <NA>         92
```

We can add it to the regression like we did before. We will also keep mental health in wave 1 in the model as a control variable.

```
m4 <- lm(mcs_2 ~ mcs_1 + degree, data = usw)

summary(m4)
```

```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + degree, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.02  -4.14   1.48   5.28  33.40
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.39689    0.27889   87.48   <2e-16 ***
## mcs_1            0.50569    0.00516   97.98   <2e-16 ***
## degreeNo degree -0.27911    0.10416   -2.68   0.0074 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.28 on 27632 degrees of freedom
##   (23359 observations deleted due to missingness)
## Multiple R-squared:  0.259,  Adjusted R-squared:  0.259
## F-statistic: 4.83e+03 on 2 and 27632 DF,  p-value: <2e-16
```

When including a categorical variable that is coded as a factor in a regression R will automatically select the first category as a reference and include a dummy variable in

the model where the reference is coded as 0 and the other category as 1. So, in the background such a variables is created:

```
usw %>%
  mutate(no_degree_dummy = ifelse(degree == "Degree", 0, 1)) %>%
  count(degree, no_degree_dummy)
```

```
## # A tibble: 3 x 3
##   degree    no_degree_dummy     n
##   <fct>               <dbl> <int>
## 1 Degree                  0 16491
## 2 No degree               1 34411
## 3 <NA>                   NA    92
```

As a result, when we interpret the slope for the "degree" variable we should interpret how the category coded as 1 (in this case "No degree") is different compared to the reference (in this case "Degree"). So, the interpretation here would be that people that do not have a degree, have on average lower mental health compared to those that have a degree. The expected difference is -0.279.

If we write down the regression formula it would look like this:

$$mcs\_2 = 24.397 + 0.506 * mcs_1 - 0.279 * No\_degree + 8.277$$

If we wanted to understand how this difference between the two groups comes about we can write down the expected regression formulas for the two groups. For people with no degree the dummy variable becomes 0 and the expected value will be:

$$mcs\_2_{Degree} = 24.397 + 0.506 * mcs_1 - 0.279 * 0 = 24.397 + 0.506 * mcs_1$$

While for those people that do not have a degree the formula is:

$$mcs\_2_{No\_degree} = 24.397 + 0.506 * mcs_1 + -0.279 * 1 = 24.118 + 0.506 * mcs_1$$

It is sometimes useful to write the regression model down to see what are the expected values under different circumstances.

Sometimes it might be useful to change the reference category to make the interpretation easier. For example, here it might be easier to talk about people that have a degree versus the rest. We could do this in a few different ways. We could change the order of the levels of the factor. We could create our own dummy variable coded the other way around. Alternatively, we can use the `relevel()` command in the regression to change the reference.

```
m4b <- lm(mcs_2 ~ mcs_1 + relevel(degree, ref = 2), data = usw)

summary(m4b)
```

```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + relevel(degree, ref = 2), data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.02  -4.14   1.48   5.28  33.40
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    24.11777    0.26860   89.79   <2e-16 ***
## mcs_1                           0.50569    0.00516   97.98   <2e-16 ***
## relevel(degree, ref = 2)Degree  0.27911    0.10416    2.68   0.0074 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.28 on 27632 degrees of freedom
##   (23359 observations deleted due to missingness)
## Multiple R-squared:  0.259,  Adjusted R-squared:  0.259
## F-statistic: 4.83e+03 on 2 and 27632 DF,  p-value: <2e-16
```

Most of the coefficients are the same with two exceptions. Firstly, the coefficient for degree has been reversed (0.279 vs. -0.279). This is because now we are comparing people that have a degree to those that don't. The size of the effect is the same, just the direction is different. Secondly, the value of the intercept changed (24.397 vs. 24.118). Because the reference category has changed the intercept also changed. The new value indicated the expected value of mental health in wave 2 **for people with no degree** (and mental health of 0 in wave 1).

We can also include categorical variables with more than two categories. Let's see how marital status in wave 1 impacts mental health in wave 2.

```
count(usw, marstatus_fct_1)
```

```
## # A tibble: 7 x 2
##   marstatus_fct_1           n
##   <fct>                 <int>
## 1 Married/Civil partner 25958
## 2 Living as couple       5727
## 3 Widowed                3005
## 4 Divorced               3148
## 5 Separated              1153
## 6 Never married         11967
## 7 <NA>                     36
```

When we include this variable in the regression R will again select the first category as the reference and create dummy variables for all the other variables. This is the equivalent of doing something like this:

```
usw %>%
  mutate(Living_as_c =
           ifelse(marstatus_fct_1 == "Living as couple", 1, 0),
         Widowed =
           ifelse(marstatus_fct_1 == "Widowed", 1, 0),
         Divorced =
           ifelse(marstatus_fct_1 == "Divorced", 1, 0),
         Separated =
           ifelse(marstatus_fct_1 == "Separated", 1, 0),
         Never_married =
           ifelse(marstatus_fct_1 == "Never married", 1, 0)) %>%
  count(marstatus_fct_1, Living_as_c, Widowed, Divorced,
        Separated, Never_married)
```

```
## # A tibble: 7 x 7
##   marstatus_fct_1       Living_as_c Widowed Divorced Separated Never_mar~1     n
##   <fct>                       <dbl>   <dbl>    <dbl>     <dbl>       <dbl> <int>
## 1 Married/Civil partner           0       0        0         0           0 25958
## 2 Living as couple                1       0        0         0           0  5727
## 3 Widowed                         0       1        0         0           0  3005
## 4 Divorced                        0       0        1         0           0  3148
## 5 Separated                       0       0        0         1           0  1153
## 6 Never married                   0       0        0         0           1 11967
## 7 <NA>                           NA      NA       NA        NA          NA    36
## # ... with abbreviated variable name 1: Never_married
```

Let's see how the output would look if we add this variable

```
m5 <- lm(mcs_2 ~ mcs_1 + marstatus_fct_1, data = usw)

summary(m5)
```

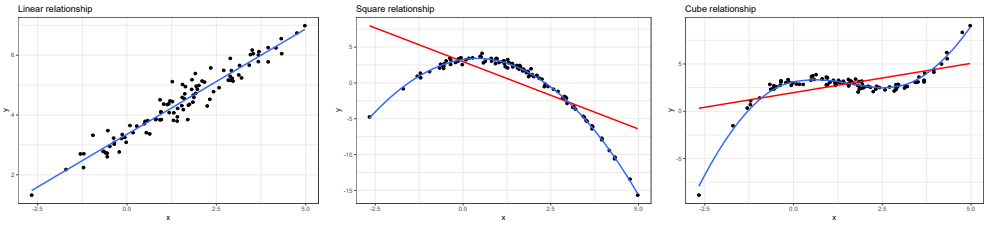
```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + marstatus_fct_1, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.48  -4.12   1.55   5.18  32.90
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     24.88195    0.27801   89.50  < 2e-16 ***
## mcs_1                            0.50037    0.00519   96.38  < 2e-16 ***
## marstatus_fct_1Living as couple -1.05987    0.15952   -6.64  3.1e-11 ***
## marstatus_fct_1Widowed           0.48227    0.22338    2.16  0.03086 *
## marstatus_fct_1Divorced         -1.05621    0.20350   -5.19  2.1e-07 ***
## marstatus_fct_1Separated        -1.18606    0.35207   -3.37  0.00076 ***
```

```
## marstatus_fct_1Never married    -1.01292    0.13135   -7.71  1.3e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.26 on 27615 degrees of freedom
##   (23372 observations deleted due to missingness)
## Multiple R-squared:  0.262,  Adjusted R-squared:  0.262
## F-statistic: 1.63e+03 on 6 and 27615 DF,  p-value: <2e-16
```

We see that now we have five different variables included in the regression instead of "marstatus_fct_1". The first category is excluded, and used as a reference while the others are included as dummy variables as shown above. The interpretation of these variables is always in comparison with the reference category. For example, people who live as a couple have lower mental health **compared to those that are married** (the reference category). The difference in expected mental health is -1.06. Similarly, people who were never married lower mental health **compared to those that are married** of -1.013.

5.2.2 Non-linear Relationships

So far we have assumed that there is a linear relationship between predictors and the the outcome in the regression. This means that an increase of the independent variable will always lead to the same effect on the outcome. That may not be always the case. For example, looking at the relationship between income and happiness we might observe a positive relationship at lower incomes but the effect might flatten out or even start to decrease after a certain point. Here are some examples of different types of relationships:



The first graph, on the left, shows a linear relationship that can be modeled using the approach we have used so far. The other two graphs show non-linear relationships. This first one shows an initial positive relationship that then becomes negative while the graph on the right shows a more complex pattern with an initial increase, then a plateau and decrease and then another increase. For these latter two relationships, using a straight line to represent it (the red line in the graph) would lead to incorrect conclusions regarding how x is effecting y.

We can explicitly model such relationships in regressions using two main strategies. The first one is to include polynomials in the regression. For example, if we investigate the relationship between income and mental health we could include both income and income squared. By including a polynomial we allow the relationship between income and mental health to bend once. If the coefficient of income squared on mental health is positive it means the relationship bends upward, i.e., as income increases the effect on mental health is larger. If, on the other hand, the effect is negative, the bend is downwards, implying that the effect of income is smaller for people with larger incomes.

Let's look at an example. We will investigate if there is a non-linear relationship between income in wave 1 and mental health in wave 2. In this regression we include both income (the logged version saved as "logincome_1") as well as the square effect. This could be a variable we save in advance or we can create it directly in the regression using the `I()` command. Here we use the latter approach:

```
m6 <- lm(mcs_2 ~ logincome_1 + I(logincome_1^2),
         data = usw)

summary(m6)
```

```
##
## Call:
## lm(formula = mcs_2 ~ logincome_1 + I(logincome_1^2), data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.04  -4.93   2.15   6.92  27.61
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        52.3311     0.5691   91.95  < 2e-16 ***
## logincome_1        -1.4743     0.2104   -7.01  2.5e-12 ***
## I(logincome_1^2)    0.1594     0.0187    8.50  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.59 on 28289 degrees of freedom
##   (22702 observations deleted due to missingness)
## Multiple R-squared:  0.0044, Adjusted R-squared:  0.00433
## F-statistic: 62.4 on 2 and 28289 DF,  p-value: <2e-16
```

Looking at the results we see that main effect of logincome is negative (-1.474) while the effect of the squared income is positive (0.159). To understand how a non-linear relationship is created by including the polynomials we can write down the equation and work out two scenarios:

$$mcs\_2 = 52.331 - 1.474 * logincome\_1 + 0.159 * logincome\_1^2$$

Let's look at how the regression looks like when logincome is small, say 2, and when it is large, for example 8:

$$mcs\_2_{logincome=2} = 52.331 - 1.474 * 2 + 0.159 * 2^2 = 52.331 - 2.948 + 0.636 = 50.019$$

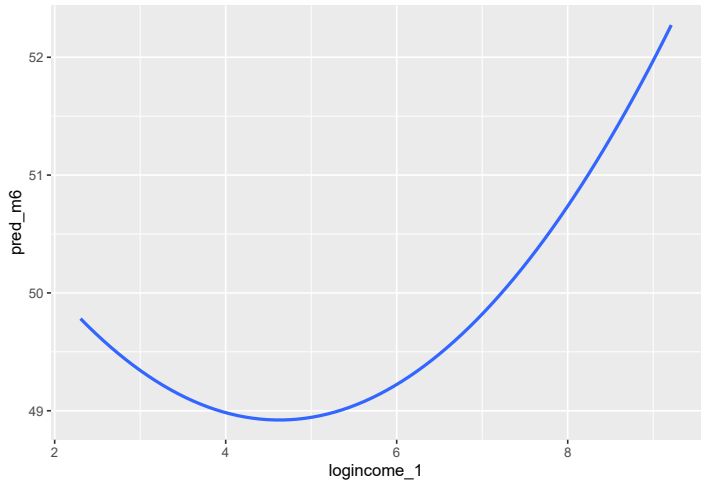
$$mcs\_2_{logincome=8} = 52.331 - 1.474 * 8 + 0.159 * 8^2 = 52.331 - 11.792 + 10.176 = 50.715$$

We can observe that for smaller values of logincome the linear effect is more important. That means, that early on the relationship is mainly defined by the linear effect. For larger values of logincome the squared effect becomes more important. So, looking at the results we see that initially logincome has a negative effect on mental health but this becomes positive for larger values of logincome.

We can also use the predicted values to see the expected relationship based on our model. Below we save the predicted score and then use `geom_smooth()` to represent the relationship. We estimate a non-linear relationship with two polynomials using the `poly(x, 2)` command:

```
usw <- mutate(usw, pred_m6 = predict(m6, usw))

ggplot(usw, aes(logincome_1, pred_m6)) +
  geom_smooth(method = lm,
              formula = y ~ poly(x, 2),
              se = F)
```



If we want to include additional "bends" in the relationship between income and mental health we can add more polynomials. For each additional polynomial we allow for an additional "bend". If the coefficient for the polynomial is positive, then the "bend" will be upward, if it is negative it will be downard. The larger the coefficient the stronger the change.

Let's explore to see if the relationship between income and mental health is actually more complex. Let's include also the cube:

```
m6b <- lm(mcs_2 ~ logincome_1 + I(logincome_1^2) +
            I(logincome_1^3), data = usw)

summary(m6b)
```

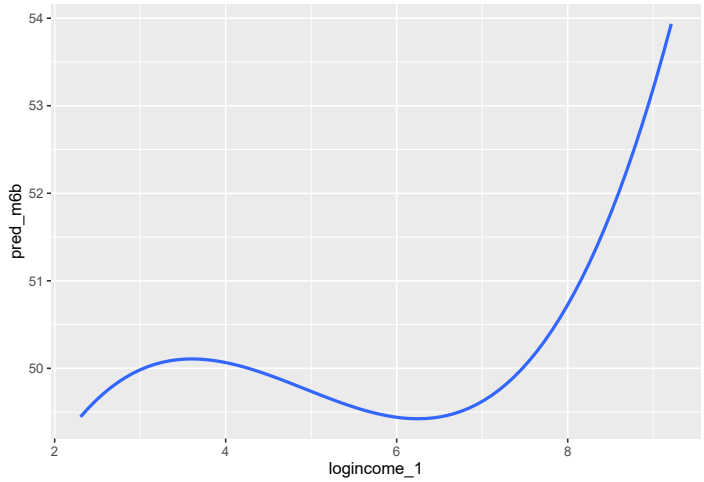
```
##
## Call:
## lm(formula = mcs_2 ~ logincome_1 + I(logincome_1^2) + I(logincome_1^3),
##     data = usw)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -49.81  -4.98   2.12   7.02  27.65
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       42.8206     1.7402   24.61  < 2e-16 ***
## logincome_1        5.0057     1.1401    4.39  1.1e-05 ***
## I(logincome_1^2)  -1.0950     0.2177   -5.03  4.9e-07 ***
## I(logincome_1^3)   0.0741     0.0128    5.78  7.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.59 on 28288 degrees of freedom
##   (22702 observations deleted due to missingness)
## Multiple R-squared:  0.00557,    Adjusted R-squared:  0.00547
## F-statistic: 52.8 on 3 and 28288 DF,  p-value: <2e-16
```

Based on these results we expect an initial positive relationship between income and mental health. This then plateaus or decreases before going up again. Let's look at the predicted scores to see how this looks like.

```
usw <- mutate(usw, pred_m6b = predict(m6b, usw))

ggplot(usw, aes(logincome_1, pred_m6b)) +
  geom_smooth(method = lm,
              formula = y ~ poly(x, 3),
              se = F)
```



An alternative way to model non-linear relationships is to convert the predictor into distinct ranges or categories and estimate separate effects for each one of these. Let's look at the relationships between age and mental health as an example:

```
ggplot(usw, aes(age, mcs_2)) +
  geom_smooth()
```

Keep in mind that the outcome varies between 0 and 100 so these fluctuations may be less important from a substantive point of view than they seem in this graph.

We see that there are quite different relationships depending on the age range. Before 25 there seems to be a negative relationship between age and mental health while between 50 and 75 there seems to be a positive relationship. Also note that the confidence interval (gray area around the line) is very large for more advanced ages because there are fewer cases in that range.

We can try to recreate this relationship by dividing the age variable into a categorical one. A useful command in this context is `cut()` which creates a factor variable with categories based on the cutoff points we give it. For simplicity sake we create a new variable with categories made of ranges of 10 years (we allow the last category to include everyone over 75).

```
usw <- mutate(usw,
              age_cat =
                cut(age, c(15, 25, 35, 45, 55, 65, 75, 101)))

count(usw, age_cat)
```

```
## # A tibble: 7 x 2
##   age_cat      n
##   <fct>    <int>
## 1 (15,25]   8049
## 2 (25,35]   8876
## 3 (35,45]   9952
## 4 (45,55]   8491
## 5 (55,65]   7255
## 6 (65,75]   5124
## 7 (75,101]  3247
```

We see that the first category includes individuals with ages between 15 and 25 (including those that are 25, "]" indicating that it includes this age). The last category includes everyone over 75.

Given that this is now a factor variable, if we included in the regression R will automatically create dummy variables for each category and make the first category (the youngest age group) the reference:

```
m7 <- lm(mcs_2 ~ age_cat,
         data = usw)

summary(m7)
```

```
##
## Call:
## lm(formula = mcs_2 ~ age_cat, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.57  -4.98   2.44   6.77  28.12
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49.329      0.157  313.81  < 2e-16 ***
## age_cat(25,35]    -0.359      0.211   -1.70    0.088 .
## age_cat(35,45]    -0.473      0.202   -2.34    0.019 *
## age_cat(45,55]    -0.182      0.206   -0.88    0.377
## age_cat(55,65]     2.224      0.210   10.58  < 2e-16 ***
## age_cat(65,75]     3.241      0.233   13.90  < 2e-16 ***
## age_cat(75,101]    2.376      0.291    8.16  3.6e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.52 on 28285 degrees of freedom
##   (22702 observations deleted due to missingness)
## Multiple R-squared:  0.0204, Adjusted R-squared:  0.0202
## F-statistic: 98.4 on 6 and 28285 DF,  p-value: <2e-16
```

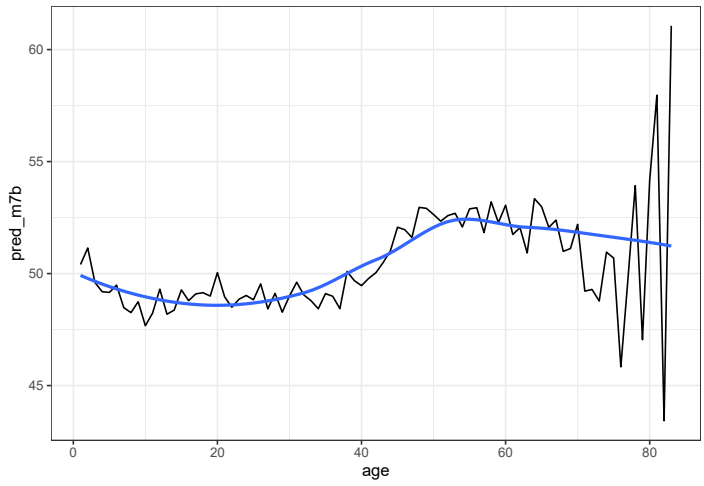
The interpretation is similar to the one for categorical predictors discussed earlier. For example, we expect a mental health score of **49.329** for those between 15 and 25 and a score of **51.553** (49.329 + 2.224) for those between 55 and 65. Note, that now we do not assume a linear relationship between age and mental health. Nevertheless we do assume that respondents within a certain age range have the same expected mental health. We could relax this assumption by making the age ranges smaller but that would result in a more complex model. At one extreme we could make a dummy for each age to estimate the expected mental health. This would result in a very complex model but with no assumptions regarding the shape of the relationship with the outcome. We can run such a model by considering age a factor in our regression.

```
m7b <- lm(mcs_2 ~ as.factor(age),
          data = usw)

summary(m7b)
```

```
##
## Call:
## lm(formula = mcs_2 ~ as.factor(age), data = usw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.34  -4.85   2.31   6.81  27.05
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         50.4060     0.4085  123.41  < 2e-16 ***
## as.factor(age)17     0.7357     0.6091    1.21  0.22714
## as.factor(age)18    -0.8065     0.6467   -1.25  0.21239
## as.factor(age)19    -1.2152     0.6546   -1.86  0.06338 .
## as.factor(age)20    -1.2493     0.6563   -1.90  0.05699 .
## as.factor(age)21    -0.9250     0.7002   -1.32  0.18652
## as.factor(age)22    -1.9265     0.6869   -2.80  0.00504 **
## as.factor(age)23    -2.1500     0.6581   -3.27  0.00109 **
## as.factor(age)24    -1.6616     0.6399   -2.60  0.00942 **
## as.factor(age)25    -2.7362     0.6511   -4.20  2.7e-05 ***
....
```

If we predict the scores based on this model we would get the following relationship:



We see that the predicted score now is allowed to change considerably from one age to another. On the one hand, this model makes fewer assumptions and may be closer to the observed data. On the other hand this model is much more complex and can be susceptible to random noise in the data. When modeling relationships we will need to

find a balance between a good representation of the data and parsimony. In this example using fewer polynomials (exemplified by the blue line) or wider age ranges might strike a better balance towards parsimony.

5.2.3 Interactions

In addition to assuming linearity, by default, the regression model assumes that the effects of the different variables do not depend on others. For example, if we investigate how having a degree and sex impact mental health we assume the effects are the same for all the cases:

```
m8 <- lm(mcs_2 ~ degree + gndr,
         data = usw)

summary(m8)
```

```
##
## Call:
## lm(formula = mcs_2 ~ degree + gndr, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48.95  -4.98   2.28   6.63  28.14
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       51.467      0.116  444.98  < 2e-16 ***
## degreeNo degree   -0.765      0.119   -6.43  1.3e-10 ***
## gndrFemale        -1.748      0.115  -15.24  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.57 on 28272 degrees of freedom
##   (22719 observations deleted due to missingness)
## Multiple R-squared:  0.00955,    Adjusted R-squared:  0.00948
## F-statistic:  136 on 2 and 28272 DF,  p-value: <2e-16
```

In this model respondents with no degree have lower mental health compared to those without a degree (by -0.765). And while we control for the effect of sex we do assume that this difference in having a degree is the same both for male and females. The same with the effect of sex. Females have lower mental health compared to males (by -1.748) while controlling for having a degree. Again, we assume that the effect of sex is not different for those with a degree compared to those without a degree. When this assumption is not met, and the effect of a variable is different for levels of another variable we say that we have a **moderated** relationship.

These types of effects can often be found in the real world as things such as interventions might have different effects for different types of people. For example, offering meals to children in elementary school may have a larger effect for children from disadvantaged backgrounds compared to those that have more affluent parents. In this case we could

say that the effect of the intervention is moderated by the socio-economic status of the pupils.

When we want to estimate such relationships using a regression we can use two main strategies. The first one is to run the regression separately for each group. For example, we can look at the effect of the intervention separately for children with different socio-economic background. Then, we can compare the effect of the interventions in the different groups.

Alternatively, we can create interactions and include them in the model. Interactions are variables that explicitly account for the different conditions in which variables can influence each other. For example, if we go back to the effect of sex and degree on mental health, we could create four different conditions: males with a degree, males without a degree, females with a degree and females without a degree. We could make a dummy variable for each of these conditions and include them in the regression. In this way we can see if the effect of having a degree is different for males and females. Here is an example how to create such variables:

```
usw %>%
  mutate(
    degree_m =
      ifelse(degree == "Degree" & gndr == "Male", 1, 0),
    nodegree_m =
      ifelse(degree == "No degree" & gndr == "Male", 1, 0),
    degree_f =
      ifelse(degree == "Degree" & gndr == "Female", 1, 0),
    nodegree_f =
      ifelse(degree == "No degree" & gndr == "Female", 1, 0)
  ) %>%
  count(degree, gndr, degree_m, nodegree_m, degree_f, nodegree_f)
```

```
## # A tibble: 6 x 7
##   degree    gndr   degree_m nodegree_m degree_f nodegree_f     n
##   <fct>     <fct>     <dbl>      <dbl>    <dbl>      <dbl> <int>
## 1 Degree    Male          1          0        0          0  7559
## 2 Degree    Female        0          0        1          0  8932
## 3 No degree Male          0          1        0          0 15593
## 4 No degree Female        0          0        0          1 18818
## 5 <NA>      Male         NA         NA        0          0    50
## 6 <NA>      Female        0          0       NA         NA    42
```

We can also create interactions directly in the regression using : (colon). Here we run a model where we include the interaction between “degree” and “gndr”.

```
m9 <- lm(mcs_2 ~ degree:gndr,
         data = usw)

summary(m9)
```

```
##
## Call:
## lm(formula = mcs_2 ~ degree:gndr, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48.85  -4.95   2.29   6.60  28.24
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  48.8531     0.0944  517.48  < 2e-16 ***
## degreeDegree:gndrMale         2.3756     0.1733   13.71  < 2e-16 ***
## degreeNo degree:gndrMale      1.9777     0.1426   13.87  < 2e-16 ***
## degreeDegree:gndrFemale       1.0479     0.1583    6.62  3.7e-11 ***
## degreeNo degree:gndrFemale       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.57 on 28271 degrees of freedom
##   (22719 observations deleted due to missingness)
## Multiple R-squared:  0.00981,    Adjusted R-squared:  0.0097
## F-statistic: 93.4 on 3 and 28271 DF,  p-value: <2e-16
```

In this regression the cases that have no degree and are female are used as the reference (no effects are estimated for them). The expected mental health in wave 2 for this group is equal to the intercept (**48.853**). The three slopes say how each particular group is different to this reference. For example, those that have a degree and are male have **2.376** higher mental health compared to the reference. When using interaction the interpretation can be a little tricky. Because of that, my recommendation is to write down the formula and see what happens under different conditions. The regression formula based on this model (ignoring the residual) is:

$$mcs\_2 = 48.853 + 2.376 * Degree * Male + 1.978 * Nodegree * Male + 1.048 * Degree * Female$$

Now we can calculate the expected value for different combinations. For example, if we want to calculate the expected value for females who have no degree we can replace with 1 when these categories are present in the formula and with 0 when they are not. This will result in the following equation:

$$mcs\_2_{No\_degree\_Female} = 48.853 + 2.376 * 0 * 0 + 1.978 * 1 * 0 + 1.048 * 0 * 1 = 48.853$$

So the expected mental health in wave 2 for females with no degrees is **48.853**.

We can do the same for the other conditions:

$$mcs\_2_{Degree\_Female} = 48.853 + 2.376 * 1 * 0 + 1.978 * 0 * 0 + 1.048 * 1 * 1 = 48.853 + 1.048 = 49.901$$

$$mcs\_2_{No\_degree\_Male} = 48.853 + 2.376 * 0 * 1 + 1.978 * 1 * 1 + 1.048 * 0 * 0 = 48.853 + 1.978 * 1 * 1 = 50.831$$

$$mcs\_2_{Degree\_Male} = 48.853 + 2.376 * 1 * 1 + 1.978 * 0 * 1 + 1.048 * 1 * 0 = 48.853 + 2.376 = 51.229$$

It appears that males with a degree have the highest mental health in wave 2 and females with no degree have the lowest values.

We can also predict the expected values based on our model and then calculate the average predicted score for the four groups:

```
usw <- mutate(usw, pred_m8 = predict(m8, usw))

usw %>%
  group_by(degree, gndr) %>%
  summarise(pred_m8 = mean(pred_m8, na.rm = T))
```

```
## # A tibble: 6 x 3
## # Groups:   degree [3]
##   degree    gndr   pred_m8
##   <fct>     <fct>    <dbl>
## 1 Degree    Male      51.5
## 2 Degree    Female    49.7
## 3 No degree Male      50.7
## 4 No degree Female    49.0
## 5 <NA>      Male     NaN
## 6 <NA>      Female   NaN
```

These results are consistent with our findings from the equations.

In this example, we explored the interaction between two categorical variables. We can also have interactions between a categorical variable and a continuous one or between two continuous variables. In these situations we can calculate the interactions by multiplying the variables of interest.

Let's imagine we want to explore if the effect of age on mental health is moderated by having a degree. Maybe we expect that ageing has less of an impact on mental health for those with a degree compared to those without one. We can create the interaction by multiplying age with degree. Because degree is a factor we first need to convert degree to a numeric value. We also subtract the value 1 as it is coded by default as 1 and 2.

```
usw %>%
  mutate(age_nodegree = age*(as.numeric(degree) - 1)) %>%
  select(age, degree, age_nodegree) %>%
  head()
```

```
## # A tibble: 6 x 3
##   age       degree    age_nodegree
##   <dbl+lbl> <fct>            <dbl>
## 1 39        No degree           39
## 2 59        Degree               0
## 3 39        No degree           39
## 4 17        Degree               0
## 5 72        No degree           72
## 6 57        Degree               0
```

We see that the new variables gets the value 0 for those with a degree (because age is multiplied by 0) and the age value for those without a degree. If we include this new variable in the model it will indicate how the effect of age on the outcome is different for those without a degree compared to those with a degree.

When we run the regression we can include the variable we created or use : (colon) to make the interactions directly in the regression. To make the intercept easier to interpret we use here "age_center" where 0 is the average age.

```
m10 <- lm(mcs_2 ~ degree + age_center + degree:age_center,
          data = usw)

summary(m10)
```

```
##
## Call:
## lm(formula = mcs_2 ~ degree + age_center + degree:age_center,
##     data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.74  -4.94   2.27   6.81  28.06
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 50.51298    0.09540  529.50  < 2e-16 ***
## degreeNo degree             -0.89385    0.11884   -7.52  5.6e-14 ***
## age_center                   0.09291    0.00647   14.36  < 2e-16 ***
## degreeNo degree:age_center -0.03777    0.00749   -5.04  4.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.54 on 28271 degrees of freedom
##   (22719 observations deleted due to missingness)
## Multiple R-squared:  0.016,  Adjusted R-squared:  0.0159
## F-statistic:  153 on 3 and 28271 DF,  p-value: <2e-16
```

To facilitate the interpretation, again I recommend to write down the equation based on this regression:

$$mcs\_2 = 50.513 - 0.894 * Nodegree + 0.093 * age\_center - 0.038 * Nodegree * age\_center$$

If we wanted to estimate the expected mental health for respondents with a degree and average age we would get:

$$mcs\_2_{average\_age\_Degree} = 50.513 - 0.894 * 0 + 0.093 * 0 - 0.038 * 0 * 0 = 50.513$$

While the expected value for those without a degree and with an average age is:

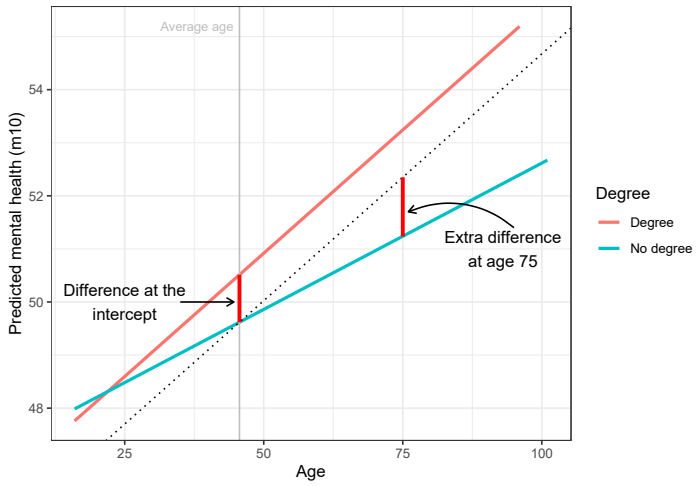
$$mcs\_2_{average\_age\_No\_degree} = 50.513 - 0.894 * 1 + 0.093 * 0 - 0.038 * 1 * 0 = 50.513 - 0.894 = 49.619$$

So the main effect of the degree tells us what is the difference between those with a degree and those without a degree in mental health when "age_center" is 0. Let's see how this difference would look like when age is higher. For example, what would be the expected mental health when age is 75? We know that average age is around 45. That means if we add 30 years to "age_center" we would get the target age. Here are the two equations in this condition:

$$mcs\_2_{age=30\_No\_degree} = 50.513 - 0.894 * 1 + 0.093 * 30 - 0.038 * 1 * 30 = 50.513 - 0.894 + 0.093 * 30 - 0.038 * 30 = 51.269$$

$$mcs\_2_{age=30\_Degree} = 50.513 - 0.894 * 0 + 0.093 * 30 - 0.038 * 0 * 30 = 50.513 + 0.093 * 30 = 53.303$$

So we see that the difference between those with a degree and those without a degree is larger at age 75 compared to that observed at average age. As age increases respondents with a degree have a higher mental health advantage compared to those without a degree. We can visualize this relationship using the predicted scores from our model:



5.3 Further Reading

For a general introduction to regression modeling and statistical inference I recommend Agresti (2018). For a more in depth introduction to modeling categorical outcomes I recommend Agresti (2007).