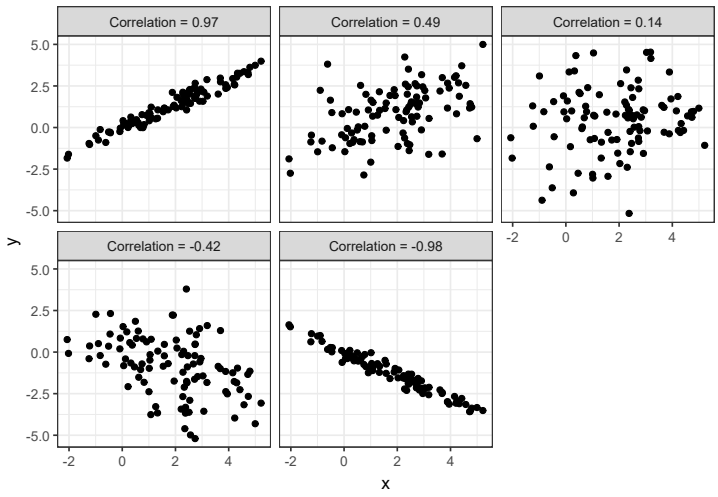
Chapter 5

Introduction to Regression Models

It's important to have a good grasp of regression modelling to analyze longitudinal data. This technique forms the foundation of many statistical models used to understand longitudinal data. We'll begin by exploring correlations and basic regression models and then progress to more advanced topics like interactions and non-linear relationships.

5.1 Correlation and Regression

Probably one of the most intuitive ways 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 standardised to 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.



In these graphs, we see that for positive correlations, an increase in x is associated with an increase in y, while for negative correlations, it's associated with a decrease. We also observe 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 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 covariance estimate. If x and y vary together, this will be a large number; otherwise, it will be small. The covariance has no scale (it can be any number). 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 to analyse some real data. Looking at the previously prepared data, we can explore the relationship between mental health in wave 1 and wave 2 using a correlation.

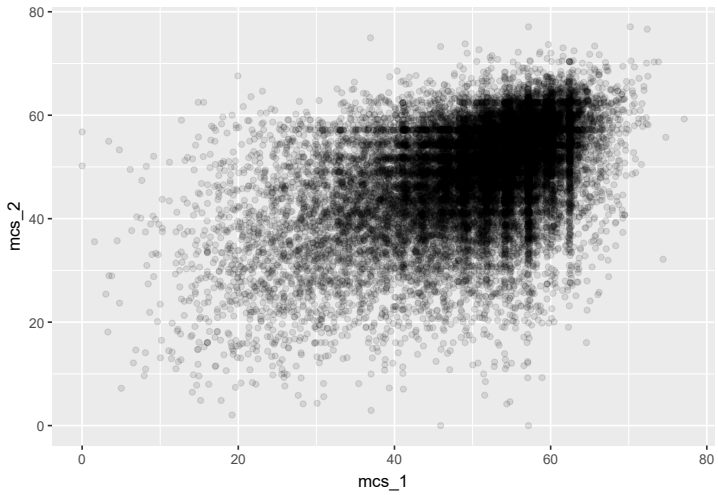
```
cor(usw$mcs_1, usw$mcs_2, use = "complete.obs")
```

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

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

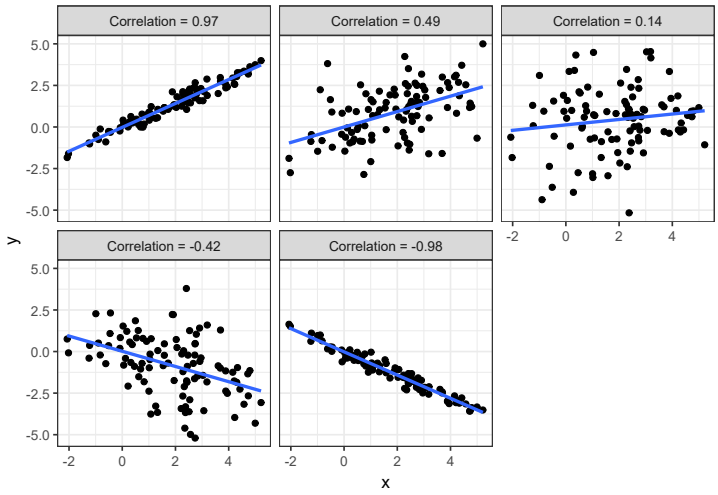
We see a positive, moderate-sized correlation. This implies that people who have good mental health in wave 1 also have good mental health in wave 2. We can also think of this as "stability," indicating that mental health is fairly stable over time. We can also explore the relationship using a scatter 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" 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, suggesting some change in time for mental health.

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



A regression also represents the relationship between variables but in a different way. Typically, the regression is described 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 our model. This value varies by individual (i)
- α (alpha), also known as the **intercept/constant**, is the expected value when the predictor is 0

- β (beta) is also known as the **slope** and tells us the expected change in the dependent variable (y) with an increase of 1 in the independent variable (x)
- σ (sigma) is also know as **the residual**. This represents the unexplained variance of y and summarises 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 (which we add on the right side). We also indicate what 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)
```

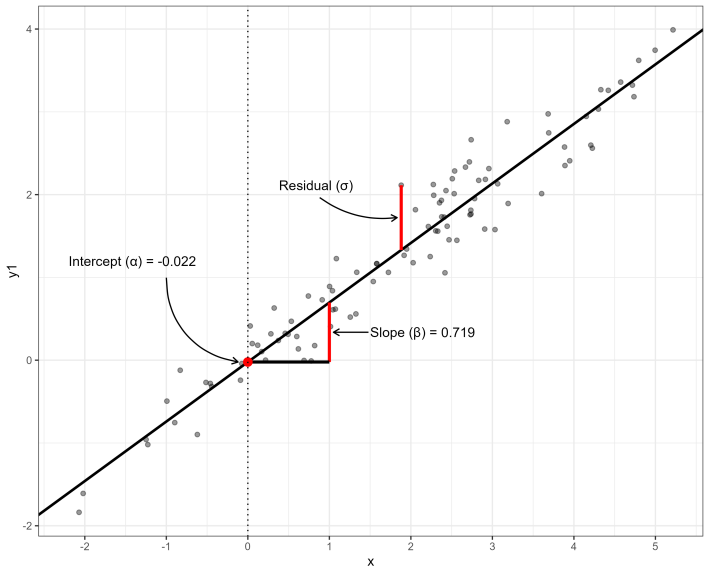
```
##
## Call:
## lm(formula = y1 ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6610 -0.1808 -0.0033  0.1857  0.7829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0217     0.0456   -0.48     0.63
## x             0.7189     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* is 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, 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:

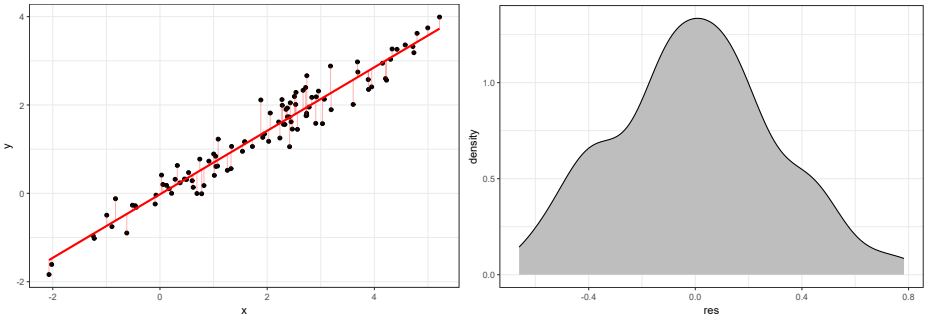
$$y1 = -0.022 + 0.719 * x + 0.303$$

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



We can see how the regression line represents the relationship between x and *y1*. More precisely, the slope indicates the angle of this line. It shows how the value *y1* increases by **0.719** when x increases by 1. We also see that the intercept is 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, the residual indicates the distance between the observed scores (the dots in the scatterplot) and the predicted value (the line).

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



The left graph 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 this variable is.

```r
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") increases 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 suggested by the correlation). Through the intercept, we see that for respondents with 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 standardised, it can be hard to interpret. The R-squared is an alternative indicator we can use to understand how well our predictors explain the outcome. It indicates what proportion of the total variation of the dependent variable is explained by our model. This can range from 0, where we explain no variation, to 1, where 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. Often, we want to make inferences about the general population. For example, suppose we use a sample of the general population, as we are doing when using the Understanding Society data. In that case, 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, and p values come in. The **standard error** estimates the amount of uncertainty we have around the coefficient in the population. 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 to determine whether the coefficient 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 hypothesis for p-values below that. While this may be useful in some conditions, the p-value is a contentious topic in

statistics, given its 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 we get:

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

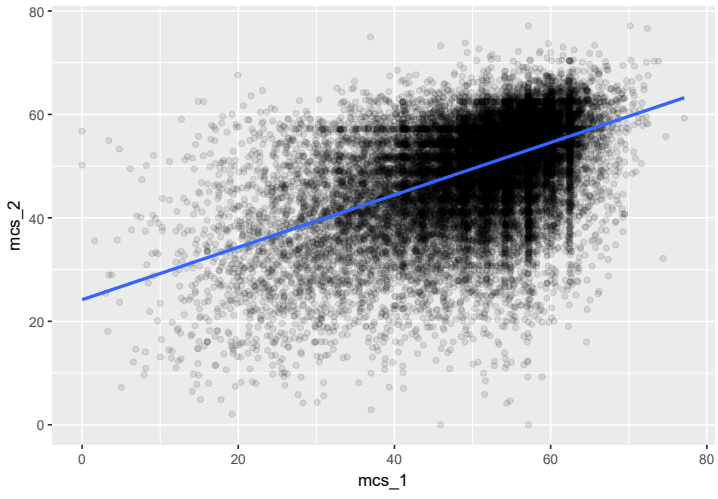
This would indicate that we expect the relationship between mental health in wave 1 and mental health in wave 2 to be somewhere between **0.496** and **0.516** in the population. Again, this comes with some assumptions. For example, because we use the 95% confidence interval, it means that 5% of the time, 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 visualise 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 several ways. One way to expand it is by using multiple predictors. We rarely expect one variable to

explain another perfectly. As such, typically, we want to include different predictors in our regression models. For example, we might expand the previous model explaining mental health in wave 2 by 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 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 modelling that makes 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 add it to 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, our model is slightly better now, with the R-squared being larger than before (0.263 vs. 0.259). It appears 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 predicted 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 with a score of 0 for mental health in wave 1 and age 0. While this might be mathematically sound, it does not make much 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.

Recoding the age variable can make the intercept more useful. In chapter 3, we created a centred 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 with 0 on the mental health score in wave 1 and **average age**. As you can see, the interpretation is much more interesting now. If we wanted, we could also centre mental health in wave 1, so the intercept would refer to the expected value for those with average mental health in wave 1. Remember that the intercept can be a valuable tool to understand your data, but you must be careful how you code the predictors.

5.2.1 Categorical Predictors

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

Let's see how education impacts mental health. We have coded education differently, but let's use the degree variable. 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 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 variable 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 from the reference (in this case, "Degree"). So, the interpretation here would be that people who do not have a degree have, on average, lower mental health than those who 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 want 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 down the regression model to see 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 who have a degree versus the rest. We could do this in a few different ways. We could change the order of the factor's levels. 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 who have a degree to those who don't. The effect size is the same, just the direction is different. Secondly, the intercept value changed (24.397 vs. 24.118). Because the reference category has changed, the intercept has 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. For example, 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 categories. 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_married     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
```

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)
```

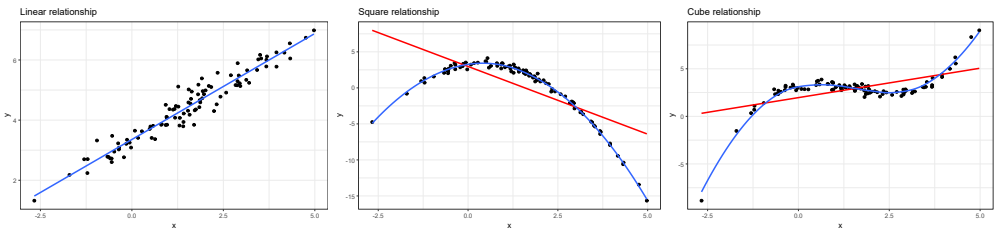
```
##
## Call:
## lm(formula = mcs_2 ~ mcs_1 + marstatus_fct_1, data = usw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.48  -4.12   1.55   5.18  32.90
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     24.88195    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 now have five variables 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 who are married** (the reference category). The difference in expected mental health is -1.06. Similarly, people who were never married have lower mental health **compared to those who are married** of -1.013.

5.2.2 Non-linear Relationships

So far, we have assumed a linear relationship between predictors and the outcome in the regression. This means that an increase in 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. However, the effect might flatten out or 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 modelled 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. In contrast, 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 affecting 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 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 impact 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 income (the logged version saved as "logincome_1") and the square effect. This could be a variable we save in advance or 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
## -49.84  -4.98   2.14   7.03  27.58
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        100.241     10.787    9.29  < 2e-16 ***
## logincome_1        -14.165      2.764   -5.12  3.0e-07 ***
## I(logincome_1^2)     0.988      0.177    5.59  2.3e-08 ***
## ---
## 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:  0.00531,    Adjusted R-squared:  0.00523
## F-statistic: 75.4 on 2 and 28289 DF,  p-value: <2e-16
```

Looking at the results, we see that the main effect of income is negative (-14.165) while the effect of the squared income is positive (0.988). To understand how a non-linear relationship is created by including the polynomials, we can write down the equation and work out two scenarios:

$$mcs\_2 = 100.241 - 14.165 * logincome\_1 + 0.988 * logincome\_1^2$$

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

$$mcs\_2_{logincome=2} = 100.241 - 14.165 * 2 + 0.988 * 2^2 = 100.241 - 28.33 + 3.952 = 75.863$$

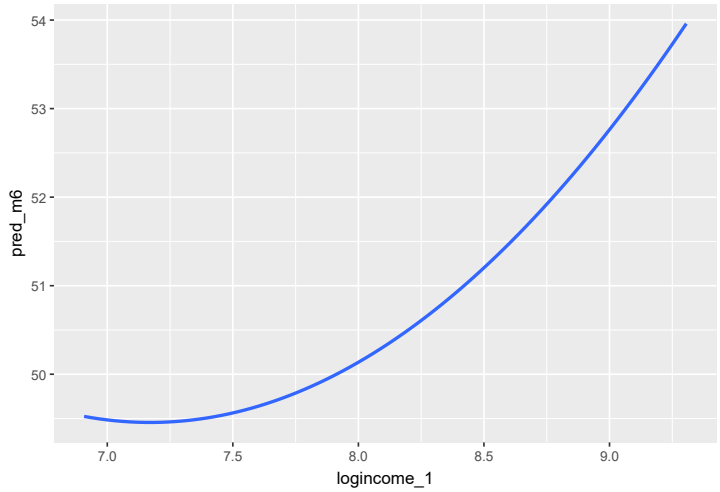
$$mcs\_2_{logincome=8} = 100.241 - 14.165 * 8 + 0.988 * 8^2 = 100.241 - 113.32 + 63.232 = 50.153$$

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 logincome has a negative effect on mental health initially, but this becomes positive for larger values of logincome.

Based on our model, we can also use the predicted values to see the expected relationship. Below, we save the predicted score and 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 downward. The larger the coefficient, the stronger the change.

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

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

summary(m6b)
```

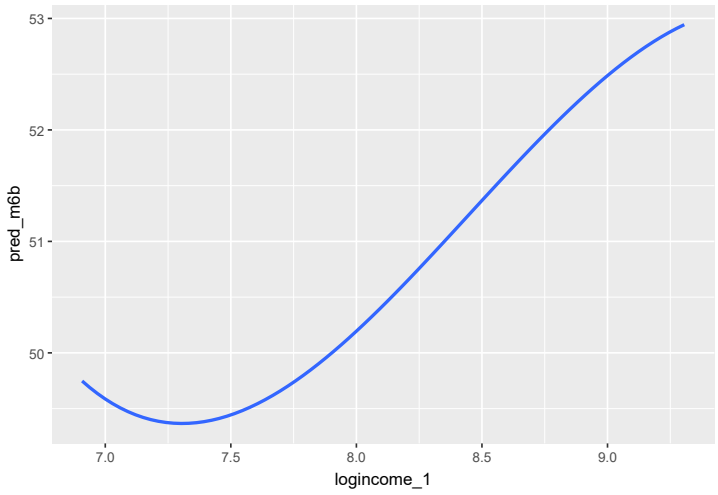
```
##
## Call:
## lm(formula = mcs_2 ~ logincome_1 + I(logincome_1^2) + I(logincome_1^3),
##     data = usw)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -49.81  -4.98   2.13   7.00  27.70
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        410.185    131.641    3.12   0.0018 **
## logincome_1       -132.422     50.134   -2.64   0.0083 **
## I(logincome_1^2)    15.971      6.345    2.52   0.0118 *
## I(logincome_1^3)    -0.630      0.267   -2.36   0.0182 *
## ---
## 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:  0.0055, Adjusted R-squared:  0.0054
## F-statistic: 52.2 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 relationship then plateaus or decreases before going up again. Let's look at the predicted scores to see how this looks.

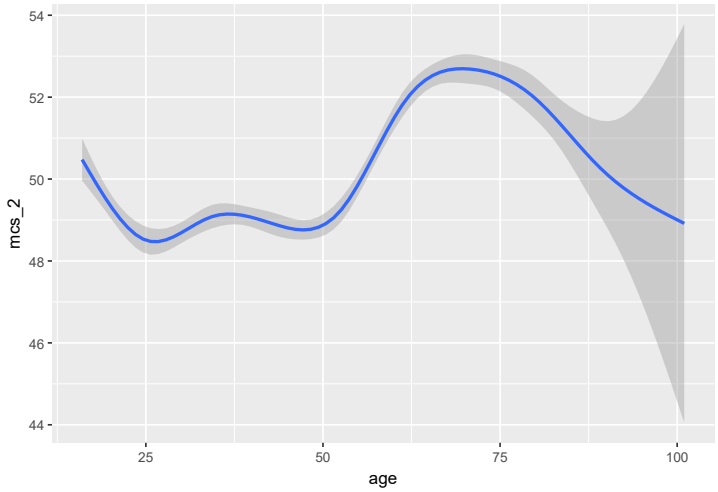
```
usw <- mutate(usw, pred_m6b = predict(m6b, usw))

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. Let's look at the relationships between age and mental health as an example:

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



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

We see 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 (the grey area around the line) is considerable for more advanced ages because there are fewer cases in that range.

We can 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, 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 between 15 and 25 (including those who are 25; "]" indicates that it includes this age). The last category includes everyone over 75.

Given that this is now a factor variable, if we include it 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
```

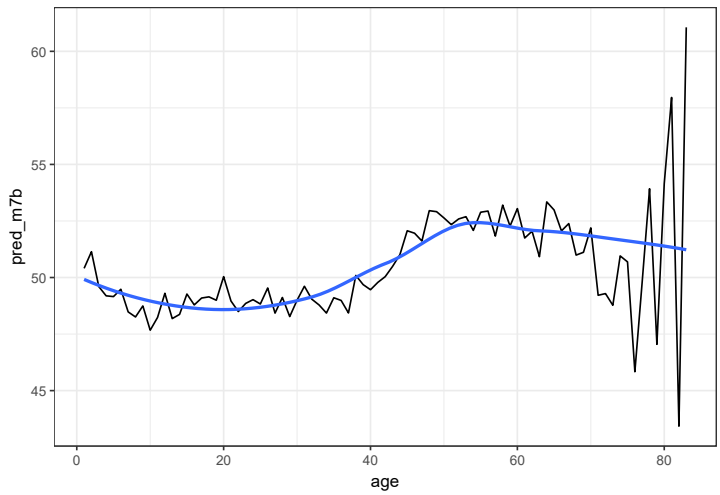
The interpretation is similar to the one for categorical predictors discussed earlier. For example, we expect a mental health score of **49.329** for those between 15 and 25 and a score of **51.553** (49.329 + 2.224) for those between 55 and 65. Note that we do not assume a linear relationship between age and mental health. Nevertheless, we assume respondents within a specific age range have the same expected mental health. We could relax this assumption by making the age ranges smaller, resulting 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 as 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:



The predicted score can 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, it is much more complex and can be susceptible to random noise in the data. When modelling 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). While we control for the effect of sex, we assume that this difference in having a degree is the same for males and females. The same is true for the impact 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 impact of a variable is different for levels of another variable, we say that we have a **moderated** relationship.

These effects can often be found in the real world, as interventions might have different effects on different types of people. For example, offering meals to elementary school children may have a larger effect on children from disadvantaged backgrounds than

on those with 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 impact 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 condition 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 of 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 including 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). This group's expected mental health in wave 2 equals the intercept (**48.853**). The three slopes show how each particular group is different from this reference. For example, those with a degree and are male have higher mental health than the reference (by **2.376**). 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 with no degree, we can replace it with 1 when these categories are present in the formula and 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:

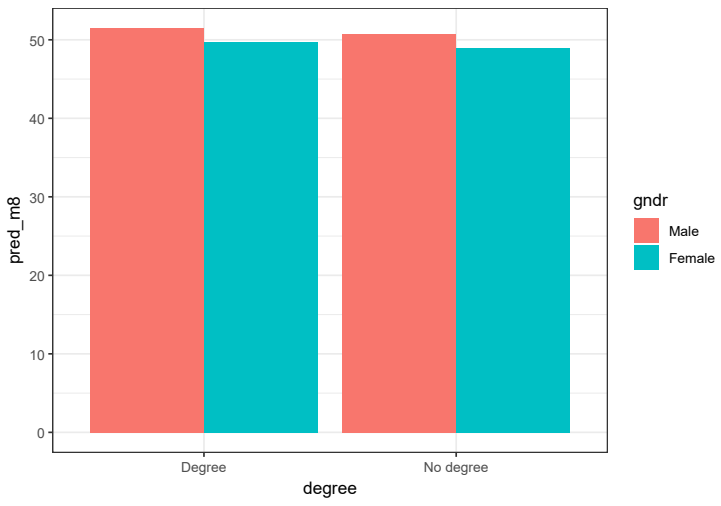
```
usw <- mutate(usw, pred_m8 = predict(m8, usw))

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
```

We could also make the results easier to present by using a graph:

```
usw |>
  group_by(degree, gndr) |>
  summarise(pred_m8 = mean(pred_m8, na.rm = T)) |>
  na.omit() |>
  ggplot(aes(degree, pred_m8, fill = gndr)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_bw()
```



These results are consistent with our findings from the equations.

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

Let's imagine we want to explore whether 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
```

The new variable gets 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 "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 writing 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 the following:

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

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

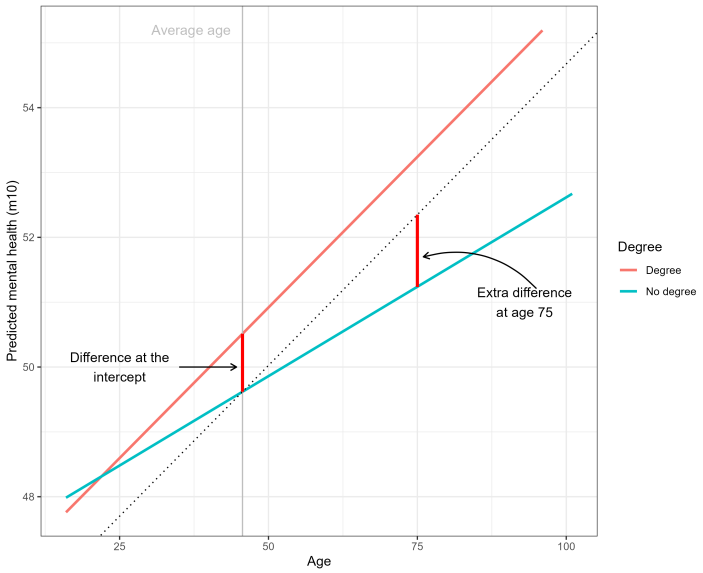
$$mcs\_2_{average\_age\_No\_degree} = 50.513 - 0.894 * 1 + 0.093 * 0 - 0.038 * 1 * 0 = 50.513 - 0.894 = 49.619$$

So, the degree's main effect tells us 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 the 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 for these conditions:

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

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

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 visualise this relationship using the predicted scores from our model:



5.3 Introduction to Generalized Linear Models (GLM)

So far, we have discussed how to model continuous outcomes. Nevertheless, we often need to model different types of variables, such as dichotomous, ordinal, nominal, and so on. Similarly, sometimes outcomes may not be normally distributed, but instead, they may be skewed, have a floor or ceiling effect, or have a different shape. We need to expand the regression model we discussed so far to analyse such variables. This is where the Generalised Linear Model framework comes in.

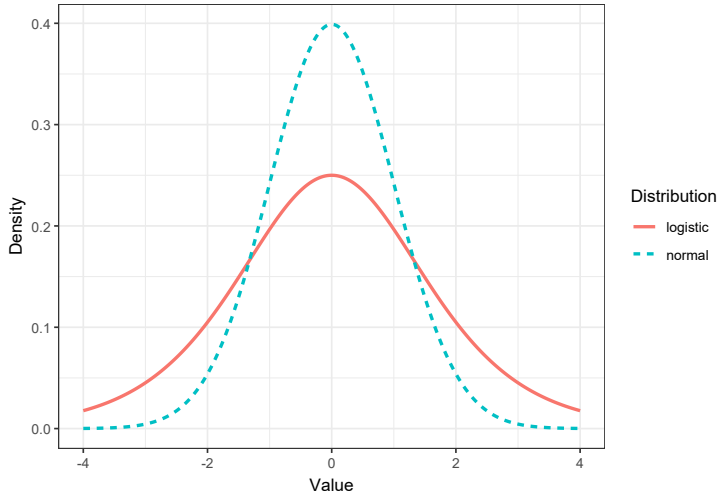
Generalised Linear Models (GLMs) are a flexible extension of ordinary linear regression that allow for the response variables to have error distributions other than a normal distribution. GLMs are used to model the relationship between a dependent variable and one or more independent variables, particularly when the dependent variable follows a distribution from the exponential family, such as normal, binomial, or Poisson distributions. GLMs consist of three key components: **the random component**, which specifies the distribution of the response variable; the **systematic component**, which is a linear predictor combining the predictors; and the **link function**, which connects the mean of the response variable to the linear predictor. The link function transforms the expected value of the response variable to the linear predictor scale.

For example, in the case of Ordinary Least Squares (OLS) regression, the random component specifies that the response variable Y follows a normal distribution with mean μ and variance σ^2; for example, house prices can be modelled as $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$. The systematic component combines the predictors in a linear predictor, such as $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$, where house prices might depend on the size and number of bedrooms. The link function for OLS is the identity link, meaning the expected value of the response variable is directly equal to the linear predictor, $g(\mu_i) = \mu_i = \eta_i$, so the predicted house price is the output of the linear model.

Here, we will focus on logistic and probit models for dichotomous outcomes, but the approaches we will discuss can also be extended to other types of GLMs. The probit and logit models are both types of regression used for binary outcome variables, but they

differ in the link function they use. The logit model uses the logistic function, which assumes that the underlying distribution of the error terms is logistic. This results in an S-shaped curve that maps any real-valued number into the (0, 1) interval, making it suitable for modelling probabilities. The logistic distribution has heavier tails than the normal distribution, predicting more extreme values.

In contrast, the probit model uses the cumulative distribution function (CDF) of the standard normal distribution as its link function. This model assumes that the underlying distribution of the error terms is normal, leading to a similar S-shaped curve but with slightly different characteristics. The normal distribution is symmetric and has lighter tails than the logistic distribution, often resulting in slightly different predictions, particularly in the distribution's tails. Here are the hypothetical distributions for the logistic and normal (probit) distributions:



When choosing between probit and logit models, it's important to consider the specific context of the data and the theoretical implications. While both models generally yield similar results, the logit model's practicality and interpretability make it the more commonly used option in real-world applications.

5.3.1 Logistic Regression

Logistic regression is a type of GLM used when the dependent variable is binary. It models the probability that a given outcome occurs rather than directly predicting the outcome itself.

For logistic regression, the random component specifies that the response variable Y follows a binomial distribution appropriate for binary outcomes like disease presence (Yes/No), expressed as $Y_i \sim \text{Binomial}(n_i, p_i)$. The systematic component remains a linear predictor combining the predictors, $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$, for instance, using age and cholesterol level to predict heart disease probability. The link function in logistic regression is the logit link, which transforms the probability into the log-odds scale, $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$. This means the log-odds of having a heart disease is modeled as a linear function of the predictors.

So, when doing a logistic regression, the probability of the event happening is first transformed into odds and then into log odds. Odds take any value between 0 and infinity

and represent the ratio of the probability of the event happening to the probability of the event not happening. Log odds are the natural logarithm of the odds and can take any real value. The logit link function maps the probability of the event occurring to the log-odds scale, allowing for a linear relationship between the predictors and the log-odds of the event. Here are some typical values in the three different scales so you can get an intuition of their range:

Probability	Odds	Log_Odds
0.10	0.11	-2.2
0.25	0.33	-1.1
0.50	1.00	0.0
0.75	3.00	1.1
0.90	9.00	2.2

Given what we have covered so far, we can write the equation for the logistic regression model as:

$$\text{logit}(P(Y=1|X)) = \log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Where:

- $P(Y=1|X)$ is the probability of the event occurring.
- β_0 is the intercept.
- $\beta_1, \beta_2, ..., \beta_k$ are the coefficients for the predictors $X_1, X_2, ..., X_k$.

To better understand how this works, let's consider a model using our data. Suppose we want to predict whether an individual is single or not in wave 2 using age, having a degree, and living in an urban area. We can fit a logistic regression model using the `glm()` function in R with the family set to `binomial`. This will run a logistic model.

```
logit_model <- glm(single_2 ~ age_center + urban_1 + degree,
                   data = usw,
                   family = binomial)

summary(logit_model)
```

```
##
## Call:
## glm(formula = single_2 ~ age_center + urban_1 + degree, family = binomial,
##     data = usw)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.662167   0.038347  -17.27   <2e-16 ***
## age_center      -0.016840   0.000622  -27.09   <2e-16 ***
## urban_1         -0.260484   0.027092   -9.61   <2e-16 ***
## degreeNo degree  0.589410   0.023912   24.65   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50039  on 38322  degrees of freedom
## Residual deviance: 48578  on 38319  degrees of freedom
##   (12671 observations deleted due to missingness)
## AIC: 48586
##
## Number of Fisher Scoring iterations: 4
```

We can write the result as an equation:

$$\text{logit}(P(Y = 1|X)) = -0.66 - 0.02 * \text{age\_center} - 0.26 * \text{urban\_1} + 0.59 * \text{degreeNo\_degree}$$

The logistic regression model's coefficients indicate the effects of different predictors on the likelihood of being single in wave 2 of the study. Overall, the interpretation is the same as in linear regression, but the scale now is the log odds. For example, the intercept (-0.66) represents the log odds of being single for people of average age who live in rural areas and have a degree. Age has a small negative effect, meaning that when age increases by 1, the likelihood of being single decreases by 0.016 log-odds. People living in urban areas have 0.26 lower log odds of being single than those in urban areas. Conversely, individuals without a degree are likelier to be single, with a positive coefficient of 0.59. All these predictors are highly statistically significant (p-values $<$ 2e-16), indicating strong evidence that they are associated with the likelihood of being single in wave 2.

We can convert the log odds to odds and probabilities to make the coefficient easier to interpret. For example, to convert the log odds of the intercept (-0.66) to a probability, we first exponentiate the number. Specifically, we calculate $e^{-0.66}$, which is approximately 0.516. This means the odds of the event occurring are 0.516 to 1. To convert these odds to a probability, we can use the formula $P = \frac{\text{odds}}{1+\text{odds}}$. Substituting the calculated odds, we get $P = \frac{\text{odds}}{1+\text{odds}} \approx 0.34$. Therefore, a log-odds of -0.66 corresponds to a probability of approximately 34%. This implies that the likelihood of being single in wave 2 is around 34% for people with average age, no degree and who lived in an urban area in wave 1.

We can exponentiate all the coefficients to odds ratios to make them easier to interpret:

```
exp(coef(logit_model))
```

```
##     (Intercept)      age_center         urban_1 degreeNo degree
##          0.5157          0.9833          0.7707          1.8029
```

The odds of 1 imply no difference between the groups; values above mean increased chances of the event happening, while below 1 imply the opposite. For example, based on the output above, each one-unit increase in age slightly decreases the odds of being single by approximately 1.67% (1 - 0.983). Living in an urban area decreases the odds of being single by around 23% (1-0.77). Conversely, individuals without a degree are around 80% more likely to be single than those with a degree.

One strategy to make this easier to interpret is to think of hypothetical individuals with different values on the predictors and calculate their model-implied probabilities. For example, we can calculate the predicted probability of being single in wave 2 for a person of average age living in an urban area and without a degree. We can also calculate the expected probability for a person ten years above average age, living in a rural area and with a degree. Finally, we can look at an individual who is five years below average age, in an urban area and has no degree. We can put these hypothetical cases in a dataset:

```
new_data <- data.frame(
  age_center = c(0, 10, -5),
  urban_1 = c(1, 0, 1),
  degree = c("No degree", "Degree", "No degree")
)
```

We can then predict the probabilities for these cases using the logistic regression model and visualise the results:

```
new_data <- new_data %>%
  mutate(predicted_probabilities = predict(logit_model,
                                           newdata = new_data,
                                           type = "response"))

new_data
```

```
##   age_center urban_1    degree predicted_probabilities
## 1          0       1 No degree                 0.4175
## 2         10       0    Degree                 0.3035
## 3         -5       1 No degree                 0.4381
```

The second individual has the lowest probability of being single (30%), while the last one has the highest one (43%).

Alternatively, we can calculate the marginal effects. Marginal effects represent the change in the probability of the outcome for a one-unit change in the predictor, holding all other predictors constant. In a logistic regression, the relationship between predictor variables and the probability of the outcome is not linear, making it difficult to interpret coefficients directly. The `margins` package addresses this by averaging the predicted changes across all observations in the dataset, providing a more intuitive understanding of the predictor's impact on the outcome.

```
library(margins)

marginal_effects <- margins(logit_model)

summary(marginal_effects)
```

```
##          factor     AME     SE       z      p   lower   upper
```

```
##       age_center -0.0037 0.0001 -28.0886 0.0000 -0.0040 -0.0035
##  degreeNo degree  0.1277 0.0049  25.8204 0.0000  0.1180  0.1373
##          urban_1 -0.0577 0.0060  -9.6550 0.0000 -0.0694 -0.0460
```

Average Marginal Effects (AME) quantify the average change in the predicted probability of an outcome for a one-unit change in a predictor variable, holding all other variables constant. AMEs are particularly useful in logistic regression models because they provide an intuitive measure of the impact of predictor variables on the probability of the outcome despite the non-linear nature of these models. Positive AMEs indicate an increase in the likelihood of the outcome as the predictor variable increases, while negative AMEs indicate a decrease in the probability of the outcome.

The AME coefficients from the table provide insights into how each predictor variable affects the probability of being single in wave 2. For "age_center", the AME of -0.0037 suggests that for each unit increase in age, the likelihood of being single decreases by approximately 0.37%. The AME for "No degree" is 0.1277, meaning that individuals without a degree have a 12.77% higher probability of being single than those with a degree. Lastly, the AME for "urban_1" is -0.0577, indicating that living in an urban area decreases the probability of being single by 5.77% compared to living in a non-urban area.

5.3.2 Probit Regression

Probit regression is another type of GLM used for binary outcomes. For probit regression, the random and the systematic components are the same as for a logistic regression. The only difference is the link function used. The link function in probit regression is the probit link, which transforms the probability using the cumulative distribution function (CDF) of the standard normal distribution, $g(\mu_i) = \Phi^{-1}(\mu_i) = \eta_i$. This means the probability of having a heart disease is modelled using the inverse of the standard normal CDF, connecting it to a linear function of the predictors.

The probit model can be written as:

$$\text{probit}(P(Y = 1|X)) = \Phi^{-1}(P(Y = 1|X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Where:

- $P(Y = 1|X)$ is the probability of the event occurring.
- Φ^{-1} is the inverse cumulative distribution function (CDF) of the standard normal distribution.
- β_0 is the intercept.
- $\beta_1, \beta_2, ..., \beta_k$ are the coefficients for the predictors $X_1, X_2, ..., X_k$.

Using the same dataset, we can fit a probit regression model with the `glm()` function in R by setting the family to `binomial(link = "probit")`.

```
probit_model <- glm(single_2 ~ age_center + urban_1 + degree,
                    data = usw,
                    family = binomial(link = "probit"))

summary(probit_model)
```

```
##
## Call:
## glm(formula = single_2 ~ age_center + urban_1 + degree, family = binomial(link = "probit")
##     data = usw)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.404255   0.023201  -17.42   <2e-16 ***
## age_center      -0.009828   0.000378  -26.02   <2e-16 ***
## urban_1         -0.158844   0.016353   -9.71   <2e-16 ***
## degreeNo degree  0.356718   0.014415   24.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50039  on 38322  degrees of freedom
## Residual deviance: 48618  on 38319  degrees of freedom
##   (12671 observations deleted due to missingness)
## AIC: 48626
##
## Number of Fisher Scoring iterations: 5
```

We can write up the formula for the probit model as:

$$\Phi^{-1}(P(Y = 1|X)) = -0.40 - 0.01 * \text{age\_center} - 0.16 * \text{urban\_1} + 0.36 * \text{degreeNo\_degree}$$

The probit model results indicate that age has a small negative effect (-0.0098), implying that the probability of being single decreases as age increases. Living in an urban area (-0.158) also reduces the likelihood of being single compared to living in a rural area. Conversely, individuals without a degree (0.356) are more likely to be single than those with a degree.

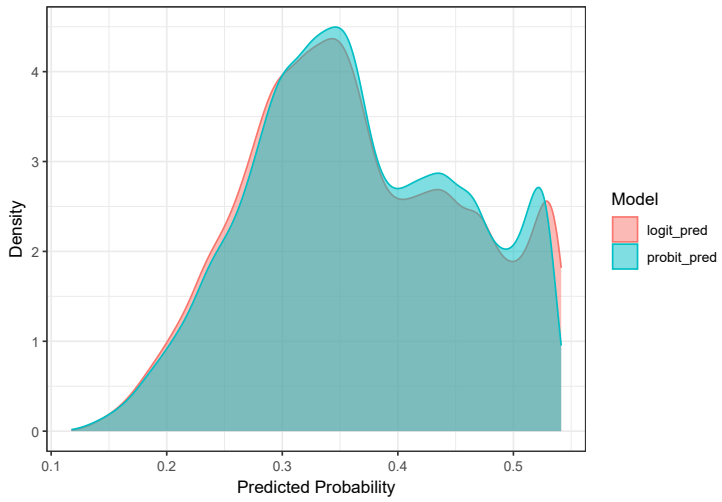
We can transform the coefficients into probabilities to facilitate the interpretation of the results (no odds interpretation is possible in this type of model). As we did before, we could create hypothetical cases and predict the probabilities or calculate the marginal effects.

```
marginal_effects2 <- margins(probit_model)

summary(marginal_effects2)
```

```
##           factor     AME     SE        z      p   lower   upper
##       age_center -0.0036 0.0001 -26.7256 0.0000 -0.0038 -0.0033
##  degreeNo degree  0.1272 0.0050  25.6664 0.0000  0.1175  0.1369
##          urban_1 -0.0577 0.0059  -9.7475 0.0000 -0.0693 -0.0461
```

The results are very similar to the logistic regression model. We can further explore the similarities of probit and logit by predicting the probabilities based on the two models and visualising their density distributions:



Both models provide similar predicted values. Often, the choice between them depends on ease of interpretation. Some people find interpreting odds intuitive, while others prefer to focus on the probabilities. Also, there seem to be some preferences depending on the field, so check the literature to see what model is more common.

The discussion of GLMs could be expanded to include how to estimate the fit of the models and explore other types of link functions. For now, we will stop here. This should give you a good starting point to explore this topic further. It should also provide a solid foundation when we discuss expanding some of the longitudinal models to include different types of outcomes.

5.4 Further Reading

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