Data Analysis
Week 6: Multiple regression

1 Introduction

In Week 3 we introduced regression modelling where we modelled the relationship between an outcome variable y and an explanatory variable x. We only included one explanatory variable x, which was either a numerical or categorical variable. Here, we shall now examine fitting regression models with more than one explanatory variable. This is known as **multiple regression**.

When fitting regression models with multiple explanatory variables, the interpretation of an explanatory variable is made in association with the other variables. For example, if we wanted to model income then we may consider an individuals level of education, and perhaps the wealth of their parents. Then, when interpreting the effect an individuals level of education has on their income, we would also be considering the effect of the wealth of their parents simultaneously, as these two variables are likely to be related.

Note: Additional information and examples can be found in Chapter 7 of An Introduction to Statistical and Data Science via R.

2 Regression modelling with two numerical explanatory variables

Before we begin we need to load the following packages into R:

```r
library(ggplot2)
library(dplyr)
library(moderndive)
library(ISLR)
library(skimr)
library(plotly)
library(tidyr)
```

Let's start by looking at fitting a regression model with two numerical explanatory variables. We shall examine a data set within the ISLR package, which is an accompanying R package related to the textbook An Introduction to Statistical Learning with Applications in R. We will take a look at the Credit data set, which consists of predictions made on the credit card balance of 400 individuals, were the predictions are based on information relating to income, credit limit and the level of education of an individual.

Note: This is a simulated data set and is not based on credit card balances of actual individuals.

The regression model we will be considering contains the following variables:

- the numerical outcome variable y, the credit card balance of an individual; and
- two explanatory variables x_1 and x_2, which are an individuals credit limit and income (in thousands of dollars), respectively.

2.1 Exploratory data analysis

Task: Start by subsetting the Credit data set so that we only have the variables we are interested in, that is, Balance, Limit and Income. Note, it is best to give your new data set a different name than Credit as to not overwrite the original Credit data set. We can the glimpse function to take a look at our new data set (named Cred in this case):

```r
glimpse(Cred)
```

```
Observations: 400
Variables: 3
$ Balance <int> 333, 903, 580, 964, 331, 1151, 203, 872, 279, 1350, 14...
$ Limit   <int> 3606, 6645, 7075, 9504, 4897, 8047, 3388, 7114, 3300, ...
$ Income  <dbl> 14.891, 106.025, 104.593, 148.924, 55.882, 80.180, 20....
```

Note: the `View` function can also be used within RStudio to examine a spreadsheet of the data.

Now, let's take a look at summary statistics relating to our data set using the `skim` function:

```r
Cred %>%
  skim()
```

```
Skim summary statistics
 n obs: 400 
 n variables: 3 

-- Variable type:integer ---------------------------------------------
 variable missing complete   n    mean      sd  p0     p25    p50     p75
  Balance       0      400 400  520.01  459.76   0   68.75  459.5  863   
    Limit       0      400 400 4735.6  2308.2  855 3088    4622.5 5872.75
  p100     hist
  1999         
 13913         

-- Variable type:numeric ---------------------------------------------
 variable missing complete   n  mean    sd    p0   p25   p50   p75   p100
   Income       0      400 400 45.22 35.24 10.35 21.01 33.12 57.47 186.63
```

Now that we are looking at the relationship between an outcome variable and multiple explanatory variables, we need to examine the correlation between each of them. We can examine the correlation between `Balance`, `Limit` and `Income` by creating a table of correlations as follows:

```r
Cred %>%
  cor()
```

```
          Balance     Limit    Income
Balance 1.0000000 0.8616973 0.4636565
Limit   0.8616973 1.0000000 0.7920883
Income  0.4636565 0.7920883 1.0000000
```

Question: Why are the diagonal components of our correlation table all equal to 1?

From our correlation table we can see that the correlation between our two explanatory variables is 0.792, which is a strong positive linear relationship. Hence, we say there is a high degree of *collinearity* between our explanatory variables.

Collinearity (or **multicollinearity**) occurs when an explanatory variable within a multiple regression model can be linearly predicted from the other explanatory variables with a high level of accuracy. For example, in this case, since `Limit` and `Income` are highly correlated, we could take a good guess as to an individual's `Income` based on their `Limit`. That is, having one or more highly correlated explanatory variables within a multiple regression model essentially provides us with redundant information. Normally, we would remove one of the highly correlated explanatory variables, however, for the purpose of this example we shall ignore the potential issue of collinearity and carry on. You may want to use the `pairs` function or the `ggpairs` function from the `GGally` package to look at potential relationships between all of the variables within a data set.

Note: When we have several potential explanatory variables a model selection technique can help to identify which explanatory variables are significant predictors (in addition to the others) and which variables should be removed from the model. One procedure that can be used is **stepwise regression**, which implements an automatic procedure for choosing which explanatory variables should be included within the final model. A common stepwise procedure compares models using the model fit criterion **Akaike Information Criterion** (AIC) and can be implemented in R using the `stepAIC` function from the `MASS` library. This procedure allows for forward selection and backward selection (or both), where forward selection starts with the simplest model before iteratively including one explanatory variable at a time until the AIC reaches a minimum. The backward selection approach starts with the most complex model before removing one explanatory variable at a time until the minimum AIC is achieved.

Let's now produce scatterplots of the relationship between the outcome variable and the explanatory variables. First, we shall look at the scatterplot of `Balance` against `Limit`:

```
ggplot(Cred, aes(x = Limit, y = Balance)) +
  geom_point() +
  labs(x = "Credit limit (in $)", y = "Credit card balance (in $)",
       title = "Relationship between balance and credit limit") +
  geom_smooth(method = "lm", se = FALSE)
```

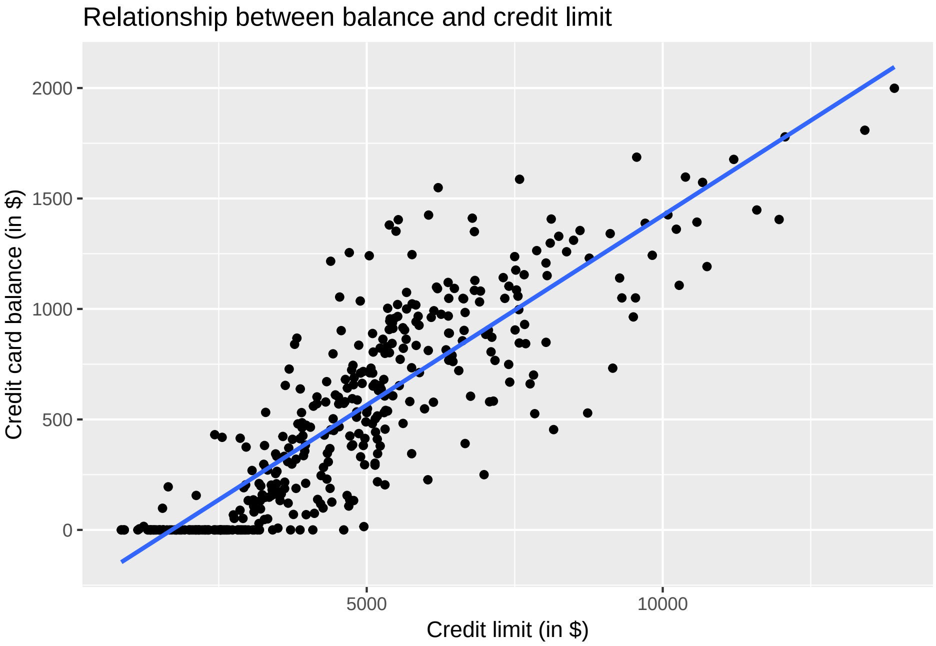

Figure 1: Relationship between balance and credit limit.

Now, let's look at a scatterplot of `Balance` and `Income`:

```
ggplot(Cred, aes(x = Income, y = Balance)) +
  geom_point() +
  labs(x = "Income (in $1000)", y = "Credit card balance (in $)",
       title = "Relationship between balance and income") +
  geom_smooth(method = "lm", se = FALSE)
```

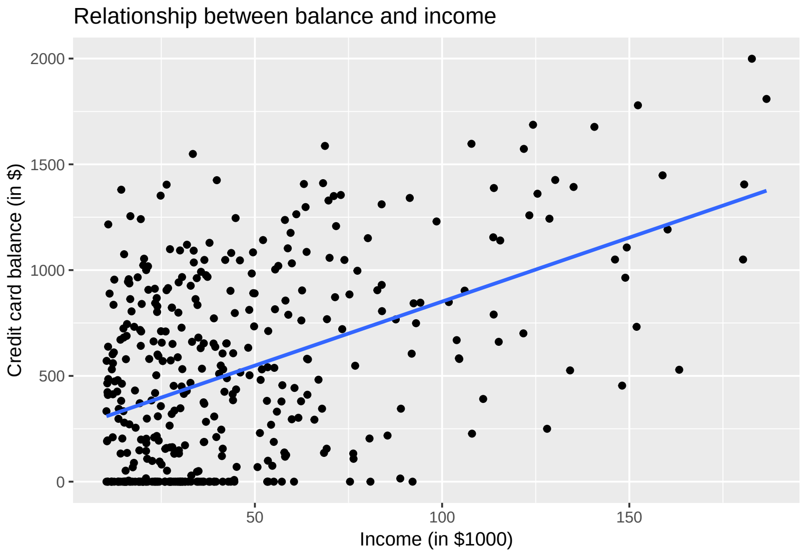

Figure 2: Relationship between balance and income.

The two scatterplots above focus on the relationship between the outcome variable `Balance` and each of the explanatory variables independently. In order to get an idea of the relationship between all three variables we can use the `plot_ly` function within the `plotly` library to plot a 3-dimensional scatterplot as follows:

```
plot_ly(Cred, x = ~Income, y = ~Limit, z = ~Balance,
        type = "scatter3d", mode = "markers")
```

In Week 3, when we fit our regression model, we were looking at the *best-fitting line*. However, now that we have more than one explanatory variable, we are looking at the *best-fitting plane*, which is a 3-dimensional generalisation of the best-fitting line.

2.2 Formal analysis

The multiple regression model we will be fitting to the credit balance data is given as:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2),$$

where

- y_i is the balance of the i^{th} individual;
- α is the intercept and positions the best-fitting plane in 3D space;
- β_1 is the coefficient for the first explanatory variable x_1;
- β_2 is the coefficient for the second explanatory variable x_2; and
- ϵ_i is the i^{th} random error component.

Similarly to Week 3, we use the `lm` function to fit the regression model and the `get_regression_table` function to view our parameter estimates:

```
Balance.model <- lm(Balance ~ Limit + Income, data = Cred)
get_regression_table(Balance.model)
```

```
# A tibble: 3 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept -385.         19.5     -19.8       0 -423.    -347.
```

```
2 Limit         0.264     0.006      45.0        0     0.253    0.276
3 Income       -7.66      0.385     -19.9        0    -8.42    -6.91
```

Note: To include multiple explanatory variables within a regression model we simply use the `+` sign, that is `Balance ~ Limit + Income`.

How do we interpret our model estimates defining the regression plane? They can be interpreted as follows:

- The **intercept** represents the credit card balance (`Balance`) of an individual who has $0 for both credit limit (`Limit`) and income (`Income`). However, the interpretation of the intercept in this case is somewhat limited as there are no individuals with $0 credit limit and income in the data set, with the smallest credit card balance being $0.
- The coefficient for credit limit (`Limit`) tells us that, *taking all other variables in the model into account*, that there is an associated increase, on average, in credit card balance of $0.26.
- Similarly, the coefficient for income (`Income`) tells us that, *taking all other variables in the model into account*, that there is an associated decrease, on average, in credit card balance of $7.66.

What do you notice that is strange about our coefficient estimates given our exploratory data analysis? Well, from our scatterplots of credit card balance against both credit limit and income, we seen that there appeared to be a positive linear relationship. Then, why do we then get a negative coefficient for income (-7.66)? This is due to a phenomenon known as **Simpson's Paradox**. This occurs when there are trends within different categories (or groups) of data, but that these trends disappear when the categories are grouped as a whole. For more details see Section 7.3.2 of An Introduction to Statistical and Data Sciences in R.

2.3 Assessing model fit

Now we need to assess our model assumptions. As a reminder from Week 3, our model assumptions are:

1. The deterministic part of the model captures all the non-random structure in the data, i.e. the residuals have mean zero.
2. The scale of the variability of the residuals is constant at all values of the explanatory variables.
3. The residuals are normally distributed.
4. The residuals are independent.
5. The values of the explanatory variables are recorded without error.

First, we need to obtain the fitted values and residuals from our regression model:

```r
regression.points <- get_regression_points(Balance.model)
```

Recall that `get_regression_points` provides us with values of the:

- outcome variable y (`Balance`);
- explanatory variables x_1 (`Limit`) and x_2 (`Income`);
- fitted values $\widehat{y}$; and
- the residual error $(y - \widehat{y})$.

We can assess our first two model assumptions by producing scatterplots of our residuals against each of our explanatory variables. First, let's begin with the scatterplot of the residuals against credit limit:

```r
ggplot(regression.points, aes(x = Limit, y = residual)) +
  geom_point() +
  labs(x = "Credit limit (in $)", y = "Residual", title = "Residuals vs credit limit")  +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```

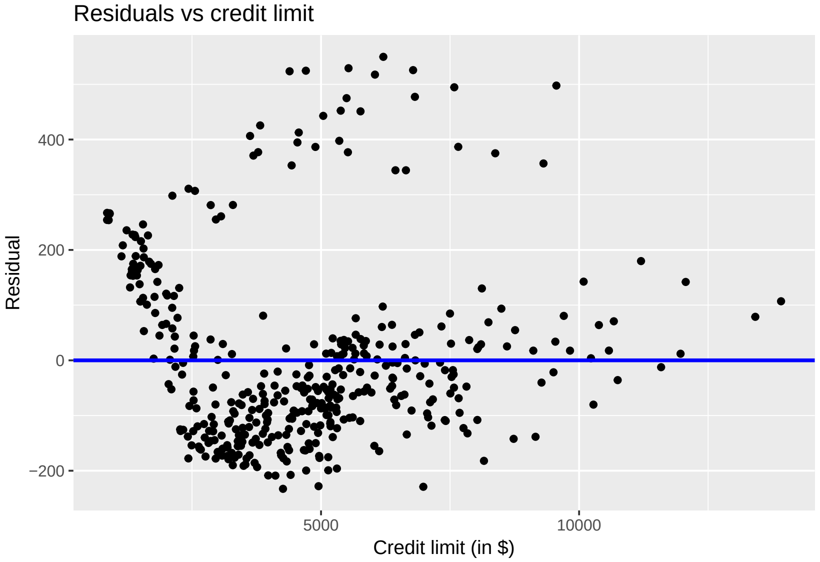

Figure 3: Residuals vs credit limit.

Now, let's plot a scatterplot of the residuals against income:

```r
ggplot(regression.points, aes(x = Income, y = residual)) +
  geom_point() +
  labs(x = "Income (in $1000)", y = "Residual", title = "Residuals vs income") +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```

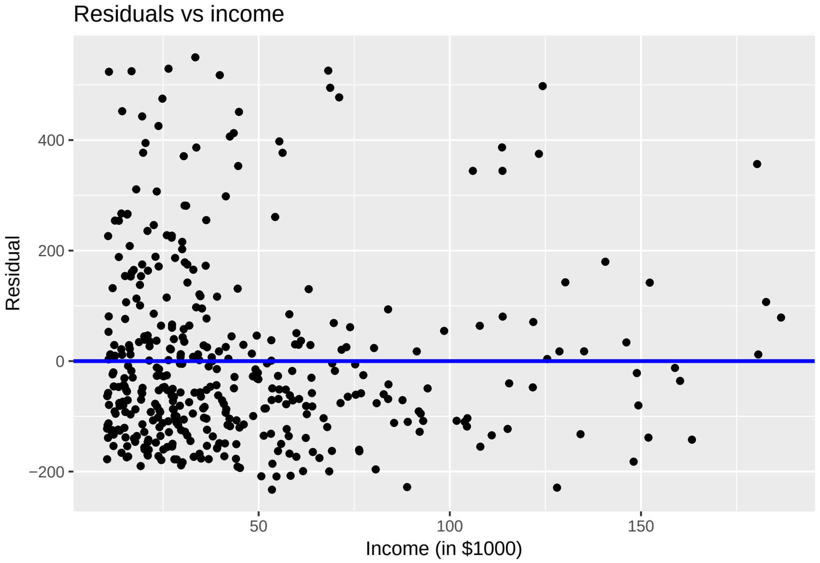

Figure 4: Residuals vs income.

Finally, we can check if the residuals are normally distributed by producing a histogram:

```r
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(color = "white") +
  labs(x = "Residual")
```

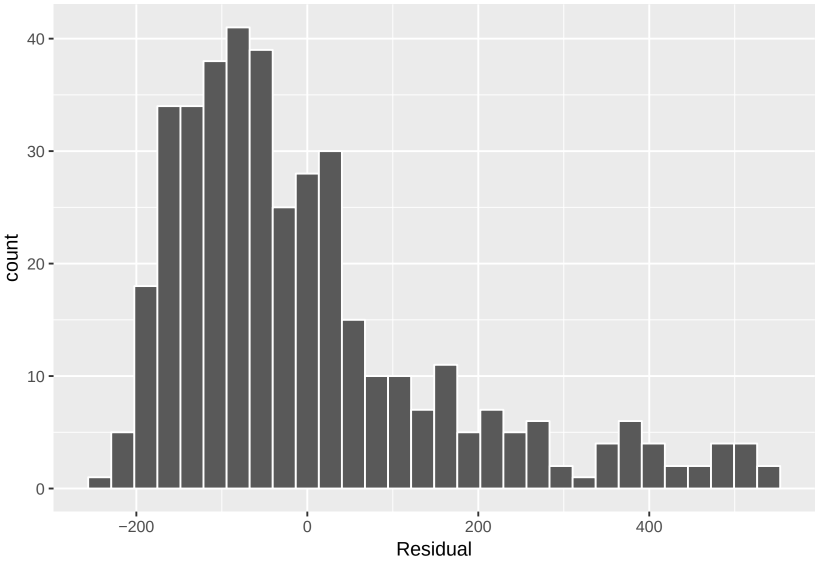

Figure 5: Histogram of the residuals.

3 Regression modelling with one numerical and one categorical explanatory variable

Let's expand upon Tasks 1, 2 and 3 from Week 3 by revisiting the instructor evaluation data set `evals`. In Week 3 you were tasked with examining the relationship between teaching score (`score`) and age (`age`). Now, let's also introduce the additional (binary) categorical explanatory variable gender (`gender`). That is, we we will be examining:

- the teaching score (`score`) as our outcome variable y;
- age (`age`) as our numerical explanatory variable x_1; and
- gender (`gender`) as our categorical explanatory variable x_2.

3.1 Exploratory data analysis

Start by subsetting the `evals` data set so that we only have the variables we are interested in, that is, `score`, `age` and `gender`. Note, it is best to give your new data set a different name than evals as to not overwrite the original `evals` data set. Your new data set should look like the one below.

```
# A tibble: 463 x 3
   score   age gender
   <dbl> <int> <fct>
 1   4.7    36 female
 2   4.1    36 female
 3   3.9    36 female
 4   4.8    36 female
 5   4.6    59 male
 6   4.3    59 male
 7   2.8    59 male
 8   4.1    51 male
 9   3.4    51 male
10   4.5    40 female
# ... with 453 more rows
```

Note: You can also view your data set using the `glimpse` function, or by opening a spreadsheet view in RStudio using the `View` function.

We can use the `skim` function to obtain some summary statistics from our data:

```
eval.score %>%
  skim()
```

```
Skim summary statistics
 n obs: 463
 n variables: 3

-- Variable type:factor ---------------------------------------------
 variable missing complete   n n_unique                top_counts ordered
   gender       0      463 463        2 mal: 268, fem: 195, NA: 0   FALSE

-- Variable type:integer --------------------------------------------
 variable missing complete   n  mean   sd p0 p25 p50 p75 p100     hist
      age       0      463 463 48.37 9.8 29  42  48  57   73

-- Variable type:numeric --------------------------------------------
 variable missing complete   n mean   sd  p0 p25 p50 p75 p100
    score       0      463 463 4.17 0.54 2.3 3.8 4.3 4.6    5
```

Now, let's compute the correlation coefficient between our outcome variable `score` and our numerical explanatory variable `age`:

```
eval.score %>%
  get_correlation(formula = score ~ age)
```

```
# A tibble: 1 x 1
  correlation
        <dbl>
1      -0.107
```

Note: The correlation coefficient only exists between numerical variables, which is why we do not include our categorical variable `gender`.

We can now visualise our data by producing a scatterplot, where seeing as we have the categorical variable `gender`, we shall plot the points using different colours for each gender:

```
ggplot(eval.score, aes(x = age, y = score, color = gender)) +
  geom_jitter() +
  labs(x = "Age", y = "Teaching Score", color = "Gender") +
  geom_smooth(method = "lm", se = FALSE)
```

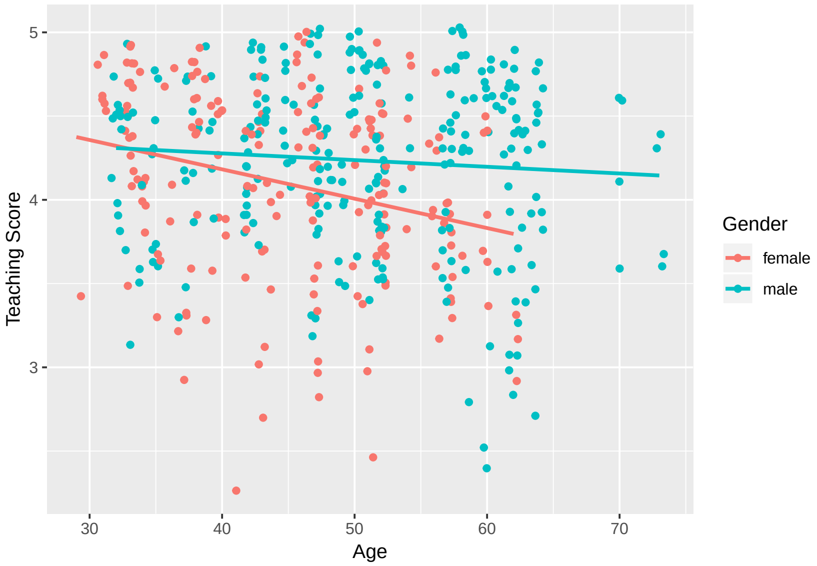

Figure 6: Instructor evaluation scores by age and gender. The points have been jittered.

Note: The above code has jittered the points, however, this is not necessary and `geom_point` would suffice. To plot separate points by gender we simply add the `color` argument to the `aes` function and pass to it `gender`.

From the scatterplot we can see that:

- There are very few women over the age of 60 in our data set.
- From the plotted regression lines we can see that the lines have different slopes for men and women. That is, the associated effect of increasing age appears to be more severe for women than it does for men, i.e. the teaching score of women drops faster with age.

3.2 Multiple regression: parallel slopes model

Here, we shall begin by fitting what is referred to as a parallel regression lines model. This model implies that the slope of relationship between teaching score (`score`) and age (`age`) is the same for both males and females, with only the intercept of the regression lines changing. Hence, our parallel regression lines model is given as:

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \\ &= \alpha + \beta_{\text{age}} \cdot \text{age} + \beta_{\text{male}} \cdot \mathbb{I}_{\text{male}}(x) + \epsilon_i, \end{aligned}$$

where

- α is the intercept of the regression line for females;
- β_{age} is the slope of the regression line for both males and females;
- β_{male} is the additional term added to α to get the intercept of the regression line for males; and
- $\mathbb{I}_{\text{male}}(x)$ is an indicator function such that

$$\mathbb{I}_{\text{male}}(x) = \left\{ \begin{array}{ll} 1 & \text{if gender } x \text{ is male,} \\ 0 & \text{Otherwise.} \end{array} \right.$$

We can fit the parallel regression lines model as follows:

```
par.model <- lm(score ~ age + gender, data = eval.score)
get_regression_table(par.model)
```

```
# A tibble: 3 x 7
  term       estimate std_error statistic p_value lower_ci upper_ci
  <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept     4.48      0.125     35.8    0         4.24     4.73
2 age          -0.009     0.003     -3.28   0.001    -0.014   -0.003
3 gendermale    0.191     0.052      3.63   0         0.087    0.294
```

Hence, the regression line for females is given by:

$$\widehat{\text{score}} = 4.48 - 0.009 \cdot \text{age},$$

while the regression line for males is given by:

$$\widehat{\text{score}} = 4.48 - 0.009 \cdot \text{age} + 0.191 = 4.671 - 0.009 \cdot \text{age}.$$

Now, let's superimpose our parallel regression lines onto the scatterplot of teaching score against age:

```
coeff  <- par.model %>%
          coef() %>%
          as.numeric()

slopes <- eval.score %>%
  group_by(gender) %>%
  summarise(min = min(age), max = max(age)) %>%
  mutate(intercept = coeff[1]) %>%
  mutate(intercept = ifelse(gender == "male", intercept + coeff[3], intercept)) %>%
  gather(point, age, -c(gender, intercept)) %>%
  mutate(y_hat = intercept + age * coeff[2])

ggplot(eval.score, aes(x = age, y = score, col = gender)) +
  geom_jitter() +
  labs(x = "Age", y = "Teaching Score", color = "Gender") +
  geom_line(data = slopes, aes(y = y_hat), size = 1)
```

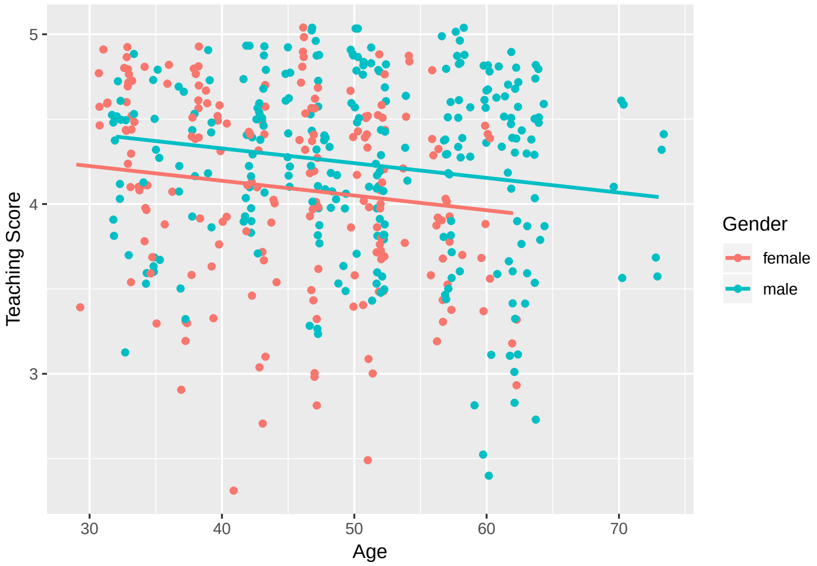

Figure 7: Instructor evaluation scores by age and gender with parallel regression lines superimposed.

Note: go through the code used to create `coeff` and `slopes` and make sure you understand it.

From the parallel regression lines model both males and females have the same slope, that is, the associated effect of age on teaching score is the same for both men and women. Hence, for every one year increase in age, there is an associated decrease in teaching score of 0.009. However, male instructors have a higher intercept term, that is, there is a vertical bump in the regression line for males in teaching scores. This is linked to the average difference in teaching scores that males obtain relative to females.

Question: What is different between our previous scatterplot of teaching score against age (Figure 6) and the one we just created with our parallel lines superimposed (Figure 7)? In the original plot we have what is referred to as an interaction effect between age and gender. Hence, gender interacts in different ways for both males and females by age, and as such we should have different intercepts **and** slopes.

3.3 Multiple regression: interaction model

There is an *interaction effect* if the associated effect of one variable depends on the value of another variable. For example, the effect of age here will depend on whether the instructor is male or female, that is, the effect of age on teaching scores will differ by gender. The interaction model can be written as:

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i \\ &= \alpha + \beta_{\text{age}} \cdot \text{age} + \beta_{\text{male}} \cdot \mathbb{I}_{\text{male}}(x) + \beta_{\text{age, male}} \cdot \text{age} \cdot \mathbb{I}_{\text{male}}(x) + \epsilon_i, \end{aligned}$$

where $\beta_{\text{age, male}} \cdot \text{age} \cdot \mathbb{I}_{\text{male}}(x)$ corresponds to the interaction term.

In order to fit an interaction term within our regression model we replace the `+` sign with the `*` sign as follows:

```
int.model <- lm(score ~ age * gender, data = eval.score)
get_regression_table(int.model)
```

```
# A tibble: 4 x 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
```

```
1 intercept         4.88      0.205     23.8    0         4.48     5.29
2 age              -0.018     0.004     -3.92   0        -0.026   -0.009
3 gendermale       -0.446     0.265     -1.68   0.094    -0.968    0.076
4 age:gendermale    0.014     0.006      2.45   0.015     0.003    0.024
```

Hence, the regression line for females is given by:

$$\widehat{\text{score}} = 4.88 - 0.018 \cdot \text{age},$$

while the regression line for males is given by:

$$\widehat{\text{score}} = 4.88 - 0.018 \cdot \text{age} - 0.446 + 0.014 \cdot \text{age} = 4.434 - 0.004 \cdot \text{age}.$$

Note: How do they compare with the teaching score values from the parallel regression lines model?

Here, we can see that, although the intercept for male instructors may be lower, the associated average decrease in teaching score with age (0.004) is not as severe as it is for female instructors (0.018).

3.4 Assessing model fit

Now we have to assess the fit of the model by looking at plots of the residuals. We shall do this for the interaction model. First, we need to obtain the fitted values and residuals from the interaction model as follows:

```r
regression.points <- get_regression_points(int.model)
```

```
# A tibble: 463 x 6
      ID score   age gender score_hat residual
   <int> <dbl> <dbl> <fct>      <dbl>    <dbl>
 1     1   4.7    36 female      4.25    0.448
 2     2   4.1    36 female      4.25   -0.152
 3     3   3.9    36 female      4.25   -0.352
 4     4   4.8    36 female      4.25    0.548
 5     5   4.6    59 male        4.20    0.399
 6     6   4.3    59 male        4.20    0.099
 7     7   2.8    59 male        4.20   -1.40
 8     8   4.1    51 male        4.23   -0.133
 9     9   3.4    51 male        4.23   -0.833
10    10   4.5    40 female      4.18    0.318
# ... with 453 more rows
```

Let's start by looking at a scatterplot of the residuals against the explanatory variable by gender:

```r
ggplot(regression.points, aes(x = age, y = residual)) +
  geom_point() +
  labs(x = "age", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ gender)
```

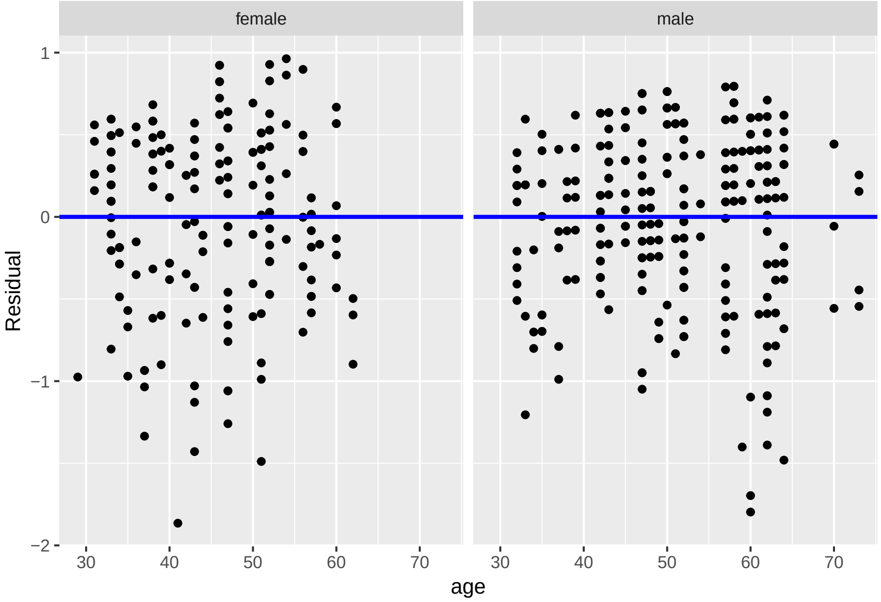

Figure 8: Residuals vs the explanatory variable age by gender.

Now, we can plot the residuals against the fitted values:

```
ggplot(regression.points, aes(x = score_hat, y = residual)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ gender)
```

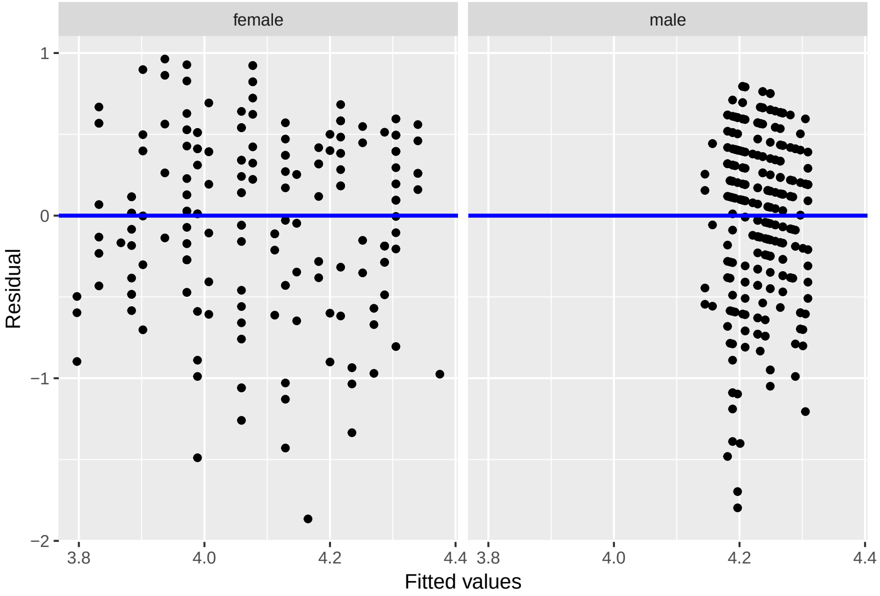

Figure 9: Residuals vs the fitted values.

Finally, let's plot histograms of the residuals to assess whether they are normally distributed with mean zero:

```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(binwidth = 0.25, color = "white") +
  labs(x = "Residual") +
  facet_wrap(~gender)
```

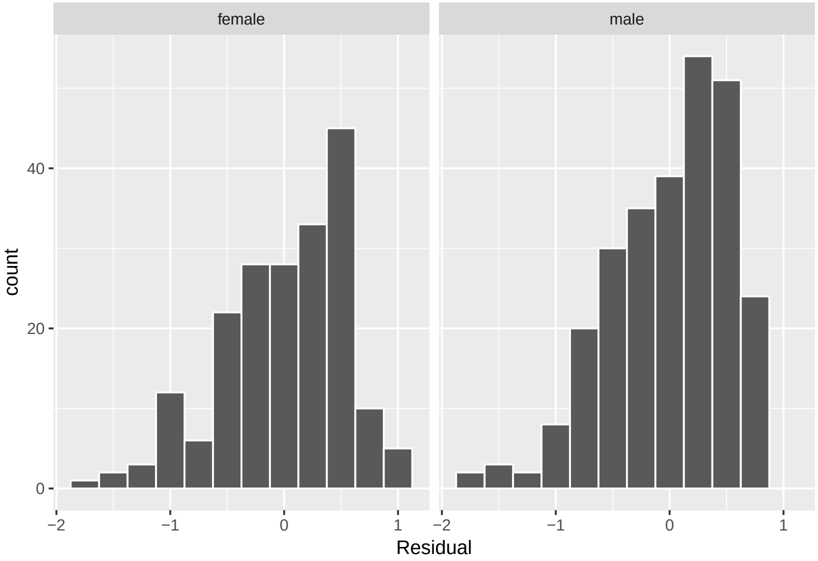

Figure 10: Histograms of the residuals by gender.

Question: Do the model assumptions hold?

4 Tasks

1. Assess the model assumptions for the parallel regression lines model. Do they appear valid?
2. Return to the `Credit` data set and fit a multiple regression model with `Balance` as the outcome variable, and `Income` and `Age` as the explanatory variables, respectively. Assess the assumptions of the multiple regression model.
3. Return to the `Credit` data set and fit a parallel regression lines model with `Balance` as the outcome variable, and `Income` and `Student` as the explanatory variables, respectively. Assess the assumptions of the fitted model.

Trickier

4. Load the library `datasets` and look at the `iris` data set of Edgar Anderson containing measurements (in centimetres) on 150 different flowers across three different species of iris. Fit an interaction model with `Sepal.Width` as the outcome variable, and `Sepal.Length` and `Species` as the explanatory variables. Assess the assumptions of the fitted model.