Multiple Regression

El Mex

Contents


```
library(tidyverse)
library(moderndive)
library(skimr)
library(ISLR)
```

1 Multiple regression

As in our previous lession, our outcome variable will still be teaching score, but we'll now include two different explanatory variables: age and gender. Could it be that instructors who are older receive better teaching evaluations from students? Or could it instead be that younger instructors receive better evaluations? Are there differences in evaluations given by students for instructors of different genders?

We will have:

1. A numerical outcome variable y, the instructor's teaching score
2. Two explanatory variables:
    - A numerical explanatory variable x_1, the instructor's age
    - A categorical explanatory variable x_2, the instructor's (binary) gender

1.1 One numerical and one categorical explanatory variable

1.1.1 Exploratory data analysis

Use the `evals` data frame and `select()` the columns: `ID`, `score`, `age`, `gender`

```
# save this data in a new data frame called: `evals_ch6`
evals_ch6 <- evals %>%
  select(ID, score, age, gender)
```

1. **Looking at the raw data values**

```r
# get a glimpse
glimpse(evals_ch6)
```

```
## Rows: 463
## Columns: 4
## $ ID     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1...
## $ score  <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, ...
## $ age    <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, ...
## $ gender <fct> female, female, female, female, male, male, male, male, male...
```

Display a random sample of only 5 rows (each row is one course)

```r
evals_ch6 %>% sample_n(size = 5)
```

```
## # A tibble: 5 x 4
##      ID score   age gender
##   <int> <dbl> <int> <fct>
## 1    77   3.5    49 male
## 2   439   3.3    35 female
## 3   462   4.4    42 female
## 4   433   3.1    62 male
## 5   312   2.9    62 female
```

2. **Computing summary statistics**

```r
# select the variables of interest in our model and then skim()
evals_ch6 %>%
  select(score, age, gender) %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	463
Number of columns	3
Column type frequency:	
factor	1
numeric	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	mal: 268, fem: 195

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
score	0	1	4.17	0.54	2.3	3.8	4.3	4.6	5	
age	0	1	48.37	9.80	29.0	42.0	48.0	57.0	73	

Observe that we have no missing data, that there are 268 courses taught by male instructors and 195 courses taught by female instructors, and that the average instructor age is 48.37. Recall that each row represents a particular course and that the same instructor often teaches more than one course

Let's compute the correlation coefficient between our two numerical variables: `score` and `age`

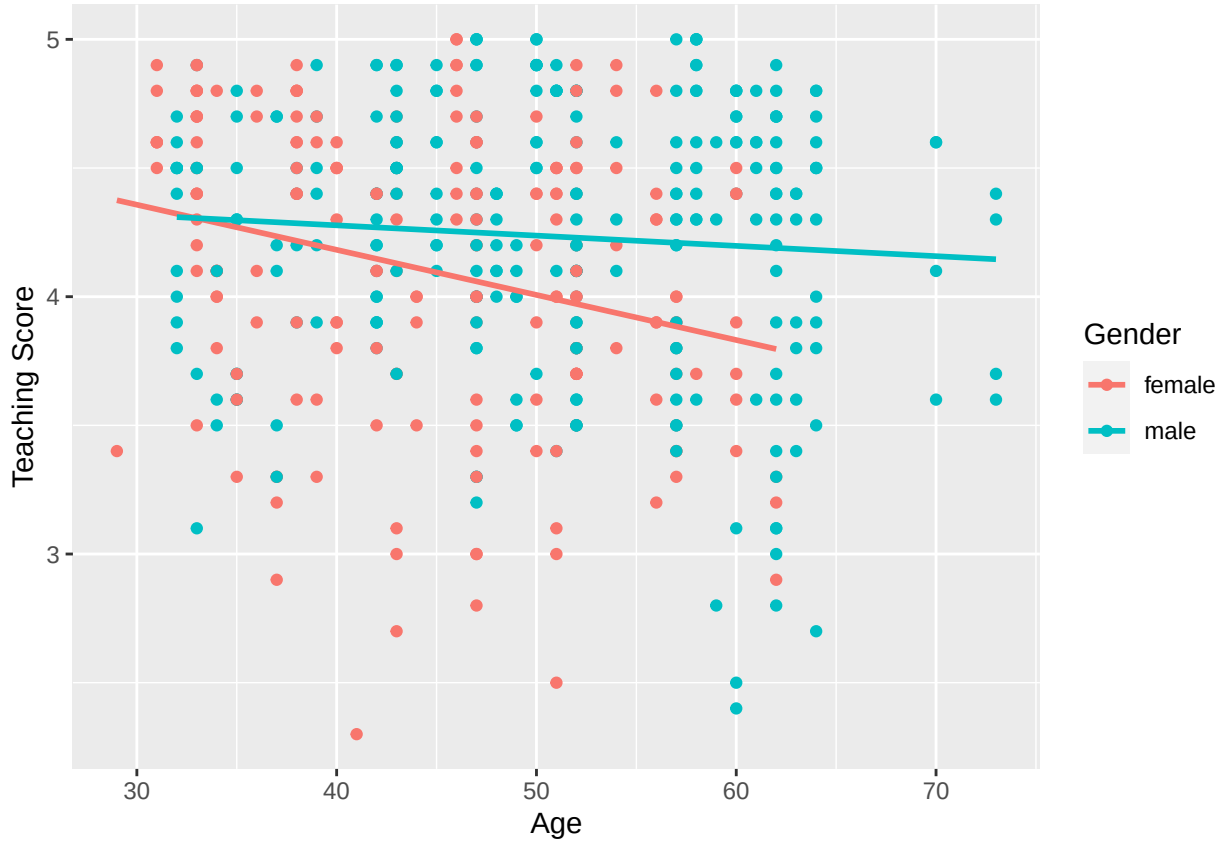
```
evals_ch6 %>%
  get_correlation(formula = score ~ age)
```

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

3. **Creating data visualizations**

The outcome variable `score` and explanatory variable `age` are both numerical, hence we'll use a scatterplot to display their relationship. How can we incorporate the categorical variable `gender`, however? By `mapping` the variable `gender` to the `color` aesthetic

```
ggplot(evals_ch6, aes(x = age, y = score, color = gender)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Score", color = "Gender") +
  geom_smooth(method = lm, se = FALSE)
```



There are almost no women faculty over the age of 60 as evidenced by lack of red dots above $x = 60$. Second, while both regression lines are negatively sloped with age (i.e., older instructors tend to have lower scores), the slope for age for the female instructors is more negative

1.1.2 Interaction model

First "fit" the model using the `lm()` "linear model" function and then we apply the `get_regression_table()` function. Our model formula won't be of the form `y ~ x`, but rather of the form `y ~ x1 * x2`

```
# fit regression model:
score_model_interaction <- lm(score ~ age * gender, data = evals_ch6)

# get regression table:
get_regression_table(score_model_interaction)
```

```
## # 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
```

Female instructors are the "baseline for comparison" group (alphabetically ordered). Thus,

`intercept`is the intercept for only the female instructors. Same for `age`, the slope for age only for the female instructors

The red regression line (female) has an intercept of 4.883 and slope for age of -0.018. Remember that while the intercept has a mathematical interpretation, it has no practical interpretation here since instructors can't have zero age

The value for `gendermale` of -0.446 is not the intercept for the male instructors, but rather the offset in intercept for male instructors relative to female instructors. The intercept for the male instructors is `intercept + gendermale` = 4.883 + (-0.446) = 4.883 - 0.446 = 4.437

Same for `age:gendermale` = 0.014 is not the slope for age for the male instructors the slope for age for the male instructors is `age + age:gendermale` $= -0.018 + 0.014 = -0.004$**. Thus, the blue regression line has intercept 4.437 and slope for age of -0.004**

Table 4: Comparison of intercepts and slopes for interaction model

Gender	Intercept	Slope for age
Female instructors	4.883	-0.018
Male instructors	4.437	-0.004

Since the slope for age for the female instructors was -0.018, it means that on average, a female instructor who is a year older would have a teaching score that is 0.018 units *lower*. For the male instructors, however, the corresponding associated decrease was on average only 0.004 units. While both slopes for age were negative, the slope for age for the female instructors is more negative

Let's now write the equation for our regression lines, which we can use to compute our fitted values $\hat{y} = \widehat{\text{score}}$.

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= b_0 + b_{\text{age}} \cdot \text{age} + b_{\text{male}} \cdot 1_{\text{is male}}(x) + b_{\text{age,male}} \cdot \text{age} \cdot 1_{\text{is male}}(x) \\ &= 4.883 - 0.018 \cdot \text{age} - 0.446 \cdot 1_{\text{is male}}(x) + 0.014 \cdot \text{age} \cdot 1_{\text{is male}}(x)\end{aligned}$$

Whoa! That's even more daunting than the equation you saw for the life expectancy as a function of continent in Subsection 5.2.2! However, if you recall what an "indicator function" does, the equation simplifies greatly. In the previous equation, we have one indicator function of interest:

$$1_{\text{is male}}(x) = \begin{cases} 1 & \text{if instructor } x \text{ is male} \\ 0 & \text{otherwise} \end{cases}$$

Second, let's match coefficients in the previous equation with values in the `estimate` column in our regression table in Table 6.3:

1. b_0 is the `intercept` = 4.883 for the female instructors
2. b_{age} is the slope for `age` = -0.018 for the female instructors
3. b_{male} is the offset in intercept = -0.446 for the male instructors
4. $b_{\text{age,male}}$ is the offset in slope for age = 0.014 for the male instructors

Let's put this all together and compute the fitted value $\hat{y} = \widehat{\text{score}}$ for female instructors. Since for female instructors $1_{\text{is male}}(x)$ = 0, the previous equation becomes

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= 4.883 - 0.018 \cdot \text{age} - 0.446 \cdot 0 + 0.014 \cdot \text{age} \cdot 0 \\ &= 4.883 - 0.018 \cdot \text{age} - 0 + 0 \\ &= 4.883 - 0.018 \cdot \text{age}\end{aligned}$$

which is the equation of the red regression line in Figure 6.1 corresponding to the female instructors in Table 6.4. Correspondingly, since for male instructors $1_{\text{is male}}(x)$ = 1, the previous equation becomes

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= 4.883 - 0.018 \cdot \text{age} - 0.446 + 0.014 \cdot \text{age} \\ &= (4.883 - 0.446) + (-0.018 + 0.014) * \text{age} \\ &= 4.437 - 0.004 \cdot \text{age}\end{aligned}$$

which is the equation of the blue regression line in Figure 6.1 corresponding to the male instructors in Table 6.4.

Before we end this section, we explain why we refer to this type of model as an "interaction model." The $b_{\text{age,male}}$ term in the equation for the fitted value $\hat{y} = \widehat{\text{score}}$ is what's known in statistical modeling as an "interaction effect." The interaction term corresponds to the `age:gendermale` = 0.014 in the final row of the regression table in Table 6.3.

We say there is an interaction effect if the associated effect of one variable *depends on the value of another variable*. That is to say, the two variables are "interacting" with each other. Here, the associated effect of the variable age *depends* on the value of the other variable gender. The difference in slopes for age of +0.014 of male instructors relative to female instructors shows this.

Another way of thinking about interaction effects on teaching scores is as follows. For a given instructor at UT Austin, there might be an associated effect of their age *by itself*, there might be an associated effect of their gender *by itself*, but when age and gender are considered *together* there might be an *additional effect* above and beyond the two individual effects.

1.1.3 Parallel slopes model

When creating regression models with one numerical and one categorical explanatory variable, another type of model we can use is known as a parallel slopes model. Unlike interaction models where the regression lines can have different intercepts and different slopes, parallel slopes models still allow for different intercepts but force all lines to have the same slope. The resulting regression lines are thus parallel

Use `geom_parallel_slopes()` included in the `moderndive` package

```
ggplot(evals_ch6, aes(x = age, y = score, color = gender)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Scores", color = "Gender") +
  geom_parallel_slopes(se = FALSE)
```

We now have parallel lines corresponding to the female and male instructors, respectively: here they have the same negative slope. Older instructors will tend to receive lower teaching scores than instructors who are younger. Since the lines are parallel, the associated penalty for being older is assumed to be the same for both female and male

On the other hand, the two lines have different intercepts as the blue line is higher than the red line. It tells us that female instructors tended to receive lower teaching scores than male instructors

To obtain the exact values of the two intercepts and slope, let's fit the model and apply `get_regression_table()`. Unlike the interaction model which had a model formula of the form `y ~ x1 * x2`, our model formula is now of the form `y ~ x1 + x2`

```
# fit regression model:
score_model_parallel_slopes <- lm(score ~ age + gender, data = evals_ch6)

# get regression table:
get_regression_table(score_model_parallel_slopes)
```

```
## # 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
```

The red regression line corresponding to the female instructors has an intercept of 4.484 while the blue regression line corresponding to the male instructors has an intercept of 4.484 +

0.191 = 4.675. Once again, since there aren't any instructors of age 0, the intercepts only have a mathematical interpretation but no practical one. However, we now only have a single slope for age of -0.009. This is telling us that an instructor who is a year older than another instructor received a teaching score that is on average 0.009 units lower. This penalty for being of advanced age applies equally to both female and male instructors

Let's now write the equation for our regression lines, which we can use to compute our fitted values $\hat{y} = \widehat{\text{score}}$.

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= b_0 + b_{\text{age}} \cdot \text{age} + b_{\text{male}} \cdot 1_{\text{is male}}(x) \\ &= 4.484 - 0.009 \cdot \text{age} + 0.191 \cdot 1_{\text{is male}}(x)\end{aligned}$$

Let's put this all together and compute the fitted value $\hat{y} = \widehat{\text{score}}$ for female instructors. Since for female instructors the indicator function $1_{\text{is male}}(x) = 0$, the previous equation becomes

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= 4.484 - 0.009 \cdot \text{age} + 0.191 \cdot 0 \\ &= 4.484 - 0.009 \cdot \text{age}\end{aligned}$$

which is the equation of the red regression line in Figure 6.2 corresponding to the female instructors. Correspondingly, since for male instructors the indicator function $1_{\text{is male}}(x) = 1$, the previous equation becomes

$$\begin{aligned}\hat{y} = \widehat{\text{score}} &= 4.484 - 0.009 \cdot \text{age} + 0.191 \cdot 1 \\ &= (4.484 + 0.191) - 0.009 \cdot \text{age} \\ &= 4.675 - 0.009 \cdot \text{age}\end{aligned}$$

which is the equation of the blue regression line in Figure 6.2 corresponding to the male instructors.

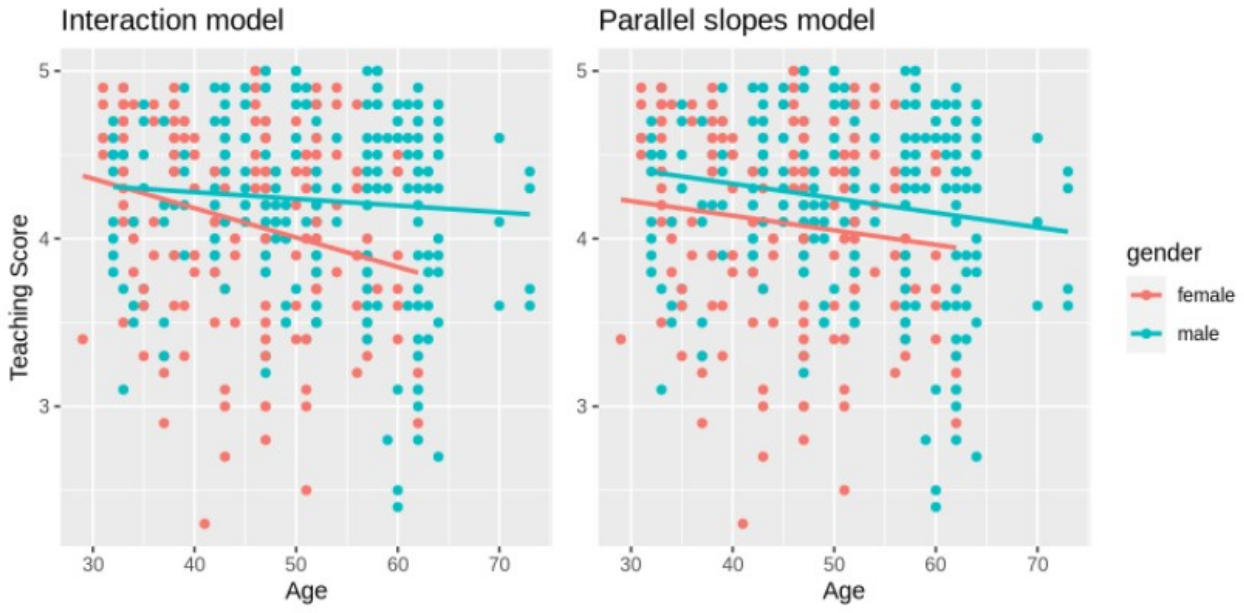


FIGURE 6.3: Comparison of interaction and parallel slopes models.

Looking at the left-hand plot in Figure 6.3, the two lines definitely do not appear to be parallel,

so why would we force them to be parallel? For this data, we agree! It can easily be argued that the interaction model on the left is more appropriate. However, later on model selection, we'll present an example where it can be argued that the case for a parallel slopes model might be stronger

1.1.4 Observed/fitted values and residuals

Say, you have a female instructor who is 36 years old and a male who is 59 years old. What would their fitted value $\hat{y}$ be?

Let's mark the points of the ages on the scatterplot by drawing dot lines

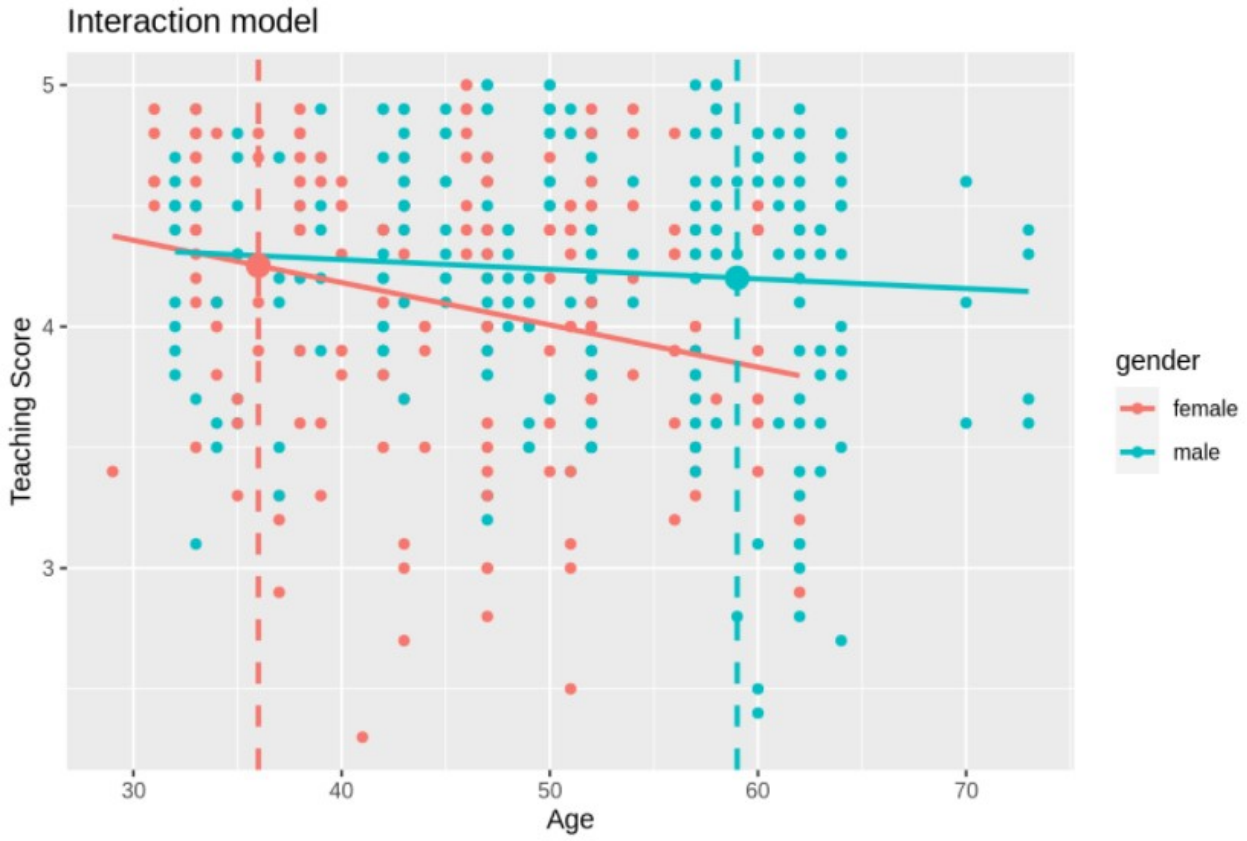


FIGURE 6.4: Fitted values for two new professors.

We can use the equations of the two regression lines:

- For all female instructors: $\hat{y} = \widehat{\text{score}} = 4.883 - 0.018 \cdot \text{age}$
- For all male instructors: $\hat{y} = \widehat{\text{score}} = 4.437 - 0.004 \cdot \text{age}$

So our fitted values would be: $4.883 - 0.018 \cdot 36 = 4.24$ and $4.437 - 0.004 \cdot 59 = 4.20$, respectively.

Now what if we want the fitted values not just for these two instructors, but for the instructors of all 463 courses included in the `evals_ch6` data frame? Doing this by hand would be long and tedious! This is where the `get_regression_points()` function from the `moderndive` package can help: it will quickly automate the above calculations for all 463 courses. We present a preview of just the first 10 rows out of 463 in Table 6.7.

```
regression_points <- get_regression_points(score_model_interaction)

head(regression_points, 10)
```

```
## # A tibble: 10 x 6
##       ID score   age gender score_hat residual
##    <int> <dbl> <int> <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
```

It turns out that the female instructor of age 36 taught the first four courses, while the male instructor taught the next 3. The resulting $\hat{y} = \widehat{\text{score}}$ fitted values are in the `score_hat` column. Furthermore, the `get_regression_points()` function also returns the residuals $y - \hat{y}$. Notice, for example, the first and fourth courses the female instructor of age 36 taught had positive residuals, indicating that the actual teaching scores they received from students were greater than their fitted score of 4.25. On the other hand, the second and third courses this instructor taught had negative residuals, indicating that the actual teaching scores they received from students were less than 4.25.

Compute the observed values, fitted values, and residuals for the parallel slopes model we saved in `score_model_parallel_slopes`

```
regression_points_parallel <- get_regression_points(score_model_parallel_slopes)

head(regression_points_parallel, 10)
```

```
## # A tibble: 10 x 6
##       ID score   age gender score_hat residual
##    <int> <dbl> <int> <fct>      <dbl>    <dbl>
##  1     1   4.7    36 female      4.17    0.528
```

```
##  2     2   4.1    36 female      4.17   -0.072
##  3     3   3.9    36 female      4.17   -0.272
##  4     4   4.8    36 female      4.17    0.628
##  5     5   4.6    59 male        4.16    0.437
##  6     6   4.3    59 male        4.16    0.137
##  7     7   2.8    59 male        4.16   -1.36
##  8     8   4.1    51 male        4.23   -0.132
##  9     9   3.4    51 male        4.23   -0.832
## 10    10   4.5    40 female      4.14    0.363
```

1.2 Two numerical explanatory variables

We'll use the `Credit` dataset from the `ISLR` package whichis the outcome variable of interest is the credit card debt of 400 individuals

In this section, we'll fit a regression model where we have:

1. A numerical outcome variable y, the cardholder's credit card debt
2. Two explanatory variables:
    - One numerical explanatory variable x_1, the cardholder's credit limit
    - Another numerical explanatory variable x_2, the cardholder's income (in thousands of dollars)

1.2.1 Exploratory data analysis

`select()` the subset of the variables we'll consider in this chapter, and save this data in the new data frame `credit_ch6`. Also select `Balance` variable and save it with a new name: `debt` so it is easy to understand

```
credit_ch6 <- Credit %>% as_tibble() %>%
  select(ID, debt = Balance, credit_limit = Limit,
         income = Income, credit_rating = Rating, age = Age)

glimpse(credit_ch6)
```

```
## Rows: 400
## Columns: 6
## $ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ debt          <int> 333, 903, 580, 964, 331, 1151, 203, 872, 279, 1350, 1...
## $ credit_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...
## $ credit_rating <int> 283, 483, 514, 681, 357, 569, 259, 512, 266, 491, 589...
## $ age           <int> 34, 82, 71, 36, 68, 77, 37, 87, 66, 41, 30, 64, 57, 4...
```

Let's look at a random sample of five out of the 400 credit card holders

```
credit_ch6 %>% sample_n(size = 5)
```

```
## # A tibble: 5 x 6
##      ID  debt credit_limit income credit_rating   age
```

```
##   <int> <int>        <int>  <dbl>         <int> <int>
## 1   163  1086         7530   63.8           515    56
## 2   378     0         1567   28.1           142    51
## 3   352   302         5140   61.6           374    71
## 4   126     0         2531   27.6           195    34
## 5   155     0         3155   55.1           235    31
```

Let's now compute summary statistics

```
credit_ch6 %>%
  select(debt, credit_limit, income) %>% skim()
```

Table 5: Data summary

Name	Piped data
Number of rows	400
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
debt	0	1	520.02	459.76	0.00	68.75	459.50	863.00	1999.00	
credit_limit	0	1	4735.60	2308.20	855.00	3088.00	4622.50	5872.75	13913.00	
income	0	1	45.22	35.24	10.35	21.01	33.12	57.47	186.63	

`debt`: the mean and median credit card debt are \$520.01 and \$459.50, respectively, and that 25% of card holders had debts of \$68.75 or less. Let's now look at one of the explanatory variables `credit_limit`: the mean and median credit card limit are \$4735.6 and \$4622.50, respectively, while 75% of card holders had incomes of \$57,470 or less

We can compute the correlation coefficient between the different possible pairs of these variables as they are numerical

```
credit_ch6 %>% get_correlation(debt ~ credit_limit)
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.862
```

```
credit_ch6 %>% get_correlation(debt ~ income)
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.464
```

Or we can simultaneously compute them by returning a correlation matrix

```
credit_ch6 %>% select(debt, credit_limit, income) %>%
  cor()
```

```
##                  debt credit_limit    income
## debt         1.0000000    0.8616973 0.4636565
## credit_limit 0.8616973    1.0000000 0.7920883
## income       0.4636565    0.7920883 1.0000000
```

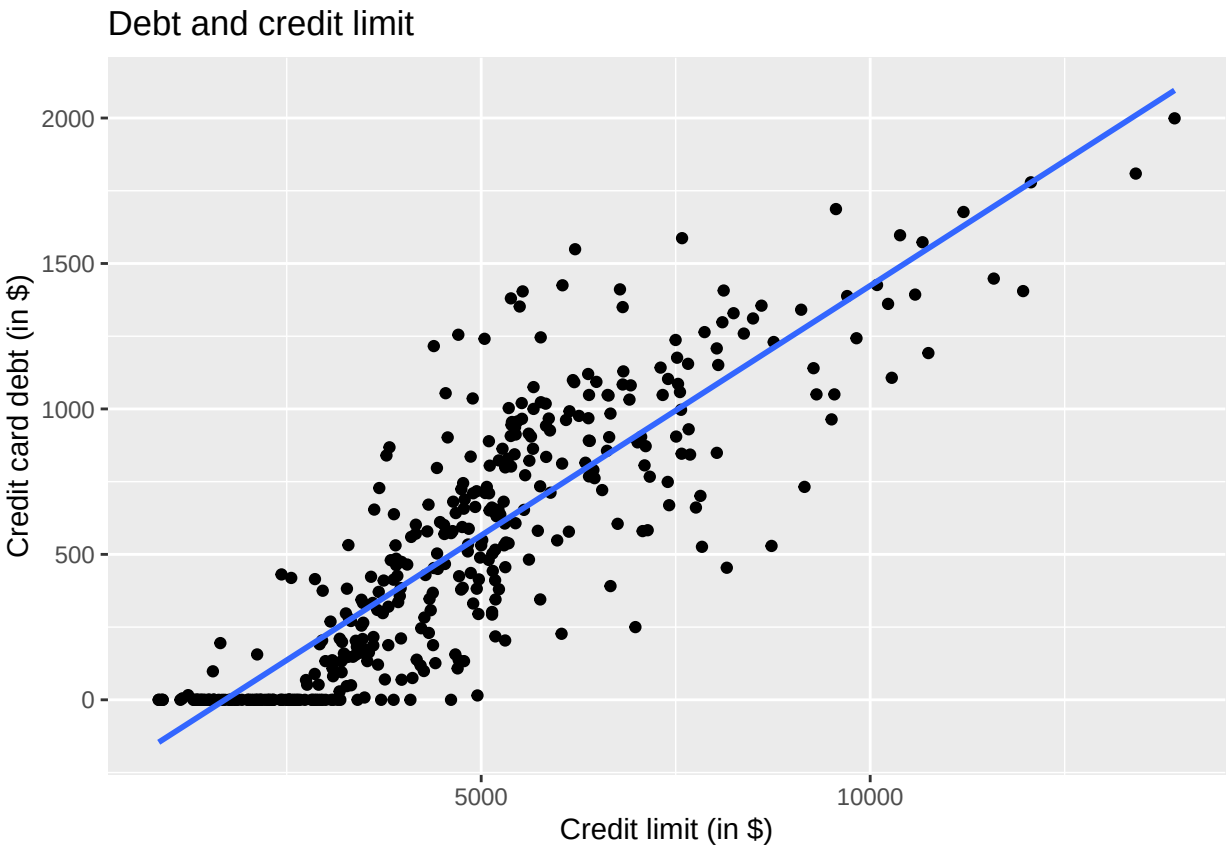
1. `debt` with itself is 1 as we would expect based on the definition of the correlation coefficient.
2. `debt` with `credit_limit` is 0.862. This indicates a strong positive linear relationship, which makes sense as only individuals with large credit limits can accrue large credit card debts.
3. `debt` with `income` is 0.464. This is suggestive of another positive linear relationship, although not as strong as the relationship between `debt` and `credit_limit`.
4. As an added bonus, we can read off the correlation coefficient between the two explanatory variables of `credit_limit` and `income` as 0.792.

We say there is a high degree of *collinearity* between the `credit_limit` and `income` explanatory variables. Collinearity (or multicollinearity) is a phenomenon where one explanatory variable in a multiple regression model is highly correlated with another.

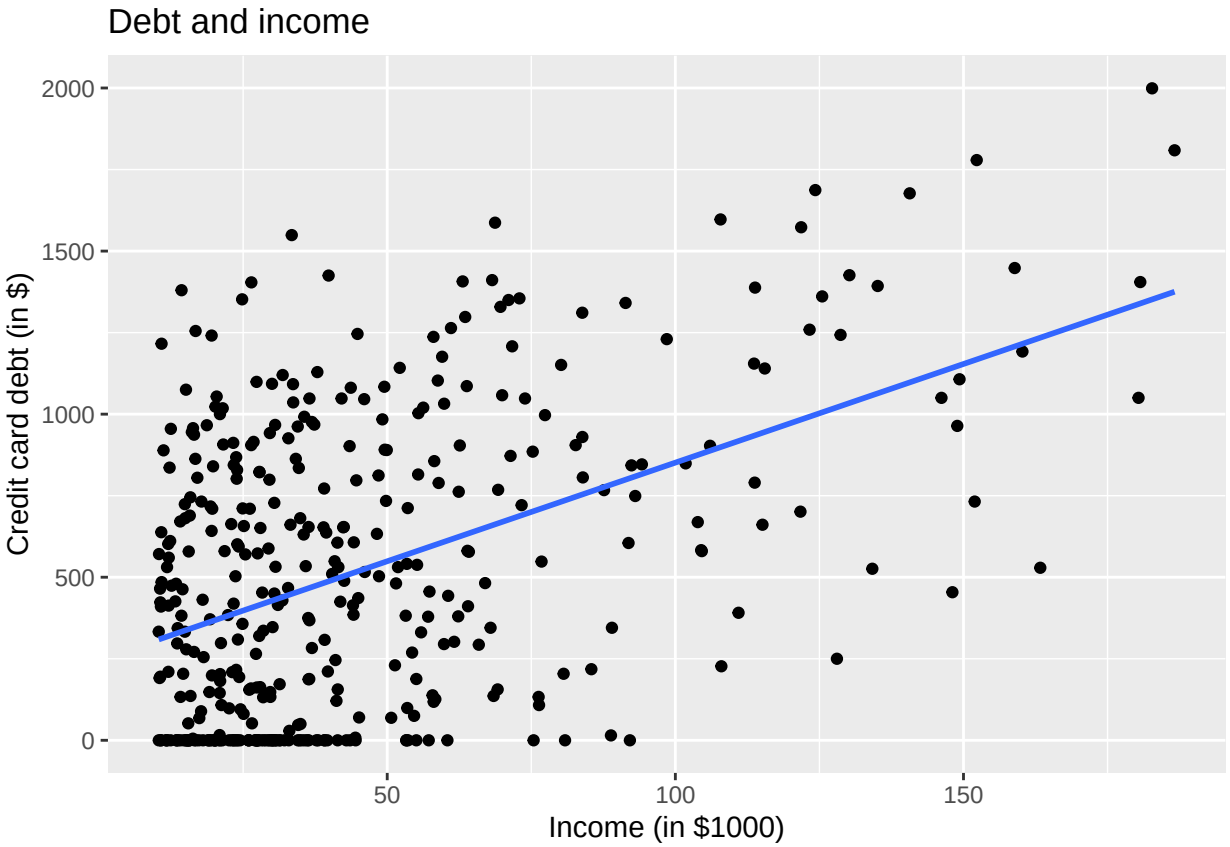
So in our case since `credit_limit` and `income` are highly correlated, if we knew someone's `credit_limit`, we could make pretty good guesses about their `income` as well. Thus, these two variables provide somewhat redundant information. However, we'll leave discussion on how to work with collinear explanatory variables to a more intermediate-level book on regression modeling.

Let's visualize the relationship of the outcome variable with each of the two explanatory variables in two separate plots

```
ggplot(credit_ch6, aes(x = credit_limit, y = debt)) +
  geom_point() +
  labs(x = "Credit limit (in $)",
       y = "Credit card debt (in $)",
       title = "Debt and credit limit") +
  geom_smooth(method = lm, se = FALSE)
```




```
ggplot(credit_ch6, aes(x = income, y = debt)) +
  geom_point() +
  labs(x = "Income (in $1000)", y = "Credit card debt (in $)",
       title = "Debt and income") +
  geom_smooth(method = "lm", se = FALSE)
```



Observe there is a positive relationship between credit limit and credit card debt: as credit limit increases so also does credit card debt. This is consistent with the strongly positive correlation coefficient of 0.862 we computed earlier. In the case of income, the positive relationship doesn't appear as strong, given the weakly positive correlation coefficient of 0.464

To visualize the joint relationship of all three variables simultaneously, we need a 3-dimensional (3D) scatterplot:

1. The numerical outcome variable y `debt` is on the vertical axis
2. The two numerical explanatory variables, x_1 `income` and x_2 `credit_limit`, are on the two axes that form the bottom plane

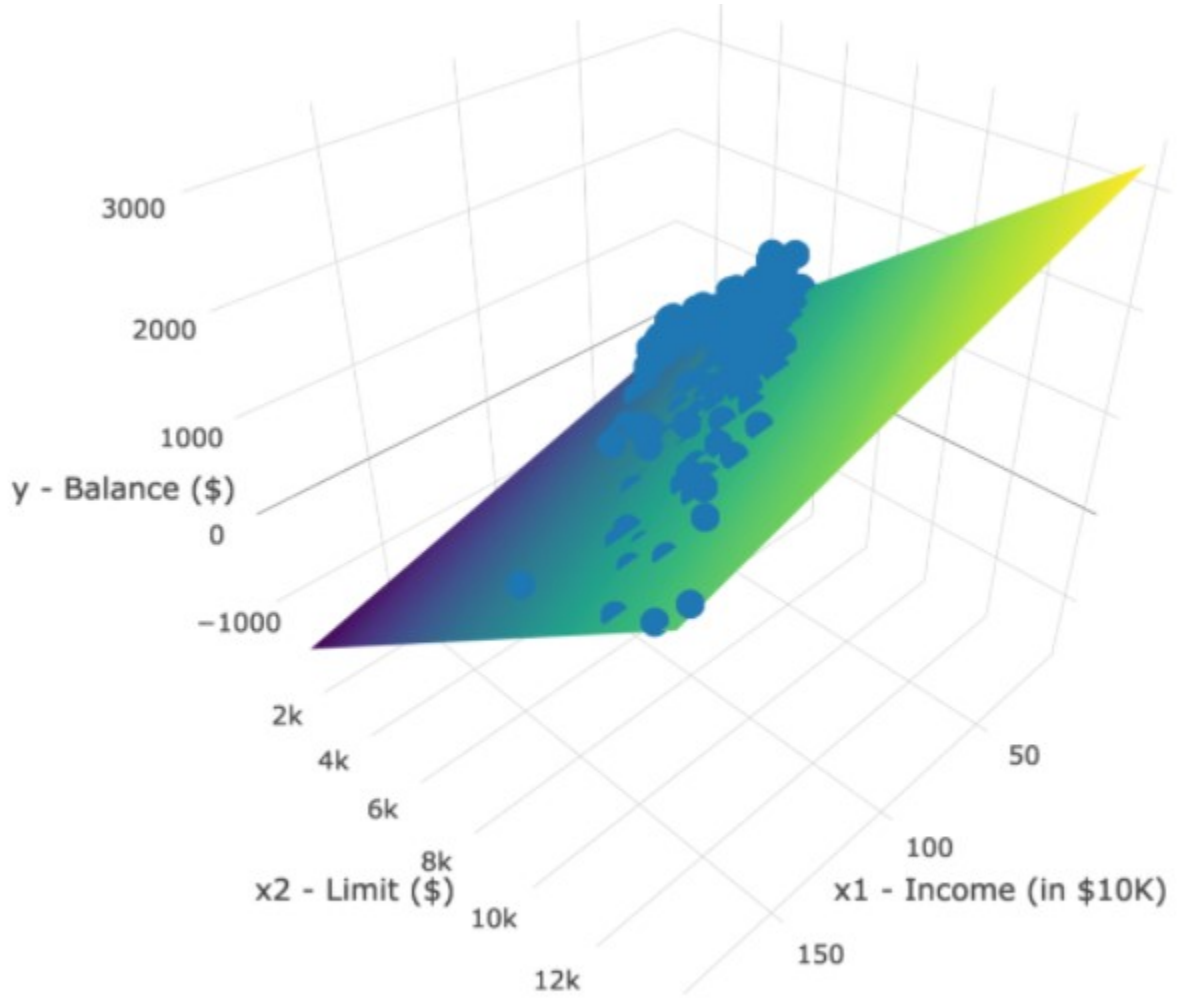


FIGURE 6.6: 3D scatterplot and regression plane.

1.2.2 Regression plane

We'll only consider a model fit with a formula of the form `y ~ x1 + x2`

```
# fit regression model:
debt_model <- lm(debt ~ credit_limit + income, data = credit_ch6)

# get regression table:
get_regression_table(debt_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 credit_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
```

The `intercept` value is -$385.179. This intercept represents the credit card debt for an individual who has `credit_limit` of $0 and `income` of $0. In our data, however, the intercept has no practical interpretation since no individuals had `credit_limit` or `income` values of $0

Second, the `credit_limit` value is $0.264. Taking into account all the other explanatory variables in our model, for every increase of one dollar in `credit_limit`, there is an associated increase of on average $0.26 in credit card debt

Third, `income` = -$7.66. Taking into account all other explanatory variables in our model, for every increase of one unit of `income` ($1000 in actual income), there is an associated decrease of, on average, $7.66 in credit card debt

Putting these results together, the equation of the regression plane that gives us fitted values $\hat{y} = \widehat{\text{debt}}$ is:

$$\begin{aligned} \hat{y} &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \\ \widehat{\text{debt}} &= b_0 + b_{\text{limit}} \cdot \text{limit} + b_{\text{income}} \cdot \text{income} \\ &= -385.179 + 0.263 \cdot \text{limit} - 7.663 \cdot \text{income} \end{aligned}$$

Recall however in the right-hand plot of Figure 6.5 that when plotting the relationship between `debt` and `income` in isolation, there appeared to be a *positive* relationship. In the last discussed multiple regression, however, when *jointly* modeling the relationship between `debt`, `credit_limit`, and `income`, there appears to be a *negative* relationship of `debt` and `income` as evidenced by the negative slope for `income` of -$7.663. What explains these contradictory results? A phenomenon known as *Simpson's Paradox*, whereby overall trends that exist in aggregate either disappear or reverse when the data are broken down into groups. In Subsection 6.3.3 we elaborate on this idea by looking at the relationship between `credit_limit` and credit card `debt`, but split along different `income` brackets.

1.2.3 Observed/fitted values and residuals

Let's also compute all fitted values and residuals for our regression model using the `get_regression_points()`

$$\hat{y} = \widehat{\text{debt}} = -385.179 + 0.263 \cdot \text{limit} - 7.663 \cdot \text{income}$$

```
get_regression_points(debt_model)
```

TABLE 6.11: Regression points (First 10 credit card holders out of 400)

ID	debt	credit_limit	income	debt_hat	residual
1	333	3606	14.9	454	-120.8
2	903	6645	106.0	559	344.3
3	580	7075	104.6	683	-103.4
4	964	9504	148.9	986	-21.7
5	331	4897	55.9	481	-150.0
6	1151	8047	80.2	1127	23.6
7	203	3388	21.0	349	-146.4
8	872	7114	71.4	948	-76.0
9	279	3300	15.1	371	-92.2
10	1350	6819	71.1	873	477.3