Math 300 NTI Lesson 15

Multiple Regression - Two Numerical Predictors

Professor Bradley Warner

June, 2022

Contents

Objectives

1. For two numerical explanatory variables in a linear regression model, conduct exploratory analysis and explain the relationship between the variables.
2. Fit a linear regression model to two numerical explanatory variables using the `lm()` function and interpret the output.

Reading

Chapter 6.2

Lesson

Remember that you will be running this more like a lab than a lecture. You want them using `R` and answering questions. Have them open the notes rmd and work through it together.

Work through the learning checks LC6.2 - 6.3.

- We use `select()` to change the names of the variables as well as to select them.
- Collinearity (or multicollinearity) is a phenomenon where one explanatory variable in a multiple regression model is highly correlated with another. This course doesn't discuss multicollinearity but it impacts the inference portion of the analysis cycle. Math 378 is a course that presents methods to handle multicollinearity.
- We preface our interpretation with the statement, "taking into account all the other explanatory variables in our model" in this section. This means we have to treat the other variables as at a constant value even though collinearity in practice may not allow this. It is only from an interpretation point of view that we use that statement.

- 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. The next lesson discusses this in more depth.

Setup

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

Recreate the analysis done in the book.

```
my_skim <- skim_with(numeric = sfl(hist = NULL))
```

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

Let's look at 5 random rows of data.

```
set.seed(507)
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   218   955         5395   12.5           392    65
## 2   160     0         3000   53.3           235    53
## 3   112     0         2959   28.6           231    60
## 4    77   532         3293   30.6           251    68
## 5   265   651         5107   28.0           380    55
```

```
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, 1~
## $ debt          <int> 333, 903, 580, 964, 331, 1151, 203, 872, 279, 1350, 1407~
## $ credit_limit  <int> 3606, 6645, 7075, 9504, 4897, 8047, 3388, 7114, 3300, 68~
## $ income        <dbl> 14.891, 106.025, 104.593, 148.924, 55.882, 80.180, 20.99~
## $ credit_rating <int> 283, 483, 514, 681, 357, 569, 259, 512, 266, 491, 589, 1~
## $ age           <int> 34, 82, 71, 36, 68, 77, 37, 87, 66, 41, 30, 64, 57, 49, ~
```

```
credit_ch6 %>%
  select(debt, credit_limit, income) %>%
  my_skim() %>%
  print()
```

```
## -- Data Summary ------------------------
##                            Values
## 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
## 1 debt                  0             1  520.   460.    0     68.8  460.   863
## 2 credit_limit          0             1 4736.  2308.  855   3088   4622.  5873.
## 3 income                0             1   45.2   35.2  10.4   21.0   33.1   57.5
##     p100
## 1  1999
## 2 13913
## 3   187.
```

```
## $numeric
##
## -- Variable type: numeric ----------------------------------------------------
##   skim_variable n_missing complete_rate   mean     sd    p0    p25    p50    p75
## 1 debt                  0             1  520.   460.    0     68.8  460.   863
## 2 credit_limit          0             1 4736.  2308.  855   3088   4622.  5873.
## 3 income                0             1   45.2   35.2  10.4   21.0   33.1   57.5
## # ... with 1 more variable: p100 <dbl>
```

```
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
```

LC 6.2 (Objective 1)

(LC6.2) Conduct a new exploratory data analysis with the same outcome variable y being `debt` but with `credit_rating` and `age` as the new explanatory variables x_1 and x_2. Remember, this involves three things:

- Most crucially: Looking at the raw data values.
- Computing summary statistics, such as means, medians, and interquartile ranges.
- Creating data visualizations.

What can you say about the relationship between a credit card holder's debt and their credit rating and age?

Solution:

- Most crucially: Looking at the raw data values.

```
credit_ch6 %>%
  select(debt, credit_rating, age) %>%
  head()
```

```
## # A tibble: 6 x 3
##    debt credit_rating   age
##   <int>         <int> <int>
## 1   333           283    34
## 2   903           483    82
## 3   580           514    71
## 4   964           681    36
## 5   331           357    68
## 6  1151           569    77
```

Computing summary statistics, such as means, medians, and interquartile ranges.

```
credit_ch6 %>%
  select(debt, credit_rating, age) %>%
  my_skim() %>%
  print()
```

```
## -- Data Summary ------------------------
##                            Values
## 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
## 1 debt                  0             1 520.  460.   0  68.8 460. 863  1999
## 2 credit_rating         0             1 355.  155.  93 247.  344  437.  982
## 3 age                   0             1  55.7  17.2 23  41.8  56   70    98
```

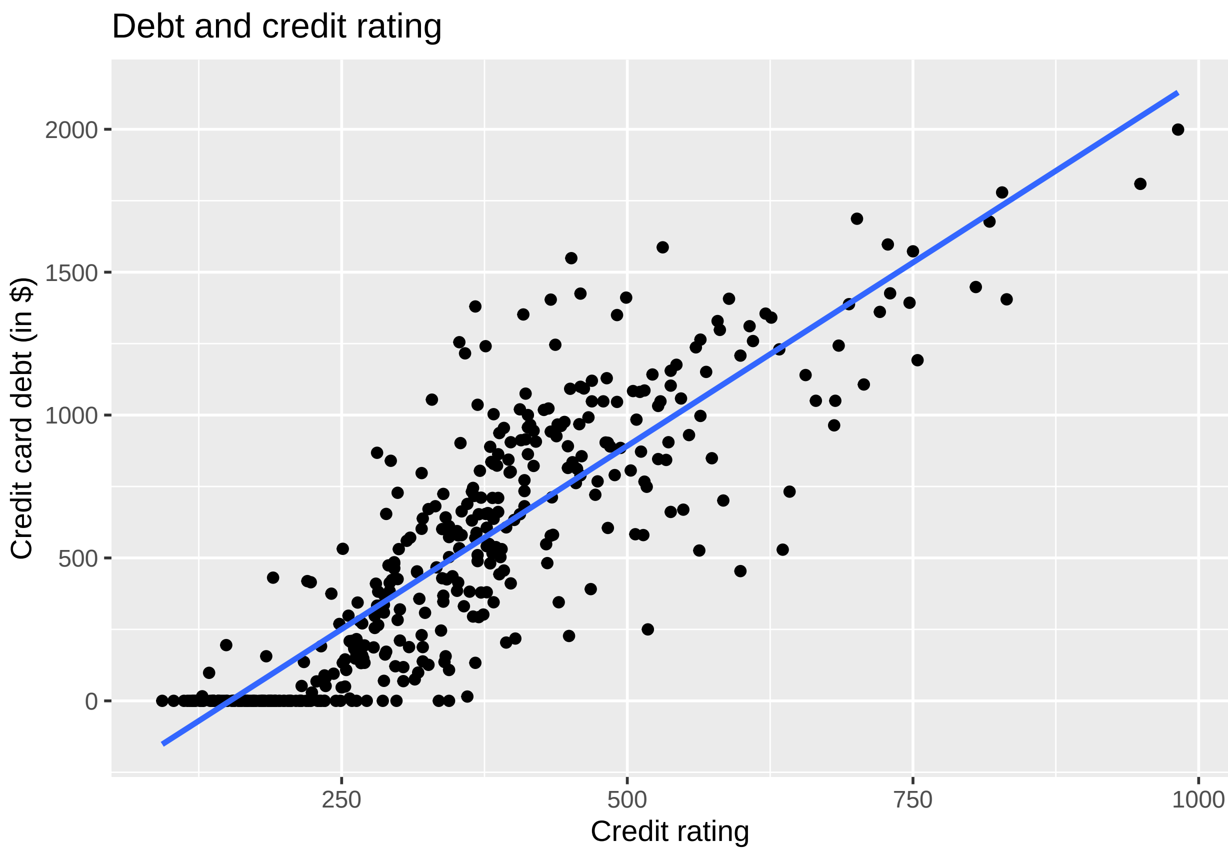
```
## $numeric
##
## -- Variable type: numeric --------------------------------------------------------
##   skim_variable n_missing complete_rate  mean    sd p0   p25  p50  p75 p100
## 1 debt                  0             1 520.  460.   0  68.8 460. 863  1999
## 2 credit_rating         0             1 355.  155.  93 247.  344  437.  982
## 3 age                   0             1  55.7  17.2 23  41.8  56   70    98
```

```
credit_ch6 %>%
  select(debt, credit_rating, age) %>%
  cor()
```

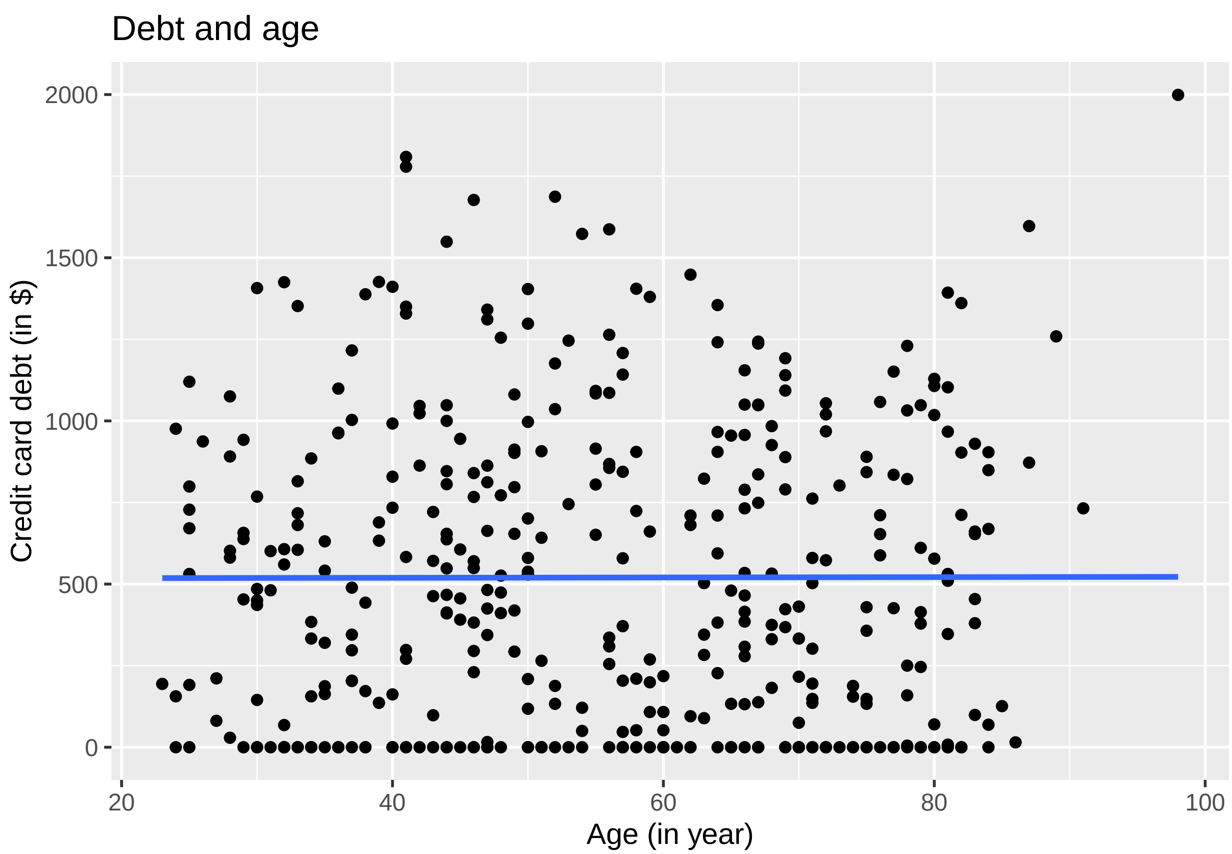
```
##                          debt credit_rating         age
## debt            1.000000000     0.8636252 0.001835119
## credit_rating 0.863625161     1.0000000 0.103164996
## age             0.001835119     0.1031650 1.000000000
```

- Creating data visualizations.

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




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



It seems that there is a positive relationship between one's credit rating and their debt, and very little relationship between one's age and their debt. The is a slight linear relationship between `age` and `credit_rating`.

LC 6.3 (Objective 2)

(LC6.3) Fit a new simple linear regression using `lm(debt ~ credit_rating + age, data = credit_ch6)` where `credit_rating` and `age` are the new numerical explanatory variables x_1 and x_2. Get information about the "best-fitting" regression plane from the regression table by applying the `get_regression_table()` function. How do the regression results match up with the results from your previous exploratory data analysis?

```
# Fit regression model:
debt_model_2 <- lm(debt ~ credit_rating + age, data = credit_ch6)
```

```
# Get regression table:
get_regression_table(debt_model_2)
```

```
## # 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      -270.      44.8       -6.02       0  -358.    -181.
## 2 credit_rating     2.59     0.074     34.8        0     2.45     2.74
## 3 age              -2.35     0.668     -3.52       0    -3.66    -1.04
```

The coefficients for both new numerical explanatory variables x_1 and x_2, `credit_rating` and `age`, are 2.59 and -2.35 respectively, which means that `debt` and `credit_rating` are positively correlated, which matches up with our explanatory analysis. However, `debt` and `age` are negatively correlated but in our exploratory analysis we surmised there was no relationship. When we account for credit rating, the debt tends to decrease with age, for one year increase in age, the debt on average decreases -2.35 with a credit rating held constant.

Documenting software

- File creation date: 2022-06-24
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `skimr` package version: 2.1.4
- `ISLR` package version: 1.4

- `moderndive` package version: 0.5.4