Multiple Regression

Robert Edwards

15 February 2019

1 Introduction

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 individual's 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.

2 Modelling with Two Continuous Covariates

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

- the continuous outcome variable y, the credit card balance of an individual; and
- two explanatory variables x1 and x2, which are an individual's credit limit and income (both in thousands of dollars), respectively.

2.1 Exploratory Data Analysis

Table 1: Summary statistics on credit scores.

Variable	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Balance	520.01	459.76	0	68.75	459.5	863	1999
Income	45.22	35.24	10.35	21.01	33.12	57.47	186.63
Limit	4735.6	2308.2	855	3088	4622.5	5872.75	13913

What is the mean credit Limit?

- Mean Credit Limit = 4735.6

What is the median credit Balance?

- Median Credit Balance = 459.5

What is the percent credit card holders with income greater than $57,470?

- Mean Credit Limit = 0.25

What is the correlation coefficient for the linear relationship between Balance and Limit?

- Cor(Balance, Limit) = 0.8616973

Table 2: Correlation coefficient for the linear relationship between Balance and Limit

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

What would be the verbal interpretation of the correlation coefficient for the linear relationship between Balance and Income?

- Weakly Positive

Collinearity (or **multicolinnearity**) 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 momre higly correlated explantory variables within a multiple regression model essentially provides us with redundant information. Normally, we would remove one of the highly correlated variables, but for the purpose of this example we will ignore the potenital issue.

```
p1 <- ggplot(credit, aes(x = Limit, y = Balance)) +
         geom_point() +
         labs(x = "Credit limit [$]",
              y = "Credit card balance [$]",
              title = "Relationship between balance and credit limit") +
         geom_smooth(method = "lm", se = FALSE)
p2 <- ggplot(credit, aes(x = Income, y = Balance)) +
         geom_point() +
         labs(x = "Credit income [$]",
              y = "Credit card balance [$]",
              title = "Relationship between income and income") +
         geom_smooth(method = "lm", se = FALSE)

grid.arrange(p1, p2, layout_matrix = matrix(seq_len(1*2), nrow = 1, ncol = 2))
```

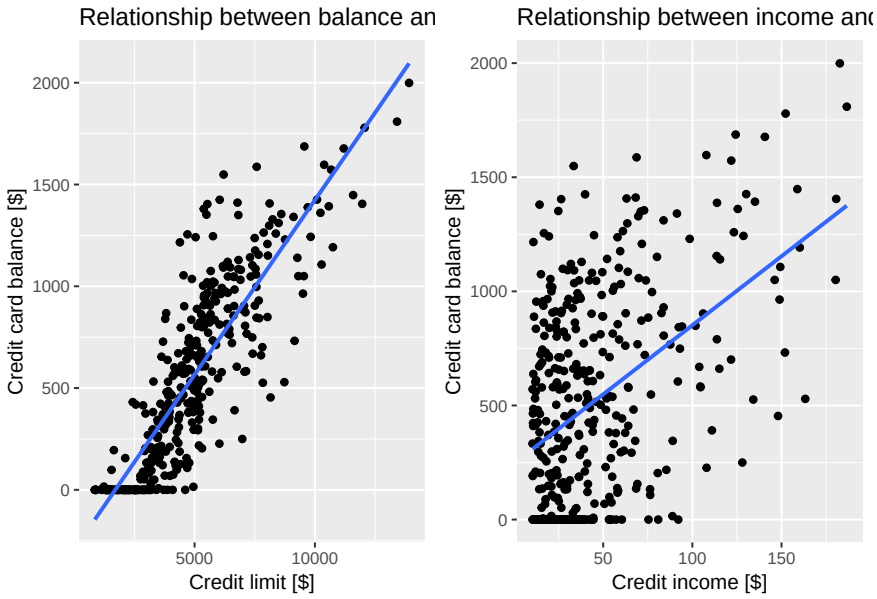

Figure 1: Relationship between balance and explanatory variables: credit limit and income.

What is the relationship between balance and credit limit?

- Positive

What is the relationship between balance and income?

- Positive

The two scatterplots in Figure 3 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 3D scatterplot as follows.

3D scatterplot between balance and explanatory variables: credit limit and income.

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 credit balance of the i^{ith} 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;
- ϵ_i is the i^{th} random error component

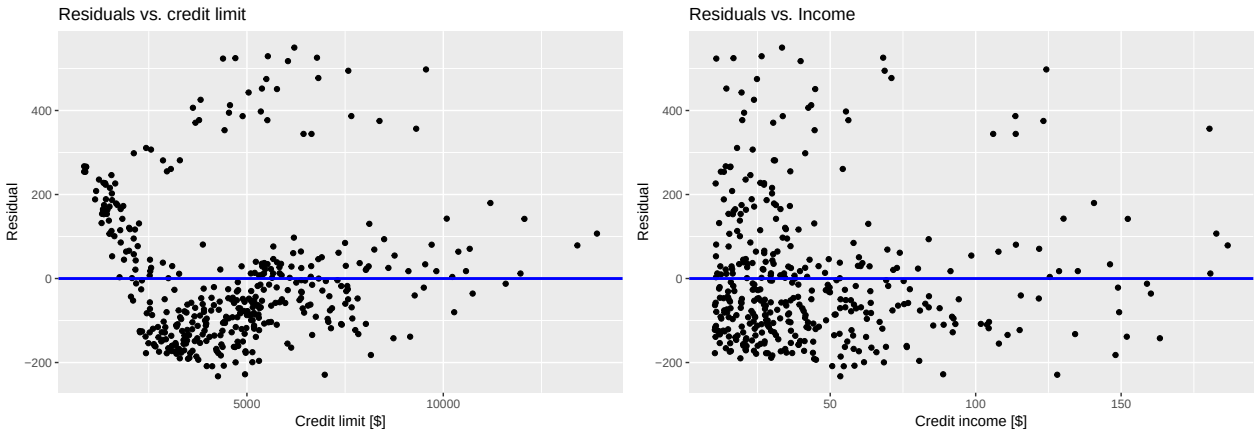

Figure 2: Residual plots of credit limit and income.

Table 3: Estimates of the parameters from the fitted linear regression model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-385.179	19.465	-19.789	0	-423.446	-346.912
Limit	0.264	0.006	44.955	0	0.253	0.276
Income	-7.663	0.385	-19.901	0	-8.420	-6.906

Simpson's Paradox: From Figure 3 we see positive relationships between credit card balance against both credit limit and income. Why do then get a negative coefficient for income ($\widehat{\beta_{income}} = -7.66$)? This is due to a phenomenon known as **Simpson's Paradox**. This occurs when there are trends within different catagories (or groups) of data, but that these trends disappear when the categories are grouped as a whole.

2.3 Assessing Model Fit

Now we need to asses our model assumptions:

1. The deterministic part of the model captures all the non-random structure in the data (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:

```
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 $\widehat{y}$
- the residual error $(y - \widehat{y})$

We can asses our first two model assumptions by producing scatterplots of our residuals against each of our explanaotry variables.

From Figure 2 wesee that the residuals do not have mean zero and constant variability across all values of the explanatory variables.

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

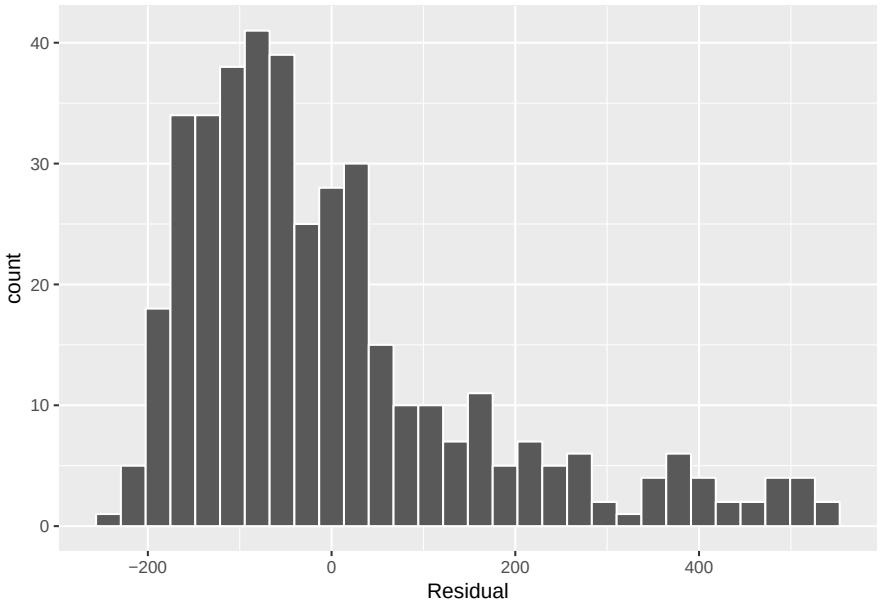

Figure 3: Histogram of residuals.

Do the residuals appear to be normally distributed?

- No, the histogram is right-skewed, which suggests that we are underestimating a lot of credit card holder's balances by a relatively large amount. That is, since the residuals are computed as $y - \widehat{y}$, a lot of the fitted values are much lower than the observed values.

3 Modelling With 1 Continuous & 1 Categorical Explanaotry Variable

We are revisiting the instructor evaluation data set evals. We have examine dthe relationship between teaching score (score) and age (age) and we now introduce the additional (binary) categorical explanatory variable gender (gender):

- the teaching score (score) as our outcome variable y
- age (age) as our continuous explanatory variable x_1
- gender (gender) as our categorical explanatory variable x_2

3.1 Exploratory Data Analysis

First, we subset the evals dataset so we only have score, age, and gender.

```
Observations: 463
Variables: 3
$ score  <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, ...
$ age    <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40,...
$ gender <fct> female, female, female, female, male, male, male, male,...
```

```
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  mean  sd p0 p25 p50 p75 p100
      age 48.37 9.8 29  42  48  57   73

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

How many males are in the data set?

- There are 268 males in the data set

What is the median age in the data set?

- The median age is 48

What is the maximum teaching score of the bottom 25% of the professors?

- Max teaching score is 3.8 of bottom 25% of professors

What is the correlation coefficient between score and age?

- Correlation(score, age) = -0.107032, which suggests a very weak negative relationship.

Figure 4: Instructor evaluation scores by age and gender. The points have been jittered.

From the scatterplot in Figure 2 we can see that:

- There are very few women over the age of 60 in our dataset

- 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

We begin by fitting a parallel regression lines model. This model implies that the slope of the 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:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$
$$y_i = \alpha + \beta_{age} \cdot age_i + \beta_{male} \cdot \mathbb{I}_{\text{male}}(i) + \epsilon_i$$

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;
- age_i is the age of the i^th observation
- β_{male} is the additional term added to α to get the intercept of the regression line for males;
- $\mathbb{I}_{\text{male}}(i)$ is an indicator function such that

$$^I male(i) = \begin{cases} 1 & if\ the\ ith\ observation\ is\ male \\ 0 & Otherwise \end{cases}$$

We fit the parallel regression lines model as follows:

```
par.model <- lm(score ~ age + gender, data = eval.score)
  get_regression_table(par.model) %>%
#    dplyr::select(term, estimate) %>% #necessary to include dplyr here
    kable(caption = "\\label{tab:reg2} Estimates of the parameters from the fitted parallel slopes regr
    kable_styling(latex_options = "HOLD_position")
```

Table 4: Estimates of the parameters from the fitted parallel slopes regression model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.484	0.125	35.792	0.000	4.238	4.730
age	-0.009	0.003	-3.280	0.001	-0.014	-0.003
gendermale	0.191	0.052	3.632	0.000	0.087	0.294

Hence, the regression line for females is given by:

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

while the regression line for males is given by:

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

From parallel regression lines model, what would be the teaching score of a female instructor aged 37?

- teaching score would be 4.1630428

From parallel regression lines model, what would be the teaching score of a male instructor aged 52?

- teaching score would be 4.2234489

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

Figure 5: Instructor evaluation scores by age and gender with parallel regression lines superimposed. The points have been jittered.

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:

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

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

```
int.model <- lm(score ~ age * gender, data = eval.score)
  get_regression_table(int.model) %>%
#    dplyr::select(term, estimate) %>% #necessary to include dplyr here
    kable(caption = "\\label{tab:reg3} Estimates of the parameters from the fitted interaction regressi
    kable_styling(latex_options = "HOLD_position")
```

Table 5: Estimates of the parameters from the fitted interaction regression model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.883	0.205	23.795	0.000	4.480	5.286
age	-0.018	0.004	-3.919	0.000	-0.026	-0.009
gendermale	-0.446	0.265	-1.681	0.094	-0.968	0.076
age:gendermale	0.014	0.006	2.446	0.015	0.003	0.024

Hence, the regression line for females is given by:

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

while the regression line for males is given by:

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

Notice how the interaction model allows for different slopes for females and males (-0.018 and -0.004, respectively). These fitted lines correspond to the fitted lines we first saw in Figure 4 repeated in Figure 6 below but without the jitter.

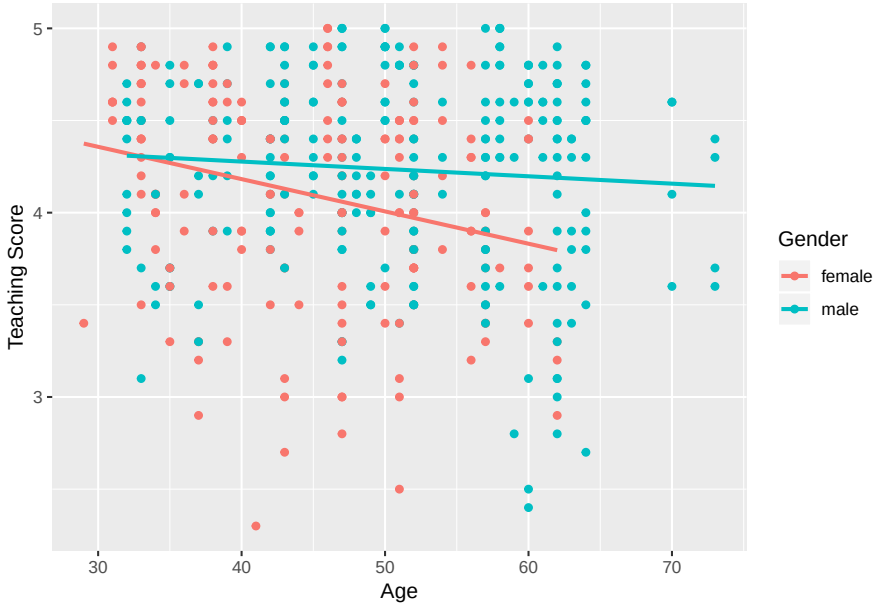

Figure 6: Instructor evaluation scores by age and gender with parallel regression lines superimposed. The points have been jittered.

From interaction regression lines model, what would be the teaching score of a female instructor aged 37?

- teaching score would be 4.2346249

From interaction regression lines model, what would be the teaching score of a male instructor aged 52?

- teaching score would be 4.2293225

Figure 7: Residuals vs. age by gender

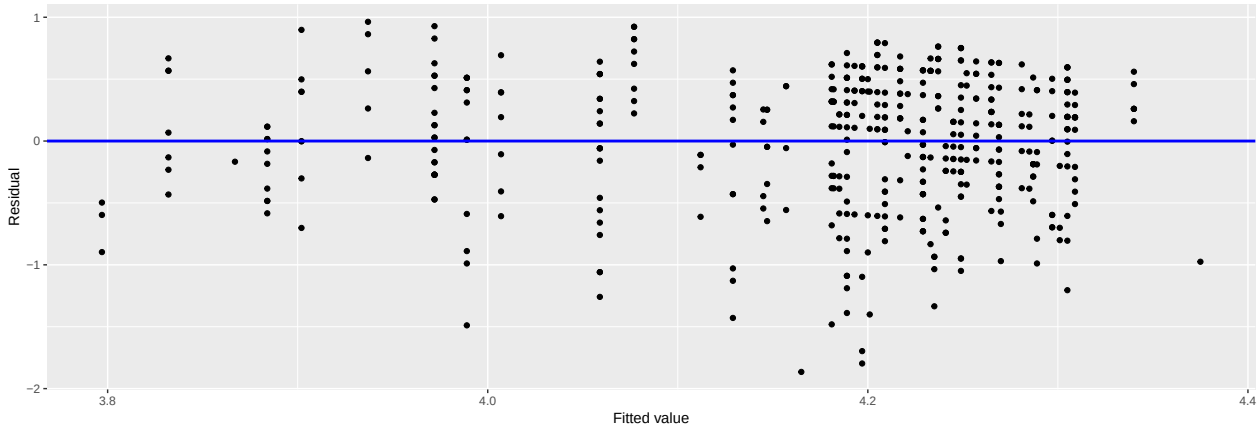

Figure 8: Residuals vs. fitted values

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

3.4 Assessing Model Fit

Now we assess the fit of the model by looking at plots of the residuals of the interaction model.

```
<ggproto object: Class FacetWrap, Facet, gg>
    compute_layout: function
    draw_back: function
    draw_front: function
    draw_labels: function
    draw_panels: function
    finish_data: function
    init_scales: function
    map_data: function
    params: list
    setup_data: function
    setup_params: function
```

```
shrink: TRUE
train_scales: function
vars: function
super:  <ggproto object: Class FacetWrap, Facet, gg>
```

From Figure 7 and Figure 8 we see that the residuals do have mean zero and constant variability across all values of the explanatory variables.

Finally, we can check if the residuals are normally distributed by producing histograms by gender:

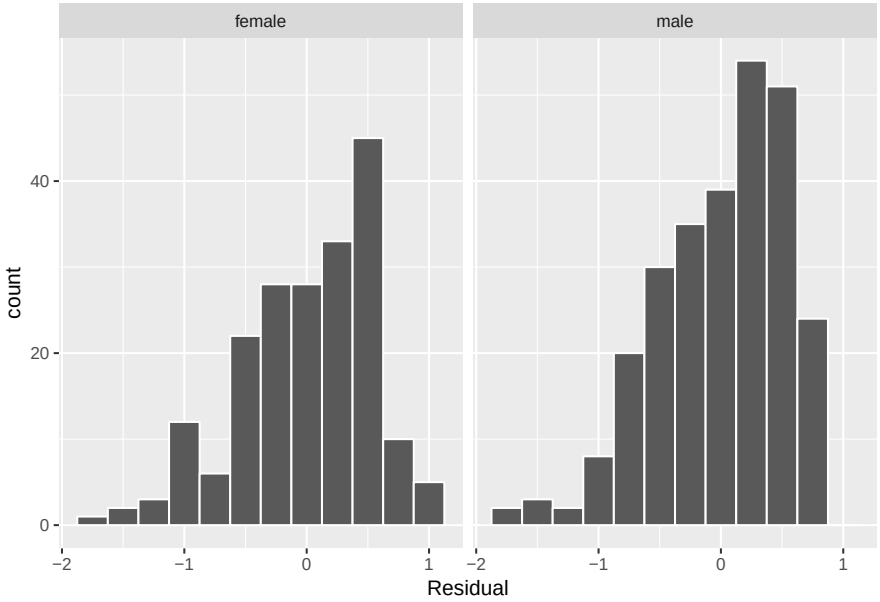

Figure 9: Histogram of residuals by gender.

The residuals do not appear to be normally distributed. Both histograms are left-skewed (and more so for males).