Notes 7 - Linear Regression with Qualitative Predictors

Jillian Morrison

November 10, 2019

What will we cover

- We know that Linear Regresion applies ONLY when:
  - The **Response Variable** (the thing you want to predict) is **Numerical/Quantitative**
- But Linear Regression can work when:
  - The **Predictor Variable** (the thing you are using to predict) is:
    * Numerical/Quantitative (which we have learnt already) **OR**
    * Categorical/Qualitative

We will learn how to interpret the results of Linear Regression when the predictor variable is categorical.

Qualitative Predictors

What happens if your predictor variable is no longer quantitative?

For example:

Consider the `Credit` Dataset which is attached and you want to predict `Balance` (how much a person owes) using `Gender`

```
> library(readr)
> Credit <- read_csv("Credit.csv")
> head(Credit)
# A tibble: 6 x 12
     ID Balance Income Limit Rating Cards   Age Education
  <int>   <int>  <dbl> <int>  <int> <int> <int>     <int>
1     1     333   14.9  3606    283     2    34        11
2     2     903  106.   6645    483     3    82        15
3     3     580  105.   7075    514     4    71        11
4     4     964  149.   9504    681     3    36        11
5     5     331   55.9  4897    357     2    68        16
6     6    1151   80.2  8047    569     4    77        10
# ... with 4 more variables: Gender <chr>, Student <chr>,
#   Married <chr>, Ethnicity <chr>
```

Qualitative Predictors

First of all, what questions are you asking?

How are these questions different from what was asked when the predictor was quantitative?

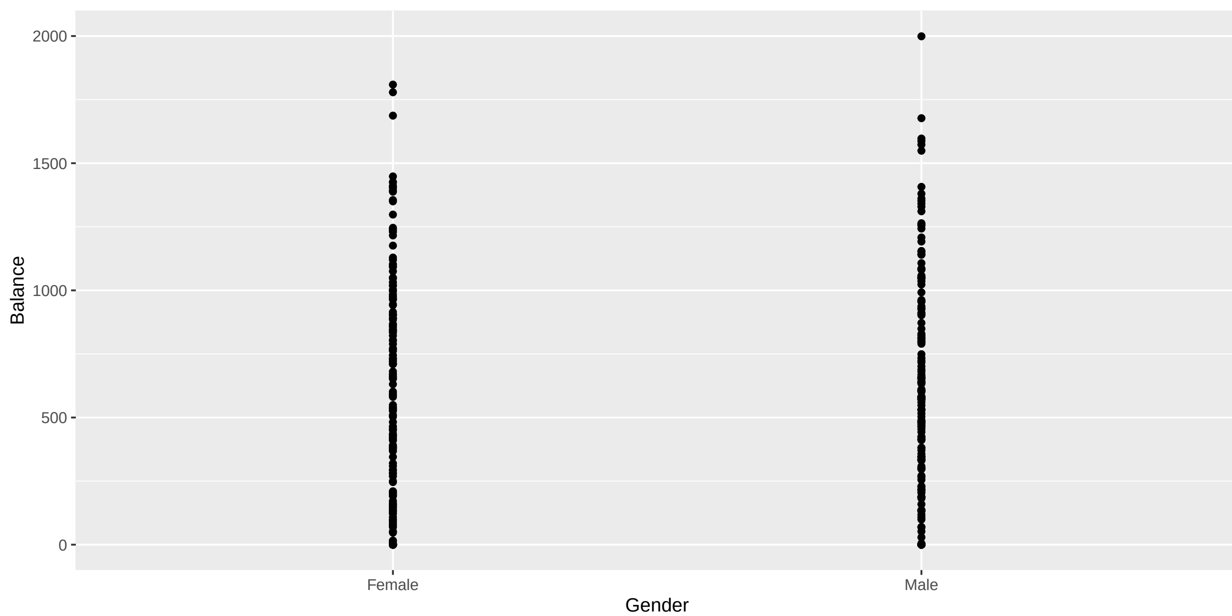
Qualitative Predictors

Questions might be:

1. How strong is the relationship between `Gender` and `Balance`?
2. What is the effect of `Gender` on `Balance`?
3. Is `Gender` a good predictor of `Balance`?
4. How good are the predictions based on your model?

Qualitative Predictors

```
> library(ggplot2)
> ggplot(Credit, aes(x=Gender, y=Balance))+geom_point()
```



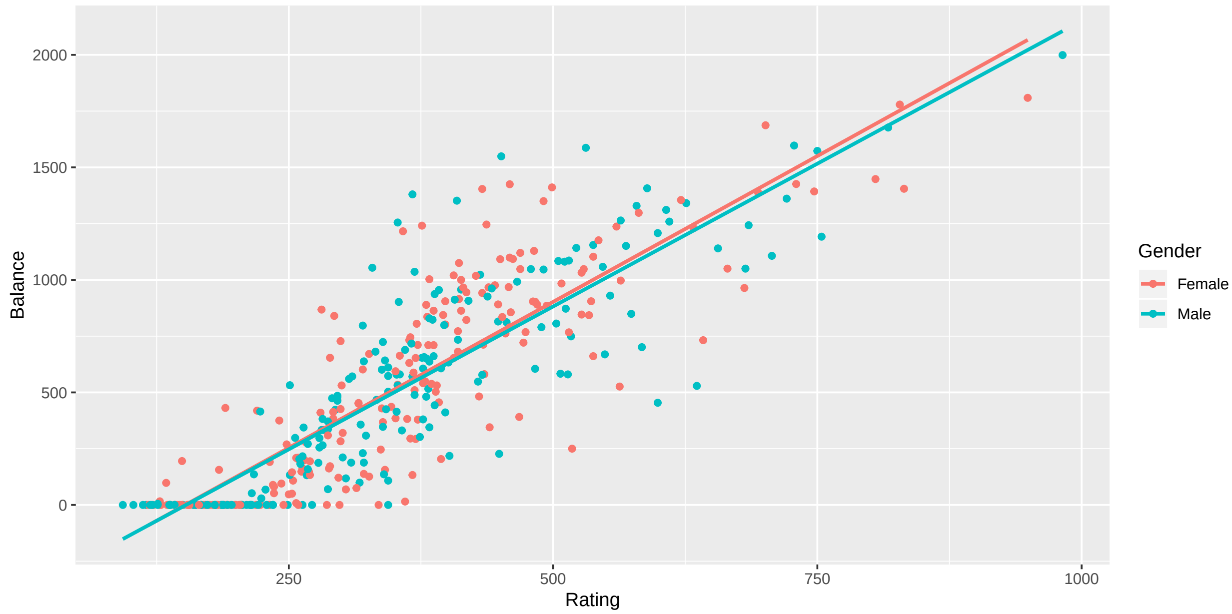
Can you see a difference between female and male?

Qualitative Predictors

What if we added another variable?

- Want to predict Balance using Gender and Income.

```
> library(ggplot2)
> ggplot(Credit, aes(x=Rating, y=Balance, color=Gender))+geom_point()+geom_smooth(method='lm', se=FALSE
```



Do you see a difference in slope for female versus male?

We will come back to this when we consider multiple linear regression (when you have multiple predictors instead of just one).

Fitting the Model

Want to predict Balance using Gender. Let's fit a linear model.

```
> mod1 = lm(Balance~Gender, data=Credit)
> summary(mod1)

Call:
lm(formula = Balance ~ Gender, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-529.54 -455.35  -60.17  334.71 1489.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   529.54      31.99  16.554   <2e-16 ***
GenderMale    -19.73      46.05  -0.429    0.669
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611, Adjusted R-squared:  -0.00205
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

Questions to Answer - Strength of Association

1. How strong is the relationship between Gender and Balance?

- Using $R^2 = 0.0004611$, this is pretty small which means that the association is small.
- If you go back to the plot, you will notice that there wasn't a big difference between Males and Females, they were both spread out over the entire range of Balance.

Questions to Answer - Effect of the qualitative predictor on the response

2. What is the effect of Gender on Balance?

Let's first understand what the coefficients mean:

- If your predictor variable has 2 groups (for example gender is either male or female), we have:
  - Female - Baseline/Comparison group (chosen alphabetically)
  - Male - other group
- So, you should expect to have:
  - Intercept coefficient β_0 - effect of the Baseline/Comparison group
  - Slope Coefficent β_1 - average effect of the difference between the other group and the Baseline group.
- Then:
  - The effect of the other group alone is $\beta_1 + \beta_0$

Questions to Answer - Effect of the qualitative predictor on the response

Let's translate this. Recall:

```
(Intercept)  GenderMale
  529.53623   -19.73312
```

- Determining which group is Baseline:
  - Since the the slope coefficeint is for Male, it implies that this is the difference between Male and the Baseline (FEMALE)
- Interpreting Intercept:
  - So, the intercept is 529.54 which means that the average credit balance for the baseline (FEMALE) is $529.54.
- Interpreting the slope coefficient:
  - Then, the slope coefficient for Male is -19.23 which means that males are estimated to have $19.23 *LESS* in credit balance than the baseline (FEMALE).
  - In other words, males are estimated to have $529.54 - $19.23 = $509.80 in credit balance.

Questions to Answer - Is your predictor variable a good predictor of the response?

- Here, again, we are asking the the slopes are sufficiently bigger than zero (0). In other words, does your variable really matter when it comes to predicting your response?
  - We look at `Pr(>|t|)` for the slope. In this case, it is 0.669.
  - Since this is bigger than 0.05, we can conclude that this slope (of -19.23) is not very different from 0

```
$coefficients
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 529.53623   31.98819 16.5541153 3.312981e-47
GenderMale  -19.73312   46.05121 -0.4285039 6.685161e-01
```

- Interpretation:
  - This slope is the difference of Male from the baseline (FEMALE)
  - Since this is not different from 0 (in other words: it is essentially 0), it means that there really is no difference between male and the baseline (FEMALE) when it comes to Balance.
  - So, `Gender` is not a good predictor of `Balance`

Final Question : How good are the predictions based on your model?

- **ANSWER** is RSE (or RMSE)

Here, the RSE is 460.2.

AGAIN - We do not expect great prediction power because R^2 is small and difference between Male and the baseline (FEMALE) is not huge.

ALSO - We can use this to select the best model if we are interested in the predictor variable that gives the best predictions.

```
> summary(mod1)

Call:
lm(formula = Balance ~ Gender, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-529.54 -455.35  -60.17  334.71 1489.20 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   529.54      31.99  16.554   <2e-16 ***
GenderMale    -19.73      46.05  -0.429    0.669    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611, Adjusted R-squared:  -0.00205 
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

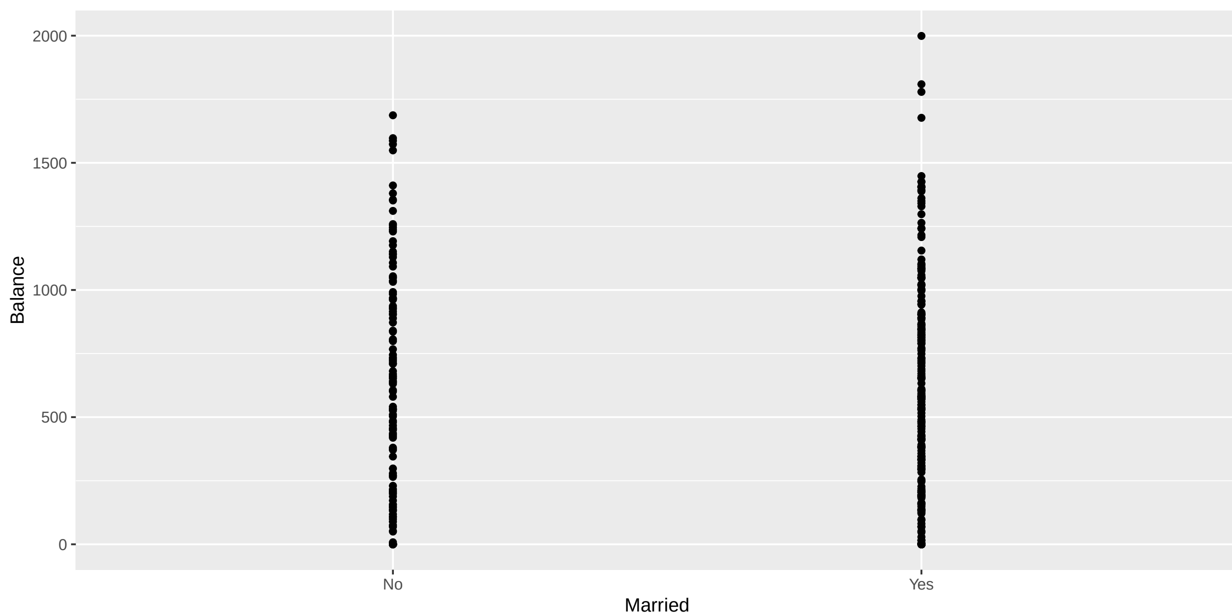
Exercise - Use Married to predict Balance.

1. How strong is the relationship between `Married` and `Balance`?
2. What is the effect of `Married` on `Balance`?
3. Is `Gender` a good predictor of `Balance`?
4. How good are the predictions based on your model? Which of `Married` or `Gender` more accurately predicts 'Balance?

Exercise Solutions

ALWAYS PLOT FIRST - EDA!!!

```
> ggplot(Credit, aes(x=Married, y=Balance))+geom_point()
```



Exercise Solutions

```
> mod2 = lm(Balance~Married, data=Credit)
> summary(mod2)

Call:
lm(formula = Balance ~ Married, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-523.29 -451.03  -60.12  345.06 1481.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  523.290     36.974  14.153   <2e-16 ***
```

```
MarriedYes    -5.347     47.244  -0.113     0.91
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.3 on 398 degrees of freedom
Multiple R-squared:  3.219e-05, Adjusted R-squared:  -0.00248
F-statistic: 0.01281 on 1 and 398 DF,  p-value: 0.9099
```

Exercise Solutions

1. How strong is the relationship between `Married` and `Balance`?

   - $R^2 = 0.000032$, which is very weak since it is close to 0

2. What is the effect of `Married` on `Balance`?

   - The Baseline is `No` since the slope coefficent is for `Yes`
   - Intercept is 523.29, so Not Married people are estimated to have an average `Balance` of $523.29
   - Slope for `Yes` is -5.35, so Married people are estimated to have an average Balance of $523.29-$5.35 =$517.90. This is $5.35 less on average than the Not Married people.

3. Is `Gender` a good predictor of `Balance`?

   - Pr(>|t|) for the slope coefficient is 0.91, which is bigger than 0.05

   - So, the slope is essentially 0 (or not different from 0)
   - This means that `Married` is not a good predictor of `Balance`

4. How good are the predictions based on your model? Which of `Married` or `Gender` more accurately predicts 'Balance?

   - RSE is 460. This is slightly better than when `Gender` was used (460.2).
   - These are both bad predictors of `Balance` and so we expect predictions to be bad

What if we had a qualitative predictor with more than 2 groups?

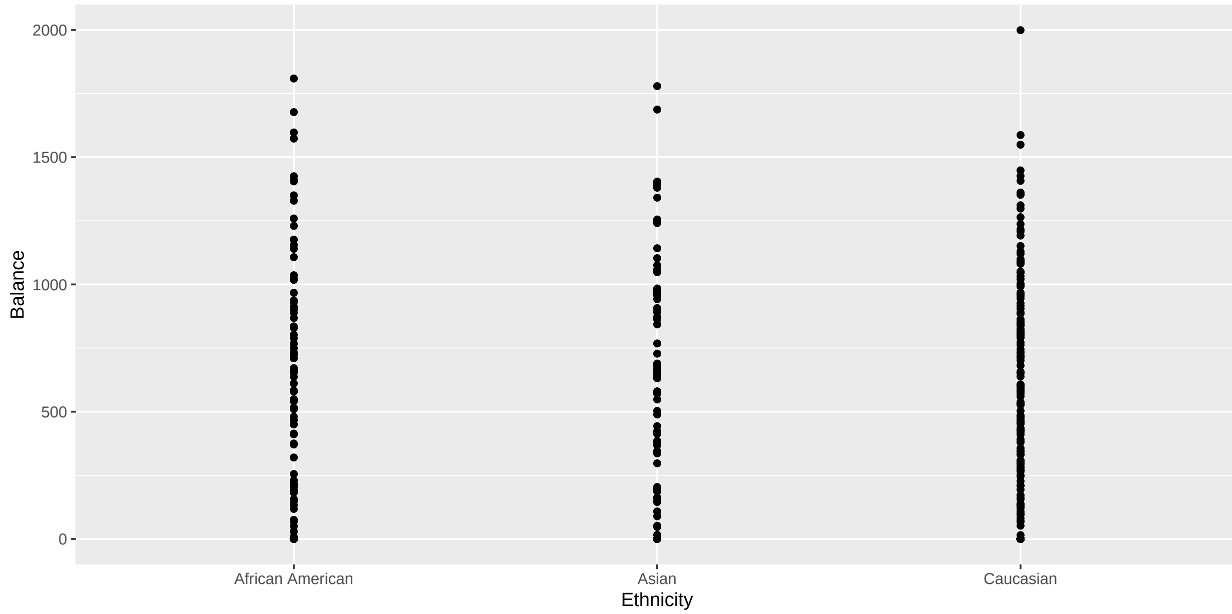
- Same thing applies

  - Intercept - The baseline group
  - slopes - difference of group from the baseline group

Predictor Variable with more than 2 groups

Example: Let's use `Ethnicity` to predict `Balance`

```
> ggplot(Credit, aes(x=Ethnicity, y=Balance))+geom_point()
```



Do you see a difference? Probably not…

Predictor Variable with more than 2 groups

```
> mod3 = lm(Balance~Ethnicity, data=Credit)
> summary(mod3)

Call:
lm(formula = Balance ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774
EthnicityCaucasian   -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Exercise Solutions

1. How strong is the relationship between `Ethnicity` and `Balance`?

- $R^2 = 0.000219$, which is very weak since it is close to 0

2. What is the effect of `Ethnicity` on `Balance`?

   - The Baseline is `African American` since the slope coefficents are for `Asian` and `Caucasian`
   - Intercept is 531.0, so African American people are estimated to have an average `Balance` of $531.0
   - Slope for `Asian` is -18.7, so Asian people are estimated to have an average Balance of $531.0-$18.7 =$512.3. This is $18.7 less on average than African Amerian people.
   - Slope for `Caucasian` is -12.5, so Caucasian people are estimated to have an average Balance of $531.0-$12.5 =$518.5. This is $12.5 less on average than African Amerian people.
   - **NOTE** - Notice that there is no comparison made directly between Caucasian and Asian people. You can make this directly after interpreting your slopes OR you can manually change your baseline. To change your baseline, use:
     - `New_datset_Name <- Credit %>% mutate(Ethnicity = relevel(Ethnicity, ref = "Caucasian"))`
     - This will set `Caucasian` as the baseline group

Exercise Solutions

3. Is `Ethnicity` a good predictor of `Balance`?

   - Pr(>|t|) for the slope coefficient is 0.77 and 0.83 for `Asian` and `Caucasion` respectively. These are bigger than 0.05

   - So, the slopes are essentially 0 (or not different from 0)
   - This means that `Asian` and `Caucasion` are no different from `African American` and so `Ethnicity` is not a good predictor of `Balance`

4. How good are the predictions based on your model? Which of `Married` or `Gender` or `Ethnicity` more accurately predicts `Balance`?

   - RSE is 461. This is WORSE than when `Gender` (460.2) and `Married` (460) was used
   - These are all bad predictors of `Balance` and so we expect predictions to be bad. Though, `Ethnicity` is slightly worse than them all.