Notes 8 - Multiple Linear Regression

Jillian Morrison

November 15, 2019

What will we cover

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

But, what if you wanted to use multiple predictor variables to predict the same response variable. For example, we want to use `Limit`, `Age` and `Gender` to predict `Balance` from the `Credit` dataset we used in notes 7.

We will learn how to interpret the results of Linear Regression when there are multiple predictor variables of the same response. This is called **Multiple Linear Regression**

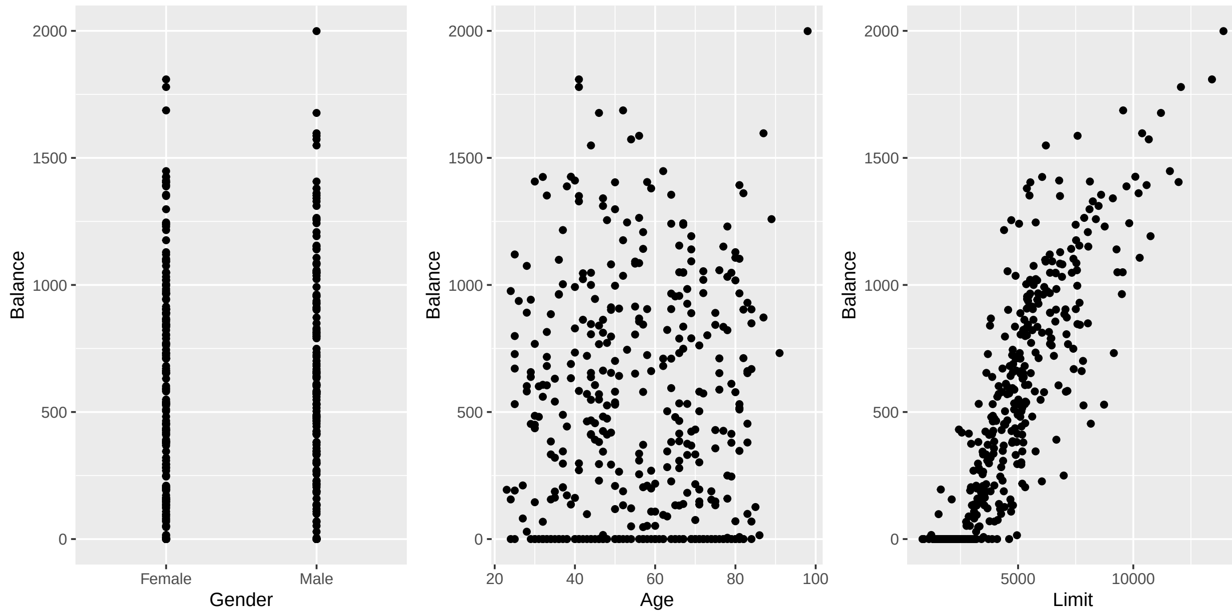
Multiple Predictors of the same Response

Let's remind ourselves of the dataset:

```
> library(readr)
> library(dplyr)
> Credit <- read_csv("Credit.csv")
> Credit2<-Credit%>%select(Balance, Limit, Gender, Age)  ##Selecting only the variables I will use in th
> head(Credit2)
# A tibble: 6 x 4
  Balance Limit Gender   Age
    <int> <int> <chr>  <int>
1     333  3606 Male      34
2     903  6645 Female    82
3     580  7075 Male      71
4     964  9504 Female    36
5     331  4897 Male      68
6    1151  8047 Male      77
```

Multiple Predictors of the same Response

```
> library(gridExtra)
> library(ggplot2)
> a=ggplot(Credit, aes(x=Gender, y=Balance))+geom_point()
> b=ggplot(Credit, aes(x=Age, y=Balance))+geom_point()
> c=ggplot(Credit, aes(x=Limit, y=Balance))+geom_point()
> grid.arrange(a,b,c, nrow=1)
```



We have already seen that `Gender` was horrible for predicting `Balance`, but we haven't actually seen the rest.

Multiple Predictors of the same Response

First, why do we care about using multiple linear regression versus simple linear regression?

1. In Simple Linear Regression, we estimated the model for each variable without consideration of other variables that might matter to predicting the response.
   - So, we know how each variable affects the response in ISOLATION.
   - This can lead to misleading estimates when you only consider one variable in ISOLATION.
2. So, we would like to know how the response is affected by these variables, but when we consider them together in the same model.

Multiple Predictors of the same Response

Let's fit the model:

```
> m1=lm(Balance~Age+Gender+Limit, data=Credit2)
> summary(m1)

Call:
lm(formula = Balance ~ Age + Gender + Limit, data = Credit2)

Residuals:
    Min      1Q  Median      3Q     Max
-690.26 -151.44   -6.67  129.49  762.07

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -167.18265   45.34243  -3.687 0.000258 ***
Age           -2.29261    0.67309  -3.406 0.000726 ***
GenderMale   -12.53603   23.08899  -0.543 0.587474
Limit          0.17334    0.00503  34.459  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 230.7 on 396 degrees of freedom
Multiple R-squared:   0.75, Adjusted R-squared:  0.7481
F-statistic: 396.1 on 3 and 396 DF,  p-value: < 2.2e-16
```

- Notice that the results look very similar to all that we have done before.
- Also remember that Gender is Qualitative/Categorical and Age, Limit are Numerical

Multiple Linear Regression - Interpretation

1. How strong is the relationship between `Gender, Age and Limit` and `Balance`?

   - This time we use the Adjusted R-squared. This is adjusted for the number of predictors in the model.
   - adjusted R^2 is 0.748. So, the predictors have a strong relationship with `Balance`

2. What are the effects of `Gender, Age and Limit` on `Balance`?

   1. Quantitative/Numerical predictors - `Age` and `Limit`
      - These are interpreted merely as slopes - like we did before.
      - For a unit increase in predictor, the response increases by the value of the slope.
        - coefficient for `Age` is -2.29. As `Age` increases by 1 unit (year in this case), the `Balance` decreases by $2.29.
        - coefficient for `Limit` is 0.173. As `Limit` increases by 1 unit (dollar in this case), the `Balance` increases by $0.173
   2. Qualitative/Categorical predictors - `Gender`
      - The baseline group is `Female`.
      - The slope for `Male` is -12.5. This means that the `Balance` for `Male` is $12.50 less on average than for the baseline (`FEMALE`)
      - Note that you cannot interpret the intercept the same here as you did in simple linear regression because the intercept also takes in consideration other variables.

NOTE: These slopes/effects are interepreted assuming we hold all other variables in the model CONSTANT

Multiple Linear Regression - Interpretation

3. Are `Gender, Age and Limit` good predictors of `Balance`?

   - Let's look at Pr(>|t|) for these coefficients.
     - Age: Pr(>|t|) = 0.00073
     - Limit: Pr(>|t|) = $< 2 \times 10^{-16}$
     - GenderMale: Pr(>|t|) = 0.58747
   - Interpretations

  - Since Pr(>|t|) for `Age` and `Limit` is less than 0.05, we have evidence to say that these coefficients/slopes are different from 0. `Age` and `Limit` contribute to `Balance` -Since Pr(>|t|) for `Gender` is more than 0.05, this slope is NOT different from 0. `Gender` does NOT contribute to `Balance`

4. How good are the predictions based on your model?

  - RSE is 231.
  - We can compare this model to the other 3 models we fit in the last notes:
    - `Ethnicity` to predict `Balance`: RSE= 461
    - `Gender` to predict `Balance`: RSE= 460.2
    - `Married` to predict `Balance`: RSE= 460

So, this model with multiple predictors better predicts `Balance` than the other models with one predictors that we previously fit, since it has the smallest RSE.

Model Selection Procedures

Notice that we have selected the best model, **for making the most accurate predictions**, time and time using RSE.

Let's formalize this a bit.

- What if we:

  1. fit many models by starting with the model with the most predictors of the response. Then, in every step, we remove a predictor. Every time we calculate the RSE. We compare all the models and choose the one with the smallest RSE.
     - This is called **BACKWARD SELECTION**
  2. fit many models by starting with the model with one predictor of the response. Then, in every step, we add a predictor until we have added all the possible predictors. Every time we calculate the RSE. We compare all the models and choose the one with the smallest RSE.
     - This is called **FORWARD SELECTION**
  3. fit every possible model to predict the respnse. Every time we calculate the RSE. We compare all the models and choose the one with the smallest RSE.
     - This is called **BEST SUBSET SELECTION**

Forward Selection

You can use Cross Validation from notes 6 to select the best model using RSE/RMSE, let's use AIC here as an example. What you need to know is: - Lower AIC is better - Higher AIC is worse

```
> library(MASS)
> m1_subsets= stepAIC(m1, direction = "forward", trace=FALSE)
> #summary(m1_subsets)
> m1_subsets$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Balance ~ Age + Gender + Limit
```

```
Final Model:
Balance ~ Age + Gender + Limit


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                        396   21082895 4357.003
```

We see that the best model using forward selection has `Limit` and `Age` ONLY!

Backward Selection

```
> library(MASS)
> m1_subsets= stepAIC(m1, direction = "backward", trace=FALSE)
> #summary(m1_subsets)
> m1_subsets$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Balance ~ Age + Gender + Limit

Final Model:
Balance ~ Age + Limit

      Step Df Deviance Resid. Df Resid. Dev      AIC
1                            396   21082895 4357.003
2 - Gender  1 15694.42       397   21098589 4355.301
```

We see that the best model using backward selection also has `Limit` and `Age` ONLY!

Appendix

Applying Cross Validation Using Leave-One-Out Cross Validation

Using function for Cross-Validation

```
> library(caret)
> data_ctrl <- trainControl(method = "LOOCV")      # Type of validation
> Best_LOOCV <- train(Balance~Age+Gender+Limit,    # model to fit
+                     data=Credit2,
+                     trControl = data_ctrl,
+                     method = "leapBackward",     # specifying selection method
+                     na.action = na.pass)         # pass missing data to model - some models will handl
```

Seeing the results - Metrics

```
> Best_LOOCV$results  ##See the details of the RSE values etc.
  nvmax     RMSE  Rsquared      MAE
1     2 231.2561 0.7463709 177.3919
```

```
2     3 231.7540 0.7452824 177.7984
3     4 231.7540 0.7452824 177.7984
```

- We see the best:
  - 2 variable model (1 response 1 predictor) - has RMSE 231.3
  - 3 variable model (1 response 2 predictors) - has RMSE 231.8
  - 4 variable model (1 response 3 predictor) - has RMSE 231.8
    - this has the same RMSE as the 3 variable model- so is not 'best'

Seeing the summary - best model

```
> summary(Best_LOOCV$finalModel)
Subset selection object
3 Variables  (and intercept)
           Forced in Forced out
Age            FALSE      FALSE
GenderMale     FALSE      FALSE
Limit          FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: backward
         Age GenderMale Limit
1  ( 1 ) " " " "        "*"
2  ( 1 ) "*" " "        "*"
```

- We see the best:
  - 2 variable model (1 response 1 predictor) with RMSE 231.3 has `Limit`
  - 3 variable model (1 response 2 predictors) with RMSE 231.8 has `Limit` and `Age`

Note: These are from the asterisx

Fitting the best model Let's choose the model with `Limit` only since that has the smallest overall RMSE

```
> model=lm(Balance~Limit, data=Credit2)
> summary(model)

Call:
lm(formula = Balance ~ Limit, data = Credit2)

Residuals:
    Min      1Q  Median      3Q     Max
-676.95 -141.87  -11.55  134.11  776.44

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.928e+02  2.668e+01  -10.97   <2e-16 ***
Limit        1.716e-01  5.066e-03   33.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 233.6 on 398 degrees of freedom
Multiple R-squared:  0.7425,    Adjusted R-squared:  0.7419
F-statistic:  1148 on 1 and 398 DF,  p-value: < 2.2e-16
```