Multiple Linear Regression: Credit Data

Prof. Kapitula

10/15/2020

read: <https://moderndive.com/6-multiple-regression.html> read: Chapter 6 - 8 in Regression and Other Stories

## Credit Data Analysis

library(ISLR) #contains the credit data  
library(tidyverse)  
library(GGally)  
library(moderndive)  
library(skimr)  
library(rstanarm)  
library(bayesplot)  
theme\_set(bayesplot::theme\_default())

data(Credit, package = "ISLR")  
Credit <- as\_tibble(Credit)  
glimpse(Credit)

## Rows: 400  
## Columns: 12  
## $ ID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …  
## $ Income <dbl> 14.891, 106.025, 104.593, 148.924, 55.882, 80.180, 20.996, …  
## $ Limit <int> 3606, 6645, 7075, 9504, 4897, 8047, 3388, 7114, 3300, 6819,…  
## $ Rating <int> 283, 483, 514, 681, 357, 569, 259, 512, 266, 491, 589, 138,…  
## $ Cards <int> 2, 3, 4, 3, 2, 4, 2, 2, 5, 3, 4, 3, 1, 1, 2, 3, 3, 3, 1, 2,…  
## $ Age <int> 34, 82, 71, 36, 68, 77, 37, 87, 66, 41, 30, 64, 57, 49, 75,…  
## $ Education <int> 11, 15, 11, 11, 16, 10, 12, 9, 13, 19, 14, 16, 7, 9, 13, 15…  
## $ Gender <fct> Male, Female, Male, Female, Male, Male, Female, Male, …  
## $ Student <fct> No, Yes, No, No, No, No, No, No, No, Yes, No, No, No, No, N…  
## $ Married <fct> Yes, Yes, No, No, Yes, No, No, No, No, Yes, Yes, No, Yes, Y…  
## $ Ethnicity <fct> Caucasian, Asian, Asian, Asian, Caucasian, Caucasian, Afric…  
## $ Balance <int> 333, 903, 580, 964, 331, 1151, 203, 872, 279, 1350, 1407, 0…

Check out 5 random cases.

Credit %>% sample\_n(size = 5)

## # A tibble: 5 x 12  
## ID Income Limit Rating Cards Age Education Gender Student Married  
## <int> <dbl> <int> <int> <int> <int> <int> <fct> <fct> <fct>   
## 1 393 26.0 2308 196 2 24 10 " Mal… No No   
## 2 340 149. 10278 707 1 80 16 " Mal… No No   
## 3 254 85.4 5182 402 6 60 12 " Mal… No Yes   
## 4 125 29.6 2529 192 1 30 12 "Fema… No Yes   
## 5 371 35.6 6135 466 4 40 12 " Mal… No No   
## # … with 2 more variables: Ethnicity <fct>, Balance <int>

Credit %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 400 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 4 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

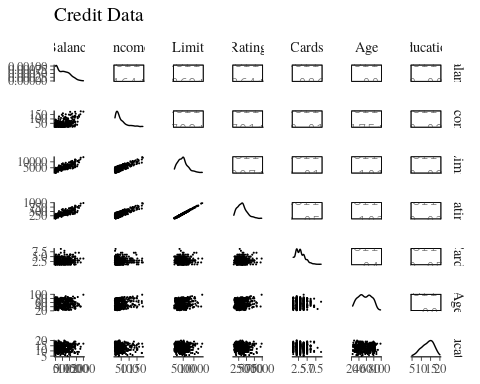
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| Gender | 0 | 1 | FALSE | 2 | Fem: 207, Ma: 193 |
| Student | 0 | 1 | FALSE | 2 | No: 360, Yes: 40 |
| Married | 0 | 1 | FALSE | 2 | Yes: 245, No: 155 |
| Ethnicity | 0 | 1 | FALSE | 3 | Cau: 199, Asi: 102, Afr: 99 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| ID | 0 | 1 | 200.50 | 115.61 | 1.00 | 100.75 | 200.50 | 300.25 | 400.00 | ▇▇▇▇▇ |
| Income | 0 | 1 | 45.22 | 35.24 | 10.35 | 21.01 | 33.12 | 57.47 | 186.63 | ▇▂▁▁▁ |
| Limit | 0 | 1 | 4735.60 | 2308.20 | 855.00 | 3088.00 | 4622.50 | 5872.75 | 13913.00 | ▆▇▃▁▁ |
| Rating | 0 | 1 | 354.94 | 154.72 | 93.00 | 247.25 | 344.00 | 437.25 | 982.00 | ▆▇▃▁▁ |
| Cards | 0 | 1 | 2.96 | 1.37 | 1.00 | 2.00 | 3.00 | 4.00 | 9.00 | ▇▇▂▁▁ |
| Age | 0 | 1 | 55.67 | 17.25 | 23.00 | 41.75 | 56.00 | 70.00 | 98.00 | ▆▇▇▇▁ |
| Education | 0 | 1 | 13.45 | 3.13 | 5.00 | 11.00 | 14.00 | 16.00 | 20.00 | ▂▅▇▇▂ |
| Balance | 0 | 1 | 520.02 | 459.76 | 0.00 | 68.75 | 459.50 | 863.00 | 1999.00 | ▇▅▃▂▁ |

## Scatterplot Matrix

ggpairs(Credit, columns = c(12,2:7), title = "Credit Data",  
 lower=list(continuous=wrap("points", size=0.1)))



## Student Only Model Using Standard LM and Modern Dive Output

lm\_fit <- lm(Balance ~ Student, data = Credit)  
# Get regression table:  
get\_regression\_table(lm\_fit, print=TRUE)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std\_error | statistic | p\_value | lower\_ci | upper\_ci |
| intercept | 480.369 | 23.434 | 20.499 | 0 | 434.300 | 526.439 |
| StudentYes | 396.456 | 74.104 | 5.350 | 0 | 250.771 | 542.140 |

get\_regression\_summaries(lm\_fit, print=T)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r\_squared | adj\_r\_squared | mse | rmse | sigma | statistic | p\_value | df | nobs |
| 0.067 | 0.065 | 196703.8 | 443.5131 | 444.626 | 28.622 | 0 | 1 | 400 |

summary(lm\_fit) #traditional output

##   
## Call:  
## lm(formula = Balance ~ Student, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -876.82 -458.82 -40.87 341.88 1518.63   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 480.37 23.43 20.50 < 2e-16 \*\*\*  
## StudentYes 396.46 74.10 5.35 1.49e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 444.6 on 398 degrees of freedom  
## Multiple R-squared: 0.06709, Adjusted R-squared: 0.06475   
## F-statistic: 28.62 on 1 and 398 DF, p-value: 1.488e-07

## Student Only Model Using Stan GLM Defaults

lm\_stan\_default <- stan\_glm(Balance ~ Student, data = Credit, refresh=0 )  
# Get regression table:  
print(lm\_stan\_default)

## stan\_glm  
## family: gaussian [identity]  
## formula: Balance ~ Student  
## observations: 400  
## predictors: 2  
## ------  
## Median MAD\_SD  
## (Intercept) 480.5 24.9   
## StudentYes 396.7 75.1   
##   
## Auxiliary parameter(s):  
## Median MAD\_SD  
## sigma 445.7 15.4   
##   
## ------  
## \* For help interpreting the printed output see ?print.stanreg  
## \* For info on the priors used see ?prior\_summary.stanreg

sims=as.matrix(lm\_stan\_default)  
quantile(sims[,2],c(0.025,0.975))

## 2.5% 97.5%   
## 248.4315 540.5746

We will use lm for now just because we will get about the same and I do not want all that output. ## Income and Student

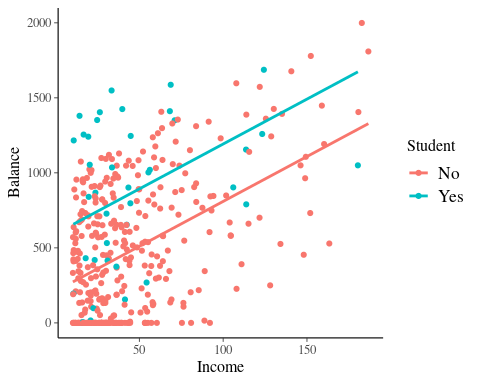
lm\_fit <- lm(Balance ~ Income+Student, data = Credit)  
# Get regression table:  
get\_regression\_table(lm\_fit, print=TRUE)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std\_error | statistic | p\_value | lower\_ci | upper\_ci |
| intercept | 211.143 | 32.457 | 6.505 | 0 | 147.333 | 274.952 |
| Income | 5.984 | 0.557 | 10.751 | 0 | 4.890 | 7.079 |
| StudentYes | 382.671 | 65.311 | 5.859 | 0 | 254.272 | 511.069 |

get\_regression\_summaries(lm\_fit, print=T)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r\_squared | adj\_r\_squared | mse | rmse | sigma | statistic | p\_value | df | nobs |
| 0.277 | 0.274 | 152347.6 | 390.3174 | 391.789 | 76.225 | 0 | 2 | 400 |

ggplot(Credit, aes(x = Income, y = Balance, color = Student)) +  
 geom\_point() +  
 labs(x = "Income", y = "Balance", color = "Student") +  
 geom\_parallel\_slopes(se = FALSE)



lm\_fit <- lm(Balance ~ Income\*Student, data = Credit)  
# Get regression table:  
get\_regression\_table(lm\_fit, print=TRUE)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std\_error | statistic | p\_value | lower\_ci | upper\_ci |
| intercept | 200.623 | 33.698 | 5.953 | 0.000 | 134.373 | 266.873 |
| Income | 6.218 | 0.592 | 10.502 | 0.000 | 5.054 | 7.382 |
| StudentYes | 476.676 | 104.351 | 4.568 | 0.000 | 271.524 | 681.827 |
| Income:StudentYes | -1.999 | 1.731 | -1.155 | 0.249 | -5.403 | 1.404 |

get\_regression\_summaries(lm\_fit, print=T)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r\_squared | adj\_r\_squared | mse | rmse | sigma | statistic | p\_value | df | nobs |
| 0.28 | 0.274 | 151836.4 | 389.6619 | 391.625 | 51.304 | 0 | 3 | 400 |

ggplot(Credit, aes(x = Income, y = Balance, color = Student)) +  
 geom\_point() +  
 labs(x = "Income", y = "Balance", color = "Student") +  
 geom\_smooth(method = "lm", se = FALSE)

