Predictive Assignment 4

Sreshtha Chatterjee - 724

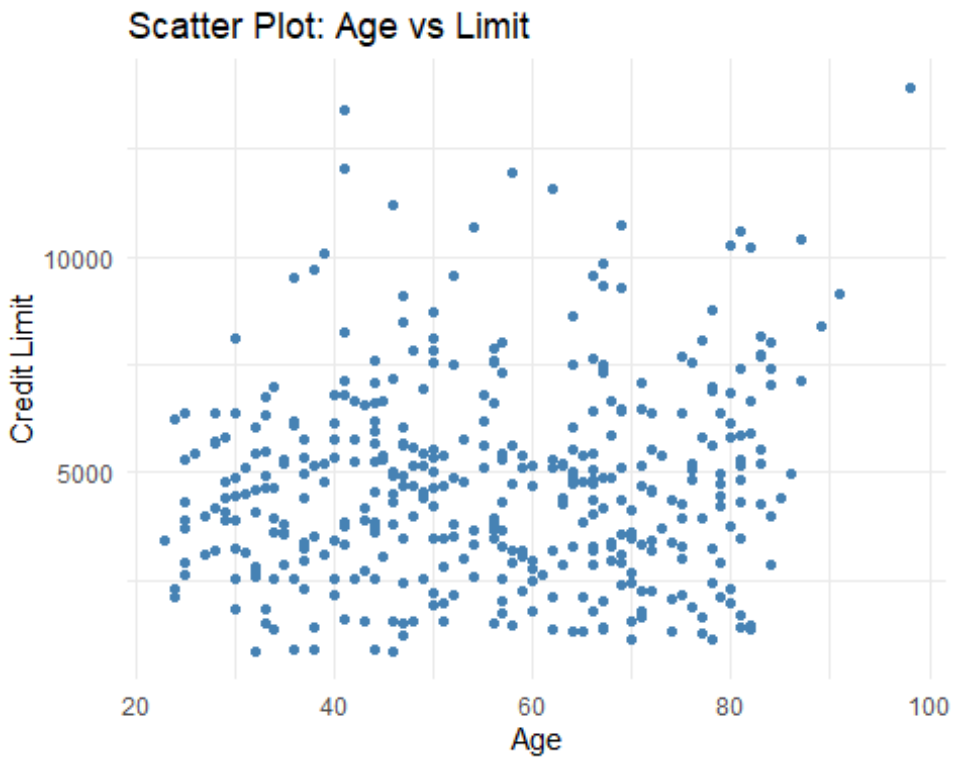
2026-02-19

```
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.5.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.5.2

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.5.2

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(car)

## Warning: package 'car' was built under R version 4.5.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.5.2

data(Credit)
```

(a) Scatter Plots

(b) Age versus Limit

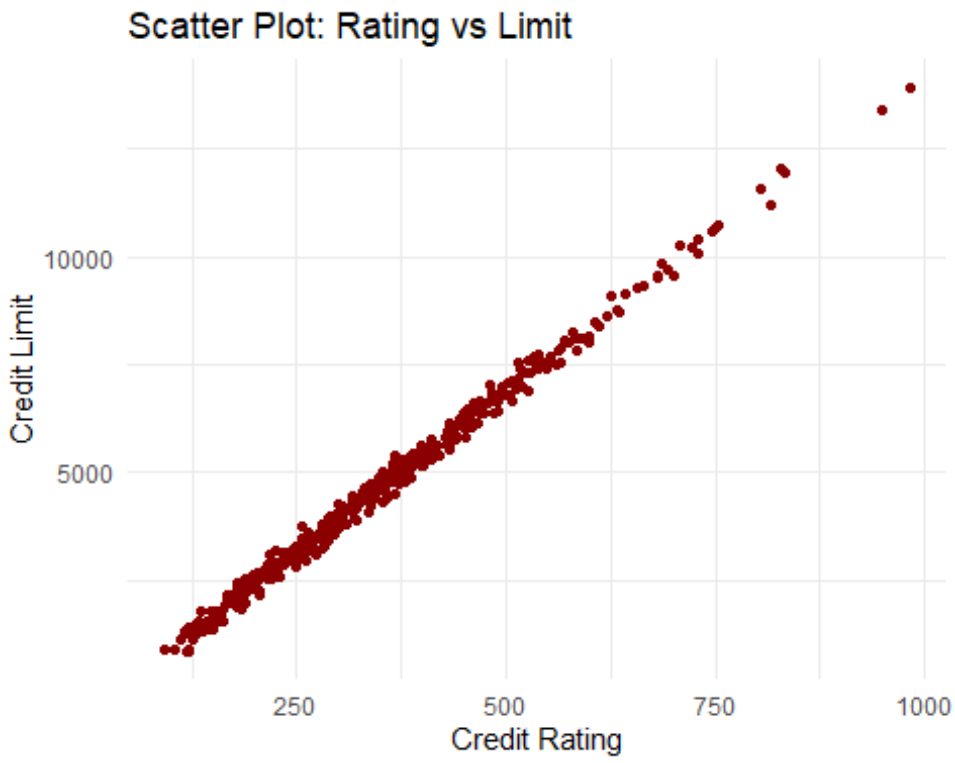
```
ggplot(Credit, aes(x = Age, y = Limit)) +
  geom_point(color = "steelblue") +
  labs(title = "Scatter Plot: Age vs Limit",
       x = "Age",
       y = "Credit Limit") +
  theme_minimal()
```



The scatter plot shows little to no clear linear relationship between Age and Limit. The points appear widely dispersed.

(ii) Rating versus Limit

```
ggplot(Credit, aes(x = Rating, y = Limit)) +
  geom_point(color = "darkred") +
  labs(title = "Scatter Plot: Rating vs Limit",
       x = "Credit Rating",
       y = "Credit Limit") +
  theme_minimal()
```



The scatter plot shows a very strong positive linear relationship between Rating and Limit. This suggests high correlation between these two predictors, which is a warning sign of multicollinearity.

(b) Regression Analysis

We estimate three regression models:

Balance ~ Age + Limit

Balance ~ Age + Rating + Limit

Balance ~ Rating + Limit

```
# Model 1
model1 <- lm(Balance ~ Age + Limit, data = Credit)

# Model 2
model2 <- lm(Balance ~ Age + Rating + Limit, data = Credit)

# Model 3
model3 <- lm(Balance ~ Rating + Limit, data = Credit)

stargazer(model1, model2, model3,
          type = "text",
          title = "Regression Results",
          column.labels = c("Balance ~ Age + Limit",
                            "Balance ~ Age + Rating + Limit",
                            "Balance ~ Rating + Limit"),
```

```
          dep.var.labels = "Balance",
          digits = 3)

##
## Regression Results
##
================================================================================
======================
##                                                  Dependent variable:
##                     ---------------------------------------------------------
-------------------------
##                                                          Balance
##                       Balance ~ Age + Limit   Balance ~ Age + Rating +
Limit Balance ~ Rating + Limit
##                                 (1)                          (2)
(3)
## --------------------------------------------------------------------------
-------------------------
## Age                           -2.291***                    -2.346***
##                                (0.672)                      (0.669)
##
## Rating                                                      2.310**
2.202**
##                                                            (0.940)
(0.952)
##
## Limit                          0.173***                     0.019
0.025
##                                (0.005)                      (0.063)
(0.064)
##
## Constant                      -173.411***                  -259.518***
-377.537***
##                                (43.828)                     (55.882)
(45.254)
##
## --------------------------------------------------------------------------
-------------------------
## Observations                     400                          400
400
## R2                              0.750                        0.754
0.746
## Adjusted R2                     0.749                        0.752
0.745
## Residual Std. Error    230.532 (df = 397)          229.080 (df = 396)
232.320 (df = 397)
## F Statistic         594.988*** (df = 2; 397)    403.718*** (df = 3; 396)
582.820*** (df = 2; 397)
##
================================================================================
```

```
======================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

Comment on Differences

The marked difference occurs in Model (ii), where both Rating and Limit are included together.

Because Rating and Limit are highly correlated:

The standard errors of their coefficients increase substantially.

One or both predictors may become statistically insignificant.

Coefficient magnitudes may change noticeably compared to Model (iii).

This instability shows multicollinearity.

(c) Variance Inflation Factor (VIF)

We compute VIF for the full model:

```
vif(model2)
```

```
##        Age     Rating      Limit
##   1.011385 160.668301 160.592880
```

Comment on Multicollinearity

If VIF > 5 → moderate multicollinearity

If VIF > 10 → severe multicollinearity

We expect Rating and Limit to have very large VIF values due to their strong correlation.

This confirms the presence of multicollinearity in the model including both variables.

Conclusion

Age and Limit show weak relationship.

Rating and Limit show extremely strong linear relationship.

Including both Rating and Limit inflates standard errors.

High VIF values confirm multicollinearity.

This example clearly demonstrates how highly correlated predictors can distort regression inference.