Data 558: Statistical Machine Learning

Spring 2023

Homework – 1

Arjun Sharma

I. Conceptual Questions

Problem 1 (8 points): Describe the bias-variance tradeoff in your own words. Accompany your explanations with a hand-drawn figure that contains on the x-axis flexibility of method and curves of bias, variance, training error, test error and irreducible error (all on the same plot). Make sure you also explain why each of the curves have the shape that they do.

The Bias-Variance trade-off is the result an inherent conflict of simultaneously minimizing the Bias and Variance in a supervised statistical model. In practical scenarios, it is usually observed that models with low bias tend to have a high variance and vice versa. The bias-variance trade-off is essentially a compromise on the model's ability to learn the relationship between the training data and the model's ability to generalize correctly on unseen data. There are two concerns associated with developing any supervised statistical model:

a. Is the model too simple for the data?

This concern is associated with the bias that the model learns as it is trained on the training set. If the model is too simple to capture the relationship between the predictors and the outcome, the bias of the model tends to increase. This happens because the training data may require a more complex model to accurately learn the inherent relationship between the predictors and the outcome, and the model is too simple to do so. An example of this scenario is if we were to fit a linear regression model to predict the square of a positive integer. The data clearly has a quadratic relationship but the model is not capable of capturing the inherent relationship between the predictor and target, and hence increases the bias in order to offset the difference in the ground truth and its own predictions.

b. Is the model too complex for the data?

This concern is associated with how sensitive the model is to fluctuations within the data. A model with high variance is problematic because it is prone to noise in the data and has not generalized well. High variance arises mainly due to 2 reasons:

1. The model is being trained on limited data.

2. The model is of a higher order polynomial than the inherent relationship between the predictors and the outcome.

The reasons above lead to the model being prone to noise in the data, and often overfitting on the training dataset, this hinders the ability of the model to generalize, and often leads to poor performance on testing and/or real-world data.

An example of this scenario is using a neural network model to predict house prices based on only the area of the house and the number of rooms with about 100 training examples.

The objective through the bias-variance trade-off is to minimize both the bias and variance such that a suitable compromise that enables generalization of the model is reached.

Below is the figure which captures the relationship between the model flexibility and the bias, variance, training error, testing error and the irreducible error.

The irreducible error remains constant regardless of the model specifics. Higher flexibility leads to lower bias and higher variance as explained earlier. The training error decreases with flexibility as the model's complexity is able to capture the relationship in the data. However, test error increases beyond a certain flexibility because the model becomes too complex for the data and this leads to higher noise sensitivity arising from higher variance.

The image below was created using MS paint.

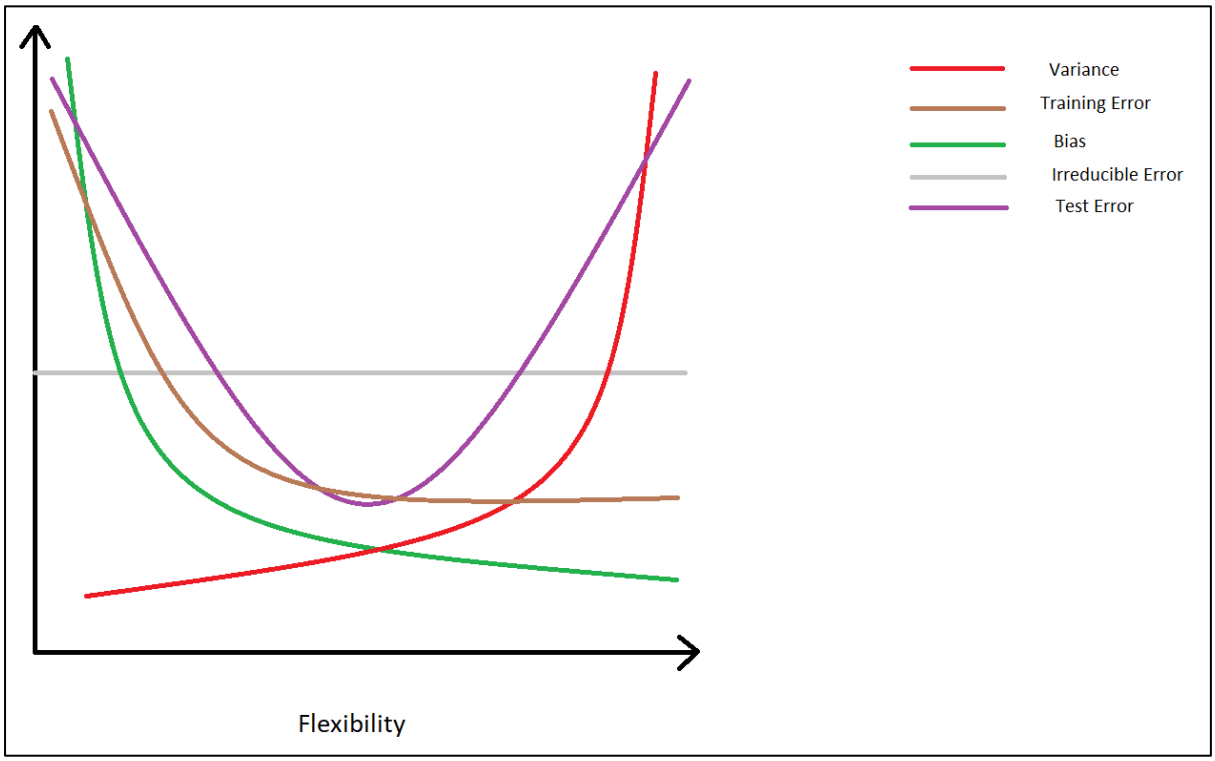


***Figure 1**: Relationship between flexibility, Bias, Variance, Train Error, Test Error and Irreducible Error*

Problem 2 (8 points): Suppose we see that our estimated regression function suffers from high variance. What can we do to alleviate this? Answer the same question but this time with the scenario being high bias. Now consider the nearest neighbor regression estimator we talked about in class. What choices, when constructing this estimator, affect the resulting bias and variance of the estimator? How would you control bias and variance with this estimator? How would you construct a nearest neighbor regression estimator with bias as low as possible? Conversely, how would you construct one with variance as low as possible?

High variance can arise as a result of one or more of the following reasons:

1. The model used is of a higher complexity than the inherent relationship between the predictors and the target

2. The model is being trained on a small training set, and as a result, the model is more prone to variance as a result of noise.

The model seems to have overfit on the training data. Since the model is tightly fit to the training data, the following can be done:

a. Use a model with a lower complexity

b. Use more training data

c. Use regularization techniques (the technique itself is model-dependent)

d. Use k-fold cross validation to segment the dataset

High bias can arise as a result of one or more of the following reasons:

1. The model used is too simple to accurately learn the patterns to capture the relationship between the predictors and the outcome.

2. The model is being trained on insufficient data.

The model seems to have underfit on the training data. Since the model has poorly generalized on the training data, the following can be done:

a. Use a model with a higher complexity

b. Use more training data; this can allow us to inspect whether the high bias was a result of insufficient data.

In the context of the K-Nearest Neighbors estimator, the only choice that we have in terms of constructing the estimator is our choice of ‘k’, and perhaps our method of calculating distances. However, the method of calculating distances is peripheral to the real concern which is the value of ‘K’. The figure represents the trend of errors as the choice of ‘K’ varies. It is observed that the training and test error rates are fairly high for a low value of K, however, at a sufficiently large K, both errors have been reduced significantly.

The bias can be minimized if a sufficiently small value for K is chosen, while the variance can be minimized if a sufficiently large value of K is chosen. In order to obtain a generalized K-NN estimator, we will have to choose a value for K that optimizes or balances the errors occurring as a result of the bias and variance.

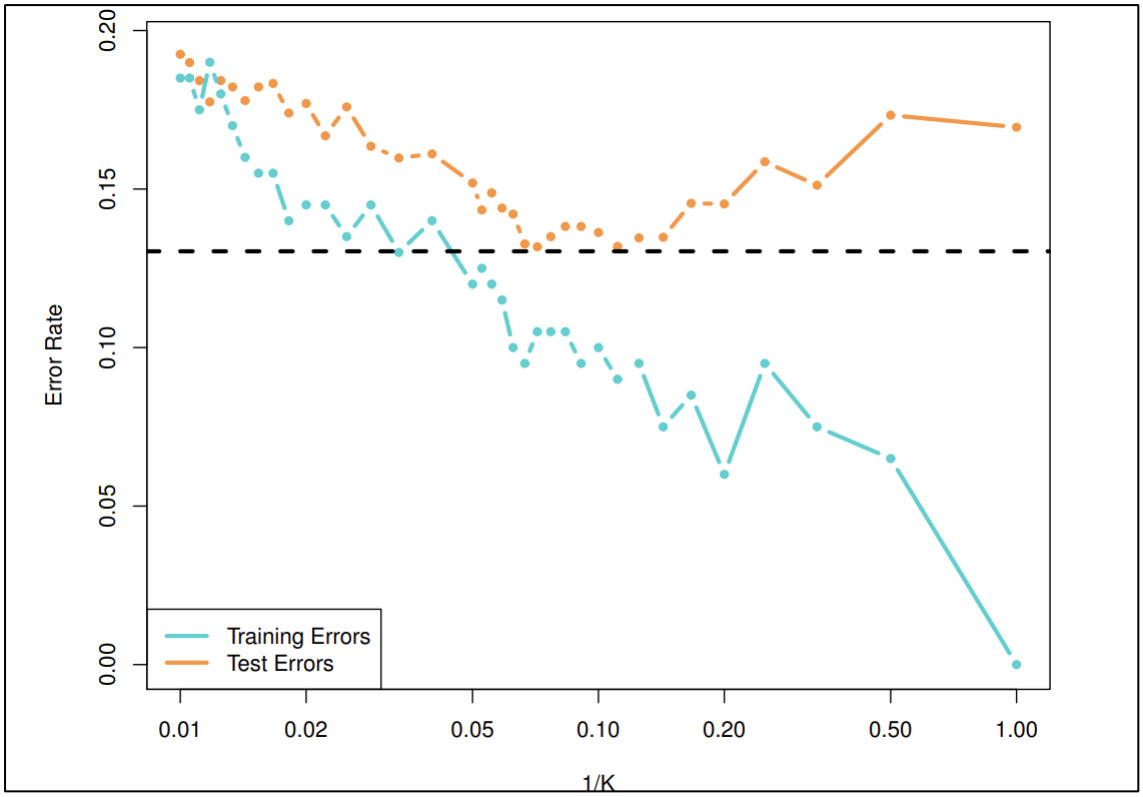


Figure 2: *Error rate as result of varying values of 'K' for the KNN estimator in an experiment.*

Problem 3 (10 points): Consider a univariate regression problem with a single predictor, i.e. $p = 1$. Suppose we observe n data points of the response and predictor $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$. We fit a linear regression model without an intercept and obtain regression parameter $\widehat{\beta}$. Let $\widehat{y}^{(i)} = x^{(i)}\widehat{\beta}$ be the prediction response variable for observation i. Show that we can write $\widehat{y}^{(i)} = \sum_{j=1}^{n} a_j y^{(j)}$, where $a_j \in \mathbb{R}$ for $j \in \{1, 2, \ldots, n\}$

PROBLEM 3

Given: Univariate regression model with a single predictor.

To prove: We can write $\hat{y}^{(i)} = \sum_{j=1}^{n} a_j y^{(j)}$, where $a_j \in R$ for $j \in \{1, 2 \ldots n$

Proof:

Let the relationship between the predictor and the target be represented as:

$$y = \beta_1 X_1 + \epsilon$$

From the question, we have no intercept, and:

$$\hat{y}^{(i)} = x^{(i)} \hat{\beta} \rightarrow \text{given.} - (1)$$

→ The RSS for this model is:

$$RSS = \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

In order to estimate $\hat{\beta}$, the RSS must be minimized.

$$\min_{\beta} RSS = \min_{\beta} \sum_{i=1}^{n} \left(y^{(i)} - x^{(i)}\beta\right)^2$$

Differentiate with respect to β:

$$\sum_{i=1}^{n} \left[-x^{(i)} \left(y^{(i)} - x^{(i)} \hat{\beta} \right) \right] = 0$$

$$\Rightarrow \boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n} x^{(i)2}}} \quad -(2)$$

Substitute (2) in (1):

$$\hat{y}^{(i)} = x^{(i)} \frac{\sum_{j=1}^{n} x^{(j)} y^{(j)}}{\sum_{j=1}^{n} x^{(j)2}}$$

$$\Rightarrow \boxed{\hat{y}^{(i)} = \sum_{j=1}^{n} \left(\frac{x^{(i)} x^{(j)}}{\sum_{j=1}^{n} x^{(j)2}} \right) y^{(j)}}$$

Now, let $\boxed{a_j = \frac{x^{(i)} x^{(j)}}{\sum_{j=1}^{n} x^{(j)2}}}$

Then, $\boxed{\hat{y}^{(i)} = \sum_{j=1}^{n} a_j y^{(j)}}$

$\therefore$ Proved

Problem 4 (10 points): Derive a formula for estimating regression parameters in a multiple linear regression model with an intercept (remember in class, we derived it without an intercept). Suppose the true relationship is linear but without an intercept term. Compared to the setting with an intercept, which will have a smaller training RSS? What about test RSS?

PROBLEM 4

To Find: Formula to estimate regression parameters in a multiple linear regression model with an intercept.

Let $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p + 0$

We need to choose $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_p$ to minimize the RSS.

$$\min RSS = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[y^{(i)} - \begin{bmatrix} 1 & X_1^{(i)} & X_2^{(i)} & \dots X_p \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \right]^2$$

$$\nabla_\beta (RSS) = \sum_{i=1}^{n} -(1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p)^T \left(y^{(i)} - (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}\right)$$

$$\Rightarrow \quad \nabla_\beta (RSS) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p)^T (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p) \beta = \sum_{i=1}^{n} (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p) y^{(i)}$$

Let $M = \sum_{i=1}^{n} (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p)^T (1 \; X_1^{(i)} \; X_2^{(i)} \dots X_p)$

Then, $$\boxed{\beta = M^{-1} \sum_{i=1}^{n} (1 \; X_1^{(i)} \; X_2^{(i)} \; X_3^{(i)} \dots X_p) y^{(i)}}$$

During training, it is expected that the model with an intercept has a lower RSS as the intercept would account for some portion of the variance that is not explained by the other parameters. It is expected that during testing, the model with the intercept still has the lower RSS. The model without an intercept would be vulnerable to overfitting because the absence of a bias term would make the model more sensitive to minor fluctuation in the data. As a result, the model without the intercept would not be able to generalize on the data.

Hence, the model without the intercept has a higher RSS during both training and testing.

II. Applied Questions

Problem 1 (OPTIONAL/BONUS: 4 points): In this question, you will implement your own linear regression estimator (with an intercept term). Write it as a function that takes in data of arbitrary number of predictors (p) and observations (n) and of a response variable, and outputs i) estimated coefficients, ii) the p-values corresponding to hypotheses $H_0 : \beta_j = 0$ for every j using t-statistics, and iii) p-value corresponding to hypothesis $H_0 : \beta = 0$ using the F-statistic. Now, take $p = 3$ and let x_1, x_2, x_3 be independent normal and $y = x_1 + 2x_2 + 0.5x_3 + \epsilon$ where ϵ is also an independent normal random variable. Generate $n = 100$ observations of (x, y). Using this data, compare the regression coefficient estimates and p-values of your code to the one obtained from using `R` or `python` packages.

Problem 2 (8 points): Consider the linear model $y = 0x_1 + 0.01x_2 + 2x_3 + \epsilon$ where ϵ, x_1, x_2, x_3 are all independent normal random variables. Compute p-values associated to each regression parameters for $n = \{10, 20, 100, 200\}$. What do you notice, especially as the sample size increases?

Source Code:

```
```{r}
n <- c(10, 20, 100, 200)
p_vals <- matrix(NA, nrow = length(n), ncol = 4)

for (i in seq_along(n)) {
  x <- matrix(rnorm(n[i]*3), ncol=3)
  eps <- rnorm(n[i], sd = 0.1)
  y <- 0*x[,1] + 0.01*x[,2] + 2*x[,3] + eps

  fit_p3 <- lm(y ~ x[,1] + x[,2] + x[,3])
  p_vals[i,] <- summary(fit_p3)$coefficients[,4]
}

colnames(p_vals) <- c("Intercept","x1", "x2", "x3")
rownames(p_vals) <- n
p_vals

```
```

Results:

```
    Intercept        x1         x2            x3
10  0.6034685 0.2115249 0.32264404  3.714261e-09
20  0.2284987 0.7934148 0.26517765  1.869505e-19
100 0.8039492 0.8560778 0.01737060 9.089186e-131
200 0.4784559 0.8949454 0.01964052 6.789323e-263
```

The p-value associated with x3 reduces as the sample size increases. It is possible that the higher weight associated with x3 leads to a more pronounced influence on the p-value as the sample size increases. The p-value associated with x1 and x2 are much larger and don't seem to follow any such pattern observed in x3, attributed to the lower weights assigned to them. It appears that x1 and x2 are more likely to be impacted by noise in the data.

Problem 3: Consider the linear model $y = 0.2x_1 + 0.4x_2 + 0.3x_3 + \epsilon$ where ϵ is a normal random variable and

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & a & 0 \\ a & 1 & a \\ 0 & a & 1 \end{pmatrix} \right),$$

where a is a scalar parameter. We consider 3 possible values of a: $a \in \{0.2, 0.5, 0.8\}$. For every value of a, we generate $n = \{10, 50, 100\}$ observations of (x, y).

(a) (8 points): For each a and n, fit a linear regression model (with an intercept term) with 80% of the data (the other 20% is used to evaluate performance on test data);

(b) (8 points): Report the accuracy of estimated coefficients and train and test RSS.

(c) (8 points): What do you notice? Does large or small values of a make inference easier, and why?

Source Code:

```
library(mvtnorm)
library(dplyr)
sample_sizes <- c(10, 50, 100)
scalars <- c(0.2, 0.4, 0.6)
sample_meanvector <- c(0, 0, 0)

report <- data.frame()

for(sample_size in sample_sizes){
  for(a in scalars){
    sample_covariance_matrix <- matrix(c(1, a, 0, a, 1, a, 0, a, 1), ncol = 3)
    data <- as.data.frame(rmvnorm(n = sample_size, mean = sample_meanvector, sigma = sample_covariance_matrix))
    colnames(data) <- c("x1", "x2", "x3")
    data$error <- rnorm(sample_size)
    data$y <- 0.2*data$x1 + 0.4*data$x2 + 0.3*data$x3 + data$error
    data <- data %>% dplyr::select(-error)
    train_idx <- sample(nrow(data), round(0.8*nrow(data)), replace = FALSE)
    train_data <- data[train_idx, ]
    test_data <- data[-train_idx, ]
    model <- lm(y ~ ., data = train_data)
    y_hat_train <- predict(model, newdata = train_data)
    y_hat_test <- predict(model, newdata = test_data)
    coefs <- coef(model)[-1]
    baseline <- c(0.2, 0.4, 0.3)
    pct_error <- 100*(coeffs - baseline)/baseline
    output <- data.frame(predictor = c("x1", "x2", "x3"),
                         true_coef = baseline,
                         estimated_coef = coefs) %>%
              mutate(n = sample_size,
                     a = a,
                     accuracy = abs(true_coef - estimated_coef)/true_coef,
                     train_rss = sum((train_data$y - y_hat_train)^2),
                     test_rss = sum((test_data$y - y_hat_test)^2)) %>%
              dplyr::select(n, a, everything())
    results <- bind_rows(results, output)
  }
}
results
```

Results:

	n	a	predictor	true_coef	estimated_coef	accuracy	train_rss	test_rss
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
x1...1	10	0.2	x1	0.2	0.367606306	0.83803153	2.427460	0.6502372
x2...2	10	0.2	x2	0.4	0.716080960	0.79020240	2.427460	0.6502372
x3...3	10	0.2	x3	0.3	0.088374226	0.70541925	2.427460	0.6502372
x1...4	10	0.4	x1	0.2	0.000607712	0.99696144	4.310518	1.1915175
x2...5	10	0.4	x2	0.4	-0.806462415	3.01615604	4.310518	1.1915175
x3...6	10	0.4	x3	0.3	0.382354179	0.27451393	4.310518	1.1915175
x1...7	10	0.6	x1	0.2	0.511297060	1.55648530	3.260530	3.5843436
x2...8	10	0.6	x2	0.4	-0.476404014	2.19101004	3.260530	3.5843436
x3...9	10	0.6	x3	0.3	0.730865833	1.43621944	3.260530	3.5843436
x1...10	50	0.2	x1	0.2	-0.025119027	1.12559513	42.375809	17.0371097

1-10 of 27 rows Previous 1 2 3 Next

	n	a	predictor	true_coef	estimated_coef	accuracy	train_rss	test_rss
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
x2...11	50	0.2	x2	0.4	0.154573706	0.61356573	42.375809	17.0371097
x3...12	50	0.2	x3	0.3	0.681661619	1.27220540	42.375809	17.0371097
x1...13	50	0.4	x1	0.2	0.216461059	0.08230530	26.688835	6.3270117
x2...14	50	0.4	x2	0.4	0.101956972	0.74510757	26.688835	6.3270117
x3...15	50	0.4	x3	0.3	0.700723022	1.33574341	26.688835	6.3270117
x1...16	50	0.6	x1	0.2	0.015774643	0.92112678	34.675629	6.3633412
x2...17	50	0.6	x2	0.4	0.671747870	0.67936968	34.675629	6.3633412
x3...18	50	0.6	x3	0.3	-0.006965812	1.02321937	34.675629	6.3633412
x1...19	100	0.2	x1	0.2	0.080537568	0.59731216	56.587508	14.6132137
x2...20	100	0.2	x2	0.4	0.303579805	0.24105049	56.587508	14.6132137

11-20 of 27 rows Previous 1 2 3 Next

	n	a	predictor	true_coef	estimated_coef	accuracy	train_rss	test_rss
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
x3...21	100	0.2	x3	0.3	0.261521916	0.12826028	56.587508	14.6132137
x1...22	100	0.4	x1	0.2	0.178183405	0.10908297	61.390578	19.6381494
x2...23	100	0.4	x2	0.4	0.433294241	0.08323560	61.390578	19.6381494
x3...24	100	0.4	x3	0.3	0.233990726	0.22003091	61.390578	19.6381494
x1...25	100	0.6	x1	0.2	0.204317566	0.02158783	90.631162	12.9720260
x2...26	100	0.6	x2	0.4	0.689614546	0.72403636	90.631162	12.9720260
x3...27	100	0.6	x3	0.3	0.257342025	0.14219325	90.631162	12.9720260

21-27 of 27 rows Previous 1 2 3 Next

No clear pattern is observed with respect to how accuracy and the train_rss or the test_rss change. For change in ‘a’ the percentage error also changes however there is no clear pattern to this as well. A lower percentage error implies that our model’s estimate is ‘close’ to the true value while a higher percentage error implies that the estimate is relatively far from the true value. As the percentage error increases, our ability to interpret or infer from the model decreases, making the model harder to infer from.

Problem 4: Consider the *Credit* dataset with balance (average credit debt) as the response variable and the predictors age, cards, education, income, limit, and rating. The dataset can be loaded in `R` using the command:

```
library (ISLR2)
```

In `Python`, the dataset can be loaded by running

```
from ISLP import load_data
```

followed by running

```
Credit = load_data("Credit")
```

(a) (8 points): For each predictor, fit a a simple linear regression model. Describe your results. In which of the models is there a statistically significant association between the predictor and response? Use both plots and quantitative evidence (e.g. R^2 value) to back up your assertions.

(b) (8 points): Fit a multiple regression model to predict the response using all of the predictors. Describe your results. Using an statistic, provide evidence for whether any predictor is useful for predicting the response. Using t-statistics, provide evidence for which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) (8 points): How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis

(d) (8 points): Is there evidence of non-linear association between any of the predictors and response? To answer this question, for each predictor X, fit a model of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

Ans.

(a) For all 6 predictors, a linear regression model is fit, where the 'Balance' is taken as the target, and one of the said columns is taken as the predictors. In addition, a scatter plot for the predictor vs target is created, with the best fit line to denote the relationship between the predictor and the target.

We take the level of significance, α to be 0.05.

For any given predictor, we state the following hypotheses:

H_0: There is no relationship between the predictor and the target. i.e., $\beta_i = 0$, where i = 0,1

H_A: There exists a relationship between the predictor and the target. i.e., at least one $\beta_i \neq 0$

Setup Code:

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
library (ISLR2)
```

```{r}
df = Credit
```

```{r}
summary(df)
```
```

i. Balance vs Age:

Source Code:

```
Predictor: Age
Target: Balance
```{r}
fit_Age <- lm(Balance~Age, data=df)
summary(fit_Age)
plot(fit_Age)

scatter <- plot(df$Age, df$Balance)
abline(fit_Age)
```
```

Model Summary:

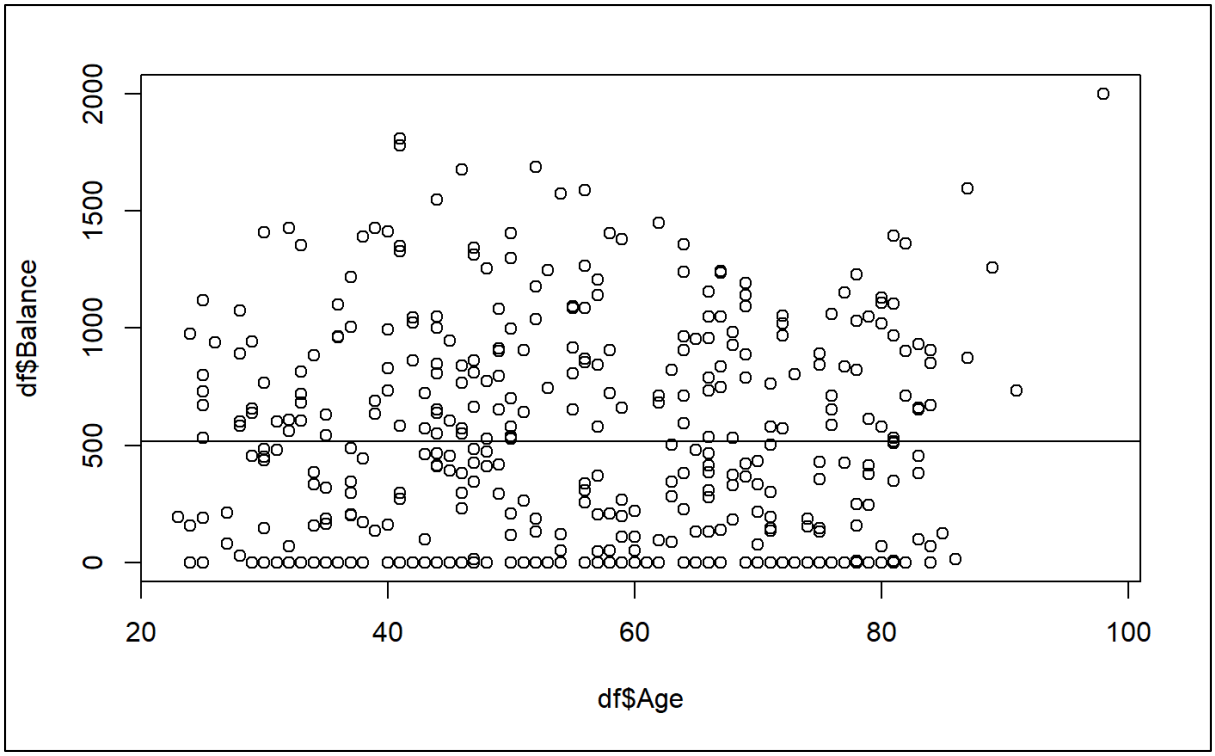
```
Call:
lm(formula = Balance ~ Age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-521.40 -451.50  -59.94  343.47 1476.91

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 517.29222   77.85153   6.645    1e-10 ***
Age           0.04891    1.33599   0.037    0.971
---
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:  3.368e-06, Adjusted R-squared:  -0.002509
F-statistic: 0.00134 on 1 and 398 DF,  p-value: 0.9708
```

Best Fit Line:



Inference: Based on the t-test results, it is observed that the p-value for the 'Age' predictor is roughly 0.971, which, by far exceeds the significance level of 0.05. We fail to reject the null hypothesis in this case.

Also, the Multiple R-Squared value is 3.368e-6, which implies that the model with 'Age' as the predictor only explains 0.0003368% of the variance in the target.

Furthermore, the regression line on the scatter plot above is observed to be almost flat. Visual observation suggests that the change in age does not explain the change in the Balance variable.

Based on the points stated above, we conclude that there is no statistically significant association between 'Age' and 'Balance'.

ii. Balance vs Cards:

Source Code:

````
Predictor: Cards
Target: Balance
```{r}
fit_Cards <- lm(Balance~Cards, data=df)
summary(fit_Cards)
plot(fit_Cards)
scatter <- plot(df$Cards, df$Balance)
abline(fit_Cards)
```
````

Model Summary:

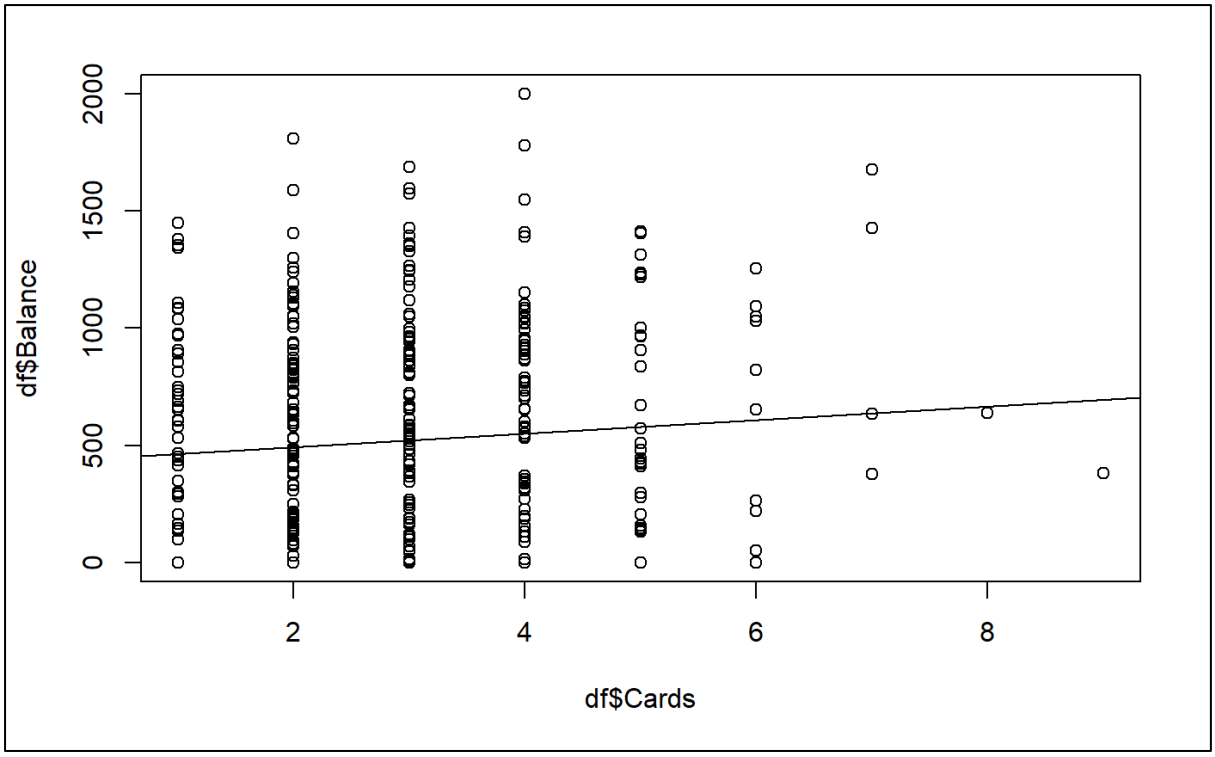
```
Call:
lm(formula = Balance ~ Cards, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-608.21 -455.24  -38.75  350.99 1448.77

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   434.29      54.57   7.958 1.83e-14 ***
Cards          28.99      16.74   1.731   0.0842 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 458.6 on 398 degrees of freedom
Multiple R-squared:  0.007475,  Adjusted R-squared:  0.004981
F-statistic: 2.997 on 1 and 398 DF,  p-value: 0.08418
```

Best Fit Line:



Inference: Based on the t-test results, it is observed that the p-value for the ‘Cards’ predictor is roughly 0.08418, which exceeds the significance level of 0.05. We fail to reject the null hypothesis in this case.

Also, the Multiple R-Squared value is 0.007475, which implies that the model with ‘Cards’ as the predictor only explains 7.475 of the variances in the target.

The regression line on the scatter plot suggests that the ‘Cards’ shares a weakly positive linear relationship with ‘Balance’. The relationship between the 2 variables is not strong, and we cannot conclude that there is a significant statistical association between Cards and Balance.

iii. Balance vs Education:

Source Code:

```
Predictor: Education
Target: Balance
```{r}
fit_Edu <- lm(Balance~Education, data=df)
summary(fit_Edu)
plot(fit_Edu)
scatter <- plot(df$Education, df$Balance)
abline(fit_Edu)
```
```

Model Summary:

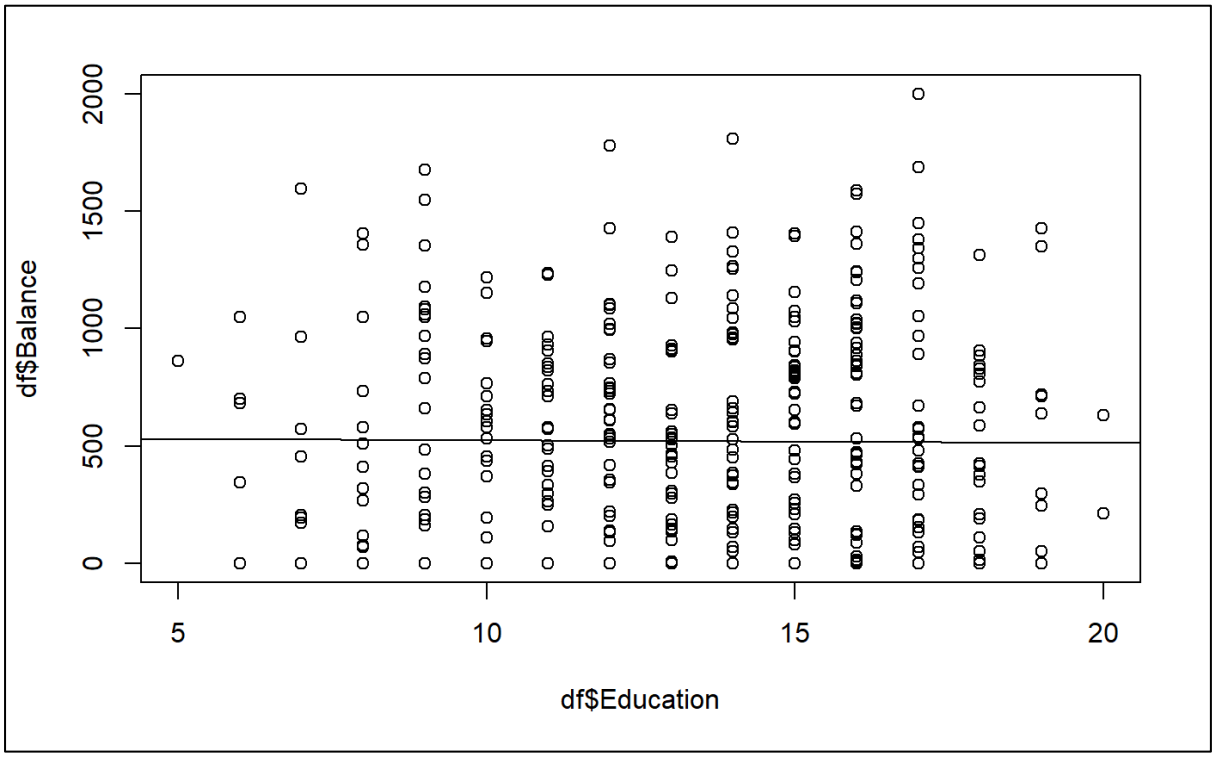
```
Call:
lm(formula = Balance ~ Education, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-528.85 -452.73  -61.05  337.20 1483.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  535.966    101.814   5.264 2.31e-07 ***
Education     -1.186      7.374  -0.161    0.872
---
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:  6.499e-05, Adjusted R-squared:  -0.002447
F-statistic: 0.02587 on 1 and 398 DF,  p-value: 0.8723
```

Best Fit Line:



Based on the t-test results, it is observed that the p-value for the ‘Education’ predictor is roughly 0.8723, which by far exceeds the significance level of 0.05. We fail to reject the null hypothesis in this case.

Also, the Multiple R-Squared value is 6.499e-5, which implies that the model with ‘Education’ as the predictor only explains 0.0006499% of the variance in the target.

Furthermore, the regression line on the scatter plot above is observed to be almost flat. Visual observation suggests that the change in Education Level explains almost none of the change in the Balance variable.

Based on the points stated above, we conclude that there is no statistically significant association between ‘Education’ and ‘Balance’

iv. Balance vs Income:

Source Code:

````
Predictor: Income
Target: Balance
```{r}
fit_Income <- lm(Balance~Income, data=df)
summary(fit_Income)
plot(fit_Income)
scatter <- plot(df$Income, df$Balance)
abline(fit_Income)
```
````

Model Summary:

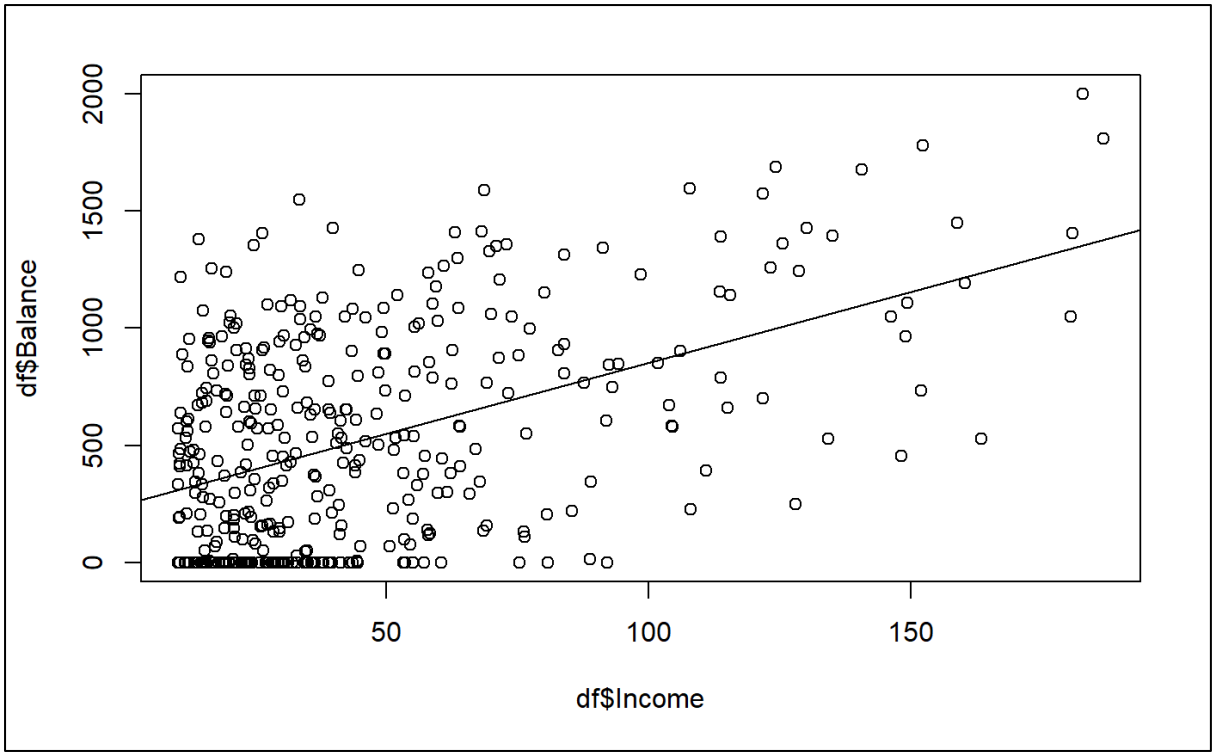
```
Call:
lm(formula = Balance ~ Income, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-803.64 -348.99  -54.42  331.75 1100.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 246.5148    33.1993   7.425  6.9e-13 ***
Income        6.0484     0.5794  10.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 407.9 on 398 degrees of freedom
Multiple R-squared:  0.215,     Adjusted R-squared:  0.213
F-statistic:   109 on 1 and 398 DF,  p-value: < 2.2e-16
```

Best Fit Line:



Inference: Based on the t-test results, it is observed that the p-value for the 'Income' predictor is <2.2e-16, which is much lower than the significance level of 0.05. We reject the null hypothesis in this case. The association between Income and Balance is statistically significant

Also, the Multiple R-Squared value is 0.215, which tells us that the model with 'Income' as the predictor explains 21.5% of the variances in the target.

The regression line on the plot above depicts a clear albeit weakly linear relationship between 'Income' and 'Balance'. The relationship between Income and Balance is hence statistically significant.

v. Balance vs Limit:

Source Code:

````
Predictor: Limit
Target: Balance
```{r}
fit_Limit <- lm(Balance~Limit, data=df)
summary(fit_Limit)
plot(fit_Limit)

scatter <- plot(df$Limit, df$Balance)
abline(fit_Limit)
```
````

Model Summary:

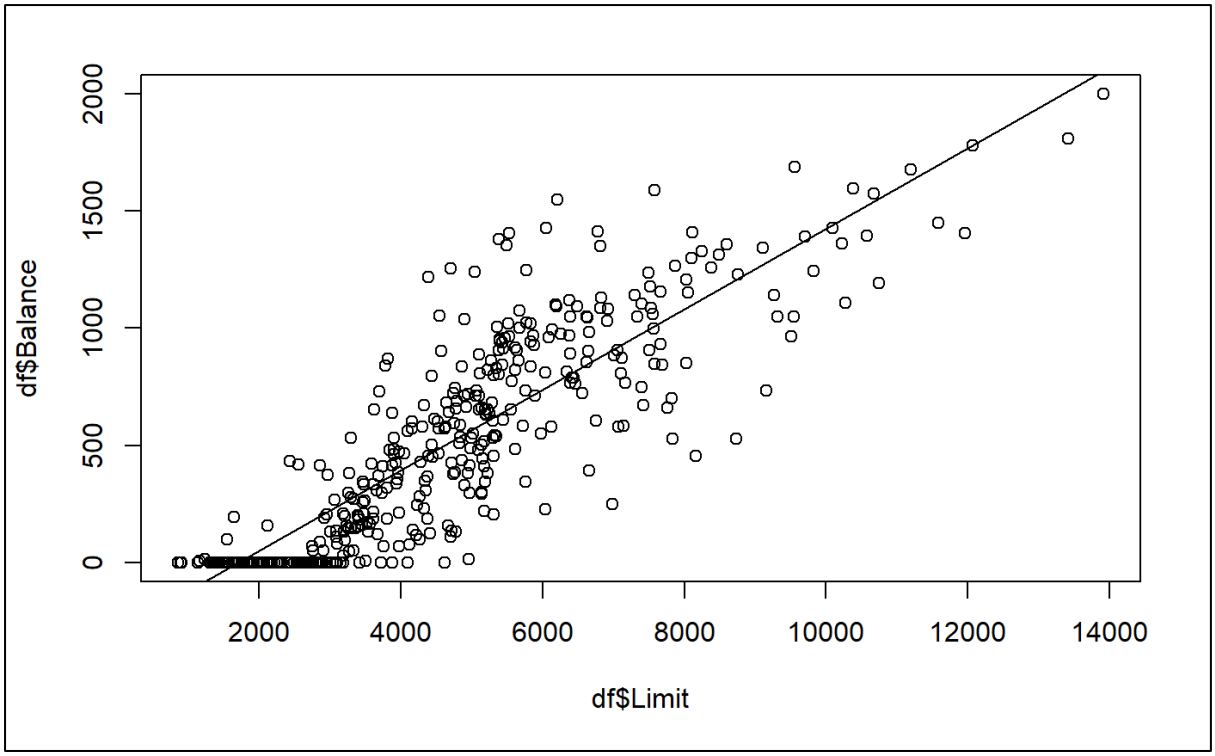
```
Call:
lm(formula = Balance ~ Limit, data = df)

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
```

Best Fit Line:



Inference: Based on the t-test results, it is observed that the p-value for the ‘Limit’ predictor is <2.2e-16, which is much lower than the significance level of 0.05. We reject the null hypothesis in this case. The association between Limit and Balance is statistically significant

Also, the Multiple R-Squared value is 0.7425, which tells us that the model with ‘Limit’ as the predictor explains 74.25% of the variances in the target, which clearly depicts the strength of the relationship between the Limit and Balance

The regression line on the plot above depicts a clear and strong linear relationship between ‘Limit’ and ‘Balance’. The relationship between Limit and Balance is statistically significant.

vi. Balance vs Rating:

Source Code:

````
Predictor: Rating
Target: Balance
```{r}
fit_Rating <- lm(Balance~Rating, data=df)
summary(fit_Rating)
plot(fit_Rating)

scatter <- plot(df$Rating, df$Balance)
abline(fit_Rating)
```
````

Model Summary:

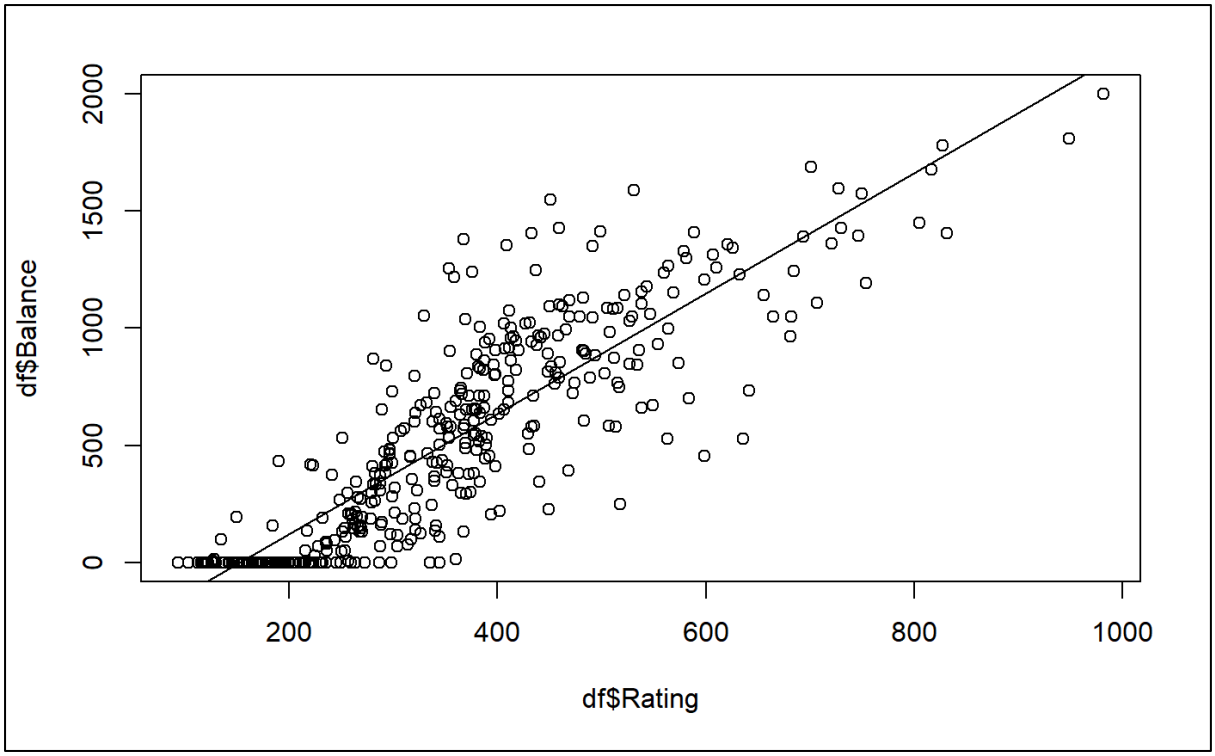
```
Call:
lm(formula = Balance ~ Rating, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-712.28 -135.32   -9.58  125.67  829.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -390.84634   29.06851  -13.45   <2e-16 ***
Rating         2.56624    0.07509   34.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 232.1 on 398 degrees of freedom
Multiple R-squared:  0.7458,    Adjusted R-squared:  0.7452
F-statistic:  1168 on 1 and 398 DF,  p-value: < 2.2e-16
```

Best Fit Line:



Inference: Based on the t-test results, it is observed that the p-value for the 'Rating' predictor is <2.2e-16, which is much lower than the significance level of 0.05. We reject the null hypothesis in this case. The association between Rating and Balance is statistically significant

Also, the Multiple R-Squared value is 0.7458, which tells us that the model with 'Rating' as the predictor explains 74.58% of the variances in the target, which clearly depicts the strength of the relationship between the Limit and Balance

The regression line on the plot above depicts a clear and strong linear relationship between 'Rating' and 'Balance'. The relationship between Rating and Balance is statistically significant.

Conclusion: It is observed that the predictors 'Limit', 'Rating' and 'Income' have a statistically significant association with the target.

(b) We take the level of significance, α to be 0.05.

The hypothesis is formulated as follows:

H_0: There is no relationship between the predictor and the target. i.e., $\beta_i = 0$, where i = 0,1, 2... n, where n is the number of coefficients

H_A: There exists a relationship between the predictor and the target. i.e., at least one $\beta_i \neq 0$

Below is the source code for fitting "Balance" as a target with the 6 predictors used earlier:

Source Code:

```r
fit_all <- lm(Balance~ Age + Cards + Education + Income + Limit + Rating, data = df)
summary(fit_all)
plot(fit_all)
```

The summary of the model is given below.

Model Summary:

```
Call:
lm(formula = Balance ~ Age + Cards + Education + Income + Limit +
    Rating, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-227.25 -113.15  -42.06   45.82  542.97

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Age           -0.89240    0.47808  -1.867  0.06270 .
Cards         11.59156    7.06670   1.640  0.10174
Education      1.99828    2.59979   0.769  0.44257
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *
Rating         2.06310    0.79426   2.598  0.00974 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,    Adjusted R-squared:  0.8764
F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

It is observed that for the model fitted on the predictors ‘Age’, ‘Cards’, ‘Education’, ‘Income’, ‘Limit’, and ‘Rating’, the p-value is observed to be < 2.2e-16, which is well within the limit of $\alpha = 0.05$. We Reject the null hypothesis and infer that not all coefficients for the given predictors are zero. The Multiple R-squared value of 0.8782 indicates that the 6 predictors explain 87.82% of the variance observed in Balance.

From the model summary, based on the t-test results, it is observed that the predictors ‘Income’, ‘Limit’, and ‘Rating’ have a statistically significant association with “Balance”, as the p-value associated with the said predictors is within the limit of $\alpha = 0.05$.

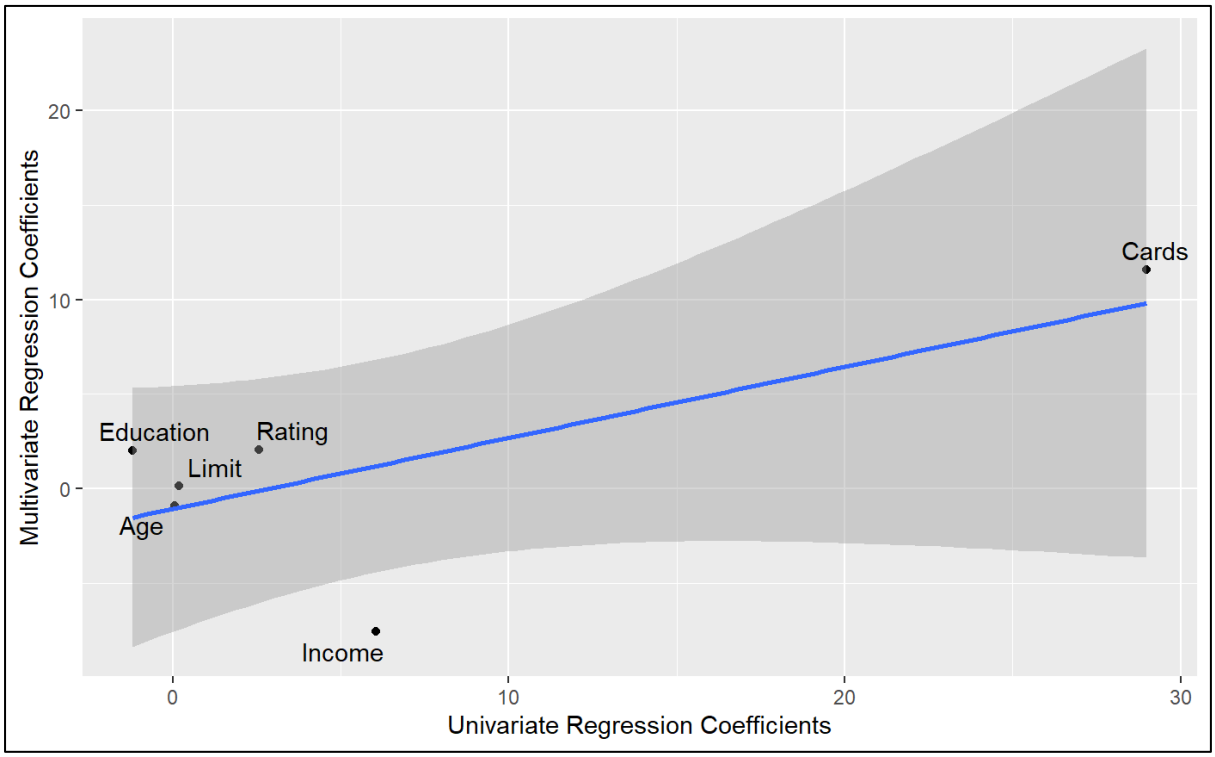
(c) Comparing our results from (a) and (b), the results are consistent. In section (a), it was observed that ‘Income’, ‘Limit’ and ‘Rating’ had a statistically significant association with Balance. This also appears to be the case in section (b) where a multiple regression model was fit against Balance. The figure below depicts the univariate regression coefficients on the x-axis and the multivariate regression coefficients on the y-axis.

Source Code:

```r
library(ggplot2)
library(ggrepel)

df_points <- data.frame(coefs, all_coef, coef_label)
ggplot(df_points, aes(coefs,all_coef)) +
  geom_point() + geom_smooth(method = "lm") + geom_text_repel(aes(label = coef_label))
```

Graph of Univariate Regression Coefficients and Multivariate Regression Coefficients:



(d) A model is fit for all 6 predictors, keeping 'Balance' as the target.

The models are of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

We take the level of significance, α to be 0.05.

For any given predictor, we state the following hypotheses:

H_0: There is no relationship between the predictor and the target. i.e., $\beta_i = 0$, where i = 0,1,2,3

H_A: There exists a relationship between the predictor and the target. i.e., at least one $\beta_i \neq 0$

i. **Balance vs Age:**

Source Code:

```r
```{r}
fit_non_lm <- lm(Balance ~ Age  + I(Age^2) + I(Age^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Age + I(Age^2) + I(Age^3), data = df)

Residuals:
   Min     1Q Median     3Q    Max
-672.3 -437.7  -51.4  346.6 1210.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.618e+03  6.588e+02  -2.456 0.014479 *
Age          1.330e+02  3.880e+01   3.429 0.000669 ***
I(Age^2)    -2.556e+00  7.200e-01  -3.550 0.000431 ***
I(Age^3)     1.536e-02  4.247e-03   3.617 0.000337 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 454 on 396 degrees of freedom
Multiple R-squared:  0.03236,   Adjusted R-squared:  0.02503
F-statistic: 4.415 on 3 and 396 DF,  p-value: 0.004543
```

It appears that there exists a non-linear relationship between Balance and Age. There appears to be statistical significance in the association of Balance and Age for all degrees of Age. The p-value is also within the limit of the level of significance for all parameters, hence we reject the null hypothesis and infer that there exists a statistically significant association between all degrees of Age and Balance. However, the Multiple R-squared value tells us that all the parameters explain only about 3.236% of the variance in Balance.

ii. **Balance vs Education:**

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Education + I(Education^2) + I(Education^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Education + I(Education^2) + I(Education^3),
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-645.01 -466.92  -56.71  348.02 1466.39

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1359.2648  1017.3233   1.336    0.182
Education      -187.5694   255.1849  -0.735    0.463
I(Education^2)   13.1928    20.5223   0.643    0.521
I(Education^3)   -0.2953     0.5314  -0.556    0.579

Residual standard error: 460.8 on 396 degrees of freedom
Multiple R-squared:  0.00285,   Adjusted R-squared:  -0.004705
F-statistic: 0.3772 on 3 and 396 DF,  p-value: 0.7695
```

The p-value for the all the indicators combined against the target is 0.7695. This means that for a significance level of 0.05, we fail to reject the null hypothesis that there is no

association between “Balance” and Education, Education^2, and Education^3. Furthermore, the p-values for all the individual parameters also greatly exceeds the significance level. Hence, we fail to reject the null hypothesis. We infer that there is no association, linear or non-linear between Education and Balance. In addition, the Multiple R-Squared statistic of 0.00285 implies that only 2.85% of the variation in Balance is explained by Education, Education^2, and Education^3.

iii. **Balance vs Limit:**

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Limit+ I(Limit^2) + I(Limit^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Limit + I(Limit^2) + I(Limit^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-690.84 -132.00   -1.41  134.21  746.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.213e+02  7.596e+01  -4.230  2.9e-05 ***
Limit        1.610e-01  4.561e-02   3.530 0.000465 ***
I(Limit^2)   6.930e-06  8.005e-06   0.866 0.387158
I(Limit^3)  -5.903e-10  4.100e-10  -1.440 0.150743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 230.9 on 396 degrees of freedom
Multiple R-squared:  0.7498,    Adjusted R-squared:  0.7479
F-statistic: 395.5 on 3 and 396 DF,  p-value: < 2.2e-16
```

At a significance level of 0.05, we see that the p-value for the association between Limit, Limit^2, and Limit^3 with Balance is <2.2e-16, indication a very strong association. Particularly, the p-value for Limit is well within the significance level of $\alpha = 0.05$. We reject the null hypothesis and infer that there is a relationship between Balance and the regression coefficients. However, it is noticed that the p-value for Limit^2 and Limit^3 exceeds the significance level. Hence, we fail to reject the null hypothesis for Limit^2 and Limit^3. In addition, the Multiple R-Squared statistic of 0.7498 indicates that ‘Limit’ in its linear and non-linear forms as a predictor explains 74.98% of the variance in Balance.

iv. **Balance vs Cards:**

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Cards + I(Cards^2) + I(Cards^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Cards + I(Cards^2) + I(Cards^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-679.12 -454.21  -56.09  352.99 1455.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  667.919    164.949   4.049 6.18e-05 ***
Cards       -189.339    152.923  -1.238    0.216
I(Cards^2)    54.856     41.604   1.319    0.188
I(Cards^3)    -3.831      3.344  -1.146    0.253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 458.4 on 396 degrees of freedom
Multiple R-squared:  0.01325,   Adjusted R-squared:  0.005778
F-statistic: 1.773 on 3 and 396 DF,  p-value: 0.1517
```

The p-value for the all the indicators combined against the target is 0.1517. This means that for a significance level of 0.05, we fail to reject the null hypothesis that there is no association between “Balance” and Cards, $Cards^2$, and $Cards^3$. Furthermore, the p-values for all the individual parameters also greatly exceeds the significance level. Hence, we fail to reject the null hypothesis. We infer that there is no association, linear or non-linear between Cards and Balance. In addition, the Multiple R-Squared statistic of 0.01325 implies that only 1.325% of the variation in Balance is explained by Cards, $Cards^2$, and $Cards^3$.

v. **Balance vs Income:**

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Income + I(Income^2) + I(Income^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Income + I(Income^2) + I(Income^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-786.14 -361.88  -55.17  312.13 1106.27

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.944e+02  8.498e+01   3.465 0.000589 ***
Income       3.812e+00  4.653e+00   0.819 0.413086
I(Income^2)  1.982e-02  6.584e-02   0.301 0.763495
I(Income^3) -3.533e-05  2.562e-04  -0.138 0.890382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 408.5 on 396 degrees of freedom
Multiple R-squared:  0.2166,    Adjusted R-squared:  0.2107
F-statistic: 36.51 on 3 and 396 DF,  p-value: < 2.2e-16
```

The p-value for the all the indicators combined against the target is < 2.2e-16, which means that at the 0.05 significance level, we reject the null hypothesis that all regression coefficients are zero. However, for a significance level of 0.05, we fail to reject the null hypothesis that there is no association between “Balance” and Income, $Income^2$, and $Income^3$ because the p-values for all the individual parameters also greatly exceeds the significance level. Hence, we fail to reject the null hypothesis. We infer that there is no association, linear or non-linear between Balance and Income. In addition, the Multiple R-Squared statistic of 0.2166 implies that 21.66% of the variation in Balance is explained by Income, $Income^2$, and $Income^3$.

vi. **Balance vs Rating:**

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Rating + I(Rating^2) + I(Rating^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Rating + I(Rating^2) + I(Rating^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-719.86 -129.97   -1.79  131.14  808.30

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.579e+02  1.038e+02  -4.411 1.33e-05 ***
Rating       2.658e+00  8.159e-01   3.258  0.00122 **
I(Rating^2)  9.159e-04  1.909e-03   0.480  0.63160
I(Rating^3) -1.403e-06  1.330e-06  -1.055  0.29226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 229.1 on 396 degrees of freedom
Multiple R-squared:  0.7536,    Adjusted R-squared:  0.7517
F-statistic: 403.7 on 3 and 396 DF,  p-value: < 2.2e-16
```

The p-value for the all the indicators combined against the target is < 2.2e-16, which means that at the 0.05 significance level, we reject the null hypothesis that all regression coefficients are zero. However, for a significance level of 0.05, we fail to reject the null hypothesis that there is no association between “Balance” and Rating^2, Rating^3 because the p-values for all the individual parameters also greatly exceeds the significance level. Hence, we fail to reject the null hypothesis. We infer that there is linear association between Rating and Balance no non-linear association between Balance and Rating. In addition, the Multiple R-Squared statistic of 0.7536 implies that 75.36% of the variation in Balance is explained by Rating, Rating^2, and Rating^3.

Upon fitting a model of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ for all indicators, we obtain the following results:

Source Code:

```
```{r}
fit_non_lm <- lm(Balance ~ Age + Education + Limit + Cards + Income + Rating + I(Age^2) + I(Age^3)
+ I(Education^2) + I(Education^3) + I(Limit^2) + I(Limit^3) + I(Cards^2) + I(Cards^3) + I(Income^2)
+ I(Income^3) + I(Rating^2) + I(Rating^3), data = df)
summary(fit_non_lm)
```
```

Model Summary:

```
Call:
lm(formula = Balance ~ Age + Education + Limit + Cards + Income +
    Rating + I(Age^2) + I(Age^3) + I(Education^2) + I(Education^3) +
    I(Limit^2) + I(Limit^3) + I(Cards^2) + I(Cards^3) + I(Income^2) +
    I(Income^3) + I(Rating^2) + I(Rating^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-256.49  -73.52  -37.52   14.95  521.51

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.113e+01  4.128e+02   0.197 0.844292
Age             1.533e+01  1.340e+01   1.144 0.253224
Education      -1.182e+02  8.216e+01  -1.439 0.150920
Limit          -1.841e-01  1.678e-01  -1.097 0.273280
Cards          -6.604e-01  4.991e+01  -0.013 0.989450
Income         -3.422e+00  1.804e+00  -1.897 0.058609 .
Rating          2.429e+00  3.047e+00   0.797 0.425835
I(Age^2)       -2.905e-01  2.509e-01  -1.158 0.247742
I(Age^3)        1.613e-03  1.492e-03   1.081 0.280240
I(Education^2)  1.061e+01  6.623e+00   1.603 0.109872
I(Education^3) -2.925e-01  1.717e-01  -1.703 0.089373 .
I(Limit^2)      7.035e-05  3.123e-05   2.252 0.024864 *
I(Limit^3)     -3.772e-09  1.757e-09  -2.148 0.032382 *
I(Cards^2)      4.326e+00  1.341e+01   0.323 0.747162
I(Cards^3)     -3.077e-01  1.074e+00  -0.286 0.774661
I(Income^2)    -9.017e-02  2.666e-02  -3.382 0.000795 ***
I(Income^3)     3.631e-04  1.088e-04   3.337 0.000929 ***
I(Rating^2)    -4.200e-03  7.522e-03  -0.558 0.576923
I(Rating^3)     4.288e-06  5.697e-06   0.753 0.452125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 145.4 on 381 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:    0.9
F-statistic: 200.4 on 18 and 381 DF,  p-value: < 2.2e-16
```

All the indicators put together, are strongly associated with Balance since the p-value associated with the model is <2.2e-16, which is far below the 0.05 significance level. The Multiple R-Squared statistic indicates that the predictors explain 90.45% of the variation in Balance, which is actually higher than that in section (b) when only linear relationships between the predictors and Balance were being examined. Furthermore, we see that there is sufficient evidence for us to conclude that there is a statistically significant association between the non-linear parameters of Income and Limit. However, the summary of this model, contradicts that of the models that were fit on the individual predictors as seen earlier.

Conclusion: When individual predictors are fit against Balance, we find insufficient evidence for non-linear relationships for all predictors except age. However, when a model with all indicators with non-linear parameters is fit, we find strong evidence of a non-linear relationship between Balance and (Income, Limit).

We can infer that the individual predictors do not share a non-linear relationship with Balance.