MS6061 Statistical Modelling

Regression Assignment

Dara Corr - 22275193

November 20, 2022

Contents

1 Introduction

This Assignment looks at creating and interpreting Regression models in R. This assignment is split into two main parts.

In Section A, we look at linear regression, using a dataset of the Wages of 5307 employees in a country across a number of industries. We take a random sample of 800 observations from this dataset to work with. We make a linear regression model to predict Annual Earnings using the *lm()* function and assess the predictive performance of the model using 10-fold cross validation. We also create a Robust Regression model to predict Annual Earnings from this dataset and we compare the results.

In Section B of the assignment, we fit a Logistic Regression Model to the Credit Default Dataset, containing data from 900 customers of a credit institution and we create a logistic regression model that tries to predict who is likely to default on their loans based on the data.

2 Linear regression

2.1 Question 1

First we imported the dataset **Wages.xls** into Rstudio and randomly selected 800 observations from it based on a column of random numbers that was added in Excel. Once the Data is in Rstudio, I categorized the categorical data in the dataset as categorical variables in R using the *factor()* function.

```r
WagesData <- data.frame(head(Wages, 800))#create dataframe of first 800 lines of sorted excel sheet

#Declare Categorical Data to R
WagesData$Gender.f <- factor(WagesData$Gender, levels =c(1,2), labels = c("Male", "Female"))
WagesData$Education.f <- factor(WagesData$Education, levels = c(0, 1), labels = c("high", "low"))
WagesData$service.f <- factor(WagesData$service, levels = c(1,2,3,4), labels = c("<2 years","2-5 years","6-10 years","10+
      years"))
WagesData$JobCategory.f <- factor(WagesData$JobCategory, levels = c(1,2,3,4), labels = c("Management","Professional","
      Assistant Professional","Clerical"))
WagesData$Sector.f <- factor(WagesData$Sector, levels = c(1,2,5,7), labels = c("Industry", "Construction & Transport", "
      Finance", "Health & Education"))
```

Then I created a linear model that contained all variables in the data.

```r
reg1 <- lm(Annual_Earnings ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + service.f +
      JobCategory.f + Sector.f, data = WagesData)
```

and I checked for multicolinearity and to see if the residuals are approximately normally distributed with no heteroscedasticity.

```r
#Check for Multicolinearity (VIFs):
car::vif(reg1)

#highest vif value = 2.99 for Time_paid_employ - some multicolinearity but not a major issue here
#second highest vif value is 2.41 for age
```

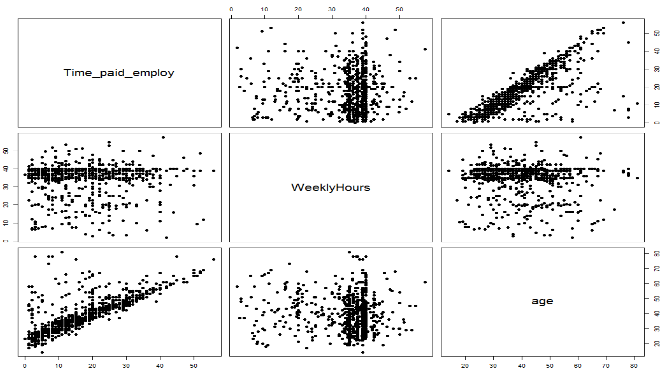

Figure 1: By plotting pairs of predictor variables, it is apparent that there is some multicolinearity between variables *Time_paid_employ* and *age*

We see there is some multicollinearity in the sample. *Time_paid_employ* has a VIF value of 2.9907 and *age* has a VIF of 2.4108. These values show that some of the predictor variables are moderately correlated since the VIFs are ≈ 2.5 which is a possible cause for concern. A

VIF value of 5 shows that there is high multicollinearity in the model, which would need to be addressed. Since none of the VIFs for the variables here are close to 5, we determine that multicollinearity is not a problem here and should not have a major impact on our model.

Now we plot the residuals to see how they are distributed.

```
res = resid(reg1)
fit = fitted(reg1)

plot(res,fit)
hist(res) #residuals not distributed normally
```

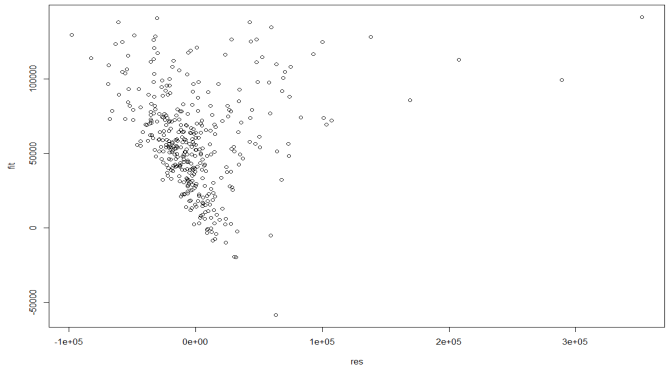

Figure 2: Plotting the residuals of the first linear regression model with all variables included in the model

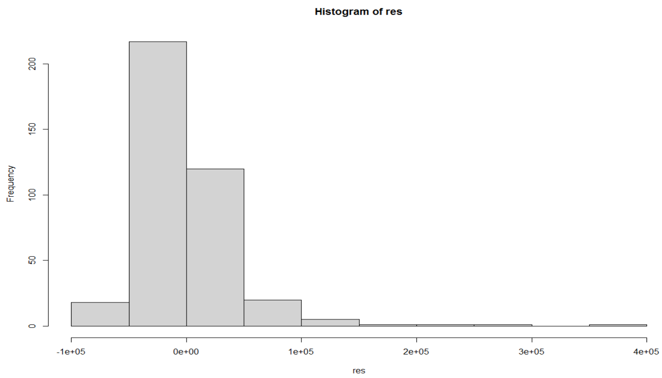

Figure 3: Histogram plot of the residuals of the first linear regression model with all variables included in the model

It is clear from these plots that the residuals in this model are not normally distributed and are clearly positively skewed as we see in the Histogram in Figure 3.
The Normality of residuals is an assumption of running a linear regression model. For our model to be vaild we need to transform the data in a way that gives us normally distributed residuals.

Since we have skewed data, we can try to use a log transformation of the dependent variable

Annual_Earnings to see if this will give us a model with normally distributed residuals.

```
#take log of Annual Earnings
reg2 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + service.f +
       JobCategory.f + Sector.f, data = WagesData)
summary(reg2) #->big R^2 improvement

res = resid(reg2)
fit = fitted(reg2)
plot(res,fit) #better distribution of residuals

hist(res)#residuals distributed approximately normally

#investigate summary to see what variables are significant and which ones can be removed:
summary(reg2)
```

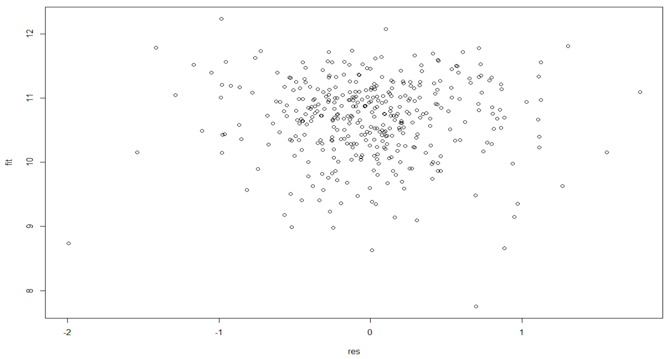

Figure 4: Plotting the residuals of the linear regression model with log transform of dependent variable Annual_Earnings

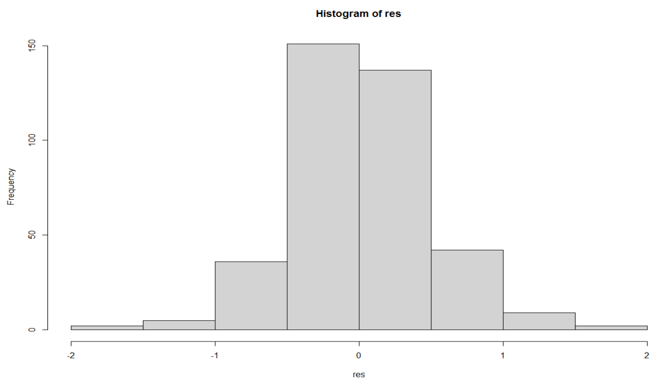

Figure 5: Histogram plot of the residuals of the linear regression model with log transform of dependent variable Annual_Earnings

We can see from the plots made in Figure 4 and Figure 5, that the log transformation of the dependant variable fixes the problem of the positively skewed distribution of residuals, and now that the residuals of this model are normally distributed, we have a linear regression model we can work with to make predictions.

Looking at the summary from this regression model, we can find out more about the model, such as the R^2 value and what values are significant or not.

```
Call:
lm(formula = log(Annual_Earnings) ~ No_Of_Weeks + Gender.f +
    Education.f + Time_paid_employ + WeeklyHours + age + service.f +
    JobCategory.f + Sector.f, data = WagesData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99466 -0.27626 -0.00485  0.25725  1.77810

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        8.183717   0.240424  34.039  < 2e-16 ***
No_Of_Weeks                        0.029942   0.003791   7.897 3.32e-14 ***
Gender.fFemale                    -0.180604   0.056112  -3.219   0.0014 **
Education.flow                    -0.474479   0.079374  -5.978 5.35e-09 ***
Time_paid_employ                   0.011579   0.004223   2.742   0.0064 **
WeeklyHours                        0.031596   0.003500   9.027  < 2e-16 ***
age                                0.004085   0.003473   1.176   0.2402
service.f2-5 years                -0.087017   0.091449  -0.952   0.3420
service.f6-10 years                0.073814   0.098347   0.751   0.4534
service.f10+ years                 0.098667   0.137917   0.715   0.4748
JobCategory.fProfessional         -0.192747   0.073507  -2.622   0.0091 **
JobCategory.fAssistant Professional -0.397341   0.093742  -4.239 2.85e-05 ***
JobCategory.fClerical             -0.583209   0.081524  -7.154 4.59e-12 ***
Sector.fConstruction & Transport  -0.084322   0.080737  -1.044   0.2970
Sector.fFinance                    0.168883   0.066281   2.548   0.0112 *
Sector.fHealth & Education        -0.247546   0.095981  -2.579   0.0103 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4987 on 368 degrees of freedom
  (416 observations deleted due to missingness)
Multiple R-squared:  0.6204,    Adjusted R-squared:  0.6049
F-statistic: 40.09 on 15 and 368 DF,  p-value: < 2.2e-16
```

Figure 6: Summary for regression model with log transformation of dependent variable and all independent variables included in the model

We note that this model has an R^2 value of 0.6204. The summary of this model also suggests that there are some insignificant variables in the model, namely *service* and *age*.
We can conduct a partial F-test to assess if the variable service is significant in our model or not.

```r
#age and service appear to be insignificant
#conduct partial f test with service to check if it is significant or not:

reg3 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + JobCategory.
    f + Sector.f, data = WagesData)
summary(reg3) #service removed from reg2 in reduced model for partial_f test

anova(reg2,reg3)
#p value is 0.06203 > 0.05 -> service is thus insignificant in this model
```

```
> anova(reg2,reg3)
Analysis of Variance Table

Model 1: log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f +
    Time_paid_employ + WeeklyHours + age + service.f + JobCategory.f +
    Sector.f
Model 2: log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f +
    Time_paid_employ + WeeklyHours + age + JobCategory.f + Sector.f
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    368 91.527
2    371 93.367 -3   -1.8394 2.4652 0.06203 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Partial F test output

The partial F test output shows that sector has a p-value of 0.06203 which is greater than the value of p = 0.05 for the 95% Confidence Interval. Because of this, we deem service to be an insignificant variable in our model and we can remove it.

After removing service from the model, we look to see what is the next insignificant variable is, so it can be removed (if there are any more insignificant variables in the model). From the summary table for *reg2*, we see that *age* has a p-value of $0.2402 > 0.05$ and thus it is also deemed insignificant and can be removed from the model.

Having removed *service* and *age* from the model we are now left with *reg4* with the following summary:

```
> summary(reg4)

Call:
lm(formula = log(Annual_Earnings) ~ No_Of_Weeks + Gender.f +
    Education.f + Time_paid_employ + WeeklyHours + JobCategory.f +
    Sector.f, data = WagesData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95847 -0.30636 -0.02611  0.24237  1.64770

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      8.286468   0.220781  37.532  < 2e-16 ***
No_Of_Weeks                      0.029789   0.003430   8.685  < 2e-16 ***
Gender.fFemale                  -0.185665   0.056095  -3.310  0.00102 **
Education.flow                  -0.462192   0.079736  -5.797 1.45e-08 ***
Time_paid_employ                 0.017699   0.002776   6.375 5.43e-10 ***
WeeklyHours                      0.030946   0.003505   8.830  < 2e-16 ***
JobCategory.fProfessional       -0.223623   0.073065  -3.061  0.00237 **
JobCategory.fAssistant Professional -0.425171   0.093618  -4.542 7.55e-06 ***
JobCategory.fClerical           -0.614452   0.081208  -7.566 3.04e-13 ***
Sector.fConstruction & Transport -0.069418   0.081001  -0.857  0.39200
Sector.fFinance                  0.183606   0.066392   2.765  0.00597 **
Sector.fHealth & Education      -0.237988   0.096433  -2.468  0.01404 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5021 on 372 degrees of freedom
  (416 observations deleted due to missingness)
Multiple R-squared:  0.6111,	Adjusted R-squared:  0.5996
F-statistic: 53.13 on 11 and 372 DF,  p-value: < 2.2e-16
```

Figure 8: Summary for regression model with log transformation of dependent variable and independent variables age and service removed from the model.

It is worth noting that about half of the data rows in this sample have missing values for *Education*. But the results of the linear regression model show that it is indeed statistically significant with a p-value of $1.45e-08$ and certainly important to include it into the model. If we were to remove *Education*, then we are omitting important data from the model and the R^2 value decreases from 0.6111 to 0.5977. So I chose to keep the *Education* variable in the model.

After checking that the residuals are normally distributed and that there is no heteroscedastcity in the residuals, we want to check that there are no influential outliers in the data. To do this I plotted the standardized residuals against leverage to see if any data point exceeded the Cook's Distance.

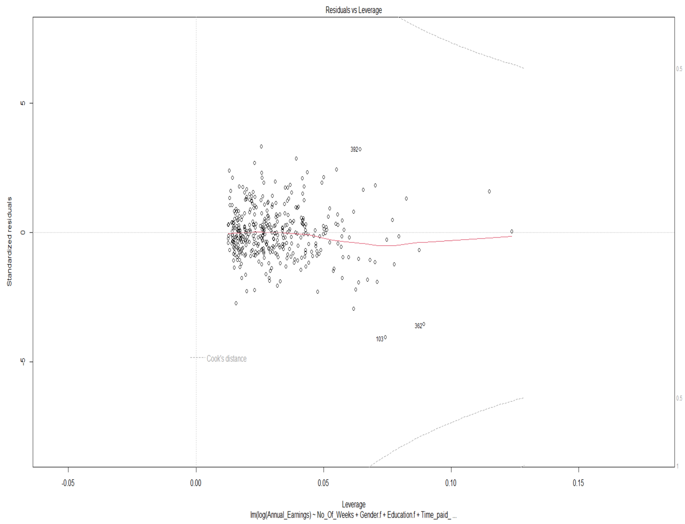

Figure 9: Plot of Standardized Residuals vs Leverage for the Linear Regression model *reg4*

We can see from the plot, that none of the data points exceed the cook's distance, so leverage is not an issue here.

Now we can use our model to make predictions and observations by taking the exponential of the regression coefficients since we applied a log transform to the dependent variable in our model:

```
> rounded_coefficients <-round((exp(reg4$coefficients) - 1)*100, 2)
> rounded_coefficients
                (Intercept)                     No_Of_Weeks                  Gender.fFemale
                  396878.82                            3.02                          -16.94
                Education.flow                Time_paid_employ                     WeeklyHours
                     -37.01                            1.79                            3.14
     JobCategory.fProfessional JobCategory.fAssistant Professional        JobCategory.fClerical
                     -20.04                          -34.63                          -45.91
  Sector.fConstruction & Transport                  Sector.fFinance       Sector.fHealth & Education
                      -6.71                           20.15                          -21.18
```

Figure 10: exponential of regression coefficients

Some of the observations I found from this model are:

- For every week worked, annual earnings are predicted to incrunderstand by 3.02% .
- Females are predicted to have 16.94% annual earnings, controlling for other variables in this model.
- Employees with low education are predicted to have 37.01% less annual earnings controlling for other variables in the model.
- For every year someone is employed, they are predicted to have an increase in annual earnings of 1.79%,controlling for other variables.

- If an employee works an additional hour every week, they are predicted to have their annual earnings increase by 3.14 %, controlling for other variables.
- This model predicts that a Professional earns 20.04% less per annum than someone working in management, controlling for other variables.
- This model predicts that an Assistant Professional earns 34.63% less per annum than someone working in management, controlling for other variables.
- This model predicts that a member of clerical staff earns 45.91% less per annum than someone working in management, controlling for other variables.
- This model predicts that someone working in Industry earns 6.71% less per annum than someone working in the Construction and Transport sector.
- This model predicts that someone working in Industry earns 20.15% more per annum than someone working in the Finance sector.
- This model predicts that someone working in Industry earns 21.18% less per annum than someone working in the Health and Education sector.

2.2 Question 2

Next, I assessed the performance of my final linear model from Question 1 using 10-fold cross validation.

I constructed the 10-fold cross validation in R as follows:

```
#Use 10-fold cross validation to assess the predictive performance of the final
#regression model in A(1)

WagesData$log_Earnings = log(WagesData$Annual_Earnings)

reg4 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f +
      Sector.f, data = WagesData)
summary(reg4)#compare residual standard error and R^2 here with one obtained from 10-fold CV

library(caret)
dataset <- data.matrix(WagesData[,c("log_Earnings","No_Of_Weeks","Gender.f","Education.f","Time_paid_employ","WeeklyHours","
      JobCategory.f","Sector.f")])

set.seed(22275193)
train_control <- trainControl(method = "repeatedcv",
                              number = 10, repeats = 3)

set.seed(22275193)
model <- train(log(Annual_Earnings)~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f +
       Sector.f, data = WagesData,
               method = "lm",
               trControl = train_control,
               na.action = na.pass)

print(model) #compare with summary(reg4)
```

which gives the following output:

```
> print(model) #compare with summary(reg4)
Linear Regression

800 samples
  7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 720, 720, 720, 720, 720, 720, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5106045  0.5930868  0.3834277

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 11: 10-fold Cross Validation output

We compare this output to the output for the summary of the model (*Summary(reg4)* shown in figure 8). The model has a Residual Standard Error of 0.5021 and an R^2 value of 0.6111. The result of the Cross Validation is a RMSE value of 0.5106 and an R^2 value of 0.5931. Since these values are quite similar, we can say that the model is consistent in the results it predictions it makes and it is suitable to make predictions with.

2.3 Question 3

Now we will carry out a Robust Regression on this dataset and compare the result of the Robust Regression with the result from the Final Model we used in Question 1 and Question 2. For the Robust Regression model, I included the same variables I included in the final regression model in Question 1, i.e. I kept the log transformation of the dependent variable and removed *age* and *service* from the model.

```
#want to keep RSE to a min value
library(MASS)

rob_reg <- rlm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f
      + Sector.f, data = WagesData)
summary(rob_reg)
summary(rob_reg)$sigma

rob_reg$w

rounded_coefficients <-round((exp(rob_reg$coefficients) - 1)*100, 2)
rounded_coefficients

plot(rob_reg)
```

```
> summary(rob_reg)

Call: rlm(formula = log(Annual_Earnings) ~ No_of_Weeks + Gender.f +
    Education.f + Time_paid_employ + WeeklyHours + JobCategory.f +
    Sector.f, data = WagesData)
Residuals:
      Min        1Q    Median        3Q       Max
-2.009551 -0.287145 -0.009492  0.272677  1.627707

Coefficients:
                                    Value   Std. Error t value
(Intercept)                          8.3743  0.2045    40.9600
No_of_Weeks                          0.0302  0.0032     9.5209
Gender.fFemale                      -0.1852  0.0519    -3.5659
Education.flow                      -0.4750  0.0738    -6.4330
Time_paid_employ                     0.0189  0.0026     7.3580
WeeklyHours                          0.0284  0.0032     8.7544
JobCategory.fProfessional           -0.2514  0.0677    -3.7155
JobCategory.fAssistant Professional -0.4389  0.0867    -5.0632
JobCategory.fClerical               -0.6484  0.0752    -8.6223
Sector.fConstruction & Transport    -0.0946  0.0750    -1.2613
Sector.fFinance                      0.1469  0.0615     2.3899
Sector.fHealth & Education          -0.2831  0.0893    -3.1701

Residual standard error: 0.4142 on 372 degrees of freedom
  (416 observations deleted due to missingness)
```

Figure 12: Summary of Robust Regression for Wages Dataset

Since the Residual Standard Error of 0.4142 is less than the value of 0.5021 that we had for the Residual Standard Error for the previous model we had (reg4), we can say that the Robust Regression Model is better at predicting the Annual Wages for this dataset.

An advantage of using Robust Regression instead of Least-Squares Regression, is that Robust Regression assigns weights to outliers which means that influential outliers will not have as much of an effect on the regression coefficients. This means that outliers have less of an impact on the model, whereas in least squares regression, outliers can have a large impact on the model which is not desirable.

One limitation of using Robust Regression, is that there is no R^2 coefficient like there is in Least Squares Regression. We can use the residual standard error to gauge the performance of the model instead.

3 Logistic Regression

Here we are given a data set called **Credit Default.xls** and it contains data from 900 customers of a credit institution. We want to create a Logistic Regression Model which will try to predict what Borrowers are likely to default on their debt.

There are 5 variables in this dataset. The Dependent Variable in our model is the *Default* variable. Then we will keep *Age*, *DebtRatio*, *YearlyIncome* and *LatePayment* as independent variables in the model.

First I declared the categorical data in R and then I created a Logistic Regression Model including all of the variables in the model.

```r

#Default and LatePayment are already coded correctly with 1s and 0s
Credit_Default$Default.f <- factor(Credit_Default$Default, labels = c("no","yes"))
Credit_Default$LatePayment.f <- factor(Credit_Default$LatePayment, labels = c("no","yes")) # in last 2 years

lreg <- glm(Default.f ~ Age + DebtRatio + YearlyIncome + LatePayment.f, data = Credit_Default, family = binomial)
summary(lreg)
```

```
> summary(lreg)

Call:
glm(formula = Default.f ~ Age + DebtRatio + YearlyIncome + LatePayment.f,
    family = binomial, data = Credit_Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0050  -0.8167  -0.6054   0.8801   2.0716

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.881047   0.328716   2.680  0.00736 **
Age            -0.030126   0.005882  -5.122 3.02e-07 ***
DebtRatio       0.194799   0.191594   1.017  0.30928
YearlyIncome   -0.008376   0.002829  -2.961  0.00307 **
LatePayment.fyes 1.859930  0.171527  10.843  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1192.3  on 899  degrees of freedom
Residual deviance: 1006.3  on 895  degrees of freedom
AIC: 1016.3

Number of Fisher Scoring iterations: 4
```

Figure 13: Logistic Regression Model output for Credit Default Data with all variables inluded in the model

We see that Debt Ratio has a p-value of $0.30928 > 0.05$, so we deem Debt Ratio insignificant and we can remove it from our model. Using a model with Debt Ratio removed, gives us a model with only significant variables remaining.

```r
summary(lreg2)
```

```
> summary(lreg2)

Call:
glm(formula = Default.f ~ Age + YearlyIncome + LatePayment.f,
    family = binomial, data = Credit_Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0312  -0.8143  -0.6082   0.8863   2.0333

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.959787   0.320475   2.995  0.00275 **
Age             -0.029895   0.005864  -5.098 3.43e-07 ***
YearlyIncome    -0.008693   0.002808  -3.096  0.00196 **
LatePayment.fyes 1.871240   0.171258  10.926  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1192.3  on 899  degrees of freedom
Residual deviance: 1007.3  on 896  degrees of freedom
AIC: 1015.3

Number of Fisher Scoring iterations: 4
```

Figure 14: Logistic Regression Model with *DebtRatio* removed

Now that we are only left with significant variables in the model, we can make predictions and observations according to the model.

It is clear that the most important variable when it comes to predicting whether someone will default on their debt or not, is *LatePayment* as it has the smallest p-value of $2e-16$. Age is the next important variable, followed by Yearly Income as third most important variable.

We can calculate odds ratios for this model by taking the exponential of the regression coefficients. And then we can get 95% Confidence Intervals of the Odds Ratios also.

```
exp(lreg2$coefficients)
#someone who was late paying their loan is 6.49 times more likely to default on their loan
#than someone who has not had a late payment in the last 2 years

#confidence interval (95%):
cbind(exp(lreg2$coefficients),exp(confint(lreg2)))
```

```
> cbind(exp(lreg2$coefficients),exp(confint(lreg2)))
Waiting for profiling to be done...
                             2.5 %    97.5 %
(Intercept)      2.6111407 1.3968565 4.9127840
Age              0.9705475 0.9592856 0.9816133
YearlyIncome     0.9913443 0.9858497 0.9967716
LatePayment.fyes 6.4963501 4.6637181 9.1321003
```

Figure 15: Logistic Regression Model output for Credit Default Data with all variables inluded in the model

Since the 95% Confidence Interval for the OR of *LatePayment* does not include the value of 1, we deem it to be statistically significant. The variables *Age* and *YearlyIncome* are also statistically significant since the Confidence Intervals of their odds ratios do not include the value of 1.

The Odds Ratios for Age and Yearly Income predict that as Age increases and as Yearly Income increases, the Odds Ratio of someone defaulting on their borrowings decreases. So, older borrowers are less likely to default on their Borrowings, compared to younger borrowers. Similarly, borrowers with larger Yearly Incomes are less likely to default on their borrowings compared to Borrowers with smaller yearly incomes.

The main conclusion from this analysis is that, borrowers who have made late payments of their debt in the last 2 years are 6.50 times more likely to default on their debt than borrowers who have not made any late payments on their debt in the last 2 years.

R code used

```r
library(carData)
library(car)
library(jtools)
library(tidyverse)
library(dplyr)

#A
#(1)
WagesData <- data.frame(head(Wages, 800))#create dataframe of first 800 lines of sorted excel sheet

#Declare Categorical Data to R
WagesData$Gender.f <- factor(WagesData$Gender, levels =c(1,2), labels = c("Male", "Female"))
WagesData$Education.f <- factor(WagesData$Education, levels = c(0, 1), labels = c("high", "low"))
WagesData$service.f <- factor(WagesData$service, levels = c(1,2,3,4), labels = c("<2 years","2-5 years","6-10 years","10+ years"))
WagesData$JobCategory.f <- factor(WagesData$JobCategory, levels = c(1,2,3,4), labels = c("Management","Professional"," Assistant Professional","Clerical"))
WagesData$Sector.f <- factor(WagesData$Sector, levels = c(1,2,5,7), labels = c("Industry", "Construction & Transport", " Finance", "Health & Education"))

WagesData <- subset(WagesData, select = -c(RANDnum))#removed Random Number column used to sort dataset into sample dataset
view(WagesData)

#create first regression model with all variables included in the model:
reg1 <- lm(Annual_Earnings ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + service.f + JobCategory.f + Sector.f, data = WagesData)
summary(reg1)
summ(reg1, vif = TRUE)

#Check for Multicolinearity (VIFs):
car::vif(reg1)

#highest vif value = 2.99 for Time_paid_employ - some multicolinearity but not a major issue here
#second highest vif value is 2.41 for age

colnames(WagesData)

pairs(WagesData[,1:10], pch = 16)
pairs(WagesData[,5:7], pch = 16) #some multicolinearity between predictor vals Time_paid_employ and age
cor(WagesData[,2:10])

#plot residuals

res = resid(reg1)
fit = fitted(reg1)

plot(res,fit)
hist(res) #residuals not distributed normally


#take log of Annual Earnings
reg2 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + service.f + JobCategory.f + Sector.f, data = WagesData)
summary(reg2) #->big R^2 improvement

res = resid(reg2)
fit = fitted(reg2)
plot(res,fit) #better distribution of residuals

hist(res)#residuals distributed approximately normally

#investigate summary to see what variables are significant and which ones can be removed:
summary(reg2)

#age and service appear to be insignificant
#conduct partial f test with service to check if it is significant or not:

reg3 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + age + JobCategory.f + Sector.f, data = WagesData)
summary(reg3) #service removed from reg2 in reduced model for partial_f test

anova(reg2,reg3)
#p value is 0.06203 > 0.05 -> service is thus insignificant in this model

res = resid(reg3)
fit = fitted(reg3)

plot(res,fit) #residuals look OK, no heteroscedasticity
summary(reg3)

#age looks to be infignificant -> remove it from model
```

```r
reg4 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f + Sector.f, data = WagesData)
summary(reg4)

anova(reg4,reg3) #pval =0.2028 > 0.05, so age is significant and can be removed from the model

res = resid(reg4)
fit = fitted(reg4)

plot(res,fit)#resid vs leverage plot shows that none of the outliers are influential
hist(res)

library(zoom)
zm()

colSums(is.na(WagesData))#416 NA values in Education - half of recorded Education vals are NA
#does model change much if education is removed?
summary(reg4)

rounded_coefficients1 <-round((exp(reg4$coefficients) - 1)*100, 2)
rounded_coefficients1

#model with education removed
reg5 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Time_paid_employ + WeeklyHours + JobCategory.f + Sector.f, data = WagesData)
summary(reg5)

rounded_coefficients2 <-round((exp(reg5$coefficients) - 1)*100, 2)
rounded_coefficients2

#keep reg 4 as full model
plot(reg4)#model looks good, there are several outliers but none exceed cooks distance

zm()
#run with reg4 as final model but flag that half data for education is NA

d<-density(reg4[['residuals']])
plot(d,main='Residual KDE Plot',xlab='Residual value')#confirming normal dist of residuals

exp(reg4$coefficients)

rounded_coefficients <-round((exp(reg4$coefficients) - 1)*100, 2)
rounded_coefficients

#can then comment on results in the report based on the coefficients above ^






########################################################################
#A(2)
#Use 10-fold cross validation to assess the predictive performance of the final
#regression model in A(1)

WagesData$log_Earnings = log(WagesData$Annual_Earnings)

reg4 <- lm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f + Sector.f, data = WagesData)
summary(reg4)#compare residual standard error and R^2 here with one obtained from 10-fold CV

library(caret)
dataset <- data.matrix(WagesData[,c("log_Earnings","No_Of_Weeks","Gender.f","Education.f","Time_paid_employ","WeeklyHours","JobCategory.f","Sector.f")])

set.seed(22275193)
train_control <- trainControl(method = "repeatedcv",
                              number = 10, repeats = 3)

set.seed(22275193)
model <- train(log(Annual_Earnings)~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f + Sector.f, data = WagesData,
               method = "lm",
               trControl = train_control,
               na.action = na.pass)

print(model) #compare with summary(reg4)

########################################################################
#A(3)
#Fit Robust Regression Model to predict Annual Earnings
#want to keep RSE to a min value
library(MASS)

```

```r
rob_reg <- rlm(log(Annual_Earnings) ~ No_Of_Weeks + Gender.f + Education.f + Time_paid_employ + WeeklyHours + JobCategory.f + Sector.f, data = WagesData)
summary(rob_reg)
summary(rob_reg)$sigma

rob_reg$w

rounded_coefficients <-round((exp(rob_reg$coefficients) - 1)*100, 2)
rounded_coefficients

plot(rob_reg)

##########################################################################
#Part B

#Default and LatePayment are already coded correctly with 1s and 0s
Credit_Default$Default.f <- factor(Credit_Default$Default, labels = c("no","yes"))
Credit_Default$LatePayment.f <- factor(Credit_Default$LatePayment, labels = c("no","yes")) # in last 2 years

lreg <- glm(Default.f ~ Age + DebtRatio + YearlyIncome + LatePayment.f, data = Credit_Default, family = binomial)
summary(lreg)

#DebtRatio is insignificant

lreg2 <- glm(Default.f ~ Age + YearlyIncome + LatePayment.f, data = Credit_Default, family = binomial)
summary(lreg2)

hist(lreg2$residuals)
#best model is lreg2

exp(lreg2$coefficients)
#someone who was late paying their loan is 6.49 times more likely to default on their loan
#than someone who has not had a late payment in the last 2 years

#confidence interval (95%):
cbind(exp(lreg2$coefficients),exp(confint(lreg2)))
```