sost20131

2023-01-08

SOST20131/SOST30031
Answering Social Research Questions with Statistical Models

Essay Assignment 2

Student ID: 10916038

Contents

Assignment 2

Question1

Getting to know the data - Exploratory data analysis

1.

First a general understating of the variables should be established so that it is intuitive to interpret whether or not the fitted model will make sense or not.

The response variable, FDI, concerns the level of investment into a particular country from foreign sources. A number of explanatory variables are employed that seek to explain the movement in FDI levels.

Intuitively, ROC (return on capital invested) should correlate positively with FDI as it implies that foreign investors will be able to achieve their goal of increasing their invested capital.

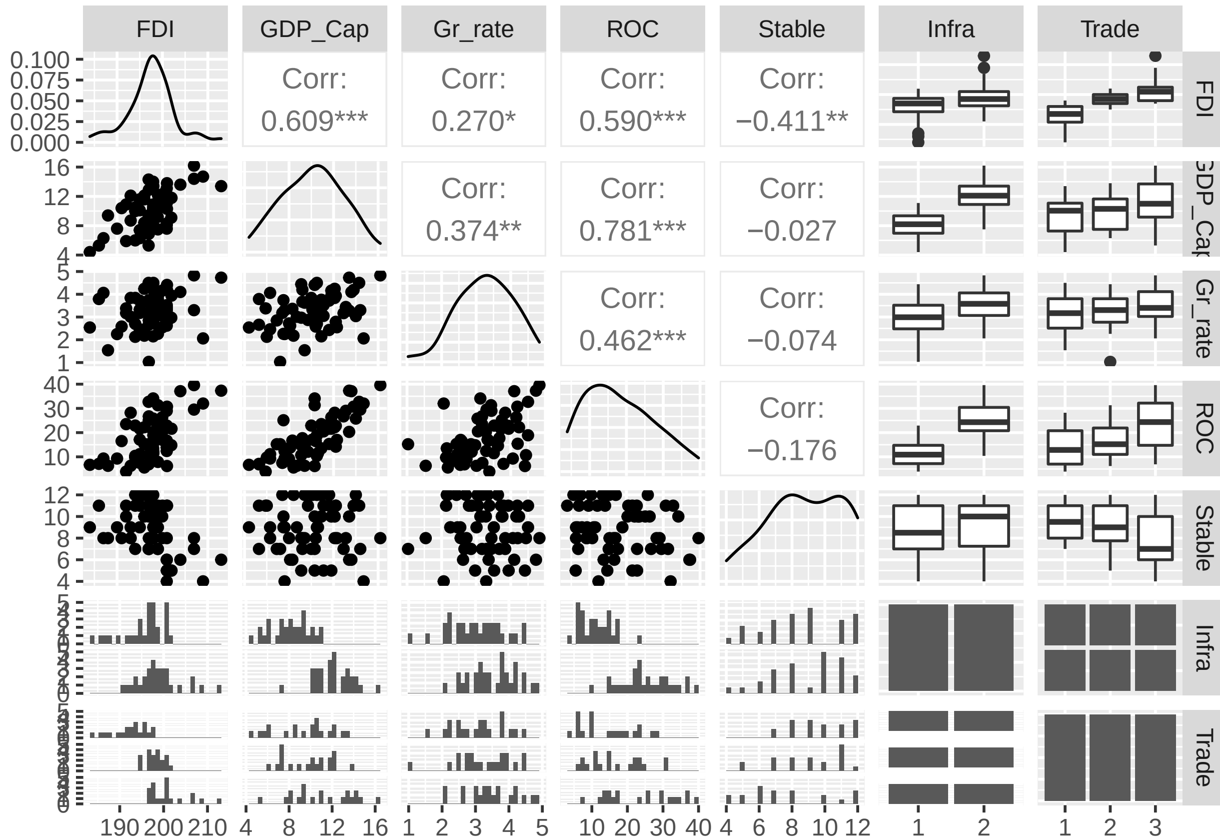
GDP per capita corresponds to individuals owning more wealth in said country. Investors may value this as an opportunity since investing in a country with a wealthier population can prove to be prosperous for their business. Thus a positive relationship between FDI and GDP per capita may be inferred.

For the same reason, a positive coefficient can be expected from variables Gr_rate, Infra and Trade as high values correspond to the country having a good track record for being able to sustain itself and encourage growth with good infrastructure and progressive trade legislation.

Finally, a negative coefficient should be expected from the variable Stable as a high number of government changes implies an unstable regime which can be unattractive for foreign investors.

Through R, the relationship between the variables can be visualized.

```r
mydata <- read.csv("https://tanjakec.github.io/SOST20131_30031/data/FDI.csv")
mydata$Infra <- as.factor(mydata$Infra)
mydata$Trade <- as.factor(mydata$Trade)
GGally::ggpairs(mydata)
```



It should be noted that the categorical variables have been converted into factors. This means that the factors will be stored as vectors of the integer values that they previously represented. A dummy variable is created for each level of the factor, the number of which is determined by k-1 with k being the number of categories for said variable, this is so that one category can be used as the reference level.

From the plot, it is evident that FDI is strongly correlated with GDP_Cap and ROC. Furthermore, strong relationships are present between other predictors such as ROC and GDP_Cap and ROC and Gr_rate. Furthermore, categorical variable Infra seems to show correlation between GDP_Cap as the medians of the distributions are quite different across the levels of Infra against GDP_Cap, this is also true for the spread as the interquartile range is not very similar. ROC also seems to be correlated with Infra for the same reason. This may prove to be problematic later as it implies multicollinearity.

Modelling

With an overview of the variables that are being dealt with, the model can now start to be put together. A stepwise approach will be taken whereby a saturated model will first be created and variables will be removed until the ideal model is achieved.

```
full_model <- lm(FDI ~., data = mydata) # includes all variables
summary(full_model)
```

```
##
## Call:
## lm(formula = FDI ~ ., data = mydata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.8594 -1.2595 -0.0808  1.4183  8.7210
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.472086   2.690418  70.425  < 2e-16 ***
## GDP_Cap       0.938713   0.236206   3.974 0.000219 ***
## Gr_rate      -0.089122   0.498508  -0.179 0.858807
## ROC           0.003144   0.088466   0.036 0.971782
## Stable       -0.539889   0.172155  -3.136 0.002816 **
## Infra2       -0.169110   1.493552  -0.113 0.910287
## Trade2        4.875539   0.891476   5.469 1.31e-06 ***
## Trade3        5.890833   1.179891   4.993 7.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.734 on 52 degrees of freedom
## Multiple R-squared:  0.7481, Adjusted R-squared:  0.7142
## F-statistic: 22.06 on 7 and 52 DF,  p-value: 1.646e-13
```

From the following results, multicollinearity can be identified. There are several reasons for this. The ROC coeffiecient is much lower than expected and is not significant. This may be due to ROC being correlated with GDP_Cap, Gr_rate and Infra.

```
car::vif(full_model)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## GDP_Cap 3.179681  1        1.783166
## Gr_rate 1.277878  1        1.130433
## ROC     5.267033  1        2.295002
## Stable  1.221857  1        1.105377
## Infra   4.476163  1        2.115695
## Trade   1.928662  2        1.178458
```

Also, with a vif value (Variance Inflation Factor) above 5 it is suggested that ROC is correlated with other explanatory variables in the model. Thus ROC will be removed from the model.

```
model1 <- lm(FDI ~. - ROC, data = mydata)
summary(model1)
```

```
##
## Call:
## lm(formula = FDI ~ . - ROC, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8633 -1.2684 -0.0897  1.4174  8.7346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.4790     2.6579  71.289  < 2e-16 ***
## GDP_Cap       0.9409     0.2258   4.167 0.000114 ***
## Gr_rate      -0.0846     0.4774  -0.177 0.860031
```

```
## Stable       -0.5413     0.1663  -3.255 0.001976 **
## Infra2       -0.1358     1.1519  -0.118 0.906598
## Trade2        4.8805     0.8723   5.595 7.92e-07 ***
## Trade3        5.9116     1.0151   5.824 3.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.708 on 53 degrees of freedom
## Multiple R-squared:  0.7481, Adjusted R-squared:  0.7196
## F-statistic: 26.23 on 6 and 53 DF,  p-value: 3.05e-14
```

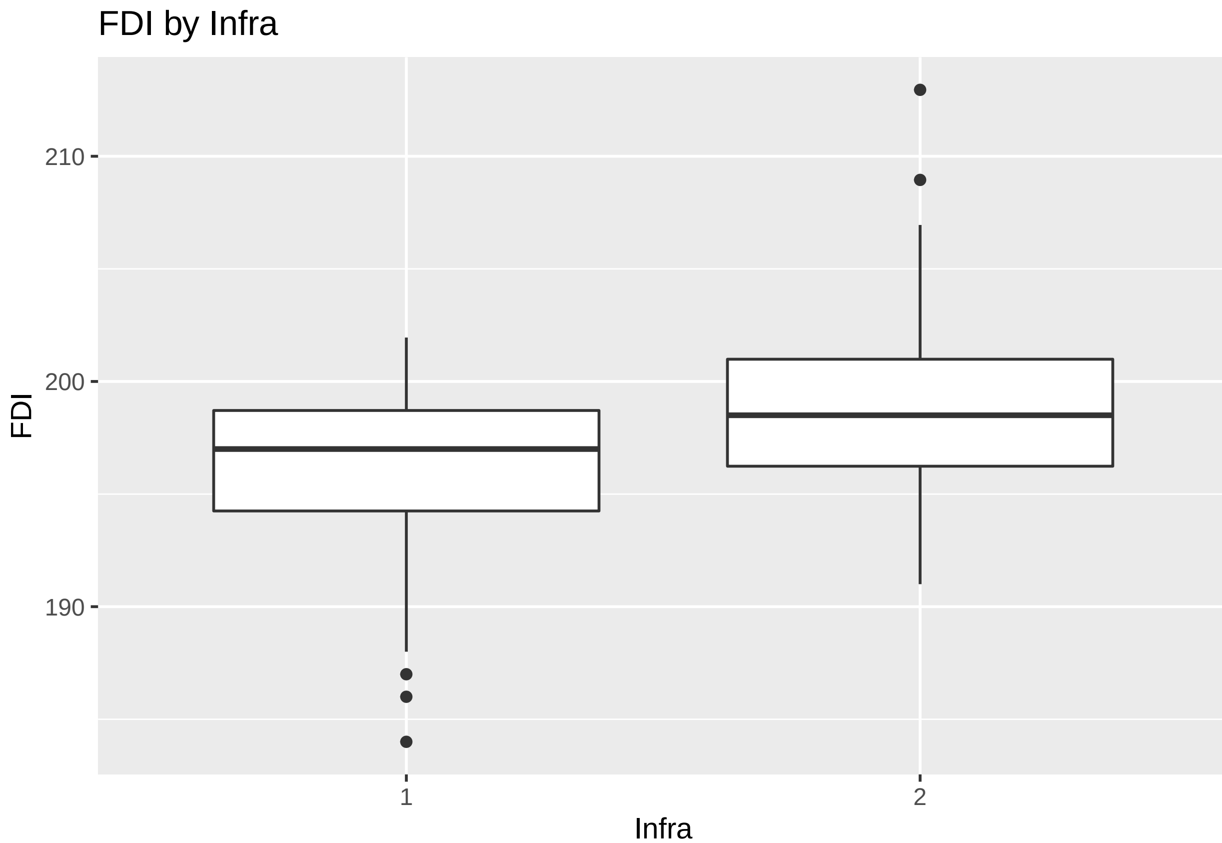
```r
car::vif(model1)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## GDP_Cap 2.961653  1        1.720945
## Gr_rate 1.194724  1        1.093034
## Stable  1.161560  1        1.077757
## Infra   2.713529  1        1.647279
## Trade   1.427570  2        1.093073
```

The R squared adjusted has changed slightly from 0.7142 to 0.7295 and the R squared value hasn't changed at all. There are no longer any vif values above 5. This proves that the explanatory power of ROC was able to be represented by other variables in the model, hence collinearity was certainly present.

```r
library(ggplot2)

ggplot(data = mydata, aes(x = Infra, y = FDI)) +
  geom_boxplot() +
  ggtitle("FDI by Infra")
```



It is hard to gauge from the box plot whether FDI and Infra are truly correlated. A t-test can be used to confirm whether these continuous and categorical variables are related.

h_0 : The means are equal

h_A : The means are different

```
t.test(mydata$FDI ~ mydata$Infra, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  mydata$FDI by mydata$Infra
## t = -2.6798, df = 58, p-value = 0.009572
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -5.8814438 -0.8518895
## sample estimates:
## mean in group 1 mean in group 2
##        195.8167        199.1833
```

With a p-value less than 0.05 we can conclude that they have significantly different means and are related.

However Infra shows up as not being significant in this regression model. Furthermore, Infras relationship from the previous box plot inspection between GDP_Cap suggests it may also be contributing to multi-collinearity as it seems to be correlated with GDP___Cap.

A t-test can be conducted to verify whether or not it should be removed from the model.

$h_0 : Infra = 0$, (Infra is unimportant)

$h_A : Infra > 0$ (Infra has a positive influence)

```
qt(0.95,53) #53 df
```

```
## [1] 1.674116
```

Infra t-value = -0.118

$-0.118 < 1.67$

$t_{calc} < tcrit$

Therefore we can reject the alternate hypothesis.

From all of the collective evidence it can be safely concluded that Infra is to be removed from the model.

```
model2 <- lm(FDI ~. - ROC - Infra, data = mydata)
summary(model2)
```

```
##
## Call:
## lm(formula = FDI ~ . - ROC - Infra, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8072 -1.3012 -0.0538  1.3545  8.7185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.63356    2.29103  82.772  < 2e-16 ***
## GDP_Cap       0.92068    0.14544   6.331 5.01e-08 ***
## Gr_rate      -0.09192    0.46906  -0.196  0.84537
## Stable       -0.54240    0.16445  -3.298  0.00173 **
## Trade2        4.89108    0.85969   5.689 5.34e-07 ***
## Trade3        5.95001    0.95263   6.246 6.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.683 on 54 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.7247
## F-statistic: 32.06 on 5 and 54 DF,  p-value: 5.104e-15
```

The R squared adjusted increases slightly, while it isn't a dramatic increase the reduction of a variable is a success in itself following the criteria of parsimony which prefers a model with fewer variables to one with many variables.

Finally, it is observed that Gr_rate has a negative coefficient, this goes against the inital analysis in which Gr_rate was expected to show a positive relationship with FDI. A t-test can also be conducted to test whether we should include this predictor by checking if the coefficient should be positive.

$h_0 : Gr_rate = 0$, (Infra is unimportant)

$h_A : Gr_rate > 0$ (Infra has a positive influence)

```r
qt(0.95, 54) #54 df
```

```
## [1] 1.673565
```

$-0.196 < 1.67$

$t_{calc} < tcrit$ Thus we can reject the alternative hypothesis and conclude that Gr_rate can be removed from the model.

```r
model3 <- lm(FDI ~. - ROC - Infra - Gr_rate, data = mydata)
summary(model3)
```

```
##
## Call:
## lm(formula = FDI ~ . - ROC - Infra - Gr_rate, data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.661 -1.300 -0.035  1.317  8.626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.4274     2.0173  93.903  < 2e-16 ***
## GDP_Cap       0.9109     0.1355   6.723 1.07e-08 ***
## Stable       -0.5411     0.1629  -3.322  0.00159 **
## Trade2        4.8837     0.8513   5.737 4.27e-07 ***
## Trade3        5.9368     0.9419   6.303 5.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.66 on 55 degrees of freedom
## Multiple R-squared:  0.7479, Adjusted R-squared:  0.7295
## F-statistic: 40.78 on 4 and 55 DF,  p-value: 7.573e-16
```

The model has improved since R squared adjusted value is now larger than before and the number of variables has been reduced since the first attempt with the most saturated model.

Given that we haven't identified correlations between the remaining predictors in our data analysis and since the vif indicates that multicollinearity isn't a concern the remaining model seems to be ideal.

```
##             GVIF Df GVIF^(1/(2*Df))
## GDP_Cap 1.105430  1        1.051394
## Stable  1.155769  1        1.075067
## Trade   1.266694  2        1.060884
```

Furthermore, if a backwards elimination approach is taken with the remaining variables we can see that the AIC for the initial model is 122.17. Removing any of the variables increases AIC which reduces the models fit. Hence it can be concluded that the current model is ideal:

```r
step(model3, direction = "backward")
```

```
## Start:  AIC=122.17
## FDI ~ (GDP_Cap + Gr_rate + ROC + Stable + Infra + Trade) - ROC -
##     Infra - Gr_rate
##
##          Df Sum of Sq    RSS    AIC
## <none>                389.10 122.17
## - Stable   1     78.08 467.18 131.14
## - Trade    2    341.80 730.90 156.00
## - GDP_Cap  1    319.80 708.90 156.16
```

```
##
## Call:
## lm(formula = FDI ~ (GDP_Cap + Gr_rate + ROC + Stable + Infra +
##     Trade) - ROC - Infra - Gr_rate, data = mydata)
##
## Coefficients:
## (Intercept)      GDP_Cap       Stable       Trade2       Trade3
##    189.4274       0.9109      -0.5411       4.8837       5.9368
```

Model comparison / evaluation / prediction

Finally, to compare the two models, the AIC's of the saturated model and reduced model can be weighed up. The AIC measures how good a model is based on goodness of fit and complexity with a lower AIC being preferable to a higher value.

```
extractAIC(model3)
```

```
## [1]   5.0000 122.1697
```

```
extractAIC(full_model)
```

```
## [1]   8.0000 128.1098
```

The extractAIC() function was used as the computation method to calculate the AIC is the same as the one previously used in the backwards elimination function allowing for easier interpretation.

The results show that the AIC of the reduced model is lower which proves that it is superior.

It can also be observed that the R squared adjusted value has risen from 0.7142 to 0.7295 which means the reduced model explains the data 1.5% better than the full model.

Furthermore, from the low p-value of the model (less than 0.05), it can be concluded that there is sufficient evidence that the observed effect exists in the larger population.

After arriving at the final model, it can be assumed that the model adheres to the principle of parsimony as it is a smaller model that doesn't have interactions between factors. In addition to this, the explanatory power of the model hasn't been reduced therefore it can be seen as a success.

2.

The equation for the reduced model is as follows:

$$FDI = b_0 + b_1 GDP\_Cap + b_2 Stable + b_3 Trade2 + b_4 Trade3 + e$$

$$FDI = 189.43 + 0.91 GDP\_Cap + -0.54 Stable + 4.88 Trade2 + 5.94 Trade3$$

As trade was converted into a factor at the start, it can be assumed that while holding all other variables constant, FDI increases by 4.88 when Trade is 2 rather than when it is at 1. Vice versa for Trade 3 whereby FDI increases by 5.94 when Trade is 3 rather than when it is 1.

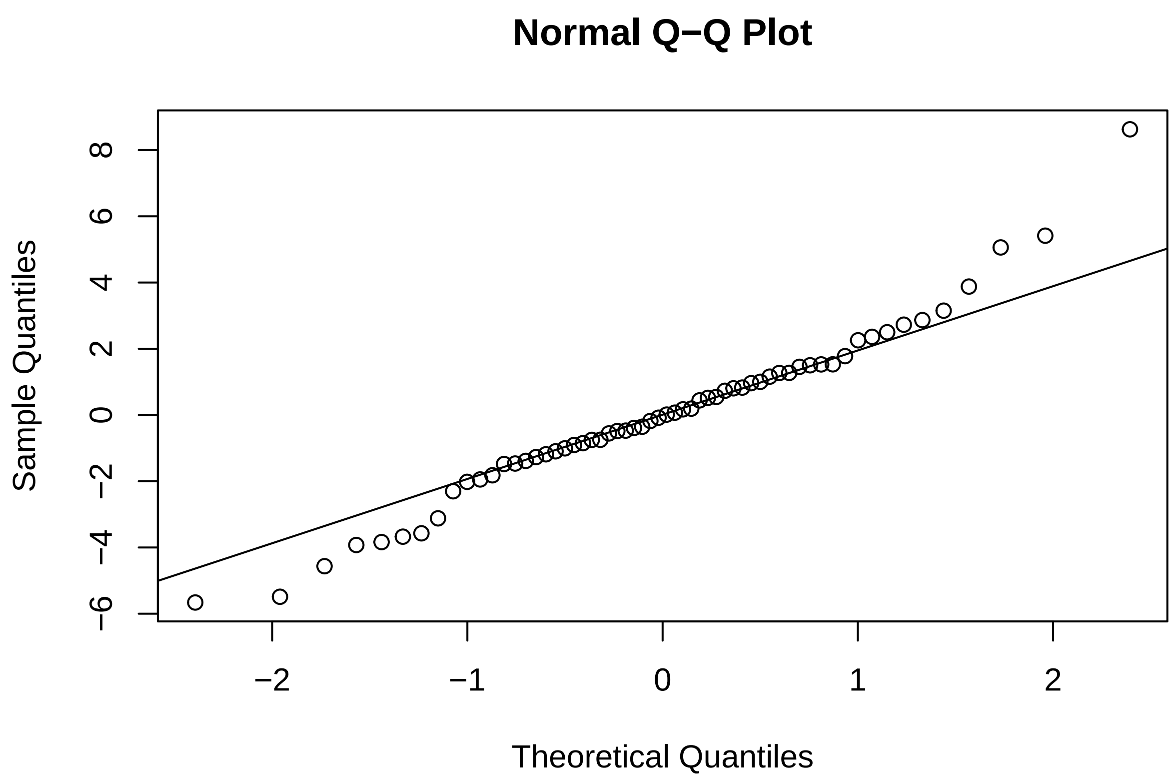
Therefore in the following situation: "The country receiving the investment has GDP per capita of 11.1 and Gr_rate per capita of 3.05; The average return on capital invested is 20.5%; There were 11 changes of government over the past 25 years and the Country has good infrastructure with some restrictions on trade."

The FDI can be calculated as such:

$$FDI = 189.43 + 0.91(11.1) + -0.54(11) + 4.88(1) + 5.94(0)$$

$= 198.471$

```
res <- resid(model3)
qqnorm(res)
qqline(res)
```




```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97026, p-value = 0.1503
```

The shapiro wilk test involves a null hypothesis of normality which is able to be rejected when a p-value is less than 0.05. This is not the case here and the model can be assumed to be normally distributed, the qq plot aids in visualising this fact. Thus the model and its predictions can be assumed to be valid as the residuals are normally distributed and good fit is also implied.

The results also make sense according to intuition. It is evident that Trade conditions strongly influence the FDI as foreign investors would be likely to invest in countries that favour relaxed trade legislation as it gives investors security in the fact that their relationship to their foreign investment won't be interrupted by the government. Furthermore the remaining predictor variables are also in line with the analysis conducted previously based on intuition.

Question 2

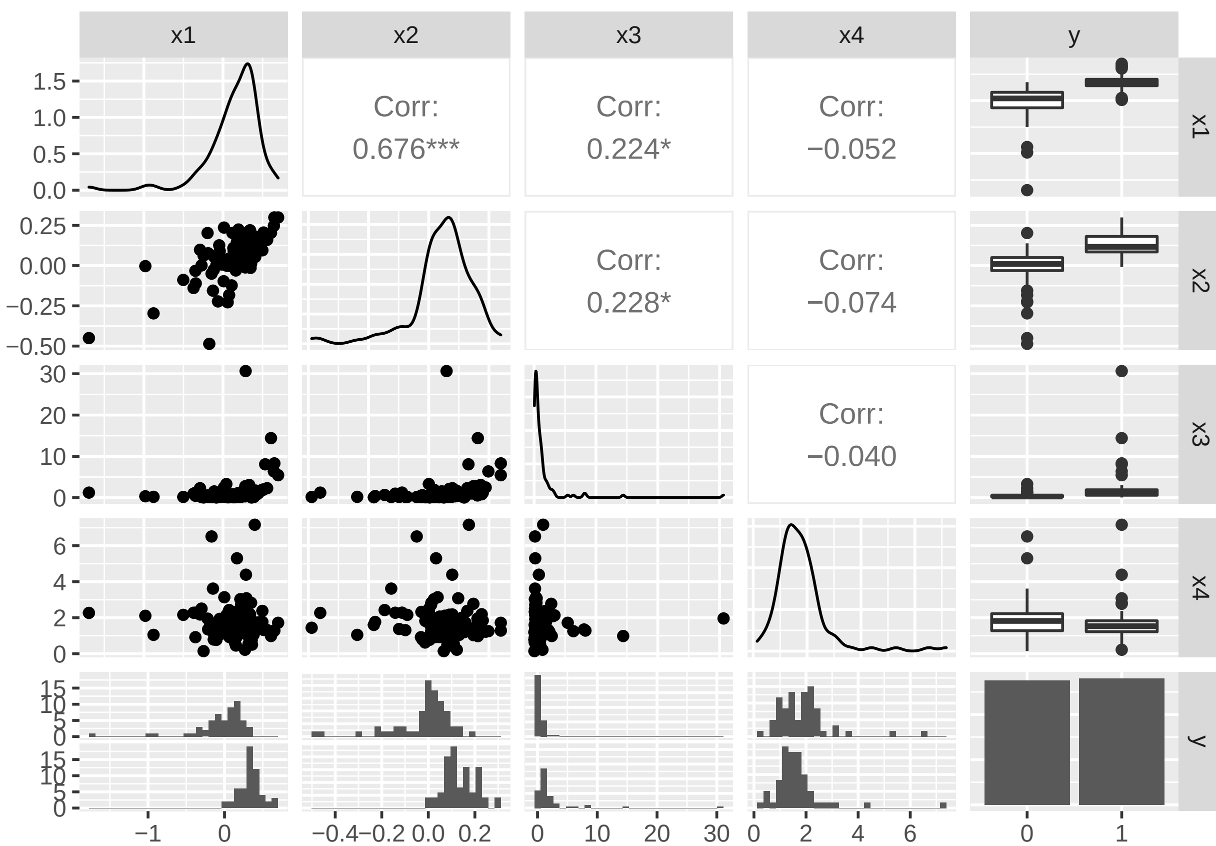
Data inspection

1.

Financial ratios can be used as tools to evaluate the performance of a company.

```r
ratios <- read.csv("https://tanjakec.github.io/SOST20131_30031/data/four_ratios.csv")
ratios$y <- as.factor(ratios$y)

GGally::ggpairs(ratios)
```



From the graph it seems like the distribution for x3 is positively skewed and all other variables seem to be negatively skewed with the exception of x4. This may lead to biased estimates and can negatively affect the validity of the model.

The box plots on the right indicate that it can be assumed that all of the financial ratios are related to y in some way. However the relationship between y and x3 and y and x4 seems to be ambiguous from the plot. A t-test can be used to clarify this.

```
t.test(ratios$x3 ~ ratios$y, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ratios$x3 by ratios$y
## t = -2.9724, df = 109, p-value = 0.003637
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -3.0697349 -0.6136852
## sample estimates:
## mean in group 0 mean in group 1
##       0.4549364       2.2966464
```

The p value is less than 0.05 hence the null hypothesis (there is no significant difference between the means) can be rejected and the alternative hypothesis (there is a significant difference between the means) can be accepted.

```
t.test(ratios$x4 ~ ratios$y, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ratios$x4 by ratios$y
## t = 1.079, df = 109, p-value = 0.283
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1751679  0.5938018
## sample estimates:
## mean in group 0 mean in group 1
##        1.878251        1.668934
```

In this case, the p value is larger than than 0.05 hence the null hypothesis (there is no significant difference between the means) can not be rejected and instead, the alternative hypothesis (there is a significant difference between the means) is rejected.

Thus it is implied that x4 may not be a significantly strong predictor of y, whereas some relationship can still be assumed between y and x3.

Logistic regression

2.

To assess the impact that the rations have on a companies' future (to stay solvent or go bankrupt) logistic regression can be used. A model will be fit and the equation should look as such:

$$y = \frac{1}{(1 + e^{(-(b0+b1x1+b2x2+b3x3+b4x4))})}$$

```r
set.seed(123)
split_idx = sample(nrow(ratios), 88) #80:20 split
ratios_train = ratios[split_idx, ]
ratios_test = ratios[-split_idx, ]
```

The model will be trained on 80% of the data and will be tested on the remaining 20% of data which is unseen, this will prevent the predictions from following a biased model.

```r
log_model <- glm(formula = y~.,data = ratios_train, family = binomial(logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(log_model)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(logit), data = ratios_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6776  -0.2707   0.0000   0.2391   2.4983
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4595     1.1610  -2.980  0.00288 **
## x1           10.9919     3.5831   3.068  0.00216 **
## x2           22.6859     7.8037   2.907  0.00365 **
## x3            1.2453     0.6162   2.021  0.04330 *
## x4           -0.6349     0.4222  -1.504  0.13263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 121.948  on 87  degrees of freedom
## Residual deviance:  41.577  on 83  degrees of freedom
## AIC: 51.577
##
## Number of Fisher Scoring iterations: 8
```

The most saturated model is fit including all of the variables. This provides a base model to which all newer models can be compared to.

x4 seems to be an insignificant predictor in the model. A chi squared test can also be conducted to confirm which variables are not to be included in the model.

```r
anova(log_model, test="Chisq")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    87    121.948
## x1    1   53.066        86     68.883 3.226e-13 ***
## x2    1   20.916        85     47.967 4.798e-06 ***
## x3    1    4.329        84     43.637   0.03746 *
## x4    1    2.060        83     41.577   0.15117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$h_0 : \beta_i = 0,$$
$$h_1 : \beta_i \neq 0$$

As x4 has a p-value that is above the threshold of 0.05 the null hypothesis isn't rejected. Thus it is safe to assume that x4 should be removed.

This can also be assumed to be an intuitive removal as a company may have high debt from investing in itslef but may have a high cashflow from the revenue of its projects. The company may have a low cash flow/debt ratio but it still may not go bankrupt as it has a steady income to pay off such debts. Furthermore, a company may have other assets to pay off its debts rather than cash flow therefore the ratio depicted by x4 seems to have weak explanatory power.

```r
log_model1 <- glm(formula = y~. - x4,data = ratios_train, family = binomial(logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(log_model1)
```

```
##
## Call:
## glm(formula = y ~ . - x4, family = binomial(logit), data = ratios_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.54752  -0.32153   0.00001   0.26181   2.59809
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.428      1.099  -4.031 5.56e-05 ***
## x1            10.084      3.229   3.123  0.00179 **
## x2            22.546      7.674   2.938  0.00330 **
## x3             1.176      0.597   1.969  0.04891 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 121.948  on 87  degrees of freedom
## Residual deviance:  43.637  on 84  degrees of freedom
## AIC: 51.637
##
## Number of Fisher Scoring iterations: 8
```

Following the removal of x4, the AIC value has increased which implies a drop in quality for the new model. However an AIC increase from 51.577 to 51.637 can be considered insignificant and the latter model can be preferred as it has fewer variables. This is in line with parsimony.

Furthermore it can be tested whether the predictor variables have explanatory power greater than 0 with the G statistic.

$$h_0 : \beta_i = 0,$$

$$h_1 :$$

at least one variable is significantly different to 0

```r
#check if model sig in predicting y
G_calc <- log_model1$null.deviance - log_model1$deviance
Gdf <- log_model1$df.null - log_model1$df.residual
pscl::pR2(log_model1)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -21.8186603 -60.9742227  78.3111247   0.6421658   0.5893028   0.7858725
```

```r
qchisq(.95, df = Gdf)
```

```
## [1] 7.814728
```

```r
1 - pchisq(G_calc, Gdf)
```

```
## [1] 1.110223e-16
```

Since 80.3715635 > 9.487729 the null hypothesis can be rejected.

Accuracy of the model

3.

To judge the accuracy of the model a confusion matrix can be constructed to represent true positive, false positive, false negative and true negative predictions. It is created based on the test dataset that was partitioned earlier.

```
#confusion matrix
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
response_pr <- round(predict(log_model1,  ratios_test, type = "response"), 2)

confusion_matrix <- table(ratios_test$y, round(response_pr))
confusion_matrix
```

```
##
##      0  1
##   0 12  0
##   1  1 10
```

From this, the accuracy can be calculated by taking the number of correct predictions in the diagonal, and dividing by the total predictions made.

```
accuracy <- function(x){
  sum(diag(x) / (sum(rowSums(x)))) * 100
}

accuracy(confusion_matrix)
```

```
## [1] 95.65217
```

The accuracy is 96%.

The following are the odds ratios for each coefficient in the model:

```
exp(coef(log_model1))
```

```
##  (Intercept)           x1           x2           x3
## 1.193309e-02 2.396377e+04 6.190297e+09 3.240324e+00
```

It should be noted that the coefficients for x1 and x2 are substantially above 1. This may suggest that the model is overfitting as the vif doesn't suggest any multicollinearity:

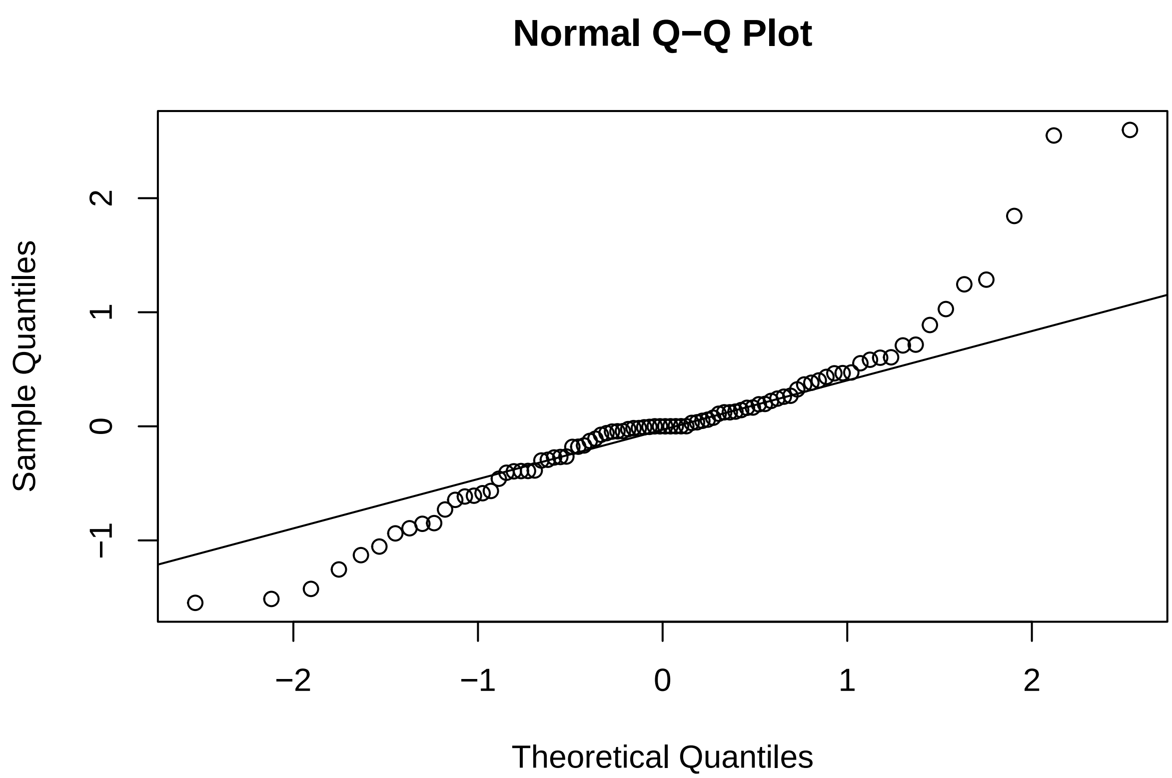
```
car::vif(log_model)
```

```
##       x1       x2       x3       x4
## 1.245393 1.045922 1.158541 1.194283
```

Therefore more data may be required to remedy this.

Furthermore, it seems that the residuals are not normally distributed. An implication of this would be to assume that the model is not a good fit.

```
res <- resid(log_model1)
qqnorm(res)
qqline(res)
```




```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.91772, p-value = 3.354e-05
```

The results from the shapiro test are significant as they are quite below 0.05 hence it can be assumed that the residuals are not normally distributed, the qq plot helps visualise this.

In the future, a dataset with a larger sample should be used to avoid these problems.