1. Create a multiple linear regression to explain third period exam (G3)

A. Are we able to use all 30 independent variables?

Answer:

No, we cannot use all 30 variables (excluding G1-G3 - done for all models and tests). The sample size is 395 and the square root of the sample size comes to ~20 (19.87 to be exact). So we should select 20 of the most significant variables to be included in the model. We shouldn't include more than 20 variables.

B. What assumptions do we need to check for? Are there any violations of these assumptions in this equation? If so, how would you correct them?

Answer:

- **Normally distributed** - Is the data normally distributed?-NO. **This assumption is violated**. We can correct this by normalizing the data or performing some kind of transformation (such as log) to make it normal. This is inferred from the Normal Q-Q Plot.
- **Linear relationship** - Do the x variables explain y variables? According to the Residual vs Fitted plot, linear relationship exists. **This assumption is NOT violated**.
- **No multicollinearity** - Are there NO strong relationships between the x variables? = checked using VIF.
  - If VIF > 10, there is multicollinearity. If VIF between 5 and 10 there is mild multicollinearity. This test is used to remove a variable from the model.
  - There is no variable with VIF > 10 or VIF between 5 and 10. So there is no multicollinearity or mild multicollinearity.
  - For this data, no VIF is greater than 10. Hence **this assumption is not violated**.
- **Homoscedasticity** - Are the variances equal? = Levine's Test. But here, by looking at the Scale-Location plot, we see equal variance does not exists. Hence, **this assumption is violated**. Can be fixed by transforming data.

C. Run a manual regression and check for VIF, what do we learn about VIF? Are there additional variables we need to remove from the model?

Answer:

All the VIF scores are < 10 and also not in the range between 5 and 10. Hence, we can say there is no multicollinearity and we do not need to remove any variables from the model, as determined by this VIF test.

```r
#1. Create a multiple linear regression to explain the third period exam(G3).

names(df)
#Check Spearman Correlations
df_filtered <- df[, c(1:30, 33)]

library(corrplot)

corrplot(cor(df_filtered, method = "spearman"))

corrplot(cor(df_filtered, method = "spearman"), method="number")


#Using Manual Multiple Linear Regression

#Create Initial Linear Regression Model with Enter Method

model1 <- lm(G3 ~ ., data=df_filtered)
model1

library(DescTools)
VIF(model1)

summary(model1)
```

```
> VIF(model1)
    school        sex        age    address    famsize    Pstatus       Medu       Fedu       Mjob       Fjob     reason   guardian traveltime
  1.446120   1.425251   1.630225   1.310588   1.117916   1.120734   2.201222   1.841650   1.451142   1.138936   1.116758   1.223092   1.246556
 studytime   failures  schoolsup     famsup       paid activities    nursery     higher   internet   romantic     famrel   freetime      goout
  1.296937   1.316950   1.146071   1.249942   1.283765   1.130280   1.110713   1.239064   1.204926   1.123389   1.102293   1.259978   1.412500
      Dalc       Walc     health   absences
  1.888434   2.245642   1.124729   1.214718
```

```
Residual standard error: 4.136 on 364 degrees of freedom
Multiple R-squared:  0.2469,    Adjusted R-squared:  0.1849
F-statistic: 3.979 on 30 and 364 DF,  p-value: 1.371e-10
```

D. Create a plot of the model diagnostics. What do we learn from these plots? Are there any issues with the model? Do you trust the reliability of this model?

Answer:

Plots:

Residual vs Fitted Plot:

- Checks for linear relationship - have linearity between x and y. The red line will be completely horizontal if it's linear. But there is no pattern. **Hence, linear relationship exists.**

Normal Q-Q Plot:

- We can see there are several outliers as some dots don't follow the straight line. **Hence, we see the data is not normal**

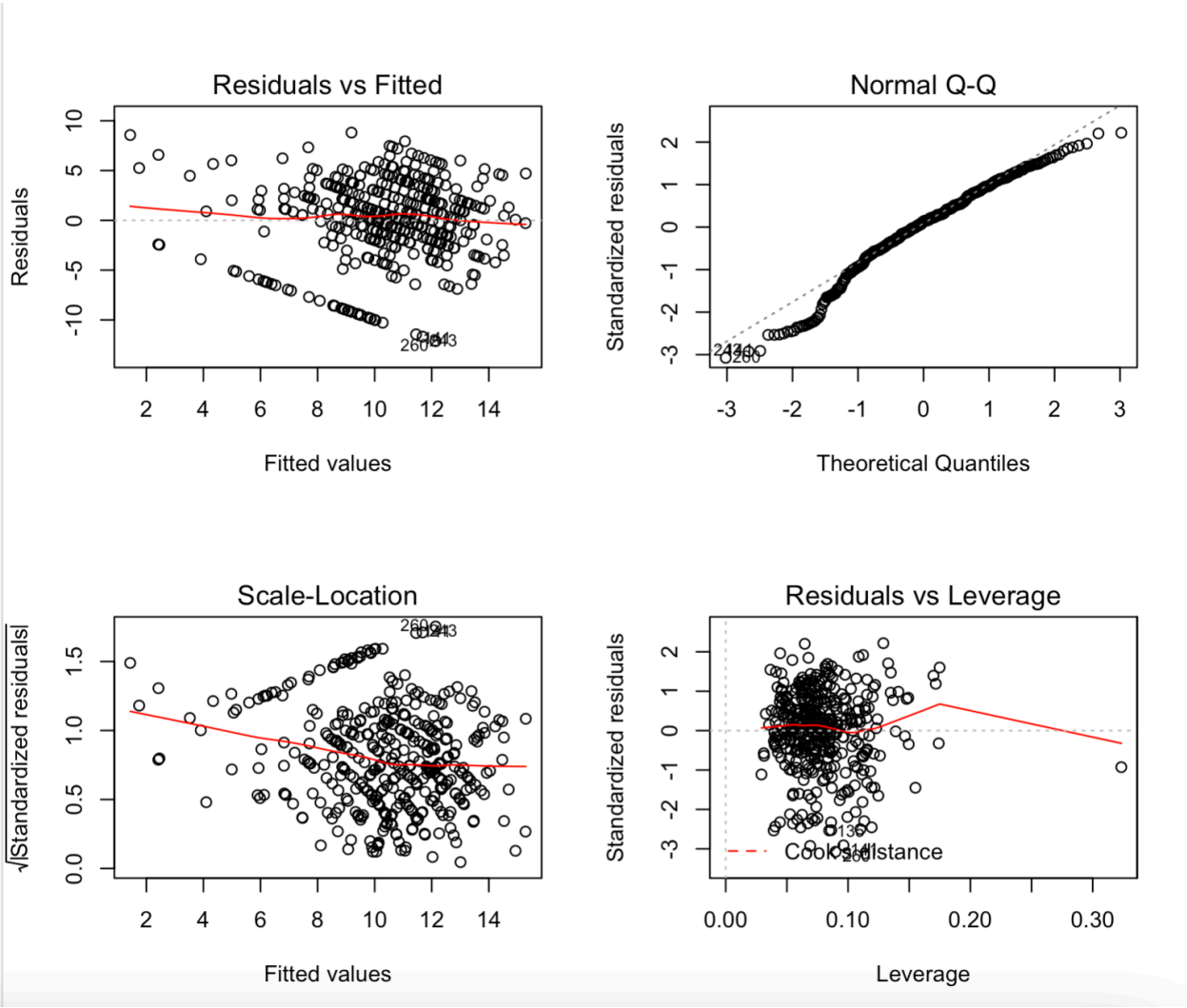
Scale-Location:

- Checks for homoscedasticity. The red line has to be straight with equal spread. In this model, the red line is approximately straight, hence **there is heteroskedasticity => no equal variances**.

Residual vs Leverage:

- Checks for influential outliers - has high leverage and high cooks distance. There are no influential outliers in this dataset - no points above the horizontal Cook's distance line with high leverage.
- Outliers do exist but are not influential. Hence they must be removed.

```
#Diagnostic Plots for Model Fit
par(mfrow = c(2, 2))
plot(model1)
```



Since the assumptions from these plots are violated, there are issues when trying to fit a linear model to this dataset and hence is not reliable. Also since it includes all 30 variables when only 20 should be included, this model is not reliable. To make it more reliable, we can transform the data, remove outliers and use a feature selection method to filter out the significant variables.

E. Using stepwise equation, explain the outputs of the regression. Comment on overall significance of the regression fit. Which predictors have coefficients that are significantly different from zero at the .05 level?

Answer:

Stepwise Regression

- Errors in this model run from-12.2556 (MIN residual) to 8.4802 (MAX residual)
- P-value = 2.843e-15 < 0.05. Hence we reject null hypothesis of "all beta coefficients = 0 or are the same" and accept the alternate hypothesis of "beta coefficients are different from 0 and are different from themselves".
- Multiple R-squared: 0.2256, and Adjusted R-squared: 0.1992
  - From this we can see that adjusted R-squared is not beyond 5% of multiple R-squared. So there is little to no overfitting or multicollinearity. **So we have good initial fit**.
- Using significance levels, the following variables are significant from 0 at the 0.05 level:

Variable	Significance Score	Beta Coefficient	Interpretation
Failures	7.81e-08 ***	-1.66755	For every increase in number of past class failures, G3 decreases by 1.66
Medu	0.0141 *	0.50817	For every increase in category of mother education, G3 increases by 0.508
Sex	0.0108 *	1.16436	For every increase in category of sex, G3 increases by 1.16436
goout	0.0109 *	-0.48421	For every increase in category of going out with friends, G3 decreases by 0.48421
Romantic	0.0162 *	-1.09284	For every increase in category of romantic relationship, G3 decreases by 1.092

```r
#Using Stepwise Multiple Linear Regression

null = lm(G3 ~ 1-G3, data=df_filtered)
null

full = lm(G3 ~ .-G3, data=df_filtered)
full

#Stepwise Regression
train_Step = step(null, scope = list(upper=full), direction="both")
summary(train_Step)
```

```
> summary(train_Step)

Call:
lm(formula = G3 ~ failures + Medu + sex + goout + romantic + 
    reason + famsize + schoolsup + address + higher + famsup + 
    studytime + absences, data = df_filtered)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2256  -1.9274   0.4789   2.7016   8.4802 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.64359    2.81875   2.002   0.0460 *  
failures    -1.66755    0.30438  -5.479 7.81e-08 ***
Medu         0.50817    0.20616   2.465   0.0141 *  
sex          1.16436    0.45442   2.562   0.0108 *  
goout       -0.48421    0.18917  -2.560   0.0109 *  
romantic    -1.09284    0.45267  -2.414   0.0162 *  
reason       0.32846    0.17609   1.865   0.0629 .  
famsize      0.67025    0.46308   1.447   0.1486    
schoolsup   -1.05676    0.62791  -1.683   0.0932 .  
address      0.91627    0.50674   1.808   0.0714 .  
higher       1.60425    1.01688   1.578   0.1155    
famsup      -0.68385    0.44323  -1.543   0.1237    
studytime    0.43085    0.26824   1.606   0.1091    
absences     0.04174    0.02676   1.560   0.1197    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.1 on 381 degrees of freedom
Multiple R-squared:  0.2256,    Adjusted R-squared:  0.1992 
F-statistic: 8.539 on 13 and 381 DF,  p-value: 2.843e-15
```

2. Create a linear regression to explain G3 using **Lasso Regression**

A. **How is Lasso regression different from manual/automatic regression?**

Answer:

Unlike manual/automatic linear regression, lasso regression does not give out p-values. Lasso regression can be used as a model selection technique - we can take the significant variables from Lasso regression and run them using a regular linear regression model. Lasso regression can be used to select the model with best coefficients.

B. Explain the output of Lasso regression.

Answer:

```
#Lasso Regression

x=model.matrix(G3 ~ ., data=df_filtered)
y=df_filtered$G3

library(glmnet)
cv.lasso <- cv.glmnet(x,y, typemeasure="mse", alpha=1)
cv.lasso

ls(cv.lasso)

Lambda.best <- cv.lasso$lambda.min

predict(cv.lasso, s = Lambda.best, type = "coefficients")


#Ordinary Least Squares Regression

model2 <- lm(G3 ~ sex+Medu+reason+failures+higher+romantic+goout, data=df_filtered)
summary(model2)
```

Continued…

```
> Lambda.best <- cv.lasso$lambda.min
> predict(cv.lasso, s = Lambda.best, type = "coefficients")
32 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept)  9.75232795
(Intercept)  .
school       .
sex          0.67402573
age         -0.09687147
address      0.30927018
famsize      0.37246698
Pstatus     -0.04936055
Medu         0.38120032
Fedu         .
Mjob         .
Fjob         .
reason       0.20785645
guardian     .
traveltime  -0.12988967
studytime    0.12247643
failures    -1.59646071
schoolsup   -0.63720831
famsup      -0.29122574
paid         .
activities   .
nursery      .
higher       0.90091735
internet     0.20937627
romantic    -0.63372798
famrel       .
freetime     0.01054110
goout       -0.29124301
Dalc         .
Walc         .
health       .
absences     0.01269533
>
```

- For Lasso regression, we separate out the x and y (dependent and independent) variables.
- Using the *glmnet* function, we specify the appropriate parameters for lasso regression: using alpha=1 and mean square error as the measure of performance.
- We use **lambda.min** to get the estimates and the *predict* function to get the coefficients for all the variables.
- We get the following significant variables as Lasso regression **output**: sex, age, address, famsize, Pstatus, Medu, reason, traveltime, studytime, failures, schoolsup, famsup, higher, internet, romantic, freetime, goout and absences. They all have coefficients > 0 or < 0.
- Note: the output of Lasso changes every time we run it due to the random selection of k-folds when training.
- Using those variables that are significant from Lasso regression, we build an ordinary least squares regression model.

C. What do you learn about student grades based upon this regression?

Answer:

From the lasso regression, the following variables are significant in the OLS model - We only look at the significant variables - those which affect G3 in positive or negative ways:

Variable	Coefficient	Interpretation
sex	1.00331	1 unit increase in category of sex results in 1.003 increase in G3
Medu	0.42038	1 increase in category of mothers education results in 0.42 increase in G3
failures	-1.62499	1 increase in number of failures results in decrease of G3 by 1.62
romantic	-1.06967	1 increase in category of romantic relationships results in decrease of G3 by 1.06
goout	-0.52001	1 increase in category of number of mins spent going out, results in decrease of G3 by 0.52

Based on the model coefficients, we can see that the interpretations make sense. Using these coefficients we build an ordinary least squares regression model to build a more efficient model with significant variables.

This model was built using the significant variables got from Lasso Regression. This is called the ordinary lease squares model.

```
Call:
lm(formula = G3 ~ sex + age + address + famsize + Pstatus + Medu +
    reason + traveltime + studytime + failures + schoolsup +
    famsup + higher + internet + romantic + freetime + goout +
    absences, data = df_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3602  -2.0206   0.4523   2.6863   9.0590

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.70852    4.93026   2.172  0.03048 *
sex          1.00331    0.46554   2.155  0.03178 *
age         -0.22989    0.18513  -1.242  0.21509
address      0.61527    0.54836   1.122  0.26257
famsize      0.66596    0.47031   1.416  0.15760
Pstatus     -0.50964    0.70346  -0.724  0.46923
Medu         0.42038    0.21203   1.983  0.04814 *
reason       0.32353    0.17649   1.833  0.06757 .
traveltime  -0.20497    0.32173  -0.637  0.52447
studytime    0.45051    0.27108   1.662  0.09737 .
failures    -1.62499    0.30873  -5.263 2.38e-07 ***
schoolsup   -1.27687    0.65058  -1.963  0.05042 .
famsup      -0.77132    0.44824  -1.721  0.08611 .
higher       1.41898    1.02555   1.384  0.16729
internet     0.50502    0.59168   0.854  0.39390
romantic    -1.06967    0.45691  -2.341  0.01975 *
freetime     0.27788    0.22374   1.242  0.21503
goout       -0.52001    0.19825  -2.623  0.00907 **
absences     0.04506    0.02758   1.634  0.10319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.1 on 376 degrees of freedom
Multiple R-squared:  0.2359,    Adjusted R-squared:  0.1993
F-statistic: 6.448 on 18 and 376 DF,  p-value: 5.626e-14
```

D. Are the lasso regression results different from the stepwise regression in the previous problem? If there is a difference, how are the two results different?

Answer:

- The result of lasso regression is different from stepwise regression. Lasso regression does not give a p-value or F-stats. It gives us only the significant variables.

Stepwise Regression Result:
The significant variables chosen by the model are: **failures, Medu, sex, goout and romantic** (taking only variables with * - with some significance relative to alpha = 0.05).

The stepwise regression is not used to select variables but takes into account significant variables when compared to manual regression.

Lasso Regression Output:

Lasso regression had non-zero coefficients for the following variables: **sex, age, address, famsize, Pstatus, Medu, reason, traveltime, studytime, failures, schoolsup, famsup, higher, internet, romantic, freetime, goout and absences.**

We can use the output of lasso regression to build a better linear model. We build the linear model with the variables selected above. The results vary because the significant variables selected by each of the models are different.

When we compare OLS to step-wise regression:

If we compare the results of the stepwise regression and the model build after variable selection via Lasso, we can compare the p-values, F-statistics and R squared. When we compare the R squared values, we can see the model trained after variable selection is similar and not too different, but the step-wise model performs better.

Since Lasso regression selects different k-folds when run each time, the outputs can vary. We might get a better model if we run the regression several times and get the best result from that.