ADM 2304 - Assignment 4

Renad Gharz

16/05/2022

```
## Loading required package: carData
```

```
## New names:
## * `` -> `...4`
```

1. Compensation of Senior Executives

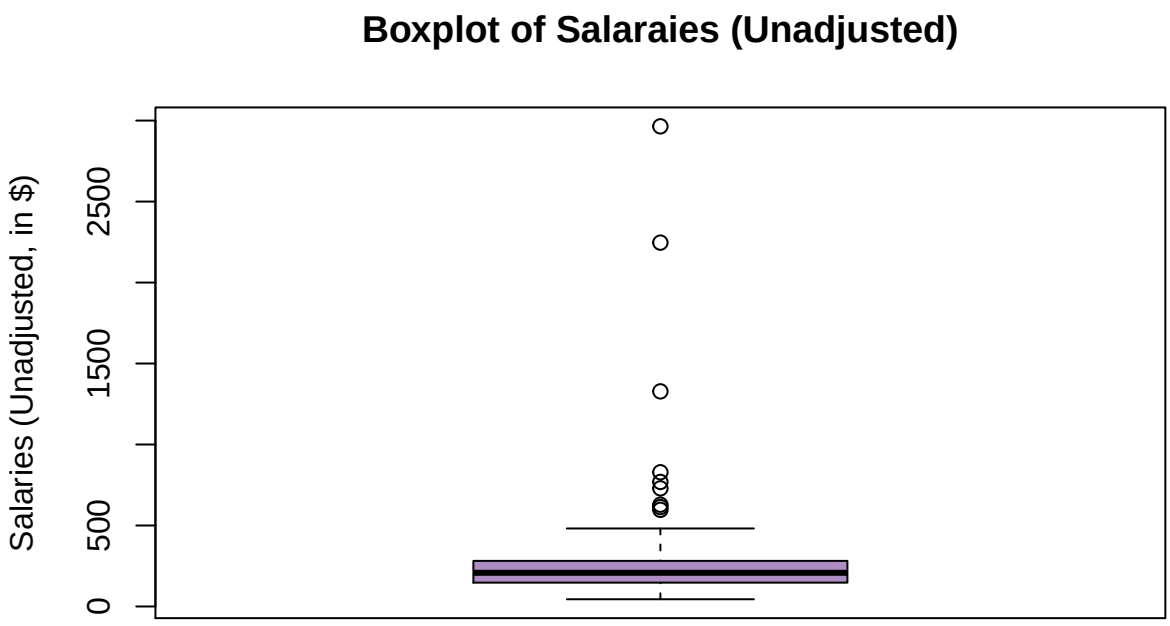
A large international bank is reviewing its compensation policy for their senior executives. They want to closely examine the relationship between senior executives' salaries and the growth in their business portfolio. The dataset **Executive Salaries** contains last year's annual compensation of the bank senior executives (in thousands of dollars), including performance pay, and the annual rate of growth of their respective business portfolio (as a percentage). These variables are called *Salary_raw* and *ROG_raw*, respectively.

1.a

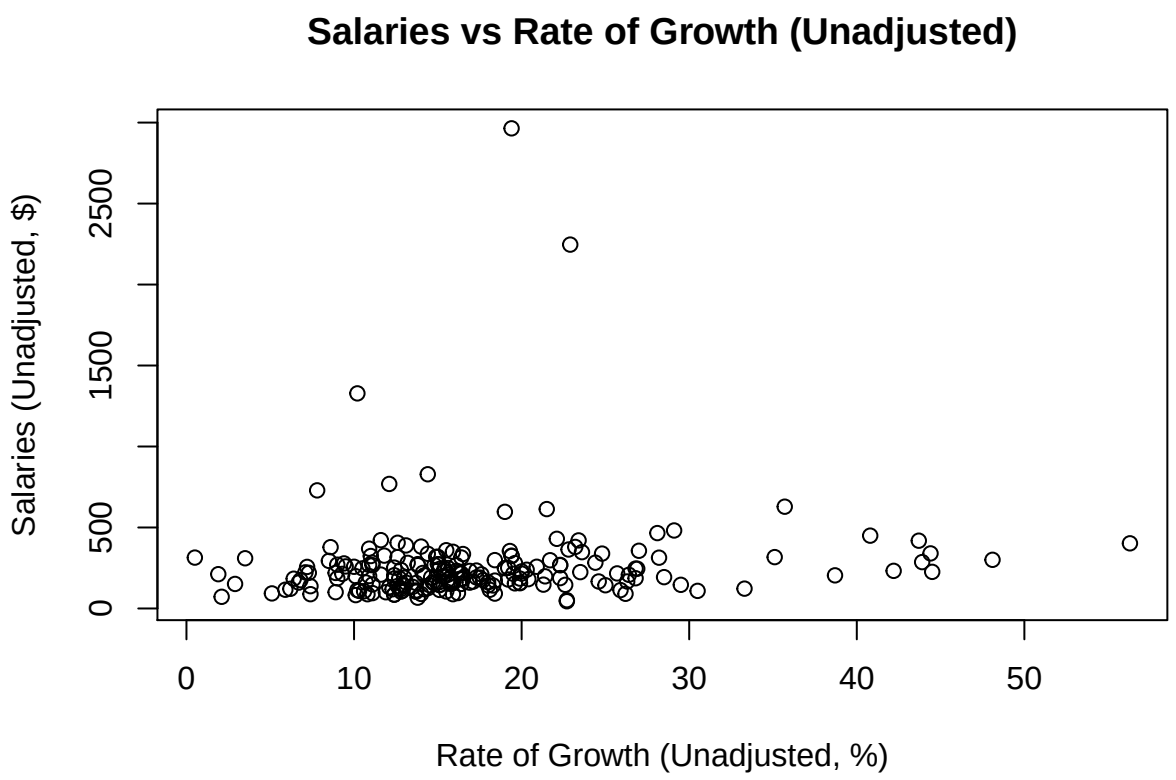
Draw a boxplot of the salary of senior executives (variable *Salary_raw*) and produce a scatter plot of this variable against the annual rate of growth (variable *ROG_raw*). Are there any apparent outliers in the data? Are there high leverage points?

```
##Boxplot
boxplot(
  exec_sals$Salary_raw,
  main = "Boxplot of Salaraies (Unadjusted)",
  ylab = "Salaries (Unadjusted, in $)",
  col = brewer.pal(1, "PRGn"))
```

```
## Warning in brewer.pal(1, "PRGn"): minimal value for n is 3, returning requested palette with 3 diffe
```




```
##Scatterplot
plot(
  exec_sals$ROG_raw,
  exec_sals$Salary_raw,
  main =
    "Salaries vs Rate of Growth (Unadjusted)",
  xlab = "Rate of Growth (Unadjusted, %)",
  ylab = "Salaries (Unadjusted, $)")
```



Yes, there are several apparent outliers. We can spot them by looking at the boxplot which shows 2 groups of outliers. The first group consists of 3 extreme outliers (the 3 furthest points above the upper fence) very distance from the rest of the data (and even distanced between themselves). The second group consists of moderate outliers (around 6 or 7 points) grouped together slightly above the upper fence of the boxplot. These outliers can be seen on the scatterplot by looking at the data spread vertically; we can see the 3 extreme outliers at the spread out far away from where most of the data points are clustered together and we can see the moderate outliers slightly above where all the data point are clustered (between 500 and 1000 on y-axis).

There are also a few high leverage points we can see on the scatterplot by looking at the data spread horizontally. These high leverage points are located on the far-right of the x-axis as we can see them stand out from the rest of the data points, the vast majority of which are grouped together on the far-left side of the plot. Although, a few high leverage points are present, these points do not appear to be influential points – at first glance – that would influence the slope of the regression should we redraw the scatterplot (with a regression line) and exclude these high leverage points. In other words, the regression line would not significantly change if we excluded these few high leverage points from our data.

1.b

Use statistical software to estimate the model below and report your results.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

```
##Linear regression model
sals_lm <- lm(
  Salary_raw~
```

```
    ROG_raw,
  exec_sals)
summary.lm(sals_lm)
```

```
##
## Call:
## lm(formula = Salary_raw ~ ROG_raw, data = exec_sals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.03 -105.20  -50.79   27.76 2699.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  192.638     42.648   4.517 1.05e-05 ***
## ROG_raw        3.700      2.225   1.663   0.0978 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.3 on 207 degrees of freedom
## Multiple R-squared:  0.01319,    Adjusted R-squared:  0.008421
## F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

```
##Regression coefficients
sals_lm_coeffs <- round(
  sals_lm[["coefficients"]],3)
sals_lm_coeffs
```

```
## (Intercept)     ROG_raw
##     192.638       3.700
```

The linear regression model is:

$$y_i = 192.638 + 3.7x_i + \varepsilon_i$$

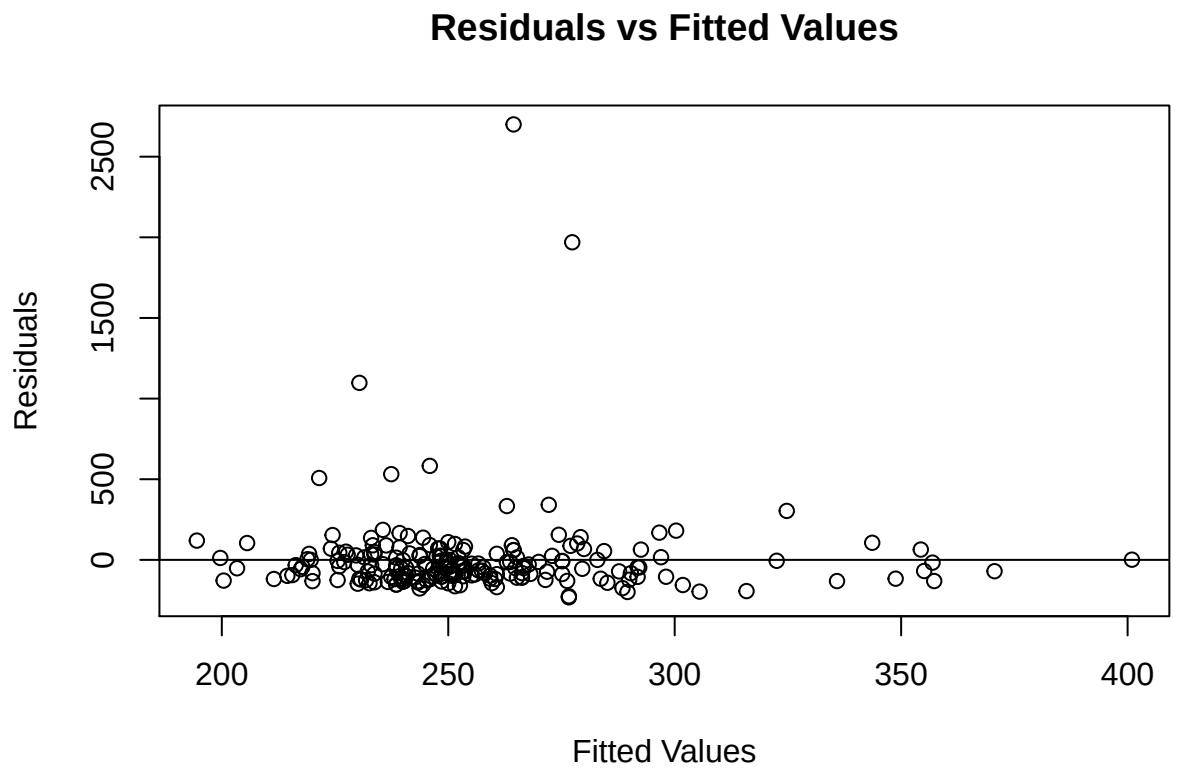
The results of the estimated model tell us that the intercept of the equation (β_0) is 192.638. This means that executives whose business portfolios experience no growth (0% rate of growth) in a given year are expected on average to have an annual compensation of $192,638 for that year. The results of the model also tell us that the slope of the equation (β_1) is 3.7. This means that for every percentage increase in x_i, y_i is expected to increase on average by that slope value (which is 3.7 in this case or $3,700 in the business context). We can also evaluate the standard deviation of the residuals (S) which is given in the model summary output from Minitab as 273.311. The standard deviation of residuals is rather large which means that the datapoints are fairly spread away from the regression line indicating that the model may not be that well fitted to the data (i.e., the model is not that good and seems to have weak predictive power because the datapoints are located too distantly from the regression line).

1.c

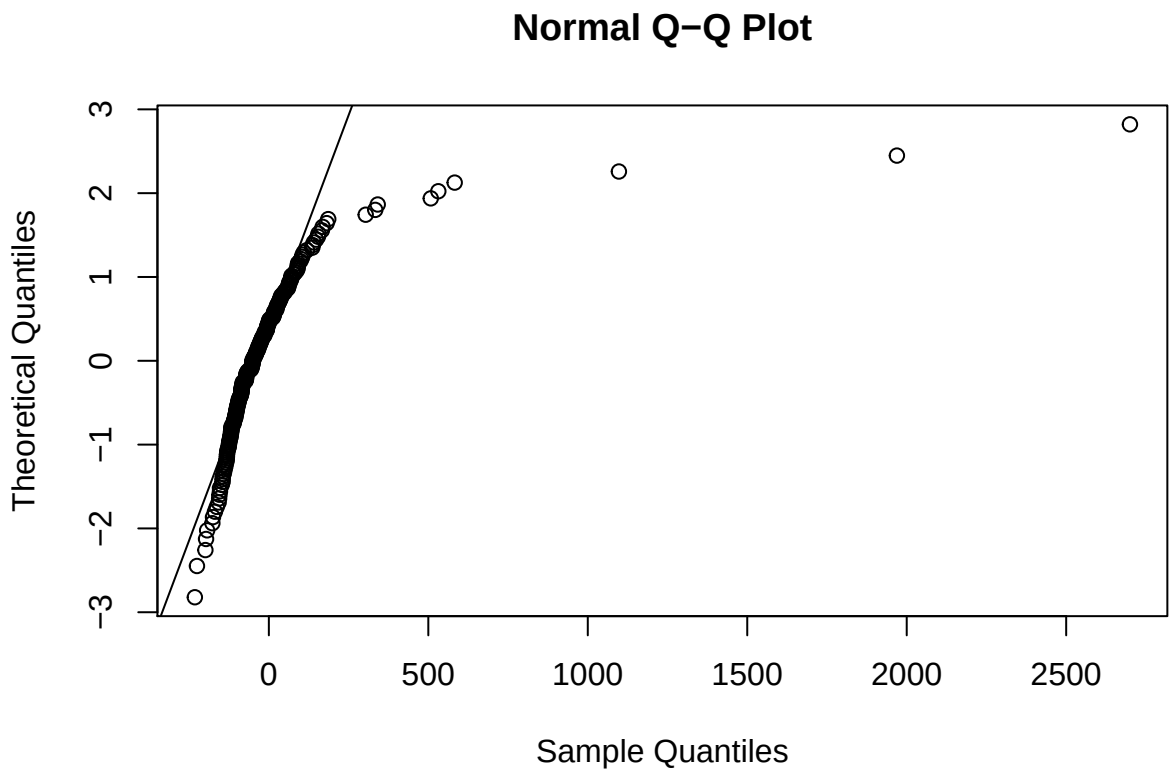
Create the corresponding residual plot (residuals against predicted values) and normal probability plot of the residuals. Looking at these two plots, does the estimated model appear satisfactory?

```
##Residuals vs fitted values plot
plot(
  fitted(sals_lm),
  resid(sals_lm),
  main = "Residuals vs Fitted Values",
  xlab = "Fitted Values",
```

```
  ylab = "Residuals")
##Adding mean line
abline(0,0)
```




```
##Q-Q plot
qqnorm(sals_lm$residuals,
       datax = TRUE) #Data
qqline(sals_lm$residuals,
       datax = TRUE) #QQ line
```



No, the estimated model does not appear satisfactory because 2 out of the 3 conditions required for linear regression (nearly normal residuals, and constant variability) are not satisfied. By looking at the residuals fit graph, we can clearly see that the data points are not equally spread across the x-axis. The data points on the left seem to be heavily clustered (densely populated/grouped together) while the further right we go on the residuals fit plot, we can see that the spread of the data points is much scarcer than those on the left; the variance of the data is heteroscedastic. Thus, the constant variability condition fails to be satisfied.

By looking at the normal probability plot of the residuals, we can clearly see that the residuals are extremely skewed as they do not line up at all with the red normal line. This is to be expected because we have found some outliers (unusual observations) in the scatterplot from part (a) that don't follow the same trend as the rest of the data. As a result, the data cannot be considered normally distributed, and thus the nearly normal residuals condition fails to be satisfied.

Because 2 of the 3 conditions required for a linear regression model fail to be satisfied, the model estimated in part (b) is not satisfactory.

1.d

> Use statistical software to estimate a new model, this time by using the adjusted variables *Salary_adj* and *ROG_adj* which exclude all the data points for which the salary of the senior executive appears extraordinarily large considering the ROG of their business portfolio. Report your results.

```
##Linear regression model
sals_adj_lm <- lm(
  Salary_adj~
    ROG_adj,
  exec_sals)
```

```r
summary.lm(sals_adj_lm)
```

```
##
## Call:
## lm(formula = Salary_adj ~ ROG_adj, data = exec_sals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -186.93  -65.17  -10.81   61.49  230.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 160.8777    13.9894  11.500  < 2e-16 ***
## ROG_adj       3.1125     0.7308   4.259 3.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.06 on 198 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.08393,    Adjusted R-squared:  0.0793
## F-statistic: 18.14 on 1 and 198 DF,  p-value: 3.169e-05
```

```r
##Regression coefficients
sals_adj_lm_coeffs <- round(
  sals_adj_lm[["coefficients"]],3)
sals_adj_lm_coeffs
```

```
## (Intercept)     ROG_adj
##     160.878       3.112
```

The new linear regression equation is:

$$y_i = 160.878 + 3.112x_i + \varepsilon_i$$

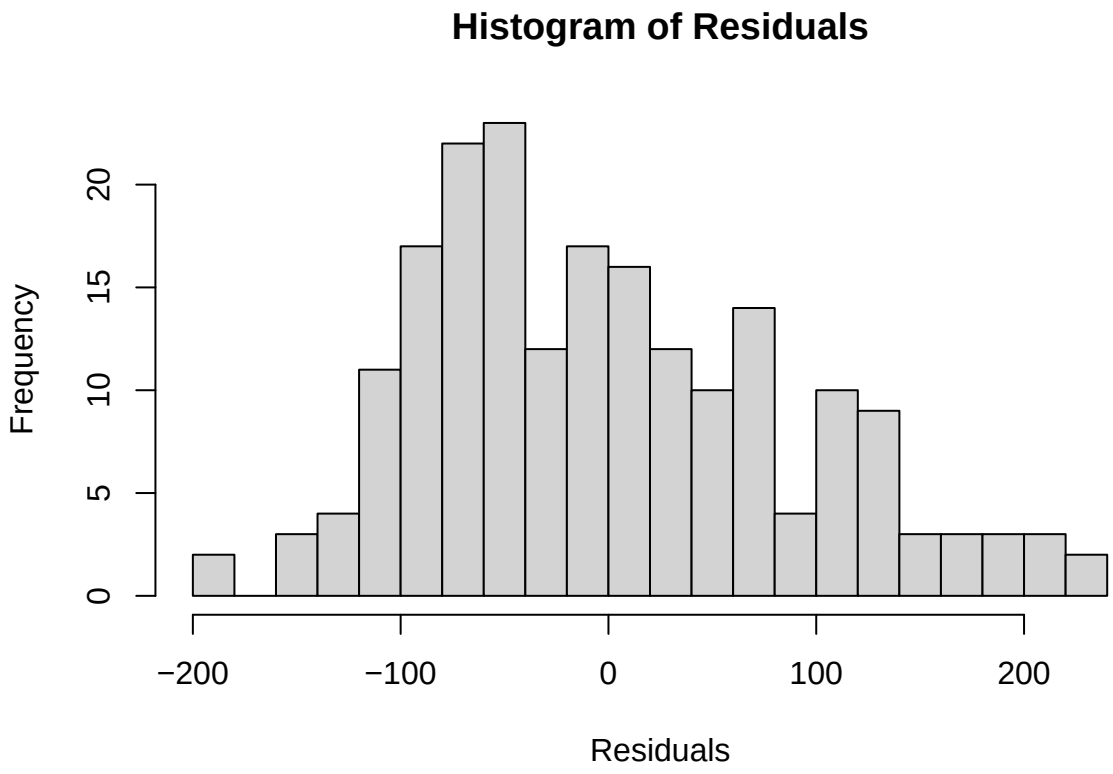
The newly calculated linear regression equation using the adjusted variables is quite different than the non-adjusted linear regression equation with both the intercept and the slope coefficients (β_0 and β_1) changing. The new intercept coefficient is 160.878 meaning that executives whose business portfolios experience no growth (0% rate of growth) in a given year are expected on average to have an annual compensation of \$160,878 for that year, instead of the previous \$192,638. The slope coefficient also changed to 3.112 meaning that for every point increase in x_i, y_i is expected to increase on average by that slope value which is now 3.112 (\$3,112)instead of the previous 3.7 (\$3,700). Furthermore, the standard deviation of the residuals which was given in the output in the model summary is 88.0583, which is significantly less than the 273.311 of the non-adjusted data. Since the standard deviation of residuals helps measure the distance of the datapoints from the regression line in the model (lower is better because it means less variability), we can compare it to the previous model; it tells us that the new model's datapoints are significantly less spread away from the regression line making it a better model than the original model because it is better fitted. As a result of that, the new model demonstrates much stronger predictive power than the first one.

1.e

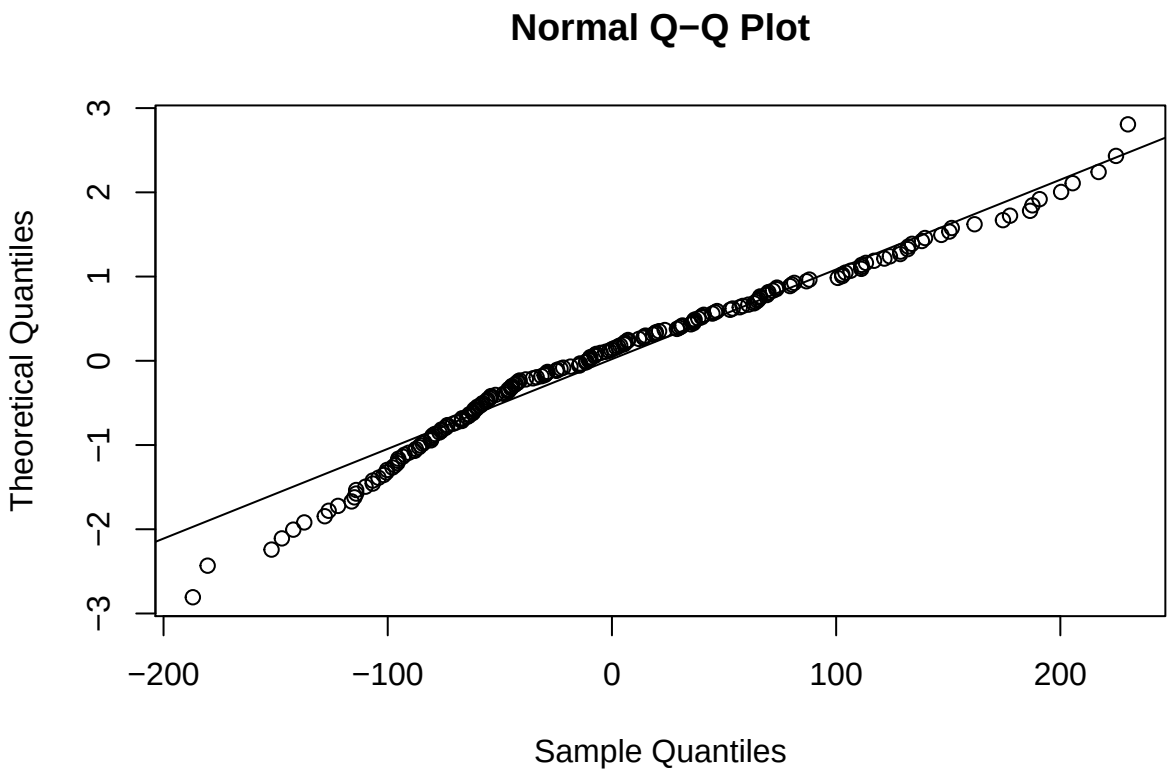
Produce a histogram and a normal probability plot of the residuals of your regression model in part d) above. Does this regression appear to meet the condition of near normality?

```r
##Histogram of residuals
hist(sals_adj_lm$residuals,
     breaks = 15,
```

```
     xlim = c(-200,250),
     main = "Histogram of Residuals",
     xlab = "Residuals")
```




```
##Q-Q plot
qqnorm(sals_adj_lm$residuals,
       datax = TRUE) #Data
qqline(sals_adj_lm$residuals,
       datax = TRUE) #QQ line
```



Although there is still some skewness present in the normal probability plot of residuals, it is significantly less than the non-adjusted data. The skewness is now moderate so even though there is still skewness present, since it is moderate, then we can say that this adjusted regression model appears to the satisfy the condition of near normality.

By looking at the histogram of residuals, we can see that the histogram bars are not perfectly symmetrically, thus also indicating that some skewness remains in the data, we can once again see that the skewness is moderate and that the bars of the histogram appear to be relatively symmetrical. Thus, we can also confirm this way that the adjusted regression model appears to satisfy the condition of near normality, just like the normal probability plot of residuals.

1.f

> What are the units of the slope coefficient b1 in this last regression equation? What is the average impact on the salary of senior executives whose ROG increases by 1% point?

The units of measurement of the slope coefficient (β_1) are in \$. For each 1% increase in the rate of growth (x_i) of the senior bank executives' business portfolios, we expect on average their annual compensation (salary) to increase by \$3,112.

1.g

> Calculate and interpret the value of the R^2 for this regression.

```
sals_adj_rsq <-round(summary(
  sals_adj_lm)$r.squared,4)
sals_adj_rsq
```

```
## [1] 0.0839
```

The correlation of determination R^2 helps analyze the strength of the fit of a linear model. Its value represents the percentage of the response variable's variation that is explained by the model (the predictor variable), thus the higher the value the better.

The value obtained from the calculation and the Minitab output shows that only 8.39% of the variation in the response variables can be explained by the predictor variable (the model), which means the model cannot account for the remaining 91.61% of the variation. This result is significantly better than the R^2 that was computed with the original data (1.32%) because it means the model accounts for far more variation than the original model.

1.h

Calculate a 95% confidence interval for the regression slope in part f).

```
sals_adj_lm_ci_cv <- round(qt(
  0.05/2,
  sals_adj_lm$df.residual,
  lower.tail = FALSE),3)

##Coefficient CI
sals_adj_lm_coeff__ci <- confint(
  sals_adj_lm,level = 0.95)[2,]
sals_adj_lm_coeff__ci
```

```
##    2.5 %   97.5 %
## 1.671366 4.553552
```

The 95% confidence interval for the regression slope β_1 is (1.671, 4.554). If we reran this model several times with different samples, we expect 95% of those samples' individual confidence intervals to contain the true slope of the model.

1.i

Use your results to calculate a 95% interval to estimate the mean salary of senior executives with an ROG of 20 per cent.

```
##CI given ROG of 20
exec_sals_mean_given_20 <- predict(
  sals_adj_lm,
  data.frame(ROG_adj = 20),
  interval = "confidence")
exec_sals_mean_given_20
```

```
##        fit      lwr     upr
## 1 223.1269 210.1757 236.078
```

The 95% confidence interval for the mean salary of senior executives with an ROG of 20% is (210.176, 236.078). This means that we are 95% certain that the population mean salary of senior executives with an ROG of 20% will fall between $210,176 and $236,078; 19 times out of 20.

2. Wages of University Professors

Your task will be to determine a regression model for estimating the impact of education, experience, and tenure on wages of university professors in Canada. The dataset **Professor Wages** provides information on 494 professors. It includes the following variables:

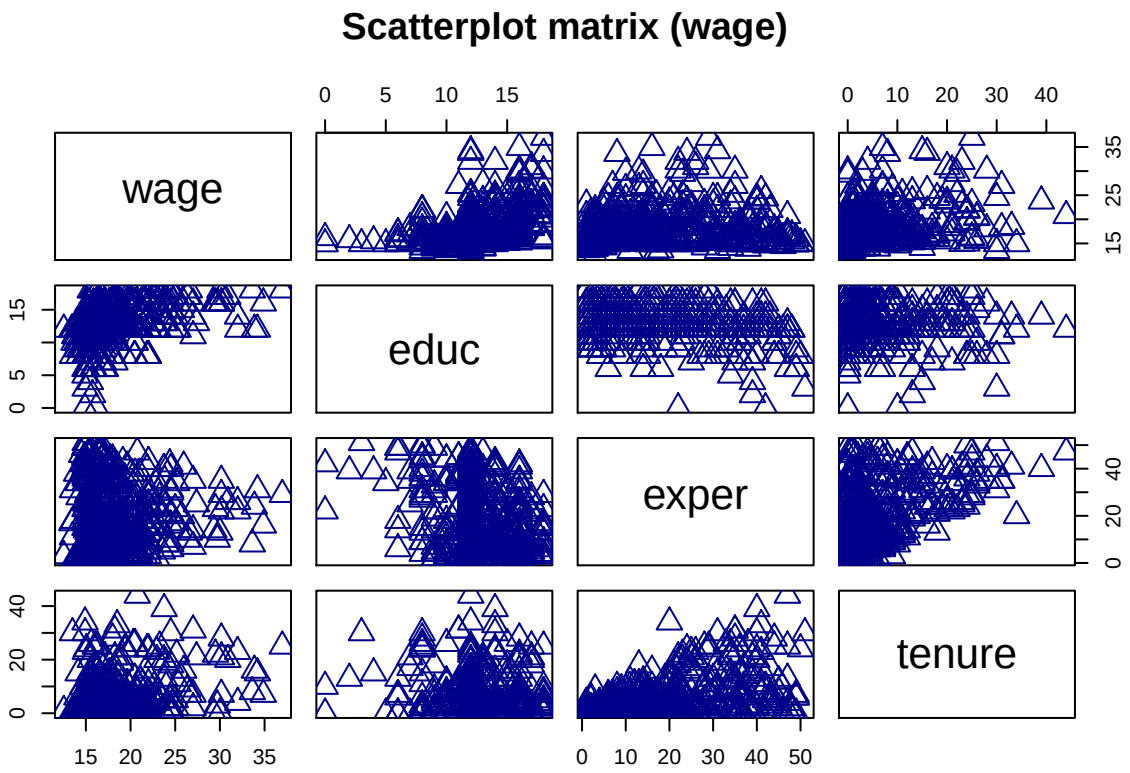
- *wage*: annual salary in CAD $10,000
- *educ*: number of years of training and education
- *exper*: number of years teaching and doing research at university level
- *tenure*: number of years in tenure

Consider a 5% significance level where needed.

2.a

Using a graphical display, explain why a log base 10 transformation of *wage* is necessary before fitting a linear regression model.

```
##Scatterplot matrix
plot(prof_wages,
     cex = 2,
     pch =2,
     col = "dark blue",
     main = "Scatterplot matrix (wage)")
```

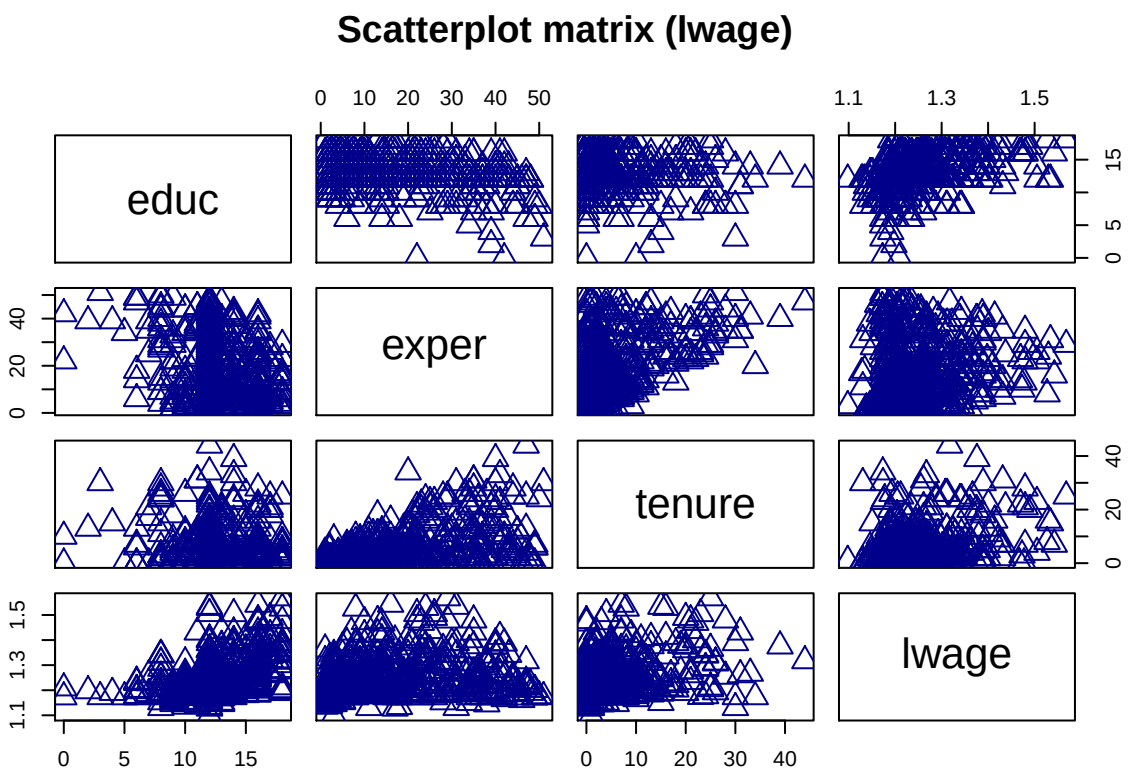


The true relationship of the X and Y variables in the data is not always linear. In case that true relationship between the response and predictors is non-linear then a log base 10 transformation will help linearize the data which would then allow us to perform a linear regression on the data. If we keep the data as it is, we won't be able to do a proper linear regression analysis since the results would be inaccurate. In this case, the relationships of the Xs and the Y appear to be linear for (*exper*, *wage*) and (*tenure*, *wage*) but non-linear for (*educ*, *wage*). A log base 10 transformation would help improve the first 2 pairs' linear relationship and make the third pair linearly related in order to perform a linear regression analysis.

2.b

Create a new variable called *lwage* as the log base 10 transformation of the variable *wage*. Using scatter plots, examine the pairwise relationship between the variables *lwage*, *educ*, *exper* and *tenure*. Comment on your results.

```
##Transforming data
prof_wages$lwage <- log(prof_wages$wage,10)
##Removing wage column
prof_wages$wage <- NULL
##Scatterplot matrix
plot(prof_wages,
     cex = 2,
     pch =2,
     col = "dark blue",
     main = "Scatterplot matrix (lwage)")
```




```
##Correlation matrix
prof_wages_corr_mat <-
  round(cor(prof_wages),4)
prof_wages_corr_mat
```

```
##           educ   exper  tenure  lwage
## educ    1.0000 -0.2860 -0.0747 0.4219
## exper  -0.2860  1.0000  0.5232 0.1222
## tenure -0.0747  0.5232  1.0000 0.3356
## lwage   0.4219  0.1222  0.3356 1.0000
```

We can look at the scattering of the data as well as the red regression lines on the scatterplots. If the regression line is moving upwards then it's a positive relationship; if it's pointing downwards, then it's a negative relationship. We can also roughly assess the strength of the relationships by looking at the angle of the regression lines; the flatter the line is the weaker the relationship, the more diagonal (angled) the line is, the stronger the relationship between the two variables is. Finally, the correlation coefficient of those lines can also give us information about the strength and direction of the variables' relationships.

The *educ* predictor variable does not have a linear relationship with any of the other 3 variables. There are also no extremely strong relationships that stand out but there are 2 scatterplots whose variables have moderate relationship (*exper*, *tenure*) and (*tenure*, *exper*). Just under half of the relationships are rather negative (5 to be exact) while the remaining 6 relationships are positive. One third of the relationships have very weak relationships (4 to be exact), half of them have weak relations (6 out of 12), and only two have moderate relationships (as mentioned previously).

2.c

> Perform a multiple linear regression analysis with *lwage* as the response variable and *educ*, *exper* and *tenure* as predictors. Report your results. Explain why you might consider dropping variable *exper*. Also, do you observe any sign of multicollinearity? Explain.

```
##Regression model
prof_wages_lm <-
  lm(lwage~educ+exper+tenure, prof_wages)
summary.lm(prof_wages_lm)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + tenure, data = prof_wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16905 -0.04152 -0.01320  0.03514  0.26657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.0500934  0.0161066  65.196  < 2e-16 ***
## educ        0.0136086  0.0011293  12.051  < 2e-16 ***
## exper       0.0005086  0.0002685   1.894   0.0588 .
## tenure      0.0034988  0.0004724   7.406 5.71e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06523 on 490 degrees of freedom
## Multiple R-squared:  0.3185, Adjusted R-squared:  0.3144
## F-statistic: 76.35 on 3 and 490 DF,  p-value: < 2.2e-16
```

```
##Regression coefficients
prof_wage_lm_coeffs <-
  summary(prof_wages_lm)$coefficients[,1]
prof_wage_lm_coeffs
```

```
##  (Intercept)         educ        exper       tenure
## 1.0500934125 0.0136085882 0.0005085951 0.0034987865
```

The regression equation is:

$$lwage = 1.0501 + 0.01361(educ) + 0.000509(exper) + 0.003499(tenure) + \varepsilon_i$$

From the regression equation, we find that professors who have 0 years of training and education, 0 years of teaching and doing university-level research, and 0 years in tenure are expected on average to earn a log wage of 1.0501 ($112,228). However, this doesn't tell us much because it's unlikely for a professor to have 0 years of training and education, teaching and research, and tenure.

For the slope coefficient of educ; with all else held constant (meaning the other predictors don't fluctuate), for each year of training and education, we would expect the log of the professors' salaries to increase on average by 0.01361 ($10,318). For the slope coefficient of exper; with all else held constant, for each year of teaching and doing research at university level, we would expect the log of the professors' salaries to increase on average by 0.000509 ($10,012). For the slope coefficient of tenure; with all else held constant, for each year in tenure, we would expect the log of the professors' salaries to increase on average by 0.003499 ($10,081).

We can also gauge the strength and predictive power of the model from the Minitab output. S represents the standard deviation of the residuals ε_i and it tells us how far the datapoints of the model are spread out (distanced) from the regression line. A smaller S value means that the model is better fitted which gives it better predictive power over a model with a larger S. In this case, S has a value of 0.0652306 meaning that the datapoints are spread around the regression line within that range. R^2 represents the percentage of the variance in the response variable that can be explained by the predictor variables (the model). The higher the better because it would mean our model accounts for more of the response variable's variance thus making our model better fitted with more predictive power. In this case, R^2 is 31.85% which means that 68.15% of the variance cannot be accounted for by the model. This isn't necessarily bad depending on the context of the regression analysis.

When performing a multiple regression analysis, sometimes certain predictor variables will not be contributing much to the predictive power to the model (i.e., they have very small coefficients). When these non-contributing predictor variables are kept in the model, there's a chance it may reduce the accuracy of the model's predictions. Thus, by removing these variables, the prediction power of the model is likely to increase (which is good) making it more parsimonious. In this case, we can see that the *exper* prediction variable does not contribute much to the model's overall output (weak coefficient compared to the other two predictors). For that reason, we could consider dropping it from the model to improve the model's strength.

There are no signs of multicollinearity in this model because the variance inflation factors (VIFs) of the model are not greater than 10, which means that none of the predictors in our model are highly correlated. As shown in the output below:

```
##Calculating model VIF
prof_wages_lm_vif <-
  round(vif(prof_wages_lm),3)
prof_wages_lm_vif
```

```
##   educ  exper tenure
##  1.098  1.504  1.389
```

2.d

> Repeat part c) but now with *lwage* as the response variable and only two independent variables, *educ* and *tenure*. Report your results.

```
##New regression model
prof_wages_lm_new <-
  lm(lwage~educ+tenure, prof_wages)
summary.lm(prof_wages_lm_new)
```

```
##
## Call:
## lm(formula = lwage ~ educ + tenure, data = prof_wages)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.18142 -0.04242 -0.01438  0.03712  0.26932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.0641137  0.0143426  74.192   <2e-16 ***
## educ        0.0129871  0.0010834  11.988   <2e-16 ***
## tenure      0.0039687  0.0004031   9.846   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0654 on 491 degrees of freedom
## Multiple R-squared:  0.3135, Adjusted R-squared:  0.3107
## F-statistic: 112.1 on 2 and 491 DF,  p-value: < 2.2e-16
```

```
##Regression coefficients
prof_wages_lm_new_coeffs <-
  summary(prof_wages_lm_new)$coefficients[,1]
prof_wages_lm_new_coeffs
```

```
## (Intercept)        educ      tenure
## 1.064113681 0.012987057 0.003968683
```

The new regression equation is:

$$lwage = 1.0641 + 0.01299(educ) + 0.003969(tenure) + \varepsilon_i$$

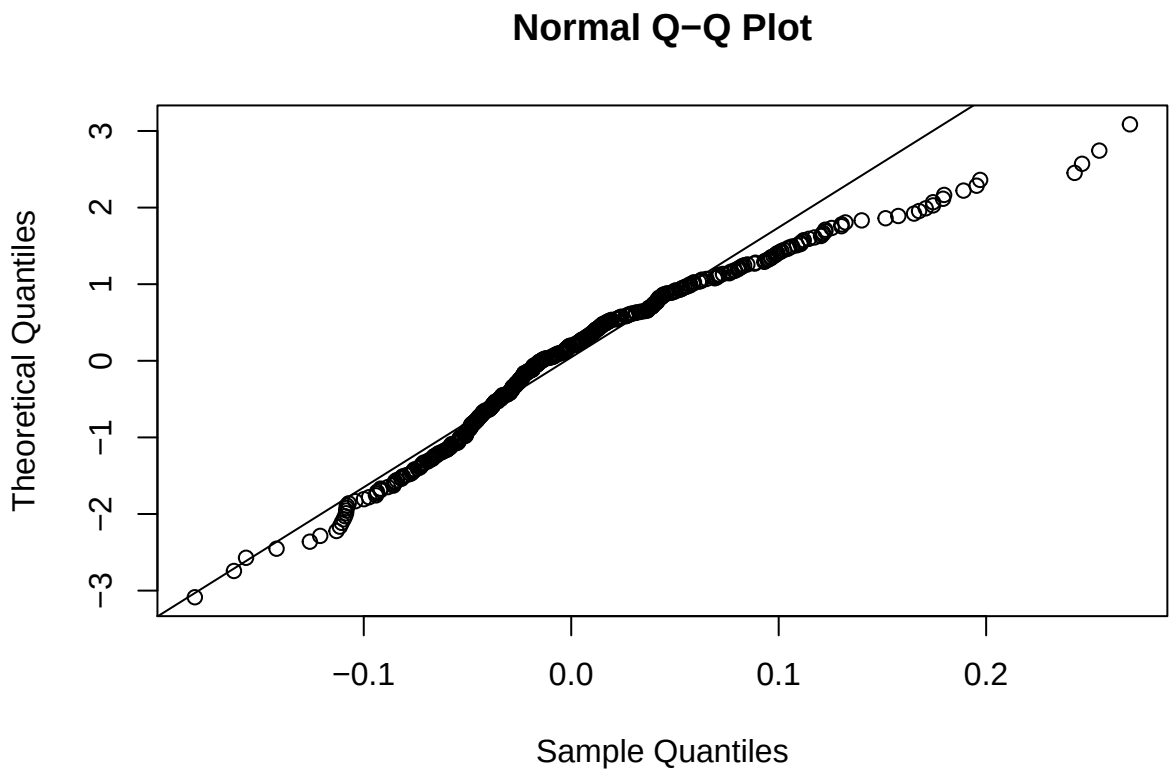
The main changes that we can notice by removing the exper predictor variable is the change in the remaining 2 slope coefficients and intercept coefficient. The *educ* coefficient decreased from 0.01361 ($10,318) to 0.01299 ($10,304) while the *tenure* coefficient increased from 0.003499($10,081) to 0.003969 ($10,092). The intercept coefficient increased from 1.0501 ($112,228)to 1.0641 ($115,904). The model's strength does not seem to have increased by much as the S and R^2 values remain relatively the same as the original model, 0.0654023 vs 0.0652306, and 31.35% vs 31.85%, respectively. Although the R^2 value decreased by 0.5% in the new model, the change is extremely small and is thus unlikely to have any significant impact on the model.

With the new model, for the slope of *educ*; all else held constant, we expect on average for the log of professors' salaries to increase by 0.01299 ($10,304) for every year of training and education they have compared to the previous 0.01361 ($10,318). For the slope of *tenure*; all else held constant, we expect on average for the log of professors' salaries to increase by 0.003969 ($10,092) for every year in tenure compared to the previous 0.003499 ($10,081). The new intercept coefficient 1.0641 also tells us that professors with 0 years of training and education, and 0 years in tenure can expect on average to earn $115,904.

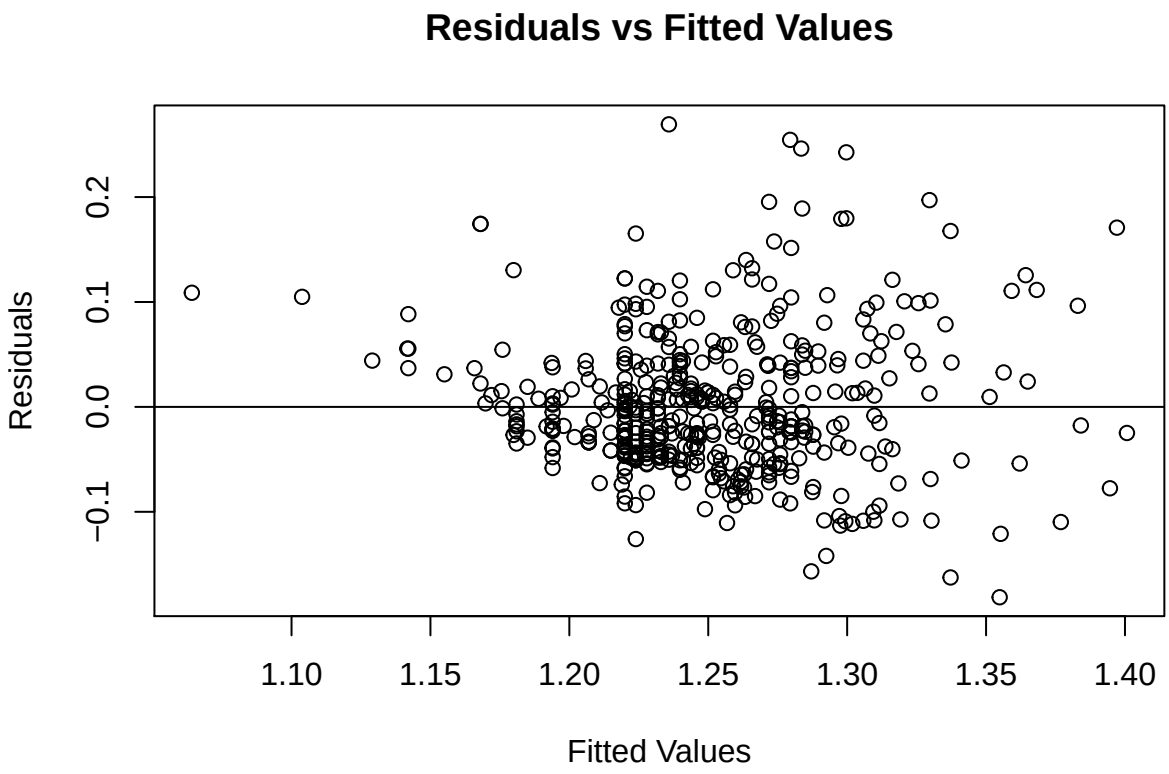
2.e

> Create the residual against fitted values plot and the normal probability plot of the residuals corresponding to the model in part d). Do the equal variance and near normality conditions appear to be satisfied? Explain

```
##Q-Q plot
qqnorm(prof_wages_lm_new$residuals,
       datax = TRUE)
qqline(prof_wages_lm_new$residuals,
       datax = TRUE)
```




```
##Residuals vs fitted values
plot(
  fitted(prof_wages_lm_new),
  resid(prof_wages_lm_new),
  main = "Residuals vs Fitted Values",
  xlab = "Fitted Values",
  ylab = "Residuals")
##Adding mean line
abline(0,0)
```



From the normal probability plot of residuals, we can see that although the datapoints' alignment with the normal diagonal line is not perfect, it's still relatively aligned to the extent that we can say that the near normality condition is satisfied. We can also visualize the near normality condition in the residuals fit plot; on it, we can see that the datapoints are rather symmetrically spread around 0 vertically (the mean), even though there are some noticeable outliers, they're very few and don't significantly impact the near normality condition, thus it's still satisfied.

As for the equal variance condition, it does not appear to be satisfied. By looking at the residuals fit plot, we can see that the datapoints are not equally distributed across the fitted values (horizontal/x-axis). The data start off vary sparse (outliers) and around the middle-fitted values, the data are densely clustered together, and then become less grouped on the far right side of the fitted values. Thus, the equal variance condition is not satisfied because the data are not equally distributed on the residuals versus fitted values plot.

2.f

> Based on the regression report for part d), test the significance of the predictors *educ* and *tenure* separately and the significance of the model as a whole. Are your conclusions consistent? Explain.

Model Hypothesis Test

$$\begin{cases} H_0 : \beta_{educ} = \beta_{tenure} = 0 \\ H_A : \text{Not all } \beta's \text{ are zero} \end{cases}$$

```
##Removing exper variable
prof_wages$exper <- NULL
##Complete model
prof_wages_lm_whole <-
```

```
  lm(lwage~1, prof_wages)
summary.lm(prof_wages_lm_whole)
```

```
##
## Call:
## lm(formula = lwage ~ 1, data = prof_wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15161 -0.05923 -0.01911  0.03969  0.31841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.249561   0.003544   352.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07878 on 493 degrees of freedom
```

```
##Test statistic - Model
prof_wages_lm_stat <-
  summary(prof_wages_lm_new)$fstatistic[1]
prof_wages_lm_stat
```

```
##    value
## 112.1341
```

```
##Critical value
prof_wages_lm_crit <-
  qf(0.05,
     summary(prof_wages_lm_new)$fstatistic[2],
     summary(prof_wages_lm_new)$fstatistic[3],
     lower.tail = FALSE)
prof_wages_lm_crit
```

```
## [1] 3.014085
```

```
##Validating test
prof_wages_lm_stat > prof_wages_lm_crit
```

```
## value
##  TRUE
```

Since $F_{stat} > F_{0.05;2,491} \rightarrow 112.13 > 3.014$, we reject the null hypothesis in favour of the alternative. There's sufficient evidence that the model as a whole is significant, meaning that in this case, the model has some predictive power to be able to predict on average the wages of professors based on years of education and in tenure.

Coefficient Test - Education

$$\begin{cases} H_0 : \beta_{educ} = 0, & \text{when all the other variables are included in the model} \\ H_A : \beta_{educ} \neq 0, & \text{when all the other variables are included in the model} \end{cases}$$

```
##Test statistic - educ
prof_wages_lm_educ <- round(summary(
  prof_wages_lm_new)$coefficients[2,3],3)
prof_wages_lm_educ
```

```
## [1] 11.988
```

```
##Crit value for coefficients
prof_wages_lm_coeff_crit <- round(qt(
  0.05/2,
  summary(prof_wages_lm_new)$df[2],
  lower.tail = FALSE),3)
prof_wages_lm_coeff_crit
```

```
## [1] 1.965
```

```
##Validating test
prof_wages_lm_educ > prof_wages_lm_coeff_crit
```

```
## [1] TRUE
```

Since $t_{stat} > t_{\alpha,n-k-1} \rightarrow 11.99 > 1.965$, we reject the null hypothesis in favour of the alternative. There is sufficient evidence that the education predictor adds significance to the model that has other predictors. All else being equal, professors are expected on average to earn a log wage of 0.01299 ($10,304) for every year of training and education they have.

Coefficient Test - Education

$$\begin{cases} H_0 : \beta_{tenure} = 0, & \text{when all the other variables are included in the model} \\ H_A : \beta_{tenure} \neq 0, & \text{when all the other variables are included in the model} \end{cases}$$

```
##Test statistic - tenure
prof_wages_lm_tenure <- round(summary(
  prof_wages_lm_new)$coefficients[3,3],3)
prof_wages_lm_tenure
```

```
## [1] 9.846
```

```
##Validating test
prof_wages_lm_tenure > prof_wages_lm_coeff_crit
```

```
## [1] TRUE
```

Since $t_{stat} > t_{\alpha,n-k-1} \rightarrow 9.85 > 1.965$, we reject the null hypothesis in favour of the alternative. There is sufficient evidence that the tenure predictor adds significance to the model that has other predictors. All else being equal, professors are expected on average to earn a log wage of 0.003969 ($10,092) for every year in tenure.

Comments

The results of the 3 hypotheses tests are consistent with each other. If our model as a whole was not significant then the 2 coefficients of the model would not be significant individually, thus there would have been no need to run the latter two hypotheses' tests in the first place. But because we found the model to be significant as whole, we ran the coefficients' individual hypotheses test and found that they are also significant with all else being equal, which is consistent with first hypothesis test and confirms that the model as a whole is significant. For the business significance, this means that our predictors (education and tenure) are significant (are capable of) in predicting the wage of professors.

2.g

> Based on the results for the model in part d), what is the estimated impact on wages of an additional year of training and education? Express your answer in the right salary units.

$log_{10}(x) = b \rightarrow x = 10^b 10^b = 10^{0.01299} = 1.03036239521.0303623952 \cdot 10,000 = 10,303.62395197554 \approx 10,304$

An additional year of training and education is estimated to impact professors' annual wages by $10,304 CAD, all else being equal.

2.h

> Obtain the standard error of the fit (SE fit) using statistical software and then manually compute the prediction interval and the confidence interval for the mean wage of professors with 13 years of training and education and 15 years in tenure. Express your answers in the right salary units. Why is the prediction interval wider than the confidence interval? Explain.

```
##SE fit
prof_wages_lm_se_fit <- predict(
  prof_wages_lm_new,
  data.frame(educ=13,
             tenure=15),
  se.fit = TRUE)
prof_wages_lm_se_fit
```

```
## $fit
##        1
## 1.292476
##
## $se.fit
## [1] 0.004929265
##
## $df
## [1] 491
##
## $residual.scale
## [1] 0.06540229
```

```
##Confidence interval
prof_wages_lm_ci <- round(predict(
  prof_wages_lm_new,
  data.frame(educ=13,
             tenure=15),
  interval = "confidence"), 4)
prof_wages_lm_ci
```

```
##      fit    lwr    upr
## 1 1.2925 1.2828 1.3022
```

```
##Prediction interval
prof_wages_lm_pi <- round(predict(
  prof_wages_lm_new,
  data.frame(educ=13,
             tenure=15),
  interval = "prediction"), 4)
prof_wages_lm_pi
```

```
##      fit    lwr    upr
## 1 1.2925 1.1636 1.4213
```

The confidence interval is based on the expected value of the annual wage of professors given 13 years of training and education, and 15 years in tenure so it only needs to factor in the uncertainty in estimating the mean wage because the mean is a fixed target and thus does not vary or fluctuate.

On the other hand, the prediction interval is predicting the future wage of professors given 13 years of training and education, and 15 years in tenure so it needs to account for increased variability and the possibility of the response variable (annual wage) fluctuating in the future. The response variable would then be considered a moving target, thus the P.I. is wider because it needs to account for that future variability of the response variable changing (moving) from its mean value. Mathematically, this is represented by adding *MSE* to the SE of fit, as well squaring the latter, which is what contributes to the wider interval and accounts for the previously mentioned fluctuations of the response variable.