STAT2008 Assignment 2

u7205329

October 20th, 2020

Part a

Given that Y is the log(wage) and the numeric variables as predictors are years of `education` (X_1), years of `experience` (X_2), and `age` (X_3), we have the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

To test whether this model is significant, we perform an overall F test with the hypotheses:

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \beta_3 = 0 \\ H_a &: \quad \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned}$$

```
Analysis of Variance Table

Response: log(wage)
            Df  Sum Sq Mean Sq F value    Pr(>F)
education    1  21.481 21.4807 97.2864 < 2.2e-16 ***
experience   1   9.915  9.9154 44.9068 5.295e-11 ***
age          1   0.028  0.0277  0.1253    0.7235
Residuals  530 117.023  0.2208
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the ANOVA table as able, we can derive the test statistics:

$$F^* = \frac{MSR}{MSE} = \frac{(21.481 + 9.915 + 0.028)/3}{0.2208} = 47.43691.$$

For $\alpha = 0.05$, we require $F(0.95; 3, 350) = 2.6217$. Since $F^* = 47.44 > 2.6217$, we can reject H_0 in favour of the alternative and conclude that the model is significant. That is, at least one of the covariates are useful for us to predict the log of `wage`.

Part b

Below are the coefficients from the summary table of the MLR model in part (a)

```
               Estimate Std. Error    t value  Pr(>|t|)
(Intercept)  0.84480019  0.7188375  1.1752311 0.2404300
education    0.13805273  0.1179131  1.1708011 0.2422049
experience   0.05352942  0.1179613  0.4537878 0.6501673
age         -0.04172549  0.1178636 -0.3540151 0.7234683
```

and the estimated regression function

$$\hat{Y} = 0.84480 + 0.13805 X_1 + 0.05353 X_2 - 0.04173 X_3$$

We can interpret the estimated coefficients as follows:

- b_0: The mean log(wage) is estimated to be 0.8448 when years of education, years of experience and years of age are equal to zero, with a standard deviation of 0.7188.
- b_1: The mean log(wage) is estimated to increase by 0.13805 when the years of education increase by 1, holding age and experience constant, with a standard deviation of 0.1179.
- b_2: The mean log(wage) is estimated to increase by 0.05353 when the years of potential work experience increase by 1, holding age and education constant, with a standard deviation of 0.11796.
- b_3: The mean log(wage) is estimated to decrease by 0.04173 when age increase by 1, holding experience and education constant, with a standard deviation of 0.11786.

We also construct the 95% Bonferroni joint confidence interval for the slope parameters. As we want $1 - m\alpha = 0.95$, where m is the number of intervals (in our case, 3), we need to use the significance level $\alpha = 0.05/3$.

```
                0.833 %  99.167 %
(Intercept) -0.8815627 2.5711631
education   -0.1451277 0.4212332
experience  -0.2297670 0.3368258
age         -0.3247871 0.2413361
```

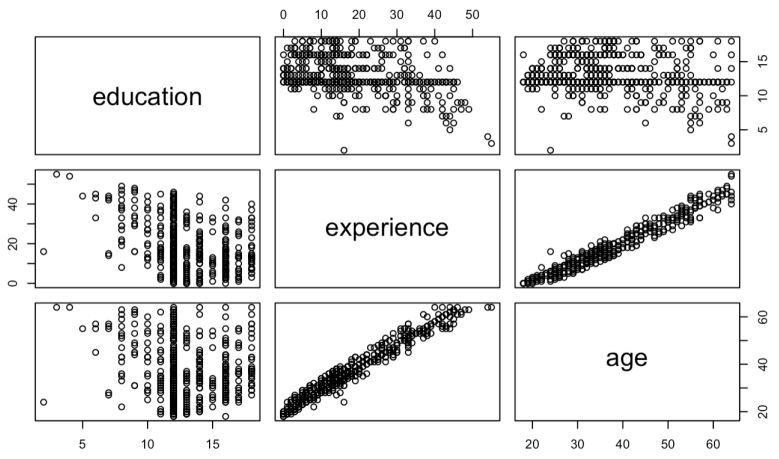
Thus, the three intervals above (except intercept) will jointly cover β_1, β_2, and β_3 simultaneously with a confidence interval of at least 0.9833.

Part c

For the summary output (as shown above), the t-values show that `education`, `experience`, and `age` are not significant additions to the model. They do not contradict the test result in part (a) because these t-tests are testing the marginal contribution of each variable, considering the other variables are already in the model. In part (a), however, we performed an overall F test to check whether there is a regression relation between log(wage) and the covariates as a whole.

A potential problem for insignificant marginal contribution is **multicolinearity** since age and experience or age and education can be greatly correlated. From the summary table, we do see that the estimated coefficient for `age` has a negative sign, which is unexpected. Based on the other coefficients, we saw that an increase in years of experience or education correspond to an increase in the estimated mean of log(wage). An increase in years of experience or education should correspond to an increase in age. By the transitive property, we expected a positive sign in b_4.

We investigate the pairwise correlation between the covariates by plotting the scatter plot matrix and the correlation matrix as follows:




```
                education experience        age
education   1.0000000 -0.3526764 -0.1500195
experience -0.3526764  1.0000000  0.9779612
age        -0.1500195  0.9779612  1.0000000
```

Figure 1: Left: Scatter plot matrix, Right: Correlation matrix

It can be seen that there exists a strong positive linear association between `experience` and `age` ($r^2_{23} = 0.978$). We further study the problem by examine the VIF values

```
education experience        age
 229.5738  5147.9190  4611.4008
```

The largest VIF value equals to 5147.9 indicates a severe multicolinearity problem. Interestingly, while $r^2_{12} = -0.35$ and $r^2_{13} = -0.15$ are not large, $VIF_{(education)} = 229.57$ shows that the combination of `experience` and `age` are also strongly correlated to `education`. One possible remedial measure is to drop one of the variables, e.g. `age` (since we can somewhat derive the age group of an individual from `education` and `experience`).

Part d

Since we now only include `education` and `experience` as potential predictors, we consider the new MLR model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where X_{i1} = years of education and X_{i2} = years of experience.

```
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 0.59416863 0.124428158  4.775194 2.325369e-06
education   0.09641369 0.008309695 11.602555 6.748580e-28
experience  0.01177396 0.001755530  6.706783 5.104845e-11
```

From the summary output, we have the response function

$$\hat{Y} = 0.5942 + 0.0964 X_1 + 0.0118 X_2$$

We are interested in how `log(wage)` and `experience` are related, given that `education` is already included in the model. If we were to hold X_1 constant and increase X_2 by one unit, on the original scale, the response variable `wage` will multiply by e^{b_2} or $e^{0.0118}$. We can visualise this relationship by generating the plot below

While a positive relationship is visible, we have shown that it is not linear. As we wish to conduct a test whether a second-order term is needed for `experience` given `education` is in the model, we are in fact considering the two following models and whether or now we can drop the second-order term.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X^2_{i2} + \epsilon_i \quad \text{Full model}$$

and

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \text{Reduced model}$$

Hypotheses:

$$H_0: \quad \beta_3 = 0$$
$$H_a: \quad \beta_3 \neq 0$$

```
                    Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)      0.5203217710 0.1236162526  4.209170 3.010737e-05
education        0.0897560821 0.0083205199 10.787317 1.160240e-24
experience       0.0349403392 0.0056492113  6.184994 1.242179e-09
I(experience^2) -0.0005362401 0.0001245024 -4.307068 1.971719e-05
```

It can be quickly be seen from the summary output that $|t^*| = 4.3071 > t(0.975, 530)$ and thus we can reject H_0 and conclude that the second-order of `experience` is a significant addition and to be retained in our model.

Part e

To investigate how marital status affect the wage, we consider the SLR model where log(wage) is the response and marital status the predictor. X is the indicator variable of marriage (1 if yes, 0 if no).

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 1.960443 0.03858151 50.813028 2.820855e-206
marriedyes  0.150657 0.04765581  3.161356  1.659834e-03
```

From the summary output, we have the regression function

$$\hat{Y} = 1.96044 + 0.1566X$$

To see whether married people earn more salary than unmarried, we perform an upper-tailed test on β_1 with the hypotheses:

$$H_0: \quad \beta_1 \leq 0$$
$$H_a: \quad \beta_1 > 0$$

Also from the summary output, we obtain the test statistics:

$$t^* = \frac{b_1}{s(b_1)} = \frac{0.15066}{0.04766} = 3.1611$$

For $\alpha = 0.05$, we require $t(0.95, n-2 = 532) = 1.6477$. Since $t^* > 1.6477$, we can reject H_0 in favour of the alternative and conclude that married people earn significantly more salary than unmarried people. Indeed, we can see the difference in the median of log(wage) when plotting the box plot as below

The 95% confidence interval of the slope coefficient is (0.0570, 0.2443). That is, from our model, we are 95% confident that a married individual's `log(wage)` is between 0.0570 to 0.2443 higher than that of an unmarried individual.

Part f

We now investigate the same relationship, but considering more covariates come into play. With the predictor variables as `education` (X_1), `experience` (X_2), and `marriage` (d_1), and including the second-order of `experience`, we have the new linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2}^2 + \beta_4 d_{i1} + \epsilon_i$$

The response function becomes for the two types of people based on their marital status. If married,

$$E\{Y\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2}^2 + \beta_4$$

If unmarried,

$$E\{Y\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2}^2$$

It appears that the one difference in the two models is β_4. To test whether married people earn more than unmarried, we perform an upper-tailed test on β_4 with the hypotheses:

$$\begin{aligned} H_0: &\quad \beta_4 \leq 0 \\ H_a: &\quad \beta_4 > 0 \end{aligned}$$

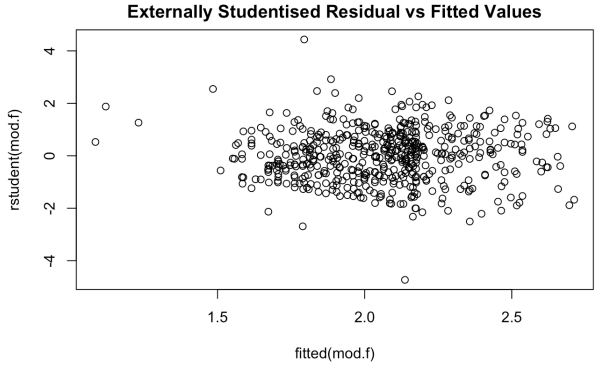
```
                    Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)      0.5091583328 0.1239001891  4.109423 4.596281e-05
education        0.0895022226 0.0083193610 10.758305 1.516662e-24
experience       0.0328343518 0.0059058497  5.559632 4.293931e-08
I(experience^2) -0.0005018299 0.0001276174 -3.932301 9.533369e-05
marriedyes       0.0546330090 0.0448906175  1.217025 2.241373e-01
```

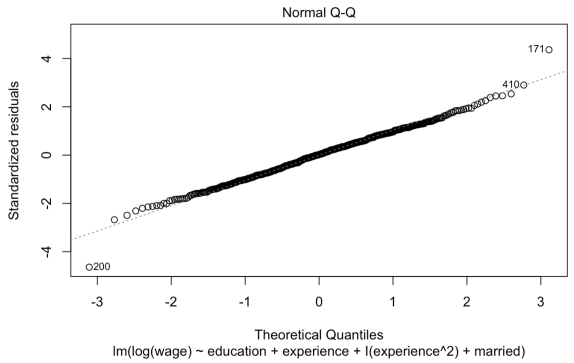
From the summary output, the test statistics can be derived:

$$t^* = \frac{b_4}{s(b_4)} = \frac{0.0546330}{0.0448906} = 1.217025.$$

For $\alpha = 0.05$, we require $t(0.95, 532) = 1.6477$ and since $1.2170 < 1.6477$, we do not reject H_0 and conclude that married people do not earn significantly higher than those unmarried given that `education` and `experience` are also considered. This conclusion is different from the result we had from part e. The reason of difference is that with this new model, the majority of changes in log(wage) have been explained by `education` and `experience`. After that, marital status appears not to have any major contribution in explaining the changes in the response function anymore.

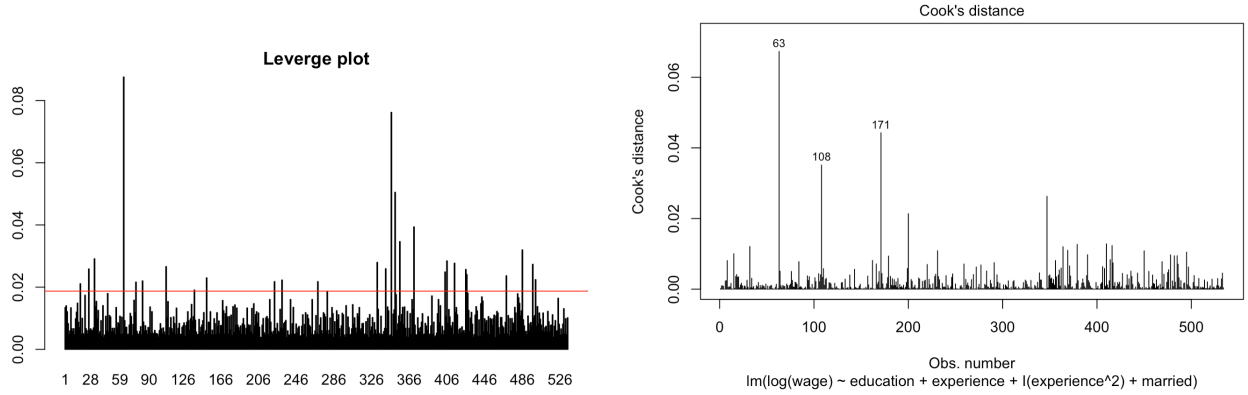
Part g





Externally studentized residual plot: Aside from a few data points with smaller fitted values, the data set seems to be randomly distributed with no discernible shape. Hence, linearity and constant variance assumption is satisfied. It appears that the majority of the data set varies within 2 residual values. However there are two observations that are 4 points away from 0 which can be identified as 171 (top) and 200 (bottom). They are potential outliers.

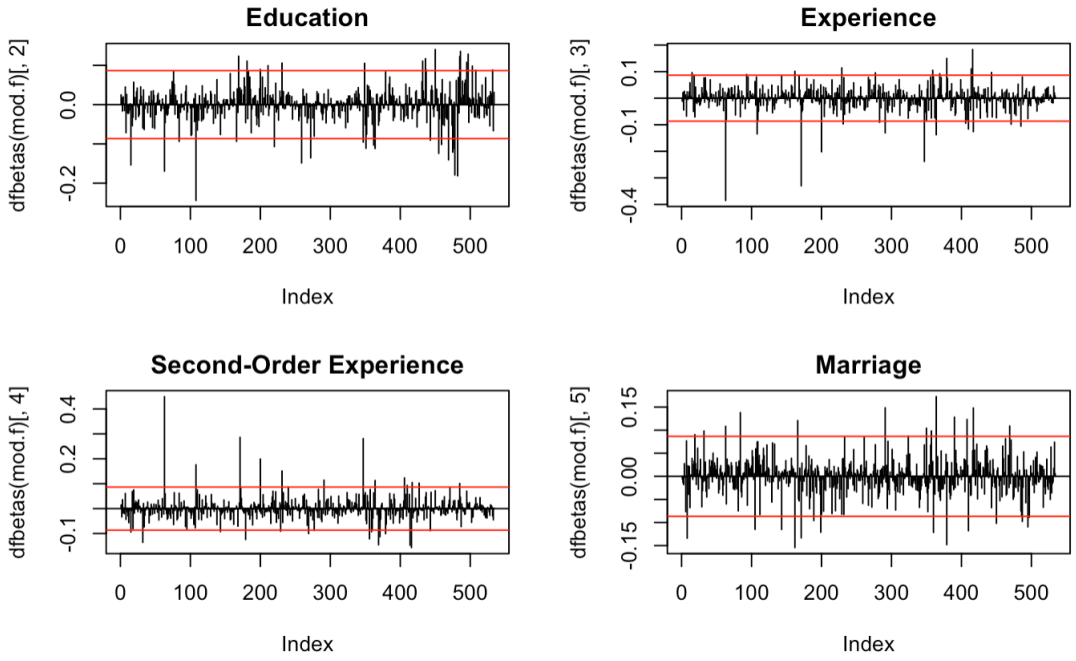
Normal QQ Plot: Overall, the data set seems to be normally distributed. However, once again, we see the observation 171 and 200 stray far away from the QQ line and the rest of the data set.



Leverage plot: There are many points exceed the threshold of $2p/n$. Distinctly, there are two observations with extreme values of leverage comparing to others, which can be identified as 63 and 347.

Cook's Distance: There are two points that have the largest influence among all observations which can be identified as 63 and 171.

Since there is no spatial or sequential order in the data set, the independence assumption is satisfied.



DFBETAS: Using the guideline of $2/\sqrt{n}$ for large data sets, the observation 108 can be considered as potentially influential for Education. Observation 63 and 171 both greatly exceed the guideline for Experience and the second-order of Experience. There are a number of observations that have greater influence on Marital Status but none distinctively stands out.

Since both influence measures (Cook's Distance and DFBETAS) identified the observations 63, 171 as influential and 63 as both an outlying X and Y observation, we decide to investigate the characteristics of these data points.

Firstly, **observation 171** is a female manager of age 21, with 14 years of education and 1 year of experience. Her wage at the point of data collection is 44.5 dollars per hour, which is the highest among all data points. **Observation 63** is a male worker of age 64, with 3 years of education and 55 years of experience. His wage is 7 dollars per hour which is lower than the average wage of the data set (9.024). Another data point of interest is **Observation 200**: a male manager of age 42, with 12 years of education and 24 years of experience. His wage is 1 dollar per hour which is the lowest among all data points.

Part h

Since `occupation` is a qualitative variable with 6 factors, we have 5 indicator variables. With log(wage) (Y) regressed on `experience` and `occupation`, we have the first-order linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 d_{i1} + \beta_3 d_{i2} + \beta_4 d_{i3} + \beta_5 d_{i4} + \beta_6 d_{i5} + \epsilon_i$$

where: X_{i1} = years of experience and d_{i1} to d_{i5} are indicator variables for Office, Sales, Service, Technical, and Worker, respectively (Management as the base line). Let us define them as follows:

$$d_{i1} = \begin{cases} 1 & \text{if office} \\ 0 & \text{otherwise.} \end{cases} \qquad d_{i3} = \begin{cases} 1 & \text{if service} \\ 0 & \text{otherwise.} \end{cases} \qquad d_{i5} = \begin{cases} 1 & \text{if worker} \\ 0 & \text{otherwise.} \end{cases}$$

$$d_{i2} = \begin{cases} 1 & \text{if sales} \\ 0 & \text{otherwise.} \end{cases} \qquad d_{i4} = \begin{cases} 1 & \text{if technical} \\ 0 & \text{otherwise.} \end{cases}$$

To test whether there is any difference in the wages of different occupations we perform a partial F test on the coefficients of the indicators. Our hypotheses are:

$$\begin{aligned} H_0 &: \quad \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \\ H_a &: \quad \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned}$$

```
Analysis of Variance Table

Response: log(wage)
            Df  Sum Sq Mean Sq F value    Pr(>F)
experience   1   1.721  1.7213  7.6317  0.005935 **
occupation   5  27.864  5.5729 24.7088 < 2.2e-16 ***
Residuals  527 118.861  0.2255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the ANOVA Table, we can derive the test statistics:

$$F^* = \frac{27.864/5}{0.2255} = 24.7088$$

With $\alpha = 0.05$, we require $F(24.7088; p-1 = 4, n-p = 529) = 2.389$. Since $F^* = 24.71 > 2.389$, we can reject H_0 in favour of the alternative. That is, we cannot drop the coefficients of the indicators from the model. In other words, controlling the `experience`, there is different in log(wage) for different types of occupation.

Part i

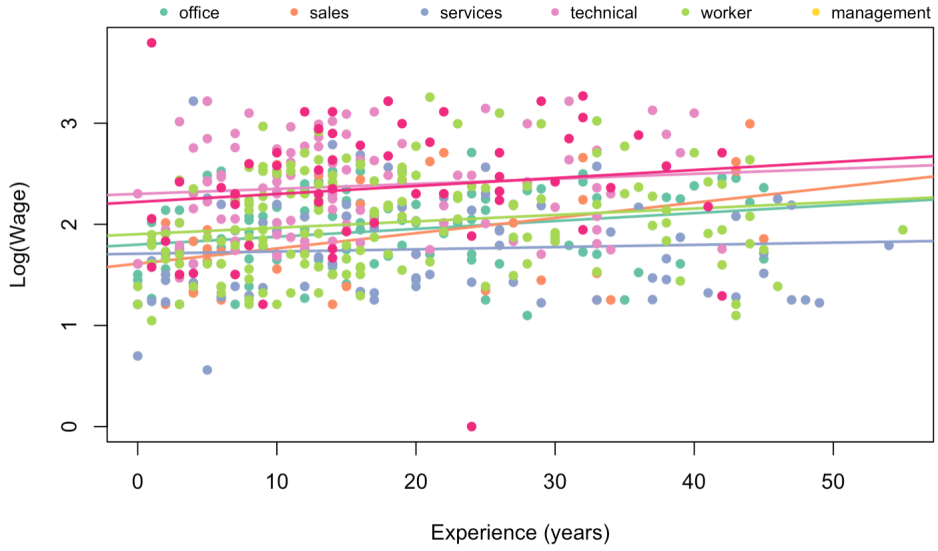
Since we are concerned that interaction effects may be present between `experience` and `occupation`, we consider the MLR model

$$\begin{aligned} Y_i = \beta_0 + \beta_1 X_1 + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \beta_5 d_4 + \beta_6 d_5 + \\ \beta_7 X_1 d_1 + \beta_8 X_1 d_2 + \beta_9 X_1 d_3 + \beta_{10} X_1 d_4 + \beta_{11} X_1 d_5 + \epsilon_i \end{aligned}$$

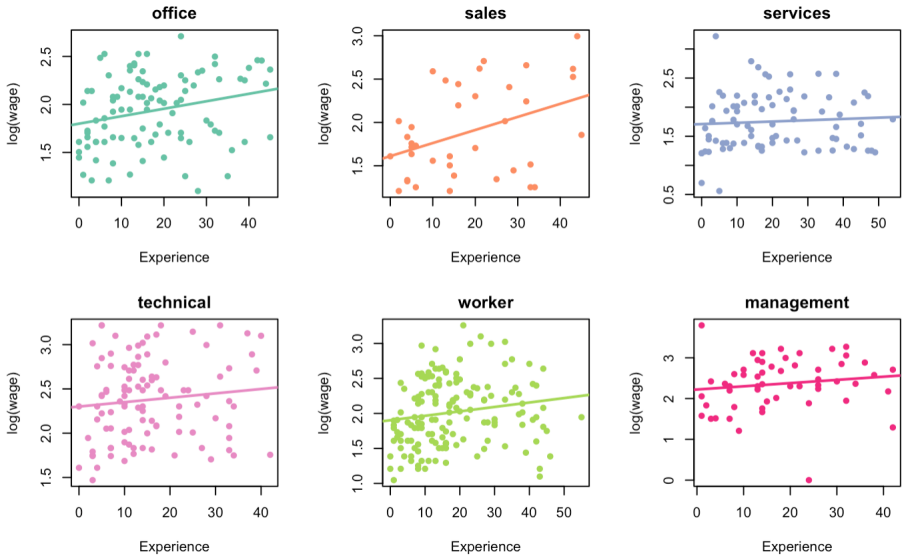
Note that both the intercept and slope now differ for each type of occupation in regression model:

$$\begin{aligned} E\{Y\} &= (\beta_0 + \beta_2) + (\beta_1 + \beta_7) X_1 \quad \text{(Office)} \\ E\{Y\} &= (\beta_0 + \beta_3) + (\beta_1 + \beta_8) X_1 \quad \text{(Sales)} \\ E\{Y\} &= (\beta_0 + \beta_4) + (\beta_1 + \beta_9) X_1 \quad \text{(Service)} \\ E\{Y\} &= (\beta_0 + \beta_5) + (\beta_1 + \beta_{10}) X_1 \quad \text{(Technical)} \\ E\{Y\} &= (\beta_0 + \beta_6) + (\beta_1 + \beta_{11}) X_1 \quad \text{(Worker)} \\ E\{Y\} &= \beta_0 + \beta_1 X_1 \quad \text{(Management)} \end{aligned}$$

We have the scatter plot of log(wage) against experience as follow



While there is a slightly linear positive relationship, the interaction is too small to be discernible. As the plot is quite messy, we can plot the scatter plots for each occupation individually,



It can be seen that for `sales`, the positive linear relationship is more visible than other types of occupation. However, it is also true that there are a lot of variation around the fitted regression lines. To formally test whether the interaction is significant, we will compare the two models: the full model as stated above, and a reduced model without the interaction terms as stated in part (h). The hypotheses are:

$$\begin{aligned} H_0: &\quad \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0 \\ H_a: &\quad \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned}$$

```
Analysis of Variance Table

Response: log(wage)
                    Df  Sum Sq Mean Sq F value    Pr(>F)
experience           1   1.721  1.7213  7.6161  0.005988 **
occupation           5  27.864  5.5729 24.6581 < 2.2e-16 ***
experience:occupation  5   0.886  0.1772  0.7838  0.561610
Residuals          522 117.975  0.2260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can use the test statistics

$$F^* = \frac{0.886/5}{0.2260} = 0.7838$$

With $\alpha = 0.05$, we require $F(0.7838; 5, 522) = 2.2312$. Since $F^* = 0.7838 < 2.2312$, we conclude H_0. That is interaction effect is not significant and that the effect of an individual's years of potential work experience on their log(wage) does not depend on their type of occupation.

Appendix

Part a

```
wage_data = read.csv('Wage.csv', stringsAsFactors = T)
attach(wage_data)

mod.a = lm(log(wage) ~ education + experience + age)
anova(mod.a)

qf(1-0.05, 3, 530)
```

Part b

```
summary(mod.a)$coef
confint(mod.a, level = 1 - 0.05/3)
```

Part c

```
dat = data.frame(education, experience, age)
cor(dat)

pairs(dat) # Scatter plot matrix

library(faraway)
faraway::vif(mod.a
```

Part d

```
mod.d = lm(log(wage) ~ education + experience)
summary(mod.d)$coef
```

```
exp.new = sort(experience)
edu.constant = rep(mean(education), length(exp.new))

dat.new = data.frame(education = edu.constant, experience = exp.new)

plot(exp.new, exp(predict(mod.d, newdata = dat.new)),
     type = 'l', lwd = 2, ylab = 'Log(Wage)', xlab = 'Experience (Years)',
     main = 'Experience vs. Log(wage), considering Education')
```

```
mod.d2 = lm(log(wage) ~ education + experience + I(experience^2))
summary(mod.d2)$coef
```

Part e

```
mod.e = lm(log(wage) ~ married)
summary(mod.e)$coef

#Box plot
plot(married, log(wage), main = 'Log(Wage) vs. Marital Status',
     ylab = 'Log(Wage)', xlab = 'Married ')

n = length(wage_data[,1])
qt(1 - 0.05, n-2)

confint(mod.e) # Confidence Interval
```

Part f

```
mod.f = lm(log(wage) ~ education + experience + I(experience^2) + married)
summary(mod.f)$coef
```

Part g

```
plot(fitted(mod.f), rstudent(mod.f),
     main = "Externally Studentised Residual vs Fitted Values")
identify(fitted(mod.f), rstudent(mod.f))

plot(mod.f, which = c(2)) ## QQ plot
```

```
barplot(hatvalues(mod.f), main = 'Leverge plot')
p = 5
threshold = (2*p) / n
abline(h = threshold, col = 'red')
id = order(abs(hatvalues(mod.f)), decreasing = T)[1:2]

plot(mod.f, which = c(4)) # Cook's Distance plot
wage_data[63,]
wage_data[171,]
wage_data[200,]
mean(wage_data$wage)
```

```
par(mfrow = c(2,2))
plot(dfbetas(mod.f)[,2], type = 'h', main = 'Education')
abline(h = 0)
abline(h = c(-2/sqrt(n), 2/sqrt(n)), col = 2)

plot(dfbetas(mod.f)[,3], type = 'h', main = 'Experience')
abline(h = 0)
abline(h = c(-2/sqrt(n), 2/sqrt(n)), col = 2)

plot(dfbetas(mod.f)[,4], type = 'h',  main = 'Second-Order Experience')
abline(h = 0)
abline(h = c(-2/sqrt(n), 2/sqrt(n)), col = 2)

plot(dfbetas(mod.f)[,5], type = 'h', main = 'Marriage')
abline(h = 0)
abline(h = c(-2/sqrt(n), 2/sqrt(n)), col = 2)
```

Part h

```
mod.h = lm(log(wage) ~ experience + occupation)
anova(mod.h)

f_stat.h = (27.864 / 5) / 0.2255
qf(1-0.05, 4, 529)
```

Part i

```
mod.i = lm(log(wage) ~ experience + occupation + experience*occupation)

ls_ = c("office" ,"sales","services","technical","worker", "management")
library("RColorBrewer")
cols = brewer.pal(n = 6, name = "Set2")
```

Acknowledgement:

The function `add_legend` below belongs to the user **Jan van der Laan** from Stack Overflow (https://stackoverflow.com/questions/3932038/plot-a-legend-outside-of-the-plotting-area-in-base-graphics) (accessed 19/10/2020) and is not of my original work. The function was employed to add the legends outside of the scatter plot.

```
add_legend <- function(...) {
  opar <- par(fig=c(0, 1, 0, 1), oma=c(0, 0, 0, 0),
    mar=c(0, 0, 0, 0), new=TRUE)
  on.exit(par(opar))
  plot(0, 0, type='n', bty='n', xaxt='n', yaxt='n')
```

```
  legend(...)
}
```

```
par(mar = c(5, 4, 1.4, 0.2))

plot(experience, log(wage), xlab = 'Experience (years)',
     ylab = 'Log(Wage)', type = 'n')

for (i in 1:5){
  points(experience[occupation == ls_[i]], log(wage[occupation == ls_[i]]),
         col = cols[i], pch = 16, ylim = c(0, 4))

  abline(a = coef(mod.i)[1] + coef(mod.i)[i+2],
       b = coef(mod.i)[2] + coef(mod.i)[i+7], col = cols[i], lwd = 2)
}

points(experience[occupation == ls_[6]], log(wage[occupation == ls_[6]]),
       col = "#F0027F", pch = 16)
abline(a = coef(mod.i)[1],
       b = coef(mod.i)[2], col =  "#F0027F", lwd = 2)
add_legend("topright", legend = ls_, pch = 16, col = cols,
           horiz = TRUE, bty = 'n', cex = 0.8)
```

```
par(mfrow=c(2,3))
for (i in 1:5){
  plot(experience[occupation == ls_[i]], log(wage[occupation == ls_[i]]),
       xlab = 'Experience', ylab = 'log(wage)',
       col = cols[i], pch = 16, main = ls_[i])

  abline(a = coef(mod.i)[1] + coef(mod.i)[i+2],
         b = coef(mod.i)[2] + coef(mod.i)[i+7], col = cols[i], lwd = 2)
}

plot(experience[occupation == ls_[6]], log(wage[occupation == ls_[6]]),
     xlab = 'Experience', ylab = 'log(wage)',
     col = "#F0027F", pch = 16, main = ls_[6])

abline(a = coef(mod.i)[1], b = coef(mod.i)[2], col =  "#F0027F", lwd = 2)
```

```
anova(mod.i)
qf(1-0.05, 4, 529)
```