Assignment 4 CS 4070

Nirmal Roy, 4724429

December 18, 2017

Question 1.a

For the given dataframe we have language is an explanatory variable which is a factor with levels **English**, **French**, and **Others**. We construct two dummy regressors for the three level factor in the following way:

Table 1: Dummy Regressors

Levels	D_1	D_2
English	0	0
French	1	0
Others	0	1

A model for the regression of wages on education(X_1), age(X_2) and type of language is given by

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \tag{1}$$

The three regression equation therefore are:

$$\text{English: } Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$
$$\text{French : } Y_i = (\alpha + +\gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$
$$\text{Others : } Y_i = (\alpha + +\gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Using summary in R we get the coefficients to be $\alpha = -6.047982$, $\beta_1 = 0.902757$, $\beta_2 = 0.256748$, $\gamma_1 = 0.130743$ and $\gamma_2 = 0.109277$. Hence, (1) becomes

$$\hat{Y} = -6.047982 + 0.902757 X_1 + 0.256748 X_2 + 0.130743 D_1 + 0.109277 D_2 \tag{2}$$

To test the hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$, we need to consider the full model and the reduced model without the dummy regressor coefficients. Let, $R_1^2 = 0.2491028$ be the correlation coefficient for the full model and $R_0^2 = 0.2490697$ be the same for the reduced model.

Now we can calculate the incremental F statistic,

$$F_0 = \frac{n-k-1}{q} * \frac{R_1^2 - R_0^2}{1 - R_1^2} = \frac{3987-4-1}{2} * \frac{0.2491028 - 0.2490697}{1 - 0.2491028} = 0.0878 \tag{3}$$

with 2 and 3982 degrees of freedom for which p-value is 0.9159 which we can find from an F distribution table. Since this value is greater than 0.05 we cannot reject the null hypothesis that the factor has no effect.

The coefficient α, therefore, gives the intercept for English language ; γ_1 represents the constant vertical difference between the parallel regression planes for English and French (fixing the values of education and age); and γ represents the constant vertical distance between the regression planes for Other and English (again, fixing education and age). Since English is coded 0 for both the dummy regressors it serves as the baseline with which the regression for the other factors(French and Others).

Also, β_1 gives the change in wage for unit change in education for all levels of language. Similarly, β_2 gives the change in wages for unit change in age for all levels of language.

The three fitted regression equation therefore are:

$$\begin{aligned}\text{English: } \hat{Y} &= -6.048 + 0.902X_1 + 0.257X_2\\ \text{French : } \hat{Y} &= -5.917 + 0.902X_1 + 0.257X_2\\ \text{Others : } \hat{Y} &= -5.939 + 0.902X_1 + 0.257X_2\end{aligned}$$

Question 1.b

If we allow interaction in the model, where each quantitative explanatory variable interact with the factor the regression equation of wages on education(X_1) and age(X_2) is given by

$$\begin{aligned}Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2}\\ +\delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} \epsilon_i\end{aligned} \tag{4}$$

Using summary in R we get the coefficients to be $\alpha = -6.726275$, $\beta_1 = 0.948728$, $\beta_2 = 0.258354$, $\gamma_1 = 5.496013$, $\gamma_2 = 1.446932$, $\delta_{11} = -0.230440$, $\delta_{12} = -0.138087$, $\delta_{21} = -0.062332$, and $\delta_{22} = 0.010594$ Hence, (4) becomes

$$\begin{aligned}\hat{Y} = -6.726275 + 0.948728X_1 + 0.258354X_2 + 5.496013D_1 + 1.446932D_2\\ -0.230440X_1D_1 - 0.138087X_1D_2 - 0.062332X_2D_1 + 0.010594X_2D_2\end{aligned} \tag{5}$$

To perform incremental F-test of each interaction, we construct the following table which shows the full model and the reduced model and the variables/interactions involved with the correlation coefficients for the respective model. Then we can compute the F values using equation (3) and test whether they have effect on the regression.

Table 2: Table showing the model and parameter to be tested

Model	**Terms**	**Parameters**	R^2	**df**
1	*E,A,L,E*L1,E*L2,A*L1, A*L2*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	0.250423	8
2	*E,A,L,E*L2,A*L1, A*L2*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{12}, \delta_{21}, \delta_{22}$	0.249929	7
3	*E,A,L,E*L1,A*L1, A*L2*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{21}, \delta_{22}$	0.2499713	7
4	*E,A,L,E*L1,E*L2, A*L2*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{22}$	0.2499207	7
5	*E,A,L,E*L1,E*L2,A*L1*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21},$	0.2503958	7

where E is Education, A is Age, L is Language, and L1 is English/French and L2 is English/Other.

Table 3: Analysis-of-Variance Table, Showing Incremental F-Tests

Source	**Models Contrasted**	**df**	**F**	**p**
English	1-2	1	2.6213	0.1055
French	1-3	1	2.397	0.1216
Others	1-4	1	2.6657	0.1026
Others	1-5	1	0.1441	0.7042

As we can see the p-value for all the reduced model is greater than 0.05 and hence the null models that they individually don't have any influence on the regression can't be rejected.

Regression equation for each category of the factor is given by

$$\text{English: } \hat{Y} = -6.726 + 0.949X_1 + 0.258X_2$$
$$\text{French : } \hat{Y} = -1.23 + 0.718X_1 + 0.196X_2$$
$$\text{Others : } \hat{Y} = -5.279 + 0.811X_1 + 0.269X_2$$

We can see from fig 1, 2, and 3 that these values are equal to values we obtain by performing the regression on the quantitative explanatory variables separately for each category of the factor. I have used code [#1b.3] from appendix for the same.

```
Call:
lm(formula = wages ~ education + age, data = eng)

Residuals:
    Min      1Q  Median      3Q     Max
-22.079  -4.537  -0.776   3.639  37.790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.726275   0.708074  -9.499   <2e-16 ***
education    0.948728   0.042577  22.282   <2e-16 ***
age          0.258354   0.009964  25.928   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.83 on 3241 degrees of freedom
  (2593 observations deleted due to missingness)
Multiple R-squared:  0.2538,    Adjusted R-squared:  0.2534
F-statistic: 551.3 on 2 and 3241 DF,  p-value: < 2.2e-16
```

Figure 1: The coefficients obtain by regressing wages on education and age for language English

```
Call:
lm(formula = wages ~ education + age, data = French)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7668  -4.5991  -0.7982   3.4709  25.8577

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.23026    2.52209  -0.488    0.626
education    0.71829    0.13862   5.182 4.46e-07 ***
age          0.19602    0.03762   5.211 3.86e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.959 on 256 degrees of freedom
  (359 observations deleted due to missingness)
Multiple R-squared:  0.1517,    Adjusted R-squared:  0.145
F-statistic: 22.88 on 2 and 256 DF,  p-value: 7.197e-10
```

Figure 2: The coefficients obtain by regressing wages on education and age for language French

```
Call:
lm(formula = wages ~ education + age, data = Other)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9603  -4.3622  -0.6674   3.5589  29.5021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.27934    1.60088  -3.298  0.00105 **
education    0.81064    0.07672  10.566  < 2e-16 ***
age          0.26895    0.02551  10.544  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.673 on 481 degrees of freedom
  (728 observations deleted due to missingness)
Multiple R-squared:  0.2741,    Adjusted R-squared:  0.2711
F-statistic: 90.81 on 2 and 481 DF,  p-value: < 2.2e-16
```

Figure 3: The coefficients obtain by regressing wages on education and age for language Other

Question 1.c

In fig 4 and 5 we plot the "effect display" of education by type and age by type respectively. The solid lines give fitted values under the model, while the broken lines give 95% pointwise confidence intervals around the fit. To compute the fitted values in each figure the other variable has been set to its mean. As we can see in both the plots the effect does not change much with different types of language.

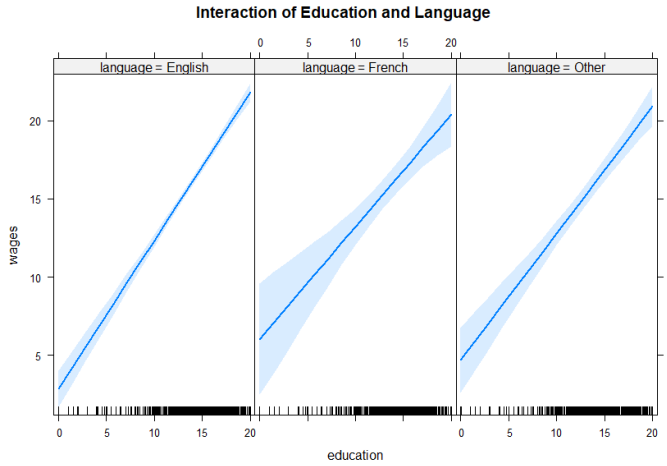

Figure 4: Education-by-type "effect display" for the regression of wages on education, age and type of Language

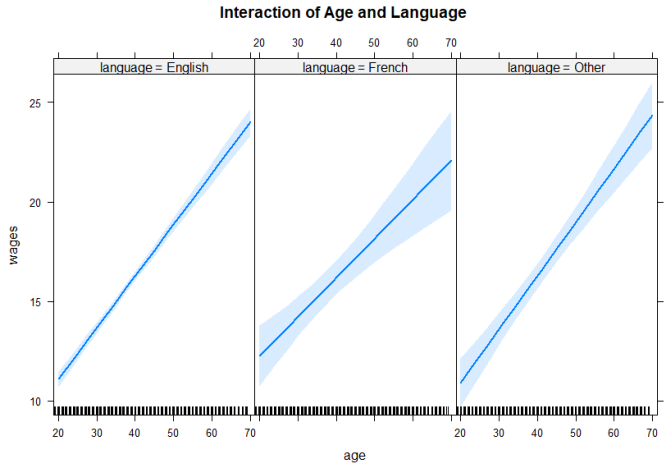

Figure 5: Age-by-type "effect display" for the regression of wages on education, age and type of Language

Going by the null test done in sections 1.a and 1.b we can conclude that neither the dummy regressors nor the interaction between the quantitative variable and the factors are essential for our regression of wages.

Hence the final regression equation for all categories is given by :

$$\hat{Y} = -6.036836 + 0.903962X_1 + 0.256568X_2 \tag{6}$$

Appendix

```r
#1a
langauge = factor(SLID$language)
summary(lm(wages ~ education + age +language, SLID))

#1b.1
summary(lm(wages ~ education + age +language + education:language
           + age:language, na.omit(SLID)))
#1b.2
wages = na.omit(SLID$wages)
TSS = sum((wages-mean(wages))^2)
c1 = as.numeric(SLID$language=="French")
c2 = as.numeric(SLID$language == "Other")
delta11 = SLID$education*c1
delta12 = SLID$education*c2
delta21 = SLID$age*c1
delta22 = SLID$age*c2

fitmod1 = lm(wages~education + age + language + delta11 + delta12 + delta21 + delta22, SLID)
fitmod2 = lm(wages~education + age + language + delta12 + delta21 + delta22, SLID)
fitmod3 = lm(wages~education + age + language + delta11 + delta21 + delta22, SLID)
fitmod4 = lm(wages~education + age + language + delta11 + delta12 + delta22, SLID)
fitmod5 = lm(wages~education + age + language + delta11 + delta12 + delta21, SLID)

anova(fitmod2, fitmod1)
anova(fitmod3, fitmod1)
anova(fitmod4, fitmod1)
anova(fitmod5, fitmod1)
#1b.3
eng <- SLID[SLID$language=="English",]
summary(lm(wages~ education + age, eng))

#1c
plot(effect("education:language", lm(wages ~ education + age + language+
                                       education:language + age:language, SLID)))
summary(lm(wages~education + age , SLID))
```

Figure 6: Relevant Codes