Ec122 - HW6

Ben Juarez

11/7/2021

HGL 7.1

```
setwd("~/Desktop/R")
metrics = read.table("dat/metrics.dat")
colnames(metrics) = c("salary", "gpa", "metrics", "female")
```

Part a

```
SAL = metrics$salary
GPA = metrics$gpa
METRICS = metrics$metrics

metrics_lm = lm(SAL ~ GPA + METRICS)
summary(metrics_lm)
```

```
##
## Call:
## lm(formula = SAL ~ GPA + METRICS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3238.1  -828.7  -225.4   731.0  3967.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24199.7     1078.4  22.440  < 2e-16 ***
## GPA           1643.3      352.3   4.665 2.59e-05 ***
## METRICS       5033.1      456.3  11.030 1.23e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1435 on 47 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7257
## F-statistic: 65.81 on 2 and 47 DF,  p-value: 2.364e-14
```

With this estimated equation, we can interpret several things. With a 1 unit increase in GPA, we estimate a 1,643 dollar increase in salary. Those who take econometrics are estimated to have a 5033 dollar increase in salary relative to those who did not take econometrics. Someone who did not take econometrics and had a 0.0 GPA would theoretically have a starting salary of 24,200 dollars. With $R^2 = 74$, we can determine that GPA and METRICS explains 74% of the variation.

Part b

We modify the equation as follows to take gender into account:

$$S\hat{A}L = \beta_1 + \beta_2 \cdot GPA + \beta_3 \cdot METRICS + \beta_4 \cdot FEMALE + e$$

$$\implies S\hat{A}L = \begin{cases} (\beta_1 + \beta_4) + \beta_2 \cdot GPA + \beta_3 \cdot METRICS & FEMALE = 1 \\ \beta_1 + \beta_2 \cdot GPA + \beta_3 \cdot METRICS & FEMALE = 0 \end{cases}$$

Part c

Let us modify the equation in order to see if the value of econometrics was the same for men and women by using a slope-indicator variable:

$$S\hat{A}L = \beta_1 + \beta_2 \cdot GPA + \beta_3 \cdot METRICS + \beta_4 \cdot FEMALE + \beta_5(METRICS \times FEMALE) + e$$

$$\implies S\hat{A}L \begin{cases} (\beta_1 + \beta_4) + \beta_2 \cdot GPA + (\beta_3 + \beta_5)METRICS & FEMALE = 1 \\ \beta_1 + \beta_2 \cdot GPA + \beta_3 \cdot METRICS & FEMALE = 0 \end{cases}$$

This way, we can see that β_5 will help us determine the change in slope between females and males.

HGL 7.5

```
setwd("~/Desktop/R")
utown = read.table("dat/utown.dat")
colnames(utown) = c("price", "sqft", "age", "utown", "pool", "fplace")
```

Part a

Let us estimate the model as follows:

$$ln(PRICE) = \beta_1 + \delta_1 \cdot UTOWN + \beta_2 \cdot SQFT + \gamma(SQFT \times UTOWN) + \beta_3 \cdot AGE + \delta_2 \cdot POOL + \delta_3 FPLACE + e$$

```
PRICE = utown$price
UTOWN = utown$utown
SQFT = utown$sqft
AGE = utown$age
POOL = utown$pool
FPLACE = utown$fplace

price_llr = lm(log(PRICE) ~ UTOWN + SQFT + (SQFT * UTOWN) + AGE + POOL + FPLACE)
summary(price_llr)
```

```
##
## Call:
## lm(formula = log(PRICE) ~ UTOWN + SQFT + (SQFT * UTOWN) + AGE +
##     POOL + FPLACE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.298458 -0.039371  0.002029  0.044092  0.183979
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4637921  0.0263762 169.236  < 2e-16 ***
## UTOWN        0.3334175  0.0358795   9.293  < 2e-16 ***
## SQFT         0.0359599  0.0010444  34.430  < 2e-16 ***
## AGE         -0.0009043  0.0002181  -4.146 3.67e-05 ***
## POOL         0.0189899  0.0050978   3.725 0.000206 ***
## FPLACE       0.0065561  0.0041405   1.583 0.113642
## UTOWN:SQFT  -0.0034284  0.0014145  -2.424 0.015539 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06486 on 993 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.8611
## F-statistic:  1033 on 6 and 993 DF,  p-value: < 2.2e-16
```

Part b

Let us interpret SQFT and AGE as follows using partial derivatives:

$$\frac{\partial PRICE}{\partial AGE}(\frac{1}{PRICE}) = \beta_3$$

With this, we can determine that 1 year increase in the age of the house implies that it is estimated to sell for 0.09% less.

$$\frac{\partial PRICE}{\partial SQFT}(\frac{1}{PRICE}) = \beta_2 + \gamma \cdot UTOWN$$

With this, we can determine that an increase in SQFT implies either a 3.6% (not close to university) or $0.0359 - 0.0034 \implies 3.25\%$ (close to university).

Part c

Let us find both a rough and exact calculation of the percentage change in price due to the presence of a pool:

$$ln(PRICE)_{\text{pool}} - ln(PRICE)_{\text{no pool}} = \delta_2 \implies 0.019 \cdot 100 = 1.90\%$$

$$\frac{PRICE_{\text{pool}} - PRICE_{\text{no pool}}}{PRICE_{\text{pool}}} = e^{\delta_2} - 1 \implies 0.01917 \cdot 100 = 1.92\%$$

Part d

Let us find both a rough and exact calculation of the percentage change in price due to the presence of a fireplace:

$$ln(PRICE)_{\text{fireplace}} - ln(PRICE)_{\text{no fireplace}} = \delta_3 \implies 0.0065561 \cdot 100 = 0.66\%$$

$$\frac{PRICE_{\text{fireplace}} - PRICE_{\text{no fireplace}}}{PRICE_{\text{fireplace}}} = e^{\delta_3} - 1 \implies 0.006577638 \cdot 100 = 0.66\%$$

Part e

Let us compute the percentage change in price of 2500-square-foot home close to the university relative to another house in another location as follows:

$$ln(\hat{PRICE}_{\text{close to U}})|_{SQFT=25} - ln(\hat{PRICE}_{\text{NOT close to U}})|_{SQFT=25} = \beta_2 - \gamma(25) = 0.0333 - 0.00343(25) = 0.248$$

$$\implies (e^{0.0248} - 1) \cdot 100 = 28.1\%$$

Thus, we have a 28.1% increase in price for being close to the university.

HGL 7.9

```
setwd("~/Desktop/R")
star = read.table("dat/star.dat")
colnames(star) = c("id", "schid", "tchid", "tchexper", "absent", "readscore", "mathscore", "totalscore"
```

Part a

```
mean(subset(star, regular==1)$totalscore)
```

```
## [1] 918.0429
```

```
mean(subset(star, aide==1)$totalscore)
```

```
## [1] 918.3568
```

```
mean(subset(star, small==1)$totalscore)
```

```
## [1] 931.9419
```

The averages for (i), (ii), (iii) are calculated as shown, respectively. We can see that the test scores are higher in the smaller class, and that the presence of an aide does not seem to impact total test scores.

Part b

Let us estimate the following model: $TOTALSCORE_i = \beta_1 + \beta_2 \cdot SMALL_i + \beta_3 \cdot AIDE_i + e_i$ where $AIDE$ is an indicator variable.

```
TOTALSCORE = star$totalscore
SMALL = star$small
AIDE = star$aide

b_lm = lm(TOTALSCORE ~ SMALL + AIDE)
summary(b_lm)
```

```
##
## Call:
## lm(formula = TOTALSCORE ~ SMALL + AIDE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283.04  -50.94   -7.04   42.64  334.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 918.0429     1.6412 559.363  < 2e-16 ***
## SMALL        13.8990     2.4085   5.771  8.3e-09 ***
## AIDE          0.3139     2.3102   0.136    0.892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 73.49 on 5783 degrees of freedom
## Multiple R-squared:  0.0073, Adjusted R-squared:  0.006957
## F-statistic: 21.26 on 2 and 5783 DF,  p-value: 6.3e-10
```

We can see that $931.94 - 918.04 = 13.899$, relating the coefficient of SMALL to the averages such that it is the difference between the average score of small classes and the average score of regular classes. Similarly, with $918.3568 - 0.3139 = 918.0429$, we see that the coefficient of AIDE is the difference between average score of classes with aides and average score of classes without aides (both for regular classes).

Let us test the statistical significance of β_3 at a 5% level of significance with $t_c = 1.96$:

$$t = \frac{b_3}{se(b_3)} = 0.136 < 1.96 = t_c$$

Thus, we cannot conclude that there is significant difference between total test scores in regular classes with no aide and in regular classes with an aide (fail to reject null hypothesis $\beta_3 = 0$).

Part c

Let us estimate the following model: $TOTALSCORE_i = \beta_1 + \beta_2 \cdot SMALL_i + \beta_3 \cdot AIDE_i + \beta_4 \cdot TCHEXPER + e_i$ where $AIDE$ is an indicator variable.

```
TOTALSCORE = star$totalscore
SMALL = star$small
AIDE = star$aide
TCHEXPER = as.integer(star$tchexper)
```

```
## Warning: NAs introduced by coercion
```

```
c_lm = lm(TOTALSCORE ~ SMALL + AIDE + TCHEXPER)
summary(c_lm)
```

```
##
## Call:
## lm(formula = TOTALSCORE ~ SMALL + AIDE + TCHEXPER)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.13  -51.41   -7.44   42.57  341.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 904.7212     2.2280 406.071  < 2e-16 ***
## SMALL        14.0061     2.3953   5.847 5.27e-09 ***
## AIDE         -0.6006     2.3065  -0.260    0.795
## TCHEXPER      1.4690     0.1672   8.784  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.08 on 5762 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.02033,    Adjusted R-squared:  0.01982
## F-statistic: 39.86 on 3 and 5762 DF,  p-value: < 2.2e-16
```

Let us determine if this variable is statistically significant (5% level of significance):

$$t = \frac{b_4}{se(b_4)} = 8.78 > t_c = 1.96$$

Thus, we can say that this variable is statistically significant since we reject the null hypothesis that $\beta_4 = 0$. We can also see that there is some change to β_2, β_3 since β_2 went from 13.9 to 14.0 while β_3 changed signs (very little impact).

Part d

Let us estimate the following model: $TOTALSCORE_i = \beta_1 + \beta_2 \cdot SMALL_i + \beta_3 \cdot AIDE_i + \beta_4 \cdot TCHEXPER + \beta_5 \cdot BOY + \beta_6 \cdot FREELUNCH + \beta_7 \cdot WHITE_ASIAN + e_i$ where $AIDE$ is an indicator variable.

```
TOTALSCORE = star$totalscore
SMALL = star$small
AIDE = star$aide
TCHEXPER = as.integer(star$tchexper)
```

```
## Warning: NAs introduced by coercion
```

```
BOY = star$boy
FREELUNCH = star$freelunch
WHITE_ASIAN = star$white_asian

d_lm = lm(TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH + WHITE_ASIAN)
summary(d_lm)
```

```
##
## Call:
## lm(formula = TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH +
##     WHITE_ASIAN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.27  -47.92   -8.07   40.29  323.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 923.2498     3.1210 295.822  < 2e-16 ***
## SMALL        13.8960     2.2936   6.059 1.46e-09 ***
## AIDE          0.6983     2.2093   0.316    0.752
## TCHEXPER      1.1140     0.1613   6.908 5.43e-12 ***
## BOY         -14.0452     1.8457  -7.610 3.19e-14 ***
## FREELUNCH   -34.1170     2.0639 -16.531  < 2e-16 ***
## WHITE_ASIAN  11.8373     2.2108   5.354 8.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.98 on 5759 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1023, Adjusted R-squared:  0.1013
## F-statistic: 109.3 on 6 and 5759 DF,  p-value: < 2.2e-16
```

We can say that these new variables are statistically significant considering the summary (all at 1% level). We also again see very small changes to β_2, β_3, so we can say that these estimates are not meaningfully affected.

Part e

Let us estimate the following model: $TOTALSCORE_i = \beta_1 + \beta_2 \cdot SMALL_i + \beta_3 \cdot AIDE_i + \beta_4 \cdot TCHEXPER + \beta_5 \cdot BOY + \beta_6 \cdot FREELUNCH + \beta_7 \cdot WHITE_ASIAN + \beta_8 \cdot TCHWHITE + \beta_9 \cdot TCHMASTERS + \beta_10 \cdot SCHURBAN + \beta_11 \cdot SCHRURAL + e_i$ where $AIDE$ is an indicator variable.

```
TOTALSCORE = star$totalscore
SMALL = star$small
AIDE = star$aide
TCHEXPER = as.integer(star$tchexper)
```

```
## Warning: NAs introduced by coercion
```

```
BOY = star$boy
FREELUNCH = star$freelunch
WHITE_ASIAN = star$white_asian
TCHWHITE = star$tchwhite
TCHMASTERS = star$tchmasters
SCHURBAN = star$schurban
SCHRURAL = star$schrural

e_lm = lm(TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH + WHITE_ASIAN + TCHWHITE + TCHMASTERS +
summary(e_lm)
```

```
##
## Call:
## lm(formula = TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH +
##     WHITE_ASIAN + TCHWHITE + TCHMASTERS + SCHURBAN + SCHRURAL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.91  -47.12   -8.43   39.97  314.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 931.7553     3.9401 236.482  < 2e-16 ***
## SMALL        13.9803     2.3023   6.072 1.34e-09 ***
## AIDE          1.0023     2.2166   0.452  0.65114
## TCHEXPER      1.1562     0.1657   6.979 3.31e-12 ***
## BOY         -14.0083     1.8426  -7.602 3.37e-14 ***
## FREELUNCH   -32.5325     2.1260 -15.302  < 2e-16 ***
## WHITE_ASIAN  16.2326     2.7804   5.838 5.57e-09 ***
## TCHWHITE     -7.6683     2.8420  -2.698  0.00699 **
## TCHMASTERS   -3.5598     2.0193  -1.763  0.07798 .
## SCHURBAN     -5.7499     2.8580  -2.012  0.04428 *
## SCHRURAL     -7.0061     2.5585  -2.738  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.86 on 5755 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1059, Adjusted R-squared:  0.1043
## F-statistic: 68.14 on 10 and 5755 DF,  p-value: < 2.2e-16
```

Following the summary, we can see that TCHWHITE, SCHRURAL are significant at the 1% level, SCHURBAN is significant at the 5% level, and TCHMASTERS is significant at the 10% level (failed to reject at 5% level).

Again we see minimal changes to β_2 and β_3.

Part f

Considering the previous parts, it is interesting to see that the added variables are statistically significant themselves, but we still saw very little impact to the main factors of SMALL and AIDE, which reinforces that these are the key variables. This also tells us that the added variables are uncorrelated to our key variables and treatment effects.

Part g

```
library("car")
```

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

```
SCHID = star$schid
g_lm = lm(TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH + WHITE_ASIAN + TCHWHITE + TCHMASTERS +
linearHypothesis(g_lm, matchCoefs(g_lm,"SCHID"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## as.factor(SCHID)123056 = 0
## as.factor(SCHID)128068 = 0
## as.factor(SCHID)128076 = 0
## as.factor(SCHID)128079 = 0
## as.factor(SCHID)130085 = 0
## as.factor(SCHID)159171 = 0
## as.factor(SCHID)161176 = 0
## as.factor(SCHID)161183 = 0
## as.factor(SCHID)162184 = 0
## as.factor(SCHID)164198 = 0
## as.factor(SCHID)165199 = 0
## as.factor(SCHID)166203 = 0
## as.factor(SCHID)168211 = 0
## as.factor(SCHID)168214 = 0
## as.factor(SCHID)169219 = 0
## as.factor(SCHID)169229 = 0
## as.factor(SCHID)169231 = 0
## as.factor(SCHID)169280 = 0
## as.factor(SCHID)170295 = 0
## as.factor(SCHID)173312 = 0
## as.factor(SCHID)176329 = 0
## as.factor(SCHID)180344 = 0
## as.factor(SCHID)189378 = 0
## as.factor(SCHID)189382 = 0
## as.factor(SCHID)189396 = 0
## as.factor(SCHID)191411 = 0
## as.factor(SCHID)193422 = 0
## as.factor(SCHID)193423 = 0
## as.factor(SCHID)201449 = 0
## as.factor(SCHID)203452 = 0
## as.factor(SCHID)203457 = 0
```

```
## as.factor(SCHID)205488 = 0
## as.factor(SCHID)205489 = 0
## as.factor(SCHID)205490 = 0
## as.factor(SCHID)205491 = 0
## as.factor(SCHID)205492 = 0
## as.factor(SCHID)208501 = 0
## as.factor(SCHID)208503 = 0
## as.factor(SCHID)209510 = 0
## as.factor(SCHID)212522 = 0
## as.factor(SCHID)215533 = 0
## as.factor(SCHID)216536 = 0
## as.factor(SCHID)218562 = 0
## as.factor(SCHID)221571 = 0
## as.factor(SCHID)221574 = 0
## as.factor(SCHID)225585 = 0
## as.factor(SCHID)228606 = 0
## as.factor(SCHID)230612 = 0
## as.factor(SCHID)231616 = 0
## as.factor(SCHID)234628 = 0
## as.factor(SCHID)244697 = 0
## as.factor(SCHID)244708 = 0
## as.factor(SCHID)244723 = 0
## as.factor(SCHID)244727 = 0
## as.factor(SCHID)244728 = 0
## as.factor(SCHID)244736 = 0
## as.factor(SCHID)244745 = 0
## as.factor(SCHID)244746 = 0
## as.factor(SCHID)244755 = 0
## as.factor(SCHID)244764 = 0
## as.factor(SCHID)244774 = 0
## as.factor(SCHID)244776 = 0
## as.factor(SCHID)244780 = 0
## as.factor(SCHID)244796 = 0
## as.factor(SCHID)244799 = 0
## as.factor(SCHID)244801 = 0
## as.factor(SCHID)244806 = 0
## as.factor(SCHID)244818 = 0
## as.factor(SCHID)244831 = 0
## as.factor(SCHID)244839 = 0
## as.factor(SCHID)252885 = 0
## as.factor(SCHID)253888 = 0
## as.factor(SCHID)257899 = 0
## as.factor(SCHID)257905 = 0
## as.factor(SCHID)259915 = 0
## as.factor(SCHID)261927 = 0
## as.factor(SCHID)262937 = 0
## as.factor(SCHID)264945 = 0
##
## Model 1: restricted model
## Model 2: TOTALSCORE ~ SMALL + AIDE + TCHEXPER + BOY + FREELUNCH + WHITE_ASIAN +
##     TCHWHITE + TCHMASTERS + as.factor(SCHID)
##
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1   5757 28129270
```

```
## 2   5679 22271315 78   5857955 19.15 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing the joint significance of these school "fixed effects" gives us an F-value of 19.15 which is greater than $F_c = 1.28$ (5% level), so we reject the null hypothesis implying that that (at least several of) the school effects are non-zero. We have a very slight increase to β_2 and β_3 such that β_2 is significant while β_3 is not significant.

HGL 8.19

Part a

Let us estimate the given wage equation as follows:

```
setwd("~/Desktop/R")
data = read.table("dat/cps4_small.dat")
colnames(data) = c("wage", "educ", "exper", "hrswk", "married", "female", "metro", "midwest", "south",

WAGE = data$wage
EDUC = data$educ
EXPER = data$exper
EXPER_2 = EXPER^2
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
wage_llr = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER*EDUC))
summary(wage_llr)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER * EDUC))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28227 -0.32856 -0.02725  0.33751  1.47088
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.297e-01  2.267e-01   2.336  0.01969 *
## EDUC         1.272e-01  1.472e-02   8.642  < 2e-16 ***
## EXPER        6.298e-02  9.536e-03   6.604 6.48e-11 ***
## EXPER_2     -7.139e-04  8.804e-05  -8.109 1.49e-15 ***
## EDUC:EXPER  -1.322e-03  4.949e-04  -2.672  0.00766 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5057 on 995 degrees of freedom
## Multiple R-squared:  0.2445, Adjusted R-squared:  0.2415
```

```
## F-statistic: 80.52 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
coeftest(wage_llr, vcov.=hccm(wage_llr,type="hc1"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  5.2968e-01  2.5283e-01  2.0950   0.03642 *
## EDUC         1.2720e-01  1.6960e-02  7.4999 1.413e-13 ***
## EXPER        6.2981e-02  1.1378e-02  5.5355 3.969e-08 ***
## EXPER_2     -7.1394e-04  9.2013e-05 -7.7591 2.114e-14 ***
## EDUC:EXPER  -1.3224e-03  6.3679e-04 -2.0766   0.03809 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part b

Let us now add MARRIED to the given wage equation:

```
WAGE = data$wage
EDUC = data$educ
EXPER = data$exper
EXPER_2 = EXPER^2
MARRIED = data$married

wage_llr_b = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER*EDUC) + MARRIED)
summary(wage_llr_b)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER * EDUC) +
##     MARRIED)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29834 -0.32252 -0.02409  0.33333  1.45621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5410611  0.2268944   2.385  0.01728 *
## EDUC         0.1261199  0.0147433   8.554  < 2e-16 ***
## EXPER        0.0613731  0.0096289   6.374 2.82e-10 ***
## EXPER_2     -0.0006934  0.0000897  -7.729 2.64e-14 ***
## MARRIED      0.0402895  0.0337911   1.192  0.23342
## EDUC:EXPER  -0.0013091  0.0004949  -2.645  0.00829 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5056 on 994 degrees of freedom
## Multiple R-squared:  0.2456, Adjusted R-squared:  0.2418
## F-statistic: 64.73 on 5 and 994 DF,  p-value: < 2.2e-16
```

Let us now test (5% level of significance) a null hypothesis that wages of married workers are less than or

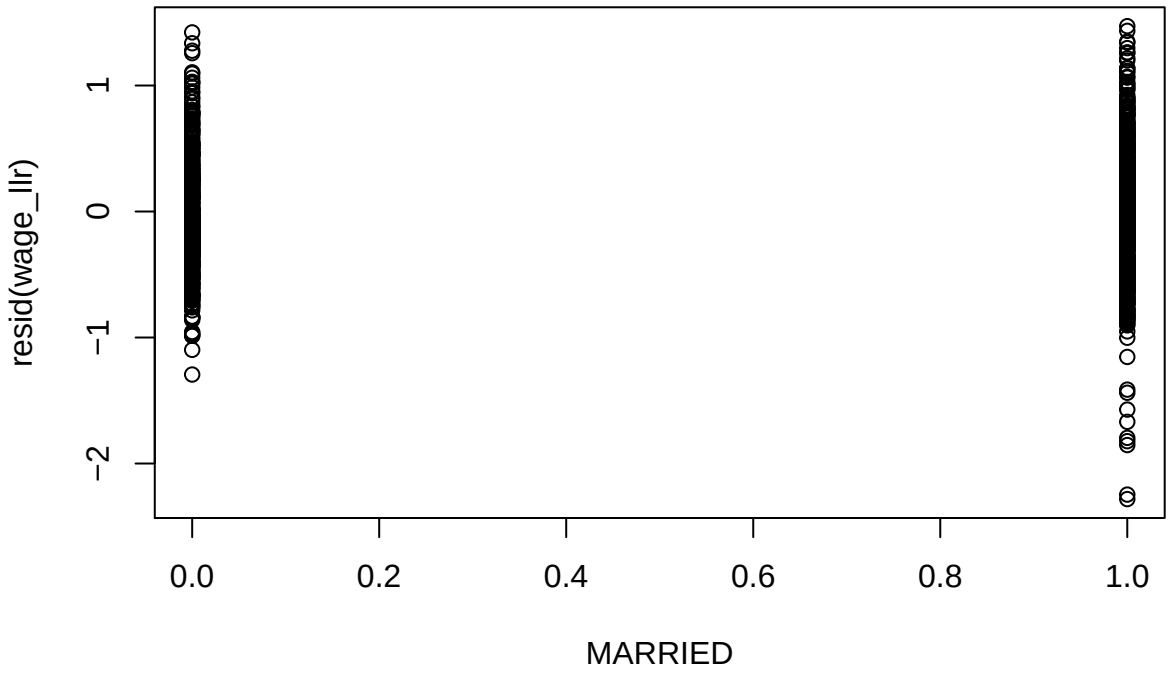
equal to those of unmarried workers against the alternative that wages of married workers are higher:

$$H_0 : \beta_6 \leq 0 \text{ vs } H_1 : \beta_6 > 0 \implies t = \frac{b_6}{se(b_6)} = 1.19 < 1.65 = t_c$$

Thus, we fail to reject H_0 meaning that we cannot conclude that married workers have meaningfully higher wages than unmarried workers.

Part c

```
plot(MARRIED, resid(wage_llr))
```



We can determine that there is evidence of heteroskedasticity since we can see that variance for married workers is greater than that of unmarried workers (in a systematic way).

Part d

Let us estimate the model from part (a) using observations from married workers:

```
WAGE = subset(data, married==1)$wage
EDUC = subset(data, married==1)$educ
EXPER = subset(data, married==1)$exper
EXPER_2 = EXPER^2

wage_llr_d1 = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER*EDUC))
summary(wage_llr_d1)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER * EDUC))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.37423 -0.34481  0.00957  0.34195  1.44652
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.9196961  0.3557963   2.585 0.009986 **
## EDUC         0.1008275  0.0221957   4.543 6.77e-06 ***
## EXPER        0.0506938  0.0149271   3.396 0.000731 ***
## EXPER_2     -0.0007088  0.0001379  -5.141 3.75e-07 ***
## EDUC:EXPER  -0.0004620  0.0007478  -0.618 0.536990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5353 on 576 degrees of freedom
## Multiple R-squared:  0.2109, Adjusted R-squared:  0.2054
## F-statistic: 38.48 on 4 and 576 DF,  p-value: < 2.2e-16
```

Let us now estimate the same model with observations from unmarried workers

```
WAGE = subset(data, married==0)$wage
EDUC = subset(data, married==0)$educ
EXPER = subset(data, married==0)$exper
EXPER_2 = EXPER^2

wage_llr_d2 = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER*EDUC))
summary(wage_llr_d2)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER * EDUC))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26823 -0.30372 -0.06065  0.29208  1.44465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1974877  0.2944715   0.671  0.50282
## EDUC         0.1512920  0.0194232   7.789 5.48e-14 ***
## EXPER        0.0728360  0.0127057   5.733 1.91e-08 ***
## EXPER_2     -0.0007014  0.0001193  -5.880 8.45e-09 ***
## EDUC:EXPER  -0.0021448  0.0006538  -3.280  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4614 on 414 degrees of freedom
## Multiple R-squared:  0.2753, Adjusted R-squared:  0.2683
## F-statistic: 39.32 on 4 and 414 DF,  p-value: < 2.2e-16
```

Let us now use the Goldfeld-Quandt test and a 5% significance level to test whether the error variances for married and unmarried workers are different:

$$H_0 : \sigma_M^2 = \sigma_{UM}^2 \text{ vs } H_1 : \sigma_M^2 \neq \sigma_{UM}^2$$

```
sigma(wage_llr_d1)^2
```

```
## [1] 0.2865846
```

```r
sigma(wage_llr_d2)^2
```

```
## [1] 0.2128502
```

$$\implies F = \frac{\hat{\sigma_{UM}}^2}{\hat{\sigma_M}^2} = \frac{0.213}{0.287} = 0.743 < 0.835 = F_c$$

Thus, we reject H_0 implying that the error variances for married and unmarried workers are not equal.

Part e

```r
library("nlme")
WAGE = data$wage
EDUC = data$educ
EXPER = data$exper
EXPER_2 = EXPER^2

w = rep(1:1000)
for (i in 1:1000) {
  if (data$married[i] == 1) {
    w[i] = 1/(sigma(wage_llr_d1)^2)
  }
  else {
    w[i] = 1/(sigma(wage_llr_d2)^2)
  }
}

glsm = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER*EDUC), weights = w)
summary(glsm)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EXPER * EDUC),
##     weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2360 -0.6559 -0.0528  0.6657  3.0876
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.780e-01  2.212e-01   2.161  0.03090 *
## EDUC         1.309e-01  1.441e-02   9.083  < 2e-16 ***
## EXPER        6.452e-02  9.324e-03   6.920 8.06e-12 ***
## EXPER_2     -7.128e-04  8.622e-05  -8.268 4.34e-16 ***
## EDUC:EXPER  -1.443e-03  4.842e-04  -2.979  0.00296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 995 degrees of freedom
## Multiple R-squared:  0.2495, Adjusted R-squared:  0.2465
## F-statistic: 82.71 on 4 and 995 DF,  p-value: < 2.2e-16
```

Looks like the coefficient estimates are very similar. The standard errors in the generalized least squares model slightly less than those from part (a).

Part f

Let us find two 95% interval estimates for a worker with 16 years of education and 10 years of experience for the given marginal effect with $t_c = 1.96$. Let us use part (a) first:

$$\frac{\partial E(ln(WAGE))}{\partial EDUC} = \beta_2 + EXPER \cdot \beta_5 = 0.127 - 0.00132 \cdot 10 = 0.114$$

```
library("sandwich")
a = lm(log(WAGE) ~ EDUC + EXPER + EXPER_2 + (EDUC*EXPER))
var_b2 = vcov(a)[2,2]
var_b5 = vcov(a)[5,5]
cov_b2_b5 = vcov(a)[2,5]
sqrt(var_b2 + 10^2*var_b5 + 2*10*cov_b2_b5)
```

```
## [1] 0.01038919
```

$$\implies 0.114 \pm t_c \cdot se = 0.114 \pm 1.96 \cdot 0.0104 = (0.0936, 0.134)$$

Now, let us use part (e) first:

$$\frac{\partial E(ln(WAGE))}{\partial EDUC} = \beta_2 + EXPER \cdot \beta_5 = 0.131 - 0.001443 \cdot 10 = 0.116$$

```
library("sandwich")
var_b2 = vcov(glsm)[2,2]
var_b5 = vcov(glsm)[5,5]
cov_b2_b5 = vcov(glsm)[2,5]
sqrt(var_b2 + 10^2*var_b5 + 2*10*cov_b2_b5)
```

```
## [1] 0.01019315
```

$$\implies 0.116 \pm t_c \cdot se = 0.116 \pm 1.96 \cdot 0.010193 = (0.0960, 0.136)$$

We can see that the interval estimates are very similar, but the interval from part (e) is just barely tighter than the interval from part (a) which makes sense.