Exercise set 2

Ti xi vXi q t ăm

63553645=

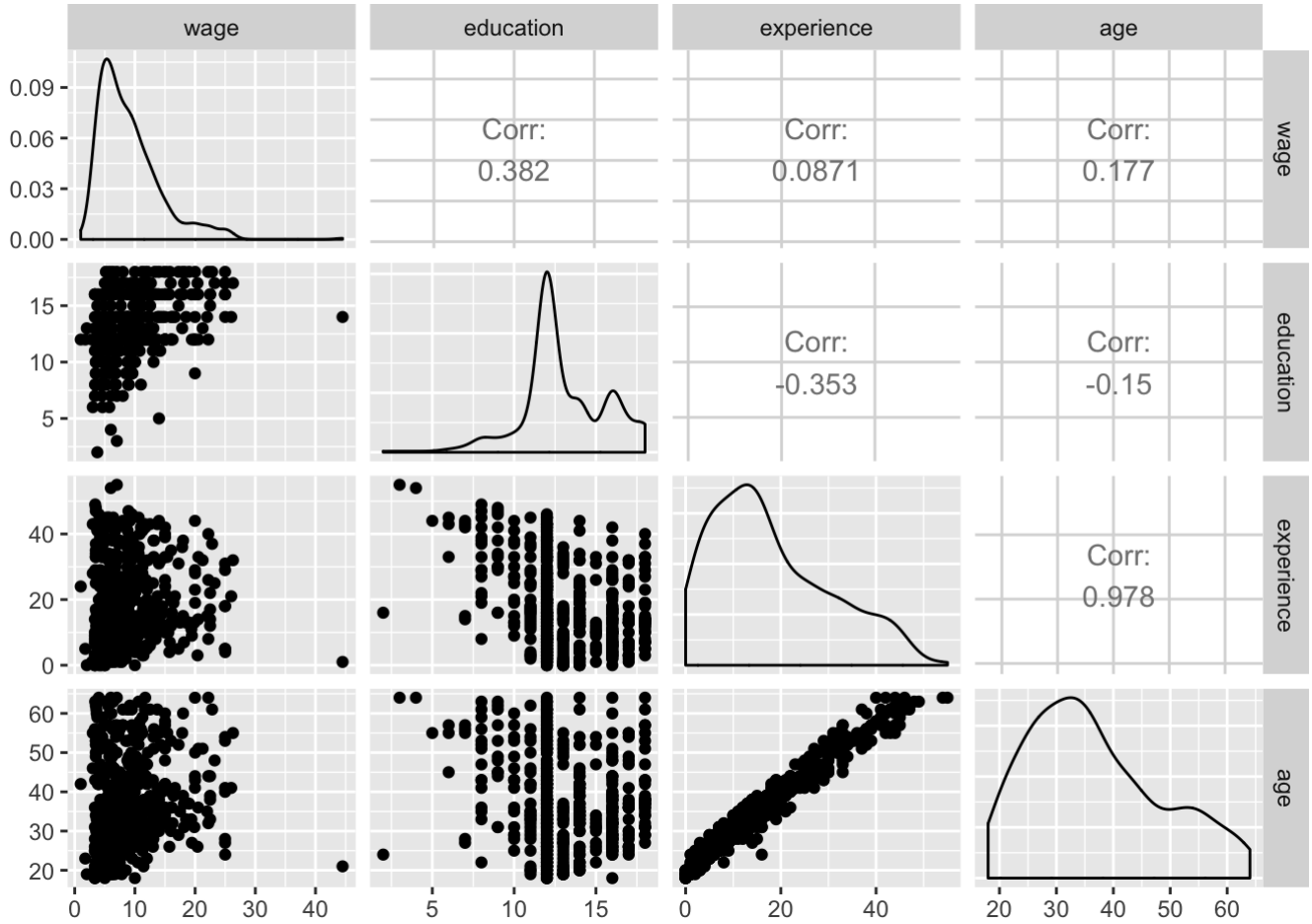
1. Simple regression with one regressor.

a, Summary of the data

```
summary(CPS1985)
```

```
##       wage           education       experience         age       
##  Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.00  
##  1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.00  
##  Median : 7.780   Median :12.00   Median :15.00   Median :35.00  
##  Mean   : 9.024   Mean   :13.02   Mean   :17.82   Mean   :36.83  
##  3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.00  
##  Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.00  
##     ethnicity     region       gender         occupation  
##  cauc    :440   south:156   male  :289   worker    :156  
##  hispanic: 27   other:378   female:245   technical :105  
##  other   : 67                            services  : 83  
##                                          office    : 97  
##                                          sales     : 38  
##                                          management: 55  
##            sector    union     married  
##  manufacturing: 99   no :438   no :184  
##  construction : 24   yes: 96   yes:350  
##  other        :411                      
##                                         
##                                         
##
```

```
ggpairs(CPS1985[c('wage','education','experience','age')])
```



b, A regression with wage as dependent variable and education as regressor

```
mod1 = lm(wage ~ education, data = CPS1985)
mod1
```

```
##
## Call:
## lm(formula = wage ~ education, data = CPS1985)
##
## Coefficients:
## (Intercept)    education
##     -0.7460       0.7505
```

c, What is the estimated average wage of an individual with 10 years of education according to the regression results?

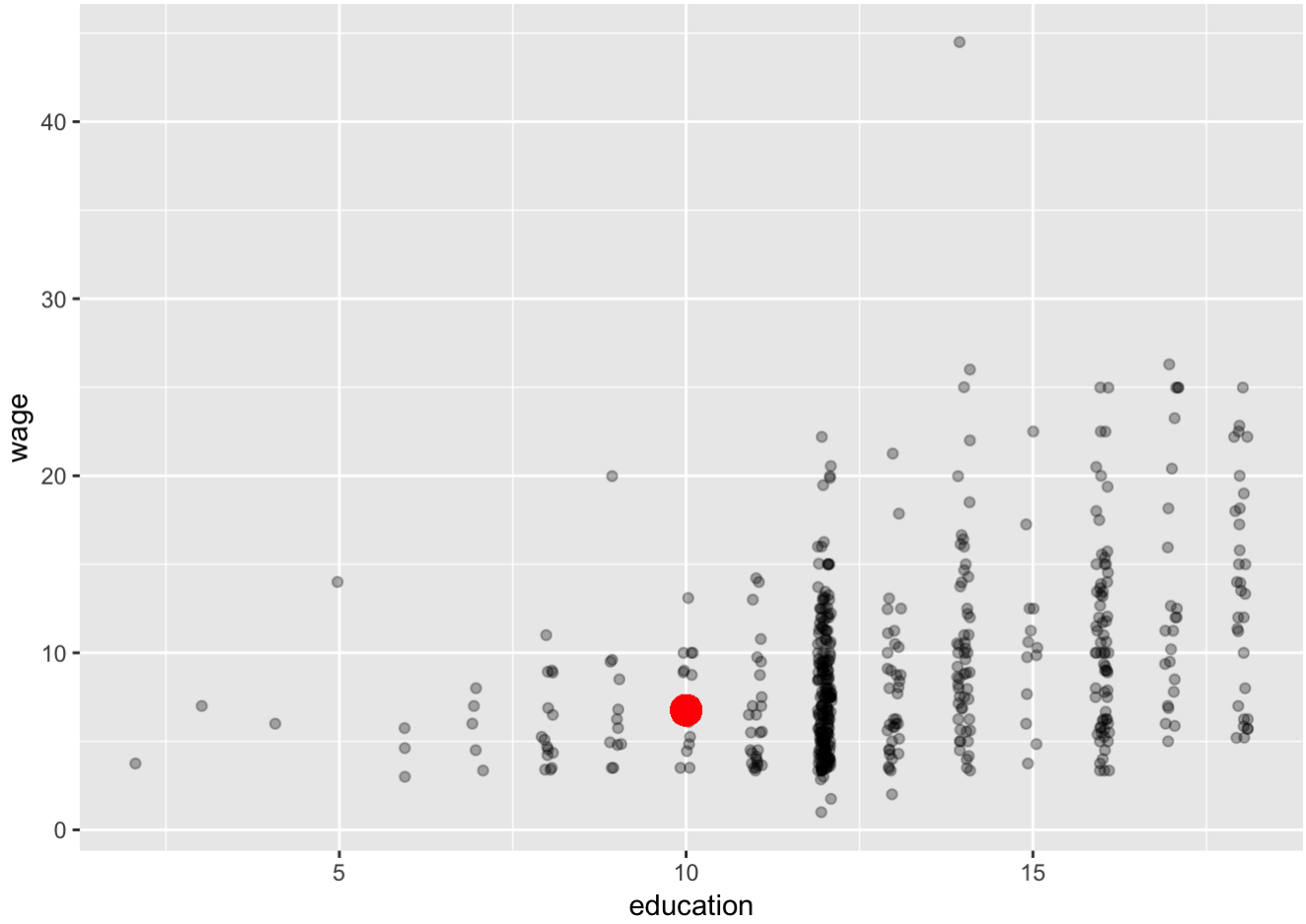
```
beta0 <- mod1$coefficients[1]
beta1 <- mod1$coefficients[2]

prediction_ten_years <- beta0 + beta1 * 10
prediction_ten_years
```

```
## (Intercept)
##    6.758628
```

Plotting the result:

```
ggplot(data = CPS1985, aes(x=education, y=wage)) +
  geom_jitter(alpha=0.3, width=0.1) +
  geom_point(aes(x=10, y = prediction_ten_years), color='red', size=5)
```



d, P-values and other information about the regression.

```
summary(mod1)
```

```
##
## Call:
## lm(formula = wage ~ education, data = CPS1985)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.911 -3.260 -0.760  2.240 34.740
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.74598    1.04545  -0.714    0.476
## education    0.75046    0.07873   9.532   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.754 on 532 degrees of freedom
## Multiple R-squared:  0.1459, Adjusted R-squared:  0.1443
## F-statistic: 90.85 on 1 and 532 DF,  p-value: < 2.2e-16
```

e, Explanation: beta1 = 0.75; so every year in education adds on everage 0.75$/hour; adjusted with beta0 (which is -0.74)

T-Testing:

H0: no signigicance, so P is about 1. We want to prove that education does NOT have effect to the wage.

This is a 2-tailed H-test, where L4 >Q,Lrkl i vl hygexrsr-1Q,Er} l hygexrsr-A4

H1: there IS significant evidence, so P is closer to 0. If it is true, it means education HAS effect on the wage.

We have a very low P-value (2.2e-16), so we can decline H0.

Simple regression with a dummy regressor. Continue to use the data set CPS1985.

a, Run a regression with wage as dependent variable and gender as regressor (X-variable).

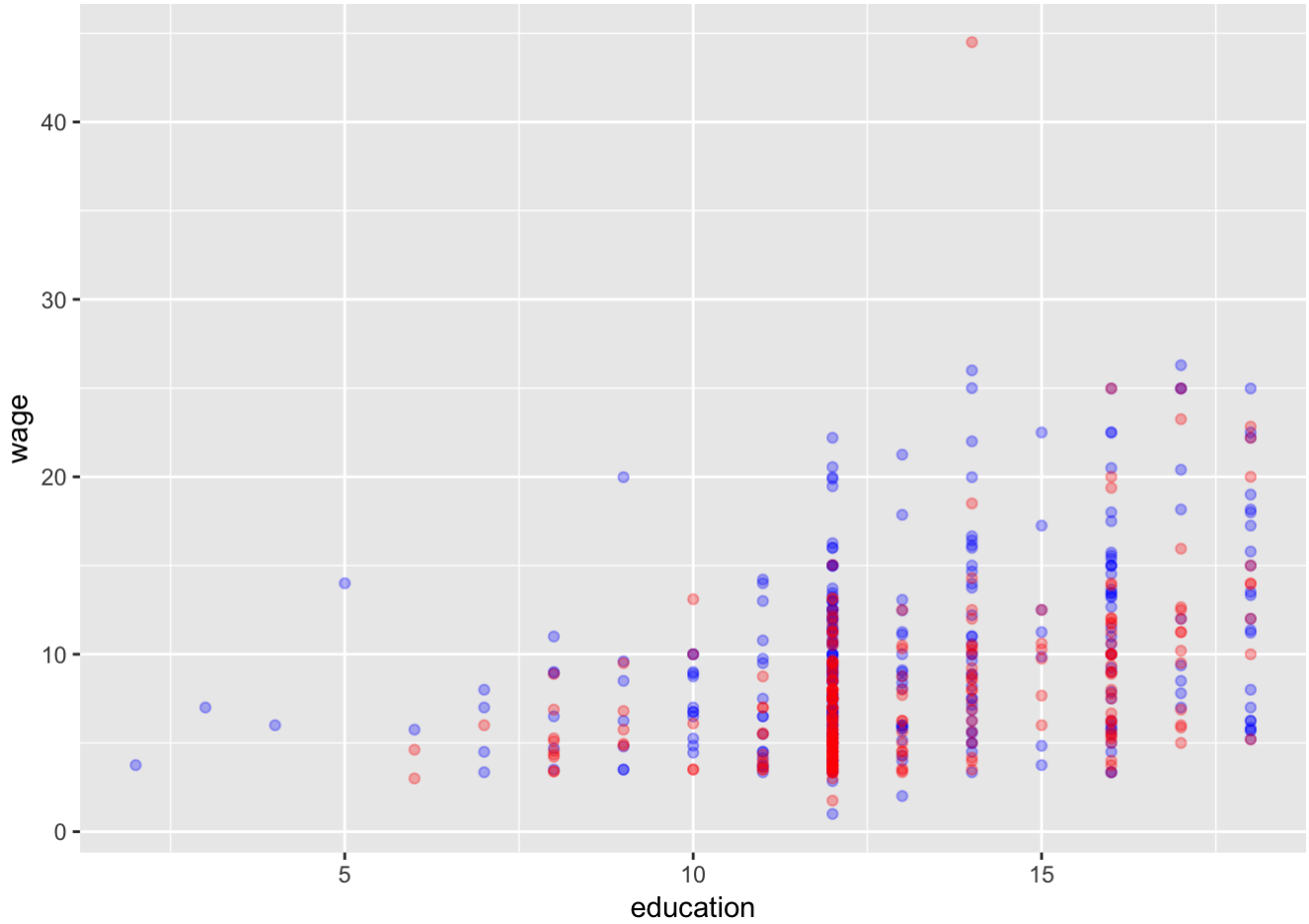
```
wageGenderMod <- lm(wage ~ gender, data=CPS1985)
summary(wageGenderMod)
```

```
##
## Call:
## lm(formula = wage ~ gender, data = CPS1985)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.995 -3.529 -1.072  2.394 36.621
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.9949     0.2961   33.75  < 2e-16 ***
## genderfemale  -2.1161     0.4372   -4.84  1.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.034 on 532 degrees of freedom
## Multiple R-squared:  0.04218,    Adjusted R-squared:  0.04038
## F-statistic: 23.43 on 1 and 532 DF,  p-value: 1.703e-06
```

b, According to the regression estimates, do women in the popula- tion earn less or more than men?

Yes. Male-female multiplier is -2.1161, and as ‘female’ value is the second in the factor (2), this means these records will yield lower values.

```
ggplot() +
  geom_point(data=subset(CPS1985, gender=='male'), aes(x=education, y=wage), color="b
lue", alpha=0.3) +
  geom_point(data=subset(CPS1985, gender=='female'), aes(x=education, y=wage), color=
"red", alpha=0.3)
```



c, Is the gender difference significant?

Yes. P = 1.703e-06, which is very slow. This is very high efidence.

b, Construct a (numerical) dummy variable coded 1 if female and 0 if male. Use this dummy in the regression instead of the factor variable gender. Do you get the same regression results?

```
CPS1985$gender_num <- ifelse(CPS1985$gender == 'female', 1, 0)
wageGenderMod <- lm(wage ~ gender_num, data=CPS1985)
summary(wageGenderMod)
```

```
##
## Call:
## lm(formula = wage ~ gender_num, data = CPS1985)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.995 -3.529 -1.072  2.394 36.621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.9949     0.2961   33.75  < 2e-16 ***
## gender_num   -2.1161     0.4372   -4.84  1.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.034 on 532 degrees of freedom
## Multiple R-squared:  0.04218,    Adjusted R-squared:  0.04038
## F-statistic: 23.43 on 1 and 532 DF,  p-value: 1.703e-06
```

Yes. Looks like LM uses also 0 and 1 when converting factorials.

e, It is possible to show that the sample average ($\bar{Y}$) is the least squares estimator of the population average. In the previous re- gression we have computed the least squares estimators of the averages of the populations for men and women. Compute the sample averages for both women and men. Compare the sample averages to the fitted values for men and women from the pre- vious regression. Are the predicted averages from the regression equal to the sample averages?

```
d_male <- subset(CPS1985, gender=="male")
d_female <- subset(CPS1985, gender=="female")


b0 = wageGenderMod$coefficients[1]
b1 = wageGenderMod$coefficients[2]

mean(CPS1985$wage)
```

```
## [1] 9.024064
```

```
mean( b0 + b1 * CPS1985$gender_num)
```

```
## [1] 9.024064
```

```
mean(d_male$wage)
```

```
## [1] 9.994913
```

```
mean( b0 + b1 * d_male$gender_num)
```

```
## [1] 9.994913
```

```
mean(d_female$wage)
```

```
## [1] 7.878857
```

```
mean( b0 + b1 * d_female$gender_num)
```

```
## [1] 7.878857
```

As you can see in the last 4 line, mean male population is the same as the predicted male population. Same for female population.