Team Based Learning

27/10/2026

Consider the dataset wage (included in the files `wageTBL.csv`, or `wageTBL.data`, or `wageTBL.txt`, or `wageTBL.RData`). The data come from the 1976 U.S. Current Population Survey (CPS) and widely used to estimate the economic returns to education and work experience (Mincer model). The variables of interest are:

- `wage`: average hourly wage (in dollars)
- `educ`: years of education
- `tenure`: years of tenure with the current employer
- `female`: gender (= 1 if female)
- `married`: marital status (= 1 if married)
- `area`: geographic indicator for regions of living (levels `north`, `south`, `west`, and `east`)
- `profocc`: occupational indicator (= 1 if in a professional occupation)
- `lwage`: logarithm of the wage `log(wage)`

1. Load the dataset into the environment and select only the variables of interest specified above.

```
data <- read.csv("wageTBL.csv")
# Equivalently
# data <- read.csv(file = "wageTBL.csv")
# data <- read.table(file = "wageTBL.data", header = T)
# data <- read.table(file = "wageTBL.txt", header = T)
# load(file = "wageTBL.RData")

vars <- c("wage", "educ", "tenure", "female", "married", "area", "profocc", "lwage")

data <- data[, vars]
```

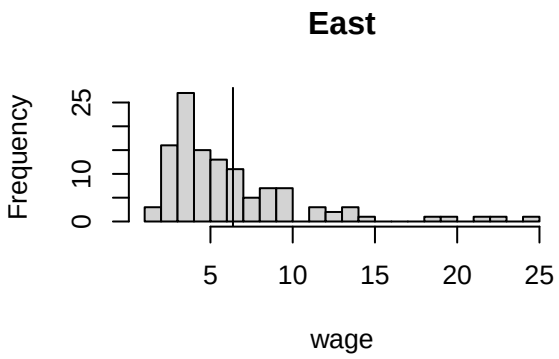
2. Compute the third quartile of `wage` for married individuals who have studied at least 10 years.

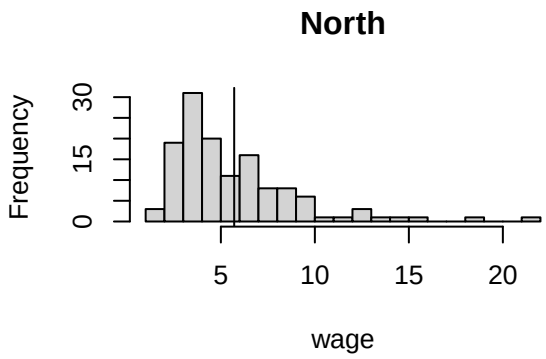
```
quantile(data[data$married == 1 & data$educ >= 10,]$wage, prob = 0.75)
```

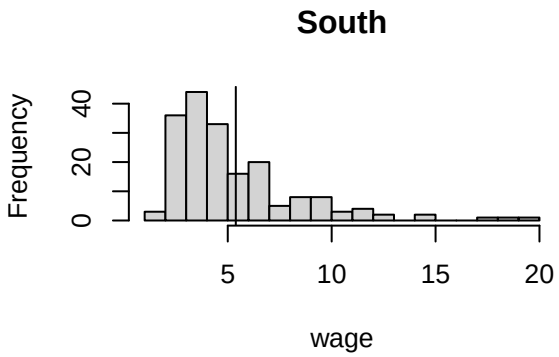
```
##   75%
## 8.605
```

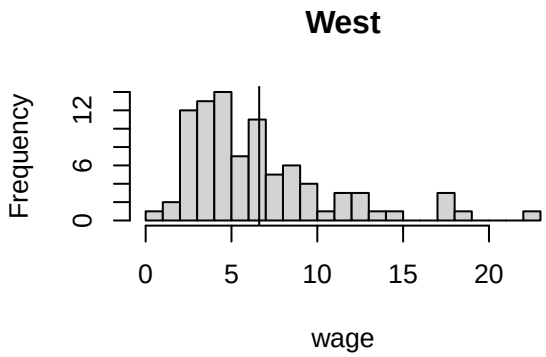
3. Split the graphics window into four parts. Then, represent the conditional distributions of `wage` with respect to the categories of the variable `area` using histograms. On each plot, indicate the respective conditional mean and put the name of the area in the title.

```
par(mfrow=c(2,2))
lev <- c("East", "North", "South", "West")
for(i in 1:length(lev)){
  hist(data$wage[data$area == lev[i]], main = lev[i], breaks = 20, xlab = "wage")
  abline(v = mean(data$wage[data$area == lev[i]]))
}
```







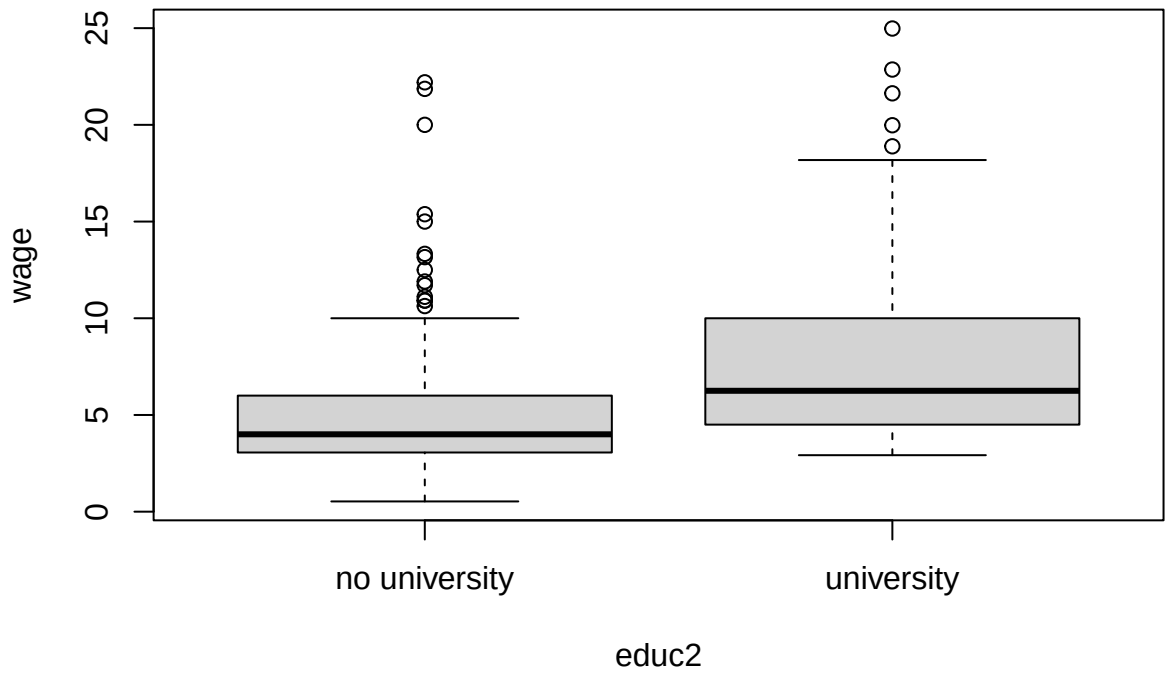



```
par(mfrow=c(1,1))
```

4. Build the variable `educ2` with two categories from `educ` according to the following rule: [0 - 13] for individuals who did not attend university, and (13,18] for those who started a university career. Explore the conditional distribution of `wage` given `educ2` using boxplot. In addition, we want to test whether the average wages of individuals who did not attend university differ from those of individuals who did.

```
data$educ2 <- factor(ifelse(data$educ <= 13, "no university", "university"))

boxplot(wage ~ educ2, data = data)
```




```
ttest <- t.test(data$wage[data$educ2 == "no university"],
                data$wage[data$educ2 == "university"],
                var.equal = T)
ttest
```

```
##
##  Two Sample t-test
##
## data:  data$wage[data$educ2 == "no university"] and data$wage[data$educ2 == "university"]
## t = -8.8521, df = 524, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.461093 -2.203887
## sample estimates:
## mean of x mean of y
##  4.964504  7.796994
```

5. Fit a multiple linear regression model with the logarithm of income `lwage` as the response variable and `educ`, `female`, and `tenure` as explanatory variables. Print the model output.

```
mod1 <- lm(lwage ~ educ + female + tenure, data = data)
summary(mod1)
```

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

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96883 -0.25262 -0.03383  0.24687  1.29983
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.633125   0.091382   6.928 1.26e-11 ***
## educ         0.081354   0.006643  12.246  < 2e-16 ***
## female      -0.297052   0.037470  -7.928 1.36e-14 ***
## tenure       0.021634   0.002588   8.359 5.78e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4188 on 522 degrees of freedom
## Multiple R-squared:  0.3828, Adjusted R-squared:  0.3793
## F-statistic: 107.9 on 3 and 522 DF,  p-value: < 2.2e-16
```

5.1 Which of the following statements is TRUE?

a. [ ] The average income for a man with 0 years of education and just hired (`tenure` = 0) is 0.63 dollars per hour
b. [x] Holding the other variables constant, women earn on average 25% less than men
c. [ ] The variability explained by the model is about 62%

```
(exp(predict(mod1, newdata = data.frame(educ = 5, female = 1, tenure = 0)))/exp(predict(mod1, new
```

```
##         1
## -25.69946
```

```
(exp(coef(mod1)[3])-1)*100
```

```
##    female
## -25.69946
```

The linear regression model is

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 female_i + \beta_3 tenure_i + \varepsilon_i$$

$$\widehat{\log(wage)}_{\text{woman}} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 + \hat{\beta}_3 tenure$$
$$\widehat{\log(wage)}_{\text{man}} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_3 tenure$$

$$\log(\widehat{wage}_{\text{woman}}) - \log(\widehat{wage}_{\text{man}}) = \hat{\beta}_2 \quad \Longrightarrow \quad \frac{\widehat{wage}_{\text{woman}}}{\widehat{wage}_{\text{man}}} = e^{\hat{\beta}_2}$$

$$\frac{\widehat{wage}_{\text{woman}} - \widehat{wage}_{\text{man}}}{\widehat{wage}_{\text{man}}} = (e^{\hat{\beta}_2} - 1) \times 100$$

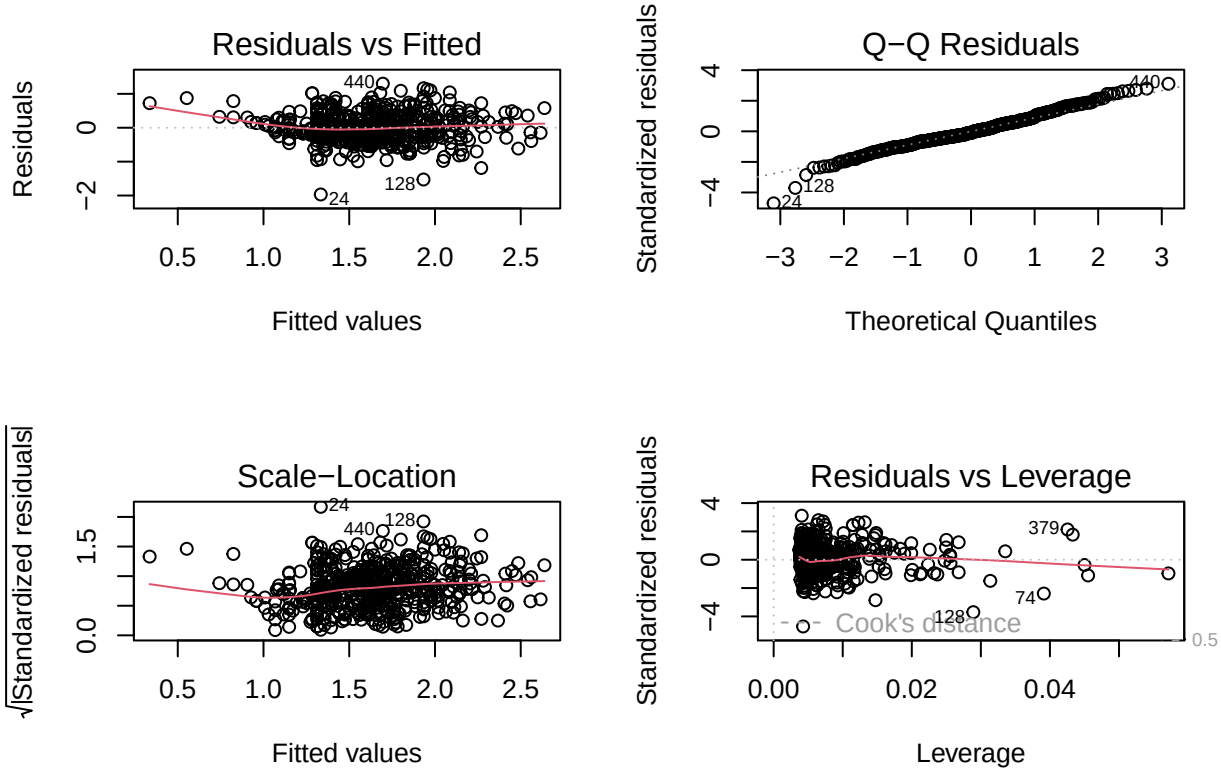
By assuming the normality of errors, the result can be interpreted also in terms of conditional expectation.

5.2 Plot the residuals and select the TRUE answer:

a. [ ] The distribution of the residuals does not approximate a normal distribution well

b. [ ] There is evidence of two distinct groups
c. [x] Heteroskedasticity can be suspected

```
par(mfrow=c(2,2))
plot(mod1)
```



5.3 Add the interaction between `female` and `tenure` to the model estimated in point 6. Which of the following statements is TRUE?

a. [ ] The interaction between `tenure` and `female` is definitely significant because the two coefficients are significant
b. [x] Average income (on the log scale) increases as tenure increases, but more slowly for women
c. [ ] After ten years of work (`tenure` = 10), a woman earns the same wage as a man

```
mod2 <- lm(lwage ~ educ + female*tenure, data = data)
summary(mod2)
```

```
##
## Call:
## lm(formula = lwage ~ educ + female * tenure, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9918 -0.2510 -0.0331  0.2429  1.3086
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    0.603353   0.092165   6.546 1.41e-10 ***
## educ           0.081982   0.006628  12.369  < 2e-16 ***
## female        -0.243048   0.045257  -5.370 1.19e-07 ***
## tenure         0.024991   0.003030   8.248 1.32e-15 ***
## female:tenure -0.012201   0.005775  -2.113   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4174 on 521 degrees of freedom
## Multiple R-squared:  0.3881, Adjusted R-squared:  0.3834
## F-statistic:  82.6 on 4 and 521 DF,  p-value: < 2.2e-16
```