Instituto Superior Técnico | Departamento de Matemática

Linear Model Analysis (LMAC/MECD/MMAC)

Exam (resit) — 1st semester – 2023/24

Duration: 2 hours — 31/01/24 – 15h30m

Please explain your answers appropriately!

A researcher collected the following data to investigate factors affecting wages in the United States (*wage*: salary in dollars per hour, *city*: city size categorized as small/med/big, *educ*: education in years, *exper*: job experience in years, *female*: 1 for female/0 for male).

Data:

```
     wage  city educ exper female
1    4.28 small   12    28      1
2    7.96 small   10    35      1
3   11.42 small   16    27      0
4    8.76   big   16    12      0
5    5.00   med   16     5      1
................................
486  6.73 small   10    34      0
487  5.31 small   17     3      0
488 10.58 small   16    19      0
489  4.17   med   10    13      0
```

Summaries:

```
      wage           city          educ          exper          female
 Min.   : 1.160   big  :102   Min.   : 5.00   Min.   : 0.00   Min.   :0.0000
 1st Qu.: 3.750   med  :153   1st Qu.:12.00   1st Qu.: 7.00   1st Qu.:0.0000
 Median : 5.330   small:234   Median :12.00   Median :14.00   Median :0.0000
 Mean   : 6.569               Mean   :12.63   Mean   :17.03   Mean   :0.2843
 3rd Qu.: 7.960               3rd Qu.:13.00   3rd Qu.:25.00   3rd Qu.:1.0000
 Max.   :38.860               Max.   :17.00   Max.   :48.00   Max.   :1.0000
```

She then considered several regression models with response variable $y = \log(wage)$ and candidate explanatory variables:

$$x_1 = \begin{cases} educ - 12, & \text{if } educ > 12 \\ 0, & \text{if } educ \leq 12 \end{cases}, \; x_2 = exper - 14, \; x_3 = female$$

$$x_4 = \begin{cases} 1, & \text{if } city = \text{small} \\ 0, & \text{otherwise} \end{cases}, \; x_5 = \begin{cases} 1, & \text{if } city = \text{big} \\ 0, & \text{otherwise} \end{cases}, \; x_6 = \begin{cases} 1, & \text{if } city = \text{med} \\ 0, & \text{otherwise} \end{cases}.$$

The models were fit in R and the outputs shown in Figure 1 were obtained.

1. The researcher concluded that the best model is Model 2. Do you agree? Can you find a reason to justify the researcher's choice? (2.0)

2. Present the ANOVA table for Model 2 and conclude whether or not the regression is significant. Use $\alpha = 0.05$. (2.0)

3. Write down the equation for Model 2 ($y_i = \cdots\cdots$). Describe all its components and assumptions. (2.0)

4. Present the fitted regression equation for Model 2 and interpret each of the estimated regression coefficients. (2.0)

5. (a) Why is x_6 not included in any model? (1.0)

(b) What is the name of a model in which all explanatory variables are categorical? (1.0)

6. The researcher used the predict function in R and Model 2 to produce the following table.

city	educ	exper	female	$\hat{\mu}_{y\|\mathbf{x}_0}$	$s.e.\left(\hat{\mu}_{y\|\mathbf{x}_0}\right)$
med	12	14	0	1.66	0.0435
med	12	14	1	1.23	0.0525

Assuming Model 2, obtain joint confidence intervals for the average log-wages of people with the characteristics listed. Use the most efficient simultaneous procedure and 90% global confidence. Interpret the intervals. How would prediction intervals compare with the ones you have computed? (2.0)

7. The main conclusion from this study was that it provides strong evidence of gender discrimination. What type of discrimination is at stake and what might this evidence be? (2.0)

8. The researcher also produced the diagnostic plots for Model 2 shown in Figure 2. What can you conclude from each plot? (2.0)

9. Using the additional information that $SSPE = 78.707$, with 342 degrees of freedom, test Model 1 for lack-of-fit. (2.0)

10. How would you validate this study and its conclusions? (2.0)

Figure 1: R outputs.

Model 0:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49454    0.02775  53.850   <2e-16 ***
x1           0.12994    0.01424   9.125   <2e-16 ***
x2           0.02306    0.00191  12.076   <2e-16 ***
---
Residual standard error: 0.497 on 486 degrees of freedom
Multiple R-squared:  0.3096,    Adjusted R-squared:  0.3068
F-statistic:   109 on 2 and 486 DF,  p-value: < 2.2e-16
```

Model 1:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.6020774  0.0326565  49.059  < 2e-16 ***
x1           0.1131246  0.0141700   7.983 1.04e-14 ***
x2           0.0320200  0.0030809  10.393  < 2e-16 ***
I(x2^2)     -0.0008417  0.0001582  -5.320 1.59e-07 ***
I(x1 * x2)   0.0034817  0.0012808   2.718  0.00679 **
---
Residual standard error: 0.4802 on 484 degrees of freedom
Multiple R-squared:  0.3582,    Adjusted R-squared:  0.3529
F-statistic: 67.54 on 4 and 484 DF,  p-value: < 2.2e-16
```

Model 2:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.6570710  0.0434919  38.101  < 2e-16 ***
x1           0.0864054  0.0132115   6.540 1.57e-10 ***
x2           0.0255946  0.0029002   8.825  < 2e-16 ***
I(x2^2)     -0.0007182  0.0001449  -4.957 9.94e-07 ***
I(x1 * x2)   0.0029334  0.0011756   2.495   0.0129 *
x3          -0.4310024  0.0464595  -9.277  < 2e-16 ***
x4           0.0954291  0.0457031   2.088   0.0373 *
x5           0.2503456  0.0574415   4.358 1.60e-05 ***
---
Residual standard error: 0.4364 on 481 degrees of freedom
Multiple R-squared:  0.4734,    Adjusted R-squared:  0.4657
F-statistic: 61.77 on 7 and 481 DF,  p-value: < 2.2e-16
```

Model 3:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.6387642  0.0441417  37.125  < 2e-16 ***
x1           0.1061541  0.0168667   6.294 7.01e-10 ***
x2           0.0279461  0.0033109   8.441 3.77e-16 ***
I(x2^2)     -0.0007689  0.0001491  -5.156 3.70e-07 ***
I(x1 * x2)   0.0027483  0.0011821   2.325  0.02050 *
x3          -0.4389382  0.0464393  -9.452  < 2e-16 ***
x4           0.1404270  0.0511457   2.746  0.00627 **
x5           0.2356655  0.0579780   4.065 5.62e-05 ***
I(x2 * x3)  -0.0063927  0.0043875  -1.457  0.14577
I(x1 * x4)  -0.0480148  0.0258922  -1.854  0.06429 .
---
Residual standard error: 0.4348 on 479 degrees of freedom
Multiple R-squared:  0.4792,    Adjusted R-squared:  0.4694
F-statistic: 48.98 on 9 and 479 DF,  p-value: < 2.2e-16
```

Model 4:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7136901  0.0341222  50.222  < 2e-16 ***
x1           0.0869895  0.0132545   6.563 1.36e-10 ***
x2           0.0253457  0.0029078   8.716  < 2e-16 ***
I(x2^2)     -0.0007203  0.0001454  -4.954 1.01e-06 ***
I(x1 * x2)   0.0031725  0.0011741   2.702 0.007135 **
x3          -0.4264969  0.0465708  -9.158  < 2e-16 ***
x5           0.1913489  0.0501859   3.813 0.000155 ***
---
Residual standard error: 0.4379 on 482 degrees of freedom
Multiple R-squared:  0.4686,    Adjusted R-squared:  0.462
F-statistic: 70.84 on 6 and 482 DF,  p-value: < 2.2e-16
```

Some useful sums:

```
  sum(y)  sum(Y^2)
831.5573  1587.997
```

Figure 2: Diagnostic plots for Model 2.