Ma 3b Practical – Recitation 8

March 14, 2025

Derive estimators for coefficients a and b, test hypothesis that a=0 or b=0, logistics model.

Exercise 1. Consider the maximum temperatures example. We found that

$$\hat{a} = -2.47 \text{ and } \hat{b} = 0.289,$$

with the following data :

$$\begin{aligned} &SS_R = 8.341676, && S_{xx} = 2480, && S_{yy} = 214.7437, \\ &S_{xy} = 715.46, && \bar{x} = 16, && \bar{y} = 2.146452. \end{aligned}$$

In this case, the coefficient of determination satisfies

$$R^2_{x,y} = 0.98^2 = 96.04\%.$$

The regression line explains 96% of the variation between the y_i 's. How would you test the hypothesis:

$$H_0 : b = 0 \quad \text{versus} \quad H_1 : b \neq 0.$$

Exercise 2. Consider again the maximum temperatures example. We found that

$$\hat{a} = -2.47 \quad \text{and} \quad \hat{b} = 0.289,$$

with the following data :

$$\begin{aligned} &SS_R = 8.341676, && S_{xx} = 2480, && S_{yy} = 214.7437, \\ &S_{xy} = 715.46, && \bar{x} = 16, && \bar{y} = 2.146452. \end{aligned}$$

We now test the hypothesis

$$H_0 : a = 0 \quad \text{versus} \quad H_1 : a \neq 0$$

Exercise 3. Spam filters are built on principles similar to those used in logistic regression. We fit a probability that each message is spam or not spam. We have several email variables for this problem: to multiple, cc, attach, dollar, winner, inherit, password, format, re subj, exclaim subj, and sent email. We won't describe what each variable means here for the sake of brevity, but each is either a numerical or indicator variable.

	Estimate	Std. Error	z value	Pr(> |z|)
Intercept)	-0.8124	0.0870	-9.34	0.0000
to multiple	-2.6351	0.3036	-8.68	0.0000
winner	1.6272	0.3185	5.11	0.0000
format	-1.5881	0.1196	-13.28	0.0000
re subj	-3.0467	0.3625	-8.40	0.0000

(a) Write down the model using the coefficients from the model fit. (b) Suppose we have an observation where to multiple $= 0$, winner $= 1$, format $= 0$, and re subj $= 0$. What is the predicted probability that this message is spam?

Solution. Under H_0(i.e. when $b = 0$), we have

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{b}-b) = \sqrt{\frac{(31-2)2480}{8.3417}}(\hat{b}-0) \sim t_{n-2}.$$

At a significance level of $\alpha = 5\%$, we find the critical value $t_{\alpha/2,n-2} = 2.045$. For an estimate $\hat{b}$ equal to 0.289, the test statistic, when $b = 0$, is

$$\sqrt{\frac{(31-2)2480}{8.3417}}(0.289-0) = 26.83466.$$

Since $26.8 \gg 2.045$, we reject H_0 and conclude that there is a non-zero slope in the linear model (no matter what the significance level is). A 95% confidence interval for $\hat{b}$ is

$$-t_{n-2,\alpha/2} \leqslant \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{b}-b) \leqslant t_{\alpha/2,n-2}.$$

This implies that

$$\hat{b} - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{\alpha/2,n-2} \leqslant b \leqslant \hat{b} + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{\alpha/2,n-2},$$

and we find the interval

$$[0.267; 0.311].$$

Solution. With the above data, we deduce that

$$\sum_{i=1}^{n} x_i^2 = S_{xx} + n\bar{x}^2 = 2480 + 31 \cdot (16)^2 = 10416.$$

At the significance level $\alpha = 5\%$, we find the critical value $t_{0.025,29} = 2.045$. For an estimate $\hat{a}$ equal to -2.47, the test statistic, when $a = 0$, is

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_{i=1}^{n} x_i^2}}(\hat{a}-0) = \sqrt{\frac{31(31-2)\cdot 2480}{8.3417(10416)}} \cdot (-2.47) = -12.51195.$$

Since $-12.5 \ll -2.045$, we reject H_0 and conclude that there is a non-zero y-intercept in the linear model (no matter what the significance level is). A 90% confidence interval for a is

$$-2.47 \pm t_{0.05,29}\sqrt{\frac{8.3417(10416)}{31(31-2)\cdot 2480}}.$$

where $t_{0.05,29} = 1.699$, so we find the interval [-2.805427;-2.134573].

Solution.

Solution to Logistic Regression Problem

Step 1: Write Down the Logistic Regression Model

A logistic regression model follows the form:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where:

- p is the probability that the message is spam.
- β_0 is the intercept.
- β_i are the estimated coefficients for each variable.
- x_i are the values of the variables for a given observation.

From the table, the estimated model is:

$$\log\left(\frac{p}{1-p}\right) = -0.8124 - 2.6351(\text{to multiple}) + 1.6272(\text{winner}) - 1.5881(\text{format}) - 3.0467(\text{re subj})$$

Step 2: Substitute the Given Values

We are given:

$$\text{to multiple} = 0, \quad \text{winner} = 1, \quad \text{format} = 0, \quad \text{re subj} = 0$$

Substituting these into the model:

$$\log\left(\frac{p}{1-p}\right) = -0.8124 - 2.6351(0) + 1.6272(1) - 1.5881(0) - 3.0467(0)$$

$$\log\left(\frac{p}{1-p}\right) = -0.8124 + 1.6272$$

$$\log\left(\frac{p}{1-p}\right) = 0.8148$$

Step 3: Convert Log-Odds to Probability

The logistic function is:

$$p = \frac{e^{\log\left(\frac{p}{1-p}\right)}}{1 + e^{\log\left(\frac{p}{1-p}\right)}}$$

Substituting:

$$p = \frac{e^{0.8148}}{1 + e^{0.8148}}$$

Computing the exponent:

$$e^{0.8148} \approx 2.2588$$

$$p = \frac{2.2588}{1 + 2.2588}$$

$$p = \frac{2.2588}{3.2588} \approx 0.6932$$