

## Assignment 7

1. Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2). \quad (1)$$

Assuming that there are replicate observations, a lack-of-fit test can be performed based on comparison with the model having  $\mathbf{X}$  replaced by  $\widetilde{\mathbf{X}}$ , the matrix for the (cell) means parameterization of the one-way model in which the levels of the class variable are the replicate groups (i.e. groups of observations having identical rows in  $\mathbf{X}$ ). Recall that the column space of  $\widetilde{\mathbf{X}}$  contains the column space of  $\mathbf{X}$ .

Let  $\mathbf{P}$  and  $\widetilde{\mathbf{P}}$  be the projections onto the column spaces of  $\mathbf{X}$  and  $\widetilde{\mathbf{X}}$ , respectively. Also, let  $n$  be the total number of observations and  $c$  be the number of replicate groups.

Suppose  $\mathbf{X}$  has rank  $p'$ , where  $p' < c$ .

- Show that  $\widetilde{\mathbf{P}} - \mathbf{P}$  is a projection. Also, find the corresponding degrees of freedom. (Hint: First explain why each column of  $\mathbf{P}$  is in the column space of  $\widetilde{\mathbf{P}}$ .)
- Write  $\text{SS(LF)}$ , the sum of squares for lack of fit, as a quadratic form in  $\mathbf{Y}$ , using  $\mathbf{P}$  and  $\widetilde{\mathbf{P}}$ . (Recall that  $\text{SS(LF)}$  is the difference in residual sums of squares between the model using  $\mathbf{X}$  and the model using  $\widetilde{\mathbf{X}}$ .)
- Using the expression from part (b), show that

$$\text{SS(LF)} = \sum_{\ell=1}^c n_{\ell} (\bar{Y}_{\ell} - \widehat{Y}_{(\ell)})^2$$

where  $n_{\ell}$  is the number of observations in replicate group  $\ell$ ,  $\bar{Y}_{\ell}$  is the average of the dependent variable values for the observations in group  $\ell$ , and  $\widehat{Y}_{(\ell)}$  is the least squares predicted value, based on model (1), that is common to all observations in group  $\ell$ . (Hint:  $\widetilde{\mathbf{P}} - \mathbf{P} = (\widetilde{\mathbf{P}} - \mathbf{P})'(\widetilde{\mathbf{P}} - \mathbf{P})$ . Why?)

- Assuming the error vector  $\boldsymbol{\epsilon}$  satisfies the conditions stated in (1) (but *not* assuming that model (1) is correct), use the result of part (b) to find an expression for  $E(\text{SS(LF)})$  in terms of  $\mathbf{P}$ ,  $\widetilde{\mathbf{P}}$ ,  $\boldsymbol{\mu} = E(\mathbf{Y})$ , and any scalars you may need. *Do not make any assumptions about  $\boldsymbol{\mu}$  yet.*
- Recall that  $\text{MS(LF)}$  is  $\text{SS(LF)}$  divided by the degrees of freedom for lack of fit. Use the expression from part (d) to show the following:
  - If  $\boldsymbol{\mu}$  is in the column space of  $\mathbf{X}$ , then  $E(\text{MS(LF)}) = \sigma^2$ .
  - If  $\boldsymbol{\mu}$  is in the column space of  $\widetilde{\mathbf{X}}$ , but *not* in the column space of  $\mathbf{X}$ , then  $E(\text{MS(LF)}) > \sigma^2$ .

2. An experiment is conducted to examine the effect of temperature on the maturation rate of a certain species of insect. To obtain one observation, an experimenter grows a collection of newly-hatched insects at a specified constant temperature and defines the response to be the average number of days that the insects in this collection require to become mature. (All

collections have approximately the same number of insects.) The temperatures (°C) used in the experiment are 10, 15, 20, 25, 28, 30, and 32. There are two replicate observations at each temperature. Assume the data are as listed after your name below, where  $y_{tt\_r}$  labels the response from the  $r^{\text{th}}$  replicate at temperature  $tt$ .

name	y10_1	y10_2	y15_1	y15_2	y20_1	y20_2	y25_1	y25_2
BAI LEI	22.6	23.3	15.0	14.3	14.7	12.6	8.5	7.8
BROWN STEPHEN V	23.0	21.9	15.0	13.0	11.7	14.2	8.3	7.8
CHANG CHE-SHUN	21.0	21.3	15.2	12.9	12.4	13.2	9.8	8.8
CHEN OU	24.8	23.5	14.4	16.5	11.9	13.1	8.3	11.6
DONOHOE MICHAEL	23.4	22.7	14.6	14.7	10.7	14.5	9.1	7.2
GAO HAIBING	21.9	21.2	17.5	16.2	14.5	11.7	8.2	9.4
GARG DIVYA	25.5	23.6	14.1	16.7	13.5	13.1	8.8	10.2
GLUCK MATHEW R	22.6	22.0	13.9	14.5	14.9	12.9	9.0	7.3
GORDON ROBERT F	19.2	22.1	14.5	13.8	12.4	13.9	9.3	9.4
GUCI LEDIA	23.7	23.6	16.3	17.9	12.2	13.3	7.7	6.6
HARIHARAN POOJA	22.0	21.8	15.3	16.6	11.6	14.0	7.7	9.4
HUANG LEI	23.4	23.1	17.3	13.1	13.7	11.7	10.4	8.7
KIM CHANMIN	22.7	25.2	15.4	15.3	12.3	10.0	10.4	10.5
KIRPICH ALEX	23.0	22.8	14.5	14.4	13.4	11.9	10.9	9.0
LEARY EMILY V	23.8	22.3	16.5	18.0	13.2	13.5	9.5	10.7
LI KE	22.7	25.4	15.7	17.2	11.7	14.8	6.8	7.2
LIN TONG	23.6	21.5	12.2	14.6	13.9	12.1	10.0	5.8
LIU MINZHAO	26.7	22.3	18.7	16.6	14.6	13.0	9.5	7.4
LU CUIE	21.7	24.5	12.3	16.8	12.9	12.0	8.1	9.4
LUO XUAN	23.8	21.5	15.7	14.8	14.7	12.6	6.7	8.4
MA LU	21.9	24.1	14.6	14.8	12.2	14.6	10.6	8.3
MALLICK PRANJAL	20.1	21.1	18.3	15.6	13.7	13.3	10.1	9.5
MARCUS GABRIEL	22.6	24.3	12.7	14.4	13.4	11.9	8.5	8.5
NAMKOONG YOUNG	23.2	21.7	14.7	15.8	13.0	13.3	7.3	7.8
NEAL DANIEL W	21.7	21.0	13.7	15.1	13.9	11.7	9.1	6.4
PETTERSON SONIA	22.1	22.6	14.0	16.3	12.5	13.0	8.0	8.7
PRANO BRIJIDA A	24.0	20.6	16.9	15.9	14.8	15.3	7.7	12.6
SHAO ANQI	21.7	22.9	17.0	17.7	15.8	15.1	8.8	11.4
SINHA AMIT	21.3	23.1	17.1	17.5	13.7	16.2	7.6	7.0
THAYER LAURA K	23.4	21.9	14.6	15.2	15.0	12.4	11.2	10.3
YE RONGZHONG	21.9	23.1	14.3	17.1	13.0	15.6	7.0	9.1
ZHOU ZHUO	25.0	20.6	16.4	16.8	14.9	15.4	7.8	8.6
ZHU XIAOYU	23.4	24.6	15.8	16.6	13.5	13.0	8.7	8.2
demo	20.0	23.6	16.2	16.6	13.4	13.0	5.4	8.7

CONTINUED ...

name	y28_1	y28_2	y30_1	y30_2	y32_1	y32_2
BAI LEI	8.5	6.2	8.5	6.8	9.3	7.2
BROWN STEPHEN V	8.5	4.9	8.2	6.2	6.1	6.8
CHANG CHE-SHUN	7.2	8.8	4.9	7.6	9.1	7.3
CHEN OU	6.2	6.7	8.3	8.8	9.7	7.8
DONOHOE MICHAEL	6.1	8.3	8.6	7.4	7.0	9.8
GAO HAIBING	8.0	9.4	8.7	9.3	9.7	9.6
GARG DIVYA	3.2	7.8	7.0	8.1	8.3	7.4
GLUCK MATHEW R	6.4	7.9	10.1	6.2	6.9	8.1

GORDON ROBERT F	7.7	5.7	6.5	7.4	7.1	11.1
GUCI LEDIA	10.2	6.0	4.3	6.6	7.4	10.3
HARIHARAN POOJA	6.9	6.6	7.0	4.2	7.3	7.5
HUANG LEI	5.0	7.6	9.5	5.8	9.5	7.6
KIM CHANMIN	8.6	7.2	8.9	7.3	6.9	8.5
KIRPICH ALEX	8.6	8.3	5.9	8.7	9.4	6.5
LEARY EMILY V	6.5	8.3	8.7	5.9	8.7	7.8
LI KE	6.3	6.2	7.9	7.2	9.1	9.4
LIN TONG	8.4	7.6	5.8	6.8	5.4	8.4
LIU MINZHAO	5.8	7.5	8.7	8.0	8.3	8.7
LU CUIE	6.9	7.4	8.2	7.4	9.2	10.4
LUO XUAN	6.6	5.1	6.2	7.6	7.5	7.8
MA LU	7.0	9.3	7.3	7.2	6.6	8.8
MALLICK PRANJAL	7.8	7.3	6.6	7.8	8.4	9.5
MARCUS GABRIEL	7.4	8.7	6.8	7.5	11.2	7.0
NAMKOONG YOUNG	7.3	10.3	5.8	8.3	9.1	7.8
NEAL DANIEL W	8.5	7.1	7.0	8.5	6.6	8.9
PETTERSON SONIA	6.5	6.4	9.0	9.4	8.9	9.6
PRANO BRIJIDA A	9.7	7.5	7.5	6.8	8.5	7.3
SHAO ANQI	7.8	6.2	6.8	7.3	4.9	6.0
SINHA AMIT	10.3	8.8	8.8	7.9	9.4	10.7
THAYER LAURA K	8.6	6.1	5.4	8.7	8.1	6.3
YE RONGZHONG	6.3	8.5	5.6	7.3	4.5	7.0
ZHOU ZHUO	8.2	9.2	5.7	7.4	7.7	8.0
ZHU XIAOYU	8.1	10.5	10.4	10.1	9.1	8.1
demo	10.2	6.3	7.8	9.2	10.2	6.7

(You may use SAS® or R to perform any required computations.)

- Fit simple linear, quadratic, and cubic regression models using least squares, under the usual error assumptions. In each case, write out the estimated regression equation and produce a plot of the data and the fitted curve.
- Assuming that the quadratic model is correct, estimate the mean number of days required for an insect to become mature if the temperature is held constant at 17°C. Give the corresponding 95% two-sided confidence interval.
- Fit the highest degree of polynomial that has a unique least squares fit for these data. Give a basic ANOVA table for this model. (Note: You may need to use orthogonal polynomials to avoid numerical instability.)
- Compute the  $C_p$ , AIC, and SBC statistics for the simple linear, quadratic, and cubic models. (Use the formulas given in the textbook. For  $C_p$ , use  $s^2$  from the full model that you fit in part (c).)
- Perform a lack-of-fit test (based on *pure error*) for the quadratic model.