

Homework 3

1. Problems from the textbook:

- 4.5 4.9 4.13 4.23 4.26 4.29 4.59**
4.63 (Make sure that you know how to complete the ANOVA tables on Page 237)
4.81 4.82

2. An equal number of families from eight different cities of various sizes were asked their family incomes and how much money they spent for food, clothing, and housing per year. The city sizes (x_1), average annual family income (x_2), and average family expenditure (y) are summarized below. (City size in 1000s, family income and expenditure in \$100s.)

City Size (x_1):	30	50	75	100	150	200	175	120
Income (x_2):	122	230	144	280	278	197	230	190
Expenditure (y):	65	77	79	80	82	90	84	81

(a) We first study the relationship between City sizes (x_1) and expenditure (y). The following quantities can be easily computed.

$$\sum y_i^2 = 51236; \bar{y} = 79.75; \sum x_{1i}^2 = 126550; \bar{x}_1 = 112.5; \sum x_{1i}y_i = 74445$$

- i. Plot the sample data, y vs. x_1 . Does the plot suggest that the city size and expenditure have a straight-line relationship?
- ii. Complete the ANOVA table.

Source	Df	SS	MS	F Value	P-value
Model	(1)	(4)	(6)	(8)	.003
Error	(2)	73.7253	(7)		
Total	(3)	(5)			

- iii. If model $y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$ with $\epsilon_i \sim_{\text{i.i.d.}} N(0, \sigma^2)$ is fit to the data using LS techniques, compute $\hat{\beta}_0, \hat{\beta}_1$.
 - iv. Find the 99% confidence interval for β_1 and **interpret**.
 - v. Given a city that has a population of 100,000 people, find the 95% prediction interval for its average family annual expenditure.
- (b) Now consider model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} + \epsilon_i$ with $\epsilon_i \sim_{\text{i.i.d.}} N(0, \sigma^2)$, which gives the following fit.
- i. Complete the ANOVA table.
 - ii. In a city that has a population of 100,000 people, what is the effect of the average family income (x_2) on the expenditure (y)? Base your answer on the above model fit.
 - iii. Test if all the terms involving x_2 can be dropped from the above model, namely, $H_0: \beta_2 = \beta_3 = 0$. ($\alpha = 0.05$)
 - iv. Compute the coefficient of determination R^2 and the adjusted one R_a^2 .

Parameter	Estimates	Standard Error	t	P-value
β_0	45.8404	6.2542	7.3295	0.0018
β_1	0.3359	0.0695	4.8344	0.0084
β_2	0.1230	0.0339	3.6314	0.0221
β_3	-0.0012	0.0003	-3.4874	0.0252

Source	Df	SS	MS	F Value	P-value
Model	(1)	(4)	(6)	(8)	.0042
Error	(2)	16.8957	(7)		
Total	(3)	(5)			

Solutions

2. Solution:

- (a) i. Omitted.
 ii. To solve the ANOVA table, first compute blank (5),

$$SSTo = \sum y_i^2 - n \cdot \bar{y}^2 = 51,236 - 8 \times 79.75^2 = 355.5.$$

The rest of the ANOVA follows easily.

Source	Df	SS	MS	F Value	P-value
Model	1	281.78	281.78	22.93	0.003
Error	6	73.72	12.29		
Total	7	355.50			

- iii. $\hat{\beta}_1 = .1055$ and $\hat{\beta}_0 = 67.9775$.
 iv. It can be found that $s.e.(\beta_1) = 0.022$. So 95% CI for β_1 is

$$.1055 \pm t_{.975}^{(6)} \times .022 = (.0517, .1593)$$

where $t_{.975}^{(6)} = 2.447$.

- v. Here $x_0 = 100$, corresponding predicted value $\hat{y} = 78.43$. The 95% PI is (69.31, 87.55), where the standard error for \hat{y} is 1.27.
 (b) i. To solve the ANOVA table, first copy the $SSTo$ first the ANOVA table in part (a) ii. Namely, $SSTo = 355.50$.
 ii. Rewrite the model as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \\ &= (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) \cdot x_2 + \varepsilon \end{aligned}$$

Therefore, given that $x_1 = 100$, every one unit increase in x_2 corresponds to $\hat{\beta}_2 + \hat{\beta}_3 x_1 = .123 + (-.0012) \times 100 = .003$ increase in the expenditure y .

- iii. Here, the reduced model is $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$. So $SSE_{\text{reduced}} = 73.7253$. And $SSE_{\text{full}} = 16.8957$. It can be found that the observed F test statistic is equal to 6.727, which is less than $F_{.95}(2, 4) = 6.944$. Hence, we cannot reject the null hypothesis.

Source	Df	SS	MS	F Value	P-value
Model	3	338.6043	112.8681	26.72	0.0042
Error	4	16.8957	4.2239		
Total	7	355.50			

iv.

$$R^2 = \frac{SSR}{SSTo} = 95.25\%,$$

which means that 95.25% of the total variation in y can be explained by the regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$. ¶

```
# Homework 3

y <- c(65, 77, 79, 80, 82, 90, 84, 81)
x1 <- c(30, 50, 75, 100, 150, 200, 175, 120)
x2 <- c(122, 230, 144, 280, 278, 197, 230, 190)

sum(y); mean(y); sum(x1); mean(x1)
sum(x1^2); sum(y^2); sum(x1*y)

var(y)*(8-1)

fit1 <- lm(y~x1); summary(fit1); anova(fit1)
fit2 <- lm(y~x1+x2 + x1:x2); summary(fit2); anova(fit2)
```