Assignment #2 — Due Thu, 14 Oct.

*This homework covers Chapter 3 (*Problem* 1-6) and Chapter 4 (*Problem* 7-11). Submit your homework on Canvas or send it to our TA, Mr. LYU Zhongyuan (zlyuab@connect.ust.hk).

*No late homework will be accepted for credit.

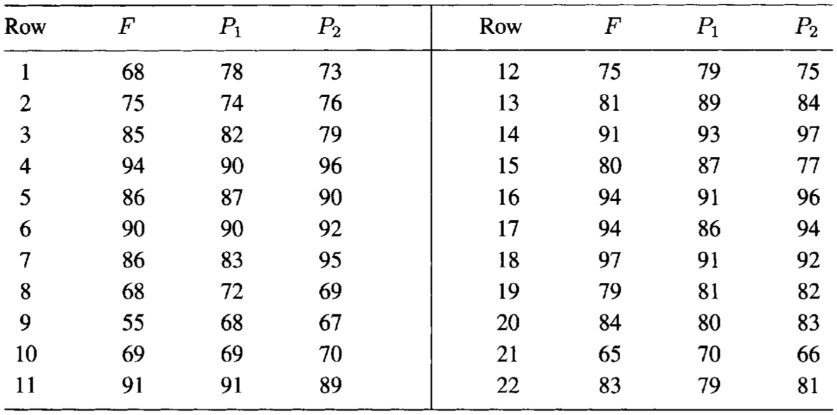
*Append the R codes you used to your submission. *If the problem does not need R or is not explicitly stated to complete in R, then you should just do it by hand with a calculator.*

*In case of rounding error, keep 3 figures after the decimal point.

Problem 1 Using the supervisor data, verify that the coefficient of X_1 in the fitted equation $\hat{Y} = 15.3276 + 0.7803X_1 - 0.0502X_2$ in (3.12), *equation on lecture slides of Chapter 3*, can be obtained from a series of simple regression equations, as outlined in Section 3.2 for the coefficient of X_2.

Problem 2 (*use* R) The following Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data is given in the file *Examination_Data.txt*.

Table 3.10 Examination Data: Scores in Final (F), First Preliminary (P_1), and Second Preliminary (P_2) Examinations



Row	F	P_1	P_2	Row	F	P_1	P_2
1	68	78	73	12	75	79	75
2	75	74	76	13	81	89	84
3	85	82	79	14	91	93	97
4	94	90	96	15	80	87	77
5	86	87	90	16	94	91	96
6	90	90	92	17	94	86	94
7	86	83	95	18	97	91	92
8	68	72	69	19	79	81	82
9	55	68	67	20	84	80	83
10	69	69	70	21	65	70	66
11	91	91	89	22	83	79	81

(a) Fit each of the following models of the data:

$$\begin{aligned} \text{Mode 1} &: F = \beta_0 + \beta_1 P_1 + \varepsilon \\ \text{Model 2} &: F = \beta_0 + \beta_2 P_2 + \varepsilon \\ \text{Model 3} &: F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon \end{aligned}$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

(c) Which variable individually, P_1 or P_2, is a better predictor of F? Why?

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction interval in this case?

Problem 3 Table 3.11 shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 20 observations. Complete the 13 missing numbers, then compute the *sample variances* $\mathrm{Var}(Y)$ and $\mathrm{Var}(X_1)$.

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	1	1848.76	69.222
Residuals	480.736	18	26.708	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	-1.84	0.0824
X_1	1.271	0.1528	8.32	< 0.0001
n = 20	R^2 = 0.794	R_a^2 = 0.783	$\hat{\sigma}$ =5.168	df =18

Problem 4 (*use* R) Using the Supervisor Performance data *Supervisor.txt*, test the hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$ in each of the following models:

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Problem 5 Table 3.14 shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

Gender : An indicator variable 1=man and 0=woman
Education : Years of schooling at the time of hire
Experience : Number of months of pervious work experience
Months : Number of months with the compnay

Table 3.14 Regression Output When Salary is Related to Four Predictor Variables

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	3526.4	327.7	10.76	0.000
Gender	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
n = 93	R^2 = 0.515	R_a^2 = 0.489	$\hat{\sigma}$ = 507.4	df = 88

In (a)-(b) below, specify the null and alternative hypothesis, the test used, and your conclusion using a 5% level of significance. For (c)-(e), only point prediction is desired.

(a) Conduct the F-test for the overall fit of the regression

(b) Is there a *positive* linear relationship between salary and experience, after accounting for the effect of the variables Gender, Education and Months?

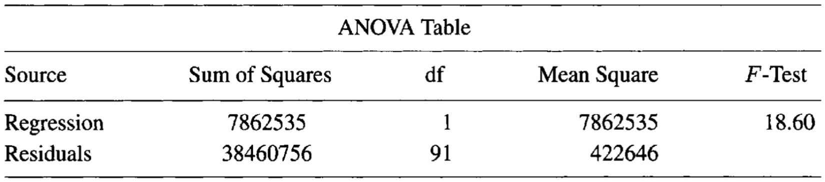
(c) What salary would you forecast for a man with 12 years of education, 10 months of experience, and 15 months with the company?

(d) What salary would you forecast, on average, for men with 12 years of education, 10 months of experience, and 15 months with the company?

(e) What salary would you forecast, on average, for women with 12 years of education, 10 months of experience, and 15 months with the company?

Problem 6 Consider the regression model that generated the output in Table 3.14 to be a full model. Now consider the reduced model in which Salary is regressed on only Education. The ANOVA table obtained when fitting the model is shown in Table 3.15. Conduct a single test to compare the full and reduced models. What conclusion can be drawn from the result of the test? (Use $\alpha = 0.05$.)

Table 3.15 ANOVA Table When the Beginning Salary is Regressed on Education



ANOVA Table

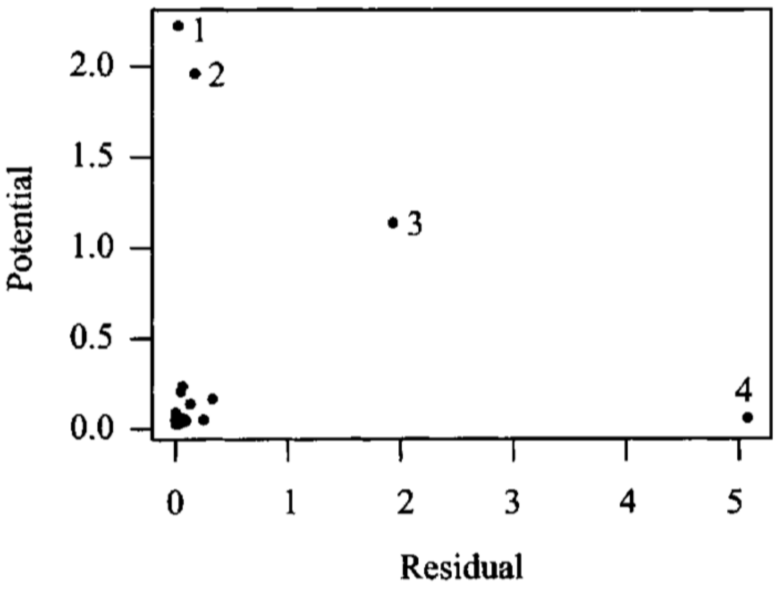
Source	Sum of Squares	df	Mean Square	F-Test
Regression	7862535	1	7862535	18.60
Residuals	38460756	91	422646	

Problem 7 (*use* R) Consider the computer repair problem discussed in Chapter 2. In a second sampling period, 10 more observations on the variables Minutes and Units were obtained. Since all observations were collected by the same method from a fixed environment, all 24 observations were pooled to form one dataset. The data appear in Table 4.6 and stored in the file *Computer_Repair.txt.*

(a) Fit a linear regression model relating Minutes to Units.

(b) Check each of the standard regression assumptions and indicate which assumptions seems to be violated.

Problem 8 In an attempt to find unusual points in a regression data set, a data analyst examines the *potential-Residual* plot. Classify each of the unusual points on this plot according to the type.



Problem 9 Name one or more graphs that can be used to validate each of the following assumptions. For each graph, sketch an example where the corresponding assumption is valid and an example where the assumption is clearly invalid.

(a) There is a linear relationship between the response and predictor variables.

(b) The observations are independent of each other

(c) The error terms have constant variance

(d) The error terms are uncorrelated

(e) The error terms are normally distributed

(f) The observations are equally influential on least squares results

Problem 10 Consider again the Examination data used in *Problem 2* and given in Table 3.10.

(a) For each of the three models, draw the *potential-Residual* plot. Identify all unusual observations (by number) and classify as outliers, high-leverage point, and/or influential observation.

(b) What model would you use to predict the final score F?

Problem 11 Identify unusual observations for the data set in Table 4.7.

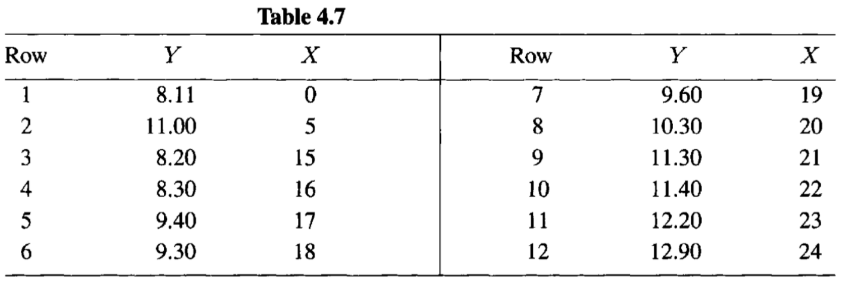


Table 4.7

Row	Y	X	Row	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24