Assignment #1 — Due Sat, 25 Sep.

*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.*

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

Problem 1 In each of the following sets of variables, identify which of the variables can be regarded as a response variable and which can be used as predictors? (Explain)

(a) Number of cylinders and gasoline consumption of cars

(b) SAT scores, grade point average, and college admission

(c) Supply and demand of certain goods

(d) Company's assets, return on stock, and net sales

(e) The distance of a race, the time to run the race, and the weather conditions at the time of running

(f) The weight of a person, whether or not the person is a smoker, and whether or not the person has a lung cancer

(g) The height and weight of a child, his/her parents' height and weight, and the gender and age of the child

Problem 2 Suppose that 20 observations are observed as follows

ID	observation
1	240
2	243
3	250
4	254
5	264
6	279
7	284
8	285
9	290
10	298
11	302
12	310
13	312
14	315
15	322
16	337
17	348
18	384
19	386
20	520

Assume the observations are i.i.d and sampled from normal distribution. Test the hypothesis

$$H_0 : \mu = 200 \quad \text{V.S.} \quad H_1 : \mu \neq 200$$

with significance level $\alpha = 0.02$. Then, use the function `t.test()` in R to repeat the test and report the result.

Problem 3 Explain why you would or wouldn't agree with each of the following statements:

(a) $\mathrm{Cov}(Y, X)$ and $\mathrm{Cor}(Y, X)$ can take values between $-\infty$ and $+\infty$.

(b) If $\mathrm{Cov}(Y, X) = 0$ or $\mathrm{Cor}(Y, X) = 0$, one can conclude that there is no relationship between Y and X.

(c) The least squares regression line passes the point $(\bar{x}, \bar{y})$.

(d) The least squares line fitted to the points in the scatter plot of Y versus $\hat{Y}$ has a zero intercept and a unit slope.

Problem 4 Based on the regression output in the following table (sample size $n = 14$), test the following hypothesis using $\alpha = 0.05$:

Table 1: Regression Output for Computer Repair Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	4.162	3.355	1.24	0.2385
Units	15.509	0.505	30.71	< 0.0001

(a) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 \neq 15$

(b) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 > 15$

(c) $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$

(d) $H_0 : \beta_0 = 5$ versus $H_1 : \beta_0 \neq 5$

Problem 5 Using the regression output in Problem 4, construct the 98% confidence interval for β_0.

Problem 6 When fitting a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data using the least squares method, suppose that $H_0 : \beta_1 = 0$ was not rejected. This implies that the model can be written simply as: $Y = \beta_0 + \varepsilon$.

(a) Show that the least squares estimate of β_0 is $\hat{\beta}_0 = \bar{y}$.

(b) What are the ordinary least squares residuals in this case?

(c) Show that the ordinary least squares residuals sum up to zero.

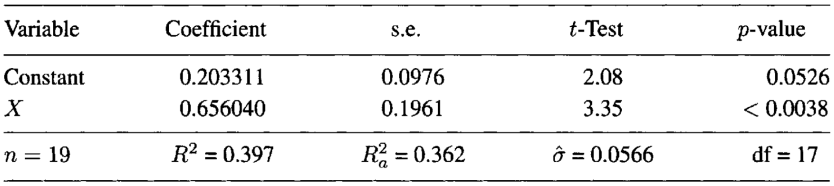
Problem 7 Using the data in Table 2.5, and the fitted values and residuals in Table 2.7 (*tables on lecture slides*), verify that

(a) $\text{Cor}(Y, X) = \text{Cor}(Y, \hat{Y}) = 0.994$

(b) SST=27768.348

(c) SSE=348.848

Problem 8 Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in the following Table. It was also found that SSR=0.0358 and SSE=0.0544. Suppose that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ satisfies the usual regression assumptions.



Variable	Coefficient	s.e.	t-Test	p-value
Constant	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	< 0.0038
$n = 19$	$R^2 = 0.397$	$R_a^2 = 0.362$	$\hat{\sigma} = 0.0566$	df = 17

(a) Compute $\text{Var}(Y)$ and $\text{Cor}(Y, X)$.

(b) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?

(c) Suppose further that the mean and variance of the participation rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (b).

(d) Construct the 95% confidence interval for the slope of the true regression line, β_1.

(e) Test the hypothesis $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 > 1$ at the 2% significance level.

(f) If Y and X were reversed in the above regression, what would you expect R^2 to be?

Problem 9 Consider fitting a simple linear regression model through the origin, $Y = \beta_1 X + \varepsilon$, to a set of data using the least squares method.

(a) Give an example of situation where fitting the model $Y = \beta_1 X + \varepsilon$ is justified by theoretical or other physic and material considerations.

(b) Show that the least squares estimate of β_1 is as given by $\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}$.

(c) Show that the residuals $e_1, e_2, \cdots, e_n$ will not necessarily add up to zero.

Problem 10 (Use R to do this problem) One may wonder if people of similar heights tend to marry each other. For this purpose, a sample of newly married couples was selected. Let X be the height of the husband and Y be the height of the wife. The heights (in centimeters) of husbands and wives are found in file: Heights of husband and wife.txt

(a) Compute the covariance between the heights of the husbands and wives.

(b) What would the covariance be if heights were measured in inches rather than in centimeters?

(c) Compute the correlation coefficient between the heights of the husband and wife.

(d) What would the correlation be if every man married a woman exactly 5 centimeters shorter than him?

(e) Use wife's height as the response variable and husband's wife as the predictor variable, fit a simple linear regression model and report the R output.

(f) Based on part (e), these the null hypothesis that the slope is zero at the significance level 0.05.

(g) Based on part (e), these the null hypothesis that the intercept is zero at the significance level 0.05.