

Assignment 1

For every problem that asks you to derive a formula or verify a statement, please remember to carefully show *all* of your steps. You may assume that the least squares estimates exist and are unique, unless the problem indicates otherwise.

1. For data that follow a simple linear regression model, show the following:
 - (a) $E(\bar{Y}) = \beta_0 + \beta_1\bar{X}$
 - (b) $E(Y_i - \bar{Y}) = \beta_1(X_i - \bar{X})$
 - (c) $\hat{\beta}_1$ is unbiased
 - (d) $\hat{\beta}_0$ is unbiased
2. With the aid of the formulas for the least squares estimates and/or normal equations, show algebraically that (with probability 1)
 - (a) $\sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n e_i = 0$
 - (b) $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$
 - (c) $\sum_{i=1}^n (\hat{Y}_i - \bar{Y}) = 0$

[Note that the results of parts (b) and (c) prove

$$\sum_{i=1}^n \bar{Y}(Y_i - \bar{Y}) = 0 \quad \text{and} \quad \sum_{i=1}^n \bar{Y}(\hat{Y}_i - \bar{Y}) = 0,$$

which were two facts used during lecture to derive ANOVA relationships.]

3. Perform the following exercises from the textbook, Section 1.11:
 - (a) Exercise 1.14
 - (b) Exercise 1.16
 - (c) Exercise 1.22 (In part (c), use Problem 2(a) above, instead of Exercise 1.8.)
4. Suppose the data pairs (X_i, Y_i) , $i = 1, \dots, n$, follow the simple linear regression model with all of the usual assumptions.
 - (a) Briefly explain why Y_1, \dots, Y_n are independent. Are they identically distributed?
 - (b) Derive $\text{Cov}(Y_k, \bar{Y})$.
 - (c) Derive $\text{Cov}(Y_k, \hat{\beta}_1)$.
[Suggestion: Use the formula $\hat{\beta}_1 = \sum_i x_i Y_i / \sum_i x_i^2$, where $x_i = X_i - \bar{X}$.]
 - (d) Using the previous parts, derive $\text{Cov}(Y_k, \hat{Y}_k)$.
[Suggestion: First show that $\hat{Y}_k = \bar{Y} + \hat{\beta}_1 x_i$.]
 - (e) Using the previous part, derive $V(e_k)$, the variance of the k^{th} residual.
[You may use the formula for $V(\hat{Y}_k)$ given during lecture and in the textbook.]

5. Data sets that follow were generated (with rounding) using the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim i.i.d. N(0, \sigma^2) \quad i = 1, \dots, 6$$

	X1	Y1	X2	Y2	X3	Y3	X4	Y4	X5	Y5	X6	Y6
BAI LEI	22	41	52	49	60	69	42	55	47	60	65	62
BROWN STEPHEN V	47	58	72	65	61	55	65	76	69	67	28	38
CHANG CHE-SHUN	45	42	45	29	20	29	41	50	31	38	24	36
CHEN OU	54	62	46	56	39	52	59	66	45	42	71	72
DONOHUE MICHAEL	54	60	44	56	54	58	64	66	27	31	79	75
GAO HAIBING	68	63	36	59	70	64	54	59	51	48	63	62
GARG DIVYA	65	75	32	30	77	64	28	58	56	54	67	49
GLUCK MATHEW R	78	71	59	60	24	36	75	69	62	55	37	44
GORDON ROBERT F	42	48	38	37	69	74	34	45	53	53	63	67
GUCI LEDIA	57	60	76	56	73	82	54	58	40	61	47	47
HARIHARAN POOJA	26	49	36	52	34	34	29	40	26	53	51	61
HUANG LEI	43	58	79	88	43	50	27	42	46	55	52	53
KIM CHANMIN	51	66	73	76	64	62	42	50	75	74	48	44
KIRPICH ALEX	21	30	70	80	56	65	38	50	31	24	36	36
LEARY EMILY V	58	66	77	62	45	48	44	67	69	57	72	72
LI KE	57	63	64	60	58	58	58	40	64	61	76	75
LIN TONG	48	56	51	49	59	63	22	34	49	48	38	43
LIU MINZHAO	38	53	22	28	23	39	60	71	73	60	59	56
LU CUIE	74	65	71	84	52	56	28	38	64	52	57	51
LUO XUAN	35	33	80	73	32	51	30	28	43	85	79	87
MA LU	74	62	55	80	60	80	74	70	71	67	28	56
MALLICK PRANJAL	63	63	37	62	59	61	62	79	29	28	78	71
MARCUS GABRIEL	61	56	78	62	25	52	76	70	27	47	21	31
NAMKOONG YOUNG	70	61	50	77	25	34	62	69	66	78	31	46
NEAL DANIEL W	26	40	79	81	42	44	38	39	58	54	50	61
PETTERSON SONIA	46	48	27	26	60	65	72	65	79	74	22	34
PRANO BRIJIDA A	35	39	22	37	67	63	29	27	32	46	46	74
SHAO ANQI	69	61	35	58	54	61	39	52	37	41	36	27
SINHA AMIT	63	67	58	45	46	49	73	70	20	40	52	29
THAYER LAURA K	45	63	42	53	58	72	42	59	63	68	49	68
YE RONGZHONG	73	79	30	36	47	57	69	78	21	18	37	47
ZHOU ZHUO	63	56	66	62	32	36	72	71	78	86	22	30
ZHU XIAOYU	33	43	72	68	60	68	24	32	41	64	67	79
EXTRA	47	78	22	42	24	41	65	63	63	68	28	37
demo	20	31	40	60	60	57	69	60	52	73	57	62

Use the data set listed after your name to complete the following parts. (If your name does not appear on the list, you should contact the instructor.) Answer each of parts (a) through (e) *using only a calculator*, then check your answers using SAS® or R. *Attach printouts of your SAS® or R code and output!*

- (a) Compute

- (i) the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
- (ii) the unbiased estimate s^2 of σ^2 .

- (b) Compute an ANOVA table that includes the corrected total, regression, and residual sources of variation.

- (c) Compute the coefficient of determination.
- (d) Compute estimates of the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also, compute individual 95% two-sided confidence intervals for β_0 and β_1 .
- (e) Test the null hypothesis $\beta_1 = 0$ versus the alternative $\beta_1 \neq 0$ at level $\alpha = 0.05$. Perform both the F -test and the two-sided t -test.
- (f) Suppose $(X_{\text{new}}, Y_{\text{new}})$ follows the same model (with the same true parameters) and is independent of the data. If $X_{\text{new}} = 25$, compute (using only a calculator)
 - (i) a 95% two-sided confidence interval for $E(Y_{\text{new}})$.
 - (ii) a 95% two-sided prediction interval for Y_{new} .
- (g) Create a plot of your data, along with the estimated regression line. (You may either do this carefully by hand or using a computer.)