SM-4331 Exercise 5

1. Let a_i, b_j, c and d be any real numbers. Show that

$$\sum_{i=1}^{n}(a_i - c)(b_i - d) = \sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b}) + n(\bar{a} - c)(\bar{b} - d),$$

where $\bar{a} = n^{-1}\sum_{i=1}^{n} a_i$ and $\bar{b} = n^{-1}\sum_{i=1}^{n} b_i$.

2. For the simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, the ordinary least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions to

$$\arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Use calculus (i.e. differentiate the sum of squared errors with respect to β_0 and β_1) to derive the LSE solutions.

3. Let the observations $\{(y_i, x_i) | i = 1, \ldots, n\}$ be taken from the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Suppose n is a large integer.
   (a) Construct a Wald test for $H_0 : \beta_1 = 2\beta_0$ against $H_1 \beta_1 \neq 2\beta_0$.
   (b) For a given x, construct a confidence interval for $\mu(x) := \mathrm{E}(y) = \beta_0 + \beta_1 x$.

4. Consdider a linear model $y_i = \beta x_i + \epsilon_i$, where $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 > 0$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, and $x_1, \ldots, x_n$ are constants.
   (a) Find the LSE $\hat{\beta}$. Suggest an estimator for σ^2.
   (b) Show that the LSE $\hat{\beta}$ is unbiased, and find $\mathrm{SE}(\hat{\beta})$.
   (c) If in addition $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ find a confidence interval for β.
   (d) Based on the interval for β, find a confidence interval for $\mu(x) = \mathrm{E}(y)$, where $y = \beta x + \epsilon$.

5. The table below lists the USA social security costs in 7 years between 1965 to 1992.

Year	1965	1970	1975	1980	1985	1990	1992
x = number of years from 1960	5	10	15	20	25	30	32
y = social security cost (\$ billions)	17.1	29.6	63.6	117.1	186.4	246.5	285.1

   (a) Plot the data y against x.
   (b) Compute $\sum_i x_i$, $\sum_i y_i$, $\sum_i x_i^2$, $\sum_i y_i^2$, and $\sum_i x_i y_i$, and therefore fit the data to a simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$. Superimpose the fitted regression line onto the plot in (a).
   (c) Test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$. What can be concluded on the social security costs from the test?

Velocity (mph)	20.5	20.5	30.5	40.5	48.8	57.8
Stopping distance (ft)	15.4	13.3	33.9	73.1	113.0	142.6

(d) Plot the residuals against x. Are you happy with the fitted model? If not, discuss what you may try to do to achieve a better fitting.

6. The stopping distance (y) of a car was studied in relation to the velocity (x) of the car. The table below lists the stop distances at 6 different velocities.

   (a) Plot y against x, and $z := \sqrt{y}$ against x.

   (b) Compute the sample correlation coefficients of y and x, and z and x.

   (c) Fit the linear regression model for y on x, and examine the residuals.

   (d) Fit the linear regression model for z on x, and examine the residuals.

   (e) For a given x, a predictive interval for $y = \beta_0 + \beta_1 x + \epsilon$ with coverage probability $1 - \alpha$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}(\alpha/2)\hat{\sigma}\sqrt{1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n\sum_{j=1}^{n}(x_j - \bar{x})^2}}.$$

   Based on this formula, compute the predictive intervals with coverage probability 0.95 for y and z when $x = 35$.

   (f) Which model is better?

7. In a regression analysis, three possible models have been tried:

   - **Model 1**: Regress y on x_1.
   - **Model 2**: Regress y on x_2.
   - **Model 3**: Regress y on x_1 and x_2.

   The numerical output of these models are shown below.

   (a) Find the missing values **A1–A8**.

   (b) What can be concluded from these three fitted regression models?

Model 1: $y = \beta_0 + \beta_1 x_1 + \epsilon$

	Estimate	SE	T	$P(t > \|T\|)$
β_0	1.1398	0.1019	11.183	$< 2e^{-16}$
β_1	0.8604	0.1025	**A1**	$1.6e^{-12}$

$\hat{\sigma} = 0.905$ on 78 degrees of freedom.
$R^2 = 0.4746$, $\tilde{R}^2 =$ **A2**

Model 2: $y = \beta_0 + \beta_2 x_2 + \epsilon$

	Estimate	SE	T	$P(t > \|T\|)$
β_0	1.04989	0.20152	5.210	$1.5e^{-6}$
β_2	-0.01336	**A3**	-0.092	**A4**

$\hat{\sigma} = 1.248$ on 78 degrees of freedom.

Model 3: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

	Estimate	SE	T	$P(t > \|T\|)$
β_0	1.16464	0.14762	7.890	$1.66e^{-11}$
β_1	0.86067	0.10314	8.345	$2.20e^{-12}$
β_2	-0.02493	0.10635	-0.234	0.815

$\hat{\sigma} =$ **A5** on **A6** degrees of freedom.
$R^2 =$ **A7**

ANOVA for Model 3: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

Source	d.f.	SS	Mean SS	T	$P(F > T)$
$\sum_{i=1}^{79}(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} - \bar{y})^2$	1	57.695	57.695	**A8**	$2.225e^{-12}$
$\sum_{i=1}^{79}(\hat{\beta}_0 + \hat{\beta}_2 x_{i2} - \bar{y})^2$	1	0.046	0.046	0.055	0.8153
$\sum_{i=1}^{79}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$	77	63.833	0.829		