StatsI — Exercise 9

1. Let a_i, b_j, c, d are 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 $\bar{a} = n^{-1} \sum_i a_i$, $\bar{b} = n^{-1} \sum_i b_i$.

2. Find unknown c or α in the following expressions using Murdoch and Barnes "Statistical Tables":

$$P(F_{7,8} > c) = 0.01, \qquad P(F_{5,3} \le 28.2) = \alpha, \qquad P(F_{6,10} \le c) = 0.05.$$

3. 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: Year-1960	5	10	15	20	25	30	32
y: social security cost ($ Billion)	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 $\sum_i x_i y_i$, therefore fit the data with a simple regression model $y = \beta_0 + \beta_1 x + \varepsilon$. Superimpose the fitted regression line in the plot (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?

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

4. 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.

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

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

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

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

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

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

$$\hat{y} \pm t_{\alpha/2, n-2}\, \hat{\sigma} \left\{ 1 + \frac{\sum_{i=1}^n (x_i - x)^2}{n \sum_{j=1}^n (x_j - \bar{x})^2} \right\}^{1/2}.$$

   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?

5. Let the observations $\{(y_i, x_i), i = 1, \cdots, n\}$ be taken from the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Suppose n is a large integer.

   (a) Construct a Wald test for $H_0 : \beta_1 = 2\beta_0$ against $H_1 : \beta_1 \ne 2\beta_0$.

(b) For a given x, construct a confidence interval for $\mu(x) = Ey = \beta_0 + \beta_1 x$.

6. For linear model $y_i = \beta x_i + \varepsilon_i$, where $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$, and $\mathrm{Cov}(\varepsilon_i,\ \varepsilon_j) = 0$ for $i \neq j$, and $x_1, \cdots, x_n$ are constants.

   (a) Find the LSE $\widehat{\beta}$. Suggest an estimator for σ^2.

   (b) Show the LSE $\widehat{\beta}$ is unbiased, and find $\mathrm{SE}(\widehat{\beta})$.

   (c) If in addition $\varepsilon_i \sim N(0,\ \sigma^2)$, find a confidence interval for β. Based on the interval for β, find a confidence interval for $\mu(x) = E(y)$, where $y = \beta x + \varepsilon$.

7. In a regression analysis, three possible models have been tried: regress y on x_1, or on x_2, or on x_1 and x_2 together.

   (a) Find the missing values A1, A2, A3, A4, A5, A6 and A7 in the R outputs below.

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

```
> lmr1 <- lm(y ~ x1)
> summary(lmr1)
Call: lm(formula = y ~ x1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1398     0.1019  11.183  < 2e-16
x1            0.8604     0.1025      A1  1.62e-12
---
Residual standard error: 0.905 on 78 degrees of freedom
Multiple R-squared: 0.4746,       Adjusted R-squared: A2

> lmr2 <- lm(y ~ x2)
> summary(lmr2)
Call: lm(formula = y ~ x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04989    0.20152   5.210  1.5e-06
x2          -0.01336         A3  -0.092       A4
---
Residual standard error: 1.248 on 78 degrees of freedom

> lmr12 <- lm(y ~ x1 + x2)
> summary(lmr12)
Call: lm(formula = y ~ x1 + x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.16464    0.14762   7.890 1.66e-11
x1           0.86067    0.10314   8.345 2.20e-12
x2          -0.02493    0.10635  -0.234    0.815
---
Residual standard error: A5 on 77 degrees of freedom
Multiple R-squared: A6
```

```
> anova(lmr12)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 57.695  57.695      A7    2.225e-12
x2         1  0.046   0.046   0.055    0.8153
Residuals 77 63.833   0.829
```

8. The passenger car mileage data are saved in the file 'carMileage.txt' available from the ST425 moodle page. Perform the following regression analysis using *R*.
   (a) Fit a simple linear regression model to predict MPG (miles per gallon) from HP (horsepower).
   (b) Fit a multiple regression model for MPG using the other 4 variables in the data.
   (c) Using the *R*-function `step` to search an optimum model for predicting MPG.