STAT 3008
Exercises 3

Problems refer to the problem sets in the textbook: Applied Linear Regression, 3rd edition by Weisberg.

1. (i) Problem 2.4.2.
   Write the mean function in the deviations from the mean form as in Problem 2.3. For this particular problem heights data in the file heights.txt, give an interpretation for the value of β_1. In particular, discuss the three cases of $\beta_1 = 1$, $\beta_1 < 1$ and $\beta_1 > 1$. Obtain a 99% confidence interval for β_1 from the data.

   (ii) Problem 2.4.3.
   Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

   (iii) From the regression $E(Dheight|Mheight) = \beta_o + \beta_1 Mheight$, express the relation in the form $Mheight = \alpha_o + \alpha_1 E(Dheight|Mheight)$.

   (iv) Fit the model $E(Mheight|Dheight) = \beta_o + \beta_1 Dheight$. Is it the same as (iii)?

   Remarks. Part (iii) and (iv) show that regression treats x and y differently. Note that $\hat{\beta}_1 < 1$ no matter $Mheight$ is chosen to be x or y. This is an example of **regression to the mean**. The mathematical reason is that $SXX \approx SYY$, so no matter how you do the regression, you have $\hat{\beta}_1 = SXY/SXX$ or SXY/SYY, both are $\approx SXY/\sqrt{SXX\ SYY} = r_{xy} < 1$.

2. Problem 2.7.
   Regression through the origin
   Occasionally, a mean function in which the intercept is known a priori to be zero may be fit. This mean function is given by

$$E(y|x) = \beta_1 x \tag{2.30}$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is RSS $=\sum(y_i - \hat{\beta}_1 x_i)^2$.

2.7.1. Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum(x_i y_i)/\sum(x_i^2)$ Show that $\hat{\beta}_1$ is unbiased and that $\text{Var}(\hat{\beta}_1)=\sigma^2/\sum(x_i^2)$. Find an expression for $\hat{\sigma}^2$. How many df does it have?

2.7.2. Derive the analysis of variance table with the larger model given by (2.16), but with the smaller model specified in (2.30). Show that the F-test derived from this table is numerically equivalent to the square of the t -test (2.23) with $\beta_0^* = 0$ 2.7.3. The data in Table 2.6 and in the file snake.txt give X = water content of snow on April 1 and Y = water yield from April to July in inches in the Snake River watershed in Wyoming for n = 17 years from 1919 to 1935 (from Wilm, 1950).

TABLE 2.6 Snake River Data for Problem 2.7

X	*Y*	*X*	*Y*
23.1	10.5	32.8	16.7
31.8	18.2	32.0	17.0
30.4	16.3	24.0	10.5
39.5	23.1	24.2	12.4
52.5	24.9	37.9	22.8
30.5	14.1	25.1	12.9
12.4	8.8	35.1	17.4
31.5	14.9	21.1	10.5
27.6	16.1		

Fit a regression through the origin and find $\hat{\beta}_1$ and σ^2. Obtain a 95% confidence interval for β_1. Test the hypothesis that the intercept is zero.

2.7.4. Plot the residuals versus the fitted values and comment on the adequacy of the mean function with zero intercept. In regression through the origin, $\sum(\hat{e}_i \neq 0)$.

3. Problem 2.8.
   2.8. Scale invariance

2.8.1. In the simple regression model (2.1), suppose the value of the predictor X is replaced by cX, where c is some non zero constant. How are $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2, and the t -test of NH: $\beta_1 = 0$ affected by this change?
2.8.2. Suppose each value of the response Y is replaced by dY, for some $d \neq 0$. Repeat 2.8.1.

4. Problem 2.10.1 and 2.10.2.
   2.10. Zipfs law
   Suppose we counted the number of times each word was used in the written works by Shakespeare, Alexander Hamilton, or some other author with a substantial written record (Table 2.7). Can we say anything about the frequencies of the most common words? Suppose we let fi be the rate per 1000 words of text for the ith most frequent word used. The linguist George Zipf (19021950) observed a law like relationship between rate and rank (Zipf, 1949),

$$E(f_i|i) = a/i^b$$

   and further observed that the exponent is close to b = 1. Taking logarithms of both sides, we get approximately

$$E(log(f_i)|log(i)) = log(a) - blog(i)$$

   Zipfs law has been applied to frequencies of many other classes of objects besides words, such as the frequency of visits to web pages on the internet and the frequencies of species of insects in an ecosystem. The data in MWwords.txt give the frequencies of words in works from four different sources: the political writings of eighteenth-century American political figures Alexander Hamilton, James Madison, and John Jay, and the book Ulysses by twentieth-century Irish writer James Joyce. The data are from Mosteller and Wallace (1964, Table 8.1-1), and give the frequencies of 165 very common words. Several missing values occur in the data; these are really words that were used so infrequently

TABLE 2.7 The Word Count Data

Word	The word
Hamilton	Rate per 1000 words of this word in the writings of Alexander Hamilton
HamiltonRank	Rank of this word in Hamilton's writings
Madison	Rate per 1000 words of this word in the writings of James Madison
MadisonRank	Rank of this word in Madison's writings
Jay	Rate per 1000 words of this word in the writings of John Jay
JayRank	Rank of this word in Jay's writings
Ulysses	Rate per 1000 words of this word in *Ulysses* by James Joyce
UlyssesRank	Rank of this word in *Ulysses*

that their count was not reported in Mosteller and Wallaces table.
2.10.1. Using only the 50 most frequent words in Hamiltons work (that is, using only rows in the data for which HamiltonRank ≤ 50), draw the appropriate summary graph, estimate the mean function (2.31), and summarize your results.
2.10.2. Test the hypothesis that b = 1 against the two-sided alternative and summarize.

5. Problem 2.12.
2.12. Old Faithful Use the data from Problem 1.4, page 18. 2.12.1. Use simple linear regression methodology to obtain a prediction equation for interval from duration. Summarize your results in a way that might be useful for the nontechnical personnel who staff the Old Faithful Visitors Center. 2.12.2. Construct a 95% confidence interval for

$$E(interval|duration = 250)$$

2.12.3. An individual has just arrived at the end of an eruption that lasted 250 seconds. Give a 95% confidence interval for the time the individual will have to wait for the next eruption.
2.12.4. Estimate the 0.90 quantile of the conditional distribution of

$$interval|(duration = 250)$$

assuming that the population is normally distributed.

6. Show that $Cov(\bar{y}, \hat{\beta}_1) = 0$.

7. Let

$$X = \begin{pmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 5 \\ 1 & 1 \\ 1 & 2 \\ 1 & 8 \\ 1 & 0 \end{pmatrix}.$$

Find $X'X$, XX', $(X'X)^{-1}$, $tr(X'X)$ and $tr(XX')$.