Practice Problems

Please use $\alpha = 0.05$ for all statistical tests.
Please keep at least four decimal digits in all your numerical calculations.

1. Use not more than 3 sentences to answer each of the following questions.

   (a) "It is always preferable to include as many terms as possible in a regression model". Do you agree? Explain.

   (b) "If $R^2 = 0$, then it means that there is no relation between X and Y". Please comment.

   (c) "Since most statistical studies could only lead to association but not causation, there is no use in studying statistics". Please comment.

2. This question is concerned with the following data set:

i:	1	2	3	4	5	6	7	8	9	10	11	12
X_i:	16	16	16	24	24	24	32	32	32	40	40	40
Y_i:	199	196	200	218	220	223	237	234	235	250	248	253

   (a) Fit a simple linear regression to this data set.

   (b) Estimate the sub-population mean $E(Y|X = x)$ when $x = 28.0$. Attach a 95% confidence interval to your estimate.

   (c) Construct a *prediction* interval for a new observation if you know that its x-value is 42.0.

   (d) Construct the corresponding ANOVA table. Use the computed F-value to test if $\beta_1 = 0$.

3. The regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ was fitted to a set of n data points. Part of the outputs from the R command `anova` are:

```
Response: Y
          Df Sum Sq Mean Sq   F value
X1         1  0.001   0.001    0.0644
X2         1 35.003  35.003      (**)
X3         1  0.511   0.511   62.0219
X4         1 32.041  32.041 3892.4496
Residuals 31  0.255     (*)
```

   (a) Find `(*)` and `(**)`.

   (b) Test "$H_0 : \beta_3 = \beta_4 = 0, \beta_0, \beta_1, \beta_2$ arbitrary" against "$H_1 : \beta_j$ arbitrary for all i".

   (c) For "$H_0 : Y = \beta_0$", with the above R outputs, which of the following alternative hypotheses can it be tested against with? "$H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$" or "$H_1 : Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4$"? Perform the corresponding test.

4. As an alternative to the ordinary least squares (OLS) principle, one could use the least absolute error (LAE) method to estimate the parameters β_0 and β_1 in the simple linear regression model. LAE estimates these parameters by minimizing the following *sum of absolute errors* (SAE):

$$\text{SAE}(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|. \tag{1}$$

   (a) One attractive property of LAE over OLS is that it is less sensitive to outliers. Can you explain why?

(b) Describe how you would modify (1) to perform weighted LAE regression.

(c) LAE also has shortcomings. For example, unlike OLS, no closed-form expressions are available for $\hat{\beta}_0$, $\hat{\beta}_1$, $\mathrm{Var}(\hat{\beta}_0)$ nor $\mathrm{Var}(\hat{\beta}_1)$. To calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, one has to use iterative numerical procedures.

Suppose now that you have written such a *R* routine for calculating $\hat{\beta}_0$ and $\hat{\beta}_1$. Describe how you would use this routine and the bootstrap method to construct 95% confidence intervals for these LAE estimates.

5. Consider the multiple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \quad \mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{I}.$$

Suppose we use $\hat{\boldsymbol{\beta}} = (0.5 + \alpha)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to estimate $\boldsymbol{\beta}$, where α is a pre-specified constant between 0 and 1.

(a) Calculate the bias and variance of $\hat{\boldsymbol{\beta}}$.

(b) How should we choose α if we want to minimize the variance of the estimator?

—— End of Midterm Exam ——