STAT 513 fa 2020 hw 6

Do NOT use fancy R functions like `lm()` or `t.test()` to do any part of this homework.

1. The following data from [2] are body surfaces (cm^2) and metabolic rates (kcal/day) measured on a set of dogs:

Body surface (cm^2)	Metabolic rate (kcal/day)
10750	1113
8805	982
7500	908
7662	842
5286	626
3724	430
2423	281

Regard the metabolic rate as the response variable and the body surface area as the covariate and assume that the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\text{Normal}(0, \sigma^2)$ random variables, is appropriate for these data.

(a) Compute the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1.

(b) Make a scatterplot of the metabolic rates versus the body surface areas with the least-squares line overlaid.

(c) Compute the residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_1 x_i$, for $i = 1, \ldots, n$, and use these to compute the unbiased estimator $\hat{\sigma}^2$ of σ^2. Report the value of $\hat{\sigma}$.

(d) Give a 99% confidence interval for β_1.

(e) Compute the p-value of these data for testing H_0: $\beta_1 \leq 0$ versus H_1: $\beta_1 > 0$. Interpret your answer.

(f) Construct a 95% confidence interval for the average metabolic rate of dogs with body surface area equal to 6000cm^2.

(g) Construct an interval which will contain with probability 0.95 the metabolic rate of a randomly selected dog among dogs with body surface area equal to 6000cm^2.

(h) Make a scatterplot of the metabolic rates versus the body surface areas with the least-squares line overlaid. Then, for the sequence of x values in `x.seq <- seq(2423,10750,length=500)`, add to the plot the upper and lower values of 99% confidence intervals for $\beta_0 + \beta_1 x$ as well as the upper and lower bounds of 99% prediction intervals for new values of Y at these values of x.

2. Suppose we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$ on n subjects, where

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{Normal}(0, \sigma^2)$ random variables. Moreover, suppose that each of the values $x_1, \ldots, x_n$ is equal to 0 or 1, according to whether the subject was placed into a control or a treatment group. That is, suppose

$$x_i = \begin{cases} 0 & \text{if subject } i \text{ in control group} \\ 1 & \text{if subject } i \text{ in treatment group} \end{cases}, \quad i = 1, \ldots, n.$$

(a)   i. Give the parameter in Model (1) which is the expected value of Y_i when subject i belongs to the control group, that is when $x_i = 0$.

     ii. Use the parameters in Model (1) to express the expected value of Y_i when subject i belongs to the treatment group, that is when $x_i = 1$.

     iii. Give the parameter in Model (1) which represents the difference in the expected values of the responses between the control and treatment groups.

     iv. Let $n_1 = \#\{x_i = 0\}$ and $n_2 = \#\{x_i = 1\}$ be the numbers of subjects in the control and treatment groups and let

$$\bar{Y}_1 = \frac{1}{n_1} \sum_{\{i:x_i=0\}} Y_i \quad \text{and} \quad \bar{Y}_2 = \frac{1}{n_2} \sum_{\{i:x_i=1\}} Y_i,$$

so that $\bar{Y}_1$ and $\bar{Y}_2$ are the means of the response variable in the control and treatment groups. Now, using the fact that

$$\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{\{i:x_i=0\}} (Y_i - \beta_0)^2 + \sum_{\{i:x_i=1\}} [Y_i - (\beta_0 + \beta_1)]^2,$$

show that the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{Y}_1 \quad \text{and} \quad \hat{\beta}_1 = \bar{Y}_2 - \bar{Y}_1.$$

(b) Use the following R code to read in the data from the study in [1], which investigated the efficacy of procyanidin B-2 from apples as a hair growing agent:

```
apple_hair <- read.table(file=url("http://users.stat.ufl.edu/~winner/data/apple_hair.dat"))
x <- apple_hair$V1 - 1 # treatment indicator equal to 0 or 1
Y <- apple_hair$V4 # total hair increase
```

The covariate in x is an indicator of belonging to the control group (0) or to the treatment group (1). The response in Y is the total hair increase over six months.

     i. Compute the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1.

     ii. Compute the residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_1 x_i$, for $i = 1, \ldots, n$, and use these to compute the unbiased estimator $\hat{\sigma}^2$ of σ^2. Report the value of $\hat{\sigma}$.

     iii. Compute the p-value of these data for testing $H_0 \colon \beta_1 = 0$ versus $H_1 \colon \beta_1 \neq 0$. Interpret your answer.

(c) Let μ_1 and μ_2 represent the expected total hair increase in the control and treatment groups, respectively, and separate the data into control and treatment data with

```
ctrl <- Y[x==0]
trt <- Y[x==1]
```

   i. Make boxplots of the control and treatment data with the command

```
boxplot(ctrl,trt,names=c("control","treatment"))
```

   ii. Give the sample means $\bar{Y}_1$ and $\bar{Y}_2$ of the control and treatment groups.

   iii. Compute the value of S_{pooled} for the control and treatment data. Compare this to the value of $\hat{\sigma}$ that you computed earlier.

   iv. Compute the p-value of these data for testing H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$ with the equal-variances two-sample t-test. Interpret your answer.

   v. Keep your head from exploding.

References

[1] A. Kamimura, T. Takahashi, and Y. Watanabe. Investigation of topical application of procyanidin b-2 from apple to identify its potential use as a hair growing agent. *Phytomedicine*, 7(6):529–536, 2000.

[2] Knut Schmidt-Nielsen. *Scaling: why is animal size so important?* Cambridge University Press, 1984.