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

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

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

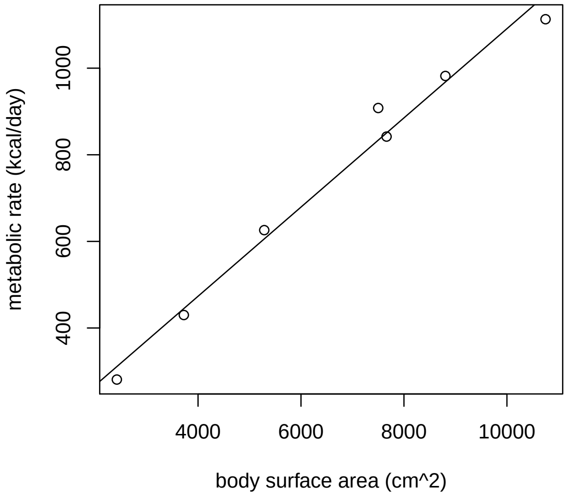
(a) Compute the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1.

```
x <- c(10750,8805,7500,7662,5286,3724,2423) # cm^2
Y <- c(1113,982,908,842,626,430,281) # kcal/day

Y.bar <- mean(Y)
x.bar <- mean(x)

Sxx <- sum( (x - x.bar)^2 )
SYY <- sum( (Y - Y.bar)^2 )
beta1.hat <- cor(x,Y)*sqrt(SYY/Sxx)
beta0.hat <- Y.bar - beta1.hat * x.bar
```

This gives $\hat{\beta}_0 = 61.40537$ and $\hat{\beta}_1 = 0.1029721$.

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

```
plot(Y ~ x, xlab="body surface area (cm^2)",ylab="metabolic rate (kcal/day)")
abline(beta0.hat,beta1.hat)
```



(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 $i = 1, \dots, n$, and use these to compute the unbiased estimator $\hat{\sigma}^2$ of σ^2. Report the value of $\hat{\sigma}$.

```
Y.hat <- beta0.hat + beta1.hat * x
e.hat <- Y - Y.hat
n <- length(x)
sigma.hat <- sqrt(sum(e.hat^2)/(n-2))
```

This gives $\hat{\sigma} = 45.55539$.

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

```
lower <- beta1.hat - qt(.995,n-2) * sigma.hat * sqrt( 1 / Sxx )
upper <- beta1.hat + qt(.995,n-2) * sigma.hat * sqrt( 1 / Sxx )
```

This gives the interval $(0.07736797, 0.1285762)$.

(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.

```
T1 <- ( beta1.hat - 0 )  / (sigma.hat / sqrt(Sxx))
pval <- 1-pt(T1,n-2)
```

This gives the p-value 8.128529×10^{-6}; there is very strong evidence that the slope is positive, so that higher body surfaces areas are associated with higher metabolic rates.

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

```
Y.hat6000 <- beta0.hat+beta1.hat*6000
lo95ci6000 <- Y.hat6000 - qt(.975,n-2) * sigma.hat * sqrt(1/n+(6000-x.bar)^2/Sxx)
up95ci6000 <- Y.hat6000 + qt(.975,n-2) * sigma.hat * sqrt(1/n+(6000-x.bar)^2/Sxx)
```

This gives the interval $(633.9313, 724.5447)$.

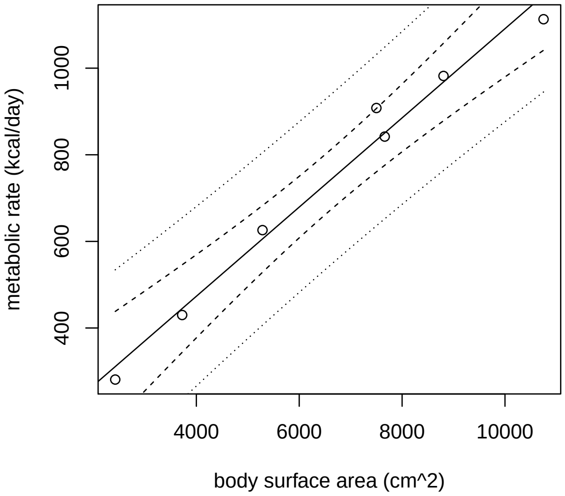
(g) Construct an interval which will contain with probability 0.95 the metabolic rate of a randomly selected dog among dogs with body surface area equal to 6000cm^2.

```
lo95pi6000 <- Y.hat6000 - qt(.975,n-2) * sigma.hat*sqrt(1+1/n+(6000-x.bar)^2/Sxx)
up95pi6000 <- Y.hat6000 + qt(.975,n-2) * sigma.hat*sqrt(1+1/n+(6000-x.bar)^2/Sxx)
```

This gives the interval $(553.6752, 804.8008)$.

(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.

```
x.seq <- seq(min(x),max(x),length=500)
Y.hat.x.seq <- beta0.hat + beta1.hat * x.seq

se.x.seq <- sigma.hat * sqrt(1/n + (x.seq - x.bar)^2 / Sxx)

lo99conf <- Y.hat.x.seq - qt(.995,n-2) * se.x.seq
up99conf <- Y.hat.x.seq + qt(.995,n-2) * se.x.seq

pred.se.x.seq <- sigma.hat * sqrt(1 + 1/n + (x.seq - x.bar)^2 / Sxx)

lo99pred <- Y.hat.x.seq - qt(.995,n-2) * pred.se.x.seq
up99pred <- Y.hat.x.seq + qt(.995,n-2) * pred.se.x.seq

plot(Y ~ x, xlab="body surface area (cm^2)",ylab="metabolic rate (kcal/day)")
abline(beta0.hat,beta1.hat)
lines(lo99conf~x.seq,lty=2)
lines(up99conf~x.seq,lty=2)
lines(lo99pred~x.seq,lty=3)
lines(up99pred~x.seq,lty=3)
```



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 $\text{Normal}(0, \sigma^2)$ 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 $x_i = 0$.

If $x_i = 0$ then $\mathbb{E}Y_i = \beta_0$.

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 $x_i = 1$.

If $x_i = 1$ then $\mathbb{E}Y_i = \beta_0 + \beta_1$.

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

The parameter β_1 represents the expected value of the response under the treatment minus the expected value of the response under the control.

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

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

Let

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

Now set the partial derivatives of $Q_n(\beta_0, \beta_1)$ with respect to β_0 and β_1 equal to zero and solve the system of equations:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} Q_n(\beta_0, \beta_1) &= -2 \sum_{\{i:x_i=0\}} (Y_i - \beta_0) - 2 \sum_{\{i:x_i=1\}} [Y_i - (\beta_0 + \beta_1)] \\
&= -2[n_1(\bar{Y}_1 - \beta_0) + n_2(\bar{Y}_2 - \beta_0 - \beta_1)] = 0 \\
\frac{\partial}{\partial \beta_1} Q_n(\beta_0, \beta_1) &= -2 \sum_{\{i:x_i=1\}} [Y_i - (\beta_0 + \beta_1)] \\
&= -2n_2(\bar{Y}_2 - \beta_0 - \beta_1) = 0.
\end{aligned}$$

The second equation gives $\beta_1 = \bar{Y}_2 - \beta_0$. Plugging this into the first equation gives

$$-2[n_1(\bar{Y}_1 - \beta_0) + n_2(\bar{Y}_2 - \beta_0 - \beta_1)] = -2n_1(\bar{Y}_1 - \beta_0) = 0,$$

which gives $\beta_0 = \bar{Y}_1$. Therefore, we have

$$(\hat{\beta}_0, \hat{\beta}_1) = (\bar{Y}_1, \bar{Y}_2 - \bar{Y}_1).$$

(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.

   We get $\hat{\beta}_0 = 0.08$ and $\hat{\beta}_1 = 6.598947$.

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 $\hat{\sigma}$.

   We get $\hat{\sigma} = 5.229006$.

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

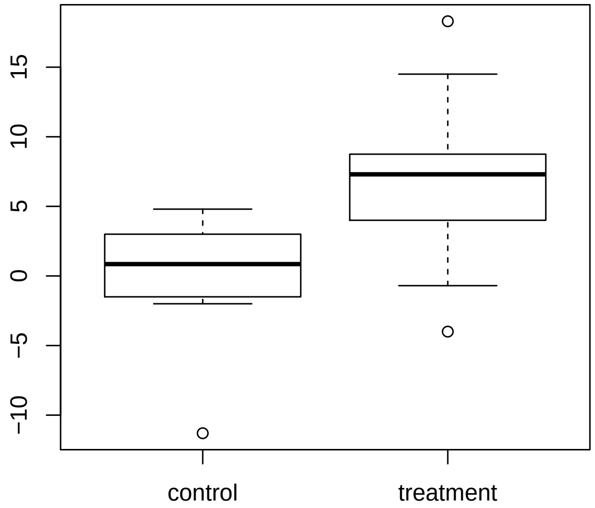
   The p-value is 0.003243934. There appears to be fairly strong evidence in favor of the apple treatment for hair growth! Will definitely bookmark this homework for later in life.

(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.

We have $\bar{Y}_1 = 0.08$ and $\bar{Y}_2 = 6.678947$.

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.

```
n1 <- length(ctrl)
n2 <- length(trt)

Y1.bar <- mean(ctrl)
Y2.bar <- mean(trt)

S.pooled <- sqrt(  (var(ctrl)*(n1-1) + var(trt)*(n2-1))/(n1+n2-2)  )
```

We get $S_{\text{pooled}} = 5.229006$. Wow! This is the same value that we got for $\hat{\sigma}$. My head is about to explode!

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.

```
T <- (Y1.bar - Y2.bar)/(S.pooled * sqrt(1/n1 + 1/n2) )
2*(1 - pt(abs(T),n1+n2-2))
```

The p-value is 0.003243934. That's the same p-value we got before! It seems we can use the framework of the linear regression model to do some of the things which we have already done. Sure is pretty neat stuff!

v. Keep your head from exploding.

It was hard, but I managed.

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.