Tutorial Problems

Question 1

The effect of height (H) and weight (W) on catheter length (L) on $n = 12$ children with congenital heart disease was analyzed using two models:

$$\text{Model 1:} \qquad \mathtt{L}_i = \beta_0 + \beta_1 \mathtt{W}_i + \beta_2 \mathtt{H}_i + \epsilon_i, \quad \epsilon_i \sim NID(0, \sigma_1^2)$$

and

$$\text{Model 2:} \qquad \mathtt{L}_i = \gamma_0 + \gamma_1 \mathtt{W}_i + \varepsilon_i, \quad \varepsilon_i \sim NID(0, \sigma_2^2).$$

Assume that both models pass the diagnostic tests. Use the output after this question only (i.e. do not use R) to answer the following questions.

(a) Write down the fitted models, including the estimates of the error variances. (Lecture 4 & 6)
(b) Calculate a 90% confidence interval for β_2. What can you conclude from your interval? (Lecture 4)
(c) What is the multiple correlation coefficient between L and (H, W)? (Lecture 7)
(d) Give a 95% confidence interval for σ_2. (Lecture 4)
(e) Calculate the sample coefficient of correlation between the L and W values. (Assumed knowledge)
(f) Calculate a 90% confidence interval for the expected catheter length when the weight is 90 for Model 2. (Lecture 4)

Question 1 Output

```
dat <- data.frame(
  "H"=c(42.8, 63.5, 37.5, 39.5, 45.5, 38.5, 43, 22.5, 37, 23.5, 33, 58),
  "W"=c(40, 93.5, 35.5, 30, 52, 17, 38.5, 8.5, 33, 9.5, 21, 79),
  "L"=c(37, 49.5, 34.5, 36, 43, 28, 37, 20, 33.5, 30.5, 38.5, 47))

fit1 <- lm(L ~ W + H, data = dat)
summary(fit1)
```

```
##
## Call:
## lm(formula = L ~ W + H, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.048 -1.258 -0.259  1.899  7.004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0084     8.7512   2.401   0.0399 *
## W             0.1908     0.1652   1.155   0.2777
## H             0.1964     0.3606   0.545   0.5993
## ---
```



```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.943 on 9 degrees of freedom
## Multiple R-squared:  0.8053, Adjusted R-squared:  0.7621
## F-statistic: 18.62 on 2 and 9 DF,  p-value: 0.0006336
```

```
fit2 <- lm(L ~ W, data=dat)
summary(fit2)
```

```
##
## Call:
## lm(formula = L ~ W, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.994 -1.481 -0.135  2.091  7.040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.63746    2.00421  12.792 1.60e-07 ***
## W            0.27727    0.04399   6.303 8.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.802 on 10 degrees of freedom
## Multiple R-squared:  0.7989, Adjusted R-squared:  0.7788
## F-statistic: 39.73 on 1 and 10 DF,  p-value: 8.871e-05
```

```
c(qt(0.95, 9), qt(0.95, 9), qt(0.95, 10), qt(0.95, 11))
```

```
## [1] 1.833113 1.833113 1.812461 1.795885
```

```
c(qchisq(0.025, 9), qchisq(0.975, 9), qchisq(0.025, 10), qchisq(0.975, 10))
```

```
## [1]  2.700389 19.022768  3.246973 20.483177
```

```
X <- model.matrix(~ W, data=dat)
solve(t(X) %*% X)
```

```
##              (Intercept)             W
## (Intercept)  0.277938162 -0.0051043889
## W           -0.005104389  0.0001338856
```

Question 2

The aim of a study of weekly fuel consumption (Y in tons) was to develop a prediction equation to predict Y on the basis of x_1 (average hourly temperature, $^\circ F$) and x_2 (the wind chill index). Observations were taken over 8 weeks. Assume

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \;\; \epsilon_i \sim NID(0, \sigma^2), \;\; i = 1, 2, \ldots, 8.$$

The fitted least squares multiple regression equation is

$$Y_i = 13.109 - 0.09 x_{i1} + 0.083 x_{i2}.$$

The residual sum of squares is 0.674, $\bar{x}_1 = 43.975$, $\bar{x}_2 = 12.875$ and

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 5.434 & 0.0859 & 0.1189 \\ 0.0859 & 0.00147 & 0.0016 \\ 0.1189 & 0.0016 & 0.00359 \end{pmatrix},$$

where $\mathbf{X}$ is the 8×3 design-matrix associated with the above multiple regression. Note that

```
qt(0.975, 5)
```

```
## [1] 2.570582
```

(a) Give an estimate for σ^2. (Lecture 4)
(b) Obtain a 95% CI for β_1. (Lecture 4)
(c) Test the hypothesis that $\beta_2 = 0$ against the alternative that β_2 is positive using a false positive rate of 0.05. (Lecture 4)
(d) Can the above model be simplified? (Lecture 4)
(e) Give an estimate for the average fuel consumption if the average hourly temperature is $40(^\circ F)$ and the chill index is 20. (Lecture 4)

Computer Problems

For the following questions, use the `olympic.txt` dataset that consists of the winning heights or distances (in inches) for the High Jump, Discus and Long Jump events at the Olympics up to 1996.

Question 1

In an experiment to determine the source from which corn plants in various soils obtain their phosphorous, the concentration of inorganic phosphorous
(x_1) and of two types of organic phosphorous (x_2, x_3) in the soil, and also the phosphorous content (y) of the plants, were measured. The data are

```
x1 <- c(0.4, 0.4, 3.1, 0.6, 4.7, 1.7, 9.4, 10.1, 11.6, 12.6, 13.8, 10.9, 23.1, 29.9)
x2 <- c(53, 23, 19, 34, 24, 65, 44, 31, 29, 58, 55, 37, 46, 51)
x3 <- c(158, 163, 37, 157, 59, 123, 46, 117, 173, 112, 117, 111, 114, 124)
y <- c(64, 60, 71, 61, 54, 77, 81, 93, 93, 51, 60, 76, 96, 99)
```

(a) Create a data-frame `dat`, produce the empirical covariance and correlation matrix (`var` and `cor`), and have a look at the pairwise scatterplots (`pairs` or otherwise). Comment on these in terms of obvious outliers, shape of plots and homoscedasticity.

(b) Use `lm()` to fit the regression model

$$Y_i = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \epsilon_i\,, \quad \epsilon_i \sim NID(0, \sigma^2),$$

show the summary of the `lm` output and give the fitted value and residual for the first observation.

(c) Calculate a 95% confidence interval for β_3.

(d) What is the square of the multiple correlation coefficient for this model?

(e) What observation has fourth largest leverage and are there any high leverage points? (a) Using the `lm()` output it seems that x_3 can be dropped from the model. It also seems that x_2 can be dropped. Why?

(f) Determine a reasonable model with x_1 alone for predicting the expected phosphorous content of the plants.What is the R^2 value for your final model? How does this compare with your answer in (d)? Estimate the error standard deviation and compare it with the estimate in (b).

(g) For the regression of phosphorous content on inorganic phosphorous only calculate a 94% prediction interval for $Y|x_1 = 40$.

(h) Show that for the full model the largest Cook's distance is 0.63 and for the simple linear regression model 0.19 (2dp). Why are the values different?

