UNIVERSITY OF MASSACHUSETTS

Department of Mathematics and Statistics

Basic Exam - Applied Statistics

Tuesday, January 17, 2017

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. This question is about the relationship between a baby's age and the length of its foot. Figure 1 shows data for $n = 39$ babies.

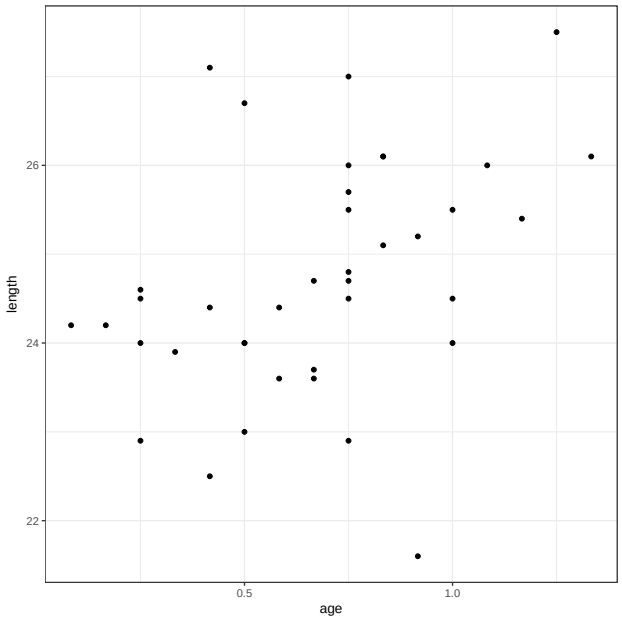


Figure 1: foot length (cm) versus age (yr)

Linear regression output is shown below.

```
Call:
lm(formula = length ~ age, data = feet)


Residuals:
    Min      1Q  Median      3Q     Max
-3.5355 -0.6058 -0.0159  0.5841  2.8037


Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.5970     0.4828  48.880   <2e-16 ***
age           1.6783     0.6568   2.555   0.0149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.231 on 37 degrees of freedom
Multiple R-squared:   0.15,Adjusted R-squared:  0.127
F-statistic: 6.529 on 1 and 37 DF,  p-value: 0.01485
```

(a) (5pts) The parameters are the intercept β_0, the slope β_1, and the error SD σ. From the regression output, what are the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}$?

(b) (5pts) The X matrix for this regression is

$$X = \begin{bmatrix} 1 & \text{age}_1 \\ 1 & \text{age}_2 \\ \vdots & \vdots \\ 1 & \text{age}_{39} \end{bmatrix}$$

Using matrix notation in terms of X and $\hat{\sigma}^2$, what is the estimated covariance matrix for the vector $(\hat{\beta}_0, \hat{\beta}_1)^t$?

(c) (5pts) We're interested in estimating θ, the average footlength of babies who are 0.75 years old. Write a simple formula for $\hat{\theta}$ in terms of $(\hat{\beta}_0, \hat{\beta}_1)^t$. For these data it turns out that $\hat{\theta} \approx 24.86$.

(d) (5pts) Using your answers to parts (b) and (c) write a formula for the estimated SD of $\hat{\theta}$. For these data it turns out that the estimated SD of $\hat{\theta}$, according to this formula, is about 0.20.

(e) (5pts) Another way to estimate θ, call it $\hat{\theta}_2$, is to calculate the mean footlength of the eight babies who are 0.75 years old. It turns out $\hat{\theta}_2 \approx 25.14$. Assuming the SD of the sample is the same as the residual SD from the regression, show how to derive that the estimated SD of $\hat{\theta}_2 \approx 0.43$.

(f) (7pts) Suppose that we have also recorded gender of the 39 babies. It's possible that the relationship between age and length is different for boys than for girls. If so, then we should include a second predictor in the linear model: `boy`, a variable that would be 1 for a boy and 0 for a girl.

(i) Write a linear model, using notation similar to `length ~ age`, but expanded to include `boy`, that allows for two parallel lines, one for boys and one for girls.

(ii) Write a linear model using similar notation that allows for two nonparallel lines, one for boys and one for girls. Denote this linear model as model $(*)$.

(g) (7pts) Consider the following statements

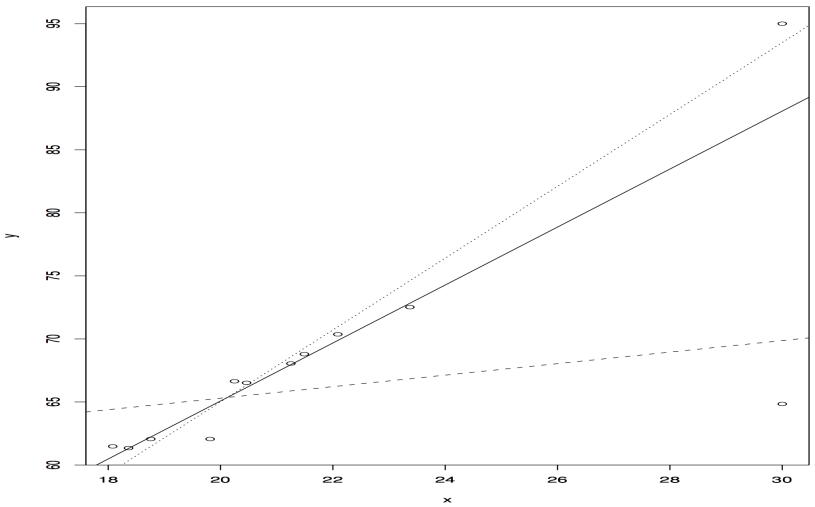
(a) For every age, there is no difference in the mean foot length for boys and girls

(b) The effect of age on the foot length does not depend on gender.

Translate (a) and (b) into two hypotheses on the parameters of model $(*)$.

(h) (7pts) Let $(y_i, x_{1i}, x_{2i}), i = 1, ..., 39$, be the 39 observations on foot length, age and gender. Assume that the error term is normally distributed with a mean of zero and a variance of σ^2. How would you test the hypotheses for (a) and (b) in the previous item? What are the test statistics and their null distributions?

2. (15pts) Consider the figure below. The solid line shows the regression of y on x based on only the scatter of data points in the left-hand portion of the plot. The dotted line shows the same regression as the first, but with the point in the upper right-hand corner included too. Similarly, the dashed line shows the same regression as the first, but with the point in the lower right-hand corner included too. Comment on the degree of (i) outlying-ness, (ii) leverage, and (iii) influence of the point in the upper right-hand corner. Do the same for the point in the lower right-hand corner. Justify your answer through appropriate description of the likely values of the statistics t_i, h_{ii}, and D_i. (That is, the studentized residual value, the hat-matrix entry, and Cook's distance.)



3. Suppose that in an experimental study you suspect that many observations were tainted by a technician and now you want to test them *jointly* for being outliers. To this end, you organize the suspected observations as the last q observations from a total of n and adopt a *mean shift outlier model* (MSOM) on these last observations:

$$\begin{aligned} y_1 &= \mathbf{x}_1^T\beta + \epsilon_1 \\ &\vdots \\ y_{n-q} &= \mathbf{x}_{n-q}^T\beta + \epsilon_{n-q} \\ y_{n-q+1} &= \mathbf{x}_{n-q+1}^T\beta + \delta_1 + \epsilon_{n-q+1} \\ &\vdots \\ y_n &= \mathbf{x}_n^T\beta + \delta_q + \epsilon_n \end{aligned}$$

This model can be specified in matrix form by

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & I_q \end{bmatrix} \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \epsilon$$

where $E(\epsilon|X) = 0$ and $\text{Var}(\epsilon|X) = \sigma^2 I_n$, $X = \begin{bmatrix} X_1 & 0 \\ X_2 & I_q \end{bmatrix}$, and $\delta = (\delta_1, \cdots, \delta_q)^T$. After some algebra, we can show that

$$(X^TX)^{-1} = \begin{bmatrix} (X_1^TX_1)^{-1} & -(X_1^TX_1)^{-1}X_2^T \\ -X^2(X_1^TX_1)^{-1} & I_q + X_2(X_1^TX_1)^{-1}X_2^T \end{bmatrix}.$$

Now consider $\hat{\beta}$ and $\hat{\delta}$, the Least Square Estimator (LSE) for β and δ under this model, and $\hat{\beta}_1$, the LSE for β when regressing only $\mathbf{y}_1$ on X_1, that is, when ignoring the last q observations.

(a) (6pts) Show that (i) $\hat{\beta} = \hat{\beta}_1$ and (ii) $\hat{\delta} = \mathbf{y}_2 - X_2\hat{\beta}_1$, that is, the LSE for δ is the difference between the (removed) observed values and the fitted values for X_2 in the model without the last q suspected observations. *Hint: The general formula for the LSE of β in the regression model $\mathbf{y} = X\beta + \epsilon$ is $\hat{\beta} = (X^TX)^{-1}X\mathbf{y}$.*

(b) (6pts) Show that the last q observations are *perfectly* fit by the MSOM: $\hat{\mathbf{y}}_2 = X_2\hat{\beta} + \hat{\delta} = \mathbf{y}_2$. What can you say about the relation between the LSE $\hat{\sigma}^2$ for σ^2 under the MSOM and the LSE $\hat{\sigma}_1^2$ for σ^2 under the model without the last q observations?

(c) (6pts) Find the hat matrix for the MSOM and comment on the leverage for the suspected data

points in light of the results from the previous item. *Hint: The hat matrix H for a general regression model $\mathbf{y} = X\beta + \epsilon$ is the matrix such that $\hat{\mathbf{y}} = H\mathbf{y}$. Because $\hat{\mathbf{y}} = X\hat{\beta} = X(X^TX)^{-1}X\mathbf{y}$, we have $H = X(X^TX)^{-1}X$.*

(d) (6pts) Conduct a joint outlier test by testing $\delta_1 = \cdots = \delta_q = 0$. State the test statistic and its distribution under the null.

4. (15pts) In the previous question, assume we have observations $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$ and we suspect the last $q = 2$ observations might be tainted. Write R code to perform the hypothesis test in part (d) of the question 3.