Homework 12
Due Friday Dec 7

(1) Show that the least squares estimates that minimize $Q = \sum_i [Y_i - (\beta_0 + \beta_1 x_i)]^2$ are given by

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}.$$

(2) Show that

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x}) Y_i}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i x_i (Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2}.$$

(3) Suppose that

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. Find the Fisher information matrix $I(\beta_0, \beta_1, \sigma)$ based on the joint density of the data $\prod_i f(y_i | x_i;\ \beta_0, \beta_1, \sigma)$. Use $I(\beta_0, \beta_1, \sigma)$ to find the asymptotic standard error of $\hat{\beta}_1$. How does this compare to the exact formula for the standard error?

(4) If the Universe is expanding, as predicted by general relativity, then all galaxies should be receding from us. Furthermore, the velocity with which each galaxy is moving away from us should be proportional to its distance from us. (Think of dots on a balloon and imagine blowing up the balloon.) This leads to Hubble's law:

$$\text{Recession Velocity} = H_0 \times \text{Distance}$$

for some constant H_0. In 1929 Edwin Hubble investigated this relationship by collecting data on 24 galaxies. The data can be obtained from:

http://lib.stat.cmu.edu/DASL/Datafiles/Hubble.html

The variables are distance (in megaparsecs = 3.26 light years) and velocity (in km/sec).

(4a) Do a linear regression of velocity on distance. Get $\hat{\beta}_0$, $\hat{\beta}_1$, their standard errors and 95 per cent confidence intervals for β_0 and β_1. First, write your own R program using the formulas we derived in class. Then compare your answers to the built-in R programs.

Here is how to use the built-in routines:

```
data <- scan("hubble.dat")           ### get data and put it in a matr
data <- matrix(data,ncol=2,byrow=T)
distance <- data[,1]
velocity <- data[,2]
plot(distance,velocity)
out <- lm(velocity ~ distance)       ### do the linear regression
abline(out)                          ### add regression line to the pl
names(out)
temp <- summary(out)
print(temp)
names(temp)
sigma.hat <- temp$sigma              ### get sigma.hat
### plot the residuals (standardized to have st. dev 1)
### These should look roughly like random N(0,1) numbers
residuals <- residuals(out)
plot(distance,residuals/sigma.hat,ylim=c(-4,4))
beta0.hat <- out$coeff[1]
beta1.hat <- out$coeff[2]
var.covar <- sigma.hat^2*temp$cov  ### this is how to extract the
                                   ### variance-covariance matrix
```

Note: The age of the universe, in billions of years, is $978/H_0$. With your estimate of H_0 this would make the Universe 2 billion years old, which is far too small. Even the Earth is older than 2 billion years. Thus, people were skepical about Hubble's value for H_0. However, Hubble made some errors in calibrating distances. Correcting for these errors has led to a new estimates of H_0 around 65 suggesting that the age of the Universe is 15 billion years.

(4b) Hubble's law says that $Y = \beta_0 + \beta_1 X$ where $\beta_1 = H_0$ and $\beta_0 = 0$. This implies that there should be no intercept in the model. Is your fit consistent with this, i.e. does your confidence interval for β_0 include 0?

(4c) We can force the regression line to omit the intercept. This is called "regression through the origin." The model is:

$$Y = \beta_1 x + \epsilon.$$

Derive the least squares estimate $\hat{\beta}_1$ for this model and its standard error.

(4d) Fit the Hubble data to the regression-through-the-origin model. Get the estimate, standard error and fitted value. First use your formulas. The use the R built-in functions. To get R to fit this the model, you have to do this:

```
D    <- matrix(distance,ncol=1)
junk <- lm.fit(x=D,y=velocity)
beta1.hat <- junk$coef
plot(distance,velocity)
abline(a=0,b=beta1.hat)
```