STAT 8230 — Applied Nonlinear Regression
Homework 1 – Due Tuesday, Sept. 6

Homework Guidelines:

- Homework is due by 4:30 on the due date specified above. You may turn it in at the beginning of class or place it in my mailbox in the Statistics Building. **No late homeworks will be accepted without permission granted prior to the due date.**

- Use only standard (8.5 × 11 inch) paper and use only one side of each sheet.

- Homework should show enough detail so that the reader can clearly understand the procedures of the solutions.

- Problems should appear in the order that they were assigned.

Assignment:

1. The following table contains a partial listing of data on the average life expectancies of men and women in the US over the period 1940–1992. The full data set is contained on the course website in the file hwk1-1.dat. Use R to do the computational aspects (model fitting, inferences) of the following problems.

Year (x_3)	Gender (x_2)	Life Expectancy (years) (y)
1940	1	65.2
1940	0	60.8
1950	1	71.1
1950	0	65.6
⋮	⋮	⋮
1991	1	78.9
1991	0	72.0
1992	1	79.1
1992	0	72.3

Consider the following classical linear model for these data

$$y_i = \alpha_1 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots e_n \overset{iid}{\sim} N(0, \sigma^2)$ and $x_{i4} = x_{i2}x_{i3}$. In this data set, both male and female life expectancies are available in 26 of the years between 1940 and 1992 (not every year), so there is a total of $n = 52$ observations. Here $x_2 = 1$ for females, $x_2 = 0$ for males.

a. What is the expectation function for $i = 1, 3, 5, \ldots, 51$ (the odd values of i, corresponding to females)?

b. What is the expectation function for $i = 2, 4, 6, \ldots, 52$ (the even values of i, corresponding to males)?

c. Interpret each of the parameters $\alpha_1, \alpha_2, \alpha_3$, and α_4.

d. Suppose we replace x_3 =year by $x_3^* \equiv x_3 - 1940$ and x_4 by $x_4^* \equiv x_2 x_3^*$ and consider the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}^* + \beta_4 x_{i4}^* + e_i \quad i = 1, \ldots, n \tag{$*$}$$

Interpret each of the parameters $\beta_1, \beta_2, \beta_3$, and β_4.

e. Describe in words what this model assumes about how average life expectancies differ between men and women and how they have changed over time during the period 1940–1992.

f. Fit model (*) to the data set in hwk1-1.dat using ordinary least-squares. Report the following results from your fitted model:

   i. regression parameter estimates and corresponding standard errors and the estimated error variance;
   ii. the estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$, the least-squares regression parameter estimate;
   iii. 95% confidence intervals for each regression parameter;
   iv. a plot of the data along with the estimated regression function and 95% simultaneous confidence bands for the regression function. Your plot should use different plotting symbols for males and females and should display separate regression functions and 95% bands for males and females on the same graph.

g. Express each of the following hypotheses in mathematic form in terms of the parameters of model (*), and test them at significance level $\alpha = 0.05$. State the conclusions of your tests.

   i. Hypothesis: males and females have the same mean life expectancy and the same linear rate of increase over time in mean life expectancy over the period 1940–1992.
   ii. Hypothesis: male and female life expectancies have changed linearly over time at the same rate over the period 1940–1992.

h. Plot the Pearson residuals versus fitted values, and the Pearson residuals versus year separately for each gender.

   i. Are there any observations with extremely large (in magnitude) residuals? Can you explain why the data point that has the largest magnitude residual is so poorly fit by the model?

ii. Is there any evidence in the residual plots that suggest that one or more model assumptions are violated for this data set? If so, which one(s)?

i. Now define y_{ij} = the life expectancy for the jth gender ($j = 1$ for males and $j = 2$ for females) in the ith year ($i = 1, \ldots, 26$), and define z_i = the year in which life expectancy y_{ij} was measured minus 1940. Consider the model

$$y_{ij} = \gamma_j + \delta_j z_i + e_{ij},$$

where $i = 1, \ldots, 26$, $j = 1, 2$ and $e_{11}, e_{12}, \ldots, e_{26,2} \overset{iid}{\sim} N(0, \sigma^2)$. Repeat part (e) of this problem for this model and then demonstrate that this model is equivalent to model (*) by fitting it to the data set and comparing the vector of fitted values to those from model (*).

2. (From Gelman & Hill, 2007) The Stata data file child.iq.dta (on the course website) contains data on children's IQ test scores at age 3, mother's educational level, and the mother's age at the time of the child's birth for a random sample of 400 children. Read these data into R by placing the .dta file in your working directory and typing

```
library(foreign)
iq.data <- read.dta("‘child.iq.dta"’)
```

You will first have to install the **foreign** package.

a. Fit a regression of child test scores on mother's age, display the data and fitted model, check assumptions of the model, and interpret the slope coefficient. At what age do you recommend that mother's give birth? What are you assuming in making such a recommendation?

b. Repeat part (a) but now using a model that further includes mother's education, interpreting both slope coefficients in the model. Have your conclusions about the timing of birth changed? In this data set mother's educational level is an ordered categorical variable (1= did not complete high school, 2=high school degree but no college degree, 3= college degree but no graduate degree, 4=one or more graduate degrees), but you may treat it as continuous for this part of the problem.

c. Now create an indicator variable reflecting whether the mother completed high school or not. Consider interactions between high school completion and mother's age. Also create a plot that shows the separate regression lines for each high school completion status group.

d. Finally, fit a regression of child test score on mother's age and education level for the first 200 children in the data set and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.