BIOSTAT 653 Homework #1

Due Monday September 26th, 3:10am, in class.

Problem 1

Suppose we conduct a cross-sectional study of test scores among students in elementary and junior high schools. On one day of testing, investigators administer grade-level mathematics tests to students who are 8 years old, 10 years old, 12 years old, and 14 years old. They hypothesize that performance may depend both on age and on gender. They fit the model

$$y_i = \beta_0 + \beta_1 I(boy_i) + \beta_2 age_i + \beta_3 age_i I(boy_i) + \epsilon_i$$

where $I(boy_i)$ is an indicator function that equals to 1 when i'th individual is a boy, and equals to 0 otherwise. Describe in words to a non-statistician the hypothesis tested by each choice of L and θ_0 below.

1) $L = (0\ 0\ 0\ 1), \theta_0 = 0$

2) $L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

3) $L = (1\ 1\ 10\ 10), \theta_0 = 80$

4) $L = (0\ 0\ 4\ 4), \theta_0 = 5$

5) $L = (0\ 0\ 4\ 0), \theta_0 = -5$

6) $L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \theta_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

Solution

1) To test whether there is an interaction effects between age and gender.
2) To test whether there is a gender effect.
3) To test whether the average test score for a 10 year old boy is 80.
4) To test whether for a boy, a four year increase in age results in an average of 5 unit increase in test score.
5) To test whether for a girl, a four year increase in age results in an average of 5 unit decrease in test score.
6) To test whether gender or age affects test score (or alternatively, whether there is gender or age or gender by age effects).

Problem 2

It is well-known that lead exposure may have a negative effect on IQ. However, it is not known if the effect of lead exposure is persistent and irreversible. To answer this question, investigators studied a group of children who lived near a lead smelter. Based on the blood-lead measurement, these children can be classified into three categories: unexposed (=1), currently exposed (=2), and previously exposed (=3). The investigators collected important information from these children (gender and age) and performed test to determine the IQ for each child.

You can find the data file (leadiq.txt) on canvas. Each row represents: ID, lead exposure category, gender (boy=0, girl=1), age (in years) and IQ.

The investigators wish to use a linear regression model to answer the following research questions:

a) Research Question 1: Is there an effect of lead exposure on IQ?
b) Research Question 2: Does the effect of lead exposure on IQ depend on gender?
c) Research Question 3: Does the effect of lead exposure on IQ decay with time?

Your task is to help the investigators answer these research questions. To do so,

1) Describe, justify and fit a linear regression model that can be used to address all these research questions.
2) Do we need to use age as a covariate? Why or why not?
3) Provide parameter estimates and interpret the results in language someone without a statistics background can understand.
4) What is the expected IQ for a boy without lead exposure? What is the expected IQ for a boy who is currently exposed with lead? What is the expected IQ for a boy who had lead exposure before?
5) Provide evidence that a linear regression model does properly fit the data.

Solution

1) We can fit the following linear model:

$$IQ = \beta_0 + \beta_1 I(exp = 2) + \beta_2 I(exp = 3) + \beta_3 I(gender = 0) + \beta_4 I(gender = 0) I(exp = 2) \\ + \beta_5 I(gender = 0) I(exp = 3)$$

For question 1, we test $H_0: \beta_1 = \beta_2 = \beta_4 = \beta_5 = 0$.

For question 2, we test $H_0: \beta_4 = \beta_5 = 0$.

For question 3, we test $H_0: \beta_1 = \beta_2, \beta_4 = \beta_5$, one-side test.

```
DATA leadiq;
      INFILE 'leadiq.txt';
      INPUT id $ lead gender age iq;
      leadexp=1; IF lead=1 THEN leadexp=0;
      curexp=0; IF lead=2 THEN curexp=1;
      prevexp=0; IF lead=3 THEN prevexp=1;
      cg=curexp*gender;
      pg=prevexp*gender;
RUN;


PROC REG data=leadiq;
      MODEL iq = curexp prevexp gender cg pg;
      TEST curexp=prevexp=cg=pg=0;
      TEST cg=pg=0;
      TEST curexp=prevexp;
RUN;
```

2) We don’t need to include age as age effect is not significant in the model. Alternatively, you can say that after consulting with the scientists, they believe age does not have an effect on IQ or influence the lead effect on IQ. Notice that it is not sufficient/correct to say that age is not related to the scientific questions. For example, if age is a confounding factor, then we have to include age in the model even if age is not related to any of the scientific questions.

3) The estimates are listed below. β_0 represents the average IQ of a boy who is not exposed to lead and is estimated to be 103.70. β_1 represents the average increase in IQ of a boy who is currently exposed to lead compared to a boy who is not exposed. β_2 represents the average increase in IQ of a boy who has been previously exposed to lead compared to a boy who is not exposed. β_3 represents the gender effect on IQ, or the average IQ difference between a girl and a boy. β_4 represents the average IQ increase in a girl who is currently exposed to lead compared with a boy who is currently exposed to lead. β_5 represents the average IQ increase in a girl who has been previously exposed to lead compared with a boy who has been previously exposed to lead.

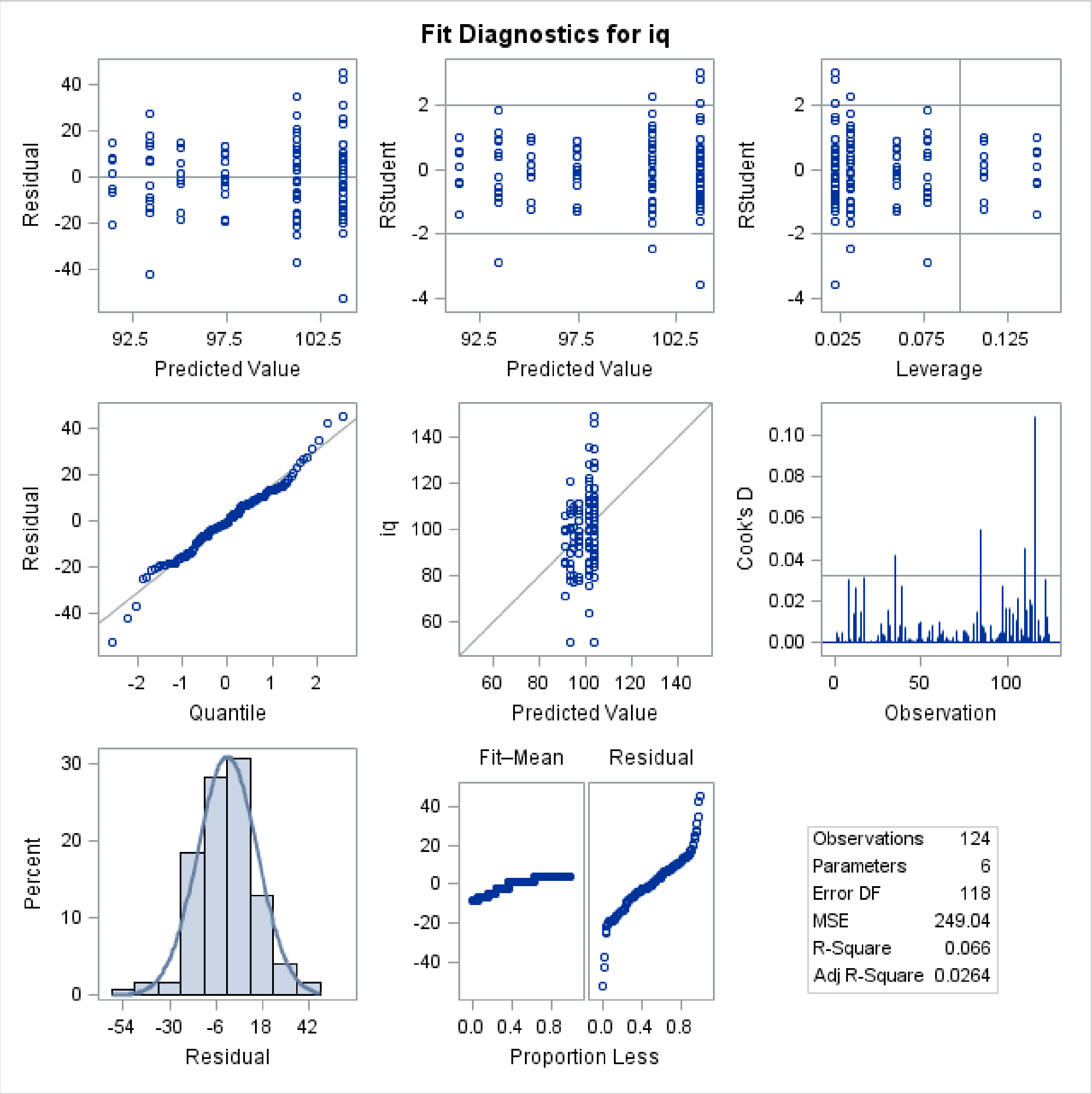
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	103.69565	2.32676	44.57	<.0001
curexp	**1**	-6.28389	4.47917	-1.40	0.1633
prevexp	**1**	-10.23411	4.95685	-2.06	0.0411
gender	**1**	-2.41440	3.63265	-0.66	0.5076
cg	**1**	-3.56879	7.96378	-0.45	0.6549
pg	**1**	4.06397	7.74747	0.52	0.6009

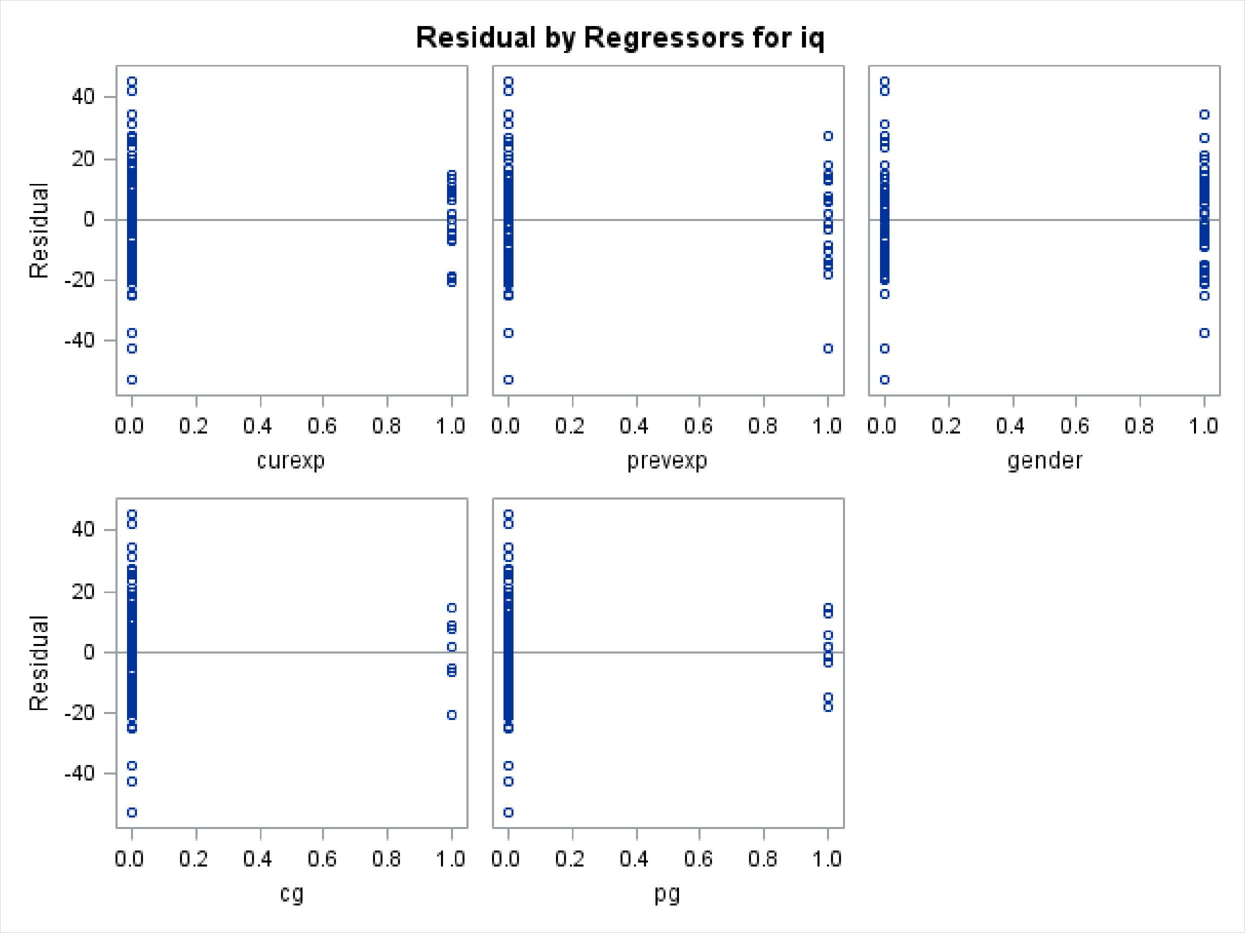
4) The expected IQ for a boy without lead exposure is (103.70). The expected IQ for a boy who is currently exposed with lead is (97.41). The expected IQ for a boy who had lead exposure before is (93.46).

5) The residual diagnostic plots are below. We want to show that the effect of each variable on IQ is linear (this is easy here because each variable is binary); the residual errors follow a normal

distribution (by looking at the QQ plot); the errors are heterogeneous (by looking at the residual plot).


Fit Diagnostics for iq

Observations	124
Parameters	6
Error DF	118
MSE	249.04
R-Square	0.066
Adj R-Square	0.0264



Problem 3

As we know from problem 2, lead exposure is dangerous for children. To help these children who lived near a lead smelter, the investigators decided to find out treatments that could reduce the blood-lead levels. One particular trial they performed was a placebo-controlled, randomized study of succimer (a chelating agent). They performed this study in children with relatively high blood lead levels. They collected measurements of blood lead levels in 100 children at four different time points: week 0 (a.k.a. baseline), week 1, week 4, and week 6. These 100 children were randomly assigned to chelation treatment with succimer or to placebo. For simplicity, however, we will focus only on the 50 children assigned to chelation treatment with succimer.

You can find the data file (lead.txt) for these 50 samples on canvas. Each row represents: ID, blood lead levels at week 0, blood lead levels at week 1, blood lead levels at week 4, blood lead levels at week 6.

1) Calculate the sample mean, standard deviation and variance of the blood lead levels at each occasion (i.e. time point).
2) Construct a time plot of the blood lead levels for all individuals over time. Construct a time plot of the mean blood lead level over time. Describe the general characteristics of the time trend.
3) Calculate the 4 by 4 covariance and correlation matrices for the four repeated measures of blood lead levels. Are the diagonal elements in the covariance matrix identical to the variance computed from (1)?

Solution

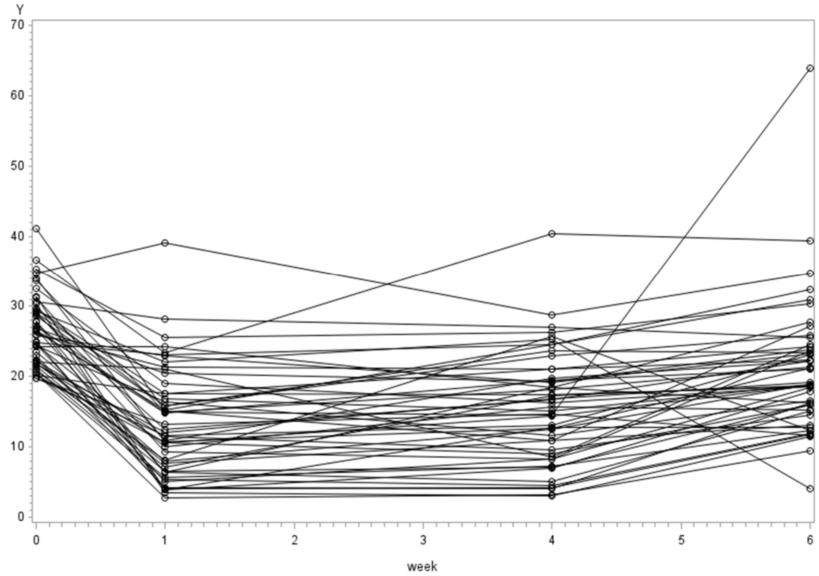
1) See SAS code and summary below.

```
DATA lead;
      INFILE 'lead.txt';
      INPUT id $ Y1 Y2 Y3 Y4;
RUN;

PROC MEANS DATA=lead MEAN STD VAR;
RUN;
```

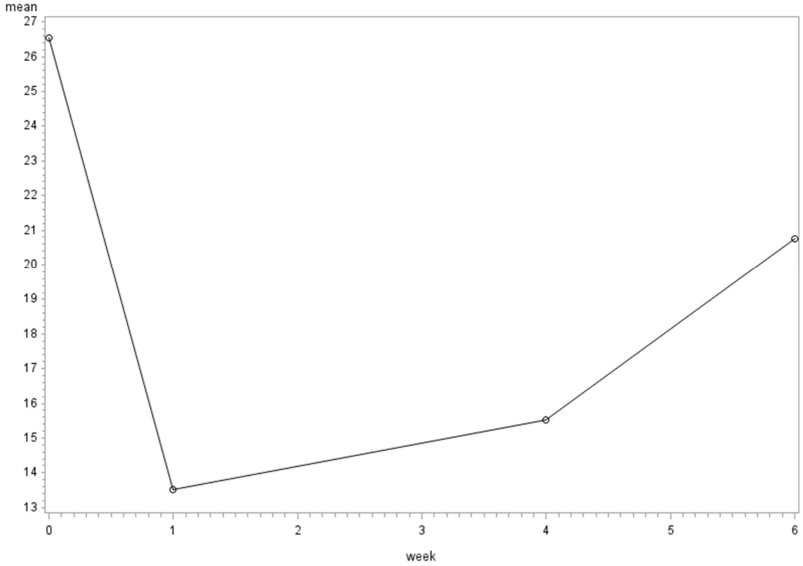
Variable	Mean	Std Dev	Variance
Y1	26.5400000	5.0209358	25.2097959
Y2	13.5220000	7.6724870	58.8670571
Y3	15.5140000	7.8522065	61.6571469
Y4	20.7620000	9.2463316	85.4946490

2) See SAS code and figure below. Pattern: a sharp decrease from week 0 to 1 and a slow increase from week 1 to 6.

```
DATA lead_long;
      SET lead;
      week=0;
      Y=Y1;
      output;
      week=1;
      Y=Y2;
      output;
      week=4;
      Y=Y3;
      output;
      week=6;
      Y=Y4;
      output;
      DROP Y1 Y2 Y3 Y4;
RUN;

symbol1 value = circle color = black interpol = join repeat = 50;
PROC GPLOT DATA=lead_long;
  PLOT Y*week = id / nolegend;
RUN;
```



```
PROC SORT DATA=lead_long;
      BY week;
RUN;

PROC MEANS DATA=lead_long NOPRINT;
      BY week;
      VAR Y;
      OUTPUT OUT=lead_mean mean=mean;
RUN;

symbol1 value = circle color = black interpol = join repeat = 4;
PROC GPLOT DATA=lead_mean;
  PLOT mean*week;
RUN;
```



3) See SAS code and summary below.

```
PROC CORR DATA=lead COV;
RUN;
```

```
                  Covariance Matrix, DF = 49

              Y1           Y2           Y3           Y4

Y1   25.20979592  15.46542857  15.13800000  22.98542857

Y2   15.46542857  58.86705714  44.02907347  35.96595510

Y3   15.13800000  44.02907347  61.65714694  33.02197143

Y4   22.98542857  35.96595510  33.02197143  85.49464898
```

```
        Pearson Correlation Coefficients, N = 50
              Prob > |r| under H0: Rho=0

             Y1          Y2          Y3          Y4

  Y1    1.00000     0.40146     0.38397     0.49511
                     0.0039      0.0059      0.0003

  Y2    0.40146     1.00000     0.73082     0.50697
         0.0039                  <.0001      0.0002

  Y3    0.38397     0.73082     1.00000     0.45482
         0.0059      <.0001                  0.0009

  Y4    0.49511     0.50697     0.45482     1.00000
         0.0003      0.0002      0.0009
```