BIOS 6611 Homework 11 – Exam Prep (Not Graded)

Nothing needs to be turned in, answer key posted on HW 11 Canvas page

A biologist wished to study the effects of the temperature of a certain medium on the growth of human amniotic cells in a tissue culture. Using the same parent batch, she conducted an experiment in which five cell lines were cultured at each of four temperatures. The procedure involved initially inoculating a fixed number (0.25 million) of cells into a fresh culture flask and then, after 7 days, removing a small sample from the growing surface to use in estimating the total number of cells in the flask. The results are given in the following table:

Number of cells ($\times 10^6$) after 7 Days. [i.e., 1.13 = 1.13 million cells]

Temperature			
40º	**60º**	**80º**	**100º**
1.13	1.75	2.30	3.18
1.20	1.45	2.15	3.10
1.00	1.55	2.25	3.28
0.91	1.64	2.40	3.35
1.05	1.60	2.49	3.12

1) Examine the relationship between temperature and cell growth.

   A) Fit a straight-line regression model regressing cell growth (number of cells after 7 days -- the dependent variable Y) on temperature (the independent variable X). Write a brief summary of the relationship between cell growth and temperature for the straight-line model.

   B) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals). Is there any evidence that any of the regression assumptions are violated? What possible remedies would you recommend exploring if you detect a violation? [*See example code below for producing these plots in SAS*]

   C) Test for lack-of-fit of the straight-line model. State your conclusion. What is the sum of squares due to pure error? Due to lack of fit of the straight-line model?

   D) Fit the quadratic regression model. Test for the significance of adding temperature^2 (X^2) to the model. Test for lack-of-fit of the quadratic model. State your conclusions.

   E) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals) for the quadratic model. Is there any evidence that any of the regression assumptions are violated?

F) Based on parts A-E, which model is most appropriate -- straight-line, quadratic, or cubic?

G) Use an orthogonal polynomial model to choose between the straight-line, quadratic, and cubic models. Discuss any similarities and/or differences between using this model versus the "lack-of-fit" tests in parts C and D for making this choice.

2) Create a new variable containing the natural logarithm of the number of cells. Perform a straight-line linear regression of the natural logarithm of number of cells on temperature (NOTE: it is a straight-line regression on the log scale, not on the original scale).

   A) Write the regression equation <u>in the log scale</u>. What are the estimates of the intercept and slope and how would you interpret them? Next, transform the estimate of the slope and its 95% confidence interval to the original (not logged) scale. How would you interpret these?

   B) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals). Is there any evidence that any of the regression assumptions are violated?

   C) Write a brief summary describing the relationship between cell growth and temperature for this model (***<u>use transformed results!</u>***).

   D) Test for lack-of-fit of the straight-line model (on the log scale). Which model (using the log transformed outcome) is most appropriate – straight line, quadratic, or cubic?

   E) Do you prefer the model you chose in part 1F or 2D? Explain.

Generic Example for Producing Diagnostic Plots in SAS

```
/* Create Output Data Set for Studentized Deleted (Jackknife)
Residuals */
PROC REG DATA = dataset;
    MODEL yvar = xvar(s);
    OUTPUT OUT= outdataset RSTUDENT = jackknife;
RUN;


/* Y-X scatterplot with linear regression line */
PROC GPLOT DATA = dataset;
    PLOT yvar*xvar;
    SYMBOL INTERPOL=rl VALUE=dot COLOR=black;
RUN;


/* Jackknife Residual Plot */
PROC GPLOT DATA = outdataset;
    PLOT jackknife*xvar;
    SYMBOL VALUE=dot INTERPOL=rl COLOR=black;
RUN;


/* Histogram of Jackknife Residuals */
PROC GCHART DATA= outdataset;
    VBAR jackknife;
RUN;


/* Normal Probability Plot of Jackknife Residuals */
PROC UNIVARIATE NORMAL PLOT DATA= outdataset;
    VAR jackknife;
RUN;
```

```
*************************************;
***  BIOS 6611                     ***;
***  Assignment #11 SAS code       ***;
*************************************;

DATA amniotic;
    INPUT cells temp;
    temp2=temp**2;
    temp3=temp**3;

    IF temp=40 THEN
        DO;
            lin=-3;
            quad=1;
            cubic=-1;
        END;

    IF temp=60 THEN
        DO;
            lin=-1;
            quad=-1;
            cubic=3;
        END;

    IF temp=80 THEN
        DO;
            lin=1;
            quad=-1;
            cubic=-3;
        END;

    IF temp=100 THEN
        DO;
            lin=3;
            quad=1;
            cubic=1;
        END;
    lncells=log(cells);
    cellsn=cells*10**6;
    DATALINES;
 1.13 40
 1.20 40
 1.00 40
 0.91 40
 1.05 40
 1.75 60
 1.45 60
```

```
 1.55 60
 1.64 60
 1.60 60
 2.30 80
 2.15 80
 2.25 80
 2.40 80
 2.49 80
 3.18 100
 3.10 100
 3.28 100
 3.35 100
 3.12 100
;
```

NOTE: Some plots are automatic in PROC REG, so be sure to check those out. It's always good, though, to build your SAS (and R!) coding repertoire, so try out some of the code below for producing graphs. ☺

```
/* QUESTION 1A */
PROC REG DATA=amniotic;
    MODEL cells=temp /clb;
    OUTPUT OUT=resids1 PREDICTED=pred RSTUDENT=jackknife;
RUN;

/* QUESTION 1B */
/* Y-X scatterplot with LINEAR regression line */
PROC GPLOT DATA=amniotic;
    PLOT cells*temp;
    SYMBOL INTERPOL=rl VALUE=dot COLOR=black;
RUN;

/* -OR- */
PROC SGPLOT DATA=amniotic;
    REG Y=cells X=temp;
RUN;

/* Jackknife Residual Plot versus Predictor, versus Predicted */
PROC GPLOT DATA=resids1;
    PLOT jackknife*(temp pred);
    SYMBOL VALUE=dot INTERPOL=rl COLOR=black;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids1;
    REG Y=jackknife X=temp;
RUN;
```

```
PROC SGPLOT DATA=resids1;
    REG Y=jackknife X=pred;
RUN;

/* Histogram of Jackknife Residuals */
PROC GCHART DATA=resids1;
    VBAR jackknife;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids1;
    histogram jackknife;
    density jackknife;
RUN;

/* Normal Probability Plot of Jackknife Residuals */
PROC UNIVARIATE NORMAL PLOT DATA=resids1;
    VAR jackknife;
RUN;

/* QUESTION 1C & 1D */
PROC REG DATA=amniotic;
    MODEL cells=temp temp2 temp3;
    LOF_linear: TEST temp2, temp3;  /* 1c */
    LOF_quad: TEST temp3; /* 1d */
RUN;

/* QUESTION 1D */
PROC SORT DATA=amniotic;
    BY temp cells;
RUN;

PROC REG DATA=amniotic;
    MODEL cells=temp temp2;
    OUTPUT OUT=resids2 PREDICTED=pred RSTUDENT=jackknife;
RUN;

***ALL;
PROC REG DATA=amniotic;
    MODEL cells=temp;
    MODEL cells=lin;
    MODEL cells=temp temp2;
    MODEL cells=lin quad;
    MODEL cells=temp temp2 temp3;
    MODEL cells=lin quad cubic;
RUN;
```

```
/* QUESTION 1E */
/* Y-X scatterplot with QUADRATIC regression line */
PROC GPLOT DATA=resids2;
    PLOT cells*temp /overlay;
    SYMBOL INTERPOL=rq VALUE=dot COLOR=black;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids2;
    REG Y=cells X=temp / degree=2;
RUN;

/* Jackknife Residual Plot versus Predicted */
PROC GPLOT DATA=resids2;
    PLOT jackknife*(pred);
    SYMBOL VALUE=dot INTERPOL=rl COLOR=black;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids2;
    REG Y=jackknife X=pred;
RUN;

/* Histogram of Jackknife Residuals */
PROC GCHART DATA=resids2;
    VBAR jackknife;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids2;
    histogram jackknife;
    density jackknife;
RUN;

/* Normal Probability Plot of Jackknife Residuals */
PROC UNIVARIATE NORMAL PLOT DATA=resids2;
    VAR jackknife;
RUN;

/* QUESTION 1G */
PROC REG DATA=amniotic;
    MODEL cells=lin quad cubic;
RUN;

/* QUESTION 2A */
PROC REG DATA=amniotic;
    MODEL lncells=temp /clb;
    OUTPUT OUT=resids3 PREDICTED=pred RSTUDENT=jackknife;
RUN;
```

```
/* QUESTION 2B */
/* Y-X scatterplot with LINEAR regression line */
PROC GPLOT DATA=amniotic;
    PLOT lncells*temp;
    SYMBOL INTERPOL=rl VALUE=dot COLOR=black;
RUN;
/* -OR- */
PROC SGPLOT DATA=amniotic;
    REG Y=lncells X=temp;
RUN;

/* Jackknife Residual Plot versus Predictor, versus Predicted */
PROC GPLOT DATA=resids3;
    PLOT jackknife*(temp pred);
    SYMBOL VALUE=dot INTERPOL=rl COLOR=black;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids3;
    REG Y=jackknife X=temp;
RUN;

PROC SGPLOT DATA=resids3;
    REG Y=jackknife X=pred;
RUN;

/* Histogram of Jackknife Residuals */
PROC GCHART DATA=resids3;
    VBAR jackknife;
RUN;
/* -OR- */
PROC SGPLOT DATA=resids3;
    histogram jackknife;
    density jackknife;
RUN;

/* Normal Probability Plot of Jackknife Residuals */
PROC UNIVARIATE NORMAL PLOT DATA=resids3;
    VAR jackknife;
RUN;

/* QUESTION 2D */
PROC REG DATA=amniotic;
    MODEL lncells=temp temp2 temp3;
    LOF_linear: TEST temp2, temp3;
    LOR_quad: TEST temp3;
RUN;
```

```
/*Compare with*/
PROC REG DATA=amniotic;
    MODEL lncells=lin quad cubic;
    LOF_linear: TEST quad, cubic;
    LOF_quad: TEST cubic;
RUN;
```