BIOS 6611 Homework 11 – Exam Prep *Answer Key*

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

Generate data set and create relevant variables needed for analyses:

```
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
;
```

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.

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

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	12.83072	12.83072	630.72	<.0001
Error	18	0.36618	0.02034		
Corrected Total	19	13.19690			

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-0.46240	0.10481	-4.41	0.0003	-0.68260	-0.24220
temp	1	0.03582	0.00143	25.11	<.0001	0.03282	0.03882

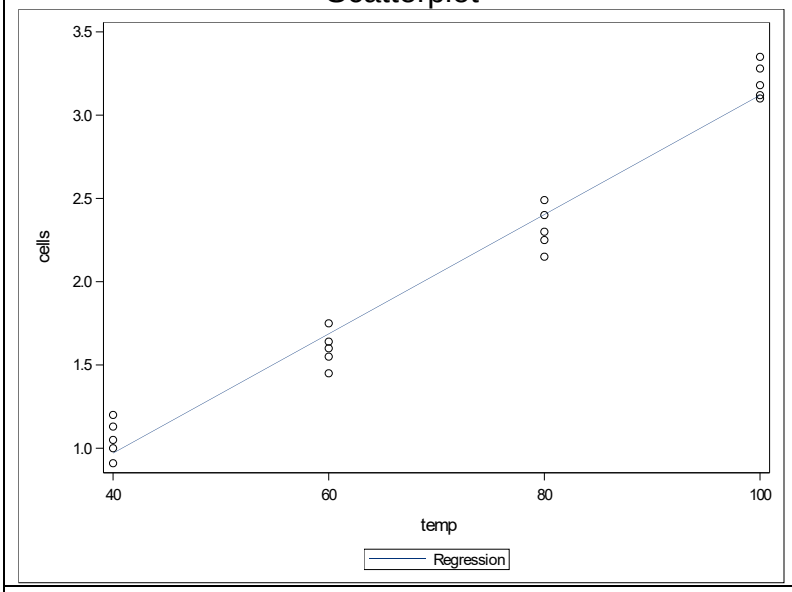
There is a significant relationship between temperature and the total number of cells ($p<0.0001$). On average, for every one degree increase in temperature, the number of cells increases by 35,820 (95% CI: 32,820 to 38,820 cells).

Recall, the scale is 10^6, so $0.03582 \times 10^6 = 0.03582 \times 1{,}000{,}000 = 35{,}820$

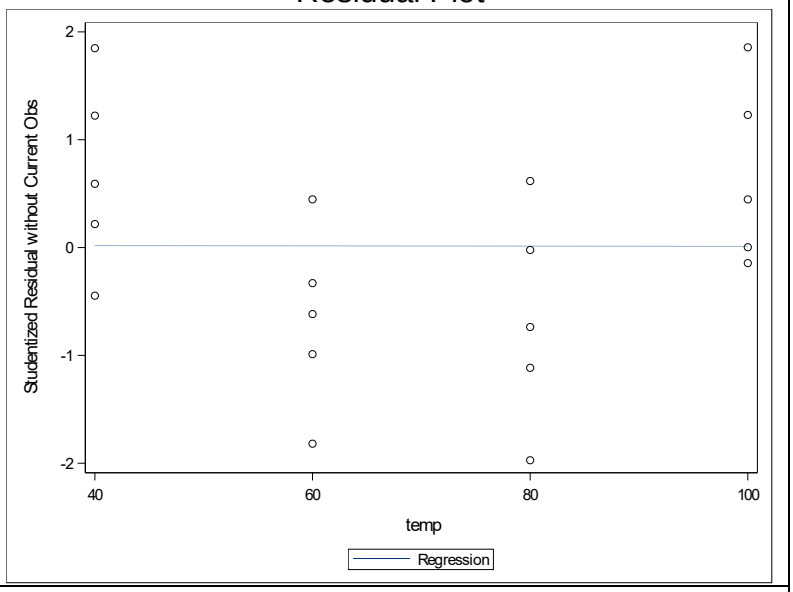
B) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals). Is there any evidence that any of the regression assumptions are violated? What possible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

```
/* 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;
```

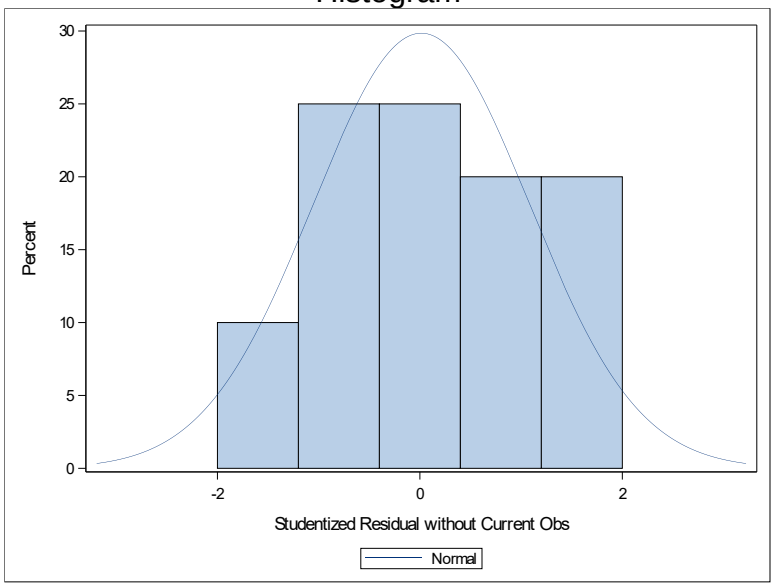
Scatterplot



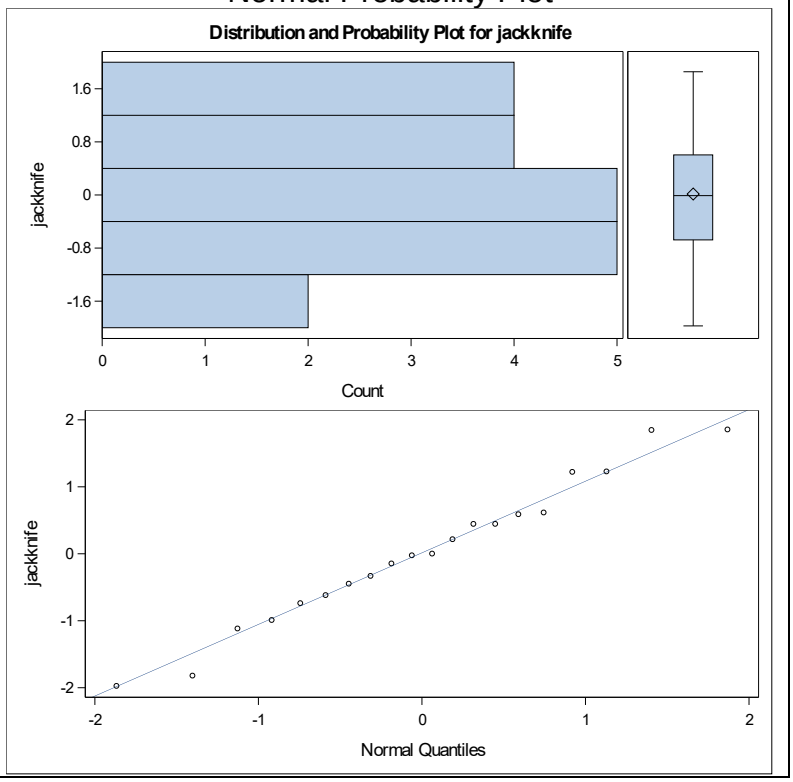
Residual Plot



Histogram



Normal Probability Plot



Given the four figures, it appears we may have a subtle departure from linearity. The scatterplot shows our straight-line regression going through all the points, but a slight curved relationship may exist. The residual plot does show some pattern and the normal probability plot has some deviation near the tails. We may wish to take a log transformation of our outcome to examine if this remedies these subtle departures from our assumptions. Alternatively, we may be able to include polynomial terms for temperature to see if this addressed the curved data.

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?

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

Test LOF_linear Results for Dependent Variable cells				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	0.07571	5.64	0.0140
Denominator	16	0.01342		

Our linear LOF test suggests that the linear model is NOT adequate (p=0.014 < 0.05), adding higher order polynomials may be beneficial.

Estimating Pure Error and LOF SS for our Straight-Line Model

The estimated sums of squares due to pure error comes from our LOF_linear table and the "Denominator" source:

$$SS(\text{pure error}) = DF_d \times MS_d = 16 \times 0.01342 = 0.21476$$

The estimated sums of squares due to lack of fit for our straight-line model comes from the "Numerator" source in our LOF_linear table:

$$SS(\text{lack of fit}) = DF_n \times MS_n = 2 \times 0.07571 = 0.15142$$

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.

```
/* 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;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	12.98210	6.49105	513.73	<.0001
Error	17	0.21480	0.01264		
Corrected Total	19	13.19690			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.49460	0.28856	1.71	0.1047
temp	1	0.00537	0.00887	0.61	0.5528
temp2	1	0.00021750	0.00006284	3.46	0.0030

Test LOF_quad Results for Dependent Variable cells *(code for test in 1c)*				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	0.00003600	0.00	0.9593
Denominator	16	0.01342		

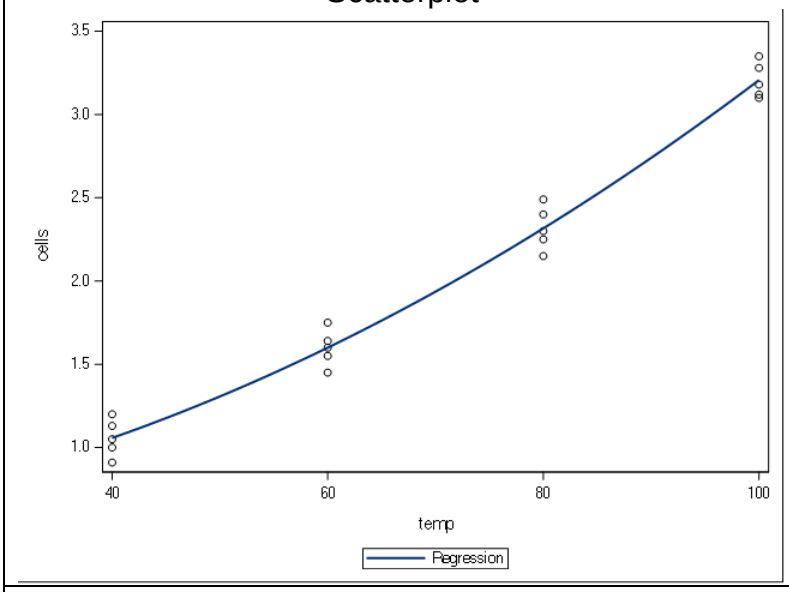
Adding temp^2 to the model is significant when adjusting for the inclusion of temp in the model (p=0.003), so it appears meaningful to include the second order polynomial in addition to temp.

Testing for a LOF in our quadratic model, p=0.9593, so we fail to reject the null hypothesis that there is a lack of fit. (In other words, this means that our second order polynomial regression model has adequate fit.)

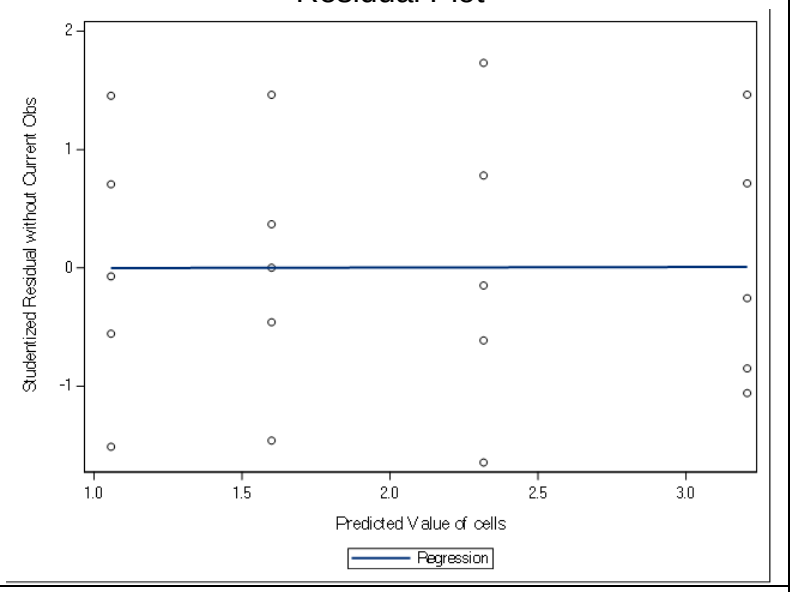
E) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals) for the quadratic model. Is there any evidence that any of the regression assumptions are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

```
/* 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;
```

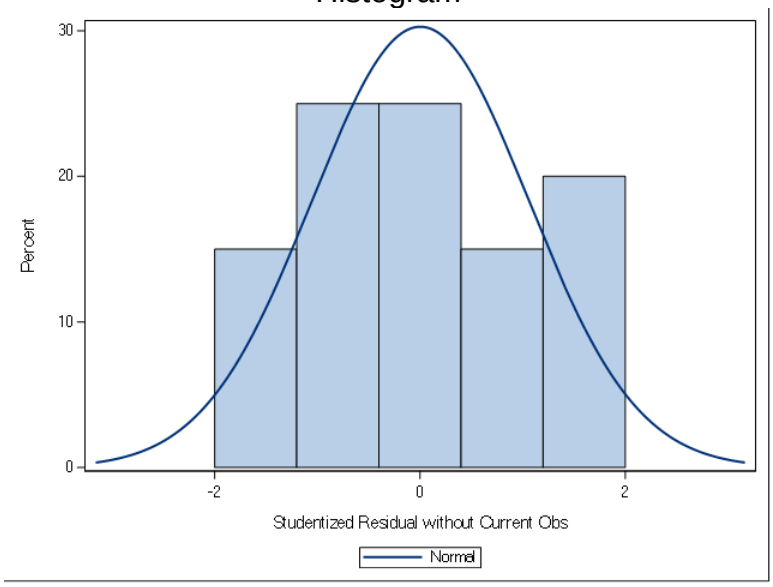
Scatterplot



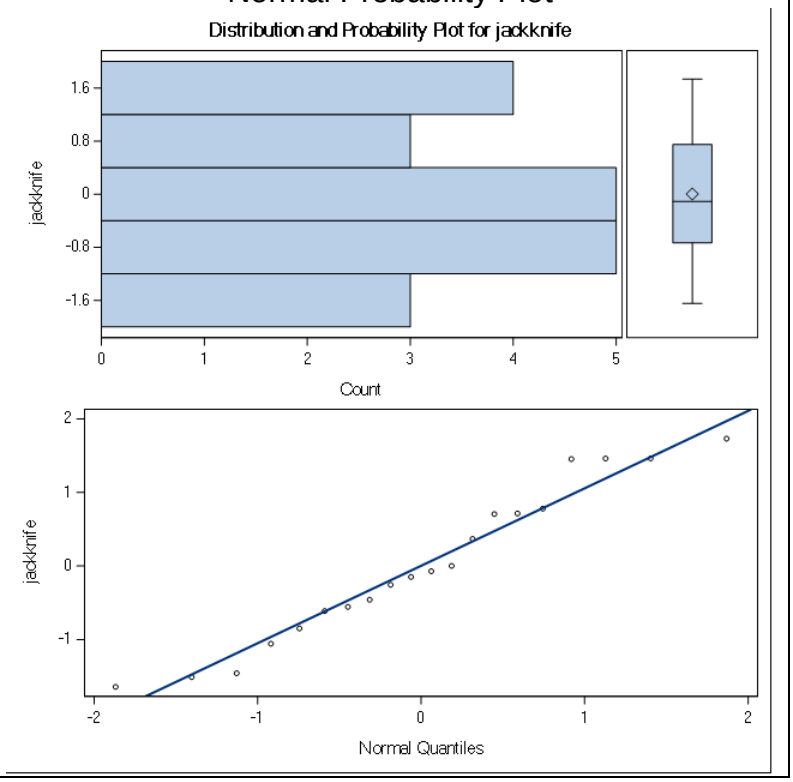
Residual Plot



Histogram



Normal Probability Plot



Adding temp^2 to our model appears to have corrected the potentially non-linear trend in the scatterplot and residual plot. However, relative to our previous model, it appears we may have some concerns about the normality assumption since the jackknife residuals appear to deviate more from the straight line.

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

The quadratic model appears to be the most appropriate. This is based on the significant LOF for the linear model (p=0.014 in part C), but an insignificant LOF for the quadratic mode (p=0.9593 in part D). Additionally, the quadratic term for power in the regression model is significant (p=0.003 from part D) and the diagnostic plots are generally improved with the potential exception of the normal probability plot.

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.

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

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	2.04500	0.02591	78.94	<.0001
lin	1	0.35820	0.01159	30.92	<.0001
quad	1	0.08700	0.02591	3.36	0.0040
cubic	1	-0.00060000	0.01159	-0.05	0.9593

The orthogonal polynomial model suggests that a cubic term does not contribute significantly to explaining the relationship with number of cells (p=0.9593), however the quadratic term is significant (p=0.004). Since the quadratic term is the highest order polynomial term that is significant, our decision on the model to use is the same as before.

Some differences from our LOF tests in parts C and D are that the orthogonal polynomial model is constructed so that each variable is independent (no collinearity), whereas our LOF tests previously used the simple polynomial variables that were likely correlated. Additionally, our LOF test used different degrees of freedom for the numerator and denominator of the reference F distribution, whereas our results for part G use DF=1 for the "numerator" part of the corresponding F-test (i.e., $t_k^2 = F_{1,k}$). In this example, our p-value from LOF_quad matches our p-value for the cubic variable term (and the F-value and t-value are extremely similar, but differ slightly due to rounding in the output).

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?

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

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-0.66817	0.05275	-12.67	<.0001	-0.77898	-0.55735
temp	1	0.01855	0.00071778	25.85	<.0001	0.01705	0.02006

The regression equation on the log scale is: $E[\log(cells)] = \beta_0 + \beta_1 \times temp$

Based on our output, the regression fit is: $\hat{Y} = -0.668 + 0.01855 \times age$, where $\hat{Y}$ is log(cells).

The intercept is the mean of the log(cells) when temperature is 0. If we exponentiate the intercept, it represents the geometric mean of total cell counts when temperature is 0: $\exp(-0.66817) = 0.513$.

The slope is the average increase in log(cells) for a one-unit increase in temperature. If we exponentiate the slope our interpretation changes to the percent increase (or multiplicative change).

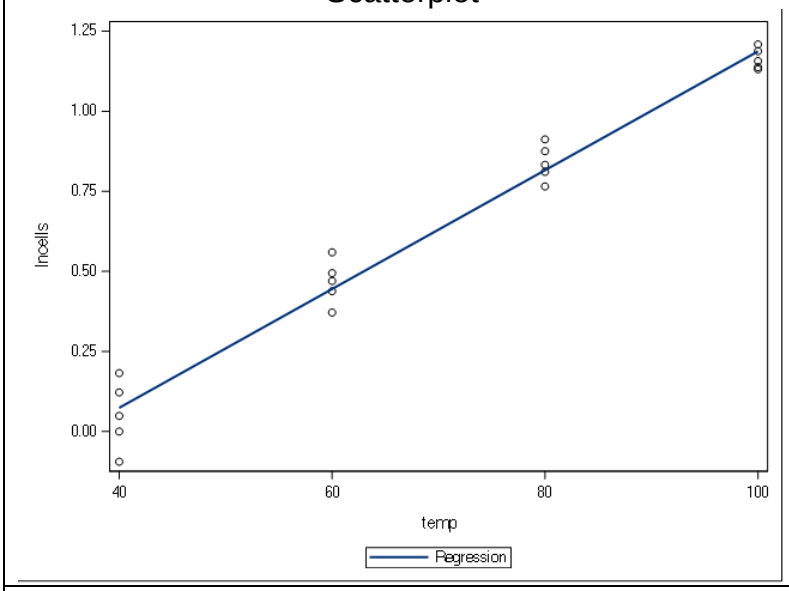
$\exp(0.01855) = 1.0187$ → One average, a one-unit increase in temperature results in a cell count that is 1.0187 times higher (1.87% higher).

For the 95% CI: $\exp\big((0.01705, 0.02006)\big) = (1.017, 1.020)$. We are 95% confident that the total cell count is between 1.7% and 2.0% higher for a one-unit increase in temperature.

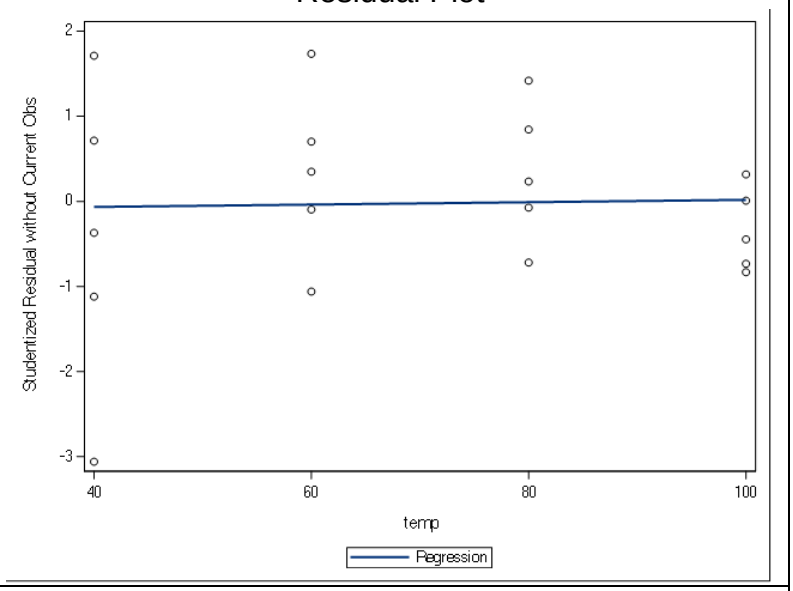
B) Produce the four diagnostic plots discussed in lecture (Y-X scatterplot; scatterplot of the Studentized deleted residuals; histogram of the Studentized deleted residuals; normal probability plot of residuals). Is there any evidence that any of the regression assumptions are violated?

```
/* 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;
```

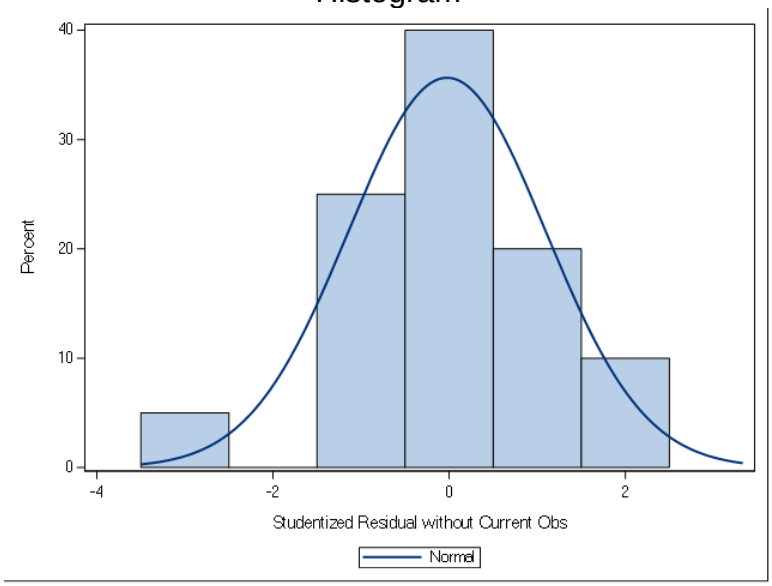
Scatterplot



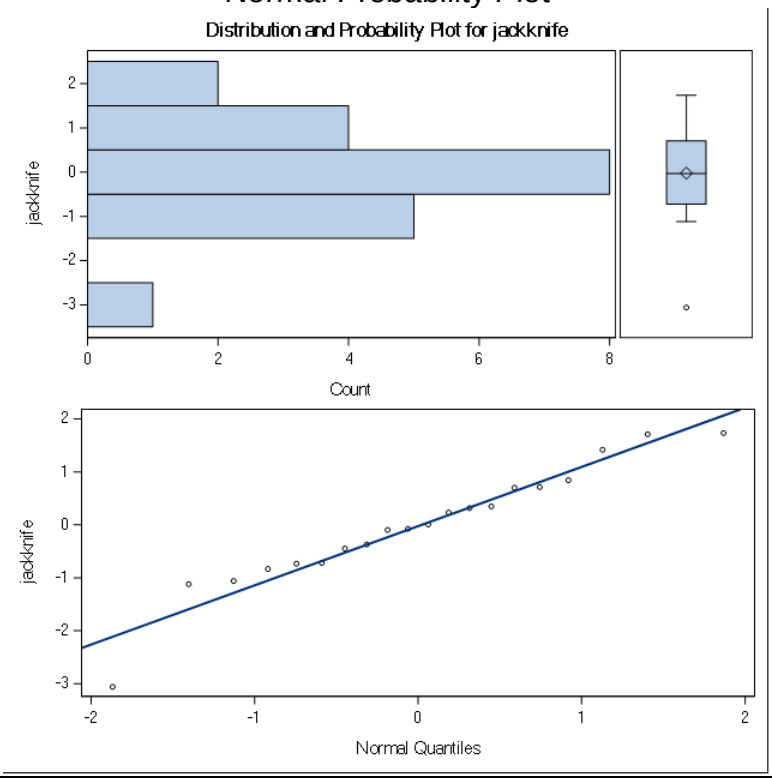
Residual Plot



Histogram



Normal Probability Plot



The scatter plot suggests that the assumption of linearity may be appropriate. However normality may be violated (based on the histogram left skew and the normal probability plot having some sharp deviations from the straight-line at the tails). Additionally, the residual plot suggests homoscedasticity is violated since the variability by temperature decreases as temperature increases.

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

There is a significant relationship between cell growth and temperature (p<0.001). On average, for every one-unit increase in temperature, total cell counts are 1.0187 times higher (1.87% higher) (95% CI: 1.7%, 2.0% higher).

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?

```
/* 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;
```

Test LOF_linear Results for Dependent Variable lncells				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	0.00505	0.98	0.3976
Denominator	16	0.00517		

Test LOR_quad Results for Dependent Variable lncells				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	0.00000619	0.00	0.9728
Denominator	16	0.00517		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.63066	0.01607	39.24	<.0001
lin	1	0.18555	0.00719	25.82	<.0001
quad	1	-0.02246	0.01607	-1.40	0.1813
cubic	1	-0.00024872	0.00719	-0.03	0.9728

Based on either set of output (the LOF_linear and LOF_quad tables, or the parameter estimates table for the orthogonal polynomials), the straight-line model is most appropriate.

If we use the LOF_linear table, we come to this conclusion because p=0.3976 > 0.05, so we fail to reject the null hypothesis that there is a lack of fit.

If we use the polynomial regression table, we can note that both quad and cubic have insignificant p-values, while lin is p<0.0001. We can also note that the p-value for cubic matches the p-value for LOF_quad.

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

There is some subjectivity to this answer.

If we are concerned about the violation of homoscedasticity and the challenge of interpreting the log(cell count) model, then we should go with the model from part 1F.

If we are concerned about the potential violation of normality from the polynomial regression model and prefer interpretations on a geometric mean-scale (or with percent change), then the model from 2D may be ideal.