

# ***Stat 431/511 Practice Final, Summer 2004***

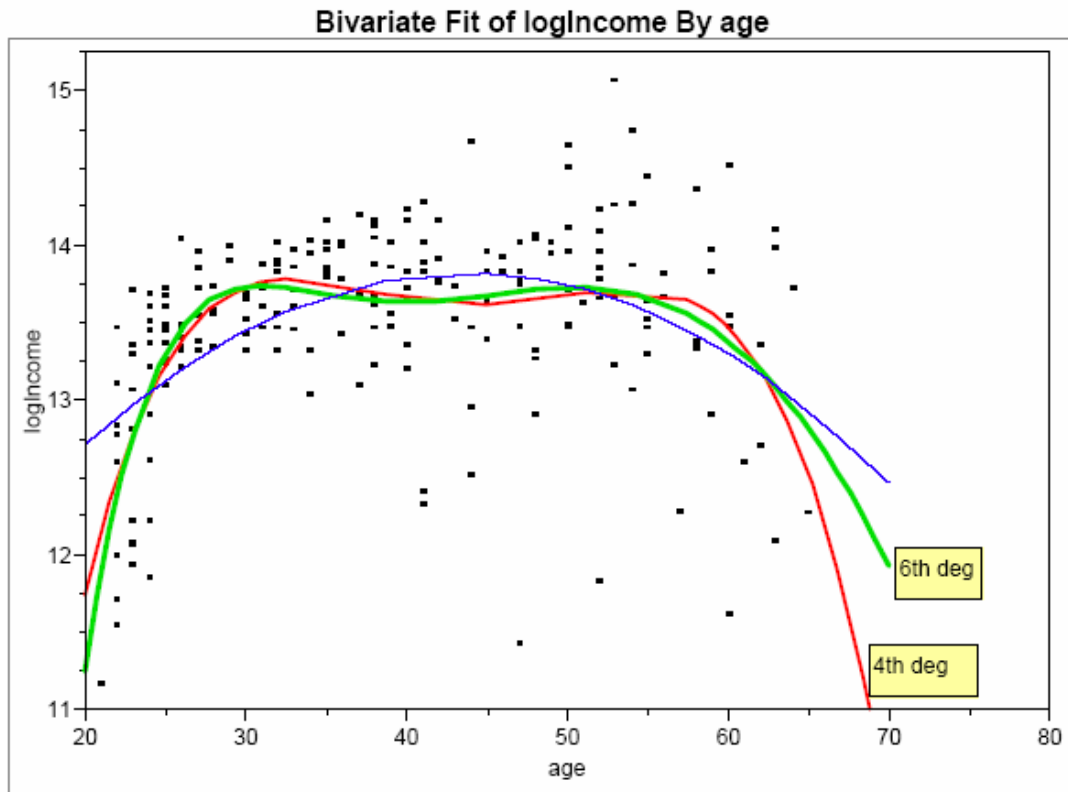
August 3<sup>th</sup>, 2004

**Time:** 90 minutes

## **Instructions:**

1. The final is closed book. Up to 3 pages double-sided A4 sized cheat sheets are allowed. Statistical tables are provided.
2. You may use a calculator.
3. Show all your work and partial credits will be given.
4. You must write the exam using pen (not pencil).
5. When performing hypothesis tests, clearly state the null and alternative hypothesis and show the critical value and/or P-value (from the table) when appropriate.

1. (8 pts) This question involves a data set that gives the earnings of a sample of Canadian men. The Y variable is the common log of their monthly earnings (in Canadian \$) and the X variable is their age. Below you will find partial output for a quadratic regression, a quartic (= 4<sup>th</sup> degree polynomial) regression, and a 6<sup>th</sup> degree polynomial regression. The plot shows the combined least squares regression plots for all three polynomial regressions.



### Polynomial Fit Degree=2

$$\text{logIncome} = 13.026 + 0.0195 \text{ age} - 0.00198 (\text{age} - 38.849)^2$$

#### Summary of Fit

RSquare	0.2308
Root Mean Square Error	0.5608
Mean of Response	13.490
Observations	205

### Polynomial Fit Degree=4

$$\text{logIncome} = 14.504 - 0.0210 \text{ age} + 0.000797 (\text{age}-38.849)^2 + 0.000201 (\text{age}-38.849)^3 - 0.0000101 (\text{age}-38.849)^4$$

#### Summary of Fit

RSquare	0.3150
Root Mean Square Error	0.5319
Mean of Response	13.490
Observations (or Sum Wgts)	205

Other tables omitted

### Polynomial Fit Degree=6

$$\text{logIncome} = 13.843 - 0.00512 \text{ age} + 0.00223 (\text{age}-38.849)^2 + 0.000013 (\text{age}-38.849)^3 - 0.0000142 (\text{age}-38.849)^4 + 0.0000005 (\text{age}-38.8488)^5 - 5.5367\text{e-}9 (\text{age}-38.849)^6$$

#### Summary of Fit

RSquare	0.3262
Root Mean Square Error	0.5302
Mean of Response	13.490
Observations	205

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	26.942	4.490	15.97
Error	198	55.659	0.281	Prob > F
C. Total	204	82.601		<.0001

#### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	13.843	0.6197	22.34	<.0001
age	-0.00512	0.0161	-0.32	0.7502
(age-38.8488)^2	0.00223	0.00161	1.39	0.1655
(age-38.8488)^3	0.000013	0.000175	0.07	0.9409
(age-38.8488)^4	-0.000014	0.000009	-1.58	0.1151
(age-38.8488)^5	0.0000005	4.505e-7	1.08	0.2809
(age-38.8488)^6	-5.537e-9	1.794e-8	-0.31	0.7580

- a) (2 pts) The entries for “**Model**” and “**F-Ratio**” have been deleted from the following table. Fill in the correct values and show *detailed* calculation procedures.

<b>Analysis of Variance</b>				
Source	DF	Sum of Squares	Mean Square	F Ratio
<b>Model</b>				
Error	202	63.539	0.315	Prob > F
C. Total	204	82.601		<.0001

<b>Parameter Estimates</b>				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	13.026	0.1308	99.60	<.0001
age	0.01951	0.00339	5.75	<.0001
(age-38.8488)^2	-0.00198	0.00029	-6.82	<.0001

- b) (4 pts) Use the information above to fill in the following entries in the **Summary of Fit** and **Analysis of Variance** tables for a **LINEAR REGRESSION** using this data. You do not need to fill in the entries marked “xxxxxx”. Show detailed calculation procedures (Assume the sample correlation coefficient is 0.232).

Linear Fit  
 $\text{logIncome} = 13.022 + 0.01205 \text{ age}$

**Summary of Fit**

RSquare	_____
Root Mean Square Error	_____
Observations	205

<b>Analysis of Variance</b>				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	_____	_____	xxxxxx	xxxxxx
Error	_____	_____	xxxxxx	Prob > F
C. Total	_____	_____		0.0008

- c) (2pts) Which of the 4 possibilities for a (polynomial) regression prediction equation is the best choice: a linear regression, a quadratic regression, a quartic regression, or a 6<sup>th</sup> degree regression? Explain your reasoning clearly, giving the important  $R^2$  values,  $P$ -values and/or  $F$ - or  $t$ -statistics that support your choice.  
[The “best” choice will have a good balance of prediction accuracy and simplicity. Thus, for example, you should not choose a quadratic regression if a linear regression provides nearly the same accuracy.]

2 (9 pts) In an experiment to study factors influencing wood specific gravity, a sample of 20 mature wood samples was obtained, and measurements were taken on number of fibers/ $mm^2$  in springwood ( $x_1$ ), number of fibers/ $mm^2$  in summerwood ( $x_2$ ), % springwood ( $x_3$ ), light absorption in springwood ( $x_4$ ), and light absorption in summerwood ( $x_5$ ).

a) (4 pts) Fitting the regression function  $\mu_{Y|x_1, x_2, x_3, x_4, x_5} = \beta_0 + \beta_1 x_1 + \dots + \beta_5 x_5$  resulted in  $R^2 = .769$ . Does the data indicate that there is a linear relationship between specific gravity and at least one of the predictors? Test using  $\alpha = .01$ .

b) (2 pts) When  $x_2$  is dropped from the model, the value of  $R^2$  remains at .769. Compute adjusted  $R^2$  for both the full model and the model with  $x_2$  deleted.

c) (3 pts) When  $x_1, x_2$  and  $x_4$  are all deleted, the resulting value of  $R^2$  is .654. The total sum of squares is  $SST = .0196610$ . Does the data suggest that all of  $x_1, x_2$  and  $x_4$  have zero coefficients in the true regression model? Test the relevant hypotheses at level .05.

3. (8 pts) The data analyzed below are from a study of the length of hospital stay following an operation for herniorrhaphy. After some preliminary analysis it was decided to use the *cube root of length of stay (called  $Length^{(1/3)}$ )* as the response variable, rather than the originally recorded variable - Length of Stay.

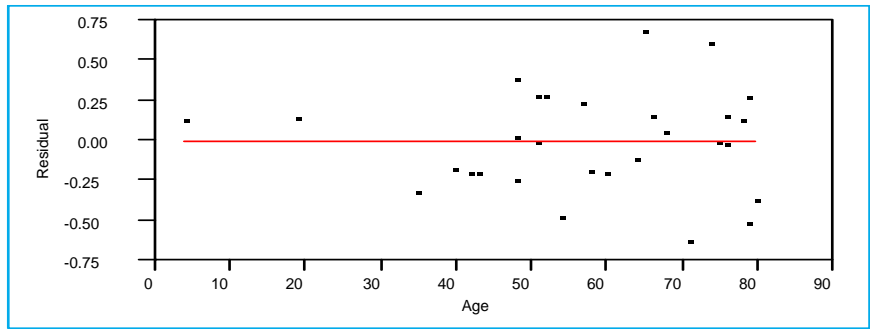
Following this are tables for a simple regression of  $Length^{(1/3)}$  on age.

Simple Regression:					
Linear Fit					
$Length^{(1/3)} = 0.937 + 0.0147 \text{ Age}$					
Summary of Fit					
RSquare					0.31
Root Mean Square Error					0.45
Mean of Response					1.76
Observations					32
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	1	2.752	2.752	13.50	
Error	30	6.113	0.204	<b>Prob&gt;F</b>	
C Total	31	8.865		0.0009	
Age: Moments					
Mean			56.03		
Std Dev			20.27		
N			32		

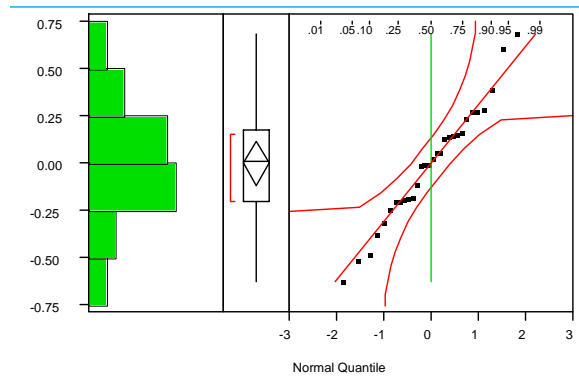
- a) (2 pts) Based on this analysis, what is the best prediction for the actual length of stay of a 60 year old?
- b) (2 pts) The hospital is trying to plan how long a stay to allow – on average – for 60-year old people after this operation. Give an appropriate 95% confidence interval.

c) (2 pts) Also give a 95% prediction interval for the length of stay for a particular 60-year old patient.

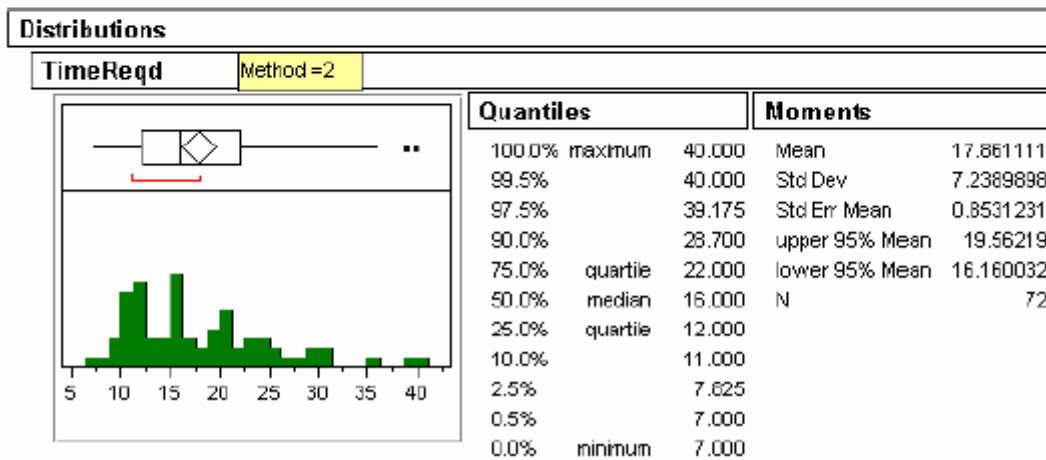
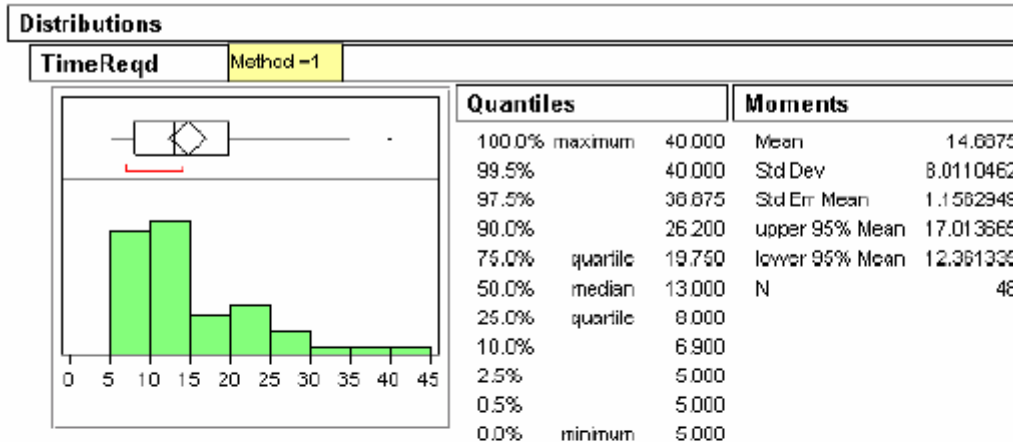
d) (2 pts) Below are a residual plot and a normal quantile plot for this simple linear regression. There is a noticeable deviation here from the desired situation in which the usual assumption of linearity and of normality and constant-variance of residuals. Which is this deviation from the assumption?



**Residuals Length<sup>(1/3)</sup>**



4. (4 pts) A large food processing center needs to be able to switch from one type of package to another quickly to react to changes in order patterns. Consultants recommended and helped implement an alternative method for changing the production line. Data was gathered for both methods. The time (in minutes) required to change the production on a food processing line was measured over several days using both methods. A summary of the data is shown below.



Is one of these changeover methods faster than the other? Conduct the appropriate hypothesis test (and check the necessary assumptions). Use  $\alpha = .05$ . Report the P-value of your test.

5. (6 pts) Prove that in linear regressions, the sum of residuals always equals 0 (Overall, there is no overestimating or underestimating).