UNIVERSITY OF TORONTO

Faculty of Arts and Science

JUNE EXAMINATION 2004

STA 302 H1F / STA 1001 H1F

Duration - 3 hours

Aids Allowed: Calculator

NAME: SOLUTIONS

STUDENT NUMBER:

- There are 17 pages including this page.
- The last page is a table of formulae that may be useful.
- Tables of the t distribution can be found on page 15 and tables of the F distribution can be found on page 16.
- Total marks: 75

1abcde	1fg	2ab	2cd	3	4

5a	5b	5cd	5ef	6ab	6cde

1. Olympic gold medal performances in track and field improve over time. A regression was run with dependent variable `longjump`, the winning distance in the long jump (in inches), and independent variable `year`, the year the Olympics was held after 1900 (counting 1900 as year 0). Data from the Olympics held from 1900 through 1984 were used (some Olympics were missed during the World Wars). Here are the data:

year	0	4	8	12	20	24	28	32	36	48
longjump	282.9	289.0	294.5	299.3	281.5	293.1	304.8	300.8	317.3	308.0

year	52	56	60	64	68	72	76	80	84
longjump	298.0	308.3	319.8	317.8	350.5	324.5	328.5	336.3	336.3

Some output from SAS is given below.

```
                              The REG Procedure
                            Descriptive Statistics

                                               Uncorrected                        Standard
Variable              Sum           Mean                SS          Variance      Deviation
Intercept        19.00000        1.00000          19.00000                 0              0
year            824.00000       43.36842             49184         747.13450       27.33376
longjump       5890.81250      310.04276           1833077         370.73440       19.25446

                              The REG Procedure
                               Model: MODEL1
                        Dependent Variable: longjump

                             Analysis of Variance
                                       Sum of           Mean
Source                   DF           Squares         Square    F Value    Pr > F
Model                     1        5054.88650     5054.88650      53.10    <.0001
Error                    17        1618.33267       95.19604
Corrected Total          18        6673.21916

             Root MSE              9.75685    R-Square     0.7575
             Dependent Mean      310.04276    Adj R-Sq     0.7432
             Coeff Var             3.14694

                             Parameter Estimates
                              Parameter        Standard
     Variable      DF          Estimate           Error     t Value    Pr > |t|
     Intercept      1         283.45427         4.28064       66.22      <.0001
     year           1           0.61308         0.08413        7.29      <.0001
```

Questions based on this output are on the next 2 pages.

Continued

(a) (2 marks) Estimate the mean change in the winning long jump distance from one Olympics to the next, assuming that the Olympics are held every 4 years.

$$4(0.61308) = 2.453$$

(b) (1 mark) Estimate the variance in winning long jump distance that is not explained by the year the Olympics were held.

$$95.196$$

(c) (1 mark) Estimate the percent of total variability in winning long jump distance that is explained by the year the Olympics were held.

$$75.75\%$$

(d) (1 mark) Estimate the correlation between winning long jump distance and the year the Olympics were held.

$$\sqrt{.7575} = .8703$$

(e) (2 marks) What is the observed value of the test statistic for the test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$? What is the p-value for this test?
Test statistic: 7.29
p-value: < 0.00005

Continued

(f) (3 marks) Construct simultaneous 99% confidence intervals for the slope and intercept of the regression line.

Using the Bonferroni method, $t_{17,0.005/2} = 3.222$

CI for slope: $.61308 \pm 3.222(.08413) = (.342, .884)$

CI for intercept: $283.454 \pm 3.222(4.2806) = (269.66, 297.246)$

(g) (5 marks) It has been suggested that the Mexico City Olympics in 1968 saw unusually good track and field performances, possibly because of the high altitude. Construct an appropriate 95% interval for the predicted winning long jump distance in 1968. Do the data support these suggestions? Explain.

$\hat{Y}_{68} = 283.454 + .613(68) = 325.14$

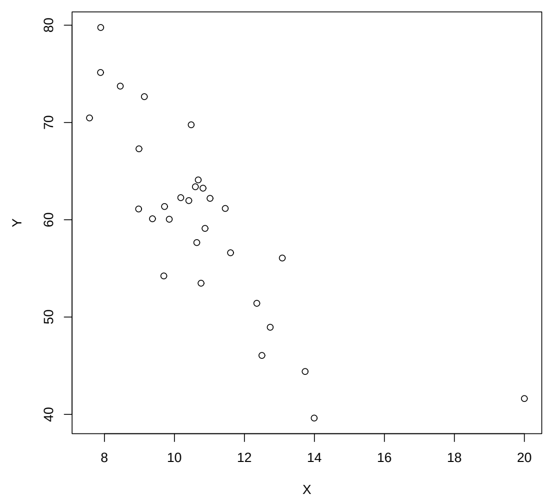
$t_{17,.025} = 2.11$

Prediction interval: $325.14 \pm 2.11(9.757)\sqrt{1 + \frac{1}{19} + \frac{(68-43.368)^2}{49184-19(43.368)^2}} = (303.57, 346.71)$

The value in 1968 was 350.5 which is not in the prediction interval. So Mexico City was unusual.

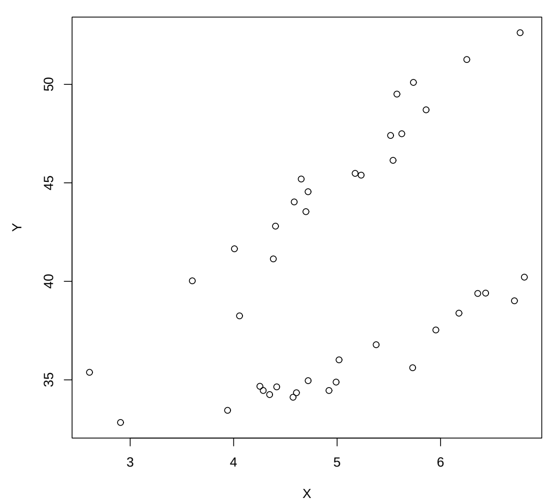
2. (8 marks (2 marks each)) Each of the following plots is a scatterplot of a dependent variable versus an independent variable. We wish to study further the relationship between the two variables. Indicate an appropriate linear regression model based on the plot.

(a)



Point in bottom right corner is influential. Fit the model $Y_i = \beta_0 + \beta_i X_i + \epsilon_i$ on X in the range from 7.5 to 14.

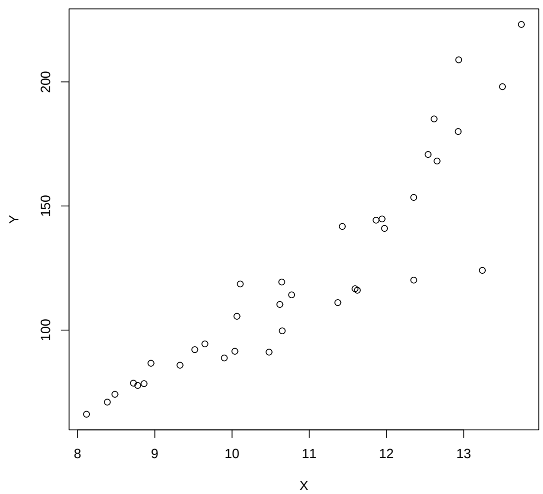
(b)



Fit model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 d_i + \beta_3 X_i * d_i + \epsilon_i$ where d_i is 1 if the ith point is in one of the groups of data and 0 otherwise.

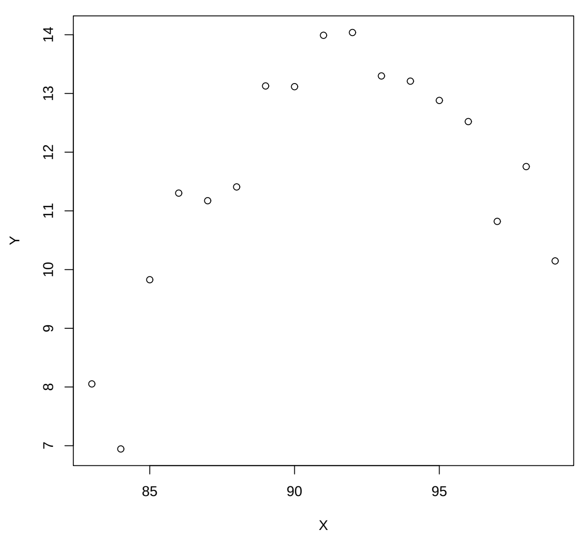
(This question continues on the next page.)

(c)



This shows curvature plus increasing variance so use $Y' = \log(Y)$ *or* $\sqrt{Y}$ *and fit* $Y_i' = \beta_0 + \beta_1 X_i + \epsilon_i$

(d)



Fit model $Y_i = \beta_0 + \beta_1 X_i' + \beta_2 (X_i')^2 + \epsilon_i$ *where* X_i' *is centred* X_i

3. (3 marks) Show, in the case of simple linear regression, that the fitted line passes through the point $(\overline{X}, \overline{Y})$.
Fitted line: $\hat{Y} = b_0 + b_1 X$ *where* $b_0 = \overline{Y} - b_1 \overline{X}$
So at $\overline{X}$, $\hat{Y} = \overline{Y} - b_1\overline{X} + b_1\overline{X} = \overline{Y}$

4. (a) (5 marks) A multiple linear regression model is be constructed to examine the relationship between a response variable Y and 3 predictor variables X_1, X_2, and X_3. Suppose measurements of these 4 variables have been taken on n items. State the multiple linear regression model in matrix terms, defining all of your matrices, including the standard assumptions.
Model: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ *where*

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \vdots \\ \boldsymbol{Y_n} \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{X_{11}} & \boldsymbol{X_{12}} & \boldsymbol{X_{13}} \\ \boldsymbol{1} & \boldsymbol{X_{21}} & \boldsymbol{X_{22}} & \boldsymbol{X_{23}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{X_{n1}} & \boldsymbol{X_{n2}} & \boldsymbol{X_{n3}} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \boldsymbol{\beta_3} \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_n} \end{pmatrix}$$

Assumptions: $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

(b) (3 marks) Derive the expression for the covariance matrix of the least squares estimators of the model coefficients of your model in part (a). (I.e., Show $\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.)

$$\begin{aligned} \text{Cov}(\mathbf{b}) &= \text{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Cov}(\boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Continued

5. The data considered in the analysis below are observations on the acceleration (`accel`) of different vehicles along with their weight-to-horsepower ratio (`whp`), the speed at which they were travelling (`speed`), and the grade of the road (`grade`) which takes values 0, 2, and 6 (a value of 0 indicates the road was horizontal). There are 50 observations in the data set.

(a) (5 marks) The first model tried for these data was

$$\texttt{accel} = \beta_0 + \beta_1\texttt{whp} + \beta_2\texttt{speed} + \beta_3\texttt{grade} + \epsilon$$

Some output from SAS is given below.

```
                         The REG Procedure
                          Model: MODEL1
                    Dependent Variable: accel

                       Analysis of Variance
                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F
Model                    3     164.99430       54.99810      25.45    <.0001
Error                   46      99.41390        2.16117
Corrected Total         49     264.40820

           Root MSE              1.47009    R-Square     0.6240
           Dependent Mean        2.60600    Adj R-Sq     0.5995
           Coeff Var            56.41183

                       Parameter Estimates
                        Parameter       Standard
     Variable     DF     Estimate          Error    t Value    Pr > |t|
     Intercept     1      7.19950        0.60087      11.98      <.0001
     whp           1     -0.01838        0.00269      -6.83      <.0001
     speed         1     -0.09347        0.01307      -7.15      <.0001
     grade         1     -0.15548        0.09040      -1.72      0.0922
```

What does this output tell you about the ability of the vehicles to accelerate under various conditions? Your answer should explain the affects of each of `whp`, `speed`, and `grade` on acceleration.

For vehicles going the same speed on a road with the same grade, as whp goes up one unit, acceleration goes down 0.0184 units on average. There is strong evidence that the effect on acceleration over and above the other variables is non-zero.

For vehicles with the same whp on roads with the same grade, as speed goes up one unit, acceleration goes down 0.0935 units on average. There is strong evidence that the effect on acceleration over and above the other variables is non-zero.

For vehicles going the same speed on a road with the same whp, as the grade of the road goes up one unit, acceleration goes down 0.155 units on average. There is only weak evidence that this effect is different from zero.

(b) (3 marks) The next model fit was

$$\texttt{accel} = \beta_0 + \beta_1\texttt{whp} + \beta_2\texttt{speed} + \beta_3\texttt{grade0} + \beta_4\texttt{grade2} + \epsilon$$

where `grade0` = 1 if `grade` is 0 and is 0 otherwise, and `grade2` = 1 if `grade` is 2 and is 0 otherwise. Some output from SAS for this model follows.

The REG Procedure
Dependent Variable: accel

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	165.00484	41.25121	18.67	<.0001
Error	45	99.40336	2.20896		
Corrected Total	49	264.40820			

Root MSE	1.48626	R-Square	0.6241
Dependent Mean	2.60600	Adj R-Sq	0.5906
Coeff Var	57.03217		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	6.25693	0.61730	10.14	<.0001	339.56180
whp	1	-0.01839	0.00272	-6.76	<.0001	52.93850
speed	1	-0.09345	0.01322	-7.07	<.0001	105.66329
grade0	1	0.93165	0.54864	1.70	0.0964	3.48062
grade2	1	0.65181	0.56668	1.15	0.2561	2.92243

What is the purpose of using the two variables `grade0` and `grade2`?
Can then model the effect of the three grades without making any assumptions about the functional form of the relationship of grade with acceleration.
Need 2 indicator variables since there are 3 levels of grade (so a grade of 6 is indicated when both are 0).

(c) (6 marks) Is the evidence for a grade effect stronger or weaker from the model in part (b) than the model in part (a)? Support your answer with appropriate statistical tests.

First model: p-value for test of no grade effect is 0.0922.

Second model:

Test $H_0 : \beta_3 = \beta_4 = 0$ *versus* H_a*: at least one of* β_3, β_4 *is not zero.*

Test statistic: $F_{obs} = \frac{(3.48+2.92)/2}{2.21} = 1.45$

Under H_0 *this is an observation from an* $F_{2,45}$ *distribution. Approximating with* $F_{2,30}$ *gives* $0.1 < p < 0.5$.

So the evidence of a grade effect is stronger for the first model.

(d) (2 marks) Note that R^2 is almost identical for the models fit in parts (a) and (b). Choose another statistic that is useful for choosing between the two models and indicate which model is preferred.

Adjusted R^2*: First model is .5995, second model is .5906 so first model is preferred.*

OR

MSE: First model is 2.16117, second model is 2.20896 so first model is preferred.

(e) (3 marks) Another model that could be fit is

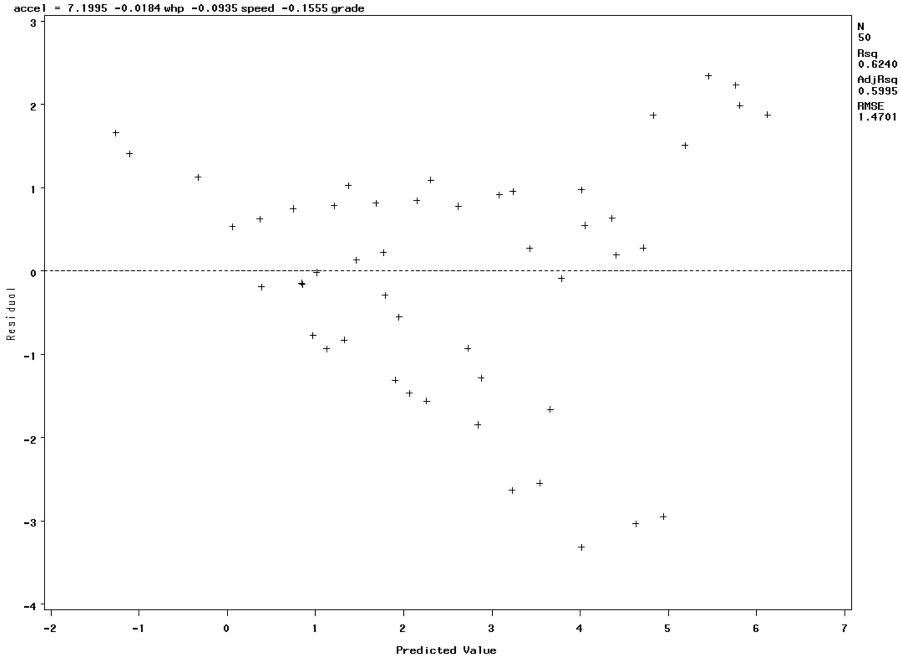
$$\texttt{accel} = \beta_0 + \beta_1\texttt{whp} + \beta_2\texttt{speed} + \beta_3\texttt{grade} + \beta_4\texttt{whp\_speed} + \epsilon$$

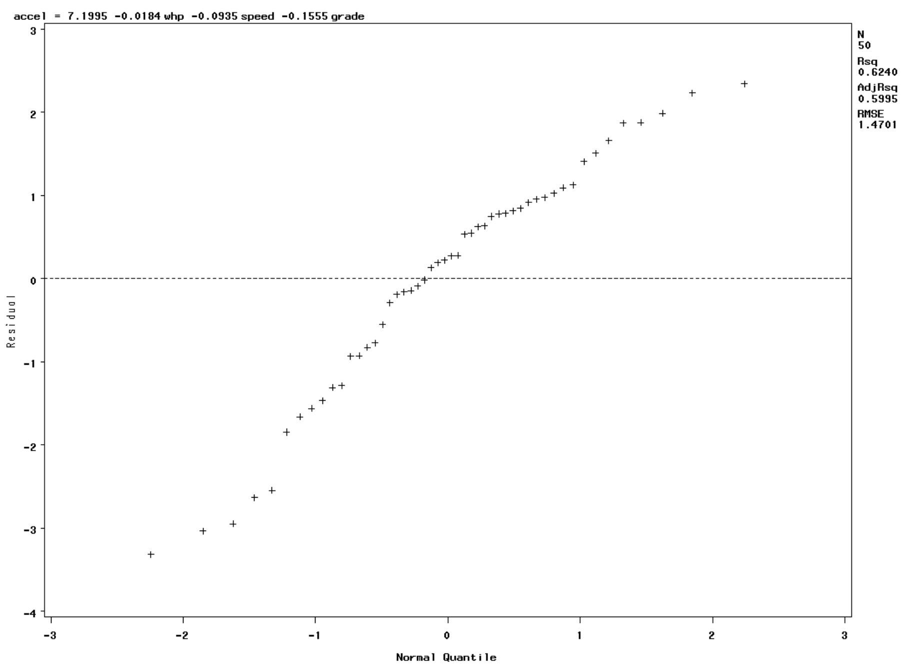
where $\texttt{whp\_speed} = \texttt{whp} * \texttt{speed}$. What additional information could be obtained from this model and how would you assess whether or not it is statistically important?

Can see whether the way whp affects acceleration varies with speed.
Test $H_0 : \beta_4 = 0$ *versus* $H_a : \beta_4 \neq 0$ *given other variables in model using usual t-test.*

(f) (5 marks) A plot of the residuals versus the predicted values and a normal quantile plot of the residuals for the regression in part (a) are on the next page. What additional information do these plots give?

First plot shows increasing variance so need transformation of Y *(or weighted least squares).*
Normal quantile plot shows tails that are a little lighter than normal so tests and confidence intervals based on normal theory are only approximately correct.


accel = 7.1995 -0.0184 whp -0.0935 speed -0.1555 grade
N
50
Rsq
0.6240
AdjRsq
0.5995
RMSE
1.4701
Residual
Predicted Value


accel = 7.1995 -0.0184 whp -0.0935 speed -0.1555 grade
N
50
Rsq
0.6240
AdjRsq
0.5995
RMSE
1.4701
Residual
Normal Quantile

Continued

6. (a) (6 marks) A simple linear regression is carried out and one point is determined to be an outlier but not an influential point. It is removed from the data and the regression line is fit to the remaining data. Which of the following quantities will differ by a substantial amount between the two regressions? If there is a substantial difference, indicate in which regression it is larger. Indicate if you don't have enough information to say.

   i. the slope
      won't differ by a substantial amount since not influential

   ii. the mean square error
      will decrease

   iii. R^2
      will increase

   (b) (3 marks) Explain the difference between an outlier and an influential point.
   An outlier has a large (in absolute value) residual.
   The point is far from the fitted line. An influential point is a point such that if it is removed and the line re-fit, the line will be different by a substantial amount. A point can be one or both of these.

(c) (2 marks) If you could choose the values of X at which to collect data before performing a simple linear regression analysis, would you prefer that $\sum_{i=1}^{n}(X_i - \overline{X})^2$ be large or small? Explain.
Large because gives smaller estimates of the variance of the regression parameter estimates.

(d) (2 marks) An investigator wishes to use multiple regression to predict a variable, Y, from two other variables, X_1 and X_2. She is also interested in the quantity that is the sum of X_1 and X_2 and includes a third predictor variable in her model, $X_3 = X_1 + X_2$. What problems might she encounter?
$\mathbf{X}'\mathbf{X}$ *will be singular so can't calculate regression coefficients*

(e) (4 marks) In a study of infant mortality, a regression model was constructed using birth weight (which is a good indicator of the baby's likelihood of survival) as a dependent variable and several independent variables, including the age of the mother, whether the mother smoked or took drugs during pregnancy, the amount of medical attention she had, her income, etc. The R^2 was 11%, but the t-test provided by SAS for the coefficient of each predictor variable had p-value less than 0.01. An obstetrician has asked you to explain the significance of the study as it relates to her practice. What would you say to her?
Since each variable has p-value < 0.01*, there is strong evidence that each helps explain birth weight over and above the other variables.*
For 2 mothers with everything else in the model the same, the coefficient of a variable gives an indication how, on average, the birth weight will change when that variable changes (if positive, good, if negative, bad).
But, since R^2 *is low, there is still a lot of variability in the data that is not explained by the model and the birth weight for any given patient may not be well predicted.*
Because this is not an experiment we can not say that any of these variables cause a change in birth weight, only that there is an indication of a relationship.