STA 302 / 1001 H - Summer 2004
Test 1 – June 2, 2004

LAST NAME:____________________ FIRST NAME:____________________

STUDENT NUMBER:____________________________________

ENROLLED IN: (circle one) STA 302 STA 1001

INSTRUCTIONS:
- Time: 60 minutes
- Aids allowed: calculator.
- A table of values from the t distribution is on the last page (page 7).
- Total points: 40

Some formulae:

$$b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \qquad b_0 = \overline{Y} - b_1\overline{X}$$

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum(X_i - \overline{X})^2} \qquad \mathrm{Var}(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right)$$

$$\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2\overline{X}}{\sum(X_i - \overline{X})^2} \qquad \mathrm{SSTO} = \sum(Y_i - \overline{Y})^2$$

$$\mathrm{SSE} = \sum(Y_i - \hat{Y}_i)^2 \qquad \mathrm{SSR} = b_1^2\sum(X_i - \overline{X})^2 = \sum(\hat{Y}_i - \overline{Y})^2$$

$$\sigma^2\{\hat{Y}_h\} = \mathrm{Var}(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right) \qquad \sigma^2\{\text{pred}\} = \mathrm{Var}(Y_h - \hat{Y}_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right)$$

Working-Hotelling coefficient: $W = \sqrt{2F_{2,n-2;1-\alpha}}$

1	2	3 abcd	3 efg

1. (a) (2 points) Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where the ϵ_i's are independent and identically distributed with the $N(0, \sigma^2)$ distribution. Assume the X_i's are fixed. What is the distribution of Y_i when X_i is 10?

$$Y_i \sim N(\beta_0 + 10\beta_1, \sigma^2)$$

(b) (3 points) The least squares estimate of the Y intercept for the model in (a) is b_0 as given on the first page. Show that b_0 is an unbiased estimate of the intercept in the model. You may take as known any results that were proved in lecture.

$$\begin{aligned} E(b_0) &= E(\overline{Y} - b_1\overline{X}) \\ &= \beta_0 + \beta_1\overline{X} - E(b_1)\overline{X} \\ &= \beta_0 \quad \text{since } E(b_1) = \beta_1 \end{aligned}$$

(c) (4 points) Show that the sum of the residuals is 0 for the least squares fit for the model above. What assumptions about the model did you use for this calculation?

$$\begin{aligned} \sum e_i &= \sum (Y_i - b_0 - b_1 X_i) \\ &= \sum (Y_i - \overline{Y} + b_1\overline{X} - b_1 X_i) \\ &= n\overline{Y} - n\overline{Y} + nb_1\overline{X} - b_1 n\overline{X} \\ &= 0 \end{aligned}$$

Don't need any statistical assumptions – just the form of the model.

2. (6 points (2 each)) For each of the following statements, say whether it is true or false. Give a brief justification of your answer.

   (a) A value of R^2 close to 1 indicates that the linear regression model is a good fit to the data.

   False. Points could be closely scattered around in a line in a curvilinear fashion, or could be caused by one overly influential point.

   (b) The estimate of the error variance, s^2, is a random variable.

   True. s^2 is a function of the Y's which are random variables.

   (c) $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 0$

   False. This is what is minimized in least squares.

3. Two species of predatory birds, collard flycatchers and tits, compete for nest holes during breeding season. Frequently, dead flycatchers are found in nest boxes occupied by tits. A field study examined whether the risk of mortality to flycatchers is related to the degree of competition between the two bird species for next sites. At each of 14 locations, the following data were collected: the number of flycatchers killed (the response variable labelled `fc_killed`) and the nest box occupancy measured as a percentage (the predictor variable labelled `tit_occ`).

The data and some SAS output are given below. Some numbers from the SAS output have been purposely deleted.

Location	1	2	3	4	5	6	7	8	9	10	11	12	13	14
fc_killed	0	0	0	0	0	1	1	1	1	2	2	3	4	5
tit_occ	24	33	34	43	50	35	35	38	40	31	43	55	57	64

```
                             The REG Procedure
                          Descriptive Statistics
                                      Uncorrected                    Standard
Variable            Sum         Mean           SS       Variance    Deviation
Intercept      14.00000      1.00000     14.00000              0            0
tit_occ       582.00000     41.57143        25844      126.87912     11.26406
fc_killed      20.00000      1.42857     62.00000        2.57143      1.60357

                             The REG Procedure
                              Model: MODEL1
                       Dependent Variable: fc_killed

                           Analysis of Variance
                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F
Model                  (A)      19.11669       19.11669
Error                  (B)      14.31188            (C)
Corrected Total         13      33.43857

             Root MSE              1.09209    R-Square     (D)
             Dependent Mean        1.42857    Adj R-Sq     0.5362
             Coeff Var            76.44618

                           Parameter Estimates
                        Parameter       Standard
     Variable      DF    Estimate          Error    t Value    Pr > |t|
     Intercept      1    -3.04686        1.15533      -2.64      0.0217
     tit_occ        1     0.10766        0.02689        (E)      0.0018
```

Questions pertaining to this output are on the next two pages.

(a) (5 points) What are the values of the 5 missing numbers (A through E) in the SAS output?

$A = 1$
$B = 12$
$C = 14.31188/12 = (1.09209)^2 = 1.1927$
$D = 19.11669/33.42857 = 0.5719$
$E = .10766/.02689 = 4.0037$

(b) (5 points) For the analysis of variance F test, state the null and alternative hypotheses, the value of the test statistic, the distribution of the test statistic under the null hypothesis, the p-value as accurately as possible, and an appropriate conclusion.

$H_0: \ \beta_1 = 0$ *versus* $H_a: \ \beta_1 \neq 0$
Test statistic: $F_{obs} = E^2 = 19.11669/C = 16.028$
Under H_0, $F_{obs} \sim F_{1,12}$
p-value = .0018
Strong evidence that the slope is not zero.

(c) (2 points) Estimate the mean change in the number of flycatchers killed when the nest box tit occupancy increases by 10%.

$0.10766(10) = 1.0766$

(d) (3 points) Give a 95% confidence interval for the slope of the line.

$t_{12,.025} = 2.179$
$.10766 \pm 2.179(.02689) = (.0491, .1662)$

(e) (5 points) Suppose an additional location was later found to have a nest box tit occupancy of 30%. Give a 90% prediction interval for this new value.

$\hat{Y} = -3.04686 + 0.10766(30) = .1829$
$t_{12,.05} = 1.782$
$\sum(X_i - \overline{X})^2 = 25844 - 14(41.57143)^2 = 13(126.87912) = 1649.4$
Prediction Interval: $.1829 \pm 1.782(1.09209)\sqrt{1 + \frac{1}{14} + \frac{(30-41.57143)^2}{1649.4}} = (-1.906, 2.272)$

(f) (3 marks) Would a 90% confidence interval for the mean number of flycatchers killed when the tit occupancy is 30% be wider or narrower than your interval in part (e). Explain why the width of the intervals differ. An answer that only points out the differences in the formulae will receive no marks.

Narrower. The s.e. of the estimate of $E(Y)$ only takes into account the variance in the estimation of the line while the s.e. of the estimate of the value of Y has this source of variance plus the model error variance (how points are scattered about the line).

(g) (2 marks) Basing your answer only on the information you have from the data and SAS output that was given, do you have any concerns about the validity of the prediction interval you found in the part (e)? Explain.

Yes. Looking at the values of the response variable, they clearly aren't samples from a normal distribution.