Material from STA302 in 1990. I have cut out a lot of blank space to save paper.

STA 302 Quiz 2

Name ______________________
Student No. ________________

1. For the MINITAB output below, n=10, H_0 is $\beta_1 \geq 0$ and H_a is $\beta_1 < 0$.

    a) What is the critical value of the test statistic at $\alpha = .01$?

    b) What is the value of the test statistic t*?

    c) Do you reject H_0? Yes or no.

    d) With this null and alternative hypothesis, what can you conclude about the presence of a linear relationship between X and Y?

```
The regression equation is
C3 = 6.66 + 2.97 C1

Predictor       Coef       Stdev    t-ratio        p
Constant      6.6567      0.6445      10.33    0.000
C1            2.9740      0.1039      28.63    0.000

s = 0.9434      R-sq = 99.0%     R-sq(adj) = 98.9%
```

2) For the MINITAB output above,

    a) Construct a 99% confidence interval for β_1.

    b) Are your results consistent with what you obtained in question 1? Explain.

Name ______________________________
Student Number ____________________

Midterm Exam
STA 302f 1990
Erindale College
Aids Allowed: Calculator, Text book

1. Let $Y_i = \beta_0 + X_i + \epsilon_i$, $i=1, \ldots, n$ where the X_i are known constants, β_0 is an unknown constant, and the ϵ_i are independent random variables with expected value zero and common variance σ^2.

a) (5 points) What is $E(Y_i)$?

b) (20 points) Find the least-squares estimate of β_0; that is, find the value of β_0 such that $Q(\beta_0)=\sum_{i=1}^{n} (Y_i-E(Y_i))^2$ is minimized. Show that it is a minimum. **Confine your answer to the space below.**

For problems 2 through 18, write T for true or F for false. These problems are worth 3 points each.

2. _______ The least-squares line is chosen so as to minimize the sum of squared vertical distances of the points (on a scatterplot) from the line.

3. _______ If your data do not include any observations with $X_i \leq 0$, it can be very misleading to interpret b_0.

4. _______ For the least-squares method to be valid, the error terms ϵ_i must be distributed normally with mean zero and common variance σ^2.

5. _______ $\frac{SSE}{SSTO}$ represents the proportion of variation in the dependent variable that is explained by the independent variable.

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

7. ______ For the simple linear regression model (2.1) on p. 31, $E(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$.

8. ______ Model (3.1) on p. 62 implies that if there is any relationship at all between X and Y, it must be linear.

9. ______ The confidence intervals and significance tests of chapter 3 are usually valid when n is large, even when the assumption of normality for the error terms is violated.

10. ______ We reject H_0 at significance level α if and only if $p > \alpha$.

11. ______ The confidence intervals and significance tests of chapter 3 are still valid when the X_i are random variables, provided that they are independent of ϵ_i, their distribution does not involve β_0, β_1 or σ^2, and all probability statements are viewed as being conditional on the particular observed values of the X_i.

12. ______ $E(Y_i) = b_0 + b_1 X_i$, where b_0 and b_1 are the least-squares estimates of β_0 and β_1 respectively.

13. ______ For simple linear regression, there is both a t-test and an F-test for H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.

14. ______ For simple linear regression, if $b_1 = 0$, then $b_0 = \overline{Y}$.

15. ______ Consider H_0: $\beta_1 \geq 0$ versus H_a: $\beta_1 < 0$. Even if H_0 is true, there could still be a negative correlation (in the population) between X and Y.

16. ______ Suppose you timed 1000 university students in the 100 meter dash, then waited a week and timed them again. Further, suppose that the mean and standard deviation of their times did not change from the first test to the second. Still, you would expect the very fastest students to run somewhat faster the second time.

17. ______ The model (3.1) on p. 64 implies that the parameter σ^2 is normally distributed.

18. ______ We reject the null hypothesis that β_1 equals a particular value (using a 2-tailed test) at significance level α if and only if the $(1-\alpha)100\%$ confidence interval for β_1 contains that particular value.

19. (10 points) For the ordinary simple linear regression model (2.1) on p. 31, show that $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$. This is a short proof.

Confine your answer to the space below.

Problems 20 through 24 refer to the following Minitab output. These questions are worth 2 points each. **Write your answers on the lines.**

```
Worksheet retrieved from file: prob2.26
MTB > regress c2 1 c1

The regression equation is
C2 = 183 + 0.262 C1

Predictor       Coef       Stdev    t-ratio        p
Constant      182.97       12.72      14.38    0.000
C1            0.2616      0.1783       1.47    0.164

s = 10.29       R-sq = 13.3%     R-sq(adj) = 7.1%

Analysis of Variance

SOURCE       DF          SS          MS        F        p
Regression    1       228.0       228.0     2.15    0.164
Error        14      1483.0       105.9
Total        15      1711.0
```

20. Is there a statistically significant relationship between X and Y at α=.20, two tailed? __________

21. What is the estimated standard deviation of b_0? _________

22. Can you reject H_0: $\beta_0=0$ at $\alpha=.01$? _________

23. What is $\sum_{i=1}^{n} (Y_i-\hat{Y}_i)^2$? _________

24. What proportion of the variation in the dependent variable is explained by the independent variable? _________

Questions 25 through 30 refer to question 2.26 (Robbery rate) on p. 58-59. Assume model (3.1) on p. 64. Some additional results are:

$n = 16$	$\bar{X} = 69.875$
$b_0 = 182.97$	$b_1 = .2616$
$SSE = 1483.0$	$MSE = 105.9$
$\sum_{i=1}^{n} (X_i-\bar{X})^2 = 3331.75$	$\sum_{i=1}^{n} (Y_i-\bar{Y})^2 = 1711.0$
$\sum_{i=1}^{n} (X_i-\bar{X})(Y_i-\bar{Y}) = 871.5$	$\sum_{i=1}^{n} X_i^2 = 81452.0$

25. (3 points) The value of the correlation coefficient r is

a) 0.187
b) 0.365
c) 0.262
d) 0.133
e) -0.133

26. (3 points) What is the predicted increase in the robbery rate for a population density increase of one person per unit area?

a) 182.97

b) 183.2316

c) .2616

d) 69.875

e) .0684

27. (3 points) A 95% confidence interval for β_1 is given by

a) (.2253, 1.687)

b) (.1484, 3748)

c) (-.1208, .6440)

d) (.8977, 1.0564)

e) (-.1129, .8071)

28. (3 points)Is the t-test for H_0: $\beta_1=0$ statistically significant at $\alpha=.05$, two-tailed? Do you reject H_0?

a) Yes, Yes

b) Yes, No

c) No, Yes

d) No, No

e) Sample size is too small for uniform convergence.

29. (3 points) A new city with a population density of 50 is to be observed. Give a 95% confidence interval for its predicted robbery rate.

a) (184.87, 207.23)

b) (189.29, 202.81)

c) (186.66, 205.44)

d) (185.702,206.398)

e) (.2253, 1.687)

30. (4 points) Explain why b_0 is meaningless here. Use the words "city" and "robbery" in your answer or you will receive no credit.

Erindale College - University of Toronto

Faculty of Arts and Science

December Examinations 1990

STA 302F

Duration - 3 hours

Aids allowed: Calculator, Textbook

Name (Please print) ______________________

Signature ______________________________

Student Number ________________________

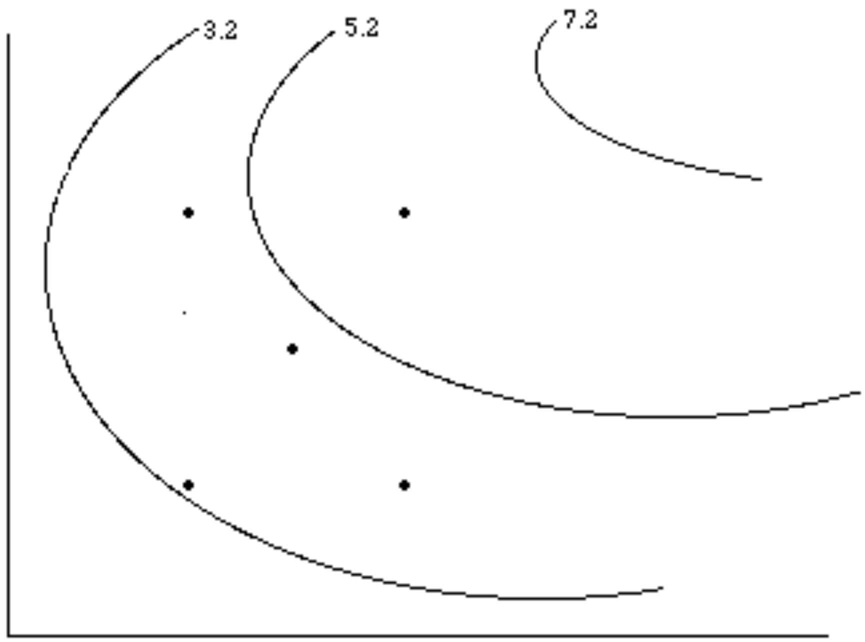
1. (5 pts) For model 7.18 on p. 237, derive the variance-covariance matrix of the vector $\widehat{\mathbf{Y}}$. Use only equations 6.45 through 6.47 on p. 203, the usual rules of matrix algebra and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

2. (7 pts) For model 7.18 on p. 237, show that $\mathbf{X}'(\mathbf{Y}-\widehat{\mathbf{Y}}) = \mathbf{0}_{p\times 1}$. Use only equations 6.45 through 6.47 on p. 203, the usual rules of matrix algebra and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

3. (3 pts) Refer to model 7.18 on p. 237. In addition, $\mathrm{Var}(\epsilon_i) = \frac{\sigma^2}{w_i}$. Using only equations 6.45 through 6.47 on p. 203, the usual rules of matrix algebra and equation (11.61) on p. 419, show that for weighted least squares, **b** is unbiased for **β**.

4. (10 pts) Suppose we have a three-category independent variable and a single quantitative independent variable. The model is $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_{12} X_{i1} X_{i2} + \beta_{13} X_{i1} X_{i3} + \epsilon_i$, where X_{i2} and X_{i3} are indicator variables for the first two categories of the qualitative independent variable. Show that the test for equality of the three slopes is the same as testing H_0: $\beta_{12}=\beta_{13}=0$.

5. (5 pts) In a response surface drug study, the contour plot of the (predicted) response surface looks something like the picture below. The five dots • are the points at which sample data were collected. Your boss shows you the contour plot and asks what should be done next. How do you respond?



6. (7 pts) Consider the SMSA data set described on p. 1161-1162, using only variables 10, 11 and 12. It is claimed that even if we control for income, crime rate still varies by geographic region. Give the matrices **C**, **β** and **h** for H_0: **Cβ**=**h**. Assume that the rate at which crime rate changes as a function of income is the same in each geographic region.

7. (7 points) Refer again to the SMSA data set, this time restricting your attention to variables 4,6,7,10 and 11, and assuming that the independent variables do not interact. You want to simultaneously determine whether (a) controlling for all other variables, number of doctors and number of hospital beds are related to crime rate, and (b) the regression coefficient for percent of population in cities is equal to one. Give the matrices **C**, **β** and **h** for H_0: **Cβ**=**h**.

Questions 8 through 11 refer to the following MINITAB output.

```
MTB > regress c4 5 c2 c3 c7 c10 c11;
SUBC> dw.

The regression equation is
C4 = 1.63 + 0.380 C2 - 0.0231 C3 + 0.00042 C7 - 0.00362 C10 + 0.00506 C11

Predictor       Coef       Stdev    t-ratio        p
Constant       1.631       1.298       1.26    0.212
C2           0.37975     0.06808       5.58    0.000
C3          -0.02308     0.02408      -0.96    0.340
C7          0.000419    0.003034       0.14    0.890
C10        -0.003622    0.003890      -0.93    0.354
C11         0.005057    0.001893       2.67    0.009

s = 1.098       R-sq = 35.9%     R-sq(adj) = 32.9%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    5      72.275      14.455     11.98    0.000
Error       107     129.105       1.207
Total       112     201.380

SOURCE       DF      SEQ SS
C2            1      57.305
C3            1       2.075
C7            1       4.018
C10           1       0.263
C11           1       8.614

Durbin-Watson statistic = 2.00
```

Continued on page 8

Questions 8-11 refer to the MINITAB output on page 7.

8. (8 pts) Fill in the blanks.

H_0	F^* or t^*	Reject H_0 at α=.05? (Yes or No)
$\beta_1=0$	________	___________
$\beta_5=0$	________	___________
$\beta_5=0$	________	___________
$\beta_3=\beta_4=\beta_5=0$	________	___________

9. (2 pts) Once X_1 has been taken into account, what proportion of the remaining variation in Y is explained by X_2, X_3,X_4 and X_5 together? ___________

10. (1 pt) Is there evidence of autocorrelation in the error terms? (Yes or No) ___________

11. (5 pts) Give a 95% Bonferroni joint confidence interval for β_2 and β_3. Use the closest df in the table as an approximation.

12. (5 pts) Consider the model for which $E(Y_i)=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_{12}X_{i1}X_{i2}$. Demonstrate that $\beta_{12}\neq0$ means that the relationship of Y to X_1 depends upon the value of X_2.

Questions 13 through 55 are worth one point each. Answer each one T for true or F for false. Scores on this section will be corrected for guessing, so guess only if you think your chances of guessing right are better than 50-50.

13. ___It is impossible for a collection of error terms to be both independent and autocorrelated.

14. ___In a situation where we are comparing regression models with differing numbers of independent variables, it makes sense to use the adjusted R^2 (adjusted for degrees of freedom) rather than the usual $R^2 = SSR/SSTO$.

15. ___Suppose that $\hat{Y} = 2.2 - 4X_1 + 11.2X_2 + 3.1X_1X_2$. For a change of one unit in X_2, the predicted change in Y is 11.2 units.

16. ___Weighted least squares is used in situations where the variance of the Y's does not appear to be equal for all levels of X.

17. ___When there is multicolinearity, it is especially important to use two-tailed rather than one-tailed t-tests.

18. ___In stepwise regression, it is possible that an X variable brought in at an early stage will be subsequently dropped if it is no

longer helpful in conjunction with X variables added at later stages.

19. ___For testing the significance of the difference between two group means, suppose we use a simple linear regression model with normal error terms and X_i a dummy variable for group membership. The t-test for H_0: $\beta=0$ is identical to the usual independent (2-sample) t-test for H_0: $\mu_1=\mu_2$.

20. ___The concept of an interaction between two quantitative independent variables is undefined.

21. ___Simple linear regression was used to test the difference in pain tolerance between males and females -- larger values of Y indicate more tolerance. The dummy variable X = 1 if the subject is female, and 0 if the subject is male. A positive value of b_1 would indicate that in the sample, males had a higher pain tolerance than females.

22. ___For the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{12} X_{i1} X_{i2} + \epsilon_i$, the least-squares method fits a plane through the cloud of (X_1,X_2,Y) points so that the sum of squared vertical distances of the points from the plane is minimized.

23. ___Suppose X is nationality (country of origin) and Y is freshman grade point average. It does not make sense to do simple

linear regression in this situation.

24. ___For the general regression model (7.18) on p. 237, the coefficient of multiple determination R^2 is exactly equal to the squared correlation r^2 between the Y_i and $\hat{Y}_i$ variables.

Continued on page 12

25. ___In stepwise regression, "tolerance" is the minimum proportion of explained variation that will allow an independent variable to be added to the model.

26. ___In the extra sum of squares approach to multiple regression, SSE(R)-SSE(F) = SSR(F)-SSR(R)

27. ___In multiple regression, the standardized regression coefficient for X_k equals the simple correlation between X_k and Y.

28. ___In well-designed experiments involving quantitative independent variables, a procedure for reducing the number of independent variables after the data are obtained is not necessary.

29. ___For the regression model 7.18 on p. 237, $(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})$ is minimized if $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, provided that $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

30. ___In multiple regression with a given significance level α, it is possible to reject $H_0:\beta_1=\beta_2=0$, while failing to reject either $H_0:\beta_1=0$ or $H_0:\beta_2=0$.

31. ___Suppose you fit a multiple regression model, obtain the residuals, and prepare a normal probability plot. If the pattern is not linear, you suspect that the assumption of normal error terms may be incorrect.

32. ___If the columns of $\mathbf{X}$ are linearly independent, $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist.

33. ___An advantage of polynomial regression is that, provided Y is an analytic function of X, it can allow valid extrapolation of the regression relation for a moderate distance beyond the range of the observed data.

34. ___In stepwise regression, the "F" statistics printed by a computer program such as SAS or MINITAB do not actually have an F distribution, and the p-values associated with these quantities are invalid.

35. ___When error terms in the regression model are positively autocorrelated, estimates of the regression coefficients are biased.

36. ___If two different dummy variables are used to represent sex of respondent in a survey, $X'X$ will <u>not</u> be singular if the constant term β_0 is omitted from the regression equation.

37. ___A market researcher would like to predict price of automobile purchased from the purchaser's age, income, sex and number of children. Multiple regression is a reasonable technique to apply.

38. ___There can be problems with the numerical accuracy of $(\mathbf{X'X})^{-1}$ when the determinant of $\mathbf{X'X}$ is close to zero.

39. ___In the presence of positively correlated error terms, the usual tests of statistical significance are still approximately valid for large samples.

40. ___If you use the extra sum of squares test H_0: $\beta_k=0$, the F^* statistic is exactly the square of $t^* = \frac{b_k}{s\{b_k\}}$.

41. ___If the independent variables are highly correlated among themselves, it is still possible to predict mean responses and new observations accurately, provided that R^2 is large and the predictions are made within the region of observations.

42. ___When severe multicolinearity exists, adding or deleting an independent variable from the model may substantially change the values of the other regression coefficients.

43. ___For the full model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ and the reduced model $Y_i = \beta_0 + \epsilon_i$, $SSE(R) = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

44. ___ When error terms are positively autocorrelated, MSE may seriously underestimate their variance.

45. ___It is possible that $b_k=0$, and yet H_0: $\beta_k=0$ is rejected.

46. ___In a study with one independent variable that takes on 5 distinct values, the pure error lack of fit test is equivalent to fitting a 5th degree polynomial and then testing H_0: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0$.

47. ___Let Y_i =sales, X_{i1}=advertising expenditures, and X_{i2} be a dummy variable for nationality (1 if the firm is Canadian, 0 if the firm is U.S.). The situation where Canadian and U.S. firms have the same slope but different intercepts is represented by the model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{12} X_{i1} X_{i2} + \epsilon_i$.

48. ___When independent variables are strongly correlated with each other, the residuals will not sum to exactly zero.

49. ___In simple linear regression with autocorrelated error terms, a good remedial measure is to take the natural log (ln) of both the X and the Y variables, adding a constant if necessary to ensure that all the numbers are positive.

50. ___ There can be problems with the numerical accuracy of $(X'X)^{-1}$ when the magnitudes of the X variables differ greatly from one another.

51. ___ When the dependent variable is measured with error, multiple regression can be extremely misleading.

52. ___ When one or more independent variable are measured with error, multiple regression can be extremely misleading.

53. ___ To represent a qualitative independent variable with k categories in a multiple regression model with an intercept, only k-1 dummy variables are needed.

54. ___The estimated standard deviations of the regression coefficients become small when the independent variables in the model are highly correlated with one another.

55. ___ Suppose you weigh yourself every morning and record the result. You would expect this time series to be negatively autocorrelated.