Material from STA302 in 1992. I have cut out a lot of blank space to save paper.

Quiz 1

1. (3 pts) A student playing a marketing game receives the following results relating advertising expenditures (X) to sales (Y): $\hat{Y} = 350.7 - .18X$, 2–sided p value for the slope = .0091. The student stated "The message I get is that the more we spend on advertising the less we sell." Comment.

2. (4 pts) Defining $k_i = \dfrac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$, show that $\sum_{i=1}^{n} k_i^2 = \dfrac{1}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$.

3. (3 pts) Computer assignment: given that $\dfrac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$, calculate a 95% confidence interval for β_1 when the independent variable is weight. You need not derive the confidence interval if you can remember it. Use $t(.975;72) = 1.9935$.

Quiz 2

1. (3 pts) Suppose we obtain this 99% confidence interval for β_1: $(-1.05727, -.452886)$.

   a) Can you conclude that there is a linear relationship between x & Y? (Yes or No)

   b) What is the implied level of significance?

   c) Is b_1 positive, or negative?

2. (4 pts) This is an exercise suggested in the lecture. Given that $\frac{b_1-\beta_1}{s\{b_1\}} \sim t(n-2)$, show that the $(1-\alpha)100\%$ confidence interval for β_1 contains zero if and only if $H_0{:}\beta_1=0$ cannot be rejected using a two–sided test with significance level α.

3. (3 pts) Hand in your log file for the computer assignment. Print your name on it neatly. Hand in <u>only</u> your log file. If you hand in your lst file instead (or as well), you get zero points. If you have 2 log files, just hand in one of them.

Name ________________________

Student Number ________________

Midterm Exam
STA 302f 1992
Erindale College
Aids Allowed: Calculator

1. Consider the following model for "regression through the origin."

$Y_i = \beta_1 x_i + \varepsilon_i$ for i = 1, ..., n, where x_i's are known constants, β_1 is an unknown constant, and ε_i's are independent $N(0,\sigma^2)$ random variables. **Do not confuse this with the model we have been considering in class.**

a) (2 pts) What is $E\{Y_i\}$?

b) (15 pts) Derive the least squares estimate of β_1. Call it $\hat{\beta}_1$. Be careful; later in the test, you will need to use it quite a bit.

1c) (5 pts) Show that the point $\hat{\beta}_1$ you have found is actually the place where the minimum occurs.

d) (3 pts) Write down the expression for SSE for this model. If you use the notation $\hat{Y}_i$, be sure you specify what it means.

e) (3 pts) What should the error degrees of freedom be for this model? If you are not sure, go ahead and guess.

f) (5 pts) For what null hypothesis is this model the reduced model? (If you think that depends on the full model, you're right. See part g on the next page).

g) (5 pts) Write down the full model for (f). Be sure to define **all** the terms you use.

2. (5 pts) Is your estimator $\hat{\beta}_1$ of problem 1 an unbiased estimator? Answer "yes" or "no" and prove your answer. (Note: If you have a computational error on question 1, you can still get full marks for this one. However, if you answer this question for the usual linear regression model instead of the model of question 1, you get a zero.)

3. (5 pts) Is the least squares estimator $\hat{\beta}_1$ of question 1 a linear combination of the form $\sum_{i=1}^{n} k_i Y_i$? Answer "yes" or "no" and prove your answer by saying what k_i is.

4. (7 pts) Find the variance of the least squares estimator $\hat{\beta}_1$ of problem 1. Show your work.

5. (5 pts) State what the probability distribution of $\hat{\beta}_1$ (still from question 1) is, and give the parameters of the probability distribution. You don't need to justify your answer; just write it down.

Questions 6 through 15 refer to the data analysis carried out with the SAS command file below. It should be familiar, except that the data file contains a smaller number of lines than you used in your homework.

```
options linesize=79 pagesize=35;
title 'STA 320F92:  Midterm SAS Job';

proc format; /* Used to label values of the categorical variables */
     value carfmt     1 = 'US'
                      2 = 'Japanese'
                      3 = 'Other' ;
data auto;
     infile 'smallcar.dat';
     input country mpg weight length;
     label country = 'Country of Origin'
           mpg = 'Miles per Gallon';
     format country carfmt.;
proc reg simple;
     model mpg=weight;
```

6. (3 pts) What is the dependent variable?

7. (3 pts) What is the independent variable?

8 (5 pts) What is wrong with these statements?

```
proc reg simple;
      model mpg=country;
```

Now refer to the tear–away sheets: Last 3 pages. Start with the SAS output.

9. (3 pts) What is the sample size n? (Give a number)

10. (3 pts) What is the estimated slope b_1? (Give a number)

11. (3 pts) What is MSR? (Give a number)

12. (5 pts) What conclusion do you draw about the weights of cars and their fuel efficiency (gas mileage)?

13. (5 pts) What conclusion can you draw about β_0? What is the meaning of this result?

14. (5 pts) What is the value of the correlation coefficient r? (Give a number)

15. (5 pts) Give a 95% confidence interval for β_1. You don't need to derive it, but show some calculations.

STA 320F92: Midterm SAS Job

1

20:27 Thursday, October 22, 1992

Descriptive Statistics

Variables	Sum	Mean	Uncorrected SS
INTERCEP	62	1	62
WEIGHT	194240	3132.9032258	644844600
MPG	1274	20.548387097	27812

Variables	Variance	Std Deviation
INTERCEP	0	0
WEIGHT	595237.33474	771.51625695
MPG	26.776308831	5.1745829621

Model: MODEL1
Dependent Variable: MPG Miles per Gallon

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1242.93183	1242.93183	191.013	0.0001
Error	60	390.42300	6.50705		
C Total	61	1633.35484			

Root MSE	2.55089	R-square	0.7610
Dep Mean	20.54839	Adj R-sq	0.7570
C.V.	12.41407		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	38.878310	1.36525496	28.477	0.0001
WEIGHT	1	-0.005851	0.00042333	-13.821	0.0001

STA 302F 92 Midterm Formula Sheet

$$b_0 = \bar{Y} - b_1\bar{X} \qquad b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

And for the normal model,

$$b_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}) , \qquad \frac{b_1-\beta_1}{s\{b_1\}} \sim t(n-2)$$

$$b_0 \sim N(\beta_0 , \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]), \qquad \frac{b_0-\beta_0}{s\{b_0\}} \sim t(n-2).$$

$\hat{Y}_h \sim N(E[Y_h], \sigma^2\{\hat{Y}_h\})$, where $E[Y_h] = \beta_0 + \beta_1 x_h$,
and $\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$. $\frac{\hat{Y}_h-E[Y_h]}{s\{\hat{Y}_h\}} \sim t(n-2)$.

$\widehat{\bar{Y}}_{h(new)} \sim N(\beta_0 + \beta_1 x_h , \sigma^2\left[\frac{1}{m} + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right])$, $\frac{\hat{\bar{Y}}_{h(new)}-\bar{Y}_{h(new)}}{s\{\hat{\bar{Y}}_{h(new)}\}} \sim t(n-2)$, leading to
the PREDICTION interval $\hat{Y}_h \pm t(1-\alpha/2;n-2)s\{\hat{Y}_{h(new)}\}$

$$SSTO = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 \quad SSE = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \qquad SSR = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$$

dfto = n–1, dfr = p–1, dfr = n–p, MS=SS/df

$$F^* = \frac{SSE(R)-SSE(F)}{dfe_R-dfe_F} \div \frac{SSE(F)}{dfe_F} \sim F(dfe_R-dfe_F,dfe_F)$$

Erindale College – University of Toronto
Faculty of Arts and Science
December Examinations 1992
STA 302F
Duration – 3 hours
Aids allowed: Calculator
Note: **Formula sheet is on the last page**

Name (Please print) ____________________

Signature ____________________________

Student Number _______________________

1. (5 points) You may recall that in simple linear regression, $\sum_{i=1}^{n} x_i e_i = 0$. For multiple linear regression, show that $\mathbf{X'e} = \mathbf{0}$.

2. (5 points) For the multiple linear regression model, show that $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$. Hint: You can use intermediate results from question 1.

3. (3 points) For the "hat matrix" $\mathbf{H}$ defined in the formula sheet, show $\mathbf{HH'}=\mathbf{H}$.

4. (5 points) For the multiple linear regression model, obtain an expression for the variance–covariance matrix of the fitted values $\hat{Y}_i$, i = 1, ..., n, in terms of the hat matrix $\mathbf{H}$. Keep simplifying!

5. (5 points) Let W_1 and W_2 be independent normal random variables, with $W_1 \sim N(\mu_1,\sigma^2)$ and $W_2 \sim N(\mu_2,\sigma^2)$. Find the joint distribution of $Y_1=W_1+W_2$ and $Y_2=W_1-W_2$. Do **not** leave the answer in terms of matrix products; perform all matrix multiplications explicitly.

6. (8 points) Suppose that we do a multiple linear regression, obtaining predicted values $\hat{Y}_i$, i = 1, ..., n. Then we treat the $\hat{Y}_i$ as the <u>independent</u> variable in a simple linear regression with the same dependent variable Y_i. Show that the R^2 value from this regression is the same as the R^2 value from the original regression. Hint: It is enough to show that SSE for the two regressions is the same. Start by writing SSE for the simple regression model.

7. (3 points) A professor stated: "Adding independent variables in a multiple regression can never reduce R^2, so we should include all available independent variables in the model." Comment.

8. For the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$ with the usual assumptions,

a) (1 point) Find $E[Y_i]$.

b) (5 points) Explain why $\beta_{12} \neq 0$ means that the rate at which the expected value of the dependent variable is changing as a function of the first independent variable <u>depends</u> on the value of the other independent variable.

9. (5 points) Make up your own example of a model with two independent variables that *cannot* be transformed into a linear regression model by taking any function of the dependent and/or the independent variables. You need not explain why it's not a linear model, but if I can transform it into one, you get zero points. I will assume Y is the observed value of the dependent variable, x_1 and x_2 are known constants (the values of the independent variable), and ε is a $N(0,\sigma^2)$ random variable.

10. Recall the car data of your very last computer assignment. There are two independent variables — Country of Origin (1=US, 2=Japan, 3=Other), and Weight. We needed a quadratic term for weight.

a) (6 points) Write a linear regression model for this problem, with country, weight (including the quadratic term), and the interaction between country and weight (again, including the quadratic term). Specify exactly what the x variables mean. I will assume you mean the usual things by the other terms. You only need to describe <u>three</u> x variables; the rest is obvious.

10b) (6 points) Give the **C**,**β** and **h** matrices for testing the null hypothesis that all regressions are linear (slopes & intercepts need not be the same).

Questions 11 through 17 refer to the computer printout starting on page 15

Please take a moment to examine the SAS command file on page 15. If you wish, you may tear off all the remaining pages starting at that page. The questions you have to answer end on page 14.

11. (1 point) Write a regression equation for the model in terms of Y_i, x's, β's and ε_i. Just write the equation. I will be able to figure out what all the terms mean. The purpose of this question is mostly to establish notation for later questions. Your x variables MUST be in same order as those in the SAS job!

12. (4 points) What is the expected number of crimes for each of the four geographic regions? Use your notation from question 11.

13. Only one of the t–tests is non–significant.

a) (1 point) What is the null hypothesis for this test? Use your notation from question 11.

b) (2 points) Of course you don't have to fit a full and reduced model to perform the test, but please state the reduced model anyway. Use your notation from question 11.

c) (4 points) What do you conclude from this test? Say something about crime or get zero points.

14. We are interested in knowing whether crime is related to geographic region, once we control for population size.

a) (3 points) What is the null hypothesis for this test? Use your notation from question 11.

b) (2 points) State the reduced model for this test.

c) (4 points) What is the value of the test statistic F*? Show your work and Circle your answer.

14d) (2 points) Approximately, what is the critical value for this test at α=.05? I know it is not in the table exactly, but you can get it with a margin of error less than a tenth.

14e) (1 point) Are the results statistically significant? YES or NO.

14f) (2 points) What do you conclude? Say something about crime or get zero points.

15. We are interested in testing whether, for a given population size, the expected number of crimes is the same for the Northern part of the U.S (average of Northeast and North Central) as for the rest of the country (average of the other 2 regions).

a) (4 points) State the null hypothesis, using your notation of question 11. Simplify as much as possible in order to get full marks.

b) (3 points) Give the p–value for this test..

16. (7 points) Give a 95% confidence interval for the expected number of serious crimes for an SMSA district in the Northeast region with a population of 900 thousand.

17. (3 points) I guess it's strange to think of a "new" SMSA being sampled from this population; maybe we are talking about data from the next year. Anyway, please obtain a 95% prediction interval for a single new SMSA in the Northeast region with a population of 900 thousand.

TOTAL MARKS = 100 POINTS: Remaining pages are printout, tables and formulas.

STA302F92 Final Exam Supplement:

Printout, Tables and Formula Sheet

Printout

Each of the cases in the "SMSA" data file is a **S**tandard **M**etropolitan **S**tatistical **A**rea — an urban census region in the United States. We are interested in predicting total number of serious crimes from total population and region of the country. The data file has other variables, but in the SAS command file below, only the ones we will use are labelled.

final.sas

```
options linesize=79 pagesize=100;
title 'STA 320F92:  Final SAS Job -- SMSA Data';

proc format; /* Used to label values of the categorical variables */
     value regfmt    1 =  'Northeast'
                     2 =  'North central'
                     3 =  'South'
                     4 =  'West';
data census;
     infile 'smsa.raw';
     input id landarea totpop urban oldfolks doctors hospbeds hsgrads
           labforce income crimes region;
     label
       totpop   = 'Estimated 1977 population in thousands'
       crimes   = 'Total serious crimes 1977'
       region   = 'Geographic region' ;
     format region regfmt.;
     if region=2 then r1=1;  else r1=0;
     if region=3 then r2=1;  else r2=0;
     if region=4 then r3=1;  else r3=0;
proc reg;
     model crimes = totpop r1 r2 r3 / covb ss1;
     Mystery: test r1=r2+r3;
```

Continued on page 16

```
                 STA 320F92:  Final SAS Job -- SMSA Data

Model: MODEL1
Dependent Variable: CRIMES     Total serious crimes 1977

                          Analysis of Variance

                             Sum of         Mean
       Source          DF      Squares       Square      F Value       Prob>F

       Model            4 1.0332623E12 258315572123      856.164       0.0001
       Error          136  41032915463  301712613.7
       C Total        140 1.0742952E12

           Root MSE   17369.87662     R-square       0.9618
           Dep Mean   56207.91489     Adj R-sq       0.9607
           C.V.          30.90290

                           Parameter Estimates

                       Parameter      Standard    T for H0:
     Variable  DF       Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1         -17511  3620.9092593        -4.836        0.0001
     TOTPOP     1      66.803749    1.15230794        57.974        0.0001
     R1         1    6626.130027  4456.2340387         1.487        0.1393
     R2         1          12988  4172.9220562         3.113        0.0023
     R3         1          22353  4689.0062999         4.767        0.0001

                                 Variable
     Variable  DF     Type I SS     Label

     INTERCEP   1  445465487233  Intercept
     TOTPOP     1  1.0255476E12  Estimated 1977 population in thousands
     R1         1     798204982
     R2         1      59726936
     R3         1    6856753717
```

Continued on page 17

Covariance of Estimates

COVB	INTERCEP	TOTPOP	R1	R2	R3
INTERCEP	13110983.864	-1603.507043	-11524153.29	-11965168.04	-11441446.69
TOTPOP	-1603.507043	1.3278135973	289.50270684	654.693103	221.01597837
R1	-11524153.29	289.50270684	19858021.807	11317283.743	11222729.28
R2	-11965168.04	654.693103	11317283.743	17413278.487	11283515.608
R3	-11441446.69	221.01597837	11222729.28	11283515.608	21986780.081

STA 320F92: Final SAS Job -- SMSA Data

```
Dependent Variable: CRIMES
Test: MYSTERY  Numerator:6770718768.3  DF:    1   F value:  22.4410
               Denominator:  3.0171E8  DF:  136   Prob>F:    0.0001
```

Formulas

Let $\mathbf{X}$ and $\mathbf{Y}$ be random matrices; let $\mathbf{A}$ and $\mathbf{C}$ be constant matrices (of the right sizes). Then

$E(\mathbf{X}+\mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$ $\quad E(\mathbf{AY}) = \mathbf{A}E(\mathbf{Y})$ $\quad \sigma^2\{\mathbf{X}+\mathbf{C}\} = \sigma^2\{\mathbf{X}\}$ $\quad \sigma^2\{\mathbf{AX}\} = \mathbf{A}\sigma^2\{\mathbf{X}\}\mathbf{A}'$

Linear Combination Theorem: If $\mathbf{Y}$ is a $k\times 1$ random matrix with $\mathbf{Y} \sim MVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$, & $\mathbf{A}$ is a $p\times k$ constant matrix of rank $p\leq k$, then $\mathbf{AY} \sim MVN(\mathbf{A}\boldsymbol{\mu},\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

For the normal multiple linear regression model in matrix form, the least squares estimator of $\boldsymbol{\beta}$ is the px1 matrix $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim MVN(\boldsymbol{\beta},\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

$MSE = (\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})/n-p$ $\quad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ $\quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$SSTO = \mathbf{Y}'[\mathbf{I} - \frac{1}{n}\mathbf{J}]\mathbf{Y}$ $\quad SSE = \mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}$ $\quad SSR = \mathbf{Y}'[\mathbf{H} - \frac{1}{n}\mathbf{J}]\mathbf{Y}$

$\sigma^2\{\hat{Y}_h\} = \sigma^2\, \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h$ and $(1-\alpha)100\%$ CI for $E(Y_h)$ is $\hat{Y}_h \pm t(1-\alpha/2;n-p)s\{\hat{Y}_h\}$

For mean of m new observations at $\mathbf{x}_h$, use prediction interval

$$\mathbf{x}_h'\mathbf{b} \pm t(1-\alpha/2;n-p)\sqrt{\frac{MSE}{m} + s^2\{\hat{Y}_h\}}\ .$$

$$F^* = \frac{SSE(R)-SSE(F)}{dfe_R-dfe_F} \div \frac{SSE(F)}{dfe_F} \sim F(dfe_R-dfe_F,dfe_F)$$

For H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{h}$, where $\mathbf{C}$ is $s\times p$ of rank $s\leq p$. $F^* = (\mathbf{Cb}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')(\mathbf{Cb}-\mathbf{h})/(s\cdot MSE)$