Graduate Statistics Fall, 2001
First midterm exam

There are two major questions, each having multiple sub-questions. Assume that each sub-question is worth 3 points.

Question A
For the first time this semester, we have several students from architecture and planning. So this problem is based (loosely!) on an urban planning problem. Over the years there has been interest in designing communities from scratch (e.g., Reston, VA., Levittown, NJ, etc.) An important issue in designing these communities is whether residential and commercial activities should be unified or separated. In a community in which these activities are unified, many people live in the same neighborhoods in which they work and most streets have both retail shops and living units. In a community in which they are separated, people travel to work in other areas and live in neighborhoods containing only residential living units. The debate continues as to which is the better plan. A researcher is interested in the effects of these two planning strategies on crime rates. She collects data from a number of census tracts from these planned communities. For each census tract, the following information is available:

TYPE: either "UNIFIED" or "SEPARATE"

INCOME median family income in $1000s

CRIME crime rate per 100,000 residents

Use these variables to specify the MODEL A/MODEL C pair you would use to answer the question. Also state the null hypothesis that each comparison tests. You may need to define new variables. If you do so, be sure to specify definitions for these new variables in a way that would allow someone else to construct the numerical values. Also, specify PA and PC (the number of parameters to be estimated from the data) for each pair of models.

1. The Department of Justice has a goal of having no more than 1500 crimes per 100,000 residents. Ignoring a census tract's TYPE and INCOME, do these communities have a crime rate that differs significantly from 1500?

2. Is a census tract's median family income related to its crime rate?

3. Assuming that median family income is related to crime rate, do a more powerful test of the question in (1) above.

4. From a previous study for which the individual data are no longer available, the following model was estimated:

CRIME = 1000 + 40 INCOME

Is this model rejected by the present census tract data?

5. Do the two types of communities—unified and separate—have different crime rates?

6. [This is beyond Chapters 1-7] Is it useful to make predictions of crime rate conditional simultaneously on both income and community type?

7. [This is not a MODEL A/MODEL C question] Some of the communities are small and so are represented by only a single census tract while other larger communities are represented by as many as six census tracts. What problems will this cause for the analysis?

8. [This is not a MODEL A/MODEL C question] Assume that the correlation between income and crime was found to be .23 in a previous study that used data from 78 census tracts. What true PRE would you use to calculate the researcher's chances of finding the same effect in the present data? How many observations should the researcher sample if she wants to have an 80% chance of finding a significant relationship again?

Question B
Data are collected on flight delays on planes departing during two different weeks, before and after September 11. From the week of September 2-8 fifteen flights were randomly sampled and from the week of September 23-29 an additional 15 flights were sampled.

The following variables were recorded for each of these flights:

LATE	Number of minutes the flight arrived later than scheduled at its destination
DISTANCE	The distance of the flight (in hundreds of miles)
PREPOST	Week (-1 if pre Sept. 11; +1 if post Sept. 11)
DPREPOST	Alternative coding of Week (0 if pre Sept. 11; +1 if post Sept. 11)

The SAS output on the following pages contains descriptive statistics on LATE and DISTANCE as well as the following three simple regression models:

MODEL LATE=DISTANCE
MODEL LATE=PREPOST
MODEL LATE=DPREPOST

Based on this output, answer the following questions.

1. FAA standards specify that planes should not be more than 7 minutes late on average to their destination. Ignoring all other variables, is there evidence from these 30 flights that the average lateness of flights is something greater

than 7 minutes? (Specify Models A and C, PRE, F*, N-PA, PA-PC, and a one sentence five-o'clock news summary conclusion.)

2. In the model where LATE is regressed on DISTANCE, provide one sentence interpretations of each of the parameter estimates.

3. Is there a significant relationship between LATE and DISTANCE? (Specify Models A and C, PRE, F*, N-PA, PA-PC, and a one sentence five-o'clock news summary conclusion.)

4. Given that DISTANCE is a predictor of LATE, provide a more powerful test of question 1: whether there is evidence from these 30 flights that the average lateness of flights is something greater than 7 minutes? (Specify Models A and C, PRE, F*, N-PA, PA-PC, and a one sentence five-o'clock news summary conclusion.)

5. What is the mean value of LATE for flights in the week before Sept. 11? What is the mean value of LATE for flights in the week after Sept. 11?

6. Is there a significant difference between the mean LATE of flights before and after Sept. 11 in these data? (Specify Models A and C, PRE, F*, N-PA, PA-PC, and a one sentence five-o'clock news summary conclusion.)

7. For the flights prior to Sept. 11, can we conclude that the average lateness of flights is something greater than 7 minutes? (Specify Models C and A and whatever you use to answer this question.)

The MEANS Procedure

Variable	Mean	Std Dev	Corrected SS	USS
late	10.5000000	10.1599280	2993.50	6301.00
distance	12.9666667	6.4030345	1188.97	6233.00

The REG Procedure
Model: MODEL1
Dependent Variable: late

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	652.06222	652.06222	7.80	0.0093
Error	28	2341.43778	83.62278		
Corrected Total	29	2993.50000			

Root MSE	9.14455	R-Square	0.2178
Dependent Mean	10.50000	Adj R-Sq	0.1899
Coeff Var	87.09094		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20.10258	3.82266	5.26	<.0001
distance	1	-0.74056	0.26520	-2.79	0.0093

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	12.27222	27.93294
distance	1	-1.28380	-0.19732

The REG Procedure
Model: MODEL2
Dependent Variable: late

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	124.03333	124.03333	1.21	0.2806
Error	28	2869.46667	102.48095		
Corrected Total	29	2993.50000			

Root MSE	10.12329	R-Square	0.0414
Dependent Mean	10.50000	Adj R-Sq	0.0072
Coeff Var	96.41226		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.50000	1.84825	5.68	<.0001
prepost	1	-2.03333	1.84825	-1.10	0.2806

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	6.71403	14.28597
prepost	1	-5.81930	1.75264

The REG Procedure
Model: MODEL3
Dependent Variable: late

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	124.03333	124.03333	1.21	0.2806
Error	28	2869.46667	102.48095		
Corrected Total	29	2993.50000			

Root MSE	10.12329	R-Square	0.0414
Dependent Mean	10.50000	Adj R-Sq	0.0072
Coeff Var	96.41226		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.53333	2.61382	4.80	<.0001
dprepost	1	-4.06667	3.69650	-1.10	0.2806

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	7.17916	17.88750
dprepost	1	-11.63861	3.50527