73-360, Fall 2000
Final Exam, Solution

The SAS output is displayed below. Only the relevant parts have been retained. Also, I have only answered the questions which are in the scope of this class.

Before we begin, some explanatory notes about the SAS output and my assumptions in answering the questions. We first note that since this is data from the US Census, the number of data points is sufficiently high that we can use the normal distribution for all of our hypothesis tests and confidence intervals.

Secondly, note that only 1.8% of the households are father-only households. Hence we ignore these households when we answer the questions; we stick to classifying all households as two-parent or mother-only. (The main reason for this assumption is that I couldn't find anything in the SAS output which uses the POPONLY variable; the preceding line is only an attempt to justify this.)

The Maximum Likelihood Estimates are precisely the Least Squares estimates, since we studied in class that LS is the Best Linear Unbiased Estimator under various assumptions which we expect to be satisfied here.

```
                                The SAS System                                1
                                        23:28 Monday, December 18, 2000

                                The MEANS Procedure

Variable       N            Mean         Std Dev         Minimum         Maximum
-------------------------------------------------------------------------------
HHINCOME    1219        39602.43        34610.17               0       353022.00
EAST        1219       0.1903199       0.3927146               0       1.0000000
MIDWEST     1219       0.2543068       0.4356496               0       1.0000000
SOUTH       1219       0.3634126       0.4811796               0       1.0000000
WEST        1219       0.1919606       0.3940039               0       1.0000000
MOMONLY     1219       0.1591468       0.3659631               0       1.0000000
POPONLY     1219       0.0180476       0.1331781               0       1.0000000
BOTH        1219       0.8228056       0.3819899               0       1.0000000
EMPPAR      1219       0.8777687       0.3276872               0       1.0000000
COLPAR      1219       0.2649713       0.4414990               0       1.0000000
-------------------------------------------------------------------------------
                                The SAS System                                2
                                        23:28 Monday, December 18, 2000
```

The REG Procedure
Model: MODEL1
Dependent Variable: HHINCOME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.563083E11	1.563083E11	144.62	<.0001
Error	1195	1.291596E12	1080833248		
Corrected Total	1196	1.447904E12			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44695	1038.07536	43.06	<.0001
MOMONLY	1	-31009	2578.54724	-12.03	<.0001

The REG Procedure
Model: MODEL2
Dependent Variable: HHINCOME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.179007E11	54475181299	52.79	<.0001
Error	1192	1.230003E12	1031882006		
Corrected Total	1196	1.447904E12			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	58917	2185.28970	26.96	<.0001
MOMONLY	1	-31317	2523.84801	-12.41	<.0001
MIDWEST	1	-19692	2816.87040	-6.99	<.0001
SOUTH	1	-18078	2630.92220	-6.87	<.0001
WEST	1	-13278	3024.84012	-4.39	<.0001

The REG Procedure
Model: MODEL3
Dependent Variable: HHINCOME

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	35700	3407.49162	10.48	<.0001
MOMONLY	1	-20586	2493.86931	-8.25	<.0001
EMPPAR	1	14823	2835.30947	5.23	<.0001
COLPAR	1	28262	1943.97632	14.54	<.0001
MIDWEST	1	-17344	2558.05447	-6.78	<.0001
SOUTH	1	-18247	2387.21705	-7.64	<.0001
WEST	1	-11712	2755.85150	-4.25	<.0001

1. What is your best estimate and a 90% confidence interval for the difference in income between a family with both parents and one with only the mother present?

   We assume that we are using a regression model where we have the region, employment and college education variables factored in. The output of this regression is shown in MODEL3 in the SAS output.

   The coefficient of MOMONLY is -20,586. Thus our best estimate is that holding other factors constant, a mother-only family earns $20,586 less than a family with both parents.

   A 90% confidence interval is given in the usual way as $b_M \pm z_{0.05} s_M$, where b_M and s_M are the coefficient and standard error estimates of the MOMONLY variable. Using $z_{0.05} = 1.64$, a 90% confidence interval for the difference is $[-24676, -16496]$.

   Note that using the other models would have given a significantly different estimate. This suggests a correlation between MOMONLY and EMPPAR or COLPAR. Since we want to hold all other factors constant, it is advisable to use the more complete model where we do have EMPPAR and COLPAR.

2. How much, on average, does a family with both parents present make (estimate a 90% confidence interval)?

   MODEL1 measures the effect of only the coefficient MOMONLY. In other words, it measures the difference in incomes between an average (with respect to location, employment and college status) family which is two-parent versus mother-only. Hence the intercept in this regression equation is our best estimate for the income of a family with both parents, and it is $44,695.

   A 90% confidence interval is computed in the same way as in Question 1. The interval is [42993, 46397].

   Note that a more accurate way to compute the average income would be to use the "more complete" model, MODEL3, with the mean values of the other variables. This gives an estimate of annual income to be $42,911, which is just outside the 90% confidence interval computed above. However, using MODEL3 makes it harder to compute a 90% confidence interval, since we will have to incorporate the standard deviations of all variables.

3. Holding constant region, test the theory that the difference between the incomes of a family with both parents present and one with the mother only present is \$35,000 in favour of the two-parent family.

   We will first do this test on MODEL1. The null hypothesis is that $\beta_M = -35,000$.. The alternate hypothesis is $\beta_M \neq -35,000$, so we will do a two-sided test. We have $b_M = -31,009$ and $s_M = 2579$ from MODEL1.

   Our z-statistic is $\frac{b_M - (-35000)}{s_M} = \frac{3991}{2579} = 1.547$. Say we do a 90% test, in which case we compare the z-statistic with $z_{0.05} = 1.64$. Since our z-statistic is less than $z_{0.05}$, we accept the null hypothesis.

   We find no statistical evidence to doubt the fact that the difference is in fact \$35,000 in favour of the two-parent family.

   We could repeat the test with MODEL3, in which case we would reject the null hypothesis. In my opinion, this is because of the correlation between EMPPAR, COLPAR and MOMONLY.

8. Would it be a good idea to include the dummy variable East in the regressions in which I already have West, South and Midwest? What would happen, good or bad - what would be the advantages and disadvantages?

   The population has been classified into exactly four regions: East, West, South and Midwest. This can be verified by checking that their means add up to 1. Hence we have the following perfect linear relationship between these four variables: `EAST + WEST + SOUTH + MIDWEST = 1`. Having all four variables in our model would violate the assumption of no perfect relationship, and our model would fail. In particular, Eviews would report a "Near singular matrix" error.

   Essentially, when we use dummy variables we have to keep an "omitted contrast", which in this case is being served by the EAST variable.

9. Does region affect income?

   To test this hypothesis, we need to test the null hypothesis that the coefficients of WEST, SOUTH and MIDWEST are all zero, against the alternative that at least one variable is non-zero. To do this, we need an unrestricted model which includes all three variables, and a restricted model which has none of these three but all the other variables of the unrestricted model. Hence we use MODEL2 as our unrestricted model, and MODEL1 as our restricted model.

   We do the test at a 95% significance level.

   The sum of squared errors of the fitted regression is shown as `Error` in the SAS output.

   Hence we have $SSE_{UR} = 1.230 \times 10^{12}$ and $SSE_R = 1.292 \times 10^{12}$. Our F-statistic is therefore $\frac{(1.292-1.230)/3}{1.230/(1196-4-1)} = \frac{0.0207}{0.001033} = 20.04$. Note that the DF coloumn in the SAS output stands for "degrees of freedom", and hence the total degrees of freedom gives us $n = 1196$. Since $F_{3,1191,0.05} = 2.60$, we reject the null hypothesis.

   We conclude that we have sufficient statistical evidence to assert that region does affect income.