Name Mr. Key
ID 777777

ECO 5350
Intro. Econometrics

Prof. T. Fomby
Fall 2007

Mid-Term Exam II

Instructions: Put your name and student ID in the upper right-hand-corner of this exam. This exam is worth a total of 78 points. The breakout of these points by questions is as follows:

Q1: a=3, b=3, c=3, d=2, e=2, f=2, g=2, h=4, i=2, j=2 (25 points)
Q2: 4 points
Q3: a=2, b=2, c=2, d=2, e=2, f=4 (14 points)
Q4: a=3, b=4, c=4 (11 points)
Q5: 6 points
Q6: 4 points
Q7: a=6, b=4, c=2, d=2 (14 points)

You have one hour and twenty minutes to take this test. A word from the wise: Don't get hung up on any one question. Answer the easy questions first and then go back and pick up the hard ones later. Good luck.

Here is a bonus questions worth **2** points. Did you hear about the Aggie who was such a poor reader he belonged to the Page of the Month Club.

+2

1. Let's start off with some easy short answer questions.

a) In prediction analysis using regression models, there are three types of predictions: They are

Case a. Prediction of Y_0 given $X = X_0$

Case b. Prediction of $E(Y_0 \mid X = X_0)$

Case c. Prediction of Y_0 given $\hat{X}_0 = X_0 + \varepsilon$

(3) Given these three separate prediction problems, the most uncertain one is Case C. The next most uncertain one is Case a. The least uncertain one is Case b.

b) For Case a and Case b the optimal unbiased predictor (3) is $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ while for Case C the optimal unbiased predictor is $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_0$.

c) Let q represent the quantity demanded of a product and p represent its price. Consider the regression equation $\hat{\log}(q) = 5 - 0.7\log(p)$. In this case the data is (3) suggesting that the elasticity of demand for the product is -0.7. This demand for this product is (elastic / inelastic / unitary elastic). Given a one percent change in prices one would expect a 0.7% change in quantity demanded.

d) (True) or False. The Frisch-Waugh theorem tells us that to get the OLS estimate of β_2 (2) in the regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ we could regress (through the origin) the residuals obtained from the regression of y on a constant and x_1, say $\hat{u}_{y|x_1}$, on the residuals obtained from the regression of x_2 on a constant and x_1, say $\hat{u}_{x_2|x_1}$, and in so doing obtain $\hat{\beta}_2$, the OLS estimate of $\hat{\beta}_2$.

e) Assume that the true consumption function for consumers is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

where y_i = the annual consumption of the i-th individual, x_{i1} = the annual wage income of the i-th individual, x_{i2} = the annual rents, interest, and royalties of the i-th individual, and u_i satisfies the classical assumptions of the multiple regression model. Now economic theory tells us that $\beta_1 > 0$ and $\beta_2 > 0$. Furthermore, assume that $\text{cov}(x_{i1}, x_{i2}) > 0$ and that you naively apply OLS to the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i$$

resulting in the estimated equation $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{u}_i$. In this case the estimated marginal propensity to consume out of wage income ($\hat{\beta}_1$) is probably too (high / low) given that the true consumption function is dependent on both wage income and income from rents, etc. (2)

f) Consider the following regression equation: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$.
Suppose that you are interested in testing the following null hypothesis: $H_0 : \beta_1 - \beta_2 = 0$. Furthermore, assume that you have the following information $\hat{\beta}_1 = 8$, $\hat{\beta}_2 = 4$, $\hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 2$, $\hat{Var}(\hat{\beta}_1) = 14$, $\hat{Var}(\hat{\beta}_2) = 15$. Then the appropriate t-statistic for testing the above null hypothesis is 4/5. You will not need a calculator for this computation. The numbers I have chosen make it an easy number to compute. (2)

$se(\hat{\beta}_1 - \hat{\beta}_2) = [\hat{Var}(\hat{\beta}_1) + \hat{Var}(\hat{\beta}_2) - 2\hat{Cov}(\hat{\beta}_1, \hat{\beta}_2)]^{1/2}$
$= [14 + 15 - 4]^{1/2} = \sqrt{25} = 5$
$\therefore t = \frac{8-4}{5} = \frac{4}{5}$

g) Consider the following hypothetical wage equation concerning the wages of female and male elementary school teachers (y) as a function of the number of years of schooling the teacher has (x) and the number of years of experience of the teacher (w). Let D be a dummy variable that is 1 if the observation is for a male teacher and 0 if the observation is for a female teacher. The **reference group** here is (female teachers / male teachers). (2)

h) Given the definition of variables in part g) above, consider the following estimated wage equation: yhat = 20 +3D + 0.5x + 0.1Dx + 0.4w + 0.05Dw where D is the **additive dummy**, Dx represents the **multiplicative dummy** for years of schooling (i.e. D times x) and Dw represents the **multiplicative dummy** for years of experience (i.e. D times w). Given this information I want you to write out the estimated wage equation for females and the estimated wage equation for males. They are:

(4)

Females: yhat = $20 + 0.5x + 0.4w$
Males: yhat = $(20+3) + (.5+.1)x + (.4+.05)w = 23 + 0.6x + 0.45w$

i) (True / False). Multicollinearity often manifests itself when multiple regression is applied to observational data in economics and in other disciplines. One of its tell-tale signs is found in highly significant overall F-statistics but insignificant individual t-statistics on the coefficients in multicollinear models. (2)

j) True or (False.) Thus, two major consequences of multicollinearity are biasedness and large sampling variances for the OLS coefficient estimates. (2)

2. (Working with models that have quadratic terms). Consider the computer output # 1 that you have been given. It reports a regression of profits of a firm on its level of output denoted by q. Given the information provided by this output, at what level of output, say q*, would you expect to maximize profits? q* = 50.65. Show me how you got your answer. (4)

$\hat{profit} = 19.61067 + 0.41736q + (-0.00412)q^2$

$\frac{d\hat{profit}}{dq} = 0.41736 - 2(0.00412)q^* = 0$

$\therefore q^* = \frac{0.41736}{2(0.00412)} \doteq 50.65$

3. (Tests of Linear Hypothesis in Linear Regression Models). Consider the computer output # 2 that you have been given. It reports a regression of a Cobb-Douglas production function of a firm where q = output, and L and K denote, respectively, the units of labor and units of capital used in producing the output q. Also logq = log(q), logk = log(K), and logl = log(L).

(2) a) True or False. This regression model is an iso-elastic regression model.

(2) b) According to the regression coefficient estimates we have here, if there is a one percent increase in capital, other things held constant, we will have a 0.3667% change in output.

(2) c) According to the regression coefficient estimates we have here, if there is a one percent increase in labor, other things held constant, we will have a 0.701% change in output.

(2) d) According to the regression coefficient estimates we have here, if there is a one percent increase in capital and labor, other things held constant, we will have a 1.067% change in output. $1.067 = 0.366 + 0.701$

(2) e) According solely to the regression coefficient estimates we have here, there appears to be (constant / increasing / decreasing) returns to scale in production in this firm.

(4) f) Now we would like to conduct a test of the null hypothesis of constant returns to scale, $H_0 : \beta_1 + \beta_2 = 1$ against the alternative hypothesis of $H_1 : \beta_1 + \beta_2 \neq 1$. Use the Variance-Covariance Matrix approach to test these hypotheses at the 5% level of significance via a t-statistic. What is your conclusion? Be sure and show how you calculated your t-statistic. If you need them, you will find t and F tables following the formula sheet that you have been given.

$$\widehat{Var}(\hat{\beta}_1 + \hat{\beta}_2) = \widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_2) + 2\,\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$
$$= 0.003105863 2 + 0.001086774 4 + 2(-0.000011305)$$
$$= 0.004170027 6$$
$$\therefore se(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{0.004170027 6} = 0.064575 7509$$
$$t = (0.3663 7 + 0.701 01)/0.064575 7509 = 1.043 < 1.96$$

Accept H_0 at c.v.t.s.

4. (Prediction in Multiple Linear Regression). Consider the computer output # 3 that you have been given. It reports a regression School District Score on a 5th grade test as a function of the student-to-teacher ratio average across the district in the 5th grade classes tested (st_ratio), expenditures per pupil on average across the district in the 5th grades classes tested (expend_pupil), and the average percent of students taking English as a second language in the 5th grade classes tested (English). Use this output to answer the following questions.

Note: $\sqrt{F_{1,97}}$

$= \sqrt{1.09}$

$= 1.04403$

The p-value of the F-statistic is 0.299 which is equal to the two-sided p-value for $t = 1.043$.

(3) a) From this output we can tell that we are interested in forecasting at the points of st_ratio = 20, expend_pupil = 5550, and English = 50.

b) At the above forecasting points what is the unbiased and efficient prediction of $E(Y \mid X = X_0)$? $\hat{E}(Y \mid X = X_0) =$ 632.329. What is the standard error of this prediction? $se(\hat{E}(Y \mid X = X_0)) =$ 1.52102. It follows that the 95% confidence interval for the prediction $E(Y \mid X = X_0)$ is [629.347, 635.310]. Show your work to get credit. Note: $t_{416,0.025} = 1.96$.

(4) $632.329 \pm 1.52102(1.96) = [629.347, 635.310]$

c) At the above forecasting points what is the unbiased and efficient prediction of $(Y_0 \mid X = X_0)$? $(\hat{Y}_0 \mid X = X_0) =$ 632.329. What is the standard error of this prediction? $se((\hat{Y}_0 \mid X = X_0)) =$ 14.43469. It follows that the 95% confidence interval for the prediction $(Y_0 \mid X = X_0)$ is [604.037, 660.621]. Show your work to get credit. Note: $t_{416,0.025} = 1.96$.

(4) $se(\hat{Y}_0 \mid X = X_0) = \sqrt{MSE(Error) + [se(\hat{E}(Y \mid X = X_0))]^2} = \sqrt{206.04694 + 1.52^2}$

$= 14.43469576$

$632.329 \pm 14.43469576(1.96) = [604.037, 660.621]$

5. (Calculation of a "Subset" F-statistic). Consider the computer output # 4 that you have been given. It reports some regressions involving the variables described in question 4 above. Use this output to test the following hypotheses at the 5% level.

(6) $H_0 : \beta_2 = \beta_3 = 0$ (that is, the variables "expend_pupil" and "english" are jointly insignificant) versus $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both. Show how you calculate the F-statistic for this test. What is the p-value of your F-statistic? What conclusion do you draw from evaluating your F-statistic?

$$F = \frac{(SSR_r - SSR_u)/2}{SSR_u/(420-4)} = \frac{(R_u^2 - R_r^2)/2}{(1-R_r^2)/(402-4)}$$

$$= \frac{(144312 - 85716)/2}{85716/416} = \frac{(0.4365 - 0.0512)/2}{(1-0.4365)/416} = 142.19$$

Since p-value of F-statistic is less than 0.0001 we reject H_0 and accept H_1.

6. (The Frisch-Waugh Theorem). Consider the computer output # 5 that you have been given. It demonstrates the use of the Frisch-Waugh theorem to get an OLS coefficient estimate and its standard error. Given only the results of this program you should be able to fill in one of the coefficients estimates and its standard error below:

(4)

$$score = \underset{(\quad)}{\_\_\_\_\_} + \underset{(0.47882)}{-0.28612} * st\_ratio + \underset{(\quad)}{\_\_\_\_\_} * expend\_pupil + \underset{(\quad)}{\_\_\_\_\_} * english$$

This is the only information you can extract from the computer output.

7. (Chow Test of Difference of Population Regression Functions Across Groups). Consider the computer output # 6 that you have been given. It reports a regression analysis of the potential differences male versus female population regressions involving the student's cumulative gpa (cumgpa) as a function of the student's SAT score (sat) and ranking as a percentage of his/her graduating class (hsperc). The variable female = 1 for a female student and 0 otherwise.

a) Conduct a Chow test of significant difference between these groups of students. Be sure and show how you calculated the Chow F-statistic. What is null hypothesis of the Chow test in this case? What is the alternative hypothesis of the Chow test? How is the numerator degrees of freedom of your test determined? How is the denominator degrees of freedom of your test determined? What conclusion do you draw from this test?

$$F = \frac{(87.78752 - 80.54193)/3}{80.54193/(366-6)} = \frac{(0.3893 - 0.3344)/3}{(1-0.3893)/(366-6)}$$

(6) numdf = 3
dendf = 360 $= 10.80$

since the p-value of the F-statistic is less than 0.0001, we reject H_0 and accept the alternative hypothesis that the male and female PRFs are different.

b) In the latter part of the computer output # 6, you will see the results of a "backward" selection analysis of the additive/multiplicative form of the regression equation. In the below area write out the "final" equations of the female and male cumulative GPA equations. You need not provide the standard errors of the estimates.

(4)

$$\widehat{cumgpa}_{female} = 1.63092 + (0.00105 + 0.00035005)\,sat - 0.00908\,hsperc$$
$$= 1.63092 + 0.00140005\,sat - 0.00908\,hsperc$$

$$\widehat{cumgpa}_{male} = 1.63092 + 0.00105\,sat - 0.00908\,hsperc$$

(2) c) The reference group in the model is male students

(2) d) Since the above "final" model had been obtained by "data dredging" we should adjust the p-values of our coefficient t-statistics by a factor of 5/3. Briefly explain how you got your answer.

$\alpha_{Adj} = \frac{c}{k}\alpha = \frac{5}{3}\alpha \therefore \frac{5}{3}$ is the "adjustment" factor

c = no. of candidate regressors (not including intercept) that you started with = 5

k = no. of regressors (not including intercept) left in the final "data dredged" model

Therefore we should adjust all of the "naive" computer p-values of the final model by the factor 5/3.

Output #1

```
data profits;
  input profit q;
datalines;
.
.
.
;

data profits;
  set profits;
  q2 = q*q;

proc gplot data=profits;
  symbol1 c=black i=spline v=dot h=.5;
  title 'Profits as a Function of Output';
  axis1 order=(0 to 120 by 10)
  label=(f=duplex 'Output');
  axis2 order=(0 to 130 by 10)
  label=(f=duplex 'Profits');
  plot profit*q / haxis=axis1 vaxis=axis2;

run;

proc reg data = profits;
    model profit = q q2;

      run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: profit

Number of Observations Read	120
Number of Observations Used	120

Analysis of Variance

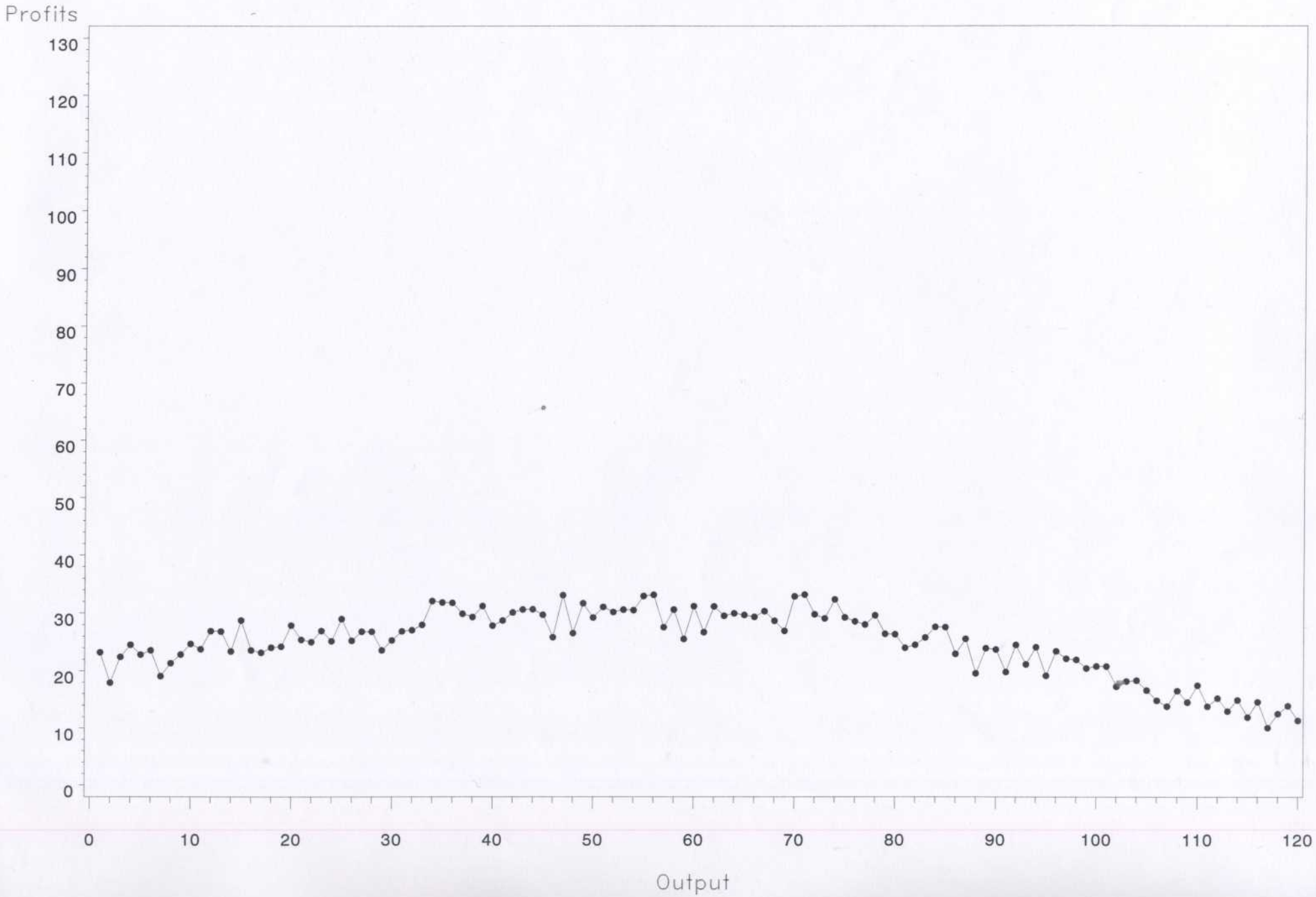
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3307.82761	1653.91381	384.67	<.0001
Error	117	503.04700	4.29955		
Corrected Total	119	3810.87462			

Root MSE	2.07353	R-Square	0.8680
Dependent Mean	24.82042	Adj R-Sq	0.8657
Coeff Var	8.35415		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.61067	0.57746	33.96	<.0001
q	1	0.41736	0.02203	18.94	<.0001
q2	1	-0.00412	0.00017639	-23.38	<.0001

Profits as a Function of Output



Output #2

```
/* Testing Cobb-Douglas Production Function for Constant Returns to
Scale.  Q = output, L = labor, K = capital.  */

data production;
  input q k l;
datalines;
.
.
.
;

data production;
  set production;
  logk = log(k);
  logl = log(l);
  logq = log(q);


  title 'Cobb-Douglas Production Function';
proc reg data = production;
   model logq = logk logl/covb;
   test logk + logl = 1;
   run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: logq

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.38226	0.69113	248.57	<.0001
Error	97	0.26970	0.00278		
Corrected Total	99	1.65196			

Root MSE	0.05273	R-Square	0.8367
Dependent Mean	7.68056	Adj R-Sq	0.8334
Coeff Var	0.68654		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.60366	0.27942	12.90	<.0001
logk	1	0.36637	0.05573	6.57	<.0001
logl	1	0.70101	0.03297	21.26	<.0001

Covariance of Estimates

Variable	Intercept	logk	logl
Intercept	0.078077471	-0.01425283	-0.003654859
logk	-0.01425283	0.0031058632	-0.000011305
logl	-0.003654859	-0.000011305	0.0010867744

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable logq

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00303	1.09	0.2993
Denominator	97	0.00278		

Output #3

```
Data Combined;
  input obs score st_ratio expend_pupil english;
  datalines;
.
.
.
;

Data combined;
set combined;
  xstar1 = st_ratio - 20;
  xstar2 = expend_pupil - 5500;
  xstar3 = english - 50;

  title 'Original Regression';
proc reg data = combined;
   model score = st_ratio expend_pupil english;

   run;

   title 'Translated Regression';
proc reg data = combined;
   model score = xstar1 xstar2 xstar3;

   run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66386	22129	107.40	<.0001
Error	416	85716	206.04694		
Corrected Total	419	152101			

Root MSE	14.35434	R-Square	0.4365
Dependent Mean	654.16024	Adj R-Sq	0.4324
Coeff Var	2.19431		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	649.58142	15.20658	42.72	<.0001
st_ratio	1	-0.28612	0.48055	-0.60	0.5519
expend_pupil	1	0.00387	0.00141	2.74	0.0064
english	1	-0.65598	0.03911	-16.77	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66386	22129	107.40	<.0001
Error	416	85716	206.04694		
Corrected Total	419	152101			

Root MSE	14.35434	R-Square	0.4365
Dependent Mean	654.16024	Adj R-Sq	0.4324
Coeff Var	2.19431		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	632.32914	1.52102	415.73	<.0001
xstar1	1	-0.28612	0.48055	-0.60	0.5519
xstar2	1	0.00387	0.00141	2.74	0.0064
xstar3	1	-0.65598	0.03911	-16.77	<.0001

Output #.4

```
Data Combined;
  input obs score st_ratio expend_pupil english;
  datalines;
.
.
.
;


  title 'Unrestricted Regression';
proc reg data = combined;
   model score = st_ratio expend_pupil english;
      test expend_pupil = 0, english = 0;

   run;

  title 'Restricted Regression';
proc reg data = combined;
   model score = st_ratio;

   run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66386	22129	107.40	<.0001
Error	416	85716	206.04694		
Corrected Total	419	152101			

Root MSE	14.35434	R-Square	0.4365
Dependent Mean	654.16024	Adj R-Sq	0.4324
Coeff Var	2.19431		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	649.58142	15.20658	42.72	<.0001
st_ratio	1	-0.28612	0.48055	-0.60	0.5519
expend_pupil	1	0.00387	0.00141	2.74	0.0064
english	1	-0.65598	0.03911	-16.77	<.0001

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable score

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	29298	142.19	<.0001
Denominator	416	206.04694		

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7789.39479	7789.39479	22.56	<.0001
Error	418	144312	345.24414		
Corrected Total	419	152101			

Root MSE	18.58075	R-Square	0.0512
Dependent Mean	654.16024	Adj R-Sq	0.0489
Coeff Var	2.84040		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	698.92218	9.46719	73.83	<.0001
st_ratio	1	-2.27906	0.47981	-4.75	<.0001

output #5

```
Data Combined;
  input obs score st_ratio expend_pupil english;
  datalines;
.
.
.
;

proc reg data = combined;
  model score = expend_pupil english;
   output out=result1 r=residscore;
   title 'Regression Producing Residscore';
run;

proc reg data = combined;
   model st_ratio = expend_pupil english;
     output out=result2 r=residst_ratio;
       title 'Regression Producing Residst_ratio';
run;

data together;
  merge combined result1 result2;

run;

proc reg data = together;
   model residscore = residst_ratio/noint;
   title 'Final Frisch-Waugh Theorem Regression';
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	66313	33156	161.17	<.0001
Error	417	85789	205.72798		
Corrected Total	419	152101			

Root MSE	14.34322	R-Square	0.4360
Dependent Mean	654.16024	Adj R-Sq	0.4333
Coeff Var	2.19262		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	641.26371	6.00189	106.84	<.0001
expend_pupil	1	0.00439	0.00111	3.96	<.0001
english	1	-0.66024	0.03842	-17.18	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: st_ratio

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	607.39263	303.69632	141.93	<.0001
Error	417	892.26116	2.13972		
Corrected Total	419	1499.65379			

Root MSE	1.46278	R-Square	0.4050
Dependent Mean	19.64050	Adj R-Sq	0.4022
Coeff Var	7.44776		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.07101	0.61210	47.49	<.0001
expend_pupil	1	-0.00182	0.00011302	-16.10	<.0001
english	1	0.01492	0.00392	3.81	0.0002

The REG Procedure
Model: MODEL1
Dependent Variable: residscore Residual

Number of Observations Read	420
Number of Observations Used	420

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	73.04308	73.04308	0.36	0.5505
Error	419	85716	204.57166		
Uncorrected Total	420	85789			

Root MSE	14.30286	R-Square	0.0009
Dependent Mean	9.9893E-13	Adj R-Sq	-0.0015
Coeff Var	1.431817E15		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
residst_ratio	Residual	1	-0.28612	0.47882	-0.60	0.5505

Output # 6

```
/* cumgpa = cumulative GPA
   female = 1 if female, 0 otherwise
   sat = SAT Score
   hsperc = 100*(rank in class/size of High School graduating class) =
            high school percentile in graduating class */

data gpa;
  input cumgpa sat hsperc female;
datalines;
.
.
.
;

 data gpa;
   set gpa;
   satfem = sat*female;
   hspercfem = hsperc*female;

     title 'Additive/Multiplicative Dummy Variable Equation';
 proc reg data = gpa;
  model cumgpa = sat hsperc female satfem hspercfem;
      test satfem = 0, hspercfem = 0, female = 0;

  run;

/*  Pooled regression  */
   title 'Combined (Pooled) Regression';
  proc reg data = gpa;
     model cumgpa = sat hsperc;

  run;

  /* A "refined" model  */
   title 'A refined model';

proc reg data = gpa;
  model cumgpa = sat hsperc female satfem;

  run;

 /* An even more refined model  */
  title 'An Even More Refined Model';

proc reg data = gpa;
  model cumgpa = sat hsperc satfem;

  run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: cumgpa

Number of Observations Read	366
Number of Observations Used	366

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	51.35176	10.27035	45.91	<.0001
Error	360	80.54193	0.22373		
Corrected Total	365	131.89369			

Root MSE	0.47300	R-Square	0.3893
Dependent Mean	2.33415	Adj R-Sq	0.3809
Coeff Var	20.26425		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.73803	0.18652	9.32	<.0001
sat	1	0.00093556	0.00017792	5.26	<.0001
hsperc	1	-0.00916	0.00136	-6.73	<.0001
female	1	-0.48952	0.39669	-1.23	0.2180
satfem	1	0.00087715	0.00038696	2.27	0.0240
hspercfem	1	-0.00024825	0.00318	-0.08	0.9377

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable cumgpa

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	2.41520	10.80	<.0001
Denominator	360	0.22373		

The REG Procedure
Model: MODEL1
Dependent Variable: cumgpa

Number of Observations Read	366
Number of Observations Used	366

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	44.10616	22.05308	91.19	<.0001
Error	363	87.78752	0.24184		
Corrected Total	365	131.89369			

Root MSE	0.49177	R-Square	0.3344
Dependent Mean	2.33415	Adj R-Sq	0.3307
Coeff Var	21.06851		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.71872	0.17033	10.09	<.0001
sat	1	0.00110	0.00016420	6.68	<.0001
hsperc	1	-0.01057	0.00124	-8.50	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: cumgpa

Number of Observations Read	366
Number of Observations Used	366

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	51.35039	12.83760	57.54	<.0001
Error	361	80.54329	0.22311		
Corrected Total	365	131.89369			

Root MSE	0.47235	R-Square	0.3893
Dependent Mean	2.33415	Adj R-Sq	0.3826
Coeff Var	20.23633		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.74172	0.18020	9.67	<.0001
sat	1	0.00093336	0.00017546	5.32	<.0001
hsperc	1	-0.00920	0.00123	-7.50	<.0001
female	1	-0.50582	0.33700	-1.50	0.1342
satfem	1	0.00088739	0.00036362	2.44	0.0151

The REG Procedure
Model: MODEL1
Dependent Variable: cumgpa

Number of Observations Read	366
Number of Observations Used	366

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	50.84775	16.94925	75.71	<.0001
Error	362	81.04594	0.22388		
Corrected Total	365	131.89369			

Root MSE	0.47316	R-Square	0.3855
Dependent Mean	2.33415	Adj R-Sq	0.3804
Coeff Var	20.27132		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.63092	0.16467	9.90	<.0001
sat	1	0.00105	0.00015823	6.62	<.0001
hsperc	1	-0.00908	0.00123	-7.40	<.0001
satfem	1	0.00035005	0.00006379	5.49	<.0001

FORMULA SHEET FOR
MID-TERM II

SOME OLS REGRESSION FORMULAS:

1. $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$; $\quad Var(\hat{\beta}_0) = \dfrac{\sigma^2 \sum_1^N X_i}{N \sum_1^N (X_i - \overline{X})^2}$

2. $\hat{\beta}_1 = \dfrac{\sum_1^N (X_i - \overline{X}) Y_i}{\sum_1^2 (X_i - \overline{X})^2} = \sum_1^N w_i Y_i$; $\quad Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_1^N (X_i - \overline{X})^2}$

3. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

4. SST = SSE + SSR ; $\sum_1^N (Y_i - \overline{Y})^2 = \sum_1^N (\hat{Y}_i - \overline{Y})^2 + \sum_1^N (Y_i - \hat{Y}_i)^2$; $R^2 = \dfrac{SSE}{SST}$

5. $t = \dfrac{\hat{\beta}_i - \beta_i^0}{se(\hat{\beta}_i)}$

6. one-tailed p-value: $\Pr(t_0 < t)$ or $\Pr(t < t_0)$

7. two-tailed p-value: $\Pr(|t_0| < t)$

8. $\Pr(\hat{\beta}_i - t_{N-K,\alpha/2} \cdot se(\hat{\beta}_i) < \beta_i < \hat{\beta} + t_{N-K,\alpha/2} \cdot se(\hat{\beta}_i)) = 1 - \alpha$

9. $F_{overall} = \dfrac{R^2/(K-1)}{(1-R^2)/(N-K)}$

10. $F = \dfrac{(SSR_R - SSR_U)/J}{SSR_U/(N-K)} = \dfrac{(R_U^2 - R_R^2)/J}{(1-R_U^2)/(N-K)}$

11. $\Pr(\hat{E}(Y \mid X = X_0) - term < E(Y \mid X = X_0) < \hat{E}(Y \mid X = X_0) + term) = 1 - \alpha$
    where $term = \hat{E}(Y \mid X = X_0) \cdot se(\hat{E}(Y \mid X = X_0))$

12. $\Pr((\hat{Y}_0 \mid X = X_0) - term < (Y_0 \mid X = X_0) < (\hat{Y}_0 \mid X = X_0) + term) = 1 - \alpha$
    where $term = (\hat{Y}_0 \mid X = X_0) \cdot se((\hat{Y}_0 \mid X = X_0)$.

13. When predicting, say, $(Y_0 \mid X = X_0)$ with some uncertainty as to the eventual value of X and that our uncertainty is expressed as E(X) = X_0 with $\Pr(X_l < X < X_u) = 1 - \alpha$, we take the **union** of the $(1-\alpha)$% confidence intervals that one would derive for $(Y \mid X = X_l)$ and $(Y \mid X = X_u)$. Our point prediction is still computed as $(\hat{Y}_0 \mid X = X_0)$.

14. Rule-of-Thumb (Adjusted) p-value that adjusts for Data Dredging: $\hat{\alpha}_{act} = \frac{c}{k}\alpha$, where c = number of candidate variables, k = number of final variables, α = the p-value the computer is giving us on the final regression. In terms of counting the number of variables, the intercept variable is not included.

15. $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$