Economics 102: Analysis of Economic Data  
Cameron Winter 2009  
Department of Economics, U.C.-Davis  

Final Exam (A) Tuesday March 17  
Compulsory. Closed book. Total of 60 points and worth 45% of course grade.  
Read question carefully so you answer the question.

| Question | 1a | 1b | 1c | 1d | 2a | 2b | 2c | 2d | 2e | 3a | 3b | 3c | 3d | 4a | 4b | 4c | 4d |
|----------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Points   | 1  | 1  | 2  | 4  | 2  | 2  | 4  | 2  | 2  | 2  | 2  | 4  | 2  | 2  | 2  | 4  |

<table>
<thead>
<tr>
<th>Question</th>
<th>5a</th>
<th>5b</th>
<th>5c</th>
<th>Multi Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
</tr>
</tbody>
</table>

Multiple Choice Questions (circle one part)

1. a b c d e
2. a b c d e
3. a b c d e
4. a b c d e
5. a b c d e

6. a b c d e
7. a b c d e
8. a b c d e
9. a b c d e
10. a b c d e

11. a b c d e
12. a b c d e
13. a b c d e
14. a b c d e
Questions 1-5
Consider data on family income and expenditure from the 1998 Consumer Expenditure Survey.

**Dependent Variable**
- Food = Annual family expenditure on food in 1998 in dollars.
- LnFood = Natural logarithm of Food.

**Regressors**
- Income = Annual family income in 1998 in dollars.
- LnIncome = Natural logarithm of Income.
- Famsize = number of members in family (averaged over year so can be non-integer)
- Drural = 1 if live in rural (non-urban) area and 0 if urban

Use the two pages of output provided at the end of this exam on:
1. Descriptive statistics output for all variables.
2. Histogram and scatterplot.
3. Various TINV values.
4. Three regressions.

Part of the following questions involves deciding which output to use. You can use the output that gets the correct answer in the quickest possible way.

1.(a) Which (if any) of income or food appears to be less symmetrically distributed, income or food? Explain your answer.

(b) On average, what percentage of family income is spent on food?

(c) Give a 95% confidence interval for the population mean family expenditure on food.

(d) Perform a test at significance level .05 of the claim that the population mean family expenditure on food is less than $5,000. State clearly the null and alternative hypotheses of your test, and your conclusion.
2. In this question the regression studied is a linear regression of food on income.

(a) According to the regression results, by how much does food expenditure increase in response to a one thousand dollar increase in income?

(b) Give a 99 percent confidence interval for the population slope parameter.

(c) Test the hypothesis at significance level 5% that the marginal propensity to consume food out of income is 0.06. State clearly the null and alternative hypothesis in terms of population parameters and your conclusion.

(d) Does heteroskedasticity appear to be a problem? Explain your answer.

(e) What is the sample correlation coefficient between family income and expenditure on food? Explain your answer.
3. In this question the regression studied is a linear regression of food on income.

(a) Predict the actual food expenditures of a family with income of $20,000.

(b) Using relevant output show that the sum of squares \( \sum_{i=1}^{n} (x_i - \bar{x})^2 = 7.7 \times 10^{11} \) for income.

(c) Give a 95 percent confidence interval for the conditional mean of food expenditures of a family with income of $20,000, using the result given in part (b) (even if your answer in (b) was different). Give answer as an expression involving numbers only, though you need not complete all the calculations.

(d) What five assumptions on the model are necessary in order for your answer in question 2 part (b) to be correct?
[Half point off for each assumption missing].
4. In this question consider both the regressions where Food is the dependent variable.

(a) Is living in a rural area an important determinant of food expenditures? Explain your answer.

(b) Are income, family size and living in an urban area jointly statistically significant at 5 percent? Perform an appropriate test. State clearly the null and alternative hypotheses of your test, and your conclusion.

(c) Using measures of goodness-of-fit, which model explains the data better - the multivariate regression or the bivariate regression? Explain your answer.

(d) Are family size and living in an urban area jointly statistically significant at 10 percent? Perform an appropriate test. State clearly the null and alternative hypotheses of your test, and your conclusion.

[Note: FINV(.10,1,696)=6.67; FINV(.10,2,696)=4.64; FINV(.10,3,696)=3.81.]
5. In this question consider the regression where LnFood is the dependent variable. For each of the following regressors provide a meaningful interpretation of the various slope coefficients in terms of impacts on food expenditures (rather than on LnFood).

(a) Provide a meaningful interpretation of the estimated coefficient for lnincome.

(b) Provide a meaningful interpretation of the estimated coefficient for famsize.

(c) Provide a meaningful interpretation of the estimated coefficient for drural.
Multiple choice questions (1 point each)

1. The Excel command to find the p-value for a two-tailed t-test in a regression with three regressors (including the intercept) when \( t = 1.57 \) and there are 40 observations is

   a. TINV(1.57,37,1)
   b. TINV(1.57,37,2)
   c. TDIST(1.57,37,1)
   d. TDIST(1.57,37,2)
   e. none of the above

2. Let \( \tilde{y}_i = b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} \). Then the ordinary least squares estimator minimizes

   a. \( \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \)
   b. \( \sum_{i=1}^{n}(y_i - \bar{y})^2 \)
   c. \( \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 \)
   d. none of the above

3. A good source for data on individuals is

   a. Michigan Panel Survey of Income Dynamics
   b. Federal Reserve Bank of St. Louis (FRED)
   c. both a. and b.
   d. neither a. nor b.

4. The standard error of the regression

   a. is normally distributed
   b. is the estimated standard deviation of the slope coefficient
   c. both a. and b.
   d. neither a. nor b.

5. The standard error of the slope coefficient

   a. increases with increase in the sample size
   b. decreases with increase in the variability of the regressors
   c. both a. and b.
   d. neither a. nor b.
6. The $T$-distribution arises in statistical inference because

a. underlying data are $T$-distributed

b. replacing unknown variance parameter $\sigma^2$ by an estimate converts a standard normal distribution to a $T$-distribution.

c. both of the above

d. neither of the above.

7. Let $d$ be an indicator variable for whether female. The regression model $y = \beta_1 + \beta_2 x + \beta_3 d + \beta_4 d \times x + \varepsilon$ is one with:

a. different intercept coefficient by gender

b. different slope coefficient by gender

c. both a. and b.

d. neither a. nor b.

8. In linear OLS regression a major problem arises if

a. important regressors are omitted

b. unnecessary (or irrelevant) regressors are included

c. neither a. nor b.

d. both a. and b.

9. Multicollinearity is a problem that arises when

a. the dependent variable is highly correlated with the regressors

b. the regressors are highly correlated with each other

c. the error is highly correlated with the dependent variable

d. the error is highly correlated with the regressors

10. Suppose we perform an OLS regression of a variable $y$ on variable $x$. The Stata command that will give the same output as Excel is

a. regress $y$ $x$

b. ols $y$ $x$

c. regress $y$ $x$, vce(robust)

d. ols $y$ $x$, vce(robust)
11. Suppose in regression of food on an intercept and income we wish to also control for region. The U.S. can be split into four regions – the north-east, south, central or west – and we compute four separate indicator variables for whether or not the family lives in each of these regions. We should

a. include just one of the four indicator variables in the regression
b. include two of the four indicator variables in the regression
c. include three of the four indicator variables in the regression
d. include all four indicator variables in the regression

12. According to the article by Kreuger, without any control for other determinants, those working with a computer on average have an hourly wage that is

a. more than twenty percent higher than those who do not work with a computer
b. between ten and twenty percent higher than those who do not work with a computer
c. between zero and ten percent higher than those who do not work with a computer
d. less than those who do not.

13. Using the more exact computational method of Kreuger, if \( \ln w = 5 + 0.1c \) where \( c \) is the computer dummy variable then the wage rate for those using a computer is higher than for those who do not use a computer by an amount equal to

a. \( \exp(0.1) \) times the wage for those who do not
b. \( (\exp(0.1) - 1) \) times the wage for those who do not
c. neither a. nor b.

14. Correlated errors are likely to arise with

a. cross section data
b. time series data
c. neither a. nor b.
d. both a. and b.
Univariate Data

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

\[ \bar{x} \pm t_{\alpha/2,n-1} \times \left( \frac{s}{\sqrt{n}} \right) \]

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \]

TDIST(\(t, df, 1\)) = Pr[\(T > t\)] \quad \text{and} \quad TDIST(\(t, df, 2\)) = Pr[|T| > t]

\(t_{\alpha/2}\) such that Pr[|T| > \(t_{\alpha/2}\)] = \(\alpha\) is calculated using TINV(\(\alpha, df\)).

Bivariate Data

\[ r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \times \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}} \]

\[ \hat{y} = b_1 + b_2 x \quad b_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad b_1 = \bar{y} - b_2 \bar{x} \]

\[ \text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{ErrorSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{RegSS} = \text{TSS} - \text{ErrorSS} \]

\[ R^2 = 1 - \frac{\text{ErrorSS}}{\text{TSS}} \]

\[ b_2 \pm t_{\alpha/2,n-2} \times s_{b_2} \]

\[ t = \frac{b_2 - \beta_20}{s_{b_2}} \quad s_{b_2}^2 = \frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\[ y|x = x^* \in b_1 + b_2 x^* \pm t_{\alpha/2,n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}} + 1 \]

\[ E[y|x = x^*] \in b_1 + b_2 x^* \pm t_{\alpha/2,n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \]

Multivariate Data

\[ \hat{y} = b_1 + b_2 x_2 + \cdots + b_k x_k \]

\[ R^2 = 1 - \frac{\text{ErrorSS}}{\text{TSS}} \quad \bar{R}^2 = R^2 - \frac{k-1}{n-k} (1 - R^2) \]

\[ b_j \pm t_{\alpha/2,n-k} \times s_{b_j} \]

\[ t = \frac{b_j - \beta_j0}{s_{b_j}} \]

\[ F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad \text{and} \quad F = \frac{(SSE_e - SSE_u)/(k-g)}{SSE_u/(n-k)} \]
<table>
<thead>
<tr>
<th></th>
<th>food</th>
<th>food info</th>
<th>income</th>
<th>income info</th>
<th>famsize</th>
<th>drural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4897.343</td>
<td>8.339223</td>
<td>41118.78204</td>
<td>10.27804</td>
<td>2.469429</td>
<td>0.1285714</td>
</tr>
<tr>
<td>Standard Error</td>
<td>107.933</td>
<td>0.021571</td>
<td>1250.928445</td>
<td>0.034932</td>
<td>0.053454</td>
<td>0.0126605</td>
</tr>
<tr>
<td>Median</td>
<td>4322.5</td>
<td>8.371585</td>
<td>32150</td>
<td>10.37817</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Mode</td>
<td>5265</td>
<td>8.568836</td>
<td>40000</td>
<td>10.59663</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2855.653</td>
<td>0.570706</td>
<td>33096.45573</td>
<td>0.924208</td>
<td>1.414257</td>
<td>0.3349643</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>8154753</td>
<td>0.325705</td>
<td>1095375382</td>
<td>0.854161</td>
<td>2.000123</td>
<td>0.1122011</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.949653</td>
<td>-0.06721</td>
<td>2.703104289</td>
<td>2.658088</td>
<td>0.889339</td>
<td>2.9549425</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.560394</td>
<td>-0.16728</td>
<td>1.513918361</td>
<td>-0.94241</td>
<td>1.026037</td>
<td>2.2240746</td>
</tr>
<tr>
<td>Range</td>
<td>20180</td>
<td>3.569412</td>
<td>191973.1</td>
<td>7.156589</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Minimum</td>
<td>585</td>
<td>6.371612</td>
<td>150</td>
<td>5.010635</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>20765</td>
<td>9.941024</td>
<td>192123.1</td>
<td>12.16589</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Sum</td>
<td>3428140</td>
<td>5837.456</td>
<td>28783147.43</td>
<td>7194.628</td>
<td>1728.6</td>
<td>90</td>
</tr>
<tr>
<td>Count</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
</tr>
<tr>
<td>Largest(175)</td>
<td>6136</td>
<td>8.721929</td>
<td>56900</td>
<td>10.94905</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Smallest(175)</td>
<td>2821</td>
<td>7.944847</td>
<td>16533</td>
<td>9.713114</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Confidence Level(95.0%)</td>
<td>211.9127</td>
<td>0.042351</td>
<td>2456.026114</td>
<td>0.068584</td>
<td>0.104949</td>
<td>0.0248571</td>
</tr>
</tbody>
</table>

Figure 1  Descriptive statistics, charts and TINV values
## Food is dependent variable

**Regression Statistics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.537478769</td>
</tr>
<tr>
<td>R Square</td>
<td>0.28883427</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.28764636</td>
</tr>
<tr>
<td>Standard Error</td>
<td>2409.831485</td>
</tr>
<tr>
<td>Observations</td>
<td>700</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1646665262</td>
<td>1.65E+09</td>
<td>283.555</td>
<td>1.19E-53</td>
</tr>
<tr>
<td>Residual</td>
<td>698</td>
<td>4053486873</td>
<td>5807288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>699</td>
<td>5700172136</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2990.453628</td>
</tr>
<tr>
<td>income</td>
<td>0.046375139</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1439.750336</td>
</tr>
<tr>
<td>income</td>
<td>0.035120583</td>
</tr>
<tr>
<td>famsize</td>
<td>816.851901</td>
</tr>
<tr>
<td>drural</td>
<td>-28.62611305</td>
</tr>
</tbody>
</table>

## Lnfood is dependent variable

**Regression Statistics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.672888443</td>
</tr>
<tr>
<td>R Square</td>
<td>0.452778856</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.450420144</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.423085235</td>
</tr>
<tr>
<td>Observations</td>
<td>700</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>103.0832845</td>
<td>34.36109</td>
<td>191.9602</td>
<td>1.08E-90</td>
</tr>
<tr>
<td>Residual</td>
<td>696</td>
<td>124.584777</td>
<td>0.179001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>699</td>
<td>227.6680615</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.381465659</td>
</tr>
<tr>
<td>lnincome</td>
<td>0.246369861</td>
</tr>
<tr>
<td>famsize</td>
<td>0.172511582</td>
</tr>
<tr>
<td>drural</td>
<td>-0.003479947</td>
</tr>
</tbody>
</table>

**Figure 2** Regression output