**Economics 102: Analysis of Economic Data**
**Cameron Spring 2016**
**Department of Economics, U.C.-Davis**

**Final Exam (A) Tuesday June 7**

Compulsory. Closed book. Total of 58 points and worth 45% of course grade.
Read question carefully so you answer the question.

**Question scores**

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

| Question | 5a | 5b | 5c | 5d | 6a | 6b | 6c | 6d | 7a | 7b | 7c | 7d | 7e | Mult Choice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Points | 1 | 3 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 10 |

**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**

**Questions 1-4**
Consider data on sales and advertising for 200 regional markets
**Note: pay attention to the units of measurement.**

**Dependent Variable**

`sales` = sales in units
`lnsales` = Natural logarithm of `sales`.

**Regressors**

`tv` = TV advertising in thousands of dollars
`radio` = radio advertising in thousands of dollars
`newspaper` = newspaper advertising in thousands of dollars
`tvbynews` = `tv` $\times$ `news`
`region1` = 1 if region 1 and = 0 otherwise
`region2` = 1 if region 2 and = 0 otherwise
`region3` = 1 if region 3 and = 0 otherwise

**Use the two pages of output provided at the end of this exam on:**
$t$ critical values, summary statistics, correlations and 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)** Give a 95% confidence interval for population mean sales.

**(b)** Perform a test at significance level .05 of the claim that population mean sales exceed 13,000 units. State clearly the null and alternative hypotheses of your test, and your conclusion.

**(c)** Suppose we give Stata command `summarize sales, detail`
Provide three different statistics that this provides in addition to command `summarize sales` (Two points for three correct; 1 point for 2 correct; 0 points for 1 or 0 correct).

**2.** In this question the regression studied is a linear regression of `sales` on `tv`.

**(a)** Give a 95 percent confidence interval for the population slope parameter.

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

**(c)** Test the hypothesis at significance level 1% that the population slope coefficient is equal to 50. **State clearly** the null and alternative hypothesis in terms of population parameters and state your conclusion.

**(d)** Predict the actual sales when `tv` advertising equals $100,000.
(Hint: Be careful with units here).

**(e)** A statistician states that a 95 percent confidence interval for actual sales given `tv` advertising equals $100,000 will have width of at least 10,000 units. Is she correct? **Explain your answer.**

**(f)** Suppose we regress `tv` on `sales` rather than `sales` on `tv`? How well will the model fit? **Explain**.

**3.** This question and the next consider all three models given in the second page of Stata output. Pay attention to the units of measurement used in defining the variables.
Note that there are only three regions: region 1, region 2 and region 3.

**(a)** In the second model what is the effect on sales of increasing TV advertising by $1,000. Evaluate this at the sample mean value of relevant variables.

**(b)** In the second model provide an interpretation of the coefficient of variable `region1`.

**(c)** Are the additional regressors in the second model, compared to the first model, jointly statistically significant at significance level 0.05? Perform an appropriate test. **State clearly** the null and alternative hypotheses of your test, and your conclusion given that the critical value for the test statistic is 2.261.

**(d)** Suppose in the second model we replaced regressors `region1` and `region2` with `region2` and `region3`. How would the output differ from that of the second model? **Explain.**

**4.(a)** In the third model what is the effect on the number of units sold of $1,000 more spending on TV advertising?

**(b)** Suppose we estimate the third model and then predict sales using the Stata commands

```
predict lnsaleshat
gen saleshat = exp(lnsaleshat)
```

Will this provide a good prediction of the level of sales? Explain your answer.

**(c)** Given the output provided is it possible to prefer the third model to the first model? **Explain your answer**.

**5.(a)** Calculate $\sum_{i=1}^{n} z_i$ for $z_i = 6/i$ and $n = 3$.

**(b)** Suppose $Y_i$ is distributed with mean 10 and variance 100, though is not necessarily normally distributed. We obtain $10,000$ samples each of size 100 and for each sample compute the sample mean $\bar{y}$. What distribution do you expect the sample means to have? Provide the mean, standard deviation and, if appropriate, the distribution.

**(c)** Let $X$ be the number of students who miss a midterm exam due to illness. Suppose $X = 1$ with probability 0.5, $X = 2$ with probability 0.3 and $X = 3$ with probability 0.2. What is the mean of $X$? **Show all workings.**

**(d)** Consider a simple random sample of size 4 with values $18, 20, 28, 30$.
Compute the sample standard deviation. **Show all workings.**

**6.** You are given the following partial Stata output

```
. regress y x z
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,     18) =     (C)
       Model |      540                                Prob > F      =
    Residual |                                         R-squared     =     (B)
-------------+------------------------------           Adj R-squared =
       Total |      720                                Root MSE      =
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          3       (A)       1.5
           z |          2       1.0
       _cons |         -4       1.0
------------------------------------------------------------------------------
```

**(a)** Calculate missing entry **(A)**.

**(b)** Calculate missing entry **(B)**.

**(c)** Calculate missing entry **(C)**.

**(d)** Suppose we perform an F test of $H_0 : \beta_z = 0$ against $H_a : \beta_z \neq 0$. What will the value of the F statistic be?

**7.** For each of the following conditions state whether or not OLS estimates of $\beta_1$, $\beta_2$ and $\beta_3$ in the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + u_i$ are likely to be biased.

**(a)** The sample comprises six observations.

**(b)** We should not have included variable $z$ in the model.

**(c)** We should have included variable $w$ in the model.

**(d)** The correlation of $x$ and $z$ equals 0.98.

**(e)** The error $u$ is heteroskedastic.

**Multiple choice questions (1 point each)**

**1.** A pie chart is best used for summarizing

**a.** categorical data

**b.** continuous data

**c.** both a. and b.

**d.** neither a. nor b.

**2.** The skewness statistic is approximately

**a.** $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$ where $s$ is the sample standard deviation

**b.** $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$

**c.** $s^3$ where $s$ is the sample standard deviation

**d.** none of the above.

**3.** For monthly data a 11 month moving average

**a.** reduces variation in the original data

**b.** can help control for seasonal variation in the data

**c.** neither a. not b.

**d.** both a. and b.

**4.** A correlation coefficient equal to 1.1

**a.** indicates strong association between $x$ and $y$ that may be positive or negative

**b.** indicates strong positive association between $x$ and $y$

**c.** indicates strong negative association between $x$ and $y$

**d.** is not possible.

**5.** If $X_i$ are independent and identically distributed as $N(\mu, \sigma^2)$ then

**a.** $(\bar{X}-\mu)/\sigma$ is $T(n-1)$ distributed

**b.** $(\bar{X}-\mu)/\sigma$ is standard normal distributed

**c.** neither a. nor b.

**6.** For a hypothesis test with size 0.05

**a.** the probability of not rejecting $H_0$ when $H_0$ is false is 0.05

**b.** the probability of not rejecting $H_0$ when $H_0$ is false is 0.95

**c.** the probability of rejecting $H_0$ when $H_0$ is true is 0.05

**d.** the probability of rejecting $H_0$ when $H_0$ is true is 0.95

**7.** Let $b$ be the slope coefficient from OLS regression of $y$ on an intercept and $x$ and let $c$ be the slope coefficient from regression of $x$ on an intercept and $y$

**a.** if $b > c$ than necessarily $x$ causes $y$

**b.** if $c > b$ than necessarily $y$ causes $x$

**c.** neither of the above.

**8.** In the linear regression model the conditional mean of $y$ given $x$ is

**a.** $\beta_1 + \beta_2 x + u$

**b.** $\beta_1 + \beta_2 x$

**c.** $b_1 + b_2 x + e$ where $b_1$ and $b_2$ are estimated coefficients and $e$ is the residual

**d.** $b_1 + b_2 x$ where $b_1$ and $b_2$ are estimated coefficients and $e$ is the residual.

**9.** The main lesson from regression analysis of school scores on the California Academic performance Index is that

**a.** by far the biggest determinant is teacher quality

**b.** by far the biggest determinant is educational attainment of parents

**c.** by far the biggest determinant is student disadvantage (English learner, free meals)

**d.** all of a., b. and c. are substantial determinants.

**10.** Let $Q$, $K$ and $L$ denote the level of output, capital and labor. A Cobb-Douglas production is estimated by regressing

**a.** $\ln Q$ on $\ln K$ and $\ln L$

**b.** $Q$ on $K$ and $L$

**c.** $\ln Q$ on $K$ and $L$

**d.** $Q$ on $\ln K$ and $\ln L$

**SOME USEFUL FORMULAS**

## Univariate Data

$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ *and* $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$

$\bar{x} \pm t_{\alpha/2;n-1} \times (s_x/\sqrt{n})$ *and* $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$\text{ttail}(df, t) = \Pr[T > t]$ where $T \sim t(df)$

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

## Bivariate Data

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

$\widehat{y} = b_1 + b_2 x_i$ $\quad b_2 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ $\quad b_1 = \bar{y} - b_2\bar{x}$

$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y}_i)^2$ $\quad \text{ResidualSS} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$ $\quad$ Explained SS = TSS - Residual SS

$R^2 = 1-$ ResidualSS/TSS

$b_2 \pm t_{\alpha/2;n-2} \times s_{b_2}$

$t = \dfrac{b_2 - \beta_{20}}{s_{b_2}}$ $\quad s_{b_2}^2 = \dfrac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ $\quad s_e^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \widehat{y}_i)^2$

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

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

## Multiple Regression

$\widehat{y} = b_1 + b_2 x_{2i} + \cdots + b_k x_{ki}$

$R^2 = 1-$ ResidualSS/TSS $\quad \bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$

$b_j \pm t_{\alpha/2;n-k} \times s_{bj}$ *and* $t = \dfrac{b_j - \beta_{j0}}{s_{b_j}}$

$$F = \frac{R^2/(k-1)}{(1 - R^2)/(n-k)} \sim F(k-1, n-k)$$

$$F = \frac{(\text{ResSS}_r - \text{ResSS}_u)/(k-g)}{\text{ResSS}_u/(n-k)} \sim F(k-g, n-k)$$

$\text{Ftail}(df1, df2, f) = \Pr[F > f]$ where $F$ is $F(df1, df2)$ distributed.

$F_\alpha$ such that $\Pr[F > f_\alpha] = \alpha$ is calculated using $\text{invFtail}(df1, df2, \alpha)$.

```
Degrees of freedom:  200     199     198     197     196     195     194     193
t_.05:             1.6525  1.6525  1.6526  1.6526  1.6527  1.6527  1.6527  1.6528
t_.025:            1.9719  1.9720  1.9720  1.9721  1.9721  1.9722  1.9723  1.9723
t_.01:             2.3451  2.3452  2.3453  2.3454  2.3455  2.3456  2.3457  2.3458
t_.005:            2.6006  2.6008  2.6009  2.6010  2.6011  2.6013  2.6014  2.6015


.
. summarize sales tv radio newspaper tvbynews region1 region2 lnsales

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       sales |       200     14022.5    5217.457       1600      27000
          tv |       200    147.0425    85.85424         .7      296.4
       radio |       200      23.264    14.84681          0       49.6
   newspaper |       200      30.554    21.77862         .3        114
    tvbynews |       200    4598.126    4870.717       6.09   29906.76
-------------+--------------------------------------------------------
     region1 |       200         .23    .4218886          0          1
     region2 |       200        .445    .4982129          0          1
     lnsales |       200    9.471746    .4143507   7.377759   10.20359

. correlate sales tv radio newspaper tvbynews region1 region2 lnsales
(obs=200)

             |    sales       tv    radio newspa~r tvbynews  region1  region2  lnsales
-------------+------------------------------------------------------------------------
       sales |   1.0000
          tv |   0.7822   1.0000
       radio |   0.5762   0.0548   1.0000
   newspaper |   0.2283   0.0566   0.3541   1.0000
    tvbynews |   0.6185   0.6031   0.2502   0.7109   1.0000
     region1 |  -0.3245  -0.2269  -0.2673  -0.1584  -0.2137   1.0000
     region2 |  -0.0522  -0.0487   0.0037   0.0339  -0.0271  -0.4894   1.0000
     lnsales |   0.9541   0.7846   0.4712   0.2114   0.5710  -0.3323  -0.0286   1.0000
```

```
. regress sales tv

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =    312.14
       Model |  3.3146e+09         1  3.3146e+09   Prob > F        =    0.0000
    Residual |  2.1025e+09       198  10618841.6   R-squared       =    0.6119
-------------+----------------------------------   Adj R-squared   =    0.6099
       Total |  5.4171e+09       199    27221853   Root MSE        =    3258.7

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |   47.53664   2.690607    17.67   0.000     42.23072    52.84256
       _cons |   7032.594   457.8429    15.36   0.000     6129.719    7935.468
------------------------------------------------------------------------------

. regress sales tv radio newspaper tvbynews region1 region2

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(6, 193)       =    304.00
       Model |  4.8988e+09         6   816466409   Prob > F        =    0.0000
    Residual |   518350292       193  2685752.81   R-squared       =    0.9043
-------------+----------------------------------   Adj R-squared   =    0.9013
       Total |  5.4171e+09       199    27221853   Root MSE        =    1638.8

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |   38.80747    2.31232    16.78   0.000     34.24681    43.36813
       radio |   187.3695   8.701045    21.53   0.000     170.2081    204.5308
   newspaper |  -32.16059   10.46367    -3.07   0.002    -52.79842   -11.52276
    tvbynews |   .2010003   .0568861     3.53   0.001      .088802    .3131985
     region1 |   -404.474   346.3489    -1.17   0.244    -1087.589    278.6409
     region2 |  -308.8007   275.7715    -1.12   0.264    -852.7135    235.1121
       _cons |   4246.044   493.7597     8.60   0.000     3272.187    5219.902
------------------------------------------------------------------------------

. regress lnsales tv

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =    317.09
       Model |   21.032308         1   21.032308   Prob > F        =    0.0000
    Residual |  13.1333104       198  .066329851   R-squared       =    0.6156
-------------+----------------------------------   Adj R-squared   =    0.6137
       Total |  34.1656184       199  .171686525   Root MSE        =    .25755

------------------------------------------------------------------------------
     lnsales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |   .0037867   .0002127    17.81   0.000     .0033673     .004206
       _cons |   8.914947   .0361853   246.37   0.000     8.843589    8.986306
------------------------------------------------------------------------------
```