Final Exam - Answers

Tomas Dvorak

Econ 253 Spring 2002

You have 120 minutes to answer the following 5 questions. The maximum number of points on this exam is 120. The number of points that a question is worth is the number of minutes that you should spend on that question. You may use a calculator and a double-sided sheet of paper with notes and formulas. Any collaboration is considered a violation of the honor code. Good luck!

1. The investigation of consumer product complaints by the Federal Trade Commission has generated much interest by manufacturers in the quality of their product. A manufacturer of an electromechanical kitchen aid conducted an analysis of a large number of consumer complaints and found that they fell into the six categories shown below.

	reason for complaint		
	electrical	mechanical	appearance
during guarantee	18%	15%	30%
after guarantee	12%	22%	3%

Answer the following questions:

(a) What is the probability that the complaint is either for electrical or mechanical reasons.

Answer: P(E)=0.3, P(M)=0.37. The two events are mutually exclusive, hence P(E+M)=0.67.

(b) What is the probability that a complaint which arrived after the guarantee period is due to appearance?

Answer: $P(A|after) = \frac{P(A \cdot after)}{P(after)} = \frac{0.03}{0.37} = 0.08$

2. A nutritionist made a new sports drink and wants to know if it will improve running times in 10K races. He knows that among men, running times are distributed normally, with a mean of 40 minutes. He picks 20 men at random and has them run a 10K race after consuming the drink. The 20 men take 38 minutes on average with a standard deviation 5 minutes. Would you consider this decisive evidence that the new drink is beneficial?

Answer: This is the test if population mean is equal to some number.

$H_0 : \mu = 40$

$H_1 : \mu < 40$

We know that $t = \frac{\overline{x}-\mu}{s/\sqrt{n}} = \frac{-2}{1.12} = -1.79$ is distributed t with $n-1$ degrees of freedom. The critical region for this test is on the left side, $(-\infty, -1.729)$. Our test statistic falls into the critical region hence we reject the null. Thus, the evidence suggest that the new drink is beneficial.

3. In the second midterm you considered a model for explaining the variation in women's labor force participation rate across U.S. states. Consider the following model extension to the model. In addition to the data from 1990 census we collected the data from 1980 census.

   - WLFP = percent in labor force who are female
   - YF = median earnings ($) by females
   - YM = median earnings ($) by males
   - MR = percent of females now married
   - D90 = 1 for 1990 Census and 0 for 1980 Census

```
regress wlfp  yf d90_yf ym d90_ym mr d90_mr

  Source |       SS       df       MS              Number of obs =     100
---------+------------------------------           F(  6,    93) =   36.82
   Model |  2122.22777     6  353.704628           Prob > F      =  0.0000
Residual |  893.406238    93  9.60651869           R-squared=0.7037
---------+------------------------------           Adj R-squared =  0.6846
   Total |  3015.63401    99  30.4609496           Root MSE      =  3.0994

------------------------------------------------------------------------------
    wlfp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      yf |   .0054334   .0012624      4.304   0.000        .0029265    .0079403
  d90_yf |  -.0031733   .0013232     -2.398   0.018       -.0058008   -.0005457
      ym |    .000207   .0004766      0.434   0.665       -.0007394    .0011535
  d90_ym |  -.0011678   .0005957     -1.961   0.053       -.0023507     .000015
      mr |   .2865163   .1136595      2.521   0.013        .0608111    .5122215
  d90_mr |   .2372676   .0713678      3.325   0.001        .0955453      .37899
   _cons |   14.15023    7.30004      1.938   0.056       -.3462061    28.64666
------------------------------------------------------------------------------
```

where `d0_yf` denotes interaction of the 1990 census dummy and female wages

   (a) Holding other variables constant, what was the effect of women's earnings on women's labor force participation in 1990? What was the effect in 1980?

   Answer: In 1990 $\Delta E(WLFP) = (B_2 + B_3) \cdot \Delta YF = (0.0054 - 0.0031) \cdot \Delta YF$. Holding other variables constant, in 1990 an increase in female earnings by a dollar increases the labor force participation rate by 0.0023percentage points. In 1980 $\Delta E(WLFP) = B_2 \cdot \Delta YF = 0.0054 \cdot \Delta YF$. Thus, in 1980, a dollar increase in female earnings increases the labor force participation by 0.0054 percentage points. This could be that a dollar in 1980 is worth a lot more than a dollar in 1990.

(b) Has the effect of marriage rate on women labor force participation changed between 1980 and 1990?

Answer: Yes, the effect of a marriage rate on female labor force participation increased. In 1980 a percentage point increase in marriage rate resulted in a 0.28 percentage point increase in women labor force participation. In 1990 a percentage point increase in marriage rate resulted in 0.28+0.23=0.51 percentage point increase in women labor force participation rate.

(c) How would you test whether the relationship between marriage rate and female labor force participation changed between 1980 and 1990?

Answer:
$E(WLFP|year = 1980) = ... + B_7 \cdot MR...$
$E(WLFP|year = 1980) = ...(B_7 + B_8) \cdot MR...$
If the null hypothesis is that the relationship is the same:
$H_0 : B_7 = B_7 + B_8 \Leftrightarrow B_8 = 0$
$H_1 : B_8 \neq 0$
Hence, to test whether the relationship is different we simply need to test if B_8 is equal to zero.

(d) Is the relationship between male earning and women labor force participation rate statistically different in 1980 and 1990?

Answer: Following the test outlined above, we see that the coefficient on the interaction is statistically insignificant at 5% (barely). Hence, we accept the null hypothesis that the relationship is the same in 1990 and 1980.

4. Suppose you were asked to explain variations in the telephone bills of households. Write down a regression model to reflect the ideas below. Start from a simple model which reflects idea (a), next incorporate idea (b) and so on. Explain what sign you might expect for each coefficient.

(a) bills rise as the number of teenagers in the household increases

Answer: $BILL_i = B_1 + B_2 \cdot TEEN_i + u_i$ where $BILL_i$ is the telephone bill in dollars, $TEEN_i$ is the number of teenagers in the household. I expect $B_2 > 0$.

(b) expenditure on phone calls increase at an increasing rate as household income increases

Answer: $BILL_i = B_1 + B_2 \cdot TEEN_i + B_3 \cdot INC_i + B_4 \cdot INC_i^2 + u_i$ where INC_i is the household income, $\frac{\Delta BILL}{\Delta INC} = B_3 + 2 \cdot B_4 \cdot INC_i$ If bills are are to increase at an increasing rate, it has to be that $B_3 > 0$ and $B_4 > 0$.

(c) bills are higher if adults have parents in different states

Answer: $BILL_i = B_1 + B_2 \cdot TEEN_i + B_3 \cdot INC_i + B_4 \cdot INC_i^2 + B_5 \cdot OOS_i + u_i$ where OOS_i is the number of adults with parents living out of state. I would expect $B_5 > 0$.

(d) adult sons call their parents less than adult daughters

Answer:$BILL_i = B_1 + B_2 \cdot TEEN_i + B_3 \cdot INC_i + B_4 \cdot INC_i^2 + B_5 \cdot OOS_i + B_6 \cdot SOOS_i + u_i$ where $SOOS_i$ is the number of male adults with parents out of state. I would expect $B_6 < 0$.

5. Using 15 years of annual data, planners in San Diego County estimated the following model for water consumption (t-statistics in parentheses):

$$\begin{aligned} water &= -327 + 0.31 House + 0.36 Pop + 0.05 LInc - 17.87 Pwater - 1.12 Rain \\ & \quad (-1.17)(0.9) \qquad (1.4) \qquad (0.6) \qquad (-1.2) \qquad (-0.8) \end{aligned}$$

$R^2 = 0.93$

- water = Total water consumption (million cubic meters)
- House = Total number of housing units (thousands)
- Pop = Total population (thousands)
- IncPC = Logarithm of income per capita (dollars)
- Pwater = Price of water (dollars/100cubic meters)
- Rain = Rainfall in inches

Answer the following questions:

(a) Based on economic theory and/or intuition, what signs would you expect for the regression coefficients (excluding the constant) and why? Do the observed signs agree with your intuition?

Answer: We would expect the signs on house and population to be positive, since both are proxies for population size, and increased population would lead to increased water consumption. We would also expect income to have a positive sign (since water is likely to be a normal good). The price of water would be expected to have a negative sign, indicating that higher water prices would lead to lower consumption. Rainfall would also have be expected to have a negative effect, since rainfall reduces dependence on irrigation. All of the signs except the sign on income agree with the above intuition.

(b) What are the interpretations of coefficients on log of per capita income and rain?

Answer:
$\frac{\Delta water}{\Delta Inc/Inc} = 0.05$. This means that a percent increase in per capita income leads to 0.0005 million of cubic feet increase in water consumption.
$\frac{\Delta water}{\Delta Rain} = -1.12$. This means that an inch increase in rainfall decreases water consumption by 1.12 million of cubic feet.

(c) According to the t-statistics every coefficient is insignificant, but the R^2 is very high. What might be the reasons for this paradoxical result? Are the estimates biased? Inefficient? Explain why or why not.

Answer: Insignificant t-statistics but a high R 2 is a sign that multicollinearity might be present. Another possibility is that the equation is misspecified. If the problem is multicollinearity (most likely the reason), the estimates are not biased, and they are not inefficient. None of the assumptions of the classical linear regression model have been violated. The standard errors are correct, although they are large. If the problem is misspecification, the estimates may be biased or inefficient.

(d) What solutions might you suggest for the planners to try and get a better idea about the true underlying values of the variables?

Answer: One possible source of the multicollinearity is the inclusion of both total population and total number of housing units, since both are proxies for the population density, and thus are likely to be highly correlated. If removing one of those two variables does not help, the only solution is likely to be to obtain more data.

6. Suppose that the data generating process for wages of an individual worker is

$$W_i = B_1 + B_2 AGE_i + u_i \tag{1}$$

where $var(u_i) = \sigma^2$ and i indexes individual workers. Suppose now that your data source has information on 100 firms. For each firm j you have data on number of employees N_j, average wage $\overline{W_j}$, and average age of employees $\overline{A_j}$. Note that you do not have data on individual workers W_is and A_is. Answer the following questions:

(a) Can you find unbiased estimates coefficients B_1 and B_2 using the data on average wages and average age of workers?

Answer:

If equation (1) holds for each worker in each firm. I can sum the right and left sides for across workers for each firm and divide by N_j to obtain:

$$\begin{aligned} \frac{1}{N_j}\sum_{i=1}^{N_j} W_i &= \frac{1}{N_j}\sum_{i=1}^{N_j} B_1 + B_2 \frac{1}{N_j}\sum_{i=1}^{N_j} AGE_i + \frac{1}{N_j}\sum_{i=1}^{N_j} u_i \\ \overline{W}_j &= B_1 + B_2 \overline{A}_j + \overline{u}_j \end{aligned}$$

Thus, we found the data generating process for average wage and average age. We see that the coefficients relating average wage and age are the same as those relating individual wages and ages. Thus, the second equation can be estimated using data on averages and the estimated coefficient will be unbiased estimates of B_1 and B_2.

(b) Is the estimation you proposed in part a. efficient? If not can you suggest a different method to obtain efficient estimates of B_1 and B_2?

Answer: No, the estimation is inefficient because $var(\frac{1}{N_j}\sum_{i=1}^{N_j} u_i) = var(\overline{u}_j) = \frac{\sigma^2}{N_j}$. Hence, the variance of the error term $\overline{u}_j$ varies with observations. We have a case of heteroskedasticity. Fortunately, there is a simple remedy. The appropriate transformation of the error term is $\overline{u}_j^* = \overline{u}_j \cdot \sqrt{N_j}$ because $var(\overline{u}_j^*) = var(\overline{u}_j \cdot \sqrt{N_j}) = \sigma^2$

$\overline{W}_j^* = \overline{W}_j \cdot \sqrt{N_j}$
$\overline{A}_j^* = \overline{A}_j \cdot \sqrt{N_j}$
Running OLS on $\overline{W}_j^* = B_1 + B_2 \overline{A}_j^* + \overline{u}_j^*$ will give us efficient estimates of B_1 and B_2. This transformation makes also intuitive sense: observations from large firms should be weighted more heavily than observations from small firms.