Practice exam questions

Nathaniel Higgins
nhiggins@jhu.edu, nhiggins@ers.usda.gov

1.

The following question is based on the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

(a)

Discuss the following two hypotheses. Give an example of a situation in which each hypothesis might be useful.

$$H_{0,1} : \beta_1 = 1, \beta_2 = 1, \beta_3 = 1$$
$$H_{0,2} : \beta_1 + \beta_2 + \beta_3 = 0$$

(b)

What test statistic would you use to complete each hypothesis test? ($H_{0,1}$? $H_{0,2}$?)

2.

The results of a regression are given below. Test the null hypothesis that $\beta_2 = 1$. Can you reject the null at standard significance levels?

```
. reg y x2 x3
```

Source	SS	df	MS
Model	23070.7615	2	11535.3807
Residual	13050.1337	497	26.2578142
Total	36120.8951	499	72.3865634

Number of obs	500
F(2, 497)	439.31
Prob > F	0.0000
R-squared	0.6387
Adj R-squared	0.6373
Root MSE	5.1242

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.9452643	.0466014	20.28	0.000	.8537043	1.036824
x3	.9413209	.0459289	20.50	0.000	.8510821	1.03156
_cons	.9578987	.2292015	4.18	0.000	.5075753	1.408222

3.

How would you test the null hypothesis that $\beta_1 = \beta_2$? Can you do it based on the above output? Regardless of your answer to the previous question, provide evidence that you could use to provide intuition for whether β_1 and β_2 appear statistically similar. Can you provide any information on the conditions under which the null hypothesis would be accepted (rejected) at conventional levels? Assume a two-sided alternative hypothesis.

4.

Suppose you run the regression

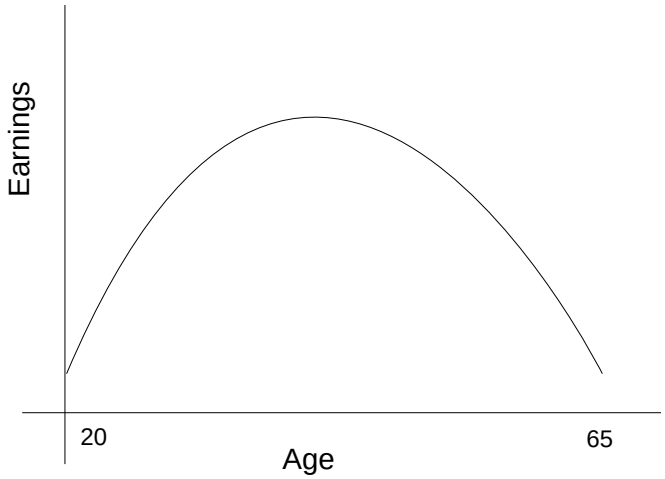
$$y = \beta_0 + \beta_1 x_1 + u$$

and obtain estimated coefficients $\widehat{\beta_0}$ and $\widehat{\beta_1}$. You then create a new variable $y*$, where $y* = y - \bar{y}$. What is the relationship between $\widehat{\beta_0}$ and $\widehat{\beta_1}$ and $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$, where

$$y* = \alpha_0 + \alpha_1 x_1 + v?$$

5.

Suppose the age-earnings profile looks something like what is displayed in the figure below.



(a)

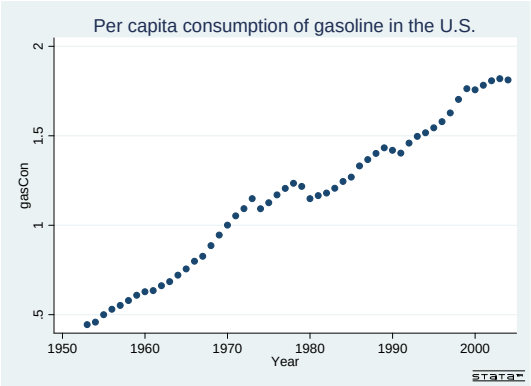
What model specification would you use to fit the regression of `earnings` on `age`? (Do not include other control variables)

(b)

What is the marginal effect of an increase in age? Is it constant over `age`?

6.

The per capita consumption of gasoline in the U.S. between 1953 and 2004 is plotted below.



Using data from 1953 to 2004, I regressed per capita consumption of gasoline on the following variables:

Table 1: independent variables

income
price of new cars (index)
price of used cars (index)
price of public transportation (index)
price of consumer durables (index)
price of consumer nondurables (index)
price of consumer services (index)
year

Index numbers are nothing fancy — indexes are simply composite measures meant to represent an aggregate. For example, there is no *single* "price of new cars." Of course, every new car has its own price. An "index" of new car prices is meant to capture the price of all new cars as a whole. It's a general number that is helpful in representing trends. For instance, although the price of some new cars may have gone down from 2005 to 2006, if *most* new cars increased in price, the index of the price of new cars would increase.

I include `year` in the regression to control for the general upward trend in gasoline consumption (evident in the scatterplot).

The results of the regression (exactly as they are reported by Stata) are in the figure below.

```
. reg gasCon income pnc puc ppt pd pn ps year
```

Source	SS	df	MS		
Model	8.52460148	8	1.06557518	Number of obs =	52
Residual	.03803733	43	.000884589	F(8, 43) =	1204.60
				Prob > F =	0.0000
Total	8.56263881	51	.167894879	R-squared =	0.9956
				Adj R-squared =	0.9947
				Root MSE =	.02974

gasCon	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0000706	.0000115	6.13	0.000	.0000473	.0000938
pnc	.0064415	.0027322	2.36	0.023	.0009315	.0119514
puc	.0000635	.0013256	0.05	0.962	-.0026098	.0027368
ppt	.0023999	.0012614	1.90	0.064	-.0001439	.0049438
pd	-.0067177	.0031152	-2.16	0.037	-.013	-.0004354
pn	-.000604	.0024084	-0.25	0.803	-.0054611	.0042531
ps	-.0056064	.0021818	-2.57	0.014	-.0100064	-.0012065
year	.0165505	.0035655	4.64	0.000	.0093598	.0237411
_cons	-32.48513	6.854403	-4.74	0.000	-46.30835	-18.66191

(a)

Do the signs of the coefficients make sense? Are they consistent with your expectations? Why or why not?

(b)

Do consumers seem to differentiate between the price of new cars and the price of used cars when it comes to gasoline consumption? That is, does it seem that the price of new and used cars have the same effect on gasoline consumption? Give as much evidence as you can.

(c)

How would you test the hypothesis that the price of new and used cars have the same effect on gasoline consumption? Give as much detail as you can. What would be the outcome of this test. If you can't calculate it exactly, can you guess what the outcome would be? What are you basing your guess on? Be as specific as possible.

(d)

"Elasticities" are a common way to express relationships in economics. For instance, the income elasticity of gasoline consumption is the percentage change in gasoline consumption due to a 1% increase in income. If we had estimated a log-log model, the income elasticity of gasoline consumption would have been given directly by the coefficient on $\log(income)$. Given that we estimated a linear model, how could you develop an estimate of the average income elasticity of gasoline consumption? Imagine that you

have the dataset loaded into Stata. What procedure would you use to calculate the average income elasticity of gasoline consumption in the data? There is no need to list Stata commands. You can describe the steps in any way you see fit. If you know Stata commands and wish to use them in your description, that is fine too.

(e)

I also re-ran the model as a log-log specification (the dependent variable and all the independent variables except for `year` have been expressed in logarithms). The results of the regression are printed below.

The results of the regression (exactly as they are reported by Stata) are in the figure below.

```
. reg lGasCon lincome lPnc lPuc lPpt lPd lPn lPs year
```

Source	SS	df	MS	Number of obs =	52
Model	8.29583465	8	1.03697933	F(8, 43) =	844.83
Residual	.05277999	43	.001227442	Prob > F =	0.0000
				R-squared =	0.9937
				Adj R-squared =	0.9925
Total	8.34861464	51	.163698326	Root MSE =	.03503

lGasCon	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lincome	.7831195	.270542	2.89	0.006	.2375196	1.328719
lPnc	-.4315087	.2757007	-1.57	0.125	-.9875121	.1244947
lPuc	-.5159915	.0983908	-5.24	0.000	-.7144154	-.3175675
lPpt	.0051545	.156962	0.03	0.974	-.3113895	.3216986
lPd	2.082664	.3003297	6.93	0.000	1.476991	2.688336
lPn	-.6926594	.2749374	-2.52	0.016	-1.247124	-.1381952
lPs	-.9996042	.4076378	-2.45	0.018	-1.821684	-.1775242
year	.0608358	.0087269	6.97	0.000	.0432364	.0784352
_cons	-125.6048	15.25337	-8.23	0.000	-156.3662	-94.84346

Compare the log-log model to the linear model. Compare the two models using all the information you can glean from the regression output. What are the differences? What might lead you to prefer one model to the other? Is it possible that a mixed model (a model with some linear and some logarithmic relationships) might be preferable?

7.

There is an urn containing 3 balls. Each ball in the urn might be either black or red.

(a)

Write down all the possibilities for the number of black and red balls.

(b)

You pass the urn around a room with 10 people in it. Each person draws a ball (without looking in the urn), records whether it is red or black, and then returns the ball to the

urn (again without looking in the urn). The first person records a red and the second person records a black. We record this data as {red, black}. Using that same recording scheme, all of the data looks like this:

{red, red, red, black, red, black, black, red, red, red}.

Your task is to estimate the number of red balls in the urn. I used Stata to create the datset displayed below.

	color	red
1.	red	1
2.	red	1
3.	red	1
4.	black	0
5.	red	1
6.	black	0
7.	black	0
8.	red	1
9.	red	1
10.	red	1

I then ran the following regression and obtained the results you see below.

```
. reg red
```

Source	SS	df	MS	Number of obs =	10
				F(0, 9) =	0.00
Model	0	0	.	Prob > F =	.
Residual	2.1	9	.233333333	R-squared =	0.0000
				Adj R-squared =	0.0000
Total	2.1	9	.233333333	Root MSE =	.48305

red	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	.7	.1527525	4.58	0.001	.3544498	1.04555

Using these results, estimate the number of red balls in the urn. Show your work and explain how you use the results to obtain your estimate.

(c)

Estimate the number of red balls in the urn using maximum likelihood. Show your work.

8.

This question is based on a dataset describing admissions to graduate school. Each observation represents an individual who applied to a particular graduate school.

The `admit` variable is an indicator ("dummy") variable. The `gre` and `gpa` variables are continuous variables. The `rank` variable is a categorical variable, indicating whether

Table 2: graduate school admissoins

`admit`	1 if the individual was admitted, 0 otherwise
`gre`	the individual's GRE score
`gpa`	the individual's undergraduate gpa
`rank`	the prestige of the individuals undergraduate institution

the undergraduate institution from which the applicant graduated was of the highest prestige rank (indicated by a 1), the next-highest rank (indicated by a 2), and so on. To represent this categorical variable we include a series of dummy variables. For instance, we could include dummy variables for rank 2 schools (equal to 1 if the individual went to a rank 2 school and 0 otherwise), rank 3 schools, and rank 4 schools. Recall that we need to omit one category to avoid perfect multicollinearity.

I ran the following regression in Stata (note the nifty trick using the "`i.`" syntax in Stata to automatically create dummy variables for all but one of the categories of `rank`).

```
. probit admit gre gpa i.rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.29667
Iteration 2:   log likelihood = -229.20659
Iteration 3:   log likelihood = -229.20658

Probit regression                                 Number of obs   =        400
                                                  LR chi2(5)      =      41.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.20658                       Pseudo R2       =     0.0831

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |    .0013756   .0006489     2.12   0.034     .0001038    .0026473
         gpa |    .4777302   .1954625     2.44   0.015     .0946308    .8608297
             |
        rank |
           2 |   -.4153992   .1953769    -2.13   0.033    -.7983308   -.0324675
           3 |    -.812138   .2085956    -3.89   0.000    -1.220978   -.4032981
           4 |    -.935899   .2456339    -3.81   0.000    -1.417333   -.4544654
             |
       _cons |   -2.386838   .6740879    -3.54   0.000    -3.708026   -1.065649
------------------------------------------------------------------------------
```

(a)

What can you tell about the marginal effects of `gre` and `gpa` on the probability of being admitted to graduate school? What can you not tell?

(b)

What can you tell about the relationship between prestige of undergraduate institution and admission to graduate school based on the coefficients on each of the rank variables.

9.

Based on the results of the regression estimates of the wage equation below, what is the effect on wage of one additional year spent at the same firm?

$$\widehat{\log(wage)} = 0.284 + 0.092educ + 0.0041exper + 0.022tenure.$$

(Assume that the variable names have the obvious meanings)

10.

Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u.$$

Explain (in steps) how you would go about testing the following null hypothesis against the alternative that the null hypothesis is not true.

$$H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

You do not need to actually estimate anything or calculate a statistic (I haven't given you any data!). But you should specify what test you will use, and how you will generate the necessary inputs to the test statistic.

11.

In a linear probability model, all of the following statements are true, except:
(a) the estimated coefficients can be interpreted directly as marginal effects.
(b) R2 is a good measure of how well the model fits the data.
(c) the predicted probability can be negative.
(d) the errors are always heteroskedastic.

12.

Suppose you run a regression of the log of household expenditures on a dummy for the household being situated in an urban area, along with a few control variables. You find a t-statistic of 2 for the urban dummy. Adding another variable to the regression will result in
(a) a larger t-statistic for the urban dummy if the additional variable is highly correlated with the log of household expenditures
(b) a smaller t-statistic for the urban dummy if the additional variable is not correlated with the log of household expenditures
(c) a smaller t-statistic for the urban dummy if the additional variable is highly correlated with the log of household expenditures
(d) a and b

13.

A weak instrument
(a) is not exogenous
(b) is not highly correlated with the dependent variable
(c) is not highly correlated with the endogenous explanatory variable of interest

14.

True or False: Since x^2 is an exact function of x, you will be faced with a perfect multicollinearity problem if you include both x and x^2 in a multivariate regression model.

15.

A researcher wishing to investigate the extent to which math background influences final exam scores in an undergraduate introductory economics course regresses final exam score on several explanatory variables, including the students score on a pop math quiz taken during an early class. Some students had to be dropped from the data set because they missed the class in which the quiz was given. The researcher argues that this is not of concern for the study since most students were there that day; in particular, only about 10% of the class has to be dropped, so he still has data on 90% of the students. Do you believe that this situation could compromise the validity of the estimates? Explain why or why not.