

**Suggested Answers, Problem Set 3**  
**ECON 30331**

**Dan Hungerman**

1. This is not a very good idea. We know from the second FOC in problem 1b) that

$$(2) \quad \partial SSE / \partial \hat{\beta}_1 = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1) x_i = 0$$

Which can be reduced to read

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1) x_i = \sum_{i=1}^n \hat{\varepsilon}_i x_i = 0$$

The OLS model chooses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  such that  $x_i$  is by construction uncorrelated with  $\hat{\varepsilon}_i$ . Therefore, the estimate for  $\hat{\gamma}_1$  will be by construction 0 and it does not inform us at all about whether  $x_i$  and  $\varepsilon_i$  are correlated.

2. The three 1<sup>st</sup> order conditions are:

$$(1) \quad \frac{\partial sse}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_{1i} \hat{\beta}_1 - x_{2i} \hat{\beta}_2) x_{1i} = 0$$

$$(2) \quad \frac{\partial sse}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_{1i} \hat{\beta}_1 - x_{2i} \hat{\beta}_2) x_{2i} = 0$$

$$(3) \quad \frac{\partial sse}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_{1i} \hat{\beta}_1 - x_{2i} \hat{\beta}_2) = 0$$

Equation (3) can be reduced to read  $\sum_{i=1}^n (y_i - \hat{\beta}_0 - x_{1i} \hat{\beta}_1 - x_{2i} \hat{\beta}_2) = 0$ . Dividing by n and solving for  $\hat{\beta}_0$  we find that  $\hat{\beta}_0 = \bar{y} - \bar{x}_1 \hat{\beta}_1 - \bar{x}_2 \hat{\beta}_2$  and because we have assumed that  $\bar{y} = \bar{x}_1 = \bar{x}_2 = 0$  then  $\hat{\beta}_0 = 0$ . Equation

(1) can be re-written to read  $\sum_{i=1}^n y_i x_{1i} - \hat{\beta}_1 \sum_{i=1}^n x_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^n x_{1i} x_{2i} = 0$ . Since  $\hat{\beta}_0 = 0$  and

$\sum_{i=1}^n x_{1i} x_{2i} = 0$  this reduces to  $\sum_{i=1}^n y_i x_{1i} - \hat{\beta}_1 \sum_{i=1}^n x_{1i}^2 = 0$  and therefore  $\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_{1i}}{\sum_{i=1}^n x_{1i}^2} = 60 / 20 = 3$ . Using the

same procedure, you can also demonstrate that  $\hat{\beta}_2 = \frac{\sum_{i=1}^n y_i x_{2i}}{\sum_{i=1}^n x_{2i}^2} = 80 / 40 = 2$ .

3. True, by adding more variables, no matter how irrelevant the variables are, the  $R^2$  can never fall. This is because if the  $R^2$  was  $R_a$  with only 1 variable, the worst that would ever happen by adding more variables is that the computer would set the estimated coefficients for the new variables to zero and obtain an  $R^2$  of  $R_a$ .

4. A sample program that generates results for this question is under the Chapter 3 heading in the STATA portion of the class web page. The program is called aps3\_q3.do.

Source	SS	df	MS	Number of obs = 95		
Model	5.34106991	4	1.33526748	F( 4, 90)	=	95.30
Residual	1.2609981	90	.01401109	Prob > F	=	0.0000
				R-squared	=	0.8090
				Adj R-squared	=	0.8005
Total	6.60206802	94	.070234766	Root MSE	=	.11837

  

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lcost	-.0070438	.0361431	-0.19	0.846	-.0788483	.0647607
lsat	.0178983	.0042339	4.23	0.000	.0094868	.0263097
rank	-.0036089	.0004302	-8.39	0.000	-.0044635	-.0027543
age	.0002676	.0003653	0.73	0.466	-.0004581	.0009934
_cons	8.038384	.7234791	11.11	0.000	6.601066	9.475701

- A 10% increase in cost is estimated to reduce salaries by .07 percent.
- A one unit increase in rank (moving from 5<sup>th</sup> to 6<sup>th</sup> for example) is estimated to reduce salaries by .36 percent.
- Below is the matrix of correlation coefficients. Just like is predicted by the first order conditions, the covariance between the estimated residuals and the x's is by construction equation to zero

	res1	lsat	lcost
res1	1.0000		
lsat	0.0000	1.0000	
lcost	0.0000	0.4930	1.0000

- The correlation coefficient between actual and predicted y is 0.8994 and this number squared is 0.908 which is exactly the R<sup>2</sup> in the model

	lsalary	pred
lsalary	1.0000	
pred	0.8994	1.0000

- Below are the results when LSAT is removed from the model. Note that the correlation coefficient between lsat and rank is -0.73. We know that ln(salaries) are negatively related to rank and negatively correlated with the lsat so taking rank out of the model would put more weight on the lsat variable in the regression and increase its value, which is exactly what happens. Notice that the coefficient on lsat doubles when school rank is eliminated from the model

```
. * run model deleting lsat from basic model
. reg lsalary lcost lsat age
```

Source	SS	df	MS	Number of obs = 95		
Model	4.35484336	3	1.45161445	F( 3, 91)	=	58.78
Residual	2.24722465	91	.024694776	Prob > F	=	0.0000
				R-squared	=	0.6596
				Adj R-squared	=	0.6484
Total	6.60206802	94	.070234766	Root MSE	=	.15715

  

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsalary						

lcost	.0847587	.0457317	1.85	0.067	-.0060817	.1755991
lsat	.0388551	.0045385	8.56	0.000	.0298399	.0478703
age	.0015209	.0004426	3.44	0.001	.0006418	.0024001
_cons	3.469744	.6323767	5.49	0.000	2.213605	4.725882

5. A sample program that generates results for this question is under the Chapter 3 heading in the STATA portion of the class web page. The program is called `aps3_q4.do`.

**Model 1:**

Source	SS	df	MS	Number of obs =	114
Model	942250.712	4	235562.678	F( 4, 109) =	8.32
Residual	3086043.86	109	28312.329	Prob > F =	0.0000
Total	4028294.57	113	35648.6246	R-squared =	0.2339
				Adj R-squared =	0.2058
				Root MSE =	168.26

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bedrooms	26.05118	18.12206	1.44	0.153	-9.866149	61.96852
bathrooms	109.7691	27.9523	3.93	0.000	54.3685	165.1696
otherrooms	32.03491	13.73668	2.33	0.022	4.809249	59.26057
age	.3275602	.4960419	0.66	0.510	-.6555788	1.310699
_cons	-14.03946	72.17339	-0.19	0.846	-157.0848	129.0058

- a) Remember, house prices are measured in thousands of dollars. Each additional bedroom increase house prices by \$26,000. Every year increase in age increase house prices by \$320.
- b) Notice that when `sq_feet` is added to the model, the coefficients on bedrooms, bathrooms and otherrooms decline so much that the signs are all now negative. This makes sense because `sq_feet` is positively correlated with these three variables so adding it to the model should decrease the coefficients on these three variables.

**Model 2**

Source	SS	df	MS	Number of obs =	114
Model	1604241.53	5	320848.306	F( 5, 108) =	14.29
Residual	2424053.05	108	22444.9356	Prob > F =	0.0000
Total	4028294.57	113	35648.6246	R-squared =	0.3982
				Adj R-squared =	0.3704
				Root MSE =	149.82

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bedrooms	-21.91485	18.39449	-1.19	0.236	-58.37592	14.54622
bathrooms	-.9638506	32.17371	-0.03	0.976	-64.73772	62.81002
otherrooms	-5.301832	14.03055	-0.38	0.706	-33.11282	22.50915
age	-.1375338	.449888	-0.31	0.760	-1.02929	.7542222
sq_feet	.2027686	.0373365	5.43	0.000	.1287611	.2767761
_cons	80.73887	66.58876	1.21	0.228	-51.25161	212.7293

- c) Notice that the  $R^2$  for model 3 is 0.3903 while the  $R^2$  for model 2 is 0.3982, not much of a change. In this sample, once one controls for `sq_feet`, adding information about the number of rooms does not add much explanatory power to the model.

**Model 3**

Source	SS	df	MS	Number of obs = 114		
Model	1572268.9	2	786134.448	F( 2, 111)	=	35.53
Residual	2456025.68	111	22126.3575	Prob > F	=	0.0000
				R-squared	=	0.3903
				Adj R-squared	=	0.3793
Total	4028294.57	113	35648.6246	Root MSE	=	148.75

  

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.2359865	.4178868	-0.56	0.573	-1.064057	.5920842
sq_feet	.1796559	.0214987	8.36	0.000	.1370547	.222257
_cons	40.32538	46.32445	0.87	0.386	-51.46961	132.1204

5. a) If  $x_{1i}$  is a linear combination of  $x_{2i}$  where  $x_{1i}=a+bx_{2i}$ , then a regression of  $x_{1i}$  on  $x_{2i}$  would generate an  $R^2$  of 1 and the denominator in the variance calculation would be zero, making the variance undefined. We cannot estimate models where covariates are linear combinations of each other.

b) If a regression of  $x_{1i}$  on  $x_{2i}$  produces an  $R_1^2$  like 0.999, then the denominator approaches zero and the variance would explode. When two highly correlated variables are added to a model, it is difficult to discern anything precise about the exact impact of  $x_{1i}$  on  $y$  because it is hard to separate the exact effect of  $x_{1i}$  from that of  $x_{2i}$ .

7. a) Since  $x_{1i}$  is randomly assigned then we expect it to be uncorrelated with all of the possible covariates. As a result, adding these new variables to the model is not expected to change the estimate on  $\hat{\beta}_1$ .

b) in a simple bivariate model, the variance on  $\hat{\beta}_1$  would be  $\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$ . In the multivariate

model where  $\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{(1 - R_1^2) \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$ , since we expect that  $x_{1i}$  will be uncorrelated with all of the

possible covariates, then  $R_1^2$  should be pretty close to zero and the variance in the multivariate case should

look a lot like the variance in the simple bivariate regression model, or  $\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$ . However,

recall that  $\hat{\sigma}_\varepsilon^2 = SSE / (n - k - 1)$  and adding covariates to the model should reduce the SSE and therefore, reduce the estimated variance on  $\hat{\beta}_1$ . In Random Assignment Clinical Trials, we typically add covariates because they reduce the objective function (SSE) which directly reduces estimated variances.

8. In a bivariate regression model, we know that  $Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$  whereas in a multivariate

regression model, we know that  $Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{(1 - R_1^2) \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$  where  $R_1^2$  is the  $R^2$  from a regression of

$x_{1i}$  on  $x_{2i}$ . Note that in top of the results on page 8, we see the correlation coefficient between  $x_{1i}$  on  $x_{2i}$  is 0.9994 which means that  $R_1^2$  should be very close to 1. Therefore, by adding  $x_{2i}$  to the model, a variable highly correlated with  $x_{1i}$ , the numerator in  $Var(\hat{\beta}_1)$  in model (2) blows up because  $1 - R_1^2$  approaches zero.