Suggested Answers, Problem Set 4
ECON 30331

Bill Evans
Spring 2018

1. The three 1st order conditions are:

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

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

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

Equation (3) can be reduced to read $\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - x_{1i}\hat{\beta}_1 - x_{2i}\hat{\beta}_2\right) = 0$. 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 $\sum_{i=1}^{n} y_i x_{1i} - \hat{\beta}_0 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^{n} x_{1i}x_{2i} = 0$. Since $\hat{\beta}_0 = 0$ and $\sum_{i=1}^{n} x_{1i}x_{2i} = 0$ this reduces to $\sum_{i=1}^{n} y_i x_{1i} - \beta_1 \sum_{i=1}^{n} x_{1i}^2 = 0$ and therefore $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_{1i}}{\sum_{i=1}^{n} x_{1i}^2} = 120/40 = 3.$ Using the same procedure, you can also demonstrate that $\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} = 160/80 = 2.$

2. Below are the results for this regression. Given the regression $\ln(weekly\ earn) = \beta_0 + age_i\beta_1 + age_i^2\beta_2 + educ\beta_3 + \varepsilon_i$, the derivative with respect to age is

$$\frac{\partial \ln(weekly\ earn)}{\partial age} = \beta_1 + 2\beta_2 age$$

This means that the derivative is a function of age. Given estimates the three derivatives are:
At age 21: 0.071 – 2(0.00071)21 = 0.041 -- an additional year of age increases ages by 4.1%
At age 35: 0.071 – 2(0.00071)35 = 0.021 – and additional year of age increases wages by 2.1%
At age 50: 0.071 – 2(0.00071)50 = -0.001 – an additional year of age decreases wages by 0.1%

```
. reg ln_weekly_earn age age2 years_educ

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  3, 19902) = 2644.18
       Model |  1493.15489     3  497.718296           Prob > F      =  0.0000
    Residual |   3746.1838 19902  .188231524           R-squared     =  0.2850
-------------+------------------------------           Adj R-squared =  0.2849
       Total |  5239.33869 19905  .263217216           Root MSE      =  .43386
```

```
------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0712533   .0020266    35.16   0.000      .067281    .0752256
        age2 |  -.0007069   .0000248   -28.51   0.000    -.0007555   -.0006583
  years_educ |   .0719499   .0011135    64.62   0.000     .0697674    .0741324
       _cons |   3.522066   .0397997    88.49   0.000     3.444055    3.600077
------------------------------------------------------------------------------
```

3. True. Remember, the definition of the R^2 is 1-SSR/SST – by adding more variables to the system SSR can never go up -- no matter how irrelevant the variables are that are added to the system. 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 the original SSR and hence the original and R^2. Therefore, the R^2 can only increase when more variables are added to the system.

4. A sample program that generates results for this question is called house_price.do.

Model 1:

```
      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  4,   109) =    8.32
       Model |  942250.712     4  235562.678           Prob > F      =  0.0000
    Residual |  3086043.86   109   28312.329           R-squared     =  0.2339
-------------+------------------------------           Adj R-squared =  0.2058
       Total |  4028294.57   113  35648.6246           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 $26K. Every year increase in age increase house prices by $328.
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 the other three variables. To many this was counterintuitive – why would more bedrooms be bad? Remember that the coefficients are assuming all else is held constant. Therefore, how do you get another bedroom "holding square feet" constant? You can only do this by having smaller bedrooms – which home buyers find a negative attribute.
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 2

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

.

Model 3

```
      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  2,   111) =   35.53
       Model |   1572268.9     2  786134.448           Prob > F      =  0.0000
    Residual |  2456025.68   111  22126.3575           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 sample program that generates results for this question is on the class web page. The program is called law_school.do.

```
      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F(  4,    90) =   95.30
       Model |  5.34106991     4  1.33526748           Prob > F      =  0.0000
    Residual |   1.2609981    90   .01401109           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) The elasticity of salaries with respect to the cost of law school is -0.007 or a 10% increase is cost is estimated to reduce salaries by 0.07 percent.
b) A one unit increase in rank (moving from 5th to 6th for example) is estimated to reduce salaries by .36 percent.
c) 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
```

d) The correlation coefficient between actual and predicted y is 0.8994 and this number squared is 0.908 which is exactly the R^2 in the model

```
             |  lsalary     pred
-------------+------------------
     lsalary |   1.0000
        pred |   0.8994   1.0000
```

E) 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 ot 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
-------------+------------------------------           F(  3,    91) =   58.78
       Model |  4.35484336     3  1.45161445           Prob > F      =  0.0000
    Residual |  2.24722465    91  .024694776           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]
-------------+----------------------------------------------------------------
       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
```

F) Below are the results of part f). Note that when we use the residuals from a regression of lcost on the other covariates from the model in part a) we obtain the exact same coefficient as we do for the beta on lcost in model a. When estimating beta, the regression only uses the portion of x that is NOT predicted by other covariates in the model.

```
predict error_lcost, residual

.
. reg lsalary error_lcost

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F(  1,    93) =    0.01
       Model |  .000532152     1  .000532152           Prob > F      =  0.9312
    Residual |  6.60153586    93  .070984257           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0107
       Total |  6.60206802    94  .070234766           Root MSE      =  .26643

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 error_lcost |  -.0070438   .0813524    -0.09   0.931    -.1685935    .1545059
       _cons |   10.55491    .027335   386.13   0.000     10.50063    10.60919
```

---

6. 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 $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}$. In the multivariate model where $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_\varepsilon^2}{(1-R_1^2)\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}$, 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 $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}$. However, recall that $\hat{\sigma}_\varepsilon^2 = SSE/(n-k-1)$ and adding covariates to the model should reduce the SSE and therefore, if the reduction in SSE is larger than the increase in the change in degrees of freedom, it should 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 – hopefully, reduces estimated variances.

7. In a bivariate regression model, we know that $Var(\hat{\beta}_1) = \dfrac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}$ whereas in a multivariate regression model, we know that $Var(\hat{\beta}_1) = \dfrac{\sigma_\varepsilon^2}{(1-R_1^2)\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}$ where R_1^2 is the R^2 from a regression of x_{1i} on x_{2i}.

Note that in results, 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.

8. If Model (2) is the correct model, we know the expected bias generated in model (1) is $E[\tilde{\beta}_1] = \beta_1 + \beta_2\hat{\delta}_1$ where $\hat{\delta}_1$ is the coefficient from the regression $x_{2i} = \delta_0 + \delta_1 x_{1i} + \phi_i$. In this case, we expect that $\hat{\delta}_1 < 0$ – people with more medical conditions are less likely to take advantage of the free exercise classes. We are also expect that $\beta_2 > 0$ (more poor health conditions tend to increase medical care costs). Therefore, because the product $\beta_2\hat{\delta}_1$ is a negative value, the estimate for $\tilde{\beta}_1$ would be biased down –by ignoring the fact that healthier people tend to enroll in the exercises classes, we are attributing too much to the exercise class.

9. If Model (2) is the correct model, we know the expected bias generated in model (1) is therefore $E[\tilde{\beta}_1] = \beta_1 + \beta_2\hat{\delta}_1$ where $\hat{\delta}_1$ is the coefficient from the regression $x_{2i} = \delta_0 + \delta_1 x_{1i} + \phi_i$. In this case, we expect that $\hat{\delta}_1$ >0 -- Higher skilled students will attend better schools. We are also expect that $\beta_2>0$ (more skilled students will earn more in the workforce). Therefore, the estimate for $\tilde{\beta}_1$ would be biased up –by ignoring the fact that higher test score kids both attend better schools and tend to make higher earnings, we overstate the impact of school quality on earnings.

10. The correlation coefficients at the end of the printout indicate that x_2, x_3 and x_4 are weakly correlated with x_1 at best and therefore, the inclusion of these variables in the model, no matter how well correlated they are with Y, will not change the coefficient on β_1.