Homework 21: ECO220Y

Required Exercises: Chapter 20: 9, 27

Required Problems:

(1) In predicting percent body fat for a sample of 65 males using weight, height, abdominal circumference, and age an R^2 of 0.1345 is obtained. Is the multiple regression statistically significant overall? If so, at which significance levels?

(2) Consider a regression with five RHS variables (i.e. x variables) to explain y.

(a) Using a random sample with 21 observations, how big does the R^2 need to be for the model overall to be statistically significant at the 0.1% level?

(b) Using a random sample with 126 observations, how big does the R^2 need to be for the model overall to be statistically significant at the 0.1% level?

(c) Using a random sample with 100,000 observations, how big does the R^2 need to be for the model overall to be statistically significant at the 0.1% level?

(d) Why are the answers so different for parts (a), (b) and (c)?

(3) Suppose for a random sample of 500 houses the mean selling price, measured in thousands of dollars, is 201.420 with a s.d. of 63.553. A multiple regression model uses five variables (living area, lot size, age, number of bathrooms, and number of bedrooms) to predict the selling price. The R-squared is 0.4782. What is the value of s_e (i.e. the Root MSE, s.d. of the residuals)? Make sure to include the units of measurement.

(4) Consider the correlation matrix below.

```
correlate pct_body_fat height_cm abdomen_cm age weight_kg if (case_number~=39 &
case_number~=42)

(obs=250)
             | pct_body_fat  height_cm  abdomen_cm      age  weight_kg
-------------+------------------------------------------------------
pct_body_fat |       1.0000
   height_cm |      -0.0294     1.0000
  abdomen_cm |       0.8237     0.1867      1.0000
         age |       0.2951    -0.2459      0.2428   1.0000
   weight_kg |       0.6173     0.5129      0.8737  -0.0161     1.0000
```

(a) Is the correlation between age and pct_body_fat statistically significant? If so, at which significance levels?

(b) Is the correlation between age and weight statistically significant? If so, at which significance levels?

(5) Consider the R^2 and the adjusted R^2: $R^2 = 1 - \frac{SSE}{SST}$ and $Adj\ R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$.

(a) Can the Adjusted R^2 ever be bigger than the R^2?

(b) Can the Adjusted R^2 ever be negative? How about the R^2? Explain.

(c) When will the Adjusted R^2 and R^2 be most different?

(6) Consider twenty x variables (x1 – x20) that are entirely unrelated to y in the population. For example, you have data for 100 Canadian adults and try to predict years of education using measures such as the circumference of the person's maternal grandfather's knee when he was 18, the last two numbers of the person's mother's SIN, the number of the house/apt where the person grew up, the temperature in Miami Florida on the day the person lost their first tooth, etc. What would happen if you regress y on all of these irrelevant variables? You might think that none of the estimated coefficients would come out to be statistically significant. However, because there are twenty chances for Type I error you would likely reject at least one true null hypothesis (H_0: $\beta_j = 0$). Here is an example of a regression of y on twenty completely irrelevant variables. Which coefficient is statistically significant in this example? Is the model overall statistically significant?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F( 20,    79) =    0.90
       Model |  13.8586384    20   .69293192           Prob > F      =  0.5906
    Residual |   60.982375    79  .771928798           R-squared     =  0.1852
-------------+------------------------------           Adj R-squared = -0.0211
       Total |  74.8410134    99  .755969833           Root MSE      =  .87859
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0798302   .1022369     0.78   0.437    -.1236672    .2833276
          x2 |   .1233008   .1048957     1.18   0.243    -.0854888    .3320905
          x3 |  -.1387776   .0915604    -1.52   0.134     -.321024    .0434689
          x4 |   .0262483   .1018933     0.26   0.797    -.1765652    .2290619
          x5 |   .1242398   .0914512     1.36   0.178    -.0577893    .3062688
          x6 |   .1557071   .0989298     1.57   0.120    -.0412078     .352622
          x7 |   .0527839   .1035481     0.51   0.612    -.1533234    .2588913
          x8 |   -.098754   .0994973    -0.99   0.324    -.2967984    .0992903
          x9 |  -.2459072   .1053007    -2.34   0.022    -.4555029   -.0363114
         x10 |  -.0156936    .088371    -0.18   0.860    -.1915917    .1602045
         x11 |   .0827041   .0981579     0.84   0.402    -.1126743    .2780826
         x12 |  -.1263454   .0974091    -1.30   0.198    -.3202333    .0675426
         x13 |  -.1245131   .1036529    -1.20   0.233    -.3308291    .0818028
         x14 |  -.0410782   .1008314    -0.41   0.685    -.2417781    .1596218
         x15 |  -.0513262   .1108718    -0.46   0.645     -.272011    .1693586
         x16 |  -.0435848   .0912515    -0.48   0.634    -.2252164    .1380469
         x17 |  -.0407015   .0962939    -0.42   0.674    -.2323697    .1509668
         x18 |  -.0575465   .1011599    -0.57   0.571    -.2589003    .1438074
         x19 |   .0440475   .0888704     0.50   0.622    -.1328445    .2209396
         x20 |  -.0191734   .1011402    -0.19   0.850    -.2204878     .182141
       _cons |   .2698414   .1033971     2.61   0.011     .0640347    .4756481
```

(7) We can use a Monte Carlo simulation to illustrate a key conceptual point from lecture: for testing the overall statistical significance of a multiple regression model the F test should be used and NOT k different t tests. Recall the definition of a Monte Carlo simulation: A problem solving method where a computer generates many random samples and the researcher makes an inference based on patterns in outcomes. If you are uncomfortable with the concept of simulation, review Lecture 10. Consider the following set-up for a Monte Carlo simulation:

1. Theoretical model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{20} x_{20i} + \varepsilon_i$
2. Computer generates a random sample:
   - $y, x_1, x_2, \ldots, x_{20}$ are <u>independently</u> drawn from N(0,1)
   - 100 observations
   - A computer estimates the regression parameters $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_{20} x_{20i}$ and the standard errors and test statistics
   - It records the t statistics and the F statistic

- To systematically record the outcomes, each regression results appear as an row in a new data set with the following variables:

Variable	Description
num_sig90	Number of significant coefficients, $\alpha = 0.10$
num_sig95	Number of significant coefficients, $\alpha = 0.05$
num_sig99	Number of significant coefficients, $\alpha = 0.01$
joint_sig90	=1 if regression significant, =0 otherwise, $\alpha = 0.10$
joint_sig95	=1 if regression significant, =0 otherwise, $\alpha = 0.05$
joint_sig99	=1 if regression significant, =0 otherwise, $\alpha = 0.01$
any_sig90	=1 if any coefficients significant, =0 otherwise, $\alpha = 0.10$
any_sig95	=1 if any coefficients significant, =0 otherwise, $\alpha = 0.05$
any_sig99	=1 if any coefficients significant, =0 otherwise, $\alpha = 0.01$

3. A computer repeats Step **2** 9,999 more times.

4. The simulation results are summarized here.

Variable	Obs	Mean	Std. Dev.	Min	Max
num_sig90	10000	2.0195	1.543489	0	10
num_sig95	10000	1.0076	1.117975	0	8
num_sig99	10000	.2033	.4783221	0	4
joint_sig90	10000	.1051	.3066976	0	1
joint_sig95	10000	.0562	.230319	0	1
joint_sig99	10000	.0117	.1075373	0	1
any_sig90	10000	.8383	.3681938	0	1
any_sig95	10000	.5937	.4911664	0	1
any_sig99	10000	.1741	.3792144	0	1

(a) What are the true parameter values in this simulation? Explain.

(b) Are the first 6 rows of results what we expected?

(c) Of the 10,000 regressions estimated, how many had at least one statistically significant coefficient estimate at the 5% level? What is the relevant test statistic? What is the critical value of the test statistic? When should you infer statistical significance?

(d) Of the 10,000 regressions estimated, how many were statistically significant overall at the 5% significance level? What is the relevant test statistic? What is the critical value of the test statistic? When should you infer statistical significance?

(e) Explain why the answers to **(c)** and **(d)** are very different.

(f) Analytically compute the expected value of each of the 9 variables: num_sig90 – any_sig99. (In this example a Monte Carlo simulation is not necessary: the probabilities and means can all be found analytically.) Hint: You'll need to remember your probability rules to do this. Explain why the Monte Carlo Simulation results are similar but not exactly the same.