Homework 21: ECO220Y – SOLUTIONS

Required Problems:

(1) $H_0: \beta_{weight} = \beta_{height} = \beta_{abdominal} = \beta_{age} = 0$ versus H_1: Not all slope coefficients are zero. With the given information, we can compute the F test statistic as: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.1345/4}{(1-0.1345)/(65-4-1)} = 2.33$. Looking at the F table, we see that for $\nu_1 = k = 4$ and $\nu_2 = n - k - 1 = 65 - 4 - 1 = 60$ the critical value for $\alpha = 0.10$ is 2.04 and the critical value for $\alpha = 0.05$ is 2.53. Hence the P-value for our test statistic is between 0.10 and 0.05, which means that our model is statistically significant overall at a 10% significance level but not a 5% significance level.

(2) For both parts, we must test:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

H_1: Not all slope coefficients are zero

using the $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ test statistic.

(a) The F table gives the critical value of 7.57. Hence, we need an F test statistic at least that big for the regression to be statistically significant overall at 0.1% significance level. Solving $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ for R^2 yields 0.716.

(b) The F table gives the critical value of 4.42. Hence, we need an F test statistic at least that big for the regression to be statistically significant overall at 0.1% significance level. Solving $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ for R^2 yields 0.155.

(c) The F table gives the critical value of 4.10. Hence, we need an F test statistic at least that big for the regression to be statistically significant overall at 0.1% significance level. Solving $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ for R^2 yields 0.000205.

(d) With a small sample (like in part (a)) we need a large R-squared of at least 0.716 – very good fit – even to rule out no relationship at all between any of the x variables and y. With small sample sizes, which are subject to lots of sampling error, we need quite dramatic evidence to rule out the null. In contrast, with a larger sample size (like in part (b)), we have less sampling error and hence we can prove that there is a relationship between y and the combination of the five x variables even if we have an R-squared as small as 0.155. In part (c) we have an extremely large sample size and there we would have a highly statistically significant regression even if it had a ridiculously tiny R-squared of 0.000205. Remember that with very large sample sizes even results that are tiny and not at all economically significant will become statistically significant.

(3) We can easily find the SST because the $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and we have been given $s_y = 63.553$. Recall that $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{SST}{n-1}}$ so $SST = 2{,}015{,}452.9$. Given that the $R^2 = \frac{SSR}{SST} = 0.4782$, the $SSR = 963{,}789.6$. Because $SST = SSR + SSE$ we obtain the $SSE = 1{,}051{,}663.3$ and then the $s_e = Root\ MSE = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{1{,}051{,}663.3}{500-5-1}} = 46.140$. Hence the standard deviation of the residuals is about $46,140, which means that there is a lot of variability in housing prices that is not being explained by our five explanatory variables.

(4) **(a)** We can have simple regression model of $pctbodyfat_i = \alpha + \beta age_i + \varepsilon_i$ in mind (or you could write the hypotheses like in Section 18.5 of the textbook), $H_0: \beta = 0$ versus $H_1: \beta \neq 0$. With the given information, we can compute the F test statistic as: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(0.2951^2)/1}{(1-0.2951^2)/(250-1-1)} = 23.65$. Looking at the F table, we see that for $\nu_1 = k = 1$ and $\nu_2 = n - k - 1 = 250 - 1 - 1 = 248$ the critical value for $\alpha = 0.001$ is 10.83. Hence the P-value for our test statistic is below 0.001, which means that our model is highly statistically significant overall at even a 0.1% significance level.

(b) We can have simple regression model of $weight_i = \alpha + \beta age_i + \varepsilon_i$ in mind (or you could write the hypotheses like in Section 18.5 of the textbook), $H_0\colon \beta = 0$ versus $H_1\colon \beta \neq 0$. With the given information, we can compute the F test statistic as: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{((-0.0161)^2)/1}{(1-(-0.0161^2)/(250-1-1)} = 0.06$. Looking at the F table, we see that for $\nu_1 = k = 1$ and $\nu_2 = n - k - 1 = 250 - 1 - 1 = 248$ the critical value for $\alpha = 0.10$ is 2.71. Hence the P-value for our test statistic is well above 0.10, which means that our model is not even close to being statistically significant overall at even a 10% significance level.

(5) (a) No. The reason is that (n – k – 1) is always less than (n – 1) whenever even one explanatory (x) variable is included.

(b) It is possible that the Adjusted R^2 could be negative. If you only included explanatory variables that are completely irrelevant a slightly negative value of the Adjusted R^2 is not implausible. The idea is that the Adjusted R^2 adjusts for the fact that even irrelevant variables will not have a zero correlation with the dependent (y) variable because of pure chance. However, if they have a below average correlation (again due to pure chance) for irrelevant variables, then the adjusted R^2 would be negative. The interpretation of a negative Adjusted R^2 would simply be that you had included only entirely irrelevant variables that are not explaining any of the variation in y. The regular R^2 must be between 0 and 1 (this can be mathematically proven).

(c) These two measures of model fit will be most different when you have included many explanatory variables that are entirely irrelevant. These will have spurious correlations that will drive the R^2 up as more and more are included, but these will not increase the adjusted R^2. This difference diminishes as the sample size grows. So, a second condition for a large difference would be a small sample size such that the degrees of freedom correction has an impact.

(6) The coefficient on x9 is statistically significant at a 5% significance level (t test). However the model overall is NOT statistically significant (F test).

(7) (a) $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \ldots, \beta_{20} = 0$. We know this because the data generating process is: $y, x_1, x_2, \ldots, x_{20}$ are independently drawn from N(0,1). If all of these variables are independently drawn from a Standard Normal distribution then in expectation there is no relationship between the variables. Hence the expected value of the slope and intercept coefficient estimates is 0.

(b) Yes. Let's start with the first row (num_sig90), which reports the number of statistically significant coefficients at a 10% significance level. We see that on average in our 10,000 estimated regressions, there were 2.0195 statistically significant slope estimates. Given that there are 20 slopes estimated, we would expect that by pure chance 10% would be significant. Recall, when we choose a significance level of 10% we are accepting the fact that 10% of the time we will make a Type I error: reject a true null hypothesis. In this simulation the null hypothesis is true: the really is no relationship between Y and X$_1$ – X$_{20}$ because they are just independent draws from the Standard Normal distribution. Hence, by chance, 10% of the time we will get a statistically significant slope estimate. Since, 0.10*20 = 2 our result of 2.0195 is what we would expect. Of course, 2.0195 is not exactly 2 but we only did 10,000 simulation draws, which means there will be a little simulation error. So the result is what we would expect. For the second row (num_sig95), which reports the number of statistically significant coefficients at a 5% significance level, we again get what we'd expect 1.0076 $\approx$ 1 = 0.05*20. For the third row (num_sig99), which reports the number of statistically significant coefficients at a 1% significance level, we again get what we'd expect 0.2033 $\approx$ 0.2 = 0.01*20.

For the fourth row (joint_sig90), which reports whether or not the regression model is jointly significant at the 10% level (overall a statistically significant regression model), we get what we'd expect 0.1051 $\approx$ 0.10. With a 10% significance level we would expect to make a Type I error 10% of the time: reject a true null hypothesis. In this case the null hypothesis is true: in fact, all of the beta parameters are zero. However, just due to pure chance for some of the samples we reject this null hypothesis and find that the model is statistically significant. Given that we recorded a model as statistically significant with a 1 and not statistically significant with a 0, if we take the average of this variable we have the fraction of the time the model is found to be statistically significant: 0.1051. For the fifth row (joint_sig95), which reports whether or not the regression model is jointly significant at the 5% level (overall a statistically significant regression model), we get what we'd expect $0.0562 \approx 0.05$. For the sixth row (joint_sig99), which reports whether or not the regression model is jointly significant at the 1% level (overall a statistically significant regression model), we get what we'd expect $0.0117 \approx 0.01$. Note, these numbers are not exactly equal to the significance level, but rather approximately ($\approx$) equal, because of simulation error. By doing this simulation much more than 10,000 times we could drive the simulation error to zero. Given the availability of inexpensive and fast computing power, Monte Carlo simulations are being used more and more by researchers. For real research many, many simulation draws are used to drive that simulation error to zero.

(c) 5,937 (=0.5937*10000). The relevant test statistic is the t test statistic. n = 100, α = 0.05, and k = 20. The degrees of freedom ν = 100 – 20 – 1 = 79. The test for statistical significance is a two-sided test so the critical value is: t = 1.99 (from statistical table). Hence infer statistical significance if t test statistic is less than -1.99 or greater than 1.99.

(d) 562 (=0.0562*10000). The relevant test statistic is the F test statistic. n = 100, α = 0.05, and k = 20. The numerator degrees of freedom ν_1 = 20 and the denominator degrees of freedom ν_2 = 100 – 20 – 1 = 79. The critical value is 1.7 (from statistical table). Hence infer statistical significance if the F test statistic is greater than 1.7.

(e) The answers are very different because the statistical tests are different. In part **(c)** we are doing 20 t tests for each regression. Hence we have 20 chances to make a Type 1 error: Reject a true null hypothesis (recall H_0: β_k = 0). In part **(d)** we are doing 1 F test for each regression. Hence the probability of making a Type 1 error is exactly 0.05.

(f)

E[num_sig90] = n*p = 20*0.10 = 2
E[num_sig95] = n*p = 20*0.05 = 1
E[num_sig99] = n*p = 20*0.01 = 0.2

E[joint_sig90] = probability of Type I error*1 + (1 - probability of Type I error)*0 = 0.10
E[joint_sig95] = probability of Type I error*1 + (1 - probability of Type I error)*0 = 0.05
E[joint_sig99] = probability of Type I error*1 + (1 - probability of Type I error)*0 = 0.01

E[any_sig90] = (P(1 slope est. is significant) + P(2 slope est. are significant) + P(3 slope est. are significant) + P(4 slope est. are significant) + P(5 slope est. are significant) + …)*1 + P(0 slope est. are significant)*0

Of course, it is quicker to use the complement rule to find E[any_sig90] = (1 – P(0 slope est. are significant))*1

To find probabilities, recognize that this is a Binomial experiment. Each explanatory variable is independently drawn so whether any particular coefficient is statistically significant is independent of the other coefficients.

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 0,1,2,\dots,n$$

$p(0) = \frac{20!}{0!(20-0)!} 0.10^0 (1-0.10)^{20-0} = 0.1216$; Hence, E[any_sig90] = 1 – 0.1216 = 0.8784

$p(0) = \frac{20!}{0!(20-0)!} 0.05^0 (1-0.05)^{20-0} = 0.3585$; Hence, E[any_sig95] = 1 – 0.3585 = 0.6415

$p(0) = \frac{20!}{0!(20-0)!} 0.01^0 (1-0.01)^{20-0} = 0.8179$; Hence, E[any_sig99] = 1 – 0.8179 = 0.1821

The Monte Carlo results are similar but not exactly the same as the analytical results we just found because we only repeated the Monte Carlo Simulation 10,000 times. There is simulation noise. If we were to do the Monte Carlo Simulation 100,000,000 times then the simulation noise would be very close to zero and we'd get simulation results that were extremely close to the exact analytical results.