Solution for test Exam April 11, 2019

Exercise 1:

Q1. We recall that the statistical model (i.e. all its assumptions) defines a *set* of hypothetical joint data densities. That is, for each value that the parameter $(\beta_1, \beta_2, \beta_3, \sigma^2)$ may take, there is such a potential or hypothetical data density. The model is correctly specified if one of these densities coincides with the DGP which is the density that we view as having generated the data **(LS2, Slide 12, 13)**.

Q2: We can estimate and infer about $\beta_1, \beta_2, \beta_3$ and σ^2 based only on the conditional model of Y_i given $X_{2,i}, X_{3,i}$. I.e. we do not have to specify the distribution of these. We would say that conditional modeling (conditional on $X_{2,i}, X_{3,i}$) is valid. **(see e.g. HN4, HN10, LS4, LS7 and LS8)**

Q3: It measures the partial derivative, i.e. $\frac{\partial E[Y_i|X_{2,i},X_{3,i}]}{\partial X_{2,i}} = \beta_2$, assuming that $X_{2,i}$ does not vary with $X_{3,i}$. The interpretation is that when comparing two individuals that differ in the value of $X_{2,i}$ by one unit while having the same values of X_{3i}, the conditional expectations of Y_i differ by β_2. In other words this is the average difference between such two individuals **(see HN §7.1, LS5 and LS6)**. Note that, this does not imply a causal effect, i.e. that the difference is *caused or due to* the difference in $X_{2,i}$.

Q4: Setting $\beta_3 = 1$ you have two possibilities. You can either minimize the SSD wrt. β_1 and β_2 where $\beta_3 = 1$ is inserted. I.e. this is eq. **7.2.3. on p. 100**, with $\beta_3 = 1$ inserted. Or, much easier you can realize that when $\beta_3 = 1$, the model reduces to a two-variable model (HN5) but where the regressand is not Y_i but $Z_i \equiv Y_i - X_{3,i}$. That is, $Z_i = \beta_1 + \beta_2 X_{2,i} + u_i$. Then you can just use the formulas from **§5.2 (eq. 5.2.1)** replacing Y_i by Z_i and X_i by $X_{2,i}$ etc. So you get $\beta_1 = \overline{Z} - \widehat{\beta}_2 \overline{X}_2$ and $\widehat{\beta}_2 = \frac{\sum_{i=1}^n Z_i (X_{2,i} - \overline{X}_2)}{\sum_{i=1}^n (X_{2,i} - \overline{X}_2)^2}$.

Q5: **See §7.6.7**

Exercise 2:

Q1. Easy and most did fine.

Q2: The Jarque Bera test (JB) test for normality is $\chi^2(2)$, i.e. 2 degrees of freedom (**see § 9.2 and PS6, E2**). Hence, the critical value (95% quantile) is 5.99. As $3.3202 < 5.99$ we cannot reject the null of correct specification with respect to normality of the errors.

White's test statistic is 8.028658 which is less than the critical value 19.67514 (the 95% quantile in $\chi^2(11)$). So again we cannot reject the null of correct specification, now meaning homoschedastic errors (constant error variance). We can thus conclude that the model is well-specified. You may also say that we have not been able to show that it is mis-specified. Note it is just as fine to compute the p-values and instead of using critical values. See also the R script.

Q3: You need to compute either the exact or the approximate/asymptotic CI. See the R script. You also need to give the correct interpretation, which is that a 99% CI has a 99% chance of including the true value. Or you could say that in a large number of hypothetically repeated samples 99% of the corresponding CIs will include the true value **(see LS3 and e.g. E1, Q1 in PS3 and also p.82** for the one-variable model case).

Q4: See the R script.

Q5: See **PS4 E2** (I also wrote this on the blackboard in that connection)

Q6: See the R script.

Q7: If you calculate it exactly you get 0.01695 i.e. 1.695%. Using the approximation (just as fine) you get 0.05/3 =0.01667 i.e. 1.66%. **See PS7, E1 Q3**.

Q8: This is the general question as to what extent a regression parameter reflects a causal effect. See my answer to **Q3 in Exercise 1** above and also lecture slides.