AAEC 4804/5804G, STAT 4804: Fundamentals of Econometrics

Your Name Here

Spring 2025 – Homework #2

Instructions

This homework is intended to help you review the material covered in Lecture 3. There is a joint emphasis on both the theoretical and practical aspects of the material. **You are strongly encouraged to work with your classmates, but you must submit your own answers.**

Question 1: Partialling Out

(Adopted from Page 109, C1): A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put the infant at risk for contracting various illnesses. Since factors other than cigarette smoking that effect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

where *bwght* is birth weight, *cigs* is the number of cigarettes smoked per day during pregnancy, $faminc$ is family income, and u is the error term.

(i) What is the most likely sign for β_2? **Why?**

(ii) Do you think that *cigs* and $faminc$ are likely to be correlated? **Explain and be sure to discuss the sign/nature of this correlation.**

(iii) Now estimate the equation with and without $faminc$ using the `bwght` data from the `wooldridge` package. Report the results in a single table using the `stargazer` package. **Discuss** and **interpret** your results, focusing on whether adding $faminc$ to the model substantially changes the estimated effect of *cigs* on *bwght*.

(iv) Confirm the partialling out interpretation of the OLS estimates by explicitly doing it. That is, regress *cigs* on $faminc$ and save the residuals, $\hat{r}_1$. Then regress *bwght* on the $\hat{r}_1$ to obtain $\hat{\beta}_1$. Repeat the process by regressing $faminc$ on *cigs* and save the residuals, $\hat{r}_2$. Then regress *bwght* on $\hat{r}_2$ to obtain $\hat{\beta}_2$. **In both cases, be sure to include a constant term in the regression.**

Compare these estimates with the OLS estimates from the full model in part (iii).

Note:

a. Present your findings from this step along with those of the full model in a single table using the `stargazer` package.

b. Use the `omit` argument to remove the constant term and the `keep.stat` argument to omit all the model statistics except the number of observations. The `stargazer` vignette is a good place to start for help with this.

c. Use the `covariate.labels` argument to properly label the four (4) coefficients in the merged table. **Hint: You can rename the residuals from the auxiliary regressions to make the table more readable. How about `"$\\hat{r}_1$"`? This will get converted properly to math as long as you keep `type = "latex"` (p.s. This is the default option if not stated otherwise).**

Question 2: Omitted Variable Bias

For this question, we are focused on illustrating the impact of an omitted, and relevant, variable from the model.

Use the dataset in `wage2` for this problem. **As usual, ensure that your auxiliary regressions contain an intercept.**

(i) Run a simple regression of IQ on $educ$ to obtain the slope coefficient, say $\widetilde{\delta}_1$. **Report and interpret this coefficient.**

(ii) Run the simple regression of $log(wage)$ on $educ$ to obtain the slope coefficient, say $\widetilde{\beta}_1$. **Report and interpret this coefficient. I would like you to speak to percent changes, where appropriate.**

(iii) Run the multiple regression of `log(wage)` on $educ$ and IQ to obtain the slope coefficients, say $\hat{\beta}_1$ and $\hat{\beta}_2$. **Report and interpret these coefficients.**

(iv) Verify that $\widetilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \widetilde{\delta}_1$. Do this using a simple logical test (i.e. `==`).

Question 3: Conducting OLS Estimation via Matrix Algebra

In class, we explored the matrix algebra representation of the OLS estimator. In this question, we will dive in deeper.

For the questions that follow, we will use the `gpa1` dataset from the `wooldridge` package. Our model of interest is:

$$colGPA = X\beta + u, \quad X = \{\mathbf{1}, hsGPA, ACT\} \tag{1}$$

Recall, you can use the `help()` function **in your R console** to learn more about the dataset and aid your interpretation. **Unless explicitly stated, store but do not report the result.**

(i) Using the `nrow()` function, determine the number of observations in the dataset. **Store this result in a variable called `N`.**

(ii) Create a dataframe that contains a column of ones (call it `Intercept`), the `hsGPA` variable, and the `ACT` variable. Pipe the dataframe to the `as.matrix()` function to convert to a `matrix` for use later on. **Ensure that ALL columns have appropriate names then store the result in a variable called `X`.**

(iii) Create a vector `y` that contains the dependent variable. **Store this as `y`.**

(iv) Using the matrix algebra representation of the OLS estimator, estimate the coefficients (β) of the model. **Store the result in a variable called `beta_hat`.**

(v) Using the identity in Equation (1), calculate the residuals, $\hat{u}$. **Store them in a variable called `resids`.**

(vi) Using the formula from our class notes, calculate $\hat{\sigma}^2$. **Store the result in a variable called `sigma_hat_sq`. I would love to see you use the `crossprod()` and `ncol()` functions here in your calculations.**

(vii) **Calculate** the Residual Standard Errors (RSE) and the R^2 of the model. Recall that $R^2 = 1 - \frac{SSR}{SST}$, where $SSR = u'u$ and $SST = \tilde{y}'\tilde{y}$ where $\tilde{y} = y - \bar{y}$. **Store the results in variables called `RSE` and `R2`, respectively. Report both.**

(viii) Estimate the variance-covariance matrix of the OLS estimator. **Hint: If you get a `non-conformable arguments` error, chances are you will need to convert the `sigma_hat_sq` to a numeric value using the `as.numeric()` function.**

Store the result in a variable called `vcov_beta_hat`.

(ix) If your calculations are correct, you will notice that the variance-covariance matrix is symmetric: the upper and lower triangular elements are the same such that `cov(x,y) = cov(y,x)`. Armed with the knowledge that the standard errors (SE) are the square root of the diagonal elements (the variances, which should all be positive), compute and report the SEs of the OLS estimator. **Store the result in a variable called `se_beta_hat`. Hint: You can use the `diag()` function to extract only the diagonal elements of a matrix.**

(x) We might be curious about the t-statistics of the OLS estimator. We will use this extensively in our lecture on Inference.

Using the formula $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$, calculate and report the t-statistics for each of the coefficients. **Store the result in a variable called `t_stats`.**

(xi) Combine your coefficients, SE, and `t_stats` into a single dataframe. Call the columns `Estimate`, `Std.Error`, and `tvalue`, respectively. **You can store this dataframe if you please but I would like you to report the results.**

(xii) Finally, using the `lm()` function, estimate the model and report the model summary in order to validate your manual calculations. Do the variables in (xi), your R^2 and RSE, from (vii) match your manual calculations?

@Gabe & Isaac, Now lean back and enjoy the feeling of satisfaction that comes from knowing you can do this by hand! Isn't that just the best?

Question 4: Adding Irrelevant Variables – Monte Carlo Simulation

This exercise should highlight the impact of adding irrelevant variables to the model. The main idea is to compare the sampling distribution of the OLS estimators when the model is correctly specified versus when it is incorrectly specified. We will vary the sample size to see how this affects the results.

Consider the following data generating process (DGP):

$$y_i = 10 + 2x_i + u_i,\ u_i \sim N(0,3), i = 1, 2, \ldots, n.$$

Also assume that x_i is generated as

$$x_i = 5 + 5\nu_i$$

where ν_i is generated randomly (but only once, i.e. **it is fixed in repeated samples**) from a uniform distribution with a [-10,10] support.

(i) Consider a sample size of $n = 25$. Perform 10,000 Monte Carlo simulations where you estimate both the correct model of $y_i = \beta_0 + \beta_1 x_i + u_i$, and the incorrect model $y_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_i + \widetilde{\beta}_2 x_i^2 + e_i$.

- Report the mean and variances of the sampling distribution for the slope parameters β_1, $\widehat{\widetilde{\beta}}_1$, and $\widehat{\widetilde{\beta}}_2$.
- Report the histograms (with (i) density plots overlaid, and (ii) a vertical line indicating the mean of the simulations) for each of these three(3) sampling parameters. **Hint: This should be easy to achieve with a simple Google search (I have found the `R Chart` website particularly helpful). In short, your density plot should be that of a normal (Gaussian) distribution with the same mean and variance as the sampling distribution.**
- Briefly discuss the differences/similarities between the sampling distributions of $\hat{\beta}_1$, and $\widehat{\widetilde{\beta}}_1$.

(ii) Repeat the above exercise for a sample size of $n = 2500$.