

Before you start, make sure you have installed and loaded these packages: haven, ggplot2, dplyr, tidyverse, stargazer, AER, MASS, and car

### Question 1: Frisch-Waugh-Lovell Theorem

In this question you will apply the FWL theorem to estimate a regression coefficient using the `birthweight_smoking` dataset.

- Regress the variable 'birthweight' on the the variables 'smoker', 'unmarried', and 'nprevist'. Save the residuals as `resids_1`
- Regress the variable 'drinks' on the variables 'smoker', 'unmarried', and 'nprevist'. Save the residuals as `resids_2`
- Use your results from (a) and (b) to estimate the coefficient on 'drinks' in the regression of 'birthweight' on 'drinks', 'smoker', 'unmarried', and 'nprevist'. Note: You must use the results from (a) and (b), and the Frisch-Waugh theorem to answer this part of the question. Compare the coefficient you obtained using the Frisch-Waugh theorem to that when estimating the full model using the `lm()` function.
- Regress 'birthweight' on 'drinks', 'smoker', 'unmarried', and 'nprevist'. Report the results of the regression in a table (using the `stargazer()` function)

### Question 2: Dummy Variables, Interaction Terms

In this question you will need to use the `CollegeDistance` dataset contained in the `AER` package. You must install and load the `AER` package to access the dataset.

*# After installing and loading the AER package, you can access the data using this command:*

```
data("CollegeDistance")
```

- Create a dummy variable called 'hispanic' that is equal to 1 if 'ethnicity' = "hispanic", and 0 otherwise. Create a dummy variable called 'afam' that is equal to 1 if 'ethnicity' = "afam", and 0 otherwise. Create a dummy variable called 'male' that is equal to 1 if 'gender' = "male", and 0 otherwise.
- Use the dummy variables you created in (a) to run a regression of 'education' on 'score', 'afam', 'hispanic', 'male', and 'distance'. Report the results of this regression in a table using the `stargazer()` function.
- Suppose you believe the effect of score on educational attainment differs based on gender. Adjust the regression in part (b) to include an interaction term between 'male' and 'score'. Report the results of this new regression in a table using the `stargazer()` function.

### Question 3: Joint Hypothesis Testing and $R^2$

In this question you will need to use the `MASchools` dataset contained in the `AER` package. You must install and load the `AER` package to access the dataset.

*# After installing and loading the AER package, you can access the data using this command:*

```
data("MASchools")
```

- Create a new data frame called 'MASchools\_subset' containing only the 'expreg', 'stratio', 'income', 'score8', and 'english' columns. Drop any observations with an NA value from 'MASchools\_subset'. You can access the help file for the `MASchools` dataset using the `?MASchools` command
- Regress 'score8' on 'expreg', 'stratio', 'income', and 'english' using observations from `MASchools_subset`.
- Calculate the SST (Total Sum of Squares) of the regression estimated in (b), where  $SST = \sum_{i=1}^N (y_i - \bar{y})^2$ , save the result. Calculate the SSR (Residual Sum of Squares) of the regression estimated in (b), where  $SSR = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ . Save the result. Hint: use the `residuals()` function

- (d) Use your results from (c) to calculate the  $R^2$  of the regression in (b). You must write the calculation you used as R code. You cannot directly report the  $R^2$  calculated automatically by R.
- (e) Suppose you wish to test if the coefficients on 'stratio' and 'income' are jointly significant, ie.  $H_0 : \beta_{stratio} = \beta_{income} = 0$  vs  $H_A$  : at least one of the coefficients is non-zero. Estimate the appropriate restricted model and save the  $R^2$ .
- (f) Use the  $R^2$  from the regressions in (b) and (e) to calculate the F-statistic for the joint test described in part (e). At the 5% significance level, would you reject the null hypothesis that the coefficients on 'stratio' and 'income' are jointly zero? (Note: be careful with BIDMAS when answering this question)

#### Question 4: Probit, Logit, Linear Probability Model

In this question you will need to use the Boston dataset contained in the MASS package. You must install and load the MASS package to access the dataset. We will use the Boston dataset to estimate the probability of a town having a high crime rate.

*# After installing and loading the MASS package, you can access the data using this command:*

```
data("Boston") # loads the dataset
glimpse("Boston") # will show you the structure of the dataset
```

- (a) Find the median value of the 'crim' variable. Hint: use the median() function. Create a dummy variable called 'HighCrime', where  $HighCrime_i = 1$  if  $crim_i >$  median of 'crim', and 0 otherwise.
- (b) Estimate a linear probability model in which you regress 'HighCrime' on 'indus', 'dis', 'tax', and 'ptratio'. What is the predicted probability of 'HighCrime' for a town with 'indus' = 10, 'dis' = 4, 'tax' = 330, and 'ptratio' = 18?
- (c) How does crime rate vary with pupil-teacher ratio, distance to employment centers, etc. ? Estimate a probit model using 'HighCrime' as the dependent variable and 'indus', 'dis', 'tax', and 'ptratio' as explanatory variables. Hint: Use the glm() function. Report the results of this regression in a table using the stargazer() function
- (d) Estimate the same model as in (c), but now use logit instead of probit. In one or two sentences, briefly compare the coefficients estimated in the logit and probit models estimated in (c) and (d).
- (e) Use the logit model estimated in (d) to find the predicted probability that 'HighCrime' = 1, for a town with 'indus' = 10, 'dis' = 4, 'tax' = 330, and 'ptratio' = 18?. What is the predicted probability using the probit model estimated in (c)? (Make sure you report the probability, not the Z-Score)