Prof. James MacKinnon *QUEEN'S UNIVERSITY* *Department of Economics*

Economics 882 Winter, 2025

Assignment 1

Due: February 4, 2025

This assignment uses the file `men2015b.csv`, which is in the data directory on the class website. It contains 32,437 observations on four variables. They are `age` (in years, from 21 to 80), `educ` (education level, which takes five values from 1 to 5), `marry` (1 if the person is married), and `earn`, weekly earnings in dollars. The data come from the Current Population Survey and are for men in 2015.

1. Create `logearn`, the natural logarithm of weekly earnings. Then estimate one or more linear regression models to explain `logearn`. The explanatory variables may be whatever functions of `age`, `educ`, and `marry` seem appropriate to you. Choose what you think is the best model, and explain why you chose it. [20]

2. Using the model you chose in question 1, predict `logearn` for 20 hypothetical individuals with ages 30, 40, 50, 60, or 70, education levels of 4 (4-year degree) or 5 (graduate degree), and marital status either 0 or 1. For which of the 20 hypothetical individuals are predicted earnings maximized?

In addition to reporting the 20 predicted values of `logearn`, report a 99% prediction interval for each of them. [15]

3. Estimate a KNN model using the `knn.reg` function and all the data as training data, without scaling the data. Using the leave-one-out predictive R^2 as a criterion, find what seems to be the best value of K. This value will turn out to be considerably larger than the ones we have encountered in the book and the lectures. Why do you think it is so large in this case? [20]

4. Using the preferred KNN model from question 3, predict `logearn` for the hypothetical individuals of question 2. Discuss how your predictions differ from those for the linear regression model. [10]

5. Estimate another KNN model. This time, scale the data in whatever way seems appropriate. Using the leave-one-out predictive R^2 as a criterion, find what seems to be the best value of K. How does this value compare with the one from question 3? What do you think explains the difference? [20]

6. Using the preferred KNN model from question 5, predict `logearn` for the hypothetical individuals of question 2. Discuss how your predictions differ from the ones you obtained in question 4. [15]