EXERCISE: CONTROLLING MODEL COMPLEXITY

In this exercise you will explore models of various complexity for the relation between education and log wages in the AK data.

You will explore linear regression models with `educ` and `yob` as explanatory variables. The point of including `yob` (year of birth) is that wages for any given individual tend to rise over time and that educational attainment might have tended to increase over cohorts. Without `yob` in the regression, the general rise in wages over time, through its correlation with the general rise in educational attainment, might lead to a spurious positive relation between `educ` and `logwage`.

At one extreme, you will fit a simple linear model, which is what Angrist and Krueger did. (Though they also used instrumental variables and the `qob` variable to account for a possible bias we will ignore.) At the other extreme, you will fit a model with each year of schooling and each birth year a separate dommy variable. This allows the effect of schooling (and age) to be positive or negative, increasing or decreasing. In fact, the fully extreme model will also use all interaction terms between these variables — allowing each birth cohort to have its own, completely unrestricted, pattern of education effects.

But there are reasonable models between these extremes. We might think that 8th grade graduation, highschool (12th year) graduation, four-year college (16 year) graduation, or the top-coded 20-year education class, which probably distinguishes PhD from Masters degrees, might be different from other yearly education milestones. They could be different just additively for the milestone year, or persistently, for that year and all subsequent years, or both.

To be precise, you will consider at least these sets of explanatory variables (in every case also including an intercept):

(1) `educ` and `yob` as numeric variables;
(2) `educ` and `yob` as sets of dummy variables. (In R, this can be done by `logwage` ∼ **`as.factor`**`(educ) + `**`as.factor`**`(yob)`.)
(3) `educ` and `yob` as sets of dummies, along with all their interactions (In R, `logwage` ∼ `educ *yob`.) This will give you 210 variables in the regression.
(4) `educ` and `yob` as numeric variables, plus dummies for `educ` at the values 8, 12,16 and 20.
(5) `educ` and `yob` as numeric variables, plus dummies for `educ >= 8`, `educ >= 12`, `educ >= 16`, and `educ == 20`.
(6) All the variables in items 4 and 5.

Then the problem will be to decide which model is best. A traditional tool for deciding whether a group of variables adds explanatory power to a regression is the F test for a linear restriction. The general form tests H0: $R\beta = \gamma$ against the alternative that this restriction does not hold. For testing whether a set of variables can be excluded, R is a matrix with a

Date: December 3, 2020.

single 1 in each row and γ is zero. We will discuss in lecture the F distribution and ways of constructing the F test statistic, but for this exercise, you need only to know there are two ways to construct an F test of an exclusion restriction.

(a) In R, after estimatng a regression model via **`lm()`** e.g. by

```
lmout <- with(akdataf,
      lm(logwage ~ as.factor(educ) * as.factor(yob))
```

run `Anova(lmout)`. Note the capital A in the function name. This function is in the package **car**. It automatically calculates F tests on groups of dummy variables. The **car** package is big, but can be downloaded in binary form for Windows. If you have trouble getting it or setting it up, an alternative is to use R's built-in **`anova()`** function. It tests blocks incrementally, however, so it gives the test you are interested in (excluding one block while all others are left free) only when the block you are testing is the last block entered in the call to **`lm()`**.

(b) Instead, or when testing groups of variables not generated by R factors, you can apply **`lm()`** to the smaller model, obtain its residual sum of squares (RSSR), then estimate the larger model, and obtain its residual sum of squares (RSSU). Then

$$\frac{RSSR - RSSU}{RSSU}\left(\frac{N-k_u}{k_u-k_r}\right) \qquad (*)$$

will have the $F(k_u - k_r, N - k_u)$ distribution under H0, where k_u is the number of variables in the larger model, and k_r is the number in the smaller model and N is sample size. One rejects for values of the statistic in the upper tail of its distribution. The first posted version of these notes showed a formula for the F test statistic that was right only for the $k_r = 0$ case.

As we have discussed, frequentist tests at fixed conventional 5% or 1% level are inconsistent. To get consistency, one needs to follow some rule for requiring smaller significance levels as sample size increases. This problem will be reflected in this exercise, in that these tests at conventional levels will tend to reject most of the H0's you will consider, suggesting that a large model is needed. Models arrived at this way tend to be "overfitted", meaning that they predict less well by an RMSE criterion than smaller models.

Another problem is that in comparing many models, as you do here, the logic of the frequentist tests is undermined. The tests' theory assumes there is a single H0 and you will accept or reject. But a second test, after you have accepted or rejected another model, will have an outcome related to the first test's outcome, so the usual significance level will not apply. There is a literature on how to handle this issue from a frequentist perspective, but in applied work on model selection this issue is usually ignored.

And a final problem for the frequentist F tests is that the theory underlying them only works when we compare a model with a restricted version of the same model. I.e., it works only on "nested" models. There is no simple frequentist test to compare two models like 4 and 5 above, which are not nested.

Comparing models by Bayesian methods is quite possible, and we will discuss how to do that in the remaining lectures. However there is a kind of quasi-Bayesian shortcut that

is widely applied, called the BIC, for "Bayesian Informatin Criterion". It is also sometimes called the Schwarz criterion, after the statistician who suggested it. The BIC, as we shall see in lecture, does provide consistent estimation of model size when the list of models includes the truth. In a regression model like those you are considering here, it rejects the restrictions in H0 when the F statistic exceeds $\log N$. In this sample, $\log N = 12.7$. Contrast this with the standard F test at the 5% or 1% level, which, for group of variables of size $k = 10$ (like the `yob` dummies in the AKdata), rejects for any value of the statistic greater than 1.83 or 2.32, respectively.

The Bayesian approach avoids the problems listed above for the frequentist testing in model selection. It is consistent. The results of multiple tests can be interpreted directly as providing information about the shape of the likelihood function. The BIC can be applied to compare non-nested models. In the form $(*)$, the BIC can be applied where RSSR and RSSU are just residual sums of squares for two non-nested models, with the choice depending on comparison of the statistic to $\log N$ and k interpreted as the difference in parameter count between the two models. When the two models have the same parameter count, BIC simply picks the one with lower sum of squared residuals.

So for this exercise, you are to try to arrive at the best model based on frequentist F tests at the 5% or 1% level. (It probably won't matter much which level you choose.) Then do the same thing using BIC.

Note on software: What you are asked to do here are standard calculations, so they can undoubtedly be done with Stata or other statistical software. You can also in principle do them in Matlab or Python without any statistical package add-ons, though programming the generation of dummy variables might be tedious.