Assignment 3

Question 1 The data set `fish` from the wooldridge package contains daily data spanning 97 days and includes data on prices and quantities of fish in a certain wholesale market. We will use the variables `lavgprc` (ln ave. prices), `wave2` and `wave3` (measures of ocean wave heights, indicating roughness of the oceans), and `ltotqty` (ln total quantity). We will also use the time index `t` (a sequence of integers from 1 to 97), and the day dummies `mon`, `tues`, `wed`, `thurs`. Use the following code to load up all the necessary libraries and data for this question, and to convert your data frame into a "tsibble":

```
```{r warning=FALSE, message=FALSE}
library(fpp3); library(car); library(sandwich); library(lmtest)
library(ivreg); library(wooldridge)
data(fish)
dat <- fish %>% as_tsibble(index=t)
```
```

(a) Regress `laveprc` on the time trend `t` and the day dummies `mon`, `tues`, `wed` and `thurs` (take friday as the base). What does the coefficient on `t` say. Is there a "day-of-the-week" effect?

(b) Drop the day dummies and add the variables `wave2` and `wave3`. What happens to the coefficient on `t`? Why do you think the coefficient on `t` changed in this manner when `wave2` and `wave3` were added?

(c) Plot the ACF of the residuals. Is there any indication of autocorrelation in the residuals? Run a regression of the residuals against the lagged residuals. What is your estimate of the lag 1 autocorrelation of the residuals?

(d) Present the results of the model in (b) with HAC standard errors. Are there any important changes to the conclusions?

(e) Estimate the demand function for fish in this market by running a regression of `ltotqty` on `laveprc` and the day dummies, first using OLS and then using 2SLS, with `wave2` and `wave3` as instruments for `laveprc`. Report your results using HAC standard errors. Are there important differences in the OLS and 2SLS estimate of the price elasticity of demand. Why does it make sense to estimate the demand equation by 2SLS? Does it make sense to use `wave2` and `wave3` as instruments for `aveprc`?

Question 2 The data set `driving` is a state-level annual panel data set covering the 48 continental US states, and spanning the years 1980 to 2004. We will use the variables `bac08` (legal blood alcohol limit of 0.08), `bac10` (legal blood alcohol limit of 0.10), `perse` (licenses can be revoked without trial), `sbprim` (a primary seat belt law where officers can stop a vehicle and issue a ticket solely for not wearing a seat belt), `sbsecon` (a secondary seat

belt law where officers can issue a ticket for not wearing a seat belt only if the driver was stopped for some other violation, and subsequently found not to be wearing a seat belt), `perc14_24` (percent population aged 14 to 24), `unem` (unemployment rate), `vehicmilespc` (vehicle miles per capita), and `totfatrte` (total fatalities per 100,000 population). The variables `bac08`, `bac10`, `perse`, `sbprim` and `sbsecon` are indicator variables indicating the presence of such laws in that state in that year. If the law was enacted during a year, the fraction of the year during which the law was in effect is recorded. The state indicator is `state`. There is a `year` variable, and year dummies `d80` through to `d04`.

(a) Run a (pooled) regression of `totfatrte` on the year dummies `d81` through to `d04`, and interpret the results.

(b) Add the variables `bac08`, `bac10`, `perse`, `sbprim`, `sbsecon`, `perc14_24`, `unem` and `vehicmilespc` to the pooled regression. Report estimates first using default standard errors, then using panel-robust standard errors, clustering by group. Are there any surprising results? Retain the year dummies but to save space, omit their estimates from your output, e.g., you can use `coeftest(modelname) %>% round(4) %>% tail(8)` to report the results in the default standard errors case.

(c) Re-estimate the regression using fixed-effects estimator, and report estimates first using default standard errors, then using panel-robust standard errors, clustering by group (again, omit the year dummies in the output). Are there important changes in the results between the default standard errors and the panel-robust standard errors? Are there important changes in the results compared with part (b)? Do you think the estimates using the fixed-effects estimators are more appropriate? Why or why not?

Question 3 The data set `smoke` from the `wooldridge` library contains data on the smoking habits and other characteristics of 807 people. We will use the variables `cigs` (number of cigarettes per day), `lcigpric` (log cigarette prices), `lincome` (log income), `white`, `educ`, `age`, `agesq`.

(a) Count the number of observations for each value of `cigs` and plot these counts against `cigs`. How many people in the sample claim to not smoke? Are there any other unique features in the counts?

(b) Estimate Poisson regressions of `cigs` on `lcigpric`, `lincome`, `white`, `educ`, `age` and `agesq`, using MLE and QMLE. Use `summary(modelname)` to show the output for each, where `modelname` is the name of the estimated model. What are the estimated price and income elasticities in both cases? Are they statistically significant? In the output, the dispersion parameter (σ^2 on slide 45) is given. Do the results suggest that QMLE approach is more appropriate than MLE, which assumes $\sigma^2 = 1$?

Question 4 (Part of past examination question)

(a) Suppose that the three variables X, Y and Z are related in population according to the equations

$$Y = \beta_0 + \beta_1 X + \epsilon\,,$$
$$X = Y + Z + v$$

where ϵ and v are independent zero-mean noise terms. Suppose that you have an iid sample $\{X_i, Y_i, Z_i\}_{i=1}^n$ of the three variables. Explain in detail why estimating β_1 by regressing Y_i on X_i using OLS leads to an inconsistent estimator for β_1. (You are not required to derive the expression to which the OLS estimator converges, but you must show that the required conditions for consistency are not met).

(b) Find a consistent estimator for β_1. Be sure to show that your proposed estimator is consistent, and state all assumptions that are required for your proof to hold.

Question 5 (Part of past examination question)

Consider the problem of deciding whether or not a variable X_{ik} should be included or excluded in a regression equation. That is, consider the problem of choosing between the models

$$\text{[A]} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

$$\text{[B]} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{k-1} X_{i,k-1} + \beta_k X_{ik} + \epsilon_i \,, \;\; i = 1, 2, ..., n.$$

One approach is to evaluate the statistical significance of $\hat{\beta}_k^{ols}$ in [B] using either the t-test or the F-test. Another approach is to include X_{ik} if doing so increases the adjusted-R^2.

(a) The F-statistic for evaluating if $\hat{\beta}_k^{ols}$ in [B] is statistically significant is

$$F = \frac{R_B^2 - R_A^2}{(1 - R_B^2)/(n - k - 1)}$$

where R_A^2 and R_B^2 are the R^2s from OLS estimation of equations [A] and [B] respectively. Show that this statistic is the square of the t-statistic for evaluating the statistical significance of $\hat{\beta}_k^{ols}$ in [B]. You may use the fact that for any given set of restrictions $\mathcal{R}\beta = r$, the difference between the restricted OLS sum of squared residuals, SSR_{res}, and the unrestricted OLS sum of squared residuals, SSR_{unres}, is

$$SSR_{res} - SSR_{unres} = (\mathcal{R}\hat{\beta}^{ols} - r)^{\mathrm{T}} [\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}]^{-1} (\mathcal{R}\hat{\beta}_{ols} - r)$$

where $\hat{\beta}^{ols}$ is the unrestricted OLS estimator of $\beta = \begin{bmatrix}\beta_0 & \beta_1 & ... & \beta_k\end{bmatrix}^{\mathrm{T}}$ in [B].

Question 6 (Part of past exam question)

(a) Consider the linear regression model

$$y = X\beta + \epsilon$$

where y and ϵ are $n \times 1$ and X is $n \times K$ with full column rank. Suppose ϵ satisfies the conditions $E(\epsilon \mid X) = 0$ and $E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2\Omega$ where Ω is a known matrix of constants.

i. The OLS estimator for β is $\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. Show that its variance-covariance matrix is

$$Var(\hat{\beta}^{ols} \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}\Omega X)(X^{\mathrm{T}}X)^{-1} \,.$$

ii. Under what conditions will the OLS estimator for β be best linear unbiased? (Just state the required conditions, you don't have to prove the result).

(b) Suppose, in the vain hope of getting more precise estimates, you "extend" the data set by including the negative of every observation. That is, you estimate (using OLS) the regression

$$\tilde{y} = \widetilde{X}\beta + \tilde{\epsilon}$$

where

$$\tilde{y} = \begin{bmatrix} y \\ -y \end{bmatrix}, \ \widetilde{X} = \begin{bmatrix} X \\ -X \end{bmatrix}, \ \text{and} \ \ \tilde{\epsilon} = \begin{bmatrix} \epsilon \\ -\epsilon \end{bmatrix}.$$

Show that doing so gives you exactly the same estimator and variance-covariance matrix as in part (a).

(c) Show that the rule "include X_{ik} if doing so increases the adjusted-R^2" is equivalent to the rule "include X_{ik} if the t-statistic for testing the hypothesis $H_0 : \beta_k = 0$ vs $H_A : \beta_k \neq 0$ is greater than one in absolute value". Explain why this means that the adjusted-R^2 rule is far more liberal than the usual t-test for deciding whether or not to include a variable.