Andrew ElHabr
ECON 7023
2/6/18

Homework 2

Note: I did not have access to STATA for this homework, so I did all the coding in R. I am still working on obtaining access to STATA, which I do not automatically have because I am an ISyE student. All of the code is attached at the end.

Wooldridge Panel Data Book

4.11)

(a) The estimated equation when using both IQ and KWW as proxies for ability is

$$\begin{aligned}\widehat{\log(wage)} = &\underbrace{5.176}_{(0.128)} + \underbrace{0.013}_{(0.003)} exper + \underbrace{0.011}_{(0.002)} tenure + \underbrace{0.192}_{(0.039)} married - \underbrace{0.082}_{(0.026)} south \\ &+ \underbrace{0.176}_{(0.027)} urban - \underbrace{0.130}_{(0.040)} black + \underbrace{0.050}_{(0.007)} educ + \underbrace{0.003}_{(0.001)} IQ + \underbrace{0.003}_{(0.002)} KWW.\end{aligned}$$

The return to education, i.e. the estimated coefficient for $educ$, is approximately 5% when using both proxies for ability, 6.5 % without any proxies for ability, an 5.4% when only using IQ as a proxy for ability. Thus, we obtain a lower estimate for return to education when using both proxies than without using a proxy and just using IQ as a proxy for ability. In all cases, education is statistically significant and the difference between all three cases is relatively minimal.

(b) The F-statistic when testing for joint significance of KWW and IQ is approximately 8.59, which gives us a p-value of 0.0002, indicating that the two proxies for ability are jointly significant.

(c) The estimated wage differential between blacks and nonblacks is approximately 13% (13% lower for blacks than nonblacks). This is a reduction from the 18.4% wage differential that is seen when no proxy for ability is used, but this differential is still significant and does not disappear.

(d) The estimated equation with the two interaction terms is

$$\begin{aligned}\widehat{\log(wage)} = &\underbrace{6.080}_{(0.561)} + \underbrace{0.012}_{(0.003)} exper + \underbrace{0.011}_{(0.002)} tenure + \underbrace{0.198}_{(0.039)} married - \underbrace{0.081}_{(0.026)} south \\ &+ \underbrace{0.178}_{(0.027)} urban - \underbrace{0.138}_{(0.04)} black + \underbrace{0.045}_{(0.008)} educ + \underbrace{0.005}_{(0.006)} IQ - \underbrace{0.025}_{(0.011)} KWW \\ &- \underbrace{0.0001}_{(0.0004)} educ(IQ - 100) + \underbrace{0.0022}_{(0.0008)} educ(KWW - 35.74439).\end{aligned}$$

The F-statistic when testing for joint significance of $educ(IQ - 100)$ and $educ(KWW - \overline{KWW})$ is approximately 4.19, which gives us a p-value of 0.015, indicating that the two proxies for ability are jointly significant at a 2% significance level. We also notice that the KWW interaction term by itself is statistically significant at a 5% significance level, and neither the IQ nor the IQ interaction terms are significant at a 5% significance level anymore. In this case, the return to education is approximately 4.5%, which is not significantly smaller than it was for the other cases. Finally, we notice that the return to education appears to increase for KWW above its mean value and slightly decrease for IQ above 100, although this decrease is not statistically significant.

4.13)

(a) The estimated equation is

$$\widehat{\log(crmrte_{87})} = -\underbrace{4.868}_{(0.432)} - \underbrace{0.724}_{(0.115)}\log(prbar) - \underbrace{0.473}_{(0.083)}\log(prbconv) + \underbrace{0.160}_{(0.206)}\log(prbpris) + \underbrace{0.076}_{(0.163)}\log(avgsen).$$

(b) With the inclusion of $\log(crmrate)$ from 1986 as an explanatory variable, the estimated equation is

$$\widehat{\log(crmrte_{87})} = -\underbrace{0.767}_{(0.313)} - \underbrace{0.185}_{(0.063)}\log(prbar) - \underbrace{0.039}_{(0.047)}\log(prbconv) - \underbrace{0.127}_{(0.099)}\log(prbpris) - \underbrace{0.152}_{(0.078)}\log(avgsen) + \underbrace{0.780}_{(0.045)}\log(crmrte_{86}).$$

The elasticities of crime rate in 1987 with respect to *prbar* and *prbconv* are much smaller now and only the elasticity with respect to *prbar* is statistically significant now. The signs of the elasticities of crime rate in 1987 with respect to *prbpris* and *avgsen* have switched with the inclusion of the lagged explanatory variable and the elasticity with respect to *avgsen* is now statistically significant for a 6% significance level. We also note that the elasticity with respect to the previous year's crime rate is large and statistically significant.

(c) The estimated equation for the model that includes all wage variables is

$$\begin{aligned}\widehat{\log(crmrte_{87})} = &-\underbrace{3.793}_{(1.957)} - \underbrace{0.173}_{(0.066)}\log(prbar) - \underbrace{0.068}_{(0.050)}\log(prbconv) - \underbrace{0.216}_{(0.102)}\log(prbpris)\\ &- \underbrace{0.196}_{(0.084)}\log(avgsen) + \underbrace{0.745}_{(0.053)}\log(crmrte_{86}) - \underbrace{0.285}_{(0.178)}\log(wcon) + \underbrace{0.064}_{(0.134)}\log(wtuc)\\ &+ \underbrace{0.254}_{(0.232)}\log(wtrd) - \underbrace{0.084}_{(0.196)}\log(wfir) + \underbrace{0.113}_{(0.085)}\log(wser) + \underbrace{0.099}_{(0.119)}\log(wmfg)\\ &+ \underbrace{0.336}_{(0.245)}\log(wfed) + \underbrace{0.039}_{(0.207)}\log(wsta) - \underbrace{0.037}_{(0.329)}\log(wloc).\end{aligned}$$

The F-statistic when testing for joint significance of all 9 wage variables is approximately 1.50, which gives us a p-value of 0.163, indicating that we fail to reject all of the wage variables are jointly insignificant even at a 15% significance level.

(d) The heteroskedasticity-robust F-statistic when testing for joint significance of all 9 wage variables is approximately 2.19, which gives us a p-value of 0.032, indicating that all of the wage variables are jointly significant even at a 4% significance level. Thus, when we do not assume homoskedasticity, we find that the additional wage variables are more likely to be jointly significant.

4.14)

(a) The estimated model is

$$\widehat{stndfnl} = -\underbrace{0.502}_{(0.196)} - \underbrace{0.290}_{(0.116)}\,frosh - \underbrace{0.118}_{(0.099)}\,soph + \underbrace{0.008}_{(0.002)}\,atndrte.$$

The estimated coefficient of *atndrte* indicates that for every one percent increase in lectures attended, the standardized final exam score increases by approximately 0.008 points, and that *atndrte* is significant for even a 0.03% significance level, so we can safely assume that *atndrte* affects *stndfnl*.

(b) I am not confident that these OLS estimates are estimating the causal effect of attendance because I think we have omitted variable bias. For example, the model does not include student ability or teacher quality, both of which I think must be controlled for when attempting to estimate the effect of attendance rate on final score.

(c) The estimated model with the proxies for student ability is

$$\widehat{stndfnl} = -\underbrace{3.297}_{(0.309)} - \underbrace{0.049}_{(0.108)} frosh - \underbrace{0.160}_{(0.090)} soph + \underbrace{0.005}_{(0.002)} atndrte + \underbrace{0.427}_{(0.082)} priGPA + \underbrace{0.084}_{(0.011)} ACT.$$

The effect of attendance rate decreases by almost a half when we estimate the model with the proxies (decreased from 0.008 to 0.005). More importantly, this effect is now only significant for a minimum of a 3% significance level. Thus, this effect's magnitude and significance have been reduced when we incorporate the proxies for student ability.

(d) The significance of $frosh$ decreases when we add the proxies (significant at 2% significance level to 65% significance level), and the significance of $soph$ increases when we add the proxies (significant at 24% significance level to 8% significance level). I believe that this indicates that the two proxies were much more highly correlated with the dummy variable $frosh$ rather than $soph$, which is expected since $priGPA$ and ACT are metrics that help define a freshman more than a sophomore. Since this is the case, $frosh$ is less useful when determining $stndfnl$ since we are able to get more useful information from the proxies that are correlated with it.

(e) The estimated model with the proxies and their respective squares for student ability is

$$\begin{aligned}\widehat{stndfnl} = \underbrace{1.385}_{(1.239)} - \underbrace{0.105}_{(0.107)} frosh - \underbrace{0.181}_{(0.089)} soph + \underbrace{0.006}_{(0.002)} atndrte - \underbrace{1.526}_{(0.474)} priGPA - \underbrace{0.112}_{(0.098)} ACT \\ + \underbrace{0.368}_{(0.089)} priGPA^2 + \underbrace{0.004}_{(0.002)} ACT^2.\end{aligned}$$

The effect of attendance rate increases from the last model we studied, but is still not as high in magnitude as the original endogenous model (increased from 0.005 to 0.006). This effect is now significant for a minimum of a 1% significance level. Thus, adding the squares of the proxies to the model does not change our previous understanding of the effect of attendance rate on standardize final exam score very much.

The F-statistic when testing for joint significance of the two squared terms is approximately 11.28, which gives us a p-value of 0.000015, indicating that the two terms are jointly significant at all reasonable significance levels.

(f) The estimated model when we add the square of $atndrte$ to the model is

$$\begin{aligned}\widehat{stndfnl} = \underbrace{1.394}_{(1.267)} - \underbrace{0.105}_{(0.107)} frosh - \underbrace{0.181}_{(0.089)} soph + \underbrace{0.006}_{(0.011)} atndrte - \underbrace{1.525}_{(0.476)} priGPA - \underbrace{0.112}_{(0.098)} ACT \\ + \underbrace{0.368}_{(0.089)} priGPA^2 + \underbrace{0.004}_{(0.002)} ACT^2 + \underbrace{0.000003}_{(0.000079)} atndrte^2.\end{aligned}$$

The standard error and t-statistic (0.036) of $atndrte^2$ indicate that there is no quadratic effect of attendance rate on standardized final exam score.

4.15)

(a) First, note that since we assume that $\mathbb{E}(x_j u) = 0$ for all $j = 1, 2, ..., K$, it must be that $\text{Cov}(\boldsymbol{x\beta}, u) = 0$.

Then,

$$\begin{aligned}\sigma_y^2 =&\text{Var}(y)\\ =&\text{Var}(\boldsymbol{x\beta}) + \text{Var}(u) + 2\text{Cov}(\boldsymbol{x\beta}, u)\\ =&\text{Var}(\boldsymbol{x\beta}) + \text{Var}(u)\\ =&\text{Var}(\boldsymbol{x\beta}) + \sigma_u^2.\end{aligned}$$

(b) This statement does not make much sense in most contexts because for it to be true, either all $\boldsymbol{x}$ are deterministic or $\boldsymbol{\beta} = 0$. This statement is also implicitly assuming that the variation of y_i does not depend on the choice of the explanatory variables, which is not true in almost any realistic context.

(c) If we use the facts that SSR/N is a consistent estimator of σ_u^2 and SST/N is a consistent estimator of σ_y^2, then

$$\begin{aligned}\text{plim}(R^2) =&\text{plim}\Big(1 - \frac{SSR}{SST}\Big)\\ =&\text{plim}\Big(1 - \frac{SSR/N}{SST/N}\Big)\\ =&1 - \text{plim}\Big(\frac{SSR/N}{SST/N}\Big)\\ =&1 - \frac{\text{plim}(SSR/N)}{\text{plim}(SST/N)}\\ =&1 - \frac{\sigma_u^2}{\sigma_y^2}\\ =&\rho^2.\end{aligned}$$

(d) This statement is not true because our proof in part (c) shows that R^2 is a consistent estimator the population R^2, and we made no assumptions about the conditional variance of u given $\boldsymbol{x}$. Therefore, no assumptions were made about homoskedasticity, and it must be that R^2 is still useful in yielding a consistent estimate of the population R^2.

5.1)

As the hint suggests, we partition the explanatory variables of (5.52) so that $\boldsymbol{x_1} = (\boldsymbol{z_1}, y_2)$ and $\boldsymbol{\beta_1} = (\boldsymbol{\delta_1'}, \alpha_1)$, and (5.52) reduces to

$$y_1 = \boldsymbol{x_1\beta_1} + \rho_1\hat{v}_2 + error.$$

Then, we find $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ by regressing $\boldsymbol{x_1}$ on $\hat{v}_2$, labeling the residuals as $\ddot{\boldsymbol{x}}_{\boldsymbol{1}}$, and then regressing y_1 on $\ddot{\boldsymbol{x}}_{\boldsymbol{1}}$. Note than when we regress $\boldsymbol{x_1}$ on $\hat{v}_2$, we can use the fact that $\boldsymbol{z_1}$ on $\hat{v}_2$ are orthogonal to see that the residual is simply $\boldsymbol{z_1}$. Then, we observe that when we regress y_2 on $\hat{v}_2$, we can use the fact that y_2 can be written as the sum of two orthogonal components, $\hat{y}_2$ and $\hat{v}_2$. Thus, when we regress y_2 on $\hat{v}_2$, the residuals must be $\hat{y}_2$.

In summary, when we regress $\boldsymbol{x_1} = (\boldsymbol{z_1}, y_2)$ on $\hat{v}_2$, the residuals are $\ddot{\boldsymbol{x}}_{\boldsymbol{1}} = (\boldsymbol{z_1}, \hat{y}_2)$. At this point, our partitioned regression method dictates that we regress y_1 on $\ddot{\boldsymbol{x}}_{\boldsymbol{1}} = (\boldsymbol{z_1}, \hat{y}_2)$ to find $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} = (\hat{\boldsymbol{\delta}}_{\boldsymbol{1}}', \hat{\alpha}_1)$, which is exactly what we do to find the 2SLS estimator, showing that the two estimators are the identical.

5.3)

(a) I would expect that *packs* would be correlated with other unobserved unhealthy habits, such as the average number of pints of beer drank per day during pregnancy, that can also affect the weight of a newborn.

(b) I would say that *cigprice* satisfies the assumption of relevancy for an instrument because there should be some significant relationship between the price of cigarettes and how many packs are smoked on average per day. For example, if the price is higher, then we would expect a lower amount of packs smoked, i.e. negative correlation between *packs* and *cigprice*.

However, it is more difficult to argue that this instrument is not correlated with the original error term. I don't think the price of cigarettes is correlated with the unobserved variable of average amount of alcohol drank, but the price of cigarettes include how much it is taxed which is usually correlated with the general health standards, which is included in the original error term. For this reason, I do not think this information satisfies the properties of a good instrumental variable for packs very well.

(c) Using OLS, the estimated equation is

$$\begin{aligned}\widehat{\log(bwght)} = &\underbrace{4.676}_{(0.022)} + \underbrace{0.026}_{(0.010)}\, male + \underbrace{0.015}_{(0.006)}\, parity + \underbrace{0.018}_{(0.006)} \log(faminc) \\ &- \underbrace{0.084}_{(0.017)}\, packs.\end{aligned}$$

Now, using 2SLS where *cigprice* is an instrument for packs, the estimated equation is

$$\begin{aligned}\widehat{\log(bwght)} = &\underbrace{4.468}_{(0.259)} + \underbrace{0.030}_{(0.018)}\, male - \underbrace{0.001}_{(0.022)}\, parity + \underbrace{0.064}_{(0.057)} \log(faminc) \\ &+ \underbrace{0.797}_{(1.086)}\, packs.\end{aligned}$$

The differences between the two estimates is significant. For the OLS estimation, an extra pack of cigarettes per day reduces baby weight by approximately 8.4% and is significant. On the other hand, for the 2SLS estimation, an extra pack of cigarettes per day increases baby weight by approximately 79.7% and is not significant. The direction of this effect is no longer intuitive, and its magnitude also seems to be exaggerated.

(d) The estimated equation for the reduced form for *packs* is

$$\begin{aligned}\widehat{packs} = &\underbrace{0.137}_{(0.104)} - \underbrace{0.005}_{(0.016)}\, male - \underbrace{0.018}_{(0.009)}\, parity - \underbrace{0.053}_{(0.009)} \log(faminc) \\ &+ \underbrace{0.0008}_{(0.0008)}\, cigprice.\end{aligned}$$

This estimation indicates that *cigprice* does not have a significant effect on the average number of packs smoked per day, telling us that *cigprice* is a weak instrument because it not very relevant. In addition, it has the opposite effect on the number of packs smoked per day than what we would expect. Both of these observations lead to the conclusion that the 2SLS estimation in part (c) is poor.

5.5)

Because we did not assume anything about the relationship between y_2 and q, they may be correlated, and thus y_2 and u_1 may be correlated, and we generally do not obtain consistent estimates of $\boldsymbol{\delta_1}, \alpha_1$ and $\boldsymbol{\psi_1}$. Thus, even if it is the case that $\text{Cov}(q, \boldsymbol{z_2}) = 0$, this method does not work since it is not valid to use $\hat{\boldsymbol{\psi}}_{\mathbf{1}}$ to test $H_0 : \boldsymbol{\psi_1} = \mathbf{0}$.

Other

Show that the IV estimation can be implemented as the 2SLS procedure.

The estimator that we obtain from the 2SLS procedure is

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y},$$

where $\hat{\boldsymbol{x}}_i = (1, x_{i,1}, ..., x_{i,K-1}, \hat{x}_{i,K}), i = 1, ..., N$ and $\hat{x}_{i,K}$ are the fitted values from first regressing the endogenous variable x_K on $\boldsymbol{z}$ where $\boldsymbol{z} = (1, x_1, x_2, ..., x_{K-1}, z_1, ..., z_M)$. We want to show that

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y},$$

since $(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y}$ is the IV estimator that uses the instruments $\hat{x}_i$ that we obtain from the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{K-1} x_{K-1} + \beta_k \hat{x}_K.$$

Thus, all we need to do is show this is to show that $\hat{\boldsymbol{X}}'\boldsymbol{X} = \hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}$. In the first stag,e we estimate the reduced model

$$x_K = \boldsymbol{z}\boldsymbol{\delta} + r_K$$

and find that

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{x}_K.$$

Then, we observe that

$$\hat{\boldsymbol{x}}_K = \boldsymbol{Z}\hat{\boldsymbol{\delta}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{x}_K.$$

Finally, we can see that

$$\hat{\boldsymbol{X}} = (\boldsymbol{1}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{K-1}, \hat{\boldsymbol{x}}_K) = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$$

since regressing any exogenous variable $x_i, i = 1, ..., K-1$ on $\boldsymbol{z}$ yields a fitted value that is exactly equal to itself, i.e.

$$\hat{\boldsymbol{x}}_i = \boldsymbol{Z}\hat{\boldsymbol{\delta}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{x}_i = \boldsymbol{x}_i, \;\; i = 1, ..., K-1.$$

Now, let $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. Then we can see that $\hat{\boldsymbol{X}} = \boldsymbol{P}_Z\boldsymbol{X}$. Since $\boldsymbol{P}_Z$ is idempotent, it must be that

$$\begin{aligned}
\hat{\boldsymbol{X}}'\boldsymbol{X} =& (\boldsymbol{P}_Z\boldsymbol{X})'\boldsymbol{X} \\
=& \boldsymbol{X}'\boldsymbol{P}_Z'\boldsymbol{X} \\
=& \boldsymbol{X}'\boldsymbol{P}_Z'\boldsymbol{P}_Z'\boldsymbol{X} \\
=& \boldsymbol{X}'\boldsymbol{P}_Z'\boldsymbol{P}_Z\boldsymbol{X} \\
=& (\boldsymbol{P}_Z\boldsymbol{X})'\boldsymbol{P}_Z\boldsymbol{X} \\
=& \hat{\boldsymbol{X}}'\hat{\boldsymbol{X}},
\end{aligned}$$

and we have shown what we desired to show. Finally, we prove that $\boldsymbol{P}_Z$ is symmetric, i.e.

$$\begin{aligned}
\boldsymbol{P}_Z' =& (\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}')' \\
=& (\boldsymbol{Z}')'((\boldsymbol{Z}'\boldsymbol{Z})^{-1})'\boldsymbol{Z}' \\
=& \boldsymbol{Z}((\boldsymbol{Z}'\boldsymbol{Z})')^{-1}\boldsymbol{Z}' \\
=& \boldsymbol{Z}(\boldsymbol{Z}'(\boldsymbol{Z}')')^{-1}\boldsymbol{Z}' \\
=& \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' \\
=& \boldsymbol{P}_Z,
\end{aligned}$$

and idempotent, i.e.

$$\begin{aligned}
\boldsymbol{P}_Z\boldsymbol{P}_Z =& (\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}')(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}') \\
=& \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' \\
=& \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' \\
=& \boldsymbol{P}_Z.
\end{aligned}$$

Appendix: R Code and Output

Tue Feb 06 00:07:20 2018

4.11

```
rm(list = ls())

library('haven')
nls80 = read_dta(file = 'nls80.dta', encoding = NULL)
```

a) and c)

```
model_nls80_1 = lm(lwage ~ exper + tenure + married + south + urban + black + educ
, data = nls80)
summary(model_nls80_1)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ, data = nls80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98069 -0.21996  0.00707  0.24288  1.22822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.395497   0.113225  47.653  < 2e-16 ***
## exper        0.014043   0.003185   4.409 1.16e-05 ***
## tenure       0.011747   0.002453   4.789 1.95e-06 ***
## married      0.199417   0.039050   5.107 3.98e-07 ***
## south       -0.090904   0.026249  -3.463 0.000558 ***
## urban        0.183912   0.026958   6.822 1.62e-11 ***
## black       -0.188350   0.037667  -5.000 6.84e-07 ***
## educ         0.065431   0.006250  10.468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3655 on 927 degrees of freedom
## Multiple R-squared:  0.2526, Adjusted R-squared:  0.2469
## F-statistic: 44.75 on 7 and 927 DF,  p-value: < 2.2e-16
```

```
model_nls80_2 = lm(lwage ~ exper + tenure + married + south + urban + black + educ
+ iq, data = nls80)
summary(model_nls80_2)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq, data = nls80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01203 -0.22244  0.01017  0.22951  1.27478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.1764392  0.1280006  40.441  < 2e-16 ***
## exper        0.0141458  0.0031651   4.469 8.82e-06 ***
## tenure       0.0113951  0.0024394   4.671 3.44e-06 ***
## married      0.1997644  0.0388025   5.148 3.21e-07 ***
## south       -0.0801695  0.0262529  -3.054 0.002325 **
## urban        0.1819463  0.0267929   6.791 1.99e-11 ***
## black       -0.1431253  0.0394925  -3.624 0.000306 ***
## educ         0.0544106  0.0069285   7.853 1.12e-14 ***
## iq           0.0035591  0.0009918   3.589 0.000350 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3632 on 926 degrees of freedom
## Multiple R-squared:  0.2628, Adjusted R-squared:  0.2564
## F-statistic: 41.27 on 8 and 926 DF,  p-value: < 2.2e-16
```

```
model_nls80_3 = lm(lwage ~ exper + tenure + married + south + urban + black + educ
+ kww, data = nls80)
summary(model_nls80_3)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + kww, data = nls80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04494 -0.21931 -0.00048  0.24163  1.26464
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.358797   0.113600  47.172  < 2e-16 ***
## exper        0.012228   0.003241   3.773 0.000172 ***
## tenure       0.011072   0.002456   4.507 7.40e-06 ***
## married      0.189461   0.039077   4.848 1.46e-06 ***
## south       -0.091601   0.026156  -3.502 0.000484 ***
## urban        0.175545   0.027032   6.494 1.36e-10 ***
## black       -0.164267   0.038530  -4.263 2.22e-05 ***
## educ         0.057628   0.006838   8.428  < 2e-16 ***
## kww          0.005028   0.001819   2.764 0.005820 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3642 on 926 degrees of freedom
## Multiple R-squared:  0.2587, Adjusted R-squared:  0.2523
## F-statistic: 40.39 on 8 and 926 DF,  p-value: < 2.2e-16
```

```r
model_nls80_4 = lm(lwage ~ exper + tenure + married + south + urban + black + educ
+ iq + kww, data = nls80)
summary(model_nls80_4)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq + kww, data = nls80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05704 -0.21621  0.00824  0.23725  1.24895
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.175644   0.127776  40.506  < 2e-16 ***
## exper        0.012752   0.003231   3.947 8.51e-05 ***
## tenure       0.010925   0.002446   4.467 8.92e-06 ***
## married      0.192145   0.038909   4.938 9.35e-07 ***
## south       -0.082029   0.026222  -3.128  0.00181 **
## urban        0.175823   0.026910   6.534 1.06e-10 ***
## black       -0.130399   0.039901  -3.268  0.00112 **
## educ         0.049837   0.007262   6.863 1.24e-11 ***
## iq           0.003118   0.001013   3.079  0.00214 **
## kww          0.003826   0.001852   2.066  0.03913 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3625 on 925 degrees of freedom
## Multiple R-squared:  0.2662, Adjusted R-squared:  0.2591
## F-statistic: 37.28 on 9 and 925 DF,  p-value: < 2.2e-16
```

b)

```
fstat_1 = (sum(model_nls80_1$residuals^2) - sum(model_nls80_4$residuals^2))/sum(mod
el_nls80_4$residuals^2)*925/2
fstat_1
```

```
## [1] 8.59499
```

```
pval_1 = 1-pf(fstat_1, 2, 925)
pval_1
```

```
## [1] 0.0002002184
```

d)

```
nls80$educiqadj = nls80$educ*(nls80$iq - 100)
nls80$educkwwadj = nls80$educ*(nls80$kww - mean(nls80$kww))

model_nls80_5 = lm(lwage ~ exper + tenure + married + south + urban + black + educ
+ iq + kww + educiqadj + educkwwadj, data = nls80)
summary(model_nls80_5)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq + kww + educiqadj + educkwwadj, data = nls80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03083 -0.21490  0.00886  0.23928  1.27862
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0800046  0.5610875  10.836  < 2e-16 ***
## exper        0.0121544  0.0032358   3.756 0.000183 ***
## tenure       0.0107206  0.0024383   4.397 1.23e-05 ***
## married      0.1978269  0.0388272   5.095 4.23e-07 ***
## south       -0.0807609  0.0261374  -3.090 0.002063 **
## urban        0.1784310  0.0268710   6.640 5.34e-11 ***
## black       -0.1381481  0.0399615  -3.457 0.000571 ***
## educ         0.0452410  0.0076469   5.916 4.64e-09 ***
## iq           0.0048228  0.0057333   0.841 0.400458
## kww         -0.0248007  0.0107382  -2.310 0.021132 *
## educiqadj   -0.0001138  0.0004228  -0.269 0.787970
## educkwwadj   0.0021610  0.0007957   2.716 0.006735 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3613 on 923 degrees of freedom
## Multiple R-squared:  0.2728, Adjusted R-squared:  0.2641
## F-statistic: 31.48 on 11 and 923 DF,  p-value: < 2.2e-16
```

```
fstat_2 = (sum(model_nls80_4$residuals^2) - sum(model_nls80_5$residuals^2))/sum(mod
el_nls80_5$residuals^2)*923/2
fstat_2
```

```
## [1] 4.19454
```

```
pval_2 = 1-pf(fstat_2, 2, 923)
pval_2
```

```
## [1] 0.01536607
```

4.13

```
rm(list = ls())

library('haven')
cornwell = read_dta(file = 'cornwell.dta', encoding = NULL)
```

a)

```r
cornwell_87 = cornwell[which(cornwell$d87 == 1),]
model_cornwell_1 = lm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen, data = cor
nwell_87)
summary(model_cornwell_1)
```

```
##
## Call:
## lm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen,
##     data = cornwell_87)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36105 -0.19129  0.07939  0.27754  0.86843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.86792    0.43153 -11.281  < 2e-16 ***
## lprbarr     -0.72397    0.11532  -6.278 1.39e-08 ***
## lprbconv    -0.47251    0.08311  -5.686 1.80e-07 ***
## lprbpris     0.15967    0.20644   0.773    0.441
## lavgsen      0.07642    0.16347   0.467    0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 85 degrees of freedom
## Multiple R-squared:  0.4162, Adjusted R-squared:  0.3888
## F-statistic: 15.15 on 4 and 85 DF,  p-value: 2.171e-09
```

b)

```r
cornwell_86 = cornwell[which(cornwell$d86 == 1),]
cornwell_87$lcrmrte86 = cornwell_86$lcrmrte
model_cornwell_2 = lm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lcrmrte8
6, data = cornwell_87)
summary(model_cornwell_2)
```

```
##
## Call:
## lm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen +
##     lcrmrte86, data = cornwell_87)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28023 -0.08931  0.03055  0.11019  0.32422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.76663    0.31310  -2.449  0.01643 *
## lprbarr     -0.18504    0.06276  -2.948  0.00414 **
## lprbconv    -0.03868    0.04660  -0.830  0.40890
## lprbpris    -0.12669    0.09885  -1.282  0.20351
## lavgsen     -0.15202    0.07829  -1.942  0.05552 .
## lcrmrte86    0.77981    0.04521  17.248  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2025 on 84 degrees of freedom
## Multiple R-squared:  0.8715, Adjusted R-squared:  0.8638
## F-statistic: 113.9 on 5 and 84 DF,  p-value: < 2.2e-16
```

c)

```
model_cornwell_3 = lm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lcrmrte8
6 + lwcon +lwtuc +lwtrd +lwfir + lwser + lwmfg +lwfed + lwsta + lwloc, data = corn
well_87)
summary(model_cornwell_3)
```

```
##
## Call:
## lm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen +
##     lcrmrte86 + lwcon + lwtuc + lwtrd + lwfir + lwser + lwmfg +
##     lwfed + lwsta + lwloc, data = cornwell_87)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08551 -0.08893  0.01248  0.10860  0.35101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.79253    1.95747  -1.937   0.0565 .
## lprbarr     -0.17251    0.06595  -2.616   0.0108 *
## lprbconv    -0.06836    0.04973  -1.375   0.1733
## lprbpris    -0.21556    0.10240  -2.105   0.0386 *
## lavgsen     -0.19605    0.08446  -2.321   0.0230 *
## lcrmrte86    0.74534    0.05303  14.054   <2e-16 ***
## lwcon       -0.28500    0.17752  -1.605   0.1126
## lwtuc        0.06413    0.13433   0.477   0.6344
## lwtrd        0.25371    0.23174   1.095   0.2771
## lwfir       -0.08353    0.19650  -0.425   0.6720
## lwser        0.11275    0.08474   1.331   0.1874
## lwmfg        0.09874    0.11861   0.832   0.4078
## lwfed        0.33613    0.24531   1.370   0.1747
## lwsta        0.03951    0.20721   0.191   0.8493
## lwloc       -0.03699    0.32915  -0.112   0.9108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1973 on 75 degrees of freedom
## Multiple R-squared:  0.8911, Adjusted R-squared:  0.8707
## F-statistic: 43.81 on 14 and 75 DF,  p-value: < 2.2e-16
```

```
fstat_1 = (sum(model_cornwell_2$residuals^2) - sum(model_cornwell_3$residuals^2))/
sum(model_cornwell_3$residuals^2)*75/9
fstat_1
```

```
## [1] 1.498106
```

```
pval_1 = 1-pf(fstat_1, 9, 75)
pval_1
```

```
## [1] 0.164286
```

d)

```r
library('lmtest')
library('sandwich')

waldtest(model_cornwell_3,model_cornwell_2,vcov=vcovHC(model_cornwell_3, type = 'HC
1'))
```

```
## Wald test
##
## Model 1: lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lcrmrte86 +
##     lwcon + lwtuc + lwtrd + lwfir + lwser + lwmfg + lwfed + lwsta +
##     lwloc
## Model 2: lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lcrmrte86
##   Res.Df Df      F  Pr(>F)
## 1     75
## 2     84 -9 2.1907 0.03189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.14

```r
rm(list = ls())

library('haven')
attend = read_dta(file = 'attend.dta', encoding = NULL)
```

a)

```r
model_attend_1 = lm(stndfnl ~ frosh + soph + atndrte, data = attend)
summary(model_attend_1)
```

```
##
## Call:
## lm(formula = stndfnl ~ frosh + soph + atndrte, data = attend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2825 -0.6719 -0.0295  0.6791  2.5455
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.501731   0.196314  -2.556 0.010813 *
## frosh       -0.289894   0.115724  -2.505 0.012478 *
## soph        -0.118446   0.099027  -1.196 0.232078
## atndrte      0.008163   0.002203   3.705 0.000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9772 on 676 degrees of freedom
## Multiple R-squared:  0.02904,    Adjusted R-squared:  0.02473
## F-statistic: 6.739 on 3 and 676 DF,  p-value: 0.0001752
```

c)

```
model_attend_2 = lm(stndfnl ~ frosh + soph + atndrte + priGPA + ACT, data = attend
)
summary(model_attend_2)
```

```
##
## Call:
## lm(formula = stndfnl ~ frosh + soph + atndrte + priGPA + ACT,
##     data = attend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1904 -0.5668 -0.0252  0.5887  2.2915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.297342   0.308831 -10.677  < 2e-16 ***
## frosh       -0.049469   0.107890  -0.459   0.6467
## soph        -0.159648   0.089772  -1.778   0.0758 .
## atndrte      0.005225   0.002384   2.191   0.0288 *
## priGPA       0.426585   0.081920   5.207 2.55e-07 ***
## ACT          0.084412   0.011168   7.559 1.33e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8851 on 674 degrees of freedom
## Multiple R-squared:  0.2058, Adjusted R-squared:  0.1999
## F-statistic: 34.93 on 5 and 674 DF,  p-value: < 2.2e-16
```

e)

```
attend$priGPAsq = attend$priGPA^2
attend$ACTsq = attend$ACT^2
model_attend_3 = lm(stndfnl ~ frosh + soph + atndrte + priGPA + ACT + priGPAsq + A
CTsq, data = attend)
summary(model_attend_3)
```

```
##
## Call:
## lm(formula = stndfnl ~ frosh + soph + atndrte + priGPA + ACT +
##     priGPAsq + ACTsq, data = attend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.14241 -0.54902 -0.02163  0.56155  2.36547
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.384812   1.239361   1.117  0.26424
## frosh       -0.105337   0.106975  -0.985  0.32513
## soph        -0.180729   0.088635  -2.039  0.04184 *
## atndrte      0.006232   0.002358   2.642  0.00842 **
## priGPA      -1.526140   0.473972  -3.220  0.00134 **
## ACT         -0.112433   0.098172  -1.145  0.25251
## priGPAsq     0.368218   0.088985   4.138 3.95e-05 ***
## ACTsq        0.004182   0.002169   1.928  0.05425 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8718 on 672 degrees of freedom
## Multiple R-squared:  0.2316, Adjusted R-squared:  0.2236
## F-statistic: 28.94 on 7 and 672 DF,  p-value: < 2.2e-16
```

```
fstat_1 = (sum(model_attend_2$residuals^2) - sum(model_attend_3$residuals^2))/sum(m
odel_attend_3$residuals^2)*671/2
fstat_1
```

```
## [1] 11.27926
```

```
pval_1 = 1-pf(fstat_1, 2, 671)
pval_1
```

```
## [1] 1.520629e-05
```

f)

```
attend$atndrtesq = attend$atndrte^2
model_attend_4 = lm(stndfnl ~ frosh + soph + atndrte + priGPA + ACT + priGPAsq + A
CTsq + atndrtesq, data = attend)
summary(model_attend_4)
```

```
##
## Call:
## lm(formula = stndfnl ~ frosh + soph + atndrte + priGPA + ACT +
##     priGPAsq + ACTsq + atndrtesq, data = attend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.14165 -0.54816 -0.02205  0.55940  2.36637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.394e+00  1.267e+00   1.100  0.27159
## frosh       -1.054e-01  1.071e-01  -0.984  0.32537
## soph        -1.808e-01  8.875e-02  -2.038  0.04199 *
## atndrte      5.843e-03  1.092e-02   0.535  0.59282
## priGPA      -1.525e+00  4.757e-01  -3.205  0.00141 **
## ACT         -1.123e-01  9.828e-02  -1.143  0.25339
## priGPAsq     3.679e-01  8.944e-02   4.113 4.38e-05 ***
## ACTsq        4.180e-03  2.171e-03   1.925  0.05461 .
## atndrtesq    2.873e-06  7.871e-05   0.036  0.97089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8725 on 671 degrees of freedom
## Multiple R-squared:  0.2316, Adjusted R-squared:  0.2225
## F-statistic: 25.28 on 8 and 671 DF,  p-value: < 2.2e-16
```

5.3

```r
rm(list = ls())

library('haven')
bwght = read_dta(file = 'bwght.dta', encoding = NULL)
```

c)

```r
model_bwght_1 = lm(lbwght ~ male + parity + lfaminc + packs, data = bwght)
summary(model_bwght_1)
```

```
##
## Call:
## lm(formula = lbwght ~ male + parity + lfaminc + packs, data = bwght)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63729 -0.08845  0.02034  0.12271  0.84409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.675618   0.021881 213.681  < 2e-16 ***
## male         0.026241   0.010089   2.601  0.00940 **
## parity       0.014729   0.005665   2.600  0.00942 **
## lfaminc      0.018050   0.005584   3.233  0.00126 **
## packs       -0.083728   0.017121  -4.890 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1876 on 1383 degrees of freedom
## Multiple R-squared:  0.03504,    Adjusted R-squared:  0.03225
## F-statistic: 12.55 on 4 and 1383 DF,  p-value: 4.905e-10
```

```
model_bwght_iv <- AER::ivreg(lbwght ~ male + parity + lfaminc + packs | male + par
ity + lfaminc + cigprice, data = bwght)
summary(model_bwght_iv, diagnostics = TRUE)
```

```
##
## Call:
## AER::ivreg(formula = lbwght ~ male + parity + lfaminc + packs |
##     male + parity + lfaminc + cigprice, data = bwght)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19538 -0.06910  0.07829  0.19077  0.89686
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.467861   0.258829  17.262   <2e-16 ***
## male         0.029821   0.017779   1.677   0.0937 .
## parity      -0.001239   0.021932  -0.056   0.9550
## lfaminc      0.063646   0.057013   1.116   0.2645
## packs        0.797106   1.086275   0.734   0.4632
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 1383     1.002   0.317
## Wu-Hausman          1 1382     1.919   0.166
## Sargan              0   NA        NA      NA
##
## Residual standard error: 0.3202 on 1383 degrees of freedom
## Multiple R-Squared: -1.812,  Adjusted R-squared: -1.82
## Wald test: 2.391 on 4 and 1383 DF,  p-value: 0.04896
```

d)

```r
model_bwght_2 = lm(packs ~ male + parity + lfaminc + cigprice, data = bwght)
summary(model_bwght_2)
```

```
##
## Call:
## lm(formula = packs ~ male + parity + lfaminc + cigprice, data = bwght)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36386 -0.11365 -0.08285 -0.04761  2.36602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1374075  0.1040005   1.321   0.1866
## male        -0.0047261  0.0158539  -0.298   0.7657
## parity       0.0181491  0.0088802   2.044   0.0412 *
## lfaminc     -0.0526374  0.0086991  -6.051 1.85e-09 ***
## cigprice     0.0007770  0.0007763   1.001   0.3171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2945 on 1383 degrees of freedom
## Multiple R-squared:  0.03045,    Adjusted R-squared:  0.02765
## F-statistic: 10.86 on 4 and 1383 DF,  p-value: 1.137e-08
```