Problem Set 4

Scott Cohn

December 02 2020

Question 1

Part A

Just as with fixed-effects, the first difference estimator is consistent (and unbiased) if the covariates are strictly exogenous. Since we remove the a_i value, correlation with a_i is okay. Since the panel is unbalanced, the fixed effects estimator can aptly handle this problem. However, in the proposed first difference model, we should be concerned as to why these second period data are missing. If they are "missing at random", this may not be problematic. If there is a non-random explanation for why values are missing in the panel, we may have a problem.

Part B

Since we are confident that a_i is uncorrelated, we may consider using a *random effects* formulation. That is, we could use an intercept, β_0, to allow for a zero-mean assumption on a_i and thus use a composite error term including a_i: $(a_i + u_i)$. The parameter a_i only causes bias and inconsistency when it is *correlated* with one (or more) of the covariates.

Question 2

Part A

Let y^* be a latent random variable where,

$$y^* = X^T\beta + u,$$

where $U \sim \mathcal{N}(0, \sigma^2)$. Then for $y = 1(y^* > 0)$, we have

$$P(y = 1 \mid x) = P(y^* > 0 \mid x) = P(x\beta + u > 0 \mid x) = P(-u < x\beta) = \Phi(x\beta).$$

Note that we have assumed that $\sigma = 1$. If we had not made this assumption, then equivalently we have:

$$P(y_i = 1) = P\left(\frac{u_i}{\sigma} > -x_i\frac{\beta}{\sigma}\right).$$

It follows that we cannot estimate β and σ separately as they enter the likelihood function as a ratio. Therefore, to make the distribution on the residuals a standard normal, we set $\sigma = 1$ which allows

$$\hat{\beta} = \arg\max_{\beta} \left[\ln \mathcal{L}(\beta)\right],$$

where $\hat{\beta} = \frac{\beta}{\sigma}$. Therefore it follows that cannot obtain the scale of the latent variable. Setting $\sigma = 1$ allows us to interpret the β coefficients in units of standard deviation of the latent variable.

Part B

B1

You can use OLS, however it would *not* make sense because the estimates of the coefficients would be inconsistent because our data is censored. It would make more sense to use a tobit model.

B2

Let α_i denote the income for y_i^* where $i \in \{1, 2, 3\}$. First,

$$\begin{aligned} P(y = 1 \mid x) &= P(x'\beta + u < \alpha_1) \\ &= \Phi(\alpha_1 - x'\beta - \alpha_1) \\ &= 1 - \Phi\left(\frac{x'\beta - \alpha_1}{\sigma}\right). \end{aligned}$$

Then,

$$\begin{aligned} P(y = 3 \mid x) &= P(x'\beta + u < \alpha_3) \\ &= \Phi(\alpha_3 - x'\beta - \alpha_3) \\ &= 1 - \Phi\left(\frac{x'\beta - \alpha_3}{\sigma}\right). \end{aligned}$$

Finally, we have the intermediate value:

$$\begin{aligned} P(y = 2 \mid x) &= P(\alpha_1 < x'\beta < \alpha_3) \\ &= 1 - \big(1 - \Phi(x'\beta - \alpha_1)\big) - \Phi(x'\beta - \alpha_3) \\ &= \Phi(x'\beta - \alpha_1) - \Phi(x'\beta - \alpha_3) \\ &= \Phi\left(\frac{\alpha_3 - x'\beta}{\sigma}\right) - \Phi\left(\frac{\alpha_1 - x'\beta}{\sigma}\right). \end{aligned}$$

B3

Then, for an i.i.d. sample, the log-likelihood function is the following:

$$\begin{aligned} S(b, s) = \sum_{i=1}^{n} 1(y_i = 1) \cdot \ln\left(1 - \Phi\left(\frac{x'b - \alpha_1}{s}\right)\right) + 1(y_i = 2) \cdot \ln\left(\Phi\left(\frac{\alpha_3 - x'b}{s}\right) - \Phi\left(\frac{\alpha_1 - x'b}{s}\right)\right) \\ + 1(y_i = 3) \cdot \ln\left(1 - \Phi\left(\frac{x'b - \alpha_3}{s}\right)\right). \end{aligned}$$

B4

In the estimated model, the slope β_j has no direct meaning. However, it does have the same *sign* as the partial effect which is measured as $\phi(x\beta) \cdot \beta_j$.

Question 3

$$P(y = 1 \mid x, z) = \Phi(\beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz).$$

Part A

The partial effect of x evaluated at some given value of x and z is

$$\frac{\partial P(y = 1 \mid x, z)}{\partial x} = \phi(x\beta) \cdot (\beta_2 + \beta_4 z),$$

per Greene section 6.3.3.

Part B

The conditional mean of the dependent variable is:

$$\begin{aligned} E[y \mid x, z, X] &= \Phi(\beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + X\beta) \\ &= \Phi(u), \end{aligned}$$

Observe that

$$\frac{\partial \Phi(u)}{\partial (xz)} = \beta_4 \phi(u).$$

Then, the full-interaction effect is the cross-partial derivative of the expected value of y:

$$\frac{\partial^2 \Phi(u)}{\partial x \partial z} = \phi(u)\beta_4 + \phi'(u)(\beta_2 + \beta_4 z)(\beta_3 + \beta_4 x).$$

Part C

Given $P(y = 1 \mid x, z) = \Phi\left(\beta_1 + \beta_2 x + \beta_3 x^2\right)$, the partial effect of x is:

$$\frac{\partial P(y = 1 \mid x, z)}{\partial x} = \phi(x\beta)(\beta_2 + 2\beta_3 x).$$

Question 4

Part A

We use the model

$$\ln(\texttt{price}) = \beta_1 + \beta_2 \ln(\texttt{dist}) + \delta_1 \texttt{y81} + \delta_2 \texttt{y81} \cdot \ln(\texttt{dist}) + u.$$

```
model_4a <- lm(lprice ~ ldist + y81 + y81ldist, data = kielmc)
```

If building the incinerator reduces the value of homes closer to the site, the sign of δ_2 would be positive. That is, in the second time period, increasing distance from the incinerator site would be associated with higher home prices, all else fixed.

If $\beta_2 > 0$, then an increase in log-distance from the incinerator site would be associated with an increase in home prices by the percentage value equal to the value of β_2.

Part B

As expected, the sign of δ_2 is positive, however it is not statistically significant. We would interpret this coefficient as: In the second time period a 1-unit increase in log-distance from the incinerator site would be associated with a 4.8% increase in home prices, all else fixed.

Part C

Table 1: Robust coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	8.058	0.375	21.501	0.000
ldist	0.317	0.038	8.437	0.000
y81	-0.011	0.762	-0.015	0.988
y81ldist	0.048	0.077	0.629	0.530

```
model_4c <- lm(lprice ~ ldist + y81 + y81ldist + age + agesq + rooms
                + baths + lintst + lland + larea, data = kielmc)
```

Table 2: Robust coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	9.952	0.416	23.907	0.000
ldist	-0.021	0.048	-0.433	0.665
y81	-0.297	0.542	-0.549	0.583
y81ldist	0.073	0.054	1.343	0.180
age	-0.007	0.002	-4.110	0.000
agesq	0.000	0.000	2.831	0.005
rooms	0.069	0.018	3.885	0.000
baths	0.203	0.022	9.356	0.000
lintst	-0.066	0.039	-1.720	0.086
lland	0.119	0.040	2.995	0.003

The variables that demonstrate the greatest magnitude and are of statistical significance are values that typically contribute to home prices – land, number of bed/baths – rather than those associated with the construction of, or proximity to, the incinerator. It is likely that the incinerator plays little role in the home prices, and the drivers of change in home value are due to the standard lot-size and room assortment. This latter model controls for these traditional hedonic price measures whereas the first model does not, further illuminating the effect (or lack thereof) of the incinerator.

Question 5

Part A

```
# FD model vote2
model_5a <- lm(cvote ~ clinexp + clchexp + cincshr,
               data = vote2)
```

The only variable that is significant at the 5% level is the change in the income share, `cinchsr`.

Part B

```
linearHypothesis(model_5a, c("clinexp=0", "clchexp=0"))
```

```
## Linear hypothesis test
##
```

Table 3: FD – Robust coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	-2.556	0.585	-4.372	0.000
clinexp	-1.292	1.291	-1.000	0.319
clchexp	-0.599	0.577	-1.038	0.301
cincshr	0.156	0.053	2.950	0.004

```
## Hypothesis:
## clinexp = 0
## clchexp = 0
##
## Model 1: restricted model
## Model 2: cvote ~ clinexp + clchexp + cincshr
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    155 9282.3
## 2    153 9102.3  2    179.98 1.5126 0.2236
```

The p-value is 0.2236. That is, we fail to reject the null hypothesis that the joint coefficients are different from 0.

Part C

```
# FD model vote2
model_5c <- plm(cvote ~ cincshr,
                data = vote2)
```

Table 4: FD – Robust coefficient summary

term	estimate	std.error	statistic	p.value
cincshr	0.219	0.035	6.347	0

We can interpret the estimated slope on the differenced `incshr` as the difference in vote return for a percent change in the income share from 1988 to 1990. That is, a 1-unit change in `incshr` from 1988 to 1990 is associated with an increase in vote difference by 2.2 %.

Similarly, increasing the incumbent vote share by 10 percentage points is predicted to increase the incumbent's share of the vote by 2.4 percentage points (1.1×2.18).

Part D

```
# FD model vote2
model_5d <- plm(cvote ~ cincshr + factor(state),
                data = vote2)
```

Clustering by state changes the standard error only slightly. The statistical significance of the variable is not affected. Moreover the change only occurs in the thousandth decimal place.

Table 5: FD – Cluster robust coefficient summary

term	estimate	std.error	statistic	p.value
cincshr	0.219	0.035	6.276	0

Part E

```
# FD model vote2
model_5e <- plm(cvote ~ cincshr,
                data = vote2 %>% filter(rptchall == 1))
```

Table 6: FD – Robust coefficient summary

term	estimate	std.error	statistic	p.value
cincshr	-0.04	0.104	-0.385	0.706

Yes, the results change; the sign of the coefficient slope is now negative and no longer significant.

Question 6

Part A

The only parameter that you *cannot* estimate is β_2 for $\texttt{educ}_i$, which is time-invariant.

Part B

```
model_6b_fe <- plm(lwage ~ union + educ + year + year*educ,
                   model = "within",
                   index = c("year"),
                   effect = "time",
                   data = wagepan)
```

```
model_6b_pool <- plm(lwage ~ union + educ,
                   model = "pooling",
                   index = c("year"),
                   data = wagepan)

pFtest(model_6b_fe, model_6b_pool)
```

```
##
##  F test for time effects
##
## data:  lwage ~ union + educ + year + year * educ
## F = 9.4138, df1 = 557, df2 = 3800, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

We estimate the fixed-effects model versus the pooled OLS model. We have evidence to reject the null hypothesis there are no time effects at the 1% level.

Table 7: FE – Robust coefficient summary

term	estimate	std.error	statistic	p.value
union	0.085	0.018	4.662	0.000
year1981	-0.016	0.003	-5.837	0.000
year1982	-0.005	0.002	-2.596	0.009
year1983	0.015	0.001	13.259	0.000
year1984	0.089	0.001	138.089	0.000
year1985	0.050	0.002	23.512	0.000
year1986	0.068	0.003	20.038	0.000
year1987	0.095	0.000	259.239	0.000
educ:year1981	0.012	0.000	50.682	0.000
educ:year1982	0.015	0.000	96.498	0.000
educ:year1983	0.018	0.000	197.490	0.000
educ:year1984	0.018	0.000	323.623	0.000
educ:year1985	0.025	0.000	185.796	0.000
educ:year1986	0.029	0.000	127.135	0.000
educ:year1987	0.032	0.000	667.189	0.000

Part C

```
model_6c_fe <- plm(lwage ~ union + educ + year + year*educ + year*union,
                   model = "within",
                   index = c("year"),
                   effect = "time",
                   data = wagepan)
```

The estimated union differential in 1980 is just the coefficient for `union` as the year effects zero out. That is, the union differential (union versus non-union) is 0.167.

The estimated union differential in 1987 is the difference of `union` and `union:year1987` per Table 8. That is, the differential in 1987 is 0.315. Next we test whether the difference between the two is statistically significant.

```
linearHypothesis(model_6c_fe, c("union:year1987 - union = 0"), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## - union  + union:year1987 = 0
##
## Model 1: restricted model
## Model 2: lwage ~ union + educ + year + year * educ + year * union
##
##   Res.Df Df      F    Pr(>F)
## 1   3794
## 2   3793  1 14.788 0.0001223 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have evidence to suggest that the union differential between 1987 and 1980 is statistically significant at the 1% level.

Table 8: FE – Robust coefficient summary

term	estimate	std.error	statistic	p.value
union	0.167	0.016	10.641	0.000
year1981	0.007	0.003	2.048	0.041
year1982	0.021	0.003	7.272	0.000
year1983	0.050	0.003	18.895	0.000
year1984	0.118	0.003	38.100	0.000
year1985	0.075	0.004	21.321	0.000
year1986	0.106	0.003	32.271	0.000
year1987	0.143	0.004	33.992	0.000
educ:year1981	0.011	0.000	58.617	0.000
educ:year1982	0.015	0.000	107.944	0.000
educ:year1983	0.017	0.000	209.611	0.000
educ:year1984	0.017	0.000	178.202	0.000
educ:year1985	0.025	0.000	211.879	0.000
educ:year1986	0.029	0.000	137.903	0.000
educ:year1987	0.031	0.000	432.969	0.000
union:year1981	-0.055	0.009	-6.158	0.000
union:year1982	-0.072	0.009	-7.849	0.000
union:year1983	-0.108	0.009	-12.328	0.000
union:year1984	-0.085	0.008	-10.405	0.000
union:year1985	-0.070	0.010	-7.001	0.000
union:year1986	-0.149	0.011	-13.756	0.000
union:year1987	-0.148	0.013	-11.115	0.000

Part D

```
model_6c_pool <- plm(lwage ~ union + educ,
                 model = "pooling",
                 index = c("year"),
                 data = wagepan)

pFtest(model_6c_fe, model_6c_pool)
```

```
##
##  F test for time effects
##
## data:  lwage ~ union + educ + year + year * educ + year * union
## F = 9.3321, df1 = 564, df2 = 3793, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

We estimate the fixed-effects model versus the pooled OLS model. We have evidence to reject the null hypothesis there are no time effects at the 1% level. That is, the we reject the null hypothesis that the union differential has not changed over time.

Question 7

Part A

```
# LPM
model_7a_lpm <- lm(approve ~ white, data = loanapp)

# Probit
model_7a_probit <- glm(approve ~ white,
                       family = binomial(link = "probit"),
                       data = loanapp)
```

Table 9: LPM – Robust coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	0.708	0.026	27.300	0
white	0.201	0.027	7.467	0

Table 10: Probit – Coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	0.547	0.075	7.251	0
white	0.784	0.087	9.041	0

```
## Average marginal effects

## glm(formula = approve ~ white, family = binomial(link = "probit"),     data = loanapp)

##   white
##  0.1507
```

The estimates are the same. It is 70% if you're not white and 90% if you're white.

Part B

```
# LPM
model_7b_lpm <- lm(approve ~ white + hrat + obrat + loanprc + unem + male
                   + married + dep + sch + cosign + chist + pubrec
                   + mortlat1 + mortlat2 +  vr,
                   data = loanapp)

# Probit
model_7b_probit <- glm(approve ~ white + hrat + obrat + loanprc + unem + male
                       + married + dep + sch + cosign + chist + pubrec
                       + mortlat1 + mortlat2 +  vr,
                       family = binomial(link = "probit"),
                       data = loanapp)
```

B1

Yes, in both models we find that being white increases the probability of a loan approval by a statistically significant margin.

Table 11: LPM – Robust coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	0.937	0.059	15.773	0.000
white	0.129	0.026	4.980	0.000
hrat	0.002	0.001	1.249	0.212
obrat	-0.005	0.001	-4.081	0.000
loanprc	-0.147	0.038	-3.893	0.000
unem	-0.007	0.004	-1.966	0.049
male	-0.004	0.019	-0.215	0.830
married	0.046	0.017	2.658	0.008
dep	-0.007	0.007	-0.989	0.323
sch	0.002	0.017	0.102	0.919
cosign	0.010	0.040	0.247	0.805
chist	0.133	0.025	5.403	0.000
pubrec	-0.242	0.043	-5.654	0.000
mortlat1	-0.057	0.066	-0.865	0.387
mortlat2	-0.114	0.091	-1.249	0.212
vr	-0.031	0.014	-2.171	0.030

Table 12: Probit – Coefficient summary

term	estimate	std.error	statistic	p.value
(Intercept)	2.062	0.337	6.125	0.000
white	0.520	0.097	5.368	0.000
hrat	0.008	0.007	1.064	0.287
obrat	-0.028	0.007	-4.126	0.000
loanprc	-1.012	0.257	-3.933	0.000
unem	-0.037	0.019	-1.948	0.051
male	-0.037	0.110	-0.335	0.738
married	0.266	0.098	2.712	0.007
dep	-0.050	0.040	-1.254	0.210
sch	0.015	0.093	0.157	0.875
cosign	0.086	0.214	0.403	0.687
chist	0.585	0.094	6.214	0.000
pubrec	-0.779	0.131	-5.946	0.000
mortlat1	-0.188	0.277	-0.678	0.498
mortlat2	-0.494	0.322	-1.535	0.125
vr	-0.201	0.082	-2.448	0.014

B2

If the loan applicant is white (`white` $= 1$), the probability of a loan approval increases by 12.9 percentage points in the LPM model.

B3

```r
# PEA
summary(margins(data = loanapp, model = model_7b_probit,
                variables = "white", atmeans = TRUE)) %>%
  kbl(format = "latex",
         digits = 3,
         booktabs = TRUE,
         caption = "PEA"
         ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 13: PEA

factor	AME	SE	z	p	lower	upper
white	0.086	0.016	5.421	0	0.055	0.118

```r
# APE
summary(margins(data = loanapp, model = model_7b_probit,
                variables = "white", atmeans = FALSE))  %>%
  kbl(format = "latex",
         digits = 3,
         booktabs = TRUE,
         caption = "APE/AME"
         ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 14: APE/AME

factor	AME	SE	z	p	lower	upper
white	0.086	0.016	5.421	0	0.055	0.118

The PEA is the partial effect at the average. That is, it is the effect of some x on y for a case with all sample averages of y. The APE (or AME) is the average partial (marginal) effect of some x on y. That is, it is the effect of x on y averaged across all cases in the sample.

In this scenario, oddly enough, I am getting that the PEA and APE are the same for `white` in the probit model. However, this value does differ quite significantly from the partial effect of `white` in the LPM model.

B4

A 1-percent increase in `obrat` is associated with the probability of a loan approval decreasing by 0.5 percentage points in the LPM model.

```r
# PEA
summary(margins(data = loanapp, model = model_7b_probit,
                variables = "obrat", atmeans = TRUE))   %>%
  kbl(format = "latex",
         digits = 3,
         booktabs = TRUE,
         caption = "PEA"
         ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 15: PEA

factor	AME	SE	z	p	lower	upper
obrat	-0.005	0.001	-4.527	0	-0.007	-0.003

```
# APE
summary(margins(data = loanapp, model = model_7b_probit,
                variables = "obrat", atmeans = FALSE))  %>%
  kbl(format = "latex",
        digits = 3,
        booktabs = TRUE,
        caption = "APE/AME"
        ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 16: APE/AME

factor	AME	SE	z	p	lower	upper
obrat	-0.005	0.001	-4.527	0	-0.007	-0.003

B5

Table 17: Predicted Approval Probabilities

at(obrat)	Prediction
10	0.956
20	0.930
30	0.894
40	0.844
50	0.781

These are the average predicted approval probabilities at `obrat` values of 10, 20, 30, 40, and 50.

B6

Table 18: dydx

obrat	AME
10	-0.0021
20	-0.0031
30	-0.0043
40	-0.0056
50	-0.0071

The average partial (marginal) effect increases in magnitude as `obrat` increases. This makes sense because if you are committed to other loan opportunities, you have debt and thus likely at greater risk to be denied (as seen in the table).

Part C

C1

The null hypothesis is

$$H0 : \texttt{hrat} = \texttt{unem} = \texttt{male} = \texttt{dep} = \texttt{sch} = \texttt{cosign} = \texttt{mortlat1} = \texttt{mortlat2} = 0.$$

C2

```
model_7b_lpm_r <- lm_robust(approve ~ white + hrat + obrat + loanprc + unem + male
                   + married + dep + sch + cosign + chist + pubrec
                   + mortlat1 + mortlat2 +  vr,
                   data = loanapp)

wald.test(b = coef(model_7b_lpm_r), Sigma = vcov(model_7b_lpm_r),
          Terms = c(3,6,7,9,10,11,14,15))
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 9.6, df = 8, P(> X2) = 0.3
```

The p-value associated with the Wald test statistic for this null hypothesis is 0.062.

C3

```
# Probit
model_7bc3_probit <- glm(approve ~ white + hrat + obrat + loanprc + unem + male
                         + married + dep + sch + cosign + chist + pubrec
                         + mortlat1 + mortlat2 +  vr,
                         family = binomial(link = "probit"),
                         data = na.omit(loanapp))
# Probit
model_7c3_probit <- glm(approve ~ white + obrat + loanprc
                         + married + chist + pubrec + vr,
                         family = binomial(link = "probit"),
                         data = na.omit(loanapp))

lrtest(model_7c3_probit, model_7bc3_probit)
```

```
## Likelihood ratio test
##
## Model 1: approve ~ white + obrat + loanprc + married + chist + pubrec +
##     vr
## Model 2: approve ~ white + hrat + obrat + loanprc + unem + male + married +
##     dep + sch + cosign + chist + pubrec + mortlat1 + mortlat2 +
##     vr
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   8 -551.58
## 2  16 -547.34  8 8.479     0.3881
```

The p-value for the likelihood ratio test for the null hypothesis is 0.3881.

C4

In both tests, we do not have evidence to reject the null hypothesis that the models are statistically different. That is, the tests do not differ in their result.

Question 8

Part A

```
model_8a <- plm(bmi ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                model = "pooling",
                data = married_bmi_pdata)
```

Table 19: Robust standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	30.360	0.964	31.495	0
male1	1.663	0.072	23.027	0
educ	-0.255	0.019	-13.187	0
age	0.186	0.037	5.094	0
agesq	-0.002	0.000	-4.036	0
smoke	-1.241	0.109	-11.354	0
logfaminc	-1.001	0.171	-5.864	0
withkid	-0.453	0.112	-4.025	0

Above is the pooled OLS with robust (but not clustered) standard errors.

Part B

```
# Obtain the residuals
married_bmi_pdata$e <- residuals(model_8a)

# Create df of f/m spousal resid
spouse_resid <- married_bmi_pdata %>%
  dplyr::select(e) %>%
  mutate(ind = rep(c(1, 2), length.out = n())) %>%
  group_by(ind) %>%
  mutate(id = row_number()) %>%
  spread(ind, e) %>%
  dplyr::select(-id) %>%
  rename("female" = 1,
         "male" = 2)

# Correlation between f/m spousal resid
cor.test(spouse_resid$female ~ spouse_resid$male) %>%
  tidy() %>%
  kbl(format = "latex",
         digits = 3,
         booktabs = TRUE,
         caption = "Correlation between spouse's residuals"
```

```
        ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 20: Correlation between spouse's residuals

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.231	19.914	0	7053	0.209	0.253	Pearson's product-moment correlation	two.sided

We have evidence to reject the null hypothesis that the true correlation is equal to 0 at the 1% level. This is what I expected – I would have guessed that the spouses residuals would be correlated given that have the same household identifier.

Part C

```
model_8c <- plm(bmi ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                index = "hhid",
                model = "pooling",
                data = married_bmi_pdata)
```

Table 21: Pooled OLS – Cluster robust standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	30.360	0.964	31.495	0
male1	1.663	0.072	23.027	0
educ	-0.255	0.019	-13.187	0
age	0.186	0.037	5.094	0
agesq	-0.002	0.000	-4.036	0
smoke	-1.241	0.109	-11.354	0
logfaminc	-1.001	0.171	-5.864	0
withkid	-0.453	0.112	-4.025	0

We re-ran part (a) with appropriately clustered standard errors. A percentage point increase in `faminc` (`logfaminc`) is associated with a 1 point decrease in `bmi`. If the household has a kid (`withkid` $= 1$), then `bmi` is associated with a 0.453 point decrease.

Part D

```
model_8d <- plm(bmi ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                model = "within",
                data = married_bmi_pdata)
```

We are no longer able to estimate `logfaminc` and `withkid`.

```
# FE v Pooled
pFtest(model_8d, model_8c) %>%
  tidy() %>%
    kbl(format = "latex",
        digits = 3,
        booktabs = TRUE,
        caption = "Fixed effects test: H0:'No fixed effects'"
```

Table 22: FE – Robust standard errors

term	estimate	std.error	statistic	p.value
male1	1.784	0.078	22.737	0.000
educ	0.054	0.031	1.725	0.085
age	0.279	0.083	3.351	0.001
agesq	-0.003	0.001	-3.828	0.000
smoke	-1.367	0.176	-7.776	0.000

```
        ) %>%
  kable_styling(latex_options = "hold_position")
```

```
## Multiple parameters; naming those columns df1, df2
```

Table 23: Fixed effects test: H0:'No fixed effects'

df1	df2	statistic	p.value	method	alternative
7052	7050	1.628	0	F test for individual effects	significant effects

We can compare to part (c), and see from the above that the null hypothesis of no fixed effects is rejected.

Question 9

Part A

```
model_9a <- plm(obese ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                model = "pooling",
                data = married_bmi_pdata)
```

Table 24: Pooled OLS – Cluster robust standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	0.536	0.079	6.791	0
male1	0.054	0.007	8.027	0
educ	-0.018	0.002	-10.847	0
age	0.014	0.003	4.355	0
agesq	0.000	0.000	-3.652	0
smoke	-0.078	0.009	-8.206	0
logfaminc	-0.073	0.014	-5.157	0
withkid	-0.043	0.010	-4.483	0

A1

```
pred_9a <- predict(model_9a)

pred_9a_df <- data.frame(
  "min" = min(pred_9a),
```

```
  "max" = max(pred_9a)
)

pred_9a_df %>%
  kbl(format = "latex",
         digits = 3,
         booktabs = TRUE,
         caption = "In-sample predicted obesity probabilities"
         ) %>%
  kable_styling(latex_options = "hold_position")
```

Table 25: In-sample predicted obesity probabilities

min	max
0.004	0.594

No, we do not have a problem with predicted values outside of the [0,1] range. See the table above to for the minimum and maximum values of the in-sample predicted obesity probabilities.

A2

A 1-unit increase in `smoke` is associated with a 7.8% decrease in the probability of obesity, holding all else equal.

A3

A 1-unit increase in `logfaminc` is associated with a 0.073% decrease in the probability of obesity, holding all else equal.

Part B

```
model_9b <- glm(obese ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                family = binomial(link = "probit"),
                data = married_bmi_pdata)
```

Table 26: Probit – Cluster standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	0.166	0.244	0.682	0.495
male1	0.174	0.023	7.437	0.000
educ	-0.057	0.005	-11.305	0.000
age	0.044	0.010	4.501	0.000
agesq	0.000	0.000	-3.848	0.000
smoke	-0.243	0.031	-7.761	0.000
logfaminc	-0.227	0.040	-5.684	0.000
withkid	-0.134	0.028	-4.758	0.000

B1

```
pred_9b <- predict(model_9b)

cor(pred_9a, pred_9b)
```

```
## [1] 0.9998661
```

They are almost perfectly correlated.

B2

Table 27: dydx

factor	AME
logfaminc	-0.071

The average partial effect of `logfaminc` is almost identical to part (a).

Part C

```
model_9c <- plm(obese ~ male + educ + age + agesq + smoke + logfaminc + withkid,
                model = "within",
                data = married_bmi_pdata)
```

Table 28: FE – Robust standard errors

term	estimate	std.error	statistic	p.value
male1	0.067	0.007	9.187	0.000
educ	0.003	0.003	0.981	0.327
age	0.017	0.008	2.072	0.038
agesq	0.000	0.000	-2.760	0.006
smoke	-0.095	0.016	-5.953	0.000

C1

In this model, controlling for couples eliminates variance in education as class-replication and shared behavior among couples will likely deter any education effects for obesity.

C2

Smoking (`smoke` $= 1$) is associated with a 10% decrease in the probability of being obese. I prefer this estimate from part (b) because the structure of the model controls for couples who will likely influence the habits of each other.

C3

You cannot consistently estimate the obesity probability for a specific individual because the panel construction, and the model parameterization, is such that we are controlling for couple groups, rather than individuals.

C4

We are able to consistently estimate the difference between the husband's obesity probability and his wife's obesity probability because we are controlling for sex and couple group. This allows us to parse out within-group effects.