GVPT728 HW#3

Yael Beshaw

2025-01-09

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
##
## The following object is masked from 'package:dplyr':
##
##     add_rownames
##
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```
library(tidycensus)
library(fixest)
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

Download Replication Data (Sorensen, 2019)

```
library(tidyverse)

turnout_data<- dataverse::get_dataframe_by_name(
  filename = 'AggregateReplicationTVData.dta',
  .f = haven::read_dta,
  dataset = '10.7910/DVN/QGMHHQ',
  server = "dataverse.harvard.edu")

# filtering for 1963 ONLY
turnout_data<-turnout_data|>
  filter(nationalelection==0)|>
  filter(year == 1963)|>
  mutate(CountyId = factor(CountyId),
         knr = factor(knr)
         )
```

Baseline Model

```
model0<-lm(turnout ~ TVdummy + logpop + education + settlement + voterpct, data=turnout_data)
summary(model0)
```

```
##
## Call:
## lm(formula = turnout ~ TVdummy + logpop + education + settlement +
##     voterpct, data = turnout_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33227 -0.03602  0.00658  0.04110  0.16886
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3463048  0.0712102   4.863 1.60e-06 ***
## TVdummy      0.0374761  0.0064089   5.848 9.63e-09 ***
## logpop       0.0057265  0.0046588   1.229 0.219651
## education    0.0112788  0.0041475   2.719 0.006793 **
## settlement  -0.0554851  0.0145962  -3.801 0.000164 ***
## voterpct     0.0060616  0.0009477   6.396 4.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06331 on 448 degrees of freedom
## Multiple R-squared:  0.338,  Adjusted R-squared:  0.3306
## F-statistic: 45.74 on 5 and 448 DF,  p-value: < 2.2e-16
```

Question #1 Estimate 3 new versions of model0 that account for correlations across levels of CountyId:

A model with cluster robust standard errors A fixed effects model A random effects model

```
# Cluster Robust Standard Error
model0_robust <- coeftest(model0,
```

```
                          vcov = vcovCL,
                          type='HC2',
                          cluster = ~CountyId
                          )

# Fixed Effects
model0_fixed <-feols(turnout ~ TVdummy + logpop + education +
                       settlement + voterpct | CountyId,
                     data=turnout_data)

(model0_fixed)
```

```
## OLS estimation, Dep. Var.: turnout
## Observations: 454
## Fixed-effects: CountyId: 19
## Standard-errors: Clustered (CountyId)
##             Estimate Std. Error   t value Pr(>|t|)
## TVdummy     0.017321   0.010772  1.607970 0.125240
## logpop      0.010558   0.007306  1.445033 0.165633
## education   0.006387   0.006042  1.057005 0.304490
## settlement -0.032702   0.014454 -2.262419 0.036278 *
## voterpct   -0.000541   0.001657 -0.326576 0.747755
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.050902     Adj. R2: 0.543073
##                  Within R2: 0.155053
```

```
# Random Effects
model0_random <- lmer(turnout ~ TVdummy + logpop + education +
                        settlement + voterpct + (1| CountyId),
                      data=turnout_data)
summary(model0_random)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: turnout ~ TVdummy + logpop + education + settlement + voterpct +
##     (1 | CountyId)
##    Data: turnout_data
##
## REML criterion at convergence: -1291.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7141 -0.5351  0.0487  0.6562  3.2061
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  CountyId (Intercept) 0.001821 0.04267
##  Residual             0.002736 0.05230
## Number of obs: 454, groups:  CountyId, 19
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.682127   0.078732   8.664
## TVdummy      0.020509   0.006662   3.078
## logpop       0.010109   0.004097   2.467
```

```
## education    0.007219   0.004004   1.803
## settlement  -0.033223   0.013307  -2.497
## voterpct     0.000227   0.001051   0.216
##
## Correlation of Fixed Effects:
##            (Intr) TVdmmy logpop eductn sttlmn
## TVdummy    -0.004
## logpop     -0.492 -0.259
## education   0.042  0.022 -0.229
## settlement -0.355 -0.062  0.502  0.390
## voterpct   -0.858  0.093  0.025 -0.067 -0.005
```

Include your output in a formatted regression table and briefly discuss the differences between your results.

```
library(modelsummary)
```

```
## `modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
##   backend. Learn more at: https://vincentarelbundock.github.io/tinytable/
##
## Revert to `kableExtra` for one session:
##
##   options(modelsummary_factory_default = 'kableExtra')
##   options(modelsummary_factory_latex = 'kableExtra')
##   options(modelsummary_factory_html = 'kableExtra')
##
## Silence this message forever:
##
##   config_modelsummary(startup_message = FALSE)
```

```
##
## Attaching package: 'modelsummary'
```

```
## The following object is masked from 'package:tidycensus':
##
##     get_estimates
```

```
model_list<-list("Original" = model0, "Robust SE" = model0_robust,
                 "Fixed Effects" = model0_fixed,
                 "Random Effects" = model0_random)
modelsummary(model_list, output = "huxtable")
```

Differences between the results

Model One: Cluster Robust SE

Highest AIC: -327.5 Highest BIC: 1517.4

The model with cluster robust standard errors does not change the value of our estimates. However, it impacts the statistical significance of our predictors. For example, education in the model with cluster robust SE's is no longer significant as it was in the original model. Following this pattern, the level of signifiance for the rest of our predictors are also decreased here. Compared to the original model, our AIC is decreased while our BIC is increased. Compared to the other models, the cluster model has the overall highest AIC and BIC.

Model Two: Fixed Effects

Lowest AIC: -1367.5 Lowest BIC: -1268.7

RMSE: 0.050902 Adj. R2: 0.543073 Within R2: 0.155053 R2 Within Adj: 0.145 RMSE:0.05

The model of fixed effects has a varied effect on our estimates. We see here that in this model, the only

predictor that remains significant is settlement. Compared to the original model we have a greater R2 and Adjusted R2, which means the model is able to explain a larger percentage of the variation. Additionally, in comparison with the original model, we see that this model has a lower AIC and BIC, and is the lowest among all model versions. Lastly, it has a lower RMSE than the original model, indicating that it is a more accurate model. Overall,this model controls for county and we see that this model explains about 15% of the within variation, allowing us to see a more accurate model.

Model Three: Random Effects

Middle Range BIC and AIC; -1242.7 and -1275.7 CountyId (Intercept): 0.04267 Residual: 0.05230

R2 Marg: 0.138
R2 Cond: 0.482 RMSE:0.05

The model of random effects has a varied effect on our estimates, simillar to the fixed effects. As expected, we see that the estimates are pulled up or down according to their value towards the global mean. Compared to the original model the BIC and AIC of the random effects model are both lower, but are in the middle compared to the the rest of the models. Similar to the fixed effects model, the RMSE is also lower, indicating more precision compared to the original model. This model also shows us that the fixed effects only explain about 14% of the variance compared to the fixed+random which explains almost 50%. When we directly analyze the random effects, we see that the variance for CountyID is 0.002 while that of the residual is 0.003. This indicates to us that there isn't much variance across CountyID. When we assess the standard deviation, we see that the standard deviation of the residual is greater than that of CountyId, meaning it is not more predictive.

Question #2

Which of the models above seems like a better approach for this analysis? Briefly discuss some pros and cons for each one.

Best Approach: Fixed Effects Model

The fixed effects model is the better appraoch as we have many pros. For example, it has both a lower BIC and AIC overall, indicating that there is a better fit. Additionally, it has a greater R2 and adjusted R2, which means that this model is able to explain more of the variance compared to the original model or any other one. Lastly, it has a greater RMSE than the original model, indicating that it is more accurate.

The main con is that the fixed effects alone only explains about 10-15% of the within variation, meaning that random effects may have an important impact on our results.

Pros and Cons of Model One: Cluster Robust SE

Pros: It keeps the large values of our estimates that were in the original model and their significance except for education. Additionally, it seems to pull our estimates towards the mean, which helps us standardize our results.

Cons: Highest AIC and BIC which means it has the worst model fit across the board.

Pros and Cons of Model Three: Random Effects

Pros: The AIC and BIC are lower than the original model but not the lowest out of all models. Allows us to see the the fixed and random effects between and within the CountyId, which provides more information and context.

Cons: It pulls the estimate values towards the global mean, which is useful but our data is not so small that it benefits greatly from this. When we assess the standard deviation, we see that the standard deviation of the residual is greater than that of CountyId, meaning it is not more predictive model overall.

	Original	Robust SE	Fixed Effects	Random Effects
(Intercept)	0.346	0.346		0.682
	(0.071)	(0.138)		(0.079)
TVdummy	0.037	0.037	0.017	0.021
	(0.006)	(0.012)	(0.011)	(0.007)
logpop	0.006	0.006	0.011	0.010
	(0.005)	(0.008)	(0.007)	(0.004)
education	0.011	0.011	0.006	0.007
	(0.004)	(0.007)	(0.006)	(0.004)
settlement	-0.055	-0.055	-0.033	-0.033
	(0.015)	(0.016)	(0.014)	(0.013)
voterpct	0.006	0.006	-0.001	0.000
	(0.001)	(0.002)	(0.002)	(0.001)
SD (Intercept CountyId)				0.043
SD (Observations)				0.052
Num.Obs.	454	454	454	454
R2	0.338		0.566	
R2 Adj.	0.331		0.543	
R2 Marg.				0.138
R2 Cond.				0.482
R2 Within			0.155	
R2 Within Adj.			0.145	
AIC	-1209.5	-327.5	-1367.5	-1275.7
BIC	-1180.7	1517.4	-1268.7	-1242.7
ICC				0.4
Log.Lik.	611.746			
F	45.741			
RMSE	0.06		0.05	0.05
Std.Errors			by: CountyId	
FE: CountyId			X	