# Model Specification and Results

In this chapter I do a step-by-step construction and fitting of a series of models. I begin with a thorough analysis of my notation, and specification of models. I then extract results from the models that best fit the data, and draw inferences on my hypotheses. I will start with a disclaimer: why the data available to me is not necessarily enough to get the necessary causal leverage for significant results. This is particularly true of individual level models. This disclaimer should be considered a large part of the analytical results of my thesis; given several months of exploratory data analysis, such explanations should serve future research into Colorado voter files. I will then proceed to construct some actual models–disclaimer notwithstanding–starting with county and proceeding to individual level modelling. Some of these models may be speculative, as given the initial disclaimer they require more data, or more processing power to actually run. I will still include them as they may be used for future research.

## Salt

The following models should be taken with a grain of salt because of a series of reasons. In this section, I will tackle these reasons one by one and then analyze what steps could be made to compensate.

### Causal Leverage

Causal leverage usually means having enough data to draw significant conclusions about correlation of variables, or in the best cases make safe causal inferences. Data that presents causal leverage should have certain characteristics. It should, first, be extensive enough. By this I mean that the raw number of observations should be as high as possible, and in the best cases significantly larger than the set of predictors that are present. This not only guarantees no issues with model matrices when running statistical models, but also that there will be enough data-per-variable to draw conclusions. Second, the data should be varied. To put it very simply, it's not enough to have hundreds of thousands of observations if they are all similar to eachother. If, for example, my data included a thousand people in Jefferson county, and 63 in all other counties of Colorado combined–one in each remaining county–, then I would not be able to leverage my data to draw conclusions on county-level effects.

As previously stated, I have registration files going back to 2012. From these files, I have extracted data for elections going back to 2010 [1]. In order to make inferences on VBM and turnout effects, given the previous criteria for causal leverage, I had to have extensive and varied data. I have extensive data–over 35 million observations at the individual level–but the data substantially lacks variance in voting method. Put simply, most people in Colorado from 2010 onward either did not vote at all, or voted by mail. If you recall the changes in Colorado election law, in 2008 counties were allowed to conduct all mail elections, and no-excuse permanent absentee voting was implemented state-wide; then in 2013 Colorado transitioned to full VBM for all elections. This means that few people were still using traditional polling places or vote centers to cast their ballots. Figure 4.1 shows how, after 2013, and even before that in 2011–the coordinated, local election for which mail ballots were more convenient for counties–over 95% of ballots cast were mail ballots. Only in the general elections of 2010 and 2012 there is some variance, but mail ballots account for well over two thirds of total votes.

This issue is not completely fatal for my county level models. There is still variance between counties that have 100% mail ballots and those that are around the 75-85% margin. For individual level models–where I am estimating voting probability–VBM will be an almost perfect predictor for voting, and therefore will not present me with any substantial analytical result on how it affects voting probability. There are some ways to compensate for this issue, which I outline; due to time or data constraints, not all of these will be implemented in this thesis:

[1] See section 3.3.1; I extracted data limited to this time period to avoid accuracy issues with migration and removal of inactive/unavailable voters

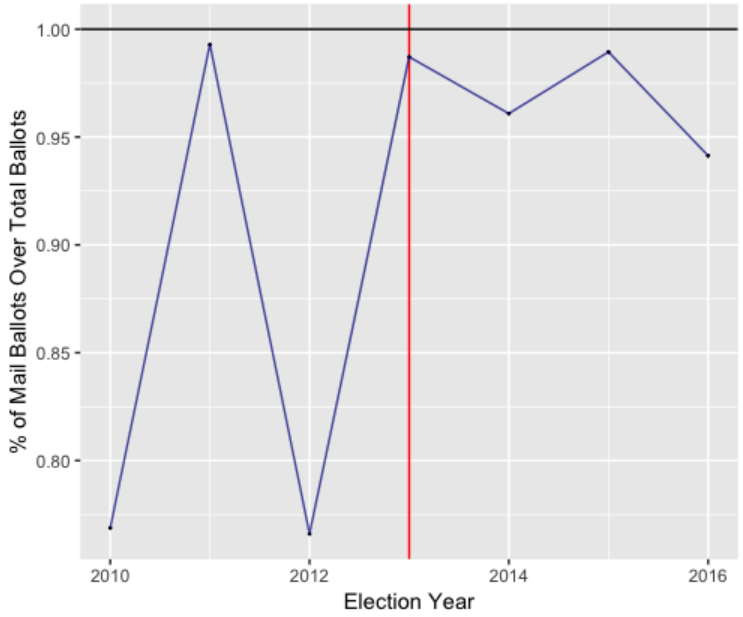



Figure 1: Percentage of mail ballots over total ballots by year

- *More (Diverse) Data*:
- *Localized, Natural Experiment Studies*:
- *Synthetic Control Group*:

**Lack of Individual Data**

**Processing Power**

## Models

### Variable Specification

I will not go through each individual variable in this section, but will briefly describe my procedure on notation for the following models. I will include more comments whenever they seem necessary under each model. In this thesis I include predictors on a series of variables that can be divided into five categories based on unit of observation: county, election, individual, local result, and ballot. The last two are functions of other units: local result units are equal to the product of elections and counties, while ballot units are equal to the number of unique individuals multiplied by the number of elections each of them was registered in. For notation, I follow this set of rules:

1. If the variable is a response, it is coded $y$.
2. If the variable is a predictor, it is coded according to Table 4.1.
3. The variable's superscript will provide information on what it represents, else it will be explained.
4. All variables represent a single value of that variable unless stated otherwise.
5. Unit of observation will also be specified in subscript, according to the indices described in Table 4.1. These indices are also used in sum notation.
6. All Greek characters represent coefficients to be calculated.
7. By $k[j]$ I represent the k-value of the j-observation. In this case, this would be the county that an individual is registered in.

8. Note that for Local Result level variables, I use $k, l$ as an indice. This is because there are very few variables at this level, it is a direct multiplicative product of two other units, and this notation avoids confusion with even more indice types.

Table 1: Variable names and indices per unit of observation

| Units | Variable | Index |
|---|---|---|
| Ballot | u | i |
| Individual | z | j |
| County | x | k |
| Election | w | l |
| Local Result | v | k,l |
| General Index | - | i' |

## County Level Models

### Specifications

In this section I will go through a step-by step creation of models at the county level. County level models use a series of variables at the election, county, and local result levels. The response variable is always turnout as a local result. If this model is considered at its most basic, it could be thought of as an assignment of voting tendencies across counties; each county independent of election has a unique range of turnout results. In this way it is possible to build a naive, baseline model of turnout as follows:

$$y_{k,l}^{turnout} = \beta_0 + (\sum_{k=1}^{64} \beta_k x_k^{county}),$$

where $x_k^{county}$ is a series of 64 dummy variables for each county of Colorado. Here differences between elections come from normally distributed error terms, rather than predictors. I name this *Model 1*, and it does not reflect the data particularly well. First off, this model includes the assumption that counties are independent of one another, which is probably false; just consider that these counties are areas of the same state, in the same country, with populations moving between them at regular intervals, and many of them covering the same metropolitan area or congressional district. Additionally, this model cannot fully calculate relevant coefficients, since a number of counties can be represented as perfect linear functions of the other variables. This means they will be dropped by *R* when the model is called in the `lm()` function.

A way to fix both these issues is to use a multilevel model with mixed effects for county. By constraining coefficients at the county level to a set distribution, this model does away with the assumption of independence. The other county level predictors help to explain some of the unexplained group level variation, which reduces the standard deviation of county coefficients and helping provide more exact estimates [@gelman_data_2006]. I call this *Model 2*, which can be written as:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban},$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

This model provides a more reasonable set of estimates for each county, but still fails at providing any sort of guess as to secular trends, time-specific effects, election type effects, or mail voting–the variable of interest. I will amend this by adding a set of variables at the election and local result levels: election type and an interaction term between election type and mail voting. This variable should reflect whether turnout effects

of mail voting are more pronounced in a specific type of election. I call this *Model 3* and it can be specified as follows:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban} + (\sum_{i'=1}^{4} \beta_{i'+3} w_{i'}^{electiontype}) * (\beta_3 v_{k,l}^{\%mail\ vote} + 1),$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

where $w_{i'}^{electiontype}$ is a series of four dummy variables for each type of election (General, Primary, Coordinated, Midterm). This model reflects nearly all the information I have available, apart from election date. For the incorporation of election dates there are two possible alternatives. First, I can simply add a dummy variable for each year. This would assume independence between each year, as it would specify different, independent "slopes" for the seven years I have data for–this is like calculating seven different models, one for each year. This is not particularly elegant as a solution nor does it reflect the fact that years actually are interconnected; of course there can be massive shifts in national or regional political climates, but those shifts happened *from some baseline*, which is reflected in previous years.

These elections can be thought of as systems for which prior condition affects future outcomes, and therefore time cannot be modeled as a series of independent effects. The solution here is adding a spline function for time, using a general additive multilevel model. The most commonly used spline function, and the default in the `gamm4` *R* package is a thin plate regression spline, which I also use here [@wood_generalized_2006]. More on the subject of splines can be found in the Wood (2006) textbook. The model, which I call *Model 4* can be written as follows:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban} + (\sum_{i'=1}^{4} \beta_{i'+3} w_{i'}^{electiontype}) * (\beta_3 v_{k,l}^{\%mailvote} + 1) + s(w_l^{year}),$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

where $s()$ is a thin plate spline function with seven knots–equal to the number of years.[2] A summary of these four models is provided in the following table:

Table 2: County level model descriptions

| Model No | Model Description |
|---|---|
| Model 1 | Naive model with only county specific effects |
| Model 2 | Multilevel model; added county level predictors |
| Model 3 | Multilevel model; added VBM, interaction terms, and election fixed effects |
| Model 4 | Multilevel General Additive model; added spline function for election year |

### Results

Calling model one results in the following:

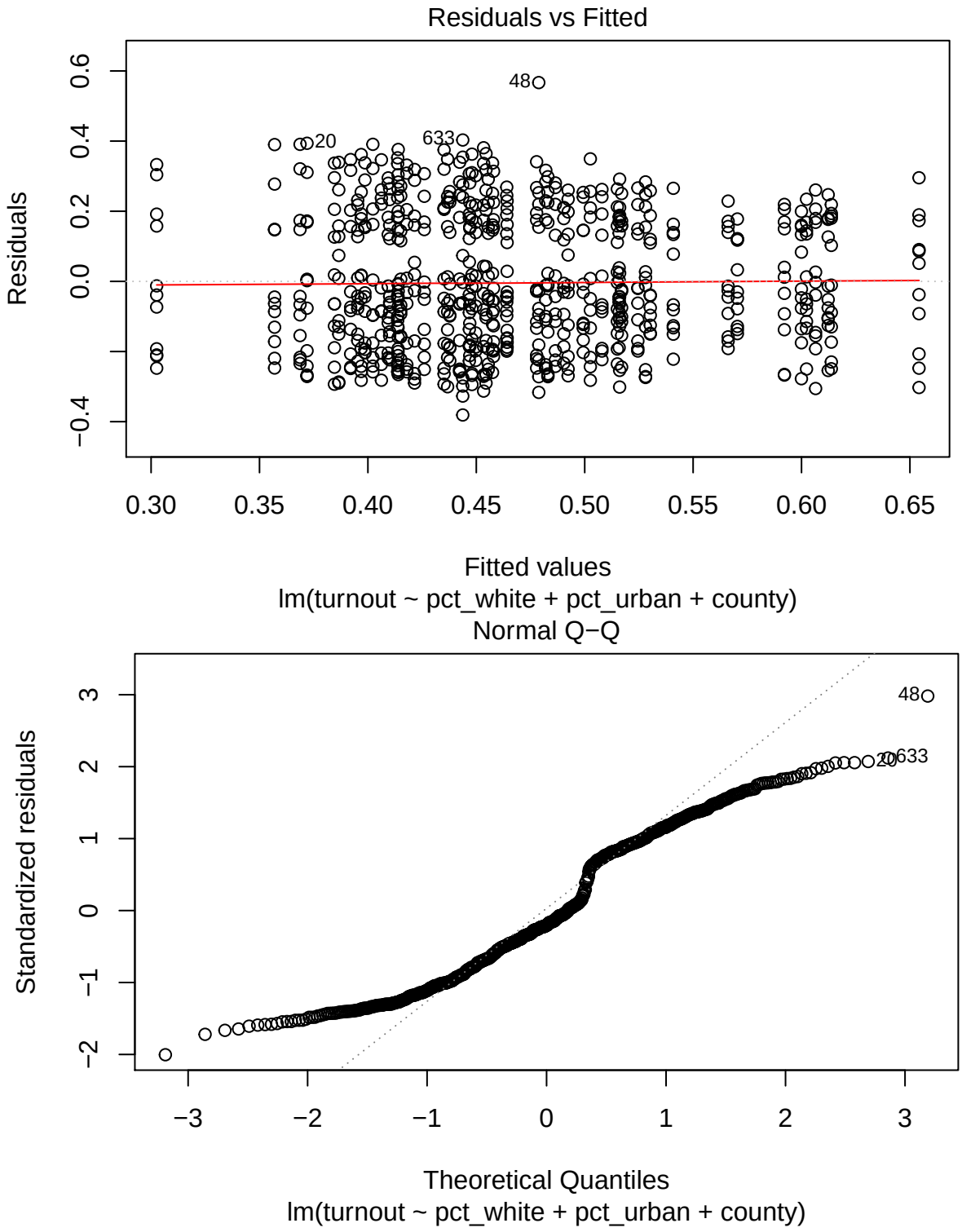
```
md_1 <- lm(data = model_dt, turnout ~ pct_white + pct_urban + county)

summary(md_1)
```

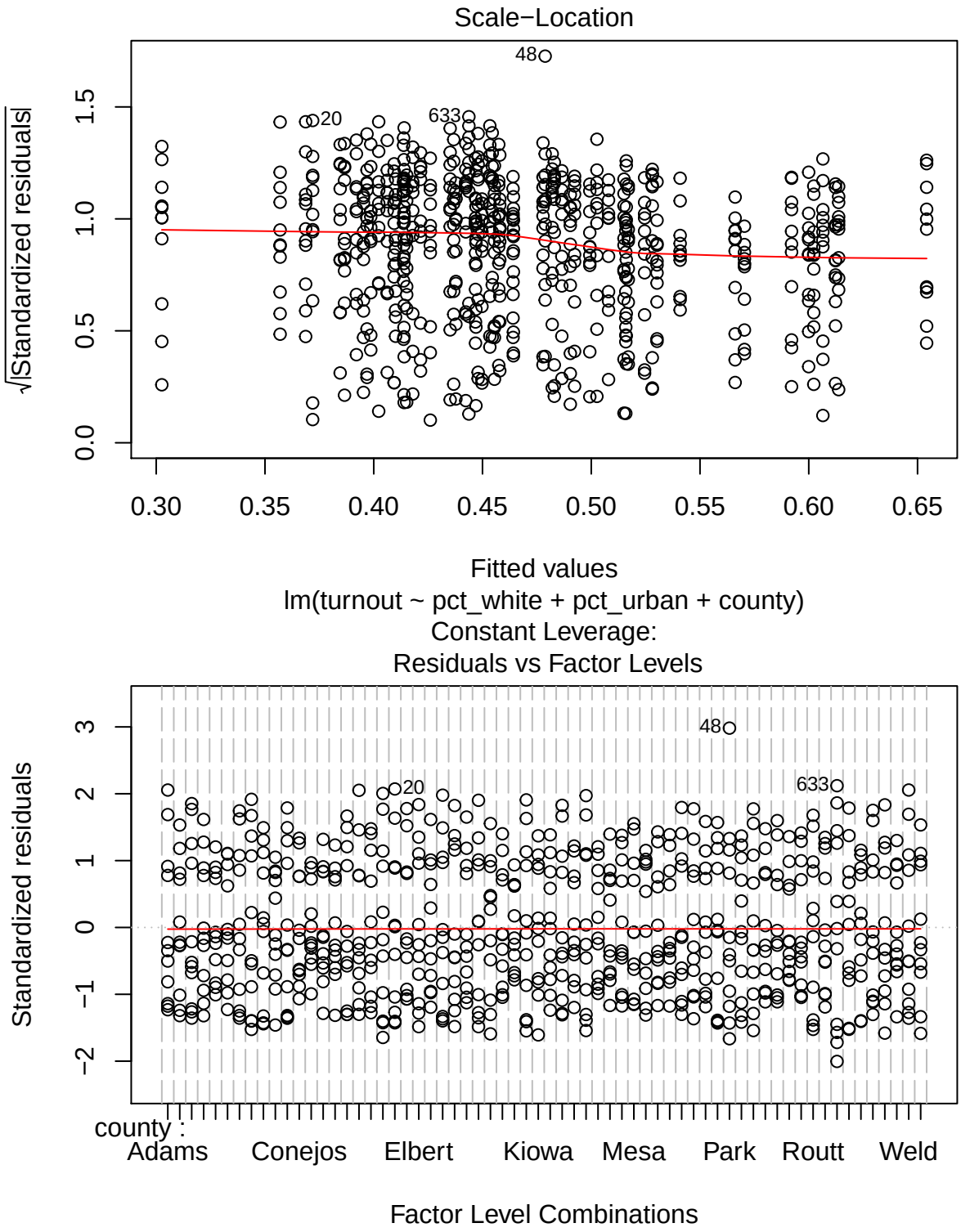
[2] I used the `gam.check()` function that is present in the `mgcv` *R* package, whose call determined that the number of knots here may be too low. However, given the data available to me, I was limited to the inclusion of seven years and as such cannot increase the number of knots any further.

```
##
## Call:
## lm(formula = turnout ~ pct_white + pct_urban + county, data = model_dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38095 -0.15980 -0.03896  0.17156  0.56682
##
## Coefficients: (2 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.684609   0.930859   0.735   0.4623
## pct_white         -0.091375   1.004528  -0.091   0.9276
## pct_urban         -0.277328   0.393461  -0.705   0.4812
## countyAlamosa     -0.008657   0.205569  -0.042   0.9664
## countyArapahoe     0.059868   0.108177   0.553   0.5802
## countyArchuleta   -0.074504   0.080693  -0.923   0.3562
## countyBaca        -0.038293   0.125250  -0.306   0.7599
## countyBent         0.031705   0.125245   0.253   0.8002
## countyBoulder      0.058465   0.221585   0.264   0.7920
## countyBroomfield   0.118109   0.252452   0.468   0.6401
## countyChaffee      0.096121   0.189179   0.508   0.6116
## countyCheyenne    -0.004180   0.124061  -0.034   0.9731
## countyClear Creek -0.158121   0.117195  -1.349   0.1777
## countyConejos     -0.147084   0.518922  -0.283   0.7769
## countyCostilla    -0.140564   0.626891  -0.224   0.8227
## countyCrowley     -0.114716   0.363623  -0.315   0.7525
## countyCuster       0.001746   0.117203   0.015   0.9881
## countyDelta       -0.024595   0.093877  -0.262   0.7934
## countyDenver      -0.002660   0.086058  -0.031   0.9753
## countyDolores     -0.115033   0.117797  -0.977   0.3292
## countyDouglas      0.095437   0.272803   0.350   0.7266
## countyEagle       -0.029321   0.084352  -0.348   0.7283
## countyEl Paso      0.032358   0.153918   0.210   0.8336
## countyElbert      -0.098869   0.117751  -0.840   0.4014
## countyFremont      0.039704   0.169325   0.234   0.8147
## countyGarfield     0.002705   0.083721   0.032   0.9742
## countyGilpin      -0.195298   0.117816  -1.658   0.0979 .
## countyGrand       -0.096441   0.107068  -0.901   0.3681
## countyGunnison    -0.091462   0.146485  -0.624   0.5326
## countyHinsdale     0.054785   0.117729   0.465   0.6418
## countyHuerfano     0.018049   0.161901   0.111   0.9113
## countyJackson     -0.034279   0.126263  -0.271   0.7861
## countyJefferson    0.094609   0.234286   0.404   0.6865
## countyKiowa        0.007118   0.117766   0.060   0.9518
## countyKit Carson   0.042564   0.080601   0.528   0.5976
## countyLa Plata    -0.108083   0.085798  -1.260   0.2082
## countyLake        -0.044992   0.110657  -0.407   0.6844
## countyLarimer      0.072687   0.260269   0.279   0.7801
## countyLas Animas  -0.006712   0.177205  -0.038   0.9698
## countyLincoln     -0.071115   0.172840  -0.411   0.6809
## countyLogan        0.091242   0.140329   0.650   0.5158
## countyMesa         0.042887   0.243655   0.176   0.8603
## countyMineral      0.016145   0.121265   0.133   0.8941
## countyMoffat       0.007592   0.187606   0.040   0.9677
```

```
## countyMontezuma   -0.114966   0.099263  -1.158   0.2472
## countyMontrose     0.016583   0.091876   0.180   0.8568
## countyMorgan       0.016310   0.089452   0.182   0.8554
## countyOtero       -0.001323   0.134714  -0.010   0.9922
## countyOuray       -0.116254   0.117889  -0.986   0.3244
## countyPark        -0.122114   0.117321  -1.041   0.2983
## countyPhillips    -0.119730   0.173567  -0.690   0.4906
## countyPitkin      -0.065013   0.178769  -0.364   0.7162
## countyProwers     -0.008041   0.096114  -0.084   0.9334
## countyPueblo       0.016798   0.096983   0.173   0.8625
## countyRio Blanco  -0.075465   0.130767  -0.577   0.5641
## countyRio Grande  -0.066331   0.247587  -0.268   0.7889
## countyRoutt       -0.028231   0.201408  -0.140   0.8886
## countySaguache    -0.195286   0.377335  -0.518   0.6050
## countySan Juan    -0.163076   0.136682  -1.193   0.2333
## countySan Miguel  -0.217148   0.122666  -1.770   0.0772 .
## countySedgwick    -0.014318   0.134360  -0.107   0.9152
## countySummit      -0.083251   0.215208  -0.387   0.6990
## countyTeller      -0.061382   0.147748  -0.415   0.6780
## countyWashington   0.009425   0.120239   0.078   0.9375
## countyWeld               NA         NA      NA       NA
## countyYuma               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1994 on 640 degrees of freedom
## Multiple R-squared:  0.1251, Adjusted R-squared:  0.03899
## F-statistic: 1.453 on 63 and 640 DF,  p-value: 0.01564
```

```
#If run this shows perfect linear relationship
#alias(md_1)

plot(md_1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(turnout ~ pct_white + pct_urban + county)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(turnout ~ pct_white + pct_urban + county)
48
20
633




Calling Model two has the following result:

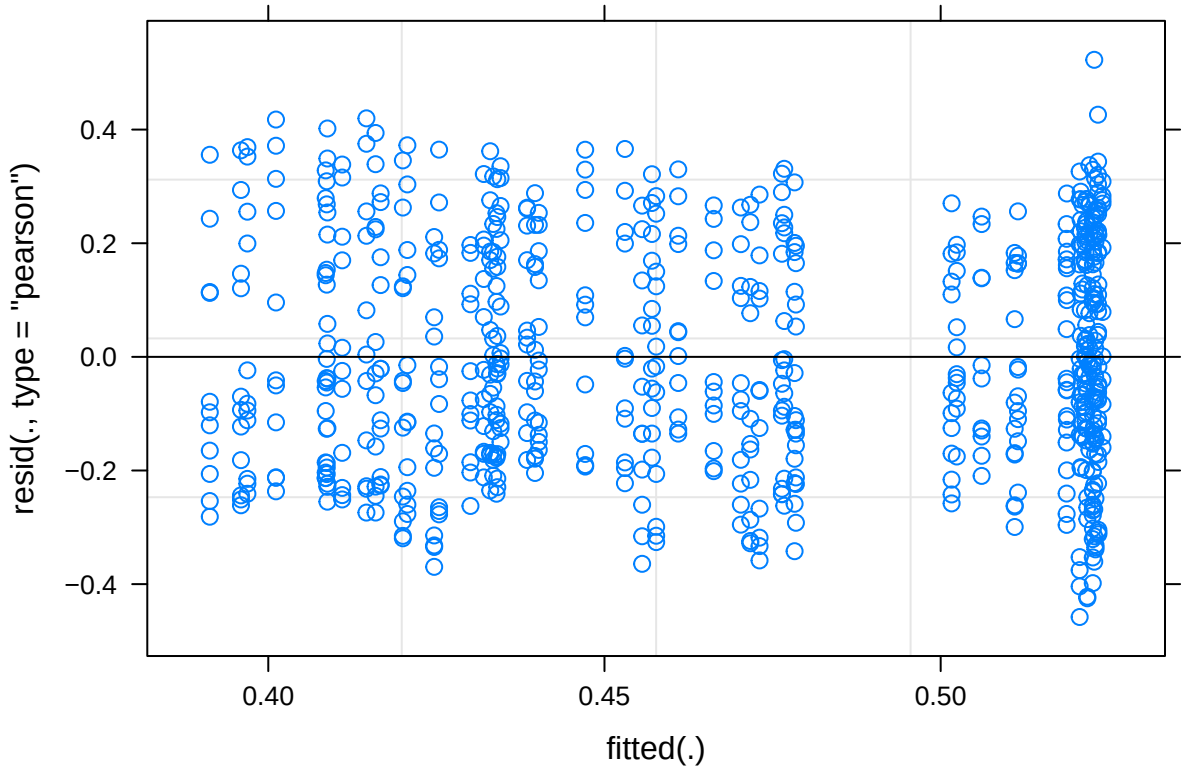
```
md_2 <- lmer(data = model_dt, turnout ~ pct_white + pct_urban + (1|county),
             REML = F)

arm::display(md_2)
```

```
## lmer(formula = turnout ~ pct_white + pct_urban + (1 | county),
##     data = model_dt, REML = F)
##             coef.est coef.se
## (Intercept)  0.49     0.05
## pct_white    0.03     0.05
## pct_urban   -0.12     0.02
##
## Error terms:
##  Groups   Name        Std.Dev.
##  county   (Intercept) 0.00
##  Residual             0.20
## ---
## number of obs: 704, groups: county, 64
## AIC = -270.8, DIC = -280.8
## deviance = -280.8
```

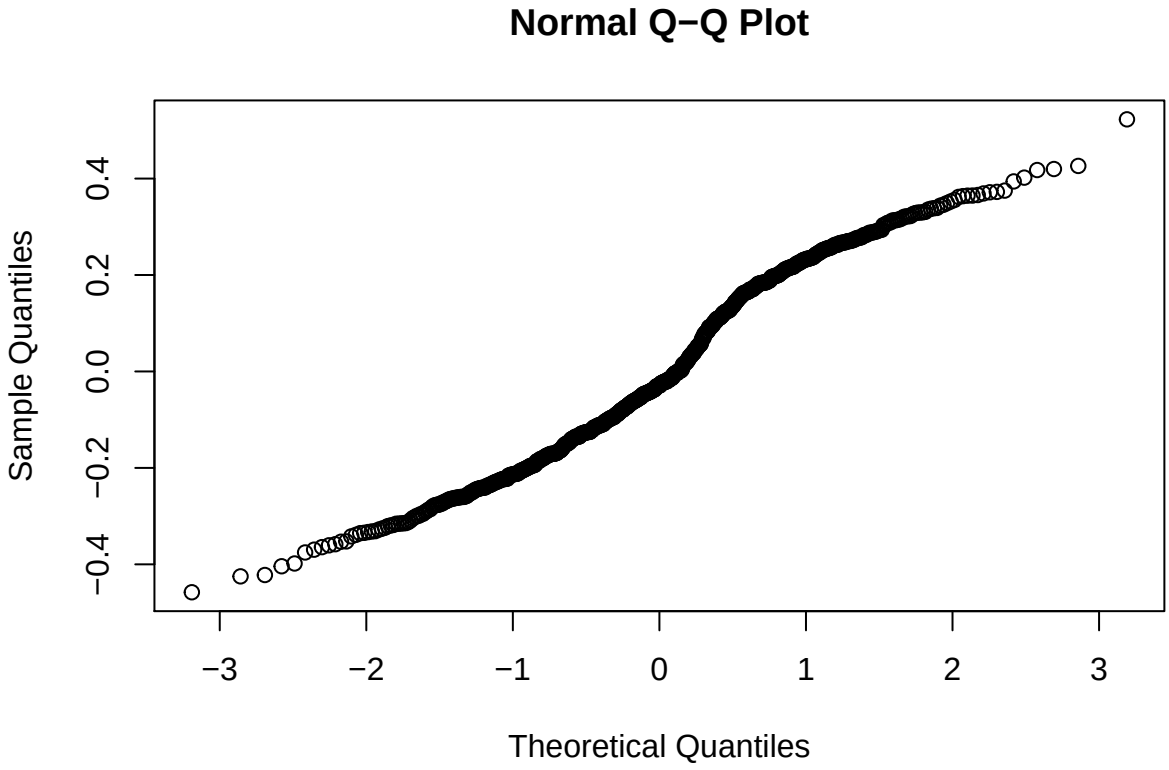
```
#Run for specific county coefs
#ranef(md_2)

fixef(md_2)
```

```
## (Intercept)   pct_white   pct_urban
##  0.49202848  0.03364891 -0.11827707
```

```
plot(md_2)
```




```
qqnorm(residuals(md_2))
```




Calling Model three:

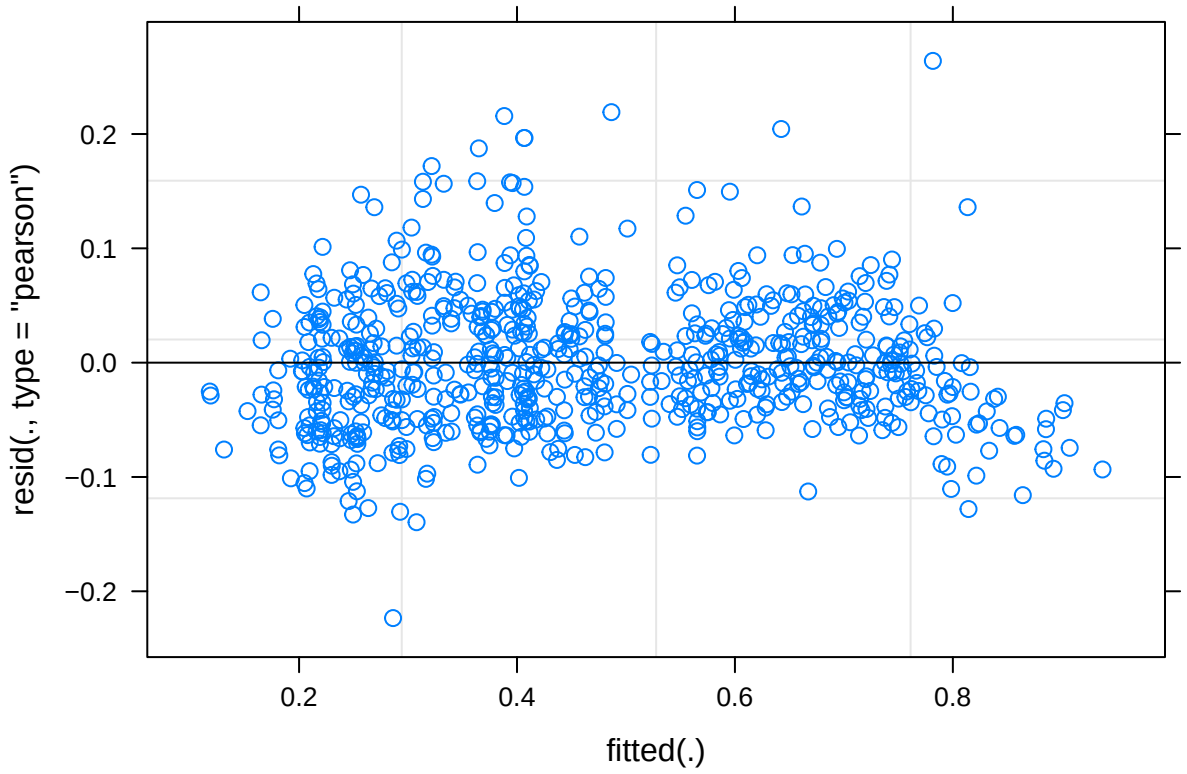
```
md_3 <- lmer(data = model_dt, turnout ~ 1 + types + pct_vbm +
               pct_urban + pct_white + pct_vbm:types + (1|county),
             REML = F)

arm::display(md_3)
```

```
## lmer(formula = turnout ~ 1 + types + pct_vbm + pct_urban + pct_white +
##     pct_vbm:types + (1 | county), data = model_dt, REML = F)
##                       coef.est coef.se
## (Intercept)            0.45     0.08
## typesGeneral           0.19     0.07
## typesMidterm           0.25     0.07
## typesPrimary          -0.07     0.07
## pct_vbm                0.00     0.07
## pct_urban             -0.12     0.02
## pct_white              0.03     0.05
## typesGeneral:pct_vbm   0.15     0.07
## typesMidterm:pct_vbm  -0.06     0.07
## typesPrimary:pct_vbm  -0.09     0.07
##
## Error terms:
##  Groups   Name        Std.Dev.
##  county   (Intercept) 0.05
##  Residual             0.06
## ---
## number of obs: 704, groups: county, 64
## AIC = -1779.2, DIC = -1803.2
## deviance = -1803.2
```

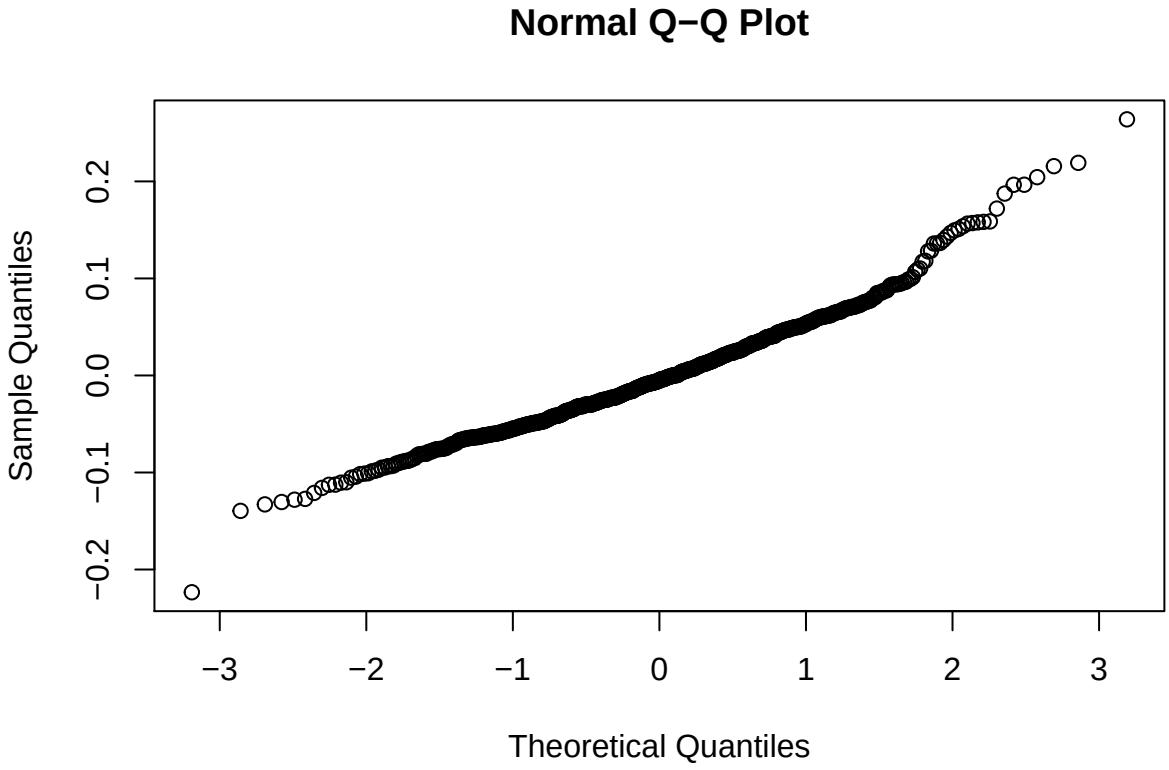
```
#Run for county coefs
#ranef(md_3)

fixef(md_3)
```

```
##          (Intercept)         typesGeneral         typesMidterm
##          0.454642317          0.190033387          0.252085742
##         typesPrimary              pct_vbm            pct_urban
##         -0.070582336         -0.001196972         -0.116841416
##            pct_white typesGeneral:pct_vbm typesMidterm:pct_vbm
##          0.032823780          0.151950039         -0.057236047
## typesPrimary:pct_vbm
##         -0.088157086
```

```
plot(md_3)
```




```
qqnorm(residuals(md_3))
```




```
anova(md_2, md_3)
```

```
## Data: model_dt
## Models:
## md_2: turnout ~ pct_white + pct_urban + (1 | county)
## md_3: turnout ~ 1 + types + pct_vbm + pct_urban + pct_white + pct_vbm:types +
## md_3:     (1 | county)
##      Df      AIC      BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## md_2  5  -270.82  -248.03 140.41  -280.82
## md_3 12 -1779.20 -1724.52 901.60 -1803.20 1522.4      7  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

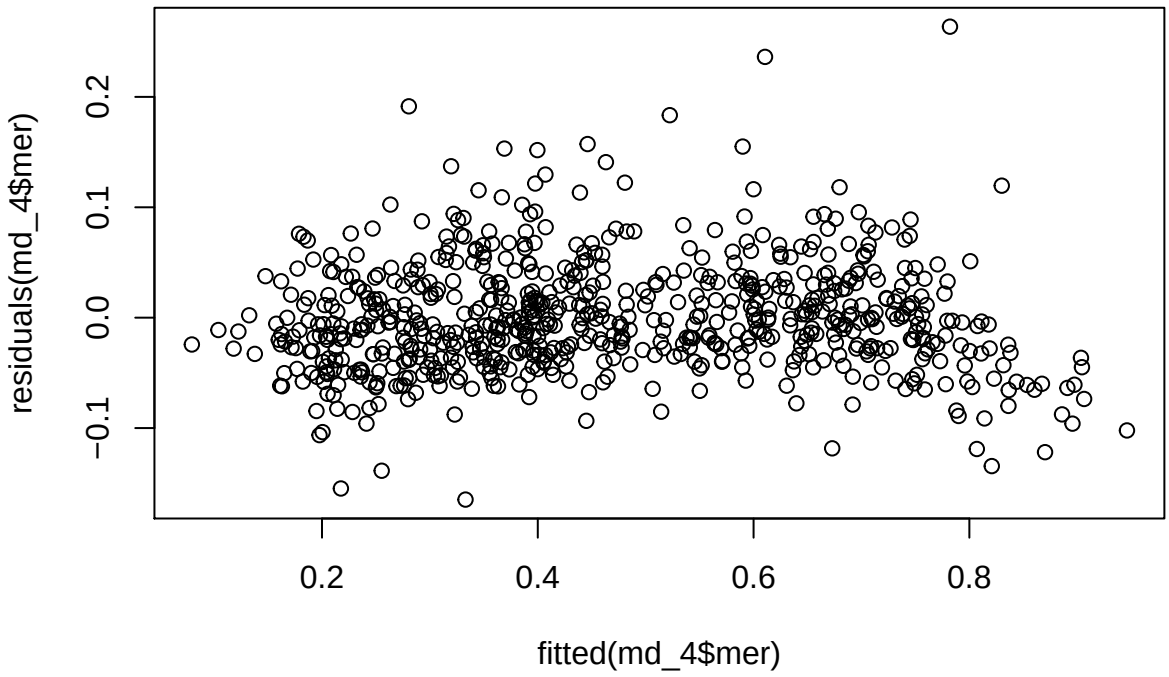
Calling Model 4:

```
model_dt$dates <-  as.integer(model_dt$dates)

md_4 <- gamm4(turnout ~ 1 + types +
                pct_urban + pct_white + pct_vbm*types + s(dates, k = 7),
              random =~ (1|county),
              data = model_dt)

summary(md_4$mer)
```

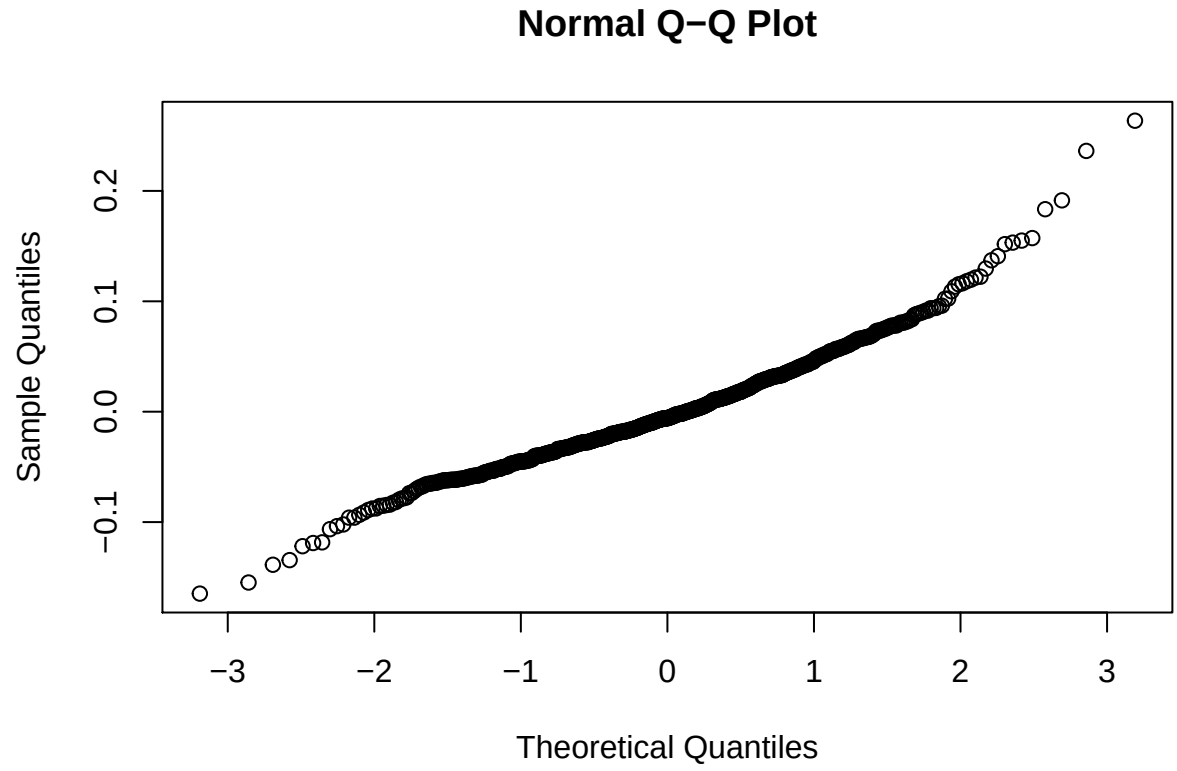
```
## Linear mixed model fit by REML ['lmerMod']
##
## REML criterion at convergence: -1899.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1206 -0.6102 -0.1110  0.5546  4.9920
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept) 0.002962 0.05442
##  Xr       s(dates)    0.967661 0.98370
##  Residual             0.002788 0.05281
## Number of obs: 704, groups:  county, 64; Xr, 5
##
## Fixed effects:
##                          Estimate Std. Error t value
## X(Intercept)             0.469534   0.072036   6.518
## XtypesGeneral            0.254063   0.064603   3.933
## XtypesMidterm            0.070291   0.062977   1.116
## XtypesPrimary           -0.170327   0.061898  -2.752
## Xpct_urban              -0.119413   0.020723  -5.762
## Xpct_white               0.031336   0.050401   0.622
## Xpct_vbm                 0.002371   0.058353   0.041
## XtypesGeneral:pct_vbm    0.085084   0.067613   1.258
## XtypesMidterm:pct_vbm    0.106871   0.064296   1.662
## XtypesPrimary:pct_vbm -0.005732   0.061585  -0.093
## Xs(dates)Fx1             -0.113090   0.019823  -5.705
##
## Correlation of Fixed Effects:
##             X(Int) XtypsG XtypsM XtypsP Xpct_r Xpct_w Xpct_v XtyG:_ XtyM:_
## XtypesGenrl -0.715
## XtypesMdtrm -0.741  0.822
## XtypesPrmry -0.736  0.833  0.882
## Xpct_urban  -0.292 -0.001  0.000  0.005
## Xpct_white  -0.571  0.001  0.000 -0.002  0.336
## Xpct_vbm    -0.792  0.864  0.887  0.883 -0.008 -0.003
## XtypsGnrl:_  0.661 -0.967 -0.747 -0.764  0.000 -0.002 -0.836
## XtypsMdtr:_  0.705 -0.779 -0.968 -0.833 -0.001 -0.001 -0.880  0.751
## XtypsPrmr:_  0.719 -0.808 -0.846 -0.972 -0.005  0.002 -0.903  0.789  0.846
## Xs(dats)Fx1 -0.015  0.116  0.138  0.107  0.011  0.004  0.013 -0.156 -0.146
##             XtyP:_
## XtypesGenrl
## XtypesMdtrm
## XtypesPrmry
## Xpct_urban
## Xpct_white
## Xpct_vbm
## XtypsGnrl:_
## XtypsMdtr:_
## XtypsPrmr:_
## Xs(dats)Fx1 -0.112
```

```
plot(fitted(md_4$mer), residuals(md_4$mer))
```




```
qqnorm(residuals(md_4$mer))
```




## Individual Level Models

### Specifications

For the rest of this write-up, assume the following:

$$y_i \sim \text{Bernoulli}(p\_vote)$$

Where $y_i \in \{0, 1\}$ is the probability that the i-th ballot was completed.

In this section I do not linearly add to the model until it reaches a final stage. The reasoning here is that there is no exact linear path to follow; there is an overarching unit of observation–the ballot–and all the rest are dependent between each other. For instance, adding a variable for Party at the ballot level would not significantly change the way I later add percentage of white residents at the county level. Therefore, the way I proceed is the following: I "build" the models step by step and separatelly for each group of variables (grouping by unit of observation). Then I present one example of what a model using two of these initial "building blocks" would look like. Since this is fairly generalizable, I then proceed directly to the full model which inlcudes all different variables.

If receiving a ballot with no information, I would predict that the probability that an additional ballot was a vote in favor would be equal to turnout, as calculated through all other ballots. Therefore:

$$\hat{p\_vote}_i = \frac{\#\text{votes cast}}{\#\text{ballots}}$$

**Estimation with only one type of data**

There are four levels of data I will go through here: County, Election, Person, and Ballot.

**County Level**

Assume that the ballot I am trying to assess completion for has the name of the county it is from written on it. There are two ways I can think of for predicting $p\_vote$. First, assume that each different county has a different, independent $p\_vote$. Therefore, in model-lingo this would look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_{k,i}\beta_k)$$

Where k counts over the 64 counties of Colorado, and $x_k$ is an indicator variable for each county. If I, quite reasonably, throw away the assumption of independance–these counties are, after all, in the same state and the same country–I could also fit a mixed effects model as such:

$$\hat{p\_vote} \sim \text{logit}^{-1}(a_{k[i]}),$$
$$a_k \sim \text{N}(\gamma_0, \sigma_\alpha^2)$$

Where $\alpha_{k[i]}$ varies by county, constrained by its standard deviation and $\gamma_0$, an intercept coefficient. Let's say now that along with the one ballot, I was given a short list of $n^{\text{county vars}}$ other county-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_k + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_{i'+64})$$

Where $x_{k[i],l}$ is the k-th value of the i'-th variable. If, as before, I do not assume independance, the model can be written as:

$$\hat{p\_vote} \sim \text{logit}^{-1}(a_{k[i]}),$$

$$a_k \sim \mathrm{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$

In the case of my specific data, for the time being I have county-level data for white population and urban population, so $n^{\text{county vars}} = 2$.

**Individual Level**

Assuming that I know the voter ID of the individual that cast their ballot, I can treat this piece of information in about the same way that I did for county as described above. This means that the following is mostly an exercise in maintaining notation constant. For these purposes, let $n^{ID}$ be the number of total unique voter IDs–individuals–that I have data on, and j an indice that sums over all individuals. Also let $z_j$ be an indicator variable for each individual. Then:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{j=1}^{n^{ID}} z_j\beta_j)$$

And the second model, not assuming independence, would be:

$$\hat{p\_vote} \sim \text{logit}^{-1}(\delta_{j[i]}),$$
$$\delta_j \sim \mathrm{N}(\zeta_0, \sigma_\delta^2)$$

Again, in a similar way to county level data, there are variables at an individual level, thus making it relativelly easy to build further models. Let's say now that along with the one ballot, I was given a short list of $n^{\text{indiv vars}}$ other individual-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{j=1}^{n^{ID}} z_j\beta_j + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_{i'+n^{ID}})$$

Where $z_{j[i],l}$ is the j-th value of the i'-th variable. If, as before, I do not assume independance, the model can be written as:

$$\hat{p\_vote} \sim \text{logit}^{-1}(\delta_{j[i]}),$$
$$\delta_j \sim \mathrm{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

In the case of my specific data, for the time being I have individual-level data for gender, so $n^{\text{indiv vars}} = 1$.

**Election Level**

Again as previously, four models come from including election level data. The first two are assuming I only knew what specific election the ballot comes from. Let $w_{i'}$ be an indicator variable for each election and $n^{elect}$ the number of elections. The model assuming independence, with $w_{i'}$ being indicator variables for each election, is:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{l=1}^{n^{elect}} w_l\beta_l)$$

Again, as previously, it would be safe to assume that each election is not held in a vacuum. Adding mixed effects this model would be:

$$\hat{p\_vote} \sim \text{logit}^{-1}(\eta_{l[i]}),$$
$$\eta_l \sim \text{N}(\nu_0, \sigma_\nu^2)$$

Again, in a similar way to county and individual level data, I add in variables at an election level. Let's say now that along with the one ballot, I was given a short list of $n^{\text{election vars}}$ other election-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{l=1}^{n^{elect}} w_l\beta_l + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\beta_{i'+n^{elect}})$$

Where $w_{l[i],i'}$ is the l-th value of the i'-th variable.

Assuming independence:

$$\hat{p\_vote} \sim \text{logit}^{-1}(\eta_{l[i]}),$$
$$\eta_l \sim \text{N}(\nu_0 + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\nu_{i'}, \sigma_\nu^2)$$

For the time being I have two different variables that describe individual elections: date and type. Note that the above models may not be the best way to describe dates! An alternative could be fitting a glm, with some smoothing spline function for year. As for type, this would include four distinct indicators; one for each election type.

### Ballot Level

In this section I assume that the ballot has some key features written on it, like the voting method, age, or party registration of the person that filled it out. A mixed effects model here would make no sense, since all the data is at the same unit of observation. Therefore, when adding ballot level variables, the model would look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\beta_0 + \sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_{l'})$$

Where $u_{i,i'}$ is the i-th value of the i'-th variable, and $n^{\text{ballot vars}}$ is the number of ballot level variables. For now, I have data on voting method, age, and party. Voting method is coded as a binary variable with value one if the method was a Mail Vote. Party includes four distinct indicators for REP, DEM, Other, and Unaffiliated. Age is tricky; for now the options would be: straight up inclusion as an integer, inclusion as a cubic polynomial, unclusion as a 2nd degree polynomial, inclusion in some form of spline function.

### Estimation with two types of data

After the work of setting up the four models at four different levels of observation, combining them in twos should be failry straightforward. To avoid being nedlessly cumulative, I will pursue this combination for County and Individual level only–instead of the six different possible combinations.

With the assumption that both counties and individuals are independent of one another, I proceed to the first type of model:

$$p\_\hat{vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_k + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_{i'+64} + \sum_{j=1}^{n^{ID}} z_j\beta_{j+n^{\text{county vars}}+64} + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_{i'+n^{ID}+n^{\text{county vars}}+64})$$

This is large and clunky. It includes variables as described above: indicators for each county and individual, and all individual or county-level variables. For the corresponding mixed-effects model, I assume the tree-like structure we discussed on Monday. The hierarchy has two "levels", with the second level consisting of two different regressions:

$$p\_\hat{vote} \sim \text{logit}^{-1}(\delta_{j[i]} + a_{k[i]}),$$
$$a_k \sim \text{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$
$$\delta_j \sim \text{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

**Estimation with the full dataset**

I now proceed to include variables from all units of observation into one model. The first model, assuming independence, is:

$$p\_\hat{vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_* + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_* + \sum_{j=1}^{n^{ID}} z_j\beta_* + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_* + \sum_{l=1}^{n^{elect}} w_l\beta_* + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\beta_* + \sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_*)$$

You will notice that I have ommited the subscript for all beta coefficients. This is because after two or three parameters, this becomes very, very large. I think it's reasonable to assume increasing indexes for different beta coefficients from left to right in this expression.

The mixed effects model will again operate on two "levels" of hierarchy, but the second level will now include three distinct regressions. Caveats for variables like age and date should be noted from previous sections.

$$p\_\hat{vote} \sim \text{logit}^{-1}(\sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_{l'} + \delta_{j[i]} + a_{k[i]} + \eta_{l[i]}),$$

$$a_k \sim \text{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$

$$\delta_j \sim \text{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

$$\eta_l \sim \text{N}(\nu_0 + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\nu_{i'}, \sigma_\nu^2)$$