

# PHP 2514 HW 6

*Blain Morin*

*May 8, 2018*

## Gelman, Hill Chapter 13 #5

(a)

We start by creating a by patient varying intercept and slopes model which predicts cd4 percentage from time, treatment, and baseline age.

```
library(lubridate)
library(lme4)
library(arm)
library(sjPlot)

### Read in data, change vdate to dates
cd4 <- read.csv("cd4.csv")
cd4$vdate <- mdy(cd4$vdate)
cd4$treatmnt <- as.factor(cd4$treatmnt)

### Extract complete cases
cd4_full <- cd4[complete.cases(cd4),]

### Run vary intercept, vary slopes model
model_135a <- lmer(cd4pct ~ treatmnt + baseage + visitno + (1 + visitno|newpid), data = cd4_full)
summary(model_135a)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cd4pct ~ treatmnt + baseage + visitno + (1 + visitno | newpid)
## Data: cd4_full
##
## REML criterion at convergence: 7116.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.9619 -0.4347 -0.0567  0.3509  3.7770
##
## Random effects:
## Groups      Name                Variance Std.Dev. Corr
## newpid      (Intercept) 142.3564 11.9313
##              visitno      0.1757  0.4192 -0.35
## Residual                    43.9336  6.6282
## Number of obs: 978, groups: newpid, 226
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 27.77502    1.62221  17.122
## treatmnt2    1.94635    1.58733   1.226
## baseage     -0.93160    0.35123  -2.652
## visitno     -0.24260    0.05169  -4.693
##
## Correlation of Fixed Effects:
##              (Intr) trtmn2 baseag
## treatmnt2 -0.452
## baseage   -0.707 -0.001
## visitno   -0.225  0.006 -0.025
```

```
sjt.lmer(model_135a)
```

	cd4pct		
	<i>B</i>	<i>CI</i>	<i>p</i>
<b>Fixed Parts</b>			
(Intercept)	27.78	24.60 – 30.95	<.001
treatmnt (2)	1.95	-1.16 – 5.06	.221
baseage	-0.93	-1.62 – -0.24	.009
visitno	-0.24	-0.34 – -0.14	<.001
<b>Random Parts</b>			
σ <sup>2</sup>	43.934		

$\tau_{00, \text{newpid}}$	142.356
$\rho_{01}$	-0.345
$N_{\text{newpid}}$	226
$\text{ICC}_{\text{newpid}}$	0.764
Observations	978
$R^2 / \Omega_o^2$	.828 / .821

From the model summary above, we see that the random intercepts vary about 27.775 with a variance of 142.36. We also see that the slope for visit number varies around -.242 with a variance of .175.

The interpretation for the binary treatment coefficient is that we expect cd4 percentage to be 1.9% higher for those in the treatment group vs the control group, all else equal.

The continuous predictor, baseline age, can be interpreted as: For every year increase in baseline age, we expect a .92% lower cd4 percentage on average, all else equal.

Visit number can be interpreted as: For each additional follow up visit, we expect a .242 decrease in cd4 percentage on average, all else equal (the actual slope is different for each person).

## Part B

We then try the model using visit number as a factor instead of using varying slopes.

```
### Run a model with varying intercepts, but fixed slopes for indicators for visit number
cd4_full$visitno_factor <- as.factor(cd4_full$visitno)
model_135b <- lmer(cd4pct ~ treatmnt + baseage + visitno_factor + (1 | newpid), data = cd
4_full)
summary(model_135b)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cd4pct ~ treatmnt + baseage + visitno_factor + (1 | newpid)
## Data: cd4_full
##
## REML criterion at convergence: 7127.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.6150 -0.4681 -0.0458  0.3939  3.5916
##
## Random effects:
## Groups Name Variance Std.Dev.
## newpid (Intercept) 125.93  11.222
## Residual 50.81  7.128
## Number of obs: 978, groups: newpid, 226
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    27.5620    1.6284  16.926
## treatmnt2       1.9017    1.5887   1.197
## baseage        -0.9231    0.3515  -2.627
## visitno_factor4 -0.6203    0.7697  -0.806
## visitno_factor7 -1.6917    0.7806  -2.167
## visitno_factor10 -2.2023    0.8086  -2.724
## visitno_factor13 -3.2870    0.8309  -3.956
## visitno_factor16 -3.5693    0.9033  -3.951
## visitno_factor19 -5.0004    0.9959  -5.021
##
## Correlation of Fixed Effects:
##              (Intr) trtmn2 baseag vstn_4 vstn_7 vst_10 vst_13 vst_16
## treatmnt2    -0.453
## baseage      -0.705 -0.001
## vistn_fctr4  -0.189  0.012 -0.001
## vistn_fctr7  -0.177  0.014 -0.016  0.414
## vstn_fctr10  -0.173  0.015 -0.015  0.401  0.420
## vstn_fctr13  -0.162  0.008 -0.021  0.389  0.409  0.410
## vstn_fctr16  -0.149  0.005 -0.015  0.361  0.376  0.370  0.371
## vstn_fctr19  -0.140  0.006 -0.008  0.329  0.340  0.336  0.346  0.316
```

```
sjt.lmer(model_135b)
```

	cd4pct		
	<i>B</i>	<i>CI</i>	<i>p</i>

## Fixed Parts

(Intercept)	27.56	24.37 – 30.75	<.001
treatmnt (2)	1.90	-1.21 – 5.02	.232
baseage	-0.92	-1.61 – -0.23	.009
visitno_factor (4)	-0.62	-2.13 – 0.89	.421
visitno_factor (7)	-1.69	-3.22 – -0.16	.031
visitno_factor (10)	-2.20	-3.79 – -0.62	.007
visitno_factor (13)	-3.29	-4.92 – -1.66	<.001
visitno_factor (16)	-3.57	-5.34 – -1.80	<.001
visitno_factor (19)	-5.00	-6.95 – -3.05	<.001

### Random Parts

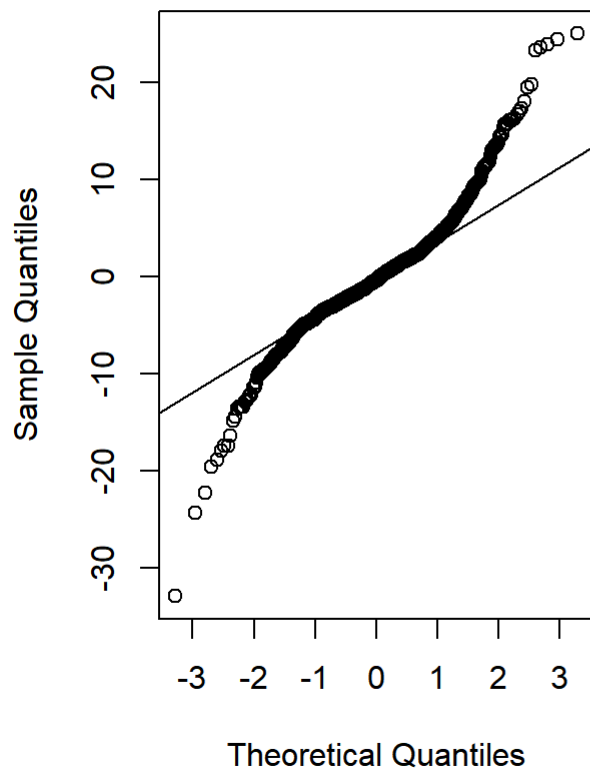
$\sigma^2$	50.814
$\tau_{00, \text{newpid}}$	125.933
$N_{\text{newpid}}$	226
$ICC_{\text{newpid}}$	0.713
Observations	978
$R^2 / \Omega_0^2$	.780 / .775

Treatment and baseline age coefficients are interpreted the same as in part (a). However, the coefficients for visit number are now in reference to visit 1. An example interpretation: At visit 7, we expect the average cd4 percentage to be 1.69 percentage points lower than on visit one, all else equal.

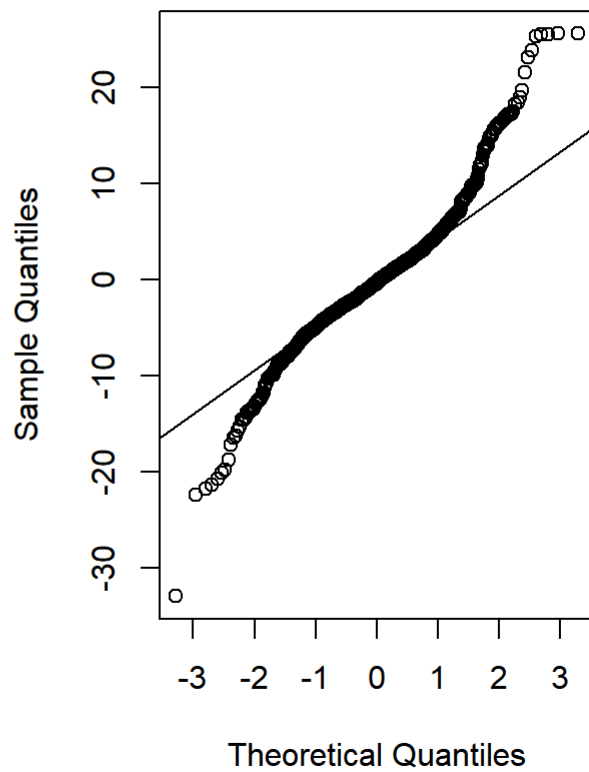
## Part C

We first compare the two models graphically using QQ plots:

**Normal Q-Q Plot Model 1**



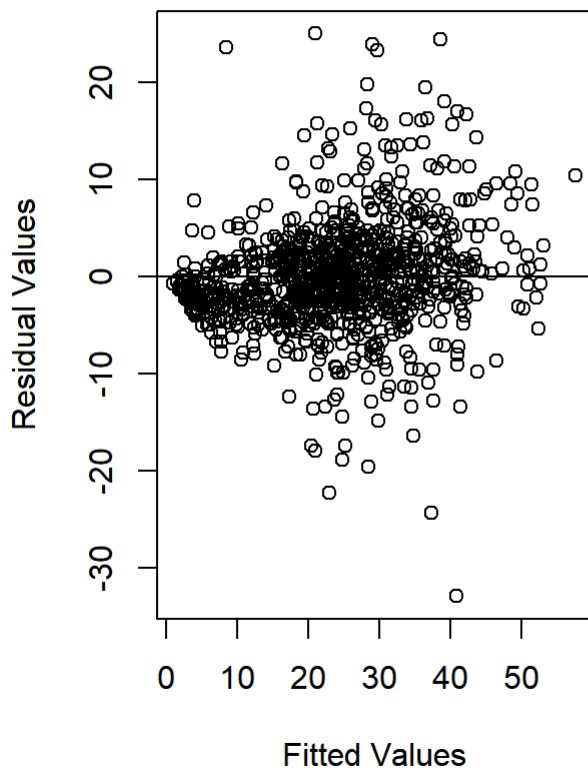
**Normal Q-Q Plot Model 2**



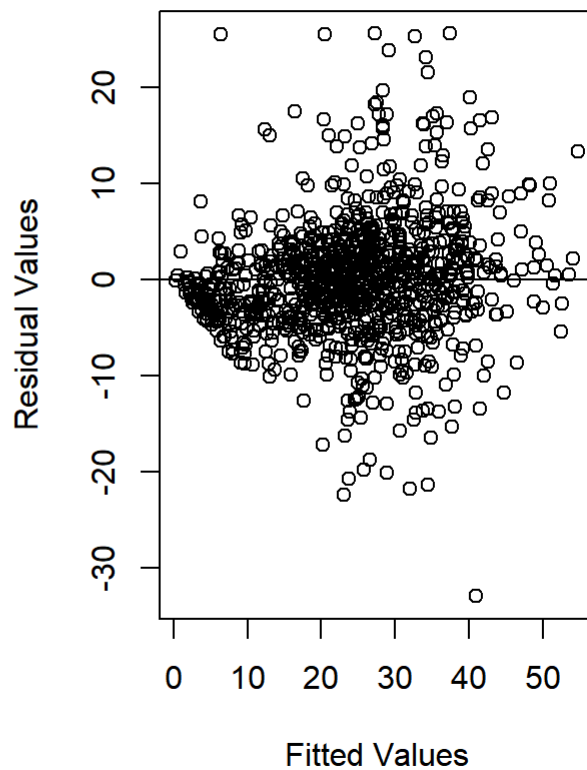
We see that neither plots are very good in the tails. However, Model 1 appears to be slightly closer to the normal line.

We also checked the residual plots.

**Fitted vs. Residuals for Model 1**



**Fitted vs. Residuals for Model 2**



The residuals plots both look randomly scattered about the zero line, and look about the same.

We also compare the models numerically, using AIC values:

```
### extract AICs  
AIC(model_135a)
```

```
## [1] 7132.335
```

```
AIC(model_135b)
```

```
## [1] 7149.121
```

We see that model 1 has a slightly lower AIC. This provides weak evidence that the varying intercept and slope model has a better fit.

## Gelman, Hill Chapter 14 #3

(a)

First, we fit a varying intercept logistic model (grouped by building). The variables that we choose to examine are:

-Missing or Worn Flooring (extflr5\_2) -Cracks in Walls (intcrack2) -Holes in the flooring (inthole2) -If there is a regular exterminator (regext) -Race (race)

```
library(readr)
library(lme4)
library(sjPlot)
library(arm)
rats = read.table("rodents.txt")

attach(rats)

### Change to factors
rats$rodent2 = as.factor(rodent2)
rats$extwin4_2 = as.factor(extwin4_2)
rats$extflr5_2 = as.factor(extflr5_2)
rats$intcrack2 = as.factor(intcrack2)
rats$inthole2 = as.factor(inthole2)
rats$old = as.factor(old)
rats$regext = as.factor(regext)
rats$housing = as.factor(housing)
rats$race = as.factor(race)

### Run Model
model.a = glmer(rodent2 ~
  extflr5_2 +
  intcrack2 +
  inthole2 +
  regext +
  as.factor(race) +
  (1 | bldg),
  family = binomial (link = "logit"),
  control = glmerControl(optimizer="bobyqa", optCtrl = list(maxfun = 100000
00)))

summary(model.a)
```



```

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## rodent2 ~ extflr5_2 + intcrack2 + inthole2 + regext + as.factor(race) +
## (1 | bldg)
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+07))
##
##      AIC      BIC   logLik deviance df.resid
##  1130.6   1191.9   -553.3   1106.6     1212
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.5346 -0.3858 -0.2036 -0.1299  4.0428
##
## Random effects:
## Groups Name             Variance Std.Dev.
## bldg    (Intercept)  2.246      1.499
## Number of obs: 1224, groups:  bldg, 742
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.0933     0.3141  -9.847 < 2e-16 ***
## extflr5_2      1.4776     0.5550   2.662  0.00776 **
## intcrack2      1.6491     0.3006   5.486 4.11e-08 ***
## inthole2       1.5121     0.3952   3.826 0.00013 ***
## regext        -0.2311     0.2030  -1.139  0.25488
## as.factor(race)2 1.7701     0.2951   5.999 1.99e-09 ***
## as.factor(race)3 1.5660     0.3414   4.587 4.50e-06 ***
## as.factor(race)4 2.2296     0.3187   6.995 2.65e-12 ***
## as.factor(race)5 0.6697     0.4275   1.566  0.11726
## as.factor(race)6 2.4110     1.2624   1.910  0.05614 .
## as.factor(race)7 1.2630     1.0053   1.256  0.20898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) ext5_2 intcr2 inth12 regext as.()2 as.()3 as.()4 as.()5
## extflr5_2    -0.265
## intcrack2    -0.379  0.062
## inthole2     -0.231  0.016 -0.246
## regext       -0.190 -0.041 -0.041 -0.042
## as.fctr(r)2  -0.679  0.089  0.173  0.097 -0.055
## as.fctr(r)3  -0.458  0.063  0.100  0.009 -0.041  0.437
## as.fctr(r)4  -0.596  0.072  0.197  0.071 -0.118  0.529  0.411
## as.fctr(r)5  -0.370 -0.097  0.118  0.034  0.053  0.328  0.262  0.309
## as.fctr(r)6  -0.155  0.037 -0.014  0.057  0.001  0.144  0.105  0.118  0.071

```

```
## as.fctr(r)7 -0.164 0.021 0.084 0.019 -0.035 0.147 0.109 0.134 0.131
##           as.()6
## extflr5_2
## intcrack2
## inthole2
## regext
## as.fctr(r)2
## as.fctr(r)3
## as.fctr(r)4
## as.fctr(r)5
## as.fctr(r)6
## as.fctr(r)7 0.030
```

```
sjt.glmer(model.a)
```

	rodent2		
	<i>Odds Ratio</i>	<i>CI</i>	<i>p</i>
<b>Fixed Parts</b>			
(Intercept)	0.05	0.02 – 0.08	<.001
extflr5_2	4.38	1.48 – 13.01	.008
intcrack2	5.20	2.89 – 9.38	<.001
inthole2	4.54	2.09 – 9.84	<.001
regext	0.79	0.53 – 1.18	.255
as.factor(race) (2)	5.87	3.29 – 10.47	<.001
as.factor(race) (3)	4.79	2.45 – 9.35	<.001
as.factor(race) (4)	9.30	4.98 – 17.36	<.001
as.factor(race) (5)	1.95	0.85 – 4.52	.117
as.factor(race) (6)	11.15	0.94 – 132.32	.056
as.factor(race) (7)	3.54	0.49 – 25.37	.209
<b>Random Parts</b>			
$\tau_{00, \text{bldg}}$	2.246		
$N_{\text{bldg}}$	742		
$ICC_{\text{bldg}}$	0.406		
Observations	1224		
Deviance	690.104		

For this model, the intercept was allowed to vary for each building. The overall mean intercept is -3.09. The random intercepts vary about this mean with a variance of 2.246.

The binary fixed effects have a similar interpretation. When exponentiated, they give an odds ratio. For example, If an apartment has holes in the flooring, they have 4.38 times the odds of having rodents than a similar apartment with no holes.

The categorical race predictor is referenced to group 1. Similar to the binary variables, exponentiating the betas give an odds ratio. For example, apartments where the tennants are of race category two, have 5.87 times the odds of having rodents than a similar apartment with tennants of race category 1.

## (b)

Next we fit a three level model. Apartments are grouped by building. Buildings are grouped by community district (cd). This model also adds community district level predictors: the mean of homeownership in the community (ownhous\_Mean) and the mean proportion of old houses in the community (old\_Mean).

```
rats$cd = as.factor(cd)

model.b = glmer(rodent2 ~
  extflr5_2 +
  intcrack2 +
  inthole2 +
  regext +
  as.factor(race) +
  old_Mean +
  ownhous_Mean +
  (1 | cd/bldg),
  family = binomial(link = "logit"),
  control = glmerControl(optimizer="bobyqa", optCtrl = list(maxfun = 100000
00)))

summary(model.b)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## rodent2 ~ extflr5_2 + intcrack2 + inthole2 + regext + as.factor(race) +
## old_Mean + ownhous_Mean + (1 | cd/bldg)
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+07))
##
##      AIC      BIC   logLik deviance df.resid
##  1096.2   1172.8   -533.1   1066.2     1209
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.4546 -0.3599 -0.2028 -0.0940  4.9028
##
## Random effects:
## Groups Name      Variance Std.Dev.
## bldg:cd (Intercept) 2.17259  1.4740
## cd      (Intercept) 0.01597  0.1264
## Number of obs: 1224, groups: bldg:cd, 742; cd, 55
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.7741     0.7030  -2.524 0.011617 *
## extflr5_2       1.3970     0.5597   2.496 0.012562 *
## intcrack2       1.4424     0.3031   4.759 1.94e-06 ***
## inthole2        1.4249     0.3958   3.600 0.000318 ***
## regext          -0.5225     0.2187  -2.389 0.016904 *
## as.factor(race)2  1.4216     0.3053   4.656 3.23e-06 ***
## as.factor(race)3  1.0243     0.3558   2.879 0.003988 **
## as.factor(race)4  1.7143     0.3249   5.276 1.32e-07 ***
## as.factor(race)5  0.6278     0.4343   1.445 0.148341
## as.factor(race)6  1.8275     1.2889   1.418 0.156238
## as.factor(race)7  1.2356     1.0134   1.219 0.222775
## old_Mean         0.4380     0.7555   0.580 0.562098
## ownhous_Mean     -3.9129     0.9054  -4.322 1.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sjt.glmer(model.b)
```

---

---

rodent2

---

*Odds Ratio*

*CI*

*p*

---

## Fixed Parts

(Intercept)	0.17	0.04 – 0.67	.012
extflr5_2	4.04	1.35 – 12.11	.013
intrcrack2	4.23	2.34 – 7.66	<.001
inthole2	4.16	1.91 – 9.03	<.001
regext	0.59	0.39 – 0.91	.017
as.factor(race) (2)	4.14	2.28 – 7.54	<.001
as.factor(race) (3)	2.79	1.39 – 5.59	.004
as.factor(race) (4)	5.55	2.94 – 10.50	<.001
as.factor(race) (5)	1.87	0.80 – 4.39	.148
as.factor(race) (6)	6.22	0.50 – 77.77	.156
as.factor(race) (7)	3.44	0.47 – 25.07	.223
old_Mean	1.55	0.35 – 6.81	.562
ownhous_Mean	0.02	0.00 – 0.12	<.001

### Random Parts

$\tau_{00, \text{bldg:cd}}$	2.173
$\tau_{00, \text{cd}}$	0.016
$N_{\text{bldg:cd}}$	742
$N_{\text{cd}}$	55
$\text{ICC}_{\text{bldg:cd}}$	0.397
$\text{ICC}_{\text{cd}}$	0.003
Observations	1224
Deviance	670.789

The three level model includes an extra random slope effect. The intercept variance within a community district is 2.17. The intercept variance between community districts is .01.

The binary and categorical interpretations are the same as in the two level model. The new continuous community level variables have similar interpretations. For example, if the proportion of old houses in a community district increases by one percent, the odds of having rodents increases by .0155. ( $\exp(.4380) / 100$ )

(c)

```
anova(model.a, model.b)
```

```
## Data: NULL
## Models:
## model.a: rodent2 ~ extflr5_2 + intcrack2 + inthole2 + regext + as.factor(race) +
## model.a:      (1 | bldg)
## model.b: rodent2 ~ extflr5_2 + intcrack2 + inthole2 + regext + as.factor(race) +
## model.b:      old_Mean + ownhous_Mean + (1 | cd/bldg)
##           Df      AIC      BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model.a 12 1130.6 1191.9 -553.29   1106.6
## model.b 15 1096.2 1172.8 -533.08   1066.2 40.435      3 8.615e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see from the ANOVA table the three level model has a better fit (p value for the chi squared test is very small).