Applied Analysis 5

Rohan Hore

August 2024

Adolescent smoking This is a longitudinal, natural history study of adolescent smoking. Students included in the longitudinal study were either in grade 8 or 10 at baseline, and self-reported on a screening questionnaire 6 to 8 weeks prior to baseline that they either had never smoked, but indicated a probability of future smoking, or had smoked in the past 90 days. The longitudinal study utilized a multi-method approach to assess adolescents at three time intervals of one week: Baseline, 6 months, and 12 months. Adolescents carried hand-held computers with them during the 7 consecutive day data collection period at each of the three time points and were trained to both respond to random prompts from the computers and to event record (initiate a data collection interview) smoking episodes. Immediately after smoking a cigarette, participants completed a series of questions on the hand held computers. Questions included ones about place, activity, companionship, mood, and other subjective items. The hand held computers dated and time-stamped each entry. For inclusion in the data here, adolescents must have smoked at least two cigarettes during the 7-day baseline data collection period; 96 adolescents met this inclusion criterion.

The 96 adolescents began the study with varying amounts of cigarette smoking experience. Adolescents were divided into three groups based on their lifetime smoking levels: those who had smoked less than 6 cigarettes in their lifetimes ($n = 16$), representing very novice smokers; those who had smoked between 6 and 99 cigarettes in their lifetimes ($n = 46$), representing a group of irregular or experimental smokers; and those who had smoked 100 or more cigarettes during their lifetimes ($n = 34$), representing more regular smokers.

Immediately after smoking the cigarette, subjects turned on their hand-held computer to complete a variety of questions. They responded with an analog ladder-type scale, by moving a stylus to the appropriate point on the ladder scale (the scale is from 1 to 10). They were first asked about their moods and feelings "right now" (after smoking) and then about how they felt just before smoking. For each subject a change score (after - before) was calculated for a number of variables: positive feeling, negative feeling, feeling of tiredness, feeling of frustration and physiological feeling. The data for this experiment are available in the file `mood.csv`

Data description The data are in spreadsheet format with 498 rows, and 10 columns. The columns are:

- *Id* - subject id (from 1 to 96)
- *Study day* - sequential day in study (from 1 to 13)
- *Week day* - day of week (1=Monday, 2=Tuesday, $\cdots$, 7=Sunday)
- *Timebin* - timing of smoking event: 1 = "$3am - 8:59am$"; 2 = "$9am - 1:59pm'$"; 3 = "$2pm - 5:59pm'$"; 4 = "$6pm - 9:59pm$";5 = "$10pm - 2:59am$"
- *SmkLevel* - 1 =smoked less than 6 cigarettes in their lifetimes; 2 =smoked between 6 and 99 cigarettes in their lifetimes; 3 =smoked 100 or more cigarettes during their lifetimes
- *cposit* - change in reported positive affect (after smoking - before smoking)
- *cnegat* - change in reported negative affect (after smoking - before smoking)
- *ctired* - change in reported tiredness/boredom (after smoking - before smoking)

- *cfrust* - change in reported frustration (after smoking- before smoking)
- *cphys* - change in reported physiological sensations (after smoking- before smoking)

The main goal of this data analysis is to study the effect of smoking history on change in physiological sensations. So, unless otherwise specified, all computational questions refer to the data from the last response, *cphys*.

Data Reading

```
#data reading
mood=read.csv("mood.csv")
head(mood)
```

```
##   Id Day Weekday Timebin SmkLevel   cposit cnegat ctired cfrust cphys
## 1  1   1       4       2        3  1.33333   -3.0   -0.5    1.0  -4.0
## 2  1   1       4       3        3  1.33333    0.4    4.0    2.0   0.0
## 3  2   2       5       5        3 -0.66667    0.8    2.0    1.0   1.0
## 4  2   3       6       3        3 -1.66667    0.2   -0.5   -0.5   0.5
## 5  2   3       6       3        3  0.33333    1.4   -1.5    0.0  -1.5
## 6  2   6       2       4        3  1.00000   -1.2   -1.0    3.0   2.5
```

Possible Questions

Problem 1

The experimenters chose to use the trichotomous ordered smoking history variable and they have chose to not report smoking history as a continuous variable. Why do you think they did that? Do you think that this *information loss* might affect our analysis?
Answer: Easier to collect data, hard to remember the exact numbers for participants ; The effect of lifetime smoking history in adolescents is substantively not the same contrasting, say, individuals who have smoked 5 and 10 cigarettes versus those who have smoked 100 and 105 cigarettes. So this kind of categorization doesn't lose any information for this particular study.

Problem 2

Do you think that the positive(*cposit*) and negative feelings(*cnegat*) represent just two opposite extremes in one single emotion-scale? How would you do some elementary analysis to answer this question?
Answer: We can look at a simple scatter-plot to study the relationship. Further, we can ignore the fact that multiple observations are coming from same individual and run a paired t-test between cposit and -cnegat to test the following hypothesis $H_0 : \mu_+ = -\mu_-$ where μ_+ and μ_- correspondingly represent the average change in positive and negative feeling. The hypothesis means that the only difference between these two measures is a mere sign-swap.

```
mood$SmkLevel=as.factor(mood$SmkLevel)
attach(mood)
#relation between cposit and cnegat
cor(cposit,cnegat)
plot(cposit,cnegat)

#t-test ignoring we have multiple data from same individual
##treating them as independent observations
t.test(cposit,-cnegat,paired = TRUE)
```

```
## [1] -0.3620395

##
##  Paired t-test
##
## data:  cposit and -cnegat
## t = 5.8827, df = 497, p-value = 7.422e-09
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.3332733 0.6675297
## sample estimates:
## mean difference
##       0.5004015
```

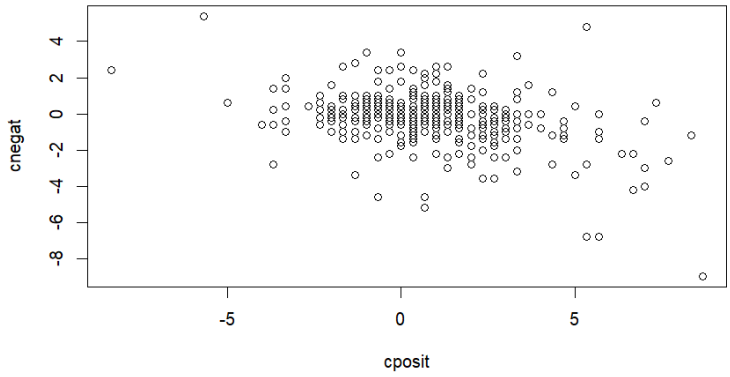

Figure 1: Caption

The scatter-plot doesn't indicate any obvious relationship between positive feeling and negative feeling. The t-test as well supports the claim that the relation between these two measurements are probably much more complicated than intended by above H_0.

Problem 3

Produce a visual summary of the change in physiological sensations for different smoking levels by looking at box-plots of subject-means. Report any notable observation and if possible, investigate them!
Answer: We first start with aggregating the data for each individual. For each individual, we now have only the mean cphys measure. We show the boxplot below.

```
id_numbers=unique(Id)
aggr.mood=matrix(nrow=length(id_numbers),ncol=3)
for(i in id_numbers){
  aggr.mood[i,1]=id_numbers[i]
  aggr.mood[i,2]=mood$SmkLevel[min(which(mood$Id==id_numbers[i]))]
  aggr.mood[i,3]=mean(mood$cphys[mood$Id==id_numbers[i]])
}
aggr.mood=as.data.frame(aggr.mood)
colnames(aggr.mood)=c("Id","SmkLevel","cphys")
boxplot(cphys~SmkLevel,data=aggr.mood)
```

Looking at the boxplot, we can't find any significant difference in the mean effect across different smoking levels. However, the variances seem to differ across different smoking levels.

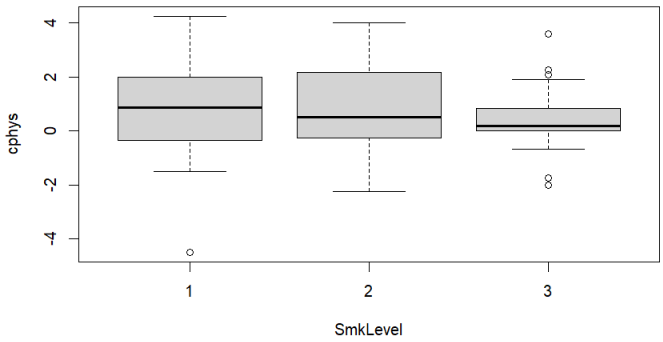

Figure 2: Caption

One possible reason could be: sample size differences. We are looking at aggregated observations for each individual. Novice smokers might have less number of observations and thus, higher individual-level variance in aggregated data. While for a regular smoker, we will typically have more observations and thus should reflect a lower variance. The following boxplot supports that claim.

```
aggr.mood=cbind(aggr.mood,0)
for(i in id_numbers){
  aggr.mood[i,4]=sum(mood$Id==id_numbers[i])
}
colnames(aggr.mood)=c("Id","SmkLevel","cphys","obs_no")
boxplot(obs_no~SmkLevel,data=aggr.mood)
```

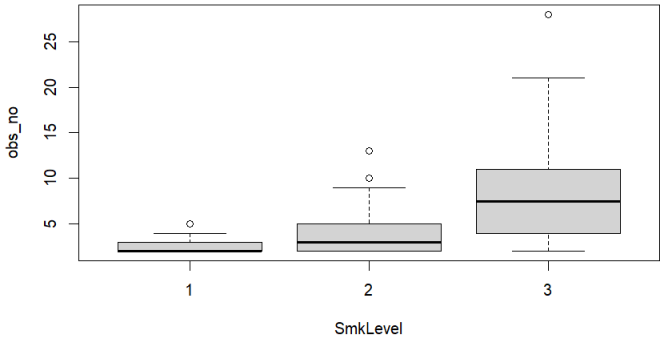

Figure 3: Caption

Problem 4

Fit a linear Gaussian model with fixed effects for smoking level and a random effect for subjects. Report the parameter estimates of the variance components and explain the meaning of the fitted model.

Answer: Let's denote cphys measure by Y. Then, we want to fit the following linear mixed model

$$Y_{ij} = \beta_0 + \beta_1 S_i + \alpha_i + \epsilon_{ij}, \quad var(\epsilon_{ij}) = \sigma^2_\epsilon, \quad var(\alpha_i) = \sigma^2_\alpha$$

where S_i is the smoking level of individual i.

```
library(lme4)
model.1=lmer(cphys~SmkLevel+(1|Id),data = mood,REML = FALSE)
summary(model.1)

model.2=lmer(cphys~1+(1|Id),data = mood,REML = FALSE)
anova(model.2,model.1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: cphys ~ SmkLevel + (1 | Id)
##    Data: mood
##
##      AIC      BIC   logLik deviance df.resid
##   1956.3   1977.3   -973.1   1946.3      493
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3137 -0.4424 -0.0735  0.4376  4.0706
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Id       (Intercept) 1.224    1.106
##  Residual             2.325    1.525
## Number of obs: 498, groups:  Id, 96
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.9282     0.3698   2.510
## SmkLevel2    -0.1217     0.4239  -0.287
## SmkLevel3    -0.4624     0.4293  -1.077
##
## Correlation of Fixed Effects:
##           (Intr) SmkLv2
## SmkLevel2 -0.872
## SmkLevel3 -0.861  0.751
```

Is there any difference in cphys for different smoking level?

```
## Data: mood
## Models:
## model.2: cphys ~ 1 + (1 | Id)
## model.1: cphys ~ SmkLevel + (1 | Id)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model.2    3 1954.1 1966.7 -974.03   1948.1
## model.1    5 1956.3 1977.3 -973.14   1946.3 1.7737  2      0.412
```

What are the other fixed/random effects you would like to include in order to strengthen the model, and discuss how you might implement them?

We might like to add an effect of Weekday or maybe, we can add an effect of weekdays/weekend (a binarized variable); Allowing random effect for days and allowing larger variances for later days (since they typically have less observations)

Problem 5

As an alternative, fit a model that allows for different between-subject variances in the three smoking-level groups. As before, explain the meaning of the fitted model. How would you assess the significance of the

results.

Answer: There might be different variance structures in different smoking level-based groups. So,we would like to accommodate that in our model. In particular, we want to fit the following linear mixed model

$$Y_{ij} = \beta_0 + \beta_1 S_i + \alpha_i + \epsilon_{ij}, \quad var(\epsilon_{ij}) = \sigma^2_\epsilon$$

where S_i is the smoking level of individual i and

$$var(\alpha_i) = \begin{cases} \sigma^2_{\alpha,1} & \text{if } S_i = 1 \\ \sigma^2_{\alpha,2} & \text{if } S_i = 2 \\ \sigma^2_{\alpha,3} & \text{if } S_i = 3 \end{cases}$$

We can approach the following model as follows.

A NOT SO RIGHT SOLUTION BELOW:

```
model.3=lmer(cphys~SmkLevel+(0+SmkLevel|Id),data = mood,REML = FALSE)
summary(model.3)

model.4=lmer(cphys~1+(0+SmkLevel|Id),data = mood,REML = FALSE)
anova(model.4,model.3)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: cphys ~ SmkLevel + ((0 + SmkLevel | Id))
##    Data: mood
##
##      AIC      BIC   logLik deviance df.resid
##   1955.7   1997.8   -967.8   1935.7      488
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3194 -0.4623 -0.1153  0.4186  4.0425
##
## Random effects:
##  Groups   Name      Variance Std.Dev. Corr
##  Id       SmkLevel1 3.5960   1.8963
##           SmkLevel2 1.4396   1.1998    0.34
##           SmkLevel3 0.4739   0.6884    0.43 -0.20
##  Residual           2.2952   1.5150
## Number of obs: 498, groups:  Id, 96
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.9497     0.5342   1.778
## SmkLevel2    -0.1406     0.5770  -0.244
## SmkLevel3    -0.4723     0.5566  -0.849
##
## Correlation of Fixed Effects:
##           (Intr) SmkLv2
## SmkLevel2 -0.926
## SmkLevel3 -0.960  0.889
## optimizer (nloptwrap) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

We also look at the anova test of whether the fixed effect of smoking level is required or not

```
## Data: mood
## Models:
## model.4: cphys ~ 1 + (0 + SmkLevel | Id)
## model.3: cphys ~ SmkLevel + ((0 + SmkLevel | Id))
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model.4    8 1953.6 1987.3 -968.82   1937.6
## model.3   10 1955.7 1997.8 -967.85   1935.7 1.9448  2     0.3782
```

Let's look at the likelihood-ratio test between the two models with smoking level fixed effect. In one we have only subject specific random effect, but in other we allow for the subject specific variance to vary across different smoking levels.

```
anova(model.1,model.3,test="LRT")
```

```
## Data: mood
## Models:
## model.1: cphys ~ SmkLevel + (1 | Id)
## model.3: cphys ~ SmkLevel + ((0 + SmkLevel | Id))
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model.1    5 1956.3 1977.3 -973.14   1946.3
## model.3   10 1955.7 1997.8 -967.85   1935.7 10.593  5    0.06008 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But why do we have additional 5 parameters in the second model? Is it the model, we wanted to fit?

-No, it also allows for some correlations. A better approach is as follows. We need to create three binary variables for three levels of smoking first and then, we can use lmer as follows

A BETTER SOLUTION NOW:

```
trans.mood=mood[,c(1,5,10)]
trans.mood=cbind(trans.mood,0);trans.mood[,4]=as.numeric(trans.mood$SmkLevel==1)
trans.mood=cbind(trans.mood,0);trans.mood[,5]=as.numeric(trans.mood$SmkLevel==2)
trans.mood=cbind(trans.mood,0);trans.mood[,6]=as.numeric(trans.mood$SmkLevel==3)
colnames(trans.mood)=c("Id","SmkLevel","cphys","Lev1","Lev2","Lev3")
trans.mood$SmkLevel=as.factor(trans.mood$SmkLevel)
head(trans.mood)
```

```
##   Id SmkLevel cphys Lev1 Lev2 Lev3
## 1  1        3  -4.0    0    0    1
## 2  1        3   0.0    0    0    1
## 3  2        3   1.0    0    0    1
## 4  2        3   0.5    0    0    1
## 5  2        3  -1.5    0    0    1
## 6  2        3   2.5    0    0    1
```

```
model.5=lmer(cphys~SmkLevel+(0+Lev1+Lev2+Lev3||Id),data = trans.mood,REML = FALSE)
summary(model.5)

model.6=lmer(cphys~1+(0+Lev1+Lev2+Lev3||Id),data = trans.mood,REML = FALSE)
anova(model.6,model.5)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: cphys ~ SmkLevel + ((0 + Lev1 | Id) + (0 + Lev2 | Id) + (0 +      Lev3 | Id))
##    Data: trans.mood
##
```

```
##      AIC      BIC   logLik deviance df.resid
##   1949.7   1979.2   -967.8   1935.7      491
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3194 -0.4623 -0.1153  0.4186  4.0425
##
## Random effects:
##  Groups   Name Variance Std.Dev.
##  Id       Lev1 3.5959   1.8963
##  Id.1     Lev2 1.4396   1.1998
##  Id.2     Lev3 0.4739   0.6884
##  Residual      2.2952   1.5150
## Number of obs: 498, groups:  Id, 96
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.9497     0.5342   1.778
## SmkLevel2    -0.1406     0.5770  -0.244
## SmkLevel3    -0.4723     0.5566  -0.849
##
## Correlation of Fixed Effects:
##           (Intr) SmkLv2
## SmkLevel2 -0.926
## SmkLevel3 -0.960  0.889
```

Is the fixed effect of smoking level necessary?

```
## Data: trans.mood
## Models:
## model.6: cphys ~ 1 + ((0 + Lev1 | Id) + (0 + Lev2 | Id) + (0 + Lev3 | Id))
## model.5: cphys ~ SmkLevel + ((0 + Lev1 | Id) + (0 + Lev2 | Id) + (0 + Lev3 | Id))
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model.6    5 1947.6 1968.7 -968.82   1937.6
## model.5    7 1949.7 1979.2 -967.85   1935.7 1.9448  2     0.3782
```

Problem 6

What does the model in problem 5 tell you about the adequacy of the model you fit in problem 4? You can formally answer this using a likelihood ratio test.

Answer: Let's compare a likelihood ratio test between the two models having smoking level as fixed effect and subject specific random effect. In one, we allow subject specific random effects to have different variances across different smoking levels and in the other, we don't. The LRT test indicates that we should use the more general model.

```
model.1p=lmer(cphys~SmkLevel+(1|Id),data = trans.mood,REML = FALSE)
anova(model.1p,model.5,test="LRT")
```

```
## Data: trans.mood
## Models:
## model.1p: cphys ~ SmkLevel + (1 | Id)
## model.5: cphys ~ SmkLevel + ((0 + Lev1 | Id) + (0 + Lev2 | Id) + (0 + Lev3 | Id))
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
```

```
## model.1p    5 1956.3 1977.3 -973.14   1946.3
## model.5     7 1949.7 1979.2 -967.85   1935.7 10.593  2   0.005009 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 7

What is the impact of non-normality? Describe in detail a model that that you think would be best for analyzing this data.

Answer: cphys is lower bounded by -10. In the data, the lowest value is -6.6. We start with a shifted version of the cphys variable so that the range is only positive numbers. We see a very big mass and a right-skewed distribution. One approach could be using glm with gamma family.

```
hist(cphys+7)
model.7=glmer(cphys+7~SmkLevel+(1|Id),data = mood,family="Gamma")
summary(model.7)
plot(model.7)
```

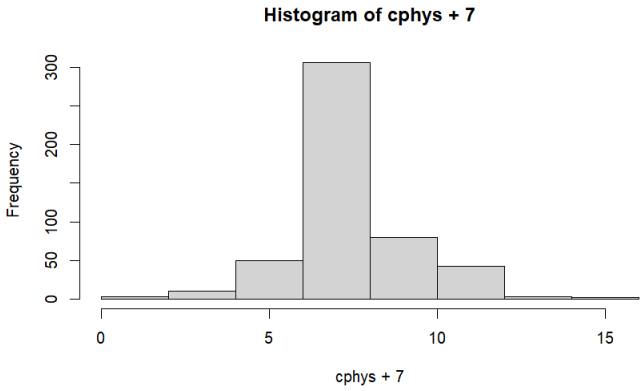

Figure 4: Caption

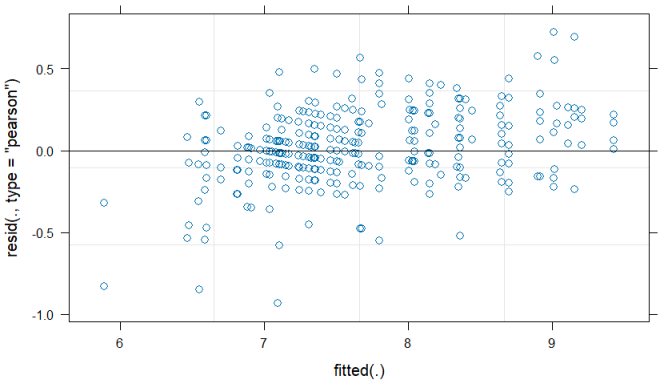

Figure 5: Caption

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: Gamma  ( inverse )
## Formula: cphys + 7 ~ SmkLevel + (1 | Id)
##    Data: mood
##
##      AIC      BIC   logLik deviance df.resid
##   1987.5   2008.6   -988.8   1977.5      493
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5433 -0.4167 -0.0568  0.4536  3.5185
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Id       (Intercept) 0.0002313 0.01521
##  Residual             0.0418577 0.20459
## Number of obs: 498, groups:  Id, 96
##
## Fixed effects:
##              Estimate Std. Error t value Pr(>|z|)
## (Intercept) 0.1323792  0.0080971  16.349   <2e-16 ***
## SmkLevel2   0.0004084  0.0092615   0.044    0.965
## SmkLevel3   0.0060883  0.0098749   0.617    0.538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) SmkLv2
## SmkLevel2 -0.852
## SmkLevel3 -0.803  0.694
```

Problem 8

The histogram of `cphys` reveals a large mass around 0-which probably means that differences near 0 i.e. in the interval $[-.5, .5]$ are not that meaningful. Suppose, we binarize the `cphys` variable- i.e. we consider the difference to be significant if it's more than 0.5. Analyze whether there is any effect of past smoking behaviour on this binarized `cphys` variable? How will you compare this result with previous conclusions?

Answer: We start with binarizing the cphys variable at 0.5 and then we fit a logit model to that with a subject random effect. We check the significance for effect of smoking level, which in this case turns out to be significant.

```
bin.mood=mood[,c(1,5,10)]
bin.mood=cbind(bin.mood,0);bin.mood[,4]=as.numeric(abs(bin.mood$cphys)>0.5)
colnames(bin.mood)=c("Id","SmkLevel","cphys","bin_cphys")
bin.mood$SmkLevel=as.factor(bin.mood$SmkLevel)
bin.mood$bin_cphys=as.factor(bin.mood$bin_cphys)
head(bin.mood)
```

```
##   Id SmkLevel cphys bin_cphys
## 1  1        3  -4.0         1
## 2  1        3   0.0         0
## 3  2        3   1.0         1
## 4  2        3   0.5         0
```

```
## 5  2        3  -1.5         1
## 6  2        3   2.5         1
```

```
model.8=glmer(bin_cphys~SmkLevel+(1|Id),data = bin.mood,family="binomial")
summary(model.8)
plot(model.8)
```

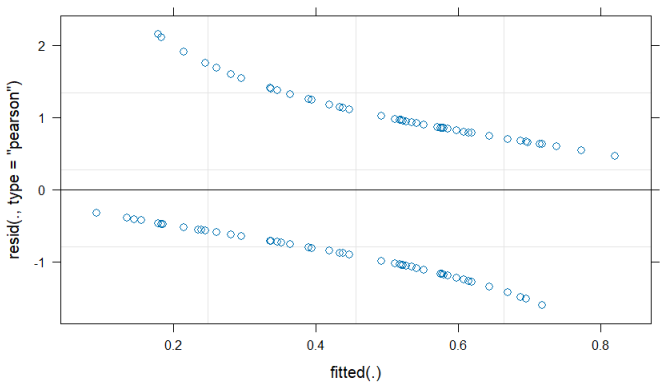

Figure 6: Caption

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: bin_cphys ~ SmkLevel + (1 | Id)
##    Data: bin.mood
##
##      AIC      BIC   logLik deviance df.resid
##    643.3    660.1   -317.6    635.3      494
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5952 -0.7366 -0.3947  0.8203  2.1487
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Id     (Intercept) 1.107    1.052
## Number of obs: 498, groups:  Id, 96
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7199     0.4532   1.588  0.11220
## SmkLevel2    -0.5525     0.5117  -1.080  0.28023
## SmkLevel3    -1.3859     0.5174  -2.679  0.00739 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) SmkLv2
## SmkLevel2 -0.884
## SmkLevel3 -0.882  0.779
```

```r
model.9=glmer(bin_cphys~1+(1|Id),data = bin.mood,family="binomial")
anova(model.9,model.8)
```

```
## Data: bin.mood
## Models:
## model.9: bin_cphys ~ 1 + (1 | Id)
## model.8: bin_cphys ~ SmkLevel + (1 | Id)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model.9    2 649.12 657.54 -322.56   645.12
## model.8    4 643.26 660.10 -317.63   635.26 9.8651  2   0.007208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```