STAT/BIOSTAT 571: Homework 8 – Key

To be handed in on Weds March 2nd, in class. Please see 'Chapter 0' of the slides for a summary of how to answer questions appropriately, and the guidelines from 570. Where solutions require use of `R`, summarize your findings in a written answer, and append your **annotated** code, to show what you did. For each question, write up your solution on your own, using **full sentences**.

1. [**Random slopes**] This question uses the models fit on slide 3.67, and variations of them

   (a) For the model with random intercepts and slopes where their covariance is unconstrained (`lme2b`) is the fitted correlation between the random intercepts and slopes random positive or negative? Without using any technical language, interpret the sign of this correlation; what two subject-level characteristics about reaction time and sleep deprivation are more likely to occur together?

   Answer: The fitted correlation between the random intercepts and slopes is 0.753. The interpretation is that the reactions of subjects who are slower on Day 1 tend to get worse quicker, as the study progresses. Alternatively, you could also state that those with fast reactions on Day 1 tend to be less badly affected, as the study goes on.

   (b) Using the fitted model with independent random slopes and intercepts (`lme2a`) make a graph showing the marginal mean and 2.5% and 97.5% quantiles of possible observations at each of the time points recorded. How do these quantities change if we fit the model where `Days` variable is not centered, but everything else is identical? Based on the data – and assuming everything else about the models is specified correctly – do you think the centered or uncentered version is more appropriate?

   Answer: The quantiles of centered case and uncentered case are shown in **Figure 1**.

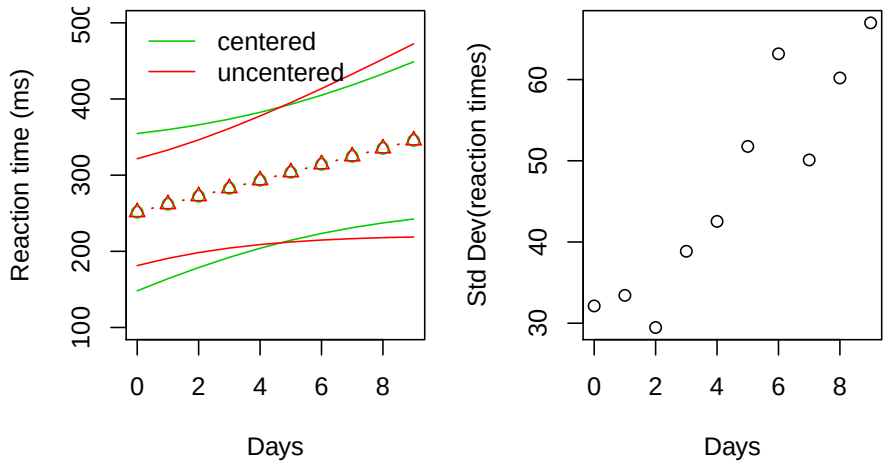


Figure 1: The marginal means and 95% prediction intervals on different time points for data with centered and uncentered variable Days.

The marginal mean values are extremely close under both models, but the quantiles of predicted observations differ. For the uncentered case, the range of the values we would expect to see become larger as the time increases, whereas those from the centered case are wider on both ends than in the middle.

To explain the different trends in these two plots, one can take a look at the marginal variance of the marginal mean.

$$Var(Y_i|X_i) = X_i G X_i + \sigma_Y^2 I \tag{1}$$

The diagonal entries of $X_i G X_i$ are the squared value of the centered days of sleep deprivation, which stay steady in the centered case and increase all the way in the uncentered case.
Based on the empirical standard deviations (see RH plot of Figure 1) the variability does seem to increase with time, suggesting informally that the uncentered version is more appropriate. (You could also look at diagnostic plots from the two models to address this question.)

2. [**Coverage of Emprical Bayes intervals**] As seen in class, we obtain the predicted random effects $\tilde{\boldsymbol{b}}_i$ as approximately the conditional means (or modes) of the true $\boldsymbol{b}_i$, conditional on the outcomes and covariates. As noted on slide 3.68, using the same approximations we can also obtain the conditional variances of the true random effects. When fitting mixed models using `lmer()` or `glmer()` from the `lme4` package, this approximate conditional standard deviations can be obtained straightforwardly using the `se.ranef()` function in the `arm` package.

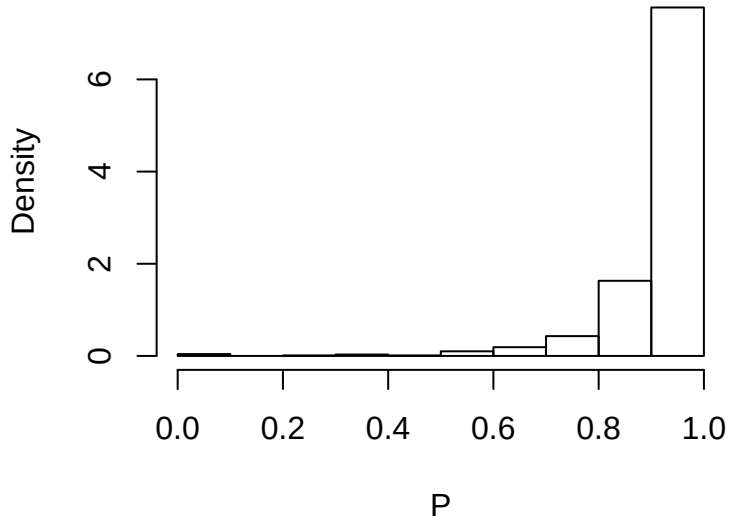
   (a) Using simulation, evaluate the proportion P of the n b_i's that are covered, using intervals of the form $\tilde{b}_i \pm 1.96 \times \tilde{s}_i$, where $\tilde{s}_i$ denotes the approximate conditional standard deviation above. Use the LMM where

$$\begin{aligned} b_i &\sim N(0, \sigma_b^2) \\ Y_{ij}|b_i &\sim N(b_i, \sigma_Y^2), \end{aligned}$$

   with $n = 1000$, $n_i = 5$, $\sigma_b = 1$, $\sigma_y = 6$. As well as stating the average of P over many replicates – which gives the average coverage for all b_i in all experiments – you should also describe the distribution of P, again over many replicates.

   Answer: We generate 1000 datasets from the stated data generated mechanisms and fit linear mixed models with only random intercepts. We calculate coverage P for each dataset, and obtain the mean of P (from 1000 datasets) is 0.916. We plot the empirical distribution of P in the following figure:

Figure 2: Histogram of P

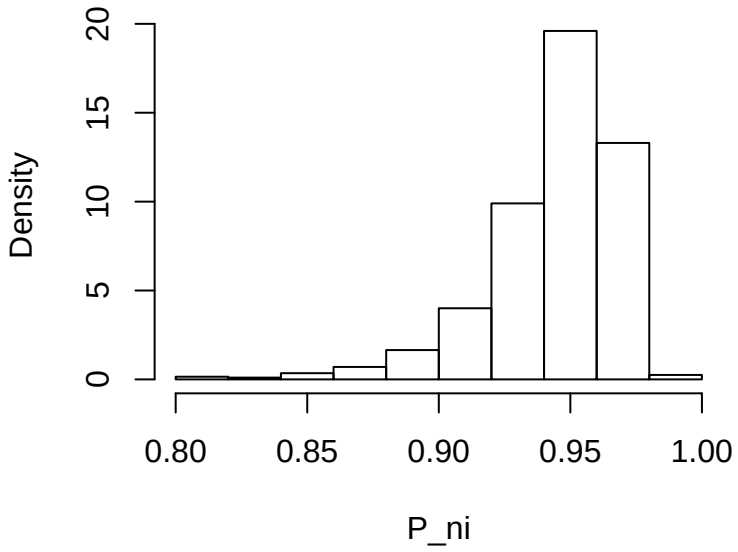


We see the empirical distribution of P is heavily left-skewed, with relatively good coverage in most cases but absolutely terribly coverage in a few extreme cases.

(b) Keen people: repeat a) but change the values of $n_i, \sigma_b^2, \sigma_Y^2$, and report what you find.

Answer: We repeat the experiment but change $n_i = 10$. Our mean coverage is improved, at 0.944.

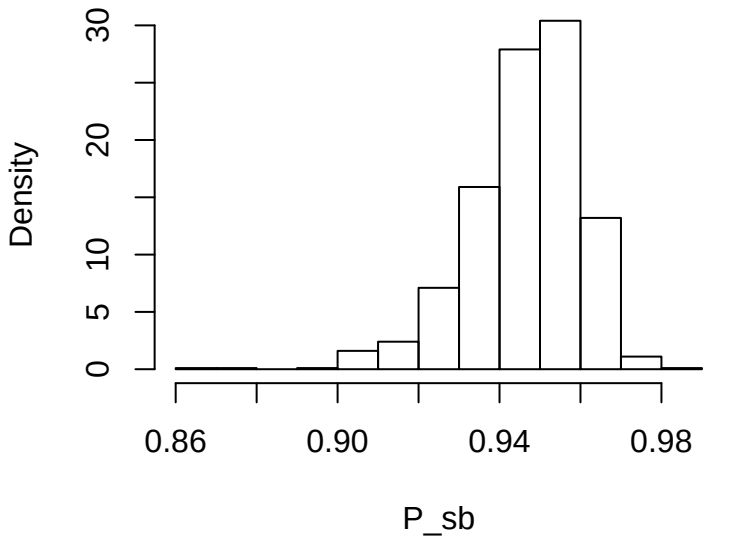
Figure 3: Histogram of P



The empirical distribution of P is also much more closely centered around its mean, although it still is left-skewed. One would expect the opposite effect to happen if we decreased n_i, with worsening coverage and more variance in coverage estimates.

We repeat the experiment but change $s_b = 2$ (resetting $n_i = 5$). Our mean coverage is improved, at 0.947.

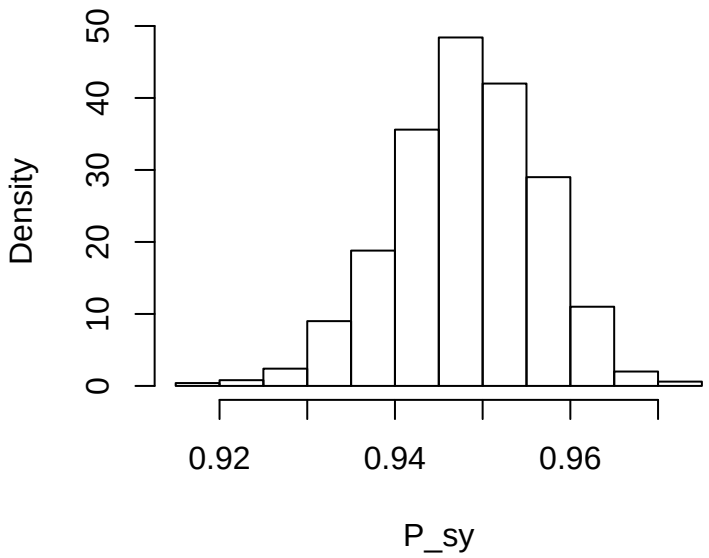
Figure 4: Histogram of P



The empirical distribution of P is also much more closely centered around its mean, although it still is left-skewed (albeit less). One would expect the opposite effect to happen if we decreased σ_b, with worsening coverage and more variance/skewness in coverage estimates.

We repeat the experiment but change $\sigma_y = 2$ (resetting $\sigma_b = 1$). Our mean coverage is improved, at 0.949.

Figure 5: Histogram of P



The empirical distribution of P is also much more closely centered around its mean, and is roughly symmetric. One would expect the opposite effect to happen if we increased σ_y, with worsening coverage and more variance/skewness in coverage estimates.

3. [**Fitting GLMMs**] Do Q9.4 from Jon's book, noting the following;

   - The data is available on the course site – the `epicalc` package is no longer available
   - To code the sine and cosine terms, for time recorded in quarters use `sin(0.5*pi*(time+1))` and `cos(0.5*pi*(time+1))`
   - I have provided `age` in years, not months. As in the book, this variable has also been roughly centered – so negative 'ages' are not data-entry errors
   - As well as implementing particular models, the 'interpret' parts of this question are very important. In these parts, you should describe the parameters being estimated in clear but accurate language

Answer:

- γ_1 is the difference in log odds of respiratory infection, comparing females to males but holding all other covariates constant, where this contrast is averaged across all subjects
- γ_2 is the difference in log odds of respiratory infection per unit difference in hfora, again assuming all other covariates are held constant and comparing across all subjects
- γ_5 is the difference in log odds of respiratory infection, comparing those on xerophthalmia to those not, assuming all other covariates are held constant and comparing across all subjects
- γ_6 and γ_7 are most easily interpreted together; they describe the linear and quadratic terms in differences in log odds, comparing subjects who are one year apart in age, assuming all other covariates are held constant and comparing across all subjects
- γ_0, γ_3 and γ_4 are trickier to interpret here as we cannot have simultaneously have $\texttt{cos}_{ij}$ and $\texttt{sin}_{ij}$ both equal to zero. Instead, you could (for example) report that $\gamma_0 - \gamma_3$ is the log odds

of respiratory infection, at time=1, averaging over observations from males with hfora=0, xero=0, and age=0. Similarly, $\gamma_0 - \gamma_4$ is the same quantity at time=2, and $\gamma_0 + \gamma_3$ is the same quantity at time=3.

4. After running a GEE with the independence working covariance assumption and present the parameter estimates and standard errors in the table below:

Table 1: Parameter estimates and standard errors from GEE

	Estimate	Robust S.E.
γ_0	-2.05	0.212
γ_1	-0.50	0.239
γ_2	-0.04	0.024
γ_3	-0.59	0.171
γ_4	-0.17	0.146
γ_5	0.58	0.420
γ_6	-0.37	0.095
γ_7	-0.16	0.063

5. In a GLMM version of the analysis;
   - γ_2 is the within-subject difference in log odds of respiratory infection per unit difference in hfora, again assuming all other covariates are held constant
   - γ_6 and γ_7 are most easily interpreted together; they describe the within-subject linear and quadratic terms in differences in log odds, comparing observations one year apart in age, assuming all other covariates are held constant
   - γ_0, γ_3 and γ_4 are again trickier to interpret here as we cannot have simultaneously have $\texttt{cos}_{ij}$ and $\texttt{sin}_{ij}$ both equal to zero. Instead, you could (for example) report that $\gamma_0 - \gamma_3$ is the log odds of respiratory infection, at time=1, for observations from males with hfora=0, xero=0, age=0 and who are otherwise 'average' (i.e. have $b_i = 0$). Similarly, $\gamma_0 - \gamma_4$ is the same quantity at time=2, and $\gamma_0 + \gamma_3$ is the same quantity at time=3.
   - γ_1 is also more difficult, as within-subject sex will not change. Hence we can interpret it as the difference in log odds of respiratory infection, comparing females to males but holding all other covariates constant, where this contrast is made in subjects who are otherwise equally 'unusual' (i.e. have the same b_i)
   - γ_5 is the within-subject difference in log odds of respiratory infection, comparing time points on xerophthalmia to those not, assuming all other covariates are held constant. (For subjects whose treatment never changes, we have to resort to language like that for γ_1 above.

- We implement GLMM logistic analysis using glmer() and obtain the following parameter estimates and standard errors:

Table 2: Parameter estimates and standard errors from GLMM

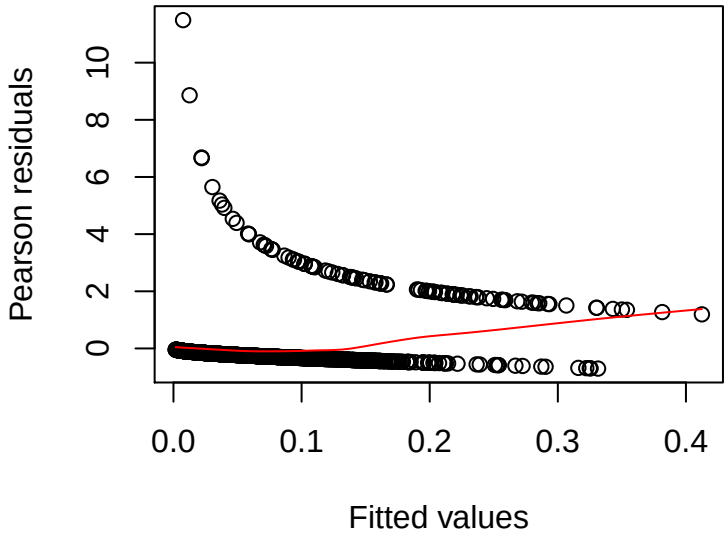
	Estimate	Standard Error
γ_0	-2.24	0.241
γ_1	-0.51	0.253
γ_2	-0.04	0.022
γ_3	-0.61	0.174
γ_4	-0.17	0.173
γ_5	0.54	0.481
γ_6	-0.37	0.095
γ_7	-0.15	0.058

- From our GEE model, we found that xeropthalmia seems to be associated with an increase in the log-odds for the presence of respiratory infection (averaging over all observations from all individuals), albeit in a statistically insignificant way. Our GLMM logistic analysis suggests that this relationship still holds when examining individual children. Namely, presence of xeropthalmia seems to be associated with an increase in the log-odds for the presence of respiratory infection within individual children. However, this relationship was also insignificant, so statistically one might say there is not enough evidence to reject the null hypothesis of no marginal or conditional relationship between respiratory infection and xeropthalmia under the GEE and GLMM models respectively. Any attentuation of the GLMM effects, due to non-collapsibility, seems mild and probably swamped by uncertainty in the point estimates.

6. [**Diagnostics**] Give diagnostics for the mean model in the GLMM fit in Q3. Say what you are plotting and what, if any, violation of the assumptions is indicated

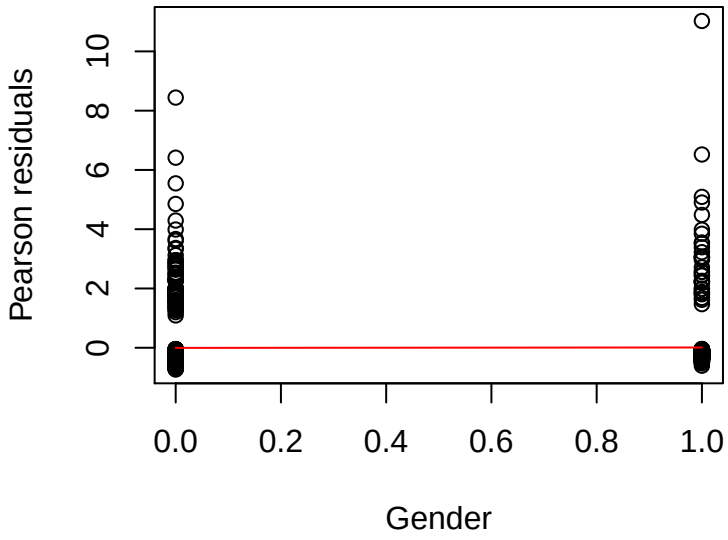
Answer: To check the appropriateness of the mean model, we can compare the Pearson residuals to the fitted values as well as against various covariates. First, we compare the Pearson residuals to the fitted values.

Figure 6: Fitted values vs Pearson residuals



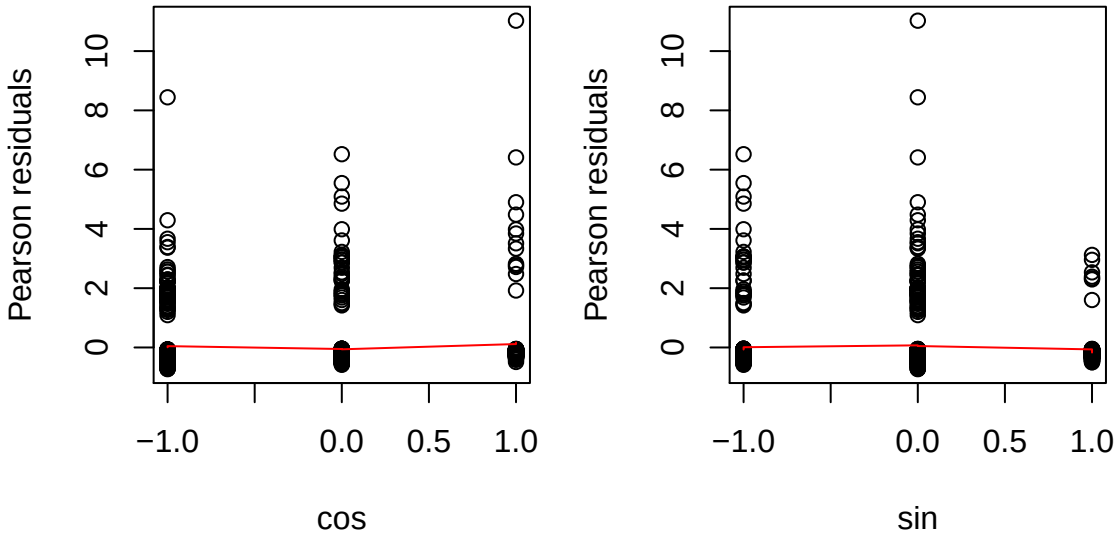
From this figure we see the distribution of residuals is decidedly right-skewed, and definitely non-normal. However, this doesn't necessarily say anything about our choice of mean model. Next, we compare our Pearson residuals to various covariates.

Figure 7: Pearson residuals vs Gender



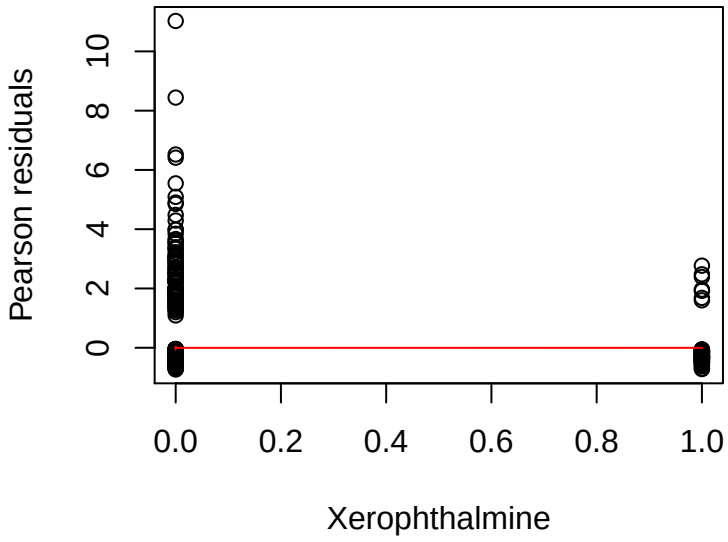
The mean of the residuals is the same for both genders, suggesting the mean model seems not to be violated.

Figure 8: Pearson residuals vs cos and sin



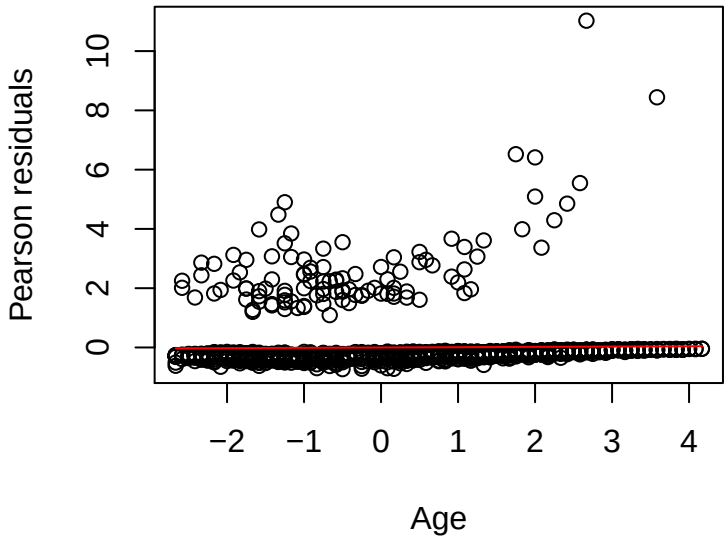
While the variance of the Pearson residuals seems to differ across the varying values of cos/sin, the mean seems to be constant. This suggests a violated of the mean-variance relationship but not necessarily of the mean model.

Figure 9: Pearson residuals vs Xerophthalmine



Again, we see a difference in the spread of residuals between the two groups of residuals but not a difference in the mean of the residuals.

Figure 10: Pearson residuals vs age



Superficially, there looks like there might be a bit of mean model misfit with respect to age (suggesting we might want to fit a more flexible higher order trend). However, the fitted spline suggests that there are enough residuals around 0 such that this overwhelms differences in the distribution of residuals away from zero.

Based on these residual plots, there seems to be little to no evidence for mean model misspecification. On the other hand, we see evidence of influential clusters and violations of our mean-variance model relationship.

R Code

```
#### problem 1 ####
library(lme4)
data(sleepstudy)
View(sleepstudy)
sleepstudy_uncentered = sleepstudy
sleepstudy$Days = with(sleepstudy,Days-mean(Days))
sleepstudy
# (a)
library(nlme)
lme2b = lme(Reaction~Days,random=~Days|Subject,data=sleepstudy)
summary(lme2b)
methods(class=(lme))
# (b)
lme2a = lme(Reaction~Days,random=reStruct(~Days|Subject,pdClass = "pdDiag"),data=sleepstudy)
summary(lme2a)
# get marginal mean
mean_marg = unique(fitted(lme2a,level = 0))
# get marginal sd
Z_i = cbind(rep(1,10),sleepstudy$Days[1:10])
G = getVarCov(lme2a)
sigma_Y = lme2a$sigma
mean_sd = sqrt(diag(sigma_Y^2+Z_i%*%G%*%t(Z_i)))
# uncentered
lme2a_uncentered = lme(Reaction~Days,random=reStruct(~Days|Subject,pdClass = "pdDiag"),
                       data=sleepstudy_uncentered)
# get marginal mean
mean_marg_uncentered = unique(fitted(lme2a_uncentered,level = 0))
# get marginal sd
Z_i_uncentered = cbind(rep(1,10),sleepstudy_uncentered$Days[1:10])
G_uncentered = getVarCov(lme2a_uncentered)
sigma_Y_uncentered = lme2a_uncentered$sigma
mean_sd_uncentered = sqrt(diag(sigma_Y_uncentered^2+Z_i_uncentered%*%G_uncentered%*%t(Z_i_uncentered)))
# plot
library(Hmisc)
par(mar=c(4,4,2,2))
ymax = max(mean_marg+mean_sd*qnorm(0.975),mean_marg_uncentered+mean_sd_uncentered*qnorm(0.975))
ymin = min(mean_marg-mean_sd*qnorm(0.975),mean_marg_uncentered-mean_sd_uncentered*qnorm(0.975))
errbar(1:10,mean_marg,mean_marg+mean_sd*qnorm(0.975),mean_marg-mean_sd*qnorm(0.975),ylim=c(ymin,ymax),
       main="Marginal means with 95% intervals at 8 time points",
       xlab="Time points",ylab="Marginal mean",col=4,lwd=3,errbar.col = 3)
errbar(1:10,mean_marg_uncentered,mean_marg_uncentered+mean_sd_uncentered*qnorm(0.975),
       mean_marg_uncentered-mean_sd_uncentered*qnorm(0.975),col=4,lwd=2,errbar.col = 2,add=T)
legend(1,470,legend=c("centered","uncentered"),col=c(3,2),lwd=2,lty=1,bty="n")

######Question 2############
library(arm)

#Parameters
n = 1000
ni = 5
s_b = 1
s_y = 6
reps = 1000

#Gen data and calc coverage
coverage <- function() {
  b = rnorm(n,0,s_b)
  id = as.numeric(sapply(1:n, function(i) rep(i,ni)))
  y = rnorm(n*ni,as.numeric(sapply(b, function(i) rep(i,ni))),s_y)
  model <- lmer(y ~ 1|id)
```

```
  mean(ranef(model)$id + qnorm(0.025)*se.ranef(model)$id < b & ranef(model)$id + qnorm(0.975)*se.ranef(model)$id > b)
}

#4 loops r 2 slow
start.time = Sys.time()
P = sapply(1:reps, function(i) coverage())
stop.time = Sys.time()
stop.time - start.time

mean(P)

pdf("P_hist.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
hist(P, freq = FALSE, main = "")
dev.off()

#Part (b)

#Increase ni
ni = 10

start.time = Sys.time()
P_ni = sapply(1:reps, function(i) coverage())
stop.time = Sys.time()
stop.time - start.time

mean(P_ni)

pdf("P_ni_hist.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
hist(P_ni, freq = FALSE, main = "")
dev.off()

#Increase s_b
ni = 5
s_b = 2

start.time = Sys.time()
P_sb = sapply(1:reps, function(i) coverage())
stop.time = Sys.time()
stop.time - start.time

mean(P_sb)

pdf("P_sb_hist.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
hist(P_sb, freq = FALSE, main = "")
dev.off()

#Decrease s_y
s_b = 1
s_y = 2

start.time = Sys.time()
P_sy = sapply(1:reps, function(i) coverage())
stop.time = Sys.time()
stop.time - start.time

mean(P_sy)

pdf("P_sy_hist.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
hist(P_sy, freq = FALSE, main = "")
```

```
dev.off()

########Homework 3############
library("geeM")
model1 <- geem(respinfect ~ sex + ht.for.age + I(cos(0.5*pi*(time+1))) +

 I(sin(0.5*pi*(time+1))) + xerop + I(age) +I(age^2),

 id=id, data=Xerop, family="binomial", corstr="exchangeable", scale.fix=1)



summary(model1)



library("lme4")

model2 <- glmer( respinfect ~ sex + ht.for.age + I(cos(0.5*pi*(time+1))) +

 I(sin(0.5*pi*(time+1))) + xerop + I(age) +I(age^2) + (1|id),

data=Xerop, family=binomial, nAGQ=10)
summary(model2)

##########Homework 4###########

which(model_coeff_norm[6,] == min(model_coeff_norm[6,]))
model_coeffs[,263]

#Examine residuals for mean-fit
resid <- resid(model2, type = "pearson")

pdf("diagnostic1.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
plot(fitted(model2),resid, ylab = "Pearson residuals", xlab = "Fitted values")
#lines(smooth.spline(fitted(model2),resid), col = "red")
lines(lowess(fitted(model2),resid, iter=0), col = "red")
dev.off()

pdf("diagnostic2.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1))
plot(xerop$sex,resid, ylab = "Pearson residuals", xlab = "Gender")
segments(0,mean(resid[xerop$sex==0]),1,mean(resid[xerop$sex==1]), col = "red")
dev.off()

pdf("diagnostic3.pdf",6,3)
par(mar = c(4.1,4.1,1.1,1.1), mfrow = c(1,2))
plot(xerop$cos,resid, ylab = "Pearson residuals", xlab = "cos")
lines(smooth.spline(jitter(xerop$cos,amount = 1e-5),resid), col = "red")
plot(xerop$sin,resid, ylab = "Pearson residuals", xlab = "sin")
lines(smooth.spline(jitter(xerop$sin,amount = 1e-5),resid), col = "red")
dev.off()

pdf("diagnostic4.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1), mfrow = c(1,1))
plot(xerop$xerop,resid, ylab = "Pearson residuals", xlab = "Xerophthalmine")
lines(smooth.spline(jitter(xerop$xerop,amount = 1e-5),resid), col = "red")
dev.off()

pdf("diagnostic5.pdf",4,3)
par(mar = c(4.1,4.1,1.1,1.1), mfrow = c(1,1))
plot(xerop$age,resid, ylab = "Pearson residuals", xlab = "Age")
```

```
lines(smooth.spline(xerop$age,resid), col = "red")
dev.off()
```