85%

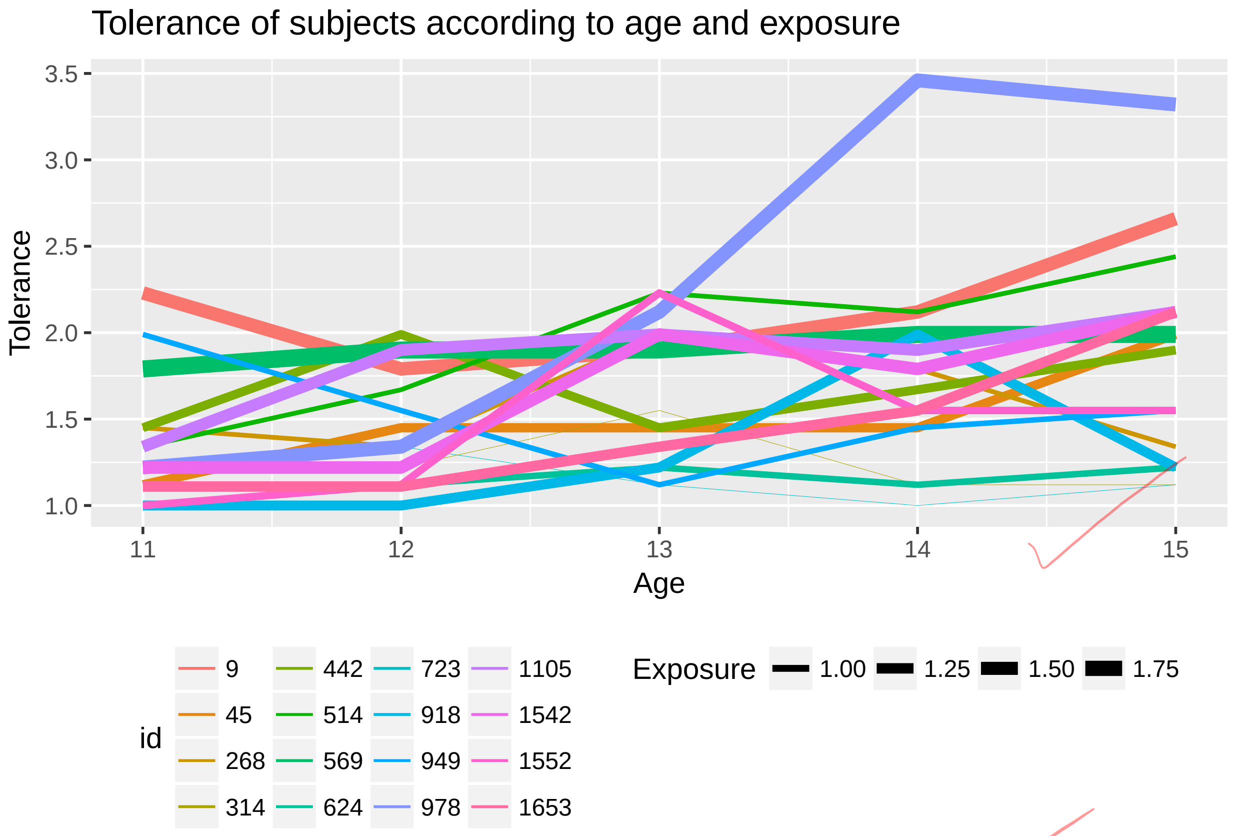
Good. Be careful to ensure you're mathematical description precisely matches what is actually done.

M6003 coursework

17974649

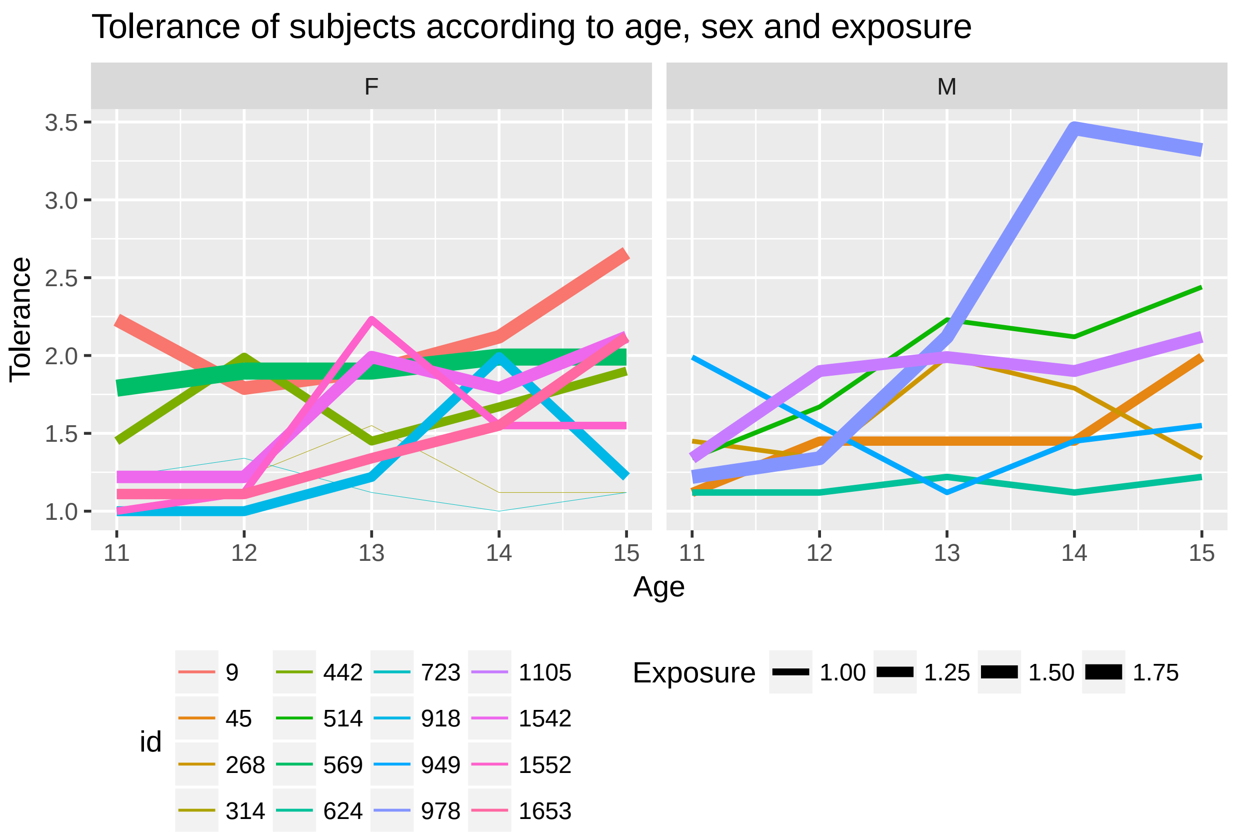
16 April 2018

The tolerance levels of youths as they progress through adolescence is assessed in this document. The dataset contains data on the tolerance of 16 youths to deviant behaviour. The average self reported exposure level to deviant behaviour at age of 11 was measured on a four point scale (0 = none, 4 = all) for nine activities, for each youth. The tolerance value represented the respondent's average score across the nine tolerance questions. Tolerance measurements were taken for each youth at ages 11, 12, 13, 14 aned 15 through self assessment. A plot of the data is displayed below



On visual inspection of the data, it appears that tolerence gradually increases with age. Subject 978 is noted to display relatively higher tolerance levels at age 14 and 15 than the other subjects. Excluding Subject 978, there appears a relatively constant variation in tolerance levels for each age group and there doesn't appear to be a relationship between exposure and tolerance. A large variation in the intercept of the plot due to subject variation was noted, and as such a random effects plot was felt appropriate for the data.

Further plots of tolerance at different ages are shown for both sexes.



There appears an increase in variation with age of tolerance levels in males compared to females (excluding subject 978). On average the sample of females may also have been exposed to higher levels of deviant behaviour than that of the males.

Model

A random effects model was constructed that took into account subject variation. As the model investiaged tolerance for each age, it was thought pertinent to include a random effect of the gradient as well as the intercept term in the model. The proposed model and the output data are demonstrated below

```
fm<- lmer(Tolerance~ Exposure + Age + Sex + (1+Age|id), data=YouthSurvey)
fm_sum <- summary(fm)
fix <- fm_sum$coefficients
fm_sum
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Tolerance ~ Exposure + Age + Sex + (1 + Age | id)
##    Data: YouthSurvey
##
## REML criterion at convergence: 69.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.597 -0.544 -0.136  0.203  2.735
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 3.3157   1.821
##           Age         0.0223   0.149    -1.00
##  Residual             0.0741   0.272
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  -0.8400     0.5829   -1.44
## Exposure      0.5759     0.1809    3.18
## Age           0.1308     0.0431    3.04
## SexM          0.1662     0.1161    1.43
##
## Correlation of Fixed Effects:
##          (Intr) Exposr Age
## Exposure -0.387
## Age      -0.913  0.000
## SexM     -0.160  0.197  0.000
```

One can see that the random effects variables *intercept* and *Age* are highly correlation (cor=-1.00). This indicates that the variance of *Tolerance* is heteroscedastic. The model is defined as

$$Tolerence = (\alpha + a_i) + \beta_1 \times Exposure + (\beta_2 + b_i) \times Age + \beta_3 \times Sex(M) + \epsilon_{ij}$$

✓

where

$\hat{\alpha} = -0.84, \hat{\beta}_1 = 0.576, \hat{\beta}_2 = 0.131, \hat{\beta}_3 = 0.166$
, $a_i \backsim N(0, 1.821), b_i \backsim N(0, 0.149), \epsilon_{ij} \backsim N(0, 0.349)$

The model you fit had correlation between a_i & b_i.

and

- *i=subject*, i=1,. . . ,16
- *j=Age*, j=11,12,. . . ,15

As $\hat{\beta}_2$ is positive in the model, and random effects correlation between intercept and slope equals -1.00, this indicates that variance of the response variable *Tolerance* decreases with increasing *Age* values.

Tolerence vs Age

Is there a tendency for tolerance of deviant behaviour to either increase or decrease as youths progress through adolesence? From our model we see that the fixed effect coefficient for Age ($\hat{\beta}_2$) = 0.131. The standard deviation for the random effect Age is 0.149. While acknowledging that the random effect for Age is correlated with the random effect for the intercept, the 95% confidence interval for $\hat{\beta}_2$ was calculated using the random effect for *Age* alone.

$$\hat{\beta}_2 = 0.131 \pm 1.96 \times 0.149$$

Therefore the 95% confidence interval for $\hat{\beta}_2$ is (-0.162, 0.423). As the 95% confidence interval includes zero, this suggests that tolerance score does not change with age.

Average tolerance between males and females

In order to compare if there is a difference in tolerance between the fixed effect of Sex, the MLE of the restricted model representing females was subtracted from the MLE of the full model representing males, excluding the Age covariate. As age has been excluded from the fixed effects, it was also excluded from the random effects. Bootstrapping was performed to obtain an appropriate H_0 distribution. The hypothesis tested was

- H_0: the average tolerance of males equals the average tolerance of females

- H_1: the average tolerance of males does not equal the average tolerance of females

The code for both models follows:

Why not include Age here?

```
lm.full <- lmer(Tolerance~ Exposure  + Sex + (1|id), data=YouthSurvey, REML=FALSE)
lm.reduced <- lmer(Tolerance ~ Exposure + (1|id), data=YouthSurvey, REML=FALSE)
obs.test.stat <- -2*(logLik(lm.reduced)-logLik(lm.full))
```

The log likelihood for the full model was -45.294 and the log likelihood for the reduced model was -47.081. The observed test statistic of $-2 * (logLik(lm.reduced) - logLik(lm.full))$ was 3.576.

In order to obtain the distribution to compare the observed test statistic, the bootstrapping technique was applied. The bootstrapping code is shown below:

```
N<-200
boot.test.stats<-rep(0,N)
for(i in 1:N){
  # Generate new data from the reduced model
  y.new<-unlist(simulate(lm.reduced))
  # Fit both models to the new data
  lm.full.new<-lmer(y.new ~ Exposure + Sex + (1|id), data=YouthSurvey, REML=FALSE)
  lm.reduced.new<-lmer(y.new ~ Exposure + (1|id), data=YouthSurvey, REML=FALSE)
  # Calculate the test statistic for the new models
  boot.test.stats[i]<- -2*(logLik(lm.reduced.new) - logLik(lm.full.new))
}
avg <- mean(boot.test.stats>obs.test.stat)
```

The p-value of 0.095 obtained from comparing the observed test statistic of the likelihood ratio to that of the bootstrap simulated data indicates that there insufficient evidence to reject the null hypothesis at the 95% level that the average tolerance of men is equal to the average tolerance of women. Therefore it is concluded that there is no difference in the average tolerance between males and females.

Tolerance between males and females at different ages

To determine if there was a difference in tolerance between males and females over time, the following hypothesis was tested

- H_0: β_2 males only model $=$ β_2 females only model
- H_1: β_2 males only model $\neq$ β_2 females only model

This isn't tested by

The MLE of the restricted model representing females was subtracted from the MLE of the full model representing males. The two models are shown below

```
lm.full.age <- lmer(Tolerance~ Exposure  + Age + Sex + (1+Age|id), data=YouthSurvey, REML=FALSE)
lm.reduced.age <- lmer(Tolerance~ Exposure  + Age + (1+Age|id), data=YouthSurvey, REML=FALSE)
obs.test.stat.age <- -2*(logLik(lm.reduced.age)-logLik(lm.full.age))
```

The log likelihood for the full model was -28.234 and the log likelihood for the reduced model was -29.295. The observed test statistic of $-2 * (logLik(lm.reduced.age) - logLik(lm.full.age))$ was 2.12.

Applying the bootstrap technique to create a distribution to compare the test statistic is shown below

You would need an Age*Sex term

```r
N<-200
boot.test.stat<-rep(0,N)
for(i in 1:N){
  # Generate new data from the reduced model
  y.new.age<-unlist(simulate(lm.reduced.age))
  # Fit both models to the new data
  lm.full.new.age<-lmer(y.new.age ~ Exposure  + Age + Sex + (1+Age|id), data=YouthSurvey, REML=FALSE)
  lm.reduced.new.age<-lmer(y.new.age ~ Exposure  + Age +  (1+Age|id), data=YouthSurvey, REML=FALSE)
  # Calculate the test statistic for the new models
  boot.test.stat[i]<- -2*(logLik(lm.reduced.new.age) - logLik(lm.full.new.age))
}
avg.age <- mean(boot.test.stat>obs.test.stat.age)
```

The p-value of 0.235 obtained from comparing the observed test statistic of the likelihood ratio to that of the bootstrap simulated data indicates that there insufficient evidence to reject the null hypothesis at the 95% level. Therefore it is concluded that there is no difference in the change in tolerance between males and females over time.

Prediction of tolerance

A male chosen at random from the study population who reported an exposure of 1.8 to deviant behaviour at age 15 represents a univariate distribution.

The expectation and variance of tolerance follow the following formulas

$$E[\mathbf{Y}] = \mathbf{XB}$$
$$Var[\mathbf{Y}] = \mathbf{X}Var(\mathbf{B})\mathbf{X}^T + \mathbf{Z}Var(\mathbf{b})\mathbf{Z}^T + \sigma^2 I_n$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & 1.8 & 15 & 1 \end{pmatrix}$$

$$\mathbf{B^T} = \begin{pmatrix} -0.840 & 0.576 & 0.131 & 0.166 \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 15 \end{pmatrix}$$

$$Var(\mathbf{B}) = \begin{pmatrix} 0.340 & -0.041 & -0.023 & -0.011 \\ -0.041 & 0.033 & 0 & 0.004 \\ -0.023 & 0 & 0.002 & 0 \\ -0.011 & 0.004 & 0 & 0.013 \end{pmatrix}$$

$$Var(\mathbf{b}) = \begin{pmatrix} 3.316 & -0.270 \\ -0.270 & 0.022 \end{pmatrix}$$

So at an age of fifteen, the expected tolerance is

$$E\left[\ Y\ \right] = \left[\ 1 \quad 1.8 \quad 15 \quad 1\ \right] \times \begin{bmatrix} -0.840 \\ 0.576 \\ 0.131 \\ 0.166 \end{bmatrix} = 2.325$$

and the $Var[\mathbf{Y}] = 0.0353 + 0.2141 + 0.0741 = 0.324$. The standard deviation $\mathbf{Y}$ is therefore 0.569. The r code for Var(B) and Var(b), E[Y] and V[Y] respectively is

```
VarB <- vcov(fm)
fm_sum <- summary(fm); varb <- unlist(fm_sum$varcor[[1]])
varb <- matrix(c(varb[1], varb[2],varb[3],varb[4]), nrow=2 )
X <- matrix(c(1,1.8,15,1), nrow=1); B <- fixef(fm); Z <- matrix(c(1,15), nrow=1)
sigma <- (0.27224); sigma2 <- sigma^2
E_Y <- X%*%B; V_Y <- X%*%VarB%*%t(X)+Z%*%varb%*%t(Z)+sigma2
```

and σ was copied directly from the residual of the random effects output of the *fm* model. The 95% prediction interval at age 15 is $(2.325 - 1.96 * \sqrt{0.3235}), (2.325 + 1.96 * \sqrt{0.3235}) = (1.21{,}3.44)$. ✓

Now suppose this individual had a tolerance response of 2.0 at age 11. This data follows a bivariate distribution

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\right)$$

Here we let μ_1 be the estimated tolerance at age 11, and μ_2 be the estimated tolerance at age 15. As we already know S_1, we can predict S_2 by the formula

$$E[S_2|S_1] = \mu_2 + V_{21}V_{11}^{-1}(S_2 - \mu_2)$$

Right idea

$$V[S_2|S_1] = V_{22} - V_{21}V_{11}^{-1}V_{12}$$

where

$$\mathbf{V} = \mathbf{A}Var(\mathbf{B})\mathbf{A^T} + \mathbf{Z}Var(\mathbf{b})\mathbf{Z^T} + \sigma^2 I_2$$

Good.

$$\mathbf{A} = \left[\begin{array}{cccc} 1 & 1.8 & 11 & 1 \\ 1 & 1.8 & 15 & 1 \end{array}\right]$$

$$\mathbf{Z} = \left[\begin{array}{cc} 1 & 11 \\ 1 & 15 \end{array}\right]$$

and

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mathbf{AB}$$

Substituting the appropriate values into the covariance matrix $\mathbf{V}$ equation above yields

$$\mathbf{V} = \begin{pmatrix} 0.161 & -0.025 \\ -0.025 & 0.324 \end{pmatrix}$$

We see that $\mathbf{V_{11}} = 0.161$, $\mathbf{V_{12}} = \mathbf{V_{12}} = -0.025$, and $\mathbf{V_{22}} = 0.324$.
Solving for $\mathbf{AB}$ we obtain $\mu_2 = 2.325$, as expected, and $\mu_1 = 1.802$. However, we wish to find the conditional mean and variance of $S_2|S_1$. We solve this from the formulas previously stated. The conditional mean is

$$E[S_2|S_1] = 2.325 - (-0.025)(0.161)^{-1}(2 - 1.802) = 2.356$$

The conditional variance is

$$V[S_2|S_1] = 0.324 - (-0.025)(0.161)^{-1}(-0.025) = 0.320$$

So the 95% conditional prediction interval for the male at age 15 given his tolerance of 2.0 at age 11 was $((2.356 - 1.96 * \sqrt{0.3201}), (2.356 + 1.96 * \sqrt{0.3201})) = (1.247,\ 3.465)$.

It can be seen that there is little change between the mean (2.325) and conditional mean (2.356) and the variance (0.324) and conditional variance (0.320) for the subject of interest. ✓