2018 First Year Exam – Methods
Statistics
210-210B-210C
June 25, 2018
9:00 – 12:00

Instructions

- There are 4 questions on the examination, each with multiple parts. Select any 3 of them to solve.
- Your solutions to each of the 3 problems you solve should be written on separate sheets of paper. Label *each sheet* with your student id number, the problem number, and the page for that problem written in the upper right hand corner. For example, the labeling on a page might be:

  ID# 912346378
  Problem 2, page 3
- You have 3 hours to complete your solution. Please be prepared to turn in your exam at 12:00 noon.

METHODS 210-210B-210C (2018), Problem 1

[Note: Students may leave any numerical computations unevaluated in expression form.]

Percentage yields from a chemical reaction for changing temperature (x_1) and agitation speed (x_2) are as follows

Average Yields (%)		x_2: Agitation Speed		Marginal mean
		Fast (1)	Slow (-1)	
x_1: Temperature	High (1)	$\bar{y}_{HF}$ = 80	$\bar{y}_{HS}$ = 74	$\bar{y}_{H\cdot}$ = 77
	Low (-1)	$\bar{y}_{LF}$ = 78	$\bar{y}_{LS}$ = 70	$\bar{y}_{L\cdot}$ = 74
Marginal mean		$\bar{y}_{\cdot F}$ = 79	$\bar{y}_{\cdot S}$ = 72	$\bar{y}_{..}$ = 75.5

The factors are defined as

$$x_1 = 1 \; if \; temperature \; is \; high \; and - 1 \; if \; low$$

$$x_2 = 1 \; if \; agitation \; speed \; is \; fast \; and - 1 \; if \; slow$$

Each listed yield is actually the average of five (5) individual independent experiments. The variance of individual measurements can be estimated from the five replications in each cell. It is found that

$$s^2 = \frac{\sum_{i=L}^{H} \sum_{j=S}^{F} \sum_{k=1}^{5} (y_{ijk} - \bar{y}_{ij})^2}{4(5-1)} = 12.5$$

a) The 20 data points are to be fitted with the following multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where the error terms are iid N(0, σ^2) with unknown σ^2. Write down the model matrix **X** associated with the data set.

b) Find the least squares estimators for the regression coefficients and express them as functions of average yields $\bar{y}_{ij}'s$. Calculate the estimate values based on the data table provided above.

c) Complete the following ANOVA table for the no-interaction model in a):

Source	d.f.	SS	MS
x_1: Temperature			
x_2: Agitation Speed			
Residual			
Total (corrected)			---

d) Based on the ANOVA table, perform hypothesis testing at 5% significance level individually for each of the 2 hypotheses below. Define the critical values and state the decision rules for each testing clearly.

   i. H_0: $\beta_1 = \beta_2 = 0$ vs. H_a: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$
   ii. H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$

e) Provide a 95% 2-sided confidence interval for the difference of mean yield percentages between fast agitation speed and slow agitation speed. (Write in the needed distributional percentile in notation form with a clear definition.)

f) What is the 95% 2-sided prediction interval for a new observation at the normal temperature (i.e., $x_1 = 0$) and with the average agitation speed (i.e., $x_2 = 0$)? (Write in the needed distributional percentile in notation form with a clear definition.)

g) A criticism was made toward the above model and analyses because a potential interaction between x_1 and x_2 was neglected. How would you respond to the criticism?

METHODS 210-210B-210C (2018), Problem 2

[Note: Students may leave any numerical computations unevaluated in expression form.]

Data of response y_i and a continuous covariate x_3 are collected for three groups A, B, and C. A one-way ANCOVA model without interaction terms is to be fitted as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

where the error terms are independently distributed as N(0, σ^2) with unknown σ^2. The groups are coded as

$$x_{i1} = \begin{cases} 1, if\ group = A \\ 0, otherwise \end{cases}$$

$$x_{i2} = \begin{cases} 1, if\ group = B \\ 0, otherwise \end{cases}$$

Sample sizes for the three groups are denoted as $n_A,\ n_B,\ and\ n_C$, respectively.

a) Write down the vector **y** and the model matrix **X**, in which x_{i3} is to be represented by its notation and x_{i1} and x_{i2} by their values.

b) Interpret the meaning of each regression coefficient in the model.

Group	$\bar{y}$	$\bar{x}_3$	n	S_{33}	S_{3y}
A	$\bar{y}_A = 5$	$\bar{x}_{3A} = 8$	$n_A = 15$	$\sum_{i=1}^{n_A}(x_{i3} - \bar{x}_{3A})^2 = 10$	$\sum_{i=1}^{n_A}(x_{i3} - \bar{x}_{3A})(y_i - \bar{y}_A) = 6$
B	$\bar{y}_B = 9$	$\bar{x}_{3B} = 6$	$n_B = 16$	$\sum_{i=n_A+1}^{n_A+n_B}(x_{i3} - \bar{x}_{3B})^2 = 12$	$\sum_{i=n_A+1}^{n_A+n_B}(x_{i3} - \bar{x}_{3B})(y_i - \bar{y}_B) = 5$
C	$\bar{y}_C = 12$	$\bar{x}_{3C} = 4$	$n_C = 14$	$\sum_{i=n_A+n_B+1}^{n_A+n_B+n_C}(x_{i3} - \bar{x}_{3C})^2 = 14$	$\sum_{i=n_A+n_B+1}^{n_A+n_B+n_C}(x_{i3} - \bar{x}_{3C})(y_i - \bar{y}_C) = 7$

Notations: For group A, $\bar{y}_A$ is the group sample mean of response y and $\bar{x}_{3A}$ is the group sample mean of the covariate x_3. For groups B and C, $\bar{y}_B$ and $\bar{y}_C$ as well as $\bar{x}_{3B}$ and $\bar{x}_{3C}$ are similarly defined.

c) Using the statistics in the above table, calculate the LS estimates for the four regression coefficients.

Note: The remainder of the question is from an alternate set of data. The data set is exactly the same size (n_A = 15, n_B = 16, n_C = 14) and the same model was applied to the data. A summary of the regression output for the data and selected summary statistics are provided below. (The table includes an additional column relative to the table above. To make the column fit the summation limits have been deleted. They are the same as in the table above.) Use these data for parts (d), (e), (f), (g).

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9077  -2.3848   0.4552   2.1183  11.6258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6879     2.8160  -0.599    0.552
X1           24.7967     1.9289  12.856 5.69e-16 ***
X2            9.8757     1.7505   5.642 1.40e-06 ***
X3            1.8173     0.1787  10.168 8.99e-13 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Group	$\bar{y}$	$\bar{x}_3$	n	S33	S3y	Syy
A	$\bar{y}_A = 57.0$	$\bar{x}_{3A} = 18.6$	$n_A = 15$	$\sum_i (x_{i3} - \bar{x}_{3A})^2 = 194.5$	$\sum_i (x_{i3} - \bar{x}_{3A})(y_i - \bar{y}_A) = 602.8$	$\sum_i (y_i - \bar{y}_A)^2 = 1944.8$
B	$\bar{y}_B = 37.0$	$\bar{x}_{3B} = 15.8$	$n_B = 16$	$\sum_i (x_{i3} - \bar{x}_{3B})^2 = 143.9$	$\sum_i (x_{i3} - \bar{x}_{3B})(y_i - \bar{y}_B) = 342.7$	$\sum_i (y_i - \bar{y}_B)^2 = 917.7$
C	$\bar{y}_C = 23.9$	$\bar{x}_{3C} = 14.1$	$n_C = 14$	$\sum_i (x_{i3} - \bar{x}_{3C})^2 = 355.3$	$\sum_i (x_{i3} - \bar{x}_{3C})(y_i - \bar{y}_C) = 315.0$	$\sum_i (y_i - \bar{y}_C)^2 = 336.6$

Notations: For group A, $\bar{y}_A$ is the group sample mean of response y and $\bar{x}_{3A}$ is the group sample mean of the covariate x_3. For groups B and C, $\bar{y}_B$ and $\bar{y}_C$ as well as $\bar{x}_{3B}$ and $\bar{x}_{3C}$ are similarly defined.

```
Residual standard error: 4.707 on 41 degrees of freedom
Multiple R-squared:  0.9195,   Adjusted R-squared:  0.9136
F-statistic: 156.1 on 3 and 41 DF,  p-value: < 2.2e-16
```

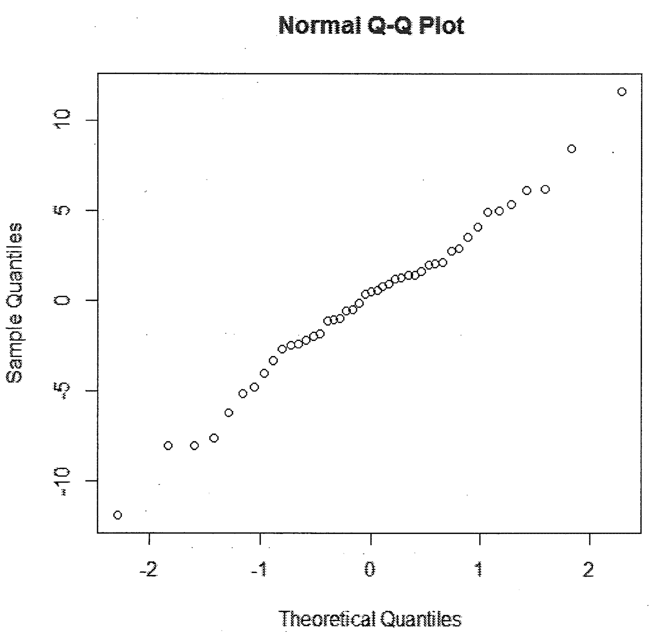
d) Calculate the Adjusted Group Means for Groups A, B, and C, respectively.

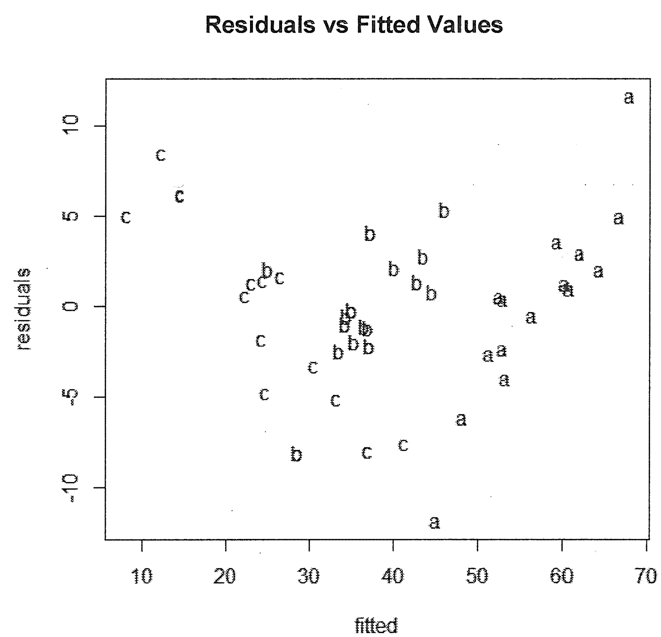
e) Assume that the (corrected) Total Sum of Squares for response y = 11284.2. Complete the following ANOVA table.

Source	d.f.	SS	MS
$SS_R(\beta_3 \mid \beta_0)$			
$SS_R(\beta_1, \beta_2 \mid \beta_3, \beta_0)$			
Residual			
Total (corrected)		11284.2	---

f) Use the ANOVA table to perform a hypothesis test at the 5% significance level for H_0: $\beta_1 = \beta_2 = 0$ vs the alternative that at least one of these coefficients is non-zero. Define the test statistic, obtain the critical value, and state your conclusion clearly.

g) Two residual plots are provided below. In the right-hand plot the observations are identified by plotting the group to which they belong. Based on these plots, comment on the appropriateness of the modeling assumptions. If you identify any possible problems with the model, explain how you would address these weaknesses.





METHODS 210-210B-210C (2018), Problem 3

Exposure to asbestos is known to increase the risk of several types of lung cancer. The Occupational Safety and Health Administration has issued regulations governing construction worker exposure to asbestos, with a legal limit of 0.1 fibers per cubic centimeter per 8-hour workday. One-half of this legal limit is considered to be an "action level" where intervention to reduce exposure is considered. We classify the exposure to asbestos at three levels (Low, < 0.05; Action Level, between 0.05 and 0.1; and Above Legal Limit > 0.1). Let

$$Y \text{ denote the exposure level, with } Y = \begin{cases} 1 & \text{for } \textit{Low Level}, \\ 2 & \text{for } \textit{Action Level}, \\ 3 & \text{for } \textit{Above Legal Level}; \end{cases}$$

$$X \text{ denote the task, with } X = \begin{cases} 0 & \text{for } \textit{Insulation}, \\ 1 & \text{for } \textit{Tile}; \end{cases}$$

and

$$Z \text{ denote the ventilation, with } Z = \begin{cases} 0 & \text{for } \textit{Ordinary}, \\ 1 & \text{for } \textit{Negative Pressure}. \end{cases}$$

The exposure levels of 83 construction workers were measured under four working conditions defined by the levels of X and Z, with the following frequency table.

		Exposure		
Task	Ventilation	Low Level	Action Level	Above Legal Limit
Insulation	Ordinary	3	3	22
Tile	Ordinary	3	1	2
Insulation	Negative Pressure	10	1	7
Tile	Negative Pressure	29	1	1

We fit the following proportional odds models (for $j = 1, 2$) to the data, and the R output is provided below.

$$\begin{aligned} &\text{Model 1: } \operatorname{logit}[P(Y \le j)] = \alpha_j \\ &\text{Model 2: } \operatorname{logit}[P(Y \le j)] = \alpha_j + \beta_1 X \\ &\text{Model 3: } \operatorname{logit}[P(Y \le j)] = \alpha_j + \beta_1 X + \beta_2 Z \\ &\text{Model 4: } \operatorname{logit}[P(Y \le j)] = \alpha_j + \beta_1 X + \beta_2 Z + \beta_3 XZ \end{aligned}$$

Use the R output to do the following (a)-(e):

(a) Test for conditional independence between exposure and ventilation given task. Use the likelihood ratio test. Interpret the results of the test in the context of the problem.

(b) Given Model 4, calculate the estimated response probability mass function (i.e. the estimated response probabilities, not the cumulative probabilities) for insulation under ordinary ventilation conditions.

(c) Given Model 4, what is the conditional odds ratio of exposure less than or equal to "Action Level" given working with tile under negative pressure to that of exposure less than or equal to "Action Level" given working with tile under ordinary ventilation conditions? Interpret your results. Calculate a 95% confidence interval for the odds ratio.

(d) Is there significant effect modification between task and ventilation? Write out your null hypothesis and use the R output to perform the test.

(e) Interpret the estimated coefficient $\hat{\beta}_1$ (or an appropriate transformation) in Model 3.

R output

Model 1:

```
Call:
vglm(formula = cbind(LowExposure, ActionLevel, AboveLegalLimit) ~
   1, family = cumulative(parallel = T), data = asbestos2)

Pearson Residuals:
  logit(P[Y<=1]) logit(P[Y<=2])
1       -3.63065       -2.94830
2       -0.61672        0.64888
3        0.22485       -0.16012
4        3.55049        2.63855

Coefficients:
                Value Std. Error t value
(Intercept):1 0.16908    0.22031 0.76744
(Intercept):2 0.46609    0.22552 2.06676

Number of linear predictors: 2
Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
Dispersion Parameter for cumulative family: 1
Residual Deviance: 49.70563 on 6 degrees of freedom
Log-likelihood: -73.80985 on 6 degrees of freedom
Number of Iterations: 4
```

```
Covariance Matrix:
               (Intercept):1 (Intercept):2
(Intercept):1    0.04853801    0.04282771
(Intercept):2    0.04282771    0.05085781
```

Model 2:

```
Call:
vglm(formula = cbind(LowExposure, ActionLevel, AboveLegalLimit) ~
   Task, family = cumulative(parallel = T), data = asbestos2)

Pearson Residuals:
  logit(P[Y<=1]) logit(P[Y<=2])
1       -1.71855        -1.0532
2       -2.33484        -1.2245
3        2.40805         1.0960
4        0.80955         0.8740

Coefficients:
                 Value Std. Error t value
(Intercept):1 -0.96123    0.32198 -2.9854
(Intercept):2 -0.52089    0.30336 -1.7171
TaskTile       2.82934    0.57638  4.9088

Number of linear predictors: 2
Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
Dispersion Parameter for cumulative family: 1
Residual Deviance: 17.45047 on 5 degrees of freedom
Log-likelihood: -57.68227 on 5 degrees of freedom
Number of Iterations: 4

Covariance Matrix:
               (Intercept):1 (Intercept):2    TaskTile
(Intercept):1    0.10367136    0.08283787 -0.10170201
(Intercept):2    0.08283787    0.09202523 -0.08370633
TaskTile        -0.10170201   -0.08370633  0.33221566
```

Model 3:

```
Call:
vglm(formula = cbind(LowExposure, ActionLevel, AboveLegalLimit) ~
   Task + Ventilation, family = cumulative(parallel = T), data = asbestos2)
```

```
Pearson Residuals:
  logit(P[Y<=1]) logit(P[Y<=2])
1       -0.49511       0.563749
2       -0.41716      -0.028693
3        0.49346      -0.859902
4        0.11898       0.357811

Coefficients:
                             Value Std. Error t value
(Intercept):1              -1.9713    0.47408 -4.1582
(Intercept):2              -1.4256    0.44027 -3.2379
TaskTile                    2.2868    0.61821  3.6991
VentilationNegativePressure 2.1596    0.56530  3.8202

Number of linear predictors: 2
Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
Dispersion Parameter for cumulative family: 1
Residual Deviance: 2.00013 on 4 degrees of freedom
Log-likelihood: -49.9571 on 4 degrees of freedom
Number of Iterations: 4

Covariance Matrix:
              (Intercept):1 (Intercept):2 TaskTile Vent:NegPres
(Intercept):1       0.22474       0.18643 -0.11330     -0.17855
(Intercept):2        0.1864       0.19383 -0.09444     -0.15853
TaskTile            -0.1133      -0.09444  0.38217     -0.01273
Vent:NegPres        -0.1785      -0.15853 -0.01273      0.31956
```

Model 4:

```
Call:
vglm(formula = cbind(LowExposure, ActionLevel, AboveLegalLimit) ~
   Task * Ventilation, family = cumulative(parallel = T), data = asbestos2)

Pearson Residuals:
  logit(P[Y<=1]) logit(P[Y<=2])
1       -0.58918        0.39182
2       -0.17220        0.19039
3        0.63654       -0.71120
4       -0.13936        0.24234
```

```
Coefficients:
                                          Value Std. Error  t value
(Intercept):1                          -1.87946    0.49694 -3.78208
(Intercept):2                          -1.33574    0.46352 -2.88173
TaskTile                                1.93976    0.91383  2.12268
VentilationNegativePressure             1.97517    0.65459  3.01743
TaskTile:VentilationNegativePressure 0.64533    1.25237  0.51529

Number of linear predictors: 2
Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
Dispersion Parameter for cumulative family: 1
Residual Deviance: 1.72919 on 3 degrees of freedom
Log-likelihood: -49.82163 on 3 degrees of freedom
Number of Iterations: 4

Covariance Matrix:
              Inter:1 Inter:2 Task:Tile Vent:NegPres     I/A
(Intercept):1  0.2469  0.2084   -0.2312      -0.2316  0.2174
(Intercept):2  0.2084  0.2148   -0.2110      -0.2109  0.2133
TaskTile      -0.2312 -0.2110    0.8350       0.2232 -0.8278
Vent:NegPres  -0.2316 -0.2109    0.2232       0.4284 -0.4208
I/a            0.2174  0.2133   -0.8278      -0.4208  1.5684
```

END OF QUESTION (3)

METHODS 210-210B-210C (2018) Problem 4

A typical primary endpoing in smoking cessation trials is point-prevalence abstinence, which is defined as abstinence status during a window (typically 7 days) immediately preceding the assessment. The following dataset contains information on the weekly abstinence status of 300 individuals enrolled in a behavioral clinical trial aimed at comparing the efficacy of an innovative experimental treatment added to a standard educational approach. The innovative treatment employs a social media supportive network, whereas the standard approach is based on the distribution of an informative leaflet. The study subjects are followed through the program over 2 months (8 weeks) and their point-prevalence abstinence is assessed at the end of each week.

(a) For the following question, you may refer to the model mod1 in the Appendix. Write the mathematical form of the assumed model (the assumed model, not the fitted model). Clearly state all the modeling assumptions, with particular regard to the mean and covariance functions.

(b) Provide an interpretation of each estimated coefficient in mod1. Based on the estimated coefficients, comment on the efficacy of the innovative treatment for smoking cessation *vs* the standard leaflet approach, as a function of time. Motivate your answer.

(c) Discuss the asymptotic distribution of the estimator used in mod1. How would you expect the estimates to change if a different working variance-covariance structure were assumed?

(d) Now refer to model mod1.age in the Appendix. Provide an interpretation of the coefficient capturing the effect of Age on the probability of smoking cessation. Discuss the Wald test, as implemented in the package GEEPACK in R, and its appropriateness for binary outcomes.

(e) Now refer to model mod2 in the Appendix. Write the mathematical form of the assumed model in matrix form (the assumed model, not the fitted model). Clearly define any variables used, and write out the elements of each vector or matrix in the model. Identify which terms in the model are fixed, and which are random. State all model assumptions.

(f) Write a sentence interpreting the effect of time in the model. Does the interpretation between the marginal and conditional models differs? If so, how?

(g) Write a sentence interpreting the effect of treatment over time for the three specific subjects with id numbers 1, 3 and 5.

(h) Weight gain is often cited as a primary reason for interrupting any attempt to quit smoking. Suppose that the investigators have also recorded the weekly changes in the individuals' weight. Comment on the interpretation of the regression parameters relating the mean response to stochastically time-varying covariates (such as individual weekly weight gain) in marginal versus conditional mixed effects models.

Appendix

mod1

```
mod1 <- geeglm( smoking ~ week +
                factor(Treatment) * week,
                family = binomial(link = "logit"),
                data = smoking.data,
                id = id, corstr = "exchangeable")
summary(mod1)
##
## Call:
## geeglm(formula = smoking ~ week + factor(Treatment) * week, family = binomial(link = "logit"),
##     data = smoking.data, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                                Estimate  Std.err    Wald Pr(>|W|)
## (Intercept)                     2.11014  0.16838 157.059   <2e-16 ***
## week                           -0.41766  0.03984 109.909   <2e-16 ***
## factor(Treatment)Treatment     -0.38871  0.23563   2.721   0.0990 .
## week:factor(Treatment)Treatment  0.13012  0.05325   5.971   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)   0.9448 0.03087
##
## Correlation: Structure = exchangeable  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.2629 0.03163
## Number of clusters:   200   Maximum cluster size: 8
```

mod1.age

```
mod1.age <- geeglm( smoking ~  factor(Treatment) * week + Age,
                    family = binomial(link = "logit"),
                     data = smoking.data,
                     id = id, corstr = "exchangeable")
```

```
## coefficient of Age
grep("Age", mod1.age.summary, value = TRUE)[2]
## [1] "Age                                 0.01423  0.00784   3.29  0.06963 .  "
anova(mod1, mod1.age, test=FALSE)
## Analysis of 'Wald statistic' Table
##
## Model 1 smoking ~ factor(Treatment) * week + Age
## Model 2 smoking ~ week + factor(Treatment) * week
##   Df   X2 P(>|Chi|)
## 1  1 3.29      0.07 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

mod2

```
mod2 <- glmer(smoking ~ factor(Treatment) * week + (week|id),
            family=binomial,
           data=smoking.data)

summary(mod2)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: smoking ~ factor(Treatment) * week + (week | id)
##    Data: smoking.data
##
##      AIC      BIC   logLik deviance df.resid
##      892      926     -439      878      895
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.613 -0.279  0.269  0.363  2.668
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 0.234    0.484
##         week        0.254    0.504    1.00
## Number of obs: 902, groups:  id, 200
##
## Fixed effects:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       2.7737     0.3014    9.20  < 2e-16 ***
## factor(Treatment)Treatment       -0.6096     0.3759   -1.62      0.1
## week                             -0.5733     0.0951   -6.03  1.7e-09 ***
## factor(Treatment)Treatment:week   0.2113     0.1290    1.64      0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) fc(T)T week
## fctr(Trtm)T -0.726
## week        -0.582  0.432
## fctr(Trt)T:  0.402 -0.530 -0.726
cbind(ranef(mod2)$id[1:5,], smoking.data$Treatment[c(1,9,17,25,33)])
##   (Intercept)    week smoking.data$Treatment[c(1, 9, 17, 25, 33)]
## 1     -0.3635 -0.3787                                       Placebo
## 2      0.0961  0.1001                                     Treatment
## 3      0.0789  0.0821                                       Placebo
## 4     -0.5828 -0.6071                                     Treatment
## 5     -0.0829 -0.0863                                     Treatment
```