First Year Qualifying Exam

Methods 210, 211, 212

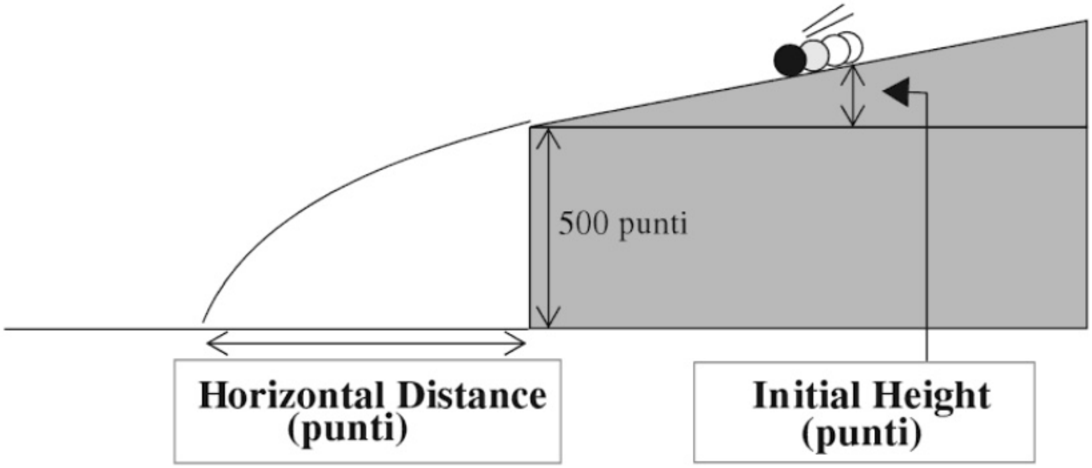
Tuesday, June 21, 2022
9:00 am-12:00 pm

- There are 4 questions on the examination. You are to do 3 of 4 questions.
- Your solutions to each problem should be written on separate sheets of paper. Only write on one side of each sheet of paper. Label each sheet with your student identification number, the problem number, and the page number of that solution written in the upper right hand corner. For example, the labeling on a page may be:

  ID# 912346378
  Problem 2, page 3

- You have 3 hours to complete your solution. Please be prepared to turn in your exam at 12:00pm.

1. As one of the pioneer of modern science, Galileo was among the first scientists who studied the laws of motion. During his inquiry on the role of inertia in motion (which finally led to Newton's first law of motion), he constructed an apparatus shown in the sketch below.
More specifically, he placed an inclined plane on a table, which was set at 500 *punti* above the floor (one punto=169/190 millimeter). Then, he released an ink-covered ball at different heights above the table, and measured the horizontal distance between the table and the ink spot left by the ball falling on the floor.



The measurements are displayed in the following table:

Horizontal Distance (*punti*)	Initial Height (*punti*)
253	100
337	200
395	300
451	450
495	600
534	800
573	1000

In this problem, we want to explore the relationship between the initial height and the horizontal distance. Answer the following questions. (*In order to receive full credit, please also include the formula/reasoning you use for obtaining the results*):

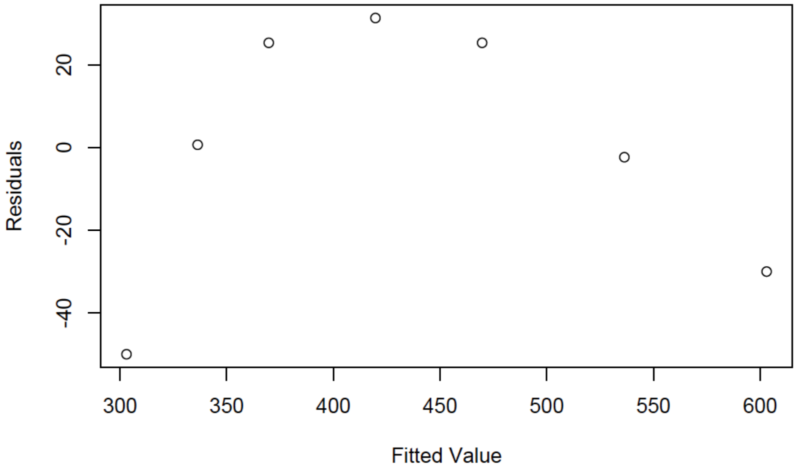
(a) First we will employ a simple regression model to analyze this data set:

$$DISTANCE = \beta_0 + \beta_1 HEIGHT + \varepsilon.$$

The output of the regression model in R and the residual plot are displayed below:

```
lm(formula = DISTANCE ~ HEIGHT)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 269.71246   24.31239  11.094 0.000104 ***
HEIGHT        0.33334    0.04203   7.931 0.000513 ***
---
Residual standard error: 33.68 on 5 degrees of freedom
Multiple R-squared:  ---,        Adjusted R-squared:  0.9116
F-statistic: 62.91 on 1 and 5 DF,  p-value: 0.0005132
```



Based on the output, do you believe that the simple linear regression proposed above is valid? If not, which assumptions are violated? State your reasoning.

(b) Please provide the 95% confidence interval for $\hat{\beta}_1$.

(c) State the null hypothesis of the F test listed in the R output in (a), and interpret the result of the test.

(d) Based on the R output above, calculate the R^2 value.

(e) Now we fit the data with a more complicated model, a model also know as polynomial regression. In particular, we will add another predictor: *Height*^{2}, the square of variable *Height*. We then fit the following regression model:

$$DISTANCE = \beta_0 + \beta_1 HEIGHT + \beta_2 HEIGHT^{\,2} + \varepsilon.$$

The above polynomial regression can be carried out easily in R: we only need to create a new variable *HEIGHT_SQUARE* that represents the square of *Height*, and then fit the response variable against both *Height* and *HEIGHT_SQUARE* using the techniques of multiple linear regression. The R output of the corresponding regression analysis is reported below:

```
lm(formula = DISTANCE ~ HEIGHT + HEIGHT_SQUARE)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.999e+02  1.676e+01  11.928 0.000283 ***
HEIGHT         7.083e-01  7.482e-02   9.467 0.000695 ***
```

```
HEIGHT_SQUARE -3.437e-04  6.678e-05  -5.147 0.006760 **
---
Residual standard error: 13.64 on 4 degrees of freedom
Multiple R-squared:  0.9903,       Adjusted R-squared:  0.9855
F-statistic:   205 on 2 and 4 DF,  p-value: 9.333e-05
```

Construct a **general F-test** to show that the introduction of the squared term does improve the fit significantly compared to the simple linear regression with only *Height*. Calculate the test statistic and explicitly compute the degrees of freedom.

(f) It is natural for us to think about adding more variables. For example, we could add another polynomial term, *HEIGHT_CUBIC*, that represents *Height*^{3}. Then, we can fit a multiple regression model that includes all the three predictors: *HEIGHT*, *HEIGHT_SQUARE*, *HEIGHT_CUBIC*. We display the ANOVA table below:

```
Analysis of Variance Table

Response: DISTANCE
             Df  SSR   Mean Sq
HEIGHT        1  71351   71351
HEIGHT_SQUARE 1   4927    4927
HEIGHT_CUBIC  1    696     696
Residuals     3     48      16
```

A general F-test shows that the new predictor, *HEIGHT_CUBIC*, should be added to the model. However, the best model is often the result of a trade-off between goodness of fit and model simplicity. Therefore, we consider the adjusted R-squared values of all the three models: (M1) simple linear regression, (M2) quadratic and (M3) cubic polynomial regression. Please, report the adjusted R^2 for the three models and then discuss your selection of the best model.

(End of Problem 1)

2. In this problem, we will consider the results of a teratology experiment in which female rats on iron-deficient diets were randomly assigned to one of the following four groups:

   I : placebo injection

   II : iron supplement injections on days 7 and 10 of the experiment

   III : iron supplement injections on days 0 and 7 of the experiment

   IV : iron supplement injections weekly for three weeks

   58 total rats were made pregnant (this defined the start of the experiment) and then sacrificed after three weeks. At that time, the total number of dead fetuses was counted for each litter. Experimenters are interested in the probability of a dead fetus conditional upon treatment group.

   (a) To begin, let us suppose that for each litter one fetus was selected randomly and labeled as alive or dead. Thus we observe independent responses $y_1, y_2, \ldots, y_{58}$, where $y_i = 1$ if the fetus was dead at the time the mother was sacrificed and 0 otherwise, $i = 1, \ldots, 58$. Based upon this information, suggest an appropriate probability model for the response variable and write down a generalized linear regression model that addresses the scientific question of interest.

   (b) Provide a precise interpretation of each parameter in the regression model you formulated in (a).

   (c) In order to obtain parameter estimates and draw inference for the above regression model, one could turn to the theory of generalized linear models provided that the probability distribution of the outcome is a member of the exponential dispersion family.

      i. Write down the form of the probability density function (pdf) for a member of the exponential dispersion family with canonical location parameter θ, dispersion parameter $a(\phi)$ and mean $b'(\theta)$.

      ii. Show that the probability model that you proposed for the teratology experiment is a member of the exponential dispersion family and identify each of the parts of the pdf.

   (d) Consider a regression model of the form $g(\mu_i) \equiv \eta_i = \boldsymbol{X}_i\boldsymbol{\beta}$, where μ_i denotes the mean of the response variable of interest, $\boldsymbol{X}_i$ is the i-th row of the design matrix, $\boldsymbol{\beta}$ is a vector of regression parameters, and $g(\cdot)$ is a differentiable function linking μ_i to the linear predictor, η_i. Using a generic likelihood pertaining to a member of the exponential dispersion family (in the form provided for (c-i)), derive the score equation used to obtain maximum likelihood estimates of $\boldsymbol{\beta}$.

   (e) Using the regression model you specified in (a) write down the score equation for estimating the model parameters (you may simply plug the relevant parts into the generic score equation you derived in (d)). Briefly explain how maximum likelihood estimates for the model parameters could be obtained in practice.

   (f) To begin to address the scientific question of interest, investigators wish to test the hypothesis that the probability of a dead fetus is the same across all treatment groups. Precisely state three different test statistics the could be used for testing this hypothesis given the regression model you formulated in (a) and state the asymptotic distribution of each statistic. Be sure to carefully define the components of each statistic you present.

Table 1: Observed counts of (litter size, number dead) for the teratology study.

Group	Response counts (Litter Size, Number Dead)
I	(10,1) (11,4) (12,9) (4,4) (10,10) (11,9) (9,9) (11,11) (10,10) (10,7) (12,12) (10,9) (8,8) (11,9) (6,4) (9,7) (14,14) (12,7) (11,9) (13,8) (14,5) (10,10) (12,10) (13,8) (10,10) (14,3) (13,13) (4,3) (8,8) (13,5) (12,12)
II	(10,1) (3,1) (13,1) (12,0) (14,4) (9,2) (13,2) (16,1) (11,1) (4,0) (1,0) (12,0)
III	(8,0) (11,1) (14,0) (14,1) (11,0)
IV	(3,0) (13,0) (9,2) (17,2) (15,0) (2,0) (14,1) (8,0) (6,0) (17,0)

(g) Since most litters are comprised of more than one fetus, the above analysis throws away a great deal of information. Suppose that we wish to use the information on all fetuses within a litter, where the observed data are now given in Table 1. Figure 1 displays four diagnostic plots (with corresponding scatterplot smoothers) produced after adding the data on the rest of each of the litters and fitting a GLM to the data assuming all observations were independent. Do any of the plots suggest a problem(s) with the model? Explain and describe implications of the problem(s).

(h) One approach to the analysis that includes all available data on each litter is to suppose that fetuses within a litter are correlated and that litters are independent. One could then consider the total number of dead fetuses per litter as the outcome variable. For a simple formalization, let Y_{ij} and Y_{ik} denote the response of fetus i and fetus k within litter i and suppose that $\text{corr}(Y_{ij}, Y_{ik}) = \rho$, $j, k = 1, \ldots, n_i$, $j \neq k$, $i = 1, \ldots, 58$. In this case, we could consider the binomial outcome $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ as the response variable. Derive the mean and variance of Y_i.

(i) A quasi-binomial regression model would assume $\text{Var}[Y_i] = \phi\mu_i(1 - \mu_i)/n_i$ where $\mu_i = \text{E}[Y_i]$ and ϕ could be estimated from the model residuals. Suppose that in fact $\text{corr}(Y_{ij}, Y_{ik}) = \rho$, $j, k = 1, \ldots, n_i$, $j \neq k$, $i = 1, \ldots, 58$ and that litters are independent. Using your result from (h) and the observed data given in Table 1, comment on the validity of the quasi-binomial model in this case. If you find the quasi-binomial model unacceptable, what other approach might you take to model these data in order to obtain consistent parameter estimates and valid inference?

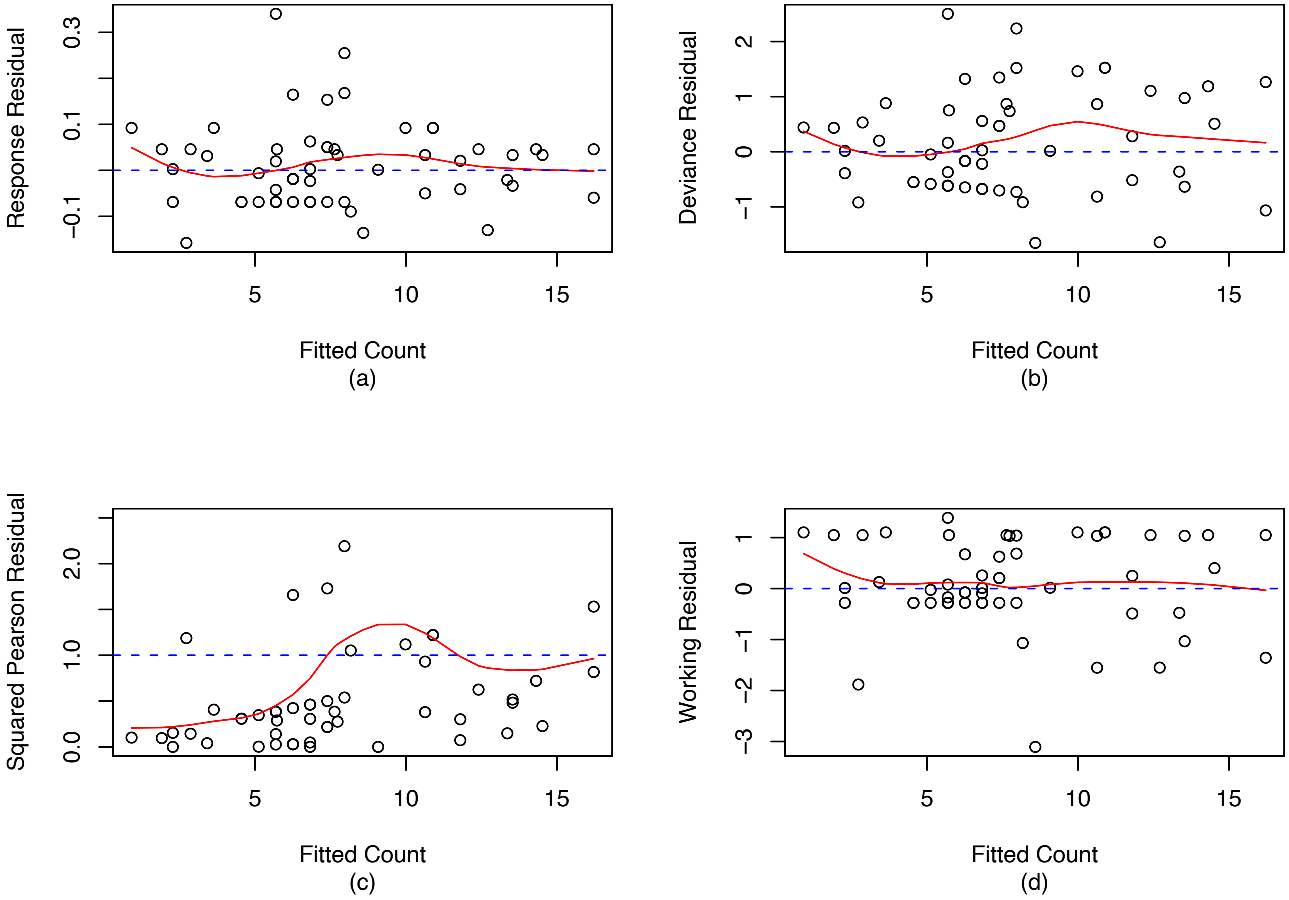


Figure 1: Residual diagnostic plots from model fitted to data in Table 1.

(End of Problem 2)

3. In the Vaccine Preparedness Study (VPS) a sample of 5,000 high risk HIV negative participants were enrolled and followed prospectively for 18 months. At baseline, participants were asked to report on the number of partners they had in the last six months that were of unknown HIV serostatus. Suppose it is reasonable to assume Z_i, the (actual) number of partners, is distributed as Poisson with a mean $\lambda(\boldsymbol{X}_i)$, that may depend on covariates $\boldsymbol{X}_i$. Interest is in whether drug and alcohol use correlate with (predict) the number of partners. Let Y_i be the reported number of partners and let $\boldsymbol{X}_i$ be covariates of interest.

   (a) Suppose we are concerned that a fraction of the participants might be uncomfortable reporting sensitive information and will choose to simply report $Y_i = 0$ even though $Z_i > 0$. Otherwise, participants report correctly, $Y_i = Z_i$. Let $\pi(\boldsymbol{X}_i)$ denote the probability that $Y_i = 0$ when $Z_i > 0$ for a subject with covariate $\boldsymbol{X}_i$. Describe a mixture model for Y_i and derive the mean and variance of Y_i in terms of $\pi(\boldsymbol{X}_i)$ and $\lambda(\boldsymbol{X}_i)$.

   (b) Suppose we are truly interested in the mean model $\log(E[Z_i]) = \log(\lambda_i) = \beta_0 + \boldsymbol{\beta}_1\boldsymbol{X}_i$, but having only observed Y_i we fit the model $\log(E[Y_i]) = \gamma_0 + \boldsymbol{\gamma}_1\boldsymbol{X}_i$. Further suppose that we obtain estimates $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_1$ by fitting a standard Poisson regression to the observed data.

      i Under what conditions, if any, will $\hat{\gamma}_0$ be a consistent estimator of β_0? Justify.

      ii Under what conditions, if any, will $\hat{\boldsymbol{\gamma}}_1$ be a consistent estimator of $\boldsymbol{\beta}_1$? Justify.

   (c) Again consider the scenario from Part (b). Under what conditions are the confidence intervals for parameters in $\boldsymbol{\gamma}_1$ which result from the standard Poisson regression asymptotically valid (ie. in large samples they obtain the correct coverage probability)? Justify. In cases where these confidence intervals are not asymptotically valid, if any, suggest a method which can be used to obtain confidence intervals with asymptotically correct coverage probability.

For the remaining questions that are based upon an analysis of the VPS data you may assume that under-reporting of past partners is not an issue...

Regression summaries (coefficient estimates and corresponding covariance matrices) for the VPS baseline data (a subset of 1000 men) are given on the following pages. We are primarily concerned with whether drug and alcohol use are associated with high risk behavior (as measured by the number of unknown HIV status partners an individual had in the past 6 months). The following mean model is of interest:

$$\log(E[Y_i]) = \beta_0 + \beta_1\texttt{age.c}_i + \beta_2\texttt{drugs}_i + \beta_3\texttt{alcohol}_i + \beta_4\texttt{drugs}_i \times \texttt{alcohol}_i$$

where

- `age.c` = age in years at entry into the study (centered to the sample mean of 26 years)
- `drugs` = 0 if no drug use; 1 if any drug use
- `alcohol` = 0 is alcohol use is < heavy; 1 if alcohol use is heavy

   (d) Consider models (1) and (2). Model (1) presents results from a standard Poisson regression analysis, and model (2) presents results from a *quasi-Poisson* (or scale overdispersion) regression analysis. Based solely upon these results which analysis would you prefer to report and why?

   (e) Based upon the model that you would choose to report, estimate the mean number of unknown HIV status partners in the past 6 months for an individual of age 26, who uses drugs and does not use alcohol. Provide a 95% confidence interval for this quantity.

(f) Provide a precise interpretation of β_4 (or some suitable transformation) in language understandable to a statistical layman. Based upon the model that you would choose to report, what is your estimate of this quantity? What are the implications of this estimate on the scientific question of interest?

(g) Consider testing $H_0 : \beta_4 = 0$ using a Deviance test based upon the *quasi-Poisson* model. Model (3) presents the output from a reduced quasi-Poisson fit omitting the `drugs` by `alcohol` interaction. Using the results from models (2) and (3) calculate the appropriate Deviance statistic for this test. What is the asymptotic distribution of this test statistic?

APPENDIX - Problem 3

```
(1) STANDARD POISSON ANALYSIS OF THE VPS DATA

> fit <- glm( y ˜ age.c + drugs + alcohol + drugs:alcohol, family=poisson )
> summary( fit )

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.291126   0.019751  65.371  < 2e-16 ***
age.c          0.009411   0.003229   2.914  0.00356 **
drugs          0.388216   0.051819   7.492 6.79e-14 ***
alcohol       -0.007742   0.060748  -0.127  0.89858
drugs:alcohol  0.349414   0.083927   4.163 3.14e-05 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3339.6  on 999  degrees of freedom
Residual deviance: 2971.0  on 995  degrees of freedom

> round( summary( fit )$cov.scaled, 4 )
              (Intercept) age.c   drugs alcohol drugs:alcohol
(Intercept)         4e-04     0 -0.0004 -0.0004        0.0004
age.c               0e+00     0  0.0000  0.0000        0.0000
drugs              -4e-04     0  0.0027  0.0004       -0.0027
alcohol            -4e-04     0  0.0004  0.0037       -0.0037
drugs:alcohol       4e-04     0 -0.0027 -0.0037        0.0070
```

```
(2) QUASI-POISSON ANALYSIS OF THE VPS DATA

> fit.quasi <- glm( y ~ age.c + drugs + alcohol + drugs:alcohol, family=quasipoisson )
> summary( fit.quasi )

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.291126   0.032584  39.624  < 2e-16 ***
age.c          0.009411   0.005327   1.766   0.0776 .
drugs          0.388216   0.085489   4.541 6.28e-06 ***
alcohol       -0.007742   0.100220  -0.077   0.9384
drugs:alcohol  0.349414   0.138460   2.524   0.0118 *

(Dispersion parameter for quasipoisson family taken to be 2.721754)

    Null deviance: 3339.6  on 999  degrees of freedom
Residual deviance: 2971.0  on 995  degrees of freedom

> round( summary( fit.quasi )$cov.scaled, 4 )
              (Intercept)  age.c   drugs alcohol drugs:alcohol
(Intercept)        0.0011  0e+00 -0.0011 -0.0011        0.0010
age.c              0.0000  0e+00 -0.0001  0.0000        0.0001
drugs             -0.0011 -1e-04  0.0073  0.0011       -0.0073
alcohol           -0.0011  0e+00  0.0011  0.0100       -0.0100
drugs:alcohol      0.0010  1e-04 -0.0073 -0.0100        0.0192
```

```
(3) QUASI-POISSON ANALYSIS OF THE VPS DATA (REDUCED MODEL)

> fit.quasi.red <- glm( y ~ drugs + alcohol + drugs:alcohol, family=quasipoisson )
> summary( fit.quasi.red )

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.270823   0.032064  39.634  < 2e-16 ***
age.c       0.008058   0.005328   1.512  0.13077
drugs       0.522743   0.064252   8.136 1.21e-15 ***
alcohol     0.171479   0.066346   2.585  0.00989 **

(Dispersion parameter for quasipoisson family taken to be 2.750395)

    Null deviance: 3339.6  on 999  degrees of freedom
Residual deviance: 2988.9  on 996  degrees of freedom
```

(End of Problem 3)

4. The world-renowned Dr. Pepper of UCI is interested in evaluating the efficacy of an innovative intervention they have developed to treat soda addiction. To understand the efficacy of the intervention, they consider a longitudinal clinical trial study where the intervention (coded "1" below) is compared to the standard of care (coded "0" below). Dr. Pepper is interested in investigating if the intervention leads to an increase in the ability to resist the temptation of drinking sodas, based on a continuous "Resist" score they have previously developed. Therefore, they enroll 200 participants (100 assigned to control, and 100 assigned to the new intervention arm) and then they compute the "Resist" score based on surveys and other assessments at baseline (week 0) and at follow-up visits at weeks 1, 2, and 3.

For all participants, information about gender is also recorded (with male coded "0" below and female "1").

Figure 1 in the Appendix reports a plot with the individual profiles ("spaghetti plot") as well as the weekly means of the resist scores from the individuals assigned to the two interventions.

Part 1. For the following questions, you can refer to the code in **Part 1** of the Appendix.

(a) Write the mathematical form of the model fit in `mod1.ML`. Write the 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.

(b) Provide the (general) expression of the maximum likelihood estimators of the fixed-effects parameters in a linear mixed effects model. Discuss their asymptotic properties and any assumption required for their validity.

(c) Based on the estimated coefficients in `mod1.ML`, determine if there is enough evidence of a treatment effect over time. Motivate your answer.

(d) Dr. Sprite is a good collaborator of Dr. Pepper. After reviewing the statistical report, Dr. Sprite suggests that Dr. Pepper should perform a formal likelihood ratio test using restricted maximum likelihood to assess the significance of the treatment effect over time. Discuss why or why not this may be a good suggestion.

(e) Discuss a test to determine if the mixed effects should consider a random slope or not. Clearly specify the hypotheses being tested, the estimation method and the relevant test statistic.

Part 2. For the following questions, you may consider the output reported in **Part 2** of the Appendix.

(a) Write the generalized estimating equation solved to obtain the estimates reported in model `mod2.ar1`. Clearly state all the assumptions of the model, in particular all the assumptions on the mean and variance-covariance structure. Identify the corresponding estimates in the reported output.

(b) Propose a test for assessing if the intervention leads to significantly different "Resist" scores than the standard of care. Clearly specify (and justify) the hypotheses being tested, the hypothesis testing approach and the relevant test statistic.

(c) With reference to the Gauss-Markov theorem for correlated data, discuss under what conditions the estimator $\hat{\boldsymbol{\beta}}_{GEE}$ from the GEE fit has minimal variance among all the linear estimators.

(d) In Pan (2001) *On the robust variance estimator in generalised estimating equations*, Biometrika 88(3): 901-906 it is proposed an alternative estimator for the variance of the outcome, say $\mathrm{Cov}\left(\mathbf{y}_i\right)$, in the expression of the sandwich estimator of the variance of $\hat{\boldsymbol{\beta}}_{GEE}$.

More specifically, Pan's alternate formulation changes the covariance of the outcome term to

$$\mathrm{Cov}\left(\mathbf{y}_i\right) = \mathbf{A}_i^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{A}_i^{-1/2}\mathbf{S}_i\mathbf{S}_i^{\mathrm{T}}\mathbf{A}_i^{-1/2}\right)\mathbf{A}_i^{1/2}$$

with $\mathbf{A}_i = \mathrm{diag}\left\{v\left(\mu_{i1}\right), \ldots, v\left(\mu_{in}\right)\right\}$ and $\mathbf{S}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$.

Explain the difference between Pan's formulation and the usual sandwich estimator of the variance. Discuss possible advantages of the new formulation.

(e) For large samples, do you think there will be a significant difference from the usual calculation?

Part 3. For the following question, you may consider also the output reported in **Part 3** of the Appendix.

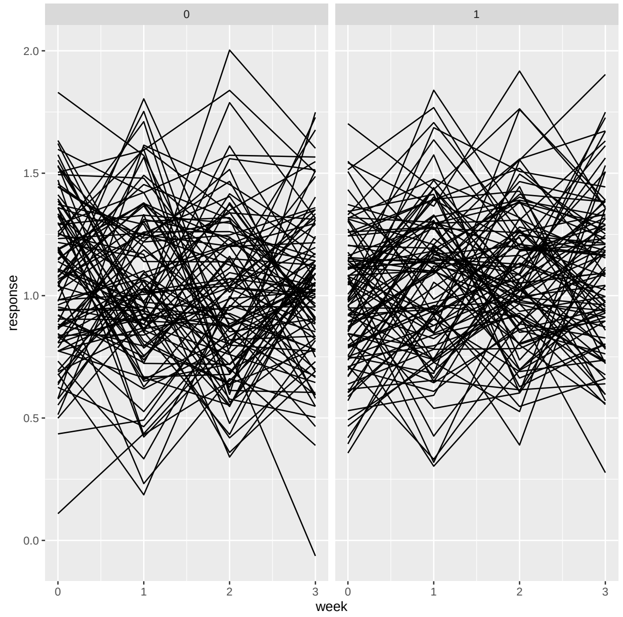
(a) Dr. Sprite suggests that the "Resist" score should be dichotomized. Hence, they suggest to fit both a conditional and a marginal model using the newly created dichotomized response. Discuss how the interpretation of the coefficients change in the new models, paying particular attention to the differences between the conditional and the marginal formulation. How would you expect the estimates in the conditional model to compare with respect to those in the marginal model?

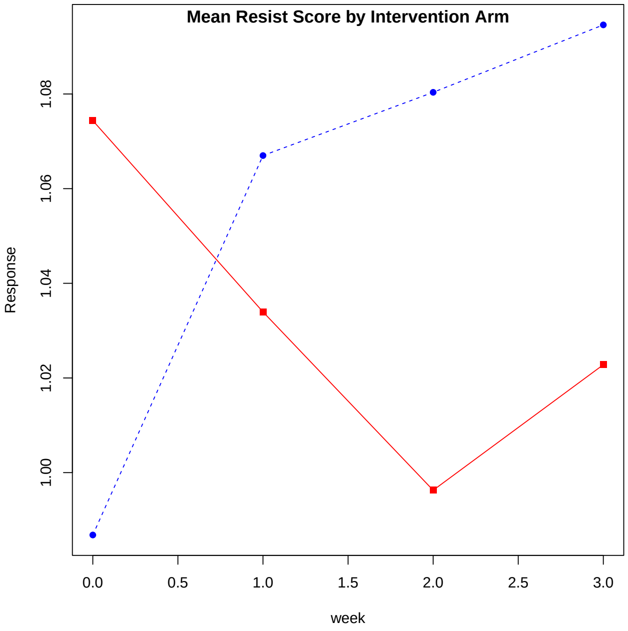
Part 4. For the following questions, you may consider the output reported in **Part 4** of the Appendix.

(a) Dr. Pepper notices that the data are affected by monotone dropout, that is some subjects stop to come to a follow-up visit and do not return in future visits. They believe that the dropout is more likely for males and it may also be associated to low "Resist" scores recorded at previous visits for those subject. Dr. Sprite suggests this may pose a problem for the validity of some of the methods used above. Identify the type of missing data mechanism discovered by Dr. Pepper. Then, briefly discuss the possible effects it may have on the estimates of the conditional and marginal models considered in the previous points, and some ways to remediate possible biases.

(b) Based on the output of `mod4.gee` in the Appendix, do you believe that Dr. Pepper was justified in their assessment about the missing data mechanism? Justify your answer.

Appendix - Problem 4





Part 1
mod1.ML

```
mod1.ML=lme(response ~  week*treatment, data=dati, random = ~ 1 + week |id, method="ML")
summary(mod1.ML)
## Linear mixed-effects model fit by maximum likelihood
##   Data: dati
##        AIC      BIC    logLik
##   395.6526 433.1295 -189.8263
##
## Random effects:
##  Formula: ~1 + week | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.16456137 (Intr)
## week        0.06738444 -0.598
## Residual    0.27329957
##
## Fixed effects:  response ~ week * treatment
##                     Value  Std.Error  DF  t-value p-value
## (Intercept)     1.0607155 0.02824254 598 37.55737  0.0000
## week           -0.0192425 0.01399181 598 -1.37527  0.1696
## treatment      -0.0540299 0.03994099 198 -1.35274  0.1777
## week:treatment  0.0529127 0.01978741 598  2.67406  0.0077
##  Correlation:
##                (Intr) week   trtmnt
## week           -0.738
## treatment      -0.707  0.522
## week:treatment  0.522 -0.707 -0.738
##
## Standardized Within-Group Residuals:
##           Min            Q1           Med            Q3           Max
## -3.0893331721 -0.5936415213  0.0005123469  0.6003593537  2.8921253882
```

```
##
## Number of Observations: 800
## Number of Groups: 200
```

Part 2
mod2.ar1

```
library(geepack)
mod2.ar1=geeglm(response ~  week*treatment, data=dati, family="gaussian", id=id,  corstr="ar1")
summary(mod2.ar1)
##
## Call:
## geeglm(formula = response ~ week * treatment, family = "gaussian",
##     data = dati, id = id, corstr = "ar1")
##
##  Coefficients:
##                Estimate  Std.err     Wald Pr(>|W|)
## (Intercept)     1.06205  0.02888 1352.691  < 2e-16 ***
## week           -0.01848  0.01469    1.583  0.20831
## treatment      -0.05906  0.03919    2.271  0.13184
## week:treatment  0.05299  0.01938    7.477  0.00625 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate  Std.err
## (Intercept)  0.09779 0.005185
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.2576 0.04502
## Number of clusters:   200  Maximum cluster size: 4
```

Part 3
mod3.glmer

```
library(lme4)

## Loading required package:  Matrix
##
## Attaching package:  ’Matrix’
## The following objects are masked from ’package:tidyr’:
##
##     expand, pack, unpack
##
## Attaching package:  ’lme4’
## The following object is masked from ’package:nlme’:
##
##     lmList

mod3.glmer=glmer(response.dico ~ week*treatment+(1|id),data=dati.dico,
        family=binomial)
summary(mod3.glmer)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: response.dico ~ week * treatment + (1 | id)
##    Data: dati.dico
##
##      AIC      BIC   logLik deviance df.resid
##   1102.1   1125.6   -546.1   1092.1      795
```

```
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.397 -0.984  0.708  0.837  1.105
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 0.351    0.593
## Number of obs: 800, groups:  id, 200
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.2046     0.1849    1.11     0.27
## week            -0.0349     0.0934   -0.37     0.71
## treatment       -0.1714     0.2614   -0.66     0.51
## week:treatment   0.1812     0.1330    1.36     0.17
##
## Correlation of Fixed Effects:
##             (Intr) week   trtmnt
## week        -0.760
## treatment   -0.707  0.537
## week:trtmnt  0.535 -0.703 -0.758
```

mod3.gee

```
mod3.gee=geeglm(response.dico ~ week*treatment,data=dati.dico,
                family = "binomial",
                id = id, corstr = "exchangeable")
summary(mod3.gee)
##
## Call:
## geeglm(formula = response.dico ~ week * treatment, family = "binomial",
##     data = dati.dico, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                Estimate Std.err Wald Pr(>|W|)
## (Intercept)      0.1885  0.1793 1.10     0.29
## week            -0.0322  0.0939 0.12     0.73
## treatment       -0.1574  0.2546 0.38     0.54
## week:treatment   0.1663  0.1331 1.56     0.21
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)        1 0.00944
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0918  0.0359
## Number of clusters:   200  Maximum cluster size: 4
```

Part 4
mod3.wgee

```
library(wgeesel)

mod4.wgee= wgee (response.dico ~  week*treatment,data=dati.dico,                          family = "binomial",
               id = id, corstr = "exchangeable", scale = NULL, mismodel =R ~ sex + lag1y)
summary(mod4.wgee)
## Call:
## wgee(model = response.dico ~ week * treatment, data = dati.dico,
##     id = id, family = "binomial", corstr = "exchangeable", scale = NULL,
##     mismodel = R ~ sex + lag1y)
```

```
##
##                 Estimates Robust SE z value Pr(>|z|)
## (Intercept)       0.12797   0.18965    0.67     0.50
## week              0.04804   0.10991    0.44     0.66
## treatment        -0.00189   0.26569   -0.01     0.99
## week:treatment    0.06186   0.15963    0.39     0.70
##
##  Estimated Scale Parameter:  1.26
##
##  Estimated Correlation:  0.0955
summary(mod4.wgee$mis_fit)
##
## Call:
## glm(formula = mismodel, family = binomial(), data = data[adjusted_idx,
##     ])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.904  -1.435   0.729   0.795   1.024
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.303      0.327    0.93    0.354
## sex            0.212      0.188    1.13    0.259
## lag1y          0.636      0.299    2.13    0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 685.57  on 599  degrees of freedom
## Residual deviance: 679.65  on 597  degrees of freedom
## AIC: 685.6
##
## Number of Fisher Scoring iterations: 4
```

(End of Problem 4)