First Year Qualifying Exam

Methods 210, 210B, 210C

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

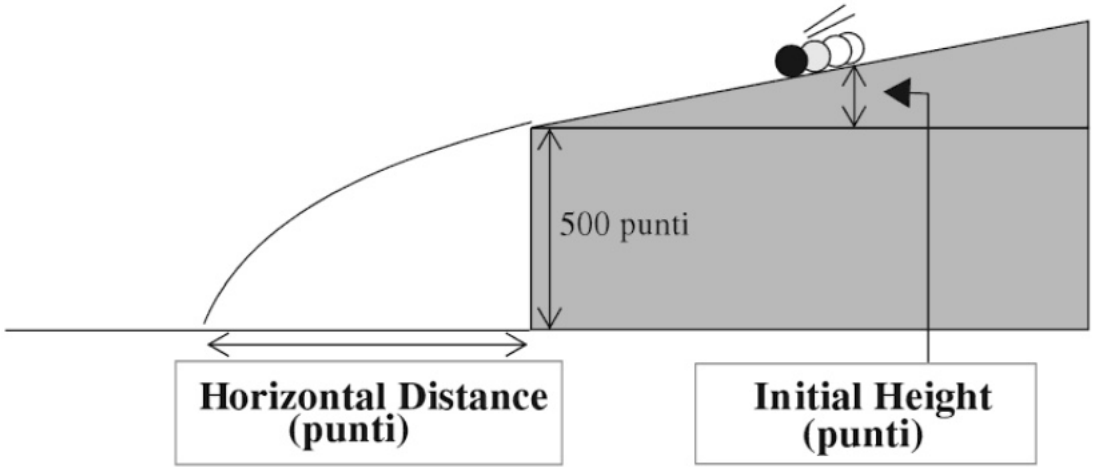
- There are 4 questions on the examination. You are to do 3 of 4 questions.
- Your solutions to each problem should be written on separate sheets of paper. Only write on one side of each sheet of paper. Label each sheet with your student identification number, the problem number, and the page number 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*):

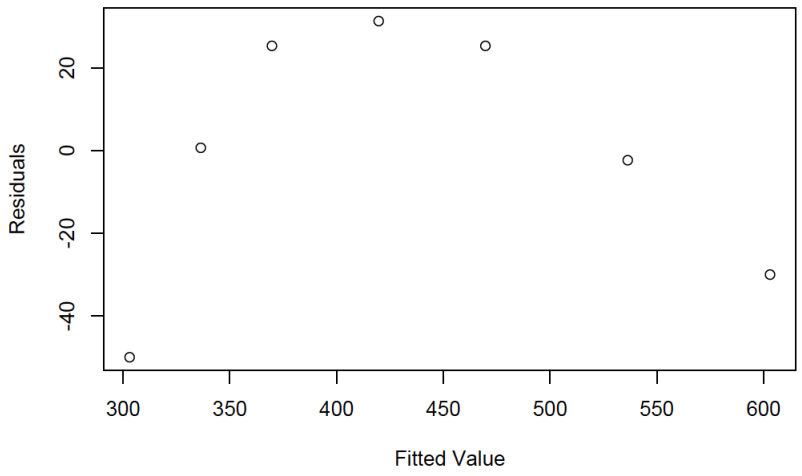
(a) First we will employ a simple regression model to analyze this data set:

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

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. Pediatric and epidemiological studies have shown that birth weight is an important predictor of children's health: for example, Lawlor et al. (2005) found an inverse relationship between birth weight and coronary heart disease and stroke. Factors that have been found to be risk factors for low birth weight (< 2,500g) are: mother's age, maternal smoking, educational level, race and socio-economic factors. Motivated by observations made on animals, epidemiologists have been investigating whether maternal exposure to air pollution during the course of the pregnancy increases the risk of low birth weight. A such study by Bell et al. (2007) used data from over 350,000 (n=358,504) singleton births registered in the states of Connecticut and Massachussets between 01/01/1999 and 12/13/2002. The study was limited to singleton births (i.e. only one child), with birthweight of at least 1,000g and gestational length of at least 32 weeks.

   In examining whether exposure to particulate matter is associated with the risk of low birth weight (<2,500g) compared to non-low birth weight (≥2,500g), the authors used the average concentration of $PM_{2.5}$ (reported in $\mu g/m^3$: microgram per m^3) during a mother's pregnancy as the exposure variable.

   The model also adjusted for the following covariates (bold text indicate the reference class):

   - Mother's marital status: **married**, unmarried
   - Tobacco use during pregnancy: Yes, **No**
   - Alcohol use during pregnancy: Yes, **No**
   - Education: < 12 yrs, **12 yrs**, 13-15 yrs, > 15 yrs
   - Mother's age: < 20, 20-24, 25-29, **30-34**, 35-39, >39 yrs, unknown
   - Mother's race: **white**, black, other
   - Child sex: male, **female**
   - First child: **yes**, no
   - Gestational length: 32-34 wks, 35-36, 37-38, **39-40**, 41-42, 43-44 wks
   - Start of prenatal care: **first trimester**, second trimester, third trimester, no care, missing

   Additionally, the paper reports the following information:

   i. IQR for $PM_{2.5}$: 2.2 $\mu g/m^3$
   ii. Odds ratio for low birth weight (< 2,500g) per IQR increase in pollution during the gestational period and 95% Confidence interval: 1.054 (1.022 to 1.087)

   In the following, use the value $\beta_0 = -2.8$ as an estimate of the intercept term.

   a) Write the generalized linear model used by the authors in their study. Clearly, specify the link function and the assumptions of the model. Identify the main parameter of interest in the study, justifying your choice.

   b) Using the information provided above, provide an estimate of the parameter of interest identified in (a) **and the corresponding 95% CI**.

   c) Interpret both the point estimate and the 95% CI obtained in (b) using a language understandable to non-statisticians.

   d) Estimate the probability of low birth weight for an infant girl whose mother is married, did not smoke nor drink alcohol during the pregnancy, has 12 years of education, is 30-34 years old, white, was

pregnant for 39-40 weeks, started prenatal care in the first trimester, is delivering her first child, and was exposed to an average PM2.5 concentration of 15 $\mu g/m^3$ during the course of the pregnancy.

e) The authors investigated also whether the effect of pollution on risk of low birth weight differs by race. Do you need to make any modifications to the model written in (a) to examine this question? Justify your answer, and in case a modification is needed, explicitly write the equation of the new model.

f) Using the equation of the model that you identified in (e), write down the expression of the risk of low birth weight for a black mother compared to that of a white mother. Assume that the two mothers have the same demographic/socio-economic and pregnancy characteristics and experienced the same level of $PM_{2.5}$ exposure during the course of their pregnancy.

g) Suppose that the authors decided to use the average $PM_{2.5}$ concentration during each trimester (i.e., three different covariates) instead of the average $PM_{2.5}$ concentration during the entire pregnancy (i.e. one single covariate) to characterize maternal exposure. How would the model in (a) and the interpretation of the parameters change? Would you have any concerns fitting this model?

(End of Problem 2)

3. A clinical investigation randomizes individuals to receive either active treatment (*TX*=1) or control (*TX*=0) after first recording a baseline measurement, Y_{i0}. Subsequent follow-up records an outcome at one follow-up time for each participant, Y_{i1}. The goal of the study is to assess whether there is an impact of treatment on Y.

Let *TX* denote the treatment assignment and let *t* denote time (*t*=0 for baseline and *t*=1 for follow-up). Assume that there are m subjects in each group with $m > 1$,and that the subjects are independent. Additionally, assume that $\sigma_0^2 = Var(Y_{i0})$, $\sigma_1^2 = Var(Y_{i1})$, and $Cov(Y_{i0}, Y_{i1}) = \rho\sigma_0 \cdot \sigma_1$.

a) Consider the model

$$E\left[Y_{ij}|\boldsymbol{X_{ij}}\right] = \beta_0 + \beta_1 \cdot t_{ij} + \gamma \cdot TX_i \cdot t_{ij} \qquad i = 1,2,\cdots,m; \quad j = 0,1$$

Provide the joint distribution of $\boldsymbol{Y}_i = \begin{pmatrix} Y_{i0} \\ Y_{i1} \end{pmatrix}$ for individuals $i = 1,2,\cdots,m$.

b) Express the joint distribution in 1. as the product of the marginal distribution of Y_{i0} times the conditional distribution of Y_{i1} given Y_{i0}.

c) Using the result obtained in (b)., write down the corresponding linear regression model and derive the MLE, $\hat{\gamma}$, for γ.

d) Calculate the variance of $\hat{\gamma}$ under the assumption that $\sigma_0^2 = \sigma_1^2$ and that σ_0, σ_1 and ρ are fixed.

(End of Problem 3)

4. Consider a longitudinal study for evaluating treatments of advanced AIDS. Individuals were randomized at baseline to two groups, `Group` $= 1$ for the treatment group, `Group` $= 0$ for the control group. The longitudinal outcome of interest is CD4 count (which measures how many CD4 cells, a type of white blood cell, is in a milliliter of blood), available at baseline and up to 9 post-randomization follow-up visits. The follow-up visits we consider all occured within 40 weeks after randomization. First 10 rows of the data set is shown in Figure 1.

First, consider the response variable to be the log transformed CD4 counts, `logcd4`, available on 1309 patients. In interpretation of the outcome, you can treat directly use the phrase "log CD4 count" as a continuous variable without further interpreting the log transformation.

(a) Consider a linear mixed effects model (LMM) with three fixed effects and two random effects: an intercept, a linear term for `Week`, a `Week-Group` interaction term, a random intercept, and a random slope for `Week`. (No main effect term for `Group`.)

(i) Write down the model and the distributional assumptions.

(ii) Define all the covariates you used in the model.

(b) Figure 2 shows the model fit result using `lmer` function in `R`. Provide an interpretation for each of the three fitted regression coefficient values.

(c) Consider the variance of the random effects.

(i) Write down the estimated value of the variances of random effects from Figure 2.

(ii) Provide an interpretation for each of the two estimated variances (or standard deviations). You don't need to simplify your answer.

(d) Consider a hypothesis test for whether the linear mixed effects model in (a) is better than a linear mixed effects model with the same fixed effects but only with a random intercept.

(i) Precisely state the null hypothesis and the alternative hypothesis in terms of parameters in the model.

(ii) What method would you use to conduct this hypothesis test? What are some considerations regarding choosing the significance level for this test?

Next, consider the binary outcome `cd4below50`, which equals 1 if the CD4 count is below 50 (a dangerously low value), and 0 otherwise.

(d) For this binary outcome, consider generalized estimating equations (GEE) with logit link, with a linear term for `Week`and a `Week-Group` interaction term. (No main effect term for `Group`.)

(i) Write down the marginal mean model.

(ii) Write down two options of working covariance matrix (including the name and the form of the matrix).

(iii) Are there any distributional assumptions made by the GEE?

(e) Figure 3 shows the model fit result using `geeglm` function in `R`. Provide an interpretation for each of the three fitted regression coefficient values.

(f) Now consider a generalized linear mixed model (GLMM) with logit link, with three fixed effects and one random effect: an intercept, a linear term for `Week`, a `Week-Group` interaction term, and a random intercept. (No main effect term for `Group`.) Write down the model and the distributional assumptions.

(g) Figure 4 shows the model fit result using `glmer` function in `R`. Provide an interpretation for the estimated coefficient for `Week`. Is this interpretation the same as / different from the corresponding term in Question (e)? Briefly explain.

APPENDIX - Problem 4

id	group	week	logcd4	cd4below50
1	0	0.0000	3.135494	1
1	0	7.5714	3.044523	1
1	0	15.5714	2.772589	1
1	0	23.5714	2.833213	1
1	0	32.5714	3.218876	1
1	0	40.0000	3.044523	1
2	1	0.0000	3.068053	1
2	1	8.0000	3.891820	1
2	1	16.0000	3.970292	0
2	1	23.0000	3.610918	1

Figure 1: First 10 rows of the CD4 data set

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: logcd4 ~ week * group - group + (1 + week | id)
   Data: long_data

REML criterion at convergence: 12078

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.240 -0.432  0.026  0.480  3.586

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.649757 0.8061
          week        0.000252 0.0159   0.20
 Residual             0.338332 0.5817
Number of obs: 5036, groups:  id, 1309

Fixed effects:
               Estimate  Std. Error          df t value Pr(>|t|)
(Intercept)    2.997493    0.025935 1299.956968  115.58  < 2e-16 ***
week          -0.013868    0.000999  983.732233  -13.88  < 2e-16 ***
week:group     0.013090    0.001923  942.733025    6.81  1.8e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) week
week       -0.194
week:group  0.003 -0.501
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0414198 (tol = 0.002, component 1)
```

Figure 2: LMM fit

```
Call:
geeglm(formula = cd4below50 ~ week * group - group, family = binomial,
    data = long_data, id = id, corstr = "exchangeable")

 Coefficients:
             Estimate   Std.err    Wald Pr(>|W|)
(Intercept)  1.755983  0.066706 692.968  < 2e-16 ***
week         0.010107  0.003594   7.907  0.00492 **
week:group  -0.024550  0.005677  18.704 1.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    1.026 0.09189
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.4488 0.05088
Number of clusters:   1309  Maximum cluster size: 9
```

Figure 3: GEE fit

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-
Hermite Quadrature, nAGQ = 50) ['glmerMod']
 Family: binomial  ( logit )
Formula: cd4below50 ~ week * group - group + (1 | id)
   Data: long_data

     AIC      BIC   logLik deviance df.resid
    3441     3467    -1716     3433     5032

Scaled residuals:
   Min     1Q Median     3Q    Max
-3.269  0.118  0.138  0.165  1.896

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 6.56     2.56
Number of obs: 5036, groups:  id, 1309

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.29063    0.16355   20.12  < 2e-16 ***
week         0.01683    0.00511    3.29    0.001 ***
week:group  -0.04281    0.00789   -5.42  5.9e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) week
week       -0.262
week:group -0.040 -0.527
```

Figure 4: GLMM fit

(End of Problem 4)