# Written Comprehensive Examination

# Methods 210, 210B, 210C

# Department of Statistics, UC Irvine Monday, June 22, 2020, 9:00 am to 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. 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. Salary inequities by gender continue to be a problem in academia, government, and industry. In this problem we will consider an analysis of monthly salaries for faculty from a single R1 university in the US during a single year. In total, monthly salary data, denoted $Y$. was obtained on $n = 1,597$ faculty members. The primary goal of the analysis is to determine whether or not evidence for gender discrimination exists with respect to pay. Along with the monthly salary values, additional covariates will be introduced into the question throughout.

   (a) We will start by considering a regression model of the following form:

   $$Y_i = \beta_0 + \beta_1 I_{\texttt{male}_i\texttt{=1}} + \epsilon_i, \qquad i = 1, \ldots, 1597.$$

   In matrix notation, this can be written as

   $$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon},$$

   with $\vec{Y}$ denoting the $n \times 1$ vector of monthly salaries, $\mathbf{X}_{n\times 2} = (\vec{1} \quad \vec{I}_{\texttt{male}_i\texttt{=1}})$, $\vec{\beta} = (\beta_0, \beta_1)$ a $2 \times 1$ column vector, and $\vec{\epsilon}$ an $n \times 1$ column vector of model residuals. Using the matrix formulation of the problem, derive the ordinary least squares estimator of $\vec{\beta}$, which we will denote by $\widehat{\vec{\beta}}$. Leaving your solution for $\widehat{\vec{\beta}}$ in matrix notation is perfectly fine.

   (b) The classical linear regression model assumes $\vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 \mathbf{I}_{n\times n})$. Under this assumption, derive the variance of $\widehat{\vec{\beta}}$.

   (c) Below is R output from fitting the model in (a) via OLS to the available salary data. Provide a precise interpretation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (or some suitable transformation of these parameters) in words that can be understood by a statistical layman.

```
##
##### OLS fit
##
> fit1 <- lm( salary ~ i.male, data=salary )
> summary(fit1)

Call:
lm(formula = salary ~ i.male, data = salary)

Residuals:
   Min     1Q Median     3Q    Max
 -3601  -1406   -419   1091   7732

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5396.9       96.5    55.9   <2e-16 ***
i.male       1334.7      111.9    11.9   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1950 on 1595 degrees of freedom
Multiple R-squared:  0.0819,Adjusted R-squared:  0.0813
F-statistic:  142 on 1 and 1595 DF,  p-value: <2e-16
```

   (d) Using the output from (c), provide a 95% confidence interval for $\beta_1$.

(e) Based upon these output, a 95% Wald-based confidence interval for the mean monthly salary among male faculty members is (6620.5, 6842.7). Use this and the model output in (c) to obtain an estimate of the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$. (You may leave your expression unevaluated, but your estimate should be a function of the output given.)

(f) Suppose that in truth $Y_i \sim Exp(\mu_i)$ with $\mu_i = \beta_0 + \beta_1 I_{\texttt{male}_i\texttt{=1}}$ but we still compute the OLS estimator for the model

$$\mathrm{E}[\vec{Y}] = \vec{\mu} = \mathbf{X}\vec{\beta},$$

where $\vec{Y}$, $\mathbf{X}$ and $\vec{\beta}$ are as defined in (a). Again using matrix notation, derive the mean and variance of the OLS estimator in this case. From these results, state which estimates from the model fit in (c) can be "trusted" and which cannot.

(g) Again consider the setting in (f). Based upon the model output in (c), state which of the following are true (in large samples like we have) and which are false. Briefly provide the reason for your response in each case.

(1) A test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ will be valid in that it will yield approximately the nominal type I error rate.

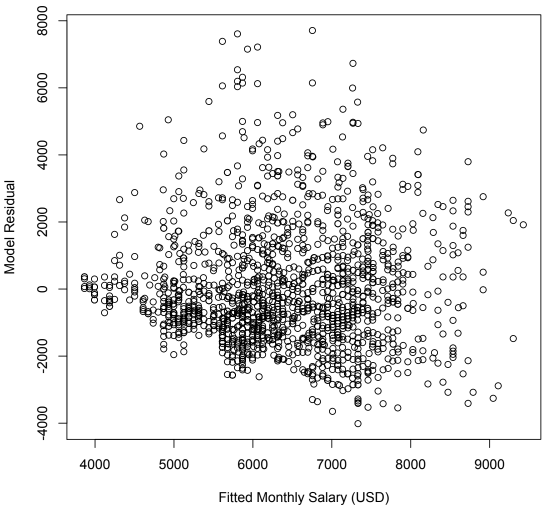
(2) A 95% confidence interval for $\beta_1$ will be valid in that it will yield approximately correct coverage probability.

(h) A colleague noted that the model in (a) is not likely to provide a fair comparison of the mean salaries between males and females with the goal of assessing whether the university is guilty of systematically paying males more the females. She suggests that, at minimum, your model should include the years since the faculty member was hired at the university and the primary school for their appointment (categorized as 1=Humanities/Social Sciences, 2=Engineering/CS, 3=Physical Sciences/Biological Sciences, 4=Business/Law/Medicine). Briefly explain why your colleague makes a valid point.

(i) Suppose you modify the model in (a) based upon the above suggestions and again use OLS to fit the following:

$$Y_i = \gamma_0 + \gamma_1 I_{\texttt{male}_i\texttt{=1}} + \gamma_2 \texttt{yr.since.hired} + \gamma_3 I_{\texttt{school}_i\texttt{=2}} + \gamma_4 I_{\texttt{school}_i\texttt{=3}} + \gamma_5 I_{\texttt{school}_i\texttt{=4}} + \epsilon_i.$$

Below is a plot of the residuals from the OLS fit of this model vs. the fitted salary values from the model. Based upon this, does it appear that the assumptions of classical linear regression are valid for these data? Explain. If not, provide two ways the model can be *changed* to address any of the violations you believe are reflected in the plot.



(j) After performing residuals diagnostics you find that the distribution of log-transformed salary is symmetric and modify the model in part (i) as follows:

$$\ln(Y_i) = \delta_0 + \delta_1 I_{\texttt{male}_i\texttt{=1}} + \delta_2 \log_2(\texttt{yr.since.hired}) + \delta_3 I_{\texttt{school}_i\texttt{=2}} + \delta_4 I_{\texttt{school}_i\texttt{=3}} + \delta_5 I_{\texttt{school}_i\texttt{=4}} + \epsilon_i.$$

(1) Provide a precise interpretation of $\delta_1$ (or a suitable transformation of this parameter) in words that can be understood by non-technical individuals. (Note that your interpretation should likely not involve $\log_e$-salary in order to be understandable!)

(2) Explain how the interpretation of $\delta_1$ and the interpretation of $\gamma_1$ (from part (i)) differ and what the relevance of this difference is in terms of your question of interest.

(3) Provide a precise interpretation of $\delta_2$ (or a suitable transformation of this parameter) in words that can be understood by non-technical individuals. (Note that your interpretation should likely not involve $\log_e$-salary or $\log_2$-years since hired in or to be understandable!)

END OF QUESTION (1)

2. A microbiology lab is studying effects of two treatments of bacterial skin infection. We will refer to them as treatments A and B.

   (a) First, lab researchers designed and performed the following experiment. They cultured bacterial strains of interest in petri dishes until each culture reached the mid-exponential growth phase, with 30 of these cultures receiving no treatment (control group), 30 receiving treatment A, and 30 receiving treatment B. Assignment to treatments and control groups was done uniformly at random. After an over-night incubation period, researchers recorded the number of colony forming units (CFUs) measured in millions per mL in each petri dish. They assume that all petri dishes had the same number of CFUs at the mid-exponential growth phase. If a treatment works, the researchers expect to see lower number of CFUs after the over-night incubation in the treated petri dishes than in the untreated/control ones. Taking a regression view of this problem, let $Y_i$ be post-treatment or control number of CFUs for petri dish $i$, $a_i$ be a binary indicator of treatment A (1 = treatment A applied, 0 = otherwise), $b_i$ be a binary indicator of treatment B (1 = treatment B applied, 0 = otherwise), $c_i$ be a binary indicator of control group (1 = no treatment applied, 0 = otherwise).

      i. We assume the following linear model:

$$Y_i = \beta_0 + \beta_1 a_i + \beta_2 b_i + \epsilon_i, i = 1, \dots, 90,$$

      where $\epsilon_i \sim N(0, \sigma^2)$. We assume that normality assumption holds here. Using ordinary least squares estimation (OLS), we obtain $\hat{\beta}_0$ = 2.2, $\hat{\beta}_1$ = -0.21, $\hat{\beta}_2$ = -0.13. Provide interpretation of these coefficients.

      ii. For the above regression, $\widehat{MSE} = 0.07$ and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.03 & -0.03 & -0.03 \\ -0.03 & 0.07 & 0.03 \\ -0.03 & 0.03 & 0.07 \end{pmatrix}.$$

      Compute 95% confidence intervals for $\beta_1$ and $\beta_2$, using the fact that $\Pr(Z > 1.99) \approx 0.975$, where $Z \sim t_{87}$. State your conclusions about anti-bacterial effects of treatments A and B.

      iii. Compute 95% confidence interval for $\beta_1 - \beta_2$ and interpret it.

      iv. Provide a different linear model formulation that would allow you to compute 95% confidence interval for $\beta_1 - \beta_2$ and to test $H_0$: $\beta_1 - \beta_2 = 0$ directly from the OLS output, without the need to know $(\mathbf{X}^T\mathbf{X})^{-1}$.

   (b) Since treatments A and B use different biological mechanisms to inhibit bacterial growth, the researchers became curious about the effect of applying both treatments simultaneously, hoping to find treatment synergy — treatments having a larger effect when applied simultaneously than the sum of their individual effects. They repeated the above experiment, but added a forth group of petri dishes that recieved both treatments A and B.

      i. Formulate a linear model that will allow the researchers to estimate an effect of synergy or competition (opposite of synergy). Provide interpretation for all model coefficients.

ii. Explain what test you would apply to check for existence of a synergistic effect between the two treatments.

(c) When the researchers looked at the data more carefully, they noticed that they were not able to "catch" all petri dishes exactly at the mid-exponential growth phase, so the starting numbers of CFUs for all petri dishes were not identical. Let's denote this starting number of CFUs for petri dish $i$ by $x_i$.

i. Formulate an analysis of covariance (ANCOVA) model that would allow the researchers to estimate effects of treatments A and B (without synergy considerations), while accounting for different initial conditions (starting numbers of CFUs). What parameter estimates, confidence intervals, and hypothesis tests would you report to the researchers?

ii. Compare and contrast assumptions of the ANCOVA model above and an ANOVA model with response being $Y_i - x_i$.

iii. Describe how you would perform model diagnostics for your ANCOVA model.

iv. Suppose your diagnostic analysis reveals problems with the ANCOVA model. You come with these results to the microbiologists and they tell you that the root of the problem could be in the anti-bacterial treatment mechanisms, which are not fully understood. The treatments can either kill some unknown number of bacteria or diminish the *rate* at which bacteria grow. Extend your ANCOVA model to allow for both anti-bacterial mechanisms and explain how you would use this new model to help the researchers decide which mechanism each treatment uses.

v. Comment on what could go wrong with the ANCOVA extension. How would you diagnose these problems?

END OF QUESTION (2)

3. Consider the logistic regression model for fitting the binary response of heart disease ($Y = 1$ for having heart disease) with three covariates $X_1$, $X_2$ and $X_3$, where $X_1 = (0, 1, 2, 3)$ indicates snoring level from "never" to "heavy," $X_2$ is gender (1 for male and 0 for female), and $X_3$ is a treatment indicator with 1 as treatment and 0 as placebo. Let $\pi = P(Y = 1)$. The fitted logistic model is

$$\text{logit}(\hat{\pi}) = -3.5 + 0.3X_1 + 0.1X_2 - 1.2X_3. \tag{1}$$

(a) Estimate the probabilities of heart disease at snoring level 0 for the placebo male group, and snoring level 3 for the treatment female group.

(b) Compute the estimated odds ratio for the treatment $X_3$, and interpret it.

(c) Compute the estimated odds ratio for the gender $X_2$, and interpret it.

(d) Suppose $-2\log(L)$ ($L$ is the likelihood function) is 24 for the logistic model fitting $X_1$ and $X_2$ only, and is 18 for the logistic model fitting $X_1$, $X_2$ and $X_3$. Is the treatment effect statistically significant?

(e) Suppose the deviance is 8 for the logistic model in (1). What is your conclusion on the goodness-of-fit of the model, based on the deviance test?

(f) Is the deviance test always appropriate to check the goodness-of-fit of a model? Provide some counter-examples if you do not think so, and alternative strategies to check the goodness-of-fit of a model.

END OF QUESTION (3)

4. In this problem we will consider growth data on $N = 200$ children randomly sampled from clinical visits in Nepal. Specifically, we will consider *growth trajectories* of the children defined as changes in height by age. By design, each child had their height (cm) measured at up to five different time points. At each visit, the age of the child (months), weight, and arm circumference were also measured. Additional available data included the sex of the child, an indicator of whether the child was currently being breast fed at the time of the visit, the mother's age (yrs) at the time of delivery of the child (`mage`), and the number of children the mother previously had that had died and that had lived. A snapshot of the first 6 lines of the dataset are available in the Appendix for your information. Also included is a scatterplot of recorded height by age for all children, and the observed growth trajectories for 25 randomly sampled children. For this problem we will only focus on a few of the available covariates in the dataset.

   (a) We will begin by considering a simple mean model regressing height on age in a linear fashion. Thus we consider a mean model of the form

   $$\mathrm{E}[HT_{ij}] = \beta_0 + \beta_1 AGE_{ij}, \qquad i = 1, \ldots, 200, \qquad j = 1, ..., n_i, \tag{1}$$

   where $n_i$ denotes the total number of observations made on subject $i$. `fit1` in the Appendix provides estimates of $\vec{\beta} = (\beta_0, \beta_1)$ based upon a GEE model utilizing an independence working correlation structure. Based upon this model fit, provide a precise interpretation of the estimated coefficient corresponding to `age` in terms that could be understood by a non-technical audience.

   (b) Based upon the the estimates given for `fit1`, provide an asymptotically valid 95% confidence interval (CI) for the first order association between age and height. You may leave your expression unevaluated, but your CI should be asymptotically valid in the sense that it will yield correct coverage probability as $N \to \infty$.

   (c) From Figures (1) and (2) in the Appendix, it is apparent that the relationship between height and age is curvilinear. As such, we consider modeling age as a quadratic term using a mean model of the form

   $$\mathrm{E}[HT_{ij}] = \gamma_0 + \gamma_1 AGE_{ij} + \gamma_2 AGE_{ij}^2, \qquad i = 1, \ldots, 200, \qquad j = 1, ..., n_i. \tag{2}$$

   `fit2` in the Appendix provides estimates of $\vec{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ based upon a GEE model utilizing an independence working correlation structure. Also provided are the model-based (naive) and robust (emprical) covariance estimates of $\widehat{\vec{\gamma}}$. From this model, we wish to test for an association between height and age. State the null and alternative hypothesis corresponding to this scientific question in terms of the model parameters in Eq. (2) and write down the test statistic and approximate distribution of the statistic that you would use to conduct an asymptotic level $\alpha$ test of null hypothesis that you wrote down. You may write your test statistic symbolically (e.g. in terms of $\widehat{\vec{\gamma}}$, $\widehat{\mathrm{Var}}[\widehat{\vec{\gamma}}]$, etc) but you should define these symbols in terms of the model output for `fit2`.

   (d) Using the model output for `fit2`, construct an asymptotically valid 95% CI for the mean height of children that are 43 months of age. You may leave your expression unevaluated, but your CI should be asymptotically valid in the sense that it will yield correct coverage probability as $N \to \infty$.

   (e) Another scientific question of interest is to determine if growth trajectories vary by the age of the mother at the time of delivery (recorded as `mage` in the dataset).

      i. Modify the mean model in Eq. (2) to allow for quadratic growth trajectories to vary by the age of the mother at the time of delivery.

      ii. State precisely what the null and alternative hypothesis would be for testing whether quadratic growth trajectories vary by the age of the mother at the time of delivery. Write your hypotheses in terms of the model parameters given in your answer to (i).

(f) To this point, we have assumed an independence working correlation structure. Explain what plots or residual diagnostics you would use to better assess the correlation structure of heights repeatedly measured on the same subject over time.

(g) An alternative approach to the GEE model is a linear mixed effects model. Write down a complete LME model specification for the model you proposed in Part e.i. that allows for each individual to have their own specific growth trajectory. Your response should include specification of all random effects and error terms along with their assumed distributions.

END OF QUESTION (4)

## APPENDIX - QUESTION 4

```
> head(nepal)
      id sex   wt    ht  arm bf mage lit died alive age
1 120011   1 12.8  91.2 14.3  0   35   0    2     5  41
2 120011   1 12.8  93.9 13.5  0   35   0    2     5  45
3 120011   1 13.1  95.2 14.5  0   35   0    2     5  49
4 120011   1 13.8  96.9 14.1  0   35   0    2     5  53
5 120011   1   NA    NA   NA  0   35   0    2     5  57
6 120012   2 14.9 103.9 13.9  0   35   0    2     5  57
```

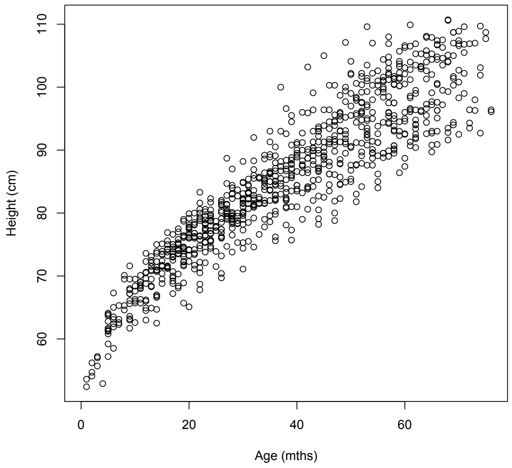


Figure 1: Scatterplot of height (cm) vs. age (mths).

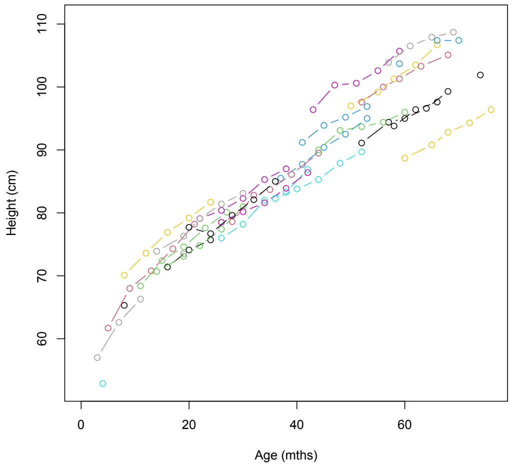


Figure 2: Observed growth trajectories for 25 randomly sampled children.

```
##
##### GEE fit regressing height on age (linear)
##
> fit1 <- gee(ht ~ age, data=nepal)
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
(Intercept)         age 
 62.4726283   0.5974161 
> summary(fit1)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998) 

Model:
 Link:                      Identity 
 Variance to Mean Relation: Gaussian 
 Correlation Structure:     Independent 

Call:
gee(formula = ht ~ age, data = nepal)

Summary of Residuals:
       Min         1Q     Median         3Q        Max 
-13.981424  -2.957127   0.377237   3.079049  15.643647 


Coefficients:
             Estimate  Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) 62.4726283 0.364674980 171.31043  0.60273202 103.64910
age          0.5974161 0.008633704  69.19581  0.01833706  32.57972

Estimated Scale Parameter:  22.26669
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    1    0    0    0
[3,]    0    0    1    0    0
[4,]    0    0    0    1    0
[5,]    0    0    0    0    1



##
##### GEE fit regressing height on age and age^2 (quadratic)
##
> fit2 <- gee(ht ~ age + I(age^2), data=nepal)
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
 (Intercept)          age     I(age^2) 
57.974929088  0.900743379 -0.003938505 
> summary(fit2)
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = ht ~ age + I(age^2), data = nepal)

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-12.8097068  -2.7658334   0.1869327   2.9463392  14.9489315


Coefficients:
                Estimate  Naive S.E.   Naive z  Robust S.E.  Robust z
(Intercept) 57.974929088 0.632687695 91.632775 0.8421500690 68.841565
age          0.900743379 0.036477232 24.693304 0.0579340769 15.547730
I(age^2)    -0.003938505 0.000461211 -8.539487 0.0008029158 -4.905253

Estimated Scale Parameter:  20.57544
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    1    0    0    0
[3,]    0    0    1    0    0
[4,]    0    0    0    1    0
[5,]    0    0    0    0    1


##
##### Naive covariance estimate of regression parameters from fit2
##
> fit2$naive.variance
              (Intercept)           age      I(age^2)
(Intercept)  0.4002937199 -2.132541e-02  2.429172e-04
age         -0.0213254060  1.330588e-03 -1.638247e-05
I(age^2)     0.0002429172 -1.638247e-05  2.127156e-07

##
##### Robust covariance estimate of regression parameters from fit2
##
> fit2$robust.variance
              (Intercept)           age      I(age^2)
(Intercept)  0.7092167388 -4.478475e-02  5.513697e-04
age         -0.0447847515  3.356357e-03 -4.476241e-05
I(age^2)     0.0005513697 -4.476241e-05  6.446738e-07
```