The University of Minnesota
School of Statistics

Ph.D. Written Examination
Applied Exam, In-Class Part

Wednesday, August 21, 2019
8:30 AM to 5:00 PM
Bruininks Hall, Room 131B

This part of the exam consists of one data analysis problem, to be solved in-class. Please submit your solution electronically, as .pdf file, to Taryn Verley (`tverley@umn.edu`) by 5:00 PM.

The applied exam is open book, open notes. You can use the internet. However, all work needs to be your own. Do not consult with any person, except the faculty member proctoring the exam, or Professor Birgit Grund (`birgit@umn.edu`), or Professor Yuhong Yang (`yangx374@umn.edu`). Academic misconduct will result in failing this examination.

Author: Grund
Graders: Grund, Chatterjee

Problem 1

A new drug for a chronic disease has been developed. In preliminary tests, it appeared that the drug is increasing cholesterol, an undesirable side effect. Statins are a drug class that is decreasing cholesterol. In order to investigate whether combining the new drug with a statin is safe and effective, a randomized study was performed as factorial design, with 2 factors:

Drug: Treatment for the chronic disease (3 levels, 1=standard treatment, 2=new drug at 400 mg/day, 3=new drug at 600 mg/day), and
Statin: Use of a statin or placebo (2 levels, 1=statin, 0=placebo).

A total of 600 participants were randomized, 100 to each of the six Drug×Statin combinations.

For this exam problem, we consider the effect of the treatments on low-density lipoprotein (LDL) cholesterol; lower LDL is better. LDL was measured at baseline, month 12, 24, 36, and 48. In order to characterize the study population, several other variables were measured at baseline, such as age, gender, race, and smoking status; all of these covariates are known to be correlated with LDL levels in the general population. The dataset is described at the end of the problem.

(a) Describe the effect of the treatments on LDL cholesterol. Provide estimates with 95% CIs.

This is an open-ended question. To help guide you, below are a few items that you should include in your report.

- Should the outcome of interest be LDL, or change in LDL from baseline, or percent change in LDL (change from baseline as percent of the baseline value)? Or some transformation of LDL? Discuss these options and justify your choice.
- Does the new drug cause higher LDL increases than the standard treatment? Does the dose matter? What is the effect of the statins?
- Justify your model choice.

(b) Does the effect of statins differ between men and women? By baseline level of LDL? If yes, interpret.

(c) Assume that you have done the analyses in part (b) for 6 different variables (including gender, age, race, ...), and obtained the following p-values: $p_1 = 0.002, p_2 = 0.25, p_3 = 0.04, p_4 = 0.01, p_5 = 0.08, p_6 = 0.60$, for variables 1-6. Considering each variable by itself, the treatment effect differs by variables 1, 3, and 4, with a significance level of 5%. In order to publish the results in a reputable journal, you need to

control for multiple comparisons. Use a method that limits the overall error rate to <5%, but is less conservative than Bonferroni. Which variables remain significant after the adjustment for multiple comparisons? What error rate does your method control?

(d) Clinicians are want to prevent levels of LDL above 150 mg/dL.

- Does the use of statins decrease the prevalence of LDL>150 mg/dL? Provide a P-value.
- Provide a point estimate and a 95% CI for the effect of statin use on the prevalence of LDL>150 mg/dL. Interpret this point estimate for your non-statistician colleagues.

When writing your answers, consider yourself the lead statistician working with a team of medical professionals. Present your main results and justifications in a way so that your non-statistician colleagues can understand. Summarize your results in at most 5 pages, plus at most 5 pages of supporting tables and figures.

The dataset is available on `http://www.stat.umn.edu/~birgit/data/cholesterol.csv`. It contains the following variables:

Name	Values	Description
id	integer	Participant identifier
visit	0, 12, 24, 36, 48	Time of measurement (in months), 0=baseline
treat	1-6	Treatment, 1=standard treatment, no statin, 2=new drug 400 mg, no statin, 3=new drug 600 mg, no statin, 4=standard treatment + statin, 5=new drug 400 mg + statin, 6 =new drug 600 mg + statin
drug	1-3	1=standard treatment, 2=new drug 400 mg, 3=new drug 600 mg
statin	0, 1	1 = statin use, 0= placebo
ldl	continuous	LDL (in mg/dL)
ldl_0	continuous	LDL at baseline (visit=0)
age	continuous	Age (in years)
gender	1 or 2	1=male, 2=female
female	0, 1	1=female, 0=male
race	1-4	Race/Ethnicity indicator
smoking	0, 1	Indicator for smoking
disease_yrs	continuous	Duration of chronic disease (years)

Disclaimer: The data for this exam problem were computer-generated.

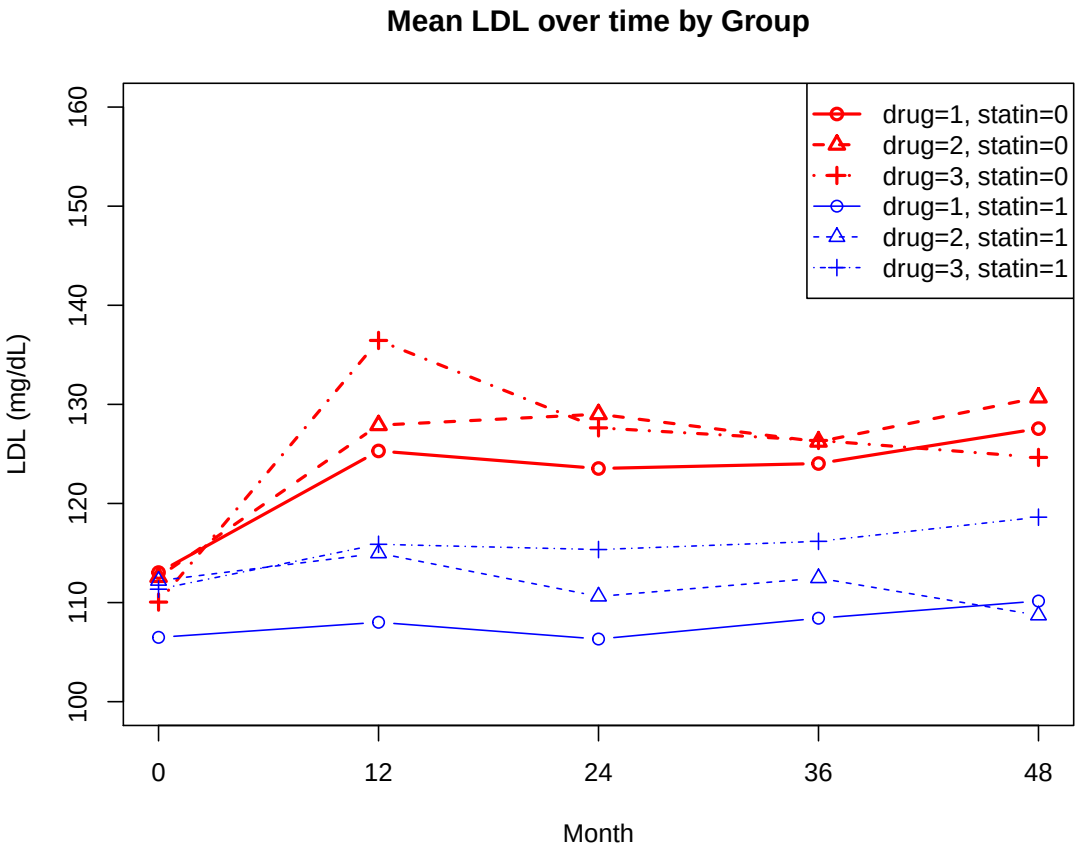

Figure 1: Mean LDL levels over time.

Solution:

(a) First, we read in the data, and define factors.

```
d = read.csv("cholesterol.csv")

d = within(d, {race = as.factor(race)
treat = as.factor(treat)
drug = as.factor(drug)
ldl_chg = ldl - ldl_0
hdl_chg = hdl - hdl_0
ldl_pchg = ldl_chg/ldl_0*100
})
```

We fit a longitudinal mixed model, with fixed effects for the factorial treatment structure, and with a random intercept per participant to model within-subject

correlation. We add covariates as fixed effects for variance reduction. After model selection, the treatment effects are estimated using contrasts.

We select as outcome "change from baseline", and then check whether "percent change from baseline" would be better based on residual plots. The figure below shows the mean change in LDL from baseline:

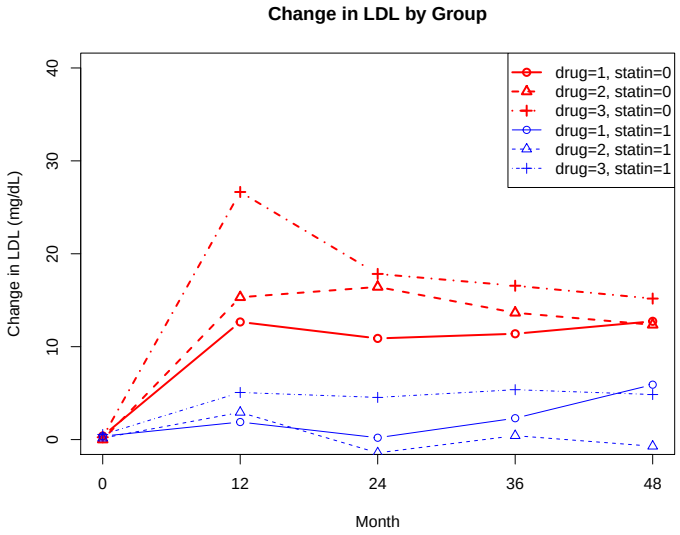

Figure 2: Mean change in LDL levels over time.

We start with a full model for the treatment factors, including covariates as additive.

```
> library(nlme)
> d.fu = d[d$visit > 0,]
> m1 = lme(ldl_chg ~ drug*statin + factor(visit) + ldl_0 + female + age +
                     race + smoking + disease_yrs, random=~1|id, data=d.fu)
> car::Anova(m1)
Analysis of Deviance Table (Type II tests)

Response: ldl_chg
Chisq Df Pr(>Chisq)
drug            28.97  2    5.1e-07 ***
statin         252.74  1    < 2e-16 ***
factor(visit)    8.37  3    0.03890 *
ldl_0           65.66  1    5.4e-16 ***
female          13.46  1    0.00024 ***
age              0.47  1    0.49258
race             2.17  3    0.53717
```

```
smoking         0.00  1    0.97518
disease_yrs     0.07  1    0.79659
drug:statin     5.67  2    0.05881 .
```

The residual plot shows no need for transformation. When fitting the same model with the "percent change in LDL" outcome, the residual plot shows problems with the model fit. We choose "change in LDL" as outcome.

A quick approach to model selection would be to select all covariates with P-values <0.05, compare the smaller model to the full model, and then try to prune some more. Note that the model needs to be re-fitted with `method="ML"` when using LRT for model comparisons.

```
> ## Variable selection
> m2 = update(m1, method="ML")
> m3 = update(m2, fixed= .~ drug*statin + factor(visit) + ldl_0 + female)
> m4 = update(m3, fixed= .~ drug + statin + factor(visit) + ldl_0 + female)
> anova(m2,m3,m4)
Model df   AIC   BIC logLik   Test L.Ratio p-value
m2     1 19 18976 19084  -9469
m3     2 13 18967 19041  -9470 1 vs 2    2.73  0.8420
m4     3 11 18969 19032  -9473 2 vs 3    6.08  0.0479
```

```
> m3a = update(m3, method="REML")
> car::Anova(m3a)
Analysis of Deviance Table (Type II tests)

Response: ldl_chg
Chisq Df Pr(>Chisq)
drug           31.10  2    1.8e-07 ***
statin        254.85  1    < 2e-16 ***
factor(visit)   8.38  3      0.039 *
ldl_0          64.99  1    7.5e-16 ***
female         18.58  1    1.6e-05 ***
drug:statin     6.02  2      0.049 *
```

The final model is **m3** or **m4**. Model **m3** contains an interaction effect between `drug` and `statin`, with p-value just below 0.05; this model is selected by the LRT and AIC. The BIC criterion selects `m4`, without interaction effect.

If we choose the model without interaction effect, we estimate the treatment effects between the standard and new drug using contrasts in the `drug` main effect. There is

no evidence for a difference between the standard treatment vs the new drug at 400 mg (p=0.48), while there is a significant difference between the standard treatment vs the new drug at 600 mg, with the new drug resulting in 5.2 mg/dL higher LDL averaged over the 4 years of follow-up (95% CI (3.2,7.1), p<0.0001). The new drug at 600 mg also results in significantly higher LDL than the old drug; estimates are given below. Figure 2 above shows that the treatment difference is driven by the high LDL in year 1 for the new drug at the 600 mg dose.

```
> m4a = update(m4, method="REML")
> library(gmodels)
> fit.contrast(m4a, "drug", coeff=c(-1,1,0), conf.int=0.95)
                 Estimate Std. Error t-value Pr(>|t|) lower CI upper CI
drug c=( -1 1 0 )    0.707          1   0.704    0.482    -1.26     2.68

> fit.contrast(m4a, "drug", coeff=c(-1,0,1), conf.int=0.95)
                  Estimate Std. Error t-value Pr(>|t|) lower CI upper CI
drug c=( -1 0 1 )     5.16          1    5.14 3.74e-07     3.19     7.13

> fit.contrast(m4a, "drug", coeff=c(0,-1,1), conf.int=0.95)
                  Estimate Std. Error t-value Pr(>|t|) lower CI upper CI
drug c=( 0 -1 1 )     4.46       1.01    4.41 1.23e-05     2.47     6.44

> fit.contrast(m4a, "drug", coeff=c(-1,0.5,0.5), conf.int=0.95)
                   Estimate Std. Error t-value Pr(>|t|) lower CI upper CI
drug c=( -1 0.5 0.5 )  2.93      0.868    3.38 0.000769     1.23     4.64
```

Since the interaction effect is borderline significant, it would be desirable to also compare the drugs separately among treatments 1-3 (with statin) and 4-6 (without statin). The results are roughly similar, not shown here.

```
> m5 = update(m4a, .~ treat + factor(visit) + ldl_0 + female)

> fit.contrast(m5, "treat", coeff=c(-1,1,0,0,0,0), conf.int=0.95)
                          Estimate Std. Error t-value Pr(>|t|) lower CI upper CI
treat c=( -1 1 0 0 0 0 )     2.75       1.42    1.94   0.0533  -0.0388     5.53

...
```

We estimate the main effect of statins immediately from the linear model, as this factor has only 2 levels. Again, we could estimate the statin effect separately for each of the 3 drugs using contrasts. Statins significantly decrease LDL (p<0.0001).

```
> intervals(m4a)
Approximate 95% confidence intervals

Fixed effects:
                 lower    est.    upper
statin          -14.684 -13.070 -11.4559
```

For full credit, at least the 95% CIs for comparing the three drugs should be adjusted for multiple comparisons, for example, by using Tukey's HSD method. The Tukey critical values depend on the number of means and the df for the "denominator"; the latter can be obtained via `anova(model)`. Here is an example of how to compute the CI halfwidth for a drug comparison with Tukey's HSD method:

```
#### Get df for denominator;
> anova(m4a)
numDF denDF F-value p-value
(Intercept)       1  1589     472  <.0001
drug              2   594      16  <.0001
....

### Tukey HSD critical value
> q=qtukey(p=0.95, nmeans=3, df=594)/sqrt(2); q
[1] 2.35

> ### Halfwidth
> Std.Error=1.0
> hw = q*Std.Error; hw
[1] 2.35
```

A quick-and-dirty alternative would be to use `linear.function{cfcdae}` with Bonferroni adjustment.

Comments:

- The **ldl_chg** values at **visit=0** need to be excluded from the dataset, since the baseline values do not contribute to the treatment difference.
- It is sensible here to include **visit** for variance reduction; it should be included as factor.
- Since this is a randomized trial, one could argue that it is unnecessary to adjust for covariates; however, covariates here provide variance reduction and thus more precise estimates if there is low model error.
- Use REML for the final 95% CI estimates, use ML for the models compared with LRT if the models differ in fixed effects.

(b) Test for the interaction effects `statin:female` and `statin:ldl_0`. The treatment effect does not differ between men and women (data not shown). However, for higher LDL values, the decrease in LDL due to statins is steeper, by -0.1 mg/dL per 1 mg/dL higher baseline LDL.

```
> m7 = update(m4a, fixed = .~drug*statin + ldl_0 + statin:ldl_0 +
                            female + factor(visit))
> summary(m7)

Fixed effects:
Value Std.Error   DF t-value p-value
        ...
        statin:ldl_0    -0.10     0.025  591   -3.91  0.0001
        ....
```

(c) Using the Benjamini-Hochberg method, Var 3 is not significant, but Var 4 and Var 1 remain significant. This method controls the FDR. Alternative methods are Holmes (strong familywise error) or Bonferroni, but these will detect fewer signals.

(d) The outcome is binary; therefore, GEE or GLMM models should be used. Despite randomization, the LDL levels at baseline differ between treatment groups, and possibly also the prevalence of LDL>150. In a GEE model, the treatment effect of statins is estimated by the interaction effect `statin:fu`, where `fu` is the indicator variable for `visit>0`.

Below we show the syntax for fitting GEE models; the parsimoneous model was selected by excluding terms with $p > 0.1$, and comparing the models using a LRT.

```
> d = within(d, {ldl_high = ifelse(ldl > 150, 1, 0)
                 fu = ifelse(visit==0, 0, 1)})

library(geepack)
> m1.gee = geepack::geeglm(formula=ldl_high ~ drug*statin*fu + age
           + female + race + smoking +disease_yrs,
           corstr = "exchangeable",
           family=binomial, data=d, id=id)
> m2.gee = geepack::geeglm(formula=ldl_high ~ drug*statin + drug*fu +
            statin*fu + age + race, family=binomial,
          corstr = "exchangeable", data=d, id=id)
> anova(m1.gee, m2.gee)
Analysis of 'Wald statistic' Table

Model 1 ldl_high ~ drug * statin * fu + age + female + race + ...
Model 2 ldl_high ~ drug * statin + drug * fu + statin * fu + ...
```

```
Df   X2 P(>|Chi|)
1  5 3.29      0.66
> summary(m2.gee)

Call:
geepack::geeglm(formula = ldl_high ~ drug * statin + drug * fu +
statin * fu + age + race, family = binomial, data = d, id = id,
corstr = "exchangeable")

Coefficients:
             Estimate  Std.err   Wald Pr(>|W|)
(Intercept)  -4.12665  0.33977 147.51  < 2e-16 ***
drug2         0.11010  0.34283   0.10  0.74809
drug3        -0.25545  0.34437   0.55  0.45821
statin       -0.41758  0.30254   1.91  0.16751
fu            0.94313  0.27870  11.45  0.00071 ***
age           0.05009  0.00507  97.74  < 2e-16 ***
race2         0.18529  0.18604   0.99  0.31924
race3         1.42718  0.18979  56.55  5.5e-14 ***
race4         0.33222  0.12355   7.23  0.00717 **
drug2:statin -0.13724  0.26795   0.26  0.60852
drug3:statin  0.64617  0.25386   6.48  0.01092 *
drug2:fu     -0.02258  0.35260   0.00  0.94893
drug3:fu      0.10337  0.34677   0.09  0.76564
statin:fu    -0.68499  0.28720   5.69  0.01708 *
```

The effect of statin use is estimated by the coefficient of `statin:fu`. Yes, there is evidence that statins decrease the prevalence of LDL>150 (p=0.017). Construct a 95% CI using the halfwidth of 1.96*Std.err = 1.96*0.29. For interpretation, use e to the power of the point estimate and CI limits. Since the coefficient is negative, there is evidence that statins lower the prevalence of LDL>150. The exp{`statin:fu`} term is a ratio of odds ratios; the odds ratio in the numerator measures the change in prevalence from baseline compared with follow-up with statins, and the odds ratio in the denominator measures the change in prevalence from baseline compared with follow-up without statins. The odds of high LDL are lowered with statins.