STAT 437 - Assignment 1
Due: Friday, January 21 on Crowdmark
Resubmit: Friday, February 4 on Crowdmark

This assignment covers the materials contained in **Lecture 002** through to **Lecture 009**. Reminder that you are permitted to discuss to these problems with classmates, but every student must submit their own solutions which are their own work (including any code, figures, etc.). **Please indicate any students that you discussed solutions with on your submission**. Please ensure that your submissions on Crowdmark are legible, and separated based on the problems included at the submission link. Submissions can be handwritten or typeset.

Part 1: True or False (20 Marks; 2 Marks Each)

For each of the following problems indicate whether the statement is true or false, and give a short justification for your answer. Correct answers without justification will receive only partial credit.

Problem 1. In an attempt to study the effects of long-term exposure to birds on individual health, researchers compared the rates of lung disease in long-term bird owners (15+ years) to the rates of lung disease in an (otherwise similar) group of new bird owners (<1 year), and a demographically-matched control group. They estimated that long-term bird owners were nearly twice as likely to have developed serious lung diseases. **True or false: assuming that the impact on health was correctly computed, this is an example of a longitudinal effect.**

Solution 1: [1] False. [1] Estimating a longitudinal effect requires a longitudinal study (following the same individuals overtime). This is instead a cohort effect, comparing the cohorts of long-term bird owners, to short-term bird owners, and non-bird owners.

Problem 2. Suppose that it is known that, as individuals age, their personality (as measured by the Big 5 Personality traits) changes independently of their past-self. That is, the correlation of these values (when measured distantly from one another) is expected to be zero, due to these measures being independent. **True or false: a study which follows individuals from childhood through adulthood, considering the development of their personality traits, can be analyzed using standard (e.g., linear or GLM) regression methods.**

Solution 2: [1] True. [1] Longitudinal methods are required because the repeated measurements tend to be dependent/correlated. If repeated measurements are truly independent, then the data could be transformed to long format, and analyzed as though it had been a set of IID individuals.

PROBLEM 3. You are given a dataset in long format. The dataset contains information on 100 individuals. It contains a column for their ID, for the time index that the measurement was taken, their outcome, and a treatment indicator. The first several rows of this dataset are shown below, **sorted by ID**.

ID	Time	Treatment	Outcome
1	1	1	30.5
1	2	1	35
1	3	1	35
2	2	0	26
2	3	0	24.5
2	5	0	20
⋮	⋮	⋮	⋮

True or false: when translated into wide format, the data frame will have 100 **rows and** 5 **columns.**

Solution 3: [1] False. [1] This is not guaranteed. If these data were balanced and each individual had 3 measurements, then it would be true, since we would have 1 column for ID, 1 column for treatment, and 3 columns for outcomes. However, these data are **not** balanced, and so we will have a column for every **unique** time measurement (which is at least 4 based on what is shown, and possibly more!).

PROBLEM 4. **True or false: in a linear regression model, in order to produce *valid* Wald-type confidence intervals for the coefficients (e.g.,** $\widehat{\beta} \pm 1.96 \times$ **s.e.**$(\widehat{\beta})$ **or similar), we must assume normality of the outcome (conditional on the covariates).**

Solution 4: [1] False. [1] Wald-type confidence intervals are valid in **small samples** only when the outcomes are assumed to be conditionally normal. If the sample size is sufficiently large, then as long as the other assumptions for OLS are met, the Wald-type intervals are (asymptotically) valid.

PROBLEM 5. Consider the following expressions for the conditional expectation of Y given X:

$$E[Y|X] = \beta_0 + \beta_1 X \tag{1}$$
$$E[Y|X] = \beta_0 + \beta_1 \log(X) \tag{2}$$
$$E[Y|X] = \beta_0 + \exp(\alpha_1 + X) \tag{3}$$
$$E[\log(Y)|X] = \exp(\alpha_0) + \beta_1 X^2 \tag{4}$$
$$E[Y|X] = \exp(\alpha_0 + \alpha_1 X) \tag{5}$$

True or false: all these models, except for (5), express a linear relationship in the conditional mean which is estimable with standard (OLS) linear regression.

Solution 5: [1] True. [1] Every model except for (5) is linear in the parameters, which is to say can be expressed as $E[W|Z] = \gamma_0 + \gamma_1 Z$ by correctly choosing W and Z; then γ_0 and γ_1 can be functionally related to the listed parameters. (5) cannot be since it requires the multiplication of two terms, rather than the summation.

[Not Required for Marks] In (1) we have $W = Y$, $Z = X$, and $\gamma = \beta$. In (2) we have $W = Y$, $Z = \log(X)$, and $\gamma = \beta$. In (3) we have $W = Y$, $Z = \exp(X)$, $\gamma_0 = \beta_0$, and $\gamma_1 = \exp(\alpha_1)$. In (4) we have $W = \log(Y)$, $W = X^2$, $\gamma_0 = \exp(\alpha_0)$, and $\gamma_1 = \beta_1$. In (5) the model simplifies to $\exp(\alpha_0)\exp(\alpha_1 X)$: if we take a $\log(\cdot)$ of both sides this can be fit using a GLM, but not standard OLS.

PROBLEM 6. Suppose we have fit a linear marginal model, with a structure given by

$$\begin{aligned} E[Y_{ij}|X_{ij}] = {} & \beta_0 + \beta_1 I(t_j = 2) + \beta_2 I(t_j = 3) + \beta_3 I(t_j = 4) + \beta_4 \text{Trt}_i + \beta_5 I(\text{Age}_i > 50) \\ & + \beta_6 I(t_j = 2)\text{Trt}_i + \beta_7 I(t_j = 3)\text{Trt}_i + \beta_8 I(t_j = 4)\text{Trt}_i \\ & + \beta_9 I(t_j = 2)I(\text{Age}_i > 50) + \beta_{10} I(t_j = 3)I(\text{Age}_i > 50) + \beta_{11} I(t_j = 4)I(\text{Age}_i > 50). \end{aligned}$$

Here, we have t_j representing the time index (that is, $I(t_j = 2)$ is 1 if it is time 2, and is 0 otherwise), the Trt_i is a binary indicator for active treatment, and $I(\text{Age}_i > 50)$ is 1 for any individual who was over 50 years old when the study began. Take

$$L = \begin{array}{c} \begin{array}{cccccccccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & \beta_9 & \beta_{10} & \beta_{11} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

True or false: testing the hypothesis $L\widehat{\beta} = \mathbf{0}$ performs a test of whether or not the time trend associated with treatment is equivalent to the time trend associated with being over 50 years old, and this hypothesis should be tested against a χ^2_3 distribution.

Solution 6: [1] True. [1] If $L\beta = 0$ then we equivalently find that $\beta_6 = \beta_9$, $\beta_7 = \beta_{10}$, and $\beta_8 = \beta_{11}$. These correspond to the terms that modify the time trends for Treatment and age, respectively. The distribution is χ^2_3 since L has rank 3.

PROBLEM 7. Two linear marginal models are fit on the same data, using the same conditional mean structure. The first makes an auto-regressive assumption for the correlation pattern matrix (i.e., $\text{cor}(Y_{ij}, Y_{i\ell}) = \rho^{|j-\ell|}$) and the second makes the exchangeable assumption (i.e., $\text{cor}(Y_{ij}, Y_{i\ell}) = \rho$). Both models are fit using maximum likelihood. The log-likelihood for the first model is found to be -1225 and the log-likelihood for the second model -1260.

True or false: we can test the hypothesis that the second model is adequate using a LRT, where the test statistic is $\Lambda = 70$, with a distribution of χ^2_1, resulting in a p-value < 0.001, and a rejection of the null hypothesis.

Solution 7: [1] False. [1] These models are not nested within one another, and as a result we cannot use the likelihood ratio test to test whether they are adequate.

PROBLEM 8. **True or false: The Poisson distribution (with parameter λ) is an exponential family distribution with: (i) canonical parameter $\theta = \log(\lambda)$, (ii) $b(\theta) = e^{\theta}$, (iii) $a(\phi) = 1$, (iv) $c(y, \phi) = -\log(y!)$, and (v) canonical link $g(x) = \log(x)$.**

Solution 8: [1] True. [1] The distribution function for a Poisson random variable is written $f(y; \lambda) = \frac{\exp^{-\lambda} \lambda^y}{y!}$. This can be re-written as

$$f(y; \lambda) = \exp\left\{y \log(\lambda) - \lambda - \log(y!)\right\}.$$

This gives $\theta = \log(\lambda)$, $a(\phi) = 1$, $b(\theta) = e^{\theta}$, and $c(y, \phi) = -\log(y!)$. Since $E[Y] = \lambda = e^{\theta}$, then $g(x) = \log(x)$ is the canonical link since $g(\mu) = \log(\lambda) = \log(e^{\theta}) = \theta$.

PROBLEM 9. Suppose that we have count data which are known to be highly over-dispersed, relative to the Poisson assumption (that is, var(Y) is much larger than $E[Y]$). Suppose further that $E[Y|X] = \exp(\beta_0 + \beta_1 X)$, for unknown β_0 and β_1. A very large sample of independent realizations is taken. Consider

$$U(\beta) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \left\{y_i - \exp(\beta_0 + \beta_1 x_i)\right\}.$$

True or false: Solving $U(\widehat{\beta}) = \mathbf{0}$, will produce a consistent estimator for β.

Solution 9: [1] True. [1] This is a quasi-likelihood estimator, with $V(\mu_i) = \mu_i$ specified, that is

$$U(\beta) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \exp(\beta_0 + \beta_1 x_i) \left(\exp(\beta_0 + \beta_1 x_i)\right)^{-1} \left(y_i - \exp(\beta_0 + \beta_1 x_i)\right).$$

The variance is incorrectly specified (does not account for over-dispersion), however, since the sample is sufficiently large, the asymptotic distribution will still be correct (since the mean model is correctly specified).

PROBLEM 10. A new, experimental treatment is being tested as a means of delaying the onset of dementia. Suppose a marginal linear model has been fit, using A_i to represent the treatment assignment (binary indicator, with $A_i = 0$ for placebo and $A_i = 1$ for experimental treatment) for individual i, and where t_j is time, treated as a continuous variate. Consider the model given by

$$E[Y_{ij}|A_i, t_j] = \beta_0 + \beta_1 t_j + \beta_2 A_i t_j.$$

True or false: β_1 is interpreted as the expected change in outcome as time passes (i.e., for a 1 unit increase in time) for those in the placebo group and β_2 represents the expected change in outcome as time passes (i.e., for a 1 unit increase in time) for those in the experimental treatment group.

Solution 10: [1] False. [1] β_1 **is** the longitudinal effect for those in the placebo group **however** β_2 is **not** the longitudinal effect for those in the active treatment group. Instead, β_2 represents the difference in the time effect between the placebo and active treatment groups, making $\beta_1 + \beta_2$ the time effect for the experimental treatment group.

Part 2: Conceptual Question (10 Marks)

For the following questions, provide your answers with justification and clear communication. The answers do not need to be long, but correct responses without complete justification will receive only partial credit.

Suppose that the viral load (a continuous outcome) is repeatedly measured on HIV positive subjects from three treatment groups: a placebo (control) group, a low-dose group, and a high-dose group. We treat time, t_{ij} as measuring the number of days since baseline enrollment in the study. At baseline, individuals had no discernible differences in viral load and were assigned to their treatment groups completely at random.

Suppose that a marginal linear model is fit to the data such that,

$$E[Y_{ij}|X_{ij}, t_{ij}] = \beta_0 + \beta_1 x_{i1} t_{ij} + \beta_2 x_{i2} t_{ij} + \beta_3 x_{i3} t_{ij},$$

where $x_{i1} = 1$ if the patient is in the control group, $x_{i2} = 1$ if the patient is in the low-dose group, and $x_{i3} = 1$ if the patient is in the high-dose group (with these indicators taking 0 otherwise). The following questions all relate to this model.

Problem 11. (2 Marks) There is no main effect of treatment included in this model. Explain why this is a natural choice and indicate what the inclusion of (non-zero) treatment effects would indicate about the underlying scenario.

Solution 11: [1] This is a natural choice since the treatment groups had no differences in observed outcome at baseline ($t_{ij} = 0$). They were randomized to their treatments without regard for their current viral load, and so when $t_{ij} = 0$ we should not expect a difference (on average) between these groups.
[1] If we had included a non-zero treatment effect, this would correspond to the groups starting with different levels of viral load. This might be reasonable if (for instance) the group receiving high-dose therapy started with higher viral loads.

Problem 12. (3 Marks) Suppose that we wish to test whether or not the response of the low-dose group changes over time. Write down the hypothesis in terms of model parameters, as a general linear hypothesis of the model (e.g., $L\beta = \mathbf{c}$ for suitable L and $\mathbf{c}$) and specify distribution under the null.

Solution 12: [1] This is a test of $H_0 : \beta_2 = 0$. [1] We would take $L = (0, 0, 1, 0)$ with $c = 0$. [1] The distribution under the null is χ^2_1, since L is rank 1.

Problem 13. (3 Marks) Suppose that we wish to test whether the responses of the low-dose and high-dose groups change at the same rate over time, while the control group exhibit no time-trend. Write down the hypothesis in terms of model parameters, as a general linear hypothesis of the model (e.g., $L\beta = \mathbf{c}$ for suitable L and $\mathbf{c}$) and specify distribution under the null.

Solution 13: [1] This is a test of the joint hypothesis, $H_0 : \beta_2 = \beta_3$ and $H_0 : \beta_1 = 0$. [1] We can take

$$L = \begin{pmatrix} 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

with $\mathbf{c} = \mathbf{0}$. [1] The distribution under the null is χ^2_2, since L is rank 2.

Problem 14. (2 Marks) Suppose that we consider two individuals, one in the low-dose group, and one in high-dose group. If it is found that $\widehat{\beta}_2 = 0.5\widehat{\beta}_3 < 0$, what is the predicted difference in viral load between these two patients after 10 days? Suppose that the first individual remains on treatment for 60 days. How long would the second individual need to remain on treatment to have an equivalent viral load prediction?

Solution 14: [1] After 10 days, the first patient has a predicted viral load of $\widehat{\beta}_1 + 10\widehat{\beta}_2 = \widehat{\beta}_1 + 5\widehat{\beta}_3$ (since $\widehat{\beta}_2 = 0.5\widehat{\beta}_3$). The second has a response of $\widehat{\beta}_1 + 10\widehat{\beta}_3$. As a result, the difference between the two is $-5\widehat{\beta}_2$. [1] If the first individual is treated for 60 days, they will have a predicted response of $\widehat{\beta}_1 + 30\widehat{\beta}_3$, which is the same predicted response that the second individual will have after **30 days**.

Part 3: Theoretical or Applied Question (20 Marks)

Please **pick one** of the following two problems and solve it. **If you solve both** only the first problem will be looked at. The first problem corresponds to an **application** problem, while the second is a **theoretical question**. For the application question, please provide the code and relevant output (consider using a software like RMarkdown, or being highly selective with what output you copy to ensure your solution is legible). For the theoretical question, please include enough of your work to justify the steps you have taken.

Problem 15. (Application) On the course website you will find a data file `dental.csv`. This reports a study of dental growth where measurements of the distance (in mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14. The data file contains these observations in wide format, recording the ID, Sex, and outcomes (Y_1, Y_2, Y_3, Y_4).

1. Generate relevant plots for the data that illustrate how the data evolve overtime, to help inform modelling. Explain what information your plots show.

2. Fit a fully saturated, marginal linear model to the data, with time and sex both treated as discrete. Use an unstructured correlation matrix and allow for variances to change over each week. Based on your fitted saturated model, is the time trend the same between the two groups?

3. Based on your model, provide an estimated 95% confidence interval for the mean distance for 8 year old boys, and 14 year old girls.

4. Determine whether it is possible to simplify the assumed correlation and variance structure. Consider: (1) a model with assumed constant variance but unconstrained correlation, (2) a model with compound symmetry but differing variance, and (3) a model which assumes constant variance and uses an auto regressive correlation structure. Are any of these models appropriate?

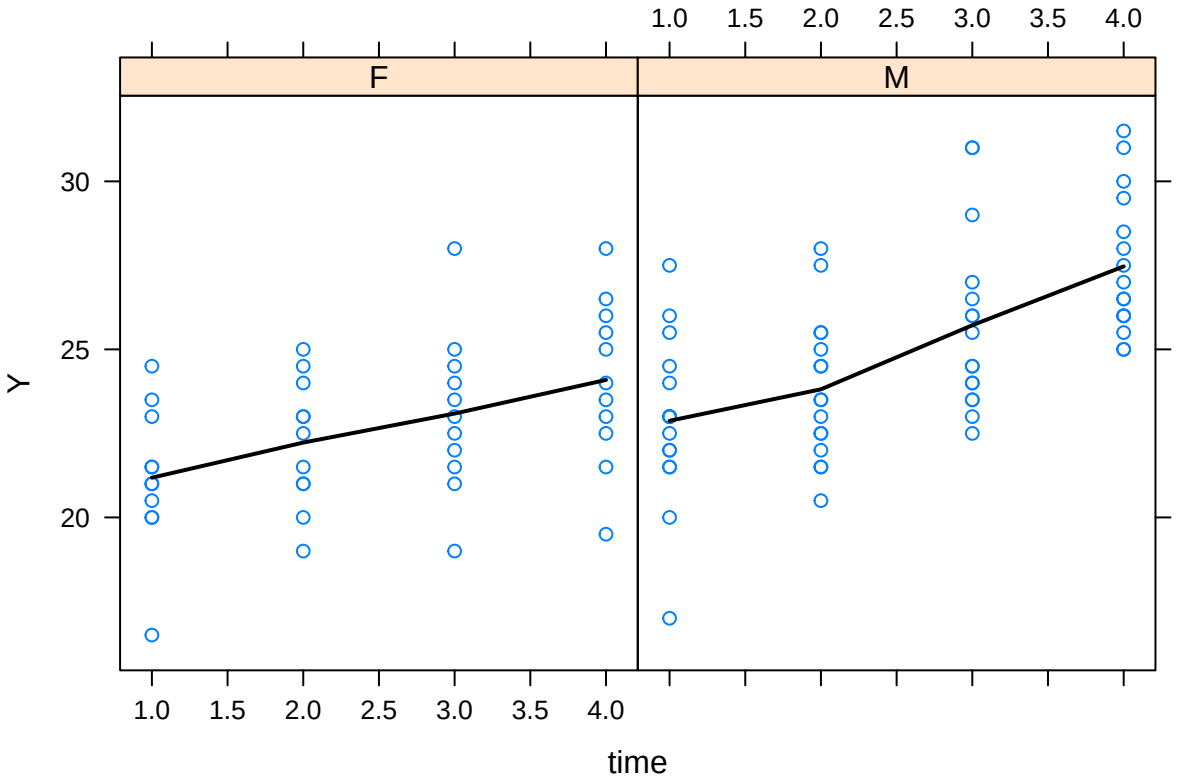
5. A doctor is interested in the distance measurements for 13 year old girls, and 15 year old boys. Provide 95% confidence interval for the mean predictions of these categories, and indicate whether or not these predictions are appropriate.

Solution 15: For complete marks, the solution must display relevant code (and output). The following are example solutions, and yours may differ slightly. However, you should be justifying choices that you make, and fully explaining what you are doing (and why).

1. (3 marks total). [2] marks for selecting plot(s) which do a good job of illustrating time trends in the data, and [1] mark for describing those trends.

```
xyplot(Y ~ time | Sex,
       groups = ID,
       data = dental_long,
       panel = function(x, y){
```

```
        panel.xyplot(x, y, type='p')
        panel.linejoin(x, y, fun=mean, horizontal=F, lwd=2, col=1)
      })
```



From this plot we see the (left-hand panel) female and (right-hand panel) male growth, over the four time points (8 through 14 years). The plotted lines illustrate the means (within groups) at each time point, to show the overall trend in the data. We can see that the spread of points appears to be fairly consistent overtime, and that both groups exhibit a (mostly linear) upward trend. The males start with larger values at baseline and appear to increase at a faster rate. Compared to the other ages, the increase from 8 to 10 year old males appears to be less pronounced, whereas the increases appear to be fairly constant over time for females.

2. (6 marks total). [1] mark for the correct formula (time as factor, with interaction), [1] mark for the correct variance structure, [1] mark for the correct correlation structure, [1] mark for indicating the correct hypothesis test, [1] mark for computing the correct test statistic/p-value, and [1] mark for getting the correct conclusion. **Note: if using LRT, you must have used ML**.

```
saturated.model <- gls(Y ~ as.factor(time)*as.factor(Sex),
                       data = dental_long,
                       weights = varIdent(form = ~ 1 | as.factor(time)),
                       correlation = corSymm(form = ~1 | ID),
                       method = 'ML')
```

Parameter	Factor	Value	Standard.Error
β_0	1	21.1818	0.7017
β_1	$I(t_j = 2)$	1.0455	0.6155
β_2	$I(t_j = 3)$	1.9091	0.6068
β_3	$I(t_j = 4)$	2.9091	0.6729
β_4	$I(\text{Sex}_i = \text{M})$	1.6932	0.9115
β_5	$I(t_j = 2)I(\text{Sex}_i = \text{M})$	-0.1080	0.7995
β_6	$I(t_j = 3)I(\text{Sex}_i = \text{M})$	0.9347	0.7883
β_7	$I(t_j = 4)I(\text{Sex}_i = \text{M})$	1.6847	0.8741

This first model can be fit using either `ML` or `REML`, but importantly should have `time` treated as factor, must have `weights = varIdent(form = ~ 1|time)` and `correlation = corSymm(form = ~1|ID)`. In order to answer whether the time trend is the same between groups, we are testing $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$. There are two main ways of conducting this test, either through the specification of

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and then testing $H_0 : L\beta = \mathbf{0}$. Alternatively you could fit the nested model without the interaction term, and conduct a LRT. To use this second method **the model must be fit using** `ML`.

Method 1:

```
L <- rbind(c(0,0,0,0,0,1,0,0),
           c(0,0,0,0,0,0,1,0),
           c(0,0,0,0,0,0,0,1))
beta.hat <- coef(saturated.model)
v.beta.hat <- saturated.model$varBeta

LB <- c(0,0,0) # Value under H0

W <- t(L%*%beta.hat - LB)%*%
      solve(L%*%v.beta.hat%*%t(L))%*%
      (L%*%beta.hat - LB)

1 - pchisq(W, df = 3) # df=3 since rank(L) = 3
```

```
##             [,1]
## [1,] 0.0322315
```

We find a test statistic value of 8.789, which corresponds to a p-value of 0.0322 and as a result we reject the null hypothesis (at a 5% significance level). That is, there is evidence that the two groups have different time trends.

Method 2:

```
reduced.model <- gls(Y ~ as.factor(time)+as.factor(Sex),
                     data = dental_long,
                     weights = varIdent(form = ~ 1 | as.factor(time)),
                     correlation = corSymm(form = ~1 | ID),
                     method = 'ML')
anova(reduced.model, saturated.model)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## reduced.model       1 15 454.6432 494.8752 -212.3216
## saturated.model     2 18 452.5093 500.7877 -208.2546 1 vs 2 8.133946  0.0433
```

Based on the LRT of the reduced model (without interactions) and the saturated model (with interactions), we find a LRT statistics of 8.134 which corresponds to a p-value (based on χ^2_3 distribution) of 0.0433 and so we reject the null hypothesis and conclude that the two groups differ substantially in their time trends.

3. (3 Marks Total) [1] mark for the correct L matrices, [1] mark for the correct standard errors, and [1] mark for using them correctly.

   Generally, when estimating with regards to variance/correlation parameters you would want to use `REML` rather than `ML`. This is not strictly required (but advisable due to the small sample size). In this case, the results of the model do not seem to materially change whether `ML` or `REML` is used. Consider that an 8 year old male is represented by $L = (1, 0, 0, 0, 1, 0, 0, 0)$ and a 14 year old female by $L = (1, 0, 0, 1, 0, 0, 0, 0)$. To generate 95% confidence intervals for these predictions we can use the following.

   With `REML`

```
saturated.model.REML <- gls(Y ~ as.factor(time)*as.factor(Sex),
                            data = dental_long,
                            weights = varIdent(form = ~ 1 | as.factor(time)),
                            correlation = corSymm(form = ~1 | ID),
                            method = 'REML')

L <- rbind(c(1,0,0,0,1,0,0,0),
           c(1,0,0,1,0,0,0,0))

beta.hat <- coef(saturated.model.REML)
se <- sqrt(diag(L%*%saturated.model.REML$varBeta%*%t(L)))
point_est <- L %*% beta.hat
```

```
    cbind(
      point_est - qnorm(0.975)*se,
      point_est,
      point_est + qnorm(0.975)*se
    )
```

```
##          [,1]     [,2]     [,3]
## [1,] 21.73474 22.87500 24.01526
## [2,] 22.77139 24.09091 25.41043
```

As a result, the estimated mean for 8 year old males is 22.875 with a 95% CI of $(21.73, 24.02)$. For 14 year old females, the estimated mean is 24.09 with a 95% CI of $(22.77, 25.41)$.

4. (4 Marks) [1] mark for each of the 3 models being fit correctly, [1] mark for all of the hypothesis tests being correctly reported.

   In order to test the fit of the differing models, we simply fit new models and use the `ANOVA` call in `R`. Note that constant variance with unconstrained correlation, compound symmetry with unconstrained variance, and constant variance with `AR(1)` correlation are all nested within the saturated model we have fit. Also, as a helpful hint, we can take a model that we have already fit, and run `update()` to change the call!

```
    model.1 <- update(saturated.model, weights = varIdent(form=~1))
    model.2 <- update(saturated.model, correlation = corCompSymm(form=~1|ID))
    model.3 <- update(saturated.model,
                      weights = varIdent(form = ~1),
                      correlation = corAR1(form = ~1|ID))

    anova(saturated.model, model.1)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## saturated.model     1 18 452.5093 500.7877 -208.2546
## model.1             2 15 448.3770 488.6090 -209.1885 1 vs 2 1.867728  0.6003
```

```
    anova(saturated.model, model.2)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## saturated.model     1 18 452.5093 500.7877 -208.2546
## model.2             2 13 450.4892 485.3569 -212.2446 1 vs 2 7.979904  0.1573
```

```
    anova(saturated.model, model.3)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## saturated.model     1 18 452.5093 500.7877 -208.2546
## model.3             2 10 458.6627 485.4840 -219.3313 1 vs 2 22.15336  0.0046
```

Based on this output, it seems that we could make either of the first simplifying assumptions (we fail to reject the null that the parameter constraints are valid), while the third model is rejected (even at a 1% significance level). Correspondingly, we could likely specify a more parsimonious model which is desirable.

5. (4 Marks) [1] mark for fitting a model with time as the continuous variate, [1] mark for using this to generate valid predictions (watch that the time scale is considered correctly), [1] mark for a discussion of whether the model is appropriate, [1] mark for a discussion of extrapolation.

   In order to make predictions when the ages are 13 or 15, we cannot treat time as a discrete variable. As a result, we need to re-fit this model, using some continuous structure of time. The plots suggest that a linear time trend will likely suffice. We also likely want different time trends for males and females, as well as different baseline values (since they appear to change at different rates, and start from different spots).

   We also may want to re-scale time. This is not strictly necessary, but you need to make sure to convert to the correct timescale if you do not. On the current timescale, an increase of time by 1 corresponds to two years, and $t = 1$ is when the age is 8. As a result, 13 will be $t = 3.5$ and 15 will be $t = 4.5$. I re-scale my times instead!

```
dental_long$time <- 2*(dental_long$time - 1) + 8
```

```
linear.model <- gls(Y ~ time*as.factor(Sex),
                    correlation = corSymm(form = ~1|ID),
                    weights = varIdent(form = ~1), # Use constant variance.
                    data = dental_long,
                    method = 'ML')
```

It is worth checking the model fit of our new model. We can simply compare the `AIC` and `BIC` from the saturated and the new linear model as a first pass. Our new model has an AIC of 443.234835 and a BIC of 472.7382785, while the old fit had values of 452.5093018 and 500.7876639, respectively. Smaller values are preferable, and so we can use this to justify the use of this model.

Alternatively, we can actually take a formal hypothesis test. The linear model is nested within the discrete time model, albeit in a way that may be non-obvious. If we restrict $2\beta_1 = \beta_2$, $3\beta_1 = \beta_3$, $2\beta_5 = \beta_6$ and $3\beta_6 = \beta_7$, then since the time gap between the factor levels corresponding to those different parameters is constant (2 years) this will be equivalent to forcing the linear relationship. As a result, we can actually perform a nested hypothesis test as to the permissibility of this assumption.

```
# Refit the model using the same variance assumption
factor.model <- gls(Y ~ as.factor(time)*as.factor(Sex),
                    correlation = corSymm(form = ~1|ID),
                    weights = varIdent(form = ~1), # Use constant variance.
                    data = dental_long,
                    method = 'ML')

# Test the nested models
anova(factor.model, linear.model)
```

```
##              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## factor.model     1 15 448.3770 488.6090 -209.1885
## linear.model     2 11 443.2348 472.7383 -210.6174 1 vs 2 2.857805  0.5819
```

With a p-value of 0.5819 we fail to reject the null and our formal test seems to agree with the `AIC` and `BIC` criteria above. (We could also test this with a compound linear hypothesis, to similar results!)

One other caveat on whether predictions from this model will be appropriate is that we are necessarily extrapolating beyond what is observed in the data. For a prediction at age 13 this is not so much an issue as we have data observed on both sides, and can be fairly confident that the trend continues in the intervening years (supposing there is nothing particularly special about even ages). However, for age 15 we are extrapolating out of the sample that we took: we should always be cautious with this. If we had scientific evidence that the trends change after the age of 14 this prediction may not be appropriate. As a result, we must bracket this prediction with the knowledge that **it is only valid if the observed trend continues**.

```
L <- rbind(c(1,13,0,0),
           c(1,15,1,15))

beta.hat <- coef(linear.model)
se <- sqrt(diag(L%*%linear.model$varBeta%*%t(L)))
point_est <- L %*% beta.hat

cbind(
  point_est - qnorm(0.975)*se,
  point_est,
  point_est + qnorm(0.975)*se
)
```

```
##          [,1]     [,2]     [,3]
## [1,] 22.38062 23.59785 24.81509
## [2,] 27.12158 28.29139 29.46119
```

As a result, we estimate that a 13 year old girls will have mean distances of 23.60 (with a 95% CI of $(22.38, 24.82)$) and that 15 year old boys will have mean distances of 28.29 (with a 95% CI of $(27.12, 29.46)$).

PROBLEM 16. (Theoretical) Consider a marginal linear model where we take the mean to be given by $\mu_i = X_i\beta$ and the variance $\Sigma_i = \sigma^2\mathbf{R}(\rho)$. Recall that we stated that the MLE for β are given by

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'\Sigma_i^{-1}Y_i.$$

1. If $\Sigma_i = \sigma^2\mathbf{R}(\rho)$, show that the MLE of β does not depend on σ^2.
2. Derive the maximum likelihood estimator for σ^2.
3. What is the expression for the profile log-likelihood of ρ (can be in terms of $\widehat{\beta}$ and $\widehat{\sigma}^2$)?
4. Prove that the MLE for the marginal linear model simplify to the OLS estimators if we assume constant variance and independent correlation.
5. Show that, if the data are balanced with equal spacing between follow-ups, assuming a first-order auto regressive correlation structure is equivalent to assuming an exponential correlation structure.

Solution 16: 1. (3 Marks Total) Here we can simply make an algebraic substitution into the MLE, as written.

$$\begin{aligned}
\widehat{\beta} &= \left(\sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'\Sigma_i^{-1}Y_i \\
&= \left(\sum_{i=1}^{n} X_i'[\sigma^2\mathbf{R}_i(\rho)]^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'[\sigma^2\mathbf{R}_i(\rho)]^{-1}Y_i \\
&= \left(\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}X_i\right)^{-1}\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}Y_i \\
&= \sigma^2\left(\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}X_i\right)^{-1}\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}Y_i \\
&= \left(\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'\mathbf{R}_i(\rho)^{-1}Y_i.
\end{aligned}$$

As a result, this expression does not depend on σ^2 and so $\widehat{\beta}$ is functionally independent of σ^2.

2. (8 Marks total) We saw in lecture that the log-likelihood is given by

$$\ell(\beta, \sigma^2, \rho) = \sum_{i=1}^{n} -\frac{k}{2}\log|2\pi| - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(Y_i - X_i\beta)'\,\Sigma_i^{-1}\,(Y_i - X_i\beta)\,.$$

Taking the assumption that $\Sigma_i = \sigma^2\mathbf{R}_i(\rho)$, and noting that $\mathbf{R}_i(\rho)$ is assumed constant over i, we get

$$\begin{aligned}
&\ell(\beta, \sigma^2, \rho) \\
&= \sum_{i=1}^{n} -\frac{k}{2}\log|2\pi| - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log|\mathbf{R}_i(\rho)| - \frac{1}{2\sigma^2}(Y_i - X_i\beta)'\,\mathbf{R}_i(\rho)^{-1}\,(Y_i - X_i\beta) \\
&= -\frac{nk}{2}\log|2\pi| - \frac{nk}{2}\log(\sigma^2) - \frac{n}{2}\log|\mathbf{R}_i(\rho)| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)'\,\mathbf{R}_i(\rho)^{-1}\,(Y_i - X_i\beta)\,.
\end{aligned}$$

Note that here we have used the fact that, for a $k \times k$ matrix M and constant c, $|cM| = c^k|M|$. From this, we can write the score function by differentiating the expression with respect to σ^2. Note the first and third terms above are independent of σ^2, so they differentiate to 0. This gives

$$S_{\sigma^2} = \frac{\partial}{\partial\sigma^2}\ell(\cdot) = -\frac{nk}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i\beta)'\,\mathbf{R}_i(\rho)^{-1}\,(Y_i - X_i\beta)\,.$$

Then by setting $S_{\sigma^2} = 0$ and isolating for $\widehat{\sigma}^2$ we get

$$\widehat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{n}(Y_i - X_i\beta)'\,\mathbf{R}_i(\rho)^{-1}\,(Y_i - X_i\beta)\,.$$

3. (2 Marks Total) Using the expression for the log-likelihood derived in the lecture, we get that

$$\ell_p(\rho) = -\frac{nk}{2}\log|2\pi| - \frac{nk}{2}\log(\widehat{\sigma}^2) - \frac{n}{2}\log|\mathbf{R}_i(\rho)| - \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n}\left(Y_i - X_i\widehat{\beta}\right)'\,\mathbf{R}_i(\rho)^{-1}\,\left(Y_i - X_i\widehat{\beta}\right).$$

4. (3 Marks Total) Consider that an independence correlation assumption, with constant variance, takes $\Sigma_i = \sigma^2 I$. In (1) we showed that $\widehat{\beta}$ does not depend on σ^2, and so taking $\mathbf{R}_i(\rho) = I$ we get

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1}\sum_{i=1}^{n} X_i'Y_i.$$

This is exactly the OLS model estimators.

You could also have argued on the basis of properties of the multivariate normal distribution. If correlation is zero between two components of a multivariate normal distribution, they are independent: as a result, the data becomes IID (by assumption), and can be fit through a standard OLS regression.

5. (4 Marks Total) If the data are balanced and equally spaced, we have that $|t_{ij} - t_{i,j+1}| = d$ for some constant d. This can be expanded to conclude that $|t_{ij} - t_{i\ell}| = d|j - \ell|$. Then, considering the exponential correlation structure we see

$$\begin{aligned} \text{corr}(Y_{ij}, Y_{i\ell}) &= \exp(-\rho|t_{ij} - t_{i\ell}|) \\ &= \exp(-\rho d|j - \ell|) \\ &= \rho_*^{|j-\ell|}, \end{aligned}$$

where $\rho_* = \exp(-\rho d)$. This is a first-order auto regressive assumption.