ST 503 Homework 5

Eric Warren

2024-03-19

Contents

1 Problem 1

Consider the following model: $Y_{ij} = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_i t_{ij}\beta_5 + M_i t_{ij}^2\beta_6 + M_i\alpha_i\beta_7 + e_{ij}$ where where t_{ij} denotes time of the j-th measurement of the i-th individual, α_i is the age of the i-th individual at the start of the study, and $M_i = 1$ if the i-th individual is a Male and $M_i = 0$ if female.

1.1 Part A

There are two groups: male and female. Write the group means $E(Y_{ij})$ for each group.

Let us remember some things which are how $E(e_{ij}) = 0$ and $E(M_i) = 0$ if female and $E(M_i) = 1$ if male. Now we can do that to calculate our model which is $Y_{ij} = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_i t_{ij}\beta_5 + M_i t_{ij}^2\beta_6 + M_i\alpha_i\beta_7 + e_{ij}$ which we can say the general form of our model is

$$E(Y_{ij}) = E(\beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_i t_{ij}\beta_5 + M_i t_{ij}^2\beta_6 + M_i\alpha_i\beta_7 + e_{ij})$$

$$= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7 + E(e_{ij})$$
$$= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7 + 0$$
$$= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7$$

So $E(Y_{ij}) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7$. Now let us look at our two cases.

- Female ($E(M_i) = 0$): So we have $E(Y_{ij}) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7$ and thus $E(Y_{ij}|Female) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + 0\beta_4 + 0t_{ij}\beta_5 + 0t_{ij}^2\beta_6 + 0\alpha_i\beta_7 = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3$. So $E(Y_{ij}|Female) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3$.
- Male ($E(M_i) = 1$): So we have $E(Y_{ij}) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + E(M_i)\beta_4 + E(M_i)t_{ij}\beta_5 + E(M_i)t_{ij}^2\beta_6 + E(M_i)\alpha_i\beta_7$ and thus $E(Y_{ij}|Male) = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + 1\beta_4 + 1t_{ij}\beta_5 + 1t_{ij}^2\beta_6 + 1\alpha_i\beta_7 = (\beta_0+\beta_4) + t_{ij}(\beta_1+\beta_5) + t_{ij}^2(\beta_2+\beta_6) + \alpha_i(\beta_3+\beta_7)$. So $E(Y_{ij}|Male) = (\beta_0+\beta_4) + t_{ij}(\beta_1+\beta_5) + t_{ij}^2(\beta_2+\beta_6) + \alpha_i(\beta_3+\beta_7)$.

1.2 Part B

Modify the model above assuming that age (α) has the same effect for both the groups and write its mathematical form.

If we were to assume that age has the same general effect for both groups that must mean that $\alpha_i\beta_4 = \alpha_i(\beta_4 + \beta_7)$. From quick algebra, we can easily find this now means that $\beta_4 = 0$. Therefore we can rewrite our model that was:

$$Y_{ij} = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_it_{ij}\beta_5 + M_it_{ij}^2\beta_6 + M_i\alpha_i\beta_7$$
$$= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_it_{ij}\beta_5 + M_it_{ij}^2\beta_6 + M_i\alpha_i0$$
$$= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_it_{ij}\beta_5 + M_it_{ij}^2\beta_6$$

So our new model is $Y_{ij} = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_it_{ij}\beta_5 + M_it_{ij}^2\beta_6$.

1.3 Part C

We might ask the question whether the individuals were similar on average at baseline ($t = 0$) or not. In the context of your model in (b), which parameter(s) are we interested in? Explain. [Hint: compare $E(Y_{ij})$ for the two groups at $t = 0$]

We should look at the β terms that are representing the t_{ij} terms. In this case, we can see how there is a t_{ij} (or t_{ij}^2) term for each of β_1, β_2, β_5, and β_6. So the parameters we should be interested in looking at are β_1, β_2, β_5, and β_6.

1.4 Part D

The next question is whether the mean response change at constant rates (i.e., they are only linear trends over time, not quadratic) for the two groups. In the context of your model in (b), which parameter(s) are we interested in? Explain your answer.

So we should only look at the linear trend terms. In our model which is $Y_{ij} = \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + \alpha_i\beta_3 + M_i\beta_4 + M_it_{ij}\beta_5 + M_it_{ij}^2\beta_6 + e_{ij}$, we can see the only linear trend terms that exist are with β_1, β_3, and β_5. However only the β_5 term is the only one that shows a difference between both groups since it has the M_i term with it, which distinguishes a difference between the groups. So the only parameter that we should be interested when trying to answer this question is β_5.

1.5 Part E

Finally, we ask the question whether the mean response show an identical pattern of change (intercepts might be different, but trends are parallel) across time for all groups. Assuming your model in (b), which parameter(s) are we interested in? Explain your answer.

Similar rationale to **Part D** we want to find the β's that have the M_i term in them which is there to show a difference between the groups but we also want to make sure they have a t_{ij} (or t_{ij}^2) term as well. If we look at our model that we have seen before which is $Y_{ij} = \beta_0+t_{ij}\beta_1+t_{ij}^2\beta_2+\alpha_i\beta_3+M_i\beta_4+M_it_{ij}\beta_5+M_it_{ij}^2\beta_6+e_{ij}$, only β_5 and β_6 meets these conditions which is why those two parameters are the only two we are interested in to answer these questions.

2 Problem 2

2.1 Read in Data

First we are going to read in our data and briefly show how it looks

```
library(tidyverse)

# Read in the data
dental <- read.table("dental.txt", header=F)

# Make column names
names(dental) <- c("obs", "child", "age", "distance", "gender")

# Change the gender column to a factor
dental$gender <- factor(dental$gender)

# Show the first couple of observations for our data
head(dental)
```

```
##   obs child age distance gender
## 1   1     1   8     21.0      0
## 2   2     1  10     20.0      0
## 3   3     1  12     21.5      0
## 4   4     1  14     23.0      0
## 5   5     2   8     21.0      0
## 6   6     2  10     21.5      0
```

2.2 Part A

From the data, compute and plot the mean response across age for each of the two groups (boys and girls). [Hint: Review Figure 1 of "The Vlagtwedde-Vlaardingen Study" in the "GLS" notes]

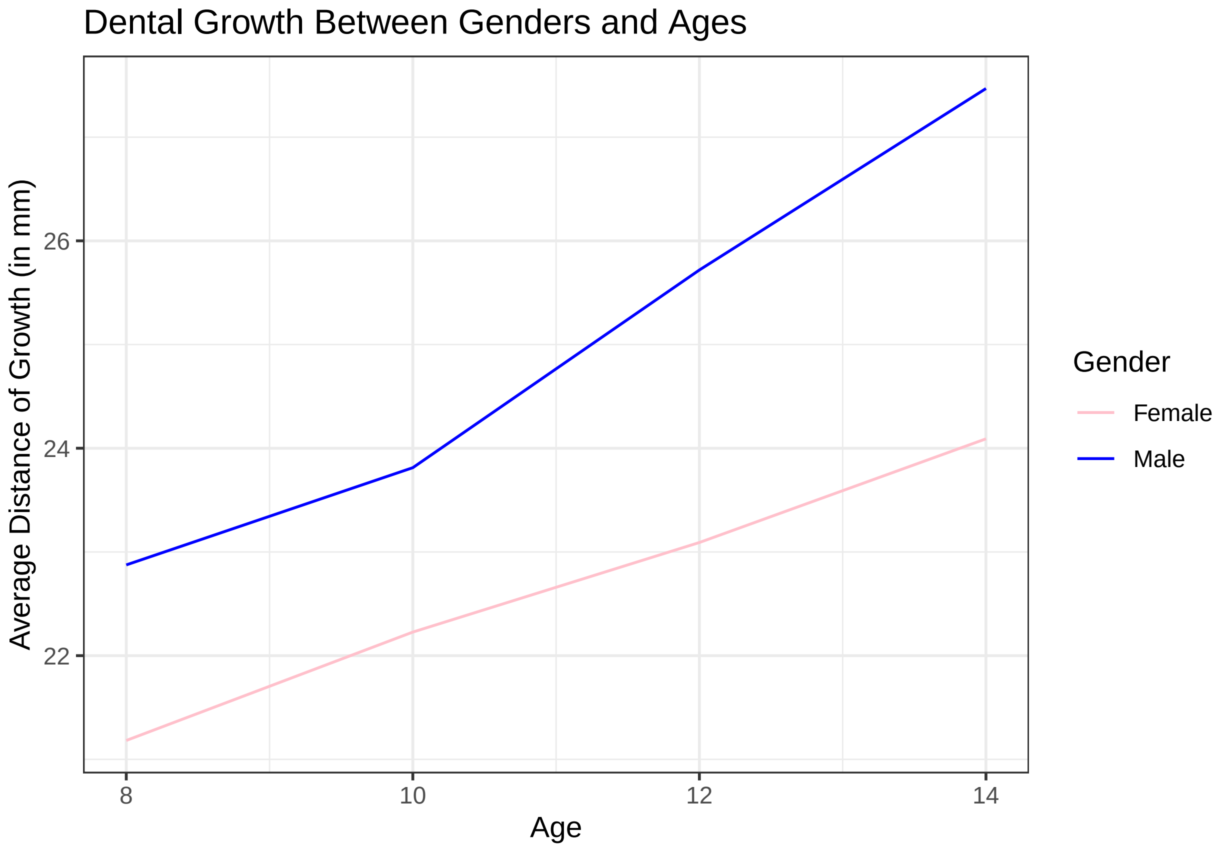
Here we are going to group our data by gender and then age and then after getting the mean values we are going to plot them. First to get the mean values of each category.

```
(dental_mean_groups <- dental %>%
  group_by(gender, age) %>%
  summarize(mean_mm = mean(distance)))
```

```
## # A tibble: 8 x 3
## # Groups:   gender [2]
##   gender   age mean_mm
##   <fct>  <int>   <dbl>
## 1 0          8    21.2
## 2 0         10    22.2
## 3 0         12    23.1
## 4 0         14    24.1
## 5 1          8    22.9
## 6 1         10    23.8
## 7 1         12    25.7
## 8 1         14    27.5
```

Now that we have seen the mean values we can make a plot of them.

```
dental_mean_groups %>%
  ggplot(aes(x = age, y = mean_mm, color = gender)) +
  geom_line() +
  labs(title = "Dental Growth Between Genders and Ages",
       x = "Age",
       y = "Average Distance of Growth (in mm)",
       color = "Gender") +
  scale_color_manual(labels = c("Female", "Male"), values = c("pink", "blue")) +
  theme_bw()
```



As we can see from our plot it seems that there is a difference between the gender of a child for growth but also a positive linear trend with age for each gender as well.

2.3 Part B

Write a model where the mean response in linear in age, where intercept and slope are different for the two groups. Explain why such a model is reasonable in this data.

First we are going to make a model for the mean response. We can write our model as $Y_{ij} = \beta_0 + t_{ij}\beta_1 + M_i\beta_2 + t_{ij}M_i\beta_3$ (and say $+e_{ij}$ to add the error term but since we are looking at mean response and $E(e_{ij}) = 0$ I left this out). In this case the t_{ij} represents each time where the measurement is taking place (or what age) while M_i represents the gender of a child where $M_i = 0$ if the child is a female and $M_i = 1$ if the child is a male. On top of that, we can also see that i represents the child we are measuring and j represents each child's individual measurement. We can also determine that a linear model is adequate in this case because in **Part A** we noticed how the trends of age and distance (growth) seemed to be linear. It is also appropriate to add the gender grouping variable M_i since the graph makes it seem there is a difference between the gender of the child.

2.4 Part C

Use the mean model in (b), and fit the following models error covariance (same for both groups). Choose a model for the covariance that adequately fits the data (you may use either AIC or BIC).

1. compound symmetry, same error variance over age
2. compound symmetry, different error variance over age

We are going to fit two models. The first is the compound symmetry but assuming same variance over age. The second is the compound symmetry but assuming difference variance over age.

```r
# Add library for gls model
library(nlme)

# Specify the model form
model_form <- distance ~ age + gender + age:gender

# Now make the first model with equal variance assumption across week
model1 <- gls(model = model_form,
              data = dental,
              correlation = corCompSymm(form = ~1 | child))

# Now make the second model with different variance assumption across week
model2 <- gls(model = model_form,
              data = dental,
              correlation = corCompSymm(form = ~1 | child),
              weights = varIdent(form = ~1 | age))

# Now display a table of the model AIC values
Model <- c("Equal Variance", "Different Variance")
`AIC Values` <- c(AIC(model1), AIC(model2)) # AIC Values
`BIC Values` <- c(BIC(model1), BIC(model2)) # BIC Values
knitr::kable(cbind(Model, `AIC Values`, `BIC Values`))
```

Model	AIC Values	BIC Values
Equal Variance	445.757249201205	461.623594596053

Model	AIC Values	BIC Values
Different Variance	449.9723761426	473.771894234872

When deciding which model is "better", we tend to look at which model has the lower AIC or BIC values. Sometimes these results can be conflicting but in this case, it is obvious as both the AIC and BIC values are lower for the Compound Symmetry Equal Variance Over Age (or Model 1). So for the rest of this problem, we are going to answer our questions using this model.

2.5 Part D

Given the choice of model for the covariance from (c), present the estimated coefficients and the corresponding standard errors in a table. Also report estimated variance components.

Since we chose this first model, we can go through and find the corresponding values we want starting with the estimated coefficients.

```
(beta_hat <- coefficients(model1))
```

```
## (Intercept)         age     gender1 age:gender1
##  17.3727273   0.4795455  -1.0321023   0.3048295
```

Here we can see that for our model, $\hat{\beta}_0 =$ 17.3727273, $\hat{\beta}_1 =$ 0.4795455, $\hat{\beta}_2 =$ -1.0321023, and $\hat{\beta}_3 =$ 0.3048295. We can say this estimated model that we had before of $Y_{ij} = \beta_0 + t_{ij}\beta_1 + M_i\beta_2 + t_{ij}M_i\beta_3$ can be estimated by $Y_{ij} =$ 17.3727273 $+t_{ij}$ 0.4795455 $+M_i$ -1.0321023 $+t_{ij}M_i$ 0.3048295.

Now we can find the estimated standard errors. Rather than using the ones that our model gives us, we should really use the ones that can be found by the robust estimators. We will show this below.

```
# Add library for robust estimators
library(clubSandwich)

# Robust covariance matrix of betahat
V.robust <- vcovCR(model1, type = "CR0")

# Robust se
se.robust <- sqrt(diag(V.robust))
round(se.robust, 4)
```

```
## (Intercept)         age     gender1 age:gender1
##      0.7252      0.0631      1.3778      0.1169
```

As we can see from the our output we can see that $\hat{SE}(\hat{\beta}_0) = 0.7252$, $\hat{SE}(\hat{\beta}_1) = 0.0631$, $\hat{SE}(\hat{\beta}_2) = 1.3778$, and $\hat{SE}(\hat{\beta}_3) = 0.1169$.

Lastly, we are going to find the estimated variance components. Please note since we used a GLS with compound symmetry with same variance over age, we know our variance components in the model structure of $\sigma^2 V = \sigma^2 \begin{pmatrix} 1 & p & p & p \\ & 1 & p & p \\ & & 1 & p \\ & & & 1 \end{pmatrix}$ where for V 1 is on the diagonals and p or the correlation is on the upper parts of the matrix above the diagonals of 1's. So we really need to find what σ^2 and p are.

```r
# Find sigma^2
sigma(model1)^2
```

```
## [1] 5.220682
```

```r
# Find the correlation (p or rho)
coef(model1$modelStruct$corStruct, unconstrained = FALSE)[[1]]
```

```
## [1] 0.6318381
```

As we can see here we get that the σ^2 is 5.220682 and that the rho or p is 0.6318381. We can plug these values into $\sigma^2 V = \sigma^2 \begin{pmatrix} 1 & p & p & p \\ & 1 & p & p \\ & & 1 & p \\ & & & 1 \end{pmatrix}$ to get our full variance structure with σ^2 being 5.220682 and rho or p being 0.6318381 as our variance components. Please note below is the full matrix structure of our variance components being combined together.

```r
getVarCov(model1, individual = 1)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 5.2207 3.2986 3.2986 3.2986
## [2,] 3.2986 5.2207 3.2986 3.2986
## [3,] 3.2986 3.2986 5.2207 3.2986
## [4,] 3.2986 3.2986 3.2986 5.2207
##   Standard Deviations: 2.2849 2.2849 2.2849 2.2849
```

2.6 Part E

Using the model in (d), test whether the two mean trends have the same rate of change for the two groups or not.

First we need to figure out what we are testing. Note our model is $Y_{ij} = \beta_0 + t_{ij}\beta_1 + M_i\beta_2 + t_{ij}M_i\beta_3$. In this case when we say rate of change we are testing the difference in slope. The only difference in slope between the groups is the β_3 since it has both the t_{ij} term representing slope and the M_i term representing the difference between the groups. So we are going to do a test on β_3. We are going to do a t-test on β_3 using the estimates and the robust standard error we found in **Part D**. Note our null hypothesis here is that $\beta_3 = 0$ and our alternative hypothesis is that $\beta_3 \neq 0$.

```r
# degrees of freedom
df <- nrow(dental) - length(beta_hat)

# t stats
t.robust <- beta_hat / se.robust

# p-values
p.value <- round(2 * pt(q = abs(t.robust),
                        df = df,
                        lower.tail = FALSE), 4)
# results
tab <- data.frame(beta_hat, se.robust, t.robust, p.value)
knitr::kable(tab)
```

	beta_hat	se.robust	t.robust	p.value
(Intercept)	17.3727273	0.7252063	23.9555672	0.0000
age	0.4795455	0.0631326	7.5958453	0.0000
gender1	-1.0321023	1.3777851	-0.7491025	0.4555
age:gender1	0.3048295	0.1168673	2.6083390	0.0104

When looking at the p-value for β_3 (which on our table is the "*age:gender1*" row), we can see that we get a p-value of 0.0104 which is less than the general α level we use of 0.05. Therefore, we reject our null hypothesis and have statistically significant evidence to conclude that $\beta_3 \neq 0$. Therefore, the two mean trends do **not** have the same rate of change for the two groups.

2.7 Part F

Using the model in (d), test whether the two mean trends are the same for the two groups.

First we need to figure out what we are testing. Note our model is $Y_{ij} = \beta_0 + t_{ij}\beta_1 + M_i\beta_2 + t_{ij}M_i\beta_3$. In this case when we say whether the two mean trends are the same for the two groups, we are looking to see if all the β's that have a M_i term are 0. In this case, the only ones that have this term are β_2 and β_3. So we are going to do a test on β_2 and β_3. We are going to do a F-test (since we are testing both β's at the same time) on β_2 and β_3 using the estimates and the robust standard error we found in **Part D**. Note our null hypothesis here is that $\beta_2 = \beta_3 = 0$ and our alternative hypothesis is that the null is not true.

```r
# L matrix
L <- rbind(c(0, 0, 1, 0), c(0, 0, 0, 1))
cc <- nrow(L)
df <- nrow(dental) - length(beta_hat)

# estimate and covariance matrix of L\beta
est <- L %*% beta_hat
varmat <- L %*% V.robust %*% t(L)

# F-test
Fstat <- c(t(est) %*% solve(varmat) %*% (est)) / cc
p.value <- pf(q = Fstat, df1 = cc, df2 = df, lower.tail=FALSE)
f_tab <- data.frame(Fstat, p.value)
knitr::kable(f_tab)
```

Fstat	p.value
7.057842	0.0013358

We can see that we get a p-value of 0.0013358 which is less than the general α level we use of 0.05. Therefore, we reject our null hypothesis and have statistically significant evidence to conclude that the statement of $\beta_2 = \beta_3 = 0$ is not true. Therefore, the two mean trends are **not** the same for the two groups.