Modeling Longitudinal Data: Heart Study

Yamar_Ba_

INTRODUCTION

The human heart has always been a source of fascination. The heart, like the rest of our organs, with the exception of our spleen, is a crucial part of human physiology. There are endless studies conducted on the heart.

For the purposes of this project we will explore the data of the prospective case-control study involving the effects of diet on pulse rate. With measurements with an emphasis on model selection.

DATA DESCRIPTION

In this data, 30 participants were assigned two different diets at random and three different exercises. The first diet or diet = 1 is a low fat diet, diet = 2 is a non-low fat diet. The first exercise or exercise = 1, is rest, exercise = 2 is walking leisurely and the third, exercise = 3 is running. Their heart heart rate or pulses were taken for each exercise at three different time points during their exercises. Measurements were taken during minute 1, minute 15 and minute 30.

Dataset Fields:

1. **id**: Subject ID
2. **diet**: 1 = low fat; 2 = not low fat
3. **exertype**: 1 = at rest, 2 = walking, 3 = running
4. **pulse**
5. **time**: 1 = 1 min, 2 = 15 min, 3 = 30 min

There is one continuous variable in the data set, “pulse”. The rest of our variables are categorical.

STATISTICAL METHODS

The dataset presented is longitudinal data. As such, there are a few modeling options. When modeling longitudinal data, mean responses profiles and covariance are modeled to give us insight into data. Although both methods are separate they are interchangeable and ultimately yield the same information. For the purposes of this model we will explore both models.

Before choosing a statistical method we can explore the data visually. We will briefly explore the following:

- Time plot
- Plot of means
- Time relative to pulse distribution

We can see from figure 1 below of the time plot that pulse readings of the two treatment groups are wide in variety with a few outliers.

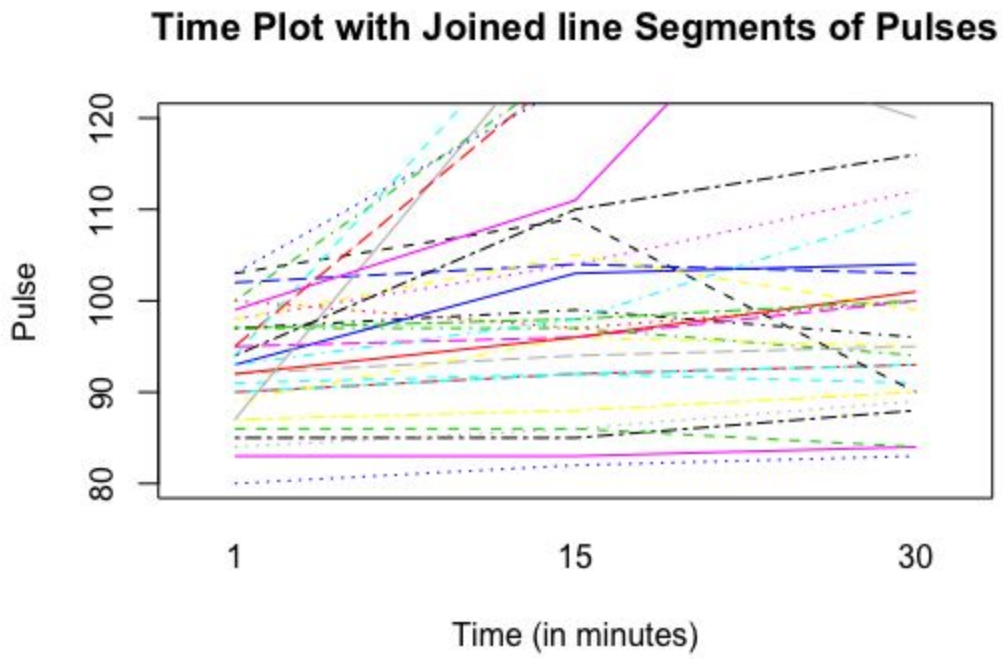

figure 1

The mean pulse rate of the diet group and non-diet group are plotted as described in the model below in figure 2.

The model for the mean for subjects in the control group:

$$E(Y_{ij}) = \beta_1 + \beta * Time_{ij},$$

The model for the mean for subjects in treatment group:

$$E(Y_{ij}) = (\beta_1 + \beta_2) + (\beta_2 * + \beta_4) \; Time_{ij}.$$

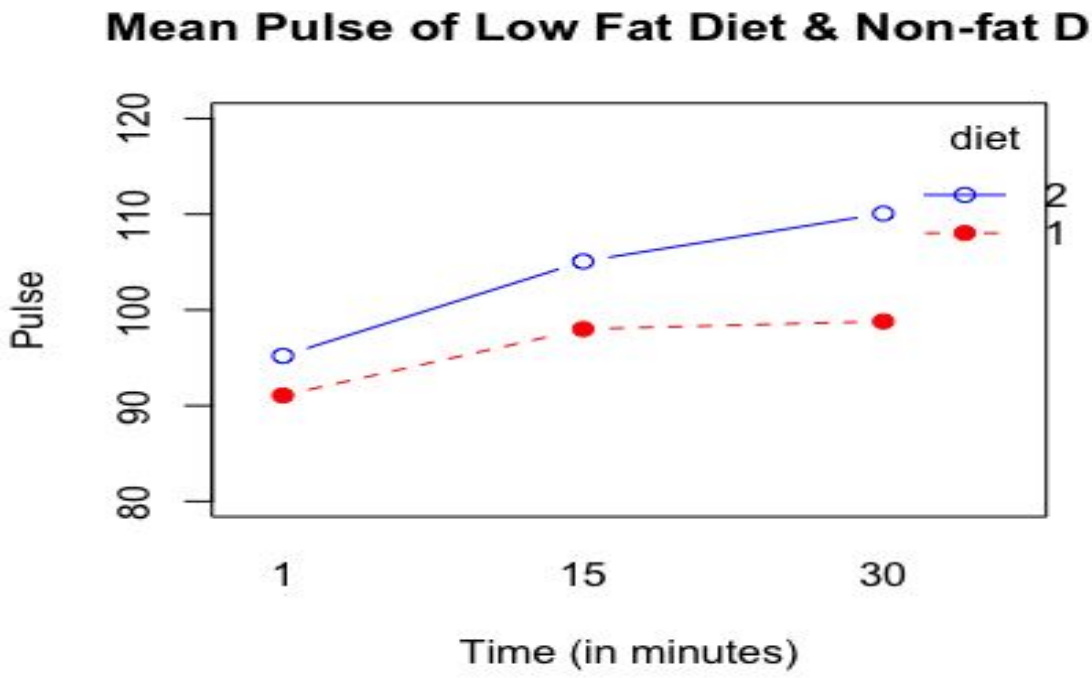

figure 2

As expected, there appears to be a linear association between time and pulse rate in both treatment groups. The pulse rate for the first group or the diet group is significantly lower than that of the non-diet group. Which suggests that there is an association between low fat diet and a lower pulse rate. The mean response of both groups widens over time reaching its peak variance at the last time point. However, both measurements seem to be relatively parallel to one another.

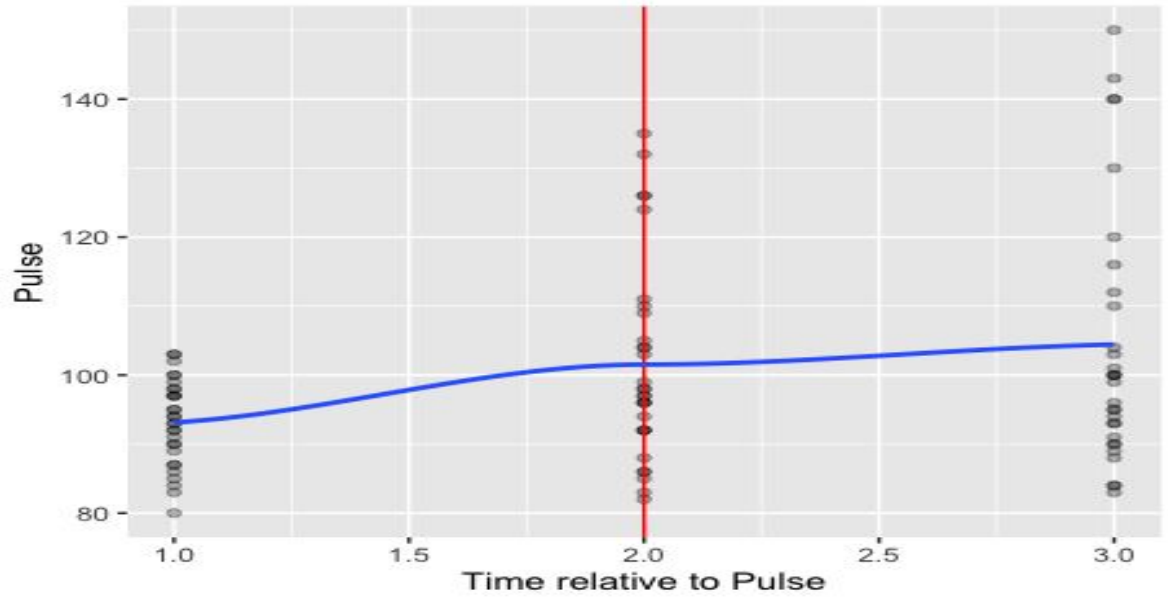

figure 3

To provide another perspective, we can see the distribution of the pulse rates relative to time. As time increases the pulse rates increase in variance and rate.

MEAN RESPONSE PROFILES

To model the mean we will model an analysis of response profiles using the following models for comparison:

1. Analysis of Response Profiles of Pulse Rate
2. Analysis of Response Profiles assuming Equal Mean Response at Baseline

The model can also be expressed Y = XM + E, where Y = (Y1,Y2,...,YN)', X = (X1,X2,...,Xp),M is a p×n matrix of unknown parameters and E = (e1, e2, . . . , en).

Since the time plot appears to be linear, there is no need to model a quadratic trend model or piecewise linear trend model. The models are used when changes in the mean response over time are non-linear and when changes in the mean response over time are non-linear and cannot be approximated with polynomials, respectively.

Analysis of Response Profiles of Pulse Rate

```
> model <- gls(pulse ~ diet*minutes.f, corr=corSymm(form= ~ time | id),
               weights = varIdent(form = ~ 1 | minutes.f))
> summary(model)
Generalized least squares fit by REML
  Model: pulse ~ diet * minutes.f
  Data: NULL
       AIC      BIC    logLik
  647.2436 676.4134 -311.6218
                                :

Coefficients:
                    Value Std.Error   t-value p-value
(Intercept)      86.93333  3.397011 25.591122  0.0000
diet              4.13333  2.148459  1.923860  0.0578
minutes.f15       4.00000  7.200088  0.555549  0.5800
minutes.f30       0.60000  9.482230  0.063276  0.9497
diet:minutes.f15  2.93333  4.553736  0.644160  0.5212
diet:minutes.f30  7.13333  5.997089  1.189466  0.2376

 Correlation:
                 (Intr) diet   mnt.15 mnt.30 dt:.15
diet             -0.949
minutes.f15       0.110 -0.105
minutes.f30       0.159 -0.151  0.795
diet:minutes.f15 -0.105  0.110 -0.949 -0.754
diet:minutes.f30 -0.151  0.159 -0.754 -0.949  0.795

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.0734912 -0.6209181 -0.1602638  0.6628373  2.3670594

Residual standard error: 5.883796
Degrees of freedom: 90 total; 84 residual

> anova(model)
Denom. DF: 84
               numDF  F-value p-value
(Intercept)        1 7615.568  <.0001
diet               1    3.044  0.0847
minutes.f          2    7.756  0.0008
diet:minutes.f     2    0.830  0.4395
```

Analysis of Response Profiles assuming Equal Mean Response at Baseline

```
> model2 <- gls(pulse ~ I(minutes.f==15) + I(minutes.f==30)
                + I(minutes.f==15 & diet=="1") + I(minutes.f==30 & diet=="1"),
                corr=corSymm(form= ~ time | id),
                weights = varIdent(form = ~ 1 | minutes.f))
> summary(model2)
                                    :

Coefficients:
                                          Value Std.Error  t-value p-value
(Intercept)                            93.13333  1.123146 82.92188  0.0000
I(minutes.f == 15)TRUE                  9.38383  3.211096  2.92231  0.0045
I(minutes.f == 30)TRUE                 13.95060  4.216270  3.30875  0.0014
I(minutes.f == 15 & diet == "1")TRUE   -1.96766  4.525988 -0.43475  0.6648
I(minutes.f == 30 & diet == "1")TRUE   -5.30119  5.920994 -0.89532  0.3731

 Correlation:
                                       (Intr) I(.==1 I(.==3 I=1&d="
I(minutes.f == 15)TRUE                  0.082
I(minutes.f == 30)TRUE                  0.118  0.793
I(minutes.f == 15 & diet == "1")TRUE    0.000 -0.705 -0.556
I(minutes.f == 30 & diet == "1")TRUE    0.000 -0.558 -0.702  0.792

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.1349038 -0.5967315 -0.2314556  0.6285504  2.3214342

Residual standard error: 6.151721
Degrees of freedom: 90 total; 85 residual

 anova(model2)
            numDF denDF  F-value p-value
(Intercept)     1    58 7115.696  <.0001
time            1    58   11.636  0.0012
time0           1    58    3.895  0.0532


model7 <- gls(pulse ~ I(minutes.f==15) + I(minutes.f==30)
                + I(minutes.f==15 & diet=="1") + I(minutes.f==30 & diet=="1"),
                corr=corSymm(form= ~ time | id),
                weights = varIdent(form = ~ 1 | minutes.f),method="ML")

model8 <- gls(pulse ~ I(minutes.f==15) + I(minutes.f==30),
                corr=corSymm(form= ~ time | id),
                weights = varIdent(form = ~ 1 | minutes.f),method="ML")
```

```
anova(model7, model8)
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
model7     1 11 669.8673 697.3652 -323.9336
model8     2  9 666.8019 689.3001 -324.4009 1 vs 2 0.9345955  0.6267
```

In comparing the Analysis of Response Profiles of Pulse Rate and the Analysis of Response Profiles assuming Equal Mean Response at Baseline we first see that p-values for Analysis of Response Profiles of Pulse Rate model are not significant. The most significant p-value being its intercept.

Analysis of Response Profiles assuming Equal Mean Response at Baseline have significant p-values concerning its slopes with the exceptions of the conditional time points. When the conditional time points are removed, as in model8 and compared with the original with the conditions, model7, the model8 a better fit with a lower AIC.

COVARIANCE

To provide an alternative perspective, we model the covariance.

We can model the covariances using the following methods:

1. Unstructured Covariance

   $Cov(Y_i) = \sum_i = \sum$

2. Autoregressive Covariance

   $Corr(Y_{ij}, Y_{i+j}) = p^K$, for all j and k and $p \geq 0$.

3. Exponential Covariance

   $Cov(Y_{ij}, Y_{ik}) = \sigma 2 \exp|-\theta |t_{ij} - t_{ik}|)$, for all $p \geq 0$ and where $\{t_{i1,..}t_{in}\}$ -> observation times of i^{th} individual.

All three methods will provide us with the same information but the methods have their own advantages and disadvantages, as well as, parameters. Which is the difference between them.

To elaborate, in Unstructured Covariance, each variance and covariance is estimated individually from the data. This makes the variance and covariance reflect the data more closely, providing a better fit. Autoregressive assumes that variances are constant across occasions. Measurements must be taken at equal intervals of time for this method to work. However, if measurements are not equally spaced, Autoregressive can be generalized by Exponential Covariance.

To demonstrate the covariance model, we will use all three methods.

1) Unstructured Covariance (REML Estimation)

```
> modelX <- gls(y ~ diet.f*mins.f, na.action=na.omit,
               corr=corSymm(form= ~ ttime | id),
              weights = varIdent(form = ~ 1 | ttime))
> summary(modelX)
                                :
Coefficients:
                    Value Std.Error  t-value p-value
(Intercept)      91.06667  1.519189 59.94425  0.0000
diet.f2           4.13333  2.148458  1.92386  0.0595
mins.f2           6.93333  3.219977  2.15322  0.0356
diet.f2:mins.f2   2.93333  4.553736  0.64416  0.5221
                                :
```

2)**Autoregressive Covariance (REML Estimation)**

```
> modelY <- gls(y ~ diet.f*mins.f, na.action=na.omit, corr=corAR1(form= ~ ttime
| id))
> summary(modelY)
                                :
Correlation Structure: AR(1)
 Formula: ~ttime | id
 Parameter estimate(s):
      Phi
0.3545084

Coefficients:
                    Value Std.Error  t-value p-value
(Intercept)      91.06667  2.833950 32.13418  0.0000
diet.f2           4.13333  4.007811  1.03132  0.3068
mins.f2           6.93333  3.219975  2.15323  0.0356
diet.f2:mins.f2   2.93333  4.553732  0.64416  0.5221
                                :
```

3) Exponential Covariance (REML Estimation)

```
> modelZ <- gls(y ~ diet.f*mins.f, na.action=na.omit, corr=corExp(form= ~ mins
| id))
> summary(modelZ)
                                :
Coefficients:
                   Value Std.Error  t-value p-value
(Intercept)     91.06667  2.833949 32.13419  0.0000
diet.f2          4.13333  4.007810  1.03132  0.3068
mins.f2          6.93333  3.219977  2.15322  0.0356
diet.f2:mins.f2  2.93333  4.553736  0.64416  0.5221

 Correlation:
                (Intr) dit.f2 mns.f2
diet.f2         -0.707
mins.f2         -0.568  0.402
diet.f2:mins.f2  0.402 -0.568 -0.707
                                :
```

In comparing the three covariance models we can see that the intercepts for all three are identical. However the p-value for the unstructured covariance slopes is lower than that of lower autoregressive covariance and exponential covariance. The standard error of the unstructured covariance is also significantly lower than that of the other two models which are equal in variance.

In general, we can see that the autoregressive covariance model and the exponential covariance model have identical outputs. The residual standard error for the unstructured covariance is lower at approximately 5.88. Compared to the autoregressive and exponential methods which are at approximemntly 10.98

```
> anova(modelX,modelY)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelX     1  7 423.9647 438.1422 -204.9824
modelY     2  6 446.3108 458.4629 -217.1554 1 vs 2 24.34611  <.0001

> anova(modelX,modelZ)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelX     1  7 423.9647 438.1422 -204.9824
modelZ     2  6 446.3108 458.4629 -217.1554 1 vs 2 24.34611  <.0001

> anova(modelY,modelZ)
       Model df      AIC      BIC    logLik
modelY     1  6 446.3108 458.4629 -217.1554
modelZ     2  6 446.3108 458.4629 -217.1554
```

```
> anova(modelX, modelY, modelZ)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelX     1  7 423.9647 438.1422 -204.9824
modelY     2  6 446.3108 458.4629 -217.1554 1 vs 2 24.34611  <.0001
modelZ     3  6 446.3108 458.4629 -217.1554
```

In comparing the three models pairwise and all together, we can see that the AIC for the unstructured covariance model is significantly lower at 423 that of the autoregressive and exponential models which are equally at 446. The BIC is also significantly lower for the unstructured covariance.

CONCLUSION

When modeling the response profile, the Analysis of Response Profiles assuming Equal Mean Response at Baseline with no conditions, provided the best model with lower p values and AIC/BIC when compared to the same model with conditions. Our final response profile equation without conditions is $\hat{Y} = 93.13333 + 8.4X_1 + 11.3X_2$

The covariance model that provided the best fit with lower p-values and AIC/BIC, was the unstructured covariance model. Our final unstructured covariance regression equation
$\hat{Y} = 91.06667 + 4.13333X_1 + 6.93333X_2 + 2.93333X_3$

Our analysis indicates a positive relationship between diet and pulse rate. An individual's pulse can be predicted based on the above equations.

Reference:

Fitzmaurice, Garrett M., and Laird, Nan M., and Ware, James H.. *Applied Longitudinal Analysis*. John Wiley and Sons, Inc. 20011

DATASET
Titled: exer.csv
Link: **https://web.stanford.edu/class/psych252/data/index.html**