Lab 9 BIS 505b

Maria Ciarleglio

5/3/2021

Goal of Lab 9

In **Lab 8**, we will analyze a **longitudinal endpoint**. We will begin by discussing how to **(1)** transform longitudinal data from wide (one row per subject) to long (multiple rows per subject) format. Next, we **(2)** summarize longitudinal data numerically and graphically. Modeling will focus on **(3)** the analysis of mean response profiles and **(4)** modeling a linear trend over time.

Data Set

In this lab, we will analyze data from n = 30 participants in a diet and exercise study. Subjects were randomly assigned to either a low fat diet or a non-low fat diet. Their pulse rate (beats per minute, BPM) was measured at three time points during exercise: 1 minute, 15 minutes, and 30 minutes. This data set is contained in `exercise.csv` and is imported as the data frame `exercise` in code chunk 3 above. The **Data Key** is provided below. We are interested in modeling pulse rate during exercise and determining if there is a significant difference in pulse rate over time in the two diet groups.

Variable Name	Definition
`id`	Subject ID
`diet`	Diet assigned
	1 = Low fat
	2 = Non-low fat
`time1`	Pulse rate at 1 minute (**Our Response**)
`time2`	Pulse rate at 15 minutes (**Our Response**)
`time3`	Pulse rate at 30 minutes (**Our Response**)

Diet group is our main exposure of interest. We begin by creating a factor version of numerical `diet`.

```
# Factor version of diet variable
exercise$diet_factor <- factor(exercise$diet, levels = 1:2,
                               labels = c("Low fat", "Non-low fat"))
```

Numerical and Graphical Summaries (Wide Data)

The data are collected in **wide format**, which has one row per subject and multiple columns containing the response (pulse rate) at each time point (`time1` , `time2` , `time3`).

```
# Printing rows 1-5 of time variables
print(exercise[1:5, c(1,3:5)], row.names = FALSE)
```

```
##  id time1 time2 time3
##   1    85    85    88
##   2    90    92    93
##   3    97    97    94
##   4    80    82    83
##   5    91    92    91
```

```
# Equivalently,
# print(exercise[,c("id", "time1", "time2", "time3")], row.names = FALSE)
```

Repeated measurements from the same individual tend to be correlated. The `cor()` function returns the **correlation matrix** between repeated measures. Similarly, the `cov()` function returns the **variance-covariance matrix** of repeated measurements.

```
# Correlation matrix of pulse measurements over time
cor(exercise[,3:5])
```

```
##           time1     time2     time3
## time1 1.0000000 0.5445409 0.5191479
## time2 0.5445409 1.0000000 0.8502755
## time3 0.5191479 0.8502755 1.0000000
```

- We observe a *positive correlation* between pulse measurements over time. The correlation decreases slightly with increasing time separation.

```
# Variance-covariance matrix of pulse measurements over time
cov(exercise[,3:5])
```

```
##          time1     time2     time3
## time1 37.84368  48.78851  60.28506
## time2 48.78851 212.11954 233.76092
## time3 60.28506 233.76092 356.32299
```

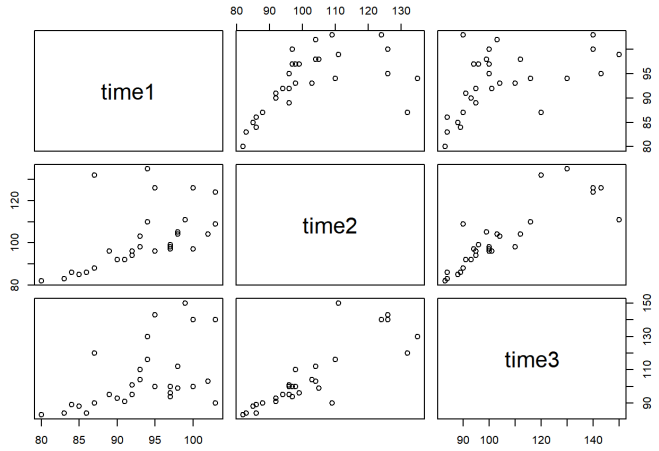
```
# Diagonal elements of v-c matrix are variances
aggregate(cbind(time1, time2, time3) ~ 1, data = exercise,
          FUN = var, na.rm = TRUE)
```

```
##      time1    time2   time3
## 1 37.84368 212.1195 356.323
```

- The *diagonal elements* of the variance-covariance matrix are the variances of the pulse measurements taken at 1, 15, and 30 minutes. We observe an *increase in the variability* of the pulse measurements as time increases. The off-diagonal elements are the pairwise covariances between measurement times.

Pairwise scatterplots between pulse measurements are produced using the `pairs()` function:

```
# Scatterplot matrix
pairs(exercise[,3:5])
```



- We observe a strong positive correlation between measurements over time.

The `aggregate()` function can be used to compute the mean pulse rate at each time point by group:

```
# Mean pulse rate over time by group
aggregate(cbind(time1, time2, time3) ~ diet_factor, data = exercise,
          FUN = mean, na.rm = TRUE)
```

```
##   diet_factor    time1    time2    time3
## 1     Low fat 91.06667  98.0000  98.8000
## 2 Non-low fat 95.20000 105.0667 110.0667
```

- Mean pulse rate tends to increase over time in the two groups. The average pulse rate is higher in those not on a low fat diet. We will plot these means over time in a **mean response profile** plot.

Data Management (Wide to Long)

The functions used to model the longitudinal pulse rate require a single response column. Each subject will then have multiple rows of data. Data from the same subject will have a common subject ID variable and a time variable will indicate when the measurement was taken. To transform the data from **wide format** to **long format**, we can use the `reshape()` function.

`reshape()` Function Arguments	Option Definition
`varying=`	Variable names containing measurements over time (wide format)
`v.names=`	New response variable name (long format)
`timevar=`	New time variable name (long format), consecutive numerical values for later modeling
`times=`	Consecutive numerical values representing time ordering
`idvar=`	ID variable (wide format)
`direction=`	Reshaping to `="long"` or `="wide"`

```
# Reshaping wide data to long data
exlong <- reshape(exercise, varying = c("time1", "time2", "time3"),
                  v.names = "pulse",
                  timevar = "time.num",
                  times = 1:3,
                  idvar = "id",
                  direction = "long")

# Sort data by id then by time
exlong <- exlong[order(exlong$id, exlong$time.num),]

# Clear row names
row.names(exlong) <- NULL

# "mins" represent actual value of time (in minutes),
# ... Numerical version of "mins" used in linear trend model
exlong$mins[exlong$time.num == 1] <- 1
exlong$mins[exlong$time.num == 2] <- 15
exlong$mins[exlong$time.num == 3] <- 30

# ... Factor version of "mins" used in response profile model
exlong$mins_factor <- factor(exlong$mins)

print(head(exlong, n = 10), row.names = FALSE)
```

```
##  id diet diet_factor time.num pulse mins mins_factor
##   1    1     Low fat        1    85    1           1
##   1    1     Low fat        2    85   15          15
##   1    1     Low fat        3    88   30          30
##   2    1     Low fat        1    90    1           1
##   2    1     Low fat        2    92   15          15
##   2    1     Low fat        3    93   30          30
##   3    1     Low fat        1    97    1           1
##   3    1     Low fat        2    97   15          15
##   3    1     Low fat        3    94   30          30
##   4    1     Low fat        1    80    1           1
```

`pulse` is the response variable nested within subject `id` taken at different points in time. Notice that we created several versions of "time" above.

- `time.num` is a numeric (integer) variable (`1:3`) that indicates the relative ordering of the measurements taken over time. This variable is necessary for future modeling that will be performed.
- `mins` is a numeric variable (`1` , `15` , `30`) that equals time in minutes. This numeric version of time will be used in modeling the linear trend of the pulse rate over time.
- `mins_factor` is a factor version of the `mins` variable that will be used in the response profile analysis.

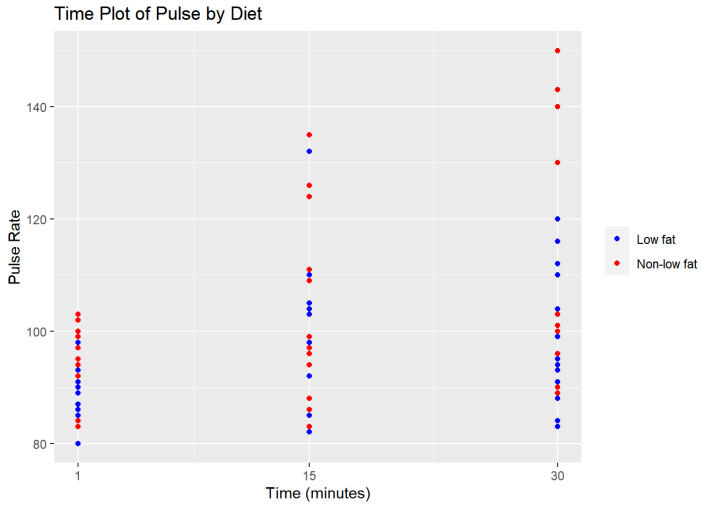
Numerical and Graphical Summaries (Long Data)

We are interested in describing the trend in the response over time. Plots allow us to visualize longitudinal data response trajectory. Three commonly-used plots are:

1. **Time plot**: Scatterplot of points over time
2. **Spaghetti plot**: Individual subject trajectories over time
3. **Mean response profile plot**: Mean response plotted over time

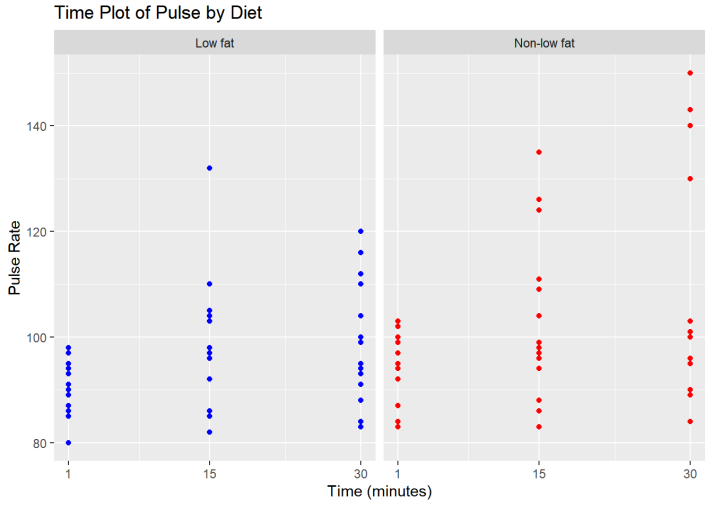
The **time plot** is essentially a scatterplot of the response `pulse` (y) over values of time `mins` (x). We can use different point colors for the two levels of `diet` :

```
# Time plot of pulse
ggplot(data = exlong, aes(x = mins, y = pulse, col = diet_factor)) +  # point colors by diet
  geom_point() +
  labs(title = "Time Plot of Pulse by Diet",                          # plot title
       x = "Time (minutes)", y = "Pulse Rate") +                      # axis labels
  scale_color_manual(values = c("blue", "red")) +                     # manually setting point colors
  theme(legend.title = element_blank()) +                             # no legend title, set legend position
  scale_x_continuous(breaks = c(1, 15, 30))                           # x-axis tick marks at values of time
```



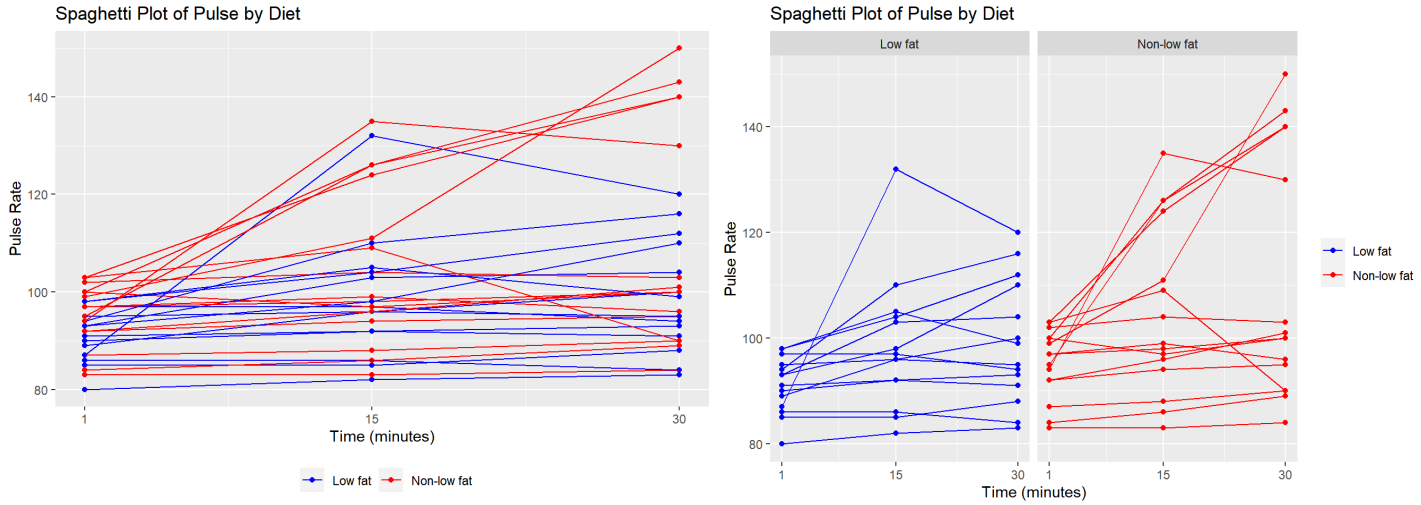
Adding a `facet_grid()` layer creates two separate subplots for those on the low fat diet and those on the non-low fat diet:

```
# Time plot of pulse, subplots by diet
ggplot(data = exlong, aes(x = mins, y = pulse, col = diet_factor)) +
  geom_point() +
  labs(title = "Time Plot of Pulse by Diet",
       x = "Time (minutes)", y = "Pulse Rate") +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.position = "none") +                                   # legend unnecessary
  scale_x_continuous(breaks = c(1, 15, 30)) +
  facet_grid(. ~ diet_factor)                                         # panel for each diet
```



A problem with the time plot is that we do not know which y-values are taken from the same individual. A **spaghetti plot** addresses this by using lines to connect points from the same individual.

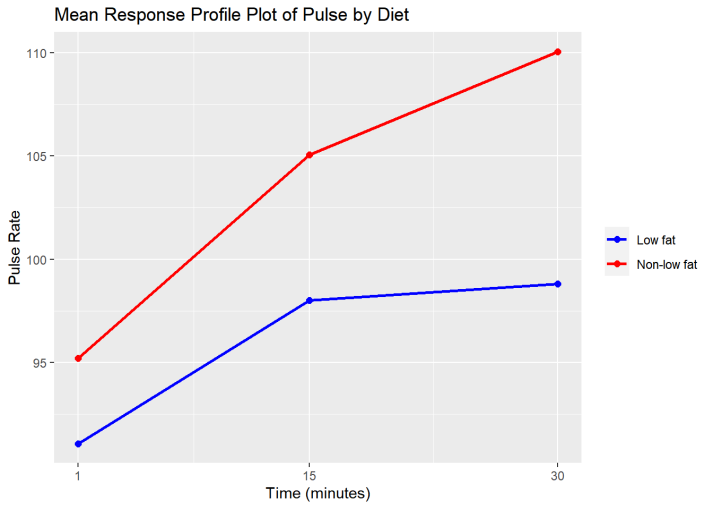
```
# Spaghetti plot of pulse
ggplot(data = exlong, aes(x = mins, y = pulse, col = diet_factor)) +
  geom_point() +
  geom_line(aes(group = id)) +                                    # lines for each subject id
  labs(title = "Spaghetti Plot of Pulse by Diet",
       x = "Time (minutes)", y = "Pulse Rate") +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.title = element_blank(), legend.position = "bottom") +
  scale_x_continuous(breaks = c(1, 15, 30))

# Spaghetti plot of pulse, subplots by diet
ggplot(data = exlong, aes(x = mins, y = pulse, col = diet_factor)) +
  geom_point() +
  geom_line(aes(group = id)) +
  labs(title = "Spaghetti Plot of Pulse by Diet",
       x = "Time (minutes)", y = "Pulse Rate") +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.title = element_blank()) +
  scale_x_continuous(breaks = c(1, 15, 30)) +
  facet_grid(. ~ diet_factor)                                     # panel for each diet
```



The spaghetti plots show us that most individuals experience a slow increase in the pulse rate while exercising over time. A few individuals experience an increase and then a decrease, and some experience a sharp increase. The non-low fat diet group has some subjects with a rapid increase in pulse over time.

Finally, the **mean response profile plot** visualizes the mean response at each time point by group and allows us to see the average trend over time. The `stat_summary()` function can be used in `ggplot()` to compute the mean of the response using the `fun=mean` argument:

```
# Mean response profile of pulse
ggplot(data = exlong, aes(x = mins, y = pulse, col = diet_factor)) +
  stat_summary(geom = "line", fun = mean, size = 1) +             # lines connecting means
  stat_summary(geom = "point", fun = mean, shape = 19, size = 2) +  # points at means
  labs(title = "Mean Response Profile Plot of Pulse by Diet",
       x = "Time (minutes)", y = "Pulse Rate") +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.title = element_blank()) +
  scale_x_continuous(breaks = c(1, 15, 30))
```



On average, those in the non-low fat diet group have a higher pulse rate while exercising compared to those in the low fat diet group. Through our modeling, we are interested in determining if there is a significant difference in the pulse rate over time in these two groups.

Analysis of Response Profiles

We will construct a linear regression model that accounts for the **correlated nature** of these data (i.e., repeated measures over time within subject are not independent). The main focus of this analysis is to determine if the changes in pulse rates are the same in the two diet groups. This hypothesis focuses on the statistical significance of the **interaction** of diet and time. This analysis does not impose a parametric (e.g., linear, quadratic) trend in the response over time. Rather, we include the **categorical** or **factor** version of time in this model (`mins_factor`). The interaction model includes the main effect of diet (`diet_factor`), the main effect of time (`mins_factor`) and the interaction of diet and time (`diet_factor:mins_factor`):

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_1 \,\text{NonLFDiet} + \beta_2 \,\text{Mins15} + \beta_3 \,\text{Mins30} + \beta_4 \,\text{NonLFDiet*Mins15} + \beta_5 \,\text{NonLFDiet*Mins30}$$

The low fat diet group and 1 minute (baseline reading) are the reference categories in the model above.

The `gls()` function in the `nlme` package allows us to account for **correlated errors** and also allows us to account for **unequal variances** of the response over levels of a covariate (i.e., time). As we saw in the correlation and variance-covariance matrices, there is a positive correlation between repeated measurements taken over time and the variance of the pulse rate tends to increase as time increases. We can also observe this increase in variability in the time plot and the spaghetti plot. The `gls()` function fits models that are analogous to the linear models fit using `lm()` to model our quantitative response of pulse rate.

There are two main arguments in the `gls()` function that allow us to construct the variance-covariance matrix that will be used to account for correlated errors and non-constant variance in the response (**heteroscedasticity**) over our three measurement times,

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

`gls()` Function Arguments	Option Definition
`formula=`	`analysis_variable ~ predictor_variable1 + predictor_variable2`
`data=`	Data frame containing sample data
`correlation=`	Specify form of the covariance matrix
`weights=`	Specify the structure of the heteroscedasticity

- The `correlation=` argument is used to construct the **variance-covariance matrix** of the linear model. We will assume an **unstructured** symmetric covariance matrix that allows each element of the matrix to freely equal any non-negative value that is suggested by the data. We will assume: `correlation = corSymm(form = ~ time.num | id)`
    - `corSymm()` assumes a general correlation structure (i.e., unstructured).
    - `(form = ~ time.num | id)` states the correlation between observation times is assumed to be different and that the responses over time are correlated within the same `id`. `id` is our *grouping variable*. That is, repeated measures are nested within subject id. Note that time here must be represented as a numerical sequence of consecutive integers (`time.num` which equals `1`, `2`, `3` at the three measurement times).
- The `weight=` argument is used to model **heteroscedasticity**, or the dependence of the variance on certain variables. In our case, the variance of the response is expected to vary over the three time points in this study. We will assume: `weights = varIdent(form = ~ 1 | mins_factor)`.
    - `varIdent()` allows different variances according to the levels of a classification factor (`mins_factor`).
    - `(form = ~ 1 | mins_factor)` specifies that the grouping variable is our factor version of time. Here we indicate that observations observed at the same time have common variance.

Interaction Model

Note that the model syntax `diet_factor*week_factor` below automatically includes the main effects of diet (`diet_factor`) and time (`week_factor`) in addition to their interaction. Thus, the main effects do not need to be listed on the right hand side of our model formula in the `gls()` function:

```
# Response Profile Model 1 (with interaction)
mod.mrp1 <- gls(pulse ~ diet_factor*mins_factor, data = exlong,
                correlation = corSymm(form = ~ time.num | id),   # numeric time sequence of consecutive integ
ers
                weights = varIdent(form = ~ 1 | mins_factor))

summary(mod.mrp1)
```

```
## Generalized least squares fit by REML
##   Model: pulse ~ diet_factor * mins_factor
##   Data: exlong
##        AIC      BIC    logLik
##   647.2436 676.4134 -311.6218
##
## Correlation Structure: General
##  Formula: ~time.num | id
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.505
## 3 0.464 0.840
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | mins_factor
##  Parameter estimates:
##        1       15       30
## 1.000000 2.441249 3.110969
##
## Coefficients:
##                                       Value Std.Error  t-value p-value
## (Intercept)                         91.06667  1.519190 59.94424  0.0000
## diet_factorNon-low fat               4.13333  2.148459  1.92386  0.0578
## mins_factor15                        6.93333  3.219977  2.15322  0.0342
## mins_factor30                        7.73333  4.240582  1.82365  0.0718
## diet_factorNon-low fat:mins_factor15  2.93333  4.553736  0.64416  0.5212
## diet_factorNon-low fat:mins_factor30  7.13333  5.997089  1.18947  0.2376
##
##  Correlation:
##                                     (Intr) dt_N-f mns_15 mns_30 d_N-f:_1
## diet_factorNon-low fat              -0.707
## mins_factor15                        0.110 -0.078
## mins_factor30                        0.159 -0.112  0.795
## diet_factorNon-low fat:mins_factor15 -0.078  0.110 -0.707 -0.562
## diet_factorNon-low fat:mins_factor30 -0.112  0.159 -0.562 -0.707  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
```

- The **fitted model** is given by the equation, $\hat{y} = 91.067 + 4.133$ NonLFDiet $+ 6.933$ Mins15 $+ 7.733$ Mins30 $+ 2.933$ NonLFDiet$\times$Mins15 $+ 7.133$ NonLFDiet$\times$Mins30
- The estimated **intercept** a = 91.067 equals the estimated mean pulse rate in the low fat diet group at 1 minute.
- The estimated **slope** associated with the **main effect of diet group** b_1 = 4.133 equals the difference in mean pulse rate in the non-low fat diet group vs. the low fat diet group (reference) at 1 minute ($\bar{y}_{nonLF,1} - \bar{y}_{LF,1}$).
- The estimated **slope** associated with the **dummy variable for 15-minutes** b_2 = 6.933 equals the mean change in pulse rate between 15 minutes and 1 minute in low fat diet group ($\bar{y}_{LF,15} - \bar{y}_{LF,1}$).
- The estimated **slope** associated with the **dummy variable for 30-minutes** b_3 = 7.733 equals the mean change in pulse rate between 30 minutes and 1 minute in low fat diet group ($\bar{y}_{LF,30} - \bar{y}_{LF,1}$).
- The estimated **slope** associated with the **non-low fat diet** $\times$ **15-minute interaction** b_4 = 2.933 equals the difference in the mean change from baseline (1 minute) at 15 minutes in the non-low fat diet group vs. the low fat diet group ($[\bar{y}_{nonLF,15} - \bar{y}_{nonLF,1}] - [\bar{y}_{LF,15} - \bar{y}_{LF,1}]$). Based on the t-test of the slope parameter $H_0 : \beta_4 = 0$ vs. $H_1 : \beta_4 \neq 0$,

there is no significant difference in the change from baseline at 15 minutes in the two diet groups (p-value = 0.521).

- The estimated **slope** associated with the **non-low fat diet** $\times$ **30-minute interaction** b_5 = 7.133 equals the difference in the mean change from baseline (1 minute) at 30 minutes in the non-low fat diet group vs. the low fat diet group ($[\bar{y}_{nonLF,30} - \bar{y}_{nonLF,1}] - [\bar{y}_{LF,30} - \bar{y}_{LF,1}]$). Again, there is no significant difference in the change from baseline at 30 minutes in the two diet groups (p-value = 0.238).
- Under $H_0 : \beta_4 = \beta_5 = 0$, the trajectories are parallel in the two diet groups. We can perform an F-test to test the overall significance of the interaction to simultaneously test $H_0 : \beta_4 = \beta_5 = 0$ vs. H_1 : At least one $\beta_4, \beta_5 \neq 0$ using the `anova()` function on our model object `mod.mrp1` .

```
# F-test of interaction diet_factor:mins_factor (beta4 and beta5)
anova(mod.mrp1, type = "marginal")
```

```
## Denom. DF: 84
##                          numDF  F-value p-value
## (Intercept)                  1 3593.311  <.0001
## diet_factor                  1    3.701  0.0578
## mins_factor                  2    2.335  0.1030
## diet_factor:mins_factor      2    0.830  0.4395
```

- We fail to reject $H_0 : \beta_4 = \beta_5 = 0$ (p-value = 0.44). Thus, there is not a significant difference in the effect of diet over time. We cannot reject the null hypothesis that the mean response profiles in the two groups are parallel.

The model above is called a **saturated model** because it estimates 6 parameters and there are 6 group means to be estimated (mean pulse rate in the low fat diet group and in the non-low fat diet group at times 1, 15, and 30 minutes). The `predict()` function is used to estimate **fitted values** from the model. We see that our model's **predicted** or **fitted values** $\hat{y}$ of mean pulse rate for each combination of diet and time are equal to the raw mean values:

```
# Values of x used to predict mean of y
pred.x <- expand.grid(diet_factor = levels(exlong$diet_factor),
                      mins_factor = levels(exlong$mins_factor))

# Fitted values of y (mean pulse rate) for given values of x in "pred.x"
predicted <- predict(mod.mrp1, newdata = pred.x, type = "response")

pred.df <- data.frame(pred.x, predicted)

# Printing fitted values
print(pred.df, digits = 5, row.names = FALSE)
```

```
##  diet_factor mins_factor predicted
##      Low fat           1    91.067
##  Non-low fat           1    95.200
##      Low fat          15    98.000
##  Non-low fat          15   105.067
##      Low fat          30    98.800
##  Non-low fat          30   110.067
```

```
# Observed mean values
aggregate(cbind(time1, time2, time3) ~ diet_factor, data = exercise,
          FUN = mean, na.rm = TRUE)
```

```
##   diet_factor    time1    time2    time3
## 1     Low fat 91.06667  98.0000  98.8000
## 2 Non-low fat 95.20000 105.0667 110.0667
```

Main Effects Model

Removing the interaction term, we fit a model that includes only the main effects of diet (`diet_factor`) and time (`week_factor`),

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_1 \,\text{NonLFDiet} + \beta_2 \,\text{Mins15} + \beta_3 \,\text{Mins30}$$

```
# Response Profile Model 2 (without interaction)
mod.mrp2 <- gls(pulse ~ diet_factor + mins_factor, data = exlong,
                correlation = corSymm(form = ~ time.num | id),
                weights = varIdent(form = ~ 1 | mins_factor))

summary(mod.mrp2)
```

```
## Generalized least squares fit by REML
##   Model: pulse ~ diet_factor + mins_factor
##   Data: exlong
##        AIC      BIC    logLik
##   654.1714 678.7149 -317.0857
##
## Correlation Structure: General
##  Formula: ~time.num | id
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.510
## 3 0.463 0.841
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | mins_factor
##  Parameter estimates:
##        1       15       30
## 1.000000 2.422702 3.130627
##
## Coefficients:
##                           Value Std.Error  t-value p-value
## (Intercept)            91.28355  1.509513 60.47219  0.0000
## diet_factorNon-low fat  3.69956  2.120444  1.74471  0.0846
## mins_factor15           8.40000  2.253781  3.72707  0.0003
## mins_factor30          11.30000  3.019923  3.74182  0.0003
##
##  Correlation:
##                        (Intr) dt_N-f mns_15
## diet_factorNon-low fat -0.702
## mins_factor15           0.080  0.000
## mins_factor30           0.114  0.000  0.796
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.0361095 -0.5988226 -0.2062959  0.6071614  2.3727351
##
## Residual standard error: 5.8853
## Degrees of freedom: 90 total; 86 residual
```

- The **fitted model** is given by the equation, $\hat{y}$ = 91.284 + 3.7 NonLFDiet + 8.4 Mins15 + 11.3 Mins30
- The estimated **intercept** a = 91.284 equals the estimated mean pulse rate in the non-low fat diet group at 1 minute.
- The estimated **slope** associated with the **main effect of diet group** b_1 = 3.7 equals the adjusted difference in mean pulse rate the non-low fat diet group vs. the low fat diet group, holding time constant. That is, the non-low fat diet group has an average pulse that is 3.7 BPM greater than the low fat diet group. A significance test of this slope parameter $H_0 : \beta_1 = 0$

vs. $H_1 : \beta_1 \neq 0$ indicates there is not a significant difference in the mean pulse in the non-low fat diet group vs. the low fat diet group (p-value = 0.085).

- The estimated **slope** associated with the **dummy variable for 15-minutes** b_2 = 8.4 equals the diet-adjusted mean change in pulse rate between 15 minutes and baseline. The average pulse rate at 15 minutes is 8.4 BPM higher than at baseline, controlling for diet. A significance test of this slope parameter $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ indicates there is a significant diet-adjusted difference in the mean pulse at 15 minutes vs. baseline (p-value <.001).
- The estimated **slope** associated with the **dummy variable for 30-minutes** b_3 = 11.3 equals the diet-adjusted mean change in pulse rate between 30 minutes and baseline. The average pulse rate at 30 minutes is 11.3 BPM higher than at baseline, controlling for diet. A significance test of this slope parameter $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$ indicates there is a significant diet-adjusted difference in the mean pulse at 30 minutes vs. baseline (p-value <.001).

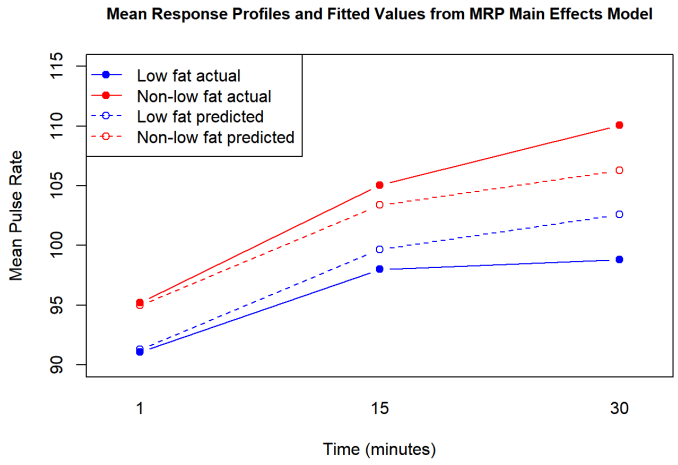
```
# Clearing previous predicted values
rm(pred.x, predicted, pred.df)

# Values of x used to predict mean of y
pred.x <- expand.grid(diet_factor = levels(exlong$diet_factor),
                      mins_factor = levels(exlong$mins_factor))

# Fitted values of y (mean pulse rate) for given values of x in "pred.x"
predicted <- predict(mod.mrp2, newdata = pred.x, type = "response")

pred.df <- data.frame(pred.x, predicted)

# Printing fitted values
print(pred.df, digits = 5, row.names = FALSE)
```

```
##  diet_factor mins_factor predicted
##      Low fat           1    91.284
##  Non-low fat           1    94.983
##      Low fat          15    99.684
##  Non-low fat          15   103.383
##      Low fat          30   102.584
##  Non-low fat          30   106.283
```

- Notice that the difference between the mean predicted pulse rate in the non-low fat diet group vs. the low fat diet group at all time points is equal to the estimated slope of `diet_factor` from our model, b_1 = 3.7 (e.g., 94.98 - 91.28).
- Similarly, the difference in the mean predicted pulse rate at 15 minutes vs. baseline is equal to the estimated slope of the 15-minute dummy variable of `time_factor` , b_2 = 8.4, and the difference in the mean predicted pulse rate at 30 minutes vs. baseline is equal to the estimated slope of the 30-minute dummy variable of `time_factor` , b_3 = 11.3.

We can add the model fitted values to the mean response profile plot to see how the observed and predicted values compare. The `interaction.plot()` function in **R** allows us plot the mean of the response for two-way combinations of factors (illustrating possible interactions). We can add points and lines to the plot using the `points()` function. The `type="b"` argument adds both points and lines connecting the plotted points. Below, we add the fitted means for the low fat diet group and the non-low fat diet group separately so that we can specify different colors for the two sets of fitted values.

```
# Picking out fitted values for each diet group (for plotting)
pred.lf <- subset(pred.df, diet_factor == "Low fat")
pred.nlf <- subset(pred.df, diet_factor == "Non-low fat")

# Base R plotting of mean response profile
interaction.plot(exlong$mins_factor,        # x-axis variable
                 exlong$diet_factor,        # lines by diet
                 exlong$pulse,              # y-axis variable
                 fun = "mean",              # summary statistic plotted for response
                 type = "b", pch = 19, lty = 1,
                 xlab = "Time (minutes)", ylab = "Mean Pulse Rate",
                 col = c("blue", "red"),
                 ylim = c(90, 115),
                 legend = FALSE)
title("Mean Response Profiles and Fitted Values from MRP Main Effects Model", cex.main = .95)

# Adding fitted means from MRP main effects model
points(factor(pred.lf$mins_factor), pred.lf$predicted, col = "blue", type = "b", lty = 2)
points(factor(pred.nlf$mins_factor), pred.nlf$predicted, col = "red", type = "b", lty = 2)

# Adding legend
legend("topleft",
       c("Low fat actual", "Non-low fat actual", "Low fat predicted", "Non-low fat predicted"),
       lty = c(1, 1, 2, 2), pch = c(19, 19, 1, 1), col = rep(c("blue", "red"),2))
```



- Notice how the fitted profiles are parallel. The constant difference between the two curves is the estimated slope of `diet_factor` . Since we are not including an interaction term in this model, we are not allowing the effect of diet to differ by time.

Linear Trend over Time

To model a linear trend over time, we use the numeric version of time (`mins`) in the model rather than `mins_factor` .

Interaction Model

The interaction model includes the main effect of diet (`diet_factor`), the main effect of (linear) time (`mins`), and the interaction of diet and time (`diet_factor:mins`):

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_1 \,\text{NonLFDiet} + \beta_2 \,\text{Minutes} + \beta_3 \,\text{NonLFDiet*Minutes}$$

The equation of the line in the *non-low fat diet group* is given by:

$$E(Y|x) = \mu_{y|x} = (\alpha + \beta_1) + (\beta_2 + \beta_3)\,\text{Minutes}$$

The equation of the line in the *low fat diet (reference) group* is given by:

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_2\,\text{Minutes}$$

The interaction term β_3 allows the slopes of the two lines to differ.

```
# Linear Trend Model 1 (with interaction)
mod.lin1 <- gls(pulse ~ diet_factor*mins, data = exlong,
                correlation = corSymm(form = ~ time.num | id),
                weights = varIdent(form = ~ 1 | mins))

summary(mod.lin1)
```

```
## Generalized least squares fit by REML
##   Model: pulse ~ diet_factor * mins
##   Data: exlong
##        AIC      BIC    logLik
##   667.3987 691.9422 -323.6994
##
## Correlation Structure: General
##  Formula: ~time.num | id
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.497
## 3 0.463 0.835
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | mins
##  Parameter estimates:
##        1       15       30
## 1.000000 2.485041 3.117276
##
## Coefficients:
##                               Value Std.Error  t-value p-value
## (Intercept)                90.82304 1.5028476 60.43397  0.0000
## diet_factorNon-low fat      3.88368 2.1253474  1.82732  0.0711
## mins                        0.20811 0.1414731  1.47103  0.1449
## diet_factorNon-low fat:mins  0.25532 0.2000732  1.27611  0.2054
##
##  Correlation:
##                             (Intr) dt_N-f mins
## diet_factorNon-low fat      -0.707
## mins                         0.068 -0.048
## diet_factorNon-low fat:mins -0.048  0.068 -0.707
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.0684029 -0.5350308 -0.0707899  0.7433586  2.6026790
##
## Residual standard error: 5.883839
## Degrees of freedom: 90 total; 86 residual
```

- The **fitted model** is given by the equation, $\hat{y} = 90.823 + 3.884$ NonLFDiet $+ 0.208$ Minutes $+ 0.255$ NonLFDiet$\times$Minutes
- The estimated **intercept** a = 90.823 equals the mean pulse rate in the low fat diet group at 0 minutes.
- The estimated **slope** associated with the **main effect of diet group** b_1 = 3.884 equals the difference in mean pulse rate in the non-low fat diet group vs. the low fat diet group (reference) at 0 minutes.

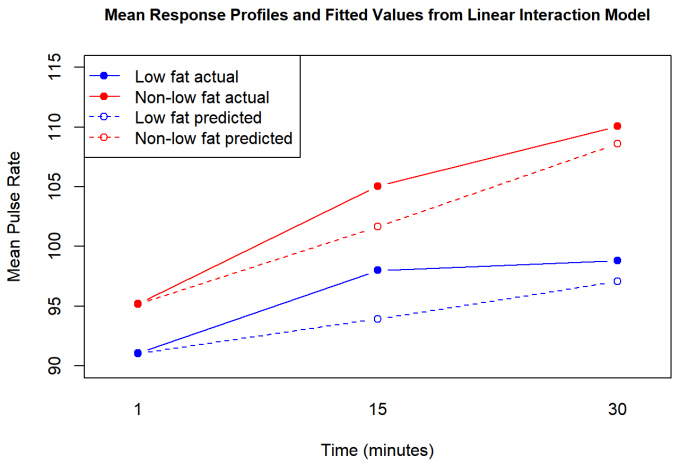
- The estimated **slope** associated with **quantitative time** b_2 = 0.208 equals the slope of the pulse rate linear trajectory in the low fat diet (reference) group. Pulse is increasing at a rate of b_2 beats per minute in this group.
- The estimated **slope** associated with the **non-low fat diet** $\times$ **quantitative time interaction** b_3 = 0.255 equals the difference in the slope in the non-low fat diet group vs. the low fat diet group. That is, the slope in the non-low fat diet group is b_3 units steeper than in the low fat diet group. Based on the t-test of the slope parameter $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$, there is no significant difference in the slopes in these two groups; there is no significant difference in the effect of diet over time (p-value = 0.205).

```
# Clearing previous predicted values
rm(pred.x, predicted, pred.df, pred.lf, pred.nlf)

# Values of x used to predict mean of y
pred.x <- expand.grid(diet_factor = levels(exlong$diet_factor),
                      mins = c(1, 15, 30))

# Fitted values of y (mean pulse rate) for given values of x in "pred.x"
predicted <- predict(mod.lin1, newdata = pred.x, type = "response")

pred.df <- data.frame(pred.x, predicted)

# Printing fitted values
print(pred.df, digits = 5, row.names = FALSE)
```

```
##  diet_factor mins predicted
##      Low fat    1    91.031
##  Non-low fat    1    95.170
##      Low fat   15    93.945
##  Non-low fat   15   101.658
##      Low fat   30    97.066
##  Non-low fat   30   108.610
```

Plotting the fitted lines from this linear interaction model:

```
# Picking out fitted values for each diet group (for plotting)
pred.lf <- subset(pred.df, diet_factor == "Low fat")
pred.nlf <- subset(pred.df, diet_factor == "Non-low fat")

# Base R plotting of mean response profile
interaction.plot(exlong$mins_factor,        # x-axis variable
                 exlong$diet_factor,        # lines by diet
                 exlong$pulse,              # y-axis variable
                 fun = "mean",              # summary statistic plotted for response
                 type = "b", pch = 19, lty = 1,
                 xlab = "Time (minutes)", ylab = "Mean Pulse Rate",
                 col = c("blue", "red"),
                 ylim = c(90, 115),
                 legend = FALSE)
title("Mean Response Profiles and Fitted Values from Linear Interaction Model", cex.main = .95)

# Adding fitted means from linear interaction model
points(factor(pred.lf$mins), pred.lf$predicted, col = "blue", type = "b", lty = 2)
points(factor(pred.nlf$mins), pred.nlf$predicted, col = "red", type = "b", lty = 2)

# Adding legend
legend("topleft",
       c("Low fat actual", "Non-low fat actual", "Low fat predicted", "Non-low fat predicted"),
       lty = c(1, 1, 2, 2), pch = c(19, 19, 1, 1), col = rep(c("blue", "red"),2))
```



- We see that the slope of the fitted line in the non-low fat diet group is indeed steeper than the fitted line in the low fat diet group. However, the test of the interaction does not suggest there is a significant difference in these two slopes.

Main Effects Model

Removing the interaction term gives the main effects model that includes the main effect of diet (`diet_factor`) and the main effect of (linear) time (`mins`). This model assumes a common slope (common effect of time), but allows for the two lines to be shifted by the effect of diet.

$$E(Y|x) = \mu_{y|x} = \alpha + \beta_1 \, \text{NonLFDiet} + \beta_2 \, \text{Minutes}$$

```
# Linear Trend Model 2 (without interaction)
mod.lin2 <- gls(pulse ~ diet_factor + mins, data = exlong,
                correlation = corSymm(form = ~ time.num | id),
                weights = varIdent(form = ~ 1 | mins))

summary(mod.lin2)
```

```
## Generalized least squares fit by REML
##   Model: pulse ~ diet_factor + mins
##   Data: exlong
##        AIC      BIC    logLik
##   665.6384 687.8316 -323.8192
##
## Correlation Structure: General
##  Formula: ~time.num | id
##  Parameter estimate(s):
##  Correlation:
##   1     2
## 2 0.498
## 3 0.463 0.834
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | mins
##  Parameter estimates:
##        1       15       30
## 1.000000 2.485072 3.139061
##
## Coefficients:
##                           Value Std.Error  t-value p-value
## (Intercept)            90.91629 1.5011515 60.56437  0.0000
## diet_factorNon-low fat  3.69959 2.1204436  1.74472  0.0846
## mins                    0.33744 0.1011115  3.33735  0.0012
##
##  Correlation:
##                         (Intr) dt_N-f
## diet_factorNon-low fat -0.706
## mins                    0.049  0.000
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.03103058 -0.50180929 -0.07520676  0.67504152  2.46295993
##
## Residual standard error: 5.885348
## Degrees of freedom: 90 total; 87 residual
```

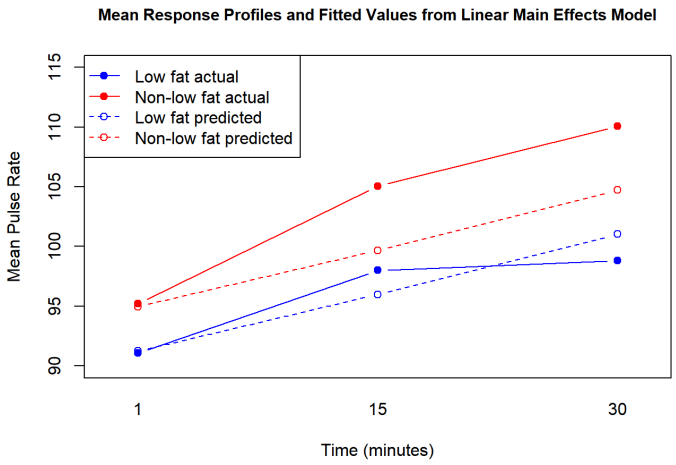
- The **fitted model** is given by the equation, $\hat{y} = 90.916 + 3.7$ NonLFDiet $+ 0.337$ Minutes
- The estimated **intercept** a = 90.916 equals the mean pulse rate in the low fat diet group at 0 minutes.
- The estimated **slope** associated with the **main effect of diet group** b_1 = 3.7 equals the difference in mean pulse rate in the non-low fat diet group vs. the low fat diet group, holding time constant. The non-low fat diet group has an average pulse that is 3.7 BPM greater than the low fat diet group. A significance test of this slope parameter $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ indicates there is not a significant difference in the mean pulse in the non-low fat diet group vs. the low fat diet group (p-value = 0.085).
- The estimated **slope** associated with **quantitative time** b_2 = 0.337 equals the slope of the pulse rate linear trajectory. Pulse is increasing at a rate of b_2 beats per minute. A significance test of this slope parameter $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ indicates there is a significant association between pulse rate and time; pulse rate is increasing significantly over time at a rate of 0.337 beats per minute (p-value = 0.001).

```
# Clearing previous predicted values
rm(pred.x, predicted, pred.df, pred.lf, pred.nlf)

# Values of x used to predict mean of y
pred.x <- expand.grid(diet_factor = levels(exlong$diet_factor),
                      mins = c(1, 15, 30))

# Fitted values of y (mean pulse rate) for given values of x in "pred.x"
predicted <- predict(mod.lin2, newdata = pred.x, type = "response")

pred.df <- data.frame(pred.x, predicted)

# Printing fitted values
print(pred.df, digits = 5, row.names = FALSE)
```

```
##  diet_factor mins predicted
##      Low fat    1    91.254
##  Non-low fat    1    94.953
##      Low fat   15    95.978
##  Non-low fat   15    99.678
##      Low fat   30   101.040
##  Non-low fat   30   104.739
```

Plotting the fitted lines from this linear main effects model:

```
# Picking out fitted values for each diet group (for plotting)
pred.lf <- subset(pred.df, diet_factor == "Low fat")
pred.nlf <- subset(pred.df, diet_factor == "Non-low fat")

# Base R plotting of mean response profile
interaction.plot(exlong$mins_factor,        # x-axis variable
                 exlong$diet_factor,        # lines by diet
                 exlong$pulse,              # y-axis variable
                 fun = "mean",              # summary statistic plotted for response
                 type = "b", pch = 19, lty = 1,
                 xlab = "Time (minutes)", ylab = "Mean Pulse Rate",
                 col = c("blue", "red"),
                 ylim = c(90, 115),
                 legend = FALSE)
title("Mean Response Profiles and Fitted Values from Linear Main Effects Model", cex.main = .95)

# Adding fitted means from linear main effects model
points(factor(pred.lf$mins), pred.lf$predicted, col = "blue", type = "b", lty = 2)
points(factor(pred.nlf$mins), pred.nlf$predicted, col = "red", type = "b", lty = 2)

# Adding legend
legend("topleft",
       c("Low fat actual", "Non-low fat actual", "Low fat predicted", "Non-low fat predicted"),
       lty = c(1, 1, 2, 2), pch = c(19, 19, 1, 1), col = rep(c("blue", "red"),2))
```



- We see the fitted linear trajectories in the two groups. The slope of the two lines is forced to be equal in this model since we are not including an interaction term. There is a constant difference between the two lines that is equal to the estimated main effect of `diet_factor`, b_1. While this linear main effects model is most parsimonious, we see that the fitted line tends to underestimate the effect in the non-low fat diet group.

Comparing Fitted Models

The **Akaike information criterion** (AIC) can be used to evaluate how well a model fits the data and can help aid with model selection. The AIC is computed based on:

- The number of independent variables included in the model (the AIC penalizes models that contain a greater number of independent variables)
- The maximum likelihood estimate of the model

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest independent variables. *Lower AIC* values indicate better model fit. The AIC of a model is extracted using the `AIC()` function. Note that all of our models were fit using the method of **Restricted Maximum Likelihood**. To compare AICs, it's best to compare models fit by maximizing the log-likelihood. Thus, we will re-fit all four models using `method = "ML"` and then compare the AIC of these four models:

```
# Response Profile Model 1 (with interaction)
mod.mrp1ML <- gls(pulse ~ diet_factor*mins_factor, data = exlong,
                  correlation = corSymm(form = ~ time.num | id),
                  weights = varIdent(form = ~ 1 | mins_factor),
                  method = "ML")        # Fit model using maximum likelihood method

# Response Profile Model 2 (without interaction)
mod.mrp2ML <- gls(pulse ~ diet_factor + mins_factor, data = exlong,
                  correlation = corSymm(form = ~ time.num | id),
                  weights = varIdent(form = ~ 1 | mins_factor),
                  method = "ML")

# Linear Trend Model 1 (with interaction)
mod.lin1ML <- gls(pulse ~ diet_factor*mins, data = exlong,
                  correlation = corSymm(form = ~ time.num | id),
                  weights = varIdent(form = ~ 1 | mins),
                  method = "ML")

# Linear Trend Model 2 (without interaction)
mod.lin2ML <- gls(pulse ~ diet_factor + mins, data = exlong,
                  correlation = corSymm(form = ~ time.num | id),
                  weights = varIdent(form = ~ 1 | mins),
                  method = "ML")

# AICs of all models
data.frame(model = c("mod.mrp1ML", "mod.mrp2ML", "mod.lin1ML", "mod.lin2ML"),
           AIC = c(AIC(mod.mrp1ML), AIC(mod.mrp2ML), AIC(mod.lin1ML), AIC(mod.lin2ML)),
           parameters = c(mod.mrp1ML$dims$p, mod.mrp2ML$dims$p, mod.lin1ML$dims$p, mod.lin2ML$dims$p))
```

```
##        model      AIC parameters
## 1 mod.mrp1ML 668.1427          6
## 2 mod.mrp2ML 665.8709          4
## 3 mod.lin1ML 668.4083          4
## 4 mod.lin2ML 668.1023          3
```

Model selection is a balance between parsimony and goodness-of-fit. While the mean response profile interaction model gave fitted means that were exactly equal to the observed means, it also required the largest number of estimated parameters. The AIC accounts for the number of parameters estimated by a model, and we see that the *mean response profile main effects model* has the lowest AIC value. This model gave fitted means that were reasonably close to the actual means and does a good job at describing the trend in the pulse rate over time.

We hope you enjoyed using **R** and **R** Markdown in BIS505b!