29.5: Analysis of Variance (Section 28.6.1)

Often we want to compare several different means, to determine whether any of them are different from the others. In this case, let's look at the data from NHANES to determine whether Marital Status is related to sleep quality. First we clean up the data:

```
NHANES_sleep_marriage <-
  NHANES_adult %>%
  dplyr::select(SleepHrsNight, MaritalStatus, Age) %>%
  drop_na()
```

In this case we are going to treat the full NHANES dataset as our sample, with the goal of generalizing to the entire US population (from which the NHANES dataset is mean to be a representative sample). First let's look at the distribution of the different values of the `MaritalStatus` variable:

```
NHANES_sleep_marriage %>%
  group_by(MaritalStatus) %>%
  summarize(n=n()) %>%
  kable()
```

MaritalStatus	n
Divorced	437
LivePartner	370
Married	2434
NeverMarried	889
Separated	134
Widowed	329

There are reasonable numbers of most of these categories, but let's remove the `Separated` category since it has relatively few members:

```
NHANES_sleep_marriage <-
  NHANES_sleep_marriage %>%
  dplyr::filter(MaritalStatus!="Separated")
```

Now let's use `lm()` to perform an analysis of variance. Since we also suspect that Age is related to the amount of sleep, we will also include Age in the model.

```
lm_sleep_marriage <- lm(SleepHrsNight ~ MaritalStatus + Age,
                        data=NHANES_sleep_marriage)
summary(lm_sleep_marriage)
```

```
##
## Call:
## lm(formula = SleepHrsNight ~ MaritalStatus + Age, data = NHANES_sleep_marriage)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.016 -0.880  0.107  1.082  5.282
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.51758    0.09802   66.49  < 2e-16 ***
## MaritalStatusLivePartner    0.14373    0.09869    1.46  0.14536
## MaritalStatusMarried        0.23494    0.07094    3.31  0.00093 ***
## MaritalStatusNeverMarried   0.25172    0.08404    3.00  0.00276 **
## MaritalStatusWidowed        0.26304    0.10327    2.55  0.01090 *
## Age                         0.00318    0.00141    2.25  0.02464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 4453 degrees of freedom
## Multiple R-squared:  0.00458,    Adjusted R-squared:  0.00347
## F-statistic:  4.1 on 5 and 4453 DF,  p-value: 0.00102
```

This tells us that there is a highly significant effect of marital status (based on the F test), though it accounts for a very small amount of variance (less than 1%).

It's also useful to look in more detail at which groups differ from which others, which we can do by examining the *estimated marginal means* for each group using the `emmeans()` function.

```
# compute the differences between each of the means
leastsquare <- emmeans(lm_sleep_marriage,
                       pairwise ~ MaritalStatus,
                       adjust="tukey")

# display the results by grouping using letters

CLD(leastsquare$emmeans,
    alpha=.05,
    Letters=letters)
```

```
##  MaritalStatus emmean    SE   df lower.CL upper.CL .group
##  Divorced         6.7 0.066 4453      6.5      6.8  a
##  LivePartner      6.8 0.073 4453      6.7      7.0  ab
##  Married          6.9 0.028 4453      6.8      7.0   b
##  NeverMarried     6.9 0.050 4453      6.8      7.0   b
##  Widowed          6.9 0.082 4453      6.8      7.1  ab
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 5 estimates
## significance level used: alpha = 0.05
```

The letters in the `group` column tell us which individual conditions differ from which others; any pair of conditions that don't share a group identifier (in this case, the letters `a` and `b`) are significantly different from one another. In this case, we see that Divorced people sleep less than Married or Widowed individuals; no other pairs differ significantly.

29.5.1 Repeated measures analysis of variance

The standard analysis of variance assumes that the observations are independent, which should be true for different people in the NHANES dataset, but may not be true if the data are based on repeated measures of the same individual. For example, the NHANES dataset involves three measurements of blood pressure for each individual. If we want to test whether there are any differences between those, then we would need to use a *repeated measures* analysis of variance. We can do this using `lmer()` as we did above. First, we need to create a "long" version of the dataset.

```
NHANES_bp_all <- NHANES_adult %>%
  drop_na(BPSys1,BPSys2,BPSys3) %>%
  dplyr::select(BPSys1,BPSys2,BPSys3, ID) %>%
  gather(test, BPsys, -ID)
```

Then we fit a model that includes a separate intercept for each individual.

```
repeated_lmer <-lmer(BPsys ~ test + (1|ID), data=NHANES_bp_all)
summary(repeated_lmer)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: BPsys ~ test + (1 | ID)
##    Data: NHANES_bp_all
##
## REML criterion at convergence: 89301
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -4.547 -0.513 -0.005  0.495  4.134
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 280.9    16.8
##  Residual              16.8     4.1
## Number of obs: 12810, groups:  ID, 4270
##
## Fixed effects:
##             Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)  122.0037     0.2641 4605.7049   462.0   <2e-16 ***
## testBPSys2    -0.9283     0.0887 8538.0000   -10.5   <2e-16 ***
## testBPSys3    -1.6215     0.0887 8538.0000   -18.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) tsBPS2
## testBPSys2 -0.168
## testBPSys3 -0.168  0.500
```

This shows us that the second and third tests are significant different from the first test (which was automatically assigned as the baseline by `lmer()`). We might also want to know whether there is an overall effect of test. We can determine this by comparing the fit of our model to the fit of a model that does not include the test variable, which we will fit here. We then compare the models using the `anova()` function, which performs an F test to compare the two models.

```
repeated_lmer_baseline <-lmer(BPsys ~ (1|ID), data=NHANES_bp_all)
anova(repeated_lmer,repeated_lmer_baseline)
```

```
## Data: NHANES_bp_all
## Models:
## repeated_lmer_baseline: BPsys ~ (1 | ID)
## repeated_lmer: BPsys ~ test + (1 | ID)
##                        Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## repeated_lmer_baseline  3 89630 89652 -44812    89624
## repeated_lmer           5 89304 89341 -44647    89294   330      2     <2e-16
##
## repeated_lmer_baseline
## repeated_lmer          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that blood pressure differs significantly across the three tests.
