MA376 Applied Statistics: Lesson 9

Paired Data

COL Nick Clark

```r
library(tidyverse)
```

I was once asked to consult on a project with DPE examining the surface on which the ACFT sled pull was conducted. They wanted to demonstrate that there was an effect due to the surface. They took 25 volunteers and had them do the sled pull on both grass and sand.

Their *initial* sources of variation diagram was:

Observed variation in:	Sources of explained variation:	Sources of unexplained variation:
Inclusion Criteria:		
Design:		

Let's specify the model used for this type of experiment:

Their statistical question was:

The data was fit using:

```r
# Load the data.
grass_sand_df <- read_csv("https://raw.githubusercontent.com/nick3703/MA376/master/ACFT.csv") %>%
  mutate(Participant = as.factor(Participant)) %>%
  mutate(Surface = ifelse(Surface %in% c("S","S "), "S", Surface)) %>%
  filter(Surface %in% c("G","S")) %>%
  mutate(Surface = as.factor(Surface)) %>%
  droplevels()

# Fit the effects model.
contrasts(grass_sand_df$Surface) = contr.sum
surface_lm <- lm(Sled ~ Surface, data = grass_sand_df)
summary(surface_lm)
```

```
##
## Call:
## lm(formula = Sled ~ Surface, data = grass_sand_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.460  -5.942  -1.160   4.346  17.964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.398      1.054  36.423   <2e-16 ***
## Surface1      -1.438      1.054  -1.364    0.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.454 on 48 degrees of freedom
## Multiple R-squared:  0.03732,    Adjusted R-squared:  0.01726
## F-statistic: 1.861 on 1 and 48 DF,  p-value: 0.1789
```

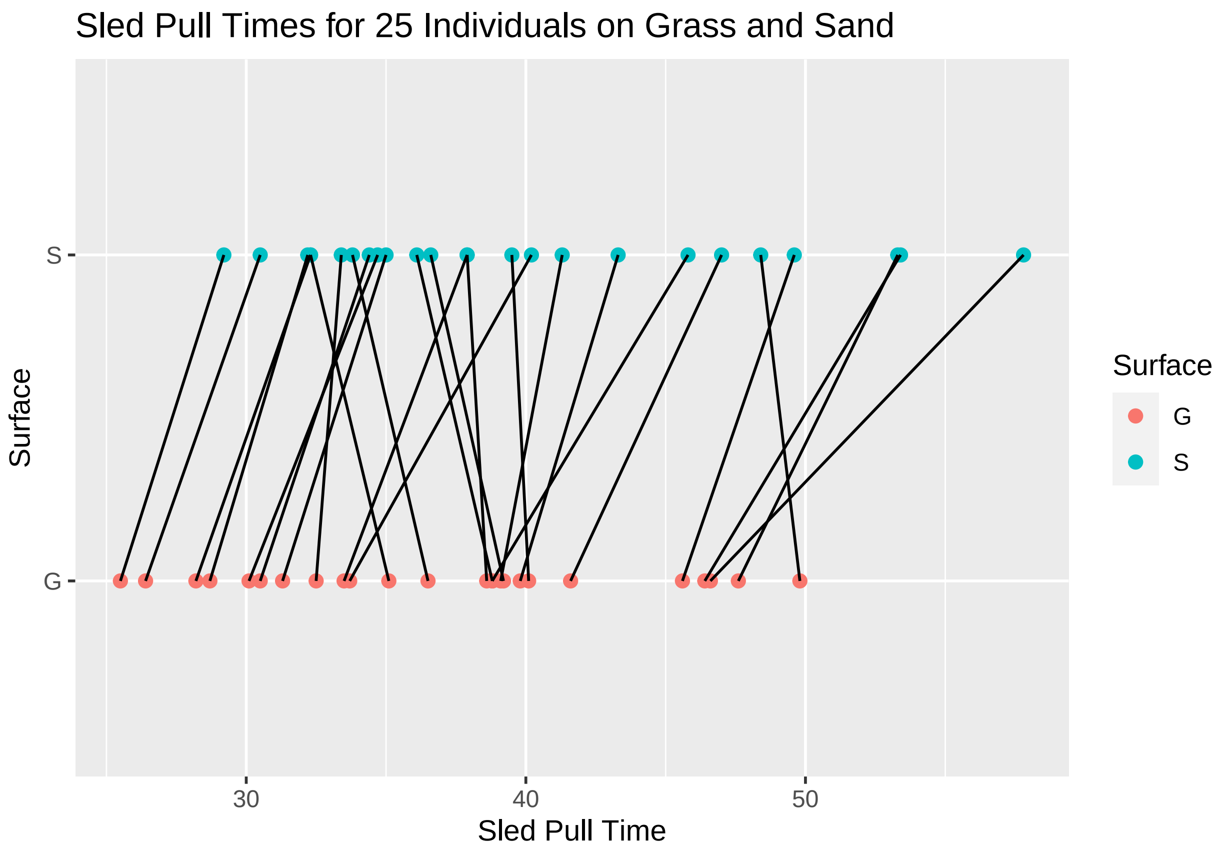
```
anova(surface_lm)
```

```
## Analysis of Variance Table
##
## Response: Sled
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Surface    1  103.39 103.392  1.8606 0.1789
## Residuals 48 2667.28  55.568
```

What is our conclusion?

Let's dig into the individual participant sled pull times.

```
grass_sand_df %>%
  ggplot(aes(x = Sled, y = Surface, group = Participant, col = Surface)) +
  geom_point(size = 2) +
  geom_line(col = "black") +
  labs(title = "Sled Pull Times for 25 Individuals on Grass and Sand",
       x = "Sled Pull Time",
       y = "Surface")
```



What can we learn by looking at this plot?

Let’s modify our sources of variation diagram. Our plot revealed a major source of unexplained variation, but it doesn’t have to remain unexplained!

Observed variation in:	Sources of explained variation:	Sources of unexplained variation:
Inclusion Criteria:		
Design:		

Let’s write out a new statistical model:

Does the hypothesis we are testing change?

There are two ways we can approach this statistical question. Let's consider a single subject and see what happens if we take the difference of our two observations.

These differences are typically analyzed using a paired *t*-test. Instead of our two observed values for each subject, we are looking at our differences.

```
diff_df <- grass_sand_df %>%
  group_by(Participant) %>%
  arrange(Surface) %>% # This is to help us compute sand minus grass.
  summarize(diff = diff(Sled))%>%
  select(diff)
```

Then we can just do a standard one sample *t*-test to see if the difference is indeed zero.

```
n <- nrow(diff_df)
y_vals <- diff_df$diff
t_stat <- mean(y_vals) / (sd(y_vals) / sqrt(n))
p_val <- 2 * (1 - pt(abs(t_stat), n-1))
p_val
```

```
## [1] 0.0005733163
```

You can do this procedure in R using `t.test()`:

```
t.test(y_vals)
```

```
##
##  One Sample t-test
##
## data:  y_vals
## t = 3.9666, df = 24, p-value = 0.0005733
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.379554 4.372446
## sample estimates:
## mean of x
##     2.876
```

The second way to analyze the data (which will be more helpful for this class) is to continue in an ANOVA framework:

Recall the original model that ignored the repeated nature of the two measures on each participants:

```
contrasts(grass_sand_df$Participant) = contr.sum
surface_lm <- lm(Sled ~ Surface, data = grass_sand_df)
anova(surface_lm)
```

```
## Analysis of Variance Table
##
## Response: Sled
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Surface    1  103.39 103.392  1.8606 0.1789
## Residuals 48 2667.28  55.568
```

Now let’s see what value is added by considering the observations in pairs.

```
full_surface_lm<-lm(Sled ~ Surface + Participant, data = grass_sand_df)
anova(full_surface_lm)
```

```
## Analysis of Variance Table
##
## Response: Sled
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Surface      1  103.39 103.392  15.734 0.0005733 ***
## Participant 24 2509.56 104.565  15.912 1.333e-09 ***
## Residuals   24  157.71   6.571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the ANOVA table we see that by adding a second factor (`Participant`) we have actually taken some of our unexplained variation (residuals) and explained it through `Participant`. This makes the F-statistic for `Surface` bigger because we are no longer comparing 103.392 to 55.56, but rather 103.392 to 6.57. Note other things that change in this new ANOVA table.

Let’s see how the sums of squares calculations have changed: