Lesson 3: Afghan Schools (continued)

CDT: I.M. Smrt

Notation

Notation	Definition
i	group
j	observation
$y_{i,j}$	the response for the j^{th} observation from the i^{th} group
$\bar{y}$	overall sample mean
$\bar{y}_i$	sample mean of the i^{th} group
n	total sample size
n_i	group size of the i^{th} group

Background

Recall that in Lesson 2 we analyzed a data set of 1394 school children in northwestern Afghanistan. Researchers grouped 31 villages into 11 equally sized village groups based on political and cultural alliances. Five of these village groups were then randomly assigned to establish a village based school.

```
library(tidyverse)
afghan = read_csv("https://raw.githubusercontent.com/jkstarling/MA376/main/afghan_school.csv")
```

Single Mean Model

A general single mean statistical model for predicting outcomes is:

$$y_{ij} = \mu + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \text{iid } F(0, \sigma)$$

For the Afghan Schools experiment, the fitted single mean model and the standard error of the residuals (SD of the response) are:

Predicted test score $= \hat{y}_i = 73.24$ points; SE of residuals $= 11.14$ points.

```
single.model = lm(test_score~1, data = afghan)
summary(single.model)
```

```
##
## Call:
## lm(formula = test_score ~ 1, data = afghan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.244  -9.244  -1.244   7.756  23.756
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.2439     0.2984   245.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 1393 degrees of freedom
```

Sum of squares total (SSTotal), Sum of squared error (SSError), and Standard Error of Residuals (SE)

The SSError (SSE) is the sum of the squared prediction errors (residuals) for a particular model.

$$SSError = \sum_{\text{all obs}} (\text{observed value - predicted value})^2 = \sum_{\text{all obs}} residuals^2$$

The SSTotal (SST) represents the sum of squared difference between the observed values and the overall mean. Note that in the case of the single mean model, SSE = SST.

$$SSTotal = \sum_{\text{all observation}} (\text{observed value - overall mean})^2$$

The Standard Error of the residuals (SE) is an estimate of the standard deviation of the residuals $\epsilon_{i,j}$. In general, the formula for SE is the square root of the SSE divided by the degrees of freedom of our model (to get an unbiased estimate of σ^2. The standard error of our residuals is one way to estimate how appropriate the model is for our data by comparing the SEs for different models.

$$SE = \sqrt{\frac{SSE}{degreesoffreedom}}$$

1) Calculate `SSE` for the single-means model and name it `SSTotal`.

```
ybar <- mean(afghan$test_score)
SSTotal <- sum( ( afghan$test_score - ybar)^2 )

SSTotal
```

```
## [1] 172961.1
```

2) Calculate the standard error of the residuals and name it `SE.single`. How many degrees of freedom do we have with the single-means model?

```r
n <- length(afghan$test_score)
dof.1 <- n - 1
SE.single <- sqrt(SSTotal/dof.1)

SE.single
```

```
## [1] 11.14291
```

Multiple (Separate) Means Model

A general statistical model for predicting outcomes depending on which treatment group the experimental unit is assigned to is:

$$y_{ij} = \mu_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \text{iid } F(0, \sigma)$$

For the Afghan Schools experiment, the fitted multiple means model and the standard error of the residuals are:

$$\text{Predicted Score} = \hat{y}_i = \begin{cases} 70.003 \text{ points}, & \text{if traditional school (Control)} \\ 76.485 \text{ points}, & \text{if village school (Treatment)} \end{cases}$$

3) Recreate the multiple-means model in R that we did last class:

```r
multi.model <- afghan %>% lm(test_score ~ 0 + as.factor(treatment), data = .)
summary(multi.model)
```

```
##
## Call:
## lm(formula = test_score ~ 0 + as.factor(treatment), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.485  -9.003  -1.485   7.997  26.997
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## as.factor(treatment)0   70.003      0.404   173.3   <2e-16 ***
## as.factor(treatment)1   76.485      0.404   189.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.66 on 1392 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9793
## F-statistic: 3.294e+04 on 2 and 1392 DF,  p-value: < 2.2e-16
```

Effects Model

This section in our book also proposes another model (see p. 47). This model incorporates the **effect** of each treatment, where the **effect** is the *difference of the mean response in the mean of each treatment group from the overall mean response*. We can calculate the effect for each group by subtracting the overall mean from each group mean.

This model can be represented by:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$
$$\alpha_i = \mu_1 - \mu$$
$$\epsilon_{ij} \sim \text{iid } F(0, \sigma)$$

How is this different than the multiple means model?

4) Calculate the effects for both treatments.

```
gp0 <- afghan %>% filter(treatment=="0")
gp1 <- afghan %>% filter(treatment=="1")
gp0.mean <- mean(gp0$test_score)
gp1.mean <- mean(gp1$test_score)
eff0 <- gp0.mean - ybar
eff1 <- gp1.mean - ybar
print(c(eff0, eff1))
```

```
## [1] -3.241033  3.241033
```

5) Create this model in R and assign it to `eff.model`. What do we notice about the values we see in the summary output?

```
afghan <- afghan %>% mutate(treatment1 = as.factor(treatment))
contrasts(afghan$treatment1) = contr.sum
eff.model <- afghan %>% lm(test_score~treatment1, data = .)
summary(eff.model)
```

```
##
## Call:
## lm(formula = test_score ~ treatment1, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.485  -9.003  -1.485   7.997  26.997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.2439     0.2856  256.42   <2e-16 ***
```

```
## treatment11  -3.2410     0.2856  -11.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.66 on 1392 degrees of freedom
## Multiple R-squared:  0.08466,    Adjusted R-squared:  0.084
## F-statistic: 128.7 on 1 and 1392 DF,  p-value: < 2.2e-16
```

We should get the result: (check the values for `contrasts` in the `afghan` dataframe)

$$\text{Predicted Test Score} = \hat{y}_i = 73.24 + \begin{cases} -3.24, \text{ points}, & \text{if traditional school (Control)} \\ 3.24, \text{ points}, & \text{if village school (Treatment)} \end{cases}$$

To see what this model is fitting, consider the model:

$$y_{ij} = \mu + \alpha_i \, x_{ij} + \epsilon_{ij}$$

$$x_{ij} = 1 \text{ if observation } j \text{ is from a government school, -1 if observation } j \text{ is from a village school}$$

$$\epsilon_{ij} \sim \text{iid } F(0, \sigma)$$

Alternative (base R model)

If we had created a model using only the linear model where the `treatment` is a categorical variable, we will get the following model:

$$y_{ij} = \mu + \beta \, x_{ij} + \epsilon_{ij}$$

$$x_{ij} = 1 \text{ if observation } j \text{ is from a village school, 0 otherwise}$$

$$\epsilon_{ij} \sim \text{iid } F(0, \sigma)$$

7) Create a model with `treatment` as a categorical variable and assign it to `base.model`. What is μ?

```
base.model <- afghan %>% lm(test_score~as.factor(treatment), data = .)
summary(base.model)
```

```
##
## Call:
## lm(formula = test_score ~ as.factor(treatment), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.485  -9.003  -1.485   7.997  26.997
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            70.0029     0.4040  173.29   <2e-16 ***
## as.factor(treatment)1   6.4821     0.5713   11.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.66 on 1392 degrees of freedom
## Multiple R-squared:  0.08466,    Adjusted R-squared:  0.084
## F-statistic: 128.7 on 1 and 1392 DF,  p-value: < 2.2e-16
```

8) Calculate and verify that the SE for the multiple means model is 10.66 points. Name the output `SE.multi`. What is the degrees of freedom for this model? What does this mean in relation to the two models?*

```
SSE0 <- sum((gp0$test_score - gp0.mean)^2)
SSE1 <- sum((gp1$test_score - gp1.mean)^2)
SSError <- SSE0 + SSE1

dof.2 <- n - 2

SE.multi <- sqrt(SSError / dof.2)

SE.multi
```

```
## [1] 10.66463
```

Sum of squares model (SSModel).

The SSModel measures the variation in the group means from the overall mean. This is also referred to as the sum of squares for the treatment.

$$SSModel = \sum_{\text{all observation}} (\text{group mean - overall mean})^2 = \sum_{allgroups} (\text{group size}) \times (effect)^2$$

9) Calculate `SSModel`. Hint use `fitted.values` from the multiple means model.

```
#summing across observations
SSModel <- sum((multi.model$fitted.values - ybar)^2)

# or...
SSModel <- length(gp0$test_score) * eff0^2 + length(gp1$jerib_cnt) * eff1^2

SSModel
```

```
## [1] 14642.99
```

10) Verify your sum of squares calculations by ensuring the SSTotal = SSError + SSModel below.

```
SSTotal == SSError + SSModel
```

```
## [1] TRUE
```

To summarize these calculations:

$$SSTotal = SSModel + SSError,$$

and given degrees of freedom (df) for a sum of squares calculation is the number of data points minus the number of parameters we need to estimate, i.e.,

$$df(SSTotal) = df(SSModel) + df(SSError)$$

Degrees of freedom represents the number of independent values in a sum.

Note: these output (SST, SSM, SSE and their df) are given in what is known as an ANOVA (analysis of variance) table. More on this will come later in the course.

Coefficient of Determination (R^2)

The **coefficient of determination**, R^2, is the proportion of the total variation in the response variable which is explained by the explanatory variable(s) in the model. Note that $0 < R^2 < 1$. Larger values of R^2 indicate that more of the variation in the response is explained by the explanatory variable(s).

$$R^2 = \frac{\text{SSModel}}{\text{SSTotal}} = 1 - \frac{\text{SSError}}{\text{SSTotal}}$$

11) Calculate the R^2 for the multiple means model.

```
rsquared <- SSModel / SSTotal

rsquared
```

```
## [1] 0.08466059
```

```
1-(SSError/SSTotal)
```

```
## [1] 0.08466059
```

Note: We should NOT use the Rsquared value from the model without the intercept (separate means models).

References

Burde, Dana, and Leigh L Linden. "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools." American Economic Journal: Applied Economics 5, no. 3 (July 2013): 27–40. https://doi.org/10.1257/app.5.3.27.