Lesson 3: Afghan Schools (continued)

LTC James K. Starling

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:

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

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) and Standard Error of Residuals

The SSTotal represents the total variation in the residuals from the single mean model.

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

1) Calculate `SSTotal`.

```
SSTotal = afghan %>% mutate(sqrd_resid = (test_score-mean(test_score))^2) %>%
  summarise(sum(sqrd_resid)) %>% pull()

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

SSTotal
```

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

2) What is the sample size? Save this value as `n`.

```
# all options below are the same
n <- afghan %>% tally() %>% pull()
n <- pull(tally(afghan))
n <- length(afghan$test_score)

n
```

```
## [1] 1394
```

3) Calculate the standard error of the residuals and name it `SE.single`.

```r
SE.single <- sqrt(SSTotal/(n-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:

$$y_{ij} = \mu_j + \epsilon_{ij} \quad \text{where } i = \text{ experimental unit and } j = \text{treatment group}$$

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

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

SE of residuals = 10.66 points. What does this mean in relation to the two models?

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

```
##
## Call:
## lm(formula = test_score ~ 0 + as.factor(treatment), data = afghan)
##
## 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
```

Effect

The **effect** of each treatment is the difference of the mean response in the treatment group from the overall mean response.

We can calculate the effect for each group by subtracting the overall mean from each group mean.

```
mean.sm <- single.model$coefficients
mean0 <- afghan %>% filter(treatment=="0") %>% summarise(mean(test_score)) %>% pull()
mean1 <- afghan %>% filter(treatment=="1") %>% summarise(mean(test_score)) %>% pull()

#Effect for treatment group
mean1 - mean.sm
```

```
## (Intercept)
##    3.241033
```

```
#Effect for control group
mean0 - mean.sm
```

```
## (Intercept)
##   -3.241033
```

We can write the multiple means model in terms of an overall mean and the effects of the treatment group. See p. 47, Sect. 1.2: "overall mean + treatment effect + random error".

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

Sum of squared error (SSError) and Standard Error of Residuals

The SSError represents the leftover variation in the response variable after accounting for the treatment group, i.e., the variation unexplained by the treatment groups.

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

4) Calculate the SSError for the above model. Hint: you can use `predict(model)` to obtain a predicted value for each observation.

```
SSError = afghan %>%
  mutate(predicted = predict(multi.model)) %>%
  summarize(sum((test_score-predicted)^2)) %>% pull()

SSError <- sum((afghan$test_score - multi.model$fitted.values)^2)

SSError
```

```
## [1] 158318.1
```

Note. The residuals are a object in each model you create. You can pull a vector of residuals from your model using `model$residuals`. The code below uses this method to calculate SSError.

```
sum(multi.model$residuals^2)
```

```
## [1] 158318.1
```

```
sum(resid(multi.model)^2)
```

```
## [1] 158318.1
```

6) Calculate the standard error of the residuals for the multiple means model and name it `SE.multi`.

```
SE.multi = sqrt(SSError/(n-2))

SE.multi
```

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

Sum of squares model (SSModel).

The SSModel measures the variation in the group means from the overall mean.

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

7) Calculate `SSModel`. Hint `predict()` provides the group mean for each observation in the multiple means model

```
#summing across observations
overall_avg = afghan %>% summarise(mean(test_score)) %>% pull()

SSModel = sum((predict(multi.model)-overall_avg)^2)

SSModel
```

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

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

```
SSTotal
```

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

```
SSError+SSModel
```

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

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.

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).

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

9) 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).

Effect Size Larger R^2 values indicate less unexplained variation in the response variable and more precise predictions. However, there is no set value for what makes an R^2 value meaningful in a given scenario. Practical significance refers to whether the group differences are large enough to be of value in the scenario, and usually requires subject matter knowledge and/or something to which to compare the group differences.

An **effect size** measure compares differences in group means to the standard error of the residuals. In this study, the difference in group means is just over half of one residual standard error which seems meaningful in this context.

```
#difference in group means
treatment_mean = as.numeric(multi.model$coefficients[2])
control_mean = as.numeric(multi.model$coefficients[1])

diff = treatment_mean - control_mean
diff
```

```
## [1] 6.482066
```

```
# Standard error of residuals for multiple mean model (same as SE.multi)
SE = summary(multi.model)$sigma
SE
```

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

```
num <- sum((multi.model$fitted.values- afghan$test_score)^2)
SE <- sqrt(num/(n-2))
SE
```

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

```
#Effect size
diff/SE
```

```
## [1] 0.6078099
```

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.