ANOVA: Indicator Variables Formulation

Sleuth3 Sections 5.2, 6.2, 9.3.2 and 9.3.3

Iris Data Again

Example 1: Sepal Width of Iris Flowers

Study Overview

We have measurements of the characteristics of 150 iris flowers, 50 each of three different species:

- Iris setosa is found in the arctic, including Alaska and Maine in the United States, Canada, Russia, northern China, Korea and other northern countries.
- Iris versicolor is found in the eastern United States and eastern Canada.
- Iris virginica is found in the eastern United States

Look at the Data:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# Calculate sample means and standard deviations separately for each species
iris %>%
  group_by(Species) %>%
  summarize(
    mean = mean(Sepal.Width),
    sd = sd(Sepal.Width)
  )
```

```
## # A tibble: 3 x 3
##   Species     mean           sd
##   <fct>      <dbl>        <dbl>
## 1 setosa     3.428 0.3790643691
## 2 versicolor 2.77  0.3137983234
## 3 virginica  2.974 0.3224966382
```

Parameters:

$\mu_1 =$ Average sepal width among all setosa flowers (in the region where the flowers in the sample were found?)

$\mu_2 =$ Average sepal width among all versicolor flowers (in the region where the flowers in the sample were found?)

$\mu_3 =$ Average sepal width among all virginica flowers (in the region where the flowers in the sample were found?)

Indicator variable parameterization

R's output (and output from most other statistical packages) directly answers 3 questions:

1. What is an estimate of μ_1?
2. What is an estimate of $\mu_2 - \mu_1$?
3. What is an estimate of $\mu_3 - \mu_1$?

Verification:

```
anova_fit <- lm(Sepal.Width ~ Species, data = iris)
summary(anova_fit)
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Species, data = iris)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.128 -0.228  0.026  0.226  0.972
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.42800    0.04804  71.359  < 2e-16 ***
## Speciesversicolor -0.65800    0.06794  -9.685  < 2e-16 ***
## Speciesvirginica  -0.45400    0.06794  -6.683 4.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3397 on 147 degrees of freedom
## Multiple R-squared:  0.4008, Adjusted R-squared:  0.3926
## F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```

1. Compare the Estimate labeled (Intercept) to:

- the group mean for setosa flowers: 3.428

2. Compare the Estimate labeled Speciesversicolor to:

- the difference in group means for setosa and versicolor flowers: 2.770 - 3.428 = -0.658
- the results of a hypothesis test that $\mu_2 - \mu_1 = 0$:

```
fit.contrast(anova_fit, "Species", c(-1, 1, 0), conf = 0.95)
```

```
##                        Estimate Std. Error   t value     Pr(>|t|)   lower CI   upper CI
## Species c=( -1 1 0 )   -0.658 0.06793755 -9.685366 1.832489e-17 -0.7922604 -0.5237396
## attr(,"class")
## [1] "fit_contrast"
```

3. Compare the Estimate labeled Speciesvirginica to:

- the difference in group means for setosa and virginica flowers: 2.974 - 3.428 = -0.454
- the results of a hypothesis test that $\mu_3 - \mu_1 = 0$:

```
model_fit <- lm(Sepal.Width ~ Species, data = iris)
fit.contrast(model_fit, "Species", c(-1, 0, 1), conf = 0.95)
```

```
##                        Estimate Std. Error   t value     Pr(>|t|)   lower CI   upper CI
## Species c=( -1 0 1 )   -0.454 0.06793755 -6.682608 4.538957e-10 -0.5882604 -0.3197396
## attr(,"class")
## [1] "fit_contrast"
```

Getting to estimates of means from the R output

We express the mean for a flower of a particular species as follows:

$\mu = \beta_0 + \beta_1 Speciesversicolor + \beta_2 Speciesvirginica$

Here, $Speciesversicolor$ and $Speciesvirginica$ are new **indicator variables**:

- $Speciesversicolor = \begin{cases} 1 \text{ if a flower is of the versicolor species} \\ 0 \text{ otherwise} \end{cases}$
- $Speciesvirginica = \begin{cases} 1 \text{ if a flower is of the virginica species} \\ 0 \text{ otherwise} \end{cases}$

In the background, R creates a new copy of our data frame with these indicator variables that looks like this:

```
iris_augmented
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species Speciesversicolor Speciesvirginica
## 1           5.1         3.5          1.4         0.2     setosa                 0                0
## 2           4.9         3.0          1.4         0.2     setosa                 0                0
## 3           4.7         3.2          1.3         0.2     setosa                 0                0
```

$$\vdots$$

```
## 48          4.6         3.2          1.4         0.2     setosa                 0                0
## 49          5.3         3.7          1.5         0.2     setosa                 0                0
## 50          5.0         3.3          1.4         0.2     setosa                 0                0
## 51          7.0         3.2          4.7         1.4 versicolor                 1                0
## 52          6.4         3.2          4.5         1.5 versicolor                 1                0
## 53          6.9         3.1          4.9         1.5 versicolor                 1                0
```

$$\vdots$$

```
## 98           6.2         2.9          4.3         1.3 versicolor                 1                0
## 99           5.1         2.5          3.0         1.1 versicolor                 1                0
## 100          5.7         2.8          4.1         1.3 versicolor                 1                0
## 101          6.3         3.3          6.0         2.5  virginica                 0                1
## 102          5.8         2.7          5.1         1.9  virginica                 0                1
## 103          7.1         3.0          5.9         2.1  virginica                 0                1
```

Express the mean for *setosa* flowers in terms of β_0, β_1, and/or β_2

Express the mean for *versicolor* flowers in terms of β_0, β_1, and/or β_2

Express the mean for *virginica* flowers in terms of β_0, β_1, and/or β_2

In the output from `summary(anova_fit)`, we have:

- The row labeled `(Intercept)` is related to β_0 (estimate and test of $H_0 : \beta_0 = 0$)
- The row labeled `Speciesversicolor` is related to β_1 (estimate and test of $H_0 : \beta_1 = 0$)
- The row labeled `Speciesvirginica` is related to β_2 (estimate and test of $H_0 : \beta_2 = 0$)

Changing the baseline category

Here's a thing you don't really need to know how to do; just showing that it is possible.

Suppose that instead of using `setosa` for the baseline species, we want to use `virginica` as the baseline.

```r
# What are the levels of the Species variable in the iris data frame, in order?
levels(iris$Species)
```

```
## [1] "setosa"     "versicolor" "virginica"
```

```r
# Update the levels to be in the order "virginica" first, "versicolor" second, and "setosa" third
iris <- iris %>%
  mutate(
    Species = factor(Species, levels = c("virginica", "versicolor", "setosa"))
  )

# Fit with updated order of levels
fit2 <- lm(Sepal.Width ~ Species, data = iris)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Species, data = iris)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.128 -0.228  0.026  0.226  0.972
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.97400    0.04804  61.908  < 2e-16 ***
## Speciesversicolor -0.20400    0.06794  -3.003  0.00315 **
## Speciessetosa      0.45400    0.06794   6.683 4.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3397 on 147 degrees of freedom
## Multiple R-squared:  0.4008, Adjusted R-squared:  0.3926
## F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```

For you to do:

Back in Lab 2 on RStudio, look at the output from calling `summary` on your linear model fit object. Answer the questions below.

- What is the baseline category for the explanatory variable in this model?

- What are the possible values of the `JudgeSpock's` variable, and in what circumstances does the variable equal each of those values?

- What is the `Estimate` labelled `(Intercept)` an estimate of? Be as precise as possible.

- What is the `Estimate` labelled `JudgeSpock's` an estimate of? Be as precise as possible.

- Use the output from `summary` conduct a test of the null hypothesis that $\mu_A = \mu_{Spock's}$.

- Could you use the output from `summary` conduct a test of the null hypothesis that $\mu_B = \mu_{Spock's}$? (The answer is no - why not?)