Two Lines

Sleuth3 Chapters 9 and 10

Example 1: Adapted from Case Study 9.1 in Sleuth3

Quote from the book:

> Meadowfoam (*Limnanthes alba*) is a small plant found growing in moist meadows of the US Pacific Northwest. It has been domesticated at Oregon State University for its seed oil… Researchers reported the results from one study in a series designed to find out how to elevate meadowfoam production to a profitable crop.In a controlled growth chamber, they focused on the effects of two light-related factors: light intensity, at the six levels of 150, 300, 450, 600, 750, and 900 μmol/m^2/sec; and the timing of the onset of the light treatment, either at photoperiodic floral induction (PFI) – the time at which the photo period was increased from 8 to 16 hours er day to induce flowering – or 24 days before PFI. … (Data from M. Seddign and G. D. Jolliff, "Light Intensity Effects on Meadowfoam Growth and Flowering," *Crop Science* 34 (1994): 497-503.)

In this experiment, the researchers planted 10 seedlings in each combination of timing and light intensity, and recorded the mean number of flowers per seedling among those 10 seedlings. They did that twice for each of the 12 combinations of timing and intensity, resulting in a total of 24 observations for our analysis.

The following R code reads the data in and displays the first few observations and the distinct values of the `Time` and `Intensity` variables.

```
## # A tibble: 6 x 3
##   Flowers  Time Intensity
##     <dbl> <dbl>     <dbl>
## 1    62.3     1       150
## 2    77.4     1       150
## 3    55.3     1       300
## 4    54.2     1       300
## 5    49.6     1       450
## 6    61.9     1       450
```

```
## [1] 24
```

```
## # A tibble: 2 x 1
##    Time
##   <dbl>
## 1     1
## 2     2
```

```
## # A tibble: 6 x 1
##   Intensity
##       <dbl>
## 1       150
## 2       300
## 3       450
## 4       600
## 5       750
## 6       900
```

`Flowers` is our response variable, and `Time` and `Intensity` are our explanatory variables.

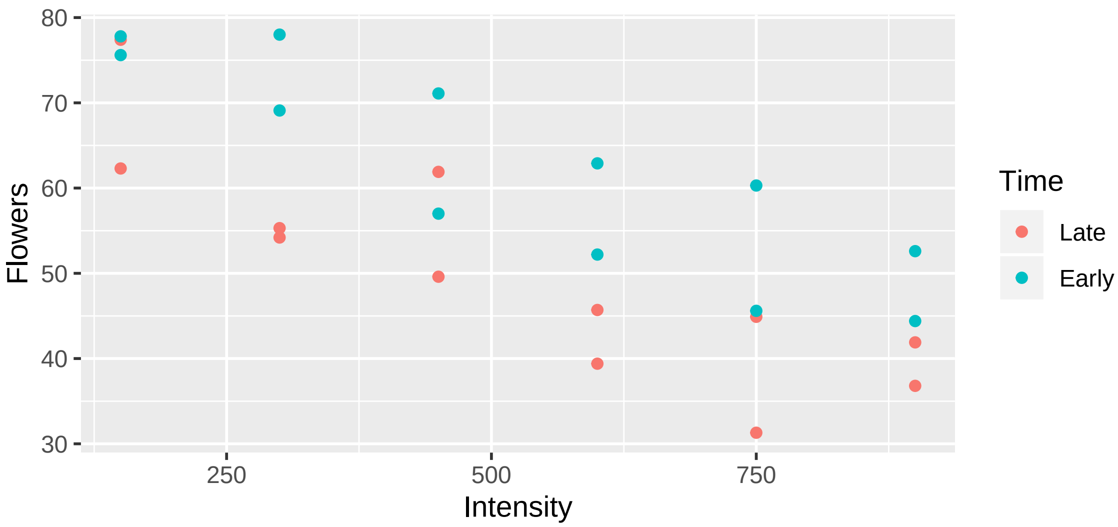
Note that `Time` is currently coded as a numeric variable, either 1 for "Late" or 2 for "Early". We need to tell R it's actually a categorical variable; this is referred to as a `factor` in R. We can do this as follows:

```
meadowfoam <- meadowfoam %>%
  mutate(
    Time = factor(Time, labels = c("Late", "Early"))
  )
head(meadowfoam)
```

```
## # A tibble: 6 x 3
##   Flowers Time  Intensity
##     <dbl> <fct>     <dbl>
## 1    62.3 Late        150
## 2    77.4 Late        150
## 3    55.3 Late        300
## 4    54.2 Late        300
## 5    49.6 Late        450
## 6    61.9 Late        450
```

1. Make a plot of the data using `Flowers` for the vertical axis, `Intensity` for the horizontal axis, and `Time` for the color of points.

```r
ggplot(data = meadowfoam, mapping = aes(x = Intensity, y = Flowers, color = Time)) +
  geom_point()
```



2. Fit a linear model using both `Intensity` and `Time` as explanatory variables, allowing for the slope of the line describing the relationship between light intensity and average number of flowers to be different for the two `Time` settings. Print a summary of the model fit.

```r
lm_fit <- lm(Flowers ~ Intensity * Time, data = meadowfoam)
summary(lm_fit)
```

```
## 
## Call:
## lm(formula = Flowers ~ Intensity * Time, data = meadowfoam)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.516 -4.276 -1.422  5.473 11.938 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         71.623333   4.343305  16.491 4.14e-13 ***
## Intensity           -0.041076   0.007435  -5.525 2.08e-05 ***
## TimeEarly           11.523333   6.142360   1.876   0.0753 .  
## Intensity:TimeEarly  0.001210   0.010515   0.115   0.9096    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.598 on 20 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7692 
## F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

3. In the model summary output, there should be a reference to a variable called `TimeEarly`. What are the possible values of that variable, and under what circumstances does the variable have each of those values?

TimeEarly = 1 for observations where Time is "Early", and 0 otherwise. In this data set, the only possible values for the Time variable are "Early" or "Late", so in practice TimeEarly is 0 if Time is "Late".

4. Write down a single combined equation for the estimated mean number of flowers as a function of the `TimeEarly` and `Intensity` variables.

$\hat{\mu} = 71.623 - 0.041 Intensity + 11.523 TimeEarly + 0.001 Intensity \times TimeEarly$

5. Write down two separate equations: one for the estimated mean number of flowers as a function of `Intensity` in the population of flowers when the light is turned on early, and a second for the estimated mean number of flowers as a function of `Intensity` in the population of flowers when the light is turned on late.

$\hat{\mu} = (71.623 + 11.523) + (-0.041 + 0.001) Intensity$

$\hat{\mu} = 71.623 - 0.041 Intensity$

6. Conduct a test of the claim that separate slopes are not needed for the two timing conditions. State your hypotheses in terms of equations involving one or more model parameters. Additionally, provide a sentence interpreting the meaning of the null hypothesis in context. Your conclusion should be in terms of strength of evidence against the null hypothesis.

H_0: $\beta_3 = 0$ (where β_3 is the population parameter corresponding to the interaction between Intensity and TimeEarly; its estimate is labeled as `Intensity:TimeEarly` in the R output above.) There is no difference in the slopes of the lines relating lighting intensity to mean number of flowers produced in the two timing conditions, in the population of meadowfoam flowers like those in this study.

H_A: $\beta_3 \neq 0$

From the summary output above, the p-value for this test is 0.9096. The data provide no evidence against the null hypothesis that the slopes are the same for the lines relating lighting intensity to mean number of flowers produced in the two timing conditions, in the population of meadowfoam flowers like those in this study.

7. Fit a model where the constraint is imposed that the two lines have the same slope, and show the summary output.

```r
lm_fit <- lm(Flowers ~ Time + Intensity, data = meadowfoam)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Flowers ~ Time + Intensity, data = meadowfoam)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.652 -4.139 -1.558  5.632 12.165
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.305833   3.273772  21.781 6.77e-16 ***
## TimeEarly   12.158333   2.629557   4.624 0.000146 ***
## Intensity   -0.040471   0.005132  -7.886 1.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.441 on 21 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:   0.78
## F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

8. Based on the model from part 7, conduct a test of the claim that separate intercepts are not needed for the two timing conditions. State your hypotheses in terms of equations involving one or more model parameters. Additionally, provide a sentence interpreting the meaning of the null hypothesis in context. Your conclusion should be in terms of strength of evidence against the null hypothesis.

H_0: $\beta_1 = 0$ (where β_1 is the population parameter corresponding to the TimeEarly variable in the model from part 7). There is no difference in the intercepts of the lines relating lighting intensity to mean number of flowers produced in the two timing conditions, in the population of meadowfoam flowers like those in this study.

H_A: $\beta_1 \neq 0$

From the summary output above, the p-value for this test is 0.000146. The data provide strong evidence against the null hypothesis that the intercepts are the same for the lines relating lighting intensity to mean number of flowers produced in the two timing conditions, in the population of meadowfoam flowers like those in this study.

9. Conduct a test of the claim that neither the timing nor the lighting intensity are associated with the mean number of flowers that grow, based on your model from part 7. State your hypotheses in terms of equations involving one or more model parameters. Additionally, provide a sentence interpreting the meaning of the null hypothesis in context. Your conclusion should be in terms of strength of evidence against the null hypothesis.

$H_0 : \beta_1 = \beta_2 = 0$ Neither the timing nor the intensity of light is associated with the mean number of flowers produced by a plant.

H_A: At least one of β_1 and β_2 is not equal to 0.

This is an F test. The p-value for this test is in the last line of output from the R summary. It is 4.786e-08 The data provide extremely strong evidence against the null hypothesis that neither the timing nor the intensity of light is associated with the mean number of flowers produced by a plant.

10. Find and interpret a 95% confidence interval for the coefficient of the `TimeEarly` variable in the model fit from part 7.

```r
confint(lm_fit)
```

```
##                   2.5 %      97.5 %
## (Intercept) 64.49765172 78.11401495
## TimeEarly    6.68987027 17.62679640
## Intensity   -0.05114478 -0.02979808
```

We are 95% confident that at a lighting intensity of 0, the mean number of flowers produced in the population of plants exposed to early lighting is between 6.7 flowers and 17.6 flowers more than the mean number of flowers produced in the population of plants exposed to late lighting.

11. Based on your model fit from part 7, find a 95% prediction interval for the number of flowers that will grow under the early lighting condition with a lighting intensity of 450 μmol/m^2/sec. Interpret your interval in context.

```r
predict_data <- data.frame(
  Intensity = 450,
  Time = "Early"
)

predict(lm_fit, predict_data, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 65.25202 51.28716 79.21689
```

We are 95% confident that the number of flowers produced by a plant exposed to early lighting with an intensity of 450 μmol/m^2/sec will be between about 51 and 79 flowers.

I didn't ask for this, but you should know what we mean by 95% confident in this context: For 95% of samples and 95% of plants in the early lighting condition with intensity of 450, an interval produced using this procedure will contain the number of flowers produced by that individual plant.