Sections 9.1-9.4: MLR models with examples

The day after Halloween, 2017

Last class we discussed three multiple linear regression (MLR) models. The list of these models form a roadmap for Chapter 9. Here the 3 models are listed again. I have also added a Model IV.

Model I. (9.3.2) A model of a response $Y$ as a function of a predictor (or covariate or explanatory variable or independent continuous variable $X_1$) and a factor (a categorical variable $X_2$) that has only two levels $X_2 = 0$ or $X_2 = 1$):

$$
\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2
$$

(1)

This is an equal slopes model, because the two lines being estimated have the same slope.

The R-code to fit this model to data is

$$\text{lm}(Y \sim X1 + \text{as.factor}(X2)).$$

The as.factor() function is not required if the levels (or categories) of $X_2$ are character strings.

Model II. (9.3.4) A model of a response $Y$ with a predictor ($X_1$) and a factor ($X_2$) that has only two levels ($X_2 = 0$ or $X_2 = 1$):

$$
\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2
$$

(2)

This is a separate slopes model, because the two lines being estimated are allowed to have different slopes.

The R-code to fit this model to data is

$$\text{lm}(Y \sim X1 * \text{as.factor}(X2)).$$

Model III. (9.2.1) A model of a response $Y$ with two predictors ($X_1$ and $X_2$):

$$
\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2
$$

(3)

The R-code to fit this model to data is

$$\text{lm}(Y \sim X1 + X2).$$

Model IV. We can add an interaction term to Model III as well:

$$
\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2
$$

In R:

$$\text{lm}(Y \sim X1 \times X2).$$

Assumptions

To reinforce that the assumptions of an MLR are the same as for SLR, we can rewrite Model I as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma)$. In fact, any of the Models I-IV, and any other MLR model, can be rewritten as

$$y = \mu\{Y|X_1, X_2, \ldots\} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma)$. Therefore:
• **normality** the residuals in an MLR are normally distributed about the model \( \mu \{ Y | X_1, X_2 \} \)
• **constant variance** the residuals in an MLR have constant variance \( \sigma^2 \)
• **linearity** the residuals have zero mean
• **random sample** the residuals are independent

We will check these assumptions using graphical approaches just as we did for ANOVA and SLR.

### 9.1.1 Effects of light on flowers

### 9.3.2 An example of Model I

Meadowfoam (\textit{Limananthes alba}) has a unique seed oil that is similar to the oil from sperm whales (long carbon chains) in that it is non-greasy and stable. A randomized experiment was conducted to explore the relationship between light intensity (that your book calls \( X_1 \) on p. 242, with 6 intensities investigated: 150, 300, 450, 600, 750, 900 \( \mu \text{mol}/\text{m}^2/\text{sec} \)); and the timing of the onset of the light treatment (\( X_2 \)) at either photoperiodic floral induction (PFI) (\( X_2 = 1 \) when Timing = “at-PFI”) or 24 days prior to PFI (\( X_2 = 2 \) when Timing = “before-PFI”).

There were 6 \( \times \) 2 = 12 treatment combinations. There were \( n_i = 2 \) randomly assigned replications to each treatment combination. The response variable is the number of flowers per plant (\( Y \)).

Your book (p. 248) writes Model I (equation (1)) as

\[
\mu \{ \text{Flowers} | \text{Intensity, Time} \} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Dummy}_2(\text{Time}).
\]

The function \( \text{Dummy}_L() \), also called an indicator function, converts the levels of any categorical variable to 0's and 1's. The level that is equal to \( L \) is assigned a 1, all other levels are assigned to 0. For example,

\[
\text{Dummy}_2(\text{Time}) = \begin{cases} 
1 & \text{if Time = 2} \\
0 & \text{otherwise (i.e., if Time = 1)}
\end{cases}
\]

An equivalent model could be written with respect to the categorical variable Timing as

\[
\mu \{ \text{Flowers} | \text{Intensity, Timing} \} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Dummy}_{\text{before-PFI}}(\text{Timing}).
\]

so now \( \text{Dummy}_{\text{before-PFI}}(\text{Timing}) = \begin{cases} 
1 & \text{if Timing = “before-PFI”} \\
0 & \text{otherwise (i.e., if Timing = “at-PFI”)}
\end{cases} 
\]

The last model allows us to write out the equation for the line for the timing condition Timing = “at-PFI”

\[
\mu \{ \text{Flowers} | \text{Intensity, Timing = “at-PFI”} \} = \beta_0 + \beta_1 \text{Intensity}.
\]

A second line with the same slope (\( \beta_1 \)) for the timing condition Timing = “before-PFI” is

\[
\mu \{ \text{Flowers} | \text{Intensity, Timing = “before-PFI”} \} = (\beta_0 + \beta_2) + \beta_1 \text{Intensity}.
\]

Unfortunately, it is common to abuse notation and drop the notation \( \text{Dummy}() \) when writing an MLR model. It is understood that if you include a factor \( X_2 \) into a MLR, what you really mean to add to the model is \( \text{Dummy}(X_2) \). Be careful with R’s \texttt{lm()} output when using factors because:

- if your factor levels have numeric labels, R may include it as a covariate instead of a factor unless you use the \texttt{as.factor()} function.
- R uses a Dummy variable that sets the level that comes alpha-numerically first to 0 just as in the example above. You can reset the reference or base level using the \texttt{relevel} command that we used earlier.
QUESTIONS OF INTEREST:

1. Do differences in intensity affect flowering production? State this question in terms of the parameters in the model above.

2. Does timing affect production? State this question in terms of the parameters in the model above.

Let’s see about answering these questions after fitting the model \( \mu \{ \text{Flowers} | \text{Intensity}, \text{Time} \} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Time} \) to the data.

```r
library(Sleuth3)
d1 = case0901
summary(d1)
```

```
#> Flowers   Time  Intensity
#> Min. :31.30 Min. :1.0 Min. :150
#> 1st Qu.:45.42 1st Qu.:1.0 1st Qu.:300
#> Median :54.75 Median :1.5 Median :525
#> Mean   :56.14 Mean   :1.5 Mean   :525
#> 3rd Qu.:64.45 3rd Qu.:2.0 3rd Qu.:750
#> Max.   :78.00 Max.   :2.0 Max.   :900
```

```r
dim(d1)  # n=24
```

```
#> [1] 24   3
```

# Let's add a categorical variable called Timing
```r
d1$Timing = character(24)
d1$Timing[d1$Time==1] = "at-PFI"
d1$Timing[d1$Time==2] = "before-PFI"
d1$Timing = as.factor(d1$Timing)
```

# Look at rows 1-3, 12-15 of the data
```r
d1[c(1:3,12:15),]
```

```
#>   Flowers Time Intensity Timing
#> 1      62.3 1       150 at-PFI
#> 2      77.4 1       150 at-PFI
#> 3      55.3 1       300 at-PFI
#> 12    41.9 1       900 at-PFI
#> 13     77.8 2       150 before-PFI
#> 14     75.6 2       150 before-PFI
#> 15     69.1 2       300 before-PFI
```

# A scatterplot
```r
plot(Flowers ~ Intensity,pch=Time,col=Time,data=d1)
```

# Let's fit Model I to the data wrt the factor Time
```r
m1 = lm(Flowers ~ Intensity + as.factor(Time),data=d1)
summary(m1)
```

```r
## Call:
## lm(formula = Flowers ~ Intensity + as.factor(Time), data = d1)
##
## Residuals:
##    Min     1Q   Median     3Q    Max
##  -10.51  -2.88    0.00   2.88   10.51
##
## Coefficients: (1 not defined because of intercept)
##            Estimate Std. Error t value
## (Intercept)  66.250      1.151   57.08
## Intensity    0.077      0.094    0.83
## Timingat-PFI -3.248      1.370   -2.38
## Timingbefore-PFI 0.854      1.576    0.54
##
## Residual standard error: 4.06 on 12 degrees of freedom
## Multiple R-squared:  0.818,   Adjusted R-squared:  0.765
## F-statistic: 23.1 on 3 and 12 DF,  p-value: 6.85e-05
```
## lm(formula = Flowers ~ Intensity + as.factor(Time), data = d1)
##
## 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 ***
## Intensity -0.040471 0.005132 -7.886 1.04e-07 ***
## as.factor(Time)2 12.158333 2.629557 4.624 0.000146 ***
## ---
## 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

# Here's Model I again, but now wrt the factor Timing  

m2=lm(Flowers ~ Intensity + Timing, data=d1)
summary(m2)

##
## Call:
## lm(formula = Flowers ~ Intensity + Timing, data = d1)
##
## 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 ***
## Intensity -0.040471 0.005132 -7.886 1.04e-07 ***
## Timingbefore-PFI 12.158333 2.629557 4.624 0.000146 ***
## ---
## 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

# Add the model fit to the scatterplot
abline(71.3,-0.0405,col=1,lty=1)
abline(71.3+12.16,-0.0405,col=2,lty=2)
legend("topright",legend=c("Timing = 'at-PFI'", "Timing = 'before-PFI'"), pch=c(1,2),lty=c(1,2),col=c(1,2))
The assumptions for MLR are the same as the assumptions for SLR. The plots from the `diagANOVA()` command below assess these assumptions (i.e., model fit).

QUESTIONS:

1. Do the MLR assumptions appear to be satisfied?

2. Do differences in intensity affect flowering production?

3. Does timing affect production?

4. Give an estimate of $\sigma$, the constant SD

5. Give the value of $R^2$ as a quantitative measure of the model’s fit.

`source("diagANOVA.r")`
`diagANOVA(m1)`
### [1] "In this sample of size n=24, correlation of the residuals in the qq-plot is r=0.981447"
### [1] "In the following table, if r < critical.r, then the qq-plot suggests the residuals are not normal:
### `n`  `critical.r`
### 1  5   0.832
### 2 10   0.880
### 3 15   0.911
### 4 20   0.929
### 5 25   0.941
### 6 30   0.949
### 7 40   0.960
### 8 50   0.966
### 9 60   0.971
### 10 75   0.976

#### 9.3.4 An example of Model II

For the flowering experiment, what if we wanted to answer the question:

Does the change in flowering production as a function of intensity depend on the timing?

Another way to ask this question is whether there is an interaction between intensity and timing that affects flowering. Let us consider Model II to estimate the interaction.

Your book (p. 250) writes Model II (equation (2)) as

\[
\mu\{\text{Flowers}|\text{Intensity, Time}\} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Dummy}_2(\text{Time}) + \beta_3 \text{Intensity} \times \text{Dummy}_2(\text{Time}).
\]
This model allows us to write the equation for a line for the timing condition at Time = 1 (Timing = “at-PFI”) as:

$$\mu\{\text{Flowers}|\text{Intensity, Time} = 1\} = \beta_0 + \beta_1\text{Intensity}$$

The second line for the timing condition at Time = 2 (Timing = “before-PFI”) is

$$\mu\{\text{Flowers}|\text{Intensity, Time} = 2\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Intensity}$$

QUESTION: Which parameter is of interest to answer the question: Does the change in flowering production as a function of intensity depend on the timing?

Let's fit Model II to the flowering data:

```r
m3 = lm(Flowers ~ Intensity*as.factor(Time),data=d1)
summary(m3)
```

```
## Call: lm(formula = Flowers ~ Intensity * as.factor(Time), data = d1)
##
## 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) 71.62333 4.34330 16.491 4.14e-13 ***
## Intensity -0.04108 0.00743 -5.525 2.08e-05 ***
## as.factor(Time)2 11.52333 6.14236 1.876 0.0753 .
## Intensity:as.factor(Time)2 0.00121 0.01051 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
```

# A scatterplot
```
#plot(Flowers ~ Intensity,pch=Time,col=Time,data=d1)
# Add the model fit to the scatterplot
#abline(71.6,-0.0411,col=1,lty=1)
#abline(71.6+11.52,-0.0411+.0012,col=2,lty=2)
#legend(x=720,y=79,legend=c("Time = 0","Time = 1"),pch=c(1,2),lty=c(1,2),col=c(1,2))
```

# diagANOVA(m3)

# Perform a lack-of fit test
anova(m1,m3)

## Analysis of Variance Table
##
## Model 1: Flowers ~ Intensity + as.factor(Time)
QUESTIONS:

1. Does the change in flowering production as a function of intensity depend on the timing?

2. Report the R2 value for this last Model II fit to the data. How does it compare to the Model I fit to the data?

3. Report and interpret the lack of fit test.