Chapter 8 Indicator Variables

- **8.1 General Concepts**
- **8.2 Analysis of Variance ANOVA**
- **8.3 Analysis of Covariance ANCOVA**

8.1 General Concepts

- Heights of four people have been recorded:

Obs.	height	gender
1	160	F
2	150	F
3	175	M
4	165	M

- How is height (y, in cm) related to gender? Gender is not a quantitative variable. It is a categorical factor.
- Use the following coded variable:

$$x_i = \begin{cases} 0, & \text{if M} \\ 1, & \text{if F} \end{cases}$$

General Concepts (Cont'd)

- We now have the following data:

Obs.	height	gender	x
1	160	F	1
2	150	F	1
3	175	M	0
4	165	M	0

- We can check if there is a relation between height and gender using the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} \text{ so } (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 650 \\ 310 \end{bmatrix} \text{ so } \widehat{\beta} = \begin{bmatrix} 170 \\ -15 \end{bmatrix}$$

General Concepts (Cont'd)

- Fitted Model:

$$\widehat{y} = 170 - 15x_i$$

- Fitted Values:

obs	gender	x	y.hat
1	F	1	170-15 = 155
2	F	1	155
3	M	0	170-0 = 170
4	M	0	170

- The fitted values are the average heights for each gender.
- Error variance estimate:

$$\widehat{\sigma}^2 = \text{MSE} = \frac{SSE}{n-p} = \frac{\mathbf{y}^\top\mathbf{y} - \widehat{\beta}^\top\mathbf{X}^\top\mathbf{y}}{2}$$

$$= \frac{105950 - [170 \;\; -15][650 \; 310]^\top}{2} = 50$$

Standard Error Estimates

- Standard error estimates for $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\text{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

so

$$\text{Var}(\widehat{\beta}_0) = \sigma^2(1/2)$$

and

$$\text{Var}(\widehat{\beta}_1) = \sigma^2(1)$$

so

$$s.e.(\widehat{\beta}_0) = \sqrt{\text{MSE}/2} = \sqrt{25} = 5$$

$$s.e.(\widehat{\beta}_1) = \sqrt{\text{MSE}} = \sqrt{50} = 7.1$$

Standard Error Estimates (Cont'd)

- We can test whether there is a difference in mean height:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

$$t_0 = \frac{\widehat{\beta}_1}{s.e.} = -15/7.1 = -2.1$$

$$\text{p-value } = 2P(t_2 > |t_0|) = .17$$

Very weak evidence against H_0.
(Can you think of another way of coming up with this test?)

R Approach

```
> gender <- factor(c("F","F","M","M"),
                          levels=c("M","F"))
> height <- c(160,150,175,165)
> y.lm <- lm(height ~ gender)
> summary(y.lm)

Call:
lm(formula = height ~ gender)

Residuals:
 1  2  3  4
 5 -5  5 -5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  170.000      5.000  34.000 0.000864
genderF      -15.000      7.071  -2.121 0.167950

Residual standard error: 7.071 on 2 degrees of freedom
Multiple R-Squared: 0.6923,     Adjusted R-squared: 0.5385
F-statistic:   4.5 on 1 and 2 DF,  p-value: 0.1679
```

It is also an example of the regression approach to Analysis of Variance.

8.2 Analysis of Variance

- The analysis of variance is usually employed to test for mean differences among several samples. In the above example, we compared the mean heights for two samples (males and females).
- In the case where there are k different groups, one fits the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$$

where the x_j's are indicator variables.

Analysis of Variance Example

- e.g. $k = 3$ groups: A, B and C

  $$x_1 = \begin{cases} 1, & \text{if in A} \\ 0, & \text{otherwise} \end{cases}$$

  and

  $$x_2 = \begin{cases} 1, & \text{if in B} \\ 0, & \text{otherwise} \end{cases}$$

  Note that $x_1 = x_2 = 0$ for an observation from group C. Suppose 2 observations are taken from each group. Then $\mathbf{X}$ takes the form

  $$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

  if the first 2 observations are from group A, and the last 2 observations are from C.

Categorical Variables

- In R, categorical variables having more than two or more ’values’ can be dealt with using `factor`:

  e.g. Customers were asked to rate an automobile model on a number of factors, and the average was recorded in the variable `rating`. The colour was also recorded. The data are below:

  ```
  > colour.test
    colour rating
      red    7.6
   yellow    7.9
     blue    7.8
      red    8.8
   yellow    7.7
     blue    7.5
  ```

  Are rating and colour related?

Categorical Variables (Cont'd)

```
> attach(colour.test)
> x <- factor(colour)  # automates the indicator variables
> rating.lm <- lm(rating ~ x)
> summary(rating.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.6500     0.3617  21.150 0.000231
xred          0.5500     0.5115   1.075 0.361062
xyellow       0.1500     0.5115   0.293 0.788456

Residual standard error: 0.5115 on 3 degrees of freedom
```

This is the output we would have obtained if we had set xred to be the indicator of red and xyellow to be the indicator of yellow.

We conclude that rating does not have a statistically discernible relation with colour.

8.3 Analysis of Covariance

- There are both quantitative and indicator variables among the predictors.
- Height example (cont'd). Suppose weights had been recorded too:

Obs.	height	gender	weight
1	160	F	64
2	150	F	60
3	175	M	65
4	165	M	63

We may fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where x_1 is the indicator for gender and x_2 is weight.
Then

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 64 \\ 1 & 1 & 60 \\ 1 & 0 & 65 \\ 1 & 0 & 63 \end{bmatrix}$$

and $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ as usual.

8.3 Analysis of Covariance (Cont'd)

- In R, we have

```
> weight <- c(64, 60, 65, 63)
> height.lm <- lm(height ~ gender + weight)
> summary(height.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.000     64.039  -0.344    0.789
genderF       -9.000      3.742  -2.405    0.251
weight         3.000      1.000   3.000    0.205

Residual standard error: 3.162 on 1 degrees of freedom
```

8.3 Analysis of Covariance (Cont'd)

Fitted model:

$$\widehat{y} = -22 - 9x_1 + 3x_2$$

Interpretation:

$$\widehat{y} = -22 + 3x_2$$

is the line relating height to weight for males.

$$\widehat{y} = -31 + 3x_2$$

is the line relating height to weight for females.
These are parallel lines.

- With more observations, one could fit lines that may or may not be parallel:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

If β_3 is nonzero, we say there is an interaction effect on height due to weight and gender.