11.1

Analysis of Covariance
(Section 15.3)

In this section, we introduce models which include both factor and quantitative input variables.

idea: we are interested in studying factor effects, using covariate information available on each experimental unit.

see R handout: **Example 11.1**

we wish to compare the weights of turkeys from three origins: Georgia, Virginia, Wisconsin. Additionally, we have information on the age of each turkey. covariate (think of the covariate as a quantitative blocking variable)

Data Layout

ANCOVA model

Factor A		
level 1	....	level a
(X_{11}, Y_{11})		(X_{a1}, Y_{a1})
$\vdots$	...	$\vdots$
(X_{1n}, Y_{1n})		(X_{an}, Y_{an})

Sample of n turkeys from each state of origin, Each turkey returns a measurement of age (x) and weight (Y).

$$Y_{ij} = \beta_{0i}^* + \beta_1 X_{ij} + \varepsilon_{ij} \quad \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

regression lines for weight with common slope (age effect) but different intercepts (need to define origin effect).

11.2

The covariate as a blocking variable is not necessarily balanced. That is, the observed levels of age may not be equal across levels of origin. Let's first see how the covariate information is used to adjust the estimate of factor effect.

Write

$$Y_{ij} = \beta_{0i} + \beta_1 (X_{ij} - \bar{X}_i) + \varepsilon_{ij}$$

$$\left(\text{Note } \beta^*_{0i} = \beta_{0i} - \beta_1 \bar{X}_i \right)$$

To see the advantage of centering the model, let's look at the estimated regression coefficients

see R handout

Back to Example 11.1

define the model

```
ancova.mod = lm(weight ~ age + origin)
summary(ancova.mod)
```
← print model estimates

The model fit by R uses indicator variables to define the factor variable. ~~we will skip the details~~ see end notes.

$$\hat{\beta}^*_{01} = -0.4875, \quad \hat{\beta}^*_{02} = -0.761, \quad \hat{\beta}^*_{03} = 1.431$$

$$\hat{\beta}_1 = 0.4868$$

(x = age)

estimated regression functions:

georgia: $\hat{Y}_1 = -0.4875 + 0.4868\, x$

virginia: $\hat{Y}_2 = -0.7610 + 0.4868\, x$

wisconsin: $\hat{Y}_3 = 1.4309 + 0.4868\, x$

$\boxed{\text{see HW}}$

see Scatterplot. We are fitting regression lines for each state, forcing a common slope (age effect).

(11.3)

For the centered model, it can be shown that the estimated regression ~~functions~~ ~~become~~ coefficients take the form

$$\hat{\beta}_{0i} = \bar{Y}_{i\cdot} \quad , \quad i = 1, \dots, a .$$

estimated regression functions:

$$\hat{Y}_1 = \bar{Y}_{1\cdot} + \hat{\beta}(X - \bar{X}_{1\cdot})$$
$$\hat{Y}_2 = \bar{Y}_{2\cdot} + \hat{\beta}(X - \bar{X}_{2\cdot})$$
$$\hat{Y}_3 = \bar{Y}_{3\cdot} + \hat{\beta}(X - \bar{X}_{3\cdot})$$

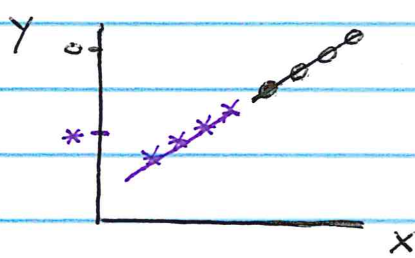
Back to Example 11.1

origin	$\bar{X}_{i\cdot}$ (age mean)	$\bar{Y}_{i\cdot}$ (weight mean)
g	25.5	11.925
v	27.5	12.625
w	25.0	13.600

(overall mean $\bar{X}_{\cdot\cdot} = 25.923$)

Because the covariate means are not equal across factor levels, using factor level means as a measure of factor effect may be misleading.

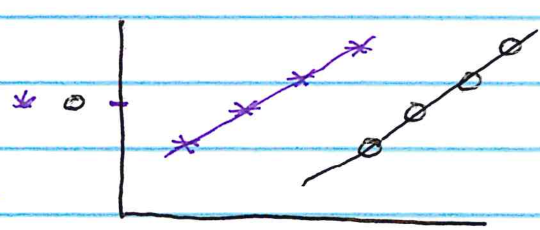
two examples:



case 1.

A comparison based on sample means has O group with larger responses than * group

But * group has smaller covariate values (i.e., * group is tested under less favorable conditions)



case 2.

A comparison based on sample means has O group and * group with no difference

But again, * group is tested under less favorable conditions

11.4

To compare factor level means, we should do so under equivalent covariate levels. Take the center covariate level $\bar{X}_{..}$

Def: least squares means: $\boxed{\bar{Y}_{i,adj} = \bar{Y}_{i.} + \hat{\beta}(\bar{X}_{..} - \bar{X}_{i.})}$

see HW(4)

Suppose $\hat{\beta} > 0$.

If $\bar{X}_{i.} < \bar{X}_{..}$, then $\bar{Y}_{i,adj} > \bar{Y}_{i.}$

(adjust upward for less favorable conditions)

If $\bar{X}_{i.} > \bar{X}_{..}$, then $\bar{Y}_{i,adj} < \bar{Y}_{i.}$

(adjust downward for more favorable conditions)

see R output

Back to Example 11.1

lsmeans(ancova.mod, pairwise~origin, adjust="none")

see HW(5)

origin	$\bar{X}_i$	$\bar{Y}_i$	$\bar{Y}_{i,adj}$	
g	25.5	11.925	12.1	
v	27.5	12.625 ↓	11.9	(more favorable, older turkeys)
w	25.0	13.600 ↑	14.0	(less favorable, younger turkeys)

($\bar{X}_{..} = 25.9$)

grouping information: $\boxed{w}$ $\boxed{g, v}$

largest ⟶ smallest

	direction	P
g − v	0	.2421
g − w	−	.000
v − w	−	.000

Let's go back to the problem of testing for factor effects.

11.5

ANOVA model: $Y_{ij} = \mu_i + \varepsilon_{ij}$

$H_0: \mu_1 = \dots = \mu_a$ (no factor A effect) factor level means are equal

Recall the sum of squares decomposition:

Let $R(A) = SS_A = SS_{TOTAL} - SSE(A)$

(variation explained by including Factor A in the model)

see R output

Back to Example 11.1

a.mod = lm (weight ~ origin)

anova (a.mod)

see HW

$F = 0.966$, $p = .4135$ 3.

The experiment finds that origin has no effect on turkey weight.

The covariate age plays no role in this analysis.

ANCOVA model: $Y_{ij} = \beta_{0i}^{*} + \beta_1 X_{ij} + \varepsilon_{ij}$

$H_0: \beta_{01}^{*} = \dots = \beta_{0a}^{*}$ (no factor A effect) Factor level regression functions are equal

Now, the sum of squares decomposition includes the variation explained by the covariate input X.

Let $R(X) = SS_{reg} = SS_{TOTAL} - SSE(X)$

(variation explained by including covariate X in the model)

Let $R(X, A) = SS_{TOTAL} - SSE(X, A)$

(variation explained by including both the covariate X and the Factor A in the model.)

Let $R(A|x) = R(x, A) - R(x)$

define the <u>sequential sum of squares</u>

(variation explained by factor A, adjusting for covariate x)

see R output

<u>Back to Example 11.1</u>

```
ancova.mod = lm(weight ~ age + origin)
anova(ancova.mod)
```

see HW

$F = 68.81$, $p = .000$ 4.

The experiment finds that origin has an effect on turkey weight, after adjusting for the effect of age.

- see case 2, page 11.3 for a sketch showing no marginal A effect, but adjusted A effect

- The problem set features data similar to case 1, with marginal A effect, but no adjusted A effect.

Unbalanced Factorial Design

(related to observational data, such as smoking and cancer, football and CTE, mask rules and virus cases)

(also related to multicollinearity)

Example 11.2 - 2^2 design with missing observations

A = temperature

B = pressure

y = yield

B \ A	1	2
2	90.6	90.8 90.9
1	90.1 90.3	

calculations: $\bar{Y}_{..} = 90.54$, $SS_{TO} = 0.4520$

models: (AB) $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$

(A) $\hat{Y} = b_0^{(A)} + b_1^{(A)} X_1$ (B) $\hat{Y} = b_0^{(B)} + b_2^{(B)} X_2$

where

$$X_1 = \begin{cases} 1, & \text{if temp} = \text{"high"} \\ 0, & \text{if temp} = \text{"low"} \end{cases}$$

and

$$X_2 = \begin{cases} 1, & \text{if pressure} = \text{"high"} \\ 0, & \text{if pressure} = \text{"low"} \end{cases}$$

fitted values:

model A,

90.33	90.85
90.33	90.85

$b_0^{(A)} = 90.33$, $b_1^{(A)} = 0.52$

$SSE(A) = (.6-.33)^2 + (.1-.33)^2 + (.3-.33)^2 + 2(.05)^2 = 0.1317$

$R(A) = SS_{TO} - SSE(A) = \underline{.3203}$

R: a.mod = lm(y~A), dummy.coef(a.mod), predict(a.mod)
anova(a.mod)

model B,

90.77	90.77
90.20	90.20

$b_0^{(B)} = 90.20$, $b_2^{(B)} = 0.57$

$SSE(B) = (.17)^2 + (.03)^2 + (.13)^2 + 2(.1)^2 = .0677$

$R(B) = SS_{TO} - SSE(B) = \underline{.3853}$

model A,B

90.60	90.85
90.20	(90.45)

$b_0 = 90.20$, $b_1 = 0.25$, $b_2 = 0.40$

$SSE(A,B) = 2(.05)^2 + 2(.10)^2 = .0250$

$R(A,B) = SS_{TO} - SSE(A,B) = .4270$

R: ab.mod = lm(y~A+B)
anova(ab.mod)

$R(B|A) = R(A,B) - R(A) = \underline{.1067}$

Anova(ab.mod)

$R(A|B) = R(A,B) - R(B) = \underline{.0417}$

End Notes, Section 11

1. The ANCOVA model defined for R computations features an alternate parameterization. Rather than define an intercept parameter for each factor level, as we have done in the notes, R defines an intercept parameter for one factor level, called the baseline level, then defines the remaining factor level parameters as comparisons to the baseline. If factor level 1 is the baseline, then the R model estimated regression functions can be written as

$$\begin{aligned}
1 &: \quad \widehat{y_1} = \widehat{\beta}_0 + \widehat{\beta}_1 x \\
2 &: \quad \widehat{y_2} = \left(\widehat{\beta}_0 + \widehat{\beta}_2\right) + \widehat{\beta}_1 x \\
3 &: \quad \widehat{y_3} = \left(\widehat{\beta}_0 + \widehat{\beta}_3\right) + \widehat{\beta}_1 x
\end{aligned}$$

Thus, we can get the parameterization we desire by defining

$$\begin{aligned}
\widehat{\beta}^*_{01} &= \widehat{\beta}_0 \\
\widehat{\beta}^*_{02} &= \widehat{\beta}_0 + \widehat{\beta}_2 \\
\widehat{\beta}^*_{03} &= \widehat{\beta}_0 + \widehat{\beta}_3
\end{aligned}$$

The R code in Handout 11 defining the estimated regression intercepts follows from above.

2. Let's think of an example which could give rise to the data seen in case 1 and case 2 in the notes.

For case 1, imagine a study comparing two types of instruction. Suppose one type is used in the lower grades and the second type is used in the upper grades. If we simply compared test scores, it may seem that the second type of instruction is superior. But if we consider covariate information on the students, say a pretest score, then the differences in post test scores can be attributed to different skill levels and not to differences in the instruction.

For case 2, imagine a standardized exam, such as the GRE, where the questions get more difficult or easy depending on how well a student performs on the previous questions. If we compare two groups simply on the number of questions answered correctly, we would see no difference between the groups. But if we consider the covariate information on the difficulty of the exam, then differences between the groups become apparent.

3. The statistic for testing factor A effects in the ANOVA model is

$$F = \frac{R\,(A)\,/\,(a-1)}{SSE\,(A)\,/\,(N-a)}$$

This matches the statistic we defined earlier in the semester for ANOVA testing, only now we use $R\,(A)$ to represent the sum of squares for the factor effect. In the computing output, the F statistic is found from the first row in the analysis of variance table.

4. The statistic for testing factor A effects in the ANCOVA model is

$$F = \frac{R\,(A\,|\,x)\,/\,(a-1)}{SSE\,(A,x)\,/\,(N-a-1)}$$

We use $R\,(A\,|\,x)$ to represent to the sequential sum of squares for the factor effect, after adjusting for the covariate effect. In the computing output, the F statistic is found from the second row in the analysis of variance table.