Unit 11: Analysis of Covariance (ANCOVA)

STA 643: Advanced Experimental Design

Derek S. Young
Learning Objectives

▶ Become familiar with the basic orthogonal designs theory of ANCOVA
▶ Understand the benefits and how we include covariates into ANOVA from a designed experiment
▶ Become familiar with the single-factor ANCOVA model
▶ Know the assumptions of ANCOVA
▶ Know how to construct an ANCOVA model when there is more than one covariate
▶ Know how to construct and analyze an ANCOVA model with a blocked design
▶ Know how to construct and analyze a multi-factor ANCOVA model
Outline of Topics

1. Linear Model Theory
2. Single-Factor ANCOVA
3. Generalizations of ANCOVA
Outline of Topics

1. Linear Model Theory
2. Single-Factor ANCOVA
3. Generalizations of ANCOVA
Overview

- **Analysis of covariance** (or ANCOVA) is a technique that combines features of ANOVA and regression.
- ANCOVA can be used for either observational or designed experiments.
- The idea is to augment the ANOVA model containing the factor effects (we will only consider fixed effects) with one or more additional quantitative variables that are related to the response variable.
  - These quantitative variables are called **covariates** (or they also called **concomitant variables**).
- This augmentation is intended to reduce large error term variances that are sometimes present in ANOVA models.
- Note that if the covariates are qualitative (e.g., gender, political affiliation, geographic region), then the model remains an ANOVA model where the original factors are of primary interest while the covariates are simply included for the purpose of error variance reduction.
In an ANCOVA model, we have observations that are taken in different categories as in ANOVA, but we also have some other predictors which are measured on an interval scale as in regression.

In the general form of the ANCOVA model, we observe

\[ Y \sim \mathcal{N}_N(\delta + X\gamma, \sigma^2 I), \quad \delta \in Q, \quad \gamma \in \mathbb{R}^s, \quad \sigma^2 > 0, \]  

where \( X \) is a known \( N \times s \) matrix (which we call a covariate matrix in ANCOVA) and \( Q \) is a known \( q \)-dimensional subspace.

We assume that \( \delta, \gamma, \) and \( \sigma^2 \) are unknown parameters.

Typically, \( Q \) is a subspace from ANOVA.

We call the ANOVA model with subspace \( Q \) the associated ANOVA model.
Coordinate-Free Version of Linear Model

Let

\[ X^* = P_{Q^\perp} X, \quad Y^* = P_{Q^\perp} Y. \]

In order to guarantee that the covariates are not confounded with the ANOVA component of the model, we need to assume that \( X^* \) has full rank; i.e., that \( \text{rank}(X^*) = s \).

Note that it is not enough that \( X \) have full column rank in order to have a possible covariance analysis.

Let \( U \) be the \( s \)-dimensional subspace spanned by the columns of \( X^* \) and let \( V \) be the subspace of possible value for \( \mu \); i.e.,

\[ V = \{ \mu = \delta + X\gamma : \delta \in Q, \gamma \in \mathbb{R}^s \}. \]

Then

\[ Y \sim \mathcal{N}_n(\mu, \sigma^2 I), \quad \mu \in V, \sigma^2 > 0, \]

which is just the ordinary linear model.
Adjusted Estimators

▶ Note that $U \perp Q$ and
\[
\mu = \delta + P_Q X \gamma + X^* \gamma, \quad \delta + P_Q X \gamma \in Q, \quad X^* \gamma \in U.
\]
▶ Therefore,
\[
\mu \in V = Q \oplus U.
\]
▶ Hence
\[
\hat{\mu} = P_V Y = P_Q Y + P_U Y = \hat{\delta}_u + X^* \hat{\gamma}
\]
\[
\hat{\delta}_u = P_Q Y
\]
\[
\hat{\gamma} = (X^* T X^*)^{-1} X^* T Y = (X^* T X^*)^{-1} X^* T Y^*
\]

▶ Note that $\hat{\delta}_u$ is just the estimator of $\delta$ based on the associated ANOVA model and that $\hat{\gamma}$ is just the least squares estimator of $\gamma$ found by regressing $Y$ (or $Y^*$) on $X^*$.
▶ The "u" subscript means the result has been unadjusted for the covariate, while the "a" subscript means the result has been adjusted for the covariate.
▶ Further results using orthogonal direct sums allow us to write
\[
\| \hat{\mu} \|^2 = \| \hat{\delta}_u \|^2 + \| X^* \hat{\delta} \|^2,
\]
and $p = \dim(V) = \dim(Q) + \dim(U) = q + s$. 
Optimal Estimators

▶ Note that
\[ \hat{\gamma} = (X^*TX^*)^{-1}X^T\hat{\mu}, \quad \gamma = (X^*TX^*)^{-1}X^T\mu. \]

▶ Therefore, \( \hat{\gamma} \) is the optimal (i.e., best unbiased) estimator of \( \gamma \); however,
\[ E[\hat{\delta}_u] = PQ(\delta + X\gamma) = \delta + (X - X^*)\gamma. \]

▶ Therefore, \( \hat{\delta}_u \) is the optimal estimator of \( \delta + (X - X^*)\gamma \) for the ANCOVA model.

▶ Note that the optimal estimator of
\[ \delta = (PQ - (X - X^*)(X^*TX^*)^{-1}X^T)\mu \]
is
\[ \hat{\delta}_a = (PQ - (X - X^*)(X^*TX^*)^{-1}X^T)\hat{\mu} = \hat{\delta}_u - (X - X^*)\hat{\gamma}. \]

▶ Thus, we have \( \hat{\delta}_a \), which has been adjusted for the covariate.
Adjusted Sums of Squares

- We can now construct the sums of squares and degrees of freedom for the error when adjusting for covariates.
- First, $SSE_a$ and $df_{e;a}$ for the ANCOVA model are given by

$$
SSE_a = \|Y - \hat{\mu}\|^2 = \|Y\|^2 - \left(\|\hat{\delta}_u\|^2 + ||X^*\hat{\gamma}\|^2\right) = SSE_u - ||X^*\hat{\gamma}\|^2
$$

$$
df_{E;a} = N - \text{dim}(V) = N - (q + s) = df_{e;u} - s,
$$

where $SSE_u$ and $df_{e;u}$ are the sum of squares and df for the error for the associated ANOVA model.
- We can compute $MSE_a = SSE_a/df_{e;a}$ in the obvious way.
- In order to produce the results presented above, it is necessary to find $X^*$.
  - Let $\hat{\delta}_u = P_Q Y$ be the estimator of $\delta$ under the associated ANOVA model.
  - Then, the residual vector for the full ANOVA model is given by

$$
Y - \hat{\delta}_u = P_{Q^\perp} Y = Y^*.
$$

- To find $X^* = P_{Q^\perp} X$, we apply the same operation to the $x$ values that we apply to the $y$ values when we find the residuals $Y^*$ for the associated ANOVA model.
Inference Setup

- Next, consider testing the null hypothesis that $\delta \in T$, a known $t$-dimensional subspace of $Q$.
- Let

$$X^{**} = P_{T\perp} X, \ Y^{**} = P_{T\perp} Y.$$ 

- Note that if $X^*$ has full rank then $X^{**}$ also does.
- As written earlier, $Y^{**}$ is the vector of residuals for the reduced form of the associated ANOVA model and $X^{**}$ is the same function of the columns of $X$ as $Y^{**}$ is of $Y$.
- Let $W$ be the subspace of possible values of $\mu$ under the null hypothesis.
- By the same argument from earlier, we see that

$$\tilde{\mu} = P_W Y = \tilde{\delta}_u + X^{**} \tilde{\gamma}$$

$$\tilde{\delta}_u = P_T Y$$

$$\tilde{\gamma} = (X^{**T}X^{**})^{-1}X^{**T}Y = (X^{**T}X^{**})^{-1}X^{**T}Y^{**}$$

$$\|\tilde{\mu}\|^2 = \|\tilde{\mu}_u\|^2 + \|X^{**}\tilde{\gamma}\|^2$$

$$k = \dim(W) = t + s.$$
Test Statistic

Therefore, for testing $\delta \in T$ for the ANCOVA model, the sum of squares, df, and mean square for the hypothesis are given, respectively, by

$$SSH_a = \|\hat{\mu}\|^2 - \|\tilde{\mu}\|^2$$

$$= \left(\|\hat{\delta}_u\|^2 + \|X^*\hat{\gamma}\|^2\right) - \left(\|\tilde{\delta}_u\|^2 + \|X^{**}\tilde{\gamma}\|^2\right)$$

$$= SSH_u + \|X^*\hat{\gamma}\|^2 - \|X^{**}\tilde{\gamma}\|^2$$

$$= SSH_u + \|X^*\hat{\gamma} - X^{**}\tilde{\gamma}\|^2$$

$$df_{h;a} = (q + s) - (t + s) = q - t = df_{h;u}$$

$$MSH_a = SSH_a / df_{h;a}.$$

$SSH_u$ and $df_{h;u}$ are the sums of squares and df, respectively, for the hypothesis in the associated ANOVA model.

Note that the df for the hypothesis in the associated ANCOVA model is the same as the df for the hypothesis in the associated ANOVA model and that if $X^{**} = X^*$, then the sum of squares and mean square for the hypothesis in the ANCOVA are the same as the sum of squares and the mean square for the hypothesis in the associated ANOVA model.

As usual, the optimal size $\alpha$ test for this hypothesis rejects if

$$F^* = \frac{MSH_a}{MSE_a} > F_{1-\alpha;df_{h;a},df_{e;a}}.$$
Outline of Topics

1. Linear Model Theory
2. Single-Factor ANCOVA
3. Generalizations of ANCOVA
Example: Goats Data

A veterinarian carried out an experiment on a goat farm to determine whether a standard worm drenching program (to prevent the goats from obtaining worms) was adequate. Forty goats were used in each experiment, where twenty were chosen completely at random and drenched according to the standard program, while the remaining twenty were drenched more frequently. The goats were individually tagged, and weighed at the start and end of the year-long study. The objective in the experiment is to compare the liveweight gains between the two treatments. This comparison could be made using an ANOVA; however, a commonly observed biological phenomenon could allow us to increase the precision of the analysis. Namely, the lighter animals gain more weight than the heavier animals, so we have a “regression to the norm” setting. Since this can be assumed to occur within both treatment groups, it is appropriate to adjust the analysis to use that covariate information.
Example: Goats Data

To understand why incorporating the covariate information might be effective, look at the plot of the data below. Here, we plotted weight gain of the goats separately for each treatment (either “standard” or “intensive”). It is evident that the error terms – as shown by the scatter around the treatment means given the respective horizontal lines – are fairly large. Thus, this indicates a large error term variance.
Example: Goats Data

Suppose that we now utilize the goats’ initial weight. The scatterplot below shows the goats’ weight gain versus their initial weight, with different colors corresponding to the different treatments. Note that the two treatment regression lines happen to be linear, although this need not be the case. The scatter around the treatment regression lines is much less than the scatter around the treatment means shown on the previous slide as a result of the goats’ weight gain being highly linearly related to the initial weight. Thus, the figure on the previous slide shows the (relatively) larger error variance under a single-factor ANOVA while the figure below shows the smaller error variance under a single factor ANCOVA.
Local Control with Covariates

- The choice of covariates is important.
  - Covariates commonly used with human subjects include prestudy attitudes, age, socioeconomic status, weight, and aptitude.
  - Covariates commonly used with business units include previous quarterly sales, employee salaries, and number of employees at individual stores.

- If the chosen covariates have no relation to the response variable, then nothing stands to be gained by using an ANCOVA and one may as well use an ANOVA.

- We know that local control practices reduce experimental error variance and increase the precision for estimates of treatment means and tests of hypotheses.

- Covariates are often used to select and group units to control experimental error variation.
Covariates Unaffected by Treatments

- For a clear interpretation of the results, a covariate should be observed before the study; or if observed during the study, then it should not be influenced by the treatments in any way.
- Whenever a covariate is affected by the treatments, an ANCOVA will fail to show some (or most) of the effects that the treatments had on the response variable.
  - Therefore, an uncritical analysis may be badly misleading.
- Typically, when a covariate is unaffected by the treatments, the distribution of subjects along the axis of the covariate will be roughly similar for all treatments and subject only to chance variation.
Example: Training Methods Study

Suppose that a testing program is interested in studying the effect of a particular training method on students’ scores. Two training methods were used and 12 students were assigned to the training method at random; i.e., six students were assigned to each training method. At the end of the program, a score was obtained to measure their amount of learning. The researcher decided to use the amount of time devoted to studying as a covariate, but found that the training method had virtually no effect. The second training method involved a computer-assisted learning program which, generally, necessitated the student needing more time to learn the program. In other words, both the learning score and the amount of study time were influenced by the treatment in this case. As a result the high correlation between the amount of study time and learning score, the marginal treatment effect of training method was small and the test for treatment effects have no significant difference between the two method. A figure of these data appears on the next slide.
Example: Training Methods Study

- To the right is a scatterplot of the training methods data.
- Method 1 is a paper-and-pencil method while Method 2 is the computer-assisted learning program.
- Method 2 needed more time devoted to learning the program, so observations for the two treatments tend to be concentrated over different intervals on the covariate’s axis.
Development of Single-Factor ANCOVA Model

- There are various ways to define a single-factor ANCOVA model and we’ll present one such model where the design could, potentially, be unbalanced.
- Let the number of subjects for the $i^{\text{th}}$ factor level be denoted by $n_i$ and the total number of cases be $N = \sum_i n_i$.
- Recall that the single-factor ANOVA model with fixed effects is given by
  \[ Y_{ij} = \mu + \tau_i + \epsilon_{ij} \]
- Let the response and covariate level for the $j^{\text{th}}$ case of the $i^{\text{th}}$ factor level be given by $Y_{ij}$ and $x_{ij}$, respectively.
- The covariance model starts with the above ANOVA model and adds another term reflecting the relationship between the response and the covariate.
- A first approximation is the linear relationship
  \[ Y_{ij} = \mu + \tau_i + \beta x_{ij} + \epsilon_{ij} \]
- In the above, $\beta$ is a regression coefficient for the relation between $Y_{ij}$ and $x_{ij}$, however, the constant $\mu$ is no longer an overall mean.
- We can make the constant an overall mean and simplify calculations by centering the covariate about the overall mean $\bar{x}$...
The single-factor ANCOVA model with fixed effects is:

\[ Y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}) + \epsilon_{ij}, \]  

(2)

where

- \( \mu \) is the overall mean (a constant);
- \( i = 1, \ldots, r \) and \( j = 1, \ldots, n_i \);
- \( \tau_i \) are the (fixed) treatment effects subject to the constraint \( \sum_{i=1}^{r} \tau_i = 0 \);
- \( \beta \) is a regression coefficient for the relationship between the response and covariate;
- \( x_{ij} \) are constants; and
- \( \epsilon_{ij} \) are the (random) errors and are iid normal with mean 0 and variance \( \sigma^2 \).

Since \( \epsilon_{ij} \) is the only random variable on the right-hand side of the ANCOVA model, it follows that

\[ \mathbb{E}(Y_{ij}) = \mu + \tau_i + \beta(x_{ij} - \bar{x}) \equiv \mu_{ij} \]
\[ \text{Var}(Y_{ij}) = \sigma^2 \]
Comparisons of Treatment Effects

▶ In ANOVA, all observations have the same mean response; i.e., $E(Y_{ij}) = \mu_i$ for all $j$.

▶ In ANCOVA, the expected response for the $i^{th}$ treatment is a regression line; i.e., $E(Y_{ij}) = \mu_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x})$.

▶ While we no longer characterize the mean response with the $i^{th}$ treatment since it varies with $x$, we can still measure the effect of any treatment compared with any other by a single number.

▶ The difference between two mean responses is the same for all values of $x$ because the slopes of the regression lines are equal.

▶ Hence, we can measure the difference at any convenient value of $x$, say, $x = \bar{x}$:

$$\mu + \tau_k - (\mu + \tau_l) = \tau_k - \tau_l \text{ for } k \neq l$$

▶ Thus, $\tau_k - \tau_l$ measures how much higher or lower the mean response is with treatment $k$ relative to treatment $l$ for any value of $x$.

▶ It follows that when all treatments have the same mean response for $x$, then the treatment regression lines must be identical and, hence, $\tau_i = 0$ for all $i$. 
The assumption in ANCOVA is that all treatment regression lines have the same slope.

Without this assumption, the difference between the effects of two treatments cannot be summarized by a single number based on the main effects; e.g., $\tau_k - \tau_l$.

When the treatments interact with the covariate – resulting in nonparallel slopes – ANCOVA is not appropriate and, instead, separate treatment regression lines need to be estimated and then compared.

Similar difficulties occur with nonlinear relationships, where the inferences regarding the responses must include a complete description involving the effects of the treatments and the covariate.
Appropriateness of ANCOVA

The key assumptions in ANCOVA are:

1. Normality of error terms
2. Equality of error variances for different treatments
3. Uncorrelatedness of error terms
4. Linearity of regression relation with covariates (i.e., appropriate model specification)
5. Equality of slopes of the different treatment regression lines

The first four assumptions are checked using the standard residual diagnostics and remedial measures.

The last assumption is particularly important for ANCOVA and will be developed in detail.
SS Partitioning for ANCOVA

- We cannot directly apply the single-factor ANOVA, but rather need to adjust the treatment SS after fitting the covariate.

- Consider the following models, which we will label $f$, $r_1$, and $r_2$, respectively:
  - **full model:**
    \[ Y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}) + \epsilon_{ij} \]
  - **reduced model without covariate:**
    \[ Y_{ij} = \mu + \tau_i + \epsilon_{ij} \]
  - **reduced model without treatment effects:**
    \[ Y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{..}) + \epsilon_{ij} \]

- The reduced model without the covariate ($r_1$) is required to assess the influence of the covariate, while the reduced model without treatment effects ($r_2$) is required to assess the significance of the treatment effects in the presence of covariates.
SS Partitioning for ANCOVA

- Least squares estimates can be derived for each of the three models: \( f \), \( r_1 \), and \( r_2 \).
- Using those least squares estimates, we have the following SSE quantities:
  - **full model:**
    \[
    SSE_f = \sum_{i} \sum_{j} \left[ y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}(x_{ij} - \bar{x}) \right]^2,
    \]
    with \( N - r - 1 \) df.
  - **reduced model without covariate:**
    \[
    SSE_{r_1} = \sum_{i} \sum_{j} \left[ y_{ij} - \hat{\mu} - \hat{\tau}_i \right]^2,
    \]
    with \( N - r \) df.
  - **reduced model without treatment effects:**
    \[
    SSE_{r_2} = \sum_{i} \sum_{j} \left[ y_{ij} - \hat{\mu} - \hat{\beta}(x_{ij} - \bar{x}) \right]^2,
    \]
    with \( N - 2 \) df.
**SS Partitioning for ANCOVA**

▶ The SS reduction due to the addition of the covariate $x$ to the model is obtained as the following difference:

$$SS(Covariate) = SSE_{r_1} - SSE_f,$$

with 1 df.

▶ The adjusted SS after fitting the covariate is:

$$SSTr(Adjusted) = SSE_{r_2} - SSE_f,$$

with $r - 1$ df.

▶ Mean squares for all of the previous sources of variability are found (as usual) by dividing the SS quantity by the respective df.

▶ Note that construction of the correct ANCOVA table (with adjustments) amounts to using the Type III SS quantities.
### ANCOVA Table and Testing

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SS(Covariate)</td>
<td>MS(Covariate)</td>
</tr>
<tr>
<td>Treatment</td>
<td>$r - 1$</td>
<td>SSTr(Adjusted)</td>
<td>MSTr(Adjusted)</td>
</tr>
<tr>
<td>Error</td>
<td>$N - r - 1$</td>
<td>SSEf</td>
<td>MSEf</td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td>SSTot</td>
<td></td>
</tr>
</tbody>
</table>

- **The test for reduction in variance due to the covariate**

  $H_0 : \beta = 0$

  $H_A : \beta \neq 0$

  has the test statistic $F^* = \frac{\text{MS(Covariate)}}{\text{MSE}_f} \sim F_{1, N-r-1}$.

- **The test for adjusted treatment effects**

  $H_0 : \tau_1 = \cdots = \tau_r = 0$

  $H_A : \text{not all } \tau_j \text{ equal } 0$

  has the test statistic $F^* = \frac{\text{MSTr(Adjusted)}}{\text{MSE}_f} \sim F_{r-1, N-r-1}$. 


Testing for Common Slopes

The linear model with different regression coefficients for each treatment group (call it model $r_3$) is

$$Y_{ij} = \mu + \tau_i + \beta_i(x_{ij} - \bar{x}..) + \epsilon_{ij},$$  \hspace{1cm} (3)

which is exactly the same as the single-factor ANCOVA model, except $\beta$ has been replaced by $\beta_i$ (the regression coefficient for the $i^{th}$ treatment).

Let $SSE_{r_3}$ be the SS for experimental error, which has $N - 2r$ df.

The SS to test homogeneity of regression coefficients for the treatment groups is

$$SS(\text{Homogeneity}) = SSE_f - SSE_{r_3}$$

The test for homogeneity (or equality) of regression coefficients $H_0 : \beta_1 = \cdots = \beta_r$

$H_A : \text{at least one } \beta_j \text{ is different}$

has the test statistic $F^* = \frac{\text{MS(Homogeneity)}}{\text{MSE}_{r_3}} \sim F_{r-1, N-2r}$. 
**ANCOVA Table with Interaction**

- The required SS for testing homogeneity of regression coefficients can be found by constructing the ANCOVA table with a treatment by covariate interaction term and, subsequently, computing the Type III SS.

- Below is the ANCOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>$r - 1$</td>
<td>SSTR(Adjusted)</td>
<td>MSTr(Adjusted)</td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
<td>SS(Regression)</td>
<td>MS(Regression)</td>
</tr>
<tr>
<td>Treatment x Regression</td>
<td>$r - 1$</td>
<td>SS(Homogeneity)</td>
<td>MS(Homogeneity)</td>
</tr>
<tr>
<td>Error</td>
<td>$N - 2r$</td>
<td>SSE$_{r3}$</td>
<td>MSE$_{r3}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td>SSTot</td>
<td></td>
</tr>
</tbody>
</table>
Treatment Means Adjusted to Covariate

- The estimated regression equation for the \( i \)th treatment group is
  \[
  \bar{y}_i = \bar{y}_i - \hat{\beta}(x - \bar{x}..),
  \]
  where \( x \) is an arbitrary value of the covariate that falls within the domain of the observed data; i.e., we are not extrapolating.

- The estimates of treatment means are adjusted to a common value for the covariate if inclusion of the covariate in the model significantly reduces the experimental error variance.

- The treatments means are found at the value of \( \bar{x}_i \), and can be adjusted to any value of the covariate, but usually we adjust to the overall mean \( \bar{x}.. \):
  \[
  \bar{y}_{i;adj} = \bar{y}_i - \hat{\beta}(\bar{x}_i - \bar{x}..)
  \]
Standard Errors for the Adjusted Treatment Means

- Denote the SS for the experimental error from an ANOVA based on the covariate \( x \) by

\[
E_{xx} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2
\]

- The standard error estimator for an adjusted treatment mean is

\[
s_{\bar{y}_i;\text{adj}} = \sqrt{\text{MSE} \left[ \frac{1}{n_i} + \frac{(\bar{x}_i - \bar{x})^2}{E_{xx}} \right]} \quad (4)
\]

- The standard error estimator for the difference between two adjusted treatment means is

\[
s_{(\bar{y}_{i1;\text{adj}} - \bar{y}_{i2;\text{adj}})} = \sqrt{\text{MSE} \left[ \frac{1}{n_{i1}} + \frac{1}{n_{i2}} + \frac{(\bar{x}_{i1} - \bar{x}_{i2})^2}{E_{xx}} \right]} \quad (5)
\]
The design matrix for ANCOVA has a special structure that fits nicely into a model composed of indicator variables.

For $k = 1, 2, \ldots, r - 1$, define the following (ternary) indicator variables:

$$I_{ij,k} = \begin{cases} 
1 & \text{if case } i \text{ is from treatment } j = k \\
0 & \text{if case } i \text{ is from treatment } j = r \\
-1 & \text{otherwise}
\end{cases}$$

The ANCOVA model can now be expressed as follows:

$$Y_{ij} = \mu + \tau_1 I_{ij,1} + \cdots + \tau_{r-1} I_{ij,r-1} + \beta(x_{ij} - \bar{x}.\,) + \epsilon_{ij}$$

Note that the treatment effects $\tau_1, \ldots, \tau_{r-1}$ are the regression coefficients for the indicator variables.

This regression format becomes helpful in understanding the test for parallel slopes.
Another Test for Parallel Slopes

- Testing for parallel slopes in an ANCOVA means that all treatment regression lines have the same slope $\beta$.
- The regression model on the previous slide can be generalized to allow for different slopes for the treatments by introducing cross-product interaction terms.
- Letting $\gamma_1, \ldots, \gamma_{r-1}$ be the regression coefficients for the interaction terms, we have our generalized model as

$$Y_{ij} = \mu + \tau_1 I_{ij,1} + \cdots + \tau_{r-1} I_{ij,r-1} + \beta(x_{ij} - \bar{x}..)$$

$$+ \gamma_1 I_{ij,1}(x_{ij} - \bar{x}..) + \cdots + \gamma_{r-1} I_{ij,r-1}(x_{ij} - \bar{x}..) + \epsilon_{ij}$$

- The test for parallel slopes is then:

$$H_0 : \gamma_1 = \cdots = \gamma_{r-1} = 0$$

$$H_A : \text{at least one } \gamma_j \text{ is not 0, for } j = 1, \ldots, r - 1.$$ 

- We can then apply the general linear $F$-test where the model above is the “full” model and the model on the previous slide without the interaction terms is the “reduced” model.
Example: Cracker Promotion Study

A company studied the effects of three different types of promotions on sales of its crackers:

- Treatment 1: Sampling of product by customers in store and regular shelf space
- Treatment 2: Additional shelf space in regular location
- Treatment 3: Special display shelves at ends of aisle in addition to regular shelf space

$N = 15$ stores were selected for the study and a CRD was used. Each store was randomly assigned one of the promotion types, with $n_i = n = 5$ stores assigned to each type of promotion. Other relevant conditions under the control of the company (e.g., price and advertising) were kept the same for all stores in the study. Data on the number of cases of the product sold during the promotional period ($y$) and data on the sales of the product in the preceding period ($x$, the covariate) are presented below.

<table>
<thead>
<tr>
<th>Treatment ($i$)</th>
<th>Store ($j$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y_{i1}$</td>
<td>$x_{i1}$</td>
<td>$y_{i2}$</td>
<td>$x_{i2}$</td>
<td>$y_{i3}$</td>
<td>$x_{i3}$</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>21</td>
<td>39</td>
<td>26</td>
<td>36</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>34</td>
<td>38</td>
<td>26</td>
<td>38</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>23</td>
<td>32</td>
<td>29</td>
<td>31</td>
<td>30</td>
</tr>
</tbody>
</table>
Example: Cracker Promotion Study

- To the right is a scatterplot of the cracker promotion data.
- It looks like there is an effect due to the different treatments.
- Moreover, the change in the response as the covariate changes appears to be similar for each treatment (i.e., parallel slopes).
Above is the ANCOVA table for the sales during the treatment period using sales from the preceding period as a covariate (you can ignore the (Intercept) row). This uses the adjusted treatment SS; i.e., the Type III SS. For testing $H_0 : \beta = 0$ – which determines the significance of the reduction in error variance due to sales from the preceding period – we have $F^* = 76.723 \sim F_{1,11}$ with a $p$-value of 2.731e-06, which is highly significant. Therefore, the addition of the covariate has significantly reduced experimental error variability. For testing $H_0 : \tau_1 = \tau_2 = \tau_3$ – which determines equality of the adjusted treatment means – we have $F^* = 59.483 \sim F_{2,11}$ with a $p$-value of 1.264e-06, which is, again, highly significant. Therefore, the treatment means adjusted for sales from the previous period are significantly different.
Example: Cracker Promotion Study

Anova Table (Type III tests)

Response: y

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>48.742</td>
<td>1</td>
<td>13.9172</td>
<td>0.004693 **</td>
</tr>
<tr>
<td>treat</td>
<td>1.263</td>
<td>2</td>
<td>0.1803</td>
<td>0.837923</td>
</tr>
<tr>
<td>x</td>
<td>243.141</td>
<td>1</td>
<td>69.4230</td>
<td>1.597e-05 ***</td>
</tr>
<tr>
<td>treat:x</td>
<td>7.050</td>
<td>2</td>
<td>1.0065</td>
<td>0.403181</td>
</tr>
<tr>
<td>Residuals</td>
<td>31.521</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Above is the ANCOVA table which includes the treatment by covariate interaction term; i.e., the interaction between promotional treatment and sales from the previous period. You can, again, ignore the (Intercept) row. The row for the interaction gives us the results for the test of $H_0 : \beta_1 = \beta_2 = \beta_3$. This test has $F^* = 1.007 \sim F_{2,9}$, which has a $p$-value of 0.403. Therefore, this test is not statistically significant and we can claim that the assumption of equal slopes is appropriate.
Example: Cracker Promotion Study

- The estimated regression equations for the three promotional treatments are:

\[
\hat{y}_{1j} = 38.2 + 0.899(x_{1j} - 23.2)
\]
\[
\hat{y}_{2j} = 36.0 + 0.899(x_{2j} - 26.4)
\]
\[
\hat{y}_{3j} = 27.2 + 0.899(x_{3j} - 25.4)
\]

- To the right, the three estimated regression equations are overlaid on the scatterplot of the data.
Example: Cracker Promotion Study

The treatment means adjusted to the mean sales data from the previous period ($\bar{x}_. = 25.0$) are

$$\bar{y}_{1; adj} = 38.2 - 0.889(23.2 - 25.0) = 39.82$$
$$\bar{y}_{2; adj} = 36.0 - 0.889(26.4 - 25.0) = 34.74$$
$$\bar{y}_{3; adj} = 27.2 - 0.889(25.4 - 25.0) = 26.84$$

The standard errors for the above adjusted treatment means are

$$s_{\bar{y}_{1; adj}} = \sqrt{3.51 \left[ \frac{1}{5} + \frac{(23.2 - 25.0)^2}{333.2} \right]} = 0.86$$
$$s_{\bar{y}_{2; adj}} = \sqrt{3.51 \left[ \frac{1}{5} + \frac{(26.4 - 25.0)^2}{333.2} \right]} = 0.85$$
$$s_{\bar{y}_{3; adj}} = \sqrt{3.51 \left[ \frac{1}{5} + \frac{(25.4 - 25.0)^2}{333.2} \right]} = 0.84$$
Example: Cracker Promotion Study

The pairwise difference between treatment means adjusted to the mean sales data from the previous period are

\[ \bar{y}_{1;adj} - \bar{y}_{2;adj} = 39.82 - 34.74 = 5.08 \]
\[ \bar{y}_{1;adj} - \bar{y}_{3;adj} = 39.82 - 36.84 = 2.98 \]
\[ \bar{y}_{2;adj} - \bar{y}_{3;adj} = 34.74 - 26.84 = 7.90 \]

The standard errors for the above differences are

\[ s(\bar{y}_{1;adj} - \bar{y}_{2;adj}) = \sqrt{3.51 \left[ \frac{1}{5} + \frac{1}{5} + \frac{(23.2 - 26.4)^2}{333.2} \right]} = 1.23 \]
\[ s(\bar{y}_{1;adj} - \bar{y}_{3;adj}) = \sqrt{3.51 \left[ \frac{1}{5} + \frac{1}{5} + \frac{(23.2 - 25.4)^2}{333.2} \right]} = 1.21 \]
\[ s(\bar{y}_{2;adj} - \bar{y}_{3;adj}) = \sqrt{3.51 \left[ \frac{1}{5} + \frac{1}{5} + \frac{(26.4 - 25.4)^2}{333.2} \right]} = 1.19 \]
Outline of Topics

1. Linear Model Theory
2. Single-Factor ANCOVA
3. Generalizations of ANCOVA
Overview

- Thus far, we have discussed only the single-factor ANCOVA setting.
- However, we can have a measured covariate (or covariates) in most of the design settings that we have discussed.
- In this lecture, we will introduce a few extensions to ANCOVA, specifically when we have more than one covariate, more than one factor, and when a blocking factor is present.
Multiple Covariates

- The single-factor ANCOVA model is usually sufficient to reduce error variability substantially.
- However, the model can be extended in a straightforward manner to include two (or more) covariates.
- The single-factor ANCOVA model with \( p \) covariates is:

\[
Y_{ij} = \mu + \tau_i + \sum_{k=1}^{p} \beta_k (x_{ijk} - \bar{x}_{..p}) + \epsilon_{ij},
\]

where all of the quantities in the model are the same as for the model with one covariate, except that

- observation \( j \) from treatment \( i \) has \( p \) measured covariates \( x_{ij1}, \ldots, x_{ijp} \);
- \( \beta_1, \ldots, \beta_p \) are the slopes estimated for the measured covariates \( x_{ij1}, \ldots, x_{ijp} \), respectively; and
- \( \bar{x}_{..1}, \ldots, \bar{x}_{..p} \) are the respective means of the \( p \) covariates.

- Assumptions and estimation for the above model are nearly identical to the single covariate model, except that the fitted regression surfaces using the estimated \( \beta_1, \ldots, \beta_k \) coefficients at different treatment levels must be parallel.
  - In other words, we had the parallel slopes assumption in the single covariate setting, now we have the parallel surfaces assumption.
Nonlinearity

▶ The linear relationship between the response \((y)\) and a covariate \((x)\) is not essential to ANCOVA.

▶ We can capture nonlinearities by defining a polynomial relationship using the previous model as follows:

\[
Y_{ij} = \mu + \tau_i + \sum_{k=1}^{q} \beta_k (x_{ij} - \bar{x}.)^k + \epsilon_{ij},
\]

(7)

which has a polynomial relationship (in terms of the covariate) of order \(q\).

▶ For example, \(q = 2\) would yield a quadratic relation.

▶ Linearity of the relation leads to simpler analysis and is often a good approximation, but if it is not, then the above model can be used.

▶ But again, ANCOVA does require that the treatment response functions be parallel; in other words, there must not be an interaction effect between the treatment and covariates.
ANCOVA with a Blocking Effect

As we noted earlier, ANCOVA can be applied to almost any experimental design with a straightforward extension of the established principles.

However, when using a blocking design, you must have more than one EU per treatment within each block in order to test for equality of slopes among treatment groups.

The ANCOVA model with a fixed effect and blocking factor is:

\[ Y_{ij} = \mu + \tau_i + \rho_j + \beta(x_{ij} - \bar{x}..) + \epsilon_{ij}, \quad (8) \]

where

- \( \mu \) is the overall mean (a constant);
- \( i = 1, \ldots, r \) and \( j = 1, \ldots, n \) (we assume balancedness for simplicity);
- \( \tau_i \) are the (fixed) treatment effects subject to the constraint \( \sum_{i=1}^{r} \tau_i = 0 \);
- \( \rho_j \) is the block effects subject to the constraint \( \sum_{j=1}^{n} \rho_j = 0 \);
- \( \beta \) is a regression coefficient for the relationship between \( y \) and \( x \);
- \( x_{ij} \) are the covariates; and
- \( \epsilon_{ij} \) are the (random) errors and are \( iid \) normal with mean 0 and variance \( \sigma^2 \).
SS Partitioning

Consider the following models, which we will label $f$, $r_1$, $r_2$, and $r_3$, respectively:

- **full model:**
  
  $$Y_{ij} = \mu + \tau_i + \rho_j + \beta(x_{ij} - \bar{x}..) + \epsilon_{ij},$$
  
  where the $\text{SSE}_f$ has $(n - 1)(r - 1) - 1$ df.

- **regular RCBD model without covariate:**
  
  $$Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$$
  
  where the $\text{SSE}_{r_1}$ has $(n - 1)(r - 1)$ df.

- **reduced model without treatment effects:**
  
  $$Y_{ij} = \mu + \rho_j + \beta(x_{ij} - \bar{x}..) + \epsilon_{ij},$$
  
  where the $\text{SSE}_{r_2}$ has $(N - n - 1)$ df.

- **reduced model without block effects:**
  
  $$Y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}..) + \epsilon_{ij},$$
  
  where the $\text{SSE}_{r_3}$ has $(N - r - 1)$ df.

The three reduced models are used to assess the influence of the covariate, treatment effects, and block effects, respectively.
SS Partitioning

- We can now further define the necessary SS quantities:
  - SS for adding the covariate $x$:
    \[ SS(\text{Covariate}) = SSE_{r_1} - SSE_f \]
    with 1 df.
  - Adjusted treatment SS after fitting the covariate and block effect:
    \[ SSTr(\text{Adjusted}) = SSE_{r_2} - SSE_f \]
    with $(r - 1)$ df.
  - Adjusted block SS after fitting the covariate and treatment effect:
    \[ SSBlk(\text{Adjusted}) = SSE_{r_3} - SSE_f \]
    with $(n - 1)$ df.
Inference

▶ For testing

\[ H_0 : \beta = 0 \]
\[ H_A : \beta \neq 0 \]
we use the test statistic
\[ F^* = \frac{\text{MS(Covariate)}}{\text{MSE}} \sim F_{1, N-r-n}. \]

▶ For testing

\[ H_0 : \tau_1 = \cdots = \tau_r = 0 \]
\[ H_A : \text{at least one } \tau_i \text{ is different} \]
we use the test statistic
\[ F^* = \frac{\text{MSTr(Adjusted)}}{\text{MSE}} \sim F_{r-1, N-r-n}. \]

▶ For testing

\[ H_0 : \rho_1 = \cdots = \rho_n = 0 \]
\[ H_A : \text{at least one } \rho_j \text{ is different} \]
we use the test statistic
\[ F^* = \frac{\text{MSBlk(Adjusted)}}{\text{MSE}} \sim F_{n-1, N-r-n}. \]
**ANCOVA Table and Standard Errors**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SS(Covariate)</td>
<td>MS(Covariate)</td>
</tr>
<tr>
<td>Block</td>
<td>n − 1</td>
<td>SSB(Adjusted)</td>
<td>MSB(Adjusted)</td>
</tr>
<tr>
<td>Treatment</td>
<td>r − 1</td>
<td>SST(Adjusted)</td>
<td>MST(Adjusted)</td>
</tr>
<tr>
<td>Error</td>
<td>N − r − n</td>
<td>SSE_f</td>
<td>MSE_f</td>
</tr>
<tr>
<td>Total</td>
<td>N − 1</td>
<td>SSTot</td>
<td></td>
</tr>
</tbody>
</table>

- Estimated standard errors for adjusted treatment means are found analogously to the single-factor ANCOVA without a blocking factor.
- Denote the SS for the experimental error from an ANOVA based on the covariate $x$ by

$$E_{xx} = \sum_{i=1}^{r} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

- The standard error estimator for an adjusted treatment mean is

$$s_{\bar{y}_i; adj} = \sqrt{\frac{1}{n} + \frac{(\bar{x}_i - \bar{x}..)^2}{E_{xx}}}$$

- The standard error estimator for the difference between two adjusted treatment means is

$$s(\bar{y}_{i1; adj} - \bar{y}_{i2; adj}) = \sqrt{\frac{2}{n} + \frac{(\bar{x}_{i1} - \bar{x}_{i2})^2}{E_{xx}}}$$
Example: Study of Soil Nutrients

Management methods on forest and range watersheds affect the nutrients in any vegetations and soil-type complex. The availability of certain soil nutrients in these watershed soils is evaluated by a pot culture technique in a greenhouse with barley plants. A researcher wants to determine the availability of nitrogen and phosphorus in a watershed dominated by a certain type of vegetation. He collected sample soils under the vegetation and composited the samples for a pot culture evaluation of nitrogen and phosphorus availability. \( r = 4 \) treatments were used for the study: (1) check (which means no fertilizer added); (2) full (a complete fertilizer); (3) nitrogen (\( N_0 \)) omitted from full; and (4) phosphorus (\( P_0 \)) omitted from full. The nutrient treatments were added as solutions to the soil, mixed, and placed in plastic pots in a greenhouse. The treatment pots were arranged on a greenhouse bench in an RCBD to control experimental error variation caused by gradients in light and temperature in the greenhouse. The barley plants were grown in the pots for seven weeks, after which they were harvested, dried, and weighed. A leaf blight infected the plants part way through the experiment, which was assumed to affect the growth of the plants. Thus, the percentage of blight was recorded. The total dry weight is the response (\( y \)) and the leaf blight is a measured covariate (\( x \)).
Example: Study of Soil Nutrients

Below is a table of the response and covariate. The last row gives the means of these variables for each treatment.

<table>
<thead>
<tr>
<th>Block ( (j) )</th>
<th>Treatment (( i ))</th>
<th>Check ( y_{1j} ) ( x_{1j} )</th>
<th>Full ( y_{2j} ) ( x_{2j} )</th>
<th>( N_0 ) ( y_{3j} ) ( x_{3j} )</th>
<th>( P_0 ) ( y_{4j} ) ( x_{4j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Check</td>
<td>23.1 13</td>
<td>30.1 7</td>
<td>26.4 10</td>
<td>26.2 8</td>
</tr>
<tr>
<td>2</td>
<td>Full</td>
<td>20.9 12</td>
<td>31.8 5</td>
<td>27.2 9</td>
<td>25.3 9</td>
</tr>
<tr>
<td>3</td>
<td>( N_0 )</td>
<td>28.3 7</td>
<td>32.4 6</td>
<td>28.6 6</td>
<td>29.7 7</td>
</tr>
<tr>
<td>4</td>
<td>( P_0 )</td>
<td>25.0 9</td>
<td>30.6 7</td>
<td>28.5 6</td>
<td>26.0 7</td>
</tr>
<tr>
<td>5</td>
<td>Mean</td>
<td>25.1 8</td>
<td>27.5 9</td>
<td>30.8 5</td>
<td>24.9 9</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>24.48 9.8</td>
<td>20.48 6.8</td>
<td>28.3 7.2</td>
<td>26.42 8.0</td>
</tr>
</tbody>
</table>
Example: Study of Soil Nutrients

- To the right is a scatterplot of the soil nutrients study.
- Different colors are used for the treatments and different plotting characters are used for the blocks.
- It does look like there is a treatment effect and that the assumption of parallel slopes will be appropriate (this will be tested in a moment); however, accounting for a blocking effect is not as easy to discern from this plot.
Example: Study of Soil Nutrients

- To the right is a boxplot of the dry weights by block.
- This does reveal that a blocking effect (at least marginally) could help explain a significant amount of experimental error.
- It is somewhat of a stretch to assume common variance here – mainly because of blocks 2 and 3 – but we will proceed to analyze the raw data as is.
  - Of course, you could explore a variance-stabilizing transformation as a remedial measure.
Example: Study of Soil Nutrients

Anova Table (Type III tests)

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>276.725</td>
<td>1</td>
<td>242.0235</td>
<td>7.759e-09 ***</td>
</tr>
<tr>
<td>x</td>
<td>26.590</td>
<td>1</td>
<td>23.2554</td>
<td>0.0005338 ***</td>
</tr>
<tr>
<td>block</td>
<td>7.786</td>
<td>4</td>
<td>1.7024</td>
<td>0.2191876</td>
</tr>
<tr>
<td>treat</td>
<td>24.860</td>
<td>3</td>
<td>7.2474</td>
<td>0.0059198 **</td>
</tr>
<tr>
<td>Residuals</td>
<td>12.577</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Above is the ANOVA table with the adjusted SS. The significance of the covariate is tested with the statistic $F^* = 23.255 \sim F_{1,11}$, which has a $p$-value of 0.001. Thus, the relationship between the percent blight damaged leaf area and dry matter production of the barley plants is highly significant. For the null hypothesis of no differences among treatment means, we have $F^* = 7.247 \sim F_{3,11}$. The corresponding $p$-value is 0.006, which means that there is at least one treatment mean significantly different from the others. If one tests the block effect, the test statistic is $F^* = 1.702 \sim F_{4,11}$ with a $p$-value of 0.219. This is not a significant effect; however, we retain the blocking effect since (a) it still helps explain some variability (albeit not a significant amount) and (b) we want to illustrate its inclusion in an experiment.
Above is the ANOVA table with the adjusted SS for testing homogeneity of the slopes. We are interested in the row for the treatment by covariate interaction. The test statistic is $F^* = 0.2502 \sim F_{3,8}$, which has a $p$-value of 0.859. This implies we would fail to reject the null hypothesis and claim that there is no significant differences between the slopes for different treatments. In other words, we can assume the slopes are parallel.
Example: Study of Soil Nutrients

The estimated regression equations for the four nutrient treatments are:

\[
\hat{y}_{1j} = 24.48 - 0.863(x_{1j} - 9.8)
\]
\[
\hat{y}_{2j} = 30.48 - 0.863(x_{2j} - 6.8)
\]
\[
\hat{y}_{3j} = 28.30 - 0.863(x_{3j} - 7.2)
\]
\[
\hat{y}_{4j} = 26.42 - 0.863(x_{4j} - 8.0)
\]

The four estimated regression equations are overlaid on the scatterplot to the right, where we have removed the different plotting characters for block to help with the visualization.
Example: Study of Soil Nutrients

We can calculate standard errors for the adjusted treatment means and their differences in the same way as for the single-factor ANCOVA. We summarize those standard errors below without repeating the formulas.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Adjusted Mean</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check</td>
<td>26.08</td>
<td>0.58</td>
</tr>
<tr>
<td>Full</td>
<td>29.49</td>
<td>0.52</td>
</tr>
<tr>
<td>N₀</td>
<td>27.65</td>
<td>0.50</td>
</tr>
<tr>
<td>P₀</td>
<td>26.46</td>
<td>0.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment Difference</th>
<th>((\bar{y}<em>{i1;adj} - \bar{y}</em>{i2;adj}))</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check-Full</td>
<td>-3.41</td>
<td>0.86</td>
</tr>
<tr>
<td>Check-N₀</td>
<td>-1.58</td>
<td>0.82</td>
</tr>
<tr>
<td>Check-P₀</td>
<td>-0.39</td>
<td>0.75</td>
</tr>
<tr>
<td>Full-N₀</td>
<td>1.83</td>
<td>0.68</td>
</tr>
<tr>
<td>Full-P₀</td>
<td>3.02</td>
<td>0.71</td>
</tr>
<tr>
<td>N₀-P₀</td>
<td>1.19</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Multifactor ANCOVA

- ANCOVA can also be employed when multiple factors are present.
- For simplicity, we will consider the case where the treatment sample size is the same for all treatments and where the number of factors is two.
- Recall that the fixed-effects ANOVA model for a two-factor balanced study is

\[
Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},
\]

where \( i = 1, \ldots, a \), \( j = 1, \ldots, b \), \( k = 1, \ldots, n \), and all of the traditional assumptions and constraints apply to each term in the above model.

- The ANCOVA model for a two-factor study with a single covariate (assuming that the relationship between \( y \) and \( x \) is linear) is

\[
Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma(x_{ijk} - \bar{x}) + \epsilon_{ijk},
\]

where \( x_{ijk} \) is the covariate value of observation \( k \) within level \( i \) of factor \( A \) and level \( j \) of factor \( B \), and \( \gamma \) is now used to represent the regression coefficient.

- Estimation and SS partitioning all follow the same logic as with previous ANCOVA models.
Example: Salable Flowers Study

A horticulturist conducted an experiment to study the effects of flower variety (factor $A$ levels: LP and WB) and moisture level (factor $B$ levels: lower and high) on yield of salable flowers ($y$). Because the plots were not the same size, the horticulturist wished to use plot size ($x$) as a covariate. Six replications were made for each treatment. The data are given in the table below.

<table>
<thead>
<tr>
<th>Factor $A$ ($i$)</th>
<th>Factor $B$ ($j$)</th>
<th>$B_1$ : low</th>
<th>$B_2$ : high</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B_1$ : low</td>
<td>$B_2$ : high</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$y_{i1k}$</td>
<td>$x_{i1k}$</td>
<td>$y_{i2k}$</td>
</tr>
<tr>
<td>$A_1$ : LP</td>
<td>98</td>
<td>15</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>77</td>
<td>7</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>14</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>5</td>
<td>55</td>
</tr>
<tr>
<td>$A_2$ : WB</td>
<td>55</td>
<td>4</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>5</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>8</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>13</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>11</td>
<td>70</td>
</tr>
</tbody>
</table>
Example: Salable Flowers Study

- To the right is a scatterplot of the salable flowers data.
- Different colors and plotting symbols are used to distinguish the four different treatments.
- It looks like there could be a treatment effect and that the assumption of parallel slopes will be appropriate, which we will test.
Example: Salable Flowers Study

Anova Table (Type III tests)

Response: y

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8218.2</td>
<td>1</td>
<td>1306.8625</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>x</td>
<td>3994.5</td>
<td>1</td>
<td>635.2118</td>
<td>4.586e-16 ***</td>
</tr>
<tr>
<td>variety</td>
<td>96.6</td>
<td>1</td>
<td>15.3617</td>
<td>0.0009211 ***</td>
</tr>
<tr>
<td>moisture</td>
<td>323.8</td>
<td>1</td>
<td>51.4988</td>
<td>8.093e-07 ***</td>
</tr>
<tr>
<td>variety:moisture</td>
<td>16.0</td>
<td>1</td>
<td>2.5511</td>
<td>0.1267191</td>
</tr>
<tr>
<td>Residuals</td>
<td>119.5</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

After adjusting for the covariate (using Type III SS), we fit the full two-factor model with the covariate. The ANCOVA table is given above (ignore the (Intercept) row). First we test the interaction term:

\[ H_0 : (\alpha \beta)_{11} = (\alpha \beta)_{12} = (\alpha \beta)_{21} = (\alpha \beta)_{22} = 0 \]
\[ H_A : \text{not all } (\alpha \beta)_{ij} \text{ equal 0} \]

The test statistic is \( F^* = 2.551 \sim F_{1,19} \), which has a \( p \)-value of 0.127. Thus, the interaction term is not statistically significant and we drop it from the model.
Example: Salable Flowers Study

Anova Table (Type III tests)

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8286.9</td>
<td>1</td>
<td>1222.944</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>x</td>
<td>3978.5</td>
<td>1</td>
<td>587.128</td>
<td>2.692e-16 ***</td>
</tr>
<tr>
<td>variety</td>
<td>97.6</td>
<td>1</td>
<td>14.396</td>
<td>0.001138 **</td>
</tr>
<tr>
<td>moisture</td>
<td>324.4</td>
<td>1</td>
<td>47.879</td>
<td>1.016e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>135.5</td>
<td>20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

After dropping the interaction term, we refit the model and then test the significance of the covariate. The ANCOVA table is given above. The test statistic is \( F^* = 587.128 \sim F_{1,20} \), which has a \( p \)-value of 2.692e-16. Thus, the relationship between the salable flowers and the plot size is highly significant. We then test the significance of each factor effect. The \( F \)-statistics for testing effects due to factor A and B are, respectively, \( F^*_A = 14.396 \) and \( F^*_B = 47.879 \). These each follow a \( F_{1,20} \) distribution, resulting in \( p \)-values of 0.001 and 1.016e-06. Therefore, both variety and moisture levels are significant effects on salable flowers in the presence of the covariate of plot size.
Example: Salable Flowers Study

Anova Table (Type III tests)

Response: y

<table>
<thead>
<tr>
<th></th>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>7736.8</td>
<td>1</td>
<td>1099.0567</td>
<td>&lt;2.2e-16***</td>
</tr>
<tr>
<td>variety</td>
<td>38.7</td>
<td>1</td>
<td>5.4927</td>
<td>0.03078 *</td>
</tr>
<tr>
<td>moisture</td>
<td>34.7</td>
<td>1</td>
<td>4.9239</td>
<td>0.03958 *</td>
</tr>
<tr>
<td>x</td>
<td>3689.6</td>
<td>1</td>
<td>524.1213</td>
<td>9.235e-15 ***</td>
</tr>
<tr>
<td>variety:x</td>
<td>4.6</td>
<td>1</td>
<td>0.6469</td>
<td>0.43172</td>
</tr>
<tr>
<td>moisture:x</td>
<td>4.2</td>
<td>1</td>
<td>0.5968</td>
<td>0.44984</td>
</tr>
<tr>
<td>Residuals</td>
<td>126.7</td>
<td>18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Finally, we need to test the reasonableness of homogeneity of the slopes. We can accomplish this by crossing the covariate with each factor. The ANCOVA table is given above. As we can see, the tests for crossing factor A and factor B with the covariate have p-value of 0.432 and 0.450, respectively. Therefore, the slopes are not considered significantly different and we can proceed to assume that the slopes are parallel.
Example: Salable Flowers Study

- To the right is a scatterplot of the salable flowers data with the fitted regression lines for each treatment:

\[
\hat{y}_{11k} = 79 + 3.263(x_{11k} - 9) \\
\hat{y}_{12k} = 70 + 3.263(x_{12k} - 9) \\
\hat{y}_{21k} = 70 + 3.263(x_{21k} - 8) \\
\hat{y}_{22k} = 61 + 3.263(x_{22k} - 7)
\]

- Clearly, plot size has an effect on the flower sales in this study.

- Further quantifications can be performed, such as standard errors and multiple comparisons, similarly to the other ANCOVA models discussed.
Final Comments About ANCOVA

- ANCOVA combines features of ANOVA and regression that partitions the total variation into components ascribable to (1) treatment effects, (2) covariate effects, and (3) random experimental error due to the experimental design (e.g., blocking).

- Using covariates makes better use of exact values for quantitative factors instead of, say, discretizing those values and defining them as classes to use as blocking factors.

- ANCOVA for comparative observational studies is at a disadvantage since EUs cannot be randomized to treatment groups and there is a possibility for an influence on the response by additional unobserved covariates that are associated with the treatment groups, thus resulting in bias.

- Since there is a regression component to ANCOVA, the same cautions regarding extrapolation apply to ANCOVA as when we perform a regression analysis.
This is the end of Unit 11.