

Analysis of Covariance Models

- A simple situation suited for ANCOVA is when we have two independent variables affecting the response: one is a factor, and the other is a continuous variable.

- ANCOVA combines the one-way ANOVA model and the SLR model:

$$Y_{ij} = \beta_0 + \tau_i + \beta_1 X_{ij} + \varepsilon_{ij}, i = 1, \dots, t, j = 1, \dots, n_i.$$

- For the t levels of the factor ($i = 1, \dots, t$), define:

→ $Y_{ij} =$

- This shows a set of t SLR lines having equal slopes, but having different intercepts for each of the t levels of the factor.

Picture (parallel lines relating $E(Y)$ and X):

Example: A study analyzing blood pressure reduction (Y) in patients. The factor is type of drug (3 different drugs). However, the weight of the patient (a continuous variable) will also affect blood pressure reduction.

- If we're interested in the effect of each drug on BP reduction, we should account for the patients' weights as well.

- One way: Break weights into categories (levels) and make weight a blocking factor.

Problems: (1) There may not be enough people in some of the weight categories.

(2) We may not know weight is affecting BP reduction until after the experiment is ongoing.

(3) Weight is inherently continuous.

(4) In some studies, there may be several continuous variables affecting the response.

- ANCOVA achieves similar benefits to blocking, but is preferable when controlling for continuous covariates.

Example: Table 11.6 data (p. 593)

- Analyzing the effect of 3 types of classes on students' post-class test score in trigonometry.

Y (POST) =

Factor (CLASSTYPE) =

● **We want to control for previous knowledge of trigonometry.**

covariate X (PRE) =

Model equation:

● **See example for ANCOVA data analysis (output similar to Table 11.7, p. 594).**

● **Important pieces of output: Overall $F^* = 8.46$ (P-value near 0) → our model is useful overall.**

(1) Does the covariate (pre-class score) have a significant effect on post-class score?

(2) Does the factor (type of class) have a significant effect on post-class score? (Look at the Type III SS)

• We also get least-squares estimates $\hat{\beta}_0, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\beta}_1$ for this model.

• Interpreting $\hat{\beta}_1 =$:

- **Adjusted estimated mean post-class scores for each type of class (at any given value of pre-class score) are:**

Class type 1: $\hat{\beta}_0 + \hat{\tau}_1 + \hat{\beta}_1 X =$

Class type 2: $\hat{\beta}_0 + \hat{\tau}_2 + \hat{\beta}_1 X =$

Class type 3: $\hat{\beta}_0 + \hat{\tau}_3 + \hat{\beta}_1 X =$

- **We can extend the ANCOVA model to have more than one covariate.**

Example: Suppose we had two continuous covariates, pre-test score and IQ.

Results from software:

- **Interpreting $\hat{\beta}_2 =$:**

Unequal Slopes Situation

- Maybe the effect of the covariate is different for each level of the factor.

Picture:

- We can formally test whether this is true by including a term for the interaction between the factor and the covariate.

Example: Are the slopes unequal for the model with factor CLASSTYPE and covariate PRE?

- **Analysis:** Include CLASSTYPE by PRE interaction term.

- Look at F-test for interaction term in output:

F* =

P-value =

Conclusion:

Logistic Regression

- In analyses we have studied, the response variable has been continuous (typically normal).
- In some studies, we have a binary response (which takes one of only two values).

→ Response represented by a dummy variable (with values 0 and 1).

Example 1: Dose-response model:

Y =

X =

Example 2: Survival model:

Y =

X =

- The standard linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ with fitted regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is one possibility for analyzing such data.

- In this situation, \hat{Y} does not represent a predicted Y -value (since Y must equal 0 or 1).
- \hat{Y} represents the estimated probability that $Y = 1$, given a certain X value.

Problem: The straight-line regression could (for certain X values) predict probabilities for Y that are outside of the range from 0 to 1.

- A more appropriate model is the logistic regression model:

- Taking the expected value for a given X , this is:

- Note again: $E(Y | X)$ represents the probability that $Y = 1$ given some X value.

Key advantage: This logistic function always lies between 0 and 1.

Note: As a function of X , the logistic function $E(Y | X)$

- has a sigmoidal shape
- approaches 0 or 1 at the left/right limits
- is monotone (always increases or always decreases)

- Value of β_1 governs the shape of the logistic curve.

Picture:

● With sample data, we estimate β_0 and β_1 to obtain the estimated logistic regression curve:

● Estimating the parameters is easier after a transformation.

Some terminology:

- This is called a logit transformation.

Note:

- With a logit transformation of both sides of the logistic regression equation, we get:

Problem with variances: Y here is a binomial variable with:

Its variance is:

- **Note the variance of Y depends on X . This implies we should not use least-squares method for estimating parameters.**
- **Alternative: Use maximum likelihood estimation: Choose values of β_0, β_1 that maximize the joint probability function, given a particular data set.**

Example (Table 13.3 data): How does a city's income level relate to the probability that it uses Tax Increment Financing (TIF)?

$Y =$

$X =$ median income of city (in thousands of dollars)

- **Parameter estimation done using software:**

$\hat{\beta}_0 =$

$\hat{\beta}_1 =$

- Software will provide a plot of the data along with the estimated logistic regression curve (see example).

Plot: For a city with median income \$11,000, the estimated probability of using TIF is about:

- Note that for any income value X^* :

the odds ratio =

is estimated to be:

- A 95% CI for this odds ratio is:

Interpretation:

Hypothesis Tests in Logistic Regression

- The Hosmer-Lemeshow Test is a test of how well the logistic model fits the data.

- Software gives Hosmer-Lemeshow P-value =

Conclusion:

Test for a Significant Effect of Income

- To test whether income is significant, we test $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.
- Instead of F-tests or t-tests, we have χ^2 tests here.
- The Wald test is a test about an individual coefficient.
- The Likelihood Ratio test is a test about the whole set of independent variables in the model.
- With simple logistic regression, these both test the same hypotheses: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.

Example: Wald test P-value =

LR test P-value =

Conclusion: