Chapter 8: Regression Models with Qualitative Predictors

• Some predictors may be binary (e.g., male/female) or otherwise categorical (e.g., small/medium/large).

• These typically enter the regression model through indicator variables (dummy variables), which take on values

• For a predictor with *c* categories, we employ

• Why not an indicator variable for each category?

Example: Table 8.2 (Insurance innovation data)
Y = Time until innovation adapted (in months)
X_1 = size of firm (continuous)

X_2 =

Model:

Mean response for mutual firms:

Mean response for stock firms:

Same

- **Why not fit two separate regressions, one for stock firms and one for mutual firms?**
- **Our model assumes same**

- **It's better to estimate these with the <u>total</u> data set.**
- **Inference for β_0 and β_2 will be more precise when we use all the data (more observations) to fit the model.**

- **We may fit our model with least squares as usual.**

<u>Example</u> (insurance innovation data). Fitted model:

- **Interpretation of b_2:**

- **95% CI for β_2:**

- **t-test for**

<u>Predictors with Several Categories</u>

• Suppose a predictor *X* has four categories:
***X* = shirt size (S, M, L, XL) of customer**
***Y* = amount spent on clothes by customer during store visit**

• Why not use a single predictor *X* defined as

Then for small size:

For medium:

For large:

For XL:

• Note the spacing between mean response functions is

• Defining *c* – 1 = 3 indicator variables here allows more

Then for small size:

For medium:

For large:

For XL:

• We can estimate the differences in mean response between the different categories by estimating

<u>Example</u> (Shirt Data): Fitted Model:

Interpretation of b_1:

Chapter 16: <u>Single-Factor ANOVA Models</u>

• An analysis of variance (ANOVA) model is a linear model in which all the predictors are represented through indicator variables.

• In an ANOVA model, the predictors are called <u>factors</u>.
• These factors may be qualitative (categorical) or quantitative, but if quantitative, we focus on several <u>discrete</u> values of the factor.

• The values that a factor may take on are called the factor <u>levels</u>.
• The response is still assumed to be continuous (typically normal).

Comparison between ANOVA Model and Regression Model

• When all predictors are qualitative, using the ANOVA model will yield identical results as using the regression model with indicators.
• The only difference is that the ANOVA model is specified with different notation.

• When the factors are quantitative (with discrete levels), there is a fundamental difference between the ANOVA model and the regression model.

• Unlike regression models, the ANOVA model does not specify the <u>functional form</u> of the relationship between the response and the predictor(s).

Picture:

ANOVA models may be used to analyze:
• <u>Experimental studies</u> (in which <u>experimental units</u> are <u>randomly</u> assigned to the different factor <u>levels</u> by the researcher)
OR
• <u>Observational studies</u> (in which the researcher does not control which <u>observational units</u> correspond to which factor levels).

<u>Note</u>: The units/individuals on which the response is measured are called experimental (or observational) units. (If humans, often called "subjects").

Example 1: <u>Response</u>:

<u>Factor</u>:

<u>Levels</u>:

<u>Subjects</u>:

Example 2: <u>Response</u>:

<u>Factor</u>:

<u>Levels</u>:

<u>Observational units</u>:

<u>Note</u>: Some studies may be a mix of experimental and observational.

• In a single-factor study, we assume that at each level of the factor, the response values follow a <u>probability distribution</u>.

Picture:

ANOVA model assumptions:

<u>Important question</u>: Are the population means for each level equal?

<u>Note</u>: If there are only two levels, we would answer this with

• The ANOVA model

The “Cell Means” Model:

• There are r levels.
Notation:

<u>Note</u>:

• The ANOVA model is a case of the <u>general linear model</u>.

<u>Example</u>: Suppose $r = 3$ and $n_1 = 1$, $n_2 = 3$, $n_3 = 2$. Then let:

• Then the ANOVA model can be stated as

Fitting the ANOVA Model

• The parameters μ_1, μ_2, …, μ_r are unknown and must be estimated from sample data.
• We may use least squares (or, equivalently if the errors are normal, maximum likelihood).

Example (Kenton Foods, Table 16.1):

Does package design significantly affect sales of breakfast cereal?
Experimental Units: 19 stores
Response:

Factor:

Some notation:

• The least squares method will choose estimators of μ_1, μ_2, …, μ_r to minimize

• For example, the LS estimator of μ_1 is found by:

• Similarly, for i = 1, 2, …, r, the least-squares estimates are the

<u>Kenton Foods Example</u>:

<u>Residuals in the ANOVA model</u>

Residual = difference between the observed *Y*-value and fitted value (in this case, the factor level sample mean).

For each observation,

For each level, i = 1, 2, …, r: