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 total data set.
• Inference for $\beta_0$ and $\beta_2$ 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.

Example (insurance innovation data). Fitted model:

• Interpretation of $b_2$:

• 95% CI for $\beta_2$:

• t-test for
Predictors with Several Categories

• 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

**Example (Shirt Data):** Fitted Model:

Interpretation of $b_1$:

---

**Chapter 16: Single-Factor ANOVA Models**

• 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 **factors**.
• These factors may be qualitative (categorical) or quantitative, but if quantitative, we focus on several **discrete** values of the factor.

• The values that a factor may take on are called the factor **levels**.
• 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 functional form of the relationship between the response and the predictor(s).

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

**Note:** The units/individuals on which the response is measured are called experimental (or observational) units. (If humans, often called “subjects”).

**Example 1:** **Response:**

**Factor:**

**Levels:**

**Subjects:**

**Example 2:** **Response:**

**Factor:**

**Levels:**

**Observational units:**
Note: 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 probability distribution.

Picture:

ANOVA model assumptions:

Important question: Are the population means for each level equal?

Note: If there are only two levels, we would answer this with

- The ANOVA model
The “Cell Means” Model:

• There are $r$ levels.

Notaion:

Note:

• The ANOVA model is a case of the general linear model.

Example: 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 $\mu_1, \mu_2, \ldots, \mu_r$ 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 $\mu_1, \mu_2, \ldots, \mu_r$ to minimize

• For example, the LS estimator of $\mu_1$ is found by:

• Similarly, for $i = 1, 2, \ldots, r$, the least-squares estimates are the

Kenton Foods Example:

Residuals in the ANOVA model

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

For each observation,

For each level, $i = 1, 2, \ldots, r$: 