1 Models with Quantitative and Qualitative Predictors

- We can include **categorical predictor variables** in a regression model by coding them using ***indicator variables*** (also called ***dummy variables***) that take the values **zero** or **one**, e.g.

$$X_i = \begin{cases} 0 & \text{if the } i\text{th individual is Male} \\ 1 & \text{if the } i\text{th individual is Female} \end{cases} \tag{1}$$

1.1 ANOVA Models as Regression Models

- ***Analysis of variance (ANOVA) models*** are models in which the predictors are **all categorical** variables (and called ***factors***).

1.1.1 One-Factor ANOVA Model with Two Levels of the Factor

- Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

where X is the gender indicator variable defined by (1) and Y is some response variable.

- The **design matrix**, when there are, say, three Females in the data set and three Males, is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \tag{3}$$

- From (2), the mean response for Males ($X = 0$) is

$$E(Y) = \beta_0$$

and the mean response for Females ($X = 1$) is

$$E(Y) = \beta_0 + \beta_1 .$$

If we write

$$\mu_1 = \beta_0 \quad \text{and} \quad \mu_2 = \beta_0 + \beta_1, \tag{4}$$

then the model (2) can be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{5}$$

where Y_{ij} is the response for the jth individual in the ith "group" ($i = 1$ for Males, $i = 2$ for Females). This is the ***one-factor ANOVA model*** (cell means version) with **two levels** of the **factor** (Male and Female).

- Note by (4) that a test of

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

in model (2) is equivalent to a test of

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

in model (5). The ***regression model F test*** of the first set of hypotheses (Class Notes 4) is equivalent to the so-called ***one-factor ANOVA F test*** of the second set (MTH 3220).

Furthermore, it can be shown that the ***t test*** of the first set of hypotheses is equivalent to the ***pooled two-sample t test*** of the second set (MTH 3220).

1.1.2 One-Factor ANOVA Models with More Than Two Levels of the Factor

- One-factor ANOVA models with **more than two levels** of the **factor** can be expressed in terms of indicator variables.

For example, suppose there are **three levels** of the **factor**: Low, Medium, and High. We define **two indicator variables** X_1 and X_2 as follows:

$$X_{i1} = \begin{cases} 1 & \text{if the } i\text{th individual is Medium} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i2} = \begin{cases} 1 & \text{if the } i\text{th individual is High} \\ 0 & \text{otherwise} \end{cases}$$

In this coding scheme, and individual who is Low is coded as $X_{i1} = 0$ and $X_{i2} = 0$.

The model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \, . \tag{6}$$

Here the **design matrix**, when there are, say, two Low individuals, two Mediums, and two Highs, is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad (7)$$

The first two rows correspond to the Lows, the next two to the Mediums, and the last two to the Highs.

From (6), the mean response for Low individuals ($X_1 = 0$, $X_2 = 0$) is

$$E(Y) = \beta_0 ,$$

the mean response for Mediums ($X_1 = 1$, $X_2 = 0$) is

$$E(Y) = \beta_0 + \beta_1 ,$$

an the mean response for Highs ($X_1 = 0$, $X_2 = 1$) is

$$E(Y) = \beta_0 + \beta_2 .$$

Now, if we write

$$\mu_1 = \beta_0, \qquad \mu_2 = \beta_0 + \beta_1, \qquad \text{and} \qquad \mu_3 = \beta_0 + \beta_2, \qquad (8)$$

then the model (6) can be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} , \qquad (9)$$

where μ_1 is the mean response for the Low "treatment group", μ_2 is the mean response for Medium, and μ_3 is the mean response for High. This is the ***one-factor ANOVA model*** (cell means version) with **three levels** of the **factor** (Low, Medium, and High).

Note by (8) that a test of

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = 0 \\ H_a &: \ \text{Not both } \beta_1 \text{ and } \beta_2 \text{ equal } 0 \end{aligned}$$

in model (6) is equivalent to a test of

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \mu_3 \\ H_a &: \ \text{Not all } \mu_i\text{'s are equal} \end{aligned}$$

in model (9). The ***regression model F test*** of the first set of hypotheses (Class Notes 11) is equivalent to the ***one-factor ANOVA F test*** of the second set (MTH 3220).

1.1.3 One-Factor ANOVA Models (in General)

- In general, if there are $\boldsymbol{a}$ **levels** of the **factor**, we will need $\boldsymbol{a-1}$ **indicator variables** to express the one-factor ANOVA model as a regression model and form the design matrix $\boldsymbol{X}$ as in (3) and (7).

- **One-factor ANOVA models** are ***general linear models*** (i.e. they're linear combinations of the β_k's), so the **least squares estimates** $\boldsymbol{b_0, b_1, \ldots, b_{p-1}}$ of the model parameters are be obtained exactly as before, i.e.

  Least Squares Estimates of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$ (Matrix Approach): The vector of estimated coefficients $\mathbf{b}$ is obtained by:

  $$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\,.$$

  It can be shown that, as we'd expect from (8),

  $$\begin{aligned} b_0 &= \bar{Y}_1 \\ b_0 + b_1 &= \bar{Y}_2 \qquad (\text{i.e. } b_1 = \bar{Y}_2 - \bar{Y}_1) \\ b_0 + b_2 &= \bar{Y}_3 \qquad (\text{i.e. } b_2 = \bar{Y}_3 - \bar{Y}_1) \\ &\vdots \\ b_0 + b_{p-1} &= \bar{Y}_p \qquad (\text{i.e. } b_{p-1} = \bar{Y}_p - \bar{Y}_1) \end{aligned}$$

  where $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_p$ are the group means in ANOVA.

- The **fitted regression model** is

  Fitted Regression Model with Categorical Predictors:

  $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_{p-1}X_{p-1}\,,$$

where the predictors $X_1, X_2, \ldots, X_{p-1}$ are **indicator** variables.

- The variance-covariance matrix of $\mathbf{b}$ is

$$\sigma^2\{\mathbf{b}\} \;=\; \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}\,,$$

and the (estimated) variance-covariance matrix $\boldsymbol{s^2\{b\}}$ is obtained by replacing σ^2 by MSE. Thus the **(estimated) standard errors** of $b_0, b_1, \ldots, b_{p-1}$ reported by statistical software are:

(Estimated) standard errors: The (estimated) standard errors $\boldsymbol{s\{b_0\}}$, $\boldsymbol{s\{b_1\}, \ldots, s\{b_{p-1}\}}$ of $b_0, b_1, \ldots, b_{p-1}$ are the square roots of the diagonal elements of the matrix

$$s^2\{\mathbf{b}\} \;=\; \text{MSE} \cdot (\mathbf{X}^T\mathbf{X})^{-1}\,.$$

1.1.4 ANOVA Models with Two or More Factors

- We can include **more than one categorical predictor variable** in a regression model.

- For example, in an experiment with **gender** (Male or Female) as a ***blocking variable*** and **group** (Treatment or Control) as the ***factor***, we could code these **two categorical predictors** using indicator variables X_1 and X_2 as

$$X_{i1} \;=\; \begin{cases} 0 & \text{if the } i\text{th individual is Male} \\ 1 & \text{if the } i\text{th individual is Female} \end{cases}$$

and

$$X_{i2} \;=\; \begin{cases} 0 & \text{if the } i\text{th individual is in the Control group} \\ 1 & \text{if the } i\text{th individual is in the Treatment group} \end{cases}$$

and then fit the (additive) model

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i\,. \tag{10}$$

If there were, say, two males and two females in each group, the **design matrix** would be

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

where the 2nd column indicates the **gender** and the 3rd indicates the **group**.

For example, the *first* row of $\boldsymbol{X}$ corresponds to a Male in the Control group. The *third* corresponds to a Male in the Treatment group. The *fifth* corresponds to a Female in the Control group. The *seventh* corresponds to a Female in the Treatment group.

Using an argument similar to (8) and (9), it can be shown that the model (10) is equivalent to the (***additive***) ***two-factor ANOVA model*** (cell means version),

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} .$$

- If the first of two categorical predictors has ***a*** **levels** and the second has ***b*** **levels**, then for the (***additive***) **two-factor ANOVA model** the **design matrix** will have $\boldsymbol{a-1}$ **columns** for the first variable and $\boldsymbol{b-1}$ **columns** for the second (in addition to the column of 1's for the intercept).

  For example, suppose subjects in an experiment are exposed to ***three*** **doses** of a medication (Low, Medium, and High) and the ***two*** **genders** (Male and Female) are used to form blocks.

  Then we code these **two** categorical predictor variables as **three** indicator variables X_1, X_2, and X_3 using

$$X_{i1} = \begin{cases} 1 & \text{if the } i\text{th individual is Medium} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{i2} = \begin{cases} 1 & \text{if the } i\text{th individual is High} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i3} = \begin{cases} 0 & \text{if the } i\text{th individual is Male} \\ 1 & \text{if the } i\text{th individual is Female} \end{cases}$$

In this case, if there are, say, one male and one female in each dose group, the **design matrix** would be

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

where the 2nd and 3rd columns indicate indicate the **dose group** and the 4th indicates the **gender**.

Example 1.1 Consider again an experiment involving ***three*** **doses** of a medication (Low, Medium, and High) administered to ***two*** **genders** (Male and Female). Let

$$X_{i1} = \begin{cases} 1 & \text{if the } i\text{th individual is Medium} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i2} = \begin{cases} 1 & \text{if the } i\text{th individual is High} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{i3} = \begin{cases} 0 & \text{if the } i\text{th individual is Male} \\ 1 & \text{if the } i\text{th individual is Female} \end{cases}$$

Suppose the **design matrix** is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

What is the **gender** and **dose group** of the *fourth* individual? How about the *sixth* individual?

- **More than two categorical predictor variables** can be represented in regression models by indicator variables in a similar manner, leading to (***additive***) ***three-factor ANOVA models***, ***four-factor ANOVA models***, etc.

> **Example 1.2** Consider data on gas mileage for cars from ***two*** **manufacturers** (Ford, Toyota), ***three*** **drive train** types (Four wheel, Front wheel, Rear wheel), and ***five*** **vehicle classes** (Suv, Pickup, Subcompact, Midsize, Compact).
>
> How many indicator variables would be needed to code the **manufacturer**? How many to code the **drive train** type? How many to code the **vehicle** class?

- **Interactions** between **categorical predictors** can also be coded using indicator variables by taking **products** of the **indicator variables** described above.

1.2 More on Coding Categorical Predictors by Indicator Variables

- We use $\boldsymbol{a-1}$ indicator variables to represent a **categorical predictor** with $\boldsymbol{a}$ levels so that the columns of the design matrix $\boldsymbol{X}$ won't be linearly dependent.

  Recall that **if** the columns of $\boldsymbol{X}$ are **linearly dependent**, we **can't** get **unique** estimates of model parameters (unless a variable is dropped from the model).

  For example, if subjects in an experiment are exposed to ***three*** **doses** of a medication (Low, Medium, and High), the following coding system **won't work**:

$$X_{i1} = \begin{cases} 1 & \text{if the } i\text{th individual is Low} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i2} = \begin{cases} 1 & \text{if the } i\text{th individual is Medium} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i3} = \begin{cases} 1 & \text{if the } i\text{th individual is High} \\ 0 & \text{otherwise} \end{cases}$$

because for, say, two individuals in each group, this would lead to the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

whose columns are **linearly dependent** (the first is the sum of the latter three).

1.3 Confounding and the Design Matrix

- Here's another example in which the columns of $\boldsymbol{X}$ are **linearly dependent**, this time the result of a poorly designed experiment.

  Suppose again in an experiment with ***blocking variable*** **gender (Male or Female)** and ***factor*** **group** (Treatment or Control), we (again) use

$$X_{i2} = \begin{cases} 0 & \text{if the } i\text{th individual is Male} \\ 1 & \text{if the } i\text{th individual is Female} \end{cases}$$

  and

$$X_{i1} = \begin{cases} 0 & \text{if the } i\text{th individual is in the Control group} \\ 1 & \text{if the } i\text{th individual is in the Treatment group} \end{cases}$$

  If all the Males are in the Treatment group and all the Females in the Control group, the **design matrix** would be

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

  Its columns would be **linearly dependent** (the second and third are identical), and we **wouldn't** be able to **uniquely** estimate the model parameters (unless we dropped one of the two variables, **gender** or **group**, from the model).

  Here, the **group** designation (Treatment or Control) is completely ***confounded*** with **gender**.