SCM 651 (Business Analytics)
Syracuse University, Whitman School of Management
Fall 2017
Notes on Statistical Analysis

Amiya K. Basu
Professor of Marketing

August 30, 2017

Contents

This Reader has been prepared by me exclusively for teaching this and similar courses.

1 Chi-Square Analysis With Cross-Tabulation

1.1 Introduction

Pivot Table: In its simplest form, a pivot table is table with a row variable and a column variable where each row has the same value of the row variable, and each column has the same value of the column variable. The row variable and the column variable have finite numbers of categories. If the row variable has R categories and the column variable has C categories, the pivot table is an $R \times C$ matrix. Each entry in the matrix is called a cell. Each cell is at the crossing of a row and a column and is designated by the row and and the column. For example, cell 11 (one-one, not eleven) is at the crossing of row 1 and column 1. Each entry in the cell is some statistic about a third variable for that cell. For example, consider the following pivot table where the row variable is class (freshman, sophomore, junior, senior, graduate), and the column variable is gender (male, female). The entries in the cells are the average number of credit hours the student is taking in Fall 2014.

	Gender	
Class	Male	Female
Freshman	15	15
Sophomore	12	16
Junior	16	16
Senior	15	18
Graduate	12	15

In the table above, cell 11 consists of freshmen males. The students in the sub-sample of male freshmen are taking 15 credits hours on the average during Fall 2014.

Cross Tabulation: A cross tabulation is a special case of a pivot table where each entry is the number of times that combination occurs in the sample. The general idea is as follows:

- You have two variables, V_1 and V_2.
- Each variable can take a fixed number of values.
- Make a table where each row corresponds to a given value of V_1, and each column corresponds to a given value of V_2.
- Each cell of the table corresponds to a given combination of values of V_1 and V_2.
- In each cell of the table, enter the number of cases in the sample that has the given combination of values of V_1 and V_2.

Example 1.1: The following data come from a sample of 20 Syracuse University undergraduate students.

Gender: 1 if male, 0 if female

Interest in watching sports on TV: 1 if low, 2 if medium, 3 if high

Student #	Gender	Interest	Student #	Gender	Interest
1	1	3	11	1	1
2	1	3	12	1	3
3	0	3	13	0	2
4	0	2	14	0	3
5	1	3	15	0	1
6	1	3	16	1	3
7	0	2	17	0	3
8	0	2	18	0	3
9	1	3	19	0	1
10	1	3	20	1	3

We now construct a cross tabulation with gender as the row variable and interest as the column variable. There are $2 \times 3 = 6$ possible combinations of values of gender and interest: female with low interest, female with medium interest, female with high interest, male with low interest, male with medium interest, and male with high interest. Thus, the cross-tabulation has two rows and three columns ($R = 2$, $C = 3$), and is given as follows:

	Interest		
Gender	Low	Medium	High
Female	2	4	4
Male	1	0	9

Note that generally speaking, men are more likely to have high interest in watching sports on television that women (9 out of 10 men, 4 out of 10 women). Thus, the proportion of men interested in watching sports on television appears to be different from the proportion of women interested in watching sports on television. In other words, the two sub-populations of men and women seem to have different levels of interest in watching sports on television. We say that gender appears to be **related** to interest in watching sports on television. If proportions were same for men and women, we would say that there is **no relationship** between gender and watching sports on television. The chi-square is used formally test if such a relationship exists in your sample.

1.2 Null Hypothesis and Chi-square Test

When we have a cross-tabulation of two variables, the chi-square test can be used to determine if the two variables are related. The null hypothesis of the chi-square test is that there is “no relationship” between the two variables. If the evidence from the sample contradicts the null hypothesis strongly enough, we reject the null hypothesis and conclude that there is a "relationship" between the variables. We now use two examples to illustrate the meaning of "no relationship" in this context and develop the chi-square test.

Example 1.2 A cross-tabulation has at least two rows and at least two columns. We first consider the simplest cross-tabulation with two rows and two columns.

Suppose we selected a simple random sample of 150 students from a college campus, and recorded (i) the gender of the student, and (ii) whether the student has attended a basketball game played by the college team during the past year. The results are expressed as the following 2×2 cross tabulation:

	Didn't Attend Game	Attended Game
Male	30	60
Female	42	18

Meaning of Null Hypothesis: The chi-square test is used to test the null hypothesis, that there is no relationship between gender and attendance, against the alternate hypothesis, that gender and attendance are related.

In this case, the overall population can be divided into two sub-populations by gender: (1) male, and (2) female. Let π_{11} and π_{12} denote the proportions of the male sub-population that did not attend, and attended a game, respectively. Since every male student must fall into one of these two categories, we have $\pi_{11} + \pi_{12} = 1$.

Similarly, we may denote the proportions of the female sub-population that did not attend, and attended the games by π_{21} and π_{22}, respectively. Once again, as attendance is dichotomous, $\pi_{21} + \pi_{22} = 1$.

If H_0 is true, that is, gender and attendance are not related, then the proportion of students that either attended or did not attend a game should not depend on gender, that is, we should have:

$$\pi_{11} = \pi_{21}, \quad \text{and} \quad \pi_{12} = \pi_{22}.$$

Thus, the null hypothesis means that for each category of gender (male or female), the proportional distribution across categories of attendance ("did not" or "did attend") is the same.

Simple Interpretation for a 2×2 Table: In the present case of a 2×2 table, H_0 has a simple interpretation. Note that since $\pi_{11} + \pi_{12} = 1$ and $\pi_{21} + \pi_{22} = 1$, it follows that if $\pi_{12} = \pi_{22}$, then we also have $\pi_{11} = \pi_{21}$. Denoting π_{21} by π_1 and π_{22} by π_2 (proportions of men and women that attended a game), respectively, it follows that for a 2×2 cross-tabulation, the null hypothesis can be expressed as $\pi_1 = \pi_2$. The alternate hypothesis is $\pi_1 \neq \pi_2$.

Expected and Observed Frequencies: The cross-tabulation obtained from the data is called the *observed* cross-tabulation, and frequencies in the cells of the observed cross-tabulation are called *observed frequencies*. The observed frequency in the ij-th cell (that is, the cell at the crossing of the i-th row and the j-th column) is denoted by O_{ij}. For example, in the cross-tabulation here, $O_{11} = 30$, $O_{12} = 60$, $O_{21} = 42$, and $O_{22} = 18$.

The *expected frequencies* are the cell frequencies we would "expect" if H_0 were true. The expected frequency in the ij-th cell is denoted by E_{ij}. In the cross-tabulation here, the total of column 1 is 72, and the total of column 2 is 78. Thus, in the aggregate sample, the proportion that did not attend a game was $\frac{72}{150}$, and the proportion that did attend a game was $\frac{78}{150}$. If H_0 is true, then the percentage distribution between did not attend and attend should be same for men and women, that is, we would expect each category of gender to have a distribution just like the aggregate sample.

In the sample, there are 90 men. Thus, if H_0 were true, we would expect that a fraction $\frac{72}{150}$ of these 90 men did not attend a game, and that a fraction $\frac{78}{150}$ of these 90 men attended a game. Restating, we have,

$$E_{11} = 90 \times \frac{72}{150} = \frac{90 \times 72}{150} = 43.2, \quad E_{12} = 90 \times \frac{78}{150} = \frac{90 \times 78}{150} = 46.8.$$

Similarly, considering the 60 women in the sample, we have:

$$E_{21} = 60 \times \frac{72}{150} = \frac{60 \times 72}{150} = 28.8, \quad E_{22} = 60 \times \frac{78}{150} = \frac{60 \times 78}{150} = 31.2.$$

Note that the expected number in cell ij can be expressed as follows:

$$(1.1) \quad E_{ij} \ = \ \frac{\text{Total of Row } i \times \text{Total of Column } j}{\text{Size of the aggregate sample}}.$$

Computed Chi-Square: The null hypothesis is rejected if the observed frequencies deviate strongly from the expected frequencies. This deviation is measured by the "computed chi-square (χ^2)," defined as follows:

$$(1.2) \quad \text{Computed } \chi^2 \ = \ \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where r is the number of rows in the cross-tabulation, and c is the number of columns.

For example, in the present case, computed χ^2

$$= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

$$= \frac{(30 - 43.2)^2}{43.2} + \frac{(60 - 46.8)^2}{46.8} + \frac{(42 - 28.8)^2}{28.8} + \frac{(18 - 31.2)^2}{31.2} = 19.39.$$

Decision Rule: From equation (1.2), the computed chi-square is zero if and only if the expected frequency is equal to the observed frequency in every individual cell. A larger value of the computed chi-square represents a greater discrepancy between what we expect from H_0 and what we observe from the data.

If H_0 is true, then the computed χ^2 follows the chi-square distribution with degrees of freedom equal to $(R - 1) \times (C - 1)$. To test H_0, we use the following decision rule:

At a confidence level of $(1 - \alpha)$, reject H_0 if the computed χ^2 exceeds χ^2_α at degrees of freedom $= (R - 1) \times (C - 1)$.

For example, suppose in the present case we wish to test H_0 at a 99% level of confidence. Since we have a 2×2 table ($r = 2$, $c = 2$), the degree of freedom of chi-square is $(2 - 1) \times (2 - 1) = 1$. Therefore, we use the following decision rule:

At a 99% level of confidence, reject H_0 if the computed $\chi^2 > \chi^2_{.01}$ at a degree of freedom of 1, that is, computed $\chi^2 > 6.63$.

Since, in the present case, the computed chi-square $= 19.39 > 6.63$, we reject H_0 at a 99% level of confidence.

Example 1.3 We now consider a cross-tabulation with two rows and three columns. A random sample of 100 students was drawn the students of Syracuse University. For each student, the following information were recorded:

- Gender: female or male
- Interest in shopping for clothes

The results are presented in the following cross-tabulation:

	Interest		
Gender	Low	Medium	High
Female	9	12	39
Male	17	10	13

To perform the chi-square test, we begin with the null hypothesis (H_0) that there is no relationship between gender and interest.

In the present case, the population can be divided into two sub-populations: (1) female, and (2) male. If H_0 were true, then the percentage distribution between the three categories of interest is same for the two sub-populations. Hence, the distribution of interest in each sub-population is same as that for the aggregate population.

In the aggregate sample of 100 students, 26 students have low interest (total of column 1), 22 students have medium interest (total of column 2), and 52 students have high interest (total of column 3). Thus, in the aggregate sample, the fraction $\frac{26}{100}$ have low interest, the fraction $\frac{22}{100}$ have medium interest, and the fraction $\frac{52}{100}$ have high interest. If H_0 is true, we would expect the following:

(1) Among the 60 students in row 1 (female):

$60 \times \frac{26}{100} = \frac{60 \times 26}{100} = 15.6$ should have low interest.

$60 \times \frac{22}{100} = \frac{60 \times 22}{100} = 13.2$ should have medium interest.

$60 \times \frac{52}{100} = \frac{60 \times 52}{100} = 31.2$ should have high interest.

(2) Among the 40 students in row 2 (male):

$40 \times \frac{26}{100} = \frac{60 \times 26}{100} = 10.4$ should have low interest

$40 \times \frac{22}{100} = \frac{40 \times 22}{100} = 8.8$ should have medium interest.

$40 \times \frac{52}{100} = \frac{40 \times 52}{100} = 20.8$ should have high interest.

Stated differently, we have the expected frequencies:

$E_{11} = 15.6 \quad E_{12} = 13.2 \quad E_{13} = 31.2$

$E_{21} = 10.4 \quad E_{22} = 8.8 \quad E_{23} = 20.8$

Note that we could get these expected frequencies by the direct application of equation (1.1).

From the observed and expected frequencies, the computed chi-square

$$= \frac{(9-15.6)}{15.6} + \frac{(12-13.2)^2}{13.2} + \frac{(39-31.2)^2}{31.2} + \frac{(17-10.4)^2}{10.4} + \frac{(10-8.8)^2}{8.8} + \frac{(13-20.8)^2}{20.8} = 12.128$$

Since the number of rows (R) is 2 and the number of columns (C) is 3, the degree of freedom of the chi-square test is $(2-1) \times (3-1) = 2$.

We can now test H_0 at any given level of confidence. For example, to test H_0 at a 99% level of confidence, we use the decision rule:

At a 99% level of confidence, reject H_0 if the computed chi-square exceeds $\chi^2_{.01}$ at degrees of freedom 2, that is, 9.21.

Since the computed chi-square = 12.218 > 9.21, we reject H_0 at a 99% level of confidence.

From the two examples given above, the following points should be noted:

Note 1: In any given row or column, the sum of the expected frequencies is same as the sum of observed frequencies.

Note 2: Since the sum of expected frequencies in any row is equal to the sum of observed frequencies, we can find one expected frequency in the row by subtraction once we compute the others using equation (1.1).

For example, in the cross-tabulation in Example (1.3), the sum of the observed frequencies and the sum of the expected frequencies are both 50. Thus, once we compute $E_{11} = 17.5$, $E_{12} = 12.5$, and $E_{13} = 8.75$ using (11.1), we can get E_{14} by subtraction:

$$E_{14} \quad = \quad 50 - (17.5 + 12.5 + 8.75) \; = \; 11.25.$$

Similarly, in any given column, we get one expected frequency by subtraction once we know the other expected frequencies. Therefore, to compute the expected frequencies in a table with r rows and c columns, we need to use equation (1.1) $(R-1) \times (C-1)$ times. This is the degree of freedom of the chi-square test.

Note 3: By using the chi-square test, we can determine if H_0 can be rejected at a given level of confidence. However, if H_0 is rejected, that is, a relationship does exist between the two variables, the chi-square test does not tell us the precise nature of the relationship.

Consider first the simplest case of a 2 × 2 cross-tabulation. As discussed in Example (1.2), this test is equivalent to testing $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 \neq \pi_2$. If H_0 is rejected, that is, we know that $\pi_1 \neq \pi_2$, the chi-square test by itself does not tell us whether $\pi_1 > \pi_2$, or, $\pi_1 < \pi_2$.

The picture gets even less clear when more than two rows or columns are involved. For example, in Example (1.3), we can reject H_0 at a 99% level of confidence, that is, conclude that the percentage distribution over interest categories is not same for men and women. However, the test does not tell us exactly where H_0 is violated, nor does it tell us the extent of the violation.

1.3 Procedure of Chi-Square Test

We now discuss how to conduct a chi-square test in the general case. Let X_1 and X_2 be two variables with r and c categories of values, respectively, and suppose we have obtained a cross-tabulation of frequencies for X_1 and X_2 where the different values of X_1 define the rows, and the different values of X_2 define the columns. Thus, we have a cross-tabulation with R rows and C columns.

We use the χ^2 test to test the null hypothesis: "X_1 and X_2 are not related," against the alternate hypothesis: "X_1 and X_2 are related."

Precise Meaning of H_0: Consider the sub-population that have $X_1 = i$. For this sub-population, let $\pi_{i1}, \pi_{i2}, \ldots, \pi_{ic}$ denote the proportions that have $X_2 = 1$, $X_2 = 2$, …, $X_2 = c$, respectively. With these notations, the null hypothesis means that all of the following are true:

$$\pi_{11} = \pi_{21} = \ldots = \pi_{r1}$$

$$\pi_{12} = \pi_{22} = \ldots = \pi_{r2}$$

$$\vdots$$

$$\pi_{1c} = \pi_{2c} = \ldots = \pi_{rc}$$

[Note that the total of each column here is 1. Consequently, the last row is redundant.]

Steps in the Chi-Square Test: We proceed as follows:

1. Compute the totals of rows 1, …, r.

2. Compute the totals of columns 1, …, c.

3. For each cell ij, compute the expected frequency by using equation (13.1):

$$E_{ij} \quad = \quad \{\text{Total of row}\, i\ \times \text{Total of column}\, j\}/n,$$

where n is the overall sample size.

4. Denoting the observed frequency in cell ij by O_{ij}, compute

$$\chi^2 \quad = \quad \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Decision Rule: At a confidence level of $(1-\alpha)$, reject H_0 if computed $\chi^2 > \chi^2_\alpha$ at degrees of freedom $= (r-1) \times (c-1)$.

P **Value:** The P Value for a chi-square test is the probability that χ^2 with degrees of freedom $(r-1) \times (c-1)$ equals or exceeds the computed χ^2. Statistical packages usually report the P Value of a chi-square test. If the P Value is available, then we can reject H_0 at any level of confidence $(1-\alpha)$ as long as $\alpha < P$ Value.

1.4 Application and Conditions for Using Chi-Square Test

The chi-square test is a simple and robust test of the relationship between any two variables, which can be used regardless of scales of measurement involved. For example, we can use a chi-square test to examine:

- Relationship between ethnicity and preferred type of music. (Two nominal variables)
- Relationship between gender and preferred color of automobile. (Two nominal variables)
- Relationship between brand of toothpaste purchased in two successive time periods. (Two nominal variables)
- Relationship between political affiliation and attitude towards President Bush. (Nominal and ordinal variables)
- Relationship between per capita advertising expenditure in sales territory and per capita sales. (Two ratio scaled variables)

In all these cases, the chi-square test can be used to determine if a relationship exists at all, before making a more sophisticated examination using advanced techniques such as logit.

However, a chi-square test is **only** valid if the expected frequencies in the cells are reasonably large. Strictly speaking, the expected frequency in **every cell** should be at least 5. In practice, the requirement is relaxed somewhat, as described below:

To conduct a chi-square test, the following conditions must **both** *be satisfied:*

(1) $E_{ij} > 1$ *in* **all** *cells.*

(2) $E_{ij} \geq 5$ *in 80% or more of the cells.*

These conditions are violated when we have too few observations in a row or a column. Then, it is necessary to modify the original cross-tabulation by combining two or more rows and/or columns so that it is valid to apply the chi-square test. The following example demonstrates how that is done.

Example 1.4 *Suppose we have collected a simple random sample of 60 students from a college campus and asked them to rate how much they like to watch professional sports on TV on a 1-7 scale (strongly dislike to strongly like). We also noted the gender of each respondent. Based on the results, we have constructed the following cross-tabulation:*

	Like to watch professional sports on TV						
Gender	1	2	3	4	5	6	7
Male	2	0	4	12	6	8	8
Female	4	3	2	6	3	1	1

At a 99% level of confidence, test the null hypothesis that gender is not related to how much one likes to watch professional sports on TV.

Answer: The cross-tabulation augmented by row totals and column totals is given below.

Gender	Like to watch sports on TV 1	2	3	4	5	6	7	**Row Totals**
Male	2	0	4	12	6	8	8	**40**
Female	4	3	2	6	3	1	1	**20**
Column Totals	**6**	**3**	**6**	**18**	**9**	**9**	**9**	

The overall sample size is 60. Hence, from equation (1.1), the expected frequencies in the cells are:

$$E_{11} = \frac{40 \times 6}{60} = 4 \quad E_{12} = \frac{40 \times 3}{60} = 2 \quad E_{13} = \frac{40 \times 6}{60} = 4 \quad E_{14} = \frac{40 \times 18}{60} = 12$$

$$E_{15} = \frac{40 \times 9}{60} = 6 \quad E_{16} = \frac{40 \times 9}{60} = 6 \quad E_{17} = \frac{40 \times 9}{60} = 6$$

$$E_{21} = \frac{20 \times 6}{60} = 2 \quad E_{22} = \frac{20 \times 3}{60} = 1 \quad E_{23} = \frac{20 \times 6}{60} = 2 \quad E_{24} = \frac{20 \times 18}{60} = 6$$

$$E_{25} = \frac{20 \times 9}{60} = 3 \quad E_{26} = \frac{20 \times 9}{60} = 3 \quad E_{27} = \frac{20 \times 9}{60} = 3$$

Since $E_{22} = 1$, chi-square test is not valid with the original cross-tabulation. Since we have only two rows, we must combine columns to make the chi-square test valid. There are different ways that can be done, and any one of them is acceptable.

For example, we may combine columns 1, 2 and 3 in the original table into new column 1, keep column 4 in the original table as new column 2, and combine columns 5, 6 and 7 in the original table as new column 3. For this new table, we have 2 rows and 3 columns, and the observed and expected frequencies as follows:

$O_{11} = 6 \quad O_{12} = 12 \quad O_{13} = 22$
$E_{11} = 10 \quad E_{12} = 12 \quad E_{13} = 18$

$O_{21} = 9 \quad O_{22} = 6 \quad O_{23} = 5$
$E_{21} = 5 \quad E_{22} = 6 \quad E_{23} = 9$

Decision Rule: At a 99% level of confidence, reject H_0 if
computed $\chi^2 > \chi^2_{.05}$ *at degrees of freedom* $= (2-1) \times (3-1) = 2$, *that is, computed* $\chi^2 > 9.21$.
Conclusion: Here, computed χ^2

$$= \frac{(6-10)^2}{10} + \frac{(12-12)^2}{12} + \frac{(22-18)^2}{18} + \frac{(9-5)^2}{5} + \frac{(6-6)^2}{6} + \frac{(5-9)^2}{9} = 7.47.$$

Since computed $\chi^2 = 7.47$ does not exceed 9.21, we cannot reject H_0 at a 99% level of confidence.

Note on Table Modification: There is no fixed rule about how rows and columns should be modified, and different combinations may make the test valid. Any time we combine multiple cells in the original table into a new cell, we can determine the observed or expected frequency in a new cell by adding the observed or expected frequencies in the cells that got merged.

For example, in Example (1.4), we got the (1,1) cell in the new table by merging the cells (1,1), (1,2) and (1,3) of the original table. Thus, for the new table,

$$O_{11} = 2 + 0 + 4 = 6, \quad \text{and} \quad E_{11} = 4 + 2 + 4 = 10$$

Thus, once we have the expected frequencies from the original table, we can quickly try out combinations of rows or columns so that we can use the chi-square test.

1.5 Using Chi-square Analysis to Test if Proportions are equal in multiple sub-populations

Suppose you have a population with two or more sub-populations, and wish to test if the same proportion of each sub-population have a property of interest. For example:

- Does the same proportion of students with different majors in a college read the Wall Street Journal?
- Does the same proportion of people in different age groups own Apple iPhones?
- Does the same proportion of stores in different countries have check-out scanners?

The chi-square test can be used to test the null hypothesis that the proportion is equal for all sub-populations. For example, suppose the population has k sub-populations, and the proportions of the sub-populations that have the property of interest are π_1, π_2, …, π_k. Then, the following two tests are equivalent:

- Test of the null hypothesis $H_0 : \pi_1 = \ldots = \pi_k$ against the alternative hypothesis that at least one of these proportions is different from the others.
- Chi-square test of the null hypothesis that X and Y are not related to each other using a $(k \times 2)$ cross tabulation (k rows, two columns) of X and Y where:
  - $X = i$ if a case is from sub-population i.
  - $Y = 1$ if the case has the property of interest, and 0 if not.

Example 1.5 Suppose you have collected simple random samples from three sub-populations: Business majors, Engineering majors, and "other" majors. For each respondent, you recorded if (s)he reads the Wall Street Journal (WSJ) every week.

Results:

(1) Business sample: $n_1 = 100$, 50 read WSJ every week.

(2) Engineering sample: $n_2 = 50$, 15 read WSJ every week.

(3) "Other: sample: $n_3 = 150$, 25 read WSJ every week.

At a 99% level of confidence, test the null hypothesis that an equal proportion of business, engineering, and other students read the Wall Street Journal every week.

Answer: The information provided can be expressed as the following cross-tabulation:

	Do not Read WSJ	Read WSJ
Business	50	50
Engineering	35	15
Other	125	25

Testing the null hypothesis that the same proportion of the three sub-populations read WSJ is equivalent to testing the null hypothesis that there is no relationship between major and reading WSJ.

Augmenting the cross-tabulation with row totals and column totals, we get

	Do not Read	Read	**Row Totals**
Business	50	50	**100**
Engineering	35	15	**50**
Other	125	25	**150**
Column Totals	**210**	**90**	Sample Size = 300

$$E_{11} = \frac{100 \times 210}{300} = 70 \qquad E_{12} = \frac{100 \times 90}{300} = 30$$

$$E_{21} = \frac{50 \times 210}{300} = 35 \qquad E_{22} = \frac{50 \times 90}{300} = 15$$

$$E_{31} = \frac{150 \times 210}{300} = 105 \quad E_{32} = \frac{150 \times 90}{300} = 45$$

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$
$$+ \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}}$$
$$= \frac{(50-70)^2}{70} + \frac{(50-30)^2}{30} + \frac{(35-35)^2}{35} + \frac{(15-15)^2}{15}$$
$$+ \frac{(125-105)^2}{105} + \frac{(25-45)^2}{45} = 31.75$$

Decision Rule: At a 99% level of confidence, reject H_0 if $\chi^2 > 9.21 = \chi^2_{.01}$ at df $= (3-1) \times (2-1) = 2$

Conclusion: Since $\chi^2 = 31.75 > 9.21$, we reject H_0 at a 99% level of confidence, and conclude that the three proportions are not all equal.

General Approach:

- Suppose you are testing if an equal proportion of k sub-populations have a property of interest (e.g., read Wall Street Journal every week), that is,

$$\pi_1 = \pi_2 = \ldots = \pi_k$$

- This is equivalent to a chi-square test with a $k \times 2$ cross-tabulation where each row comes from one sub-population, and the two columns are "do not have property," and "have property."

- Express the data as a $k \times 2$ cross tabulation. For any sub-population:

  Number who do not have property

  = Size of the sample from the sub-population − Number from sub-population who have property

- Assuming test is valid, reject H_0 if χ^2 exceeds χ^2_α at degrees of freedom $(k-1) \times (2-1) = k-1$.

1.6 Optional: Cross Tabulation and Chi-Square Test in Minitab

We provide two examples of cross-tabulation and chi-square test in Minitab using the Carrier Dome Data posted on Blackboard.

Example 1. Suppose you wish to test if there is a relation between gender (X_1), and if the student attended a football game in the Carrier Dome in the "last one year" (X_{10b}). Proceed as follows:

(1) Open the worksheet in Minitab.

(2) Click "Stat" in the menu line, drag cursor to "Tables," and click on "Cross Tabulation and Chi-Square." A dialog box opens.

(3) Mark the top line on the right ("For rows"). Click on $X1$ in the list in the left box and then click "Select."

(4) Mark the second line from the top on the right ("For columns"). Click on $X10b$ in the left box and then click "Select."

(5) Under "Display" at the right of the screen, click "Counts" and "Row percents."

(6) In the right of the dialog box, click on "Chi-square." In the box that opens, mark "Chi-square analysis," and click OK.

(7) Back in main dialog box, click OK.

Minitab prints out the following cross-tabulation:

```
Rows X1   Columns X10b
          0       1        All

0         38      37       75
          50.67   49.33    100.00

1         14      59       73
          19.18   80.82    100.00

All       52      96       148
          35.14   64.86    100.00

Pearson Chi-Square = 16.095, DF = 1, P-Value = 0.000
```

The output gives the following information:

- The cross-tabulation itself is:

	X_{10b}	
X_1	0	1
0	38	37
1	14	59

- The column at the right of the output table gives the row totals and sample size. The row at the bottom of the output table gives column totals and sample size.
- The percentage of each row ($X_1 = 0$ and $X_1 = 1$) that has $X_{10b} = 0$ and $X_{10b} = 1$. Thus, 49.33% of women and 80.82% men attended a football game at Carrier Dome.
- The computed Chi-Square.
- The degree of freedom (DF) $= (2 - 1) \times (2 - 1) = 1$.
- The P value for the null hypothesis of no relationship. Here, $P = 0.000$ means the P value is less than 0.001. Thus, H_0 can be rejected at a 99.9% level of confidence.

Note: If you wish to get the expected frequencies (E_{ij}'s) in the cells, mark "Expected cell counts" in addition to "Chi-square analysis" in Step 6 above.

Example 2. Suppose you wish to test if there is a relation between gender (X_1) and the student's interest in going to bars (X_{8d}). Proceeding as in Example 1, you get the following output:

```
Rows: X1   Columns: X8d
           1     2     3     4      5      6      7      All

0          0     2     4     8      14     19     28     75
           0.00  2.67  5.33  10.67  18.67  25.33  37.33  100.00

1          1     1     3     8      17     21     22     73
           1.37  1.37  4.11  10.96  23.29  28.77  30.14  100.00

All        1     3     7     16     31     40     50     148
           0.68  2.03  4.73  10.81  20.95  27.03  33.78  100.00

Pearson Chi-Square = 2.560, DF = 6
* WARNING * 2 cells with expected counts less than 1
* WARNING * Chi-Square approximation probably invalid
* NOTE * 6 cells with expected counts less than 5
```

Thus, in this case, it is not valid the do chi-square analysis with the original table. Clearly, the problem arises because columns 1, 2, and 3 have too few observations. You can also do this step formally by looking at expected frequencies in the cells (E_{ij}'s). To get expected frequencies, do the following:

After you click “Chi-Square” (Step 6 in Example 1), mark both “Chi-Square analysis” and “Expected cell counts” and then click OK.

Chi-Square Analysis with Recoded Data: We can recode X_{8d} and do valid chi-square analysis. There are different valid ways to recode X_{8d}, and two examples are shown below.

Recoding 1. Suppose we wish create a variable $BAR1$ which is 1 if $X_{8d} = 1, 2, 3$; 2 if $X_{8d} = 4$; 3 if $X_{8d} = 5$; 4 if $X_{8d} = 6$; and 5 if $X_{8d} = 7$. We can create the recoded variable as follows:

- Click on “Data,” drag cursor to “Code,” and click “numeric to numeric.”
- Mark the “code data from columns” in the dialog box and select X_{8d}. In the line below, enter the column where you wish to place the recoded variable.
- Enter the following in “original values” and “New:”

Original Values	New
1:3	1
4	2
5	3
6	4
7	5

- Click OK

- In the cell at the top of the column of the recoded variable, enter BAR1.

You can now do a cross-tabulation of X_1 and BAR1 as in Example 1. The results are as follows:

```
Rows: X1  Columns: BAR1
          1     2      3      4      5      All

0         6     8      14     19     28     75
          8.00  10.67  18.67  25.33  37.33  100.00

1         5     8      17     21     22     73
          6.85  10.96  23.29  28.77  30.14  100.00

All       11    16     31     40     50     148
          7.43  10.81  20.95  27.03  33.78  100.00

Pearson Chi-Square = 1.174, DF = 4, P-Value = 0.882
```

Clearly, you cannot reject H_0 (no relationship) at any reasonable level of confidence.

Recoding 2. Suppose we wish to create a variable $BAR2$ which is 1 if X_{8d} = 1, 2, 3; 2 if $X_{8d} = 4$; 3 if X_{8d} = 5, 6, or 7. We can create the recoded variable as follows:

- Click on "Data," drag cursor to "Code," and click "numeric to numeric."
- Mark the "code data from columns" in the dialog box and select X_{8d}. In the line below, enter the column where you wish to place the recoded variable.
- Enter the following in "original values" and "New:"

Original Values	New
1:3	1
4	2
5:7	3

- Click OK
- In the cell at the top of the column of the recoded variable, enter BAR2.

You can now do a cross-tabulation of X_1 and BAR2 as in Example 1. The results are as follows:

```
    Rows: X1   Columns: BAR2
               1      2      3      All

    0          6      8      61     75
               8.00   10.67  81.33  100.00

    1          5      8      60     73
               6.85   10.96  82.19  100.00

    All        11     16     121    146
               7.43   10.81  81.76  100.00

Pearson Chi-Square = 0.072, DF = 2, P-Value = 0.965
```

Once again, we cannot reject H_0 at any reasonable level of confidence.

1.7 Exercise Problems

1. We have drawn a simple random sample of size 40 from the students of a university, and asked them two questions:

(i) How many hours do you work (on a job) each week?

(ii) On the average, how many dollars do you spend on fast-food each week?

From the data, we obtained the following cross-tabulation:

	Number of hours one works			
Expenditure on Fast Food	0	1−10	11−20	over 20
$ 10 or less	2	2	0	0
$ 11 - $ 20	8	8	2	2
Over $20	2	4	8	2

At a 95% level of confidence, test the null hypothesis that the number of hours a student works is not related to expenditure on fast-food.

2. We have selected a simple random sample of 50 students at a university and recorded whether the student lives in the dorm, and how much the student spends in a month eating out. From the data, we obtained the following cross-tabulation:

	Expenditure/month (in dollars)		
Residence	0−20	21−40	> 40
Dorm	10	9	1
Not Dorm	5	13	12

At a 95% level of confidence, test the null hypothesis that the expenditure eating out is not related to whether the student lives in a dorm or not.

3. Suppose you have drawn a simple random sample from a university and asked each respondent if (s)he owned an Apple iPhone. There are four categories within the university population:

undergraduate students, graduate students, faculty and staff. The data, separated by category, are given below:

Category	Number drawn from category	Number that own Apple iPhones
1. Undergraduate Student	90	60
2. Graduate Student	40	20
3. Faculty	40	15
4. Staff	30	10

At a 99% level of confidence, test the null hypothesis that equal proportions of the four categories listed above own iPhone.

2 Regression Analysis

2.1 Introduction

2.1.1 Regression Model in General Form: Regression analysis is used to analyze how a **dependent** variable Y is related to one or more **independent** variables. The model is expressed as follows:

$$(2.1) \quad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \epsilon,$$

where:

Y = the dependent variable;

$X_1, \ldots, X_n$ are the independent variables, also called **regressors**; and

ϵ is a normally distributed **error** which captures the total effect of all variables left out of the model.

We make the following assumptions about the error term ϵ:

- ϵ is normally distributed with mean zero.
- The standard deviation of ϵ, denoted by σ_ϵ, is same for all observations and does not depend on the levels of the independent variables. (The violation of this assumption is called *heteroscedasticity.*)
- ϵ is uncorrelated from observation to observation. (The violation of this assumption is called *autocorrelation.*)

The parameters $\beta_0, \beta_1, \ldots, \beta_n$ are same for all observations and called the *regression coefficients.*

From equation (2.1) and the assumption that the expected value of ϵ is zero, it follows that:

$$(2.2) \quad E(Y|X_1, \ldots, X_n) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n,$$

that is, given $X_1, \ldots, X_n$, on the average Y is equal to $\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_n$. Equation (2.2) allows us to examine how Y change on the average if one or more of the independent variables change.

2.1.2 Nature of the dependent variable: Since we want to examine changes in Y, the measurement of Y should satisfy the properties of an interval scale.[1] For example, per capita expenditure on a brand, or the monthly dollar sales at a store, satisfy this condition. While not strictly accurate, it is common practice to use a sum of itemized rating scales, such as attitude measured on a likert summated scale, as a dependent variable in regression analysis. Sometimes,

[1] In any measurement, we assign a number or symbol to represent a property of an object. A measurement has interval scale properties if the same difference in the measured number always represents the same difference in the underlying variable. In contrast, a variable is called categorical if it is only used to place objects in different categories, such as the gender of a student.

even a single measurement on an itemized rating scale or a rank order rating scale is used as a dependent variable. Typically, the error arising from such use is not great.

2.1.3 Nature of the independent variables: The independent variables can be of three types:

(1) **Interval Scaled Variables:** These are variables that satisfy interval scale properties, such as per capita advertising expenditure, age, and household income. As with the dependent variable, a measurement on an itemized rating scale or a rank order rating scale, or a sum of itemized ratings scales, are assumed to approximately satisfy interval scale properties.

(2) **Dummy Variables:** A dummy variable is a dichotomous variable. We will only discuss dummy variables that can be 1 or 0.

Dummy variables are used to represent variables measured on a categorical scale. To represent a categorical variable with C categories, we need $(C - 1)$ dummy variables. Consider two examples:

> Example 1. Gender: Gender has two categories, male and female ($C = 2$). This can be represented by a single dummy variable D that is 1 if male, and 0 if female. We need only one dummy variable because if somebody does not belong to the first category, then (s)he automatically belongs to the second category.
>
> Example 2. Location of store: Suppose we have a categorical variable "location" that places grocery stores in Continental USA into four mutually exclusive and collectively exhaustive categories: Northeast, South, Midwest, and West. Here $C = 4$, and we can represent location by three dummy variables D_1, D_2, and D_3, defined as follows:
>
> $D_1 = 1$ if Northeast, and $D_1 = 0$ if not Northeast.
>
> $D_2 = 1$ if South, and $D_2 = 0$ if not South.
>
> $D_3 = 1$ if Midwest, and $D_3 = 0$ if not Midwest.
>
> Since the categories are mutually exclusive, no two D's can be 1 at the same time. If $D_1 = D_2 = D_3 = 0$, then we automatically know that the location is West. Thus, we do not need a fourth dummy variable for West.

(3) An **interaction** variable that is a product of two or more dummy or interval scaled variables.

2.1.4 Our Focus: In our discussion of regression analysis, we focus on three issues:

- The meaning of regression coefficients for different types of independent variables.
- The idea behind coefficient estimation, and the meaning of R^2.
- t and F tests with regression results.

2.2 Five Simple Models

We first discuss five simple regression models to illustrate the meaning of the regression coefficients. Then, we use the intuition derived to discuss more elaborate models. To make our discussion concrete, we consider the population of business school *graduates*, who completed college education and accepted a full time job in 2003. Y denotes the annual salary of a graduate. We will examine how Y depends on two independent variables:

(1) A dichotomous variable that records whether the student graduated from a program ranked in the top twenty or not. This is represented by a dummy variable D that is 1 if top 20, and 0 if not.

(2) The cumulative grade point average, denoted by X, which can range from 1 to 4.

Model 2.2.1 (Naive Model): The simplest regression model, called the *naive model* is given by:

$$(2.3)\quad Y \quad = \quad \beta_0 \, + \, \epsilon.$$

In this model, $E(Y)$ is the same, β_0, for all observations. Thus, β_0 is the population mean of Y, and ϵ represents how much Y deviates from the population mean in a given case. In the context of the graduate population, β_0 is the population mean salary, and ϵ is how much the salary of any given graduate differs from the population mean. Note that the naive model assumes that salary is normally distributed over the graduate population.

If we estimate the naive model, the estimate of the coefficient β_0 is $\overline{Y}$, the sample mean. The naive model provides a benchmark to measure the performance of regression models that incorporate independent variables.

Model 2.2.2 (One Dummy Independent Variable): Next, we extend the naive model by adding a dichotomous independent variable. Consider the model:

$$(2.4)\quad Y \quad = \quad \beta_0 + \beta_1 * D + \epsilon,$$

where $D = 1$ if the graduate comes from a top 20 program, and 0 if not.

We can express the model separately for the two levels of D as follows:

(1) $D = 0$ (not top 20): $Y = \beta_0 + \epsilon$.

(2) $D = 1$ (top 20): $Y = (\beta_0 + \beta_1) + \epsilon$.

Thus, for the sub-population of graduates of programs not ranked in the top 20, the mean of Y is β_0. For the sub-population of graduates of top 20 programs, the mean of Y is $(\beta_0+\beta_1)$. Both sub-populations have the same standard deviation, σ_ϵ. The model (2.4) therefore simply captures the possibility that the mean of Y may be different for the two sub-populations corresponding to the two levels of D. If both sub-populations have the same mean, then $\beta_1 = 0$.

Model 2.2.3 (One Interval Scaled Independent Variable): Here, we extend the naive model by adding one independent variable measured on an interval scale. This model, called the *two variable regression model*, is given as follows:

$$(2.5)\quad Y \quad = \quad \beta_0 + \beta_2 X + \epsilon.$$

For example, suppose X is the cumulative grade point average of the graduate, which can range from 1 to 4. For a given value of X, it follows from (2.5) that:

$$E(Y|X) \quad = \quad \beta_0 + \beta_2 X.$$

Figure (2.1) plots $E(Y|X)$ against X. The plot is a straight line, which is called a "regression line." The *slope* of the line is β_2, and the *intercept* of the line with the ordinate is β_0. Thus, if X changes by one unit, on the average Y changes by β_2 units.

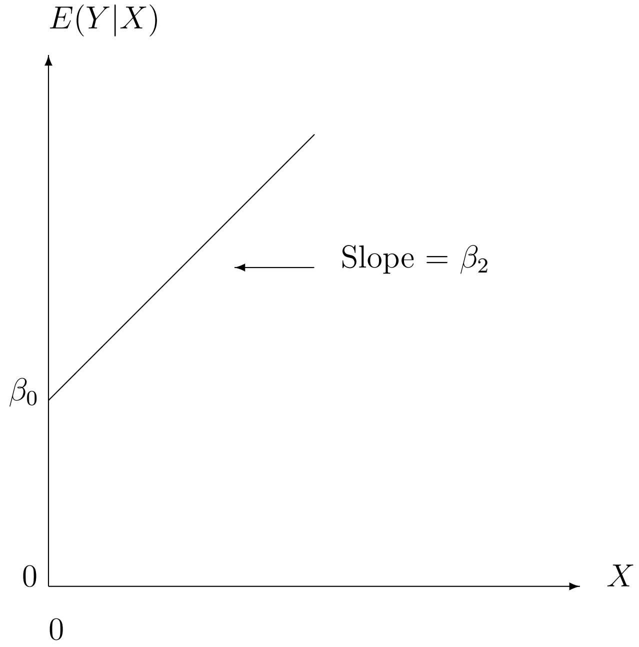


Figure 2.1

Intuitively, we are dividing the population of graduates into many sub-populations, each with a specific value of X. For a sub-population defined by a specific value of X, Y is normally distributed with a mean equal to $(\beta_0 + \beta_2 X)$, and a standard deviation equal to σ_ϵ.

Model 2.2.4 (One Dummy and One Continuous Independent Variables): Consider next the model which incorporates both a dummy and a continuous independent variables:

$$(2.6) \quad Y \quad = \quad \beta_0 + \beta_1 D + \beta_2 X + \epsilon$$

In our example:

- $D = 1$ if the graduate is from a top 20 program, and 0 if not, and
- X = cumulative GPA.

We can express the model separately for the two levels of D as follows:

(1) $D = 0$ (not top 20): $Y = \beta_0 + \beta_2 X + \epsilon$.

(2) $D = 1$ (top 20): $Y = (\beta_0 + \beta_1) + \beta_2 X + \epsilon$.

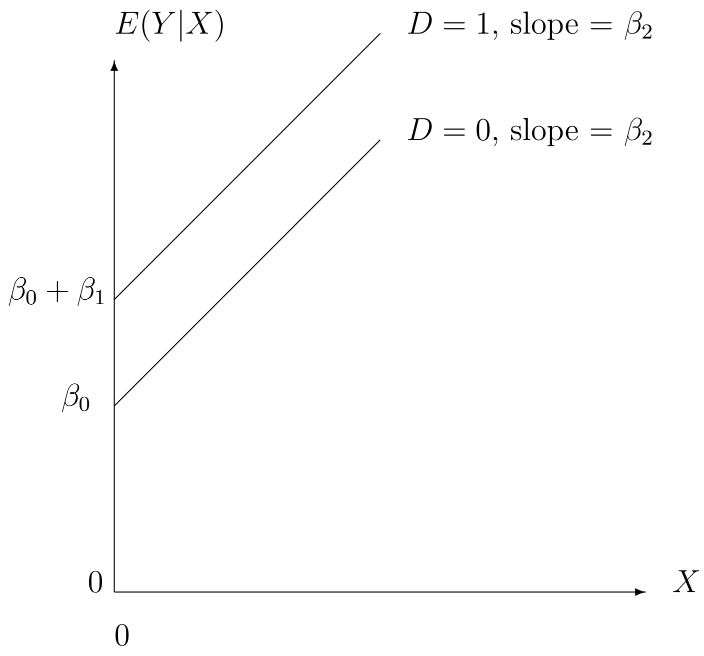


Figure 2.2

As shown in Figure (2.2), the model in equation (2.6) allows us to have two distinct regression lines relating $E(Y|X)$ to X: one for the sub-population with $D = 0$, and another for the sub-population with $D = 1$. Both regression lines have the same slope, β_2. The intercept is $(\beta_0 + \beta_1)$ for graduates of top 20 programs, and is β_0 for others. *Thus, the inclusion of the independent variable D allows the intercept term to be different for the two levels of D.* If $\beta_1 = 0$, then both lines have the same intercept, and the model reduces to the two variable regression model given by equation (2.5).

Model 2.2.5 (Model with Interaction Term): Finally, the following model has, as independent variables, a dummy variable, an interval scaled variable, and a product of a dummy and a continuous variables. Consider:

$$(2.7) \quad Y \;=\; \beta_0 + \beta_1 D + \beta_2 X + \beta_3 * D * X + \epsilon.$$

Writing the equation separately for $D = 0$ and $D = 1$, we get:

(1) $D = 0$ (not top 20): $Y = \beta_0 + \beta_1 X + \epsilon$.

(2) $D = 1$ (top 20): $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X + \epsilon$.

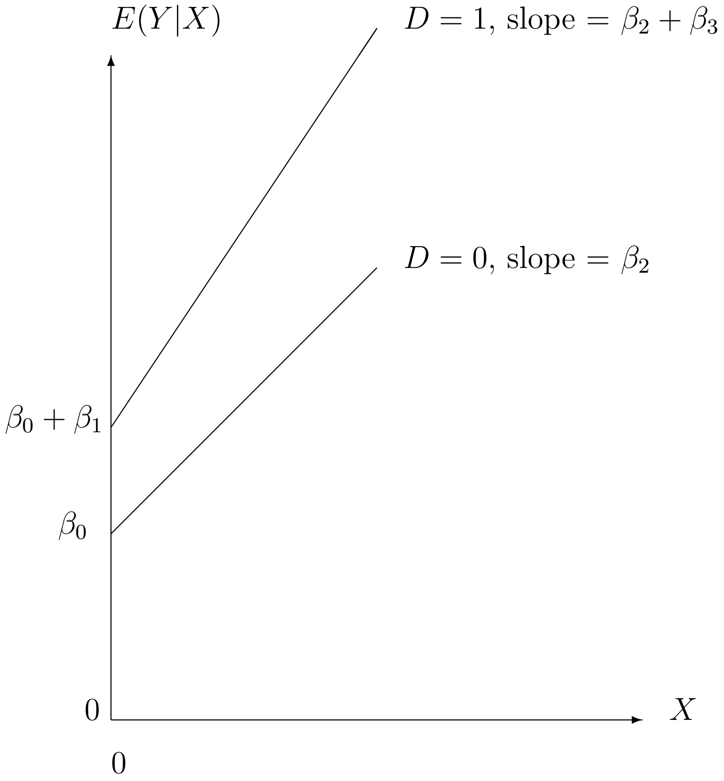

$E(Y|X)$
$D = 1$, slope $= \beta_2 + \beta_3$
$D = 0$, slope $= \beta_2$
$\beta_0 + \beta_1$
β_0
0
X
0

Figure 2.3

As shown in Figure (2.3), the model in equation (2.7) allows us to have two different regression lines relating $E(Y|X)$ to X. The regression line for $D = 0$ has an intercept of β_0, and a slope of β_2, while the regression line for $D = 1$ has an intercept of $(\beta_0 + \beta_1)$ and a slope of $(\beta_2 + \beta_3)$. Thus, β_1 is difference in intercepts, and β_3 is the difference in slopes, between the two lines.

The inclusion of D as an independent variable allows the intercept to be different. The inclusion of $D * X$ as an independent variable allows the slope to be different for the two subpopulations (top 20, and not top 20).

Note that this model includes the four previous models as special cases. First, if $\beta_3 = 0$ (interaction variable $D * X$ is removed), equation (2.7) reduces to equation (2.6) as the two regression lines have the same slope.

If we have $\beta_1 = 0$ and $\beta_3 = 0$, the intercepts become equal, and we have the two variable regression line given by equation (2.5) and shown in Figure (2.1).

If we have $\beta_2 = 0$ and $\beta_3 = 0$, then we have two regression lines parallel to the X axis with intercepts at β_0 and $(\beta_0 + \beta_1)$. This time, a change in X has no effect on $E(Y)$, and the model reduces to that given by equation (2.4).

Finally, if $\beta_1 = \beta_2 = \beta_3 = 0$, we have a regression line parallel to the X axis with intercept β_0. This is the naive model given by equation (1.3) where $E(Y)$ does not depend on D or X.

2.3 Examples of More Complex Models

We now extend the models given in Section 2.2 to include multiple continuous variables, and categorical variables with more than two categories. Once again, the focus of the discussion is on the interpretation of regression coefficients.

2.3.1 Models with only Interval Scaled Independent Variables

We first present two models with only interval scaled independent variables.

Model 2.3.1.1 (No Interaction Term): Consider again the general form of the regression model:

$$Y \ = \ \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

For this model,

$$E(Y|X_1, \ldots, X_n) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n.$$

Suppose the independent variables X_1, …, X_n all satisfy interval scale properties. Then, the regression coefficient β_i can be interpreted as follows:

If we increase X_i by a unit while keeping the other X's the same, then **on the average**, *Y will increase by β_i units.*

β_i is commonly called the "slope of Y with respect to X_i." The coefficient β_0 is called the "intercept" term.

Example. Market Build-Up Model: *A common application of regression analysis is to use it to predict future sales of an industrial product. For example, suppose we manufacture an industrial product (e.g., fiberglass insulation), and sell it primarily to four different industry categories. We can use the model*

$$Y \quad = \quad \beta_0 \ + \ \beta_1 X_1 \ + \ \beta_2 X_2 \ + \ \beta_3 X_3 \ + \ \beta_4 X_4 \ + \ \epsilon,$$

where:

- Y = *quarterly sales of our product category.*
- X_1 = *quarterly dollar sales of the output of industry category 1.*
- X_2 = *quarterly dollar sales of the output of industry category 2.*
- X_3 = *quarterly dollar sales of the output of industry category 3.*
- X_4 = *quarterly dollar sales of the output of industry category 4.*

If we can get good forecasts for X_1, X_2, X_3, X_4 using secondary data, then we can predict demand for our product category.

It is quite easy to identify situations where a market buildup model may be useful. For example, a marketer may use census data to determine how many houses, apartments, condominiums, *and office buildings will be built in 2005, and using that information predict the demand for fiberglass insulation.*

Assumptions: *In using the market build-up model, we are assuming that the demand for our product category can be expressed as the sum of demands which come from selling to the four industry categories, plus a remainder term. The term $(\beta_0 + \epsilon)$ represents this remaining demand, β_0 denoting its average, and ϵ any fluctuation from the average.*

We are further assuming that for any of these four industry categories, the demand for our product category is proportional to the output from that industry category. We are therefore making the crucial assumptions that as the output level X_1 increases, industry 1 continues to use our product category and does not switch to a substitute product category (e.g., switch from sugar to corn syrup).

Sometimes a company may use a market build up model to predict the sales of a specific brand instead of a product category. Then, Y is the sales of the specific brand. In using the market build-up model here, we make the additional assumption that the market share of the brand remains the same when product category demand changes. Since market share depends on competitive factors like advertising expenditure and price, this is a very strong assumption.

Model 2.3.1.2 (Includes Interaction Term): Next, consider the model:

$$(2.8) \quad Y \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 * X_1 * X_2 + \epsilon,$$

where X_1 and X_2 satisfy interval scale properties. To make the discussion concrete, let Y be the per capita sales, X_1 the per capita advertising expenditure, and X_2 the average retail price, in a sales region.

From equation (2.8),

$$E(Y|X_1, X_2) \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

If we hold X_2 fixed, then the plot of $E(Y|X_1, X_2)$ against X_1 gives a regression line with an intercept of $(\beta_0 + \beta_2 X_2)$, and a slope of $(\beta_1 + \beta_3 X_2)$. Thus, both the intercept and slope of the regression line depend on the level at which X_2 is held fixed.

Similarly, if we hold X_1 fixed, then the plot of $E(Y|X_1, X_2)$ against X_2 gives a regression line with intercept of $(\beta_0 + \beta_1 X_1)$, and a slope of $(\beta_2 + \beta_3 X_1)$.

Therefore, the inclusion of the *interaction variable* $X_1 * X_2$ allows the marginal effect of X_1 on Y to depend on X_2, and the marginal effect of X_2 on Y to depend on X_1.

2.3.2 Models with Only Dummy Independent Variables

We now discuss regression models with only categorical independent variables, coded as dummy variables. To make our discussions concrete, we again consider the population of business school graduates used in Section 2.2. Once again, Y is the annual salary of the graduate, and we have a dichotomous variable to indicate whether the graduate attended a top 20 program or not.

We now add another categorical variable, the major area of study (major) of the graduate. For simplicity, we assume that a graduate had exactly one major, which could be accounting, finance, marketing, or "other."

To represent these four levels of major, we define three dummy variables D_1, D_2, and D_3 as follows:

$D_1 = 1$ if the graduate had an accounting major, and 0 if not.

$D_2 = 1$ if the graduate had a finance major, and 0 if not.

$D_3 = 1$ if the graduate had a marketing major, and 0 if not.

In order to avoid confusion, we use the dummy variable D_4 to indicate whether the graduate attended a top 20 program or not: $D_4 = 1$ if top 20, and $D_4 = 0$ if not.

Model 2.3.2.1 (One Categorical Independent Variable): Consider the following model which is also called the **ANOVA** model:

$$(2.9) \quad Y \;=\; \beta_0 \,+\, \beta_1 D_1 \,+\, \beta_2 D_2 \,+\, \beta_3 D_3 \,+\, \epsilon,$$

where Y is the starting salary, and D_1, D_2, and D_3 are as defined above. Note that we cannot change any of the D's without affecting others. Thus, we need to interpret the coefficients differently from how we do it for interval scaled independent variables.

First, we write the regression equation separately for the four majors:

Accounting ($D_1 = 1$, $D_2 = 0$, $D_3 = 0$):

$E(Y) \;\equiv\; E(Y|D_1 = 1, D_2 = 0, D_3 = 0) \;=\; \beta_0 \,+\, \beta_1$, and $Y \;=\; \beta_0 \,+\, \beta_1 \,+\, \epsilon$.

Finance ($D_1 = 0$, $D_2 = 1$, $D_3 = 0$):

$E(Y) \;\equiv\; E(Y|D_1 = 0, D_2 = 1, D_3 = 0) \;=\; \beta_0 \,+\, \beta_2$, and $Y \;=\; \beta_0 \,+\, \beta_2 \,+\, \epsilon$.

Marketing ($D_1 = 0$, $D_2 = 0$, $D_3 = 1$):

$E(Y) \;\equiv\; E(Y|D_1 = 0, D_2 = 0, D_3 = 1) \;=\; \beta_0 \,+\, \beta_3$, and $Y \;=\; \beta_0 \,+\, \beta_3 \,+\, \epsilon$.

Other ($D_1 = 0$, $D_2 = 0$, $D_3 = 0$):

$E(Y) \;\equiv\; E(Y|D_1 = D_2 = D_3 = 0) \;=\; \beta_0$, and $Y \;=\; \beta_0 \,+\, \epsilon$.

Thus, β_0 is the average salary of the "other" subpopulation. $(\beta_0 + \beta_1)$ is the average salary of the "accounting" subpopulation, and β_1 is the difference between the average salaries of accounting graduates and "other" graduates. Similarly, β_2 is the difference between the average salaries finance graduates and "other" graduates, and β_3 is the difference between the average salaries of marketing graduates and "other" graduates.

For any given major, ϵ for an individual graduate represents how much that individual differs from the average of that major. The regression model assumes that the variance of ϵ and hence the variance of Y is same for all four majors.

Model 2.3.2.2 (Two Categorical Independent Variables, No Interaction Term): Consider next the model:

$$(2.10) \quad Y \;=\; \beta_0 \,+\, \beta_1 D_1 \,+\, \beta_2 D_2 \,+\, \beta_3 D_3 \,+\, \beta_4 D_4 \,+\, \epsilon,$$

where Y, D_1, D_2, and D_3 are as given in the previous example (Model 1.3.2.1), and $D_4 = 1$ if the graduate is from a top 20 program, and 0 if not.

In this case, we are placing the graduates into $4 \times 2 = 8$ categories based on major and program rank (two ranks for each of four majors). $E(Y)$ can have the following eight different values depending on major and program rank:

	Not Top 20	Top 20
Accounting	$E(Y) = \beta_0 + \beta_1$	$E(Y) = \beta_0 + \beta_1 + \beta_4$
Finance	$E(Y) = \beta_0 + \beta_2$	$E(Y) = \beta_0 + \beta_2 + \beta_4$
Marketing	$E(Y) = \beta_0 + \beta_3$	$E(Y) = \beta_0 + \beta_3 + \beta_4$
Other	$E(Y) = \beta_0$	$E(Y) = \beta_0 + \beta_4$

Note that for a given major, the difference between the averages of "top 20" and "not top 20" is β_4, and this difference is same for all four majors. We say that in this case, there is no *interaction* between the effects of major and program rank on salary.

Model 2.3.2.3 (Two Categorical Independent Variables with Interaction): Finally, consider the model:

$$(2.11) \quad Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4$$

$$+ \beta_5 * D_1 * D_4 + \beta_6 * D_2 * D_4 + \beta_7 * D_3 * D_4 + \epsilon.$$

where Y, D_1, D_2, D_3, and D_4 are as given in the previous model (Model 2.3.2.2).

Again, we are placing the graduates into $4 \times 2 = 8$ categories based on major and program rank (two ranks for each of four majors), and $E(Y)$ can have the following eight different values depending on major and program rank:

	Not Top 20	Top 20
Accounting	$E(Y) = \beta_0 + \beta_1$	$E(Y) = \beta_0 + \beta_1 + \beta_4 + \beta_5$
Finance	$E(Y) = \beta_0 + \beta_2$	$E(Y) = \beta_0 + \beta_2 + \beta_4 + \beta_6$
Marketing	$E(Y) = \beta_0 + \beta_3$	$E(Y) = \beta_0 + \beta_3 + \beta_4 + \beta_7$
Other	$E(Y) = \beta_0$	$E(Y) = \beta_0 + \beta_4$

This time, the difference in average salary between "top 20" and "not top 20" is allowed to differ from major to major, that is, the model allows interaction between major and rank.

Note that in all cases where we have only categorical independent variables, the population can be divided into a finite number of sub-populations, and the regression model is simply a device to allow the expected value of the dependent variable to be different for these sub-populations.

2.3.3 Models with Categorical and Continuous Independent Variables

We now present models that include both categorical and continuous independent variables. To make the discussion concrete, we again consider the population of business school graduates. Once again:

- Y is the salary of a graduate.

- D_1, D_2, and D_3 are the three dummy variables to represent the four categories of major: accounting, finance, marketing, and "other" ($D_1 = 1$ if accounting, and 0 if not; $D_2 = 1$ if finance, and 0 if not; $D_3 = 1$ if marketing, and 0 if not).

- X is the cumulative GPA of the graduate.

Model 2.3.3.1 (No Interaction Term): Consider first the model:

$$(2.12) \quad Y \;=\; \beta_0 \,+\, \beta_1 D_1 \,+\, \beta_2 D_2 \,+\, \beta_3 D_3 \,+\, \beta_4 X + \epsilon.$$

We can write the regression equation separately for the four majors as follows:

(1) Accounting ($D_1 = 1$, $D_2 = 0$, $D_3 = 0$):	$E(Y\|D_1, D_2, D_3, X) \;=\; (\beta_0 \,+\, \beta_1) \,+\, \beta_4 X$
(2) Finance ($D_1 = 0$, $D_2 = 1$, $D_3 = 0$):	$E(Y\|D_1, D_2, D_3, X) \;=\; (\beta_0 \,+\, \beta_2) \,+\, \beta_4 X$
(3) Marketing ($D_1 = 0$, $D_2 = 0$, $D_3 = 1$):	$E(Y\|D_1, D_2, D_3, X) \;=\; (\beta_0 \,+\, \beta_3) \,+\, \beta_4 X$
(4) Other ($D_1 = 0$, $D_2 = 0$, $D_3 = 0$):	$E(Y\|D_1, D_2, D_3, X) \;=\; \beta_0 \,+\, \beta_4 X$

Note that here, the intercept of the relation between $E(Y|D_1, D_2, D_3, X)$ and X can vary from major to major. Thus, the model in equation (2.12) is equivalent to four separate regression lines of $E(Y|X)$ against X for the four majors. For example, for an accounting major, the intercept is $(\beta_0 + \beta_1)$, and for an "other" major it is β_0. However, the slope of X is the same, β_4, for all four majors. Thus, we say that in this case, **there is no interaction** between the effects of major and GPA on starting salary. Figure 2.4 graphically shows the relationship between $E(Y|X)$ and X for the four majors.

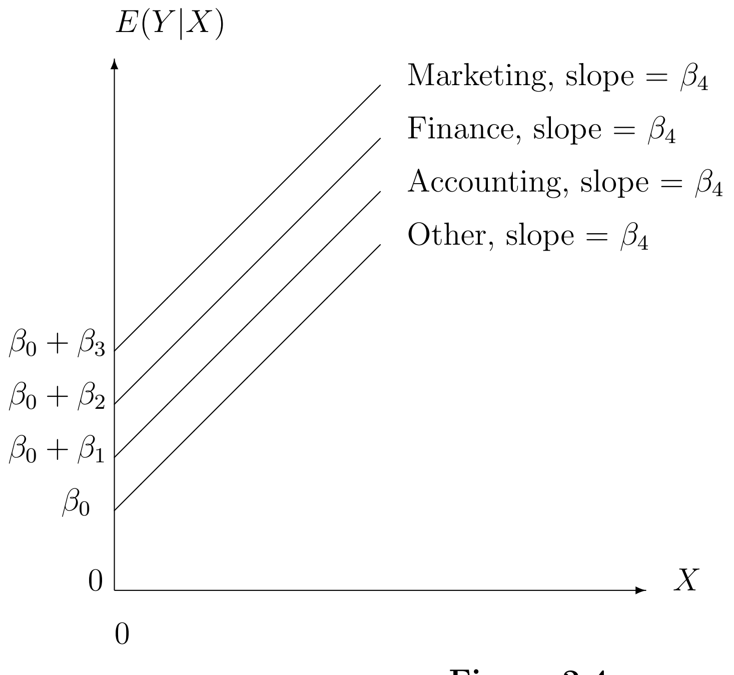


Figure 2.4

Model 2.3.3.2 (Includes Interaction Term): Consider now the model:

$$(2.13) \quad Y \;=\; \beta_0 \,+\, \beta_1 D_1 \,+\, \beta_2 D_2 \,+\, \beta_3 D_3 \,+\, \beta_4 X$$

$$+ \beta_5 * D_1 * X \; + \; \beta_6 * D_2 * X \; + \; \beta_7 * D_3 * X \; + \; \epsilon.$$

We can write the regression equation separately for the four majors as follows:

(1) Accounting ($D_1 = 1$, $D_2 = 0$, $D_3 = 0$):

$E(Y|D_1, D_2, D_3, X) \; = \; (\beta_0 \; + \; \beta_1) \; + \; (\beta_4 + \beta_5)X.$

(2) Finance ($D_1 = 0$, $D_2 = 1$, $D_3 = 0$):

$E(Y|D_1, D_2, D_3, X) \; = \; (\beta_0 \; + \; \beta_2) \; + \; (\beta_4 + \beta_6)X.$

(3) Marketing ($D_1 = 0$, $D_2 = 0$, $D_3 = 1$):

$E(Y|D_1, D_2, D_3, X) \; = \; (\beta_0 \; + \; \beta_3) \; + \; (\beta_4 + \beta_7)X.$

(4) Other ($D_1 = 0$, $D_2 = 0$, $D_3 = 0$):

$E(Y|D_1, D_2, D_3, X) \; = \; \beta_0 \; + \; \beta_4 X.$

This time, the intercept and the slope of the relation between $E(Y|D_1, D_2, D_3, X)$ and X can vary from major to major. Thus, the model in equation (2.13) is equivalent to four separate regression lines of $E(Y|X)$ against X for the four majors, where the intercepts and slopes can differ from line to line. For example, for an accounting major, the intercept is $(\beta_0 + \beta_1)$ and the slope is $(\beta_4 + \beta_5)$. For an "other" major, the intercept is β_0, and the slope is β_4. Thus, β_1 is the difference in intercepts, and β_5 is the difference in slopes, between an accounting major and an "other" major. Since the slope of GPA (X) can vary with major, this model allows interaction between major and GPA.

Equation (2.13) includes models (2.3.3.1), (2.3.2.1), and the two variable regression model as special cases.

If $\beta_5 = \beta_6 = \beta_7 = 0$, then equation (2.13) reduces to equation (2.12), where the slope is equal for all four majors.

If $\beta_5 = \beta_6 = \beta_7 = 0$, and $\beta_4 = 0$, then equation (2.13) reduces to equation (2.9), where the four regression lines are all parallel to the X axis. Here, $E(Y)$ can vary from major to major but does not depend on X.

If $\beta_5 = \beta_6 = \beta_7 = 0$, and $\beta_1 = \beta_2 = \beta_3 = 0$, we get the two variable regression model.

Clearly, it is possible to develop even more complex models with multiple categorical and interval scaled variables, and more complex interaction terms. However, most problems encountered in practice can be addressed following the examples discussed here.

2.4 Estimation of Regression Coefficients

Consider the regression model given in equation (2.1):

$$Y \; = \; \beta_0 \; + \; \beta_1 X_1 \; + \; \beta_2 X_2 \; + \; \ldots \; + \; \beta_n X_n \; + \; \epsilon.$$

The β's are parameters that remain the same for all observations, while X's and ϵ may vary from observation to observation. The objective of regression analysis is to estimate the $(n + 1)$ parameters β_0, β_1, …, β_n from a data set of n observations where we know Y, X_1, …, X_m for each observation. The sample size m should exceed $(n+1)$ by at least 30 for robust estimation of regression coefficients.

The process of estimation is logically equivalent to trying out different sets of parameter values. For example, suppose we try out a set of numbers b_0, b_1, …, b_n. (The b's are used to distinguish the trial values from true parameter values β's.) For a given observation, we compute a **predicted** Y, denoted by $\hat{Y}$, as follows:

$$\hat{Y} \quad = \quad b_0 + b_1 X_1 + \ldots + b_n X_n.$$

If this trial set of parameter values fits the data well, then the difference between the observed and predicted values of Y should be small. This difference is quantified by the **sum of squared errors**, or, SSE, defined as follows:

$$(2.14) \quad \text{SSE} \quad = \quad \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2,$$

where m is the number of observations, and Y_i and $\hat{Y}_i$ are observed and predicted values of Y for the i-th observation in the sample.

The regression estimation process finds the set of parameter values that minimizes SSE, and computes standard errors of these parameter estimates. It also computes R^2, which is defined as:

$$(2.15) \quad R^2 \quad = \quad 1 \; - \; \frac{\text{SSE}}{\text{SST}},$$

where:

SSE (sum of squared errors) $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, and

SST (total sum of squares) $= \sum_{i=1}^{n} (Y_i - \overline{Y})^2$, where $\overline{Y}$ is the average Y for the sample.

Notes on R^2: The statistic R^2 tells us how well the regression model fits the data. Three important facts about R^2 are listed below:

(1) For the naive model that includes only β_0, the estimate of β_0 is $\overline{Y}$. Thus, from (2.15), R^2 for the naive model is zero. For any other regression model, R^2 tells us how much better the model fits the data compared to the naive model.

(2) R^2 can never exceed 1. It equals 1 if and only if we have perfect fit, that is, $Y_i = \hat{Y}_i$ for every observation i.

(3) For a given set of independent variables, also called **regressors**, the estimation procedure chooses coefficients to minimize SSE. From (2.15), that is equivalent to maximizing R^2 for that set of regressors.

If we add another regressor to the model, we can always choose the coefficient of that regressor to be zero, which leaves R^2 the same. Thus, if we add a regressor to a model, R^2 either remains the same or increases.

R^2 may increase due to chance even when the true coefficient of the new regressor is zero. In Section 1.6 we discuss how to use the F test to eliminate such **non-significant** regressors.

Typical Contents of a Regression Output: When a regression model is estimated using a package such as Minitab, the output usually provides the following information:

(1) Parameter estimates: b_0, b_1, b_2, …, b_n.

(2) Approximate standard deviations of the parameter estimates: s_{b_0}, s_{b_1}, …, s_{b_n}.

(3) For each parameter $\beta_0, \beta_1, \ldots, \beta_n$, the P value for the null hypothesis that the parameter is zero.

(4) $R^2 = 1 - \dfrac{\text{SSE}}{\text{SST}}$.

The t test and the F test are used to test hypotheses about the regression parameters using the information provided by the regression output.

2.5 t-tests of Individual Regression Parameters

The t test can be used to test a hypothesis about an individual regression coefficient, or a weighted sum of regression coefficients. We only discuss the special case of tests involving a single regression coefficient.

Suppose we have estimated the regression model (2.1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \epsilon.$$

The estimated regression coefficients, denoted by b_0, b_1, …, b_m, are unbiased, normally distributed estimates of β_0, β_1, …, β_m, respectively, and $\{(b_0 - \beta_0)/s_{b_0}\}$, $\{(b_1 - \beta_1)/s_{b_1}\}$, …, $\{(b_n - \beta_n)/s_{b_n}\}$ all follow the t distribution with degrees of freedom $= (m - n - 1)$.

t statistic and P value: If you do regression with Minitab, then for each regression coefficient β_i, the program automatically tests the null hypothesis $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$. Formally, H_0 is rejected at confidence level $(1 - \alpha)$ if b_i falls outside $0 \pm t_{\alpha/2} * s_{b_i}$. Rewriting, H_0 is rejected at confidence level $(1 - \alpha)$ if $\dfrac{b_i}{s_{b_i}}$ falls outside $0 \pm t_{\alpha/2}$.

The number $\dfrac{b_i}{s_{b_i}}$ is called the t-statistic, and is simply called T in the Minitab output. The number P is the P value for the hypothesis test. For example if, for regression coefficient β_i, $P < .05$, you can reject H_0 at a 95% level of confidence.

In common practice, a coefficient is only considered significant if its P value is less than 0.1.

Other applications: We can construct a confidence interval for, or test a hypothesis on, any regression coefficient β_i as follows:

1. A $(1 - \alpha)$ confidence interval for β_i is given by $b_i \pm t_{\alpha/2} s_{b_i}$, where the degree of freedom of t is $(m - n - 1)$.

2. To test H_0: $\beta_i \leq K$ against H_a: $\beta_i > K$, we use the following decision rule:

At a confidence level of $(1 - \alpha)$*, reject* H_0 *if* $b_i > K + t_\alpha * s_{b_i}$*, where the degree of freedom of* t *is* $(m - n - 1)$.

This decision rule can also be stated as follows: *At a confidence level of* $(1 - \alpha)$*, reject* H_0 *if* $\dfrac{b_i - K}{s_{b_i}} > t_\alpha$*, where the degree of freedom of* t *is* $(n - m - 1)$.

3. To test H_0: $\beta_i = K$ against H_a: $\beta_i \neq K$, we use the decision rule:

At a confidence level of $(1-\alpha)$*, reject* H_0 *if* b_i *falls outside* $K \pm t_{\alpha/2} * s_{b_i}$*, where the degree of freedom of* t *is* $(m-n-1)$.

This decision rule can also be stated as follows: *At a* $(1-\alpha)$ *level of confidence, reject* H_0 *if* $\dfrac{b_i - K}{s_{b_i}}$ *falls outside* $0 \pm t_{\alpha/2}$*, where* t *has* $(m-n-1)$ *degrees of freedom.*

Example 2.1: *Consider the model*

$$y \quad = \quad \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \epsilon,$$

where:

Y *= starting annual salary of a business school graduate (unit=$1000);*

D_1 *= 1 if the graduate had an accounting major, and 0 if not;*

D_2 *= 1 if the graduate had a finance major, and 0 if not;*

D_3 *= 1 if the graduate had a marketing major, and 0 if not.*

We have estimated the regression model using 120 observations, and obtained the following estimates:

$b_0 = 35$, $s_{b_0} = 10$; $b_1 = 20$, $s_{b_1} = 5$; $b_2 = 15$, $s_{b_2} = 5$; $b_3 = 10$, $s_{b_3} = 4$.

Question 1. Construct a 95% confidence interval for the average salary of a graduate with an "other" major.

Answer: *Here, we have to construct a 95% confidence interval for* β_0*. The degree of freedom of the* t *variate is* $(120-3-1) = 116$*. Therefore, we can approximately use the* Z *variate instead of the* t *variate. Thus, the required 95% confidence interval is given by:*

$$35 \ \pm \ (1.96 \times 10), \quad \textit{that is,} \quad 35 \pm 19.6.$$

Question 2. Construct a 95% confidence interval for the difference in average salary between an accounting major, and an "other" major.

Answer: *Here, we have to construct a 95% confidence interval for* β_1*. Once again, the degree of freedom of* t *is 116, which exceeds 30. Thus, we can approximate the* t *variate by the* Z *variate. Hence, the required confidence interval is approximately given by:*

$$20 \ \pm \ (1.96 \times 5), \quad \textit{that is,} \quad 20 \ \pm \ 9.8.$$

Question 3. At 95% confidence, test the null hypothesis that the average salary of a finance major does not exceed the average salary of an "other" major by more than $10,000 a year.

Answer: *Here, we are testing* H_0*:* $\beta_2 \leq 10$*, against* H_a*:* $\beta_2 > 10$.

At a 95% level of confidence, we reject H_0 *if* $b_2 > 10 + (1.645 \times 5) \ = \ 18.825$.

Here, $b_2 \ = \ 15$ *does not exceed 18.225. Hence, we cannot reject* H_0 *at a 95% level of confidence.*

2.6 F Test

We only consider a simple application of the F-test to regression analysis, to perform a joint test of whether one or more β's are **all** equal to zero.

Consider the regression model in general form:

$$(2.16) \quad Y \;=\; \beta_0 \,+\, \beta_1 X_1 \,+\, \beta_2 X_2 \,+\, \ldots \,+\, \beta_n X_n \,+\, \epsilon.$$

We will test the null hypothesis that any k of the β's, which may or may not include β_0, are all equal to zero. For simplicity, we will test:

$$H_0 : \beta_1 \;=\; \beta_2 \;=\; \ldots \;=\; \beta_k \;=\; 0, \quad k \leq n,$$

against H_0 : at least one of β_1, ..., β_k is not zero. The procedure and results apply to any k regression coefficients.

We proceed as follows:

(1) Run the **full regression** using Y as the dependent variable, and X_1, X_2, ..., X_n as independent variables. Let R_f^2 represent the corresponding R^2.

(2) Run the **restricted regression** using Y as the dependent variable, and X_{k+1}, ..., X_n as independent variables. Let R_r^2 represent the corresponding R^2.

More generally, the independent variables in the restricted model are all the independent variables from the full model that do not have coefficients equal to zero under H_0. Stated differently, we get the restricted regression model by starting with the full model given by equation (2.16), and then setting some of the β's equal to zero according to H_0.

3. Compute the f-statistic, given by:

$$(2.17) \quad f \quad = \quad \frac{(R_f^2 - R_r^2)/k}{(1 - R_f^2)/(m-n-1)} \;=\; \frac{R_f^2 - R_r^2}{1 - R_f^2} \times \frac{m-n-1}{k}.$$

The restricted model uses a subset of the regressors of the full model. Hence, as discussed in Section 2.4, R_r^2 can never exceed R_f^2, that is, we always have $f \geq 0$. Also, if H_0 is true, then the difference between R_f^2 and R_r^2 should not be large, that is, the f statistic should be relatively small. Thus, a large value of the f statistic contradicts H_0.

If H_0 is true, then f follows an F-distribution with degrees of freedom $(k, n-m-1)$. We use the following decision rule:

At a confidence level of $(1-\alpha)$, reject H_0 if $f > F_\alpha(k, m-n-1)$.

A Special Case of the F-test: Suppose we wish to test $H_0 : \beta_1 = \ldots = \beta_m = 0$ against H_a : at least one of β_1, ..., β_m is not zero.

In this case, $k = n$, and the restricted model is the naive model given by equation (2.4): $Y = \beta_0 + \epsilon$. Hence, $R_r^2 = 0$, and the f statistic is given by:

$$f \quad = \quad \frac{R^2}{1 - R^2} \times \frac{m-n-1}{n},$$

where R^2 comes from the full model. This f statistic is often included in the regression output of standard packages under "F ratio."

If $f > F_\alpha(n, m-n-1)$, then we reject H_0 at a $(1-\alpha)$ level of confidence, and conclude that at least one of the independent variables has an effect on Y.

If we cannot reject H_0, it follows that none of the independent variables has a significant effect on Y, and it is not meaningful to estimate this regression model.

Example 2.2: *Consider the model:*

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X + \epsilon,$$

where:

- *Y is the salary of a business school graduate;*
- *$D_1 = 1$ if the graduate had an accounting major, and 0 if not;*
- *$D_2 = 1$ if the graduate had a finance major, and 0 if not;*
- *$D_3 = 1$ if the graduate had a marketing major, and 0 if not;*
- *X = cumulative GPA.*

We wish to test, at a confidence level of 95%, the null hypothesis that for a given value of cumulative GPA, the average salary is equal for all four majors, that is,

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0,$$

against the alternative hypothesis that at least one major has a different average salary than others.

Suppose we have a data set of 125 observations, and we have estimated the full and restricted regression models:
Full model: Y with D_1, D_2, D_3, and X. $R_f^2 = .6$.
Restricted model: Y with X. $R_r^2 = .4$.

Here, $k = 3$, and $m - n - 1 = 125 - 4 - 1 = 120$. Thus, we reject H_0 at a 95% level of confidence if $f > F_{.05}(3, 120) = 2.68$.
From the data, $f = \dfrac{(.6-.4)/3}{(1-.6)/120} = 20$. Since $20 > 2.68$, we reject H_0 at a 95% level of confidence.

Note: The F-test is quite powerful. However, it can only be used to test equality hypotheses. For example, we cannot test whether GPA being the same, on the average the salary of an accounting graduate exceeds the average salary of an "other" graduate by more than \$5000. The t-test can be used to test one sided hypotheses like this.

2.7 Exercise Problems

Problem Scenario: Consider the regression model:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2$$

$$+ \beta_5 D_1 X_1 + \beta_6 D_2 X_1 + \beta_7 D_1 X_2 + \beta_8 D_2 X_2 + \epsilon,$$

where:

Y = sales of a brand in a sales territory (unit = \$100,000) during Fall, 2003;

X_1 is the number of salespeople in the territory;

X_2 is the retail price (in dollars) in the territory;

D_1 and D_2 are dummy variables for the level of advertising in the territory. The advertising level can be low, medium, or high.

$D_1 = 1$ if the advertising level is medium, and $D_1 = 0$ otherwise;

$D_2 = 1$ if the advertising level is high, and $D_2 = 0$ otherwise.

ϵ is defined as usual.

1. Suppose we know that the true values of the regression parameters are as follows:

$\beta_0 = 5.0$, $\beta_1 = 1.0$, $\beta_2 = 1.5$, $\beta_3 = .2$, $\beta_4 = -.5$,

$\beta_5 = .1$, $\beta_6 = .15$, $\beta_7 = 0$, $\beta_8 = -.1$, $\sigma_\epsilon = 2$.

Compute the probability that the sales (unit = \$100,000) in a territory would exceed \$1100,000 if the advertising level is high, there are 20 salespeople, and the price is \$5.

2. State each of the following null hypotheses in terms of the parameters of the regression model (e.g., $H_0 : \beta_1 = 0$):

2.(a) The marginal effect of price on sales is the same for medium and high advertising levels.

2.(b) Changes in price do not affect sales if the level of advertising is low.

3. Suppose we have estimated regression models using data from 49 territories, and got the following results:

Results

Regression	Dependent Variable	Independent Variables	R^2
1	Y	$D_1, D_2, X_1, X_2, D_1X_1,$ D_2X_1, D_1X_2, D_2X_2	.75
2	Y	D_1, D_2	.2
3	Y	X_1, X_2	.2
4	Y	$D_1, D_2, X_1, X_2, D_1X_1, D_2X_1$	.6
5	Y	$D_1, D_2, X_1, D_1X_1, D_2X_1$	.55
6	Y	$D_1, D_2, X_2, D_1X_2, D_2X_2$	.5
7	Y	$D_1, D_2, D_1X_1, D_2X_1,$ D_1X_2, D_2X_2	.3
8	Y	$D_1, D_2, X_1, X_2, D_1X_2, D_2X_2$	.7

At a 99% level of confidence, test each of the following two null hypotheses:

3(a) The marginal effect of price on sales is the same for all levels of advertising.

3(b) Having an additional salesperson does not have any effect on sales at any level of advertising.

In each case, clearly state (i) H_0 in terms of the regression parameters, (ii) the decision rule, and (iii) the conclusion regarding H_0.

3 Conjoint Analysis

3.1 Idea

In conjoint analysis, we assume that the evaluation of a product alternative depends on the attribute profile of the alternative, $(X_1, \ldots, X_n)$. The evaluation score, or utility, is denoted by $U(X_1, \ldots, X_n)$. We discuss the simplest form of this utility function, called the additive part-worth utility model:

$$(3.1)\; U(X_1, X_2, \ldots, X_n) \quad = \quad U_0 + U_1(X_1) + U_2(X_2) + \ldots + U_n(X_n),$$

where U_0 is a constant term, and U_1, U_2, etc., are functions of the different attributes. This model assumes that the marginal effect of a change in an attribute (say X_1) on utility does not depend on the levels of the other attributes (i.e., X_2, X_3, etc.). Still, the conjoint model is more general than the expectancy value model.

Conjoint analysis examines products that can be expressed in terms of attributes that are easy to describe to respondents. In general, the attributes may be qualitative or quantitative. For simplicity, we only discuss quantitative attributes. The analysis is done at the level of an individual respondent who is asked to rate several hypothetical products on a given scale. From the evaluation scores, the researcher can determine how the respondent's evaluation of the product depends on attribute levels. This information is used to guide new product development and to assess the potential market shares of new product concepts.

Meaning of Part-Worths: The functions $U_1(X_1)$, $U_2(X_2)$, etc., are called the part-worth scores of the different attributes. We consider a fixed range of each attribute, and use a simple convention: At the lowest point of the range of an attribute, the part-worth score of the attribute is zero. With this convention:

- U_0 is the evaluation score of the product for which each attribute is at the lowest point of its range.
- $U_i(X_i)$ is zero if X_i is at the lowest level of its range. Otherwise, $U_i(X_i)$ is the marginal change in evaluation score if attribute X_i is higher than its lowest level.

Example: Consider a product with 2 attributes: a laptop computer with attributes weight (X_1) and hard drive size (X_2). X_1 can rage from 4 lb to 10 lb, and X_2 can range from 80 GB to 160 GB. The laptop is evaluated on a 0-100 (very poor to excellent) scale.

In this case, the additive part-worth model can be expressed as follows:

$$(3.2) \quad U(X_1, X_2) = U_0 + U_1(X_1) + U_2(X_2),$$

where $U(X_1, X_2)$ is the evaluation score of a laptop with attribute levels (X_1, X_2). By our convention, we assume that:

- $U_1(4) = 0$
- $U_2(80) = 0$

Under these assumptions:

$$U(4, 80) \quad = \quad U_0 + U_1(4) + U_2(80) \quad = \quad U_0 + 0 + 0 \;=\; U_0$$

Therefore:

- U_0 is the evaluation score of the "basic" laptop computer with each attribute at the lowest levels in the ranges considered, that is, $X_1 = 4$ and $X_2 = 80$.
- $U_1(X_1)$ is the marginal adjustment to evaluation score depending on the level of weight (X_1). If $X_1 = 4$, then no adjustment is needed.
- $U_2(X_2)$ is the marginal adjustment to evaluation score depending on the level of hard drive size (X_2). If $X_2 = 80$, then no adjustment is needed.

Estimation: To estimate the conjoint model, the researcher creates hypothetical products, and asks the respondent to rate each hypothetical product on a given scale (e.g., a 0-100 rating scale). To develop the hypothetical products, a fixed number of levels of each attribute is used. For a given attribute, the lowest and the highest point of its range are always used. Once the respondent rates the hypothetical products, the ratings are used to estimate the following:

(1) U_0, the evaluation score of the product with the lowest level of each attribute.

(2) For each attribute, the value of its part-worth function at each level used in the analysis.

Using these information, the evaluation score of any new product concept can be estimated by interpolation. Chapter 17 shows how a conjoint model is estimated using dummy variable regression. We now discuss how to use conjoint analysis results after estimation is done.

3.2 Laptop Example

We now discuss the laptop example in more detail to show how conjoint analysis is used. In the present case, the objective of conjoint analysis is to estimate:

(1) The evaluation score of the "basic" product, U_0

(2) $U_1(X_1)$ at two or more levels in the range $4 - 10$. The extreme levels 4 and 10 are always included.

(3) $U_2(X_2)$ at two or more levels in the range $80 - 160$. The extreme levels 80 and 160 are always included.

Suppose we selected the following levels of the two attributes:

(1) Four levels of X_1: 4, 6, 8, and 10.

(2) Four levels of X_2: 80, 100, 120, and 160.

Several hypothetical laptop computers are created using different combinations of these attribute levels, and the respondent provides evaluation scores for these hypothetical products. For a given respondent, we obtain the following results:

- U_0 = mean evaluation score of the "basic laptop" with $X_1 = 4$ and $X_2 = 80$.
- $U_1(4) = 0$ (by assumption)
- $U_1(6)$
- $U_1(8)$
- $U_1(10)$
- $U_2(80) = 0$ (by assumption)
- $U_2(100)$
- $U_2(120)$
- $U_2(160)$

3.3 Evaluation of a New Product Concept using Linear Interpolation

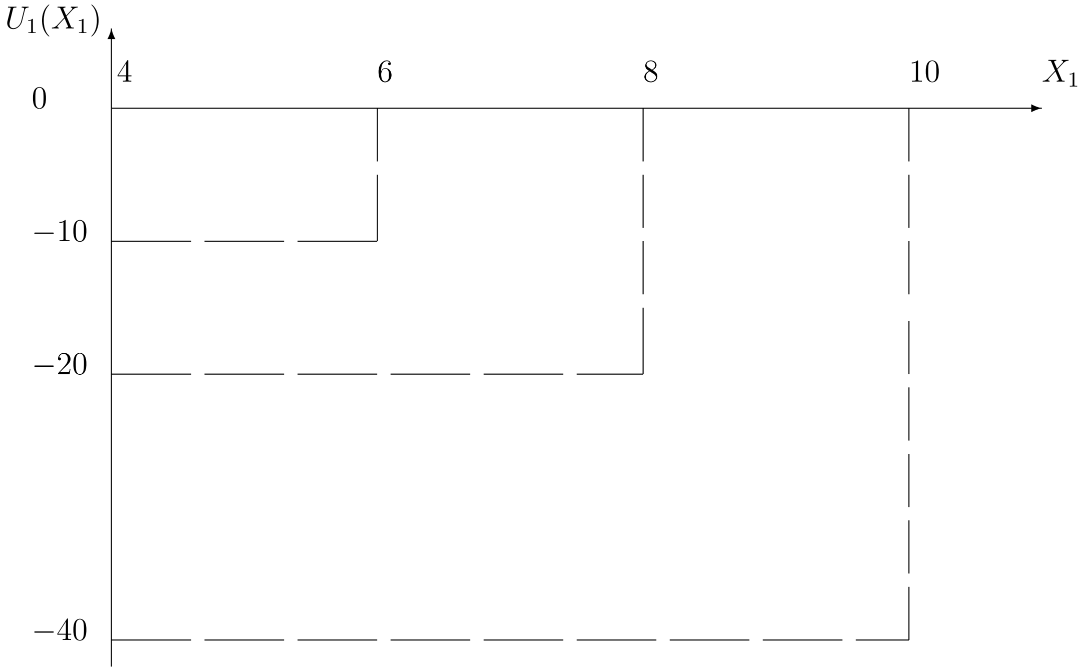
New product concepts need not have attributes exactly at the levels used in conjoint estimation. For such intermediate levels, we use linear approximation to estimate part-worth utilities. We illustrate this with an example.

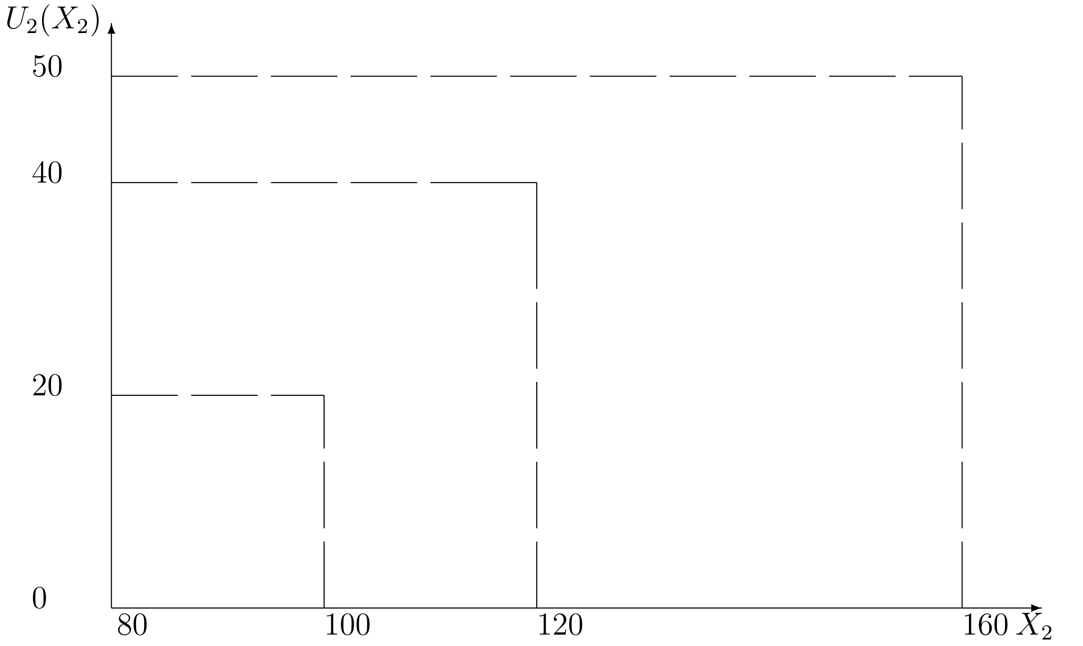
Example: Suppose, for a given individual, we have the following estimates:

- $U_0 = 60$
- $U_1(4) = 0$
- $U_1(6) = -10$
- $U_1(8) = -20$
- $U_1(10) = -40$
- $U_2(80) = 0$
- $U_2(100) = 20$

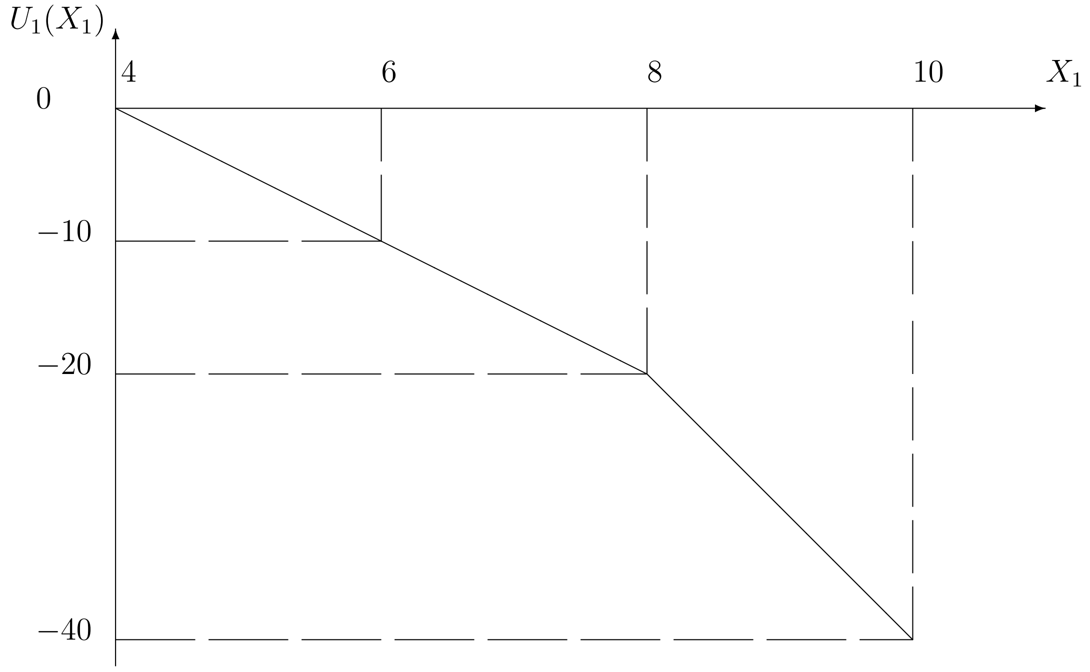
- $U_2(120) = 40$
- $U_2(160) = 50$

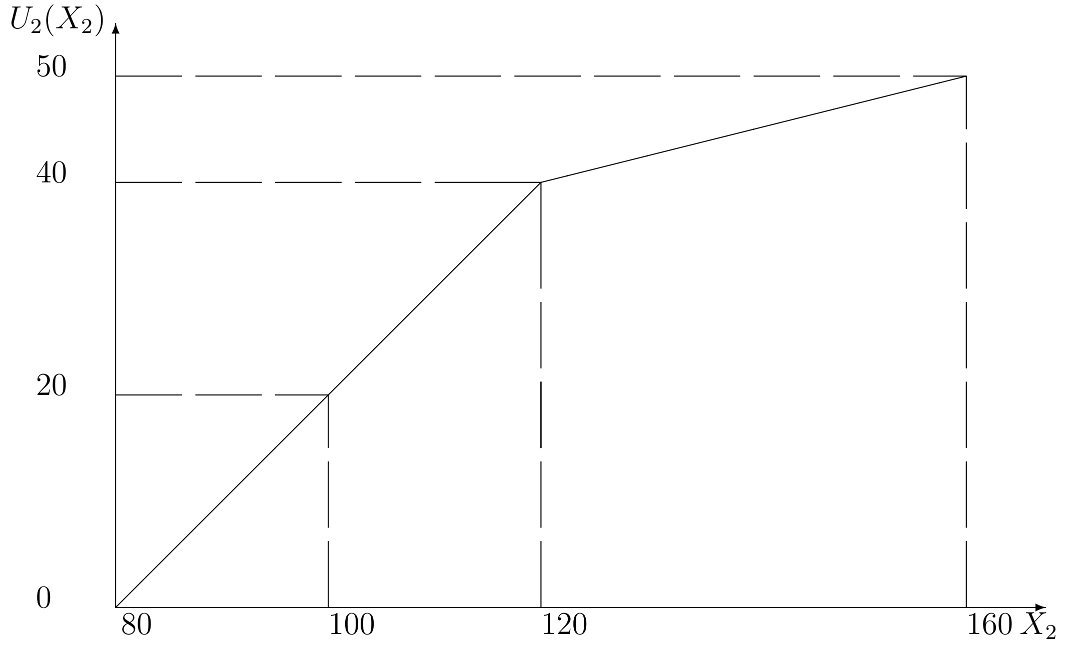
We can express the part-worth functions graphically as follows:





Note that we know $U_1(X_1)$ only at $X_1 = 4$, 6, 8, and 10, and $U_2(X_2)$ only at $X_2 = 80$, 100, 120, and 160. To approximate the curves, we join successive points by straight lines. This is called linear interpolation. We now get the following graphs:





With these graphs, we can estimate the evaluation score of any laptop with X_1 in the range 4-10, and X_2 in the range 80-160. We should not use these results to evaluate product concepts with attribute levels outside the ranges used to estimate the model.

Consider two new product concepts:

Concept 1: $X_1 = 5.5$, $X_2 = 90$

Concept 2: $X_1 = 9$, $X_2 = 150$

We can evaluate the two concepts as follows:

	Concept 1 ($X_1 = 5.5$, $X_2 = 90$)	Concept 2 ($X_1 = 9$, $X_2 = 150$)
U_0	60	60
$U_1(X_1)$	$U_1(4) + +\left\{\frac{5.5-4}{6-4}\right\} \times \left\{U_1(6) - U_1(4)\right\}$ $= 0 + \left\{\frac{1.5}{2}\right\}\left\{(-10) - 0\right\} = -7.5$	$U_1(8) + \left\{\frac{9-8}{10-8}\right\} \times \left\{U_1(10) - U_1(8)\right\}$ $= -20 + \frac{1}{2} \times \{(-40) - (-20)\} = -30$
$U_2(X_2)$	$U_2(80)$ $+\left\{\frac{90-80}{100-80}\right\} \times \{U_2(100) - U_2(80)\}$ $= 0 + \frac{10}{20} \times (20 - 0) = 10$	$U_2(120)$ $+\left\{\frac{150-120}{160-120}\right\} \times \{U_2(160) - U_2(120)\}$ $= 40 + \frac{30}{40} \times (50 - 40) = 47.5$
$U(X_1, X_2)$	$60 - 7.5 + 10 = 62.5$	$60 - 30 + 47.5 = 77.5$

Therefore, this person will prefer concept 2 to concept 1.

Note: You can compute the utilities graphically also:

- $X_1 = 5.5$ is 1.5 above 4, that is, it is 3/4 th of the way from 4 to 6. Thus, $U_1(5.5)$ is three-fourths of the way from 0 to -10, that is, -7.5.

- $X_1 = 9$ is half-way between 8 and 10. Thus, $U_1(9)$ is half-way between -20 and -40, that is, -30.

- $X_2 = 90$ is half-way between 80 and 100. Thus, $U_2(90)$ is half-way between 0 and 20, that is, 10.

- $X_2 = 150$ is three-fourths of the way from 120 to 160. Therefore, $U_2(150)$ is three-fourths of the way from 40 to 50, that is, 47.5.

When you have several new product concepts, you can proceed this way to identify the concept this person would prefer.

3.4 Prediction of Market-Share

In practice, we have parameter estimates for a representative sample of respondents. For each member of the sample, we can predict if a given new product concept will be preferred to existing products in the market. By aggregating over the sample, we can predict the market share of the new product concept.

3.5 Example 1: Credit Card

In Fall 2001, I conducted conjoint analysis for an MBA Marketing Management class (MAR 600) at Syracuse University. The conjoint survey, provided in the Appendix to this section, included two conjoint questionnaires, one on credit cards, and the other on personal computers. While the attribute levels for computers are out of date, those for credit cards are not. I discuss the credit card survey and results here, and the personal computer survey and results in the next section.

3.5.1 Model and analysis: Hypothetical credit cards were designed using 4 levels each of three attributes:

(1) X_1 (interest rate): 6, 9, 12, or 18 percent APR.

(2) X_2: (credit limit): 5, 10, 25, or 50 (unit=\$1000).

(3) X_3 (annual fee): 0, 10, 20, or 50.

We assumed that the utility for a product (X_1, X_2, X_3) can be expressed as:

$$U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3),$$

where:

- U_0 = rating of credit card with $X_1 = 6$, $X_2 = 5$, and $X_3 = 0$

- $U_1(6) = 0$

- $U_2(5) = 0$

- $U_3(0) = 0$

$U_1(X_1)$, $U_2(X_2)$ and $U_3(X_3)$ are marginal contributions of the three attributes to utility as their levels increased from the lowest points. The respondents ranked 18 hypothetical credit cards, given in the table below, from best (1) to worst (18).

Card #	Interest Rate	Credit Limit	Annual Fee	Rank
1	6%	$5000	$0	
2	12%	$25,000	$10	
3	6%	$50,000	$0	
4	12%	$5000	$50	
5	18%	$50,000	$0	
6	6%	$25,000	$20	
7	18%	$5000	$10	
8	12%	$50,000	$20	
9	18%	$5000	$50	
10	9%	$10,000	$50	
11	18%	$10,000	$20	
12	9%	$50,000	$10	
13	12%	$10,000	$0	
14	18%	$25,000	$50	
15	6%	$10,000	$10	
16	9%	$5000	$20	
17	6%	$50,000	$50	
18	9%	$25,000	$0	

Credit cards 3 (best on all attributes) and 9 (worst on all attributes) served as anchors. Every respondent should have given card 3 a rank of 1, and card 9 a rank of 18.

3.5.2 Data and Estimation: Data from 66 respondents (MBA students) were analyzed. For each respondent, data for cards 3 and 9, inserted as anchors, were dropped. The other 16 cards were re-ranked from 1 (best) to 16 (worst). Then, this rank was subtracted from 17 to get a score on a 1-16 scale (1: lowest, 16: highest). Denoting the score on the 1-16 scale by S, a final score was computed as:

$$SCORE = (\frac{100}{15}) * (S - 1)$$

This gives us a score on a 0-100 scale. For example, if rank is 1, $S = 16$ and $SCORE = 100$. Similarly, if rank is 16, $SCORE = 0$. Using these scores, we obtained the following from each respondent:

- U_0 = the rating of the basic credit card with $X_1 = 6$, $X_2 = 5$, and $X_3 = 0$ on a 0-100 scale. (Denoted by $U0$)
- $U_1(6) = 0$ (Denoted by $U11$)
- $U_1(9)$: The marginal change in utility if X_1 is 9 instead of 6. (Denoted by $U12$)
- $U_1(12)$: The marginal change in utility if X_1 is 12 instead of 6. (Denoted by $U13$)
- $U_1(18)$: The marginal change in utility if X_1 is 18 instead of 6. (Denoted by $U14$)
- $U_2(5) = 0$ (Denoted by $U21$)
- $U_2(10)$: The marginal change in utility if X_2 is 10 instead of 5. (Denoted by $U22$)

- $U_2(25)$: The marginal change in utility if X_2 is 25 instead of 5. (Denoted by $U23$)
- $U_2(50)$: The marginal change in utility if X_2 is 50 instead of 5. (Denoted by $U24$)
- $U_3(0) = 0$ (Denoted by $U31$)
- $U_3(10)$: The marginal change in utility if X_3 is 10 instead of 0. (Denoted by $U32$)
- $U_3(20)$: The marginal change in utility if X_3 is 20 instead of 0. (Denoted by $U33$)
- $U_3(50)$: The marginal change in utility if X_3 is 50 instead of 0. (Denoted by $U34$)

3.5.3 Evaluating a new product idea: For any given respondent, the utility of a product concept with attribute values within the ranges used in estimation can be estimated as follows: First get the components:
(1) U_0
(2) $U_1(X_1)$: Compute it as follows:
$6 \leq X_1 < 9$: $U_1(X_1) = U12 * (\frac{X_1 - 6}{3})$.
$9 \leq X_1 < 12$: $U_1(X_1) = U12 + (U13 - U12) * (\frac{X_1 - 9}{3})$.
$12 \leq X_1 \leq 18$: $U_1(X_1) = U13 + (U14 - U13) * (\frac{X_1 - 12}{6})$.
(3) $U_2(X_2)$: Express X_2 in units of \$1000, e.g., use 15 if credit limit is \$15,000. Compute $U_2(X_2)$ as follows:
$5 \leq X_2 < 10$: $U_2(X_2) = U22 * (\frac{X_2 - 5}{5})$.
$10 \leq X_2 < 25$: $U_2(X_2) = U22 + (U23 - U22) * (\frac{X_2 - 10}{15})$.
$25 \leq X_2 \leq 50$: $U_2(X_2) = U23 + (U24 - U23) * (\frac{X_2 - 25}{25})$.
(4) $U_3(X_3)$: Compute this as follows:
$0 \leq X_3 < 10$: $U_3(X_3) = U32 * (\frac{X_3}{10})$.
$10 \leq X_3 < 20$: $U_3(X_3) = U32 + (U33 - U32) * (\frac{X_3 - 10}{10})$.
$20 \leq X_3 \leq 50$: $U_3(X_3) = U33 + (U34 - U33) * (\frac{X_3 - 20}{30})$.
Then add: $U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$.

When you compare multiple new product concepts, a given person will select the option that yields the highest utility.

3.5.4 Benefit Segments using Cluster Analysis: I performed cluster analysis using Minitab. Each of the 66 cases in the data set is expressed in terms of 10 coordinates: $U0$, $U12$, $U13$, $U14$, $U22$, $U23$, $U24$, $U32$, $U33$, and $U34$. The program computes the distances between pairs of cases in this ten-dimensional space and tries to create clusters based on proximity. The user specifies the number of clusters to be used. From the data, we obtained a three cluster solution.

Cluster 1 (8 out of 66 respondents)

$U_0 = 39.1667$			
$U_{11} = 0$	$U_{12} = -9.7917$	$U_{13} = -16.4583$	$U_{14} = -28.75$
$U_{21} = 0$	$U_{22} = 26.25$	$U_{23} = 44.7917$	$U_{24} = 58.9583$
$U_{31} = 0$	$U_{32} = -1.875$	$U_{33} = -8.75$	$U_{34} = -21.0417$

Cluster 2 (24 out of 66 respondents)

$U_0 = 92.5$			
$U_{11} = 0$	$U_{12} = -5.9722$	$U_{13} = -11.9444$	$U_{14} = -19.5833$
$U_{21} = 0$	$U_{22} = 3.1944$	$U_{23} = 5.2083$	$U_{24} = 2.9861$
$U_{31} = 0$	$U_{32} = -25$	$U_{33} = -46.5972$	$U_{34} = -72.2917$

Cluster 3 (34 out of 66 respondents)

$U_0 = 95.5392$			
$U_{11} = 0$	$U_{12} = -20.8333$	$U_{13} = -42.6961$	$U_{14} = -70$
$U_{21} = 0$	$U_{22} = 0.098$	$U_{23} = -3.0392$	$U_{24} = -9.0196$
$U_{31} = 0$	$U_{32} = -7.451$	$U_{33} = -11.4216$	$U_{34} = -17.7941$

- Cluster 1 strongly desires a high credit limit. Possibly, these respondents are entrepreneurs who find credit availability important. As expected, cluster 1 dislikes higher interest rates or annual fees.

- Cluster 2 finds low annual fee very important. It also dislikes higher interest rates. Credit limit has almost no impact on this segment of respondents.

- Cluster 3 finds low interest rate very important. It also dislikes higher annual fees. While credit limit is not important, this segment seems to dislike higher credit limits. The reason is unclear. Given the small size, this may simply be random error. However, this may also imply sensitivity to possible identify theft.

3.5.5 Output: The output is available as an Excel file (creditf01sum2013class.xls) to the interested reader. For each of the 66 observations, you have the following data:

Case number (1 through 66)

Gender: 1 if male, 0 if female

Age: 1-5 scale

NCARD: Number of credit cards owned

CLUSTER: Cluster membership, 1 if cluster 1 and 2 if cluster 2.

$SCORE_1 - SCORE_{16}$: The scores of the sixteen credit cards used in the analysis on a 0-100 scale (higher score is better).

Conjoint Estimates: $U0$, $U11$, $U12$, $U13$, $U14$, $U21$, $U22$, $U23$, $U24$, $U31$, $U32$, $U33$, $U34$.

3.6 Example 2: Personal Computer

We used the data from the MBA class in Fall 2001 to do a second conjoint analysis on personal computers. The survey is included in the Appendix to this section.

3.6.1 Model and analysis: The hypothetical products were designed using 4 levels each of three attributes:

(1) X_1 (warranty length): 1, 2, 3, or 4 years.

(2) X_2: (RAM): 1, 2, 3, or 4 (unit= 128 MB).

(3) X_3 (hard drive size): 1, 2, 3, or 5 (unit = 20 GB).

We assumed that the utility for a product (X_1, X_2, X_3) can be expressed as:

$$U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3),$$

where:

- U_0 = utility for personal computer with $X_1 = 1$, $X_2 = 1$, and $X_3 = 1$
- $U_1(1) = 0$
- $U_2(1) = 0$
- $U_3(1) = 0$

$U_1(X_1)$, $U_2(X_2)$ and $U_3(X_3)$ are marginal contributions of the three attributes to utility as their levels increased from the lowest. Each respondent rated 18 hypothetical personal computers listed in the table below from 1 (best) to 18 (worst). Data from computers 1 and 6, inserted as anchors, were dropped. The remaining 16 products were re-ranked from 1 to 16. Each PC was then given a score of (17−rank), that is, the PC ranked 1 got a score of 16, the PC ranked 2 got a score of 15, so on.

Personal Computer #	**Warranty**	**Random Access Memory (RAM)** (in MB)	**Hard Drive Size** (in GB)	**Rank**
1	4 yr	512	100	
2	1 yr	128	60	
3	2 yr	256	20	
4	3 yr	384	60	
5	4 yr	512	20	
6	1 yr	128	20	
7	1 yr	256	100	
8	2 yr	384	40	
9	3 yr	512	100	
10	4 yr	128	40	
11	1 yr	384	20	
12	2 yr	512	60	
13	3 yr	128	20	
14	4 yr	256	60	
15	1 yr	512	40	
16	2 yr	128	100	
17	3 yr	256	40	
18	4 yr	384	100	

3.6.2 Data and Estimation: Data from 68 respondents (MBA students) were analyzed. For each respondent, data for computers 1 and 6, inserted as anchors, were dropped. The other 16 computers were re-ranked from 1 (best) to 16 (worst). Then, this rank was subtracted from 17 to get a score on a 1-16 scale (1: lowest, 16: highest). Denoting the score on the 1-16 scale by S, a final score was computed as:

$$SCORE = (\frac{100}{15}) * (S - 1)$$

This gives us a score on a 0-100 scale. For example, if rank is 1, $S = 16$ and $SCORE = 100$. Similarly, if rank is 16, $SCORE = 0$. Using these scores, we obtained the following from each respondent:

- U_0 = the rating of the basic PC with $X_1 = 1$, $X_2 = 1$, and $X_3 = 1$ on a 0-100 scale. (Denoted by $U0$)
- $U_1(1) = 0$ (Denoted by $U11$)
- $U_1(2)$: The marginal change in utility if X_1 is 2 instead of 1. (Denoted by $U12$)
- $U_1(3)$: The marginal change in utility if X_1 is 3 instead of 1. (Denoted by $U13$)
- $U_1(4)$: The marginal change in utility if X_1 is 4 instead of 1. (Denoted by $U14$)

- $U_2(1) = 0$ (Denoted by $U21$)
- $U_2(2)$: The marginal change in utility if X_2 is 2 instead of 1. (Denoted by $U22$)
- $U_2(3)$: The marginal change in utility if X_2 is 3 instead of 1. (Denoted by $U23$)
- $U_2(4)$: The marginal change in utility if X_2 is 4 instead of 1. (Denoted by $U24$)
- $U_3(1) = 0$ (Denoted by $U31$)
- $U_3(2)$: The marginal change in utility if X_3 is 2 instead of 1. (Denoted by $U32$)
- $U_3(3)$: The marginal change in utility if X_3 is 3 instead of 1. (Denoted by $U33$)
- $U_3(5)$: The marginal change in utility if X_3 is 5 instead of 1. (Denoted by $U35$)

3.6.3 Evaluating a new product idea: For a given respondent, the utility of a product concept with attribute values within the ranges used in estimation can be estimated as follows:

First obtain the components:

(1) U_0

(2) $U_1(X_1)$: Express X_1 in years. Compute $U_1(X_1)$ as follows:

$1 \leq X_1 < 2$: $U_1(X_1) = U12 * (X_1 - 1)$.

$2 \leq X_1 < 3$: $U_1(X_1) = U12 + (U13 - U12) * (X_1 - 2)$.

$3 \leq X_1 \leq 4$: $U_1(X_1) = U13 + (U14 - U13) * (X_1 - 3)$.

(3) $U_2(X_2)$: Express X_2 in units of 128 MB, e.g., use 1.5 if RAM is 192 MB. Compute $U_2(X_2)$ as follows:

$1 \leq X_2 < 2$: $U_2(X_2) = U22 * (X_2 - 1)$.

$2 \leq X_2 < 3$: $U_2(X_2) = U22 + (U23 - U22) * (X_2 - 2)$.

$3 \leq X_2 \leq 4$: $U_2(X_2) = U23 + (U24 - U23) * (X_2 - 3)$.

(4) $U_3(X_3)$: Express X_3 in units of 20 GB. For example, use 2.5 if hard drive size is 50 GB. Compute $U_3(X_3)$ as follows:

$1 \leq X_3 < 2$: $U_3(X_3) = U32 * (X_3 - 1)$.

$2 \leq X_3 < 3$: $U_3(X_3) = U32 + (U33 - U32) * (X_3 - 2)$.

$3 \leq X_3 \leq 5$: $U_3(X_3) = U33 + (U34 - U33) * (\frac{X_3 - 3}{2})$.

Then add: $U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$.

When you compare multiple new product concepts, a given person will select the option that yields the highest utility.

3.6.4 Benefit Segments using Cluster Analysis: I performed cluster analysis using Minitab. Each of the 68 cases in the data set is expressed in terms of 10 coordinates: the conjoint estimates

$U0$, $U12$, $U13$, $U14$, $U22$, $U23$, $U24$, $U32$, $U33$, and $U35$. The program computes the distances between pairs of cases in this ten-dimensional space and tries to create clusters based on proximity. The user specifies the number of clusters to be used. In this case (personal computers), three distinct clusters emerged. For each cluster, the averages of U_0, $U_1(1), \ldots, U_3(5)$ are given below.

Cluster 1 (22 respondents)			
$U_0 = -5.6818$			
$U_1(1) = 0$	$U_1(2) = 5.0000$	$U_1(3) = 10.8333$	$U_1(4) = 11.1364$
$U_2(1) = 0$	$U_2(2) = 27.7273$	$U_2(3) = 49.5455$	$U_2(4) = 70.0000$
$U_3(1) = 0$	$U_3(2) = 9.0152$	$U_3(3) = 14.9242$	$U_3(5) = 24.5455$

Cluster 2 (26 respondents)			
$U_0 = -6.0256$			
$U_1(1) = 0$	$U_1(2) = 2.8846$	$U_1(3) = 8.7179$	$U_1(4) = 8.6538$
$U_2(1) = 0$	$U_2(2) = 15.4487$	$U_2(3) = 20.8333$	$U_2(4) = 23.2051$
$U_3(1) = 0$	$U_3(2) = 23.9103$	$U_3(3) = 49.6795$	$U_3(5) = 70.7692$

Cluster 3 (20 respondents)			
$U_0 = -7.2500$			
$U_1(1) = 0$	$U_1(2) = 27.3333$	$U_1(3) = 54.7500$	$U_1(4) = 70.2500$
$U_2(1) = 0$	$U_2(2) = 9.4167$	$U_2(3) = 19.0833$	$U_2(4) = 22.1667$
$U_3(1) = 0$	$U_3(2) = 4.2500$	$U_3(3) = 10.3333$	$U_3(4) = 11.4167$

- From the results, we find that cluster 1 and cluster 2 do not find warranty length important. Both these clusters find the functional attributes of the computer more important with cluster 1 more interested in RAM, and cluster 2 more interested in hard drive size.

- In contrast, cluster 3 finds a long warranty very important.

3.6.5 Output: The output is available as an Excel file (pcf01sum2013class.xls) to the interested reader. For each of the 68 observations, you have the following data:

Case number (1 through 68)

Gender: 1 if male, 0 if female

Age: 1-5 scale

KNOW: self assessment of knowledge of PC (1-7 scale)

PPC: 1 if purchased PC in the last two years, 0 if not

OWNPC: 1 if owns PC, 0 if not

EMAIL, WORD, SURF, SPREAD, TAX, STORDAT, STAT, PROG, VIDEO, OTHER: each 1 if checked, 0 if not

CLUSTER: Cluster membership using three-cluster solution; 1 if cluster 1, 2 if cluster 2, 3 if cluster 3.

$SCORE_1 - SCORE_{16}$: The scores of the sixteen personal computers used in the analysis.

Estimates of: $U0$, $U11$, $U12$, $U13$, $U14$, $U21$, $U22$, $U23$, $U24$, $U31$, $U32$, $U33$, and $U35$.

3.6.6 Example of New Product Evaluation

Consider five new product concepts:

Concept 1: 2.5, 2.5 (320 MB), 3 (60 GB)

Concept 2: 1, 4 (512 MB), 5 (100 GB)

Concept 3: 2, 3 (384 MB), 5 (100 GB)

Concept 4: 2, 4 (512 MB), 3 (60 GB)

Concept 5: 4, 2 (256 MB), 2 (40 GB)

Thus, concept 1 provides middle of the road levels of all attributes. Concepts 3 and 4 provide high values of hard drive size and RAM, respectively. Concept 2 offers high values of functional attributes (RAM, HD) but a low warranty length. Concept 5 offers high warranty length but low levels of functionality. For each of the 68 respondents, the overall score is computed for the five product concepts, and the concept that receives the highest score is recorded. The cross-tabulation between cluster membership and concept chosen is given below.

	Choice				
Cluster	1	2	3	4	5
1	0	12	2	8	0
2	0	13	12	0	1
3	2	0	1	2	15

From the results, we note that:

- Concept 1, which offers middle of the road levels on all attributes is chosen by only 2 out of 68 respondents.

- Concept 5, which only offers a long warranty, is selected by the majority of customers in cluster 3.

- Clusters 1 and 2 select products based on functional attributes.

3.7 Example 3: Tablet Computer

3.7.1 Introduction: In the examples used so far, we only examined products with quantitative attributes such as interest rate and credit limit. However, conjoint analysis can include qualitative attributes such as brand name and country of origin to find how these attributes affect customer evaluation. We now describe a conjoint study on tablet computers where one of the attributes is brand name.

In this application, hypothetical tablet computers were designed using 4 levels each of three attributes:

(1) X_1 (brand name): Apple (1), Dell (2), HP (3), or Samsung (4)

(2) X_2: (hard drive size): 8, 16, 32, or 64 GB

(3) X_3 (battery life): 6, 8, 10, or 12 hours.

We assumed that the utility for a product (X_1, X_2, X_3) can be expressed as:

$$U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3),$$

where:

- U_0 = rating of tablet computer with $X_1 = 1$ (that is, Apple), $X_2 = 8$, and $X_3 = 6$
- $U_1(1) = 0$
- $U_2(8) = 0$
- $U_3(6) = 0$

$U_1(X_1)$, $U_2(X_2)$ and $U_3(X_3)$ are marginal contributions of the three attributes to utility as their levels differed from the base levels (Apple, 8 GB, 6 hours).

Sample: Data were collected from 167 undergraduate students at Syracuse University in Fall 2012.

Survey:

1. Gender: ____Female (0)   ____Male (1)

2. Have you ever owned a smart-phone with web browsing capability (e.g., iPhone, Galaxy, Android)?

____No (0)   ____Yes (1)

3. Have you ever owned a tablet computer?

____No (0)   ____Yes (1)

4. 18 hypothetical tablet computers are listed below. All the 18 tablet computers have the following features in common:

- Price $400
- 10.1” touch screen display with 1920 × 1280 resolution
- 1.9 MP front facing and 5.0 MP rear HD camera
- 2 GB, 1.5 GHz random access memory (RAM)
- HDMI output

- Built in WiFi (802.11 a/b/g/n)
- Bluetooth 3.0 compatible
- High speed USB 2.0 port
- Micro SD card slot
- 1 lb weight

However, these 18 tablet computers differ on three features:

- Brand name: It can be Apple, Dell, HP or Samsung
- On board memory: It can be 8 GB, 16 GB, 32 GB or 64 GB
- Battery life: It can be 6 hours, 8 hours, 10 hours or 12 hours

Instructions:

- Please examine the 18 tablet computers listed below, identify the one you like most, and give it 100 points. Again, please examine the 18 tablet computers listed below, identify the one you like least, and give it 0 points.
- For each of the sixteen other tablet computers, give it a score between 0 and 100. A tablet you like more should get a higher score.

Product	Brand Name	Hard Drive (GB)	Battery Life (Hours)	Score
1	Apple	64	12	
2	Dell	8	6	
3	HP	8	12	
4	HP	16	10	
5	Apple	64	10	
6	Dell	32	6	
7	Samsung	16	12	
8	Dell	8	10	
9	Dell	16	8	
10	Apple	16	6	
11	Samsung	64	8	
12	HP	32	8	
13	Samsung	8	6	
14	Dell	64	12	
15	Samsung	32	10	
16	Apple	32	12	
17	Apple	8	8	
18	HP	64	6	

3.7.2 Estimation: For each respondent, data for tablets 1 and 2, inserted as anchors, were dropped. The scores for tablets 3-18 were recorded as $SCORE_1$, ..., $SCORE_{16}$. Using these scores, we obtained the following from each respondent:

- U_0 = the rating of an Apple tablet computer with 8 GB hard drive and 6 hours of battery life. with $X_1 = 1$, $X_2 = 1$, and $X_3 = 1$ on a 0-100 scale. (Denoted by $U0$)
- $U_1(1) = 0$ (Denoted by $U11$)
- $U_1(2)$: The marginal change in utility if the tablet is Dell instead of Apple. (Denoted by $U12$)
- $U_1(3)$: The marginal change in utility if the tablet is HP instead of Apple. (Denoted by $U13$)
- $U_1(4)$: The marginal change in utility if the tablet is Samsung instead of Apple. (Denoted by $U14$)
- $U_2(1) = 0$ (Denoted by $U21$)
- $U_2(2)$: The marginal change in utility if hard drive size is 16 GB instead of 8 GB. (Denoted by $U22$)
- $U_2(3)$: The marginal change in utility if hard drive size is 32 GB instead of 8 GB. (Denoted by $U23$)
- $U_2(4)$: The marginal change in utility if hard drive size is 64 GB instead of 8 GB. (Denoted by $U24$)
- $U_3(1) = 0$ (Denoted by $U31$)
- $U_3(2)$: The marginal change in utility if battery life is 8 hours instead of 6 hours. (Denoted by $U32$)
- $U_3(3)$: The marginal change in utility if battery life is 10 hours instead of 6 hours. (Denoted by $U33$)
- $U_3(5)$: The marginal change in utility if battery life is 12 hours instead of 6 hours. (Denoted by $U35$)

3.7.3 Evaluating a new product idea: For a given respondent, the utility of a product concept with attribute values within the ranges used in estimation can be estimated as follows:

First obtain the components:

(1) U_0

(2) $U_1(X_1)$: 0 if Apple, U_{12} if Dell, U_{13} if HP, U_{14} if Samsung

(3) $U_2(X_2)$:

If $8 \leq X_2 < 16$: $U_2(X_2) = U22 * (\frac{X2-8}{8})$.

If $16 \leq X_2 < 32$: $U_2(X_2) = U22 + (U23 - U22) * (\frac{X_2-16}{16})$.

If $32 \leq X_2 \leq 64$: $U_2(X_2) = U23 + (U24 - U23) * (\frac{X_2-32}{32})$.

(4) $U_3(X_3)$:

If $6 \leq X_3 < 8$: $U_3(X_3) = U32 * (\frac{X_3-6}{2})$.

$8 \leq X_3 < 10$: $U_3(X_3) = U32 + (U33 - U32) * (\frac{X_3-8}{2})$.

$10 \leq X_3 \leq 12$: $U_3(X_3) = U33 + (U34 - U33) * (\frac{X_3-10}{2})$.

Then add: $U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$.

When you compare multiple new product concepts, a given person will select the option that yields the highest utility.

3.7.4 Benefit Segments using Cluster Analysis: I performed cluster analysis using Minitab. Each of the 167 cases in the data set is expressed in terms of 10 coordinates: the conjoint estimates $U0$, $U12$, $U13$, $U14$, $U22$, $U23$, $U24$, $U32$, $U33$, and $U34$. The program computes the distances between pairs of cases in this ten-dimensional space and tries to create clusters based on proximity. The user specifies the number of clusters to be used. In this case (personal computers), four distinct clusters emerged. For each cluster, the averages of U_0, $U_{11}, \ldots, U_{34}$ are given below.

Cluster 1 (40 respondents)			
$U_0 = 42.4875$			
$U_{11} = 0$	$U_{12} = -0.0875$	$U_{13} = -1.875$	$U_{14} = 1.5125$
$U_{21} = 0$	$U_{22} = 1.675$	$U_{23} = 2.35$	$U_{24} = 0.35$
$U_{31} = 0$	$U_{32} = 3.2$	$U_{33} = 8.875$	$U_{34} = 12.025$

Cluster 2 (59 respondents)			
$U_0 = 67.3453$			
$U_{11} = 0$	$U_{12} = -26.1186$	$U_{13} = -22.8305$	$U_{14} = -13.4873$
$U_{21} = 0$	$U_{22} = 2.8941$	$U_{23} = 8.1017$	$U_{24} = 10.8729$
$U_{31} = 0$	$U_{32} = 1.0551$	$U_{33} = 3.5$	$U_{34} = 5.3475$

Cluster 3 (32 respondents)			
$U_0 = 78.3398$			
$U_{11} = 0$	$U_{12} = -62.7422$	$U_{13} = -54.4375$	$U_{14} = -53.3594$
$U_{21} = 0$	$U_{22} = 4.4609$	$U_{23} = 7.25$	$U_{24} = 8.2188$
$U_{31} = 0$	$U_{32} = 1.5078$	$U_{33} = 4.8594$	$U_{34} = 4.7813$

Cluster 4 (36 respondents)			
$U_0 = 21.3646$			
$U_{11} = 0$	$U_{12} = -10.7014$	$U_{13} = -14.3194$	$U_{14} = -6.6528$
$U_{21} = 0$	$U_{22} = 16.1806$	$U_{23} = 25.7014$	$U_{24} = 35.3056$
$U_{31} = 0$	$U_{32} = 9.8958$	$U_{33} = 21.5486$	$U_{34} = 31.2986$

Summary:

- Cluster 1 is largely indifferent to brand name or attribute values. For all other clusters, everything else being same, the brand name Apple has a higher score than brand names Dell, HP or Samsung.
- Cluster 2 prefers Apple, but hard drive size and battery life are marginally important to these customers.
- Cluster 3 has very strong preference for Apple. For example, everything else being same, a Dell tablet receives 62.74 points less than Apple on a 0-100 scale. The functional attributes (HD, battery life) are only marginally important to these customers.
- Cluster 4 customers find both hard drive size and battery life more important than brand name.

16.7.5 Output: The output is available as an Excel file (tabletf12sum2013class.xls) to the interested reader. For each of the 167 observations, you have the following data:

Case number (1 through 167)

Gender: 1 if male, 0 if female

Smart: 1 if the student has ever owned a smart-phone, 0 if not

Tablet: 1 if the student has ever owned a tablet computer, 0 if not

$SCORE_1 - SCORE_{16}$: The scores of the sixteen personal computers used in the analysis.

CLUSTER: Cluster membership using three-cluster solution; 1 if cluster 1, 2 if cluster 2, 3 if cluster 3.

Estimates of: $U0$, $U11$, $U12$, $U13$, $U14$, $U21$, $U22$, $U23$, $U24$, $U31$, $U32$, $U33$, and $U34$.

3.8 Estimating Price Sensitivity

In conjoint analysis, price can be used as one of the attributes. Using the conjoint estimates from a sample of potential customers, the marketer can estimate how customers make trade offs between an attribute level and its price.

Also, for a given new product concept, the marketer can simulate multiple scenarios where the price of the product is varied, but the other attributes of the product and all attributes of the alternative products are held fixed. From such a simulation, the demands for the new product concept at different price levels can be estimated. From this demand curve, the marketer can determine how demand responds to price changes. The marketer can develop such demand curves for the overall sample as well as the different segments within the sample.

3.9 Ideal Point Model

In our discussions, we did not make any assumption about what the part-worth functions $U_1(X_1)$, etc., look like. Also, when we estimate these functions using dummy variable regression as discussed in Chapter 17, we do not make assumptions about functional form. We now discuss one specific case of the additive part-worth model called the ideal point model.

The ideal point model assumes that a given customer has an ideal point, $(X_1^*, X_2^*, \ldots, X_n^*)$, and the customer's evaluation score of a product decreases with distance from the ideal point. The distance of a product $(X_1, X_2, \ldots, X_n)$ from the ideal point is defined as follows:

$$D^2 = W_1(X_1 - X_1^*)^2 + W_2(X_2 - X_2^*)^2 + \ldots + (X_n - X_n^*)^2,$$

where W_1, W_2, ..., W_n are all greater than zero and reflect the relative importance of the different attributes. (In Assignment 2, we applied a special case of this distance where $W_1 = W_2 = \ldots = 1$.) The evaluation score of any product $(X_1, X_2, \ldots, X_n)$ can be expressed as:

$$U(X_1, X_2, \ldots, X_n) = K - D^2 = K - \sum_{i=1}^{n} W_i(X_i - X_i^*)^2,$$

where K is a constant and equals the utility of the ideal product. Rewriting:

$$U(X_1, X_2, \ldots, X_n) = \left[K - \sum_{i=1}^{n} W_i X_i^{*2}\right] + \sum_{i=1}^{n}(2W_i X_i^* X_i - W_i X_i^2)$$

If each X_i is expressed as the difference of the attribute level from the lowest level, then this becomes our usual additive part-worth utility model where:

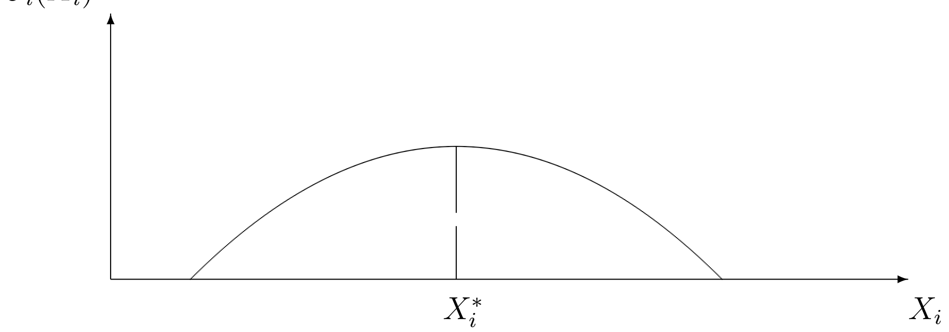
$U_0 = K - \sum_{i=1}^{n} W_i x_i^{*2}$

$U_1(X_1) = (2W_1X_1^*)X_1 - W_1X_1^2$

$\vdots$

$U_n(X_n) = (2W_nX_n^*) - W_nX_n^2$

For any attribute X_i, the part-worth function is a parabola that reaches the highest point at X_i^* as shown below. In Section 17.6 we discuss how to estimate this model.



3.10 Developing hypothetical products

To use conjoint analysis requires hypothetical new product concepts. Commercial packages (e.g., from www.sawtoothsoftware.com) can generate hypothetical products for you. You can also develop hypothetical products yourself.

Note that in the examples in Sections 3.5, 3.6 and 3.7, I used 4 levels each of three attributes. Thus, in each case, $4 \times 4 = 64$ hypothetical products were possible. Instead, we used 16 hypothetical products plus two anchors. This was done using a method called Latin Square Design that can be used for three attributes, each attribute with the same number of levels. I now show how the Latin Square design can be used to get 16 hypothetical products when there are three attributes with four levels each. You can easily extend the method for any number of levels, as long as there are three attributes each with the same number of levels. The attributes may be quantitative (such as interest rate) or categorical (such as brand name).

Step 1: First focus on attributes 1 and 2, and write the $4 \times 4 = 16$ combinations as a table:

	Attribute 2			
Attribute 1	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)
4	(4,1)	(4,2)	(4,3)	(4,4)

Step 2: For each of the 16 cells in the table above, enter the level of attribute 3 such that the following happens:

- The same level of attribute 3 appears in each row only once.

- The same level of attribute 3 appears in each column only once.

This ensures that a given level of any attribute appears with a given level of another attribute exactly once. Thus, if you treat the set of 16 hypothetical products as a sample, any pair of attributes have zero correlation. The conditions above can be satisfied in many ways. Below, I show one way of doing it.

Start with row 1, and put the four levels of the third attribute sequentially. You can use any starting point. For example, you can put them as $(1, 2, 3, 4)$, or $(2, 3, 4, 1)$, or $(3, 4, 1, 2)$, or, $(4, 1, 2, 3)$. In row 2, shift them by one place to the left. For example, if you have $(1, 2, 3, 4)$ in row 1, then it should be $(2, 3, 4, 1)$ in row 2. Again, shift them by 1 in row 3, and shift them by 1 in row 4. Thus, depending on what you chose in row 1, you may have any of the following four Latin-Square designs:

Design 1	Attribute 2			
Attribute 1	1	2	3	4
1	(1,1,1)	(1,2,2)	(1,3,3)	(1,4,4)
2	(2,1,2)	(2,2,3)	(2,3,4)	(2,4,1)
3	(3,1,3)	(3,2,4)	(3,3,1)	(3,4,2)
4	(4,1,4)	(4,2,1)	(4,3,2)	(4,4,3)

Design 2	Attribute 2			
Attribute 1	1	2	3	4
1	(1,1,2)	(1,2,3)	(1,3,4)	(1,4,1)
2	(2,1,3)	(2,2,4)	(2,3,1)	(2,4,2)
3	(3,1,4)	(3,2,1)	(3,3,2)	(3,4,3)
4	(4,1,1)	(4,2,2)	(4,3,3)	(4,4,4)

Design 3	Attribute 2			
Attribute 1	1	2	3	4
1	(1,1,3)	(1,2,4)	(1,3,1)	(1,4,2)
2	(2,1,4)	(2,2,1)	(2,3,2)	(2,4,3)
3	(3,1,1)	(3,2,2)	(3,3,3)	(3,4,4)
4	(4,1,2)	(4,2,3)	(4,3,4)	(4,4,1)

Design 4	Attribute 2			
Attribute 1	1	2	3	4
1	(1,1,4)	(1,2,1)	(1,3,2)	(1,4,3)
2	(2,1,1)	(2,2,2)	(2,3,3)	(2,4,4)
3	(3,1,2)	(3,2,3)	(3,3,4)	(3,4,1)
4	(4,1,3)	(4,2,4)	(4,3,1)	(4,4,2)

For example, in Design 4, the first product in Row 1 has level 1 of attribute 1, level 1 of attribute 2, and level 4 of attribute 4.

Suggestion: In preparing hypothetical products, avoid the extreme products which all respondents are likely to choose as best or worst. Add these extreme products as anchors and give respondents 18 products to evaluate. Drop the two extreme products when you estimate the model.

3.11 Exercise Problem

Conjoint analysis was done to determine how a Syracuse University undergraduate student's evaluation of a plug-in electric car depends on the levels of three attributes:

- X_1: Charging time after a full discharge. Range: 6 to 24 hours.
- X_2: Top speed. Range: 60 to 120 miles/hour

- X_3: Driving range, that is, the number of highway miles one can drive after a full charge. Range: 200 to 400 miles.

It is assumed that a student's rating of a plug-in electric car on a $0-100$ scale can be expressed as:

$$U(X_1, X_2, X_3) \quad = \quad U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3),$$

where $U_1(6) = 0$, $U_2(60) = 0$, $U_3(200) = 0$, and a higher score means the student likes the product better.

For a given student:

$U_0 = 40$		
$U_1(6) = 0$	$U_1(12) = -10$	$U_1(24) = -40$
$U_2(60) = 0$	$U_2(80) = 20$	$U_2(120) = 30$
$U_3(200) = 0$	$U_3(300) = 10$	$U_3(400) = 30$

Which of the following two plug-in electric cars will this student prefer?

- Car 1: $X_1 = 15$, $X_2 = 100$, $X_3 = 250$
- Car 2: $X_1 = 9$, $X_2 = 75$, $X_3 = 350$

3.12 Additional Reading

3 1. Paul E. Green, and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: issues and outlook," *Journal of Consumer Research*, volume 5, pages 103-23.

2. ____and ____(1990), "Conjoint Analysis in Marketing: new developments with implications for research and practice," *Journal of Marketing*, volume 54, pages 3-19.

3. www.sawtoothsoftware.com

4 Estimation of Conjoint Model (Optional)

Consider the additive part-worth conjoint model:

$$(4.1)\ U(X_1, X_2, \ldots, X_n) \quad = \quad U_0 + U_1(X_1) + U_2(X_2) + \ldots + U_n(X_n),$$

where U_0 is the evaluation score where each attribute is at the lowest level of its range, and each part-worth score is zero if the attribute is at the lowest level of its range. To estimate the conjoint model, we assume that when asked to rate a hypothetical product, the respondent reports the utility given by equation (4.1) plus a random error as the evaluation score Y, that is,

$$(4.2) \quad Y = U(X_1, \ldots, X_n) + \epsilon = U_0 + U_1(X_1) + \ldots + U_n(X_n) + \epsilon,$$

where ϵ is the random error. Equation (4.2) can be estimated using dummy variable regression analysis.

4.1 Review of Dummy Variable Regression

We now illustrate how dummy variable regression can be used in conjoint model estimation using two examples based on the laptop example discussed in Sections 3.2 and 3.3. Even though the attributes in the laptop example are quantitative variables, for the purpose of estimation we treat them as categorical variables with a finite number of categories.

Example 1. One Categorical Independent Variable: Consider one customer who is evaluating hypothetical laptop computers on a 0-100 scale where 0 is very poor and 100 is excellent. All the hypothetical laptop computers are identical in every way except for one attribute: weight (X_1) which can be 4 lb, 6 lb, 8 lb, or 10 lb. We will treat weight as a categorical variable that can fall in the four categories: 4 lb, 6 lb, 8 lb and 10 lb. To represent the four categories of weight, we need $(4 - 1)$, that is, 3 dummy variables. For example, we can use:

$D_1 = 1$ if weight is 6 lb, and $D_1 = 0$ if weight is 4, 8, or 10 lb

$D_2 = 1$ if weight is 8 lb, and $D_2 = 0$ if weight is 4, 6 or 10 lb

$D_3 = 1$ if weight is 10 lb, and $D_3 = 0$ if weight is 4, 6, or 8 lb

If $D_1 = D_2 = D_3 = 0$, then weight is the remaining level, 4 lb.

Let Y denote the rating of a hypothetical laptop computer by this customer. Consider the regression model:

$$(4.3) \quad Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \epsilon$$

This equation is simply a way to allow the mean of Y to be different for the four weight categories:

(1) $X_1 = 4$: $D_1 = D_2 = D_3 = 0$, $Y = \beta_0 + \epsilon$

(2) $X_1 = 6$: $D_1 = 1$, $D_2 = D_3 = 0$, $Y = \beta_0 + \beta_1 + \epsilon$

(3) $X_1 = 8$: $D_1 = 0$, $D_2 = 1$, $D_3 = 0$, $Y = \beta_0 + \beta_2 + \epsilon$

(4) $X_1 = 10$: $D_1 = D_2 = 0$, $D_3 = 1$, $Y = \beta_0 + \beta_3 + \epsilon$

Thus, β_0 is the mean of Y for laptops that weigh 4 lb. In any given assessment, the rating is the mean plus a random error ϵ that is equally likely to be positive or negative and is zero on the average.

$\beta_1 =$ Mean rating of a laptop with weight 6 lb $-$ Mean rating of a laptop with weight 4 lb

$\beta_2 =$ Mean rating of a laptop with weight 8 lb $-$ Mean rating of a laptop with weight 4 lb

$\beta_3 =$ Mean rating of a laptop with weight 10 lb $-$ Mean rating of a laptop with weight 4 lb

Example 2. Two Categorical Independent Variables: Let Y, D_1, D_2, and D_3 be as defined in Example 1. However, in any addition to weight (X_1), the hard drive size (X_2) can also vary from one hypothetical laptop computer to another, and X_2 can take one of four values: 80 GB, 100 GB, 120 GB, or 160 GB. To represent the four levels of X_2, we use $(4 - 1) = 3$ dummy variables:

$D_4 = 1$ if $X_2 = 100$, and $D_4 = 0$ if X_2 is 80, 120 or 160 GB

$D_5 = 1$ if $X_2 = 120$, and $D_5 = 0$ if X_2 is 80, 100 or 160 GB

$D_6 = 1$ if $X_2 = 160$, and $D_6 = 0$ if X_2 is 80, 100 or 120 GB

If $D_4 = D_5 = D_6 = 0$, then $X_2 = 80$ GB

Consider the regression model:

$$(4.4) \quad Y = \beta_0 + (\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) + (\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) + \epsilon$$

In this case, there are $4 \times 4 = 16$ combinations of weight and hard drive size. From Equation (4.4), the mean Y for these sixteen combinations are given as follows:

	Hard Drive Size (X_2)			
Weight (X_1)	80 GB	100 GB	120 GB	160 GB
4 lb	β_0	$\beta_0 + \beta_4$	$\beta_0 + \beta_5$	$\beta_0 + \beta_6$
6 lb	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_4$	$\beta_0 + \beta_1 + \beta_5$	$\beta_0 + \beta_1 + \beta_6$
8 lb	$\beta_0 + \beta_2$	$\beta_0 + \beta_2 + \beta_4$	$\beta_0 + \beta_2 + \beta_5$	$\beta_0 + \beta_2 + \beta_6$
10 lb	$\beta_0 + \beta_3$	$\beta_0 + \beta_3 + \beta_4$	$\beta_0 + \beta_3 + \beta_5$	$\beta_0 + \beta_3 + \beta_6$

In this case:

- β_0 is the mean evaluation score of a laptop computer with $X_1 = 4$ and $X_2 = 80$
- For any given level of X_2:

  $\beta_1 =$ Mean rating of laptop with weight 6 lb $-$ Mean rating of laptop with weight 4 lb

  $\beta_2 =$ Mean rating of laptop with weight 8 lb $-$ Mean rating of laptop with weight 4 lb

  $\beta_3 =$ Mean rating of laptop with weight 10 lb $-$ Mean rating of laptop with weight 4 lb

- For any given level of X_1:

  β_4 = Mean rating of laptop with hard drive size 100 GB $-$ Mean rating of laptop with hard drive size 80 GB

  β_5 = Mean rating of laptop with hard drive size 120 GB $-$ Mean rating of laptop with hard drive size 80 GB

  β_6 = Mean rating of laptop with hard drive size 160 GB $-$ Mean rating of laptop with hard drive size 80 GB

- For any given combination of X_1 and X_2, ϵ is a random error in evaluation that is equally likely to be positive and negative and is zero on the average.

- $(\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3)$ is the marginal adjustment in mean evaluation score depending on which of the four levels weight (X_1) is at.

  - If $X_1 = 4$, then $(\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) = 0$, that is, no adjustment is needed.
  - If $X_1 = 6$, then $(\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) = \beta_1$, that is, β_1 is the marginal adjustment if weight is 6 lb instead of 4 lb.
  - If $X_1 = 8$, then $(\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) = \beta_2$, that is, β_2 is the marginal adjustment if weight is 8 lb instead of 4 lb.
  - If $X_1 = 10$, then $(\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) = \beta_3$, that is, β_3 is the marginal adjustment if weight is 10 lb instead of 4 lb.

- $(\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6)$ is the marginal adjustment in mean evaluation score depending on which of the four levels hard drive size (X_2) is at.

  - If $X_2 = 80$, then $(\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) = 0$, that is, no adjustment is needed.
  - If $X_2 = 100$, then $(\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) = \beta_4$, that is, β_4 is the marginal adjustment if hard drive size is 100 GB instead of 80 GB.
  - If $X_2 = 120$, then $(\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) = \beta_5$, that is, β_5 is the marginal adjustment if hard drive size is 120 GB instead of 80 GB.
  - If $X_2 = 160$, then $(\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) = \beta_6$, that is, β_6 is the marginal adjustment if hard drive size is 160 GB instead of 80 GB.

4.2 Application to Conjoint Analysis

Consider a product with 2 attributes: the laptop computer in Example 2 with attributes weight (X_1) and hard drive size (X_2). Suppose we wish to examine how evaluation score Y changes if X_1 varies over the range $4 - 10$ lb, and X_2 varies over the range 80 GB to 160 GB.

In this case, the additive part-worth model can be expressed as follows:

$$(4.5) \quad U(X_1, X_2) = U_0 + U_1(X_1) + U_2(X_2),$$

where $U(X_1, X_2)$ is the mean evaluation score of a laptop with attribute levels (X_1, X_2).

Without loss of generality, we assume that:

- $U_1(4) = 0$
- $U_2(80) = 0$

Under these assumptions:

$$U(4, 80) \quad = \quad U_0 + U_1(4) + U_2(80) \quad = \quad U_0 + 0 + 0 \ = \ U_0$$

Therefore:

- U_0 is the mean evaluation score of the "basic" laptop computer with each attribute at the lowest levels in the ranges considered, that is, $X_1 = 4$ and $X_2 = 80$.
- $U_1(X_1)$ is the marginal adjustment to mean evaluation score depending on the level of weight (X_1). If $X_1 = 4$, then no adjustment is needed.
- $U_2(X_2)$ is the marginal adjustment to mean evaluation score depending on the level of hard drive size (X_2). If $X_2 = 80$, then no adjustment is needed.

The objective of conjoint analysis is to estimate:

(1) The mean evaluation score of the "basic" product, U_0

(2) $U_1(X_1)$ at three or more levels in the range $4 - 10$. The extreme levels 4 and 10 are always included.

(3) $U_2(X_2)$ at three or more levels in the range $80 - 160$. The extreme levels 80 and 160 are always included.

In this case, we use the following levels of the two attributes for estimation:

(1) Four levels of X_1: 4, 6, 8, and 10.

(2) Four levels of X_2: 80, 100, 120, and 160.

Suppose several hypothetical laptop computers are created using different combinations of these attribute levels, and the respondent provides evaluation scores for these hypothetical products. We assume that the evaluation score can be expressed as:

$$(4.6) \quad Y \ = \ U_0 + U_1(X_1) + U_2(X_2) + \epsilon$$

Consider the regression model (4.4) described already:

$$Y = \beta_0 + (\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) + (\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) + \epsilon$$

where the dummy variables are as defined before. Comparing (17.4) and (17.6), we find that for the $4 \times 4 = 16$ combinations of X_1 and X_2 used to generate the hypothetical products, we have:

$$(4.7) \quad U_0 + U_1(X_1) + U_2(X_2) \quad = \quad \beta_0 + (\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) + (\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6)$$

From (4.7), we get the following results:

- If $X_1 = 4$ and $X_2 = 80$, then $U_0 = \beta_0$.
- If $X_1 = 6$ and $X_2 = 80$, then $U_0 + U_1(6) = \beta_0 + \beta_1$. Since $U_0 = \beta_0$, we have $U_1(6) = \beta_1$.
- If $X_1 = 8$ and $X_2 = 80$, then $U_0 + U_1(8) = \beta_0 + \beta_2$. Since $U_0 = \beta_0$, we have $U_1(8) = \beta_2$.
- If $X_1 = 10$ and $X_2 = 80$, then $U_0 + U_1(10) = \beta_0 + \beta_3$. Since $U_0 = \beta_0$, we have $U_1(10) = \beta_3$.
- If $X_1 = 4$ and $X_2 = 100$, then $U_0 + U_2(100) = \beta_0 + \beta_4$. Since $U_0 = \beta_0$, we have $U_2(100) = \beta_4$.
- If $X_1 = 4$ and $X_2 = 120$, then $U_0 + U_2(120) = \beta_0 + \beta_5$. Since $U_0 = \beta_0$, we have $U_2(120) = \beta_5$.
- If $X_1 = 4$ and $X_2 = 160$, then $U_0 + U_2(160) = \beta_0 + \beta_6$. Since $U_0 = \beta_0$, we have $U_2(160) = \beta_6$.

Suppose, for a given respondent, we ran a regression using Y as the dependent variable and D_1, D_2, D_3, D_4, D_5, and D_6 as the independent variables, and obtained estimates of β_0, β_1, ..., β_6, denoted by b_1, ..., b_6. The, for this respondent:

- $U_0 \approx b_0$
- $U_1(4) = 0$, $U_1(6) \approx b_1$, $U_1(8) \approx b_2$, $U_1(10) \approx b_3$
- $U_2(80) = 0$, $U_2(100) \approx b_4$, $U_2(120) \approx b_5$, $U_2(160) \approx b_6$

Using these estimate, we can assess the evaluation score of any new product concept with X_1 in the range 4-10 lb and X_2 in the range 80-160 GB for this respondent as shown in Chapter 16.

4.3 Estimation of Conjoint Model for Credit Card Data in Section 3.5

The hypothetical credit cards were designed using 4 levels each of three attributes:
(1) X_1 (interest rate): 6, 9, 12, or 18 percent APR.
(2) X_2: (credit limit): 5, 10, 25, or 50 (unit=$1000).
(3) X_3 (annual fee): 0, 10, 20, or 50.

We assumed that the utility for a product (X_1, X_2, X_3) can be expressed as:

$$U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3).$$

U_0 is the utility of the basic product with $X_1 = 6$, $X_2 = 5$ and $X_3 = 0$. $U_1(X_1)$, $U_2(X_2)$ and $U_3(X_3)$ are marginal contributions of the three attributes to utility as their levels increased from the lowest. These are all zero at the basic levels, that is, $U_1(6) = U_2(5) = U_3(0) = 0$. Each respondent rated 18 hypothetical credit cards, and data from cards 3 and 9, inserted as anchors, were dropped. The remaining 16 cards were re-ranked from 1 to 16. For each card, we computed a score equal to $(\frac{100}{15}) \times (17 - \text{rank})$. Thus, the card ranked 1 got a score of 100, and the card ranked 16 got a score of zero. The score of a hypothetical card is the dependent variable Y.

Estimation: We use the dummy variable model:

$$Y = \beta_0 + (\beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) + (\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) + (\beta_7 D_7 + \beta_8 D_8 + \beta_9 D_9) + \epsilon.$$

The D's are only meaningful for the attribute levels used in the design, that is, $X_1 = 6$, 9, 12, 18, $X_2 = 5$, 10, 25, or 50, and $X_3 = 0$, 10, 20, or 50. β_0 is the utility of the basic product (6, 5, 0). The D's and the remaining β's may be interpreted as follows:

X_1	D_1	D_2	D_3	$U_1(X_1) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3$
6	0	0	0	0
9	1	0	0	β_1
12	0	1	0	β_2
18	0	0	1	β_3

X_2	D_4	D_5	D_6	$U_2(X_2) = \beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6$
5	0	0	0	0
10	1	0	0	β_4
25	0	1	0	β_5
50	0	0	1	β_6

X_3	D_7	D_8	D_9	$U_3(X_3) = \beta_7 D_7 + \beta_8 D_8 + \beta_9 D_9$
0	0	0	0	0
10	1	0	0	β_7
20	0	1	0	β_8
50	0	0	1	β_9

Evaluating a new product idea: Let B_0, B_1, …, B_9 represent the parameter estimates for a given person. Then, for a product concept with attribute values within the ranges used in estimation, utility can be estimated as follows:

First compute the components:

(1) U_0: Use B_0.

(2) $U_1(X_1)$: Compute it as follows:

$6 \leq X_1 < 9$: $U_1(X_1) = B_1 * (\frac{X_1 - 6}{3})$.

$9 \leq X_1 < 12$: $U_1(X_1) = B_1 + (B_2 - B_1) * (\frac{X_1 - 9}{3})$.

$12 \leq X_1 \leq 18$: $U_1(X_1) = B_2 + (B_3 - B_2) * (\frac{X_1 - 12}{6})$.

(3) $U_2(X_2)$: Express X_2 in units of \$1000, e.g., use 15 if credit limit is \$15,000. Compute $U_2(X_2)$ as follows:

$5 \leq X_2 < 10$: $U_2(X_2) = B_4 * (\frac{X_2 - 5}{5})$.

$10 \leq X_2 < 25$: $U_2(X_2) = B_4 + (B_5 - B_4) * (\frac{X_2 - 10}{15})$.

$25 \leq X_2 \leq 50$: $U_2(X_2) = B_5 + (B_6 - B_5) * (\frac{X_2 - 25}{25})$.

(4) $U_3(X_3)$: Compute this as follows:

$0 \leq X_3 < 10$: $U_3(X_3) = B_7 * (\frac{X_3}{10})$.

$10 \leq X_3 < 20$: $U_3(X_3) = B_7 + (B_8 - B_7) * (\frac{X_3 - 10}{10})$.

$20 \leq X_3 \leq 50$: $U_3(X_3) = B_8 + (B_9 - B_8) * (\frac{X_3 - 20}{30})$.

Then add: $U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$.

When you compare multiple new product concepts, a given person will select the option that yields the highest utility.

4.4 Estimation of Conjoint Model for PC Data from Section 3.6

The hypothetical personal computers were designed using 4 levels each of three attributes:

(1) X_1 (warranty length): 1, 2, 3, or 4 years.

(2) X_2: (RAM): 1, 2, 3, or 4 (unit= 128 MB).

(3) X_3 (hard drive size): 1, 2, 3, or 5 (unit = 20 GB).

We assumed that the utility for a product (X_1, X_2, X_3) can be expressed as:

$$U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$$

U_0 is the utility of the basic product with $X_1 = 1$, $X_2 = 1$ and $X_3 = 1$. $U_1(X_1)$, $U_2(X_2)$ and $U_3(X_3)$ are marginal contributions of the three attributes to utility as their levels increased from the lowest. These are all zero at the basic levels, that is, $U_1(1) = U_2(1) = U_3(1) = 0$. Each respondent rated 18 hypothetical personal computers, and data from computers 1 and 6, inserted as anchors, were dropped. The remaining PC's were re-ranked from 1 to 16. For each PC, we computed a score equal to $(\frac{100}{15}) \times (17 - \text{rank})$. Thus, the PC ranked 1 got a score of 100, and the PC ranked 16 got a score of zero. The score of a hypothetical card is the dependent variable Y.

Two models were used to estimate the model: the basic (dummy variable) model, and the ideal point model. These models are now described separately.

Estimation: We use the dummy variable regression model:

$$Y = \beta_0 + (\beta_1 D_1 + \beta_2 + \beta_3 D_3) + (\beta_4 D_4 + \beta_5 D_5 + \beta_6 D_6) + (\beta_7 D_7 + \beta_8 D_8 + \beta_9 D_9) + \epsilon$$

The D's are only meaningful for the attribute levels used in the design, that is, $X_1 = 1$, 2, 3, 4, $X_2 = 1$, 2, 3, or 4, and $X_3 = 1$, 2, 3, or 5. β_0 is the utility of the basic product (1, 1, 1). The D's and the remaining β's may be interpreted as follows:

X_1	D_1	D_2	D_3	$U_1(X_1) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3$
1	0	0	0	0
2	1	0	0	β_1
3	0	1	0	β_2
4	0	0	1	β_3

X_2	D_4	D_5	D_6	$U_2(X_2) = \beta_4 D_4 + \beta_5 D_6 + \beta_6 D_6$
1	0	0	0	0
2	1	0	0	β_4
3	0	1	0	β_5
4	0	0	1	β_6

X_3	D_7	D_8	D_9	$U_3(X_3) = \beta_7 D_7 + \beta_8 D_8 + \beta_9 D_9$
1	0	0	0	0
2	1	0	0	β_1
3	0	1	0	β_2
5	0	0	1	β_3

Evaluating a new product idea: Let B_0, B_1, …, B_9 represent the parameter estimates for a given person. Then, for a product concept with attribute values within the ranges used in estimation, utility can be estimated as follows:

First compute the components:

(1) U_0: Use B_0.

(2) $U_1(X_1)$: Express X_1 in years. Compute $U_1(X_1)$ as follows:

$1 \le X_1 < 2$: $U_1(X_1) = B_1 * (X_1 - 1)$.

$2 \le X_1 < 3$: $U_1(X_1) = B_1 + (B_2 - B_1) * (X_1 - 2)$.

$3 \le X_1 \le 4$: $U_1(X_1) = B_2 + (B_3 - B_2) * (X_1 - 3)$.

(3) $U_2(X_2)$: Express X_2 in units of 128 MB, e.g., use 1.5 if RAM is 192 MB. Compute $U_2(X_2)$ as follows:

$1 \le X_2 < 2$: $U_2(X_2) = B_4 * (X_2 - 1)$.

$2 \le X_2 < 3$: $U_2(X_2) = B_4 + (B_5 - B_4) * (X_2 - 2)$.

$3 \le X_2 \le 4$: $U_2(X_2) = B_5 + (B_6 - B_5) * (X_2 - 3)$.

(4) $U_3(X_3)$: Express X_3 in units of 20 GB. For example, use 2.5 if hard drive size is 50 GB. Compute $U_3(X_3)$ as follows:

$1 \le X_3 < 2$: $U_3(X_3) = B_7 * (X_3 - 1)$.

$2 \leq X_3 < 3$: $U_3(X_3) = B_7 + (B_8 - B_7) * (X_3 - 2)$.
$3 \leq X_3 \leq 5$: $U_3(X_3) = B_8 + (B_9 - B_8) * (\frac{X_3 - 3}{2})$.
Then add: $U(X_1, X_2, X_3) = U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3)$.

When you compare multiple new product concepts, a given person will select the option that yields the highest utility.

4.5 Estimation of Ideal Point Model

As discussed in Section 3.9, the ideal point model is an additive part-worth utility model and can therefore be estimated using dummy variable regression. However, it can also be estimated more directly as discussed here.

From Section 3.9, the utility of $(X_1, X_2, \ldots, X_n)$ can be expressed as:

$$U(X_1, X_2, \ldots, X_n) = \left[K - \sum_{i=1}^{n} W_i X_i^{*2}\right] + \sum_{i=1}^{n} (2W_i X_i^* X_i - W_i X_i^2),$$

where each X_i is expressed as the difference of the attribute level from the lowest level. Also, as shown in Section 16.8.2:
$U_0 = K - \sum_{i=1}^{n} W_i x_i^{*2}$
$U_1(X_1) = (2W_1 X_1^*)X_1 - W_1 X_1^2$
$\vdots$
$U_n(X_n) = (2W_n X_n^*) - W_n X_n^2$
Let Y be the rating score of the hypothetical product $(X_1, X_2, \ldots, X_n)$. Assuming that the respondent reports the utility plus a random error, we have:

$$Y = \beta_0 + (\beta_{11} X_1 + \beta_{12} X_1^2) + \ldots + (\beta_{n1} X_n + \beta_{n2} X_n^2) + \epsilon,$$

where:
$\beta_0 = K - \sum_{i=1}^{n} W_i x_i^{*2}$ (Utility of product with each attribute at lowest level)
$\beta_{11} = (2W_1 X_1^*)$, $\beta_{12} = -W_1$
$\vdots$
$\beta_{n1} = (2W_n X_n^*)$, $\beta_{n2} = -W_n$
This regression model can be estimated by regressing Y with X_1, X_1^2, ..., X_n, X_n^2.

Note: The regression model is more general than the ideal point model. Consider any attribute X_i.

- If $\beta_{i1} > 0$ and $\beta_{i2} < 0$, we get the classic ideal point model where $W_i = |\beta_{i2}|$, and $X_i^* = \frac{2\beta_{i1}}{|\beta_{i2}|}$.
- If $\beta_{i1} < 0$ and $\beta_{i2} < 0$, then the attribute level should be set at the lowest point of the range.

- If $\beta_{i1} > 0$ and $\beta_{i2} > 0$, then the attribute level should be set at the highest point of the range.

Estimation of Credit Card data: We estimate the regression model:

$$Y = C_0 + C_1(X_1 - 6) + C_2(X_1 - 6)^2 + C_3(X_2 - 5) + C_4(X_2 - 5)^2 + C_5X_3 + C_6X_3^2 + \epsilon,$$

where X_1, X_2, and X_3 are same as in Section 16.5, and C_0 is the rating of the card with $X_1 = 6$, $X_2 = 5$, and $X_3 = 0$ on a 0-16 scale. The results are provided in the Excel file "creditf01 ideal point.xls."

Estimation of PC data: We estimate the regression model:

$$Y = C_0 + C_1(X_1 - 1) + C_2(X_1 - 1)^2 + C_3(X_2 - 1) + C_4(X_2 - 1)^2 + C_5(X_3 - 1) + C_6(X_3 - 1)^2 + \epsilon,$$

where X_1, X_2, and X_3 are same as in Section 16.6, and C_0 is the rating of the PC with $X_1 = 1$, $X_2 = 1$, and $X_3 = 1$ on a 0-16 scale. The results are provided in the Excel file "pcf01 ideal point.xls."

5 Basics of Logit

5.1 Method

5.1.1 Context: In business management, we frequently encounter categorical dependent variables. For example:

1. We have a sample of business ventures that were launched in 2010. Some of them succeeded and the others failed. We have divided the ventures into two groups: successful and unsuccessful, and we wish to identify the characteristics which differentiate these two groups from each other. Here, the dependent variable is the dichotomous variable, successful or not.

2. We wish to determine which type of motor vehicle a given individual will choose to buy, sports car, sedan, SUV, or minivan, based on the gender, age, income, and family size of the individual. Here, the dependent variable is the type of vehicle chosen, which is a categorical variable with four levels.

For a categorical variable Y, Logit is used to determine how the probability that Y takes any given value depends on one or more independent variables.

For simplicity, we only consider **binary Logit**, where the dependent variable can take one of two values. (A variable that can take two values is also called a **dichotomous** variable.)

5.1.2 The Model: Following convention, we call dichotomous dependent variable Y, which is 1 if the object considered has the property of interest, and 0 if not. (Some statistical packages want $Y = 2$ if the object does not have the property of interest. If you coded your data 1/0, simply create a new Y equal 2 minus the original Y. This will be 1 if the original Y is 1, and 2 if the original Y is zero.)

Let X_1, X_2, ..., X_n be the independent variables. As in regression analysis, these variables may be interval scaled variables, dummy variables, or products of such variables. To model how the probability that $Y = 1$ depends on the levels of the independent variables, we first define an **indicator** function as follows:

$$(5.1) \quad I \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

From a given values of the independent variables, the probability that $Y = 1$ is given by:

$$(5.2) \quad P(Y = 1|X_1, \ldots, X_n) \quad = \quad \frac{1}{1 + e^{-I}}.$$

The function on the right hand side of (5.2) is called the **logistic function**, and the name **Logit** comes from it.

For a given value of I, we use equation (5.2) to compute the probability that $Y = 1$. For example, if $I = 2$, then

$$P(Y = 1|I) \ = \ F(I) = \frac{1}{1 + e^{-2}} = 0.881$$

Since there are just two categories of Y, Y must be equal to 2 if it is not equal to 1. Hence, in this case,

$$P(Y=2|I) \;=\; 1-P(Y=1|I) \;=\; 1-.881 \;=\; 0.119.$$

5.1.3 Interpretation of Logit Parameters: Note that $P(Y=1)$ is determined by the value of I, and increases as I increases. Suppose, for example, $\beta_1>0$. If X_1 increases and all other independent variables remain unchanged, I increases and $P(Y=1)$ increases. Thus, if $\beta_1>0$, then an increase in X_1 results in an increase in the probability that $Y=1$. Similarly, if $\beta_1<0$, then an increase in X_1 reduces the probability that $Y=1$. If $\beta_1=0$, then a change in X_1 has no effect on the probability that $Y=1$.

Note: In Logit, the same increase in X_1 does not always result in an equal change in the probability that $Y=1$. This happens because $P(Y=1)$ is bounded by 0 and 1. For example, suppose again that $\beta_1>0$. If I is already high and $P(Y=1)$ is close to 1, an increase in X_1 can only have a very small effect on $P(Y=1)$.

5.1.4 Odds Ratio: For a given set of values of predictor variables X_1, ..., X_n, the odds ratio is

$$\frac{P(Y=1|I)}{1-P(Y=1|I)}$$

For example, $Y=1$ if a person owns a hybrid vehicle and $Y=0$ if she does not, the odds ratio is the probability that she owns a hybrid vehicle, divided by the probability that she does not. Note that:

$$P(Y=1|I)=\frac{1}{1+e^{-I}}$$

$$1P(Y=1|I)=1-\frac{1}{1+e^{-I}}=\frac{e^{-I}}{1+e^{-I}}$$

Hence, odds ratio $=e^{I}$, that is, $\ln\{\text{odds ratio}\}=I=\beta_0+\beta_1X_1+\ldots+\beta_nX_n$

Thus, β_i is how much the log of odds ratio changes if X_i is changed by a unit, keeping all other X's the same.

Notes on Odds-Ratio

1. In the context of betting, let $Y=1$ if one wins a bet, and $Y=0$ if she does not, that is, she loses the bet. Then, odds-ratio is $\dfrac{\text{P(Win)}}{\text{P(Lose)}}$

2. Denoting the odds ratio by R, $R=\dfrac{P(Y=1)}{P(Y=0)}=\dfrac{P(Y=1)}{1-P(Y=1)}$. From this, it follows that

$$P(Y=1)=\frac{R}{R+1}$$

3. As $P(Y=1)$ increases from 0 to 1, odds ratio R increases from 0 to ∞. If $P(Y=1)=.5$, then $R=1$.

4. Odds ratio $R=e^{\beta_0+\beta_1X_1+\ldots+\beta_nX_n}$. Thus, if X_i increases by a unit and all other $X's$ remain unchanged, odds ratio is multiplied by e^{β_i}.

5.1.5 Naive Model: The naive model for Logit is obtained by setting $\beta_1 = \ldots = \beta_n = 0$ in equation (5.1) to get:

$$(5.3) \quad I \quad = \quad \beta_0.$$

Since I is same for all values of the independent variables, $P(Y = 1)$ is always same, and given by:

$$(5.4) \quad P(Y = 1) \quad = \quad F(\beta_0) \; = \; \frac{1}{1 + e^{-\beta_0}}$$

Since $P(Y = 1)$ is same for all observations, the proportion of the sample that has $Y = 1$ is used to estimate $P(Y = 1)$. Thus, the naive Logit model corresponds to using the proportion of past observations that have $Y = 1$ to estimate the probability of having $Y = 1$ in a future observation. If that is true, then the knowledge of the independent variables does not help us determine if Y is more or less likely to be 1.

If the naive model is **not** true, that is, some of the independent variables (X's) have non-zero coefficients, then I depends on the levels of the X's. Then, we can use the knowledge of the X's to make a better prediction of whether we have $Y = 1$.

5.2 Estimation and Testing of Logit Models

5.2.1 Model Estimation: In practice, we have to estimate the parameters of the logit model, the β's, from a sample. This is done by using the **maximum likelihood technique**, that is, the set of parameter values most likely to have generated the data is chosen. To explain the logic of estimation, consider a simple model:

$I = \beta_0 + \beta_1 X$

where X is the annual income of a person (in \$1000), and $Y = 1$ if the person owns a hybrid car, and 0 if not. The data are:

Case	1	2	3	4	5	6	7	8	9	10
Y	1	1	1	1	1	0	0	0	0	0
X	120	80	160	100	90	50	80	120	100	80

For any choice of parameter values (here β_0 and β_1), we can compute $I = \beta_0 + \beta_1 X$, $P(Y = 1)$, and $1 - P(Y = 1)$. For a given case the probability that we will observe the case is $P(Y = 1)$ if $Y = 1$, and $P(Y = 0) = 1 - P(Y = 1)$ if $Y = 0$. You can easily verify that the probability that a given case i occurs is

$[P(Y_i = 1)^{Y_i} * [1 - P(Y_i = 1)]^{1-Y_i}$

Since the cases are independent, the probability of observing the sample is

$\prod_i [P(Y_i = 1)^{Y_i} * [1 - P(Y_i = 1)]^{1-Y_i}$

where the product is over all the cases in the sample. For example, suppose we try $\beta_0 = 0$ and $\beta_1 = .01$. Then, from the sample, we get:

Case	Y	X	I	P(Y=1)	P(obs)
1	1	120	1.2	0.768524783	0.768524783
2	1	80	0.8	0.689974481	0.689974481
3	1	160	1.6	0.832018385	0.832018385
4	1	100	1	0.731058579	0.731058579
5	1	90	0.9	0.710949503	0.710949503
6	0	50	0.5	0.622459331	0.377540669
7	0	80	0.8	0.689974481	0.310025519
8	0	120	1.2	0.768524783	0.231475217
9	0	100	1	0.731058579	0.268941421
10	0	80	0.8	0.689974481	0.310025519

$P(obs)$ is the probability of observing a case. The product of the extreme right column is the probability that this sample is observed if $\beta_0 = 0$ and $\beta_1 = .01$. This is called the **likelihood** of the sample. In Logit estimation, parameters are varied to maximize likelihood. As the likelihood is a very small number for a large sample, the natural logarithm of likelihood, $\ln L$ is maximized. (For the sample given, the likelihood for $\beta_0 = 0$ and $\beta_1 = .01$ is .000518, and $\ln L = -7.56552$. The maximum likelihood estimates of β_0 and β_1 are -3.58659 and .0369917, and the corresponding $\ln L = -5.948$.)

For a large sample, the parameter estimates, denoted by b_0, b_1, . . ., b_n, are approximately normally distributed unbiased estimates of the true parameter values β_0, β_1, . . ., β_n.

5.2.2 Logit Output: The Logit output provides the following information:

1. The parameter estimates, b_0, b_1, . . ., b_n. These are unbiased and approximately normally distributed.

2. The approximate standard deviations of the parameter estimates, s_{b_0}, s_{b_1}, . . ., s_{b_m}.

3. A measure of how well the parameter estimates fit the data, denoted by **log likelihood** ($\ln L$).

The log likelihood ($\ln L$) is the natural logarithm of the probability of observing the sample of observations if the parameter values are equal to the estimated values.

4. The log likelihood of the naive model.

5.2.3. Hypothesis Testing:

1. One Parameter: For a large sample, b_0, b_1, . . ., b_n are approximately normally distributed unbiased estimates of β_0, β_1, . . ., β_n. Thus, we can construct confidence intervals for, or test hypotheses about, individual parameter values using the Z table. For example:

1.(a) A 95% confidence interval of β_1 is given by $b_1 \pm 1.96 * s_{b_1}$.

1.(b) Suppose we want to test $H_0 : \beta_i \leq 0$ against $H_a : \beta_i > 0$. At a 95% level of confidence, we can reject H_0 if $b_1 > 0 + 1.645 \times s_{b_1}$.

2. Testing if One or More Parameters are all equal to Zero: Using the Logit output, we can also test the null hypothesis that any k of the β's are all equal to zero ($k \leq m$). For simplicity, suppose we wish to test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0,$$

against the alternative hypothesis that at least one of β_1, …, β_k is not zero. To perform the test at a confidence level of $(1 - \alpha)$, proceed as follows:

(1) Perform Logit analysis with the full model: Y with X_1, …, X_n. Denote the log likelihood of the full model by $\ln L_f$.

(2) Perform Logit analysis with the restricted model: Y with X_{k+1}, …, X_m (that is, drop X_1, …, X_k from the full model as predictors.). Denote the log likelihood of the restricted model by $\ln L_r$.

(3) Compute

$$(5.5) \quad \chi^2 \quad = \quad 2(\ln L_f - \ln L_r).$$

If H_0 is true, then χ^2 follows the χ^2 distribution with k degrees of freedom.

(4) At a confidence level of $(1 - \alpha)$, reject H_0 if $\chi^2 > \chi^2_\alpha$ at degrees of freedom $= k$.

5.3 Classifying a new case

Suppose you have estimated a Logit model. For simplicity assume that the parameter estimates are perfect, that is, standard errors in estimates are zeros. You now have a new case where you know the values of the independent variables X_1, …, X_m, and want to predict if Y should be 1 or 0.

- Compute $I = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$ using the parameter estimates.
- The true value of Y is either 1 or 0. If true $Y = 1$, we say that the new case belongs to group 1. If true $Y = 0$, the new case belongs to group 0.
- If the true group is 1 and we assign the new case to group 1, there is no error and hence zero loss. If the true group is 1 but we assign the case to group 0, we denote the loss from the error as $C(0|q)$ (cost of assigning to 0 given it is really 1). Similarly, if the true group is 0, there is zero loss if we assign the new case to group 0, and loss is denoted by $C(1|0)$ if we assign the case to group 0. We summarize this is as follows:

	True Group	
Assigned to	1	0
Group 1	0	$C(1|0)$
Group 0	$C(0|1)$	0

- Given I, compute $P(Y = 1)$ and $P(Y = 0) = 1 - P(Y = 1)$.
- Suppose we assigned the new case to group 1. We incur a cost of $C(1|0)$ if the true group is 0, and we incur this cost with probability $P(Y = 0) = 1 - P(Y = 1)$. Hence the expected loss from assigning the new case to group 1 is $(1 - P(Y = 1)) * C(1|0)$.

- Suppose we assigned the new case to group 0. We incur a cost of $C(0|1)$ if the true group is 1, and we incur this cost with probability $P(Y = 1)$. Hence the expected loss from assigning the new case to group 0 is $P(Y = 1) * C(0|1)$.
- We assign the new case to group 1 if the expected loss from assigning to group 1 is less than or equal to the expected loss from assigning to group 0; otherwise we assign the new case to group 0. So, we assign the new case to group 1 if and only if

  $$(1 - P(Y = 1))C(1|0) \leq P(Y = 1)C(0|1)$$

Rewriting, the new case is assigned to group 1 if and only if the odds ratio

$$(5.6) \quad \frac{P(Y = 1)}{1 - P(Y = 1)} \geq \frac{C(1|0)}{C(0|1)}$$

Noting that ln(odds Ratio) $= I$, this condition can be restated as:

Assign the new case to group 1 if and only if

$$(5.7) \quad I \geq \ln(\frac{C(1|0)}{C(0|1)})$$

If $C(1|0) = C(0|1)$, then $\ln \dfrac{C(1|0)}{C(0|1)} = 0$, and the new case is assigned to group 1 as long as $I \geq 0$, or, equivalently, $P(Y = 1|I) \geq \dfrac{1}{2}$.

5.4 Estimating Logit Models with computers

Logit from R Commander: You can estimate a Logit model using generalized linear model (glm) with link function logit in the R Commander window. The output includes:

- The parameter estimates, standard errors of parameter estimates, and P values (that is, Pr($> |z|$)) of parameter estimates. In the column for P values, you may see numbers like 2e-6; that means 2×10^{-6}.
- Null deviance, that is, $-2 \ln L$ for the naive model with $I = \beta_0$.
- Residual deviance, that is, $-2 \ln L$ for the estimated model.

To test $H_0 : \beta_1 = \ldots = \beta_k = 0$ against the alternative hypothesis that at least one of β_1, β_k is not zero:

- Estimate full model. This gives you $-2 \ln L_{full}$.
- Estimate restricted model. This gives you $-2 \ln L_{res}$.
- Compute $\chi^2 = (-2 \ln L_{res}) - (-2 \ln L_{full})$
- If $\chi^2 > \chi^2_\alpha$ at degrees of freedom k, reject H_0 at a $(1 - \alpha)$ level of confidence.

Logit in Minitab: We use the Carrier Dome Data posted on Blackboard for illustration. Suppose you want to do logit with the following dependent and independent variables:

- Dependent Variable: X_{10b} (1 if the respondent attended a football game in the Carrier Dome in the "last one year," 0 if not)
- Independent Variables: X_1 (0 if female, 1 if male); X_4 (1 if member of fratenity/sorority, 0 if not); X_{8b} (interest in participating in sports on a 1-7 scale); X_{8k} (interest in watching sports on TV on a 1-7 scale)

Proceed as follows.

(1) Open the worksheet in Minitab.

(2) In the menu line, click "Stat," drag cursor to "Regression," and click on "Binary Logistic Regression." A dialog box opens.

(3) Mark the top middle line ("Response") of the dialog box. Mark and select X_{10b} from the variable list on the left. The dependent variable X_{10b} now shows in the "Response" line.

(4) Mark the box under "Model." Mark and select X_1, X_4, X_{8b} and X_{8k} from the variable list on the left. The independent variables X_1, X_4, X_{8b}, and X_{8k} are now listed in the box under "Model."

(5) Click OK

The results are excerpted below.

Logistic Regression Table			
Predictor	Coef	SE Coef	P
Constant	-0.564538	0.654047	0.388
X_1	1.26982	0.470844	0.007
X_4	0.272569	0.374879	0.467
X_{8b}	0.0308284	0.136365	0.821
X_{8k}	0.0734185	0.136802	0.591
Log-Likelihood = -87.192			

In the above table, the P values of X_4, X_{8b} and X_{8k} are all greater than 0.10. To test if the coefficients of these three variables are all zero, we ran logit with only X_1 as the independent variable (that is, only X_1 in the "Model" box.) The results from the restricted model are as follows:

Logistic Regression Table			
Predictor	Coef	SE Coef	P
Constant	-0.0266682	0.230961	0.908
X_1	1.46515	0.376457	0.000
Log-Likelihood = -87.661			

Therefore, in this example:

- $\ln L_f = -87.192$, $\ln L_r = -87.661$, $2(\ln L_f - \ln L_r) = 0.938$
- Since $2(\ln L_f - \ln L_r) = 0.938$ does not exceed 6.25 ($\chi^2_{.10}$ at degrees of freedom 3), we cannot reject H_0 (all three coefficients are zero) at a 90% level of confidence.

Odds Ratio in Minitab Output: The Minitab output also reports an "odds ratio" for each coefficient. This is related to the odds ratio we discussed before as follows.

- Denote the estimated constant by b_0 and the estimated coefficient of X_i by b_i. The "odds ratio" reported by Mintab for b_i is e^{b_i}. This is equal to the odds ratio of $I = b_0 + b_i$ divided by the odds ratio of $I = b_0$.
- Thus, the "odds ratio" of b_i is the factor by which odds ratio increases if $X_i = 1$ instead of $X_i = 0$, keeping all other independent variables fixed.

5.5 Probit

Probit is similar to Logit and is also used to find the probability that a respondent chooses a given alternative. In binary probit, the dependent variable Y can take two values, 1 and 0. The probability that $Y = 1$ is determined by an indicator function

(5.8) $I = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n,$

where $X_1, \ldots, X_n$ are the values of the independent variables. The probability that $Y = 1$ for a given I is given by:

(5.9) $P(Y = 1|I) = P(z \le I) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I} e^{-z^2/2} dz$

That is, $P(Y = 1|I) = F(I)$, the probability that the standard normal variate z is less than or equal to I.

Example of Probit Estimation: Probit and Logit models are estimated the same way using the maximum likelihood method. We consider the same example we used with Logit in Section 5.2.1:

$I = \beta_0 + \beta_1 X$

where X is the annual income of a person (in \$1000), and $Y = 1$ if the person owns a hybrid car, and 0 if not. The data are:

Case	1	2	3	4	5	6	7	8	9	10
Y	1	1	1	1	1	0	0	0	0	0
X	120	80	160	100	90	50	80	120	100	80

Again, given parameter values, the the probability that a given case i occurs is

$[P(Y_i = 1)^{Y_i} * [1 - P(Y_i = 1)]^{1-Y_i}$

Since the cases are independent, the probability of observing the sample is

$\prod_i [P(Y_i = 1)^{Y_i} * [1 - P(Y_i = 1)]^{1-Y_i}$

where the product is over all the cases in the sample. Again, suppose we try $\beta_0 = 0$ and $\beta_1 = .01$. Then, from the sample, we get:

Case	Y	X	I	$P(z \leq I)$	$P(obs)$
1	1	120	0.53009	0.701975234	0.701975234
2	1	80	-0.40191	0.343875131	0.343875131
3	1	160	1.46209	0.928141726	0.928141726
4	1	100	0.06409	0.525550718	0.525550718
5	1	90	-0.16891	0.432933717	0.432933717
6	0	50	-1.10091	0.135467915	0.864532085
7	0	80	-0.40191	0.343875131	0.656124869
8	0	120	0.53009	0.701975234	0.298024766
9	0	100	0.06409	0.525550718	0.474449282
10	0	80	-0.40191	0.343875131	0.656124869

For these parameter values, likelihood of the sample is .000114, and $\ln L = -9.0757$. The maximum likelihood estimates of β_0 and β_1 are -2.26591 and .0233, respectively, and the corresponding $\ln L = -5.921$.

Assigning a new case to a group: Proceeding as in Section 5.3, we first compute I and $P(Y = 1|I)$, and assign the new case to group 1 if and only if

$$(5.10) \quad \text{Odds Ratio} = \frac{P(Y = 1|I)}{1 - P(Y = 1|I)} \geq \frac{C(1|0)}{C(0|1)}$$

Output from R and Hypothesis Testing: You can estimate a Probit model using generalized linear model (glm) with link function probit in the R Commander window. The output includes:

- The parameter estimates, standard errors of parameter estimates, and P values (that is, $\Pr(>|z|)$) of parameter estimates.
- Null deviance, that is, $-2\ln L$ for the naive model with $I = \beta_0$.
- Residual deviance, that is, $-2\ln L$ for the estimated model.

To test $H_0 : \beta_1 = \ldots = \beta_k = 0$ against the alternative hypothesis that at least one of β_1, β_k is not zero:

- Estimate full model. This gives you $-2\ln L_{full}$.
- Estimate restricted model. This gives you $-2\ln L_{res}$.
- Compute $\chi^2 = (-2\ln L_{res}) - (-2\ln L_{full})$
- If $\chi^2 > \chi^2_\alpha$ at degrees of freedom k, reject H_0 at a $(1-\alpha)$ level of confidence.

Difference between Logit and Probit: In both Logit and Probit, $P(Y = 1)$ is determined by I. The difference is, in Logit,

$$P(Y = 1|I) = \frac{1}{1 + e^{-I}}$$

In Probit, $P(Y = 1|I) = P(z \leq I)$

Both the logistic distribution and the standard normal distribution (z) have mean zero and are symmetric around zero. However, the standard normal distribution drops more sharply on both sides of zero than the logistic distribution. As a result, parameter estimates in Probit are more sensitive to outlier observation. Logit estimation is more robust to outliers.

5.6 Multinomial Logit

In binary logit where the dependent variable can take one of two values. Multinomial logit extends binary logit by allowing the dependent variable to have more than two values. I briefly summarize the multinomial logit model here. We can summarize multinomial logit as follows.

Model

1. You want to find the probability that one of $(n+1)$ events, 0, 1, ..., n occurs, given the values of m predictor variables X_1, ..., X_m.

2. For each event j, there is an indicator function I_j where $I_j = \alpha V_j$.

3. $I_0 = 0$, and for $j = 1, \ldots, n$,

$$I_j = \beta_{0j} + (\beta_{1j} * X_1) + \ldots + (\beta_{mj} * X_m)$$

Note: The weights β's are allowed to be different for the different events.

4. Probability of event j is

$$P_j \quad = \quad \frac{e^{I_j}}{\sum_{k=1}^{n+1} e^{I_k}}$$

Therefore:

$$P_0 = \frac{1}{1 + e^{I_1} + \ldots + e^{I_m}}$$

For $j = 1, \ldots, n$,

$$P_j = \frac{e^{I_j}}{1 + e^{I_1} + \ldots + e^{I_m}}$$

Challenge in Estimation

- A common application of multinomial logit is to find the probability that a given brand is selected by a customer.
- In a given choice occasion, not all brands may be available to the customer. Then, the denominator in the probability expression should only include the brands available. This makes estimation difficult.

- The multinomial option included in R Commander assumes that all alternatives are available in all choice occasions.

Multinomial Logit in R Commander

- In R Commander, click Statistics ⟶ Fit Models ⟶ Multinomial Logit Model
- Select dependent and independent variables as you did for binary logit and probit.
- The estimation assumes that all alternatives are available in all choice occasions.
- The output provides estimates of β's, standard errors of estimates, and residual deviance. You can test linear hypotheses on parameters using Models as you did for logit.

Example of Multinomial Logit in R Commander We used the dataset ojdominicks.csv and estimated the multinomial logit model using BRAND (which can be HH, Minutemaid, or Tropicana) as dependent variable, and Feat (1 or 0) and price as independent variables. The output from R Commander are given below.

Output

Coefficients			
	(Intercept)	Feat	PRICE
MINUTEMAID	−7.649136	1.000171	3.734074
TROPICANA	−16.560149	2.596081	7.177186
Standard Errors			
	(Intercept)	Feat	PRICE
MINUTEMAID	0.1118496	0.03728085	0.05317353
TROPICANA	0.1652454	0.04948573	0.07063297
Value/SE (Wald statistics)			
	(Intercept)	Feat	PRICE
MINUTEMAID	−68.38767	26.8280	70.22431
TROPICANA	−100.21547	52.4612	101.61241

Residual Deviance: 48565.06, AIC: 48577.06

Note

1. In the output, HH is the default choice with $I = 0$.

For Minutemaid, $I = -7.649136 + (1.000171 * \text{Feat}) + (3.734074 * \text{Price})$

For Tropicana, $I = -16.560149 + (2.596081 * \text{Feat}) + (7.177186 * \text{Price})$

2. R Commander does not provide null deviance for multinom. However, you can use Models to test linear hypotheses on parameters. The χ^2 statistic for "Compare Two Models" and "Linear Hypothesis" are similar but not same.

3. Wald statistics are estimated coefficients divided by standard errors and approximately follow the standard normal distributions (z) for large samples. Therefore, $P < .05$ if the magnitude of the Wald statistic is greater than 1.96, and $P < .01$ if the magnitude of the Wald statistic is greater than 2.575.

Numerical example: Suppose Feat = 1 and price = 2.7.

Then:

$I_{HH} = 0$

$I_{Minutemaid} = -7.649136 + (1.000171 * 1) + (3.734074 * 2.7) = 3.4330348$

$I_{Tropicana} = -16.560149 + (2.596081 * 1) + (7.177186 * 2.7) = 5.4143342$

Probabilities:

$$P_{HH} = \frac{1}{1 + e^{3.4330348} + e^{5.4143342}} = 0.0039$$

$$P_{Minutemaid} = \frac{e^{3.4330348}}{1 + e^{3.4330348} + e^{5.4143342}} = 0.1207$$

$$P_{Tropicana} = \frac{e^{5.4143342}}{1 + e^{3.4330348} + e^{5.4143342}} = 0.8754$$

Verification:

- In the sample, I applied the filters Feat = 1 and price between 2.6 and 2.8.
- This gives me 530 cases out of which 77 are Minutemaid (14.53%) and 453 are Tropicana (85.47%).
- Thus, for Tropicana and Minutemaid, the estimated probabilities closely match the proportions of times the two brands are present in the sample.

5.7 Exercise Problems

1. **Scenario:** Suppose you have conducted logit analysis on data collected from the students of a college campus to determine how the probability that the student is member of a student organization (such as the AMA) depends on the gender of the student and the class the student is in. Suppose we have used a sample of size 500 with the indicator function given as follows:

$$I = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3$$

where:
$D_1 = 1$ if the student is a freshman, 0 if not,
$D_2 = 1$ if the student is a sophomore, 0 if not,
$D_3 = 1$ if the student is female, 0 if not.

Results:

Estimated Coefficient	Standard Deviation of Estimate
$b_0 = -2$	$s_{b_0} = .2$
$b_1 = -.6$	$s_{b_1} = .3$
$b_2 = -.4$	$s_{b_2} = .3$
$b_3 = 1.0$ $\ln L = -540$	$s_{b_3} = .4$

1.(a) Interpret in words the meaning of the following null hypothesis: $H_0 : \beta_1 = \beta_2$.

1.(b) Suppose the student population at the campus has the following composition:

	Men	Women
Freshmen	1800	1200
Sophomores	1500	1500
Juniors & Seniors	3300	2700

Based on the results of the logit analysis, estimate how many students at the campus are members of student organizations.

1.(c) Suppose we have run logit with D_3 as the only independent variable and obtained a log likelihood of -560. At a 99% level of confidence, test $H_0 : \beta_1 = \beta_2 = 0$. Interpret H_0.

2. **Problem Scenario:** Y is a binary variable defined as follows: $Y = 1$ if a person owns a digital camcorder, and $Y = 0$ if not. Let X be the income of the person (unit = \$10,000), and let $D = 1$ if the person has children, and $D = 0$ if not. Suppose Y is related to X and D according to the logit model:

$$P(Y = 1|I) = \frac{1}{1 + e^{-I}}, \quad \text{where} \quad I = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 a * D * X.$$

2.(a) Suppose we know that $\beta_0 = -2$, $\beta_1 = .1$, $\beta_2 = .5$, and $\beta_3 = .02$. If a person earns \$60,000 and has children, what is the probability that (s)he owns a digital camcorder?

2.(b) Suppose we estimated several logit models and obtained the following results:

	Dependent Variable	Independent Variables	$\ln L$
Model 1	Y	X, D, $D * X$	-150
Model 2	Y	X	-165
Model 3	Y	D	-170
Model 4	Y	X, D	-152
Model 4	Y	X, $D * X$	-160
Model 5	Y	D, $D * X$	-166
Model 6 (Naive Model)	Y	____	-180

2.(a) At a 99% level of confidence, test $H_0 : \beta_1 = \beta_3 = 0$ against H_a : at least one of β_1 and β_3 is not zero. What does the null hypothesis mean?

2.(b) At a 99% level of confidence, test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against H_a : at least one of β_1, β_2, and β_3 is not zero.

3. Let Y be a dichotomous variable representing whether a person is married or not, and x the person's age in years. Let y depend on x according to the LOGIT model with the indicator function, I, given by:

$I = -3 + 0.1 * X$

Suppose a random sample of six individuals have ages (in years) 29, 33, 42, 24, 36, and 39. Compute the probability that these six individuals are **all** married.

[Hint: We are assuming that the six cases are independent. First compute the probability for each individual. The probability that they are all married is the product of these six probabilities.]

4. Consider a simpler version of the model given above where Y and X are as defined in problem 1, and,

$$P(Y = 1|I) = \frac{1}{1 + e^{-I}}, \quad \text{where} \quad I = \beta_0 + \beta_1 X.$$

Suppose you have the following data:

Case	Y	X
1	1	8
2	1	5
3	1	6
4	1	7.5
5	0	6
6	0	4
7	0	5
8	0	3

Which of the following two sets of parameter values gives you a better fit with the data?

Set 1: $\beta_0 = -1$, $\beta_1 = .1$

Set 2: $\beta_0 = 1$, $\beta_1 = .1$.

Hint:

- For each set of parameters, you need to compute log likelihood. The set with greater log likelihood gives the better fit.
- For a given set of parameters, for each case, you need to compute the probability that you will get that observation ($P(obs)$) for that parameter set. Compute I and $P(Y = 1|I)$. If $Y = 1$, then $P(obs) = P(Y = 1|I)$. If $Y = 0$, then $P(obs) = 1 - P(Y = 1|I)$.
- For a given set of parameters, first compute $P(obs)$ for each case. Using that, compute $\ln P(obs)$ for each case and add over the observations. This is the log likelihood of the data set for that set of parameters.
- You may to use Excel to do this.

6 Principal Components and Cluster Analysis

6.1 Basics

6.1.1 Sample Mean and Sample Standard Deviation

Suppose we have drawn a sample of size n from a population and measured a variable X for each member of the sample. Let X_1, X_2, ..., X_n denote the measured values of X for first, second, ..., n-th member of the sample. The sample mean $(\overline{X})$ and the sample standard deviation (s) are defined as:

$$(6.1) \quad \overline{X} \quad = \quad \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

$$(6.2) \quad s \quad = \quad \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}}.$$

- Suppose we created a new variable Y such that $Y_i = X_i + a$, where a is a constant, for all members of the sample. Then, $\overline{Y} = \overline{X} + a$, and the standard deviation of Y is s.

- Suppose we created a new variable Y such that $Y_i = a * X_i$, where $a > 0$ is a constant, for all members of the sample. Then, the sample mean of Y is $a\overline{X}$, and the sample standard deviation of Y is $a * s$.

- If we create a new variable v such that $v_i = \dfrac{X_i - \overline{X}}{s}$ for each member of the sample, then $\overline{v} = 0$, and the sample standard deviation of v is 1.

  v is called a **standardized variable** obtained from X.

6.1.2 Correlation

Suppose we have a sample of size n and measured two variables X and Y from each member of the sample. Let X_i and Y_i denote the measured values of X and Y for member i of the sample. Then the Pearson correlation, or simply **correlation** of X and Y for the sample, denoted by $r(X, Y)$ or simply r, is given by:

$$(6.3) \quad r(X,Y) \quad = \quad \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

where $\overline{X}$ is the sample mean of X, and $\overline{Y}$ is the sample mean of Y.

- If $u = aX + b$ and $v = cY + d$ where $a > 0$, b, $c > 0$ and d are constant numbers, then $r(u, v) = r(X, Y)$. Specifically, if we standardize X and Y, the correlation remains the same.

- The correlation of a variable with itself is 1, that is, $r(X, X) = 1$.

- For any two variables X and Y, $r(X, Y) = r(Y, X)$.

- If X and Y are standardized, their covariance is equal to their correlation.
- The correlation r can never be less than -1 or more than $+1$.
- If $r = +1$, then we say that X and Y are **perfectly correlated** in the sample. In that case, the graph of Y against X is a straight line sloping upwards.
- If $r = -1$, then X and Y are **perfectly negatively correlated**, and the graph of Y against X is a straight line sloping downwards.
- If $r^2 = 1$, then the plot of Y against X is a straight line.

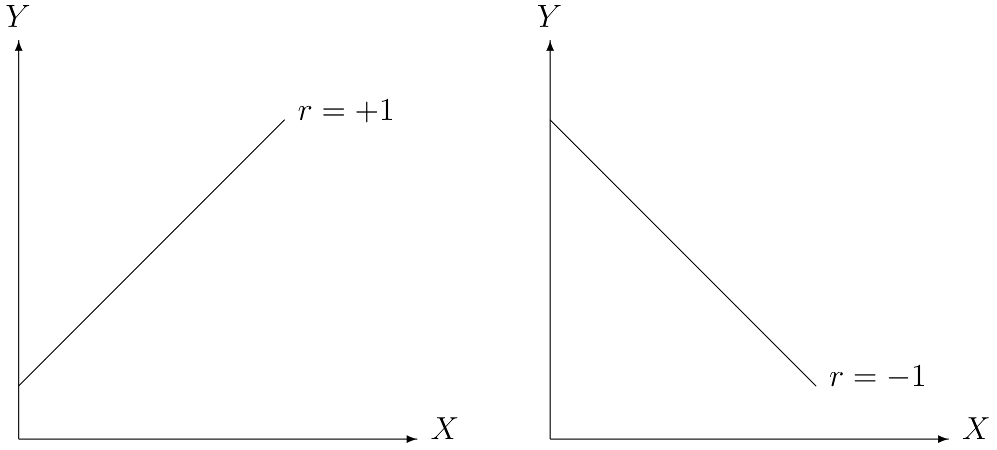


Figure 6.1

Strictly speaking, $r(X, Y)$ should be computed when X and Y both satisfy interval scale properties. In practice, we relax it to some extent. For example, we routinely compute correlations of variables measured with five and seven point scales. The error is small.

Interpretation of Positive and Negative Correlations: In the right hand side of Equation (6.3), the denominator is always positive. Thus, the sign of $r(X, Y)$ is same as the sign of the numerator.

Positive Correlation ($r(X, Y) > 0$): In this case, changes in X and Y tend to go in the same direction. Thus, if X_i exceeds $\overline{X}$, then Y_i tends to exceed $\overline{Y}$. Since $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$ are both positive, the product $(X_i - \overline{X})(Y_i - \overline{Y})$ is positive.

Similarly, if $X_i < \overline{X}$, then Y_i tends to be less than $\overline{Y}$. Hence the product $(X_i - \overline{X})(Y_i - \overline{Y})$, being the product of two negative terms, is positive.

Since $(X_i - \overline{X})(Y_i - \overline{Y})$ tends to be positive in all cases, the numerator in the right hand side of Equation (6.3) is positive.

Negative Correlation ($r(X, Y) < 0$): In this case, the changes in X and Y tend to go in opposite directions. Thus, if $X_i > \overline{X}$, we tend to have $Y_i < \overline{Y}$, and vice versa. As $(X_i - \overline{X})$ and

$(Y_i - \overline{Y})$ tend to have opposite signs, their product tends to be negative. Hence, the numerator in the right hand side of Equation (6.3) is negative.

Zero Correlation $(r(X, Y) = 0)$: In this case, if $X_i > \overline{X}$, then Y_i is equally likely to be above or below $\overline{Y}$, and vice versa.

Meaning of P Value: Let $\rho(X, Y)$ denote the correlation of X and Y computed over the whole population the sample is drawn from. Even when $\rho(X, Y) = 0$, the correlation from the sample, $r(X, Y)$, can be more than zero or less than zero. When you compute $r(X, Y)$ using a statistical package such as Minitab, the output gives you the "P value" of the null hypothesis that $\rho(X, Y) = 0$.

Formally, the P value is the probability that if $\rho(X, Y) = 0$, then the correlation of X and Y from the sample will exceed $|r(X, Y)|$, or be less than $-|r(X, Y)|$. A small P value means it is unlikely to get such data if $\rho(X, Y)$ is zero. When P is adequately small, we conclude that $\rho(X, Y)$ is not really zero, and the correlation obtained from the sample is **significant**. In practice, we usually consider a correlation to be significant if the P value is 0.1 or less.

6.2 Principal Components Analysis

6.2.1 Idea and Method: Suppose we measured m interval scaled variables $X_1, \ldots, X_m$ from each member of a sample. Principal components analysis is a method of data reduction by identifying a relatively small number of new variables that capture most of the variation in the original m variables. The procedure is outlined below. The method employs matrix algebra that is described in an optional appendix.

Steps:

1. A $m \times m$ correlation matrix R is computed so that the entry (i, j) of the matrix is the correlation between X_i and X_j. This matrix is same as the covariance matrix of the standardized variables $v_1, \ldots, v_m$ obtained from $X_1, \ldots, X_m$.

Since the correlation between a variable and itself is 1, the matrix R has a diagonal of 1's. Thus, the sum of the diagonal terms in R (that is, trace of R) is m, which is the sum of the variances of the m standardized variables.

2. Principle components analysis next finds the m eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$, and the corresponding m normalized eigenvectors of R, which we call $\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_m$. Each eigenvector $\vec{w}_i$ is a vector of m numbers, $\{w_{1i}, w_{2i}, \ldots, w_{mi}\}$.

[The term "normalized" means for each eignevector $\vec{w}_i$, the sum of squares of the m numbers in the vector is 1. We can do this without loss of generality. Any eignevector $\vec{w}_i$ for eigenvalue λ_i remains an eigenvector with the same eigenvalue if we multiply each term by the same constant number, positive or negative. Thus, if $\vec{u}_i$ is an eigenvector for eigenvalue λ_i, we get a normalized eigenvector $\vec{w}_i$ by dividing each term $\vec{u}_i$ by $\sqrt{u_{1i}^2 + u_{2i}^2 + \ldots + u_{mi}^2}$.]

3. Remember that $v_1, \ldots, v_m$ are the m original variables in standardized form. For each eigenvector i, the weighted sum

$w_{1i}v_1 + w_{2i}v_2 + \ldots + w_{mi}v_m$

is a new variable that is also a weighted sum of the original m variables $X_1, \ldots, X_m$. The m new variables we get from the m eigenvectors this way are called **principal components** and have the following properties:

- The principal components are uncorrelated with one another for the sample.
- The variance of a principal component is equal to the corresponding eigenvalue, which is a real, nonnegative number.
- The m eigenvalues add up to m, that is, together, the m principal components explain all the variance in the original m variables.

4. We arrange the eigenvalues from largest to smallest. The first principal component, corresponding to the largest eigenvalue, explains most variance in the original m variables. The second principal component, which is uncorrelated to the first principal component, explains the next most variance, and so on.

5. **Retaining principal components:** The next task is to determine the number of components we need to keep. In practice, this is done in one of two ways.

- Keep only the principal components that have eigenvalues of 1 or more, that is, which explain at least as much variance as one original variable.
- Plot the eigenvalues. There is usually a bend, called a “knee,” in the plot. Keep only the components with eigenvalues at or above the knee (Figure 6.2).

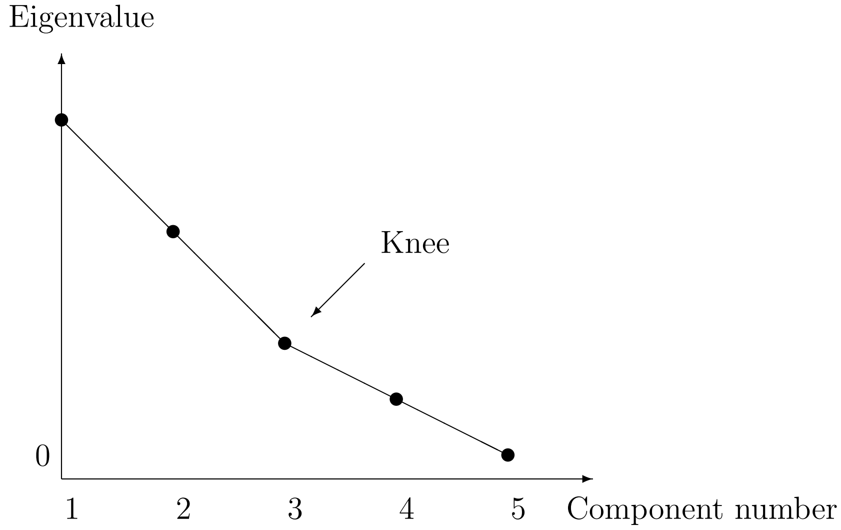


Figure 6.2

6. **Factor Rotation:** The principal components are also called factors. After we decide how many components to retain, the retained factors are “rotated.” These rotation creates “rotated factors” with the following properties.

- The number of rotated factors is same as the number of components retained.
- The rotated factors are uncorrelated with one another.
- Together, the rotated factors explain the same total variance as the retained factors.
- In a commonly used rotation method called varimax, the rotation is done such that each rotated factor has either a very high or very low squared correlation with each original variable. This makes it easy to interpret the rotated variables.

7. We now use the rotated factors instead of the original variables.

8. **Interpretation of Factors Retained:** Each factor retained is a weighted sum of the original variables X_1, …, X_m. Compute the correlations of each factor f_i with X_1, …, X_n. These correlations are called **factor loadings.** Ignore any correlation with magnitude less than 0.5. This way, you usually identify a small number of original variables that are strongly related to the factor. Interpret the factor in terms of variables it is related strongly to.

6.3 Cluster Analysis

Suppose we measured n objects on m variables. We can represent each object as a point in an m dimensional space where the coordinates of each object are the measured values of the m variables. Given these coordinates, cluster analysis tries to create groups such that members of each group are located close to one another in this space. Each group is called a cluster, and each cluster consists of objects with similar values of the variables recorded. In a widely used clustering method called K-Means Cluster Analysis, the researcher specifies the number of clusters to use. Given that number of clusters, the program assigns the objects to clusters so that "within cluster sum of squares" is minimized.

To explain within cluster sum of squares, consider a sample of n objects where each object is measured on two variables X_1 and X_2. In the two dimensional plot with orthogonal axes X_1 and X_2, the coordinates of object i in the sample are X_{i1} and X_{i2}. Suppose the n objects are assigned to G groups.

- Consider any group g. Let n_g denote the number of objects in group g. For the group g, the averages of X_1 and X_2 for the members of group g are given by

  $\overline{m}_{g1} = \frac{\sum_{i \in g} X_{i1}}{n_g}$ and $\overline{m}_{g2} = \frac{\sum_{i \in g} X_{i2}}{n_g}$, respectively.

  Call the point $(\overline{m}_{g1}, \overline{m}_{g2})$ the centroid of group g.

- For each object i in the sample, compute its squared distance from the centroid of the group it belongs to. For example, if object i belongs to group g, the squared distance is

  $(X_{i1} - \overline{m}_{g1})^2 + (X_{i2} - \overline{m}_{g2})^2$

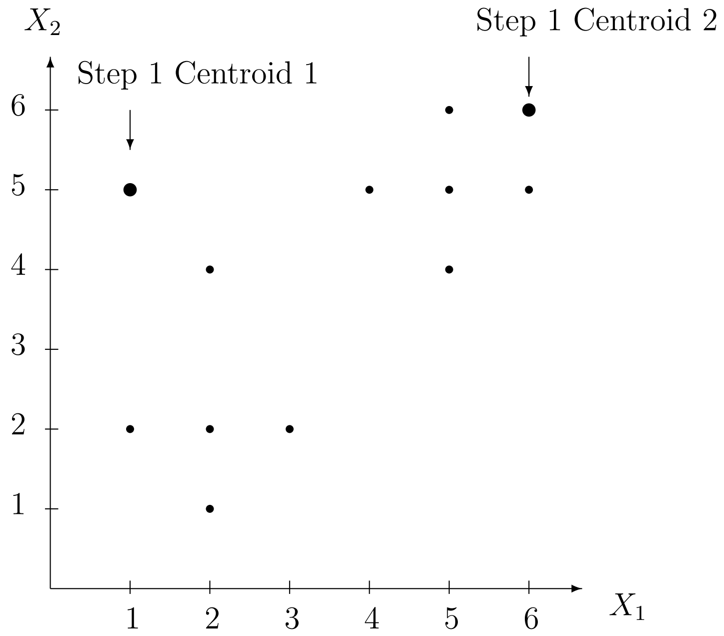
- Add the squared distances for the n objects. This is the within cluster sum of squares.

K Means Cluster Analysis proceeds as follows:

- Given G, the number of clusters specified, randomly select G distinct points as group centroids.
- For each object, assign it to the group with centroid nearest to the object.
- After all n objects are assigned, compute the centroids of the groups. With these new centroids, again assign each object in the sample to a group that has a centroid nearest to the object.
- Repeat the process until there is almost no change in centroids from one iteration to another.

Illustrative Example: Suppose you measured ten objects on two variables X_1 and X_2. You wish to create two groups from the data using K-Means Cluster Analysis.

Object	1	2	3	4	5	6	7	8	9	10
X_1	1	5	4	5	6	5	2	2	3	2
X_2	2	4	5	6	5	5	1	2	2	4



Step 1. Select two points in the (X_1, X_2) space as centroids. This choice is usually arbitrary, and may not actually be the centroid of any group of points in the data. Here, we start with the two centroids $(1, 5)$ and $(6, 6)$. With this choice, we get:

Object	X_1	X_2	Distance from Centroid 1	Distance from Centroid 2	Assign to Group
1	1	2	$\sqrt{(1-1)^2+(5-2)^2}=3$	$\sqrt{(6-1)^2+(6-2)^2}=\sqrt{41}$	1
2	5	4	$\sqrt{(1-5)^2+(5-4)^2}=\sqrt{17}$	$\sqrt{(6-5)^2+(6-4)^2}=\sqrt{5}$	2
3	4	5	$\sqrt{(1-4)^2+(5-5)^2}=3$	$\sqrt{(6-4)^2+(6-5)^2}=\sqrt{5}$	2
4	5	6	$\sqrt{(1-5)^2+(5-6)^2}=\sqrt{17}$	$\sqrt{(6-5)^2+(6-6)^2}=1$	2
5	6	5	$\sqrt{(1-6)^2+(5-5)^2}=5$	$\sqrt{(6-6)^2+(6-5)^2}=1$	2
6	5	5	$\sqrt{(1-5)^2+(5-5)^2}=4$	$\sqrt{(6-5)^2+(6-5)^2}=\sqrt{2}$	2
7	2	1	$\sqrt{(1-2)^2+(5-1)^2}=\sqrt{17}$	$\sqrt{(6-2)^2+(6-1)^2}=\sqrt{41}$	1
8	2	2	$\sqrt{(1-2)^2+(5-2)^2}=\sqrt{10}$	$\sqrt{(6-2)^2+(6-2)^2}=\sqrt{32}$	1
9	3	2	$\sqrt{(1-3)^2+(5-2)^2}=\sqrt{13}$	$\sqrt{(6-3)^2+(6-2)^2}=5$	1
10	2	4	$\sqrt{(1-2)^2+(5-4)^2}=\sqrt{2}$	$\sqrt{(6-2)^2+(6-4)^2}=\sqrt{20}$	1

Assign $(1,2)$, $(2,1)$, $(2,2)$, $(3,2)$ and $(2,4)$ to group 1. New centroid of group 1 is $(\frac{1+2+2+3+2}{5}, \frac{(2+1+2+2+4}{5})$, that is, $(2,2.2)$

Assign $(5,4)$, $(4,5)$, $(5,6)$, $(6,5)$ and $(5,5)$ to group 2. New centroid of group 2 is $(\frac{5+4+5+6+5}{5}, \frac{4+5+6+5+5}{5})$, that is, $(5,5)$.

Step 2. With the new centroids $(2,2.2)$ and $(5,5)$, again compute distances from the ten objects are assign to group with nearest centroid.

Object	X_1	X_2	Distance from Centroid 1	Distance from Centroid 2	Assign to Group
1	1	2	1.02	5	1
2	5	4	3.50	1	2
3	4	5	3.44	1	2
4	5	6	4.84	1	2
5	6	5	4.88	1	2
6	5	5	4.10	0	2
7	2	1	1.2	5	1
8	2	2	0.2	4.25	1
9	3	2	1.02	3.61	1
10	2	4	1.80	3.16	1

Since the assignment remain same, centroids also remain same. So, the final clusters are:

Cluster 1: Objects 1, 7, 8, 9 and 10.

Cluster 2: Objects 2, 3, 4, 5 and 6.

6.4 Example: Psychographic segmentation using student data

These data are excerpted from the Carrier Dome data, collected from 148 Syracuse University Undergraduate students and posted on Blackboard.

---

1. Gender ____Male ____Female

2. How interested are you in the following activities in your spare time? (please circle) (1: not interested at all, 7: very interested

	Not interested at all						**Very interested**
(a) Exercise	1	2	3	4	5	6	7
(b) Participate in sports	1	2	3	4	5	6	7
(c) Shop for clothes	1	2	3	4	5	6	7
(d) Go to bars	1	2	3	4	5	6	7
(e) Go to malls	1	2	3	4	5	6	7
(f) Watch movie	1	2	3	4	5	6	7
(g) Do volunteer work	1	2	3	4	5	6	7
(h) Study/read	1	2	3	4	5	6	7
(i) Listen to music	1	2	3	4	5	6	7
(j) Spend time with friends	1	2	3	4	5	6	7
(k) Watch sports on TV	1	2	3	4	5	6	7
(l) Watch sports at the Dome	1	2	3	4	5	6	7

3. If a **combined lacrosse/football/basketball** season ticket were available, how likely would you be to purchase it at $150?

Very unlikely						Very likely
1	2	3	4	5	6	7

The items in question 2 represent lifestyle variables (activities, interests, opinions) of the students. Each question 2(a)-2(l) is called an "item." First, we use principal components analysis to find the key underlying dimensions of the items 2(a)-2(l). Four dimensions (factors) emerge. The correlations of these four factors (Factor1-Factor4) with the original variables ($X_{2a} - X_{2l}$) are given below.

Variable	Factor1	Factor2	Factor3	Factor4
X_{2a}	**0.681**	0.059	−0.103	−0.328
X_{2b}	**0.869**	−0.152	−0.070	0.005
X_{2c}	−0.223	**0.811**	0.178	−0.075
X_{2d}	0.130	**0.516**	−0.519	−0.088
X_{2e}	−0.167	**0.801**	0.144	0.027
X_{2f}	0.172	**0.547**	−0.060	0.452
X_{2g}	0.016	0.121	**0.811**	0.002
X_{2h}	0.038	0.135	**0.691**	−0.254
X_{2i}	0.073	0.092	−0.150	**0.760**
X_{2j}	−0.055	**0.565**	0.006	0.384
X_{2k}	**0.725**	−0.285	−0.057	0.369
X_{2l}	**0.754**	0.008	0.248	0.247

Each factor is standardized, that is, it has mean zero and standard deviation 1 for the sample of 148 students. Together, these four factors explain 63.8% of the variance of the 12 original variables. To interpret the factors, only look at correlations above 0.5 in magnitude. That way:

Factor 1: Interest in watching and participating in sports.

Factor 2: Interest in shopping, bars, and malls.

Factor 3: Interest in studying and volunteer work, together with aversion to bars.

Factor 4: Interest in music.

Step 2: We use K-Means clustering to create groups of respondents with similar factor scores. There are four clusters. The average factor score of each cluster is given below. The last row of the table gives the number of respondents in each cluster.

Factor	Cluster1	Cluster2	Cluster3	Cluster4
Factor1	0.8194	−1.1010	−0.0304	0.0255
Factor2	0.9285	0.1413	−0.0869	−0.8627
Factor3	−0.3284	−0.0403	1.0193	−0.5520
Factor4	0.0721	1.0335	−0.2517	−0.5458
Size	39 (26.35%)	29 (19.59%)	37 (25%)	43 (29.05%)

The four clusters are four psychographic segments with the following characteristics:

(1) Cluster 1: Interested in sports. Also interested in shopping for clothes, and visiting bars and malls. Not interested in studies, or volunteer work.

(2) Cluster 2: Strong interest in music. Not interested in sports. Mildly interested in socializing (shopping for clothes, bars, malls, spending time with friends).

(3) Cluster 3: Strong interest in studies and volunteer work. Slightly below average interest in sports and socializing. Below average interest in music.

(4) Cluster 4: Not interested in socializing (shopping, bars, malls, spending time with friends),

studies, volunteer work, or music. Mildly interested in sports.

Once we have the segments, we can try to find who belongs to each cluster. For example, we created the following cross tabulation of gender with cluster:

	Cluster			
Gender	1	2	3	4
Women	15 (20.00%)	22 (29.33%)	22 (29.33%)	16 (21.33%)
Men	24 (32.88%)	7 (9.59%)	15 (20.55%)	27 (36.99%)

We can also check if some behavior or purchase intentions differ from cluster to cluster. Consider question 3, which is the likelihood of purchasing the combined ticket for \$150. Let Interest = 1 if the respondent circled 5-7, and 0 if the respondent circled 1-4. The cross-tabulation of Interest with Cluster is given below:

	Interest	
Cluster	0	1
1	27 (69.23%)	12 (30.77%)
2	19 (65.52%)	10 (34.48%)
3	16 (44.44%)	20 (55.56%)
4	15 (34.88%)	28 (65.12%)

Interesting, cluster 4, which is not interested in shopping, bars, malls, or participating in sports, are most likely to buy the ticket for \$150. (There is one missing case in Cluster 3.)

6.5 Additional Reading

Johannes Ledolter, Data Mining and Business Analytics in R, Wiley, 2013.

7 Solutions to Exercise Problems

7.1 Chapter 1 Exercise Problems and Solutions

1. We have drawn a simple random sample of size 40 from the students of a university, and asked them two questions:

(i) How many hours do you work (on a job) each week?

(ii) On the average, how many dollars do you spend on fast-food each week?

From the data, we obtained the following cross-tabulation:

	Number of hours one works			
Expenditure on Fast Food	0	1–10	11–20	over 20
\$ 10 or less	2	2	0	0
\$ 11 - \$ 20	8	8	2	2
Over \$20	2	4	8	2

At a 95% level of confidence, test the null hypothesis that the number of hours a student works is not related to expenditure on fast-food.

Answer: Augmenting the table by including row totals and column totals, we have:

	Number of hours one works				
Expenditure on Fast Food	0	1–10	11–20	over 20	**Row Totals**
\$ 10 or less	2	2	0	0	**4**
\$ 11 - \$ 20	8	8	2	2	**20**
Over \$20	2	4	8	2	**16**
Column Totals	**12**	**14**	**10**	**4**	$n = 40$

$$E_{11} = \frac{4 \times 12}{40} = 1.2 \quad E_{12} = \frac{4 \times 14}{40} = 1.4 \quad E_{13} = \frac{4 \times 10}{40} = 1 \quad E_{14} = \frac{4 \times 4}{40} = 0.4$$

$$E_{21} = \frac{20 \times 12}{40} = 6 \quad E_{22} = \frac{20 \times 14}{40} = 7 \quad E_{23} = \frac{20 \times 10}{40} = 5 \quad E_{24} = \frac{20 \times 4}{40} = 2$$

$$E_{31} = \frac{16 \times 12}{40} = 4.8 \quad E_{32} = \frac{16 \times 14}{40} = 5.6 \quad E_{33} = \frac{16 \times 10}{40} = 4 \quad E_{34} = \frac{16 \times 4}{40} = 1.6$$

Since E_{13} and E_{14} do not exceed 1, we can immediately tell that the original cross-tabulation should not be used to perform chi-square analysis. If we combine rows 1 and 2, and columns 3 and 4, then we get a new 2×3 cross-tabulation where $E_{ij} > 1$ in all cells and $E_{ij} \geq 5$ in 5 out of 6 cells. Hence, the chi-square test is valid with the new cross-tabulation. The new table with expected frequencies in brackets is given below:

	Number of hours one works		
Expenditure on Fast Food	0	1–10	11 or more
\$ 20 or less	10 (7.2)	10 (8.4)	4 (8.4)
Over \$20	2 (4.8)	4 (5.6)	10 (5.6)

$$\chi^2 = \frac{(10-7.2)^2}{7.2} + \frac{(10-8.4)^2}{8.4} + \frac{(4-8.4)^2}{8.4} + \frac{(2-4.8)^2}{4.8} + \frac{(4-5.6)^2}{5.6} + \frac{(10-5.6)^2}{5.6} \;=\; 9.246$$

Decision Rule: At a 95% level of confidence, reject H_0 if $\chi^2 > 5.99 = \chi^2_{.05}$ at degrees of freedom $(2-1) \times (3-1) = 2$.

Conclusion: Here, $\chi^2 = 9.246 > 5.99$. Hence we reject H_0 at a 95% level of confidence.

2. We have selected a simple random sample of 50 students at a university and recorded whether the student lives in the dorm, and how much the student spends in a month eating out. From the data, we obtained the following cross-tabulation:

	Expenditure/month (in dollars)		
Residence	0–20	21–40	> 40
Dorm	10	9	1
Not Dorm	5	13	12

At a 95% level of confidence, test the null hypothesis that the expenditure eating out is not related to whether the student lives in a dorm or not.

Answer: Augmenting the table by row totals and column totals, we have:

	Expenditure/month (in dollars)			
Residence	0–20	21–40	> 40	**Row Totals**
Dorm	10	9	1	**20**
Not Dorm	5	13	12	**30**
Column Totals	**15**	**22**	**13**	$n = 50$

$$E_{11} = \frac{20 \times 15}{50} = 6 \quad E_{12} = \frac{20 \times 22}{50} = 8.8 \quad E_{13} = \frac{20 \times 13}{50} = 5.2$$

$$E_{21} = \frac{30 \times 15}{50} = 9 \quad E_{22} = \frac{30 \times 22}{50} = 13.2 \quad E_{23} = \frac{30 \times 13}{50} = 7.8$$

Since all expected frequencies exceed 5, chi-square test is valid.

$$\chi^2 = \frac{(10-6)^2}{6} + \frac{(9-8.8)^2}{8.8} + \frac{(1-5.2)^2}{5.2} + \frac{(5-9)^2}{9} + \frac{(13-13.2)^2}{13.2} + \frac{(12-7.8)^2}{7.8} \;=\; 10.11$$

Decision Rule: At a 95% level of confidence, reject H_0 if $\chi^2 > 5.99 = \chi^2_{.05}$ at degrees of freedom $(2-1) \times (3-1) = 2$.

Conclusion: Here, $\chi^2 = 10.11 > 5.99$. Hence we reject H_0 at a 95% level of confidence.

3. Suppose you have drawn a simple random sample from a university and asked each respondent if (s)he owned an Apple iPhone. There are four categories within the university population: undergraduate students, graduate students, faculty and staff. The data, separated by category, are given below:

Category	Number drawn from category	Number that own Apple iPhones
1. Undergraduate Student	90	60
2. Graduate Student	40	20
3. Faculty	40	15
4. Staff	30	10

At a 99% level of confidence, test the null hypothesis that equal proportions of the four categories listed above own iPhone.

Answer: Testing the null hypothesis that equal proportions of all four categories own Apple iPhones is equivalent to performing a chi-square test of the H_0 that there is no relationship between category membership and iPhone ownership. From the cross-tabulation provided:

	iPhone Ownership		
Category	Owns iPhone	Does not Own iPhone	Row Totals
1. Undergraduate Student	60	30	90
2. Graduate Student	20	20	40
3. Faculty	15	25	40
4. Staff	10	20	30
Column Totals	105	95	$n = 200$

$$E_{11} = \frac{90 \times 105}{200} = 47.25 \quad E_{12} = \frac{90 \times 95}{200} = 42.75$$

$$E_{21} = \frac{40 \times 105}{200} = 21 \quad E_{22} = \frac{40 \times 95}{200} = 19$$

$$E_{31} = \frac{40 \times 105}{200} = 21 \quad E_{32} = \frac{40 \times 95}{200} = 19$$

$$E_{41} = \frac{30 \times 105}{200} = 15.75 \quad E_{42} = \frac{90 \times 105}{200} = 14.25$$

Computed χ^2

$$= \frac{(60-47.25)^2}{47.25} + \frac{(30-42.75)^2}{42.75} + \frac{(20-21)^2}{21} + \frac{(20-19)^2}{19}$$
$$+ \frac{(15-21)^2}{21} + \frac{(25-19)^2}{19} + \frac{(10-15.75)^2}{15.75} + \frac{(20-14.25)^2}{14.25} = 15.37$$

Decision Rule: At a 99% level of confidence, reject H_0 if computed $\chi^2 > 11.34 = \chi^2_{.01}$ at degrees of freedom $(4-1) \times (2-1) = 3$.

Conclusion: Here, computed $\chi^2 = 15.37 > 11.34$. Hence we reject H_0 at a 99% level of confidence and conclude that at least one proportion is different from the others.

7.2 Chapter 2 Exercise Problems and Solutions

Problem Scenario: Consider the regression model:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2$$

$$+ \beta_5 D_1 X_1 + \beta_6 D_2 X_1 + \beta_7 D_1 X_2 + \beta_8 D_2 X_2 + \epsilon,$$

where:

Y = sales of a brand in a sales territory (unit = \$100,000) during Fall, 2003;

X_1 is the number of salespeople in the territory;

X_2 is the retail price (in dollars) in the territory;

D_1 and D_2 are dummy variables for the level of advertising in the territory. The advertising level can be low, medium, or high.

$D_1 = 1$ if the advertising level is medium, and $D_1 = 0$ otherwise;

$D_2 = 1$ if the advertising level is high, and $D_2 = 0$ otherwise.

ϵ is defined as usual.

1. Suppose we know that the true values of the regression parameters are as follows:

$\beta_0 = 5.0$, $\beta_1 = 1.0$, $\beta_2 = 1.5$, $\beta_3 = .2$, $\beta_4 = -.5$,

$\beta_5 = .1$, $\beta_6 = .15$, $\beta_7 = 0$, $\beta_8 = -.1$, $\sigma_\epsilon = 2$.

Compute the probability that the sales (unit = \$100,000) in a territory would exceed \$1100,000 if the advertising level is high, there are 20 salespeople, and the price is \$5.

2. State each of the following null hypotheses in terms of the parameters of the regression model (e.g., $H_0 : \beta_1 = 0$):

2.(a) The marginal effect of price on sales is the same for medium and high advertising levels.

2.(b) Changes in price do not affect sales if the level of advertising is low.

3. Suppose we have estimated regression models using data from 49 territories, and got the following results:

Results

Regression	Dependent Variable	Independent Variables	R^2
1	Y	$D_1, D_2, X_1, X_2, D_1X_1, D_2X_1, D_1X_2, D_2X_2$	.75
2	Y	D_1, D_2	.2
3	Y	X_1, X_2	.2
4	Y	$D_1, D_2, X_1, X_2, D_1X_1, D_2X_1$	.6
5	Y	$D_1, D_2, X_1, D_1X_1, D_2X_1$	.55
6	Y	$D_1, D_2, X_2, D_1X_2, D_2X_2$	.5
7	Y	$D_1, D_2, D_1X_1, D_2X_1, D_1X_2, D_2X_2$	.3
8	Y	$D_1, D_2, X_1, X_2, D_1X_2, D_2X_2$	.7

At a 99% level of confidence, test each of the following two null hypotheses:

3(a) The marginal effect of price on sales is the same for all levels of advertising.

3(b) Having an additional salesperson does not have any effect on sales at any level of advertising.

In each case, clearly state (i) H_0 in terms of the regression parameters, (ii) the decision rule, and (iii) the conclusion regarding H_0.

Answer: The regression model can be written separately for the three levels of advertising as follows:

Advertising Level	Regression Equation
Low ($D_1 = D_2 = 0$)	$Y = \beta_0 + \beta_3 X_1 + \beta_4 X_2 + \epsilon$
Medium ($D_1 = 1$, $D_2 = 0$)	$Y = (\beta_0 + \beta_1) + (\beta_3 + \beta_5)X_1 + (\beta_4 + \beta_7)X_2 + \epsilon$
High ($D_1 = 0$, $D_2 = 1$)	$Y = (\beta_0 + \beta_2) + (\beta_3 + \beta_6)X_1 + (\beta_4 + \beta_8)X_2 + \epsilon$

1. In this case, advertising level is high, $X_1 = 20$, and $X_2 = 5$. Therefore,

$$Y = (\beta_0 + \beta_2) + (\beta_3 + \beta_6) \times 20 + (\beta_4 + \beta_8) \times 5 + \epsilon$$

$$= (5.0 + 1.5) + (0.2 + 0.15) \times 20 + (-0.5 - 0.1) \times 5 + \epsilon = 10.5 + \epsilon$$

Therefore, Y is normally distributed with mean 10.5, and the same standard deviation as ϵ, that is 2 (unit = 100,000). Hence,

$$P(Y > 11) = P\Big(\frac{Y - 10.5}{2} > \frac{11 - 10.5}{2}\Big) = P(z > 0.25) = 0.50 - 0.0987 = 0.4013$$

2.(a) The null hypothesis here is that the coefficient of X_2 is same for medium and high advertising levels, that is, $(\beta_4 + \beta_7) = (\beta_4 + \beta_8)$. Restating, we have: $H_0 : \beta_7 - \beta_8 = 0$

2.(b) The null hypothesis here is that the coefficient of price is zero at low levels of advertising, that is, $\beta_4 = 0$

3.(a) The null hypothesis is, the coefficient of X_2 is same at low, medium, and high levels of advertising, that is,

$$\beta_4 = \beta_4 + \beta_7 = \beta_4 + \beta_8, \quad \text{that is,} \quad \beta_7 = \beta_8 = 0.$$

Hence, $H_0 : \beta_7 = \beta_8 = 0$

Full model: Y with D_1, D_2, X_1, X_2, D_1X_1, D_2X_1, D_1X_2, D_2X_2 (Regression 1). $R_f^2 = 0.75$

Restricted model: Y with D_1, D_2, X_1, X_2, D_1X_1, D_2X_1 (Regression 4). $R_r^2 = 0.60$

$n = 49$, $m = 8$, $k = 2$, $n - m - 1 = 49 - 8 - 1 = 40$

$$f = \frac{(R_f^2 - R_r^2)}{(1 - R_f^2)} \times \frac{(n - m - 1)}{k} = \frac{(0.75 - 0.60)}{(1 - 0.75)} \times \frac{40}{2} = 12$$

Decision Rule: At a 99% level of confidence, reject H_0 if $f > F_{.01}(2, 40) = 5.18$

Conclusion: Since $f = 12 > 5.18$, we reject H_0 at a 99% level of confidence.

3.(b) The null hypothesis is, the coefficient of X_1 is zero at low, medium, and high levels of advertising, that is,

$$\beta_3 = \beta_3 + \beta_5 = \beta_3 + \beta_6 = 0, \quad \text{that is,} \quad \beta_3 = \beta_5 = \beta_6 = 0.$$

Hence, $H_0 : \beta_3 = \beta_5 = \beta_6 = 0$

Full model: Y with D_1, D_2, X_1, X_2, D_1X_1, D_2X_1, D_1X_2, D_2X_2 (Regression 1). $R_f^2 = 0.75$

Restricted model: Y with D_1, D_2, X_2, D_1X_2, D_2X_2 (Regression 6). $R_r^2 = 0.50$

$n = 49$, $m = 8$, $k = 3$, $n - m - 1 = 49 - 8 - 1 = 40$

$$f = \frac{(R_f^2 - R_r^2)}{(1 - R_f^2)} \times \frac{(n - m - 1)}{k} = \frac{(0.75 - 0.50)}{(1 - 0.75)} \times \frac{40}{3} = 13.33$$

Decision Rule: At a 99% level of confidence, reject H_0 if $f > F_{.01}(3, 40) = 4.31$

Conclusion: Since $f = 13.33 > 4.33$, we reject H_0 at a 99% level of confidence.

7.3 Chapter 3 Exercise Problem and Solution

Conjoint analysis was done to determine how a Syracuse University undergraduate student's evaluation of a plug-in electric car depends on the levels of three attributes:

- X_1: Charging time after a full discharge. Range: 6 to 24 hours.
- X_2: Top speed. Range: 60 to 120 miles/hour
- X_3: Driving range, that is, the number of highway miles one can drive after a full charge. Range: 200 to 400 miles.

It is assumed that a student's rating of a plug-in electric car on a 0−100 scale can be expressed as:

$$U(X_1, X_2, X_3) \quad = \quad U_0 + U_1(X_1) + U_2(X_2) + U_3(X_3),$$

where $U_1(6) = 0$, $U_2(60) = 0$, $U_3(200) = 0$, and a higher score means the student likes the product better.

For a given student:

$U_0 = 40$		
$U_1(6) = 0$	$U_1(12) = -10$	$U_1(24) = -40$
$U_2(60) = 0$	$U_2(80) = 20$	$U_2(120) = 30$
$U_3(200) = 0$	$U_3(300) = 10$	$U_3(400) = 30$

Which of the following two plug-in electric cars will this student prefer?

- Car 1: $X_1 = 15$, $X_2 = 100$, $X_3 = 250$
- Car 2: $X_1 = 9$, $X_2 = 75$, $X_3 = 350$

Answer:

$U_1(9) = U_1(6) + [(\frac{9-6}{12-6}) \times \{U_1(12) - U_1(6)\}] = 0 + [(\frac{3}{6} \times (-10 - 0)] = -5$

$U_1(15) = U_1(12) + [(\frac{15-12}{24-12}) \times \{U_1(24) - U_1(12)\}]$

$= -10 + [(\frac{3}{12} \times \{(-40) - (-10)\}] = -10 + [(\frac{3}{12}) \times (-30)] = -17.5$

$U_2(75) = U_2(60) + [(\frac{75-60}{80-60}) \times \{U_2(80) - U_2(60)\}] = 0 + [(\frac{15}{20}) \times (20 - 0)] = 15$

$U_2(100) = U_2(80) + [(\frac{100-80}{120-80}) \times \{U_2(120) - U_2(80)\}]$

$= 20 + [(\frac{10}{20}) \times (30 - 20)] = 25$

$U_3(250) = U_3(200) + [(\frac{250-200}{300-200}) \times \{U_3(300) - U_3(200)\}] = 0 + [(\frac{50}{100}) \times (10 - 0)] = 5$

$U_3(350) = U_3(300) + [(\frac{350-300}{400-300}) \times \{U_3(400) - U_3(300)\}] = 10 + [(\frac{50}{100}) \times (30-10)] = 20$

Car 1: $U(15, 100, 250) = U_0 + U_1(15) + U_2(100) + U_3(250) = 40 - 17.5 + 25 + 5 = 52.5$

Car 2: $U(9, 75, 350) = U_0 + U_1(9) + U_2(75) + U_3(350) = 40 - 5 + 15 + 20 = 70$

Since $70 > 52.5$, this person prefers car 2.

7.4 Chapter 5 Exercise Problems and Solutions

1. **Scenario:** Suppose you have conducted logit analysis on data collected from the students of a college campus to determine how the probability that the student is member of a student organization (such as the AMA) depends on the gender of the student and the class the student is in. Suppose we have used a sample of size 500 with the indicator function given as follows:

$$I \;=\; \beta_0 \,+\, \beta_1 D_1 \,+\, \beta_2 D_2 \,+\, \beta_3 D_3$$

where:
$D_1 = 1$ if the student is a freshman, 0 if not,
$D_2 = 1$ if the student is a sophomore, 0 if not,
$D_3 = 1$ if the student is female, 0 if not.

Results:

Estimated Coefficient	Standard Deviation of Estimate
$b_0 = -2$	$s_{b_0} = .2$
$b_1 = -.6$	$s_{b_1} = .3$
$b_2 = -.4$	$s_{b_2} = .3$
$b_3 = 1.0$ $\ln L = -540$	$s_{b_3} = .4$

1.(a) Interpret in words the meaning of the following null hypothesis: $H_0 : \beta_1 = \beta_2$.

1.(b) Suppose the student population at the campus has the following composition:

	Men	Women
Freshmen	1800	1200
Sophomores	1500	1500
Juniors & Seniors	3300	2700

Based on the results of the logit analysis, estimate how many students at the campus are members of student organizations.

1.(c) Suppose we have run logit with D_3 as the only independent variable and obtained a log likelihood of -560. At a 99% level of confidence, test $H_0 : \beta_1 = \beta_2 = 0$. Interpret H_0.

Answer: Here, the logit model has the indicator function:

$$I = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3,$$

where

$D_1 = 1$ if the student is a freshman, 0 if not,
$D_2 = 1$ if the student is a sophomore, 0 if not,
$D_3 = 1$ if the student is a female, 0 if not.

1.(a) According to the $H_0 : \beta_1 = \beta_2$, we have:

$$(\beta_0 + \beta_1 + \beta_3 D_3) \;=\; (\beta_0 + \beta_2 + \beta_3 D_3),$$

that is, gender being the same, the indicator function of a freshman is equal to the indicator function of a sophomore.

Therefore, H_0 means that a freshman and a sophomore of the same gender are equally likely to be a member of a student organization.

1.(b) Consider the six sub-populations separately:

(i) Freshmen, Male. Here, $D_1 = 1$, $D_2 = 0$, $D_3 = 0$.
Therefore, $I = \beta_0 + \beta_1 \;\approx\; b_0 + b_1 = -2.6 \;\longrightarrow\; P(y = 1) \;\approx\; .0691$.

(ii) Freshmen, Female. Here, $D_1 = 1$, $D_2 = 0$, $D_3 = 1$.
Therefore, $I = \beta_0 + \beta_1 + \beta_3 \;\approx\; b_0 + b_1 + b_3 \;=\; -1.6 \longrightarrow\; P(y = 1) \;\approx\; .1680$.

(iii) Sophomore, Male. Here, $D_1 = 0$, $D_2 = 1$, $D_3 = 0$.
Therefore, $I = \beta_0 + \beta_2 \;\approx\; b_0 + b_2 \;=\; -2.4 \;\longrightarrow\; P(y = 1) \;\approx\; .0832$.

(iv) Sophomore, Female. Here, $D_1 = 0$, $D_2 = 1$, $D_3 = 1$.
Therefore, $I = \beta_0 + \beta_2 + \beta_3 \;\approx\; b_0 + b_2 + b_3 \;=\; -1.4 \;\longrightarrow\; P(y = 1) \;\approx\; .1978$.

(v) Junior/Senior, Male. Here, $I = \beta_0 \;\approx\; b_0 \;=\; -2 \;\longrightarrow\; P(y = 1) \;\approx\; .1192$.

(vi) Junior/Senior, Female. Here, $I = \beta_0 + \beta_3 \;\approx\; b_0 + b_3 \;=\; -1 \;\longrightarrow\; P(Y = 1) \;\approx\; .2689$.

Therefore, the estimated total number of students that are members of student organizations is: $(.0691 \times 1800) + (.1680 \times 1500) + (.0832 \times 1500) + (.1978 \times 1500) + (.1192 \times 3300) + (.2689 \times 2700) \;=\; 1917$.

1.(c) We are testing $H_0 : \beta_1 = \beta_2 = 0$, against H_a: at least one of β_1 and β_2 is not zero.

Decision rule: At a 99% level of confidence, reject H_0 if $2(\ln L_f - \ln L_r) \;>\; 9.21 \;=\; \chi^2_{.01}$ at degrees of freedom = 2.

Conclusion: Here, $2(\ln L_f - \ln L_r) \;=\; 2\{(-540) - (-560)\} \;=\; 40 \;>\; 9.21 \;\longrightarrow$ reject H_0 at 99% confidence.

$H_0 : \beta_1 = \beta_2 = 0$ means that for a given gender, freshmen, sophomores and juniors/seniors all have the same value of I, that is, are all equally likely to be members of student organizations.

2. **Problem Scenario:** Y is a binary variable defined as follows: $Y = 1$ if a person owns a digital camcorder, and $Y = 0$ if not. Let X be the income of the person (unit = \$10,000), and let $D = 1$ if the person has children, and $D = 0$ if not. Suppose Y is related to X and D according to the logit model:

$$P(Y = 1|I) = \frac{1}{1 + e^{-I}}, \quad \text{where} \quad I = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 a * D * X$$

2.(a) Suppose we know that $\beta_0 = -2$, $\beta_1 = .1$, $\beta_2 = .5$, and $\beta_3 = .02$. If a person earns \$60,000 and has children, what is the probability that (s)he owns a digital camcorder?

Answer: For this person, $I = -2 + (.1 * 6) + (.5 * 1) + (.02 * 1 * 6) = -.78$

$$P(Y = 1) = \frac{1}{1 + e^{-I}} = \frac{1}{1 + e^{.78}} = .31432$$

2.(b) Suppose we estimated several logit models and obtained the following results:

	Dependent Variable	Independent Variables	$\ln L$
Model 1	Y	X, D, $D * X$	-150
Model 2	Y	X	-165
Model 3	Y	D	-170
Model 4	Y	X, D	-152
Model 4	Y	X, $D * X$	-160
Model 5	Y	D, $D * X$	-166
Model 6 (Naive Model)	Y	____	-180

2.(a) At a 99% level of confidence, test $H_0 : \beta_1 = \beta_3 = 0$ against H_a : at least one of β_1 and β_3 is not zero. What does the null hypothesis mean?

Answer: The null hypothesis mean $I = \beta_0 + \beta_2 D$ does not depend on income. Thus, given that a person has children or not, probability of owning a camcorder does not depend on income.

$H_0 = \beta_1 = \beta_3 = 0$.

Full model (Model 1): Y with X, D, $D * X$. $\ln L_{full} = -150$

Restricted model (Model 3): Y with D. $\ln L_{restricted} = -170$

$2(\ln L_{full} - \ln L_{restricted}) = 2 \times 20 = 40 > 9.21 = \chi^2_{.01}$ at degrees of freedom $k = 2$. Hence we reject H_0 at a 99% level of confidence.

2.(b) At a 99% level of confidence, test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against H_a : at least one of β_1, β_2, and β_3 is not zero.

Answer: $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$.

Full model (Model 1): Y with X, D, $D * X$. $\ln L_{full} = -150$

Restricted model (Model 6, naive model): $\ln L_{restricted} = -180$

$2(\ln L_{full} - \ln L_{restricted}) = 2 \times 30 = 60 > 11.3 = \chi^2_{.01}$ at degrees of freedom $k = 3$. Hence we reject H_0 at a 99% level of confidence.

3. Let Y be a dichotomous variable representing whether a person is married or not, and x the person's age in years. Let y depend on x according to the LOGIT model with the indicator function, I, given by:

$I = -3 + 0.1 * X$

Suppose a random sample of six individuals have ages (in years) 29, 33, 42, 24, 36, and 39. Compute the probability that these six individuals are **<u>all</u>** married.

[Hint: We are assuming that the six cases are independent. First compute the probability for each individual. The probability that they are all married is the product of these six probabilities.]

Answer:

Case	X	$I = -3 + .01 * X$	$P(Y = 1 \mid I)$
1	29	-0.1	0.475020813
2	33	0.3	0.574442517
3	42	1.2	0.768524783
4	24	-0.6	0.354343694
5	36	0.6	0.645656306
6	39	0.9	0.710949503

The probability that all six individuals are married is the product of the six probabilities in the extreme right column, and is given by 0.03411.

4. Consider a simpler version of the model given above where Y and X are as defined in problem 1, and,

$$P(Y = 1 \mid I) = \frac{1}{1 + e^{-I}}, \quad \text{where} \quad I = \beta_0 + \beta_1 X.$$

Suppose you have the following data:

Case	Y	X
1	1	8
2	1	5
3	1	6
4	1	7.5
5	0	6
6	0	4
7	0	5
8	0	3

Which of the following two sets of parameter values gives you a better fit with the data?

Set 1: $\beta_0 = -1$, $\beta_1 = .1$

Set 2: $\beta_0 = 1$, $\beta_1 = .1$.

Hint:

- For each set of parameters, you need to compute log likelihood. The set with greater log likelihood gives the better fit.
- For a given set of parameters, for each case, you need to compute the probability that you will get that observation ($P(obs)$) for that parameter set. Compute I and $P(Y = 1 \mid I)$. If $Y = 1$, then $P(obs) = P(Y = 1 \mid I)$. If $Y = 0$, then $P(obs) = 1 - P(Y = 1 \mid I)$.
- For a given set of parameters, first compute $P(obs)$ for each case. Using that, compute $\ln P(obs)$ for each case and add over the observations. This is the log likelihood of the data

set for that set of parameters.

- You may to use Excel to do this.

Answer: If $\beta_0 = -1$ and $\beta_1 = .1$, then $I = -1 + .1X$. From the data:

Case	Y	X	I	$P(Y=1)$	$P(obs)$
1	1	8	-0.2	0.450166003	0.450166003
2	1	5	-0.5	0.377540669	0.377540669
3	1	6	-0.4	0.40131234	0.40131234
4	1	7.5	-0.25	0.437823499	0.437823499
5	0	6	-0.4	0.40131234	0.59868766
6	0	4	-0.6	0.354343694	0.645656306
7	0	5	-0.5	0.377540669	0.622459331
8	0	3	-0.7	0.331812228	0.668187772

The likelihood of the sample is 0.004800973, and $\ln L = -5.338936762$.

If $\beta_0 = 1$ and $\beta_1 = .1$, then $I = 1 + .1X$. From the data:

Case	Y	X	I	$P(Y=1)$	$P(obs)$
1	1	8	1.8	0.858148935	0.858148935
2	1	5	1.5	0.817574476	0.817574476
3	1	6	1.6	0.832018385	0.832018385
4	1	7.5	1.75	0.851952802	0.851952802
5	0	6	1.6	0.832018385	0.167981615
6	0	4	1.4	0.802183889	0.197816111
7	0	5	1.5	0.817574476	0.182425524
8	0	3	1.3	0.785834983	0.214165017

The likelihood of the sample is 0.000645648, and $\ln L = -7.345255662$

Comparing likelihoods, $\beta_0 = -1$ and $\beta_1 = .1$ gives better fit to the data.

8 Appendix: Scales of Measurement

8.1 Introduction

A researcher typically measures different things from a respondent. For example, consider the questionnaire:

Q1. Gender: ____Male ____Female

Q2. How much money did you pay this semester to buy text books?

$ ____ dollars

Q3. How many times a month do you shop at grocery stores?

____Never ____1−2 times/month ____3−4 times/month

____5 or more times/month

This questionnaire is measuring three distinct things, which we call variables. The nature of a measurement depends on:

- the nature of the variable (whether there is a quantity involved or not), and
- how well the respondent can distinguish between levels of the variable.

Question 1 does not measure any quantity.

Question 2 measures a quantity at a fine level. Specifically, comparing the numbers we record, we can tell **how much more** one respondent has paid compared to another, and **how many times** more one has paid compared to another.

Question 3 measures a quantity at a crude level. For example, suppose we code the answers from question 3 as:

1 Never 2 1−2 times 3 3−4 times 4 5 or more times

Then, a respondent with a recorded number of 4 shops more often than a respondent with score of 3. However, we cannot tell how much more.

Clearly, with finer information, we can do a more sophisticated analysis.

Example: How far from Syracuse are Rochester, Buffalo, and Pittsburgh? How far are Chicago, St.Louis, Lincoln (Nebraska)?

You probably know the precise distances to Rochester and Buffalo. As the location becomes less familiar, you probably cannot tell me the exact distances, but still provide rank orders such as Lincoln is farther away from Syracuse than Chicago.

Depending on whether the measurement involves a quantity, and, if a quantity is involved, how finely the measurement can distinguish between different levels of the quantity, we can have four different scales of measurement: nominal, ordinal, interval, and ratio. In the discussions that follow, Z_i denotes the score given by the measurement to object i.

8.2 Definitions of Scales

1. Nominal Scale: A measurement on a nominal scale, also called a **categorical** scale only requires that if two objects get the same score, then they belong to the same category. Formally, if $Z_i = Z_j$, then objects i and j belong to the same category.

Any measurement satisfies the requirement of a nominal scale, regardless of whether the measurement has a quantitative meaning. For example, Z may measure which color a person prefers when he buys a new car: 1 for blue, 2 or green, 3 for white, 4 for red, and 5 for any other color. If two people both have $Z = 2$, then they both prefer to have green cars. Clearly, the measurement here has no quantitative meaning.

On the other hand, Z can also be a person's age measured in years. For example, if two people have $Z = 18$, then they are both 18 years old. Here, the measurement does have a quantitative meaning.

If the measurement does **not** have a quantitative meaning, we call the measured variable a **purely nominal** variable. Common examples of purely nominal variables include gender, ethnicity, and brand choice. For a purely nominal variable, it is appropriate to compute the frequency with which a given category occurs in the data, find the category that has the highest frequency (mode), and construct cross-tabulations. We can also estimate the probability that the variable will belong to a given category (e.g., the probability that a person will buy Aqua Fresh toothpaste during his next trip to a grocery store). However, it is **not** meaningful to compute a mean or a median of a purely nominal variable.

For the next three scales, the measurement does have a quantitative meaning.

2. Ordinal Scale: This scale satisfies the property that if $Z_i > Z_j$, then object i has more of some property than object j.

The respondent may or may not be able to say how much more of the property object i has than object j, that is, she may not be able to compare the sizes of differences between pairs of objects.

If the respondent **cannot** compare differences, then we call this scale a **purely ordinal** scale. If we have a variable measured on a purely ordinal scale, then we can find its median. However, since we cannot compare differences, it is not strictly correct to compute an average.

Two Commonly Used Purely Ordinal Scales:

(i) Ordered Categorical Scale: Suppose we recorded how much money a family spent on cellular telephone service during April, 2014 as follows:

<u>1</u> \$20 or less <u>2</u> \$21-\$40 <u>3</u> \$41-\$60 <u>4</u> \$61-\$80 <u>5</u> \$81-\$100 <u>6</u> \$101 or more

Here, we are dividing the range of all possible cellular service expenditures into a fixed number of categories, and a higher category represents greater expenditure. Thus, the scale satisfies the ordinal property. However, we cannot tell how much more. For example, the difference between categories 5 and 4 can be anything from \$1 to \$39. This is an example of an **Ordered Categorical Scale**. The commonly used five and seven point scales are also examples of ordered categorical scales.

Note that Z here is a coarse approximation to the underlying property X. If the number of categories is increased, the approximation improves.

(ii) Rank Orders: Suppose we administered a test to a class of 10 students, and all students got different scores. We then set $Z = 10$ for the student who had the highest score, $Z = 9$ for the student with the second highest score, . . ., $Z = 1$ for the student with the lowest score.

Clearly, if $Z_1 > Z_2$, then student 1 got a higher score than student 2. However, we cannot tell how much more.

---

3. Interval Scale: This scale requires that we have $Z_i > Z_j$ if and only if object i has more of the property than object j, and from the difference $(Z_i - Z_j)$, we know exactly how much more of the property object i has than object j.

Since the interval scale allows us to compare differences between pairs of objects, we can define a unit of difference, and compute averages and standard deviations. We can use a quantity measured on an interval scale as a dependent variable in regression analysis.

There may still not be a clearly defined origin, which is needed to compare absolute sizes rather than just differences. The final scale in the hierarchy, the ratio scale, satisfies that condition.

4. Ratio Scale: A measurement on a ratio scale satisfies all the requirements of an interval scale plus the added requirement that from the measured values, we can compare absolute sizes of the underlying properties. In this case, there is a clearly defined origin, and ratios are meaningful. For example, if $Z_i = 2Z_j$, then object i has two times the property than object j.

Note that a measurement on a ratio scale satisfies the interval, ordinal, and nominal scale properties also. A measurement on an interval scale satisfies ordinal and nominal properties as well. An ordinal scaled measurement also satisfies nominal scale properties.

Example: To make these discussions concrete, suppose you are looking at three buildings (Figure 7.1), marked 1, 2, and 3.

If we use the numbers 1, 2, and 3 to simply denote the fact that these are three different buildings, the measurement is nominal. If, in addition, we can tell from the number if a building is taller than another, then the measurement is ordinal. For example, here, building 2 is taller than building 1, and building 3 is taller than buildings 1 and 2.

If in addition we can compare the differences in heights of the buildings, we have a measurement on an interval scale. For example, we may find that building 2 is 50 feet taller than building 1, and building 3 is 100 feet taller than building 2. We, however, still cannot compare absolute heights.

If we also knew that building 1 is 100 feet tall, we would have the absolute heights of all buildings. Then, we would know that building 1 is 100 feet tall, building 2 is 150 feet tall, and building 3 is 250 feet tall. Thus, for example, the ratio of heights of buildings 3 and 1 is 2.5. The measurement is now on a ratio scale. Thus, with added information, we progress from a nominal to a ratio scale.

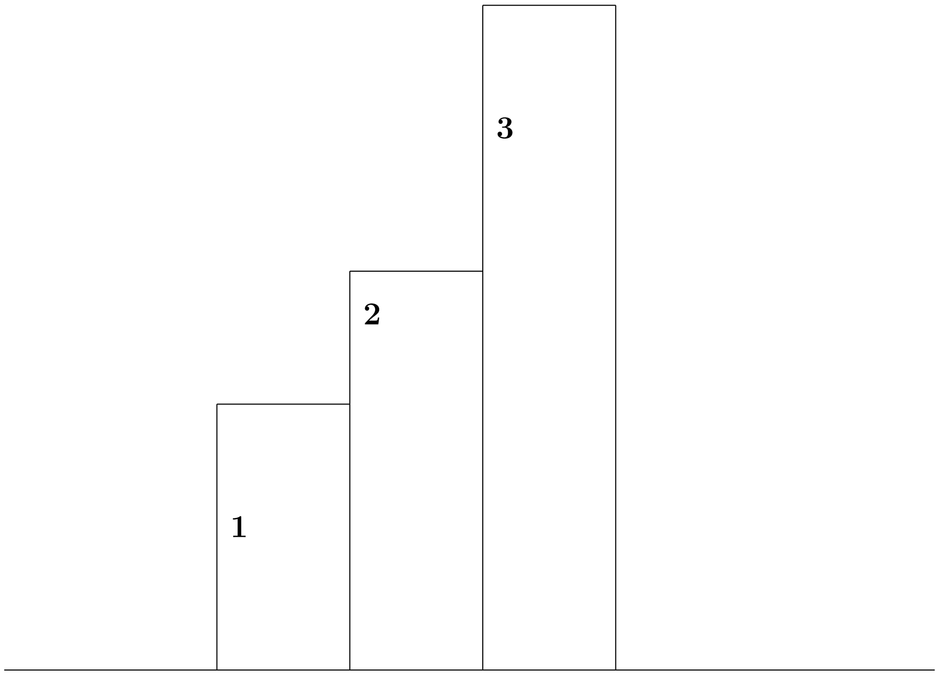


Figure 7.1

8.3 Scales of measurement and statistics

Univariate analysis

Scale of Measurement	Appropriate Statistics
Nominal	Frequency, percent, mode
Ordinal	Frequency, percent, mode, median
Interval	Frequency, percent, mode, median, mean, standard deviation
Ratio	Frequency, percent, mode, median, mean, standard deviation, percentage change

Multivariate analysis

Cross-tabulation and chi-square analysis can be used to test if two nominal variables are related. Correlation (more formally Pearson correlation) captures the strength of linear relationship between two interval scaled variables. For models with a dependent and independent variables, we have the following:

Dependent Variable	Independent Variables	Appropriate Method
Nominal	Nominal, interval	Logit
Interval	Interval	Regression
Interval	Nominal, interval	Dummy variable regression

9 Appendix: Matrix Algebra of Principal Components Analysis

9.1 Basic Results

Suppose we have n observations of m variables X_1, ..., X_m. We then standardized each variable and written the data as a matrix where each row of data comes from one respondent. Each column contains data from one variable. We call this matrix X_s.

Notations:

X_s: standardized $n \times m$ matrix of variables.
Sample correlation matrix: $R = \frac{1}{n-1} X_s' X_s$ (assumed to be full rank)
Let $\vec{z} = X_s \vec{u}$ where $\vec{u}$ is $m \times 1$ column vector. $\vec{z}$ is mean corrected since X_s is standardized.

$$\text{var}(Z) = \frac{1}{n-1} \vec{z}' \vec{z} = \frac{1}{n-1} \vec{u}' X_s' X_s \vec{u} = \vec{u}' R \vec{u}.$$

Objective of principle components analysis:

(1) $$\text{Maximize} \quad \vec{u}' R \vec{u} \quad \text{such that} \quad \vec{u}' \vec{u} = 1.$$

The Lagrangian is given by:

$$L \quad = \quad \vec{u}' R \vec{u} - \lambda(\vec{u}' \vec{u} - 1).$$

Hence:

$$\frac{\partial L}{\partial \vec{u}} \quad = \quad 2R\vec{u} - 2\lambda \vec{u} \quad = \quad \vec{0},$$

that is,

(2) $$R\vec{u} \quad = \quad \lambda \vec{u}.$$

By construction, R is a symmetric $m \times m$ matrix where the element (i, j) is the correlation between X_i and X_j. It is also positive definite since it is full rank. Hence, its eigenvalues $\lambda_1, \ldots, \lambda_m$ are all greater than zero. Also, $\sum_{i=1}^{m} \lambda_i = m$, and we can select an orthogonal set of eigenvectors $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m$ such that:

(3) $$\vec{u}_i' \vec{u}_i = 1, \quad \vec{u}_i' \vec{u}_j = 0 \quad \text{if } i \neq j.$$

Therefore:

$$\text{Cov}(z_i, z_j) = \frac{1}{n-1} \vec{z}_i' \vec{z}_j = \frac{1}{n-1} \vec{u}_i' X_s' X_s \vec{u}_j = \vec{u}_i' R \vec{u}_j = \lambda_j \vec{u}_i' \vec{u}_j.$$

Hence,

(4) $$\text{Cov}(z_i, z_j) = \lambda_i \quad \text{if } i = j, \quad 0 \quad \text{if } i \neq j.$$

Therefore,

(5) $$\text{Var}(z_i) \quad = \quad \lambda_i.$$

Also, $R\vec{u}_1 = \lambda_1$, $R\vec{u}_2 = \lambda_2$, etc. Therefore, we can write:

$$RU \quad = UD, \tag{6}$$

where

$$U = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n], \quad D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \lambda_n \end{pmatrix}$$

Thus,

$$UU' \quad = \quad U'U \quad = \quad I_{m\times m},$$

$$R = UDU' = \lambda\vec{u}_1\vec{u}_1' + \ldots + \lambda_m\vec{u}_m\vec{u}_m'.$$

Component Scores: Let

$$Z_{n\times m} \quad = \quad [\vec{z}_1, \ldots, \vec{z}_m] \quad = \quad X_sU$$

Since z_i has a variance of λ_i, $\vec{z}_{si} = \vec{z}_i/\sqrt{\lambda_i}$ has variance = 1. Let

$$Z_s = [\vec{z}_1/\sqrt{\lambda_1}, \ldots, \vec{z}_m/\sqrt{\lambda_m}] = ZD^{-1/2} = X_sUD^{-1/2}. \tag{7}$$

Therefore,

$$\frac{1}{n-1}Z_s'Z_s = \frac{1}{n-1}D^{-1/2}U'X_s'X_sUD^{-1/2} = D^{-1/2}U'RUD^{-1/2}$$

Since $RU = UD$, it follows that

$$\frac{1}{n-1}Z_s'Z_s \quad = \quad D^{-1/2}U'UDD^{-1/2} \quad = \quad I \tag{8}$$

Thus, Z_s is the $n \times m$ matrix of standardized component scores.

Since X_s and Z_s are standardized,

$$r(X_i, Z_j) = r(X_{si}, Z_{sj}) = \mathrm{cov}(X_{si}, Z_{sj}) = \frac{1}{n-1}\vec{x}_{si}'\vec{z}_{sj},$$

where $\vec{x}_{si}$ is the i-th column of X_s, $\vec{z}_{sj}$ is the j-th column of Z_s. Define the $m \times m$ matrix F as:

$$F \quad = \quad \frac{1}{n-1}X_s'Z_s. \tag{9}$$

F is called the *Factor Loading Matrix.* Note that

$$F_{ij} \quad = \quad r(X_{si}, Z_{sj}) \quad = \quad r(X_i, Z_j),$$

$$F = \frac{1}{n-1}X_s'X_sUD^{-1/2} = RUD^{-1/2} = UDU'UD^{-1/2} = UD^{1/2}$$

Rewriting,

$$F = [\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_m] = UD^{1/2}, \tag{10}$$

where

$$\vec{f_i} = \begin{bmatrix} r(x_1, z_i) \\ \vdots \\ r(x_n, z_i) \end{bmatrix} = \sqrt{\lambda_i}\vec{u}_i \tag{11}$$

The sum of squares of column j of F is:

$$\vec{f_j}'\vec{f_j} \quad = \quad \lambda_j \quad = \quad \mathrm{Var}(z_j) \tag{12}$$

The sum of squares of row i of F is:

$$(FF')_{ii} = (UD^{1/2}D^{1/2}U')_{ii} = (UDU')_{ii} = R_{ii} = 1 \tag{13}$$

This is called the *communality* of x_i, which represents the sum of squared correlation of x_i with z_1, z_2, ..., z_m. Also:

$$FF' \quad = \quad R, \tag{14}$$

and

$$Z_sF' = (X_sUD^{-1/2})(UD^{1/2})' = (X_sUD^{-1/2})(D^{1/2}U') = X_sUU' = X_s,$$

that is,

$$X_s \quad = \quad Z_sF' \tag{15}$$

Expanding,

$$\vec{x}_{si} = F_{i1}\vec{z}_{s1} + F_{i2}\vec{z}_{s2} + \ldots + F_{im}\vec{z}_{sm} \tag{16}$$

Note that column i of F' is same as row i of F, which gives us correlations of x_{si} with z_{s1}, z_{s2}, ..., z_{sm}. We also know that z_{s1}, ..., z_{sm} are mutually orthogonal in the n dimensional space of observations. They represent an m dimensional basis in an n dimensional vector space. F_{i1}, F_{i2}, ..., F_{im} represent direction cosines of $\vec{x}_{si}$ with respect to $\vec{z}_{s1}$, ..., $\vec{z}_{sm}$.

9.2 Interpretations of Principal Components Analysis

Interpretation No. 1. Note that

$$\sum_{i=1}^{m} \mathrm{Var}(z_i) \quad = \quad m \quad = \quad \sum_{i=1}^{m} \mathrm{Var}(x_{si}).$$

Thus, principal components analysis *redistributes* the total variance of the sample into mutually uncorrelated factors. The first component, corresponding to the largest eigenvalue λ_1, has the highest variance. The second component is uncorrelated with the first component, and $\vec{u}_1'\vec{u}_2 = 0$. It has the next highest variance, etc.

By retaining $\vec{z}$'s corresponding to the largest eigenvalues (λ's), we can have a reduced data set that captures *most of the information* of the original data set.

Interpretation No. 2.[2] Let $\vec{v}$ be an $(m \times 1)$ vector, and

$$\vec{z}_{n\times 1} \quad = \quad X_s\vec{v}$$

[2] This interpretation was provided by Professor V.Srinivasan in a multivariate analysis class at Stanford University.

Let $\vec{c}$ be the $(m \times 1)$ column vector of correlations of z with x_1, ..., x_m, that is:

$$\vec{c} \quad = \quad \begin{bmatrix} r(z, x_1) \\ \vdots \\ r(z, x_m) \end{bmatrix}. \tag{17}$$

As X_s is mean corrected, so is $\vec{z}$. Hence,

$$\mathrm{Var}(z) = \frac{1}{n-1}\vec{v}' X_s' X_s \vec{v} = \vec{v}' R \vec{v}$$

Therefore,

$$\vec{c} = \frac{1}{n-1} X_s' \vec{z} / \sqrt{\mathrm{Var}(z)} = \frac{1}{n-1} X_s' X_s \vec{v} / \sqrt{\mathrm{Var}(z)} = \frac{R\vec{v}}{\sqrt{\vec{v}' R \vec{v}}}$$

Hence:

$$r^2(z, x_1) + r^2(z, x_2) + \ldots + r^2(z, x_m) = \vec{c}'\vec{c} = \frac{\vec{v}' R^2 \vec{v}}{\vec{v}' R \vec{v}} \tag{18}$$

Suppose we want to choose $\vec{v}$ which maximizes the mean squared correlation of z with x_1, ..., x_m, that is, we want to maximize the ratio given by equation (18). This ratio is unchanged if $\vec{v}$ is multiplied by a scalar. Thus, we can solve the following problem:

Maximize $\vec{v}' R^2 \vec{v}$ subject to condition $\vec{v}' R \vec{v} = 1$.

The Lagrangian is:

$$L \quad = \quad \vec{v}' R^2 \vec{v} - \mu(\vec{v}' R \vec{v} - 1)$$

where μ is the Lagrangian multiplier. At the solution:

$$\frac{\partial L}{\partial \vec{v}} = 2R^2\vec{v} - 2\mu R\vec{v} = \vec{0}, \quad \text{that is,} \quad R\vec{v} = \mu\vec{v}$$

Hence, principal components analysis is equivalent to maximizing $\sum_{i=1}^{m} r^2(z, x_i)$. Note that

$$\vec{v}' R^2 \vec{v} = \vec{v}' R(R\vec{v}) = \mu \vec{v}' R \vec{v}, \quad \text{that is,} \quad \mu = \frac{\vec{v}' R^2 \vec{v}}{\vec{v}' R \vec{v}}$$

Interpretation No. 3. Let R_p be the population correlation matrix, and let $X_{n \times m}$ be standardized in the population. Hence,

$$f(\vec{x}) = \frac{1}{(2\pi)^{m/2} |R_p|^{1/2}} \, e^{-\frac{1}{2}(\vec{x}-\vec{0})' R_p^{-1} (\vec{x}-\vec{0})} = \frac{1}{(2\pi)^{m/2} |R_p|^{1/2}} e^{-\frac{1}{2}\vec{x}' R_p^{-1} \vec{x}}$$

A concentration ellipsoid for $\vec{x}$ is given by:

$$\vec{x}' R_p^{-1} \vec{x} \quad = \quad c, \tag{19}$$

where c is a given positive number.

Noting that the mean of $\vec{x}$ is $\vec{0}$, to find a major axis of the ellipsoid given by (19) we have to solve to following problem:

Maximize $(\vec{x}-\vec{0})'(\vec{x}-\vec{0})$ (this is the square of the half length of the major axis) subject to $\vec{x}'R_p^{-1}\vec{x}=c$, that is,

$$\text{Maximize} \quad \vec{x}'\vec{x} \quad \text{subject to} \quad \vec{x}'R_p^{-1}\vec{x}-c=0$$

The Lagrangian is:

$$L \quad = \quad \vec{x}'\vec{x}-\lambda(\vec{x}'R_p^{-1}\vec{x}-c)$$

At the solution:

$$\frac{\partial L}{\partial \vec{x}} \quad = \quad 2\vec{x}-2\lambda R_p^{-1}\vec{x}=\vec{0},$$

since R_p^{-1} is symmetric, that is:

$$R_p\vec{x} \quad = \quad \lambda\vec{x}. \tag{20}$$

Therefore, the major axes of the concentration ellipsoid are given by the eigenvectors of the population correlation matrix R_p.

Also,

$$R_p\vec{x}=\lambda\vec{x} \quad \longrightarrow \quad \frac{1}{\lambda}\vec{x}=R_p^{-1}\vec{x} \quad \longrightarrow \quad \vec{x}'R_p^{-1}\vec{x}=\frac{1}{\lambda}\vec{x}'\vec{x}.$$

Hence,

$$\vec{x}'\vec{x} \quad = \quad \lambda\vec{x}'R_p^{-1}\vec{x} \quad = \quad \lambda c, \tag{21}$$

that is, for a given concentration ellipsoid (c fixed), the square of the length of the major axis is proportional to the eigenvalue λ.

Thus, principal components analysis is equivalent to solving for the major axes of the concentration ellipsoid, the eigenvectors corresponding to the largest eigenvalue giving us the first major axis, etc.

Note: Solving the eigenstructure problem for R is only an approximation for solving the eigenstructure problem for R_p. On the average, solving the eigenstructure problem for R gives you biased estimates of the eigenvalues of R_p. For example, consider

$$R \quad = \quad \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Here, $\lambda_1=1+|\rho|$ is the largest eigenvalue, and $\lambda_2=1-|\rho|$ is the smallest eigenvalue. Even if

$$R_p \quad = \quad I_{2\times 2} \quad = \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$E(|\rho|)>0$, that is, $E(\lambda_1)>1$ and $E(\lambda_2)<1$. If the sample size (n) is very large, these biases tend to vanish.

9.3 On Retaining a Subset of Principal Components

Recapitulating,

$$RU=UD \quad \text{where} \quad D=\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \lambda_m \end{pmatrix}, \tag{22}$$

(23) $$R = UDU' = \lambda_1 \vec{u}_1 \vec{u}_1' + \lambda_2 \vec{u}_2 \vec{u}_2' + \ldots + \lambda_m \vec{u}_m \vec{u}_m',$$

(24) $$Z = X_s U, \quad Z_s = X_s U D^{-1/2},$$

where Z_s is the $n \times m$ matrix of standardized component scores:

(25) $$\frac{1}{m-1} Z_s' Z_s \quad = \quad I,$$

(26) $$F = UD^{1/2}, \quad \vec{f}_j' \vec{f}_j = \lambda_j = \text{var}(z_j),$$

where F is the factor loading matrix, and $\vec{f}_j$ is the j-th column of F.

Let λ_1, ..., λ_r be the r largest eigenvalues of R. Let us define the following:

(27) $$U_r \quad = \quad [\vec{u}_1, \ldots, \vec{u}_r], \quad D_r = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{bmatrix},$$

(28) $$\hat{Z}_s = X_s U_r D_r^{-1/2} = [X_s \vec{u}_1 / \sqrt{\lambda_1}, \ldots, X_s \vec{u}_r / \sqrt{\lambda_r}].$$

Thus,

$$\hat{Z}_s \quad = \quad [\vec{z}_{s1}, \ldots, \vec{z}_{sr}],$$

the matrix consisting of the first r columns of Z_s. Note also that

(29) $$RU_r = U_r D_r, \quad \text{and} \quad U_r' U_r = I_{r \times r}.$$

Note that

$$\frac{1}{m-1} \hat{Z}_s' \hat{Z}_s = \frac{1}{n-1} D_r^{-1/2} U_r' X_s' X_s U_r D_r^{-1/2}$$
$$= D_r^{-1/2} U_r' R U_r D_r^{-1/2} = D_r^{-1/2} U_r' U_r D_r D_r^{-1/2} = I_{r \times r}$$

Hence, the variables $\hat{z}_{s1}$, $\hat{z}_{s2}$, ..., $\hat{z}_{sr}$ are standardized and mutually uncorrelated.

$\vec{z}_{s1}$, ..., $\vec{z}_{sr}$ form the basis of the reduced space. The factor loading matrix in the reduced space is given by:

$$\frac{1}{n-1} X_s' \hat{Z}_s \quad = \quad \frac{1}{n-1} X_s' X_s U_r D_r^{-1/2} \quad = \quad R U_r D_r^{-1/2}$$
$$= \quad U_r D_r D_r^{-1/2} \quad = \quad U_r D_r^{1/2} \quad = \quad F_r,$$

that is,

$$(F_r)_{ij} \quad = \quad r(x_i, \hat{z}_{sj})$$

Thus,

$$F_r \quad = \quad [\vec{f}_1, \ldots, \vec{f}_r],$$

that is, it consists of the first r columns of the factor loading matrix $F = UD^{1/2}$.

Approximation of X_s in the reduced space: Define

(30) $$\hat{X}_s \quad = \quad \hat{Z}_s F_r'.$$

The covariance matrix of $\hat{X}_s$ is:

$$\frac{1}{n-1}\hat{X}_s'\hat{X}_s = F_r(\frac{\hat{Z}_s'\hat{Z}_s}{n-1})F_r' = F_rF_r'$$

$$= \lambda_1\vec{u}_1\vec{u}_1' + \ldots + \lambda_r\vec{u}_r\vec{u}_r' \quad = \quad \hat{R}^{(r)}.$$

$\hat{R}^{(r)}$ is called the r-th order approximation to R. It is a symmetric and positive semidefinite matrix.

The *communality* of x_i is defined as the sum of squares of the elements in row i of F_r. This is given by:

$$(F_rF_r')_{ii} \quad = \quad (R^{(r)})_{ii} \quad \leq \quad 1.$$

Finally, since $\vec{f}_1'\vec{f}_1 = \lambda_1$, etc., the sum of squares of the elements of F_r equals $\sum_{i=1}^{r}\lambda_i$, that is, it is a measure of the variance explained by the r principal components retained.

9.4 Bartlett's Sphericity Test

Bartlett's sphericity test addresses the issue of whether factors should be extracted at all. Principal components analysis is not meaningful if all the variables are uncorrelated in the population, that is, the population correlation matrix, R_p is $I_{m\times m}$. In that case, the concentration ellipsoid is a sphere, and every axis is of equal length.

Bartlett's sphericity test tests $H_0 : R_p = I$ against $H_a : R_p \neq I$. If H_0 is true, then $|R_p| = 1$, that is, $\ln|R_p| = 0$.

The determinant of the sample correlation matrix R is given by

$$|R| = \lambda_1 * \lambda_2 * \cdots * \lambda_m, \quad \text{that is,} \quad \ln|R| \; = \sum_{j=1}^{m}\ln\lambda_j,$$

where λ_1, ..., λ_n are the m eigenvalues of R. Since the logarithmic function is strictly concave, it can be shown that $\ln|R| < 0$ if the eigenvalues of R are not all equal.

The Bartlett test uses an approximate chi-square statistic:

$$\chi^2 \;=\; -\{n-1-\frac{2m+5}{6}\}\ln|R| \;=\; -\{n-1-\frac{2m+5}{6}\}\sum_{j=1}^{n}\ln\lambda_j.$$

Under H_0, χ^2 is approximately chi-square distributed with degrees of freedom $\frac{1}{2}(m^2-m)$. Thus, at a $(1-\alpha)$ level of confidence, $H_0 : R_p = I$ is rejected if Bartlett's $\chi^2 > \chi^2_\alpha$ for $\frac{1}{2}(m^2-m)$ degrees of freedom.

9.5 Factor Rotation

Let $J_{r\times r}$ be an orthogonal matrix, that is, $J'J = JJ' = I$. Let $\vec{f_1}$ and $\vec{f_2}$ be $r \times 1$ vectors. $\vec{f_1}'\vec{f_1}$ and $\vec{f_2}'\vec{f_2}$ are the squared lengths of $\vec{f_1}$ and $\vec{f_2}$, respectively, and the cosine of the angle θ between $\vec{f_1}$ and $\vec{f_2}$ is given by:

$$\frac{\vec{f_1}'\vec{f_2}}{\sqrt{\vec{f_1}'\vec{f_1}}\sqrt{\vec{f_2}'\vec{f_2}}} \quad = \quad \cos(\theta).$$

Let $\vec{f_1^*} = J\vec{f_1}$ and $\vec{f_2^*} = J\vec{f_2}$. Then:

$$\vec{f_1^*}'\vec{f_1^*} = \vec{f_1}'J'J\vec{f_1} = \vec{f_1}'\vec{f_1},$$

$$\vec{f_2^*}'\vec{f_2^*} = \vec{f_2}'J'J\vec{f_2} = \vec{f_2}'\vec{f_2},$$

$$\vec{f_1^*}'\vec{f_2^*} = \vec{f_1}'J'J\vec{f_2} = \vec{f_1}'\vec{f_2}.$$

Hence, pre-multiplication by the orthogonal matrix J does not change either the length of a vector, or the angle between two vectors. Thus, all vectors are rotated by the same angle.

Rotating Factors

Note that

$$\hat{X}_s = \hat{Z}_sF_r' = \hat{Z}_sJJ'F_r' = (\hat{Z}_sJ)(F_rJ)',$$

$$\hat{R}^{(r)} = F_rF_r' = F_rJJ'F_r' = (F_rJ)(F_rJ)'$$

Let us define

(31) $$\hat{Z}_s^* \quad = \quad \hat{Z}_sJ.$$

Hence,

$$\frac{1}{m-1}\hat{Z}_s^{*\prime}\hat{Z}_s^* = \frac{1}{n-1}J'\hat{Z}_s'\hat{Z}_sJ = I,$$

that is, the column vectors of $\hat{Z}_s^*$ are orthogonal and standardized.

Let F_r^* be the $n \times r$ matrix where the (i,j)-th term is $r(x_{si}, \hat{z}_{sj}^*)$. Hence,

(32) $$F_r^* = \frac{1}{n-1}X_s'\hat{Z}_s^* = \frac{1}{n-1}X_s'\hat{Z}_sJ = F_rJ.$$

Thus, F_r^* is the *factor loading matrix* of X_s with respect to the rotated factors. Also,

$$\hat{R}^{(r)} = F_r^*F_r^{*\prime} = F_rF_r'$$

Thus,

$$\text{new communality of } x_i \quad = (F_r^*F_r^{*\prime})_{ii} = (F_rF_r')_{ii} \quad = \text{old communality of } x_i.$$

Therefore, for any original variable x_i, a factor rotation leaves the communality unchanged, that is, the explained variance of x_i remains the same. Consequently,

$$\sum_{i=1}^{m}\sum_{j=1}^{r}{F_{ij}^*}^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}F_{ij}^2 = \sum_{j=1}^{r}\lambda_j,$$

that is, rotation does not affect the *total variance explained.*

Selecting Rotation Matrix

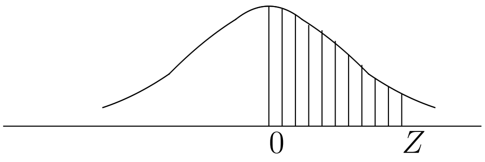
How to choose J? Let $F_r^* = F_r J$, and let G be the $(n \times r)$ matrix such that $G_{ij}^* = (F_r^*)_{ij}^2$. The rotation matrix is chosen such that the elements of G are either large or close to zero. That way, any row or any column of F_r^* has elements which are have either large or small magnitude. When that happens, we say that J has a *simple structure.* Intuitively, each original variable only has strong correlation with a small subset of factors, and vice versa.

Varimax Riotation: The objective of Varimax is to make elements of a column of G vary in magnitude. That way, each factor relates to only a subset of the original variables.

For every column j of G (that is, same factor), let $\overline{G}_{.j} = \dfrac{1}{n}\sum_{i=1}^{m} G_{ij}$. Then, Varimax chooses J to maximize $\sum_{j=1}^{r}\sum_{i=1}^{m}(G_{ij} - \overline{G}_{.j})^2$.

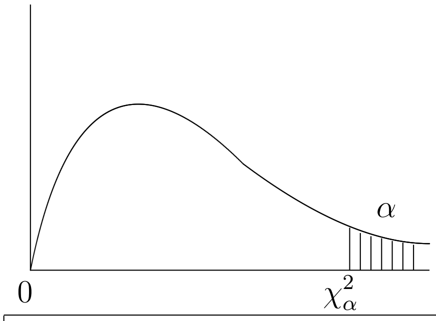
10 Statistical Tables

Table 1. Area Under the Standard Normal (Z) Curve



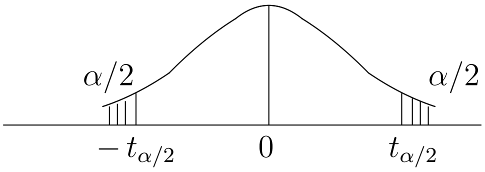
Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3079	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3943	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4773	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table 2. Chi-Square (χ^2) Distribution



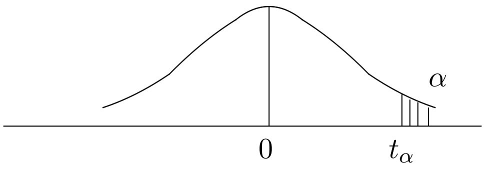
	α				
df	**.10**	**.05**	**.025**	**.01**	**.005**
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.95
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.73	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.65
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67

Table 3a. t Distribution (Two-Tail Probability)
[Table Entries are Values of $t_{\alpha/2}$]



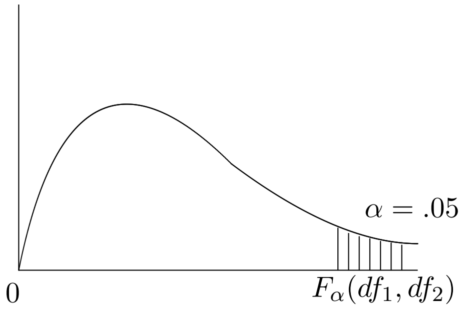
			α		
df	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Table 3b. t Distribution (One-Tail Probability)
[Table Entries are Values of t_α]



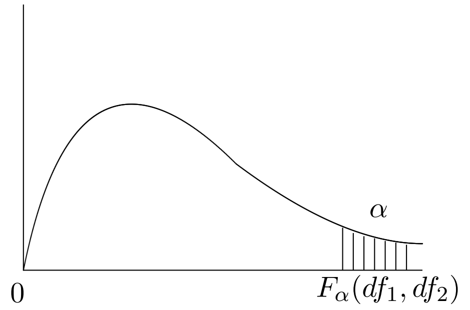
	α			
df	**0.20**	**0.10**	**0.05**	**0.01**
1	1.376	3.078	6.314	31.821
2	1.061	1.886	2.920	6.965
3	0.978	1.638	2.353	4.541
4	0.941	1.533	2.132	3.747
5	0.920	1.476	2.015	3.365
6	0.906	1.440	1.943	3.143
7	0.896	1.415	1.895	2.998
8	0.889	1.397	1.860	2.896
9	0.883	1.383	1.833	2.821
10	0.879	1.372	1.812	2.764
11	0.876	1.363	1.796	2.718
12	0.873	1.356	1.782	2.681
13	0.870	1.350	1.771	2.650
14	0.868	1.345	1.761	2.624
15	0.866	1.341	1.753	2.602
16	0.865	1.337	1.746	2.583
17	0.863	1.333	1.740	2.567
18	0.862	1.330	1.734	2.552
19	0.861	1.328	1.729	2.539
20	0.860	1.325	1.725	2.528
21	0.859	1.323	1.721	2.518
22	0.858	1.321	1.717	2.508
23	0.858	1.319	1.714	2.500
24	0.857	1.318	1.711	2.492
25	0.856	1.316	1.708	2.485
26	0.856	1.315	1.706	2.479
27	0.855	1.314	1.703	2.473
28	0.855	1.313	1.701	2.467
29	0.854	1.311	1.699	2.462
30	0.854	1.310	1.697	2.457
40	0.851	1.303	1.684	2.423
60	0.848	1.296	1.671	2.390
120	0.845	1.289	1.658	2.358
∞	0.842	1.282	1.645	2.326

Table 4a. F Distribution ($\alpha = .05$)
(df_1 = Degrees of Freedom of Numerator, df_2 = Degrees of Freedom of Denominator)



	df_1											
df_2	1	2	3	4	5	6	7	8	9	10	12	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67

Table 4b. F Distribution ($\alpha = .01$)
(df_1 = Degrees of Freedom of Numerator, df_2 = Degrees of Freedom of Denominator)



	df_1											
df_2	1	2	3	4	5	6	7	8	9	10	12	15
1	4052	4999	5403	5624	5764	5859	5928	5981	6022	6056	6107	6157
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04