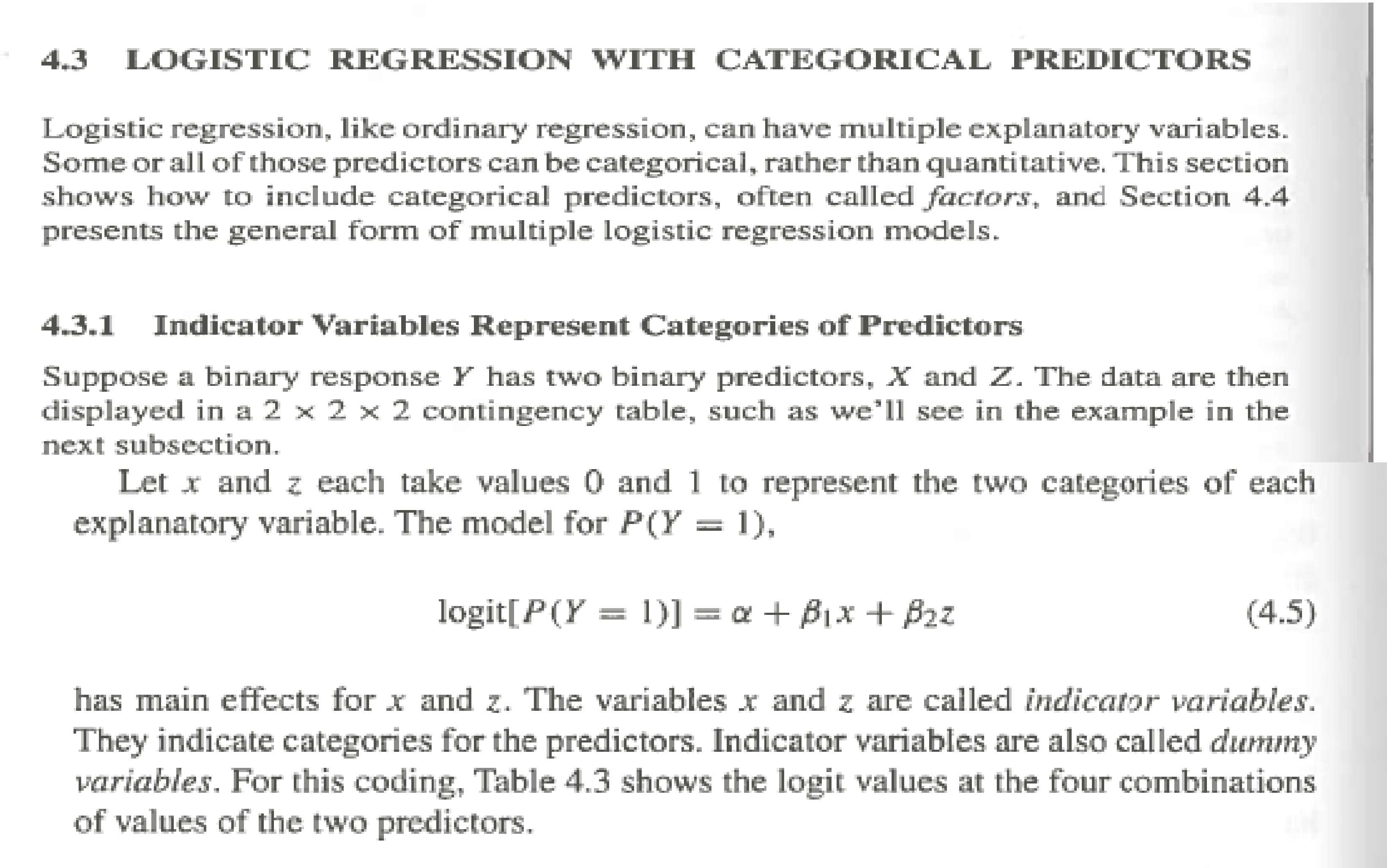
Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: John Wiley and Sons, Inc.

Chapter 4: Logistic Regression

Logistic Regression with Categorical Predictors



4.3 LOGISTIC REGRESSION WITH CATEGORICAL PREDICTORS

Logistic regression, like ordinary regression, can have multiple explanatory variables. Some or all of those predictors can be categorical, rather than quantitative. This section shows how to include categorical predictors, often called *factors*, and Section 4.4 presents the general form of multiple logistic regression models.

4.3.1 Indicator Variables Represent Categories of Predictors

Suppose a binary response Y has two binary predictors, X and Z. The data are then displayed in a 2 × 2 × 2 contingency table, such as we'll see in the example in the next subsection.

Let x and z each take values 0 and 1 to represent the two categories of each explanatory variable. The model for $P(Y = 1)$,

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z \qquad (4.5)$$

has main effects for x and z. The variables x and z are called *indicator variables*. They indicate categories for the predictors. Indicator variables are also called *dummy variables*. For this coding, Table 4.3 shows the logit values at the four combinations of values of the two predictors.

Effect of one factor (predictor) is the same at each category of the other factor

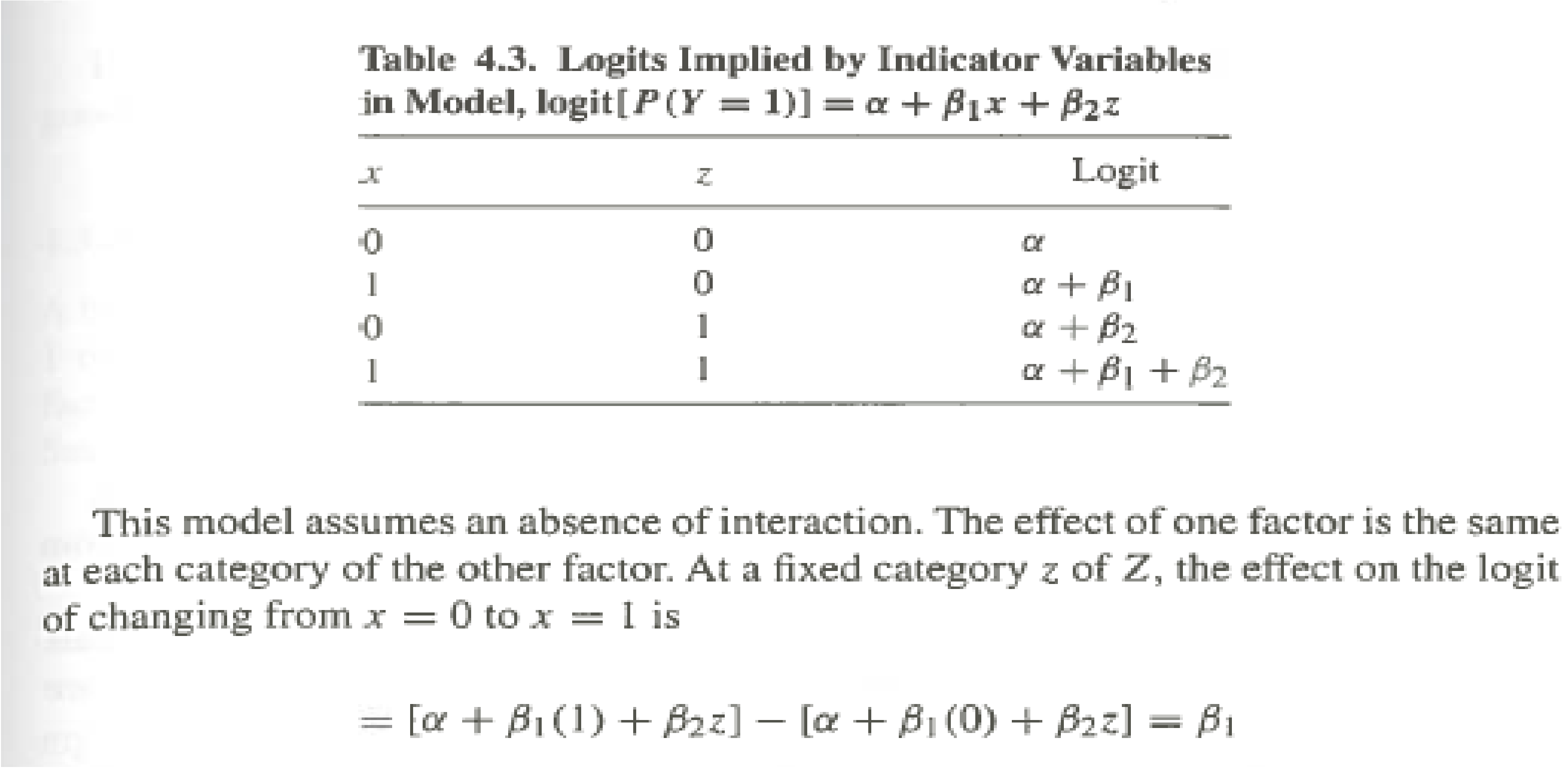


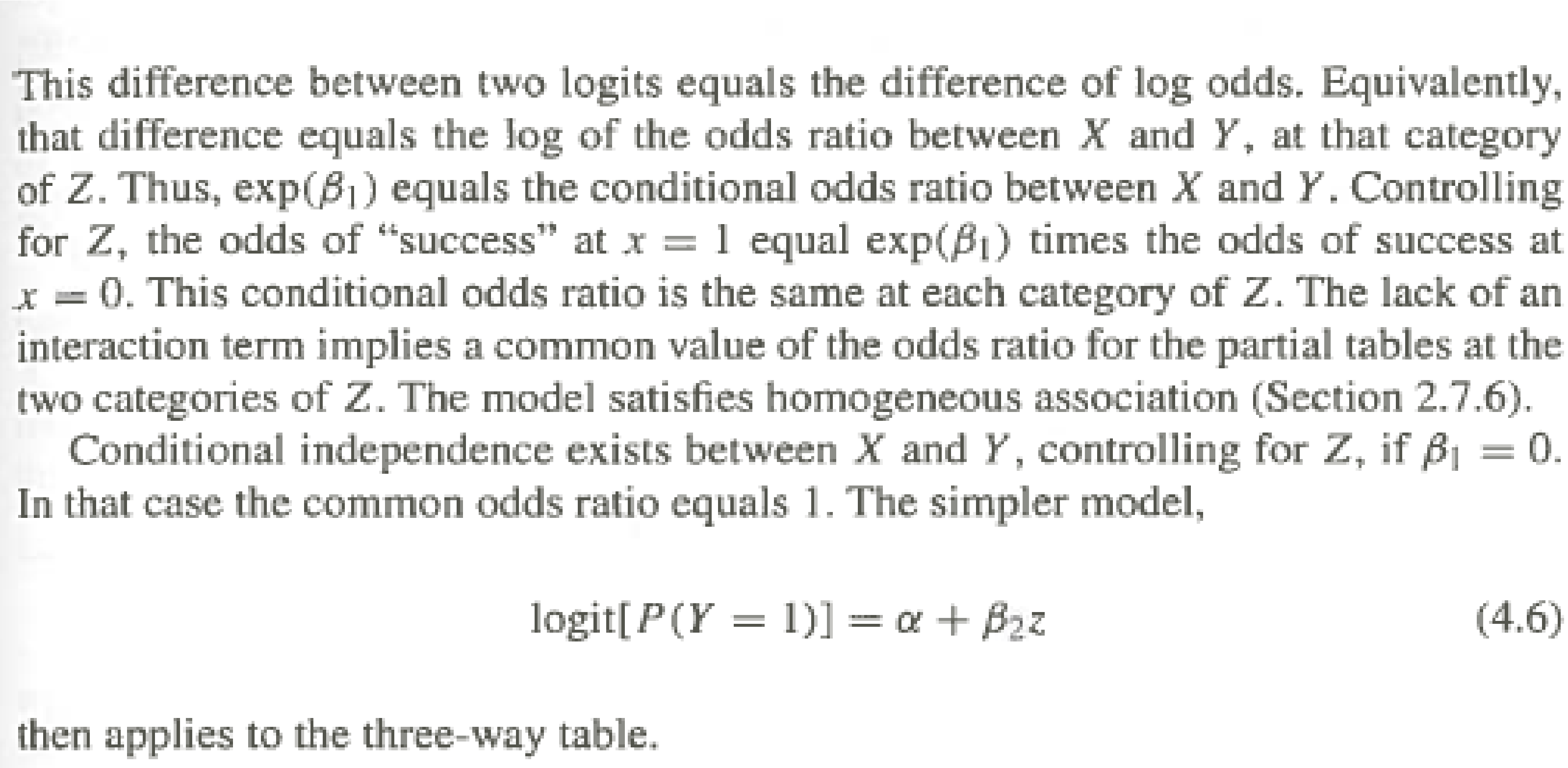
Table 4.3. Logits Implied by Indicator Variables in Model, logit$[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$

x	z	Logit
0	0	α
1	0	$\alpha + \beta_1$
0	1	$\alpha + \beta_2$
1	1	$\alpha + \beta_1 + \beta_2$

This model assumes an absence of interaction. The effect of one factor is the same at each category of the other factor. At a fixed category z of Z, the effect on the logit of changing from $x = 0$ to $x = 1$ is

$$= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1$$

Exp (B_1) equals the Conditional Odds Ratio Between X and Y (Z has no effect)



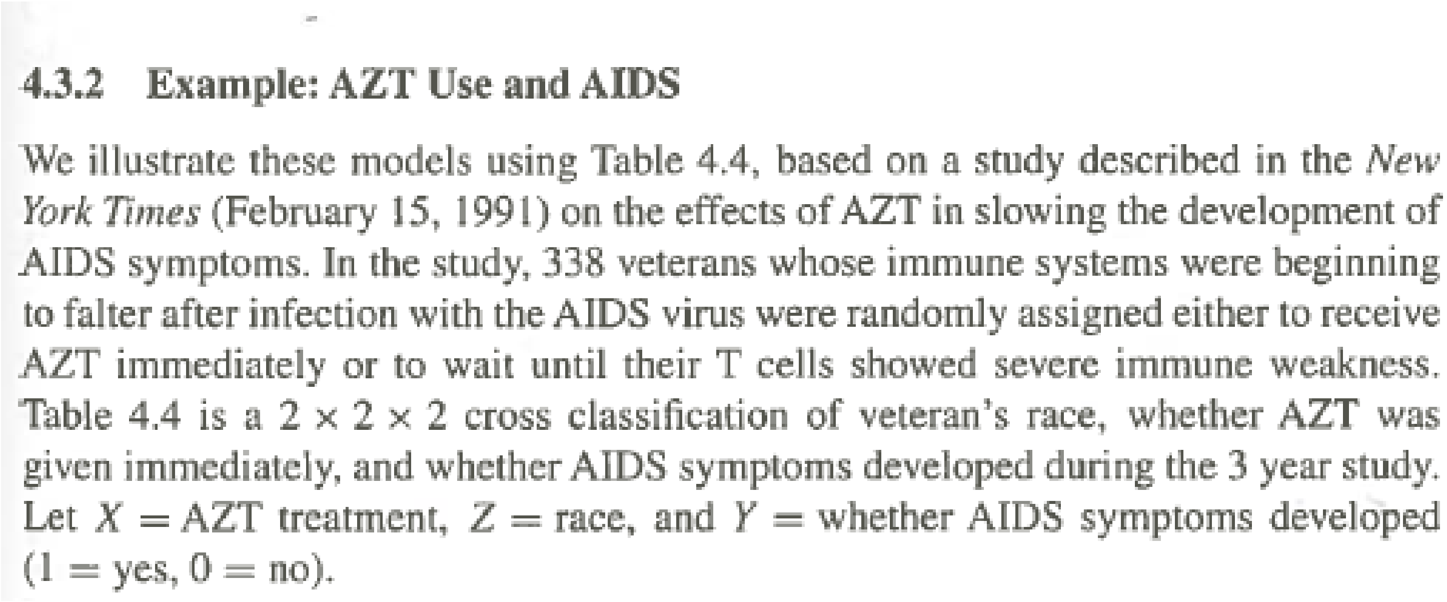
This difference between two logits equals the difference of log odds. Equivalently, that difference equals the log of the odds ratio between X and Y, at that category of Z. Thus, $\exp(\beta_1)$ equals the conditional odds ratio between X and Y. Controlling for Z, the odds of "success" at $x = 1$ equal $\exp(\beta_1)$ times the odds of success at $x = 0$. This conditional odds ratio is the same at each category of Z. The lack of an interaction term implies a common value of the odds ratio for the partial tables at the two categories of Z. The model satisfies homogeneous association (Section 2.7.6).

Conditional independence exists between X and Y, controlling for Z, if $\beta_1 = 0$. In that case the common odds ratio equals 1. The simpler model,

$$\text{logit}[P(Y = 1)] = \alpha + \beta_2 z \tag{4.6}$$

then applies to the three-way table.

Using Three – Way Contingency Table to Determine if Interaction Term Needed in Logistic Regression Model



4.3.2 Example: AZT Use and AIDS

We illustrate these models using Table 4.4, based on a study described in the *New York Times* (February 15, 1991) on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Table 4.4 is a 2 × 2 × 2 cross classification of veteran's race, whether AZT was given immediately, and whether AIDS symptoms developed during the 3 year study. Let X = AZT treatment, Z = race, and Y = whether AIDS symptoms developed (1 = yes, 0 = no).

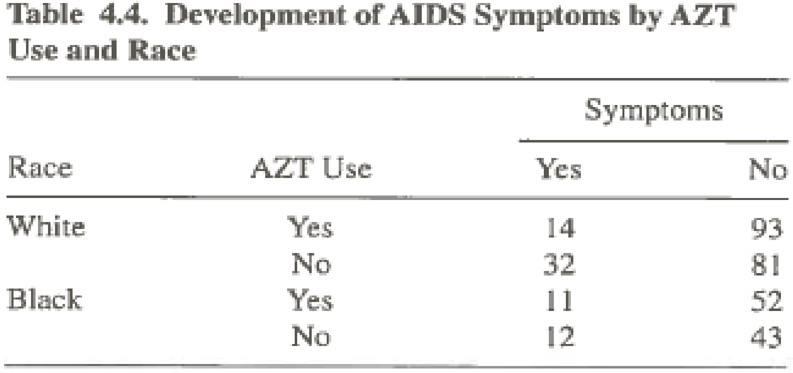
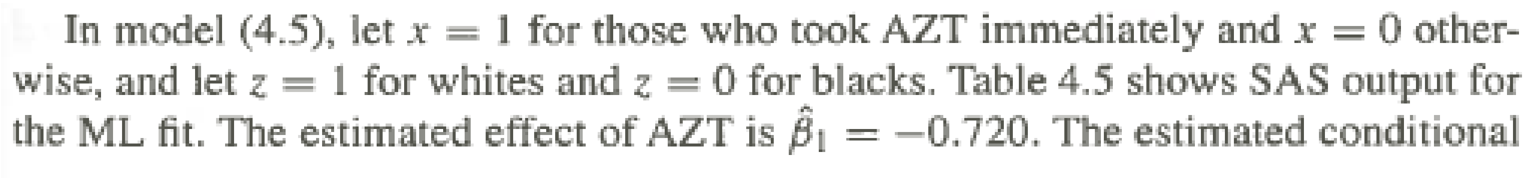


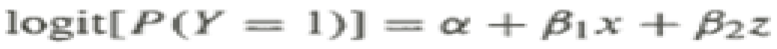
Table 4.4. Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

2 x 2 x 2 Table To Control for Confounding

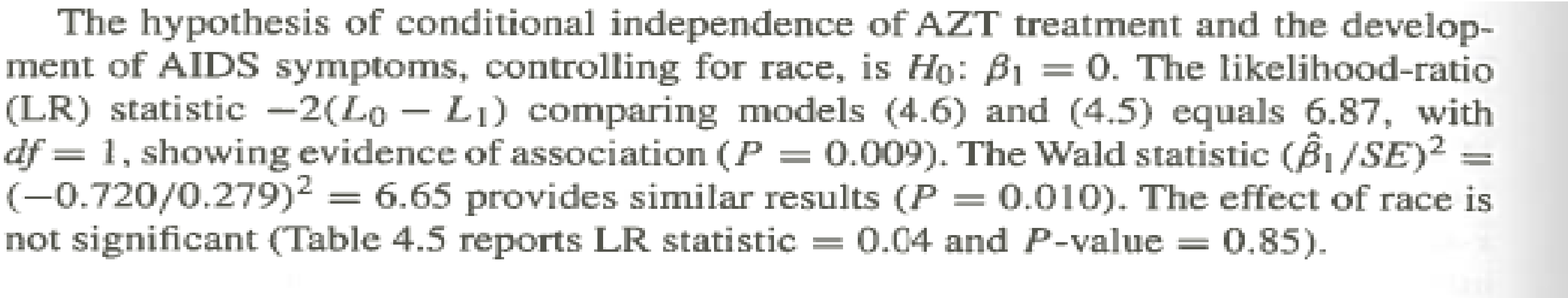


In model (4.5), let $x = 1$ for those who took AZT immediately and $x = 0$ otherwise, and let $z = 1$ for whites and $z = 0$ for blacks. Table 4.5 shows SAS output for the ML fit. The estimated effect of AZT is $\hat{\beta}_1 = -0.720$. The estimated conditional



$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$$

Controlling for Race: Not Significant



The hypothesis of conditional independence of AZT treatment and the development of AIDS symptoms, controlling for race, is H_0: $\beta_1 = 0$. The likelihood-ratio (LR) statistic $-2(L_0 - L_1)$ comparing models (4.6) and (4.5) equals 6.87, with $df = 1$, showing evidence of association ($P = 0.009$). The Wald statistic $(\hat{\beta}_1/SE)^2 = (-0.720/0.279)^2 = 6.65$ provides similar results ($P = 0.010$). The effect of race is not significant (Table 4.5 reports LR statistic = 0.04 and P-value = 0.85).

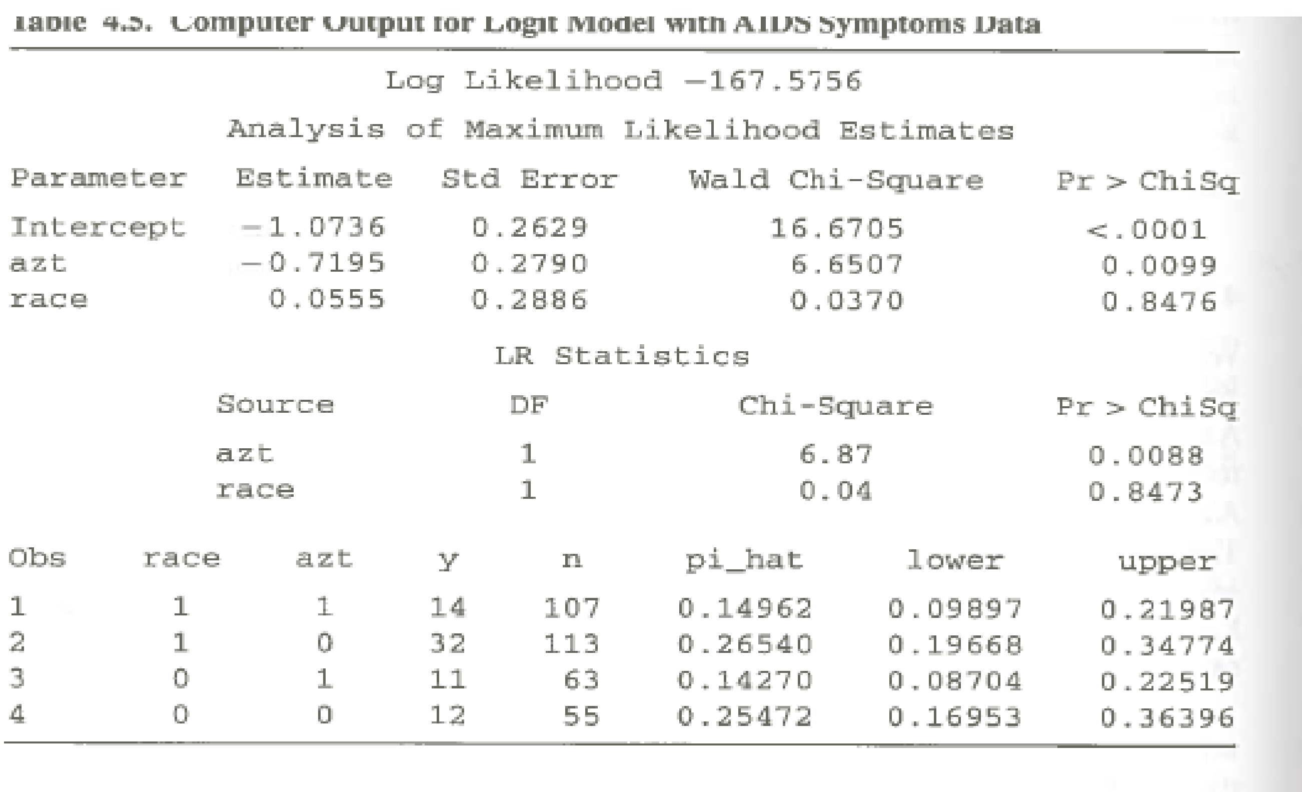
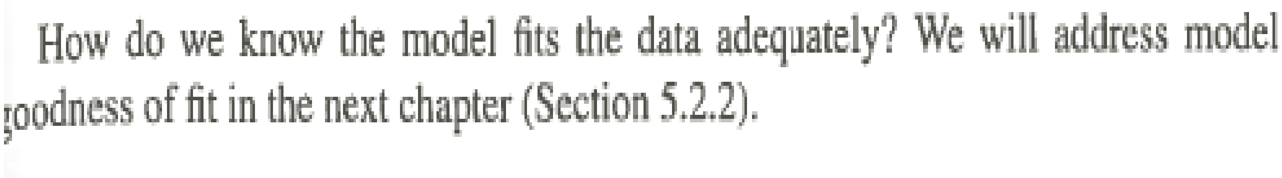


Table 4.5. Computer Output for Logit Model with AIDS Symptoms Data

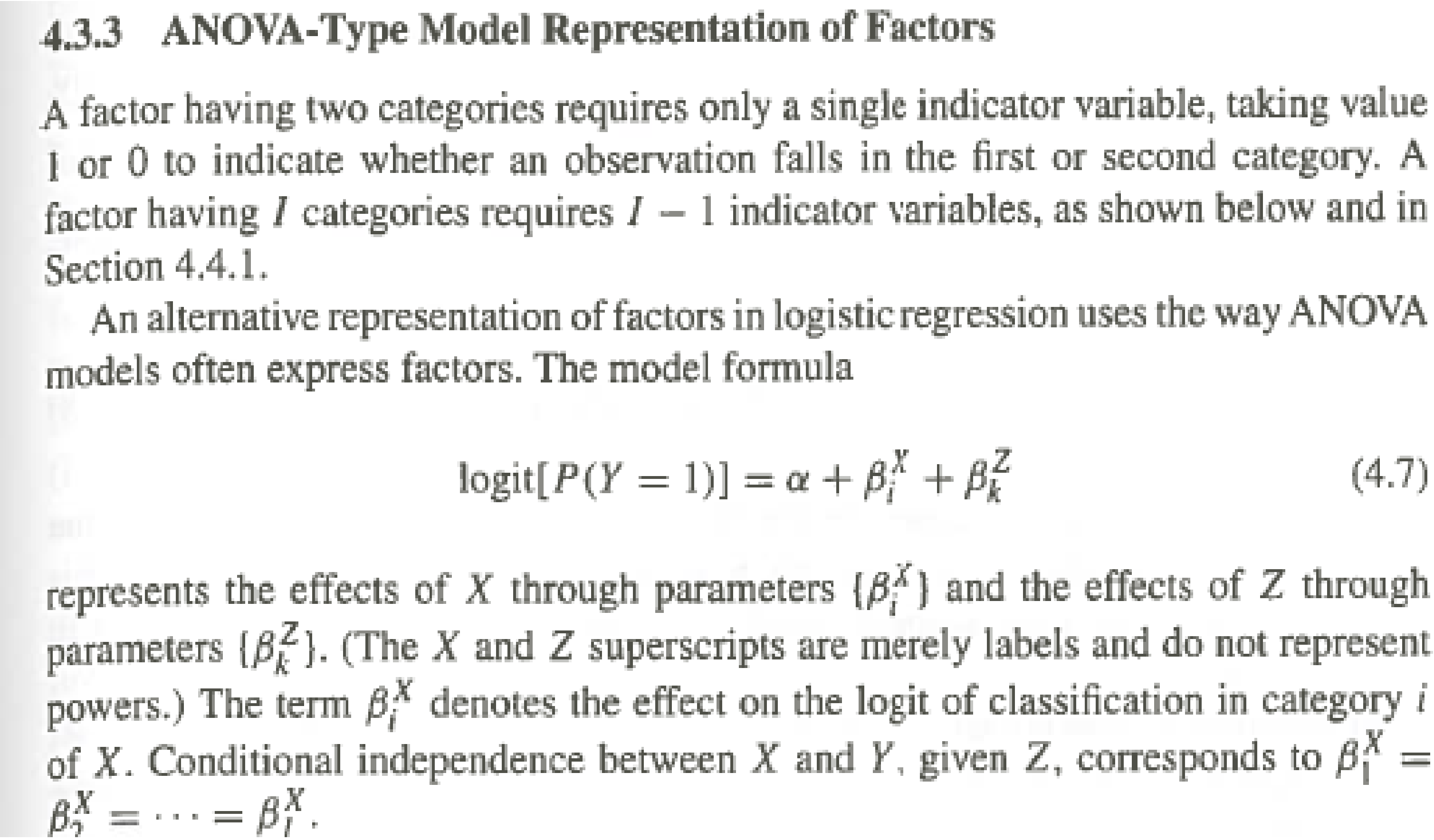
```
                    Log Likelihood -167.5756

              Analysis of Maximum Likelihood Estimates

Parameter   Estimate   Std Error   Wald Chi-Square   Pr > ChiSq
Intercept   -1.0736     0.2629        16.6705         <.0001
azt         -0.7195     0.2790         6.6507          0.0099
race         0.0555     0.2886         0.0370          0.8476

                          LR Statistics

          Source          DF          Chi-Square      Pr > ChiSq
          azt              1             6.87           0.0088
          race             1             0.04           0.8473

Obs    race    azt    y      n     pi_hat     lower      upper
1       1      1     14    107    0.14962    0.09897    0.21987
2       1      0     32    113    0.26540    0.19668    0.34774
3       0      1     11     63    0.14270    0.08704    0.22519
4       0      0     12     55    0.25472    0.16953    0.36396
```

Model Fit: Goodness of Fit Large X^2 (M) or G^2 (M) (introduced in Equation 2.7 and Deviance Using Likelihood Ratio Test Section 3.4.3)



How do we know the model fits the data adequately? We will address model goodness of fit in the next chapter (Section 5.2.2).

ANOVA-Type Model Representation of Factors



4.3.3 ANOVA-Type Model Representation of Factors

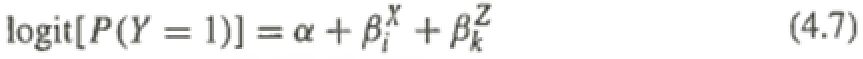
A factor having two categories requires only a single indicator variable, taking value 1 or 0 to indicate whether an observation falls in the first or second category. A factor having I categories requires $I - 1$ indicator variables, as shown below and in Section 4.4.1.

An alternative representation of factors in logistic regression uses the way ANOVA models often express factors. The model formula

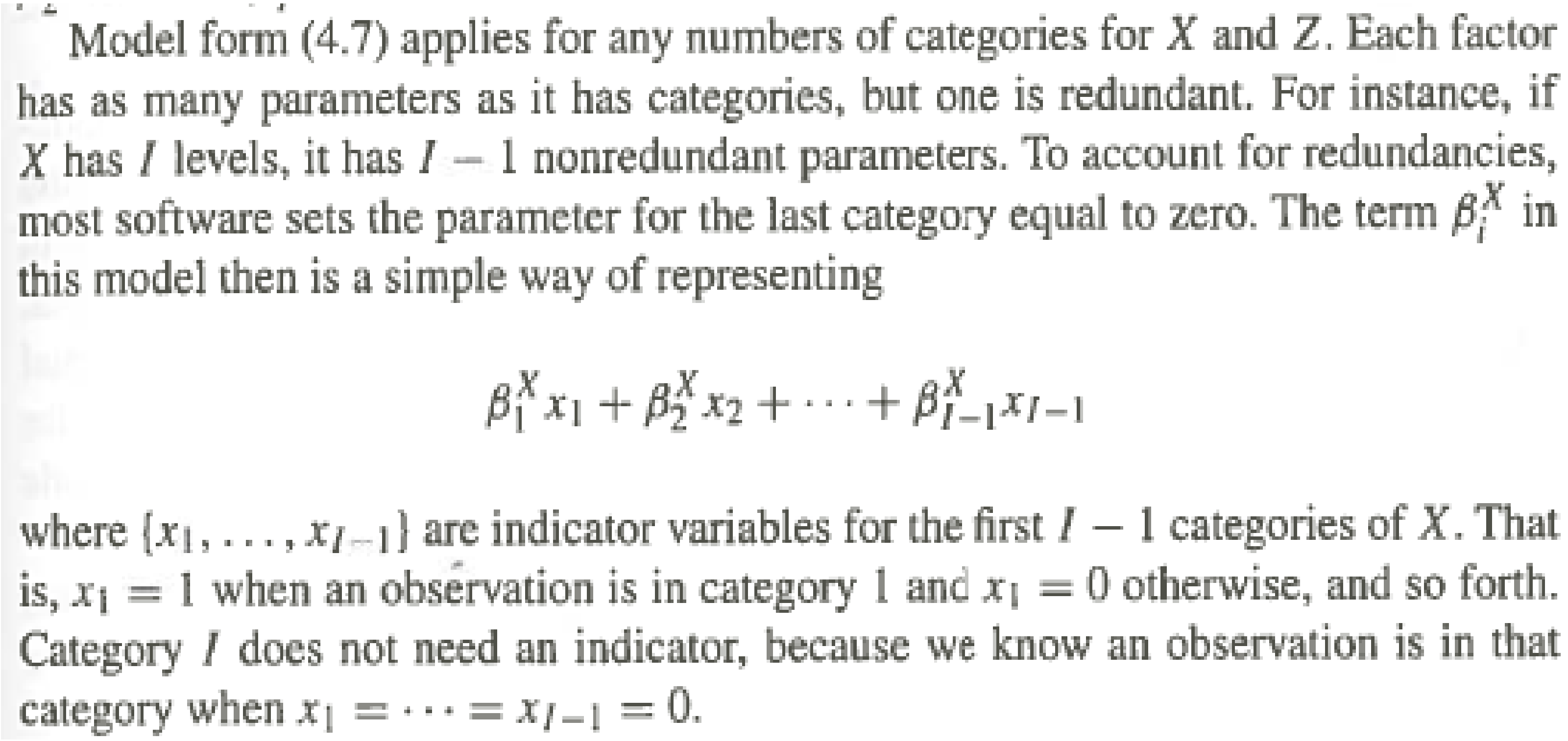
$$\text{logit}[P(Y = 1)] = \alpha + \beta_i^X + \beta_k^Z \tag{4.7}$$

represents the effects of X through parameters $\{\beta_i^X\}$ and the effects of Z through parameters $\{\beta_k^Z\}$. (The X and Z superscripts are merely labels and do not represent powers.) The term β_i^X denotes the effect on the logit of classification in category i of X. Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$.

Each variable (factor) has as many parameters as it has categories, with one redundant, so software sets parameters last category equal to zero



$$\text{logit}[P(Y = 1)] = \alpha + \beta_i^X + \beta_k^Z \qquad (4.7)$$

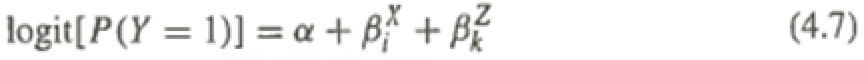


Model form (4.7) applies for any numbers of categories for X and Z. Each factor has as many parameters as it has categories, but one is redundant. For instance, if X has I levels, it has $I - 1$ nonredundant parameters. To account for redundancies, most software sets the parameter for the last category equal to zero. The term β_i^X in this model then is a simple way of representing

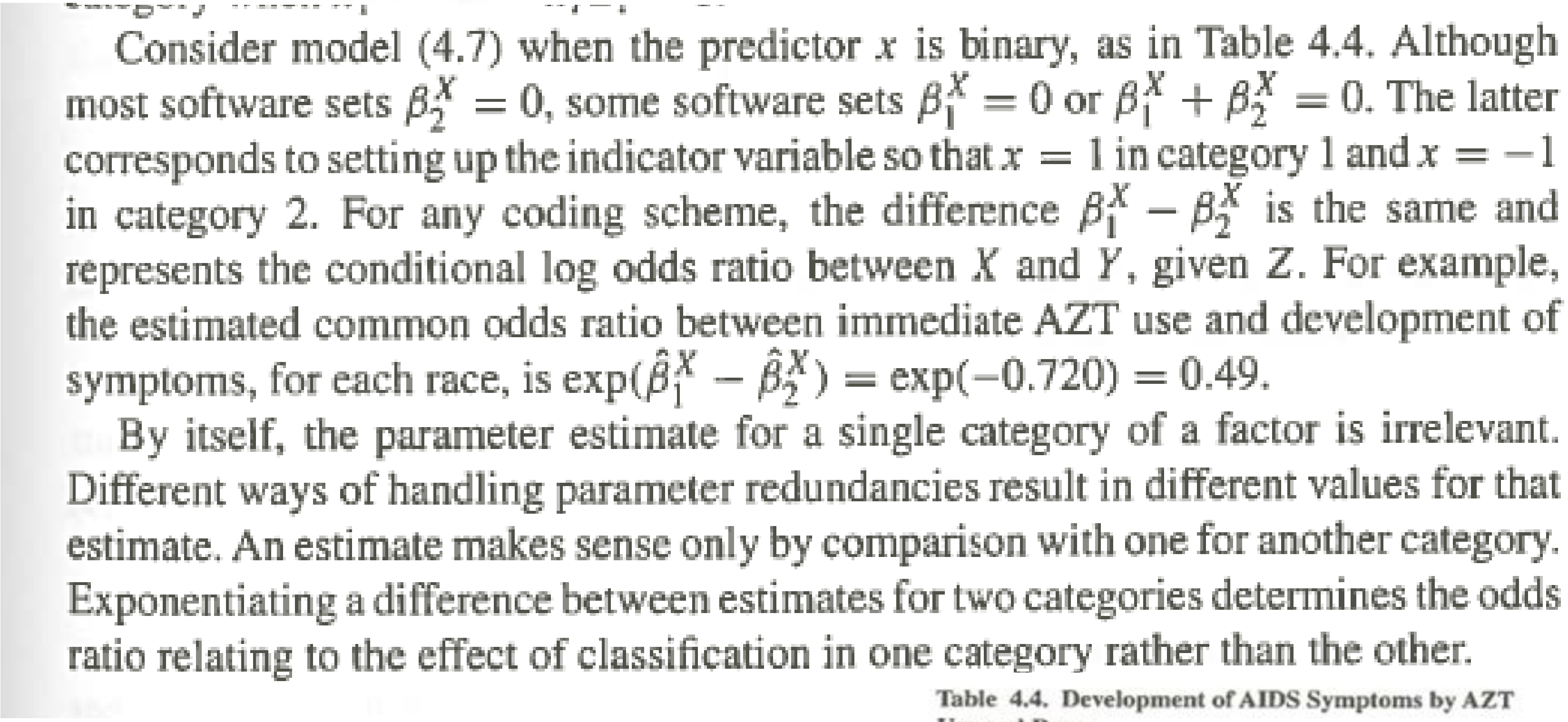
$$\beta_1^X x_1 + \beta_2^X x_2 + \cdots + \beta_{I-1}^X x_{I-1}$$

where $\{x_1, \ldots, x_{I-1}\}$ are indicator variables for the first $I - 1$ categories of X. That is, $x_1 = 1$ when an observation is in category 1 and $x_1 = 0$ otherwise, and so forth. Category I does not need an indicator, because we know an observation is in that category when $x_1 = \cdots = x_{I-1} = 0$.

Omit One Category for Comparison (Reference Level) i.e. Dummy Variable k-1 categories



$$\text{logit}[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z \qquad (4.7)$$



Consider model (4.7) when the predictor x is binary, as in Table 4.4. Although most software sets $\beta_2^X = 0$, some software sets $\beta_1^X = 0$ or $\beta_1^X + \beta_2^X = 0$. The latter corresponds to setting up the indicator variable so that $x = 1$ in category 1 and $x = -1$ in category 2. For any coding scheme, the difference $\beta_1^X - \beta_2^X$ is the same and represents the conditional log odds ratio between X and Y, given Z. For example, the estimated common odds ratio between immediate AZT use and development of symptoms, for each race, is $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = \exp(-0.720) = 0.49$.

By itself, the parameter estimate for a single category of a factor is irrelevant. Different ways of handling parameter redundancies result in different values for that estimate. An estimate makes sense only by comparison with one for another category. Exponentiating a difference between estimates for two categories determines the odds ratio relating to the effect of classification in one category rather than the other.

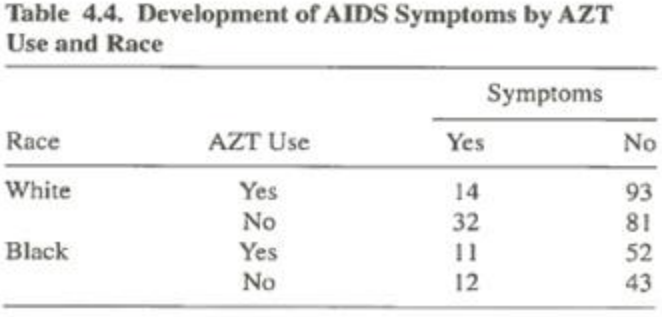


Table 4.4. Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43