Data Analysis
Binary Outcomes

James Ware, PhD
Harvard School of Public Health

Methods for Binary Outcomes

Measures of association

- Risk difference
- Relative risk
- Odds ratio

Multivariable analysis

- Logistic regression

Preventing Microbial Colonization in Hospitalized Patients

A Comparison of Two Antimicrobial-Impregnated Central Venous Catheters

NEJM, 1/7/1999

This RCT in catheterized patients compared the rates of catheter colonization in central venous catheters impregnated with either

chlorhexidine + silver sulfadiazine, or
minocycline + rifampin

Univariate Analysis of Binary Outcomes

For small to moderate sample sizes, use Fisher's exact test to test for an association between treatment assignment and outcome.

The adjusted chi-square statistic provides a good approximation to the exact test

Use the unadjusted chi-square test for large sample sizes

Univariate Analysis

	Chlorhexadine + Silver Sulfadiazine	Minocycline + Rifampin	Total
Colonized	87 (22.8%)	28 (7.9%)	115
Not Colonized	295	328	623
Total	382	356	738

Fisher's Exact Test: $P < 0.0001$
Uncorrected Chi-Square: 31.14, $P < 0.0001$
Corrected Chi-Square: 30.02, $P < 0.0001$

(Odds Ratio = 3.44; 95% CI, 2.18 to 5.37; $P<0.001$)

Multivariable Analysis of Binary Outcomes

To model the effect of covariates on the probability of a positive response, we need a generalization of multiple linear regression

The most widely-used approach, multiple logistic regression, assumes that

$$\log[p/(1-p)]$$

is a linear function of the covariates

Multiple Logistic Regression

log[p/(1-p)] is called the logit of p

We assume

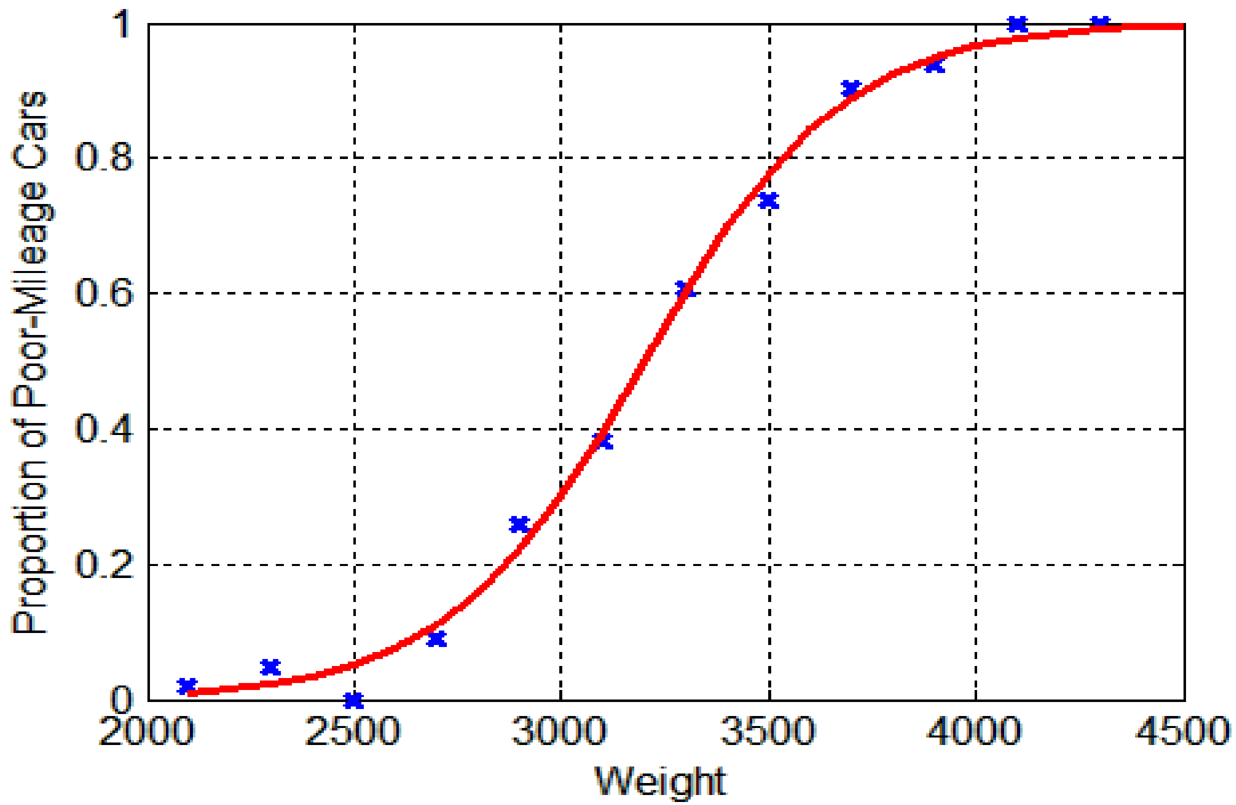
$$\log[p/(1 - p)] = \alpha + \Sigma\, \beta_j X_{ij}$$

This can be written in terms of the probability of success as

$$p(y=1\,|\,\mathbf{X}) = e^{\alpha + \beta'\mathbf{X}}/(1 + e^{\alpha + \beta'\mathbf{X}})$$

where $\beta'\mathbf{X} = \Sigma\beta_j X_{ij}$

The Logistic Regression Function



Here, cars are classified as “poor mileage” (1) or “not poor mileage” (0). The logistic regression model provides a good fit to the proportion of poor mileage vehicles as a function of weight.

Multiple Logistic Regression

If X_{i1} is an indicator variable for treatment group and we have only one other covariate, X_{i2},

$$\text{logit}(p \mid X_{i1} = 1, X_{i2}) - \text{logit}(p \mid X_{i1} = 0, X_{i2})$$

$$= (\alpha + \beta_1 + \beta_2 X_{i2}) - (\alpha + \beta_2 X_{i2})$$
$$= \beta_1$$

β_1 represents the natural logarithm of the odds ratio for an event in group 2 ($X_{i1} = 1$) relative to group 1 ($X_{i1} = 0$), adjusting for the values of X_{2i}

Antimicrobial Catheters

The investigators used multiple logistic regression to estimate the effects of several risk factors on the incidence of catheter colonization

Variables significant at a P value of 0.25 or less in univariate analysis were entered into the model in stepwise fashion.

Multiple Logistic Regression Model

Variable	OR	CI	P
chlor/silv cath	2.80	1.68,4.66	<0.001
fem/jug	3.05	1.86,5.01	<0.001
ICU hosp	2.60	1.47,4.62	<0.001
male	2.45	1.43,4.20	<0.001
Mech Vent	1.97	1.14,3.41	0.01

Modeling the Effects of Covariates

Covariates can modify treatment effects in two ways

Confounding: Failure to adjust for imbalances between treatment groups can introduce bias

Effect Modification: The effects of treatment may not be constant across subgroups defined by covariates

Effect Modification = Interaction

Prognostic Factors May be More Important than Treatment

Performance Status and Disease Extent as Determinants of Survival in Non Small Cell Lung Cancer

Median Survival (Wks) by Prognostic Factors

Perf Status	Ltd Disease	Ext Disease
0	50.9	26.4
1	30.9	18.7
2	20.9	12.1
3	4.9	6.1

Testing for Goodness of Fit

1. Test for linearity on the logit scale by adding quadratic or other nonlinear terms to see whether they enter the model

2. Use the Hosmer-Lemeshow test for goodness of fit of logistic regression. Divide the sample by estimated probability and compare fitted to observed proportions