STAT 545/GS01 1233 193 5528: Generalized Linear Models and Categorical Data Analysis (Part II)
3: Logistic regression

Ziyi Li

Department of Biostatistics
The University of Texas MD Anderson Cancer Center

ZLi16@mdanderson.org

Fall 2022

Logistic Regression Model: Overview

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{1}{1 + \exp[-(\alpha + \beta x)]} \in (0, 1)$$

1. Perhaps more widely used in clinical research/epidemiology/population health than any other regression models
2. Example: Y = hospital death after a cardiac surgery; X = age, pre-operative co-morbidities, type of operations, intra-operative characteristics, etc.
3. Y should NOT be deterministic given X, i.e., the probability $\pi(X)$ cannot be 0 or 1
4. α is the intercept, β is the **log odds ratio**. Why?
5. $\pi(x)$ takes a standard logistic function form (with a location scale shift)

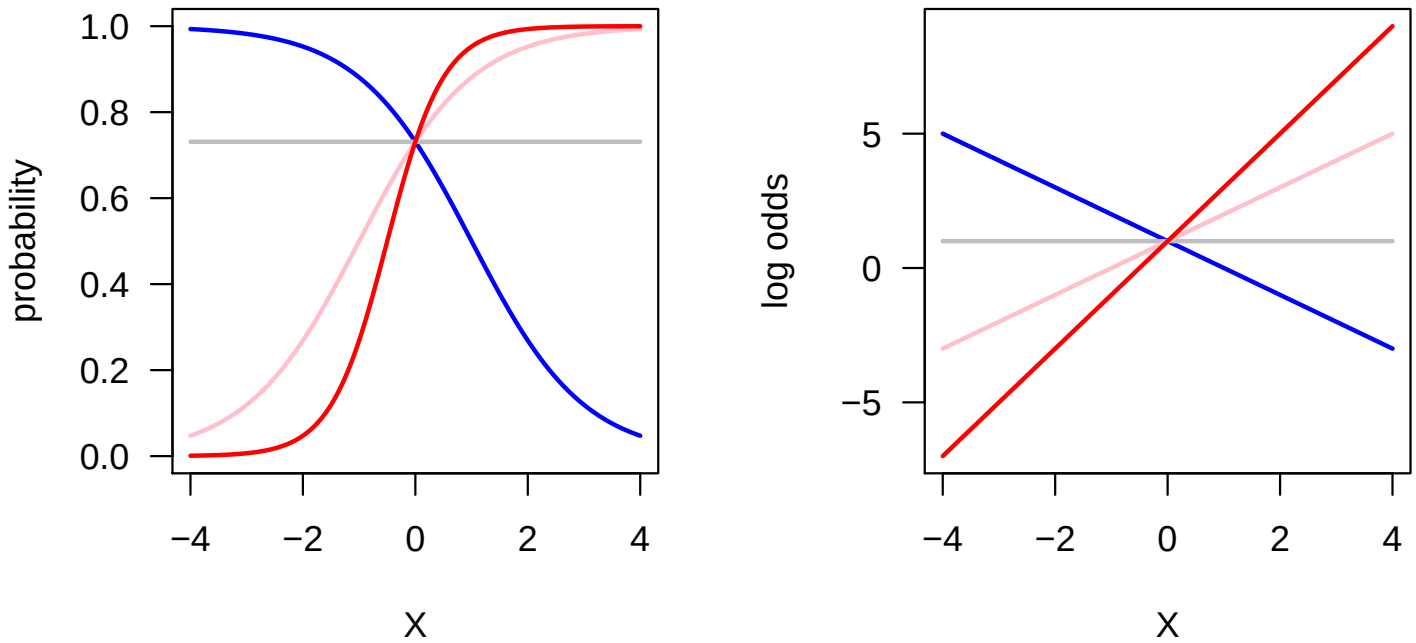
Logistic Regression Model: Interpretation

$$\text{odds} \quad \frac{\pi(X)}{1-\pi(X)} = \exp(\alpha + \beta X) \in (0, \infty)$$

$$\text{log odds} \quad log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = \alpha + \beta X \in (-\infty, \infty)$$

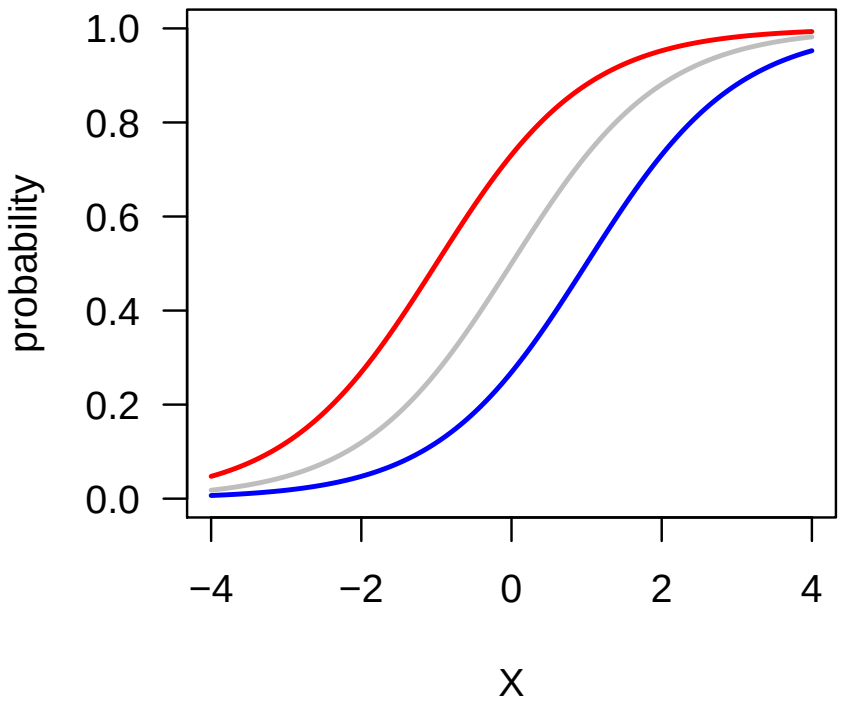
1. For $X = x$ vs. $X = x + 1$, the difference in log odds of observing $Y = 1$ is β. Therefore, $exp(\beta)$ is **odds ratio**, and β is **log odds ratio**
2. How about multiple covariates?
   $\text{logit}\,[\pi(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$

Illustrating Logistic Regression



Plot logistic regression model with $\alpha = 1$ and $\beta = -1$ (blue), 0 (gray), 1 (pink), and 2 (red). Left: probability. Right: log odds. Note that logistic function changes steepness with β but always maintains symmetry.

Illustrating Logistic Regression (cont.)



Plot logistic regression model with $\alpha = -1$ (blue), 0 (gray), 1 (red), and $\beta = 1$. Note that logistic function shift up and down with α but always maintains symmetry.

Data example

Association between snoring and heart disease:

	Heart Disease	
Snoring	Yes	No
Never	24	1355
Occasionally	35	603
Nearly every night	21	192
Every night	30	224

- Treat the rows of the tables as independent binomial samples and create a score for the level of snoring, x, $(0, 2, 4, 5)$.
- Fit a simple logistic regression model:

$$logit[\pi(x)] = \alpha + \beta x$$

Data example

Association between snoring and heart disease:

- Using R function glm or sas procedure PROC GENMOD, we obtain

$$logit[\widehat{\pi}(x)] = -3.87 + 0.40x$$

- $\widehat{\beta} = 0.40$ can be interpreted as follows: The odds ratio of having heart disease is $\exp(0.40) = 1.492$ as x increases by 1, e.g. when the snoring status changes from "Nearly every night" $(x = 4)$ to "Every night" $(x = 5)$.
- Fitted values:

	Heart Disease		Yes	Logit
Snoring	Yes	No	Proportion	Fit
Never	24	1355	0.017	0.021
Occasionally	35	603	0.055	0.044
Nearly every night	21	192	0.099	0.093
Every night	30	224	0.118	0.132

Data example

- An alternative to the logit link function is the probit link function:

$$\Phi^{-1}[\pi(x)] = \alpha + \beta x$$

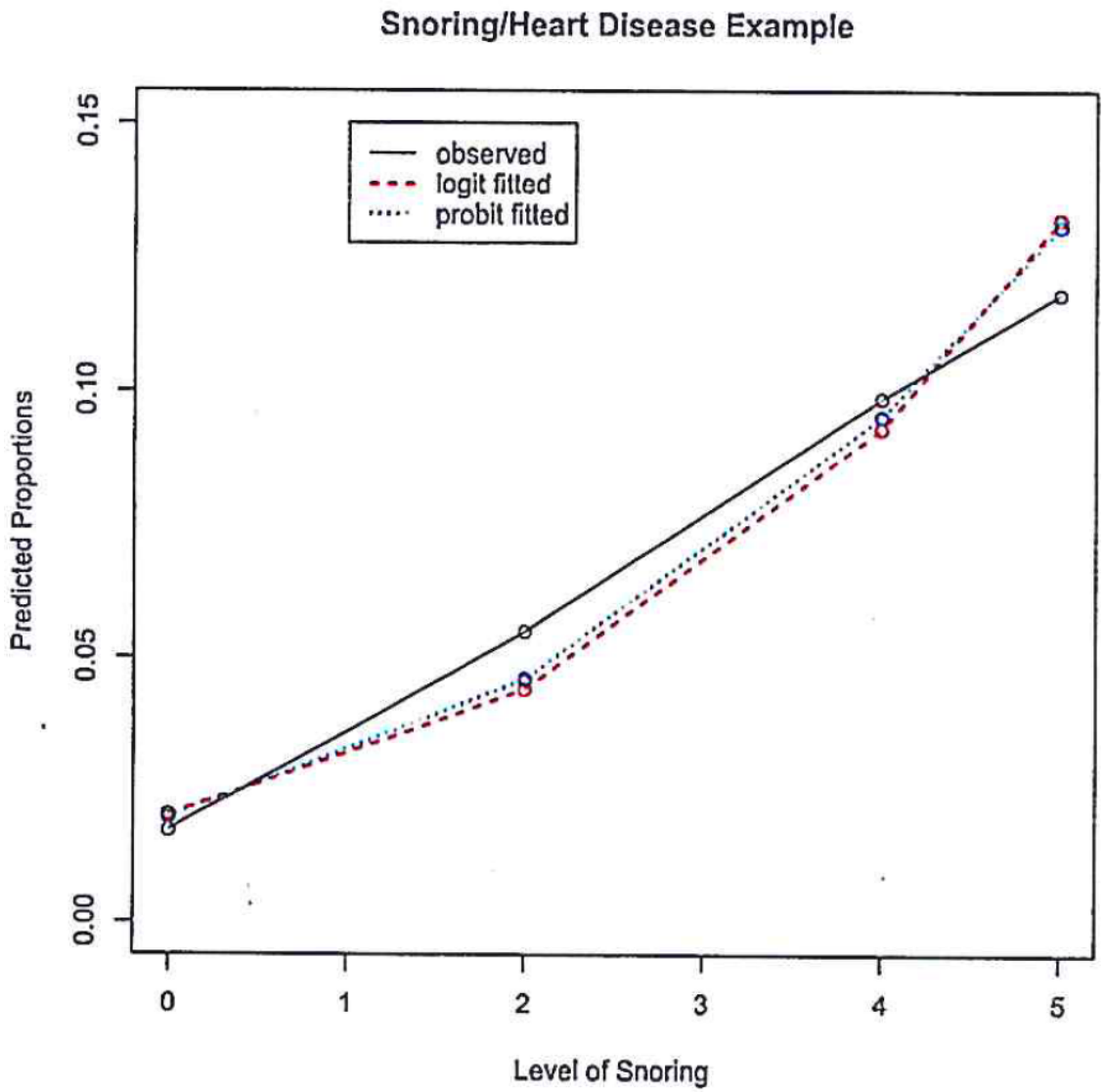
where $\Phi(\cdot)$ denotes the cdf for standard normal.

- Revisit our example on association between snoring and heart attack: Using R function *glm* with Probit link option or sas procedure PROC GENMOD with the probit link option, we have

$$\Phi^{-1}[\pi(x)] = -2.16 + 0.19x$$

	Heart Disease		Logit	Probit
Snoring	Yes	No	Fit	Fit
Never	24	1355	0.021	0.020
Occasionally	35	603	0.044	0.046
Nearly every night	21	192	0.093	0.095
Every night	30	224	0.132	0.130

Data Example



Simple Visual Model Checking by Grouping

1. Group the (continuous) covariate into 10 categories by cutoffs at the quantiles, with n_i subjects in each group ($i = 1, 2, ..., 10$)
2. May use more than 10 groups if the sample size is large
3. Calculate the average covariate within each group (x_i)
4. Calculate the number of observations in group i, n_i, and number of "1" outcomes in that group, y_i.
5. Plot x_i vs. $\log \frac{y_i + 0.5}{n_i - y_i + 0.5}$. It should be approximately a straight line because the latter estimates $logit(p_i)$
6. Note: the 0.5 is a correction needed when $y_i = 0$ or n_i.
7. Only work with a single continuous covariate; analogous to scatter plot for linear regression

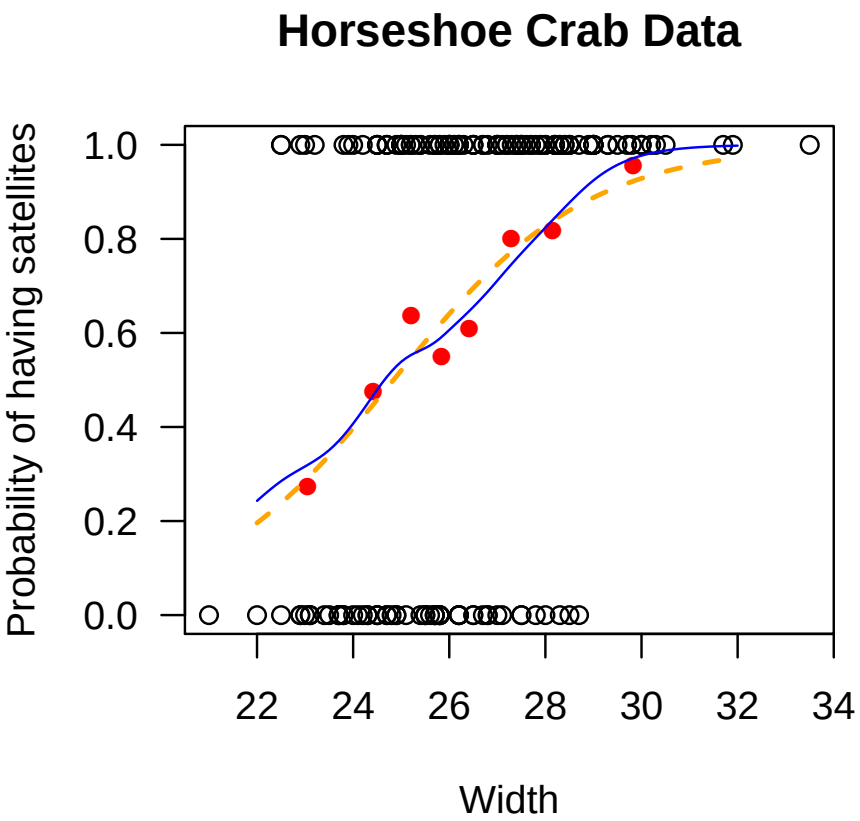
Model Checking by Generalized Additive Model

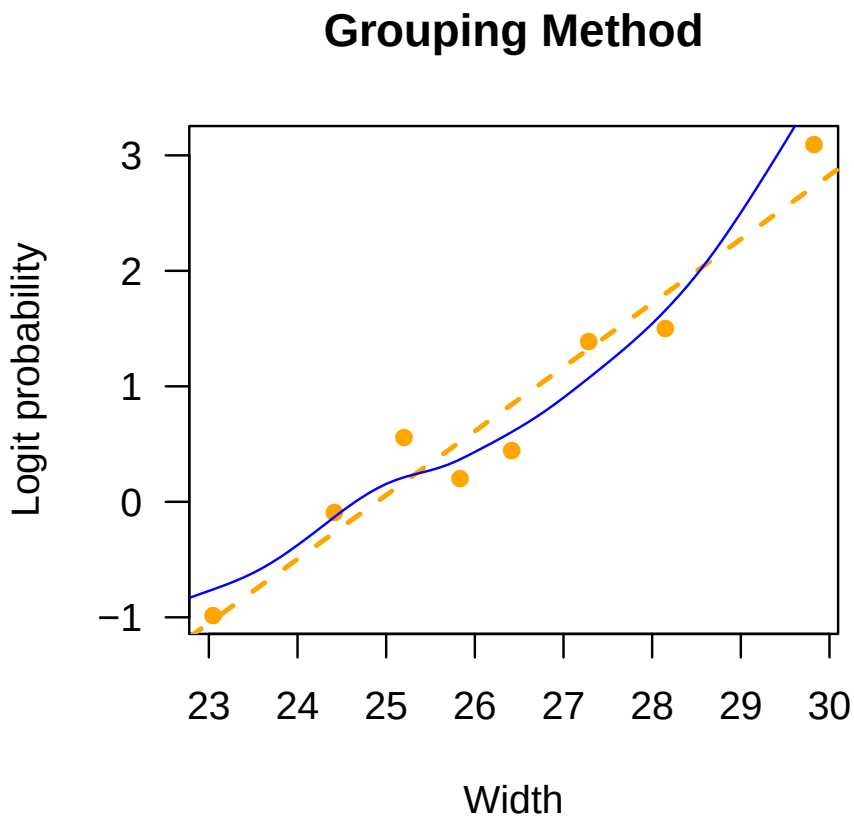
1. The logistic regression model: $logit(p) = \alpha + \beta X$
2. The generalized additive model (GAM) for binary data: $logit(p) = f(X)$, where $f(.)$ is a flexible nonlinear function (§7.4.9)
3. Can check multiple covariates, one at a time

$$logit(p) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$
$$logit(p) = \alpha + f(X_1) + \beta_2 X_2 + \beta_3 X_3$$

4. Available in SAS PROC GAM or R package `gam`
5. GAM for counts data: $log(\mu) = f(X)$ instead of $log(\mu) = \alpha + \beta X$
6. Will be discussed in more details later (§7.4.9)
7. Better than simple visual checking

Illustrating Grouping Method & GAM





Dashed orange line: logistic fit. Solid blue line: GAM fit.

Report Logistic Regression Result

	Estimate	Std. Error	Z-value	p-value
Intercept	-12.3508	2.6287	-4.698	2.62e-06
width	0.4972	0.1017	4.887	1.02e-06

1. The output above from `glm()` in `R`. Estimates are regression coefficients, i.e., intercept and log odds ratio of `width`
2. Log odds ratio of width is 0.4972. For each 1 unit increase in width, the log odds of having satellites increases by 0.4972
3. Odds ratio of width is $exp(0.4972) = 1.64$. For each 1 unit increase in width, the odds of having satellites multiply by 1.64
4. Difficult to explain on probability scale due to nonlinear link function
5. Details of model estimation will be discussed later

Logistic Regression vs. Mixture of Two Normals

1. Suppose X has normal distribution $N(\mu_1, \sigma^2)$ among subjects with $Y = 1$, and X has normal distribution $N(\mu_0, \sigma^2)$ among subjects with $Y = 0$, then **the conditional distribution of Y given X exactly satisfies a logistic regression model.**
2. For example: in a population who are referred to the clinic for persistent abdominal pain, $Y = 1$ indicates chronic pancreatitis (CP) and $Y = 0$ indicates non-CP. X is a biomarker that helps the diagnosis of CP. X has a mixture of normal distribution.
3. One of the reasons why **logit link** is so widely used

Logistic Regression vs. Mixture of Two Normals (cont.)

1. If the two normal distributions have difference variances σ_1^2 and σ_0^2, then the logistic regression result still holds, but with a quadratic term
2. The result continues to hold for multivariate covariates with multivariate normal distributions

Logistic Regression with Retrospective Studies

1. **Prospective study**: follow a cohort of chronic kidney disease (CKD) patients receiving blood pressure therapy for 5 years, and the outcome is renal failure within that time frame. Study the risk factors of renal failure within 5 years of starting therapy. Biomarkers are usually not expensive and percent renal failure is high.
2. **Retrospective study**: follow a cohort of chronic pancreatitis (CP) patients until pancreatic cancer or the end of the study. Study biomarkers that are associated with cancer. Percent pancreatic cancer is rare, and biomarker assay is expensive. Sample all the cancer cases and some controls.

Logistic Regression with Retrospective Studies (cont.)

1. Regardless of whether we use a prospective design or retrospective design, as long as we use a logistic regression model with logit link, the log odds ratio is the same. Hence, this model is very convenient to use, and the result is more easily generalizeable and facilitates the comparison of results across studies (e.g., meta-analysis).
2. We can get what we want from a prospective study (log odds ratio) using data from a retrospective study.
3. Not possible with link functions other than the logit (another reason why logit is so popular)
4. Logistic regression for **matched** case-control sampling will be discussed later

Flexible Link Function and Latent Variable

1. For regression on a binary outcome $Y = 1$ on covariate X, consider model $g(\pi(X)) = \beta_0 + \beta_1 X$ with a link function $g(.)$
2. Let T be a latent variable that satisfies $T = -\beta_0 - \beta_1 X + \epsilon$, where ϵ is a continuous variable with mean 0 and CDF $F(.)$
3. Suppose $Y = 1$ if $T \leqslant 0$ and $Y = 0$ if $T > 0$.

$$\begin{aligned}\pi(X) = Pr(Y = 1|X) &= Pr(T \leqslant 0|X) = Pr(-\beta_0 - \beta_1 X + \epsilon \leqslant 0) \\ &= Pr(\epsilon \leqslant \beta_0 + \beta_1 X) = F(\beta_0 + \beta_1 X) \\ F^{-1}(\pi(X)) &= \beta_0 + \beta_1 X\end{aligned}$$

4. Therefore, the link function $g(.)$ can be interpreted as the inverse CDF function of a continuous distribution $F^{-1}(.)$. Example: standard logistic distribution, standard normal distribution (probit regression). §4.2.5-§4.2.6

Model fitting with maximum likelihood estimation (§ 5.5)

Data: $\{y_i, \boldsymbol{x}_i; i = 1, 2, ..., n\}$. Model: $\pi(\boldsymbol{x}_i) = \dfrac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$.

The likelihood is (NOTE: $\boldsymbol{x}_i$ includes intercept term)

$$\prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{1-y_i} = \prod_{i=1}^{n} \left[\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)} \right]^{y_i} [1 - \pi(\boldsymbol{x}_i)]$$

$$= \prod_{i=1}^{n} \frac{exp\{y_i \boldsymbol{x}_i^T \boldsymbol{\beta}\}}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

The log-likelihood is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log \left[1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \right] \right\}$$

Model fitting with maximum likelihood estimation (cont.)

The likelihood equation is (compare with linear model)

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i - \frac{exp(\boldsymbol{x}_i^T \beta)}{1 + exp(\boldsymbol{x}_i^T \beta)} \boldsymbol{x}_i \right\}$$
$$= \sum_{i=1}^{n} \boldsymbol{x}_i \left[y_i - \pi(\boldsymbol{x}_i) \right] = 0$$

The point estimator $\hat{\beta}$ is the solution to this **unbiased estimating equation**.

The observed information matrix is

$$-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial \pi(\boldsymbol{x}_i)}{\partial \beta^T} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \pi(\boldsymbol{x}_i) \left[1 - \pi(\boldsymbol{x}_i) \right]$$

The variance-covariance matrix of $\hat{\beta}$ is estimated by $\left\{ -\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} \right\}^{-1}$

Predicted Probability

1. Point estimator: $\hat{\boldsymbol{\beta}}$ (p-vector).
2. Variance estimator: $\widehat{cov}(\hat{\boldsymbol{\beta}})$ ($p \times p$ matrix)
3. Wald test Z-statistic for each coefficient ($j = 1, 2, ..., p$):

$$\hat{Z}_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\beta}_j)}}, \quad \text{p} - \text{value} = 2\Phi(-|\hat{Z}_j|)$$

4. Linear predictor: $logit[\hat{\pi}(\boldsymbol{x})] = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. Its variance is estimated as $\boldsymbol{x}^T\widehat{cov}(\hat{\boldsymbol{\beta}})\boldsymbol{x}$. The confidence interval of $logit[\pi(\boldsymbol{x})]$: $logit[\hat{\pi}(\boldsymbol{x})] \pm z_{\alpha/2}\sqrt{\boldsymbol{x}^T\widehat{cov}(\hat{\boldsymbol{\beta}})\boldsymbol{x}}$.
5. The CI of $\pi(\boldsymbol{x})$ is the expit transformation of the CI of $logit[\pi(\boldsymbol{x})]$
6. Likelihood ratio and score tests are also applicable (§ 1.3.3). They are more complicated than Wald test but have slightly better performance with small sample size (§5.2.6)

Checking Goodness of Fit (§ 5.2.3 to 5.2.6)

$$\text{logit}\left[\pi(x)\right] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

1. Visual checking through grouping (discussed previously; works best with a single covariate)
2. Using a more flexible model such as Generalized Additive Model (discussed previously; R code in announcement)
3. Adding interactions, quadratic terms, etc., and testing for significance of the added terms using likelihood ratio test (§5.2.4)
4. Global goodness of fit checking by **Hosmer & Lemeshow test** (widely used for logistic regression)

Checking Goodness of Fit via Likelihood Ratio Test

$$\begin{aligned} \text{logit}\left[P(Y=1)\right] &= \alpha + \beta \text{ width} \\ \text{logit}\left[P(Y=1)\right] &= \alpha + \beta_1 \text{ width} + \beta_2 \text{ width}^2 \end{aligned} \tag{1}$$

1. Fit a logit-linear model with (centered) width: β significant
2. Fit the same model but with an additional quadratic term: β_1 significant, β_2 not.
3. Calculate the Likelihood Ratio Test (LRT) statistic, its degree of freedom, and hence its p-value: 0.825 on 1 df, $p = 0.363$.

```
fm1 <- glm( y ~ width.ctr, data = dat,
family = binomial )
fm2 <- glm( y ~ width.ctr + width.ctr2, data = dat,
family = binomial )
LRT.stat <- as.numeric( (-2)*( logLik(fm1) - logLik(fm2) ) )
pvalue <- pchisq(q = LRT.stat, df = 1, lower.tail = F)
```

Hosmer-Lemeshow Test (§ 5.2.5)

$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{n_i \left(\sum_j \hat{\pi}_{ij}/n_i\right)\left[1 - \left(\sum_j \hat{\pi}_{ij}\right)/n_i\right]} \quad \sim \quad \chi^2_{g-2}$$

1. n_i number of observations in group i. $\sum_j \hat{\pi}_{ij}/n_i$ average probability of $Y = 1$ in group i. The denominator is approximately the variance of $\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}$.
2. Analogous to chi-squared test for $g \times 2$ table. Each group roughly the same n_i. Usually we use $g = 10$ (g cannot go up with sample size).
3. A large value merely indicates *some* lack of fit but provides no insight about its nature. $d.f. \approx g - 2$ by Hosmer & Lemeshow.
4. Work with multiple covariates. Non-significance does not indicate model has no problem. `SAS Proc Logistic` with `Lackfit` option

Logistic models with categorical predictors (§ 5.3)

- When there is a single categorical predictor, the data can be arranged in an $I \times 2$ contingency table (e.g., Table 5.3)
- When the categories are unordered (e.g., nominal data), the (saturated) model is $\text{logit}(\pi_i) = \beta_i$ ($i = 1, 2, ..., I$), with I unknown parameters.
- We may write the model as $\text{logit}(\pi_i) = \alpha + \beta_i$ with set-to-zero constraint $\beta_1 = 0$ or sum-to-zero constraint $\sum_i \beta_i = 0$
- The model for subject j ($j = 1, 2, ..., n$) is $\text{logit}(\pi_j) = \alpha + \sum_{i=1}^{I} \beta_i 1\{j \in \text{group } i\}$
- When the categories are ordered (e.g., ordinal data), we may assume that $\text{logit}(\pi_i) = \alpha + \beta x_i$
  - The number of parameters reduced with the linear assumption.
  - Be careful about coding x_i (i=1,2,...,I): (1,2,3) or (1,4,9)?
  - Treat the x_i like a continuous variable.

Logistic models with categorical predictors: Example

- The study of maternal alcohol consumption and child's congenital malformation.
- X: alcohol consumption. Y: malformation status

$$logit(\pi_i) = \alpha + \beta_i$$

X	$Y = 1$	$Y = 0$	Sample logit	π_i	$\widehat{\pi}_i$
0	48	17066	-5.87	0.0028	0.0026
< 1	38	14464	-5.94	0.0026	0.0030
1-2	5	788	-5.06	0.0063	0.0041
3-5	1	126	-4.84	0.0079	0.0091
$\geqslant 6$	1	37	-3.61	0.0263	0.0231

Pearson statistic is $X^2 = 12.1(p = 0.02)$; Likelihood-ratio statistic is $G^2 = 6.2(p = 0.19)$.

Logistic models with categorical predictors: Example

Linear logit model:

$$logit(\pi_i) = \alpha + \beta x_i$$

The independence model is the special case $\beta = 0$.

Table 5.4 Software Output (Based on SAS) for Linear Logit Model Fitted to Table 5.3 on Infant Malformation and Alcohol Consumption

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	3	1.9487
Pearson Chi-Square	3	2.0523
Log Likelihood		−635.5968

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf	Limits	Wald Chi-Sq	Pr>ChiSq
Intercept	−5.9605	0.1154	−6.1930	−5.7397	2666.41	<.0001
alcohol	0.3166	0.1254	0.0187	0.5236	6.37	0.0116

The estimated multiplicative effect of a unit increase in daily alcohol consumption on the odds of malformation is $exp(0.317) = 1.37$.

Logistic models with categorical predictors: Example

- Table 5.3: fit a saturate model but the results from Pearson and likelihood ratio tests differ in statistical significance.
  - No modeling assumption with **saturated** model
  - Treat alcohol consumption as nominal data ignoring the order
  - The sparse data at higher level of alhocol may make the asymptotic property invalid
  - For logistic regression, the **effective** sample size is the number of $Y = 1$ or $Y = 0$ whichever is smaller, not the total number of subjects (n)
- Table 5.4: fit a logit-linear model
  - Impose some modeling assumption (but can be checked in this case)
  - Use monotone relationship, which is plausible from a scientific perspective
  - Higher statistical power with less model degree of freedom

Cochran-Armitage Trend Test (§ 5.3.5)

- Developed by Armitage (1955) and Cochran (1954) for $I \times 2$ tables with ordered rows
- They used a linear probability model $\pi(x) = \alpha + \beta x$
- It is a chi-square test of the independence between rows and columns under the linear assumption. $H_0 : \beta = 0$.
- This test is equivalent to the score statistic for testing $H_0 : \beta = 0$ in the linear logit model $logit\{\pi(x)\} = \alpha + \beta x$. The linear and logit-linear models are the equivalent under the null.
- Using directed models can improve inferential power
  - If the trend is indeed linear, making use of the linear trend (as in Cochran-Armitage test) is more powerful than not making use of the linear trend (as in $\text{logit}(\pi_i) = \beta_i$)
  - Similar test statistic but using less degree of freedom

Chapter 6: Building and Checking Logistic Regression Models

- Selection of covariates
- Model checking, goodness of fit, residual diagnosis
- Prediction analytics
- Cochran-Mantel-Haenszel test (CMH) with applicaiton to clinical trials and meta-analyses
- Quasi-complete separation
- Power and sample size calculation (skipped)

Model Selection (§ 6.1)

The data set is $\{Y_i, X_{1i}, X_{2i}, ..., X_{pi}; i = 1, 2, ..., n\}$. The model is

$$\pi(X_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)$$

The p covariates include interactions, quadratic terms, etc. We want to retain only the predictive covariates in the model.

- Model selection is both science and art
- Two goals: (1) complex enough to fit the data well; (2) relatively simple to interpret (avoid overfitting)
- The same principles that you learned in linear model class still apply
- Confirmatory studies vs. exploratory studies
- Start with univariate association between each covariate and the outcome, with a larger p-value threshold, such as 0.20

How many covariates can be included in the model?

$$Y_i \sim \text{Bernoulli with } \pi(\mathrm{X}_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)$$

- The effective sample size of a logistic regression is either $\sum_i Y_i$ or $n - \sum_i Y_i$, whichever is smaller
- **The rule of thumb**: no more than the effective sample size divided by 10 (or, 10 events per covariate)
- Including too many covariates may cause non-convergence, and linear additivity assumption may be more questionable.
- Exception exists, such as high-dimensional regression with penalty for $p > n$ situation

An Illustration of Multi-collinearity (§ 6.1.2)

- Avoid multicollinearity, as in linear regression (📖 Page 209, Table 6.1)
- With both covariates, the overall LR/Score/Wald test $p < 0.0001$, but `weight` has a coefficent of 0.8338 with $p = 0.2145$, and `width` has a coefficient of 0.3068 with $p = 0.0918$.
- With only `weight`, the overall LR/Score/Wald test $p < 0.0001$, and `weight` has a coefficent of 1.8151 with $p < 0.0001$.
- Both models pass Hosmer-Lemeshow test
- Correlation between `weight` and `width` is 0.887.
- `proc logistic desc ; model y = weight width / lackfit ;`
- The same phenomenon as in linear regression
- Descriptive analysis first! (e.g., `proc corr`)

Explained variation of logistic regression: pseudo R^2

- For linear regression $Y_i = \boldsymbol{X}_i^T \beta + \epsilon_i$, the R^2 is

$$R^2 = 1 - \frac{\sum_i \left(Y_i - \boldsymbol{X}_i^T \hat{\beta}\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$$

- For logistic regression $\hat{\pi}_i = expit(\boldsymbol{X}_i^T \hat{\beta})$, the analog

$$1 - \frac{\sum_i \left(Y_i - \hat{\pi}_i\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$$

may not be nondecreasing as the model gets more complex (undesirable)

Explained variation of logistic regression: pseudo R^2 (cont.)

For logistic regression, a more widely used measure is the pseudo R^2 of McFadden (1974): $\frac{L_M - L_0}{L_S - L_0} = 1 - \frac{L_M}{L_0}$

$$L = \log \prod_{i=1}^{N} \left[\pi_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}\right] = \sum_{i=1}^{N} \left[y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)\right]$$

- L_M is the log likelihood evaluated at the MLE $\hat{\pi}_i = \text{expit}(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}})$
- L_0 is the log likelihood evaluated under the MLE of the null model: $\hat{\pi}_i \equiv N^{-1} \sum_i y_i$
- L_S is the log likelihood evaluated under the saturated model with $\hat{\pi}_i = y_i$. $L_S = 0$
- A high pseudo R^2 does not necessarily indicate high accuracy in individual-level prediction

Prediction accuracy: ROC curve

- Receiver Operative Characteristics (ROC) curve: a measure of "discrimination" (higher predicted probability is associated with higher chance of $Y = 1$)
- There are other measures of "calibration" (among all subjects with predicted probability $\hat{p}$, the proportion of those with $Y = 1$ is $\hat{p}$)
- $Y = 0$ (non disease) or 1 (disease). Estimated disease risk (probability) $\hat{\pi} \in (0, 1)$. We classify the subject as a case ($Y = 1$) when $\hat{\pi} > c$ and control ($Y = 0$) when $\hat{\pi} \leq c$.
- Sensitivity $P(\hat{\pi} > c | Y = 1) \Leftarrow \frac{\sum_i 1\{\hat{\pi}_i > c\}}{\sum_i Y_i}$
- Specificity $P(\hat{\pi} \leq c | Y = 0) \Leftarrow \frac{\sum_i 1\{\hat{\pi}_i \leq c\}}{\sum_i (1 - Y_i)}$
- Illustrate with boxplot for cases and controls 

ROC Curve Example

ROC curve 📖 p225

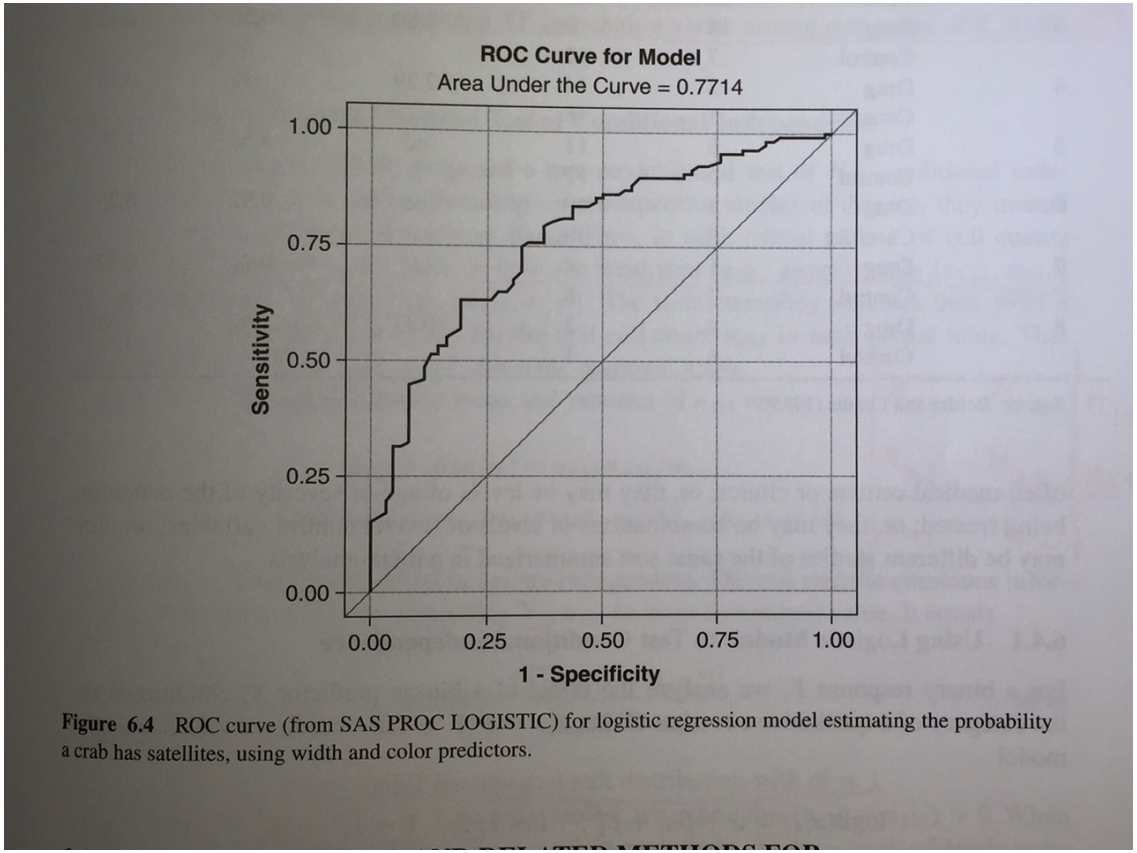


Figure 6.4 ROC curve (from SAS PROC LOGISTIC) for logistic regression model estimating the probability a crab has satellites, using width and color predictors.

Illustrate Positive Predictive Value (PPV) and Negative Predictive Value (NPV) with 2×2 table

Prediction accuracy: ROC curve (cont.)

- The area under the ROC curve (AUC) is reported as c-statistic in SAS PROC LOGISTIC. NOTE: this is the AUC for the training data (development data) only, which is often higher than the AUC from independent testing data (validation data) due to the possibility of overfitting.
- It is a number between 0 and 1. AUC $= 0.5$ is like flipping a coin. So AUC < 0.5 is unlikely. Good classification requires AUC > 0.80 (excellent, > 0.9).
- It has been shown that the AUC (c-statistic) is just the **concordance index**: randomly pick a case ($Y = 1$) and a control ($Y = 0$), the probability that the case has a higher predicted probability $\hat{\pi}$ than the control is the concordance index

Prediction error for logistic regression: Brier score

- Brier score is just squared error loss in the testing dataset: $\sum_{j=1}^{n} [y_j - \hat{\pi}_j]^2$
- A measure of both calibration and discrimination
- The smaller the better (e.g., Model 2 preferred)

y	Model 1	Model 2	Perfect
0	0.2	0.1	0
0	0.4	0.3	0
1	0.6	0.7	1
1	0.8	0.9	1

Cochran-Mantel-Haenszel Test (§ 6.4)

- Study the association between a treatment variable (e.g., binary) and a disease outcome (e.g., binary) after adjusting for a possibly confounding variable (e.g., categorical or continuous but grouped) that might influence that association
- Example in Table 6.9: multicenter randomized clinical trial comparing treatment vs. placebo on a binary outcome (cured vs. not)
- The logistic regression approach ($i = 1, 2; k = 1, 2, ..., K; x_i = 1 \text{ or } 2$):

$$\pi_{ik} = P(Y = 1 | X = i, Z = k)$$
$$logit(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$$

- Test H_0: $\beta = 0$ using Wald, score or likelihood ratio test
- What if there is interaction between X and Z, i.e., β depends on Z? Interaction always exists, more or less. But we often want an overall test which adjusts for center differences

Table 6.9

Table 6.9 Clinical Trial Relating Treatment to Response for Eight Centers, with Expected Value and Variance (of Success Count for Drug) Under Conditional Independence

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	$\text{var}(n_{11k})$
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

often medical centers or clinics; or, they may be levels of age or severity of the condition

Cochran-Mantel-Haenszel Test (§ 6.4)

Data from Center k $(k = 1, 2, ..., K)$

	cured	not	Total
Treatment	n_{11k}	n_{12k}	n_{1+k}
placebo	n_{21k}	n_{22k}	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_{++k}

- Test H_0: Treatment and outcome are independent vs. H_1: not independent
- Both the treatment (row) and outcome (column) totals fixed, $n_{11k} \sim$ hypergeometric distribution conditional on these row and column totals
- Under the null, the hypergeometric mean and variance of n_{11k} are

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$$
$$var(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/[n_{++k}^2(n_{++k} - 1)]$$

Cochran-Mantel-Haenszel Test (§ 6.4)

- Test H_0: Treatment and outcome are independent vs. H_1: not independent
- Both the treatment (row) and outcome (column) totals fixed, $n_{11k} \sim$ hypergeometric distribution conditional on these row and column totals
- Under the null, the hypergeometric mean and variance of n_{11k} are

$$\mu_{11k} = E(n_{11k}) = n_{1+k} n_{+1k} / n_{++k}$$
$$var(n_{11k}) = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / [n_{++k}^2 (n_{++k} - 1)]$$

- The CMH statistic is $CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k var(n_{11k})}$, which has a large sample chi-squared null distribution with $df = 1$.

Cochran-Mantel-Haenszel Test vs. Logistic Regression

- When the sample size per center (also called strata) is moderately large, the two produce similar results (CMH is a score test of the logistic model)
- When the number of strata is large (like matched pairs data) or some strata have small sample size, the logistic regression does not apply but CMH still applies
- An overall OR estimator is available from CMH. This is not a weighted average of ORs from individual centers 

$$\hat{\theta}_{CMH} = \frac{\sum_k (n_{11k} n_{22k}/n_{++k})}{\sum_k (n_{12k} n_{21k}/n_{++k})} = \frac{\sum_k n_{++k} p_{11|k} p_{22|k}}{\sum_k n_{++k} p_{12|k} p_{21|k}}$$

- CMH is the standard method for stratified analysis of categorical data, applicable to $I \times J \times K$ contingency table and useful in meta-analysis
- Available in `SAS PROC FREQ`
- If the treatment effect differs across strata (interaction), and that difference is of scientific interest, use logistic model with interaction

Quasi-complete Separation in Logistic Regression

- Be careful about excessively large (or small) odds ratios or excessively large standard errors
  - multi-colinearity; remove one of the correlated covariates
  - complete or quasi-complete separation: MLE does not exist or log-likelihood is flat
- Complete separation: there exists a vector $\boldsymbol{b}$ such that $\boldsymbol{b}^T \boldsymbol{x}_i > 0$ whenever $y_i = 1$ and $\boldsymbol{b}^T \boldsymbol{x}_i < 0$ whenever $y_i = 0$. (more likely with continuous covariates)
- Quasi-complete separation: $\boldsymbol{b}^T \boldsymbol{x}_i \geq 0$ whenever $y_i = 1$ and $\boldsymbol{b}^T \boldsymbol{x}_i \leq 0$ whenever $y_i = 0$. (more likely with categorical covariates)
- Solution: (1) examine inclusion/exclusion criteria of the study sample (2) Firth logistic regression (§7.4.7)

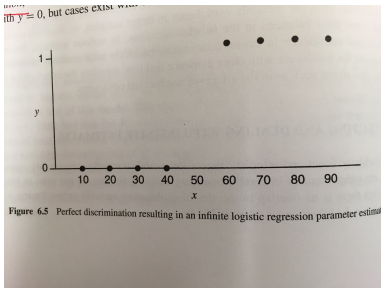


Figure 6.5 Perfect discrimination resulting in an infinite logistic regression parameter estima