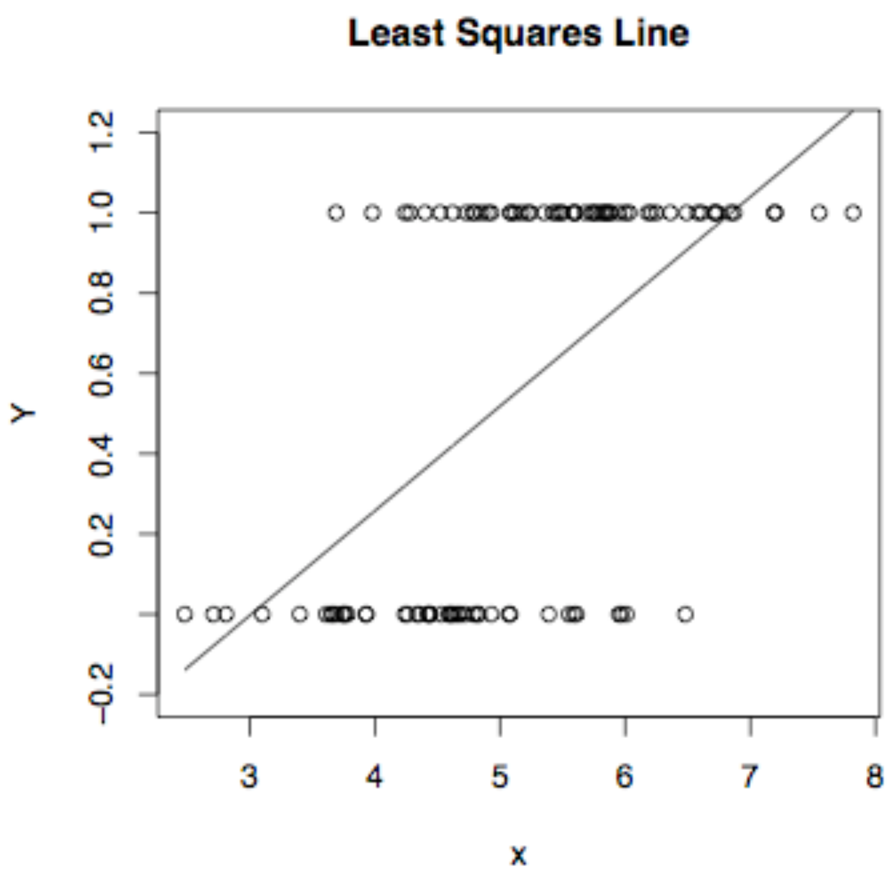
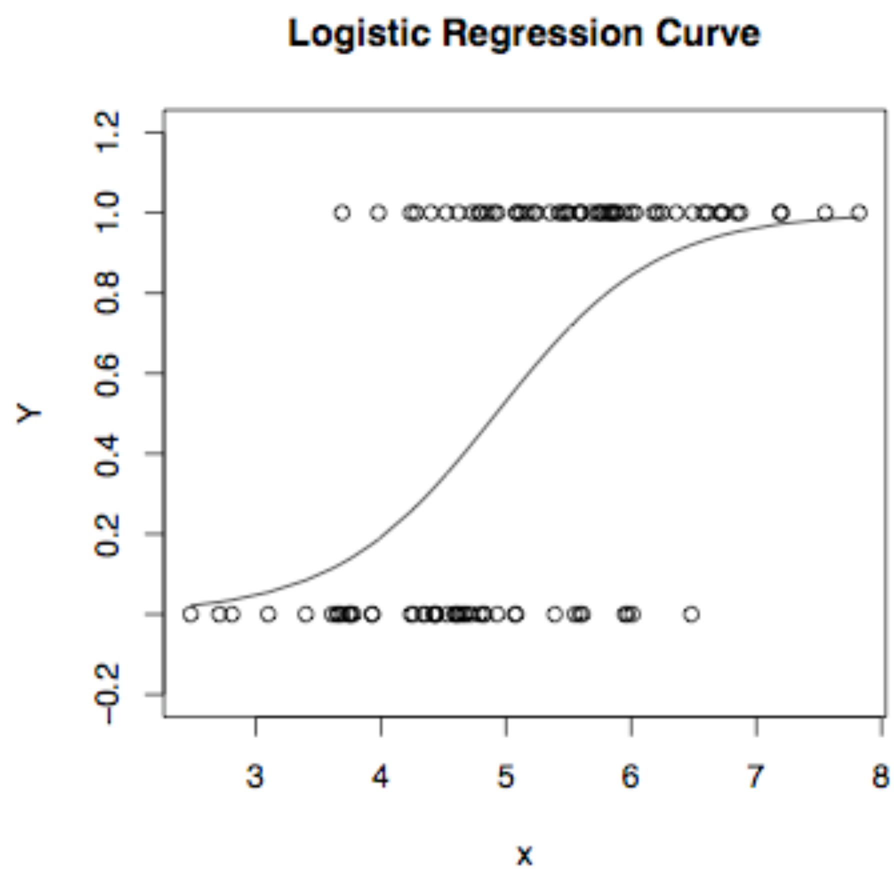
Logistic Regression

For a binary dependent variable:
1=Yes, 0=No

Least Squares vs. Logistic Regression





Linear regression model for the log odds of the event Y=1

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Equivalent Statements

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}$$

$$\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}$$

$$= e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_{p-1}x_{p-1}}$$

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

$F(x) = \frac{e^x}{1+e^x}$ is called the *logistic distribution.*

- Could use any cumulative distribution function:

$$P(Y = 1|x_1, \ldots, x_{p-1}) = F(\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1})$$

- CDF of the standard normal used to be popular
- Called probit analysis
- Can be closely approximated with a logistic regression.

In terms of log odds, logistic regression is like regular regression

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

In terms of plain odds,

- Logistic regression coefficients represent *odds ratios*
- For example, “Among 50 year old men, the odds of being dead before age 60 are three times as great for smokers.”

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = 3$$

Logistic regression

- X=1 means smoker, X=0 means non-smoker
- Y=1 means dead, Y=0 means alive

- Log odds of death = $\beta_0 + \beta_1 x$

- Odds of death = $e^{\beta_0} e^{\beta_1 x}$

$$\text{Odds of Death} = e^{\beta_0} e^{\beta_1 x}$$

Group	x	**Odds of Death**
Smokers	1	$e^{\beta_0} e^{\beta_1}$
Non-smokers	0	e^{β_0}

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Cancer Therapy Example

$$\text{Log Survival Odds} = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x$$

Treatment	d_1	d_2	**Odds of Survival** $= e^{\beta_0} e^{\beta_1 d_1} e^{\beta_2 d_2} e^{\beta_3 x}$
Chemotherapy	1	0	$e^{\beta_0} e^{\beta_1} e^{\beta_3 x}$
Radiation	0	1	$e^{\beta_0} e^{\beta_2} e^{\beta_3 x}$
Both	0	0	$e^{\beta_0} e^{\beta_3 x}$

For any given disease severity x,

$$\frac{\text{Survival odds with Chemo}}{\text{Survival odds with Both}} = \frac{e^{\beta_0} e^{\beta_1} e^{\beta_3 x}}{e^{\beta_0} e^{\beta_3 x}} = e^{\beta_1}$$

In general,

- When x_k is increased by one unit and all other independent variables are held constant, the odds of Y=1 are multiplied by e^{β_k}
- That is, e^{β_k} is an **odds ratio** --- the ratio of the odds of Y=1 when x_k is increased by one unit, to the odds of Y=1 when everything is left alone.
- As in ordinary regression, we speak of "controlling" for the other variables.

The conditional probability of Y=1

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

This formula can be used to calculate a predicted P(Y=1)
Just replace betas by their estimates

It can also be used to calculate the probability of getting
The sample data values we actually did observe, as a
function of the betas.

Maximum likelihood estimation

- Likelihood = Conditional probability of getting the data values we did observe,
- As a function of the betas
- Maximize the (log) likelihood with respect to betas.
- Maximize numerically (“Iteratively re-weighted least squares”)
- Likelihood ratio tests as usual

Wald tests

- MLEs have an approximate multivariate normal sampling distribution for large samples (Thanks Mr. Wald.)
- Approximate mean vector = the true parameter values for large samples
- Asymptotic variance-covariance matrix is easy to estimate
- H_0: $\mathbf{C}\boldsymbol{\theta} = \boldsymbol{h}$ (Linear hypothesis)
- For logistic regression, $\boldsymbol{\theta} = \boldsymbol{\beta}$

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$$

$$\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h} \text{ is multivariate normal as } n \rightarrow \infty$$

Leads to a straightforward chisquare test

- Called a Wald test
- Based on the full (maybe even saturated) model
- Asymptotically equivalent to the LR test
- Not as good as LR for smaller samples
- Very convenient, especially with SAS

$$Z = \frac{\widehat{\theta}_k}{\sqrt{\widehat{Var(\widehat{\theta}_k)}}}$$

- Approximately standard normal for large samples if θ_k=0.
- Can use to form large-sample confidence intervals
- Denominator is the square root of a diagonal element of the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$
- Square it to get a Wald test with 1 df.

Wald statistics and asymptotic standard errors

- Exist for the classical (non-conditional) log-linear models
- This is what the text is talking about in Section 5.4
- Not easy to get from R

- For logistic regression, straightforward with R as well as SAS