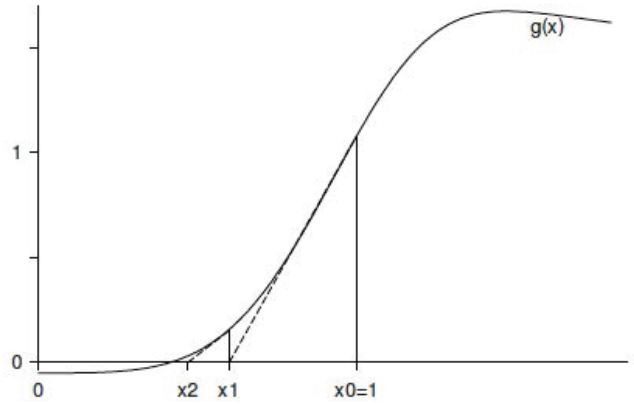
# Lecture 4: Newton's method and gradient descent

- Newton's method
- Functional iteration
- Fitting linear regression
- Fitting logistic regression

# Newton's method for finding root of a function

- solve $g(x) = 0$
- iterative method: $x_n = x_{n-1} - \frac{g(x_{n-1})}{g'(x_{n-1})}$



# Quadratic convergence of Newton's method

- let $e_n = |x_n - x_\infty|$
- Newton's method has quadratic convergence

  $\lim_{n\to\infty} \frac{e_n}{e_{n-1}^2} = \frac{1}{2} f''(x_\infty)$
- linear convergence definition:

  if $\lim_{n\to\infty} \frac{e_n}{e_{n-1}} = f'(x_\infty)$

  $f$: iteration function, $0 < |f'(x_\infty)| < 1$

# Finding the maximum

- Finding the maximum of a function $f(x)$

$$\max_{x \in \mathcal{D}} f(x)$$

- first order condition

$$\nabla f(x) = 0$$

## Newton's method for finding maximum of a function

- $g : \mathbb{R}^d \to \mathbb{R}$
- $x \in \mathbb{R}^d \triangleq \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$
- Newton's method for finding maximum

$$x_n = x_{n-1} - [H(x_{n-1})]^{-1} \nabla g(x_{n-1})$$

gradient vector

$$[\nabla g]_i = \frac{dg(x)}{d[x]_i}$$

Hessian matrix

$$[H(x)]_{ij} = \frac{d^2 g(x)}{d[x]_i d[x]_j}$$

# Gradient descent

- gradient descent for finding maximum of a function

$$x_n = x_{n-1} + \mu \nabla g(x_{n-1})$$

  $\mu$: step-size

- gradient descent can be viewed as approximating Hessian matrix as

$$H(x_{n-1}) = -I$$

# Maximum likelihood

- $\theta$: parameter, $x$: data
- log-likelihood function $\ell(\theta|x) \triangleq \log f(x|\theta)$

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \ell(\theta|x)$$

- drop dependence on $x$, but remember that $\ell(\theta)$ is a function of data $x$
- Maximize the log-likelihood function by setting $\frac{d\ell(\theta)}{d\theta} = 0$

# Linear Regression

- $n$ observations, $1$ predictive variable, 1 response variable
- finding parameters $a \in \mathbb{R}, b \in \mathbb{R}$ to minimize the **mean square error**

$$(a, b) = \arg\min_{a,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$$\hat{a} = \frac{S_{xy} - \bar{x}\bar{y}}{S_{xx} - (\bar{x})^2}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

# Multiple linear regression

- $n$ observations, $p$ predictors, 1 response variable
- finding parameters $a \in \mathbb{R}^p$, $b$ to minimize the **mean square error**

$$(a, b) = \arg\min_{a,b} \frac{1}{n} \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{p} a_j x_{ij} - b\|^2$$

$$\hat{a} = \Sigma_{xx}^{-1} \Sigma_{xy}$$

  $\Sigma_{xx}$ and $\Sigma_{xy}$ are the same covariance matrices

- difficult when $p > n$ and when $p$ is large
- use iterative method (gradient descent etc.)

# *Prostate Cancer*

The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`). The correlation matrix of the predictors given in Table 3.1 shows many strong correlations. Figure 1.1 (page 3) of Chapter 1 is a scatterplot matrix showing every pairwise plot between the variables. We see that `svi` is a binary variable, and `gleason` is an ordered categorical variable. We see, for example, that both `lcavol` and `lcp` show a strong relationship with the response `lpsa`, and with each other. We need to fit the effects jointly to untangle the relationships between the predictors and the response.

**TABLE 3.1.** *Correlations of predictors in the prostate cancer data.*

| | lcavol | lweight | age | lbph | svi | lcp | gleason |
|---|---|---|---|---|---|---|---|
| lweight | 0.300 | | | | | | |
| age | 0.286 | 0.317 | | | | | |
| lbph | 0.063 | 0.437 | 0.287 | | | | |
| svi | 0.593 | 0.181 | 0.129 | −0.139 | | | |
| lcp | 0.692 | 0.157 | 0.173 | −0.089 | 0.671 | | |
| gleason | 0.426 | 0.024 | 0.366 | 0.033 | 0.307 | 0.476 | |
| pgg45 | 0.483 | 0.074 | 0.276 | −0.030 | 0.481 | 0.663 | 0.757 |

**TABLE 3.2.** *Linear model fit to the prostate cancer data. The $Z$ score is the coefficient divided by its standard error (3.12). Roughly a $Z$ score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

| Term | Coefficient | Std. Error | $Z$ Score |
|---|---|---|---|
| `Intercept` | 2.46 | 0.09 | 27.60 |
| `lcavol` | 0.68 | 0.13 | 5.37 |
| `lweight` | 0.26 | 0.10 | 2.75 |
| `age` | −0.14 | 0.10 | −1.40 |
| `lbph` | 0.21 | 0.10 | 2.06 |
| `svi` | 0.31 | 0.12 | 2.47 |
| `lcp` | −0.29 | 0.15 | −1.87 |
| `gleason` | −0.02 | 0.15 | −0.15 |
| `pgg45` | 0.27 | 0.15 | 1.74 |

# Newton's method for fitting logistic regression model

- $n$ observations $(x_i, y_i)$, $i = 1, \cdots, n$
- parameter $a \in \mathbb{R}^p$, $b$
- predictor $x_i \in \mathbb{R}^p$, label $y_i \in \{0, 1\}$.
- $p(y_i = 1|x_i) \triangleq h(x_i; a, b) = \frac{1}{1+e^{-a^\top x - b}}$

# *South African Heart Disease*

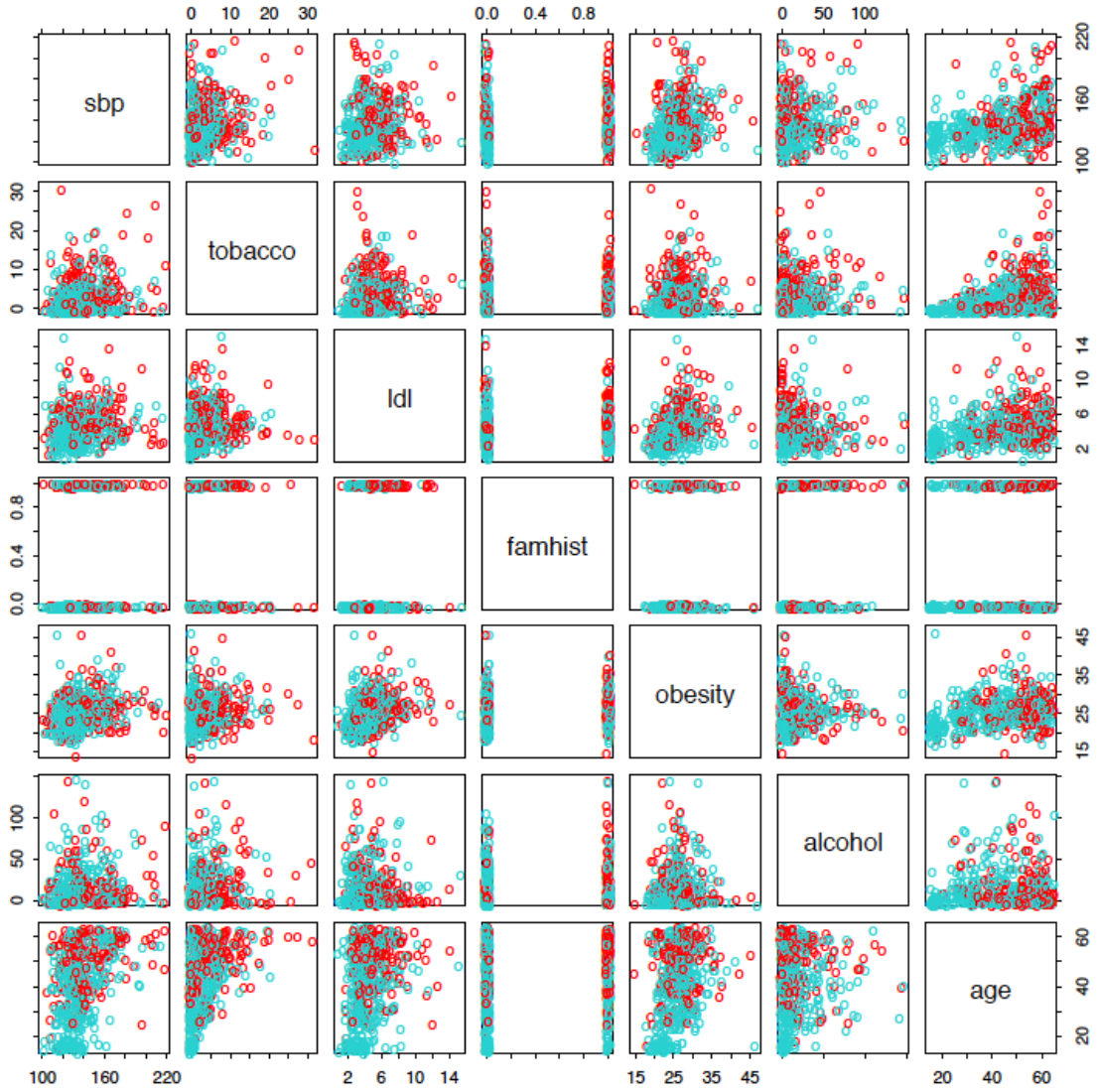
Here we present an analysis of binary data to illustrate the traditional statistical use of the logistic regression model. The data in Figure 4.12 are a subset of the Coronary Risk-Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa (Rousseauw et al., 1983). The aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region. The data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (MI) at the time of the survey (the overall prevalence of MI was 5.1% in this region). There are 160 cases in our data set, and a sample of 302 controls. These data are described in more detail in Hastie and Tibshirani (1987).

# Example: South African heart disease data

Example (from ESL section 4.4.2): there are $n = 462$ individuals broken up into 160 cases (those who have coronary heart disease) and 302 controls (those who don't). There are $p = 7$ variables measured on each individual:

- sbp (systolic blood pressure)
- tobacco (lifetime tobacco consumption in kg)
- ldl (low density lipoprotein cholesterol)
- famhist (family history of heart disease, present or absent)
- obesity
- alcohol
- age

Pairs plot (red are cases, green are controls):



Fitted logistic regression model:

| | Coefficient | Std. Error | $Z$ Score |
|---|---|---|---|
| (Intercept) | −4.130 | 0.964 | −4.285 |
| sbp | 0.006 | 0.006 | 1.023 |
| tobacco | 0.080 | 0.026 | 3.034 |
| ldl | 0.185 | 0.057 | 3.219 |
| famhist | 0.939 | 0.225 | 4.178 |
| obesity | -0.035 | 0.029 | −1.187 |
| alcohol | 0.001 | 0.004 | 0.136 |
| age | 0.043 | 0.010 | 4.184 |

The $Z$ score is the coefficient divided by its standard error. There is a test for significance called the Wald test

Just as in linear regression, correlated variables can cause problems with interpretation. E.g., sbp and obseity are not significant, and obesity has a negative sign! (Marginally, these are both significant and have positive signs)

After repeatedly dropping the least significant variable and refitting:

| | Coefficient | Std. Error | $Z$ score |
|---|---|---|---|
| `(Intercept)` | $-4.204$ | $0.498$ | $-8.45$ |
| `tobacco` | $0.081$ | $0.026$ | $3.16$ |
| `ldl` | $0.168$ | $0.054$ | $3.09$ |
| `famhist` | $0.924$ | $0.223$ | $4.14$ |
| `age` | $0.044$ | $0.010$ | $4.52$ |

This procedure was stopped when all variables were significant

E.g., interpretation of tobacco coefficient: increasing the tobacco usage over the course of one's lifetime by 1kg (and keeping all other variables fixed) multiplies the estimated odds of coronary heart disease by $\exp(0.081) \approx 1.084$, or in other words, increases the odds by $8.4\%$

# Functional iteration

- find a root of the equation $g(x) = 0$
- introduce $f(x) = g(x) + x$, $g(x) = 0 \Rightarrow f(x) = x$
- in many examples, iterates $x_n = f(x_{n-1})$ convergens to $x^* = f(x^*)$, $x^*$ called fixed point of $f(x)$
- newton's method $x_n = f(x_{n-1})$ with $f(x) = x - \frac{g(x)}{g'(x)}$

# Convergence

**Theorem.** *Suppose the function $f(x)$ defined on a closed interval $I$ satisfies the conditions:*

*1. $f(x) \in I$ whenever $x \in I$*

*2. $|f(y) - f(x)| \leq \lambda |y - x|$ for any two points $x$ and $y$ in $I$.*

Then, provided the Lipschiz constant $\lambda$ is in $[0, 1)$, $f(x)$ has a unique fixed point $x^* \in I$, and $x_n = f(x_{n-1})$ converges to $x^*$ regardless of starting point $x_0 \in I$. Furthermore, we have

$$|x_n - x_\infty| \leq \frac{\lambda^n}{1 - \lambda} |x_1 - x_0|.$$

Proof:

First consider

$$|x_{k+1} - x_k| = |f(x_k) - f(x_{k-1})| \leq \lambda |x_k - x_{k-1}| \cdots \leq \lambda^k |x_1 - x_0|$$

then for some $m > n$,

$$|x_m - x_n| \leq \sum_{k=n}^{m-1} |x_{k+1} - x_k| \leq \sum_{k=n}^{m-1} \lambda^k |x_1 - x_0| \leq \frac{\lambda^n}{1-\lambda} |x_1 - x_0|$$

So $\{x_n\}$ forms a Cauchy sequence. Since $I \in \mathbb{R}$ is closed and bounded, it is compact.

By Theorem 3.11 (b) in Rubin[1], if $X$ is a compact metric space and if $\{x_n\}$ is a Cauchy sequence in $X$, then $\{x_n\}$ converges to some point of $X$. Hence $x_n \to x_\infty$, when $n \to \infty$ and $x_\infty \in I$. Now since $f$ is continuous (it's in fact Lipschitz continuous), this means $f(x_\infty) = x_\infty$. Hence, $x_\infty$ is a fixed point. Since fixed point is unique, then $x_\infty = x^*$.

To prove the "furthermore" part, since $|x_m - x_n| \leq \frac{\lambda_n}{1-\lambda}|x_1 - x_0|$, we can send $m \to \infty$ and have $|x_\infty - x_n| \leq \frac{\lambda_n}{1-\lambda}|x_1 - x_0|$.

[1] Walter Rudin, Principles of Mathematical Analysis.

# Example

- $g(x) = -\sin(\frac{x+e^{-1}}{\pi})$, find $x$ such that $g(x) = 0$
- $f(x) = g(x) + x$, $f'(x) = -\frac{1}{\pi}\cos(\frac{x+e^{-1}}{\pi}) + 1$
- $|f'(x)| < 1$ for $[-e^{-1}, \frac{\pi^2}{2} - e^{-1}] = [-0.3679, 4.5669]$ (so we can apply function iteration to $g(x)$)

- let $I = [-0.3, 3]$, then $f(x) \in I$ whenever $x \in I$, and $\lambda < 1$ for $f(x)$ on this range