ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

09-April-2018
Lecture 10

Lecture 10

Remedial Measures and Nonlinear Regression

Announcement 1

- Assignment #5 will be released by tomorrow morning. Due on 16 April by 9:00 pm

Announcement 2

- Midterm marks will be available on IVLE by tomorrow morning.
  - Midterm paper will NOT be returned.
  - You can review your midterm paper on an appointment basis.
  - Regrading policy:
    Once you make a regrade request, not only the particular question of your interest but your entire exam will be subject to regrading. Please consider seriously before you make a request as your exam score might go down. When you would like to make a request referring to other student's grading, you need to bring the whole exam paper that you are referring to. The entire range of the referred exam paper is also subject to regrading, and its score might also go down.

Outline

- Non-constant variance: weighted regression
- Multicollinearity: Ridge regression
- Outliers: Robust regression
- Bootstrap
- Generalized Linear Model

Weighted regression

- What if the errors have non-constant variance

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \\ \epsilon_i &\sim N(0, k \cdot \sigma_i^2) \text{ (with } k > 0) \end{aligned}$$

  - Are the b_k's using least-squares estimation still unbiased?
  - DO the b_k's have minimum variance (efficient)?
- If σ_i^2 is known, we can minimize:

$$Q_w = \sum \frac{\epsilon_i^2}{\sigma_i^2} = \sum \left(\frac{Y_i - E\{Y_i\}}{\sigma_i} \right)^2 = \sum w_i \left(Y_i - E\{Y_i\} \right)^2$$

with $w_i = 1/\sigma_i^2$

  - This is called weighted least squared regression
  - We estimate the β's while taking into account difference in "information in Y's"

Weighted least-squares regression

- In matrix notation:

$$Q_w = \sum w_i \left(Y_i - E\{Y_i\}\right)^2 = (\mathbf{Y} - \mathbf{X}\beta)^t \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta)$$

  with $\mathbf{W}$ diagonal matrix with $w_i = 1/\sigma_i^2$'s on the diagonal
- Minimizing Q_w gives:

$$\mathbf{b_W} = (\mathbf{X^t W X})^{-1} \mathbf{X^t W Y} \text{ (is } k \text{ missing?)}$$

- $E\{\mathbf{b_W}\} = \beta$ and $\sigma^2\{\mathbf{b_W}\} = k(\mathbf{X^t W X})^{-1}$.
  Here, $s^2\{\mathbf{b_W}\} = MSE_W(\mathbf{X^t W X})^{-1}$

Weighted least-squares regression

- For non-constant variance model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,i} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \\ \epsilon_i &\sim N(0, k \cdot \sigma_i^2) \text{ (with } k > 0) \end{aligned}$$

  with $\sigma_i^2 = 1/w_i$ known, we estimate β by WLS estimator $\mathbf{b_W}$.
- How to estimate k?
  - For $Y_i \sim N(E\{Y_i\}, \sigma^2)$, we had $MSE = \frac{(Y_i - \hat{Y}_i)^2}{n-p}$, and $E(MSE) = \sigma^2$
  - For $Y_i \sim N(E\{Y_i\}, k/w_i)$:

$$\begin{aligned} \sqrt{w_i}\,(Y_i - E\{Y_i\}) &\sim N(0, \cdots) \\ MSE_W &= \frac{\sum w_i (Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum w_i e_i^2}{n-p} \end{aligned}$$

- Easy to get results in R: use "lm($\cdots$, weights $= 1/\sigma_i^2$)"

Weighted least-squares regression: Alternative write-up

- WLS-estimation for β in the non-constant variance model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \\ \epsilon_i &\sim N(0, k \cdot \sigma_i^2) \text{ (with } k > 0) \end{aligned}$$

  corresponds to (un-weighted) LS-estimation for β in

$$\mathbf{Y}^* = \mathbf{X}^* \beta + \epsilon^*,$$

  with

$$\begin{aligned} \mathbf{Y}^* &= \mathbf{W}^{(1/2)} \mathbf{Y}, \mathbf{X}^* = \mathbf{W}^{(1/2)} \mathbf{X}, \\ \epsilon^* &= \mathbf{W}^{1/2} \epsilon, \\ \mathbf{W}^{(1/2)} &= \text{Diagonal matrix } (\sqrt{w_i} \text{ on diagonal}) \end{aligned}$$

  with $\epsilon_i^* = \sqrt{w_i}\epsilon_i = \epsilon_i / \sigma_i \sim N(0, k)$

- Why?

Weighted least-squares regression: Alternative write-up

- For the model on the transformed scale:

$$\begin{aligned} Q &= \sum (Y_i^* - E\{Y_i^*\})^2, \\ &= (\mathbf{Y}^* - \mathbf{X}^*\beta)^t (\mathbf{Y}^* - \mathbf{X}^*\beta), \\ &= (\mathbf{Y} - \mathbf{X}\beta)^t \mathbf{W}^{1/2}\mathbf{W}^{1/2} (\mathbf{Y} - \mathbf{X}\beta) \\ &= Q_w \end{aligned}$$

What if σ_i^2 is unknown?

- Try to estimate σ_i from the data:
  1. Estimate σ_i^2 from Y replicates at same/similar X ourcomes
  2. Use relation between residuals and one/more predictors to estimate σ_i
     - E.g., use that if $\epsilon_i \sim N(0, \sigma_i^2)$, then $E\{|\epsilon_i|\} = \sqrt{2/\pi} \cdot \sigma_i$ (folded normal distribution)
- Example: suppose σ_i increases with X_i (and $X_i > 0$):

$$\epsilon_i \sim N(0, (k_0 + k_1 X_i)^2),$$

then $E\{|\epsilon_i|\} \propto (k_0 + k_1 X_i)$

What if σ_i^2 is unknown?–continued

- Use the absolute residuals to estimate σ_i:
  1. Do unweighted least-squares and get the absolute residuals $|e_i|$
  2. Esitimate σ_i by regressing the absolute residuals on X_i:
     - Fit model with $E\{|e_i|\} = \alpha_0 + \alpha_1 X_i$
     - Calculate $\hat{\sigma}_i = a_0 + a_1 X_i$
  3. Do weighted least-squared to get $\mathbf{b_W}$, with weight $w_i = 1/\hat{\sigma}_i^2$
  4. Repeat 2-3 if the regression coefficients have changed significantly (iteratively reweighted least-squares)

Example: blood pressure

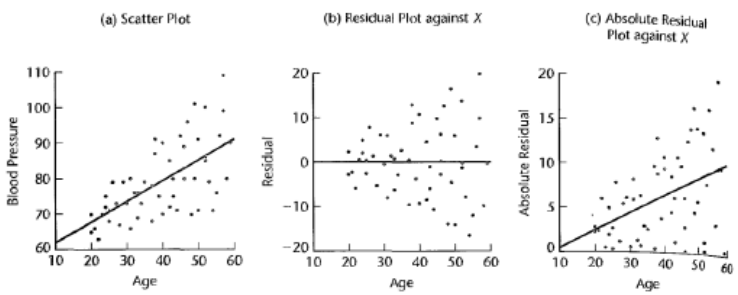
$$\hat{Y} = \underbrace{56.157}_{3.994} + \underbrace{0.58003}_{0.09695} X$$

$$\hat{s} = -1.54946 + 0.198172X$$

To estimate case 1 where $X_1 = 27$,

$$\hat{s}_1 = -1.54946 + 0.198172(27) = 3.801$$

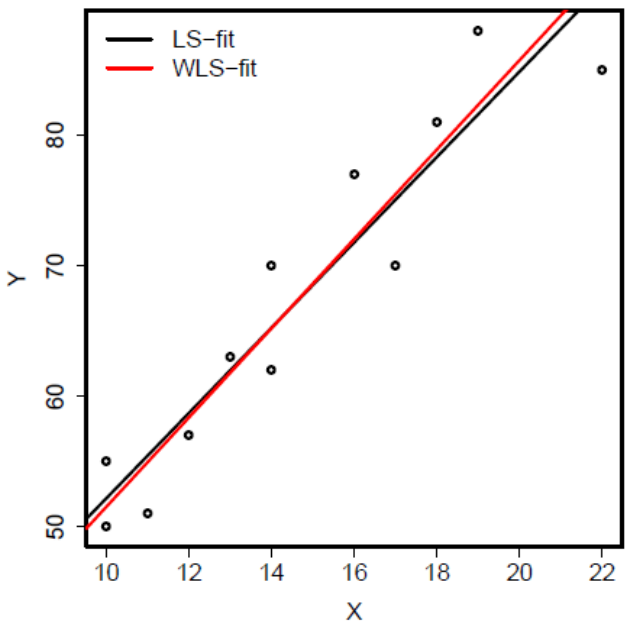
$$w_1 = \frac{1}{(\hat{s}_1)^2} = \frac{1}{(3.801)^2} = 0.0692$$

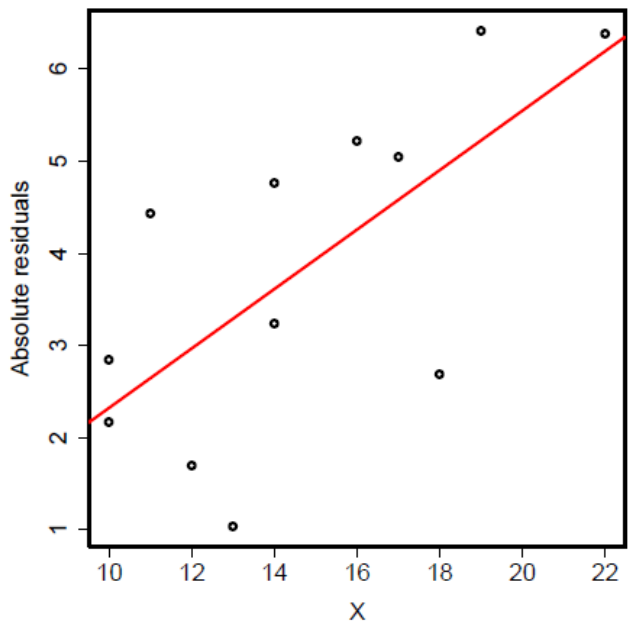


Subject i	(1) Age X_i	(2) Diastolic Blood Pressure Y_i	(3) e_i	(4) $\lvert e_i \rvert$	(5) $\hat{s}_i$	(6) w_i
1	27	73	1.18	1.18	3.801	.06921
2	21	66	−2.34	2.34	2.612	.14656
3	22	63	−5.92	5.92	2.810	.12662
...	...	...	...	...	...	...
52	52	100	13.68	13.68	8.756	.01304
53	58	80	−9.80	9.80	9.944	.01011
54	57	109	19.78	19.78	9.746	.01053

Example (Ex. 11.6): costs of computer-aided learning

- X =Number of responses, Y =Cost of computer time





Note that the stand errors of the b_w's are not correct if the σ_i's are estimated (the uncertainty in $\hat{\sigma}_i$ is not taken into account). We can estimate the standard errors with a bootstrap method.

Bootstrapping

- In statistics it refers to finding out properties of an estimator, without collecting new data, but by sampling from an approximate distribution
- The empirical distribution function is often used to sample from, which means sampling with replacement from original data set

Bootstrap example

- Interested in estimating the mean height of people world wide: how do we construct a confidence interval based on one sample?
- For constructing an confidence interval, we need to know the sampling distribution of the estimator (e.g., normal)
- But what if we would not know this sampling distribution?
  $\rightarrow$ use bootstrapping to obtain a sample from it!
- Using bootstrapping we randomly extract a new sample of n heights out of the n original heights, where each person can be selected many times:
  we create a large number of data sets that we might have seen
- We compute the mean for each of these "new" data sets:
  $\rightarrow$ we obtain a sample of means
  $\rightarrow$ we obtain a sample from the sampling distribution of the estimator

Bootstrap for regression models

- Why resampling with replacement:
  empirical distribution function will approach the underlying true distribution as $n \to \infty$
- Use empirical distribution function as our true distribution
- Statistical inference = results about distributions, some function of the true distribution (e.g., mean = $E(\cdot)$)
- Bootstrap: use approximation $\cdots$ with

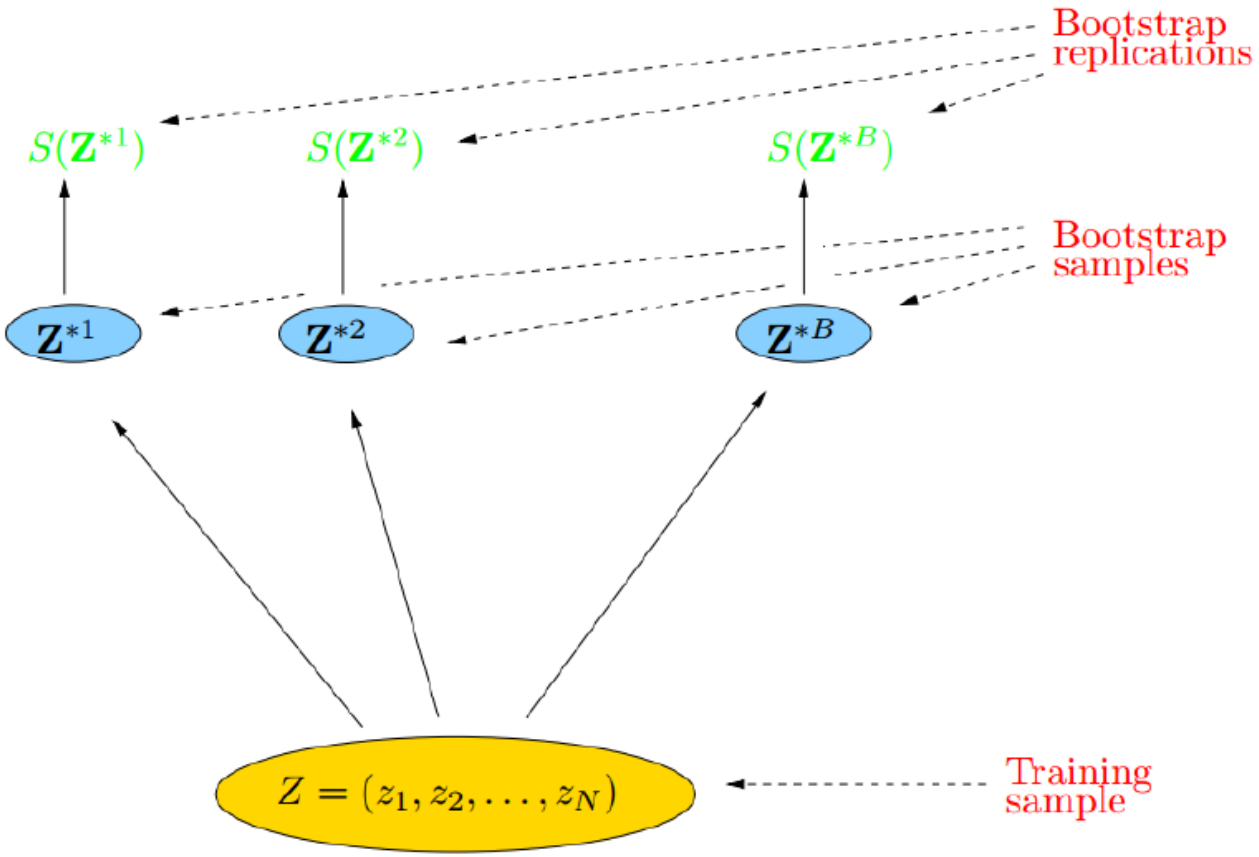
$$F_m(y) = \frac{\sum\{y_i \leq y\}}{n}$$

- Monte Carlo approximation: sample from the distribution to approximate the quantity of interest (e.g., mean).
- Sampling from the empirical distribution, is sampling from the data set with replacement

Bootstrap for regression models–continued

- If we want to estimate the variability in b_k (e.g., when σ_i was estimated from the data):
  1. Sample n bootstrap observations (X_i^b, Y_i^b) from the original data set (sample with replacement)
  2. Fit the regression model to the bootstrap sample, note the regression parameters as $\mathbf{b}^b$
  3. Repeat steps 1 and 2 B times
- We can estimate the variance of b_k by: $\frac{\sum_{b=1}^{B}(b_k^b - \bar{b}_k^b)}{B-1}$
- We can also use the sample of b_k^b's to calculate confidence intervals for β_k directly, using the sample percentiles $b_k^b(\alpha/2)$ and $b_k^b(1-\alpha/2)$ (useful when sampling distribution not normal)
- If regression model is appropriate and error terms have constant variance (e.g., for ridge regression), we can also do fixed X sampling: bootstrap the residuals, and add them to the fitted $\hat{Y}_i$'s

Schematic of the bootstrap processes



Accuracy or prediction?

- Traditional question: assessing the statistical accuracy of $S(\mathbf{Z})$: can be done by working with $S(\mathbf{Z}^{*\mathbf{1}}), \cdots, S(\mathbf{Z}^{*\mathbf{B}})$
- Modern challenge: estimating prediction error?

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

Any issue here?

- The reason is that the bootstrap data sets are acting as the training samples, while the original training set is acting as the test sample, and these two samples have observations in common.

Accuracy or prediction?

- The leave-one out bootstrap relaxes the over-fitting problem

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

(C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i, and $|C^{-i}|$ is the number of such samples)

Robust regression

- Robust regression gives automated procedure to reduce the impact of influential cases (but remember that examining outliers is important!)
- Fitted regression line not found by minimizing th esum of the squared errors, but something more robust to outliers, e.g., minimize:
  1. Sum of the absolute errors:
     least absolute residuals (LAR) regression
     $$\sum |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})|$$
  2. The median of squared errors (instead of mean):
     least median of squares (LMS) regression
     $$\mathrm{median_i}\{(\mathrm{Y_i} - (\beta_0 + \beta_1 \mathrm{X_{i1}} + \cdots + \beta_{\mathrm{p-1}} \mathrm{X_{i,p-1}}))^2\}$$
- Or use iteratively reweighted least squares estimation (IRLS)
- No analytical expression for the b_k's

Iteratively reweighted least squares (IRLS)

- Weighted linear regression, with the weights w_i for each case are based on "how far an outlying case is", measured by the scaled residuals u_i:

$$u_i = \frac{e_i}{\hat{\sigma}}$$

(to reduce influence by assigning weights varying inversely with the size of residual)

- Don't use $\sqrt{MSE}$ to scale the e_i's, use MAD (median absolute deviation) instead:

$$MAD = \frac{1}{0.6745}\text{median}\{|\text{e}_\text{i} - \text{median}_\text{l}(\text{e}_\text{l})|\}$$

(constant 0.6745 will make the MAD approximately unbiased for estimating σ from a normal distribution)

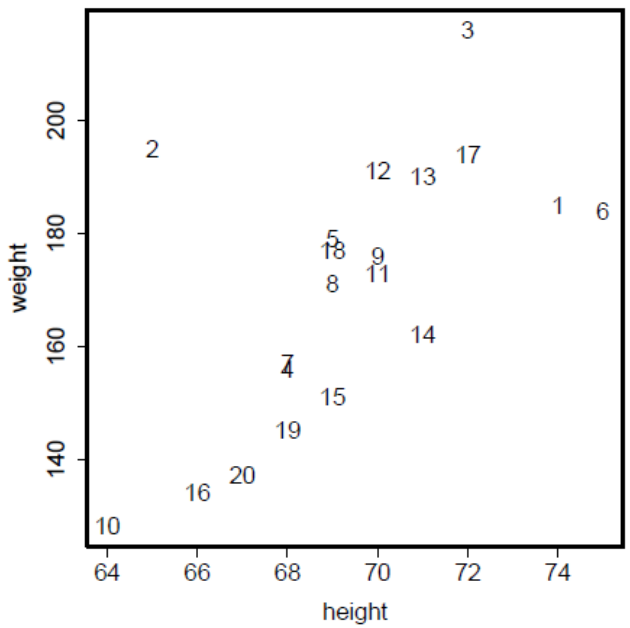
Iteratively reweighted least squares (IRLS)

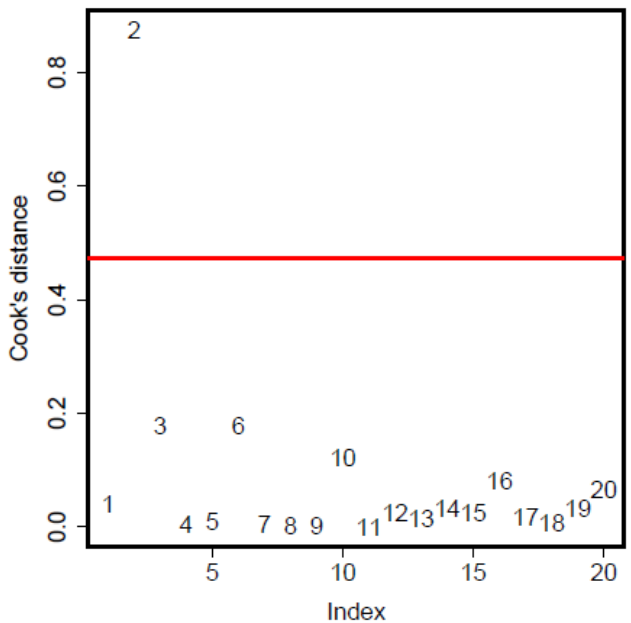
- How to assign weights?
  weight function should reduce the influence of influential cases
- Examples of weight functions: Huber, bisquare
- Huber weighting function:

$$w(u_i) = \begin{cases} 1 & |u_i| \leq 1.345 \\ 1.345/|u_i| & |u_i| > 1.345 \end{cases}$$

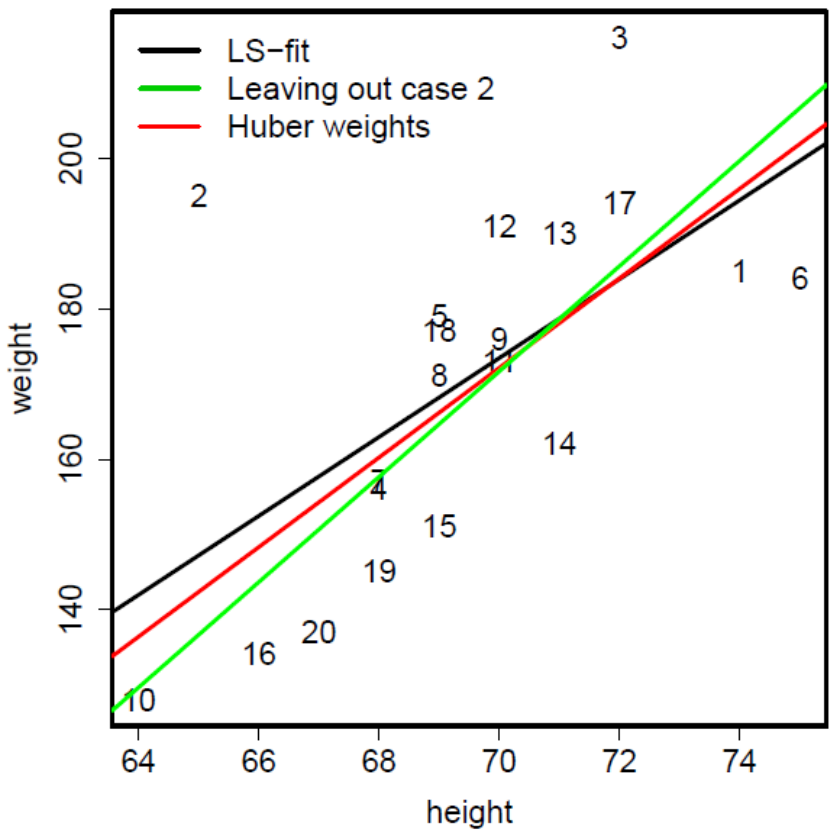
- The tuning constant 1.345 is chosen to make the procedure 95% efficient for normally distributed u_i's
- Weights are revised until the fit doesn't change much anymore

Example (Ex 11.12):
Weight (pounds) and height (inches) of 20 males students





Example (Ex 11.12): Weight (pounds) and height (inches) of 20 males students



Nonlinear models

- Regression model can be written as regression function + error terms:

$$Y_i = f(\mathbf{X}_i, \beta) + \epsilon_i$$

with here $\mathbf{X}_i$ referring to the predictor variables for the $i-$th case $\mathbf{X}_i = (1, X_{i1}, \cdots, X_{i,p-1})^t$

- For linear model, $f(\mathbf{X}_i, \beta)$ is linear in the β_k's:

$$f(\mathbf{X}_i, \beta) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} = \mathbf{X}\beta$$

- Examples of non-linear regression function:

$$\begin{aligned} Y_i &= \beta_0 \exp(\beta_1 \mathrm{X_i}) + \epsilon_\mathrm{i} \\ Y_i &= \beta_0 + \beta_1 \exp(\beta_2 \mathrm{X_i}) + \epsilon_\mathrm{i} \\ Y_i &= \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 \mathrm{X_i})} + \epsilon_i \end{aligned}$$

Estimation of regression parameters for non-linear regression

- If we can assume
$$\epsilon_i \sim N(0, \sigma^2)$$
then
$$Y_i \sim N(E\{Y_i\}, \sigma^2),$$
with $E\{Y_i\} = f(\mathbf{X_i}, \beta)$
- Least-squares estimation (or MLE) still works to find β, minimize
$$Q = \sum (Y_i - f(\mathbf{X_i}, \beta))^2$$
- However:
  - Usually no analytical solution as for linear response function
  - Use numerical search procedures
  - For inference about the β's:
    estimators for β's are not normally distributed, but we can assume normality if the sample size is large enough

Logistic regression

- What if the response variable is binary: $Y_i = 0$ or 1?
- Example: disease status Y_i, with $Y_i = 1$ if individual i has the disease, with predictors $X_{i1}, \cdots, X_{i,p-1}$.
- Probability model for binary outcomes $Y_i \sim \text{Bernoulli}(\pi_i)$:

$$\begin{cases} P(Y_i = 0) = 1 - \pi_I, \\ P(Y_i = 1) = \pi_I \end{cases}$$

  such that $E\{Y_i\} = \pi_I$, and $\text{var}(Y_i) = \pi_i(1 - \pi_i)$
- If we would write $Y_i = E\{Y_i\} + \epsilon_i$, then:

$$\epsilon_i = Y_i - E\{Y_i\} = \begin{cases} -\pi_i & \text{with probability } 1 - \pi_i \\ 1 - \pi_i & \text{with probability } \pi_i \end{cases}$$

- A general linear model will not work:
  - The mean response $E\{Y_i\} = \pi_i$ is bounded between 0 and 1
  - The error terms do not follow a normal distribution
  - The error variance is not constant

Logistic regression

- Main idea: model the expected value of Y_i, which then defines its distribution
- Logistic regression model $\rightarrow$ use the logistic function to model $E\{Y_i\}$:

$$E\{Y_i\} = \pi_i = \frac{\exp\left(\beta_0 + \sum_k \beta_k X_{ik}\right)}{1 + \exp\left(\beta_0 + \sum_k \beta_k X_{ik}\right)}$$

or similarly:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_k \beta_k X_{ik}$$

(plot!)

- Function for π_i is based on assuming underlying (latent) continuous variable W (which has a logistic distribution), that gives, for some cut-off value c : $Y_i = \begin{cases} 1 & W > c \\ 0 & W \leq c \end{cases}$

Interpretation of parameters: odds ratios

- β_1 is not just "increase" in EY associated with 1 unit increase in X"
- Use $\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 X_i$ to get interpretation of β_1
- The odds that $Y_i = 1$ are defined as:

$$\mathrm{odds}_i = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{\pi_i}{1 - \pi_i},$$

  thus $\mathrm{logit}(\pi_i)$ are the log-odds that $Y_i = 1$
- The estimated log-odds are given by: denote $\hat{\pi}'(X_i) = \widehat{\mathrm{logit}(\pi_i)}$ as the estimated log-odds at $X = X_i$:

$$\mathrm{logit}(\hat{\pi}_i) = \log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = b_0 + \sum_k b_k X_{ik},$$

  and the estimated odds:

$$\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \exp\left(b_0 + \sum_k b_k X_{ik}\right)$$

Interpretation of parameters: odds ratios

- Parameter interpretation:
  $\exp(b_k)$ gives the estimated odds ratio between two X levels, with 1 unit difference in X_k, for fixed values of the other X's with b_k's being estimators for β_k
- Then $\exp(b_1)$ gives the estimated odd ratio when X_1 differs by 1 unit:

$$\begin{aligned} b_1 &= \hat{\pi}'(X + (0, 1, 0, \cdots, 0)) - \hat{\pi}'(X) \\ &= \log(\text{odds at } (X + (0, 1, 0, \cdots, 0))) - \log(\text{odds at } X) \\ &= \log\left(\frac{\text{odds at } X + (0, 1, 0, \cdots, 0)}{\text{odds at } X}\right) \\ \exp(b_1) &= \left(\frac{odds_2}{odds_1}\right) \end{aligned}$$

Simple example

- Fitted logistic regression function for lung cancer data set given by:
$$\mathit{logit}(\hat{\pi}) = 1 + 0.05 \cdot \mathit{age} + 5 \cdot \mathit{Smoker},$$
- The estimated odds ratio between smokers and non-smokers of the same age is
$$\exp(b_2) = \exp(5) \approx 150$$
- The estimated odds ratio between non-smokers of ages 70 and 50 are
$$\exp(b_1 \cdot 20) = \exp(0.05 \cdot 20) \approx 3$$

Generalized linear model

- Both the general linear model as well as the logistic model are examples of a generalized linear model, where the mean of Y_i is "linked" to a linear function of regression parameters
- General linear model:

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \sum_k \beta_k X_{ik} \\ E\{\mathbf{Y}\} &= \mathbf{X}\beta \\ g\left(E\{\mathbf{Y}\}\right) &= \mathbf{X}\beta, \end{aligned}$$

where $g(\cdot)$ is called a link function which links the regression structure $\mathbf{X}\beta$ to $E\{\mathbf{Y}\}$. Here it is an identity function.

Generalized linear model

- Logistic regression model:

$$\begin{aligned} E\{Y_i\} &= \frac{\exp(\beta_0 + \sum_k \beta_k X_{ik})}{1 + \exp(\beta_0 + \sum_k \beta_k X_{ik})}, \\ g\left(E\{\mathbf{Y}\}\right) &= \mathbf{X}\beta \end{aligned}$$

  where link function $g(\cdot)$ is the logit function.
- Numerical search procedures are used to find maximum likelihood estimates of regression parameters in GLMs.

Likelihood function

- For GLM, LS principle is usually not applicable. Instead, maximum likelihood method is a powerful tool in this case.
- We use logit link function as an example for logistic regression.
- The log-likelihood function is given by

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \{ Y_i \log(\pi_i/(1-\pi_i)) \} + \sum_{i=1}^{n} \log(1-\pi_i) \\ &= \sum_{i=1}^{n} Y_i \cdot \mathbf{x_i^t}\beta + \sum_{i=1}^{n} \log\{1+\exp(\mathbf{x_i^t}\beta)\} \end{aligned} \tag{1}$$

Maximum Likelihood Estimation

- Maximize $l(\beta)$ in (1) to obtain $\mathbf{b}$, the maximum likelihood estimates of β.
- There is no explicit form for $\mathbf{b}$. Numerical search methods are needed to find b.
- $\hat{\pi}_i$ is then obtained by

$$\hat{\pi}_i = \frac{e^{\mathbf{x_i^t b}}}{1 + e^{\mathbf{x_i^t b}}}$$

Inferences about regression parameters

- Statistical inference on β in GLM is usually relying on large sample size.
- Let

$$G(\beta) = \frac{\partial^2 l(\beta)}{\partial \beta \beta^t}$$

  where $l(\beta)$ is the log-likelihood given by (1).
- $s^2(\mathbf{b}) = -G^{-1}(\mathbf{b})$
- The intuition of the above result can be seen by retrieving the case of "normal". The theoretical derivation relies on large sample theories.

Test concerning β_k

- $H_0 : \beta_k = 0$ versus $H_a : \beta_k \neq 0$. n must be large.
- Wald's test: $z^* = \frac{b_k}{s(b_k)} \approx N(0,1)$ when H_0 is true.
- $H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$ versus $H_a : H_0$ is not true.
- Likelihood ratio test: $G^2 = -2\{l(R) - l(F)\} \approx \chi^2_{p-q}$ when H_0 is true. The rejection criteria is one-sided.
  - $l(R)$ is the maximized log-likelihood (eq. (1)) when H_0 is true.
  - $l(F)$ is the maximized log-likelihood for the full model.

Inference about $E\{Y_h\}$ and Y_h

- Inference about $E\{Y_h\}$ and Y_h for a given $x_h = (1, x_{h1}, x_{h2}, \cdots, x_{h,p-1})^t$ can be obtained accordingly based upon $s^2(\mathbf{b})$
- For example, $\pi_h = E\{Y_h\}$:
  - $\hat{\pi}_k = \{1 + \exp(-\mathbf{x_h^t b})\}^{-1}$
  - C.I. for $\mathbf{x_h^t}\beta$ can be obtained easily, since $s^2(\mathbf{x_h^t b}) = \mathbf{x_h^t} s^2(\mathbf{b}) \mathbf{x_h}$. Assume its C.I.'s lower and upper bounds are $L \& U$ respectively.
  - C.I. for π_h is $\left(\{1 + \exp(-L)\}^{-1}, \{1 + \exp(-U)\}^{-1}\right)$

Example: disease outbreak

- Textbook p. 573
- Y =disease status: 1=with disease, 0=without disease.
  X_1: age.
  X_2: socioeconomic status. Factor variable with 3 levels.
  X_3: sector. Factor variable with 2 levels.
- Use R to fit the logistic regression:

```
X2 = as.factor(X2)
X3 = as.factor(X3)
logit_fit = glm(Y~X1+X2+X3, family = binomial("logit"))
summary(logit_fit)
logit_fit$fitted # returns are π̂_i
vcov(logit_fit)
```

Example: disease outbreak–continued

	Case i	(1) Age X_{i1}	(2) Socioeconomic Status X_{i2}	(3) Socioeconomic Status X_{i3}	(4) City Sector X_{i4}	(5) Disease Status Y_i	(6) Fitted Value $\hat{\pi}_i$
(Coded)	1	33	0	0	0	0	.209
	2	35	0	0	0	0	.219
	3	6	0	0	0	0	.106
	4	60	0	0	0	0	.371
	5	18	0	1	0	1	.111
	6	26	0	1	0	0	.136
	...	...	...	...	...	...	...
	98	35	0	1	0	0	.171

Example: disease outbreak–continued

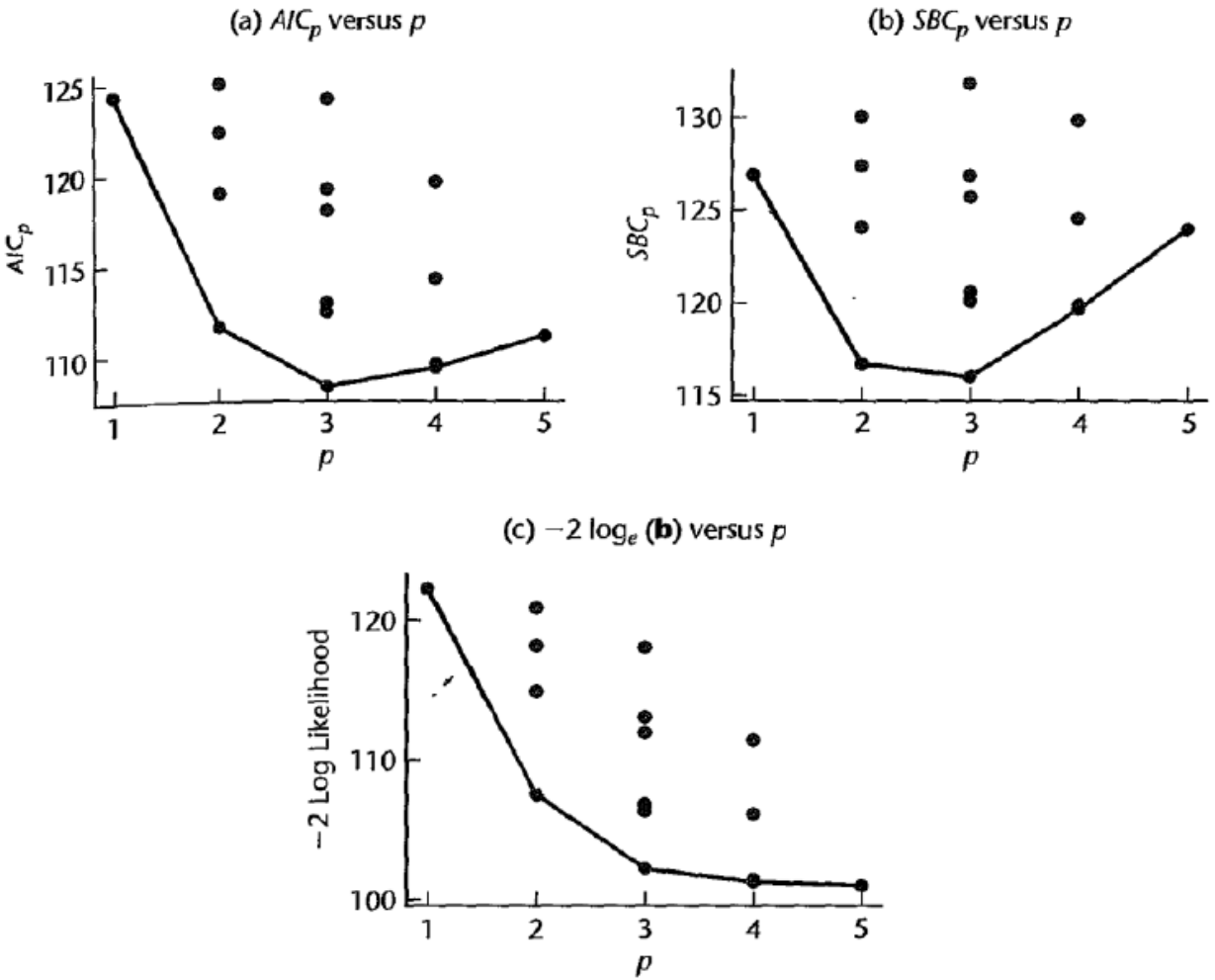
(a) Estimated Coefficients, Standard Deviations, and Odds Ratios

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	Estimated Odds Ratio
β_0	−3.8877	.9955	—
β_1	.02975	.01350	1.030
β_2	.4088	.5990	1.505
β_3	−.30525	.6041	.737
β_4	1.5747	.5016	4.829

(b) Estimated Approximate Variance-Covariance Matrix

$$\mathbf{s}^2\{\mathbf{b}\} = \begin{array}{c} \\ b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{array} \begin{array}{c} \begin{array}{ccccc} b_0 & b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{bmatrix} .4129 & -.0057 & -.1836 & -.2010 & -.1632 \\ -.0057 & .00018 & .00115 & .00073 & .00034 \\ -.1836 & .00115 & .3588 & .1482 & .0129 \\ -.2010 & .00073 & .1482 & .3650 & .0623 \\ -.1632 & .00034 & .0129 & .0623 & .2516 \end{bmatrix} \end{array}$$

Example: disease outbreak–continued



Variable selection

- AIC and SBC criteria in our last lecture can be easily adopted and used in logistic regression.
  - $AIC_p = -2 \log l(\mathbf{b}) + 2p$
  - $SBC_p = -2 \log l(\mathbf{b}) + p \log(n)$
- Stepwise regression techniques based upon Wald's test or AIC or SBC can be implemented.
- AIC, SBC have been implemented in R:

```
step(logit_fit, scope = list(lower = ~1, upper =
~X1+X2+X3), direction = "backward")
step(logit_fit, scope = list(lower = ~1, upper =
~X1+X2+X3), direction = "backward", k = log(length(Y)))
```

Poisson Regression

- Poisson regression is usually used when the response Y is a count.
- Example: in order to study conditions that affect # of times a household shops at a particular supermarket in a week.
  Response: # of times a household shops at a particular supermarket in a week.
  Predictors: the family's income, # of children, distance form the store, etc...
- Y follows a Poisson distribution if

$$P(Y = y) = f(y) = \frac{\mu^Y \exp(-\mu)}{y!}, \; y = 0, 1, 2, \cdots$$

- Clearly, $E\{Y\} = \mu$, $Var\{Y\} = \mu$.

Three-part specification of Poisson regression

- **The distribution assumption:** $P(Y_i = y_i) = f(y_i)$ as defined in the last slide.
- **Systematic component:**
  $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} = \mathbf{x_i^t}\beta$, where
  $\mathbf{x_i} = (1, x_{i1}, \cdots, x_{i,p-1})^t$
- **The link function:** once again,the form is not unique. The most popularly used one is the canonical form.
  $g(\mu_i) = \log(\mu_i)$

Fitting and inferences

- The fitting (maximum likelihood method) and inferences for Poisson model are carried out in the very similar way as for logistic regression.
- Example (textbook p. 621):
  - Y_i : Number of customers who visited store from census tract
  - X_1 : Number of housing units
  - X_2 : Average income, in dollars
  - X_3 : Average housing unit age, in years
  - X_4 : Distance to nearest competitor, in miles
  - X_5 : Distance to store, in miles

R codes for Poisson regression

```
poisson_fit = glm(Y~X1+X2+X3+X4+X5, family = poisson)
summary(poisson_fit)
poisson_fit$fitted
vcov(poisson_fit)

step(poisson_fit, scope = list(lower = ~1,
upper = ~X1+X2+X3), direction = "backward") #AIC

step(poisson_fit, scope = list(lower = ~1,
upper = ~X1+X2+X3), direction = "backward",
k = log(length(Y))) #BIC
```

Example: household shopping

(a) Fitted Poisson Response Function

$$\hat{\mu} = \exp[2.942 + .000606X_1 - .0000117X_2 - .00373X_3 + .168X_4 - .129X_5]$$

$$DEV(X_0, X_1, X_2, X_3, X_4, X_5) = 114.985$$

(b) Estimated Coefficients, Standard Deviations, and G^2 Test Statistics

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	G^2	P-value
β_0	2.9424	.207		
β_1	.0006058	.00014	18.21	.000
β_2	−.00001169	.0000021	31.80	.000
β_3	−.003726	.0018	4.38	.036
β_4	.1684	.026	41.66	.000
β_5	−.1288	.016	67.50	.000

Multiple Linear Regression as a Special Case of GLM

- Normal distribution belongs to exponential family.
- Multiple linear regression model we have studied over the semester is a special case of GLM, if we use maximum likelihood principle to estimate parameters. (Q: what's the difference between using maximum likelihood principle and LS principle?)

Three-part specification for normal response

- **The distributional assumption:** $Y_i \sim N(\mathbf{x_i^t}\beta, \sigma^2)$. Now that $E\{Y_i\} = \mu_i = \mathbf{x}_i^t\beta$
- **Systematic component:** $g(\mu_i) = \mathbf{x_i}^t\beta$, where $\mathbf{x}_i$ is the $i-$th row of $\mathbf{X}$
- **The link function**, the form is not unique. However, the canonical form is $g(\mu_i) = \mu_i$, the identity function.
- Therefore, if we use the canonical link function, the log-likelihood function in the case of GLM is the same as the classical log-likelihood for multiple linear regression models. Everything traces back to the maximum likelihood principle we have in the previous.