Advanced Methods in Statistics (Fall 2016)

J.P.Kim

Dept. of Statistics

Finally modified at October 3, 2016

Preface & Disclaimer

This note is a summary of the lecture Advanced Methods in Statistics (326.521) held at Seoul National University, Fall 2016. Lecturer was Jaeyong Lee, and the note was summarized by J.P.Kim, who is a Ph.D student. Note that in here, each section means each lecture or topic. There are few textbooks and references in this course, which are following.

- *An Introduction to Statistical Learning, James, Witten, Hastie & Tibshirani, 2013.*
- *Machine Learning: a probabilistic perspective, K.P.Murphy, 2012.*
- *Applied Bayesian Modelling, P.Congdon, 2014.*
- *The elements of Statistical Learning, Friedman, Hastie & Tibshirani, 2001.*

Also I referred to following books when I write this note. The list would be updated continuously.

- *Mathematical Statistics: Basic ideas and selected topics, Vol. I., 2nd edition, P.Bickel & K.Doksum, 2007.*
- *Linear Models in Statistics, Rencher & Schaalje, 2008.*
- *A first course in Bayesian statistical methods, P.D.Hoff, 2009.*
- *Statistical decision theory and Bayesian Analysis, James O. Berger, 2013.*
- *Introduction to Nonparametric Regression, K.Takezawa, 2005.*
- *Applied Multivariate Statistical Analysis, R.Johnson & D.Wichern, 2007.*

If you want to correct typo or mistakes, please contact to: `joonpyokim@snu.ac.kr`

Contents

1 Introduction to Bayesian Statistics

1.1 Basic concepts of Bayesian Inference

In statistical inference, we use the parametric model to obtain the information about nature. In here, parameter denotes *the nature state*, and we *observe* the observation. Often we use θ for parameter, and x for the observation. Our model is about 'the distribution of observation when the parameter is given,' i.e., $x|\theta \sim p(x|\theta)$.

Frequentists assume that the nature state is nonrandom, so parameter is unknown constant. However, Bayesians want to represent *all of uncertain information* as a probability distribution. Before we perform data analysis, we have the *belief* or *priori information* about the nature. For example, if we toss a coin, we believe that probability for head is near to 0.5. In Bayesian statistics we define such information as a form of probability distribution, and it is called a **prior distribution**, or in short, a **prior**.

There are three basic elements of Bayesian inference. First is a **prior distribution**, $\theta \sim \pi(\theta)$. Next is an **observation**. We often use the model $x|\theta \sim p(x|\theta)$. Important thing is that *every information used for estimating θ should be from the observation.* Observation is also called as a likelihood. Finally, after observing the data, we can obtain an updated information about θ, which is called a **posterior distribution**. Note that posterior distribution is $\theta|x \sim \pi(\theta|x)$, which means the distribution of the parameter after observing x.

Main goal of Bayesian inference is to obtain posterior distribution. Then how to obtain it? Next theorem makes it possible.

Proposition 1.1 (Bayes' rule)**.** *Posterior distribution is obtained as*

$$\pi(\theta|x) = \frac{p(x,\theta)}{m(x)} = \frac{\pi(\theta)p(x|\theta)}{m(x)} \propto \pi(\theta)p(x|\theta),$$

where $m(x) = \int \pi(\theta)p(x|\theta)d\theta$.

Example 1.2. Suppose that we toss a pin n times. There are two possibilities in the result: one is $\nearrow$, and the other is $\perp$. Let θ be a probability for $\nearrow$. Then our model is $x|\theta \sim Bin(n,\theta)$. Since we don't have any priori information about the simulation, we can use the noninformative prior, $\theta \sim U(0,1)$. Then posterior distribution is obtained as

$$\pi(\theta|x) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and hence $\theta|x \sim Beta(x+1, n-x+1)$.

Remark 1.3. Often we use a *point estimator* of θ, rather than posterior distribution. It would be one point summarization of the distribution. Posterior mean, median or MAP (*Maximum a Posteriori*) is frequently used.

Definition 1.4. *Interval* (L, U) *such that*

$$P(L < \theta < U|x) = 1 - \alpha$$

is called a $100(1-\alpha)\%$ *credible set or confidence interval.*

Remark 1.5. Note that frequentist and Bayesian view for confidence region is different. In Bayesian view, we can say that a **probability** the parameter contained in region is $1-\alpha$. Also note that, Bayesians represent all of uncertainty as a probability distribution.

1.2 Bayesian hypothesis testing

Consider the model $x|\theta \sim f(x|\theta)$ and hypotheses

$$H_0 : \theta = \theta_0 \; vs \; H_1 : \theta = \theta_1.$$

It can be also written as

$$H_0 : x \sim f(x|\theta_0) \; vs \; H_1 : x \sim f(x|\theta_1).$$

First, let $\pi_0 = \pi(H_0)$ and $\pi_1 = \pi(H_1)$ be prior probabilities of H_0 and H_1, respectively. Clearly $0 < \pi_0, \pi_1 < 1$ and $\pi_0 + \pi_1 = 1$ should be held. Then posterior probability of each hypothesis is obtained as

$$\pi(H_0|x) = \frac{\pi_0 f(x|H_0)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)}$$

and

$$\pi(H_1|x) = 1 - \pi(H_0|x) = \frac{\pi_1 f(x|H_1)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)}.$$

Now hypothesis procedure is that, we support H_1 if $\pi(H_1|x)/\pi(H_0|x)$ is sufficiently large.

Definition 1.6 (Bayesian Testing)**.** *Let*

$$\frac{\pi(H_1|x)}{\pi(H_0|x)} = \frac{\pi_1}{\pi_0} \cdot \frac{f(x|H_0)}{f(x|H_1)}$$

be a posterior odds. Then

$$B_{10} := \frac{f(x|H_0)}{f(x|H_1)}$$

is called a ***Bayes factor****. Then clearly we get*

$$\textit{posterior odds} = \textit{prior odds} \times \textit{Bayes factor}.$$

Note that Bayes factor has similar role to the "likelihood ratio" in frequentist's inference.

Remark 1.7. We additionally need a criterion to judge whether B_{10} is large or not. One can use *Jeffreys' criterion* following.

$\log_{10} B_{10}$	B_{10}	Strength of evidence for H_1
$0 \sim 1/2$	$1 \sim 3.2$	Not worth a bare mention
$1/2 \sim 1$	$3.2 \sim 10$	substantial
$1 \sim 2$	$10 \sim 100$	strong
> 2	> 100	decisive

Example 1.8. Consider a pin example again. Suppose that we want to test

$$H_0 : \theta = \frac{1}{2} \; vs \; H_1 : \theta = \frac{2}{3}.$$

Assume that we observed $x = 7$. Then Bayes factor is

$$B_{10} = \frac{\binom{10}{7}(2/3)^7(1/3)^3}{\binom{10}{7}(1/2)^7(1/2)^3} = 2.2197.$$

Also note that

$$\pi(H_1|x) = \frac{\pi_1 f(x|H_1)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)} = \frac{\pi_1 B_{10}}{\pi_0 + \pi_1 B_{10}} = 0.6894.$$

Thus we may support H_0.

2 Bayesian Computation

In Bayeisan analysis, every information about θ is summarized in posterior. However, in many cases, characteristic of posterior distribution (such as moments or quantiles) is difficult to find. Then we should estimate or approximate moments or quantiles. In this section, we introduce some estimation methods and random generation algorithm, which is known as Markov Chain Monte Carlo (MCMC), and give some references.

2.1 Monte Carlo

If one wants to estimate

$$I = Ef(X) = \int f(x)g(x)dx,$$

where $g(x)$ is a density and $X \sim g(x)$, then we can use a random sample $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} g$ and estimate I as

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Note that by SLLN, $\hat{I} \xrightarrow[n\to\infty]{a.s} I$. Also, it can be easily verified that

$$\hat{se}(\hat{I}) = \sqrt{\frac{\nu}{n}}$$

where

$$\nu = \widehat{Var} f(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(X_i) - \hat{I}\right)^2.$$

Note that error rate is $O(n^{-1/2}$, which does not depend on the dimension of integration. It is a strong point when we compare Monte Carlo method with other numerical approaches, such as trapezoidal rule or quadrature rules.

Example 2.1. Let $X|\theta \sim N(\theta, 1)$ and $\theta \sim Cauchy(0, 1)$. Then by Bayes' rule posterior is obtained as

$$\pi(\theta|x) \propto \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2},$$

so posterior mean (=Bayes estimator) is

$$E(\theta|x) = \frac{\int \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2}}{\int \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2}}.$$

To estimate this, we may use a random sample $\theta_i \overset{i.i.d.}{\sim} N(x,1)$, and use

$$\hat{E}(\theta|x) = \frac{\sum_{i=1}^n \frac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^n \frac{1}{1+\theta_i^2}}.$$

Importance sampling

Suppose that we want to estimate

$$I = \int f(x)g(x)dx = E[f(X)]$$

where $X \sim g$. However, random generation from g is difficult problem, so we can't handle it. Instead, there is $\pi \approx g$ such that we can easily obtain random numbers from π. Then using π we can estimate I as a *weighted average of observed values*. Precisely, let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} \pi$. Then since

$$\int f(x)g(x)dx = \int \frac{f(x)g(x)}{\pi(x)}\pi(x)dx = E\left[\frac{f(X)g(X)}{\pi(X)}\right]$$

we can estimate I as

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^n w_i f(X_i)$$

$$\hat{I}_2 = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{i=1}^n w_i}$$

where

$$w_i = \frac{g(X_i)}{\pi(X_i)}.$$

Note that $\hat{I}_2$ is biased estimator for I but has some advantages compared to $\hat{I}_1$. If we use $\hat{I}_2$, we don't have to know normalizing constant of g. To use $\hat{I}_1$, we may estimate $\int g(x)dx$ again to make g a density function.

Or we can do a further step in importance sampling, which is known as *sampling importance sampling (SIR)*. Note that, Monte Carlo can be treated as approximation of pdf to *empirical cdf*, $g(x) \approx n^{-1}\sum \delta_{x_i}$, where δ_{x_0} is Dirac measure. Importance sampling is an approximation $g(x) \approx \sum x_i \delta_{x_i}$. Sampling importance resampling is, resample $X_1^*, \cdots, X_n^*$ from the distribution estimated by (X_i, w_i), i.e., $g(x) \approx \sum w_i \delta_{x_i}$, and estimate I as

$$\hat{I}_3 = \frac{1}{n}\sum_{i=1}^n f(x_i^*).$$

2.2 Markov Chain Monte Carlo

Monte Carlo is useful when θ is high-dimensional, because its error may not depend on the dimension. However, as dimension goes high, it becomes difficult to generate a random number. Importance sampling is one alternative for this, but it may not be a good answer, because if dimension is high, $\pi \approx g$ becomes a hard goal ("curse of dimension"). In this sense, Markov Chain Monte Carlo can be other alternative.

Gibbs Sampling

Gibbs sampling is an algorithm that generates random numbers using previous numbers. Let $f(x_1, x_2, \cdots, x_p)$ be a density function, and denote $x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)$.

Algorithm 1 (systematic sweep) Gibbs Sampling

1: Initialize $(X_1^{(0)}, X_2^{(0)}, \cdots, X_p^{(0)})$.
2: **for** $t = 1, 2, 3, \cdots$ **do**
3: Sample $X_1^{(t)} \sim f_{X_1|X_{-1}}(\cdot|X_2^{(t-1)}, \cdots, X_p^{(t-1)})$.
4: Sample $X_2^{(t)} \sim f_{X_2|X_{-2}}(\cdot|X_1^{(t)}, X_3^{(t-1)}, \cdots, X_p^{(t-1)})$.
$\vdots$
5: Sample $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)}, \cdots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \cdots, X_p^{(t-1)})$
$\vdots$
6: Sample $X_p^{(t)} \sim f_{X_p|X_{-p}}(\cdot|X_1^{(t)}, \cdots, X_{p-1}^{(t)})$.
7: **end for**

Note that sequence of random variables $(X^{(0)}, X^{(1)}, X^{(2)}, \cdots)$ obtained via Gibbs sampling becomes a *Markov Chain whose stationary distribution is* $f(x_1, \cdots, x_p)$. Furthermore, under some conditions, sample moments or quantiles of $(X^{(0)}, X^{(1)}, X^{(2)}, \cdots)$ converge to those of population.

We can consider *random sweep* Gibbs sampling algorithm.

Algorithm 2 random sweep Gibbs Sampling

1: Initialize $(X_1^{(0)}, X_2^{(0)}, \cdots, X_p^{(0)})$.
2: **for** $t = 1, 2, 3, \cdots$ **do**
3: Choose $j \in \{1, 2, \cdots, p\}$ randomly.
4: Sample $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)} \cdots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \cdots, X_p^{(t-1)})$ and let $X_i^{(t)} = X_i^{(t-1)}$ $\forall i \neq j$.
5: **end for**

Example 2.2. In this example, our goal is to apply Gibbs sampling to find random numbers

with stationary distribution

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

First initialize $(x_1^{(0)}, x_2^{(0)})$. Then update x_1 and x_2 as following.

(1) $x_1^{(t)} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2^{(t-1)} - \mu_2), \sigma_1^2(1-\rho^2)\right)$

(2) $x_2^{(t)} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1^{(t)} - \mu_1), \sigma_2^2(1-\rho^2)\right)$

Finally we can estimate population characteristics. For example,

$$\int h(x_1, x_2)\pi(dx_1, dx_2) \approx \frac{1}{n}\sum_{t=1}^{n} h(x_1^{(t)}, x_2^{(t)}),$$

$$\text{and } P(X_1 \geq 0, X_2 \geq 0) \approx \frac{1}{n}\sum_{t=1}^{n} I(x_1^{(t)} \geq 0, x_2^{(t)} \geq 0).$$

Metropolis-Hastings Algorithm

Goal of Metropolis-Hastings Algorithm is to find Markov chain with stationary distribution $\pi(\theta)$, where π is given. Note that if kernel $K(x, dy)$ of Markov chain is determined, Markov chain is also determined. For this, kernel K should satisfy

$$\int \pi(dx)K(x, dy) = \pi(dy).$$

It is known that if K *satisfies detained balance condition*

$$\pi(dx)K(x, dy) = \pi(dy)K(y, dx)\ \forall x, y \in S,$$

then π is a stationary distribution of $K(x, dy)$. Now our question is: *for arbitrary proposal kernel $q(x, y)$, can we find a kernel K which satisfies detailed balance condition?* If we choose α satisfying

$$\alpha(x, y)\pi(x)q(x, y) = \pi(y)q(y, x)$$

and define *Metropolis-Hastings kernel*

$$K(x, dy) = \alpha(x, y)q(x, y)dy + (1 - \alpha(x))\delta_x(dy)$$

where

$$\alpha(x) = \int \alpha(x,y)q(x,y)dy,$$

then K satisfies detailed-balance condition. We can make this as choosing

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right).$$

Algorithm 3 Metropolis-Hastings Algorithm

1: Initialize $x^{(0)}$.
2: **for** $t = 1, 2, 3, \cdots, m$ **do**
3: Draw $x \sim q(x^{(t-1)}, \cdot)$ and $u \sim U(0,1)$ independently.
4: Obtain acceptance rate

$$\alpha(x^{(t-1)}, x) = \min\left(1, \frac{\pi(x)q(x, x^{(t-1)})}{\pi(x^{(t-1)}q(x^{(t-1)}, x)}\right).$$

5: Define

$$x^{(t)} = \begin{cases} x & \text{if} \quad u \le \alpha(x^{(t-1)}, x) \\ x^{(t-1)} & \text{if} \quad u > \alpha(x^{(t-1)}, x) \end{cases}.$$

6: **end for**

Convergence diagnostics

To judge whether generated random numbers converged or not, we may use *time series plot, cumulative sum plot, ACF plot, log-likelihood or log-posterior plot.* If first some samples did not converge and may not represent the stationary distribution, then we would *not* use such samples, and only use the numbers generated later. Such approach is called *burn-in.* Or, if ACF is too high to say that samples are independent, then we may use not all of samples, but some of them, until ACF becomes small. Such approach is called *thinning.*

Now we will see *effective sample size.* Note that for a stationary time series y_t and

$$H_m = \frac{1}{m}\sum_{t=1}^{m} h(y_t),$$

$$\sqrt{m}(H_m - Eh(y)) \xrightarrow[n\to\infty]{d} N(0, \nu^2)$$

holds for

$$\nu^2 = \sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

under some conditions. If there are m MCMC samples with ACF ρ_k, *effective sample size* is

defined as

$$M = \frac{m}{1 + 2\sum_{k=1}^{\infty} \rho_k}.$$

Note that sample mean based on m MCMC samples has the same variance with sample mean based on M i.i.d. samples.

3 Hierarchical Models

Consider a data obtained as following: There are 71 different groups rats. We are interested in the rate of *endometrial stromal polyps* in the different groups. The number of rats varies from group to group. Let y_i be the number of tumors in group i. Our model is based on binomial model,

$$y_i|\theta_i \overset{indep}{\sim} Bin(n_i, \theta_i).$$

If we assume that $\theta_1 = \theta_2 = \cdots = \theta_{71}$, then the problem becomes simple, but the assumption has many problems. First, can we believe $\theta_1 = \cdots = \theta_{71}$ holds? Under such assumption, for

$$\bar{p} = \frac{\sum x_i}{\sum n_i},\ \hat{p}_i = \frac{x_i + 0.5}{n_i + 1},$$

$$z_i = \frac{\hat{p}_i - \bar{p}}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n_i}}$$

should be $N(0,1)$ distributed approximately. In here, we used $\hat{p}_i = (x_i + 0.5)/(n_i + 1)$ because in the data $x_i = 0$ is observed for many i. However, if we plot histogram or Q-Q plot, we can easily verify that such assumption is not reasonable.

We can solve this problem by using prior distribution,

$$\theta_i \sim Beta(\alpha, \beta).$$

Even though $\theta_1 = \cdots = \theta_{71}$ did not make sense, it is reasonable to suppose that θ_is have similar values. it can be reflected on the assumption that *each θ_i was drawn from the same distribution.* There are other 70 groups that we observed, and using the observations we can estimate hyperparameters α and β. We can use MLE or MME. In this lecture, MME was used. Such approach is called *empirical Bayes.*

Although it seems very reasonable, there are some problems. First, if we wanted to infer $\theta_1, \theta_2, \cdots$ as well as θ_{71}, then we should repeat such procedures in many times. In addition, we used estimated values α and β *just as we have the exact information about prior*, even if there is uncertainty from estimation of hyperparameters.

One alternative we can use is **hierarchical model**. In here, our model is:

$$y_i|\theta_i \sim Bin(n_i, \theta_i),\ i = 1, 2, \cdots, 71,$$

$$\theta_i \sim Beta(\alpha, \beta),\ i = 1, 2, \cdots, 71.$$

We employee Gamma *hyperprior* in this lecture,

$$\alpha \sim Gamma(a_\alpha, b_\alpha),\ \beta \sim Gamma(a_\beta, b_\beta).$$

Since we don't have information about α and β, we may set $Var(\alpha)$ and $Var(\beta)$ large. For example, we can suppose that

$$E(\alpha) = \frac{a_\alpha}{b_\alpha} = 1,\ Var(\alpha) = \frac{a_\alpha}{b_\alpha^2} = 10^3,$$

and similar for β.

Before finishing, there are some remarks. First, in hierarchical model, prior can be interpreted as a part of model, and hyper-prior has a role of prior. Note that Bayesian employees prior distribution to represent *uncertainty of the nature*. Next, we can also think further-hyperprior of hyperparemeter, but it is not justified because (i) There are no data more that would be used, and (ii) uncertainty of hyperprior distribution may affects less than prior, so it may not be needed. Finally, if we choose a prior, there can be two approaches: (i) *mimic* "non-information," or (ii) reflect information that we have. One example of using latter view is an empirical Bayes approach on hyperprior.

4 Dirichlet Process

4.1 Definition and properties of Dirichlet process

Consider a Bayesian nonparametric model

$$X_1, X_2, \cdots, X_n | F \overset{i.i.d.}{\sim} F.$$

Note that in parametric model, the model is

$$X_1, X_2, \cdots, X_n | \theta \overset{i.i.d.}{\sim} f(x|\theta)$$

and we assume prior for θ on Θ. However, in nonparametric case, there is no assumption about F, so we should consider a prior distribution on

$$\mathcal{M}(\mathbb{R}) := \{F : F \text{ is a probability distribution on } \mathbb{R}\}.$$

Then it becomes *a probability distribution on the set of probability distribution.*

Definition 4.1 (Dirichlet process)**.** *Let α be a finite measure on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is a complete separable metric space, and $\mathcal{B} = \mathcal{B}(\mathcal{X})$ is a Borel σ-field on $\mathcal{X}$. Let P be a random probability measure satisfying*

$$(P(B_1), P(B_2), \cdots, P(B_k)) \sim Dirichlet(\alpha(B_1), \alpha(B_2), \cdots, \alpha(B_k)),$$

for any measurable partition $\{B_i\}_{i=1}^k$ on $\mathcal{X}$, then P is said to be distributed according to the Dirichlet Process, and denoted as

$$P \sim DP(\alpha).$$

Definition 4.2 (Alternative view: from Wikipedia)**.** *Let H be a base probability distribution on $\mathcal{X}$ and $\alpha > 0$. Then the Dirichlet process $DP(H, \alpha)$ is a stochastic process whose sample path is a probability distribution over $\mathcal{X}$ and the following holds: For any measurable finite partition of $\mathcal{X}$, say $\{B_i\}_{i=1}^n$,*

$$X \sim DP(H, \alpha) \Rightarrow (X(B_1), X(B_2), \cdots, X(B_n)) \sim Dirichlet(\alpha H(B_1), \cdots, \alpha H(B_n)).$$

Remark 4.3. Note that, if P is a random probability measure, then all of $P(B)$ becomes random

variable. $P \sim DP$ means that P is a probability measure on $\mathcal{X}$, which is random. "Dirichlet" process is termed because *every finite dimensional distribution* of P,

$$(P(B_1), \cdots, P(B_n)),$$

is Dirichlet distributed. It can be also interpreted as following briefly: We defined a "marginal" (finite-dimensional) distribution of P as a Dirichlet distribution, to define a "joint" feature of P. (Remark that: it is not a formal term!) Because P and partition B_i's should satisfy that

$$0 \leq P(B_i) \leq 1 \text{ and } \sum_{i=1}^{n} P(B_i) = 1,$$

it is natural to think Dirichlet distribution. Note that P is itself a probability measure, or distribution.

There are some important properties of DP. Proof is given in *Advanced Bayesian Analysis* course.

Proposition 4.4 (Conjugacy, or Posterior distribution)**.** *Let $P \sim DP(\alpha)$ and the model be*

$$X_1, \cdots, X_n | P \overset{i.i.d.}{\sim} P.$$

Then posterior distribution of P also becomes DP. In fact,

$$P|X_1, \cdots, X_n \sim DP\left(\alpha + \sum_{j=1}^{n} \delta_{X_i}\right),$$

where δ_c denotes Dirac measure.

Proposition 4.5 (Marginal property, Blackwell & MacQueen (1973))**.** *Let $P \sim DP(\alpha)$ and $X_1, X_2, \cdots | P \overset{i.i.d.}{\sim} P$. Then marginal $(X_1, X_2, \cdots)$ forms **Pólya urn sequence**, i.e.,*

$$\begin{aligned} X_1 &\sim \frac{\alpha}{\alpha(\mathcal{X})} \\ X_{n+1}|X_1, \cdots, X_n &\sim \frac{\alpha + \sum_{i=1}^{n} \delta_{x_i}}{\alpha(\mathcal{X}) + n}. \end{aligned}$$

(See Pólya urn scheme or Chinese restaurant process in section 4.2.)

Proposition 4.6 (Sethuramen representation)**.** *Let*

$$\theta_1, \theta_2, \cdots \overset{i.i.d.}{\sim} Beta(1, \alpha(\mathcal{X}))$$
$$Y_1, Y_2, \cdots \overset{i.i.d.}{\sim} \frac{\alpha}{\alpha(\mathcal{X})}.$$

Consider a ***stick-breaking process***

$$\begin{aligned} p_1 &= \theta_1 \\ p_2 &= \theta_2(1-\theta_1) \\ &\vdots \\ p_n &= \theta_n \prod_{i=1}^{n-1}(1-\theta_i) \\ &\vdots \end{aligned}$$

which makes $\sum_{n=1}^{\infty} p_i = 1$*. Then*

$$P = \sum_{i=1}^{\infty} p_i \delta_{Y_i} \sim DP(\alpha).$$

Remark 4.7. Note that both p_i and Y_i are random. Also, from Sethuramen representation, we can find that if $P \sim DP(\alpha)$, P is a discrete probability measure *w.p. 1.*

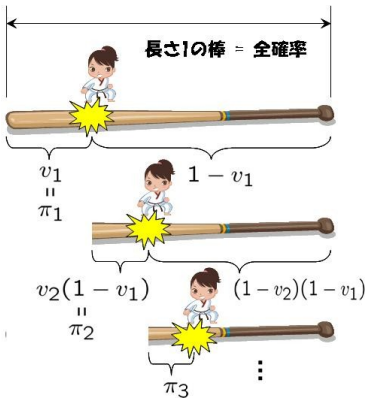

Figure 1: Stick-breaking process. Image from `http://d.hatena.ne.jp/b3s/20081021/1224569652`

Proposition 4.8 (Gamma process representation)**.** *Let* α *be a finite measure on* $[0, \infty)$ *and define a cumulative "measure" function* $A(t) = \alpha[0, t]$*. Also, let* $S(t) \sim GP(A(t), 1)$ *be a Gamma*

process. Then,

$$F(t) = \frac{S(t)}{S(\infty)} \sim DP(\alpha)$$

holds.

Also, following is known about the support of Dirichlet process.

$$supp(DP_\alpha) = \{P : supp(P) \subseteq supp(\alpha)\}.$$

4.2 Description

Contents of this section are mostly quoted from Wikipedia.

Pólya urn scheme

There is a description of Dirichlet process, which is called a Pø'lya urn scheme. Imagine that we start with an urn filled with $\alpha(\mathcal{X})$ black balls. Then we proceed as follows:

1. Each time we need an observation, we draw a ball from the urn.
2. If the ball is *black*, we generate a new (non-black) color uniformly, label a new ball this color, drop the **new ball** into the urn along with the ball we drew, and return the color we generated.
3. Otherwise, (i.e., if the ball is non-black, label a new ball with the color of the ball we drew, drop the **new ball** into the urn along with the ball we drew, and return the color we observed.

This scheme is related to Pólya urn sequence, described in proposition 4.5. Consider following situation. First, draw X_1 from the distribution $\alpha/\alpha(\mathcal{X})$. Next, for $n > 1$, with probability $\frac{\alpha}{\alpha + n - 1}$ draw X_n from $\alpha/\alpha(\mathcal{X})$, which is corresponding to *return the color generated newly*. Or, with probability $\frac{n_x}{\alpha + n - 1}$, set $X_n = x$, where n_x is the number of previous observations $X_j, j < n$ such that $X_j = x$. This procedure is corresponding to *return the color we observed*.

Chinese restaurant process

Similar description for Dirichlet process is the one so-called Chinese restaurant process. Imagine an infinitely large restaurant containing an infinite number of tables, and able to serve an infinite number of dishes. The restaurant in question operates a somewhat unusual seating policy

whereby new diners are seated either at a currently occupied table with probability proportional to the number of guests already seated there, or at an empty table with probability proportional to a constant. Guests who sit at an occupied table must order the same dish as those currently seated, whereas guests allocated a new table are served a new dish at random. The distribution of dishes after n guests are served is a sample drawn as described above.

4.3 Applications

First application we see is an estimation of cdf. Consider the model

$$P \sim DP(\alpha),\ X_1, \cdots, X_n | P \overset{i.i.d.}{\sim} P.$$

Now let the loss function be

$$L(F, G) = \int (F(t) - G(t))^2 dt.$$

We can easily find that Bayes estimator

$$\hat{F}^B(t) = \arg\min_G \mathbb{E}^\pi \left[(F - G)^2(t) | X_1, \cdots, X_n \right]$$

is obtained as

$$\hat{F}^B(t) = \mathbb{E}^\pi \left[F(t) | X_1, \cdots, X_n \right] = \int F(t) \pi(dF | X_1, \cdots, X_n).$$

Note that by definition of DP, for given t,

$$F(t) | X_1, \cdots, X_n \sim Beta(\alpha(t) + nF_n(t), \alpha(\mathbb{R}) - \alpha(t) + n(1 - F_n(t))),$$

because

$$(P((-\infty, t]), P((t, \infty))) \sim Dirichlet(\alpha((-\infty, t]) + \sum \delta_{x_i}(-\infty, t], \alpha((t, \infty)) + \sum \delta_{x_i}(t, \infty)).$$

In here, F_n denotes an empirical cdf. Then denoting $\bar{\alpha}(t) = \alpha(t)/\alpha(\mathbb{R})$, we get

$$\hat{F}^B(t) = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} \bar{\alpha}(t) + \frac{n}{\alpha(\mathbb{R}) + n} F_n(t).$$

Thus, $\alpha(\mathbb{R})$ may be interpreted as "prior size", which gives a "reliability of prior we have."

Next application is Dirichlet Process mixture model. Consider the model

$$\begin{aligned} X_i|\theta_i &\overset{indep}{\sim} f(x_i|\theta_i),\ i = 1, 2, \cdots, n \\ \theta_i|P &\overset{i.i.d.}{\sim} P \\ P &\sim DP. \end{aligned}$$

Then it can be obtained that

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} \sum_{j=1}^{\infty} P_j f(x|\theta_j)$$

where P_j's are from Sethuramen representation

$$P = \sum_{j=1}^{\infty} P_j \delta_{\theta_j} \sim DP.$$

There are some applications of this model. First, we can estimate a density $f(x|\theta)$ of continuous random variable, even though P from DP is discrete. Also, we can fit complex data using simple model f, if we employee mixture model. Furthermore, if we gather observations with the same θ_i, we can find a **clustering** algorithm without pre-determination of the number of cluster.

4.4 Sampling algorithm from the posterior

Escobar (1994) and Escobar & West (1995) suggested MCMC algorithm for Dirichlet Process mixture model. It employees a *collapsed Gibbs sampler*. First, consider the model

$$\begin{aligned} y_i|\theta_i &\overset{indep}{\sim} F(\cdot|\theta_i),\ i = 1, 2, \cdots, n \\ \theta_i|G &\overset{i.i.d.}{\sim} G,\ i = 1, 2, \cdots, n \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where $\alpha > 0$ is a constant and G_0 is a given distribution. Note that joint distribution of y, θ, and G is obtained as

$$p(dy, d\theta, dG) = \prod_{i=1}^{n} F(dy_i|\theta_i) \prod_{i=1}^{n} G(d\theta_i) DP(dG|G_0, \alpha).$$

However, since G is from infinite dimensional space, it is difficult to sample G. Thus, we will integrate out ("collapse") G. Then we get

$$\begin{aligned}
p(dy, d\theta) &= \int_G p(dy, d\theta, dG) \\
&= \int_G prod_{i=1}^n F(dy_i|\theta_i) \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha) \\
&= \prod_{i=1}^n F(dy_i|\theta_i) \underbrace{\int_G \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha)}_{=\text{marginal of } \theta} \\
&= \prod_{i=1}^n F(dy_i|\theta_i) p(d\theta|G_0, \alpha) \\
&= \prod_{i=1}^n f(y_i|\theta_i) dy_i \cdot p(d\theta|G_0, \alpha)
\end{aligned}$$

where density $f(\cdot|\theta_i)$ of $F(\cdot|\theta_i)$ exists. In here, $p(\theta|G_0, \alpha)$ denotes a joint distribution of Pólya sequence. Note that, $G(d\theta_i)$ is in fact $G(d\theta_i|G)$, and hence

$$\int_G \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha)$$

is a marginal distribution of θ_i. Using this we can find a Gibbs sampler algorithm from $[\theta_i|\theta_{-i}, y]$. First note that Pólya sequence has a conditional distribution

$$p(\theta_i|\theta_{-i}, G_0, \alpha) = \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(d\theta_i) + \frac{\alpha}{n-1+\alpha} G_0(d\theta_i).$$

Now by definition of δ_{θ_j} we get

$$\begin{aligned}
p(d\theta_i|\theta_{-i}, y) &\propto f(y_i|\theta_i) p(d\theta_i|\theta_{-i}, G_0, \alpha) \\
&= f(y_i|\theta_i) \left[\frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(d\theta_i) + \frac{\alpha}{n-1+\alpha} G_0(d\theta_i) \right] \\
&\propto \sum_{j \neq i} f(y_i|\theta_j) \delta_{\theta_j}(d\theta_i) + \alpha f(y_i|\theta_i) G_0(d\theta_i) \\
&\propto \sum_{j \neq i} q_{ij} \delta_{\theta_j}(d\theta_i) + r_i H_i(d\theta_i),
\end{aligned}$$

where

$$q_{ij} = \frac{f(y_i|\theta_j)}{\sum_{l \neq i} f(y_i|\theta_l) + \alpha \int f(y_i|\theta) dG_0(\theta)}$$

$$r_i = \frac{\alpha \int f(y_i|\theta) dG_0(\theta)}{\sum_{l \neq i} f(y_i|\theta_l) + \alpha \int f(y_i|\theta) dG_0(\theta)}$$

$$H_i(d\theta_i) = \frac{f(y_i|\theta) dG_0(\theta)}{\int f(y_i|\theta) dG_0(\theta)}.$$

It implies that, to implement this algorithm, we should know the integrated value of $\int f(y_i|\theta) dG_0(\theta)$.

Normalizing constants are determined to make them satisfy

$$\sum_{j \neq i} q_{ij} + r_i = 1.$$

So we obtain the algorithm: iterating sampler from

$$p(d\theta_i|\theta_{-i}, y) = \sum_{j \neq i} q_{ij} \delta_{\theta_j}(d\theta_i) + r_i H_i(d\theta_i).$$

5 Linear Regression I

5.1 Basic model and fitting

The goal of regression model is to find a functional relationship between covariates X and response variable y (inference), or to predict a response (prediction). The model can be represented as

$$y = f(X) + \epsilon, \ \epsilon \sim (0, \sigma^2).$$

Usually, we use a linear regression model,

$$y = X\beta + \epsilon.$$

First consider a simple linear model,

$$y = \beta_0 + \beta_x + \epsilon.$$

We often employee least square procedure to find estimates of coefficients,

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

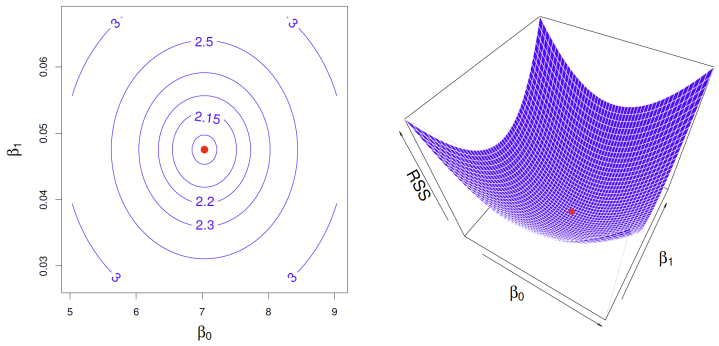

Figure 2: Square function: image from ISL

It is well known that the minimization solution is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

How this estimator is "reliable?" We can see "unbiasedness" and "standard error." Note that

$$E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$$

and

$$se(\hat{\beta}_0) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]},\ se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

Also note that we can estimate σ as

$$\hat{\sigma} = \sqrt{\frac{SS_E}{n-2}}.$$

In here, SS_E is the sum of squared residuals. With these, we can get an inference on $\hat{\beta}$. First, note that $100(1-\alpha)\%$ confidence interval is

$$\hat{\beta}_1 \pm z_{\alpha/2} se(\hat{\beta}_1) \text{ and } \hat{\beta}_0 \pm z_{\alpha/2} se(\hat{\beta}_0),$$

respectively. Next, hypothesis testing for

$$H_0 : \beta_1 = 0 \ vs\ H_1 : \beta_1 \neq 0$$

is possible with the test statistics

$$T = \frac{\hat{\beta}}{se(\hat{\beta}_1)} \underset{H_0}{\sim} t(n-2).$$

If there are several covariates, we can fit the model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon.$$

Also in this case we use a least square estimator

$$\hat{\beta} = \arg\min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

In the same way, we can perform an inference on the coefficient. Note that, coefficient β_j is interpreted as "the average effect on Y of a one unit increase in x_j, *holding all other predictors fixed.*"

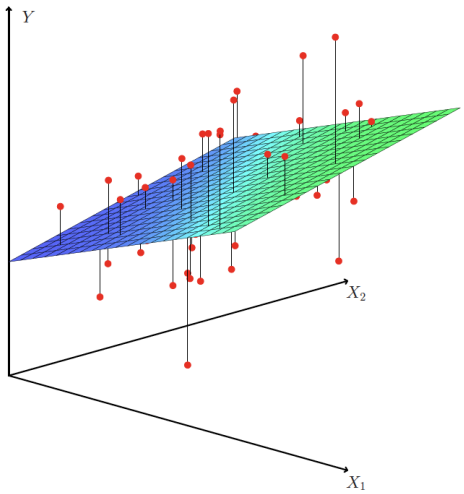

Figure 3: Square function: image from ISL

Example 5.1. Consider the advertisement data set. Our interest is to find the relationship between sales and advertisement budget. In this case, our model is

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Figure 4 shows some results of fitting simple and multiple regression models.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Simple regression of sales on radio

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Simple regression of sales on newspaper

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

Figure 4: Advertisement example: image from ISL.

We interpret these results as follows: for a given amount of TV and newspaper advertising, spending an additional \$1,000 on radio advertising leads to an increase in sales by approximately 189 units. However, while the `newspaper` regression coefficient estimate in figure 4 was significantly non-zero, the coefficient estimate for `newspaper` in the multiple regression model is close to zero, and the corresponding p-value is no longer significant. Does it make sense for the multiple regression to suggest no relationship between `sales` and `newspaper` while the simple linear regression implies the opposite? Note that the correlation between `radio` and `newspaper`

is 0.35. This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising. Now suppose that the multiple regression is correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales. Then in markets where we spend more on radio our sales will tend to be higher, and we also tend to spend more on newspaper advertising in those same markets. Hence, in a simple linear regression which only examines `sales` versus `newspaper`, we will observe that higher values of `newspaper` tend to be associated with higher values of `sales`, even if newspaper advertising does not actually affect sales. So `newspaper` sales are a surrogate for `radio` advertising; `newspaper` gets "credit" for the effect of `radio` on `sales`. (*referred from ISL, p.73-p.74*)

5.2 Variable selection

In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e., whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. We test the null

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative $H_1 : else$, with the F-statistic

$$F = \frac{(SS_T - SS_E)/p}{SS_E/(n-p-1)} \underset{H_0}{\sim} F(p, n-p-1).$$

It is related to R^2, which is defined as

$$R^2 = \frac{SS_R}{SS_T} = \max_a \widehat{Corr}(Y, a^\top X) = \widehat{Corr}(Y, \hat{Y}).$$

Sometimes, we want to test a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

In this case, we fit a *reduced model* that uses all the variables except those last q. Then the appropriate F-statistic is

$$F = \frac{(SS_E(RM) - SS_E(FM))/q}{SS_E/(n-p-1)} \underset{H_0}{\sim} F(q, n-p-1).$$

These provide information about whether each individual predictor is related to the response, after adjusting for the other predictors. In other sords, it reports the *partial effect* of adding

that variable to the model.

If we conclude that at least one of the predictors is related to the response, then it is natural to wonder *which* are the significant ones. The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as *variable selection*. Ideally, we would like to compare *all possible models* with various criteria such as Mallows' C_p, AIC, BIC, adjusted R^2, etc. Unfortunately, there are a total of 2^p models that contain subsets of p variables, so trying out every possible subset of the predictors is infeasible. Therefore, we need an automated an efficient approach to choose a smaller set of models to consider. There are 3 classical approaches for this task: forward selection, backward selection, and mixed (or stepwise) selection.

5.3 Dummy variable

If one wants to use qualitative variables as predictors, then one can use a dummy variable (indicator variable). If a qualitative predictor (factor) only has two levels, then we simply create an indicator that takes on two possible numerical values, for instance,

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

and use this variable as a predictor in the regression equation. This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

In this case, β_0 can be interpreted as the average of y_i among males, while β_1 as the average difference between females and males. When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values, so we can create additional dummy variables. For example, let

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Now β_0 can be interpreted as the average of y_i for African Americans, β_1 as the difference in y_i between the Asian and African American categories, and β_2 as the difference in y_i between the Caucasian and African American categories.

There are many different ways of coding qualitative variables besides the dummy variable approach taken here. All of these approaches lead to equivalent model fits, but the coefficients are different and have different interpretations, and are designed to measure particular *contrasts*.

5.4 Beyond additivity and linearity

The standard linear regression model provides interpretable results and works quite well on many real-world problems, but it makes several highly restrictive assumptions that are often violated in practice. Two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*. The additive assumption means that *the effect of changes in a predictor x_j on the response y is independent of the values of the other predictors.* The linear assumption states that *the change in the response y due to a one-unit change in x_j is constant, regardless of the value of x_j.*

Removing the additive assumption

Back to the advertisement example. Recall that we fitted the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

However, this model may be incorrect. Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases. It is referred to as an *interaction* effect.

Consider the standard linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Note that regardless of the value of x_2, a one-unit increase in x_1 will lead to a β_1-unit increase in y. To extend this model, we include an "interaction term,"

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Then we obtain

$$Y = \beta_0 + (\beta_1 + \beta_3 x_2) x_! + \beta_2 x_2 + \epsilon,$$

which implies that adjusting x_2 will change the impact of x_1 on y.

it is sometimes the case that an interaction term has a very small p-value, but the associated main effects do not. The *hierarchical principle* states that, if we include an interaction in a model, **we should also include the main effects, even if the p-values associated with their coefficients are not significant.** The rationale for this principle is that if $x_1 x_2$ is related to the response, then whether or not the coefficients of x_1 or x_2 are exactly zero is of little interest. Also, $x_1 x_2$ is typically correlated with x_1 and x_2, and so leaving them out tends to alter the meaning of the interaction.

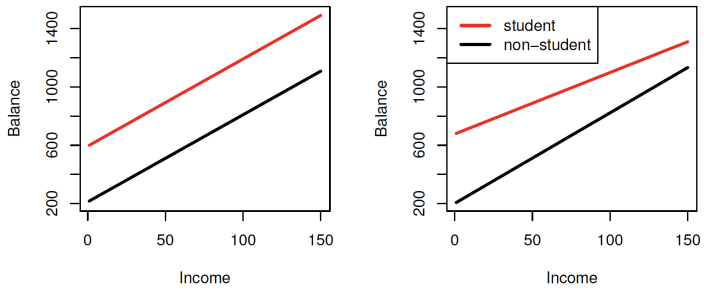

Figure 5: Interaction Effect: image from ISL.

Nonlinear relationships

In some cases, the true relationship between the response and the predictors may be nonlinear. Here we present a very simple way to directly extend the linear model to accommodate nonlinear relationships, using *polynomial regresssion*. For example, we can fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

to find a quadratic relationship between two variables. More advanced methods for nonlinear regression will be handled later.

5.5 Autocorrelation, heteroscedasticity, and outlier detection

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are following:

(a) Nonlinearity of the response-predictor relationships.

(b) Correlation of error terms.

(c) Non-constant variance of error terms.

(d) Outliers.

(e) High-leverage points.

(f) Multicollinearity.

Non-linearity of the data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. *Residual plots* are a useful graphical tool for identifying nonlinearity. We can plot the residuals e_i versus the predictor x_i, or the predicted values $\hat{y}_i$, in the case of a multiple regression model. If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use nonlinear transformation of the predictors, such as $\log x$, $\sqrt{x}$, and x^2.

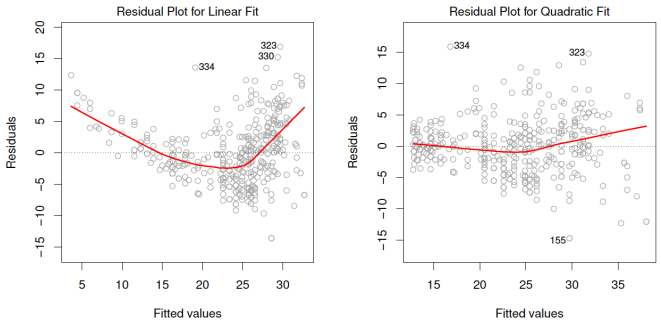

Figure 6: Residual plots of linear and quadratic model, respectively. Image from ISL.

Dependent structure of error terms

An important assumption of the linear regression model is that the error terms are uncorrelated. It means that, the fact that ϵ_i is positive provides little information about the sign of ϵ_{i+1}. The standard errors (se) are based on the assumption of uncorrelated error terms. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true se. As a result, confidence (or prediction) intervals will be narrower then they should be! in addition, p-values will be lower than they should be. In short, if the error terms are associated, *we may have an unwarranted sense of confidence in our model*. In other words, we cannot believe our inference.

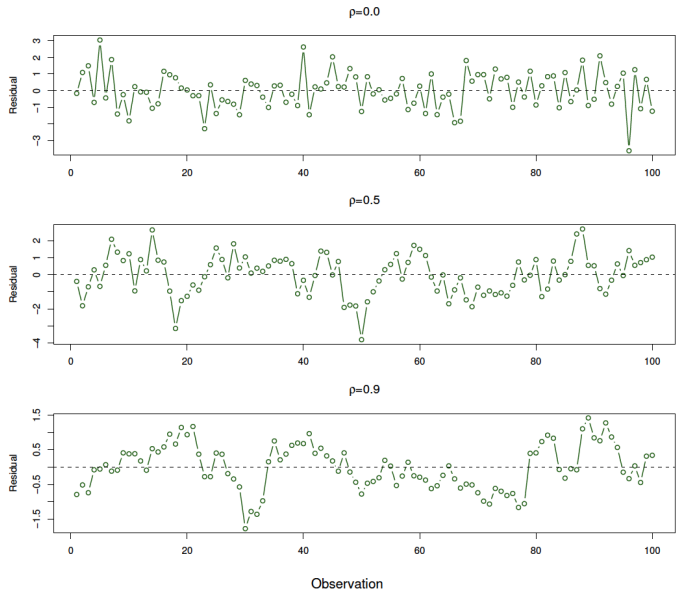

Figure 7: Autocorrelated error terms. Image from ISL.

Heteroscedasticity

Another important assumption of the linear regression model is that the error terms have a constant variance. Unfortunately, it is often the case that the variances of the error terms are non-constant. One can identify non-constant variances in the errors, or *heteroscedasticity*, from the presence of a *funnel shape* in the residual plot. When faced with this problem, one possible solution is to transform the response y using a concave function such as $\log y$ or $\sqrt{y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. Or, we have specific information of the variance of each response, then we can implement *weighted least squares*.

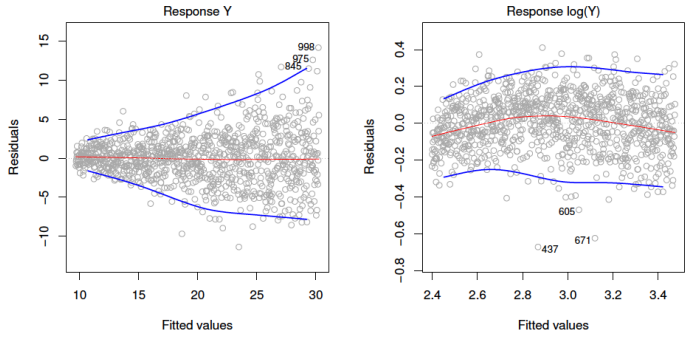

Figure 8: Heteroscedastic residuals and those of transformed responses. Image from ISL.

Outliers and Leverage points

An *outlier* is a point for which y_i is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection. Residual plots can be used to identify outliers. But in practice, it can be difficult to decide *how large a residual needs to be before we consider the point to be an outlier*. To address this problem, we can plot the *studentized residuals*. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers. Also, one solution to handle outlier is to simply remove it, but care should be taken, since an outlier may include an important information, for example, it can instead indicate a deficiency with the model, such as a missing predictor.

In contrast, observations with *high leverage* have an unusual value for x_i. High leverage observations tend to have a sizable impact on the estimated regression line. It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problems with these points may invalidate the entire fit. For this reason, it is important to identify high leverage observations. In order to quantify an observations's leverage, we compute the *leverage statistic*, h_{ii}. h_{ii} is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$.

Multicollinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes $se(\hat{\beta}_j)$ to grow. Consequently, it results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that *the power of the test is reduced* by collinearity. Further, since we cannot believe $\hat{\beta}$, confidence intervals become wider.

A simple way to detect collinearity is to look at the correlation matrix of the predictors, but it cannot detect collinearity between three or more variables, which is called *multicollinearity*. A better way to assess multicollinearity is to compute the *variance inflation factor (VIF)*. it is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own. The smallest possible value for VIF is 1. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. The VIF for each variable can be computed using the formula

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

When faced with the problem of multicollinearity, there are two simple solutions. The first is to drop one of the problematic variables from the regression. This can be usually be done without much compromise to the regression fit, since the presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables. However, dropping the variable always should be done with care. Second one is to combine the collinear variables together into a single predictor. But it can make interpretation of the fitted model harder.

6 Classification

In many situations, the response variable is not quantitative but qualitative. In this section, we study approaches for predicting qualitative responses, that is known as *classification*. We often use training error rate

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i),$$

or test error rate

$$Ave(I(y_i \neq \hat{y}_i))$$

to measure the performance of classifier.

6.1 Bayes Classifier

Let C_i be the ith group (population), and let $Y_i = j$ if $X_i \in C_j$. Then, Bayes classifier decides a new observation to classify

$$\hat{y}_0 = \arg\max_j P(Y = j|X = x_0).$$

The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate.In this case, the error rate at $X = x_0$ is

$$1 - \max_j P(Y = j|X = x_0),$$

so in general, the overall Bayes error rate is given by

$$1 - E\left(\max_j P(Y = j|X)\right),$$

where the expectation averages the probability over all possible values of X.

6.2 k-nearest neighborhood

In theory we would always like to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X. Therefore, many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the k-nearest neighbors classifier. Given a positive integer k and a test observation x_0, the k-NN

classifier first identifies the k points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

$$\hat{P}(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j).$$

Finally, k-NN classifies the test observation x_0 to the class with the largest probability.

6.3 Logistic regression and GLM

Logistic Regression

Let's denote the groups as 0 and 1. Logistic regression models the probability that Y belongs to a particular category, 0 or 1. In logistic regression, we use the logistic function,

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

to model $p(X) = P(Y = 1|X)$. To fit this model, we use a maximum likelihood estimation. This model is equivalent to

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_X},$$

where left hand side $p(X)/(1 - p(X))$ is called the *odds*, or

$$\log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X,$$

where left hand side is *log odds* or *logit*. In this case, increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}. However, because the relationship between $p(X)$ and X is not a straight line, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in X. But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing $p(X)$.

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j)).$$

To find maximization solution, we use various methods, such as *Newton-Raphson method,* or *Iteratively Re-weighted Least Squares (IRLS)*. Detail procedure is beyond the scope.

Once the coefficients have been estimated, it is a simple matter to compute the probability of

Y for any X. Also, one can use qualitative predictors with the logistic regression model using the dummy variable.

By analogy with the extension from simple to multiple linear regression, we can generalize the model as

$$\log p(X)1 - p(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X^\top \beta,$$

where $X = (X_1, \cdots, X_p)^\top$ are p predictors. It can be rewritten as

$$p(X) = \frac{e^{X^\top \beta}}{1 + e^{X^\top \beta}}.$$

We sometimes wish to classify a response variable that has more than two classes. The two-class logistic regression models discussed in above have multiple-class extensions, but in practice they tend not to be used all that often. Rather, *discriminant analysis* is popular for multiple-class classification.

Brief introduction to GLM

Let the response y_i be distributed with following density:

$$f(y_i : \theta_i, \phi) = e^{A_i(y_i\theta_i - \gamma(\theta_i))/\phi + \tau(y_i, \phi/A_i)}.$$

In here, x_i is a predictor, and $\eta_i = x_i^\top \beta$ is a linear predictor. Also $\theta_i = \theta(\eta_i)$. A_i's are weight, where ϕ is a scale. l is called a *link function* where $\eta = l(\mu)$, $\mu = Ey$.

Example 6.1 (Normal model)**.** Let $y_i \sim N(x_i^\top \beta, \sigma^2)$. Then link function is $l(\mu_i) = \mu_i$, and from

$$f(y_i : \mu_i, \sigma^2) = \exp\left(\frac{1}{\sigma^2}\left(y_i\mu_i - \frac{\mu_i^2}{2}\right)\frac{y_i^2}{2\sigma^2} \log\sqrt{2\pi}\sigma\right),$$

we get $\theta_i = \mu_i$, $\gamma(\theta_i) = \mu_i^2/2$.

Example 6.2 (Binomial model)**.** Let

$$y_i \sim Bin\left(n_i, p_i = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}\right).$$

Then density of y_i is

$$f(y_i : n_i, p_i) = \exp\left(n_i\left(\frac{y_i}{n_i}\log\frac{p_i}{1-p_i} + \log(1-p_i)\right) + \log\binom{n_i}{y_i}\right),$$

so if we see this as a density function of y_i/n_i, it forms GLM, and $A_i = n_i,\ \mu_i = \log p_i/(1-p_i), \phi = 1,\ \gamma(\theta_i) = -\log(1-p_i),$ and $\tau(y_i, \phi/n_i) = \log \binom{n_i}{y_i}$. Also, link function is a logit.

If there is no restriction on $\theta_1, \cdots, \theta_n$, i.e., there are as many estimated parameters as data points, then it is called a *saturated model*. Also, *deviance* of model M is defined as

$$\begin{aligned} D_M &= 2\phi \left(\log f(y_i : \hat{\theta}_i^S, \hat{\phi}^S) - \log f(y_i : \hat{\theta}_i^M, \hat{\phi}^M) \right) \\ &= 2 \sum_{i=1}^{n} A_i \left[(y_i \hat{\theta}_i^S - \gamma(\hat{\theta}_i^S)) - (y_i \hat{\theta}_i^M - \gamma(\hat{\theta}_i^M)) \right]. \end{aligned}$$

For inference, followings are known: Scaled deviance is chi-squared distributed, i.e.,

$$\frac{D_M}{\phi} \sim \chi^2(n-p),$$

and if other model M_0 is contained in M($M_0 \subseteq M$), and the numbers of coefficients of models are $q < p$ respectively, then

$$\frac{D_{M_0} - D_M}{\phi} \underset{M_0}{\approx} \chi^2_{p-q}.$$

6.4 LDA and QDA

Logistic regression involves directly modeling $P(Y = k|X = x)$ using the logistic function. Like this, in statistics, we model the conditional distribution of the response Y, given the predictors X. We now consider an alternative and less direct approach to estimating these probabilities. In this alternative approach, we model the distribution of the predictors X separately in each of the response classes (i.e., given Y), and then use Bayes' theorem to flip these around into estimates for $P(Y = k|X = x)$. Assume that there are $K \geq 2$ classes, and let $f_k(x) = P(X = x|Y = k)$ be the density function of X for an observation that comes from the kth class. So $f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$, and $f_k(x)$ is small if opposite. Then by Bayes theorem,

$$p_k(x) := P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)},$$

where π_i's denote prior probabilities. We refer to $p_k(x)$ as the posterior probability.

Linear Discriminant Analysis

Assume that $p = 1$, that is, we have only one predictor. We would like to obtain an estimate for $f_k(x) = P(X = x|Y = k)$ in order to estimate $p_k(x)$. In LDA, we assume that $f_k(x)$ is Gaussian. In this case,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

where μ_k and σ_k^2 are the mean and variance parameters for the kth class. For now, let us further assume that $\sigma_1^2 = \cdots = \sigma_K^2$. Then we find that

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{i=1}^K \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)}.$$

It is equivalent to assigning the observation to the class for which

$$\delta_k(x) = \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

is largest.

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, \cdots, \mu_K, \pi_1, \cdots, \pi_K$ and σ^2. The *linear discriminant analysis (LDA)* method approximates the Bayes classifier by pugging estimates for π_k, μ_k, and σ^2. In particular,

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \ \hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

are used, with training data set. In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class, i.e.,

$$\hat{\pi}_k = \frac{n_k}{n}.$$

Extension to the multivariate case is straightforward. Gaussian density becomes

$$f_k(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)\right),$$

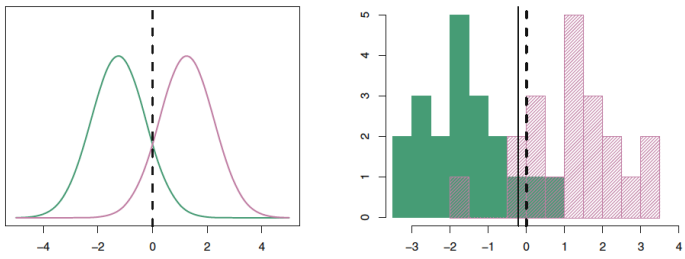

Figure 9: LDA in univariate case: image from ISL.

and therefore discriminant function becomes

$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k.$$

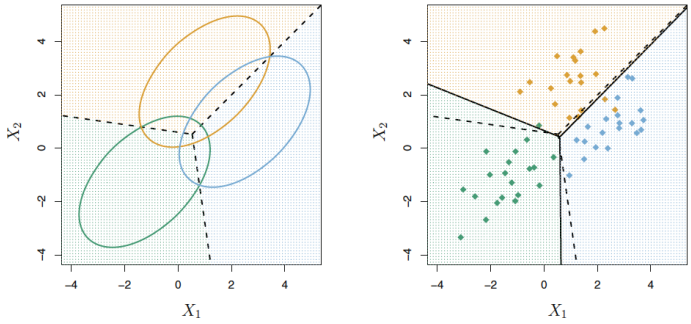

Figure 10: LDA in univariate case: image from ISL.

Quadratic Discriminant Analysis

As we have discussed, LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes. *Quadratic discriminant analysis (QDA)* provides an alternative approach. Like LDA, QDA assumes Gaussian population, but it assumes that each class has its own covariance matrix, which is the difference between LDA. That is, it assumes that

$$f_k(x) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\right),$$

which yields the discriminant function

$$\delta_k(x) = -\frac{1}{2}x^\top \Sigma_k^{-1} x + x^\top \Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log \pi_k.$$

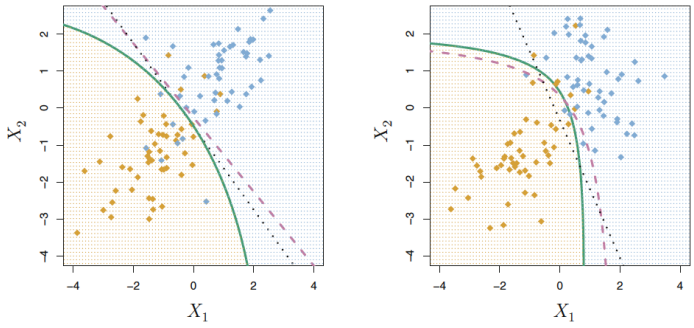

Figure 11: LDA, QDA, and Bayes classifie. Image from ISL.

See figure 11. Left one shows classification result with Bayes (purple dashed), LDA (black dotted), and QDA (green solid) for a two-class problem with $\Sigma_1 = \Sigma_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA. Right one shows classification result for the case $\Sigma_1 \neq \Sigma_2$, and details are the same as left one. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.

6.5 Error rates and Positive rates

Example 6.3. Consider the `Default` data set. we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income (`income`) and monthly credit card balance (`balance`). A *confusion matrix* compares the LDA predictions to the true default statuses.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Table 1: Confusion matrix

Table 1 reveals that LDA predicted that a total of 104 people would default. Of these people, 81 actually defaulted and 23 did not. Hence only 23 out of 9,667 of the individuals who did not default were incorrectly labeled. This looks like a pretty low error rate, but of the 333 individuals who defaulted, 252 were missed by LDA. So while the overall error rate is low, the error rate among individuals who defaulted is very high. It happens because *LDA is trying to approximate the Bayes classifier, which has the lowest total error rate* out of all classifiers. In contrast, a credit card company might particularly wish to avoid incorrectly classifying an individual who

will default, whereas incorrectly classifying an individual who will not default, though still to be avoided, is less problematic. It can be solved by *lowering the threshold*. In two-classes case, the Bayes classifier amount to assigning an observation to the *default* class if

$$P(\texttt{default} = \texttt{Yes}|X = x) > 0.5.$$

That is, Bayes classifier, and hence LDA, uses a threshold of 50% for the posterior probability of default in order to assign an observation to the *default* class. However, if we are concerned about incorrectly predicting the default status for individuals who default, then we can consider lowering this threshold, for example, we might label any customer with a posterior probability of default above 20% to the *default* class. In other words, we could instead assign an observation to the class if

$$P(\texttt{default} = \texttt{Yes}|X = x) > 0.2.$$

There are 4 possible cases: see table 2.

		Predicted class		
		- or Null	+ or Non-null	Total
True class	- or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Table 2: Possible results when applying a classifier.

We also use terms *sensitivity* and *specificity*. See table 3 for summary.

Name	Definition	Synonyms
False Pos. rate (FPR)	FP/N	Type I error, 1-Specificity
True Pos. rate (TPR)	TP/P	1-Type II error, Power, Sensitivity
Pos. Pred. value	TP/P*	Precision, 1-false discovery rate (FDR)
Neg. Pred. value	TN/N*	

Table 3: Important measures for classification and diagnostic testing

The *ROC curve* is a popular graphic for simultaneously displaying the two types of errors, FPR and TPR, or equivalently, (1-specificity) and sensitivity. The name "ROC" is an acronym for *receiver operating characteristics*. The ideal ROC curve hugs the top left corner, indicating a high TPR and a low FPR. The overall performance of a classifier, summarized over all possible thresdholds, is given by the *area under the curve (AUC)*. Considering the ideal ROC curve, higher AUC indicates that classifies perform better; in the ideal case, AUC becomes 1. Figure 12 is one example of ROC curve. In this figure, the dotted line represents the "no information"

classifier.

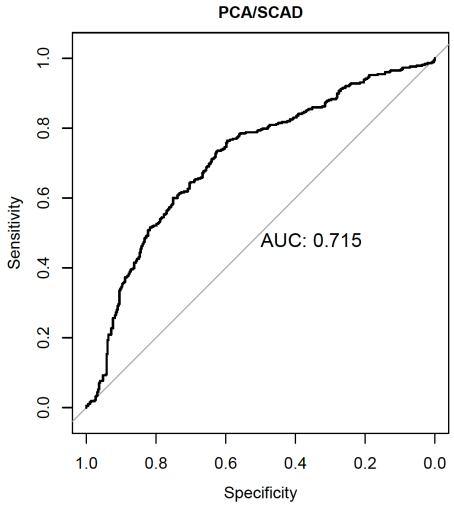

Figure 12: One example of ROC curve.