ISLR book notes: Chapter 3

Linear regression: $Y \approx \beta_0 + \beta_1 X + \epsilon$ w/ $\text{Var}(\epsilon) = \sigma^2$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- estimate β_0 and β_1 w/ least squares: $e_i = y_i - \hat{y}_i$

  min $\sum e_i^2$ → $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$

  - unbiased estimators

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

- you estimate the mean of Y as $\hat{\mu} = \bar{y}$

  that has $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$

$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$ $\quad \text{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2}$

↳ variance is smaller if the x_i's are more spread out (easier to fit a line on a larger section)

- estimate for σ: Residual Standard Error RSE

$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$

- If $\epsilon = N(0, \sigma^2)$, then the 95% confidence interval is

$[\hat{\beta}_1 - 2\,SE(\hat{\beta}_1), \hat{\beta}_1 + 2\,SE(\hat{\beta}_1)]$ for β_1

and $[\hat{\beta}_0 - 2\,SE(\hat{\beta}_0), \hat{\beta}_0 + 2\,SE(\hat{\beta}_0)]$ for β_0

- Hypothesis testing: H_0: no relationship b/w X and Y $\quad \beta_1 = 0$

  H_a: some relationship $\quad \beta_1 \neq 0$

  - we want to know how far our estimate $\hat{\beta}_1$ is away from 0

    How far is enough depends on the standard error

→ t statistic: $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ $\quad t \sim T_{n-2}$ distribution $\approx N$ is $n > 30$

(if $|t|$ is big, we have a steeper regression line and more correlation)

→ p value: probability of observing $|t|$ as big if $\beta_1 = 0$

small p value: unlikely that $\beta_1 = 0$

reject null hypothesis

so they are correlated

Assessing how good the regression is:

- RSE (residual sum of squares $\sum (y_i - \hat{y}_i)^2$)

Residual standard error:

≈ average amount the response deviates from the true line

$$RSE = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2} RSS}$$

- measures lack of fit in the context of the data (depends what is an acceptable RSE)

- R^2 is similar to RSE but scaled to the data

so is universal

$RSS = \sum (y_i - \hat{y}_i)^2$

$TSS = \sum (y_i - \bar{y}_i)^2$

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

↳ total sum of squares

- TSS measures the inherent variance in Y regardless of the regression
- RSS the variability unexplained after the regression

$0 \leq R^2 \leq 1$

↓ not good ↓ good

- model not good
- σ^2 is high

- depends what a good R^2-value is

- in simple linear regression, $R^2 = Cor(X,Y)^2$, where

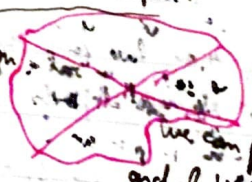
$$Cor(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Multiple regression: $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$

Q1 Is at least one of the predictors X_i useful?

H_0: $\beta_i = 0$ $\forall i$, no correlation

F-statistic: $\frac{(TSS - RSS)/p}{RSS/(n-p-1)}$



we can prove that $E\left(\frac{RSS}{n-p-1}\right) = \sigma^2$

and if H_0 is true, $E\left(\frac{TSS-RSS}{p}\right) = \sigma^2$

If H_0 is true, F-statistic is close to 1 if H_0 not true, $E\left(\frac{TSS-RSS}{p}\right) > \sigma^2$

If H_0 is not true, F-statistic is > 1

how large it needs to be depends on n, p

F-statistic $\sim F_{n,p}$-distribution, so we can calculate a p-value

- If we only want to test H_0 that only some $\beta_1, \dots \beta_q$ are 0,
  we fit another model that doesn't use them and calculate RSS_0 (RSS under hypothesis H_0)
  then we use the F-statistic $$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$
  - If $q=1$, this F-statistic is the same as the (t-statistic)^2 for that variable in the full model
  - So this reports the partial effect of adding that variable to the model

caveat:
- (If $q = p$, we get $RSS_0 = TSS$ as $\hat{y} = \hat{x}_0 = \bar{y}$)

- Why is F-statistic good:
  - If you get many predictors that are not correlated w/ y, by random chance may will have small p-values for their t-scores
  - F-statistic adjusts to number of predictors

- This only works if $p < n$, otherwise can't even fit a line

Q2 Which are the useful variables?

- Various methods to select best model: Mallow's C_p
  AIC, BIC
  adjusted R^2

How to select params to test?
- Forward: start w/ null, then add param that minimizes RSS for continue until stopping rule (greedy)
- Backward: start w/ all, remove that w/ largest p-value
  continue until stopping rule
- Mixed: start w/ null, add until one of the p-values go up (forward)
  remove largest p-value (backward)
  continue until all p value low enough and add new variables would raise them too

Q3 Model fits the data how well?

→ R^2 close to 1. model explains a large portion of the variance

$R^2 = Cor(Y, \hat{Y})^2$ in multilinear regression,

- adding more variables always increases R^2 towards 1, as we can fit better w/ more variables
- this can result in overfitting

→ RSE can increase because of the $n-p-1$ in the denominator when we add a new param

→ plot the data

you can for example spot non-linearity b/w two variables

Q4 How accurate predictions can you make?

→ we can compute a confidence interval for our reducible error

how close $\hat{Y}$ is to $f(X) = \beta_0 + \beta_1 X_1 + \beta_p X_p$, our best possible linear model

→ there is always some model bias ~~~~

e.g linearity approximates reality

→ ~~~~ error because of ε is irreducible error

we can calculate prediction intervals that also incorporate ε

- use confidence interval to qualify the uncertainty of average predictions
- use prediction interval for a particular prediction

What to do with qualitative variable?

→ use dummy variables e.g if Race has levels Asian, Caucasian and Afro-American,

do $x_{i1} = \begin{cases} 1 & \text{asian} \\ 0 & \text{not} \end{cases}$ $x_{i2} = \begin{cases} 1 & \text{caucasian} \\ 0 & \text{not} \end{cases}$

w/ model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

and interpret β_0 as avg for Afro-Americans, β_1 is diff b/w Asians and Afro-Am, and β_2 as diff b/w Caucasians and Afro-Am.

One fewer dummy variable than levels. The level w/o it is the baseline

Extension to linear model

additive: if only x_j changes, the response of Y is indep of the values of other x_i

linear: if x_j changes one unit, the change it induces in Y is constant regardless of the value of x_j

removing additive assumption:

- there may be synergy b/w variables, an interaction

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

↳ interaction term.

Hierarchical principle:

If we add an interaction term, we should also include the main effect terms even if their p-values are not significant

removing linear assumption:

polynomial regression - explored later

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

Problems:

1. Non-linearity of data

plot residual plots: $y_i - \hat{y}_i$ vs x_i for single lin regression

$y_i - \hat{y}_i$ vs $\hat{y}_i$ for multilinear

- there should be no pattern identifiable in the plot
- if there is, data might not be linear, try non-linear transformations

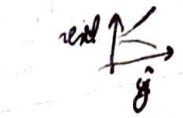
2. Correlation of error terms

- if ε_i are correlated, our model will do worse
- often a problem for time series
- plot residuals for time series, if no pattern then error terms are not correlated

Problem 3 Non-constant variance of error terms - heteroscedasticity

if $Var(\varepsilon_i) = \sigma^2$ is not true $\forall i$

- sometimes bigger y has bigger error
  can see this from funnel shape on residual plot



possible solution is to log y or $\sqrt{y}$

Problem 4 Outliers

outlier don't usually affect the regression a lot
help to see them if we plot residuals vs fitted values

to see what is an outlier, we calculate the standardized residuals

- which is $e_i / SE(e_i)$ → *
- if this is >3, then it is an outlier

$SE(e_i)$:

$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X}_{H} y$

$e = y - \hat{y} = (I - H)y$

$Var(e) = \sigma(I - H)$

$SE(e_i) = \sigma\sqrt{1 - h_{ii}}$

Problem 5 High leverage points

- high leverage points affect the regression a lot
- unusual x values
- difficult to notice a high leverage

h_i = leverage statistic quantifies this → it is the same h_{ii} as here

for simple linear regression, $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$

$\frac{1}{n} < h_i < 1$

$avg(h_i)$ is $\left(\frac{p+1}{n}\right)$ for all observations

so if $h_i >> \frac{p+1}{n}$, that point x_i has a high leverage

Problem 6 Collinearity

Predictors are closely related to one another

reduces the accuracy of the predictions

which increases the errors and makes the p-values worse

and we might even fail to reject H_0 b/c of it

looking at correlation matrix doesn't help if 3 or more are correlated

we detect collinearity w/ the Variance inflation factor $= \frac{Var(\hat{\beta}_j) \text{ when full model}}{Var(\hat{\beta}_j) \text{ when only } x_j \text{ fit}}$

no collinearity $\leftarrow 1 \leq VIF < \infty$

if VIF > 5 ~ 10, problems of collinearity

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

where $R^2_{X_j|X_{-j}}$ is from regression of X_j onto all other predictors

(you fit all other preds to determine X_j, and you take R^2 from it)

if $R^2_{X_j|X_{-j}} \approx 1$, collinearity, ~~then~~

solutions: a) drop one of the predictors

b) combine them into a single value e.g. by taking their mean

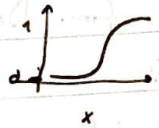
~~Exercises~~

Classification → want to predict qualitative variable (0 or 1)

trying to assess the probability of a ~~0~~ outcome

logistic regression:

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



$p(X)$ = probability of $Y=1$ given X

odds: $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X}$ e.g. if $p(X) = 0.2$ → 1 out of 5 → 1/4 odds

log odds logit: $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$

fitting it: we maximise $\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$ maximise the likelihood fn

assessing if variables are related:

z-score z-statistic for $\hat{\beta}_1$: $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$

large |z-statistic|: evidence against $H_0: \beta_1 = 0$

also you can calculate p-value from z-statistic (similar to t-statistic for linear regression)

Multiple logistic regression:

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p$$

Linear Discriminant analysis

Logistic regression is not always good enough!

- when classes are well separated, f. unstable, LDA is better
- when n is small and predictors X ~ normal, LDA is better
- when more than 2 response classes

π_k - prob. that y is category k → prior prob that an observation belongs to k

$f_k(x) = Pr(X=x \mid Y=k)$

Then $p_k(x) = Pr(Y=k \mid X=x) = \dfrac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$ from Bayes

$p_k(x)$ → posterior prob. that $X=x$ belongs to k

we know π_k from our sample (approximately)

→ we only need to estimate $f_k(x)$ $Pr(X=x \mid Y=k)$

p=1 only one predictor

- assume $f_k(x)$ ~ normal
- assume $\sigma_1^2 = \sigma_2^2 \dots = \sigma_K^2$

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

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

→ $p_k(x)$ is maximal when

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

LDA estimates μ_k, π_k and σ^2 to help maximizing $p_k(x)$:

$$\hat{\mu}_k = \frac{1}{n} \sum_{i: y_i = k} x_i$$

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

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

so we get $X=x$ assignment to the class k.

(σ not σ) $\rightarrow$ $\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$ is the largest

$\hookrightarrow$ $\hat{\delta}_k$ discriminant function

linear in x

$p > 1$:

- Assume $(X_1, \dots X_p)$ drawn from a multi-variate Gaussian

$\sim N(\mu_k, \Sigma)$ for kth class

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

$\vdots$ algebra

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad \rightarrow \text{ linear in } x$$

$P_k(x)$ is maximal when $\delta_k(x)$ is maximal

so decision boundaries are when $\delta_k(x) = \delta_\ell(x)$ $\rightarrow$ Bayes decision boundary "the theoretical best"

① estimate $\hat{\mu}_1, \dots \hat{\mu}_K$ with same formulas

$\hat{\pi}_1, \dots \hat{\pi}_K$

Σ

② plug in to $\hat{\delta}_1(x), \dots \hat{\delta}_K(x)$ $\rightarrow$ $\hat{\delta}_k(x) = \hat{\delta}_\ell(x)$ $\rightarrow$ LDA decision boundary trying to approximate Bayes boundary

③ classify x s.t. $\delta_k(x)$ is the highest

assign $Y=k$ if $P(Y=k \mid X=x) \geq P(Y=i \mid X=x) \; \forall i$

LDA approximates the Bayes classifier, eg. trying to minimise the misclassification error

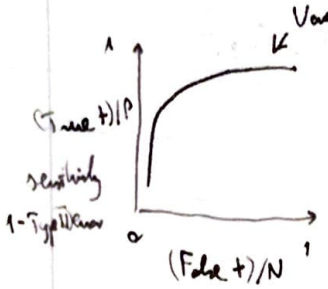
we might want to do something else, for example minimise the false positives, or the false negatives.

$\rightarrow$ If $K=2$, Bayes classifier picks a threshold 0.5 for $P_r(Y=0 \mid X=x)$ assigned: $> 0.5 \Rightarrow y=0$, $< 0.5 \Rightarrow y=1$

we can make this smaller for less false negatives

= more true positives

Roc curve:



Varying the threshold for classification (for $p(k)$)

Area under the curve measures the quality of the classifier

perfect: 1

random: 0.5

what threshold you want to pick depends on what you need the classifier for

Quadratic discriminant analysis:

- same as LDA, but assumes that $X \sim N(\mu_k, \Sigma_k)$ → so each class has unique covariance matrix

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

quadratic Bayes and QDA decision boundary

→ $= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$

- we have formulas for $\hat{\mu}_k, \hat{\pi}_k$ and $\hat{\Sigma}_k$

why use QDA and LDA? when is one better?

QDA estimates $K \cdot \frac{p(p+1)}{2}$ params for $\Sigma_1, \dots, \Sigma_K$ → quadratic in num of preds

for LDA, the model is linear in x → Kp linear coeff to estimate

→ LDA has less params → less flexible → lower variance → less overfit

→ if common Σ is wrong → can lead to high bias

LDA: when fewer observation and don't want to overfit

QDA: when large training set

or when $\Sigma_1 = \Sigma_2 = \Sigma_k$ is clearly very wrong

Comparison of LDA, QDA, Logistic regression and KNN:

Logistic regression and LDA are essentially the same model:

both fit linear decision boundary: $\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$ | $\log\left(\frac{f_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x$

where c_0 and c_1 are a function of μ_1, μ_2 and σ^2

the difference is how you fit the params:

w/ maxim likelihood	w/ assuming a normal distribution w/ a common Σ
better method if normal assumption is not true	better method if this is true

KNN is very different: non-parametric

- better when decision boundary is highly non-linear
- does not tell us which predictors are important

QDA between those above: more flexible boundary than linear, but not as flexible as KNN

- better than KNN if fewer observations

When to use which?

if decision boundary is linear (eg. predictors are not correlated differently for different classes: ~~and really depend on linear fns of the predictors~~)

if normally distributed : LDA

if not normally distributed: Logistic regression

if decision boundary is non-linear (eg predictors have different correlation esp for each class) ~~or depend polynomially on the predictors eg $X_1^2, X_1 X_2$~~

if normally distributed : QDA

if not normally distributed: KNN (cross-validated) for example

and very complex boundary :

Exercises: 1/ I $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

log. reg. $1-p(x) = \frac{1+e^{\beta_0+\beta_1 x} - e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

II $1-p(x) = \frac{1}{1+e^{\beta_0+\beta_1 x}}$

$\frac{I}{II}$: $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$

LDA p=1 2/ $p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum \pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)} = P(G=k \mid X=x)$ when $P(X=x \mid G=j) \sim N(\mu_j, \sigma^2)$

The Bayes classifier is assigning k s.t. $p_k(x)$ the largest

this means $\pi_k \exp(x-\mu_k)^2$ is the largest

$\log \pi_k - \frac{1}{2\sigma^2}(x-\mu_k)^2$ is the largest

which mean that $\log \pi_k + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$ is the largest

this is indeed $\delta_k(x)$.

QDA p=1 3/ ~~QDA p=1~~ $X \sim N(\mu_k, \sigma_k^2)$ in kth class

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

$p_k(x) = \frac{\pi_k f_k(x)}{\sum \pi_k f_k(x)}$, so Bayes maximises $\pi_k f_k(x)$

so want to maximise $\frac{\pi_k}{\sqrt{\sigma_k}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$

maximise $\log \pi_k - \frac{1}{2}\log \sigma_k - \frac{1}{2\sigma_k^2}(x-\mu_k)^2$

maximise $\log \pi_k - \frac{1}{2}\log \sigma_k \left(-\frac{1}{2\sigma_k^2}x^2\right) + \frac{1}{\sigma_k^2} x\mu_k - \frac{1}{2\sigma_k^2}\mu_k^2$

decision boundary is not linear in x

because of this term

KNN

4, $X \sim U(0,1)$

a, we want to predict X using all the observation within 10% of the range of X closest of to that test observation eg. for $X=0.6$, use $[0.55, 0.65]$

on average, what fraction of the observation will we use?

→ 10% $\frac{1}{10}$

KNN

b, $X_1, X_2 \quad U[0,1] \times [0,1]$

within 10% of x_1 and within 10% of X_2

e.g. for $(0.6, 0.35)$, we use $[0.55, 0.65] \times [0.3, 0.4]$

→ 1% of the observation $\frac{1}{100}$

c, $p = 100$

$\frac{1}{10^{100}}$ fraction of the observations

d, → when p is large, KNN gets weaker because fewer observation will be nearer the point. the space gets bigger and the points get further apart.

e, one of hypercube that contains 10% of the training observations:

$p = 1$: $1/10$ → length $= 0.1$

$p = 2$: $x_1^2 = \frac{1}{10}$ → $x_1 = \frac{1}{\sqrt{10}}$ $\quad x_1 = 10^{-\frac{1}{2}}$ length $= 0.316$

$p = 100$ $\quad x_1 = \frac{1}{10^{1/100}}$ $\quad x_1 = 10^{-1/100}$ length $= 0.977$

5/ a, If Bayes boundary is linear: LDA better on test set
QDA might be better on training set as more flexible

LDA/QDA

b, If is non linear: QDA better on both

c, if sample size bigger: we expect QDA to be better than LDA as sample size increases
QDA includes LDA within it, ~~so would not be worse, but~~

d, False: if we don't have a huge sample size, QDA will overfit and be worse than LDA
if Bayes is linear

6, x_1 = hours studied

x_2 = undergrad GPA

Y = get an A

Logistic regression: $\hat{\beta}_0 = -6$

$\hat{\beta}_1 = 0.05$

$\hat{\beta}_2 = 1$

log. reg.

$x_1 = 40$

$x_2 = 3.5$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

a, $$P(\text{get an A}) = \frac{e^{(-1/2)}}{1 + e^{-1/2}} = 0.3775$$

b, $P(\text{get an A}) = 0.5$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$0 = -6 + 0.05 x_1 + 3.5$$

$$x_1 = \frac{2.5}{0.05} = 50$$

LDA

7, predict if stock gives divident

$\bar{X} = 10$ if yes (1) x: last years profit

$\bar{X} = 0$ if no (0)

$\hat{\sigma}^2 = 36$

80% of companies gave dividends.

$\hat{\mu}_0 = 0$ $\hat{\sigma}^2 = 36$

$\hat{\mu}_1 = 10$

$\pi_0 = 0.2$

$\pi_1 = 0.8$

$$P(\text{yes} \mid x=4) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$ from Bayes.

$$P(A \mid B) = \frac{P(A)\,P(B \mid A)}{P(B)}$$

$$P(1 \mid x=4) = \frac{0.4852}{0.4852 + 0.16} = 0.752$$

log reg

KNN

8, log reg: 20% on training, 30 on test training size = test size

KNN 1: 18% over both

~~we don't know which is better could be~~ ~~log reg.~~ ~~KNN 2% on taining, 34% on test set~~ ~~or KNN 18% on training and 18% on test~~

KNN has 0% training error, so must have 36% test set error → log reg is better

9, odd of 0.37 to defaults

on average, what fraction defaults!

odds

a) odd: $\frac{p(x)}{1-p(x)} = 0.37 \rightarrow p(x) = \frac{0.37}{1.37} = 0.27$

b) 16% of default.

odd: $\frac{0.16}{1-0.16} = 0.19$

Resampling methods - Chapter 5