Logistic Regression

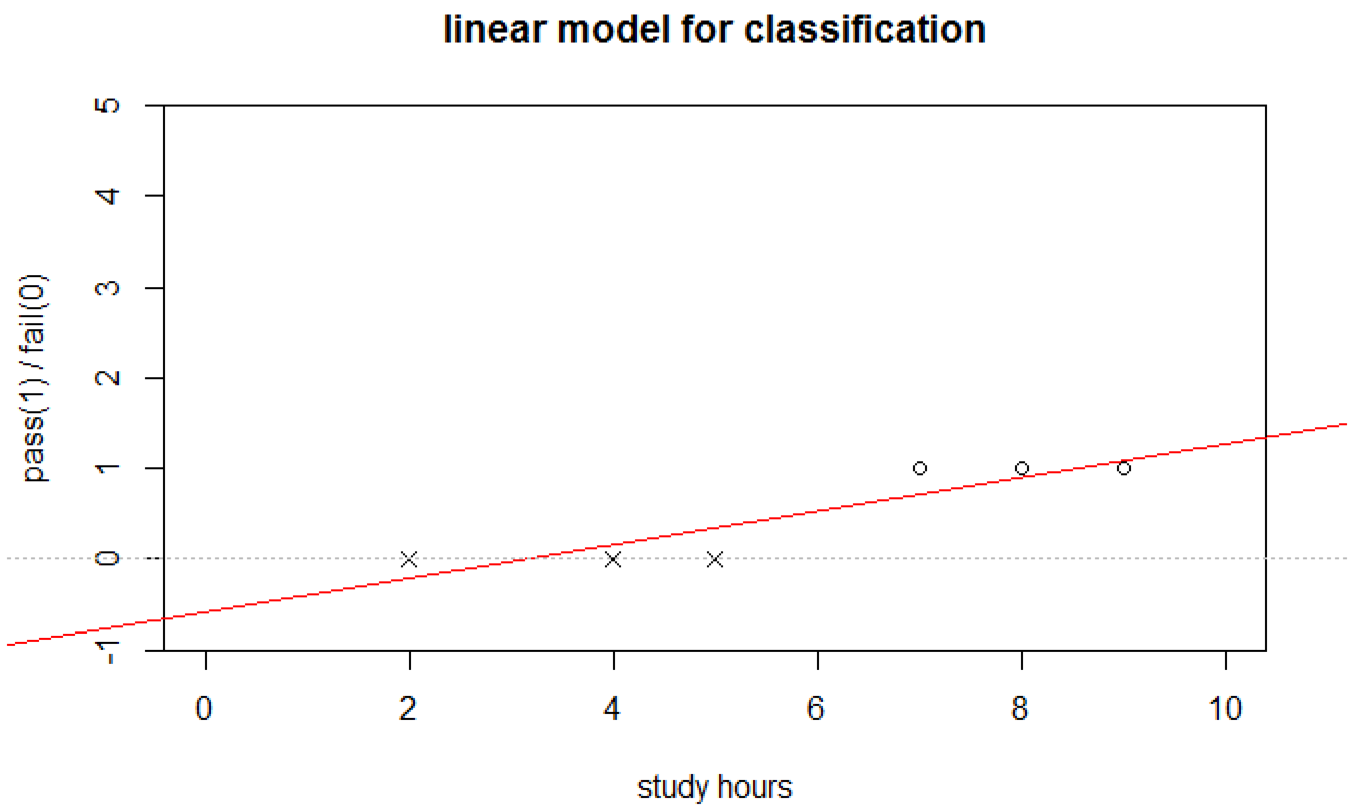
Introduction to Data Science

Presented by Hyebong Choi

Contents

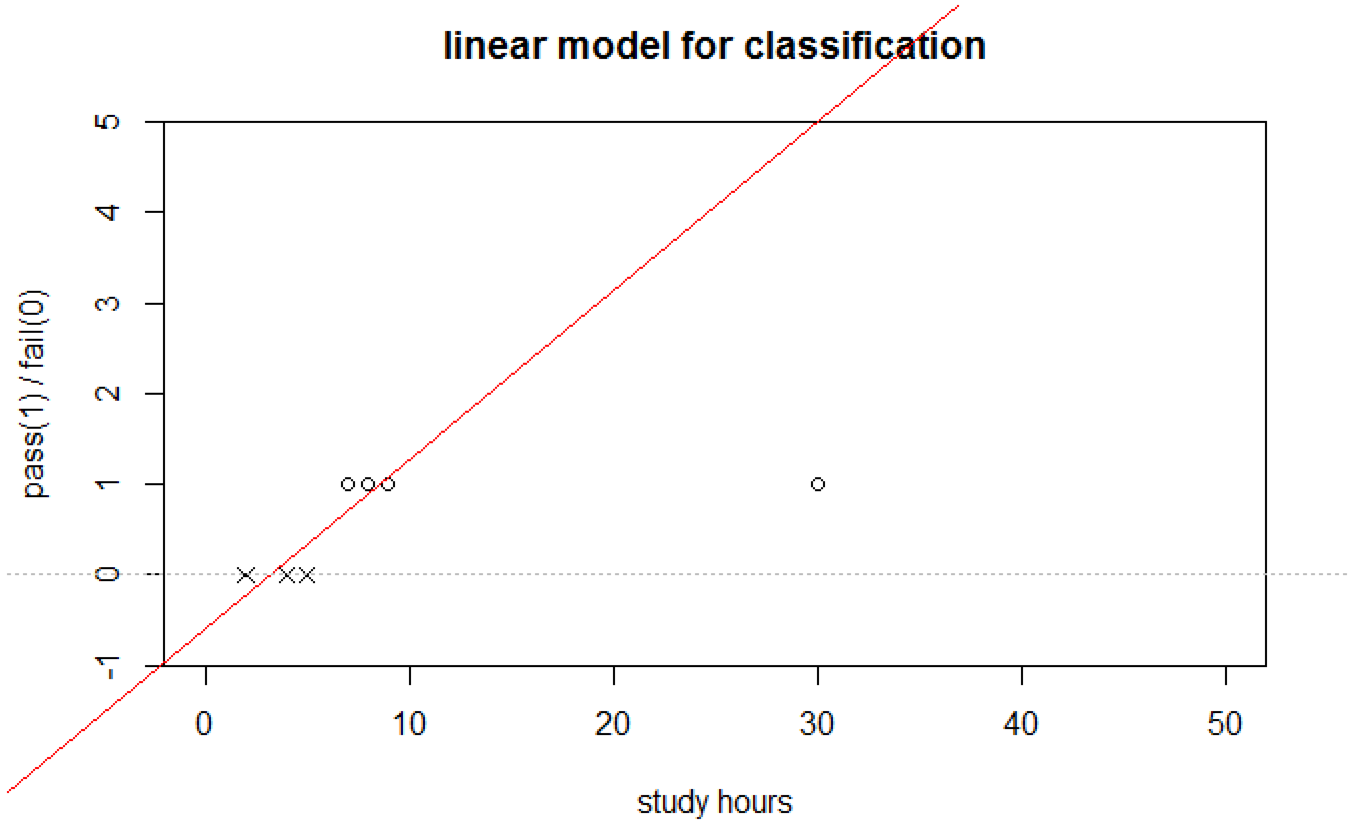
- Concept
- Example

Linear Model for Classification



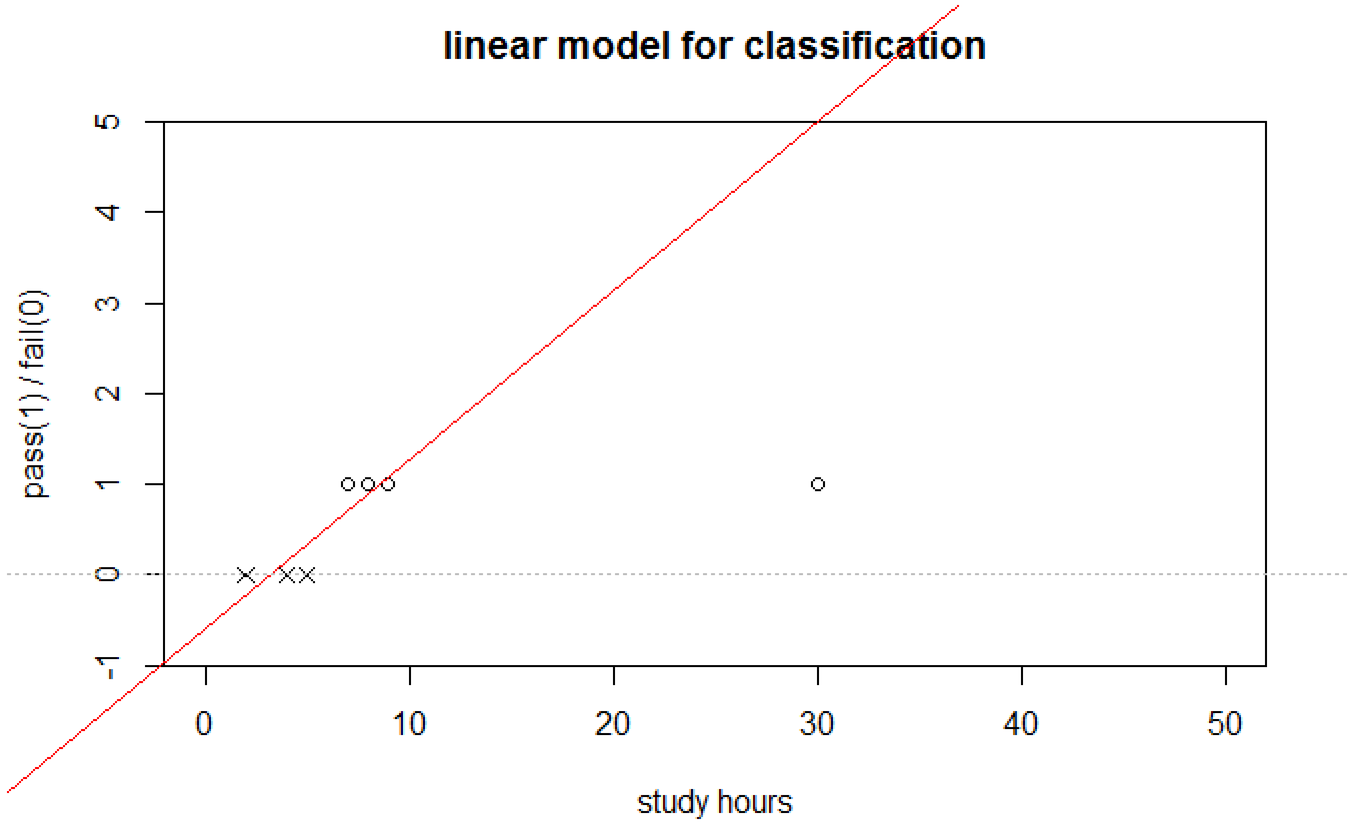
- y = 0 represents failure for the class, y = 1 for pass

Linear Model for Classification



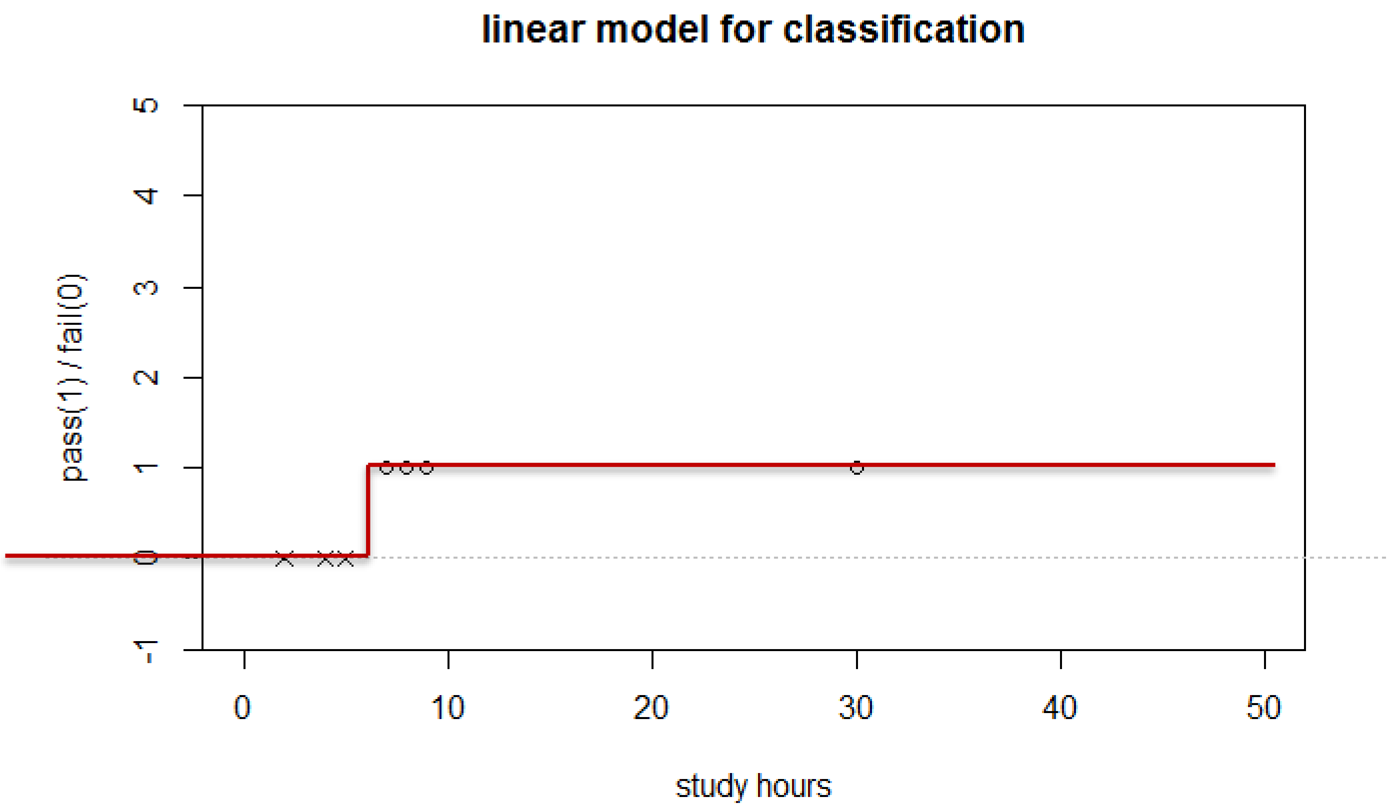
- What if there is a new student who studied 30 hours?
- The predicted value is way larger than 1 which lead to large error
- We want to have predicted value to be P(pass) which is between 0 and 1

Linear Model for Classification



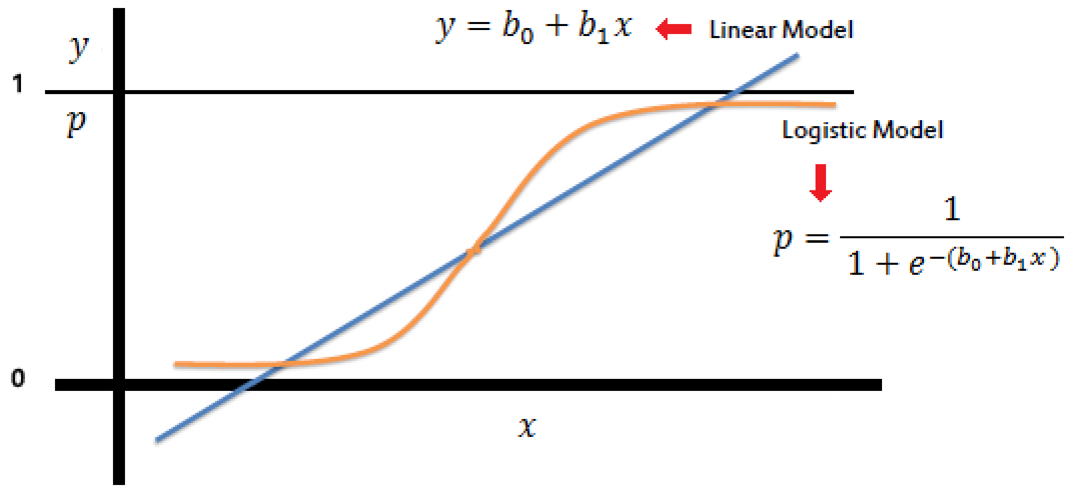
- What if there is a new student who studied 30 hours?
- The predicted value is way larger than 1 which lead to large error
- We want to have predicted value of 1, 0, or in between…

Linear Model



- We want to make something like this…
- Hypothesis of y = sign$(w \times x + b)$
- But, it is difficult to calculate in math to find best w and b

Sigmoid function



- Linear model predict value z from $-\infty \sim \infty$
- Sigmoid function g maps z to 0 ~ 1,
- When z is negative, g(z) goes close to 0
- When z is positive, g(z) goes close to 1
- When z is 0, g(z) = 0.5
- We estimate P(positive) with sigmoid function

Logistic Regression

- Linear regression to find out logit function of probability p
  - or sigmoid function of linear regression model

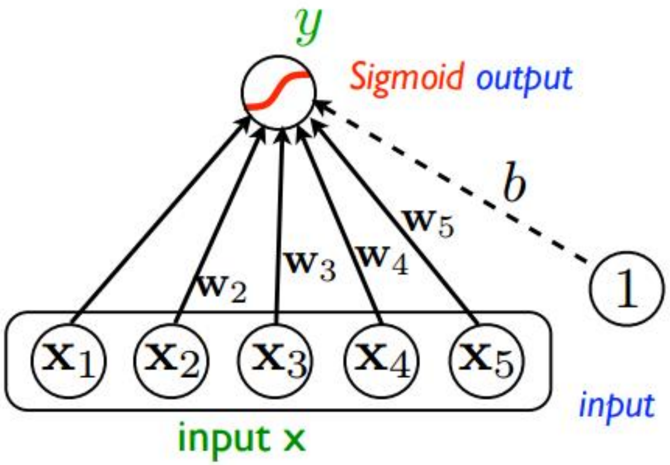
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

logit function of p

$$\hat{p} = \frac{\exp(b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p)}{1 + \exp(b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p)}$$

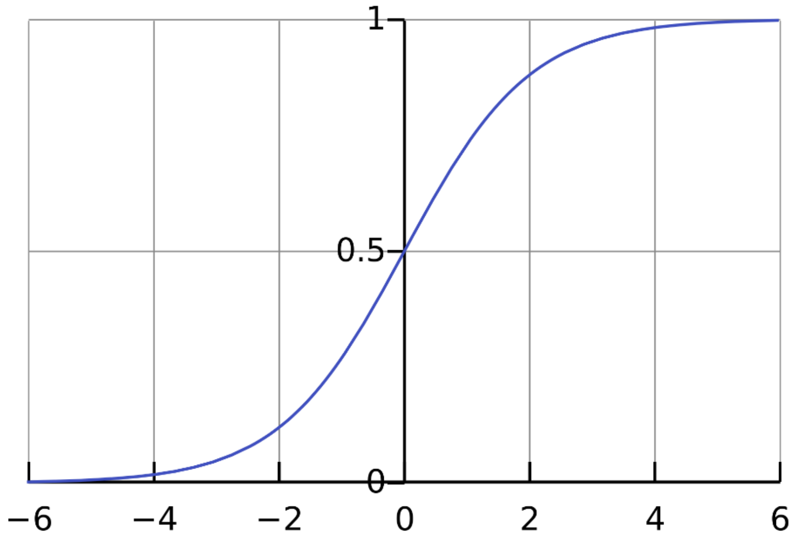
$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

sigmoid function

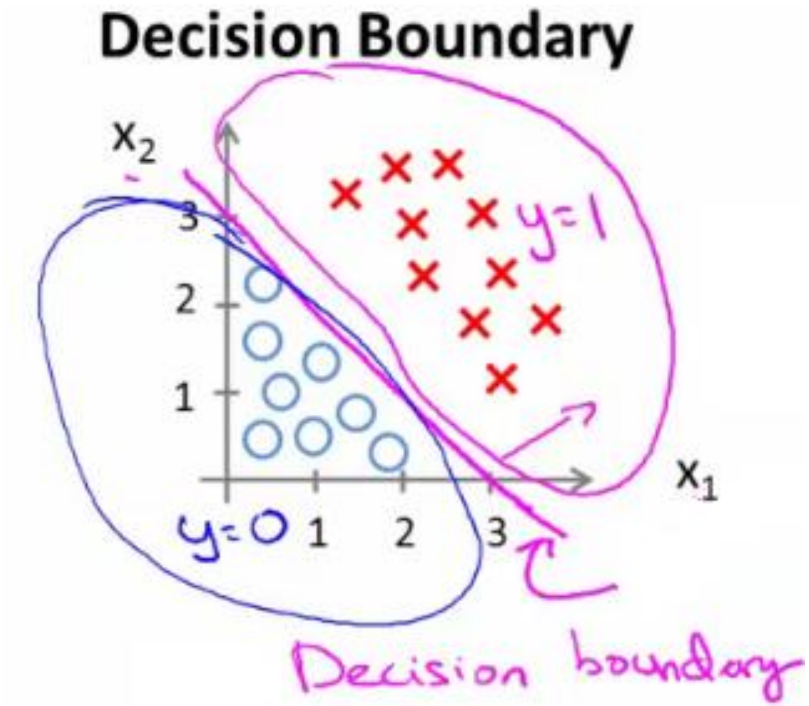


Logistic Regression

- Linear regression to find out logit function of probability p
  - or sigmoid function of linear regression model



sigmoid function



Example

- CDC 2010 natality public-use data file (http://mng.bz/pnGy)
- all births registered in the 50 US States and the District of Columbia,
  - including facts about the mother and father, and about the delivery.

```
load(url('https://github.com/hbchoi/SampleData/raw/master/NatalRiskData.rData'))

train <- sdata[sdata$ORIGRANDGROUP <= 5, ]
test <- sdata[sdata$ORIGRANDGROUP > 5, ]
```

data loading

Example

Variable	Type	Description
atRisk	Logical	TRUE if 5-minute Apgar score < 7; FALSE otherwise
PWGT	Numeric	Mother's prepregnancy weight
UPREVIS	Numeric (integer)	Number of prenatal medical visits
CIG_REC	Logical	TRUE if smoker; FALSE otherwise
GESTREC3	Categorical	Two categories: <37 weeks (premature) and >=37 weeks
DPLURAL	Categorical	Birth plurality, three categories: single/twin/triplet+
ULD_MECO	Logical	TRUE if moderate/heavy fecal staining of amniotic fluid
ULD_PRECIP	Logical	TRUE for unusually short labor (< three hours)
ULD_BREECH	Logical	TRUE for breech (pelvis first) birth position
URF_DIAB	Logical	TRUE if mother is diabetic
URF_CHYPER	Logical	TRUE if mother has chronic hypertension
URF_PHYPER	Logical	TRUE if mother has pregnancy-related hypertension
URF_ECLAM	Logical	TRUE if mother experienced eclampsia: pregnancy-related seizures

Building Model

```
# making formula for logistic regression model

complications <- c("ULD_MECO", "ULD_PRECIP", "ULD_BREECH")
riskfactors <- c("URF_DIAB", "URF_CHYPER", "URF_PHYPER", "URF_ECLAM")
y <- "atRisk"
x <- c("PWGT","UPREVIS","CIG_REC","GESTREC3", "DPLURAL", complications,
riskfactors)
fmla <- paste(y, paste(x, collapse = '+'), sep='~')

print(fmla)

## [1]
"atRisk~PWGT+UPREVIS+CIG_REC+GESTREC3+DPLURAL+ULD_MECO+ULD_PRECIP+ULD_BREECH+
URF_DIAB+URF_CHYPER+URF_PHYPER+URF_ECLAM“

# building logistic regression model
model <- glm(fmla, data = train, family = binomial(link='logit'))
```

Make Prediction

```
train$pred <- predict(model, newdata = train, type = 'response')
test$pred <- predict(model, newdata = test, type = 'response')

test[20:40, c('pred','atRisk')]
```

```
##             pred atRisk
## 2185 0.011507461  FALSE
## 2188 0.058792989  FALSE
## 2189 0.063196603  FALSE
## 2192 0.022661796  FALSE
## 2193 0.050933807   TRUE
## 2194 0.012455440  FALSE
## 2195 0.012204660   TRUE
## 2196 0.011317563  FALSE
## 2204 0.002147274  FALSE
## 2207 0.062311633  FALSE
## 2210 0.007884831  FALSE
## 2211 0.008353482  FALSE
## 2212 0.060224116  FALSE
## 2213 0.009169627  FALSE
## 2217 0.008296810  FALSE
## 2219 0.008113151  FALSE
## 2220 0.009405842  FALSE
## 2221 0.022770689  FALSE
## 2228 0.007739053  FALSE
```

```
aggregate(pred ~ atRisk, data = train, mean)

##   atRisk       pred
## 1  FALSE 0.01853135
## 2   TRUE 0.05381493

aggregate(pred ~ atRisk, data = test, mean)

##   atRisk       pred
## 1  FALSE 0.01938838
## 2   TRUE 0.04997396
```

Listing 7.11 Plotting distribution of prediction score grouped by known outcome

```
library(ggplot2)
ggplot(train, aes(x=pred, color=atRisk, linetype=atRisk)) +
       geom_density()
```

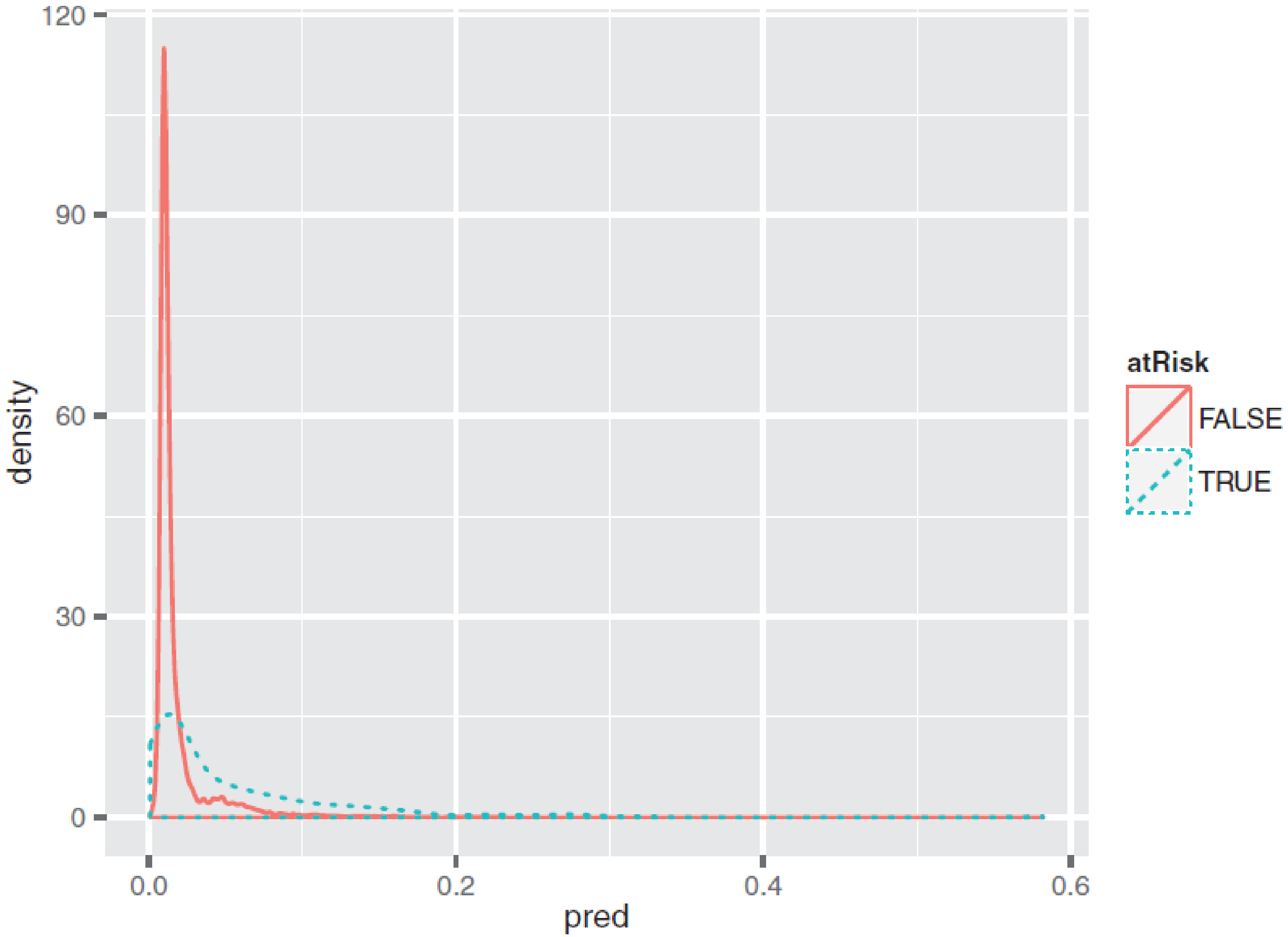


Figure 7.9 **Distribution of score broken up by positive examples (`TRUE`) and negative examples (`FALSE`)**

Precision and Recall (/w threshold)

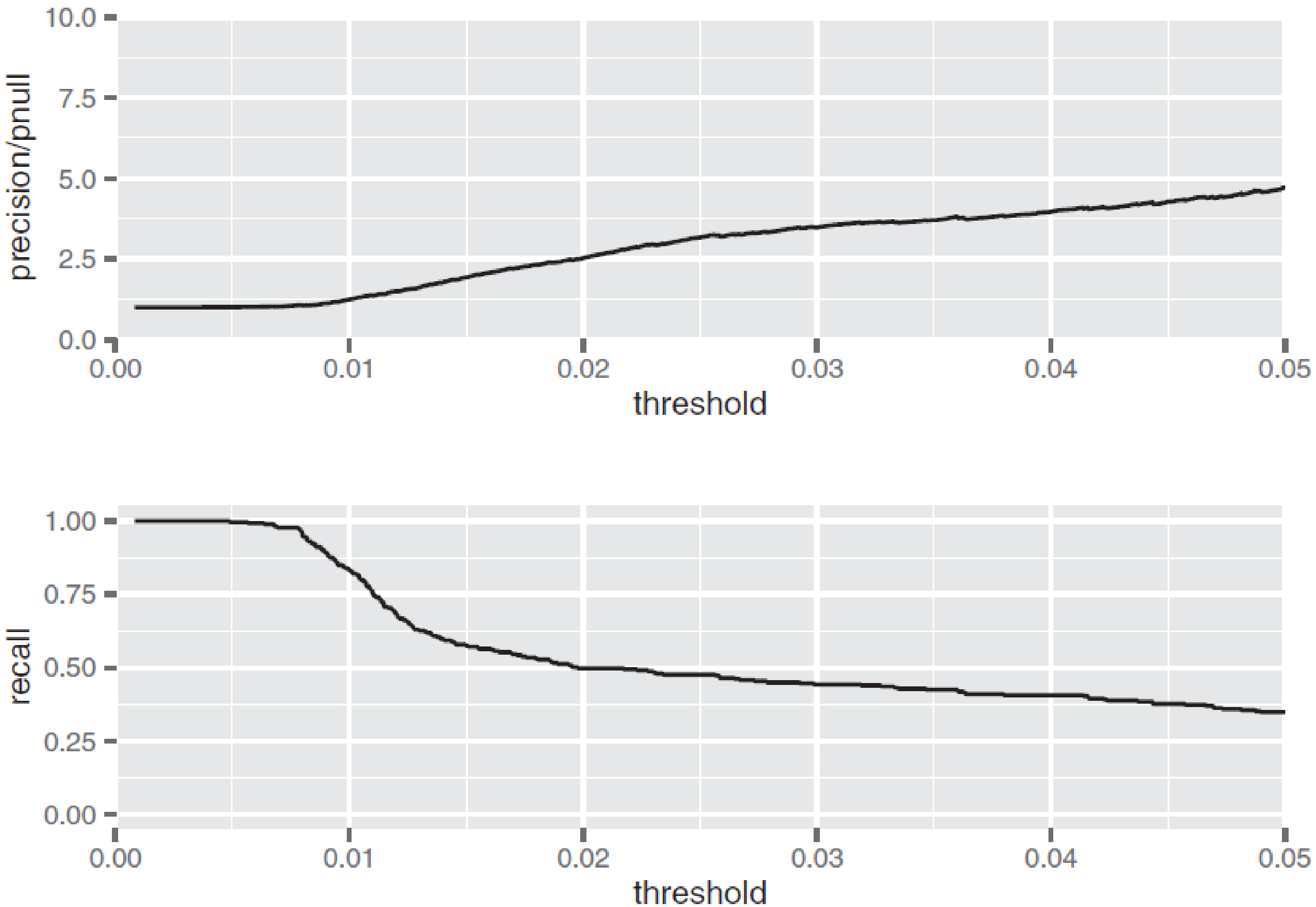


Figure 7.10 Enrichment (top) and recall (bottom) plotted as functions of threshold for the training set

Pick threshold (here 0.02)

Listing 7.13 Evaluating our chosen model

Build confusion matrix.

```
> ctab.test <- table(pred=test$pred>0.02, atRisk=test$atRisk)
 > ctab.test
        atRisk
pred     FALSE TRUE
  FALSE  9487   93
  TRUE   2405  116
> precision <- ctab.test[2,2]/sum(ctab.test[2,])
> precision
[1] 0.04601349
> recall <- ctab.test[2,2]/sum(ctab.test[,2])
> recall
[1] 0.5550239
> enrich <- precision/mean(as.numeric(test$atRisk))
> enrich
[1] 2.664159
```

Rows contain predicted negatives and positives; columns contain actual negatives and positives.

- Difficult to find definitive threshold
- when threshold = 0, recall is 1.0 and precision is 0.0173
- We set threshold to maximize the recall considering the resource (i.e. emergency equipment) that the hospital can afford

Coefficients

```
coefficients(model)
```

```
##          (Intercept)                  PWGT                  UPREVIS
##          -4.41218940            0.00376166              -0.06328943
##          CIG_RECTRUE     GESTREC3< 37 weeks DPLURALtriplet or higher
##           0.31316930            1.54518311               1.39419294
##          DPLURALtwin           ULD_MECOTRUE           ULD_PRECIPTRUE
##           0.31231871            0.81842627               0.19172008
##       ULD_BREECHTRUE           URF_DIABTRUE           URF_CHYPERTRUE
##           0.74923672           -0.34646672               0.56002503
##       URF_PHYPERTRUE          URF_ECLAMTRUE
##           0.16159872            0.49806435
```

INTERPRETING THE COEFFICIENTS

Interpreting coefficient values is a little more complicated with logistic than with linear regression. If the coefficient for the variable `x[,k]` is `b[k]`, then the odds of a positive outcome are multiplied by a factor of `exp(b[k])` for every unit change in `x[,k]`.

The coefficient for `GESTREC3< 37 weeks` (for a premature baby) is 1.545183. So for a premature baby, the odds of being at risk are `exp(1.545183)=4.68883` times higher compared to a baby that's born full-term, with all other input variables unchanged. As an example, suppose a full-term baby with certain characteristics has a 1% probability of being at risk (odds are `p/(1-p)`, or `0.01/0.99 = 0.0101`); then the odds for a premature baby with the same characteristics are `0.0101*4.68883 = 0.047`. This corresponds to a probability of being at risk of `odds/(1+odds)`, or `0.047/1.047`—about 4.5%.

Similarly, the coefficient for `UPREVIS` (number of prenatal medical visits) is about -0.06. This means every prenatal visit lowers the odds of an at-risk baby by a factor of `exp(-0.06)`, or about 0.94. Suppose the mother of our premature baby had made no prenatal visits; a baby in the same situation whose mother had made three prenatal visits would have odds of being at risk of about `0.047 * 0.94 * 0.94 * 0.94 = 0.039`. This corresponds to a probability of being at risk of 3.75%.

So the general advice in this case might be to keep a special eye on premature births (and multiple births), and encourage expectant mothers to make regular prenatal visits

References

- Practical Data Science with R, by Nina Zumel and John Mount
- R을 이용한 데이터 분석 실무, 서민구, 길벗
- Machine Learning with R, by Brett Lantz
  - 한글판, R을 활용한 기계 학습
- [DBGUIDE 연재] ggplot2를 이용한 R 시각화
  - http://freesearch.pe.kr/archives/3134