Binomial distribution → mle $\hat{\pi} = \frac{y}{n}$

s.e. $\sqrt{\hat{\pi}(1-\hat{\pi})/n}$

$w = 2\left[\ell(\hat{\pi}_1, \hat{\pi}_2) - \ell(\hat{\pi}, \hat{\pi})\right]$

since $w \sim \chi^2_{p=1}$

p-value $\Pr[\chi^2 > w]$, $\chi^2 \sim \chi^2_1$

influence of x on π

→ age influence the satisfaction (old customers tends to have higher chances to be satisfied)

↳ age is continuous : how can we see the influence of a continuous variable on the response (probability of being among satisfied customers)

π as a function of x

↓

logistic regression $\quad \pi = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}}$ $\quad \eta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$ ← linear predictor

↑

logistic function $\quad \pi \in [0; 1]$, $\eta(x) \in (-\infty; +\infty)$

in the figure $\eta(x) = \beta_0 + \beta_1 x$

→ logistic regression is a special case of GLM (Generalized Linear Models)

$g(E[Y|x]) = \eta(x)$ ← link function

→ logistic → binomial distribution

Other distributions
- gamma
- poisson
- exponential family

$\pi = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}}$

$\pi + e^{\eta(x)}\pi = e^{\eta(x)}$

$e^{\eta(x)} - e^{\eta(x)}\pi = \pi$

$e^{\eta(x)}(1 - \pi) = \pi$

$e^{\eta(x)} = \frac{\pi}{1-\pi}$

$\eta(x) = \log\left(\frac{\pi}{1-\pi}\right)$

link function : logit

$g(x)$ for binomial response

$\frac{\pi}{1-\pi}$ = odds

$\pi = \frac{\text{odds}}{1 + \text{odds}}$

Example

$x \in \{0, 1\}$ (0: old, 1: young)

$$\eta(x) = \beta_0 + \beta_1 x \overset{!}{=} 1.5299 + (-0.5466)x \overset{!}{=} 1.5299 - 0.5466x$$

old $\eta(x) = 1.5299 - 0.5466 \cdot 0 = 1.5299$ ← baseline

$\eta(x) = \log\left(\frac{\pi}{1-\pi}\right)$ ← log-odds

$$\pi = \frac{e^{\eta(x)}}{1+e^{\eta(x)}} \rightarrow \hat{\pi}_{old} = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = 0.822$$

young: $\eta(x) = 1.5299 - 0.5466 \cdot 1 = 0.9853$

$$\rightarrow \hat{\pi}_{young} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = 0.728$$

reverse:

$$\hat{\beta}_0 = \log\left(\frac{\hat{\pi}_{old}}{1-\hat{\pi}_{old}}\right) \qquad \hat{\beta}_0 + \hat{\beta}_1 = \log\left(\frac{\hat{\pi}_{young}}{1-\hat{\pi}_{young}}\right)$$

$$\hat{\beta}_1 = \log\left(\frac{\hat{\pi}_{young}}{1-\hat{\pi}_{young}}\right) - \hat{\beta}_0$$

$$= \log\left(\frac{\hat{\pi}_{young}}{1-\hat{\pi}_{young}}\right) - \log\left(\frac{\hat{\pi}_{old}}{1-\hat{\pi}_{old}}\right) \qquad \beta_1$$

$$= \log\left(\frac{\hat{\pi}_{young}}{1-\hat{\pi}_{young}} \Big/ \frac{\hat{\pi}_{old}}{1-\hat{\pi}_{old}}\right)$$ ← log odds ratio

x continuous

$$W = 2\left(\ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) - \ell(\hat{\beta}_0, \hat{\beta}_1)\right)$$

$$D(\hat{\beta}) = -2\,\ell(\hat{\beta})$$

$$= D(\hat{\beta}_0, \hat{\beta}_1) - D(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \sim \chi^2_1$$

$$\hat{\beta}_1 = \hat{\beta}_0 + \hat{\beta}_1(x+1) - \hat{\beta}_0 - \hat{\beta}_1 x$$

expected increase in the log odds

$$\hat{\beta}_1 = \log\frac{\hat{\pi}_{x+1}}{1-\hat{\pi}_{x+1}} - \log\frac{\hat{\pi}_x}{1-\hat{\pi}_x} = \log\left(\frac{\frac{\hat{\pi}_{x+1}}{1-\hat{\pi}_{x+1}}}{\frac{\hat{\pi}_x}{1-\hat{\pi}_x}}\right)$$

$$e^{\hat{\beta}_1} = \frac{\frac{\hat{\pi}_{x+1}}{1-\hat{\pi}_{x+1}}}{\frac{\hat{\pi}_x}{1-\hat{\pi}_x}}$$

$e^{\hat{\beta}_1}$ = expected increase in the odds

In this lecture

$n = 30$ (y_i, x_i) $i = 1, \dots, n$ from $y = f(x) + \varepsilon$ ← estimate $f(x)$

yesterday data

GOAL: predict the value of y^{new} when new x become available : $\hat{f}(x)$

increasing p (degree of the polynomial)

→ smaller $D(\hat{f}(x)) = \|y - \hat{y}\|^2$ → pen → $D(\hat{f}(x)) = 0$

→ larger $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ → pen → $R^2 = 1$

Our goal is not to perfectly fit the data that we have, but to predict the outcome/response for new observations

→ a polinomial of really high degree fit really well the data used to construct it, but it is useless to predict new observation

→ evaluation is done on new data (prediction)

→ "residual deviance" (computed on the new data) first decrease (the model is getting better), until a certain point, then increase (the model is getting worse) as a function of p

→ "R^2" first increase, then decrease, same about the model

→ Why we use p on the x-axis → measure (because we are using a polynomial) of the model complexity

$\hat{y}_0 = \hat{f}(x_0)$ y_0 = new response, x_0 = new input / covariate

$$E\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right]$$

$\hat{f}(x_0) = \hat{y}_0$ predicted value

$f(x_0)$ true value

our aim is to minimize the expected (squared) error

↳ $E\left[|\hat{f}(x_0) - f(x_0)|\right]$

$$E\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] = \left(E\left[\hat{f}(x_0)\right] - f(x_0)\right)^2 + Var\left[\hat{f}(x_0)\right]$$

bias^2 variance

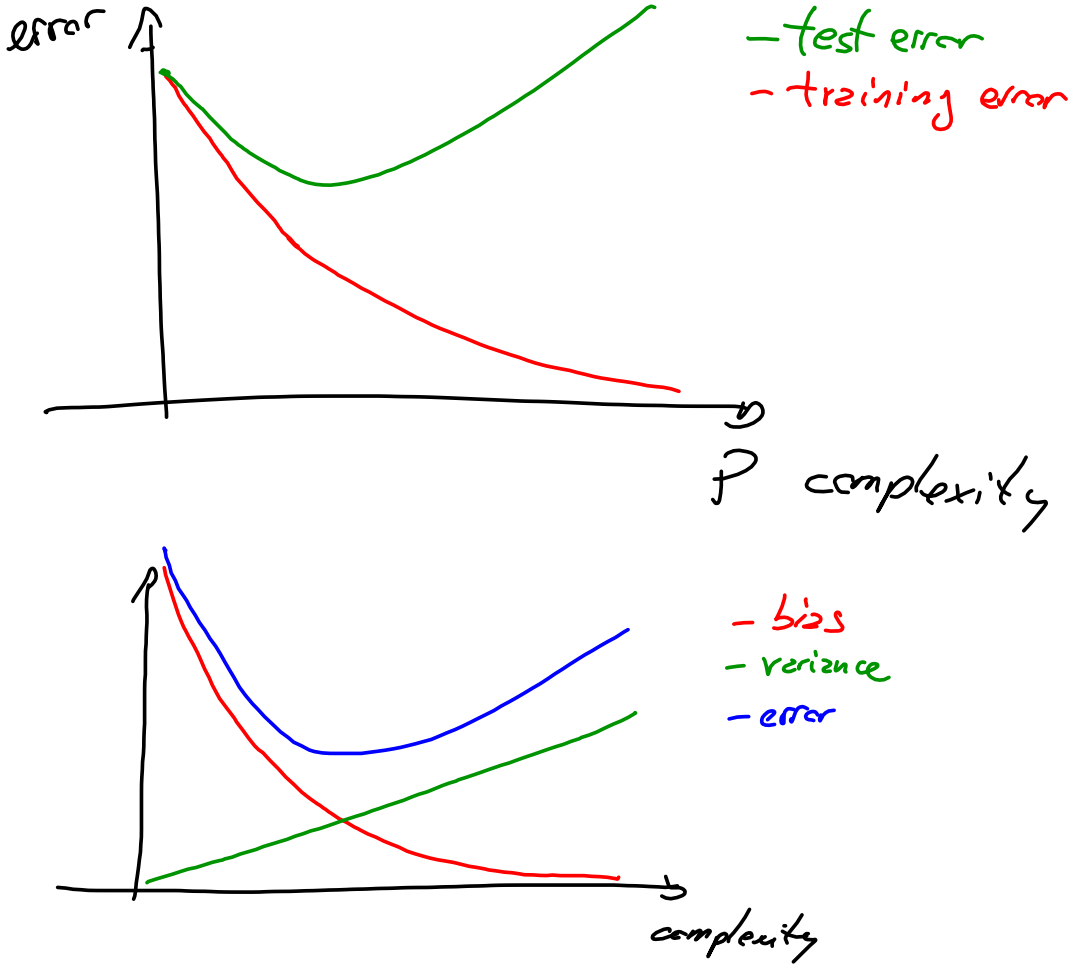
→ do exercises 3.1 and 3.2 to derive this decomposition

$$E\left[\left(\hat{f}(x_0) - y_0\right)^2\right] = \left(E\left[\hat{f}(x)\right] - f(x_0)\right)^2 + Var\left[\hat{f}(x_0)\right] + \varepsilon$$

bias^2 variance irreducible error

$y_0 = f(x_0) + \varepsilon$

mean squared error


error
- test error
- training error
p complexity
- bias
- variance
- error
complexity

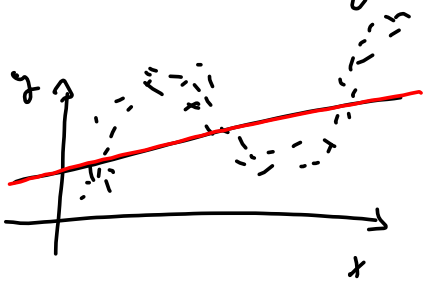
A model:

low complexity => high bias, low variance

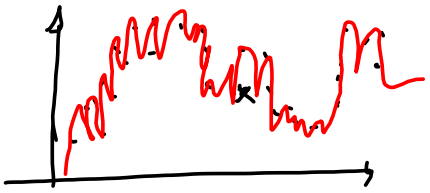
high complexity => low bias, high variance

the model does not capture well the systematic pattern of the data, but it is insensible to small changes

the model fits well the data, but small changes will lead to big variations



complexity low
high bias
low variance



complexity high
low bias
high variance

OVERFITTING

the estimated function $\hat{f}(x)$ follows random fluctuation of the data => dramatic increase in variance, negligible gain in bias

$f(x) + \varepsilon$ — not

BIAS - VARIANCE TRADE-OFF

- bias and variance are "contrasting" (one cannot decrease both simultaneously, decreasing one means an increasing other
- find the best balance between bias and variance to minimize the error

- OLS → estimator with the smallest variance among the unbiased
- statistical learning → we accept to increase the bias to have a decrease in the variance to reduce the expected prediction error

e.g.: $\hat{\beta}_{Ridge}$ $\hat{\beta}\left(X^TX + \lambda I\right)^{-1}X^Ty$

↳ increase the bias, reduce the variance

(over)optimism: when we evaluate the error in the data that we used to construct the model, we are underestimating the error

Unfortunately, $f(x)$ is unknown, so we cannot directly use the formulas that we just saw.

Goal: avoid (over)optimism in the computation of the error, in order to not risk to overfit our data

we want to avoid to model random fluctuation of the data, and only focus on the structural part of the phenomenon of interest

→ with new data, the structural part will remain the same, the random fluctuation will change

$y = f(x) + \varepsilon$ (same / different) (observations of yesterday / tomorrow)

Trick: evaluate the model on data that are different from those used to construct it!

- the evaluation of the new data is not perfect, because it is influenced by the randomness of the new data

The point here is that with new data we do not suffer from the overfitting issues ⇒ it is not perfect, but it is very good nevertheless

How to do this in practice?

- random split the data in two sets
  - training set: to construct the model
  - test set: to evaluate the model
- resampling technique
  - cross-validation (repeated cross-validation)
  - bootstrapping
- information criteria (we compute the train error and we try to estimate the (over) optimism)
  - AIC
  - BIC
  - AIC_c

Random split the data in a training and a test set

- simplest way, randomly assign each observation to one set

training set: use to construct the models (one for each p)

test set: we evaluate the models and choose the best one ("right" p)

Sometimes one splits these two operations in two different sets

→ training set: train
→ test set: choose the best
→ validation set: evaluate its performance

There is no clear rule for the proportions. In the book:

training	test	validation
50%	25%	25%
75%	25%	0%
66.6̄%	33.3̄%	0%
50%	50%	0%

also depends on the sample size