Introduction to Statistics & Machine Learning

GTIPI Summer School

Irene Schmidtmann (Irene.Schmidtmann@uni-mainz.de)

IMBEI - University Medical Center Mainz

2022/05/30

IMBEI@

JG|U UNIVERSITÄTSmedizin.
MAINZ

e:Med
SYSTEMS MEDICINE

SPONSORED BY THE
Federal Ministry
of Education
and Research

Topics

- Curse of dimensionality - or a blessing in disguise?
- Multiple testing
- Linear model selection and regularization or How to obtain parsimonious statistical models
- How to extract essential information from data: PCA

Multiple testing

Statistical tests - one and many

Classical statistical tests

- An example
- Why test hypotheses and what to do with the result?
  - Neyman Pearson paradigm
  - Fisher's approach

Testing many hypotheses

- Problems encountered when testing more than one hypothesis
- Possible solutions

Classical statistical tests

Example

- We know that a certain trait A is present in 20% of the general population.
- We study a disease D possibly related to A
- Question: Is A more frequent in patients than in the general population?
- Study to answer this question: obtain a sample of patients with disease D and determine how often trait A is present.
- Compare frequency in patient sample to reference value.

Classical statistical tests

Example (continued)

- Simulated data for 100 fictious patients with disease D
- Statistical model:
  - Binomial distribution with parameters
    - $n =$ number of patients (here: $n = 100$),
    - $p =$ probability of observing trait A
  - Random variable K describes number of observed occurences of A

Table 1: Frequency of trait

patientTraits	Freq
A	30
not A	70

- To answer research question, test hypothesis
  - $H_0 : p = 0.2$ against
  - $H_1 : p > 0.2$
- Use binomial test to compare observed frequency of trait A to reference value

Classical statistical tests

Example (continued)

- Determine test statistic $T = K$
- Reject null hypothesis if $T \geq k_0$ with
    - $k_0 =$ critical value, chosen such that $P(T \geq k_0) \leq \alpha$ if H_0 is true.
    - $\alpha =$ probability of type I error, i. e. of erroneously rejecting null hypothesis
- Choose $\alpha = 0.05 \Rightarrow k_0 = 27$
- Observed $k = 30 > k_0 = 27 \Rightarrow$ Reject null hypothesis

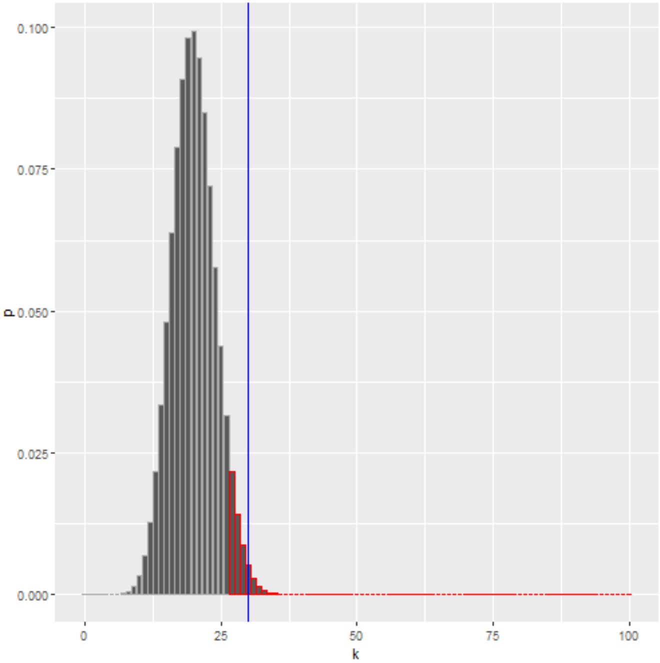


Fig. 1: Distribution of T given H_0 with rejection region

Exact Binomial test in R

One-sided test

```
binom.test(x = numA, n = numPatients, p = 0.2,
           alternative = "greater")
```

```
##
##      Exact binomial test
##
## data:  numA and numPatients
## number of successes = 30, number of trials = 100, p-value = 0.01125
## alternative hypothesis: true probability of success is greater than 0.2
## 95 percent confidence interval:
##  0.2249232 1.0000000
## sample estimates:
## probability of success
##                    0.3
```

- p-value: Probability of observing an event at least as extreme as the one actually observed, given the null hypothesis is true.

Why test hypotheses and what to do with the result?

Neyman Pearson paradigm "Hypothesis testing"

- We need to make a decision, e. g. approve new medication
- originally decision between (just) 2 distributions - simple alternative hypothesis
- in practice usually composite alternative hypothesis
    - set up two statistical hypotheses null hypothesis H_0 and alternative hypothesis H_1
    - decide on α (probability of type I error) and β (probability of type II error), necessary sample size follows
    - if data in rejection region of H_0, accept H_1
    - otherwise keep H_0

Why test hypotheses and what to do with the result?

Fisher's approach "Significance test"

- We want to learn from our data
  - require null hypothesis
  - report exact level of significance (p-value)
- p-value as measure for credibility of null hypothesis

Overall

- Unresolved dispute over formulations

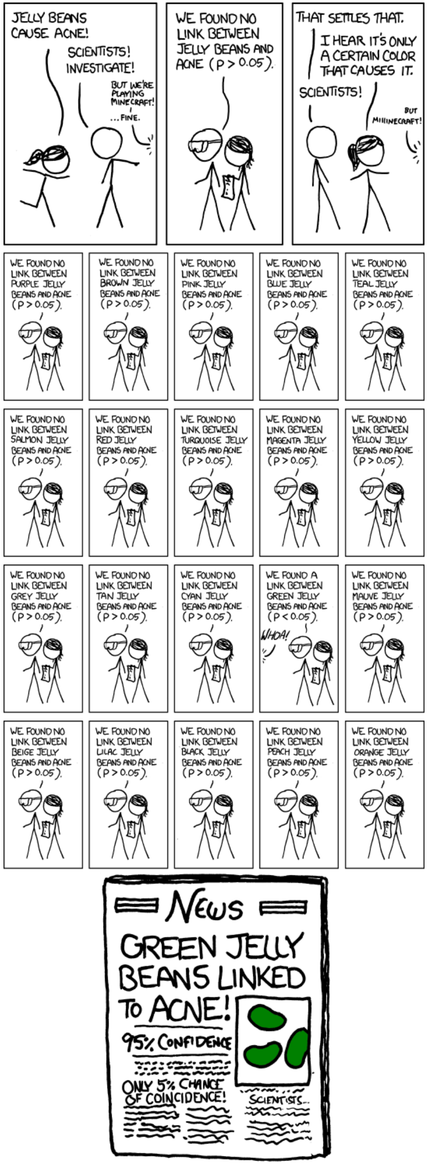

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Fig.2: https://imgs.xkcd.com/comics/significant.png

Testing many hypotheses

Problems encountered when testing more than one hypothesis

- Increase in type I error
- Temptations
    - … looking for other tests if the first one did not give the desired result
    - … changing the hypothesis if the original one did not give the desired result
    - … and even worse: HARKing – hypothesizing after the results are known
- Such proceeding violates the rules and underlying assumptions of hypothesis testing.

Testing more than one hypothesis

Table 2: Types of error in multiple testing.

Test vs reality	H_0 **is true**	H_0 **is false**	**Total**
Rejected	V	S	R
Not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

with

- m: total number of hypotheses
- m_0: number of true null hypotheses
- V: number of false positives (a measure of type I error)
- T: number of false negatives (a measure of type II error)
- S, U: number of true positives and true negatives
- R: number of rejections

m is known and R can be observed.

Family wise error rate

- Family wise error rate (FWER): is the probability that V>0, i.e. that we reject at least one true null hypothesis, i. e. we produce one or more false positive results.
- If all tests are independent we find

$$\begin{aligned} P(V > 0) &= 1 - P(\text{no rejection of any of } m_0 \text{ nulls}) \\ &= 1 - (1 - \alpha)^{m_0} \to 1 \text{ as } m_0 \to \infty \end{aligned}$$

- Worst case: $P(V > 0) = \min(1, m_0 \cdot \alpha)$
- Serious problem if thousands of genes are to be tested

Controlling family wise error rate

Bonferroni correction

- $m_0 \leq m \Rightarrow$ we are on the safe side (in terms of type I error) if we choose $\alpha = \alpha_{\mathrm{FWER}}/m$
- Drawback: Required individual α's become extremely small

  $\Rightarrow \beta$'s increase, i. e. loss of power even for moderate number of tested hypotheses m

A different concept of error control: false discovery rate (FDR)

- No longer try to keep the FWER below α, but require a boundary on the proportion of erroneously rejected null hypotheses
- FDR is the **expected** proportion of type I errors out of the rejections made
- $\mathrm{FDR} = \mathrm{E}\left[\frac{V}{\max(R,1)}\right] = \mathrm{E}\left[\frac{V}{R} \mid R > 0\right] \cdot P(R > 0)$
- FDR = FWER if **all** null hypotheses are true
- As FDR is an expectation (average), individual FDR could be much worse

How to control false discovery rate

Example

(Taken from Holmes, Huber. Modern statistics for modern biology 6.9)

- Use RNA-Seq dataset airway
    - contains gene expression measurements (gene-level counts) of four primary human airway smooth muscle cell lines with and without treatment with dexamethasone
    - interest is in differential expression
    - use the DESeq2 method (more later, see alsoe chapter 8 in Modern statistics for modern biology)
- Obtain large number of p-values
- How to decide which hypotheses should be rejected?

Controlling false discovery rate

p-value histogram

- p-value histogram exhibits mixture composed of two components:
    - p-values resulting from the tests for which the null hypothesis is true.
    - p-values resulting from the tests for which the null hypothesis is false.
- relative size of these two components depends on the fraction of true nulls and true alternatives
    - can often be visually estimated from the histogram
        - peak near 0 for false null hypotheses
        - uniform distribution for true null hypotheses

Controlling false discovery rate

p-value histogram

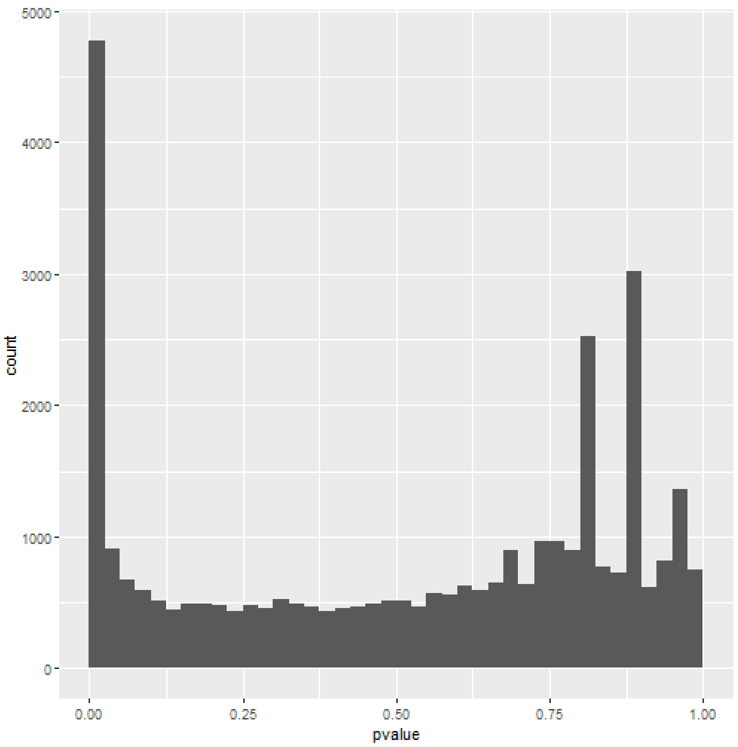


Fig. 3: p-value histogram for airway data

Controlling false discovery rate

p-value histogram

- $\alpha = 0.025$
- Given the observed distribution of p-values, 945 of the 33469 p-values are likely to correspond to true null hypotheses
- There are 4772 p-values $< \alpha$
- Estimated proportion of false rejections: 0.198

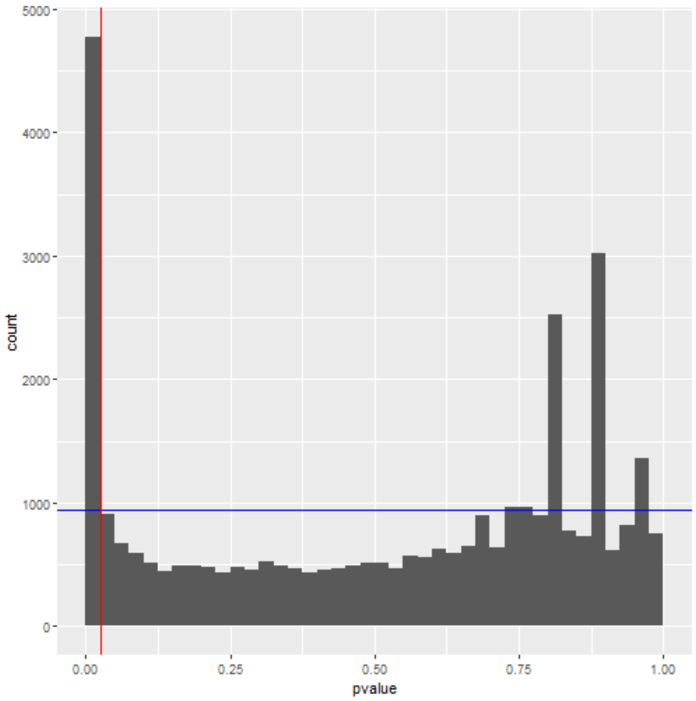


Fig. 4: Visual estimation of the FDR with the p-value histogram.

Controlling family false discovery rate

Benjamini-Hochberg algorithm

- First, order the p-values in increasing order, $p_{(1)} \ldots p_{(m)}$
- Then for some choice of φ (target FDR), find the largest value of k that satisfies: $p_{(k)} < \varphi k/m$
- Finally reject the hypotheses $1, \ldots, k$ corresponding to $p_{(1)} \ldots p_{(k)}$

Controlling false discovery rate

Benjamini-Hochberg algorithm

- Choose e. g. $\varphi = 0.05$
- Number of rejected null hypotheses: 3467
- Boundary for p-value: 0.0052
- For comparison Bonferroni boundary for $\alpha = 0.05$: 0.00000149

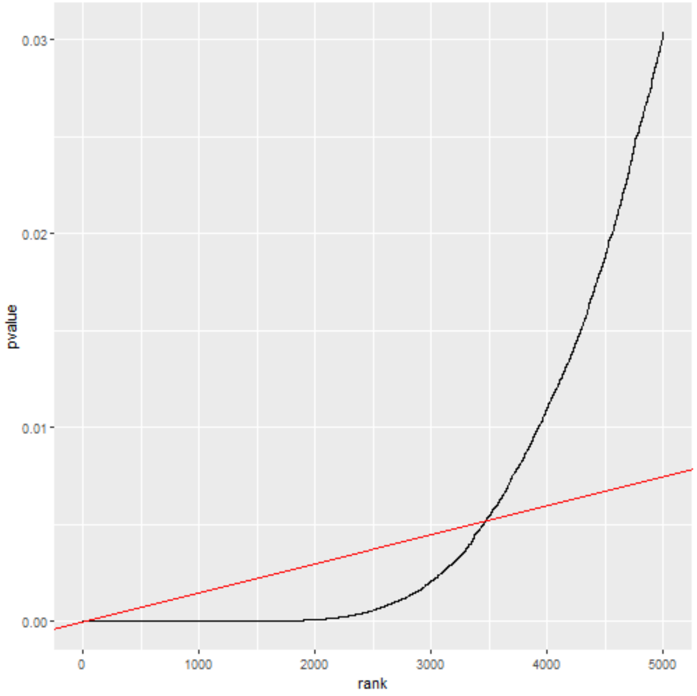


Fig. 5: Visualization of the Benjamini-Hochberg procedure.

Some extensions

- Take dependence between hypotheses into account
- Weigh p-values
- Sophisticated procedures for moderate number of hypotheses in clinical trials
    - gate-keeping
    - repeated significance testing during study
        - interim analyes
        - group sequential trials

Linear model selection and regularization or How to obtain parsimonious statistical models

Outline

- Based on Chapters 3 and 6, ISLR

Multiple linear regression

Subset Selection

- Best Subset Selection
- Stepwise Selection (Forward, Backward, Hybrid)
- Chossing Optimal Model

Shrinkage

- Ridge Regression
- The Lasso
- Selecting Turning Parameter

Linear regression

- Linear regression is commonly used to describe the relationship between
  - a quantitative response Y and
  - a set of predictor variables $X_1, X_2, \dots, X_p$
- Model is written as
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$
  where ε is usually assumed to be a normally distributed error:
  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Standard method to determine regression coefficients $\beta_1, \beta_2, \dots, \beta_p$ by minimizing least squares criterion, i. e. the squared residuals
  $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p)^2$
  where $n =$ number of observations
- Least squares estimates $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ are unbiased if the model is specified correctly

Evaluating a linear regression model

1. Is there any relationship between response and predictors?

Test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

versus the alternative $H_1 : \beta_j \neq 0$ for at least one $j \epsilon \{1, \ldots, p\}$

by computing the F-statistic

$$F = \frac{(\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2)/p}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-p-1)}$$

where

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_p x_p$
- $\sum_{i=1}^{n} (y_i - \bar{y})^2 =$ total sum of squares (about the mean)
- $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$ residual sum of squares (RSS)

Evaluating a linear regression model

2. How well does the model fit the data

Common measures for model fit

- Explained variance $R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
  - R^2 close to 1 indicates that large proportion of variance ist explained by model
  - increases when variables are added
- Residual standard error $RSE = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
  - decreases when variables are added that substantially reduce RSS
  - may increase when decrease due to addes variable is small compared to decrease in numerator
- Other criteria , e. g. AIC, BIC, ... (more details later)

Evaluating a linear regression model

3. Are all predictors necesessary?

Forward stepwise selection

- Begins with a model containing no predictors
- Adds the predictor that gives the greatest improvement to the model
- Adds further predictors until all predictors are added
- Of all models created, the “best” is chosen

Backward stepwise selection

- A model is built including all predictors
- At each step, the least-predictive is removed
- Of each of the models produced by each step, the best model is selected
- Cannot be used when n < p

Linear model selection and regularization

Motivation

Improve Accuracy

- Least-squares is ideal where $n >> p$
- Not as good if $n > p$
- Linear equations cannot be solved if $n < p$

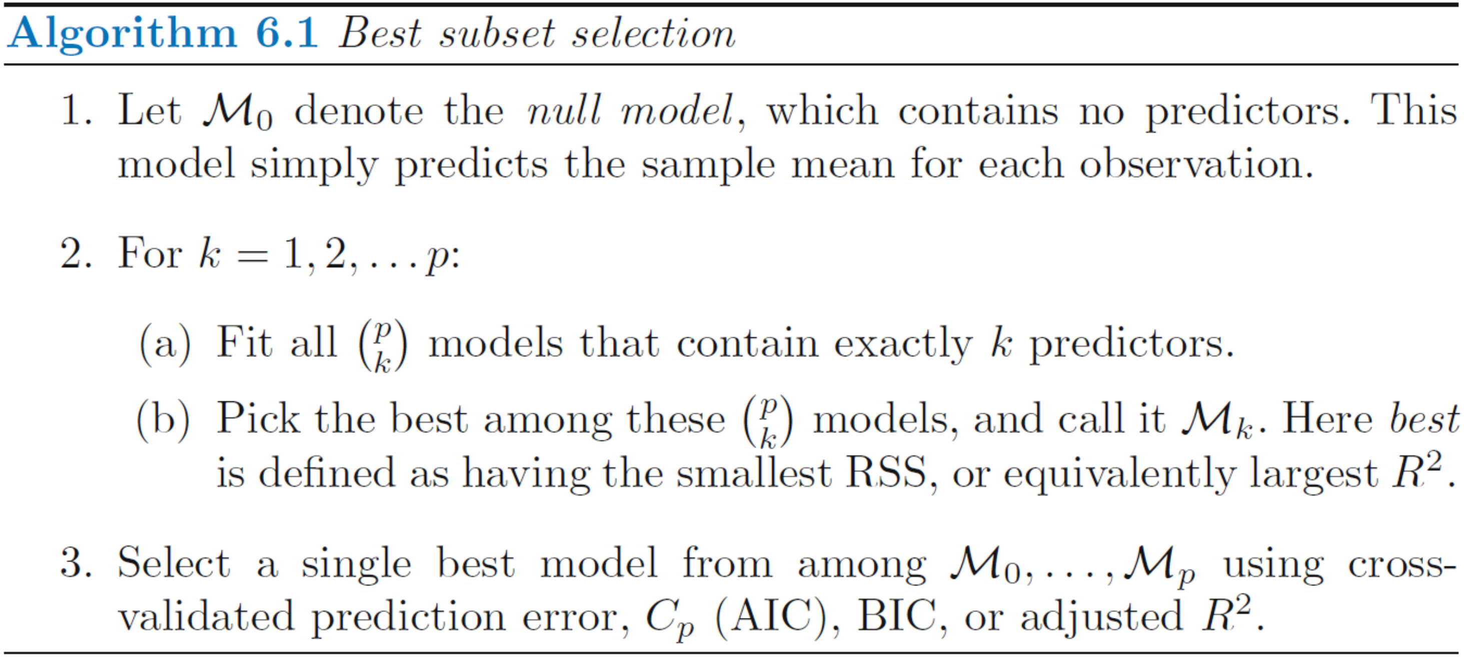
Improve Interpretability

- Remove irrelevant predictors

Subset Selection: Best Subset Selection

- Try all possible combinations of p predictors and choose the best one
- Advantages: Exhaustive & simple
- Disadvantages: Computationally intensive
  - 2^p possible models must be evaluated
  - "becomes computationally infeasible for values of p greater than around 40"

Subset Selection: Best Subset Selection



Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

    (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

    (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

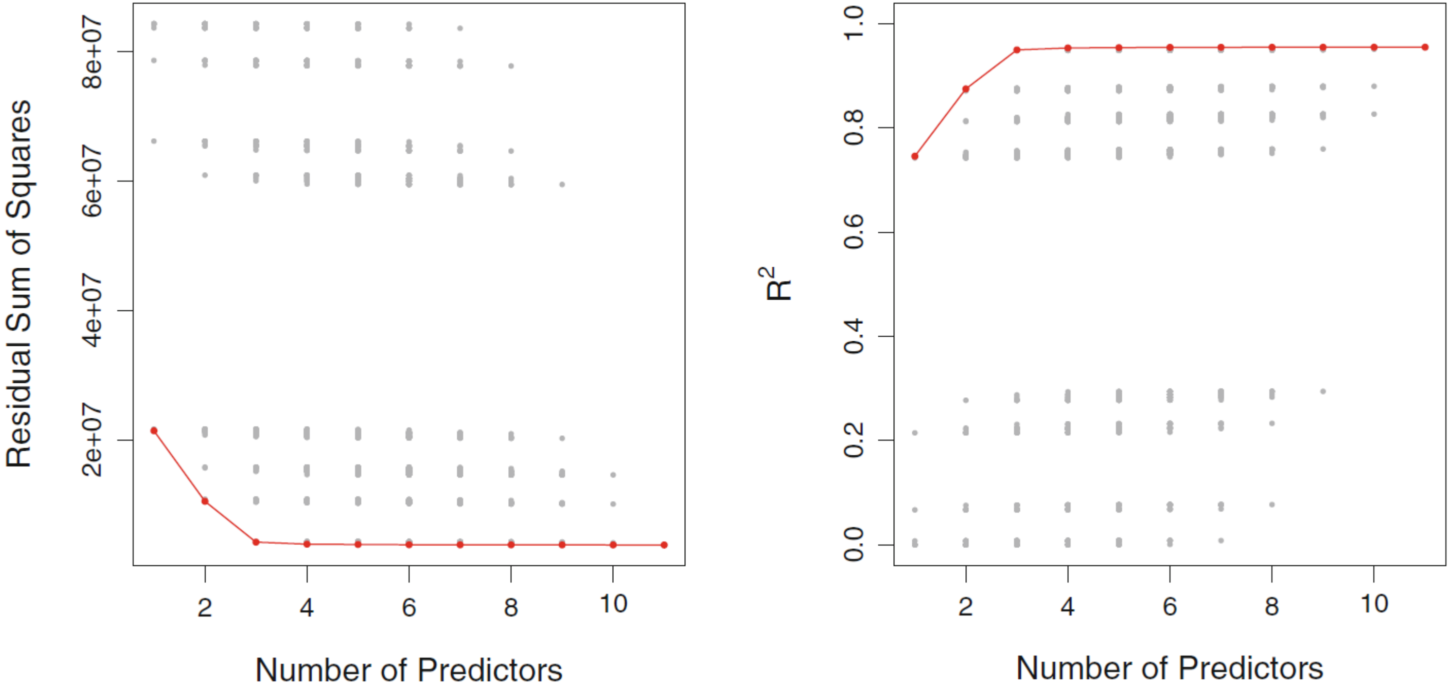
Data in regression examples

- Credit data set from `ISLR2` contains information about credit card holders and credit card debt
- Response variable is balance (average credit card debt for each individual)
- Several quantitative predictors:
    - age
    - cards (number of credit cards)
    - education (years of education)
    - income (in thousands of dollars)
    - limit (credit limit)
    - rating (credit rating)

Table 3: First lines of Credit data

Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
14.891	3606	283	2	34	11	No	No	Yes	South	333
106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
104.593	7075	514	4	71	11	No	No	No	West	580
148.924	9504	681	3	36	11	Yes	No	No	West	964
55.882	4897	357	2	68	16	No	No	Yes	South	331
80.180	8047	569	4	77	10	No	No	No	South	1151

Subset Selection: Best Subset Selection



Each possible model with all predictors of `Credit` data set. Red frontier tracks the best model for a given number of predictors, according to RSS and R^2

Subset Selection: Stepwise Selection

Stepwise methods explore a more restricted set of models, reducing overfitting and reducing time to select/fit the model.

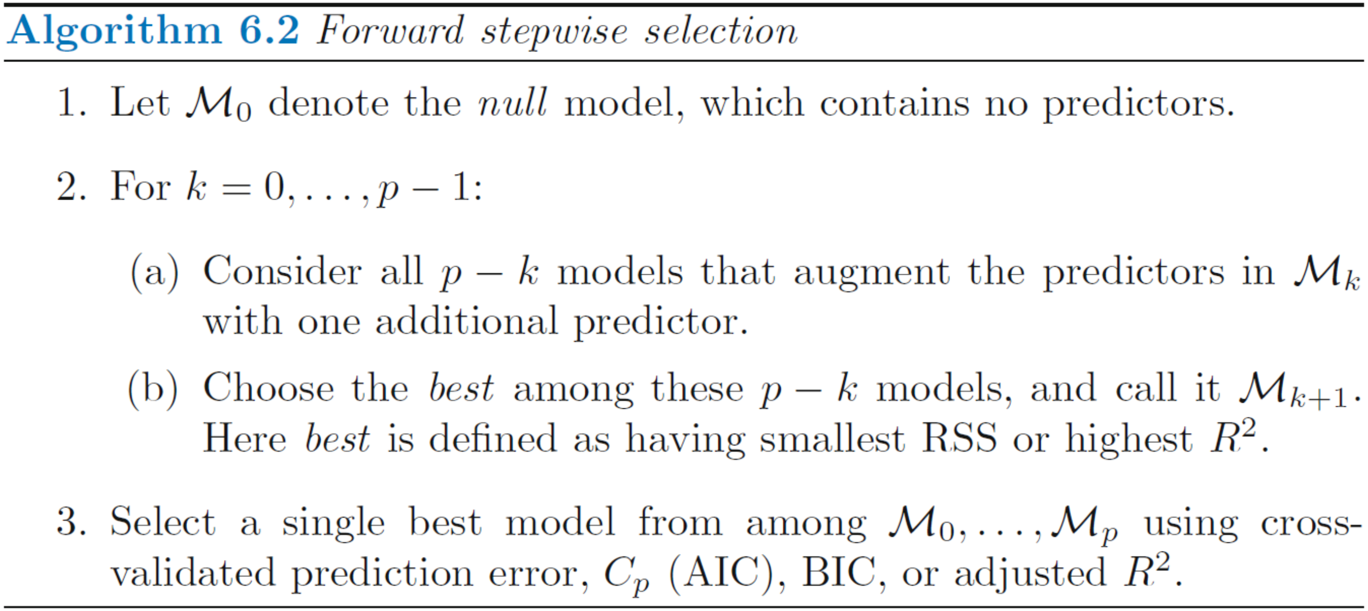
Three types:

- Forward Stepwise
- Backward Stepwise
- Hybrid Approaches

Subset Selection: Stepwise Selection

Forward Stepwise

- Begins with a model containing no predictors
- Adds the predictor that gives the greatest improvement to the model
- Adds further predictors until all predictors are added
- Of all models created, the “best” is chosen



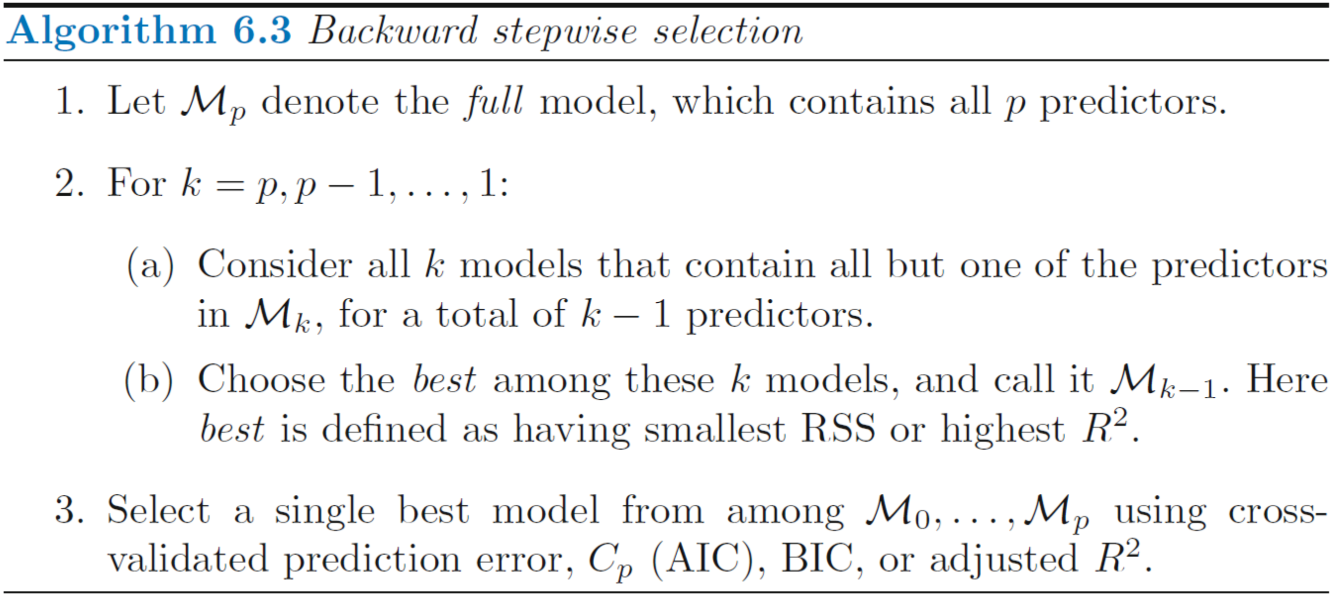
Algorithm 6.2 *Forward stepwise selection*

```
1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p - 1$:
    (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$
        with one additional predictor.
    (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$.
        Here *best* is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

Subset Selection: Stepwise Selection

Backward Stepwise

- A model is built including all predictors
- At each step, the least-predictive is removed
- Of each of the models produced by each step, the best model is selected
- Cannot be used when $n < p$



Algorithm 6.3 *Backward stepwise selection*

```
1. Let $\mathcal{M}_p$ denote the *full* model, which contains all $p$ predictors.
2. For $k = p, p-1, \ldots, 1$:
    (a) Consider all $k$ models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
    (b) Choose the *best* among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

Subset Selection: Stepwise Selection

Hybrid approaches

- Hybrid combine both forward and backward selection.
- These models begin with a null model and add predictors like forward selection.
- At each step, they also remove predictors that are less-informative, like backward selection.

Stepwise Selection vs Best Subset Selection

Stepwise Selection:

- Faster than best subset selection
- Tractable for problems with $p > 40$

Best Subset Selection

- Guaranteed to find the best possible model

Subset Selection: Optimal Model

“The model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.”

We wish to choose a model with a low test error.

Estimating test error:

- Adjust the training error to account for bias
- Directly estimate with cross-validation or a validation set

Subset Selection: Optimal Model

Adjusting with C_p

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$$

- For least-squares models with d predictors
- An unbiased estimate of MSE, if $\hat{\sigma}^2$ is an unbiased estimate of σ^2
- The penalty increases as the number of predictors in the model increases
- Choose the model with the lowest C_p value

Subset Selection: Optimal Model

Adjusting with Akaike Information Criterion AIC

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)$$

- For models fit with maximum likelihood
- Omitted a constant: Proportional to C_p

Subset Selection: Optimal Model

Adjusting with Bayesian Information Criterion BIC

$$BIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + log(n)d\hat{\sigma}^2\right)$$

- For models fit with maximum likelihood
- Omitted an additive constant
- Heavier penalty on the number of predictors than C_p

Subset Selection: Optimal Model

Adjusting with Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

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

- Regular R^2 always increases with added predictors.
- The Adjusted R^2, is corrected for the number of predictors d, such that it may decrease as additional, less-informative predictors are added to the model.
- A large value of Adjusted R^2 indicates a model with low test error.

Subset Selection: Optimal Model

Comparison: C_p vs BIC vs R^2

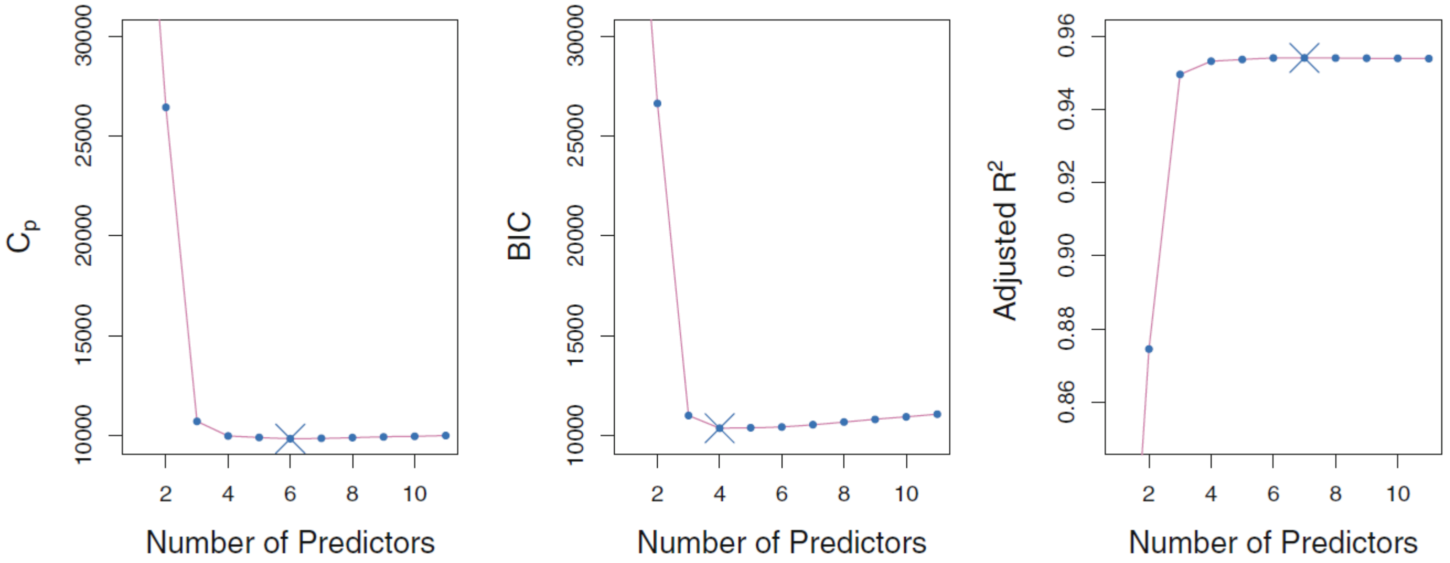


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

Subset Selection: Optimal Model

Estimating with Validation or Cross-Validation

- Compute validation set error or cross-validation error for each model
- Select model with smallest test error
- Directly estimated test error based on fewer assumptions

Subset Selection: Optimal Model

Comparison: Adjusting vs Estimating

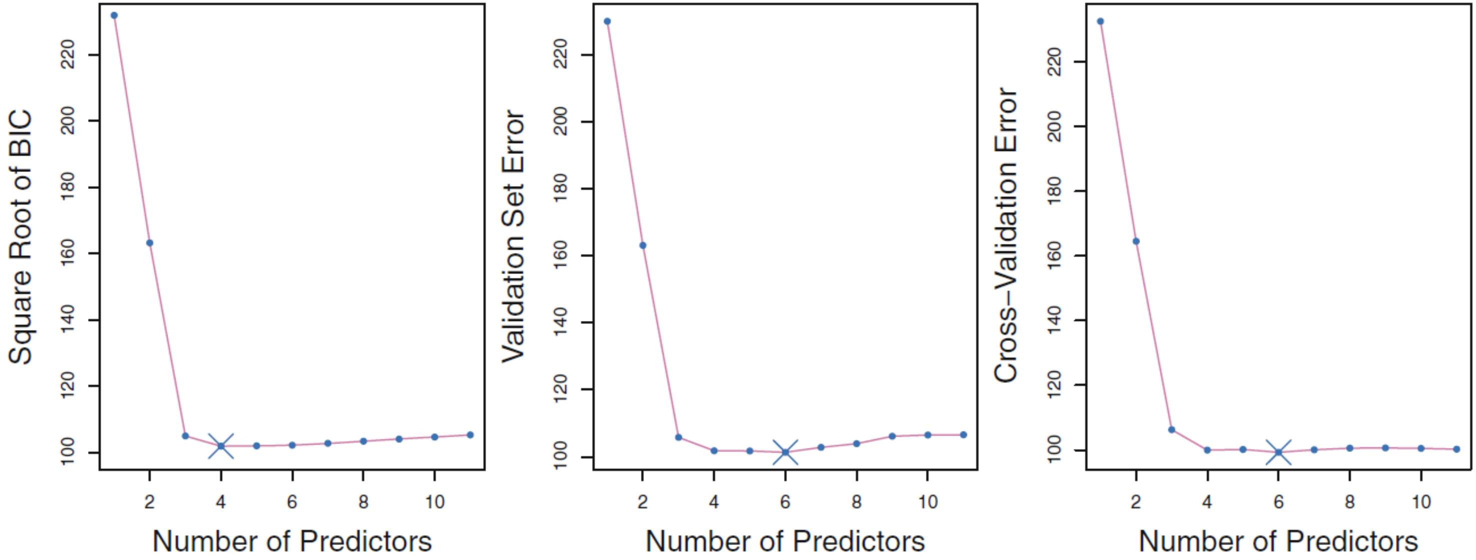


FIGURE 6.3. *For the* Credit *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11*. The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Subset Selection: Optimal Model

one-standard-error-rule

- Calculate the standard error of test MSE for each model. Select the smallest model for which the estimated test error is within one SE of the lowest point in the curve.

" The rationale here is that if a set of models appear to be more or less equally good, then we might as well choose the simplest model—that is, the model with the smallest number of predictors. "

Shrinkage methods

- Fit a model with all predictors that shrinks coefficient estimates towards zero
- Shrinking coefficient estimates can significantly reduce their variance
- Two best known shrinkage methods: ridge and lasso
- For both, ridge and lasso, predictors should be standardized, i.e.
    - substract mean
    - divide by standard deviation

Shrinkage methods: Ridge

- Very similar to least squares in that both methods select coefficients that reduce RSS

- Coefficients are estimated by minimizing slightly different quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

Shrinkage methods: Ridge

$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$

Shrinkage penalty

- Is small when coefficients close to zero
- Has the effect of shrinking β_j toward zero
- Only applied to coefficients, not to the intercept

Tuning parameter

- Controls impact of shrinkage penalty
- When $\lambda = 0$: Same results as least squares
- As $\lambda \to \infty$, coefficients approach zero
- Ridge offers a different set of coefficients for each value of λ
- Selecting a good value for λ is critical

Shrinkage methods: Ridge

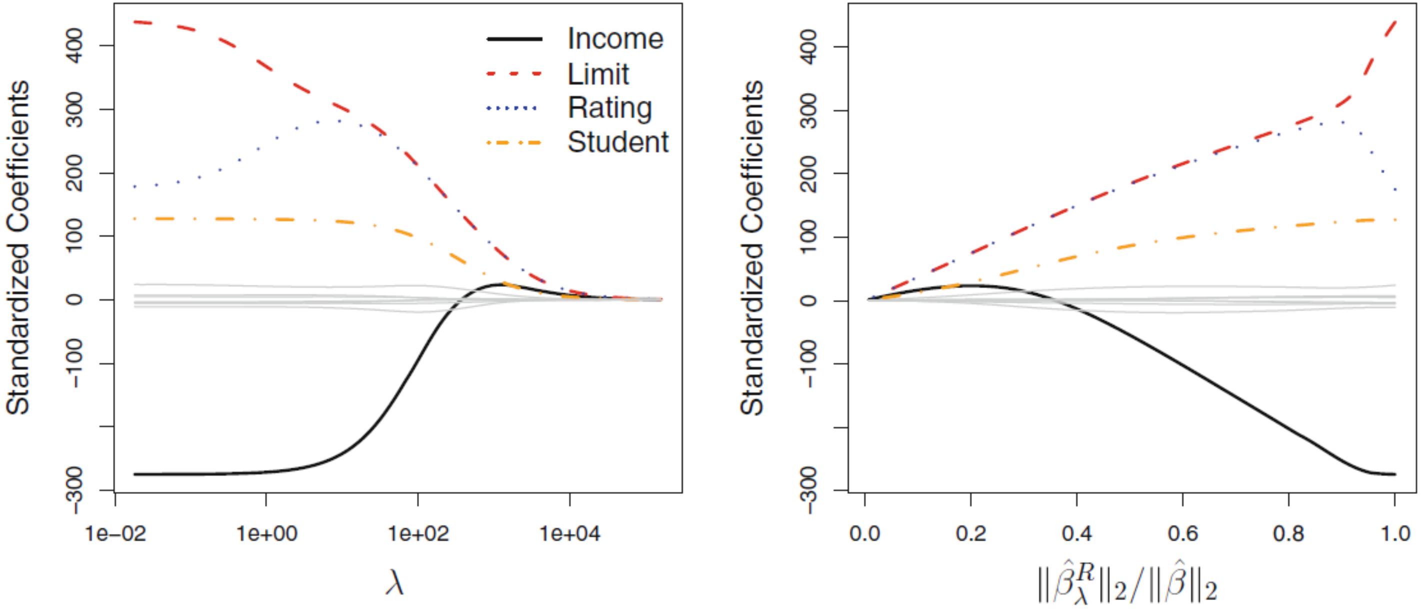


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of λ and $\|\hat\beta^R_\lambda\|_2/\|\hat\beta\|_2$.*

Shrinkage methods: Ridge

- Unlike least squares, ridge is very scale dependent
- Therefore must standardize predictor variables
- The following formula will ensure all predictor variables have a standard deviation of one

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

- And this will also center the variables

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

Shrinkage methods: Ridge

- Advantage of ridge is rooted in the bias-variance trade-off
- As λ increases, bias increases, but variance decreases

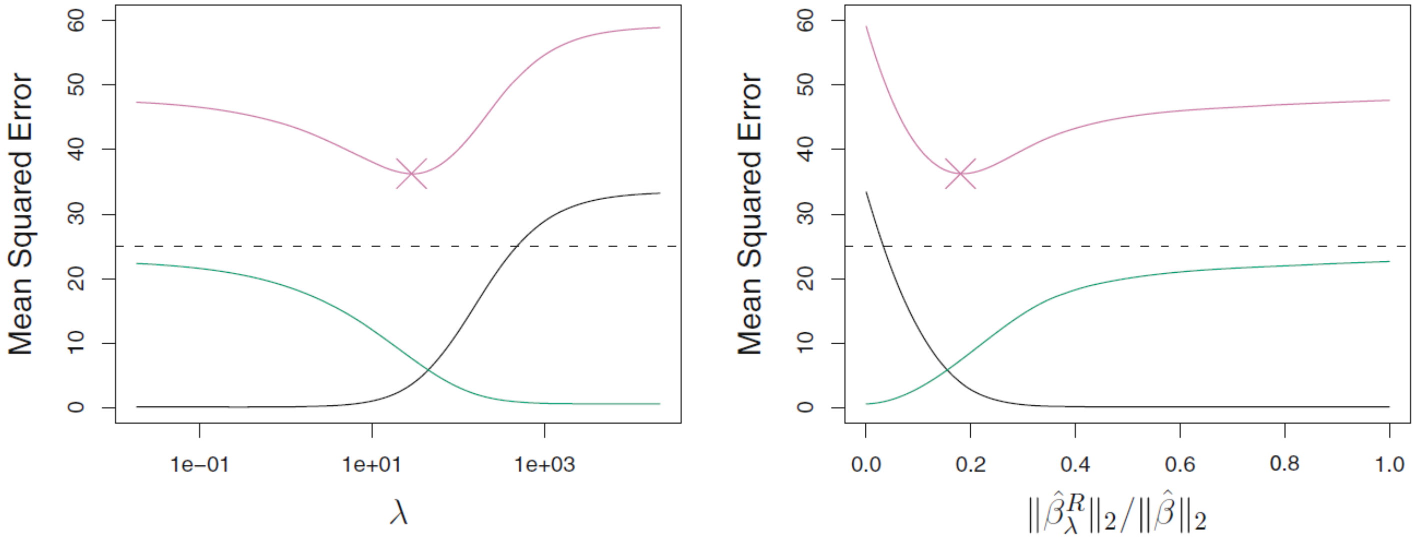


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Shrinkage methods: The Lasso

- Similar to Ridge, but with $|\beta_j|$, which forces some coefficients to be zero: Performs variable selection
- Creates models that are easier to interpret
- Shrinks coefficient estimates towards zero

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Shrinkage methods: The Lasso

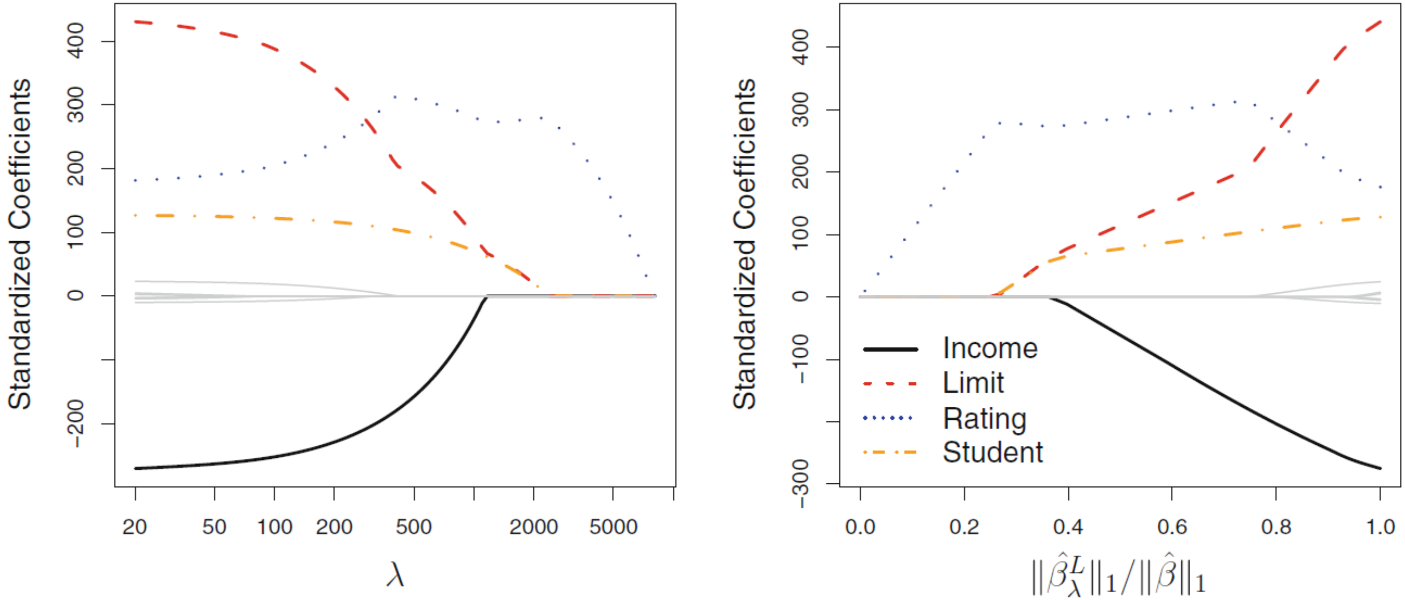


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Shrinkage methods: Alternative Formulation

Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- We are trying to find the set of estimates that lead to the smallest RSS, subject to the constraint that there is a budget s
- If s is very large, it yields least squares solution

Shrinkage methods: Graphical intuition

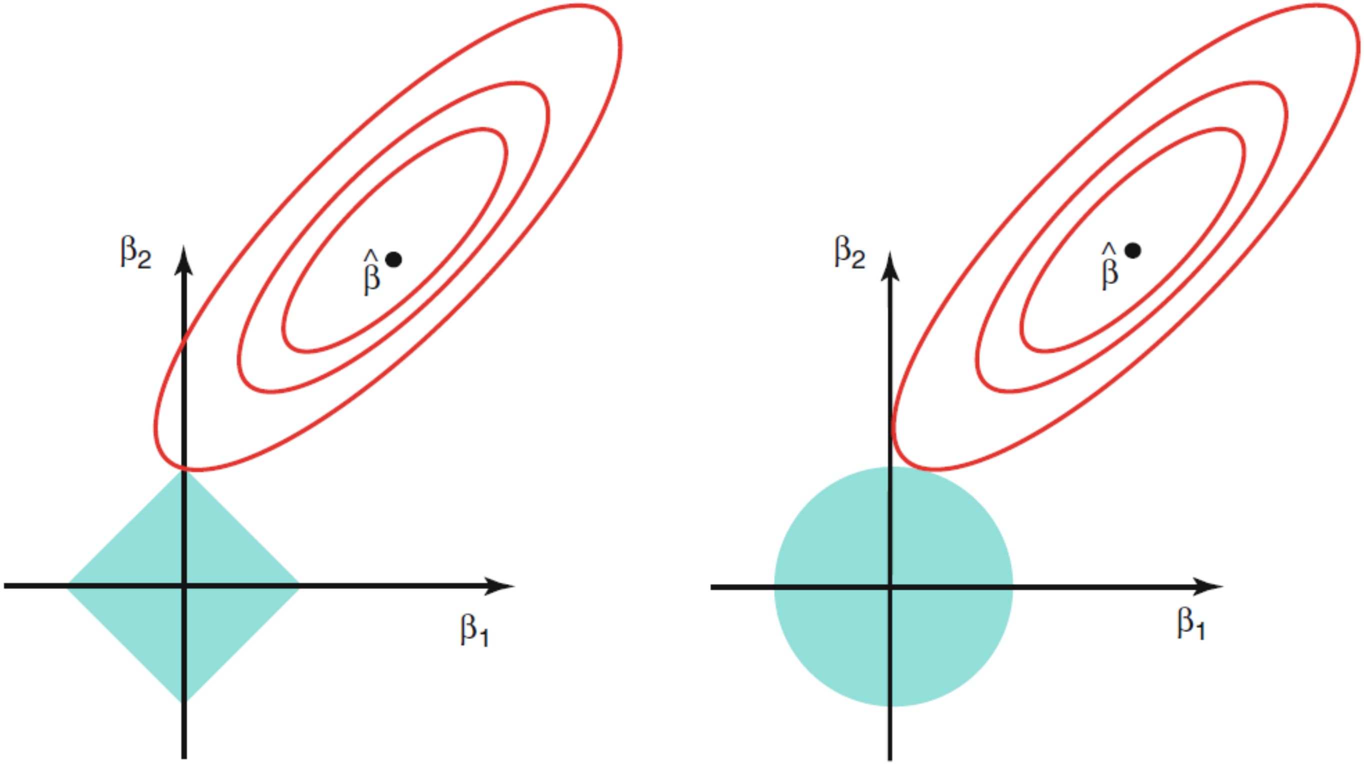


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$*, while the red ellipses are the contours of the RSS.*

Shrinkage methods: Ridge vs Lasso

- Qualitatively, both give very similar results. For both, as λ increases, variance decreases and bias increases
- If all predictors associated with outcome, ridge slightly outperforms lasso
- When not all predictors associated with outcome or when some predictors have very large coefficients, lasso slightly outperforms ridge
- Ridge regression more or less shrinks every dimension of the data by the same proportion, whereas the lasso more or less shrinks all coefficients toward or to zero by a similar amount
- Biggest advantage of lasso is variable selection, making model interpretation easier
- Use cross-validation to determine which technique is better for a particular dataset

Shrinkage methods: Bayesian point of view

- In Bayesian theory, we assume β has a prior distribution: $p(\beta)$ Multiplying that prior by the likelihood gives us the posterior distribution.
- If $p(\beta)$ follows a Gaussian distribution with mean 0 and SD that is a function of λ then the most likely posterior value for β is the ridge regression solution
- If $p(\beta)$ follows a Laplace distribution with mean 0 and a scale parameter of λ then the most likely posterior value for β is the lasso regression solution

Shrinkage methods: Selecting λ

- Cross validation is a simple way to choose the best λ
  - Choose a grid of λ values and compute cross-validation error for each value of λ
  - Choose λ for which error is smallest

How to extract essential information from data: Principal component analysis

PCA

Aim: Dimension reduction

- Reduce a two-variable scatterplot to a single coordinate (Karl Pearson 1901)
- Summarize a battery of psychological tests run on the same subjects - provide overall scores (Hotelling 1933)
- PCA is an exploratory technique to show relations between variables
- PCA is called an unsupervised learning technique
    - all variables have same status
    - no distinction between dependent and independent variables

PCA

Aim: Dimension reduction

- Project points in high-dimensional space to lower dimensions ("hyperplanes")
    - i-th principal component is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vectors
    - best-fitting line minimizes average squared distance from the points to the line
    - these directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated
- Linear technique, i. e. new variables are linear functions (linear combinations) of the old ones

PCA

Example - Turtle data

- Jolicoeur and Mossiman's 1960's Painted Turtles Dataset with size variables for two turtle populations (contained in R package ade4)

Table 4: First lines of turtle data

length	width	height	sex
93	74	37	M
94	78	35	M
96	80	35	M
101	84	39	M
102	85	38	M
103	81	37	M

PCA

Example - Turtle data

Table 5: Summary statistics on turtle data

	length	width	height
Min.	93.00	74.00	35.00
1st Qu.	106.75	86.00	40.00
Median	122.00	93.00	44.00
Mean	124.69	95.50	46.15
3rd Qu.	136.25	102.75	51.00
Max.	177.00	132.00	67.00

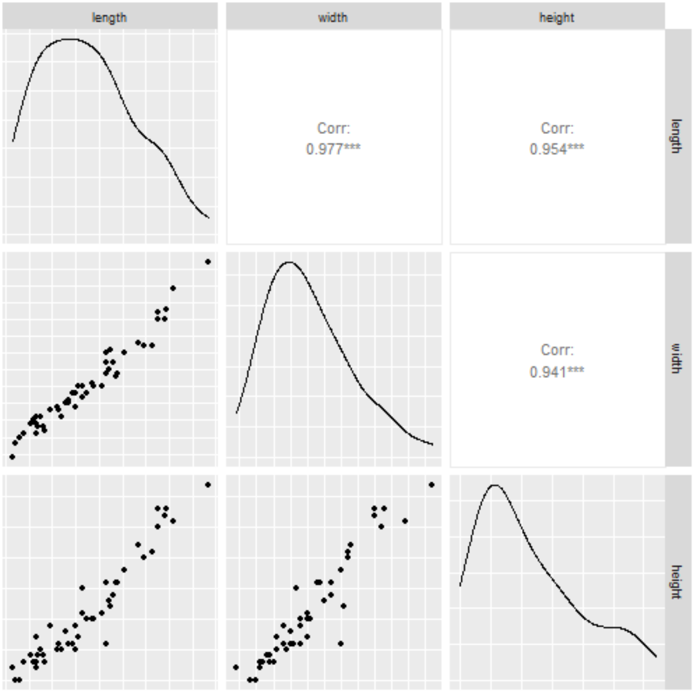


Fig. 6: All pairs of bivariate scatterplots for the three biometric measurements on painted turtles

PCA

Example - Turtle data

- Data need to be standardized to make variables in a way comparable
  - center: substract mean - new variables have mean 0
  - scale: divide by standard deviation - new variables have standard deviation 1

PCA

Summarize two-dimensional data by a line

- Aim: keep as much information as possible about both variables
- Linear regression of y on x

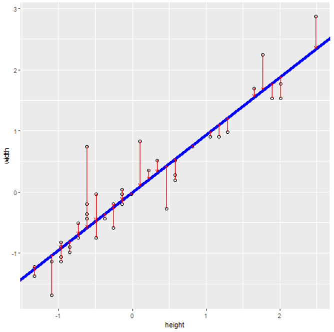


Fig. 8: Regression of width on height

  - Linear regression of x on y

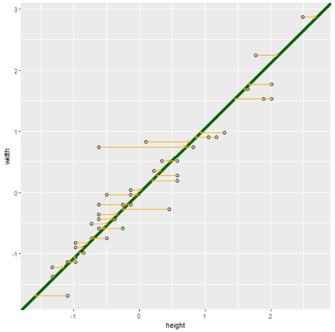


Fig. 9: Regression of height on width

A line that minimizes distances in both directions

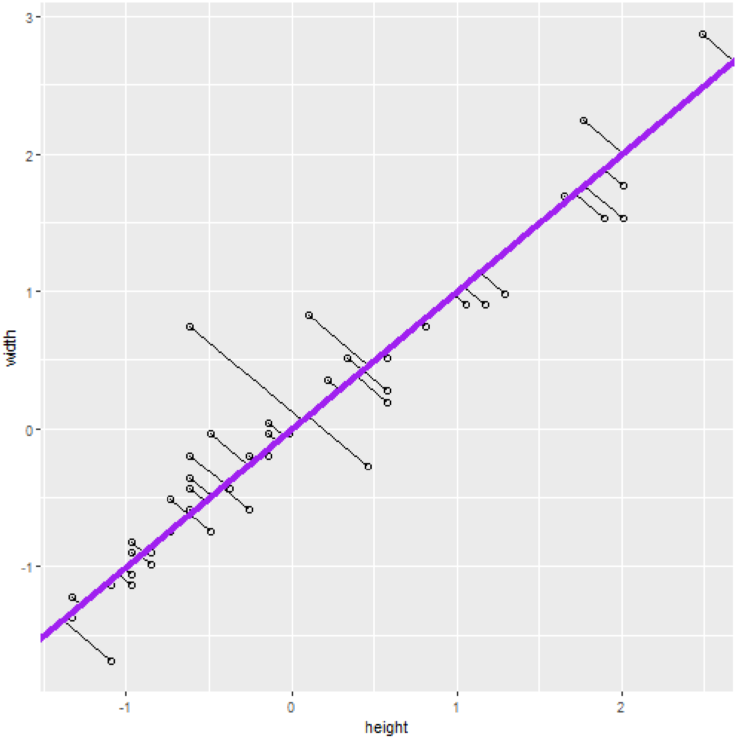


Fig. 10: A line that minimizes distances in both directions - first principal component

PCA

- Minimizing in both horizontal and vertical directions means in fact minimizing the orthogonal projections onto the line from each point.
- Total variability of the points is measured by the sum of squares of the projection of the points onto the center of gravity, i. e. the origin (0,0) - if the data are centered.
- Total variability = the total variance = inertia of the point cloud.
- Inertia can be decomposed into
  - sum of the squares of the projections onto the line plus
  - the variances along that line
- For a fixed variance, minimizing the projection distances $\longrightarrow$ maximizing the variance along that line
- Often define first principal component as the line with maximum variance

PCA

- Note, technically, a singular Value decomposition is performed, i. e. we write $m \times n$ data matrix with rank r as

- $M = U\Sigma V'$ where

  - U an $m \times m$ orthonormal matrix
  - V' an $n \times n$ transposed of an orthonormal matrix V

  - Σ an $m \times n$ matrix with rank r of the form

$$\Sigma = \left(\begin{array}{ccc|ccc}
\sigma_1 & & & & \vdots & \\
& \ddots & & \cdots & 0 & \cdots \\
& & \sigma_r & & \vdots & \\
\hline
& \vdots & & & \vdots & \\
\cdots & 0 & \cdots & \cdots & 0 & \cdots \\
& \vdots & & & \vdots &
\end{array}\right)$$

with $\sigma_1 \geq \cdots \geq \sigma_r > 0$

Singular value decomposition

turtle data

Table 6: Diagonal elements 'singular values'

x
11.704395
1.733125
1.001710

Table 7: Components of V

0.5806536	-0.2706983	0.7678306
0.5780575	-0.5270479	-0.6229526
0.5733158	0.8055699	-0.1495531

PCA results for turtle data

- First column of turtles.svd$v shows: coefficients for the three variables are practically equal
- These are the coeffcients of the first principal component:
- $Z_1 = 0.581 \cdot \text{length} + 0.578 \cdot \text{width} + 0.573 \cdot \text{height}$
- Variance of principal components
  - $\text{var}(Z_1) = 2.915$
  - $\text{var}(Z_2) = 0.064$
  - $\text{var}(Z_3) = 0.021$

PCA Extensions

See Federico's package `pcaexplorer`

References

- Susan Holmes, Wolfgang Huber. Modern Statistics for Modern Biology. Cambridge University Press, 2019. The book is also online: https://www.huber.embl.de/msmb/

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning. Second edition. Springer New York, 2021. PDF (legally) available at https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

- Trevor Hastie, Robert Tibshirani. The Elements of Statistical Learning. pringer New York, 2009 PDF (legally) available at https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf

- Frank E. Harrell. Regression Modeling Strategies. Second edition. Springer Cham 2015.

- https://en.wikipedia.org/wiki/Statistical_hypothesis_testing (especially for a comparison between Fisherian, frequentist (Neyman–Pearson) interpretation)

- https://xkcd.com/882/