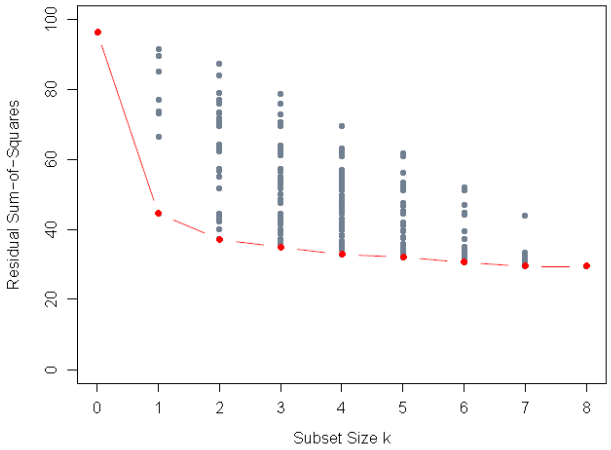
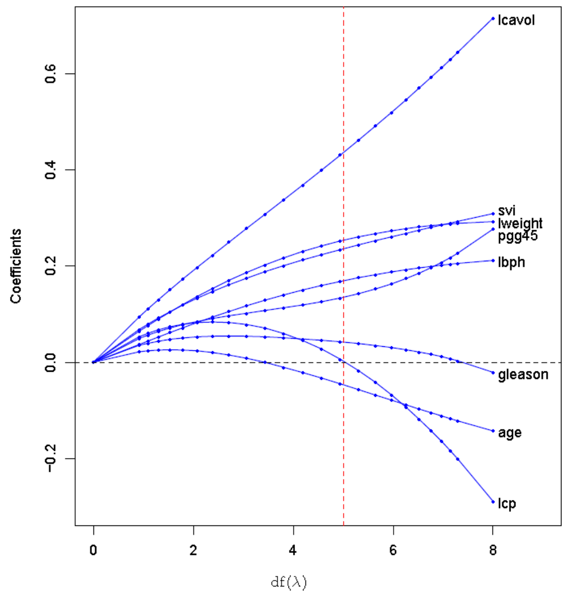
Lineární regrese, výběr atributů, regularizace

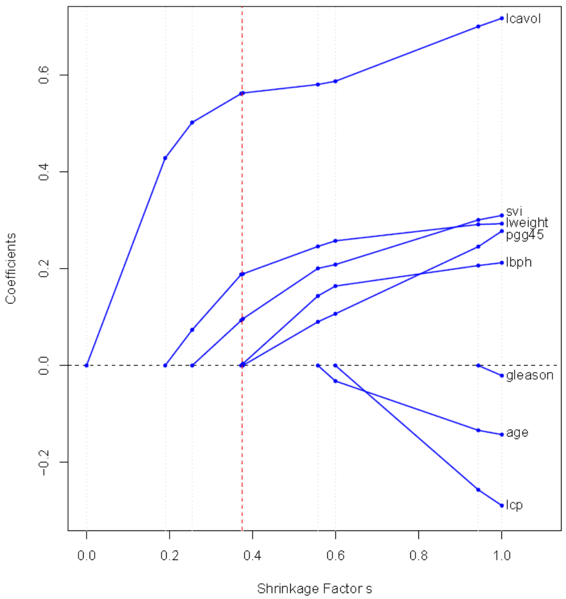
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



- AIC, BIC, krosvalidace

- exercises.RData
- toclustf.RData





Korelovaná pozorování (rezidua)

- např. u časové řady
- zpravidla podhodnocuje odhad chyby.

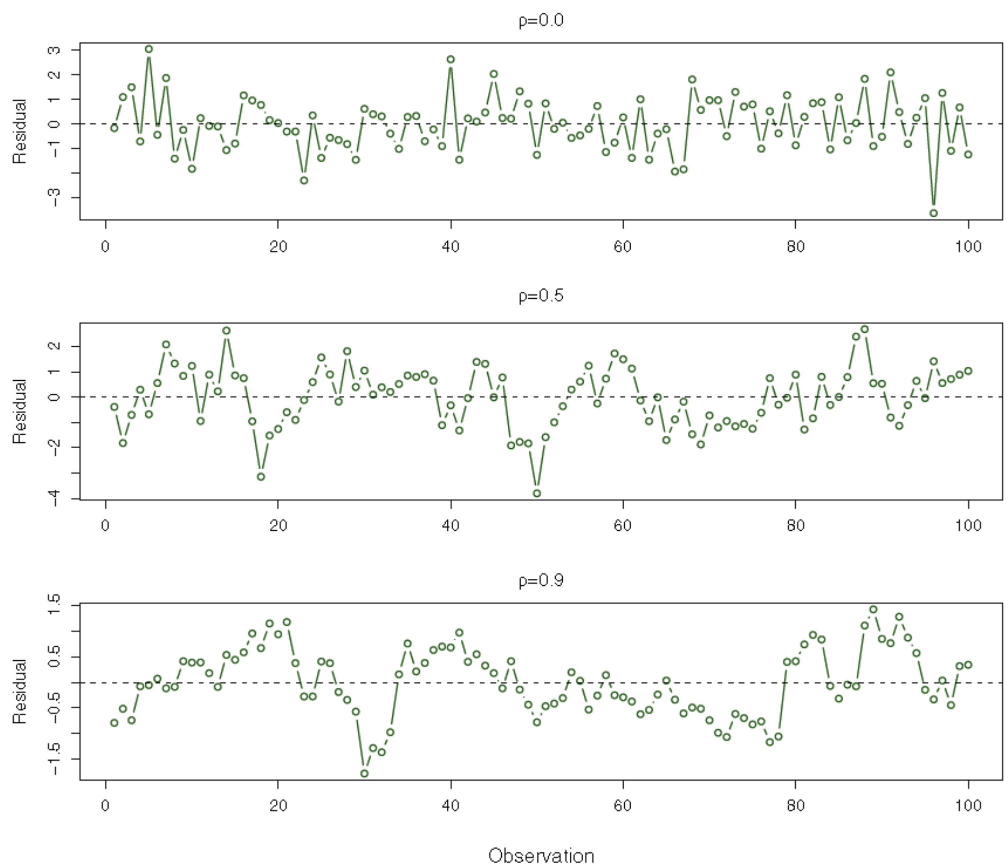


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Nekonstantní rozptyl reziduí

- log transformace, vážené nejmenší čtverce

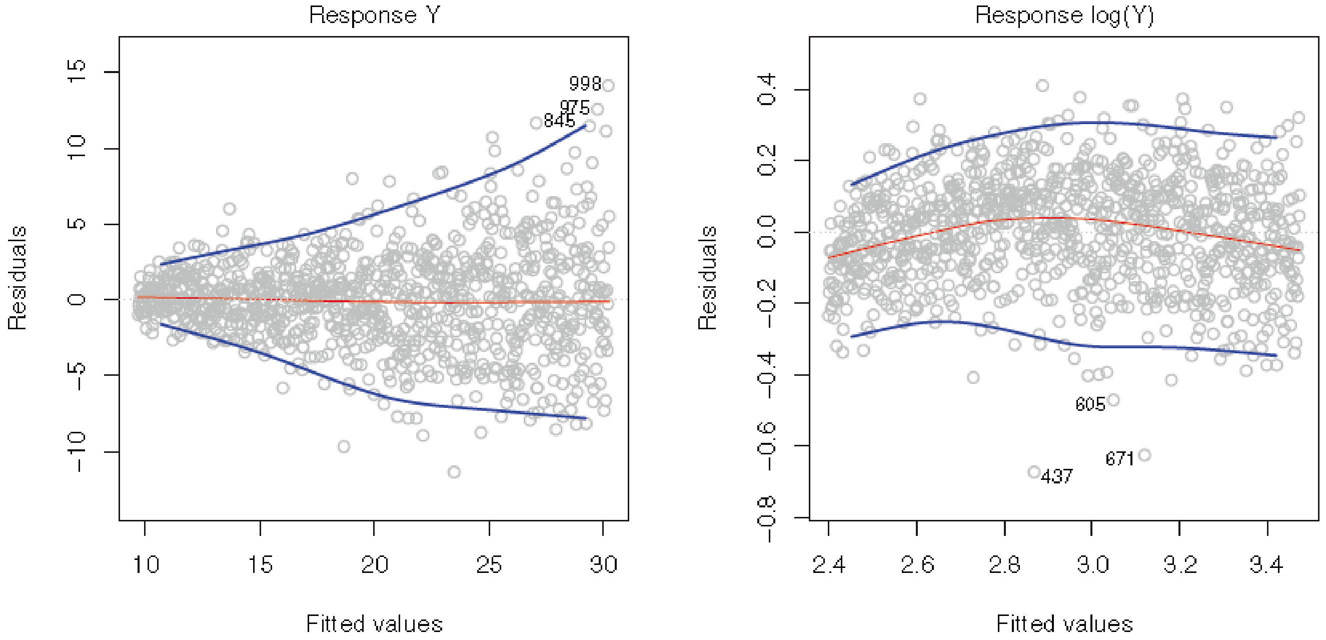


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.*

Rezidua „nerovnoměrně“ - nelinearita

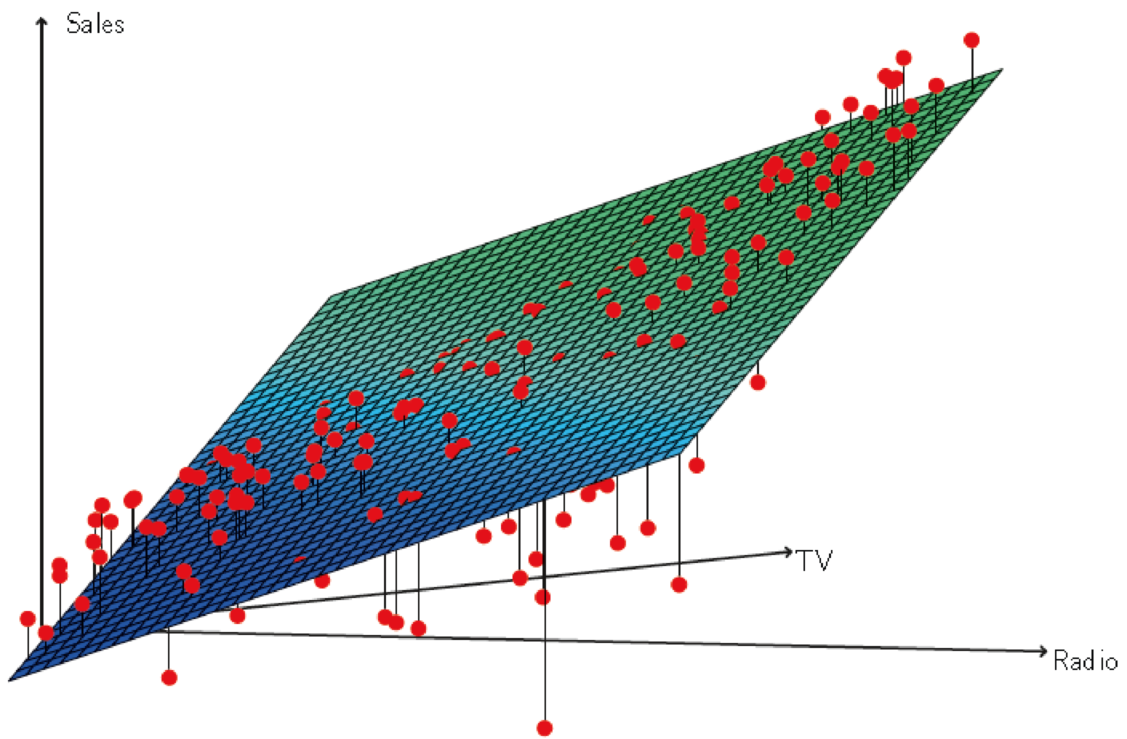
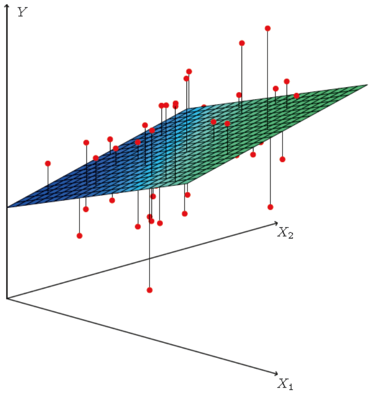


FIGURE 3.5. *For the* Advertising *data, a linear regression fit to* sales *using* TV *and* radio *as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data.*

Vícerozměrná lineární regre



- Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

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

  - p – počet vstupních proměnných
- minimalizací RSS dostaneme koeficienty β.

	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

- jednorozměrná:

	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

- Je inzerce v novinách (dle modelu) důležitá?

Kolinearita
v extrému vede k neinvertibilitě

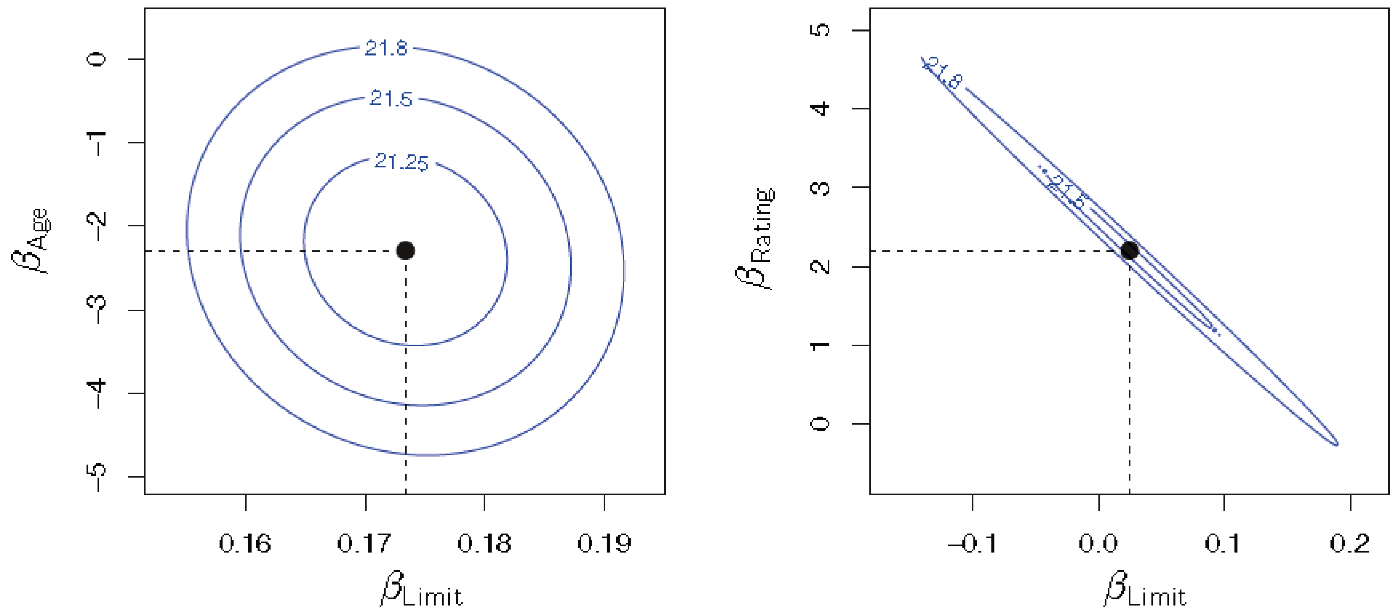


FIGURE 3.15. *Contour plots for th*
β for various regressions involving t
dots represent the coefficient values
A contour plot of RSS for the regres
minimum value is well defined. Righ
of balance *onto* rating *and* limit.
pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ *with a similar va*

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* Credit *data set are shown. Model 1 is a regression of* balance *on* age *and* limit*, and Model 2 a regression of* balance *on* rating *and* limit*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

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

Kvalitativní (diskrétní) proměnné

- Kódujeme 0/1, vícehodnotové pro každou(-1) hodnotu zvlášť.
- Př. národnost

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

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	−18.69	65.02	−0.287	0.7740
ethnicity[Caucasian]	−12.50	56.68	−0.221	0.8260

TABLE 3.8. *Least squares coefficient estimates associated with the regression of* balance *onto* ethnicity *in the* Credit *data set. The linear model is given in (3.30). That is, ethnicity is encoded via two dummy variables (3.28) and (3.29).*

Různý sklon pro třídy LR nezjistí

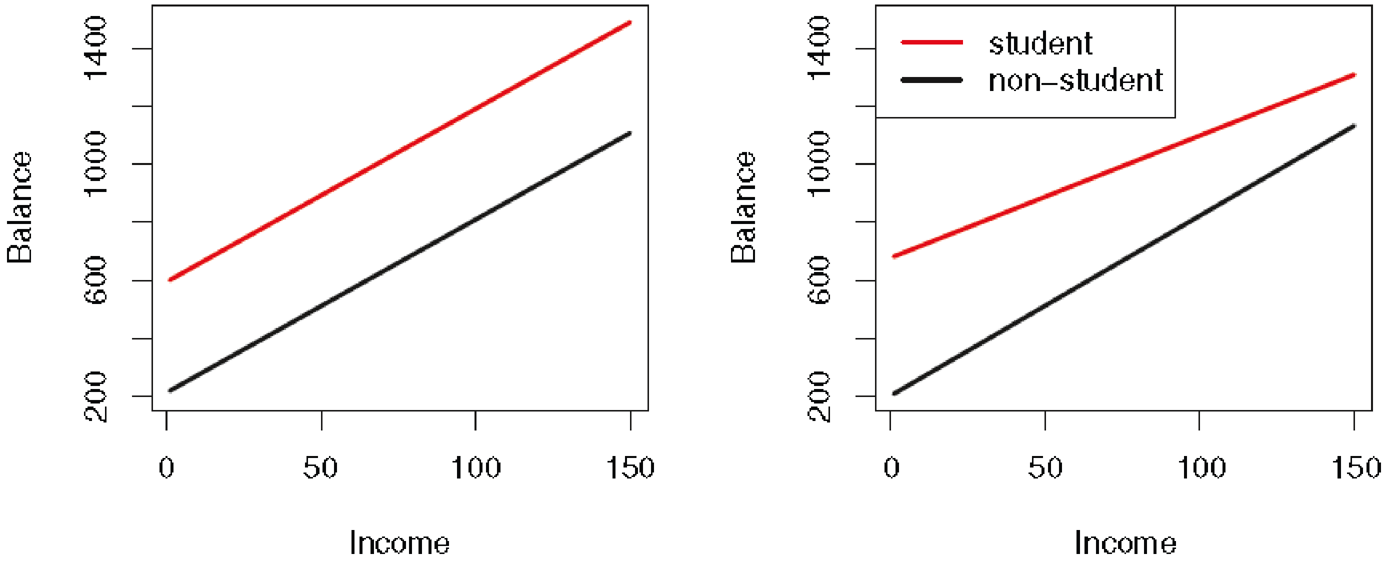


FIGURE 3.7. *For the* Credit *data, the least squares lines are shown for prediction of* balance *from* income *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* income *and* student. Right: *The model (3.35) was fit. There is an interaction term between* income *and* student.

$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases} \end{aligned}$$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not studen} \end{cases}$$

Outliers (odlehlá pozorování)

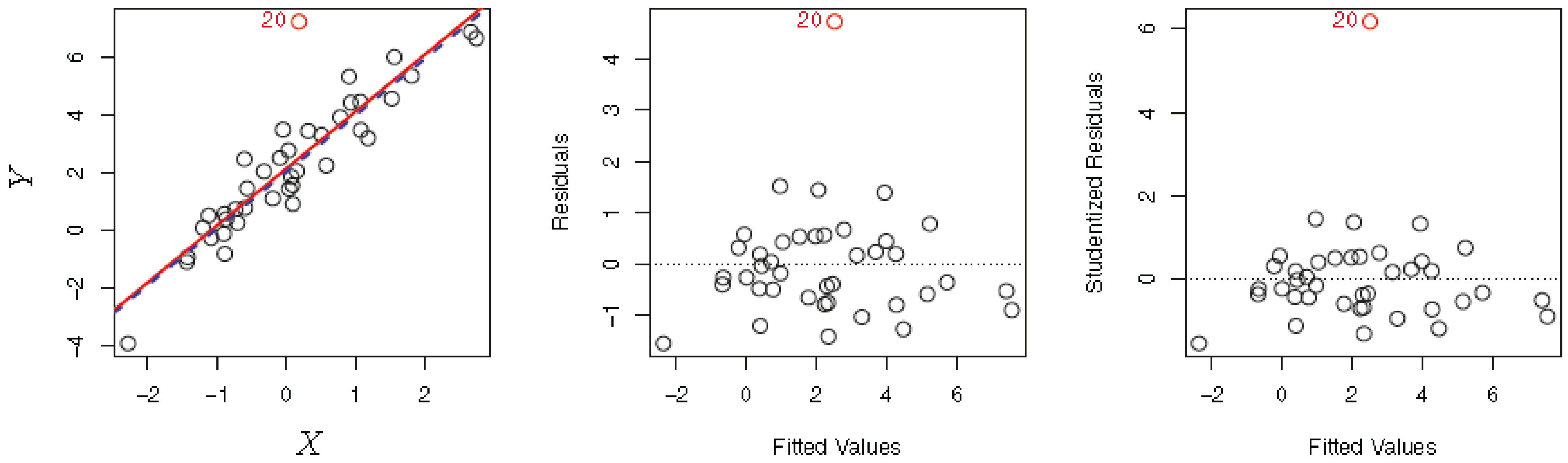


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

- Chyba v datech nebo chybějící prediktor?

High leverage – vzdálená X

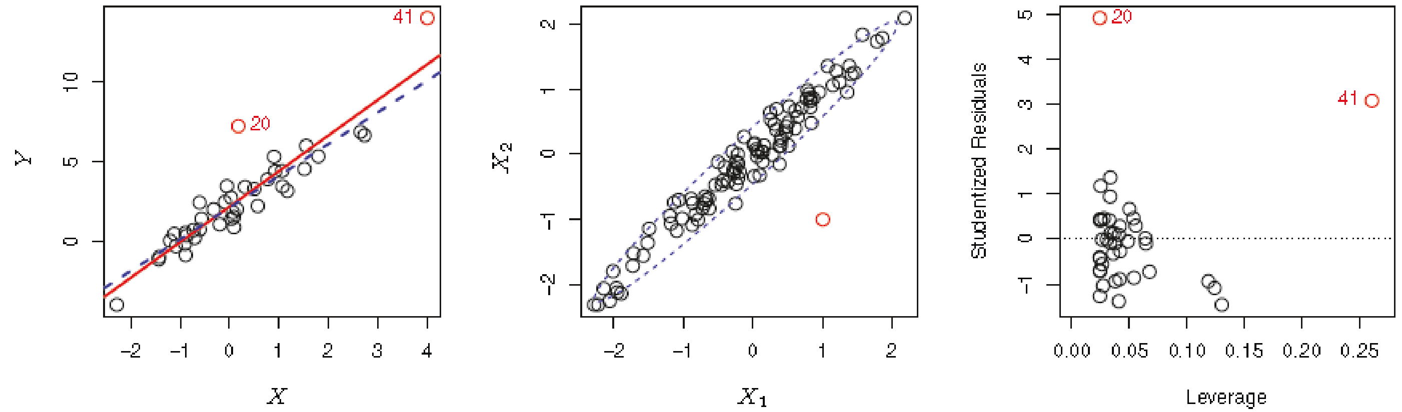


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

- leverage statistics: diagonála $H=X(X^TX)^{-1}X^T$.
- Jednorozměrně:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Why Linear Model Regularization?

- Linear models are simple, BUT
- consider p>>N,
  - we have more features than data records
  - we can (often) learn model with 0 training error
    - even for independent features!
    - it is overfitted model.
- Less features in the model may lead to smaller test error.
- We add constrains or a penalty on coefficients.
- Model with fewer features is more interpretable.

Selection, Regularization Methods

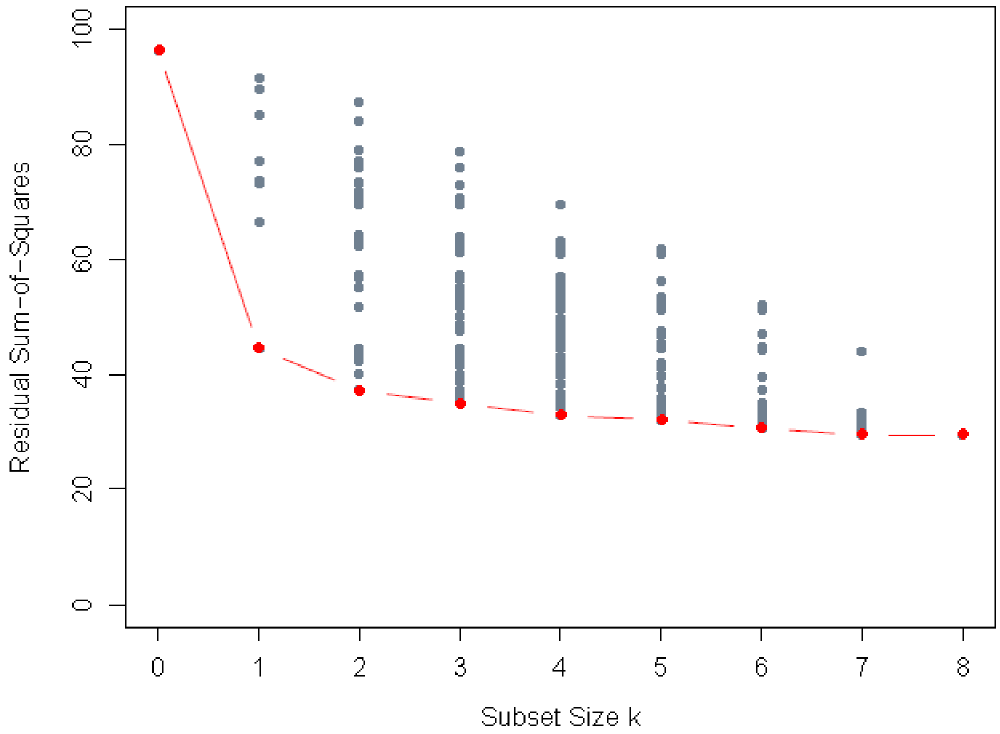
- Subset Selection
  - evaluate all subsets and select the best model (CV)
- Shrinkage (reguralization):
  - a penalty on coefficients size shrunks them towards zero
- Dimension Reduction:
  - from p dimension select M-dimensional subspace, M<p.
  - fit a linear model in this M-dim. subspace.

Best Subset Selection

- Null model $\mathcal{M}_0$ predicts $\hat{f}(x) = \bar{y}$
- for(k in 1:p)
  - fit $\binom{p}{k}$ models with exactly *k* predictors
  - select the one with smallest RSS, or equiv. largest R^2
    - denote it $\mathcal{M}_k$

- Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using crossvalidation, AIC, BIC or adjusted R^2.

Best Subset Selection



- tractable up to p=30,40.
- Simillarly, for logistic regression
  - with deviance as error measure instead of RSS,
  - again, CV for model 'size' selection.

Forward Stepwise Selection

- Null model $\mathcal{M}_0$ predicts $\hat{f}(x) = \bar{y}$
- for(k in 0:(p-1))
  - consider (p-k) adding one predictor to $\mathcal{M}_k$
  - select the one with smallest RSS, or equiv. largest R^2
    - denote it $\mathcal{M}_{k+1}$
  -
- Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$
- using crossvalidation, AIC, BIC or adjusted R^2.

Backward Stepwise Selection

- Full model $\mathcal{M}_p$ with p predictors (standard LR).
- for(k in (p-1):0)
  - consider (k+1) models removing one predictor fror $\mathcal{M}_{k+}$
  - select the one with smallest RSS, or equiv. largest R^2
    - denote it $\mathcal{M}_k$
  -
- Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$
- using crossvalidation, AIC, BIC or adjusted R^2.

Hybrid Approaches

- go Forward, any time try to eliminate useless predictor.

- Each algorithm may provide different subset for a given size k (except 0 and p ;-)
- None of these has to be optimal with respect to mean test error.

Choosing the Optimal Model

- Two main approaches:
- Analytical criteria, adjustment to the training error to reduce overfitting ('penalty')
  - should not be used for p>>N!
- Direct estimate of test error, either
  - validation set
  - or cross-validation approach.

Analytical Criteria

- Mallow 'in sample error estimate'

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

- Akaike: (more general, proportional to C_p here)

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

- Bayesian Information Criterion:

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

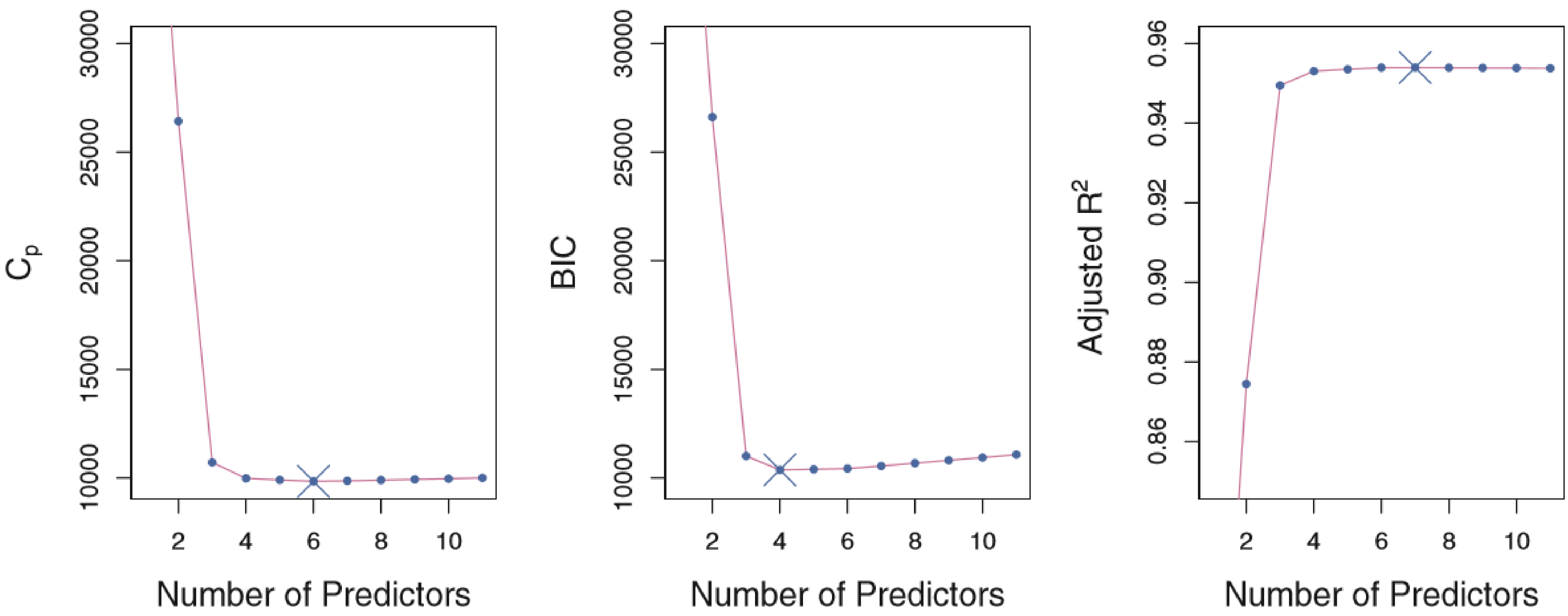
- Adjusted R^2 :

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

equiv. minimize $\frac{RSS}{n-d-1}$

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

Example



Cp
30000
25000
20000
15000
10000
2
4
6
8
10
Number of Predictors
BIC
30000
25000
20000
15000
10000
2
4
6
8
10
Number of Predictors
Adjusted R2
0.96
0.94
0.92
0.90
0.88
0.86
2
4
6
8
10
Number of Predictors

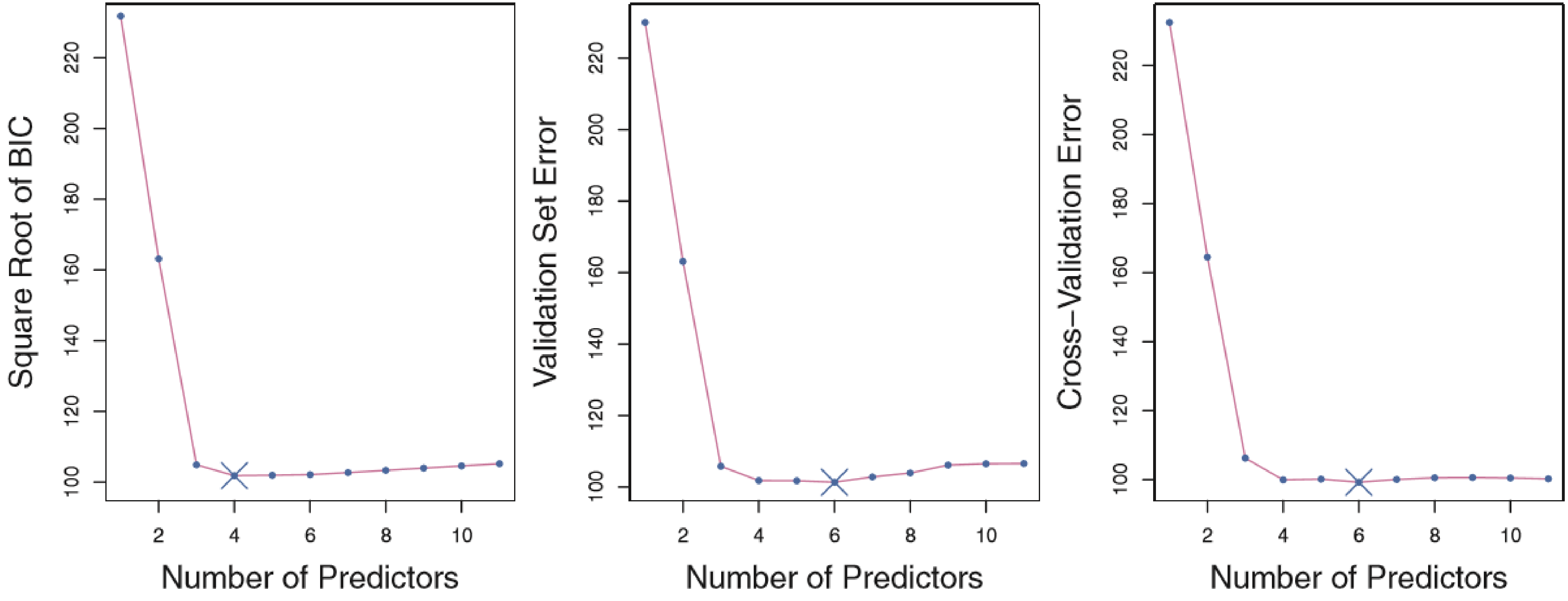
Validation and Cross-Validation

- Validation: at the beginning,
  - exclude 1/4 of data samples from training
  - use them for error estimation for model selection.

- Cross-Validation: at the beginning,
  - split data records into k=10 folds,
  - for k in 1:10
    - hide k-th fold for training
    - use it for error estimation for model selection.

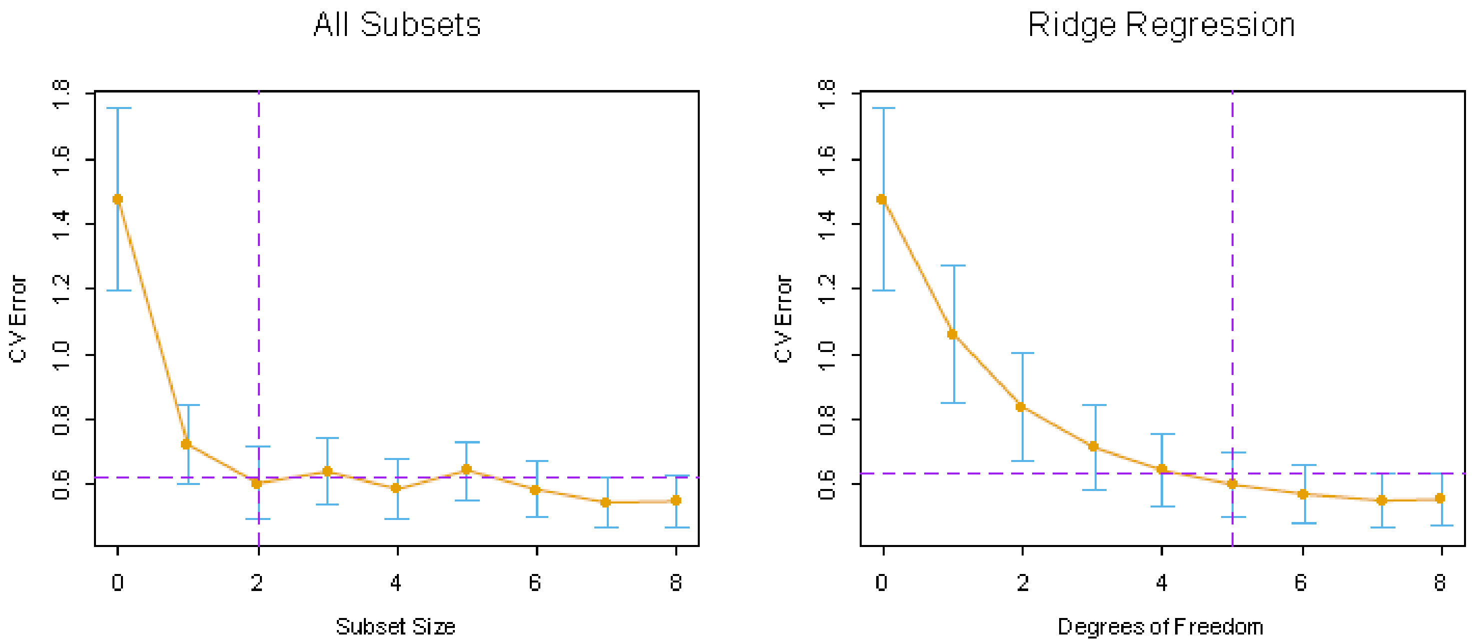
Note: different runs may provide different subsets of size 3.

Example


Square Root of BIC
Validation Set Error
Cross−Validation Error
100
120
140
160
180
200
220
2
4
6
8
10
Number of Predictors
Number of Predictors
Number of Predictors

One Standard Error Rule

- take the model size with the minimal CV error
- calculate 1 std. err. interval arround this error,
- select the smallest model with error inside this interval.



Shrinkage Methods

- Penalty for non-zero model parameters,
- no penalty for intercept.

- Ridge: $$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}.$$

- Lasso:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

Ridge

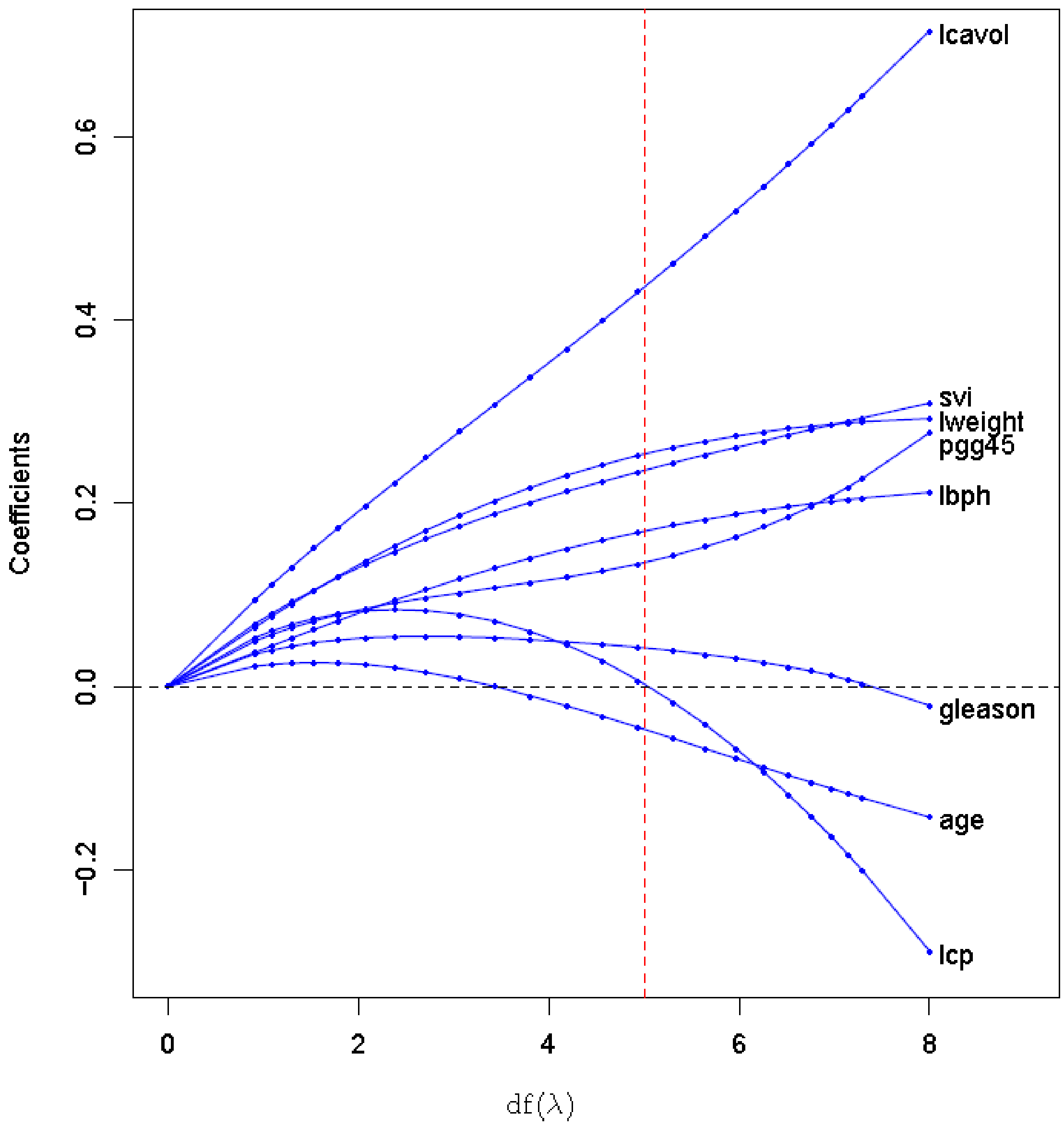
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \qquad \lambda \geq 0$$

- Parameter lambda penalizes the sum of β^2 .
- β_0 intentionally excluded from the penalty.
- we can center features and fix:

$$\beta_0 \text{ by } \bar{y} = \frac{1}{N} \sum_1^N y_i.$$

- For centered featues: $\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$
- for orthonormal features: $\hat{\beta}^{\text{ridge}} = \hat{\beta}/(1+\lambda).$
- Dependent on scale: standardization usefull.

Ridge coef. - Cancer example



Lasso regression

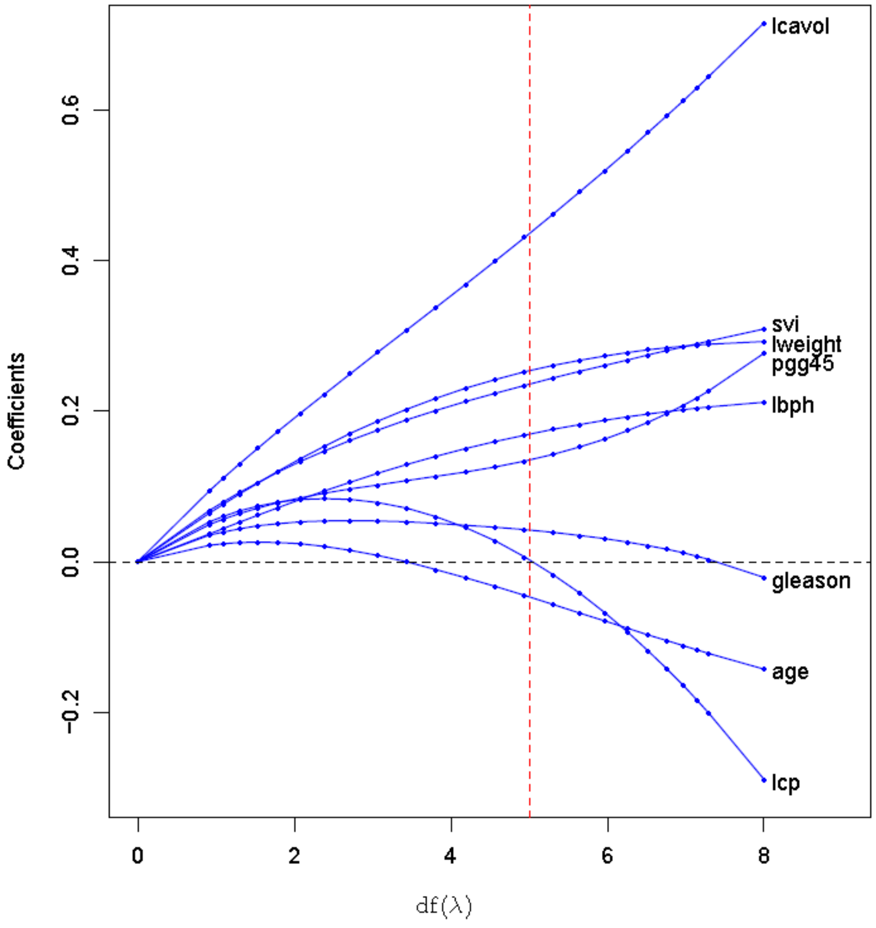
$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

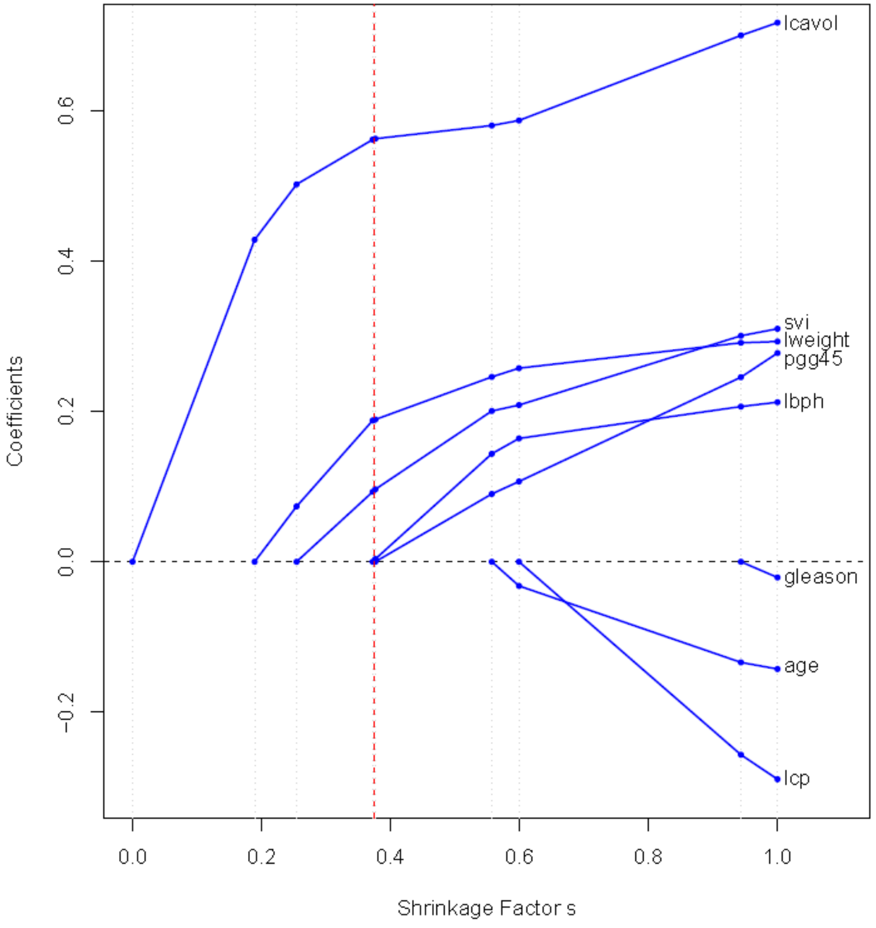
- the penalty is $\sum_1^p |\beta_j|$
- it forces some coefficients to be zero
- an equvivalent specification:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t.$$

Ridge x Lasso





Linear Models for Regression

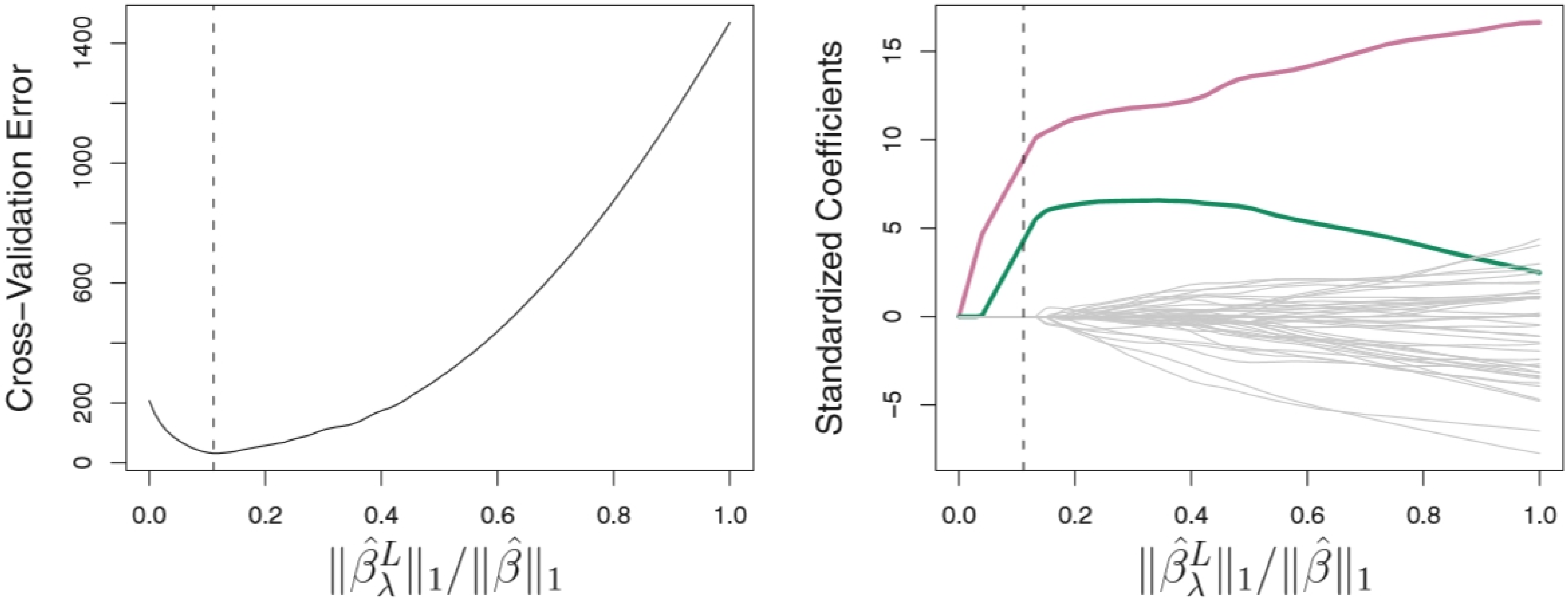
TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

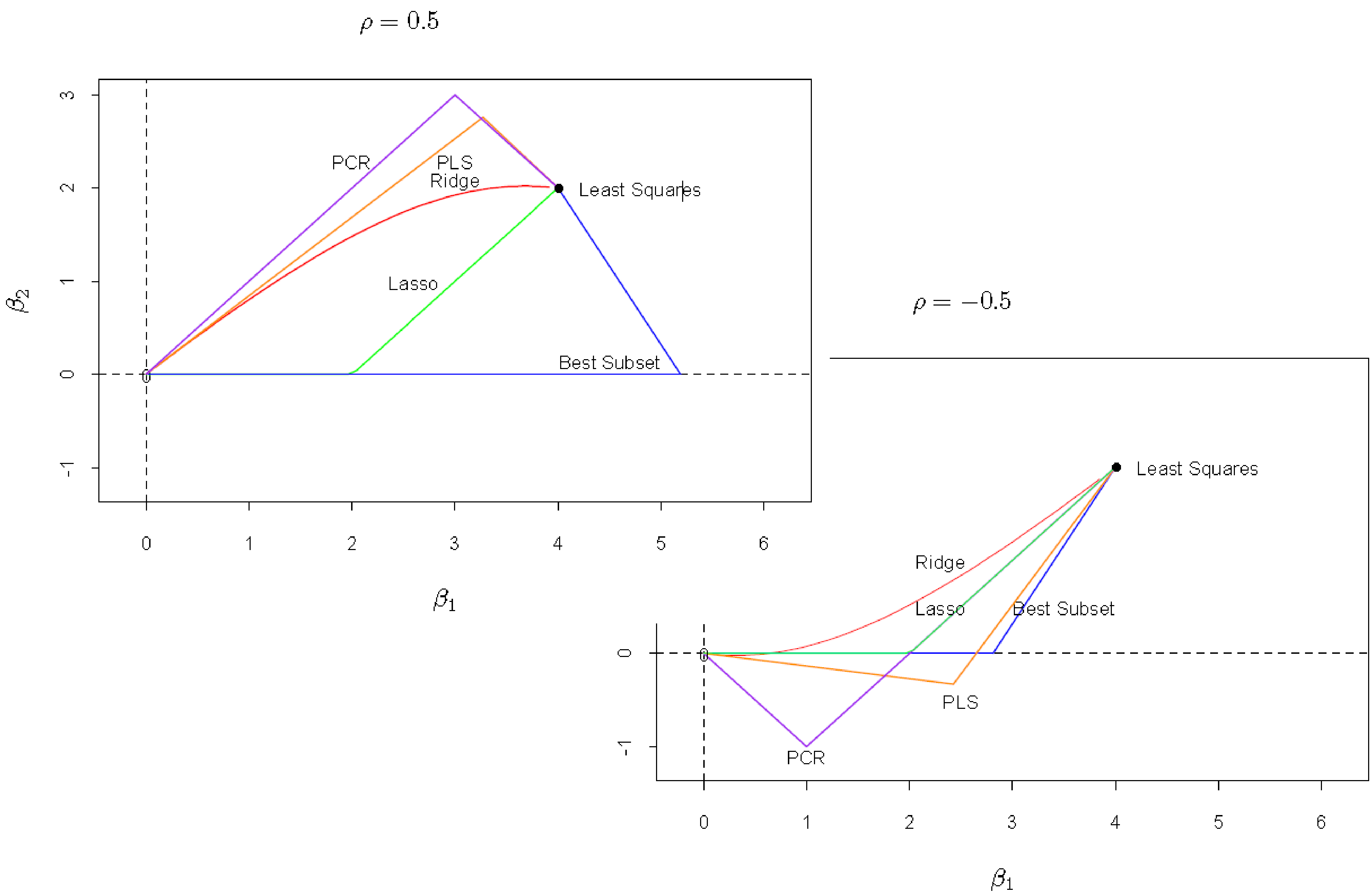
- Ridge, Lasso – penalization
- PCR, PLS – coordinate system change + dimension selection

Example

- p=45, n=50, 2 predictors relate to output.



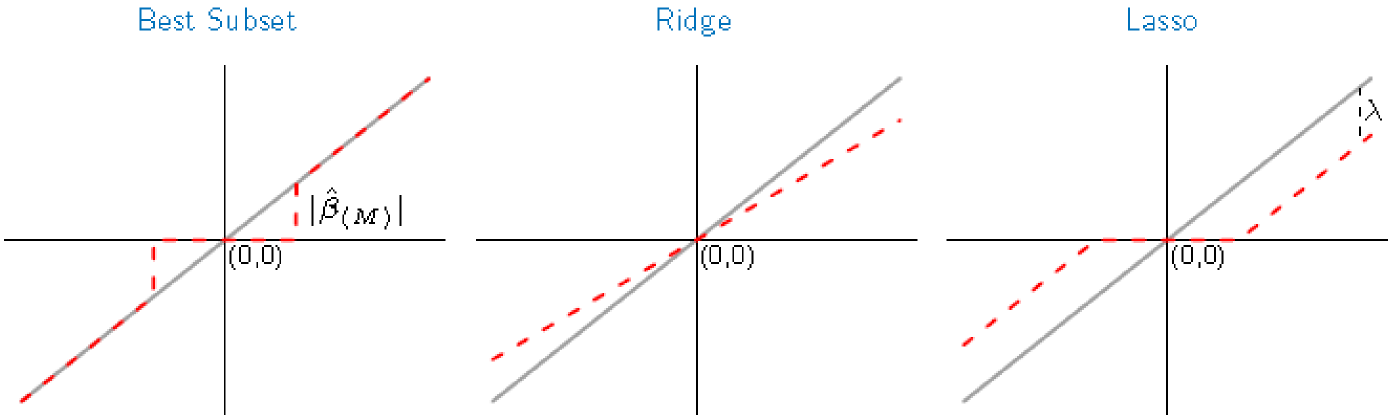
Corellated X, Parameter Shrinkage



Best subset, Ridge, Lasso

- Coefficient change for orthonormal features:

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



PCR, PLS

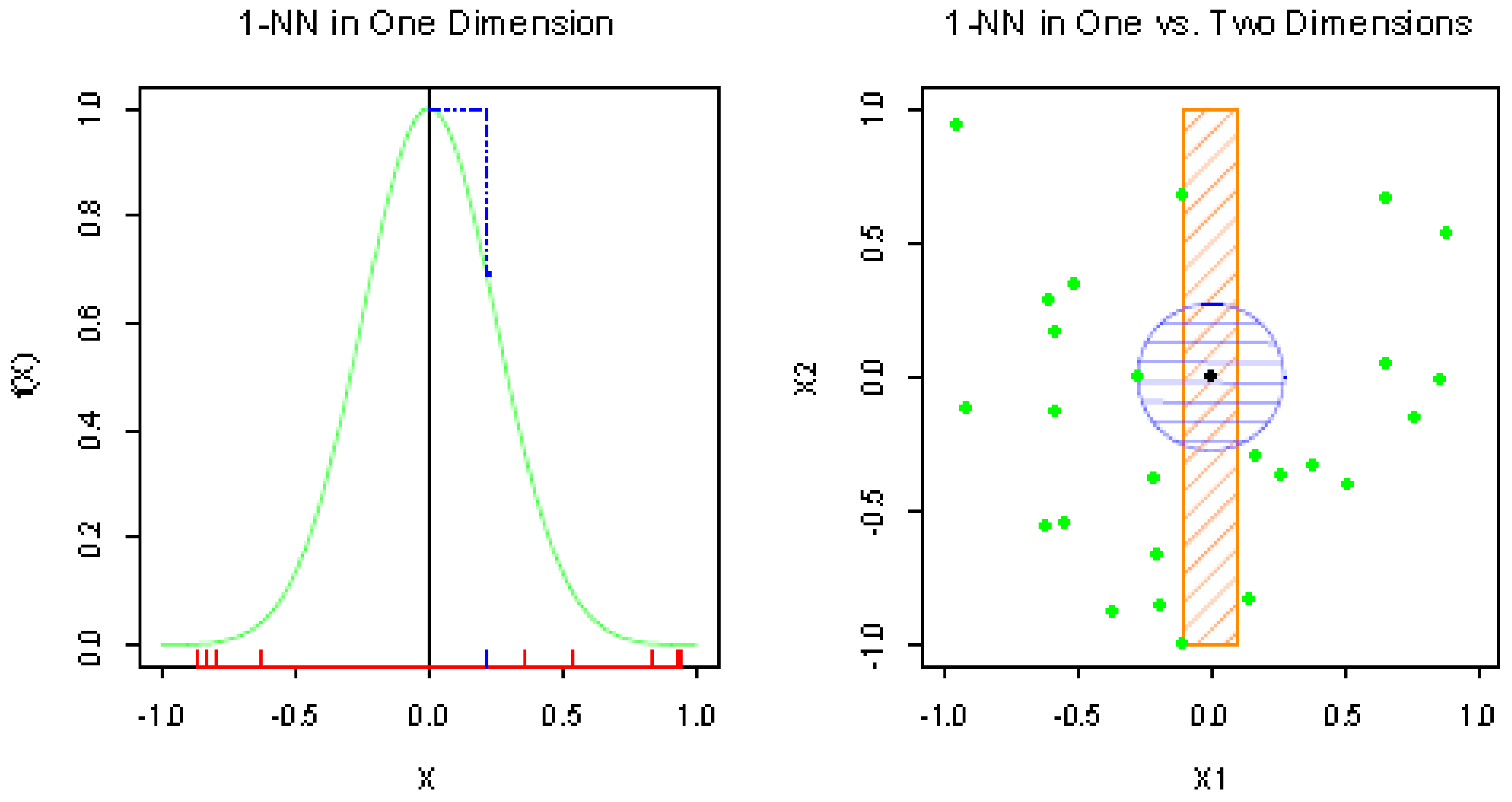
- PCR Principal component regression
  - select direction corresponding to largest eigenvalues
  - for these directions, regression coeff. are fitted.
  - For size=p equivalent with linear regression.
- Partial least squares – considers Y for selection
  - calculates regression coefficients
  - weight features and calculate eigenvalues
  - select the first direction of PLS,
  - other direction simillar, orthogonal to the first.

Bias-variance decomposition

$$\begin{aligned}\mathrm{MSE}(x_0) &= \mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\ &= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 + [\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\ &= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0).\end{aligned}$$

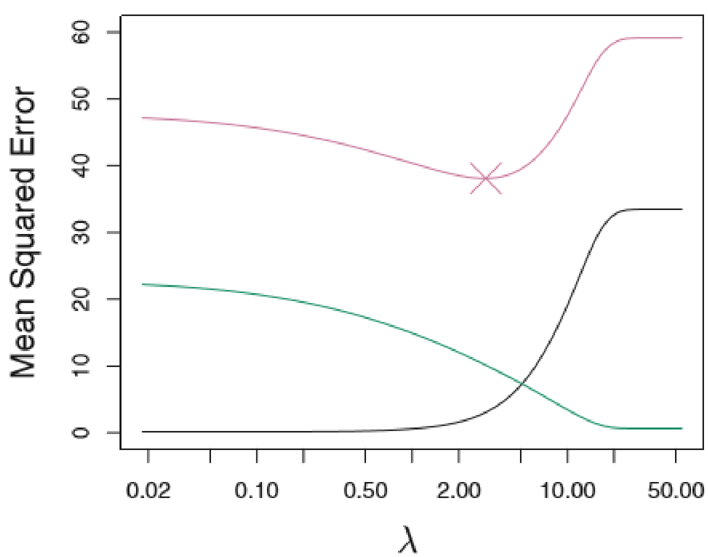
- Bias – 'systematic error',
  - usually caused by restricted model subspace
- Var – variance of the estimate
- we wish both to be zero.

Example of Bias

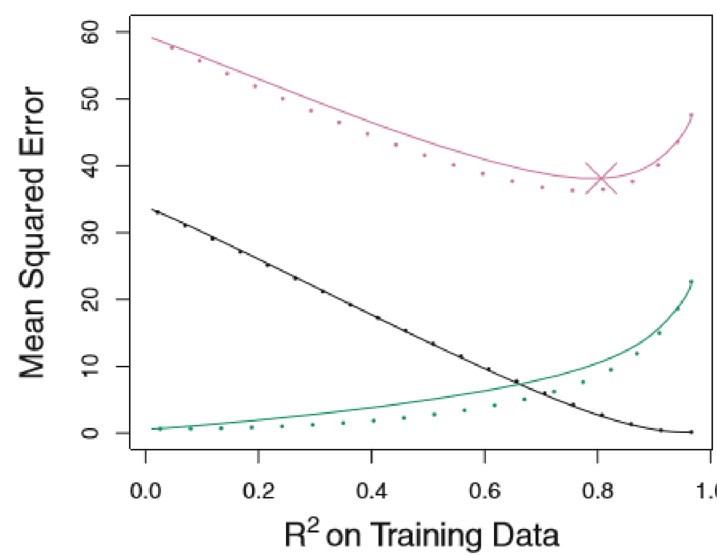


Example: Lasso, Ridge Regression

- red: MSE
- green: variance
- black: squared Bias

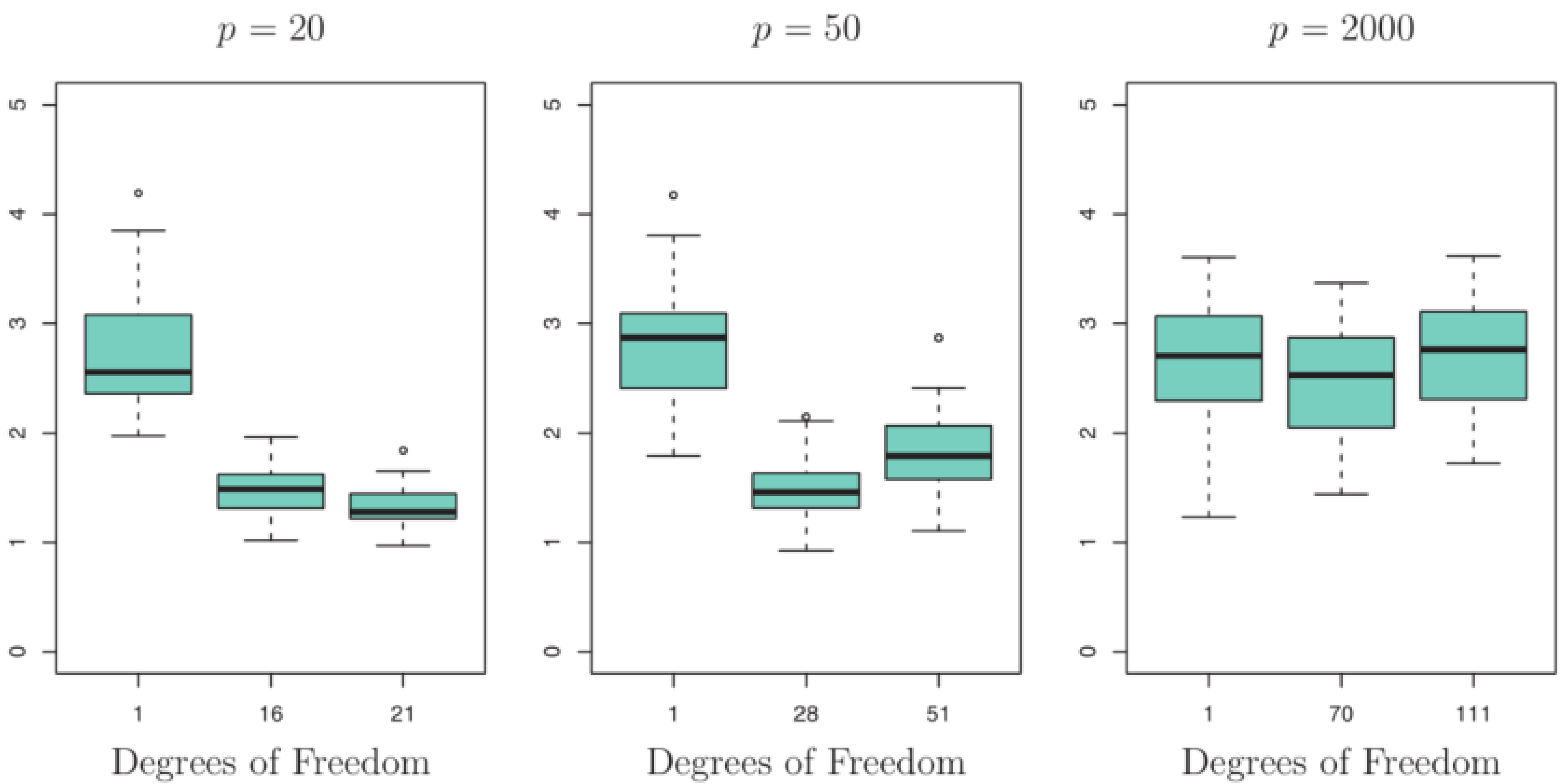


penalty



positive

MSE: 100 observations, p differs



Penalty ~ prior model probability

- Ridge $$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

- we assume prior probability of parameters
  - β_j independent, $N(0, \tau^2)$

    $y_i \sim N(\beta_0 + x_i^T \beta, \sigma^2)$

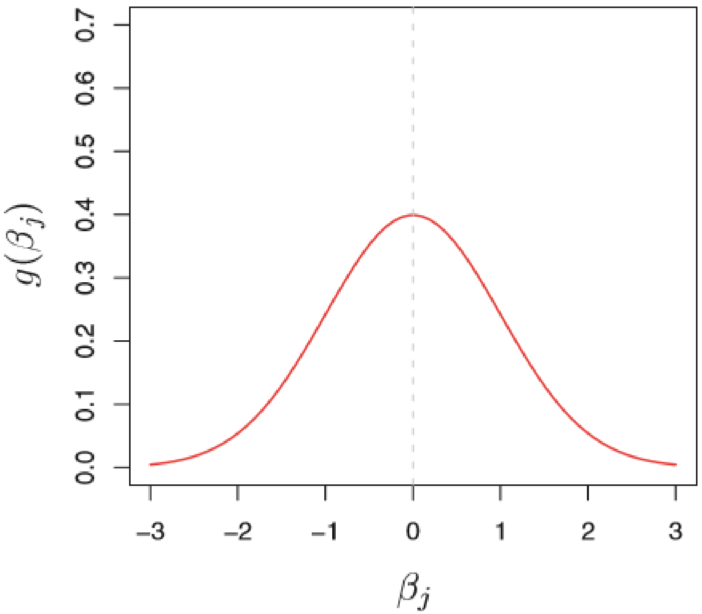
    $\lambda \cong \sigma^2 / \tau^2$

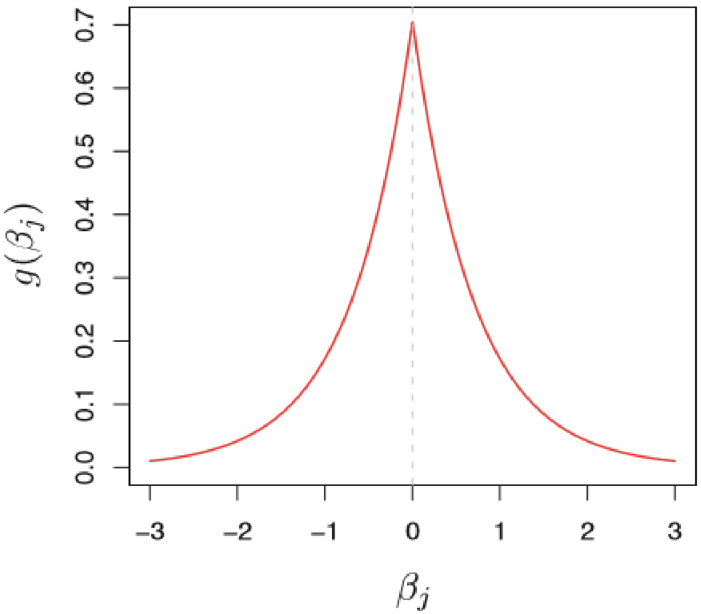
    then Ridge is most likely estimate (posterior mode).
  - Bayes formula $P(\beta / X) = \frac{P(X/\beta) \cdot P(\beta)}{P(X)}$
    - P(X) constant, $P(\beta)$ prior probability,
    - $P(X/\beta)$ likelihood, $P(\beta / X)$ posterior probability.

Prior Probability Ridge, Laso

- Ridge: Normal distribution
- Lasso: Laplace distribution

$$\frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right)$$





Principal Component Analysis PCA

sample covariance matrix is given by $\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$,

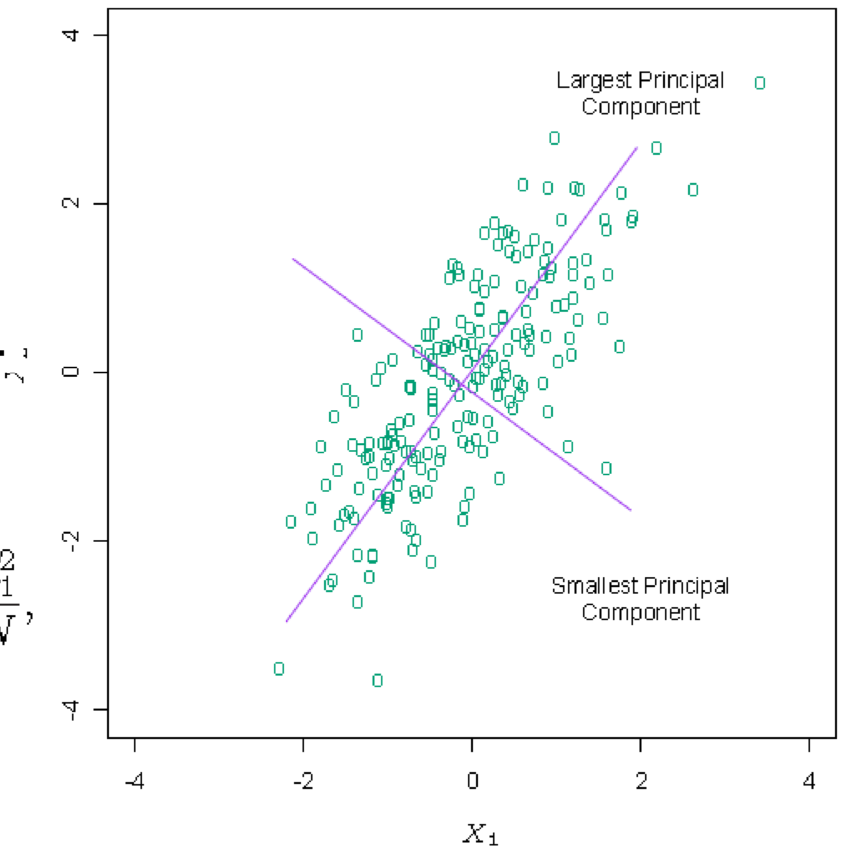
$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T,$$

eigenvectors v_j (columns of $\mathbf{V}$)

principal components

directions of $\mathbf{X}$.

$$\mathbf{z}_1 = \mathbf{X}v_1 = \mathbf{u}_1 d_1. \qquad \mathrm{Var}(\mathbf{z}_1) = \mathrm{Var}(\mathbf{X}v_1) = \frac{d_1^2}{N},$$



(vlastní čísla, vlastní vektory)