Model Assessment and Selection

TrungTin Nguyen

STATIFY team, Inria centre at the University Grenoble Alpes, France

Statistical Analysis and Document Mining

Complementary Course, Master of Applied Mathematics in Grenoble

Outline

Outline

Previous episode: generalized linear models

☛ Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression.

Previous episode: generalized linear models

☛ Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$\boxed{Y \in \text{exponential family (including Gaussian, Bernoulli \& Poisson).}}$

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0+\sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0+\sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0+\sum_{p=1}^{P} \beta_p X_p}$.

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, \dots, X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

HOW.

Previous episode: generalized linear models

Recall that conditional on $X_{[P]} \equiv (X_1, ..., X_P)$, Y belongs to a certain family of distributions: Gaussian or normal distribution for linear regression, Bernoulli distribution for logistic regression and Poisson distribution for Poisson regression. What is the common thing?

$Y \in$ exponential family (including Gaussian, Bernoulli & Poisson).

Each approach models the mean of Y as a function of the predictors.

1. Linear regression $\mathbb{E}\left[Y|X_{[P]}\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.
2. Logistic regression $\mathbb{E}\left[Y|X_{[P]}\right] = \mathbb{P}(Y = 1|X_{[P]}) = \frac{e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}{1+e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}}$.
3. Poisson regression $\mathbb{E}\left[Y|X_{[P]}\right] = \lambda(X_{[P]}) = e^{\beta_0 + \sum_{p=1}^{P} \beta_p X_p}$.

Using a link function η such that $\eta\left(\mathbb{E}\left[Y|X_{[P]}\right]\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$.

HOW. $\eta(\mu) = \mu$, $\eta(\mu) = \log(\mu/(1-\mu))$, $\eta(\mu) = \log(\mu)$.

Test error in multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Test error in multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$, $k \in [K]$, and $\mathbf{X}$?

Test error in multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$, $k \in [K]$, and $\mathbf{X}$?
- Multinomial logistic regression (LR) takes the form $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X})$, and models the probability that Y belongs to a particular category instead of the value of Y as follows:

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p.$$

- **Training error:** using non-linear LS or maximum likelihood estimation (MLE), we obtain $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) = \operatorname{argmax}_{c\in\mathcal{C}} p_c(\mathbf{x}_n)$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[y_n \neq \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right] \approx 0. \quad (1)$$

Test error in multinomial logistic regression

- We are given a **training dataset** $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $y_{[N]} \in \mathcal{C} \equiv [K]$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- How to model relationship between $p_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$, $k \in [K]$, and $\mathbf{X}$?
- Multinomial logistic regression (LR) takes the form $p_K(\mathbf{X}) = 1 - \sum_{k=1}^{K-1} p_k(\mathbf{X})$, and models the probability that Y belongs to a particular category instead of the value of Y as follows:

$$\log\left(\frac{p_k(\mathbf{X})}{p_K(\mathbf{X})}\right) = \beta_{k0} + \sum_{p=1}^{P} \beta_{kp} x_p.$$

- **Training error:** using non-linear LS or maximum likelihood estimation (MLE), we obtain $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) = \operatorname{argmax}_{c\in\mathcal{C}} p_c(\mathbf{x}_n)$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[y_n \neq \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right] \approx 0. \quad (1)$$

Test (generalization) error: for any **new sample** $(\mathbf{x}^*, y^*)$, how we guarantee

$$y^* \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}^*), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}) \equiv \mathbb{E}_{\mathbf{X},Y}\left[\mathbb{1}\left(Y \neq \widehat{r}_{\mathcal{D}}(\mathbf{X})\right)\right] \approx 0? \quad (2)$$

Outline

1. Model Assessment in Generalized Linear Models
   - Generalized linear models
   - Test error in multiple linear regression
   - Choosing the optimal model in subset selection
   - Estimating test error: two approaches
2. Generalized Linear Model Selection and Regularization
   - General model selection paradigm
   - Asymptotic approach: C_p, AIC, BIC and Adjusted R^2
   - Validation and cross-validation
   - An empirical comparison on Credit data
   - Non-asymptotic approach: slope heuristic ...
3. High-Dimensional Setting
   - Previous episode: multiple impact of high-dimensionality on statistics
4. Perspectives

Previous episode: test error in multiple linear regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.

Previous episode: test error in multiple linear regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \ldots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \ldots, \beta_P).$$

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \dots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \dots, \beta_P).$$

- **Training error:** using ordinary least squares (OLS), we obtain coefficient estimates $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or } \mathrm{RSS} \equiv \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{r}_{\mathcal{D}}(\mathbf{x}_n))^2 \approx 0. \tag{3}$$

Previous episode: test error in multiple linear regression

- We are given a **training dataset** $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$, $[N] \equiv 1, \dots, N$, i.i.d. sampled from **the true (but unknown) joint PDF** of $(\mathbf{X}, Y)$.
- Given any error term ϵ, multiple linear regression function takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon \equiv r(\mathbf{X}) + \epsilon, \quad \boldsymbol{\beta} \equiv (\beta_0, \beta_1, \dots, \beta_P).$$

- **Training error:** using ordinary least squares (OLS), we obtain coefficient estimates $\widehat{\beta}$ and $\widehat{r}_{\mathcal{D}}(\mathbf{x}_n) \equiv \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}$ such that $\forall n \in [N]$,

$$y_n \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}_n), \text{ or } \mathrm{RSS} \equiv \mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \equiv \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \widehat{r}_{\mathcal{D}}(\mathbf{x}_n)\right)^2 \approx 0. \tag{3}$$

Test (generalization) error: for **new sample** $(\mathbf{x}^*, y^*)$, how can we guarantee

$$y^* \approx \widehat{r}_{\mathcal{D}}(\mathbf{x}^*), \text{ or equivalent, } \mathcal{L}(\widehat{r}_{\mathcal{D}}) \equiv \mathbb{E}_{\mathbf{X},Y}\left[\left(Y - \widehat{r}_{\mathcal{D}}(\mathbf{X})\right)^2\right] \approx 0? \tag{4}$$

Recall from CM3, in general, it holds that $\mathcal{L}(\widehat{r}_{\mathcal{D}}, \mathcal{D}) \leq \mathcal{L}(\widehat{r}_{\mathcal{D}})$. (5)

Outline

Choosing the optimal model in subset selection

We should select a best data driven model with the smallest RSS and the largest R^2 among a collection of models with different numbers of predictors.

Choosing the optimal model in subset selection

We should select a best data driven model with the smallest RSS and the largest R^2 among a collection of models with different numbers of predictors. ❓

Choosing the optimal model in subset selection

We should select a best data driven model with the smallest RSS and the largest R^2 among a collection of models with different numbers of predictors. ❓

- 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.

Choosing the optimal model in subset selection

We should select a best data driven model with the smallest RSS and the largest R^2 among a collection of models with different numbers of predictors. ❓

- 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 low test error, not a model with low training error.** Recall that training error is usually a poor estimate of test error.

Choosing the optimal model in subset selection

We should select a best data driven model with the smallest RSS and the largest R^2 among a collection of models with different numbers of predictors. ❓

- 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 low test error, not a model with low training error.** Recall that training error is usually a poor estimate of test error.

Therefore, RSS **and** R^2 **are not suitable for selecting the best model among a collection of models with different numbers of predictors.**

Outline

1. Model Assessment in Generalized Linear Models
   - Generalized linear models
   - Test error in multiple linear regression
   - Choosing the optimal model in subset selection
   - Estimating test error: two approaches

2. Generalized Linear Model Selection and Regularization
   - General model selection paradigm
   - Asymptotic approach: C_p, AIC, BIC and Adjusted R^2
   - Validation and cross-validation
   - An empirical comparison on Credit data
   - Non-asymptotic approach: slope heuristic ...

3. High-Dimensional Setting
   - Previous episode: multiple impact of high-dimensionality on statistics

4. Perspectives

Choosing the optimal model in subset selection

In order to select the best model with respect to test error, we need to **estimate this test error:**

1. We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting: Mallows's C_p, AIC, BIC and **slope heuristic**.

Choosing the optimal model in subset selection

In order to select the best model with respect to test error, we need to **estimate this test error:**

1. We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting: Mallows's C_p, AIC, BIC and **slope heuristic**.
2. We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures (CM3 and CC4).

Outline

Outline

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.*, least-squares contrast or MLE, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[5] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

[6] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[7] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[8] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

General model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast or MLE, over S_m. ⌖ **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. **Asymptotic approach:** Mallows's C_p[1],

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[5] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

[6] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[7] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[8] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

- **Model selection problem:** let $(S_m)_{m\in\mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast or MLE, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. **Asymptotic approach:** Mallows's C_p[1], Akaike information criterion[2] (AIC),

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[5] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

[6] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[7] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[8] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

General model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast or MLE, over S_m. ⌖ **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. **Asymptotic approach:** Mallows's C_p[1], Akaike information criterion[2] (AIC), Bayesian information criterion[3] (BIC), Adjusted R^2: no finite sample guarantees, but classical and important for understanding.

---

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[5] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

[6] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[7] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[8] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

General model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast or MLE, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. **Asymptotic approach:** Mallows's C_p[1], Akaike information criterion[2] (AIC), Bayesian information criterion[3] (BIC), Adjusted R^2: no finite sample guarantees, but classical and important for understanding.
  2. **Non-asymptotic approach: slope heuristic**[4] [5] particularly useful for high-dimensional small data sets, *e.g.,* $N \ll P$.
  3. **Cross-validation procedures:**[6] [7] [8] K-Fold, leave-one-out in CC4.

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[5] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

[6] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[7] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[8] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

Outline

Overview: C_p, AIC, BIC and Adjusted R^2

1. These techniques adjust the training error for the model size, and can be used to **select among a set of models with different numbers of variables**.

Overview: C_p, AIC, BIC and Adjusted R^2

1. These techniques adjust the training error for the model size, and can be used to **select among a set of models with different numbers of variables**.
2. The next few slides display C_p, AIC, BIC and Adjusted R^2, CV, ridge and Lasso for the best model of each size produced by best subset selection on the Credit data set.

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent?} \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent?} \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

☛ C_p statistic adds a penalty of $\mathrm{pen}(p) \equiv 2p\widehat{\sigma}^2$ to the training RSS.

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent?} \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

☛ C_p statistic adds a penalty of $\mathrm{pen}(p) \equiv 2p\widehat{\sigma}^2$ to the training RSS. But why?

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent?} \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

☛ C_p statistic adds a penalty of $\mathrm{pen}(p) \equiv 2p\widehat{\sigma}^2$ to the training RSS. But why? Adjusting for the fact that the training error tends to underestimate the test error.

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent? } \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

   Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

   ☛ C_p statistic adds a penalty of $\mathrm{pen}(p) \equiv 2p\widehat{\sigma}^2$ to the training RSS. But why? Adjusting for the fact that the training error tends to underestimate the test error.

2. AIC criterion: is defined for a large class of models fit by MLE:

$$\mathrm{AIC}(p) = -2\log L(p) + 2p \text{ or equivalent? } \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right]. \tag{7}$$

   Here, $L(p)$ is the maximized value of the likelihood function for the estimated model.

Some details: C_p and AIC

1. Mallow's C_p: estimate of test MSE (unbiased one when and why?) and choosing the model with the lowest C_p value:

$$\mathrm{C_p} = \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right] \text{ or equivalent?} \frac{\mathrm{RSS}}{\widehat{\sigma}^2} + 2p - N. \tag{6}$$

Here p is the total number of parameters used (*e.g.*, number of predictors) $\widehat{\sigma}^2$ is an estimate of the variance of the error ϵ. Typically $\widehat{\sigma}^2$ is estimated using the full model containing all predictors.

C_p statistic adds a penalty of $\mathrm{pen}(p) \equiv 2p\widehat{\sigma}^2$ to the training RSS. But why? Adjusting for the fact that the training error tends to underestimate the test error.

2. AIC criterion: is defined for a large class of models fit by MLE:

$$\mathrm{AIC}(p) = -2\log L(p) + 2p \text{ or equivalent?} \frac{1}{N}\left[\mathrm{RSS}(p) + 2p\widehat{\sigma}^2\right]. \tag{7}$$

Here, $L(p)$ is the maximized value of the likelihood function for the estimated model.

In the case of the linear model with Gaussian errors, MLE and least squares are the same thing, and $\mathrm{C_p}$ and AIC are equivalent (Prove this?).

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N) p \widehat{\sigma}^2\right]. \tag{8}$$

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N)p\widehat{\sigma}^2\right]. \tag{8}$$

Compare to C_p and AIC?

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N)p\widehat{\sigma}^2\right]. \tag{8}$$

Compare to C_p and AIC?
How we should choose the best data driven model $\widehat{p}$ from $\mathrm{BIC}(p)$?

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N)p\widehat{\sigma}^2\right]. \tag{8}$$

Compare to C_p and AIC?
How we should choose the best data driven model $\widehat{p}$ from $\mathrm{BIC}(p)$?

- Notice that BIC replaces the $2p\widehat{\sigma}^2$ used by $\mathrm{C_p}$ with a $\log(N)p\widehat{\sigma}^2$ term, where N is the number of observations.

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N)p\widehat{\sigma}^2\right]. \tag{8}$$

Compare to C_p and AIC?
How we should choose the best data driven model $\widehat{p}$ from $\mathrm{BIC}(p)$?

- Notice that BIC replaces the $2p\widehat{\sigma}^2$ used by $\mathrm{C_p}$ with a $\log(N)p\widehat{\sigma}^2$ term, where N is the number of observations.
- The BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.

Some details: BIC

3. BIC criterion: is derived from a Bayesian point of view, up to irrelevant constants, given by

$$\mathrm{BIC}(p) = \frac{1}{N}\left[\mathrm{RSS}(p) + \log(N)p\widehat{\sigma}^2\right]. \tag{8}$$

Compare to C_p and AIC?
How we should choose the best data driven model $\widehat{p}$ from $\mathrm{BIC}(p)$?

- Notice that BIC replaces the $2p\widehat{\sigma}^2$ used by $\mathrm{C_p}$ with a $\log(N)p\widehat{\sigma}^2$ term, where N is the number of observations.
- The BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- Since $logN > 2$ for any $N > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than $\mathrm{C_p}$. To be verified in Figure 1 soon!

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.
Compare to C_p, AIC and BIC?

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.
Compare to C_p, AIC and BIC?
How we should choose the best data driven model $\widehat{p}$ from Adjusted $\mathrm{R}^2(p)$?

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.
Compare to C_p, AIC and BIC?
How we should choose the best data driven model $\widehat{p}$ from Adjusted $\mathrm{R}^2(p)$?

- Unlike $\mathrm{C_p}$, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.
Compare to C_p, AIC and BIC?
How we should choose the best data driven model $\widehat{p}$ from Adjusted $\mathrm{R}^2(p)$?

- Unlike $\mathrm{C_p}$, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing $\frac{\mathrm{RSS}}{N-p-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\mathrm{RSS}}{N-p-1}$ may increase or decrease, due to the presence of p in the denominator.

Some details: Adjusted R^2

4. Adjusted R^2: is given by

$$\text{Adjusted } \mathrm{R}^2(p) = 1 - \frac{N-1}{N-p-1}\frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{9}$$

Here TSS is the total sum of squares.
Compare to C_p, AIC and BIC?
How we should choose the best data driven model $\widehat{p}$ from Adjusted $\mathrm{R}^2(p)$?

- Unlike $\mathrm{C_p}$, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing $\frac{\mathrm{RSS}}{N-p-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\mathrm{RSS}}{N-p-1}$ may increase or decrease, due to the presence of p in the denominator.
- Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model. To be verified in Figure 1 soon!

Outline

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

- A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

- A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

- The one-standard-error rule.

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

- A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

- The one-standard-error rule. Do you know this rule?

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

- A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

- The one-standard-error rule. Do you know this rule?
  1. Calculate the standard error of the estimated test MSE for each model size,

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

☻ A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

☛ The one-standard-error rule. Do you know this rule?

1. Calculate the standard error of the estimated test MSE for each model size,
2. Select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

- A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

- The one-standard-error rule. Do you know this rule?
    1. Calculate the standard error of the estimated test MSE for each model size,
    2. Select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

What is the rationale for this?

Some details: validation and cross-validation

5. **CV procedure:** has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.

☺ A wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

☛ The one-standard-error rule. Do you know this rule?

1. Calculate the standard error of the estimated test MSE for each model size,
2. Select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

What is the rationale for this? 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.

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k}\left(y_n - \widehat{r}^{(-k)}(\mathbf{x}_n)\right)^2\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k}\left(y_n - \widehat{r}^{(-k)}(\mathbf{x}_n)\right)^2\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k}\left(y_n - \widehat{r}^{(-k)}(\mathbf{x}_n)\right)^2\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.
   K = ?

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k}\left(y_n - \widehat{r}^{(-k)}(\mathbf{x}_n)\right)^2\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.
   $K = ?$ Setting $K = N$ yields N-fold or leave-one out cross-validation

Previous episode: K-Fold cross-validation for linear regression in detail

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k}\left(y_n - \widehat{r}^{(-k)}(\mathbf{x}_n)\right)^2\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.

   $K = ?$ Setting $K = N$ yields N-fold or leave-one out cross-validation

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k} \mathbb{1}\left(y_n \neq \widehat{r}^{(-k)}(\mathbf{x}_n)\right)\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n \in \mathcal{D}_k} \mathbb{1}\left(y_n \neq \widehat{r}^{(-k)}(\mathbf{x}_n)\right)\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m} \in \mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k} \mathbb{1}\left(y_n \neq \widehat{r}^{(-k)}(\mathbf{x}_n)\right)\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \operatorname{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.
   K = ?

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k} \mathbb{1}\left(y_n \neq \widehat{r}^{(-k)}(\mathbf{x}_n)\right)\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \mathrm{argmin}_{\mathbf{m}\in\mathcal{M}}\, \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.
   K = ? Setting $K = N$ yields N-fold or leave-one out cross-validation (LOOCV, **high variance**).

Previous episode: CV for classification problems

1. Split the training dataset randomly into K folds so that we have $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K = \mathcal{D}$, where $\mathcal{D}_k$ denotes the indices of the observations in part k. There are N_k observations in part k: if N is a multiple of K, then $N_k = N/K$.
2. For each $k \in [K]$, fit a model $\widehat{r}^{(-k)}(\mathbf{x}, \mathbf{m})$, indexed by a tuning parameter (or a model) $\mathbf{m} \in \mathcal{M}$, on all samples from the training set except those in the kth fold.
3. Estimating test error $\mathcal{L}(\widehat{r}_{\mathcal{D}})$ via averaging the final resulting MSE estimates

$$\mathcal{L}(\widehat{r}_{\mathcal{D}}) \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K}}_{\text{average over } K \text{ folds}} \underbrace{\left[\frac{1}{N_k}\sum_{n\in\mathcal{D}_k} \mathbb{1}\left(y_n \neq \widehat{r}^{(-k)}(\mathbf{x}_n)\right)\right]}_{\text{Estimate test error for each fold}} \equiv \mathrm{CV}(\widehat{r}, K, \mathbf{m}).$$

4. Best data-driven model: $\widehat{\mathbf{m}} \equiv \mathrm{argmin}_{\mathbf{m}\in\mathcal{M}} \mathrm{CV}(\widehat{r}, K, \mathbf{m})$.
   $K = ?$ Setting $K = N$ yields N-fold or leave-one out cross-validation (LOOCV, **high variance**). A common better choice $K = 5$ or 10.

Outline

An empirical comparison on Credit data

- **Description:** the response is balance (average credit card debt for 400 individuals) and there are 6 quantitative predictors: income (in thousands of dollars), limit (credit limit), rating (credit rating), cards (number of credit cards), age, education (years of education), and 4 qualitative variables: own (house ownership), student (student status), married (Yes or No), and region (East, West or South). [James et al., 2021, Section 3.3].

An empirical comparison on Credit data

- **Description:** the response is balance (average credit card debt for 400 individuals) and there are 6 quantitative predictors: income (in thousands of dollars), limit (credit limit), rating (credit rating), cards (number of credit cards), age, education (years of education), and 4 qualitative variables: own (house ownership), student (student status), married (Yes or No), and region (East, West or South). [James et al., 2021, Section 3.3].
- **Goal:** develop an accurate model that can be used to **predict balance** on the basis of 10 predictors ← Using glmnet package in $\boldsymbol{R}$.

```
> head(Credit)                                                          Y
   Income Limit Rating Cards Age Education Own Student Married Region Balance
1  14.891  3606    283     2  34        11  No      No     Yes  South     333
2 106.025  6645    483     3  82        15 Yes     Yes     Yes   West     903
3 104.593  7075    514     4  71        11  No      No      No   West     580
4 148.924  9504    681     3  36        11 Yes      No      No   West     964
5  55.882  4897    357     2  68        16  No      No     Yes  South     331
6  80.180  8047    569     4  77        10  No      No      No  South    1151
```

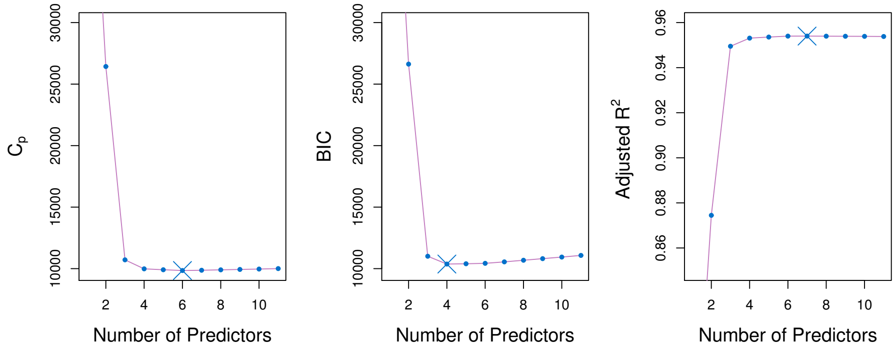


Figure 1: C_p (or AIC), BIC and Adjusted R^2 are shown for the best models of each size for the Credit data set [James et al., 2021, Figure 6.2]. Cp and BIC are estimates of test MSE.

Some comments. . .

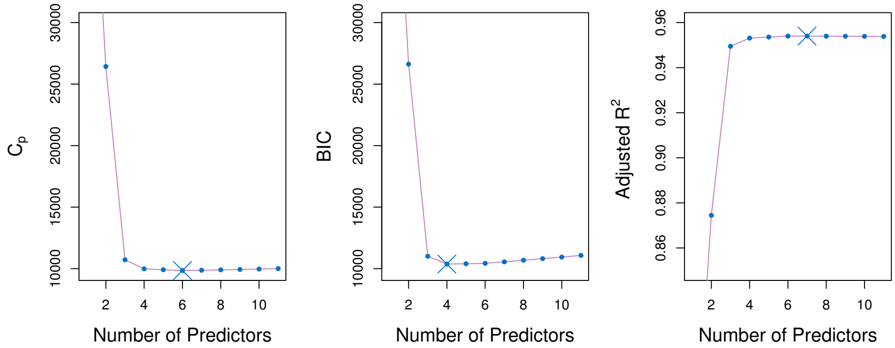


Figure 1: C_p (or AIC), BIC and Adjusted R^2 are shown for the best models of each size for the Credit data set [James et al., 2021, Figure 6.2]. Cp and BIC are estimates of test MSE.

Some comments...

In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected.

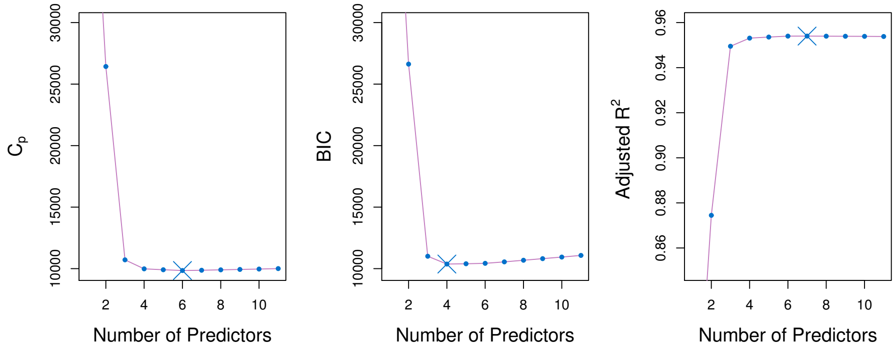


Figure 1: C_p (or AIC), BIC and Adjusted R^2 are shown for the best models of each size for the Credit data set [James et al., 2021, Figure 6.2]. Cp and BIC are estimates of test MSE.

Some comments...

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.

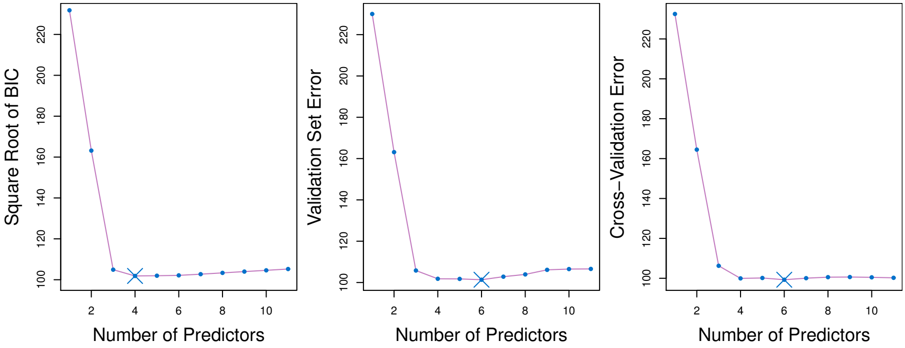


Figure 2: The overall best model, based on each of these quantities, is shown as a blue cross 'x'. [James et al., 2021, Figure 6.3].

Some comments. . .

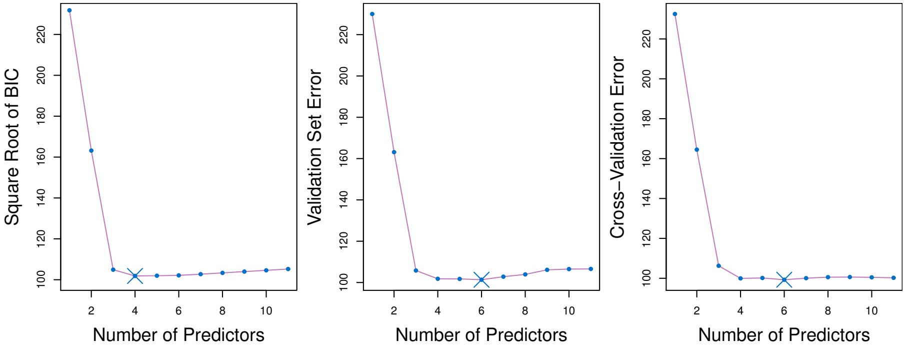


Figure 2: The overall best model, based on each of these quantities, is shown as a blue cross 'x'. [James et al., 2021, Figure 6.3].

❓ Some comments. . . However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

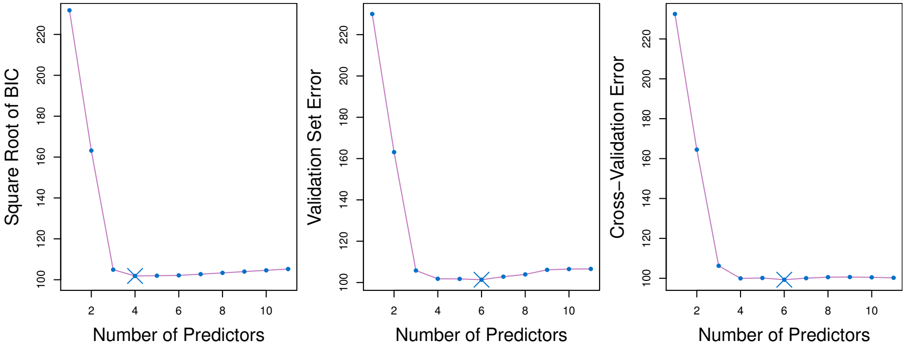


Figure 2: The overall best model, based on each of these quantities, is shown as a blue cross 'x'. [James et al., 2021, Figure 6.3].

Some comments... However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Applying the one-standard-error rule to the validation set or cross-validation approach?

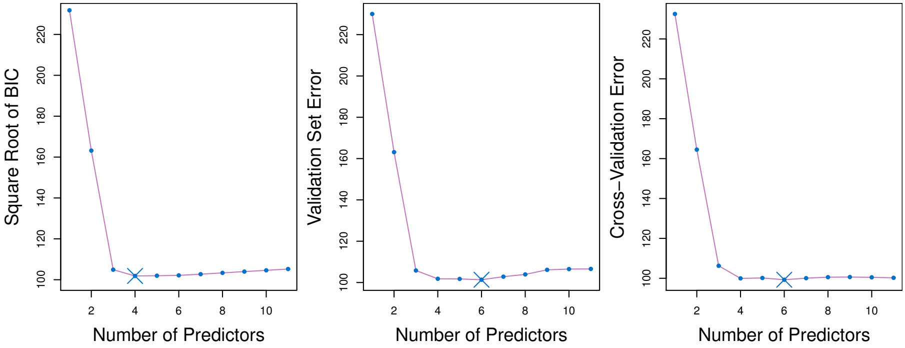


Figure 2: The overall best model, based on each of these quantities, is shown as a blue cross 'x'. [James et al., 2021, Figure 6.3].

Some comments... However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Applying the one-standard-error rule to the validation set or cross-validation approach? leads to selection of the three-variable model.

In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.

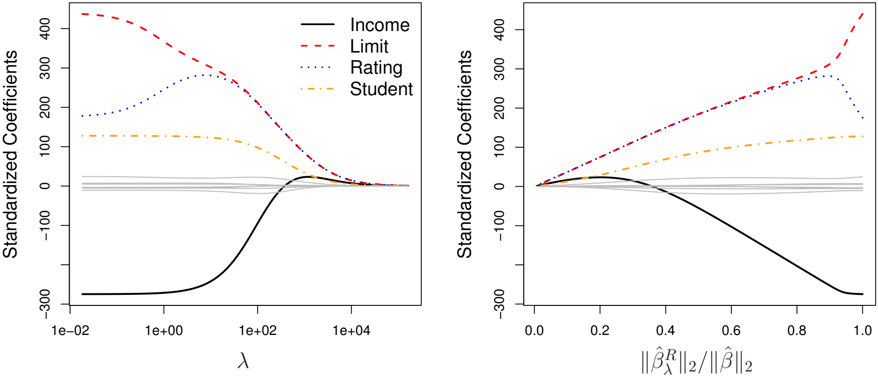


Figure 3: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\|\widehat{\beta}_\lambda^{\text{ridge}}\|_2/\|\widehat{\beta}^{\text{ls}}\|_2$ [James et al., 2021, Figure 6.4].

In the right-hand panel, a small value of the x-axis indicates that the ridge regression coefficient estimates have been shrunken very close to zero.

In the left-hand panel, each curve corresponds to the Lasso regression coefficient estimate for one of the ten variables, plotted as a function of λ.

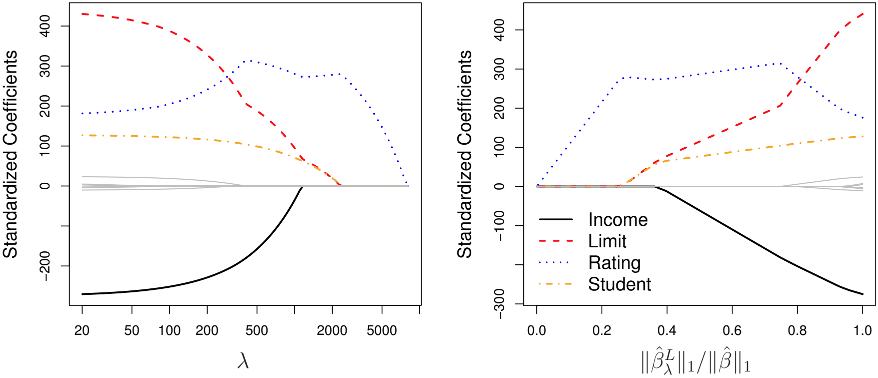


Figure 4: The standardized Lasso coefficients are displayed for the Credit data set, as a function of λ and $\|\widehat{\beta}^{\text{lasso}}_\lambda\|_1/\|\widehat{\beta}^{\text{ls}}\|_1$ [James et al., 2021, Figure 6.6].

In the right-hand panel, a small value of the x-axis indicates that the Lasso regression coefficient estimates have been shrunken to zero.

Outline

1. Model Assessment in Generalized Linear Models
   - Generalized linear models
   - Test error in multiple linear regression
   - Choosing the optimal model in subset selection
   - Estimating test error: two approaches
2. Generalized Linear Model Selection and Regularization
   - General model selection paradigm
   - Asymptotic approach: C_p, AIC, BIC and Adjusted R^2
   - Validation and cross-validation
   - An empirical comparison on Credit data
   - Non-asymptotic approach: slope heuristic ...
3. High-Dimensional Setting
   - Previous episode: multiple impact of high-dimensionality on statistics
4. Perspectives

Outline

Outline

1. Model Assessment in Generalized Linear Models
   - Generalized linear models
   - Test error in multiple linear regression
   - Choosing the optimal model in subset selection
   - Estimating test error: two approaches
2. Generalized Linear Model Selection and Regularization
   - General model selection paradigm
   - Asymptotic approach: C_p, AIC, BIC and Adjusted R^2
   - Validation and cross-validation
   - An empirical comparison on Credit data
   - Non-asymptotic approach: slope heuristic ...
3. High-Dimensional Setting
   - Previous episode: multiple impact of high-dimensionality on statistics
4. Perspectives

Multiple impact of high-dimensionality on statistics

1. High-dimensional spaces are vast and data points are isolated in their immensity (CC5).
2. The accumulation of small fluctuations in many different directions can produce a large global fluctuation.
3. An event that is an accumulation of rare events may not be rare.
4. Numerical computations and optimizations in high-dimensional spaces can be overly intensive.

For more details, see [Giraud, 2021, Chapter 1].

Outline

Perspectives

1. $\boxed{\text{Week 8 (28/03/2023):}}$ **Tree-Based Methods** (decision tree, bagging, random forests, boosting).

Perspectives

1. $\boxed{\text{Week 8 (28/03/2023):}}$ **Tree-Based Methods** (decision tree, bagging, random forests, boosting).
2. $\boxed{\text{Week 9 (04/04/2023):}}$ **Some Exercises (TD)** for the bonus grade.

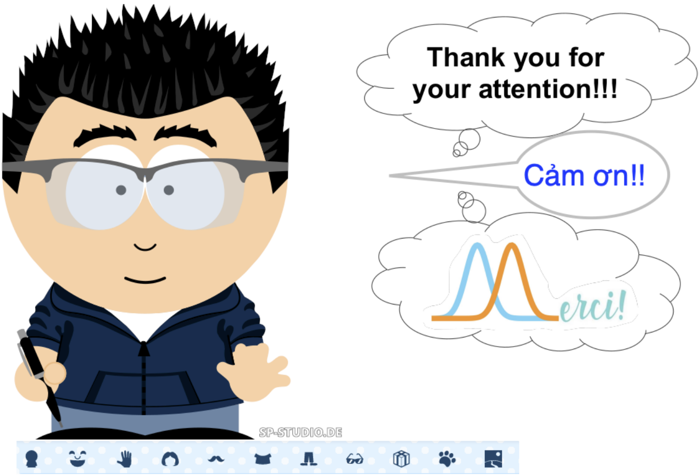
Perspectives

1. $\boxed{\text{Week 8 (28/03/2023):}}$ **Tree-Based Methods** (decision tree, bagging, random forests, boosting).
2. $\boxed{\text{Week 9 (04/04/2023):}}$ **Some Exercises (TD)** for the bonus grade.
3. $\boxed{\text{Week 10 (18/04/2023):}}$ **Some Exercises (TD)** for the bonus grade and **send the Final CC Evaluation** (Deadline 02/05/203).

Perspectives

1. $\boxed{\text{Week 8 (28/03/2023):}}$ **Tree-Based Methods** (decision tree, bagging, random forests, boosting).
2. $\boxed{\text{Week 9 (04/04/2023):}}$ **Some Exercises (TD)** for the bonus grade.
3. $\boxed{\text{Week 10 (18/04/2023):}}$ **Some Exercises (TD)** for the bonus grade and **send the Final CC Evaluation** (Deadline 02/05/203).
4. $\boxed{\text{Week 11 (25/04/2023):}}$ **Last CC with questions**.

“Essentially, all models are wrong, but some are useful”.[9]



↑ This is my best data-driven model to approximate myself.

[9] Box, G. E.P. (1979). “Robustness in the strategy of scientific model building”. In Robustness in Statistics (pp. 201-236). Academic Press.

References

Giraud, C. (2021).
Introduction to High-Dimensional Statistics, volume 2 of *Monographs on Statistics & Applied Probability*.
Taylor & Francis.

(Cited on page 97.)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).
An Introduction to Statistical Learning: with Applications in R, volume 2 of *Springer Texts in Statistics*.
Springer.

(Cited on pages 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, and 93.)