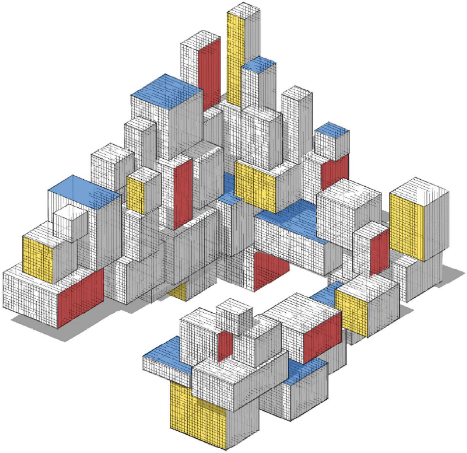
Linear Models



Purpose

In this lecture we continue the analysis of linear models. Topics include:

- Predictive residual sum of squares
- Akaike information criterion for regression models
- Categorical Features
- Nested Models

Cross-Validation and Predictive Residual Sum of Squares

We start by considering the n-fold cross-validation, also called leave-one-out cross-validation, for the linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is an $n \times p$ matrix.

We partition the data into n data sets, leaving out precisely one observation per data set, which we then predict based on the $n - 1$ remaining observations.

Let $\widehat{y}_{-i}$ denote the prediction for the i-th observation using all the data except y_i.

The error in the prediction, $y_i - \widehat{y}_{-i}$, is called a predicted residual — in contrast to an ordinary residual, $e_i = y_i - \widehat{y}_i$.

PRESS

In this way, we obtain the collection of predicted residuals $\{y_i - \widehat{y}_{-i}\}_{i=1}^n$ and summarize them through the predicted residual sum of squares (PRESS):

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_{-i})^2.$$

For linear models, the PRESS can be calculated quickly using only the ordinary residuals and the projection matrix $\mathbf{P} = \mathbf{X}\mathbf{X}^+$

The i-th diagonal element $\mathbf{P}_{ii}$ of the projection matrix is called the i-th leverage, and it can be shown that $0 \leqslant \mathbf{P}_{ii} \leqslant 1$.

Theorem: PRESS for Linear Models

Consider the linear model where the $n \times p$ model matrix $\mathbf{X}$ is of full rank. Given an outcome $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ of $\boldsymbol{Y}$, the fitted values can be obtained as $\widehat{\boldsymbol{y}} = \mathbf{P}\boldsymbol{y}$, where $\mathbf{P} = \mathbf{X}\mathbf{X}^+ = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the projection matrix. If the leverage value $p_i := \mathbf{P}_{ii} \neq 1$ for all $i = 1, \ldots, n$, then the predicted residual sum of squares can be written as

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - p_i} \right)^2,$$

where $e_i = y_i - \widehat{y}_i = y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i$ is the i-th residual.

Polynomial Regression (cont.)

In our running polynomial regression example, we estimated the generalization risk for various polynomial prediction functions using independent validation data.

The following Python code estimates the expected generalization risk via cross-validation (thus only from the training set), using the fast computation of the PRESS.

We find that the PRESS values divided by $n = 100$ for the constant, linear, quadratic, cubic, and quartic order polynomial regression models are, respectively, 152.487, 56.249, 51.606, 30.999, and 31.634.

Hence, the cubic polynomial regression model has the lowest PRESS, indicating that it has the best predictive performance.

polyregpress.py

```
import numpy as np
import matplotlib.pyplot as plt

def generate_data(beta , sig, n):
    u = np.random.rand(n, 1)
    y = u ** np.arange(0, 4) @ beta.reshape(4,1) + (sig * np.random.randn(n, 1))
    return u, y

np.random.seed(12)
beta = np.array([[10.0, -140, 400, -250]]).T;
sig=5; n = 10**2;
u,y = generate_data(beta,sig,n)

X = np.ones((n, 1))
K = 12 #maximum number of parameters
press = np.zeros(K+1)
for k in range(1,K):
    if k > 1:
        X = np.hstack((X, u**(k-1))) # add column to matrix
    P = X @ np.linalg.pinv(X) # projection matrix
    e = y - P @ y

    press[k] = np.sum((e/(1-np.diag(P).reshape(n,1)))**2)

plt.plot(press[1:K]/n)
```

In-Sample Risk

Given a fixed data set τ with associated response vector $\boldsymbol{y}$ and $n \times p$ matrix of explanatory variables $\mathbf{X}$, the in-sample risk of a prediction function g is defined as

$$\ell_{\text{in}}(g) := \mathbb{E}_{\mathbf{X}} \, \text{Loss}(Y, g(\boldsymbol{X})).$$

Here, $\boldsymbol{X}$ takes the values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with equal probability, and given $\boldsymbol{X} = \boldsymbol{x}_i$ the random variable Y is drawn from $f(y \,|\, \boldsymbol{x}_i)$.

The difference between the in-sample risk and the training loss is called the optimism:

$$\text{op}_{\tau} = \ell_{\text{in}}(g_{\tau}) - \ell_{\tau}(g_{\tau}).$$

We showed earlier that (for the squared-error loss) the expected optimism of a learner $g_{\mathcal{T}}$, i.e., $\mathbb{E}_{\mathbf{X}}[\ell_{\text{in}}(g_{\mathcal{T}}) - \ell_{\mathcal{T}}(g_{\mathcal{T}})]$ *for any linear model*, is equal to $2\ell^* p/n$, where $\ell^* = v^2$ is the irreducible risk.

Theorem: Expected In-Sample Risk for Linear Models

Let $\mathbf{X}$ be the model matrix for a linear model, of dimension $n \times p$. If $\mathbb{V}\mathrm{ar}[Y - g^*(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}] =: v^2$ does not depend on $\boldsymbol{x}$, then the expected in-sample risk (with respect to the squared-error loss) for a random learner $g_{\mathcal{T}}$ is given by

$$\mathbb{E}_{\mathbf{X}}\, \ell_{\mathrm{in}}(g_{\mathcal{T}}) = \mathbb{E}_{\mathbf{X}}\, \ell_{\mathcal{T}}(g_{\mathcal{T}}) + \frac{2\ell^* p}{n},$$

where ℓ^* is the irreducible risk.

Model comparison heuristic: Estimate the irreducible risk $\ell^* = v^2$ via $\widehat{v^2}$, using a model with relatively high complexity. Then choose the linear model with the lowest value of

$$\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 + 2\,\widehat{v^2} p.$$

Akaike Information Criterion

We can also use the Akaike information criterion (AIC) as a heuristic for model comparison.

Earlier, we discussed the AIC in the unsupervised learning setting but the arguments used there can also be applied to the supervised case, under the in-sample model for the data.

This leads to the heuristic of selecting the learner $g(\cdot \,|\, \widehat{\boldsymbol{\theta}}_n)$ with the smallest value of the AIC:

$$-2\sum_{i=1}^{n} \ln g_i(y_i \,|\, \widehat{\boldsymbol{\theta}}_n) + 2q,$$

where g_i is the pdf of Y_i under the in-sample model, and q is the total number of model parameters.

Example: AIC for the Normal Linear Model

For the normal linear model $Y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}, \sigma^2)$ with a p-dimensional vector $\boldsymbol{\beta}$, we have

$$g_i(y_i \mid \underbrace{\boldsymbol{\beta}, \sigma^2}_{=\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{\sigma^2}\right), \quad i = 1, \ldots, n,$$

so that the AIC is

$$n \ln(2\pi) + n \ln \widehat{\sigma}^2 + \frac{\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{\widehat{\sigma}^2} + 2q,$$

where $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ is the maximum likelihood estimate and $q = p + 1$ is the number of parameters (including σ^2). For model comparison we may remove the $n \ln(2\pi)$ term if all the models are normal linear models.

Categorical Features

Suppose the data is given in the form of a spreadsheet or data frame with n rows and $p + 1$ columns, where the first element of row i is the response variable y_i, and the remaining p elements form the vector of explanatory variables (features) $\boldsymbol{x}_i^\top$.

So far we have only considered continuous features.
However, linear models with categorical explanatory variables often arise in factorial experiments.

These are controlled statistical experiments in which the aim is to assess how a response variable is affected by one or more factors tested at several levels.

For categorical features, the model matrix is constructed using indicator or dummy variables.

Example: Crop Yield

The data lists the yield of a food crop for four different crop treatments (e.g., strengths of fertilizer) on four different blocks (plots).

Table: Crop yield for different treatments and blocks.

	Treatment			
Block	1	2	3	4
1	9.2988	9.4978	9.7604	10.1025
2	8.2111	8.3387	8.5018	8.1942
3	9.0688	9.1284	9.3484	9.5086
4	8.2552	7.8999	8.4859	8.9485

Data Frame for Crop Yield Data

The corresponding data frame is given below

The values 1, 2, 3, and 4 have no qualitative meaning (it does not make sense to take their average, for example) — they merely identify the category of the treatment or block.

Table: Crop yield data organized as a data frame in standard format.

Yield	Block	Treatment
9.2988	1	1
9.4978	1	2
9.7604	1	3
10.1025	1	4
8.2111	2	1
8.3387	2	2
⋮	⋮	⋮
8.4859	4	3
8.9485	4	4

Indicator Features

Suppose there are r factor (categorical) variables $u_1, \ldots, u_r$, where the j-th factor has p_j mutually exclusive levels, denoted by $1, \ldots, p_j$.

Let $x_{jk} = \mathbb{I}\{u_j = k\}$ be the indicator feature each factor j at level k.

Since $\sum_k \mathbb{I}\{u_j = k\} = 1$, it suffices to consider only $p_j - 1$ of these indicator features for each factor j.

The model for a single response Y is

$$Y = \beta_0 + \sum_{j=1}^{r} \sum_{k=2}^{p_j} \beta_{jk} \underbrace{\mathbb{I}\{u_j = k\}}_{x_{jk}} + \varepsilon,$$

where we have omitted one indicator feature (corresponding to level 1) for each factor j.

For independent responses $Y_1, \ldots, Y_n$ corresponding to the factors $\{u_{1j}\}_{j=1}^{p_1}, \ldots, \{u_{rj}\}_{j=1}^{p_r}$, let $x_{ijk} = \mathbb{I}\{u_{ij} = k\}$. Then, the linear model for the data becomes

$$Y_i = \beta_0 + \sum_{j=1}^{r} \sum_{k=2}^{p_j} \beta_{jk} x_{ijk} + \varepsilon_i,$$

where the $\{\varepsilon_i\}$ are independent with expectation 0 and variance σ^2.

By gathering the β_0 and $\{\beta_{jk}\}$ into a vector $\boldsymbol{\beta}$, and the $\{x_{ijk}\}$ into a matrix $\mathbf{X}$, we have again a linear model of the form $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

We can interpret β_0 as a baseline response when using the explanatory vector $\boldsymbol{x}^\top$ for which $x_{j1} = 1$ for all factors $j = 1, \ldots, r$.

The other parameters $\{\beta_{jk}\}$ can be viewed as incremental effects relative to this baseline effect.

Example: Crop Yield (cont.)

In this case the linear model has eight parameters: $\beta_0, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{22}, \beta_{23}, \beta_{24}$, and σ^2. Ordering $\boldsymbol{y}$ row-wise, i.e., $\boldsymbol{y} = [9.2988, 9.4978, 9.7604, 10.1025, 8.211, \dots, 8.9485]^\top$, the linear model can be written as:

$$\boldsymbol{Y} = \underbrace{\begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{C} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{C} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{C} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \\ \beta_{22} \\ \beta_{23} \\ \beta_{24} \end{bmatrix}}_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \quad \text{where} \quad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and with $\mathbf{1} = [1, 1, 1, 1]^\top$ and $\mathbf{0} = [0, 0, 0, 0]^\top$. Estimation of $\widehat{\boldsymbol{\beta}}$ and σ^2, model selection, and prediction can now be carried out in the usual manner for linear models.

Nested Models

Let $\mathbf{X}$ be a $n \times p$ model matrix of the form $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are model matrices of dimension $n \times k$ and $n \times (p - k)$, respectively.

The linear models $\boldsymbol{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and $\boldsymbol{Y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ are said to be nested within the linear model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Suppose we wish to assess whether to use the full model matrix $\mathbf{X}$ or the reduced model matrix $\mathbf{X}_1$.

Let $\widehat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ under the full model, and let $\widehat{\boldsymbol{\beta}_1}$ denote the estimate of $\boldsymbol{\beta}_1$ for the reduced model.

Let $\boldsymbol{Y}^{(2)} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ be the projection of $\boldsymbol{Y}$ onto the space Span($\mathbf{X}$) spanned by the columns of $\mathbf{X}$; and let $\boldsymbol{Y}^{(1)} = \mathbf{X}_1\widehat{\boldsymbol{\beta}_1}$ be the projection of $\boldsymbol{Y}$ onto the space Span($\mathbf{X}_1$) spanned by the columns of $\mathbf{X}_1$ only.

Projections

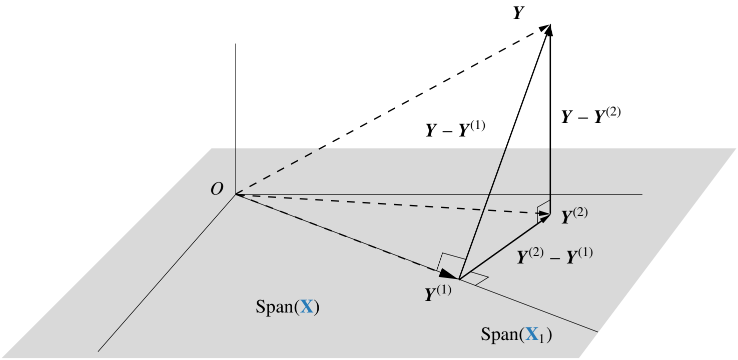

$\boldsymbol{Y}$
$\boldsymbol{Y} - \boldsymbol{Y}^{(2)}$
$\boldsymbol{Y} - \boldsymbol{Y}^{(1)}$
O
$\boldsymbol{Y}^{(2)}$
$\boldsymbol{Y}^{(2)} - \boldsymbol{Y}^{(1)}$
$\boldsymbol{Y}^{(1)}$
Span($\mathbf{X}$)
Span($\mathbf{X}_1$)

Figure: The residual sum of squares for the full model corresponds to $\|\boldsymbol{Y} - \boldsymbol{Y}^{(2)}\|^2$ and for the reduced model it is $\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2$. By Pythagoras's theorem, the difference is $\|\boldsymbol{Y}^{(2)} - \boldsymbol{Y}^{(1)}\|^2$.

Pythagoras

In order to decide whether the features in $\mathbf{X}_2$ are needed, we may compare the estimated error terms of the two models. The comparison is thus between $\|\boldsymbol{Y} - \boldsymbol{Y}^{(2)}\|^2$ and $\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2$.

If there is little difference between the model error for the full and reduced model, then it is appropriate to adopt the reduced model, as it has fewer parameters than the full model, while explaining the data just as well.

Note that $\mathrm{Span}(\mathbf{X}_1)$ is a subspace of $\mathrm{Span}(\mathbf{X})$. Consequently, the orthogonal projection of $\boldsymbol{Y}^{(2)}$ onto $\mathrm{Span}(\mathbf{X}_1)$ is the same as the orthogonal projection of $\boldsymbol{Y}$ onto $\mathrm{Span}(\mathbf{X}_1)$; that is, $\boldsymbol{Y}^{(1)}$. By Pythagoras' theorem, we thus have the decomposition

$$\|\boldsymbol{Y}^{(2)} - \boldsymbol{Y}^{(1)}\|^2 + \|\boldsymbol{Y} - \boldsymbol{Y}^{(2)}\|^2 = \|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2.$$

General Nested Models

More generally, suppose the model matrix can be decomposed into d submatrices: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_d]$, where the matrix $\mathbf{X}_i$ has p_i columns and n rows, $i = 1, \dots, d$.

This creates an increasing sequence of "nested" model matrices: $\mathbf{X}_1, [\mathbf{X}_1, \mathbf{X}_2], \dots, [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_d]$, from (say) the baseline normal model matrix $\mathbf{X}_1 = \mathbf{1}$ to the full model matrix $\mathbf{X}$.

Now, project $\boldsymbol{Y}$ onto Span($\mathbf{X}$) to yield the vector $\boldsymbol{Y}^{(d)}$, then project $\boldsymbol{Y}^{(d)}$ onto Span($[\mathbf{X}_1, \dots, \mathbf{X}_{d-1}]$) to obtain $\boldsymbol{Y}^{(d-1)}$, and so on, until $\boldsymbol{Y}^{(2)}$ is projected onto Span($\mathbf{X}_1$) to yield $\boldsymbol{Y}^{(1)} = \overline{Y}\mathbf{1}$.

The total sum of squares can now be decomposed as

$$\underbrace{\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2}_{\text{df}=n-p_1} = \underbrace{\|\boldsymbol{Y} - \boldsymbol{Y}^{(d)}\|^2}_{\text{df}=n-p} + \underbrace{\|\boldsymbol{Y}^{(d)} - \boldsymbol{Y}^{(d-1)}\|^2}_{\text{df}=p_d} + \cdots + \underbrace{\|\boldsymbol{Y}^{(2)} - \boldsymbol{Y}^{(1)}\|^2}_{\text{df}=p_2}.$$

Software packages typically report the sums of squares as well as the corresponding degrees of freedom (df): $n - p, p_d, \dots, p_2$.

Coefficient of Determination

To assess how a linear model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ compares to the default model $\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{\varepsilon}$, we can compare the variance of the original data, estimated via $\sum_i (Y_i - \overline{Y})^2/n = \|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|^2/n$, with the variance of the fitted data; estimated via $\sum_i (\widehat{Y}_i - \overline{Y})^2/n = \|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}\|^2/n$, where $\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

The sum $\sum_i (Y_i - \overline{Y})^2/n = \|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|^2$ is sometimes called the total sum of squares (TSS), and the quantity

$$R^2 = \frac{\|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}\|^2}{\|\boldsymbol{Y} - \overline{Y}\mathbf{1}\|^2}$$

is called the coefficient of determination of the linear model. In the notation of Figure 1, $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}^{(2)}$ and $\overline{Y}\mathbf{1} = \boldsymbol{Y}^{(1)}$, so that

$$R^2 = \frac{\|\boldsymbol{Y}^{(2)} - \boldsymbol{Y}^{(1)}\|^2}{\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2} = \frac{\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2 - \|\boldsymbol{Y} - \boldsymbol{Y}^{(2)}\|^2}{\|\boldsymbol{Y} - \boldsymbol{Y}^{(1)}\|^2} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

Adjusted R^2

Note that R^2 lies between 0 and 1. An R^2 value close to 1 indicates that a large proportion of the variance in the data has been explained by the model.

Many software packages also give the adjusted coefficient of determination, or simply the adjusted R^2, defined by

$$R^2_{\text{adjusted}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

The regular R^2 is always *non-decreasing* in the number of parameters, but this may not indicate better predictive power.

The adjusted R^2 compensates for this increase by decreasing the regular R^2 as the number of variables increases. This heuristic adjustment can make it easier to compare the quality of two competing models.