Regression

Maxim Panov

Skoltech

November, 2021

Skoltech

Outline

Linear Regression

Maximum Likelihood Estimation

Prediction

Multiple Regression

Logistic regression

Model Selection

Design of Experiments

Regression Modelling

Regression is a method for studying the relationship between a response variable Y and a covariate X (also predictor variable or a feature).

One way to summarize the relationship is to calculate

$$r(x) = \mathbb{E}(Y \mid X = x) = \int y \; f(y \mid x) dy.$$

Our goal is to estimate the regression function $r(x)$ from data of the form

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{XY},$$

where F_{XY} is a joint distribution of X and Y.

Linear regression

Linear regression:

$$r(x) = \beta_0 + \beta_1 x.$$

Simple linear regression problem:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where ε_i is i.i.d. noise with expectation $\mathbb{E}(\varepsilon_i \mid X_i) = 0$ and variance $\mathbb{V}(\varepsilon_i \mid X_i) = \sigma^2$.

Tasks:

- Parameter estimation: $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$
- Prediction: $\widehat{Y}_i = \widehat{r}(X_i).$

Ordinary Least Squares

Residuals:

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i).$$

Residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2.$$

Definition

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are least squares estimates of unknown parameters (OLS), if the RSS for these estimates is minimal.

Ordinary Least Squares

Theorem

The least squares estimates of β_0 and β_1 are given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$

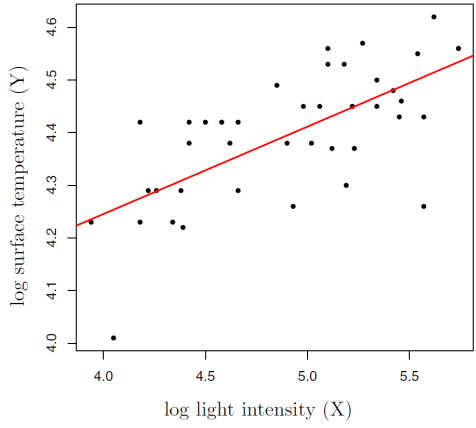
$$\widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{X}_n.$$

The unbiased estimate of noise variance σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \widehat{\varepsilon}_i^2.$$

Example: Linear Regression

The data about nearby stars: estimates of temperature and brightness.



Estimates are equal to $\widehat{\beta}_0 = 3.58$ and $\widehat{\beta}_1 = 0.166 \Rightarrow \widehat{r}(x) = 3.58 + 0.166x$.

Linear Regression

Maximum Likelihood Estimation

Prediction

Multiple Regression

Logistic regression

Model Selection

Design of Experiments

Method for Estimation Based on Likelihood Maximization

Let $\varepsilon_i \mid X_i \sim \mathcal{N}(0, \sigma^2)$. Then

$$Y_i \mid X_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{where } \mu_i = \beta_0 + \beta_1 X_i.$$

The likelihood function is

$$\begin{aligned}\mathcal{L}_0 &= \prod_{i=1}^{n} f(X_i, Y_i) = \prod_{i=1}^{n} f_X(X_i)\ f_{Y|X}(Y_i \mid X_i) \\ &= \prod_{i=1}^{n} f_X(X_i) \times \prod_{i=1}^{n} f_{Y|X}(Y_i \mid X_i) = \mathcal{L}_1 \times \mathcal{L}_2,\end{aligned}$$

where

$$\mathcal{L}_1 = \prod_{i=1}^{n} f_X(X_i), \qquad \mathcal{L}_2 = \prod_{i=1}^{n} f_{Y|X}(Y_i \mid X_i)$$

Method for Estimation Based on Likelihood Maximization

The term $\mathcal{L}_1$ does not involve the parameters β_0 and β_1.

Let's focus on $\mathcal{L}_2$, i.e. conditional likelihood:

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{Y|X}(Y_i \mid X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2\right\},$$

$$\ell(\beta_0, \beta_1, \sigma) \equiv \log \mathcal{L}(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(Y_i - (\beta_0 + \beta_1 X_i)\big)^2. \qquad (1)$$

MLE (β_0, β_1) $\Leftrightarrow$ maximization (1) $\Leftrightarrow$ minimization RSS:

$$RSS = \sum_{i=1}^{n}\big(Y_i - (\beta_0 + \beta_1 X_i)\big)^2.$$

Properties of MLE Estimates

Theorem

Let $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ *be the least squares estimate. Then*

$$\mathbb{E}(\widehat{\beta} \mid X^n) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \mathbb{V}(\widehat{\beta} \mid X^n) = \frac{\sigma^2}{n s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\bar{X}_n \\ -\bar{X}_n & 1 \end{pmatrix},$$

$$\text{where} \quad s_X^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Thus,

$$\widehat{se}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X \sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}, \qquad \widehat{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X \sqrt{n}}.$$

Properties of MLE Estimates

Theorem

Under appropriate regularity conditions[a]*:*

1. $\widehat{\beta}_0 \xrightarrow{\mathbb{P}} \beta_0,\ \widehat{\beta}_1 \xrightarrow{\mathbb{P}} \beta_1.$
2. $\frac{\widehat{\beta}_0-\beta_0}{\widehat{se}(\widehat{\beta}_0)} \rightsquigarrow \mathcal{N}(0,1),\ \frac{\widehat{\beta}_1-\beta_1}{\widehat{se}(\widehat{\beta}_1)} \rightsquigarrow \mathcal{N}(0,1).$
3. *Approximate* $1-\alpha$ *confidence intervals for parameters:*

$$\widehat{\beta}_0 \pm z_{\alpha/2}\widehat{se}(\widehat{\beta}_0) \quad \textit{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\widehat{se}(\widehat{\beta}_1).$$

4. *The Wald test for testing* $H_0\colon \beta_1 = 0$ *vs.* $H_1\colon \beta_1 \neq 0$ *has the form:* H_0 *is rejected if* $|W| > z_{\alpha/2}$*, where* $W = \widehat{\beta}_1/\widehat{se}(\widehat{\beta}_1)$.

[a] consider Bilodeau, Brenner, Theory of Multivariate Statistics

Linear Regression

Maximum Likelihood Estimation

Prediction

Multiple Regression

Logistic regression

Model Selection

Design of Experiments

Prediction

Model: $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ constructed by OLS based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Predict outcome Y_* for covariate $X = x_*$:

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*.$$

$$\mathbb{V}(\widehat{Y}_*) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = \mathbb{V}(\widehat{\beta}_0) + x_*^2 \mathbb{V}(\widehat{\beta}_1) + 2x_* \operatorname{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)$$

$\Rightarrow$ to estimate $\widehat{se}(\widehat{Y}_*)$ use $\widehat{\sigma}^2$ in place of σ^2.

Can we construct the confidence interval for Y_* using the usual form:

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{se}(\widehat{Y}_*)?$$

Prediction

Theorem

Let

$$\widehat{\xi}_n^2(x_*) = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - x_*)^2}{n \sum_{i=1}^n (X_i - \bar{X})^2} + 1 \right).$$

An approximate $1 - \alpha$ *prediction interval for* Y_* *is*

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{\xi}_n(x_*).$$

Multiple Regression

In this case, the data are of the form

$$(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n),$$

$$X_i = (X_{i1}, \ldots, X_{ik}) \in \mathbb{R}^k.$$

For $i = 1, \ldots, n$ the model is

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i,$$

$$\mathbb{E}(\varepsilon_i \mid X_{i1}, \ldots, X_{ik}) = 0.$$

To include an intercept, one usually assumes $X_{i1} = 1$ for $i = 1, \ldots, n$.

Multiple Regression

Let us define

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

As a result we obtain an equation in the matrix form:

$$Y = X\beta + \varepsilon.$$

Multiple Regression

Theorem

Assume that a matrix X^TX of size $k \times k$ is invertible,

$$\widehat{\beta} = (X^TX)^{-1}X^TY,$$
$$\mathbb{V}(\widehat{\beta} \mid X^n) = \sigma^2(X^TX)^{-1},$$
$$\widehat{\beta} \sim \mathcal{N}(\beta_*, \sigma^2(X^TX)^{-1}).$$

The estimated regression function is

$$\widehat{r}(x) = \sum_{j=1}^{k} \widehat{\beta}_j x_j, \qquad \widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2,$$

$\widehat{\varepsilon} = X\widehat{\beta} - Y$ *is a residual vector.*

Confidence Intervals

An approximate $1-\alpha$ confidence interval for β_j is

$$\widehat{\beta}_j \pm z_{\alpha/2}\widehat{se}(\widehat{\beta}_j),$$

where $\widehat{se}^2(\widehat{\beta}_j)$ is a j-th diagonal element of the matrix $\widehat{\sigma}^2(X^TX)^{-1}$.

Example: Multiple Regression

USA crime data on 47 states in 1960
http://lib.stat.cmu.edu/DASL/Stories/USCrime.html

Covariate	$\widehat{\beta}_j$	$\widehat{se}(\widehat{\beta}_j)$	W_j	p-value
Intercept	-589.39	167.59	-3.51	0.001
Age	1.04	0.45	2.33	0.025
Southern State	11.29	13.24	0.85	0.399
Education	1.18	0.68	1.7	0.093
Expenditures	0.96	0.25	3.86	0.000
Labor	0.11	0.15	0.69	0.493
Number of Males	0.30	0.22	1.36	0.181
Population	0.09	0.14	0.65	0.518
Unemployment (14-24)	-0.68	0.48	-1.4	0.165
Unemployment (25-39)	2.15	0.95	2.26	0.030
Wealth	-0.08	0.09	-0.91	0.367

Logistic Regression

So far it was assumed that Y_i takes real values.

Logistic regression is a parametric method for regression when $Y_i \in \{0, 1\}$. For a k - dimensional regressor, the model has the form

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1 \mid X = x) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}.$$

Let's introduce function $\operatorname{logit}(p) = \log \frac{p}{1-p}$. Then

$$\operatorname{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}.$$

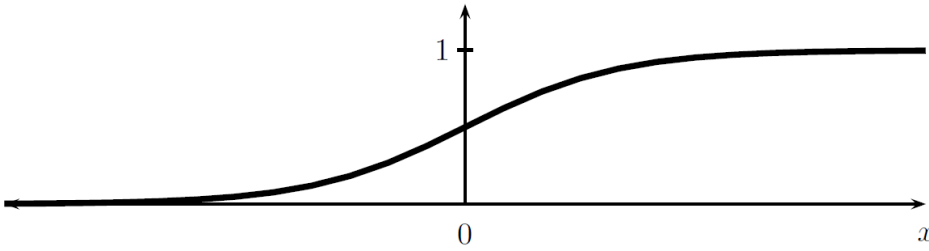
Logistic Regression

The name "logistic regression" comes from the fact that class probabilities

$$Y = 1 \quad \text{vs.} \quad Y = 0$$

are modeled using a logistic function:

$$p(x) = \frac{e^x}{1 + e^x}.$$



Maximum Likelihood Maximization Estimation

Since $Y_i \in \{0, 1\}$, then $Y_i \mid X = x \sim Bernoulli(p_i)$.

Hence, the conditional likelihood function is

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i} \big(1 - p_i(\beta)\big)^{1-Y_i}.$$

The MLE $\widehat{\beta}$ has to be obtained by maximizing $\mathcal{L}(\beta)$ numerically.

Example: Logistic Regression

Coronary heart disease data. 462 males, 15 – 64 years, from South Africa.

Output: the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease.

9 covariates (regressors):

- systolic blood pressure,
- cumulative tobacco (kg),
- ldl (low density lipoprotein cholesterol),
- adiposity,
- famhist (family history of heart disease),
- typea (type-A behavior),
- obesity,
- alcohol (current alcohol consumption)
- age.

Example: Logistic Regression

Covariate	$\widehat{\beta}_j$	$\widehat{se}$	W_j	p-value
Intercept	-6.145	1.300	-4.738	0.000
Blood pressure	0.007	0.006	1.138	0.255
Tobacco	0.079	0.027	2.991	0.003
ldl	0.174	0.059	2.925	0.003
Adiposity	0.019	0.029	0.637	0.524
famhist	0.925	0.227	4.078	0.000
Type A	0.040	0.012	3.233	0.001
Obesity	-0.063	0.044	-1.427	0.153
Alcohol	0.000	0.004	0.027	0.979
Age	0.045	0.012	3.754	0.000

Example: Logistic Regression

- Systolic blood pressure is not significant?!
- The minus sign for the obesity coefficient?!

This is caused by the dependence of regressors.

- The result doesn't mean that high systolic blood pressure is not an important marker of coronary disease.
- It just means that this covariate is not "important" compared to the other variables in the model.

Linear Regression

Maximum Likelihood Estimation

Prediction

Multiple Regression

Logistic regression

Model Selection

Design of Experiments

Model Selection

Occam's razor – "entities should not be multiplied without necessity". Many variables lead to a large variance in the forecast, but a small bias, and vice versa.

In model selection there are two problems:

1. assigning a "score" to each model which measures, in some sense, how good the model is;

2. searching through all the models to find the model with the best score.

Notations

Let $S \subset \{1, \ldots, k\}$ – a subset of the covariates.

Then

- β_S – coefficients for corresponding covariates, $\widehat{\beta}_S$ – their least squares estimate;
- X_S – the X matrix for this subset of covariates;
- $\widehat{r}_S(x)$ – estimated regression function;
- $\widehat{Y}_i(S) = \widehat{r}_S(X_i)$ – predicted values.

Prediction Risk

Prediction risk:

$$R(S) = \sum_{i=1}^{n} \mathbb{E}\big(\widehat{Y}_i(S) - Y_i^*\big)^2,$$

where $Y_i^* = X_i\beta$ is the true output value for X_i.

The problem is to select a subset of S that minimizes $R(S)$.

Estimate of Prediction Risk

Prediction risk estimation (error on the training sample):

$$\widehat{R}_{tr}(S) = \sum_{i=1}^{n} \big(\widehat{Y}_i(S) - Y_i\big)^2.$$

Theorem

Prediction risk estimation is biased compared to the actual prediction risk value:

$$\operatorname{bias}\big(\widehat{R}_{tr}(S)\big) = \mathbb{E}\widehat{R}_{tr}(S) - R(S) = -2\sum_{i=1}^{n} \operatorname{Cov}\big(\widehat{Y}_i(S), Y_i\big).$$

Estimate of Prediction Risk

- The reason for the bias is that the data are being used twice – to estimate the parameters and to estimate the risk.
- When there are many parameters, $\mathrm{Cov}(\widehat{Y}_i(S), Y_i)$ takes a large value.
- In this case, prediction for data other than the data in the training sample may be significantly worse!

C_p Mallow Statistics

C_p Mallow Statistics:

$$\widehat{R}(S) = \widehat{R}_{tr}(S) + 2|S|\widehat{\sigma}^2,$$

where $|S|$ is the number of terms, $\widehat{\sigma}^2$ – noise variance estimate obtained from the full model (i.e. including all covariates).

The criterion includes an assessment of the prediction risk on the training sample and the "complexity" of the model (regularization).

AIC

AIC (Akaike Information Criterion):

$$AIC(S) = \ell_S - |S| \to \max_S,$$

where $\ell_S = \ell_S(\widehat{\beta})$ is the log-likelihood of the model where unknown parameters were replaced with their MLE estimates.

In linear regression with normal errors (with noise variance equal to its estimate from the full model), maximizing AIC is equivalent to minimizing C_p.

Cross-Validation

Risk estimation using cross-validation (leave-one-out):

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\widehat{Y}_{(i)}(S) - Y_i\right)^2,$$

where $\widehat{Y}_{(i)}(S)$ is a prediction for Y_i obtained by fitting the model without i^{th} input.

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{\widehat{Y}_i(S) - Y_i}{1 - U_{ii}(S)}\right)^2,$$
$$U(S) = X_S (X_S^T X_S)^{-1} X_S^T.$$

K-fold Cross-Validation

1. Data is randomly divided into k disjoint groups (often taking $k = 10$).
2. One group is omitted at a time (with a return), and model is fitted with the remaining data.
3. Estimate the risk by $\sum_i (\widehat{Y}_i - Y_i)^2$ (fitted model is used to predict data in omitted group, sum over observations in omitted group).
4. Process is repeated for the remaining groups, finally, the resulting risk score is averaged.

For linear regression, Mallows C_p and K-fold cross-validation often yield essentially the same results. In more complex cases, cross-validation will be more useful.

BIC

BIC (Bayesian information criterion):

$$BIC(S) = \ell_S - \frac{|S|}{2}\log n \to \max_S .$$

This functional has a Bayesian interpretation.

- Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ – a set of possible models.
- Assume that the priori over models has the form $\mathbb{P}(S_j) = 1/m$.
- Also assume that the parameters within each model have some smooth prior.
- It can be shown that the posterior probability for a model is approximately

$$\mathbb{P}(S_j \mid \text{sample}) \approx \frac{\exp\big(BIC(S_j)\big)}{\sum_{r=1}^{m}\exp\big(BIC(S_r)\big)}.$$

BIC

Thus, choosing the model with the highest BIC is equivalent to choosing the model with the highest posterior probability.

BIC score also has an information-theoretic interpretation in terms of something called minimum description length: BIC usually selects models with fewer parameters.

Bruteforce

- If the maximum number of covariates in a model is k, then there are 2^k possible models.
- Ideally, you need to search through all these models, assign a score to each one, and select the best model according to the score.
- When there are a large number of covariates, **forward and backward stepwise regression** is used to reduce the complexity.

Forward and Backward Stepwise Regression

- **Forward stepwise regression:**
  - start with no covariates in the model;
  - then add the one variable that leads to the best score;
  - continue adding variables one at a time until the score does not improve.

- **Backwards stepwise regression:**
  - start with the biggest model;
  - drop one variable at a time that leads to the best score;
  - continue dropping variables one at a time until the score does not improve.

Other Approaches to Models Selection

The other popular approaches to model selection are mostly based on the idea of regularization:

$$\widehat{R}(\beta) = \widehat{R}_{tr}(\beta) + \lambda \ \mathrm{pen}(\beta)$$

for some $\lambda > 0$ and penalty function $\mathrm{pen}(\beta)$.

Examples:

- Ridge regression:

$$\mathrm{pen}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{k} \beta_j^2;$$

- LASSO:

$$\mathrm{pen}(\beta) = \|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|.$$

Linear Regression

Maximum Likelihood Estimation

Prediction

Multiple Regression

Logistic regression

Model Selection

Design of Experiments

Back to Linear Regression Model

Linear Regression Model:

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k} \beta_j x_j,$$

$$\mathbf{x} = (x_1, \ldots, x_k).$$

Parameters of linear regression model are estimated based on

$$D = (X, Y = f(X))$$

for given $X = (\mathbf{x}_i)_{i=1}^{N}$.

Question: how to choose design X if are allowed to do so?

Optimal Designs for Linear Models

Response Surface Model (RSM):

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k} \beta_j x_j = \mathbf{x}^T \beta.$$

$$\mathbf{x} = (x_1, \ldots, x_k).$$

Definition

Optimal design $X = (\mathbf{x}_i)_{i=1}^N$ is the one minimizing:

1. variance of parameter estimates,
2. variance of the prediction,
3. ...

Optimal designs decrease the cost of experiments allowing to estimate the model parameters based on smaller number of samples.

Optimal Designs for Linear Models

- If the noise is Gaussian, then it holds $\hat{\beta} \sim \mathcal{N}(\beta, \mathrm{Cov}(\hat{\beta} \mid X))$.
- Covariance matrix for $\hat{\beta}$ is given by

$$\mathrm{Cov}(\hat{\beta} \mid X) \sim \left(X^T X\right)^{-1}.$$

- Similarly for the prediction we get $\hat{f}(\mathbf{x}_0) \sim \mathcal{N}(f(\mathbf{x}_0), \mathrm{Var}(\hat{f} \mid X))$.
- Variance of prediction $\hat{f}(\mathbf{x}_0) = \mathbf{x}_0^T \hat{\beta}$ at arbitrary point $\mathbf{x}_0$:

$$\mathrm{Var}(\hat{f}(\mathbf{x}_0) \mid X) \sim \mathbf{x}_0 \left(X^T X\right)^{-1} \mathbf{x}_0^T.$$

D-optimality

- D-optimal design aims to minimize the determinant of the covariance matrix:

$$\det\left[\left(X^T X\right)^{-1}\right] \to \min_X .$$

- D-optimal design approximately minimizers the variance of parameter estimates.

- It is also known under the name *MaxVol*.

IV-optimality

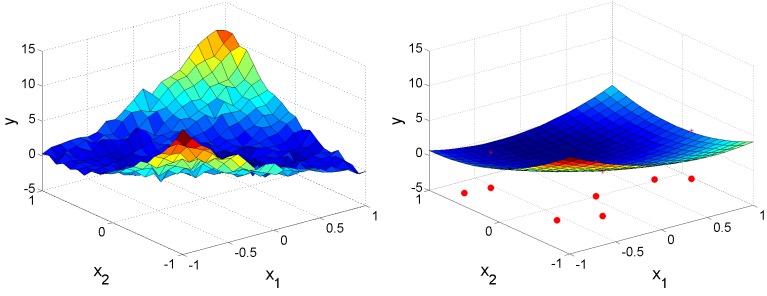
- IV-optimal design minimizes the average variance of the model prediction:

$$\int_{\mathbb{X}} \mathbf{x}^T \left(X^T X\right)^{-1} \mathbf{x}\, d\mathbf{x} \to \min_X .$$

- The integral is usually estimated via Monte-Carlo estimation.

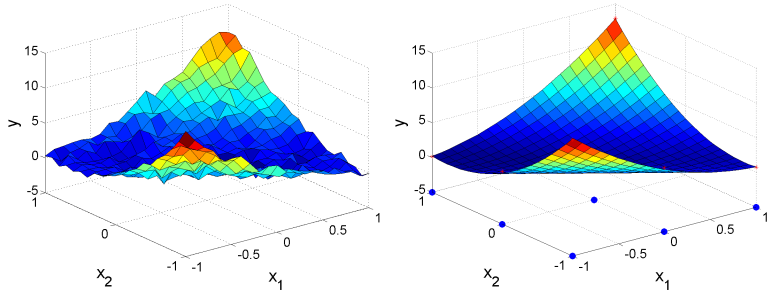
Example 1

True function with noise (left) and quadratic regression model based on random DoE (right).



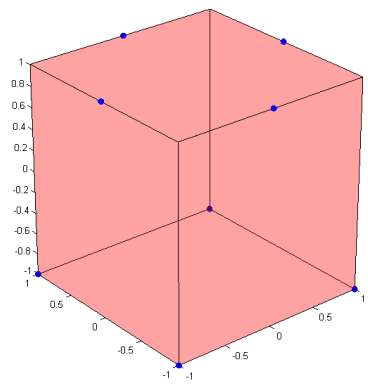
Example 1

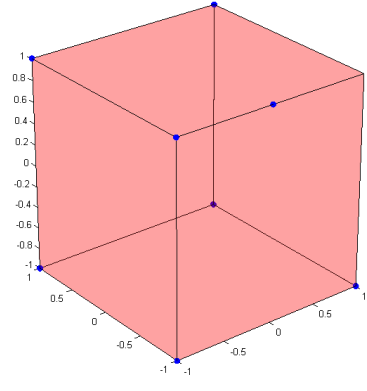
True function with noise (left) and quadratic regression model based on D-optimal design (right).



Example 2

Optimal designs in $\mathbb{X} \subset \mathbb{R}^3$:





DoE for Linear Models

Pros

- Optimal for linear models.
- In practice usually gives very reasonable designs.

Cons

- Proven optimality only for models with the a-priory specified structure.

Thank you for your attention!