Statistical Learning 1

RG chapters 1 and 2

Jakob Söhl

Delft University of Technology

Part I

Course information

Course material and software

Learning goals:

1. Get familiar with classical and modern methods in **data science** (statistics, machine learning, signal processing...).
2. Rather than mechanically applying methods to some datasets, try to **understand** methods (weaknesses/strengths).
3. Focus on methods based on a statistical model that allow for **uncertainty quantification**. Probabilistic approach to statistics/machine learning.

Course materials:

- Book: *A first course in machine learning, 2nd edition* by Rogers and Girolami (refer to as **RG**).
- Slides

Topics

Rough plan:

- Linear models
- Bayesian statistics
- Bayesian analysis of linear models
- Classification
- Bayesian computation
- Gaussian process regression

Assessment

Three assignments which will be corrected with grade $\in \{-, 0, +\}$.

Software:

- You are free to choose.
- Book is accompanied by R and Matlab scripts. `https://github.com/sdrogers/fcmlcode`
- Code will be in Julia and R.

Exam regulations:

- Three hour written exam, entrance requires all assignments to be $+$.
- You are allowed to resubmit assignments graded 0 for a second time.
- Once you are allowed to resubmit an assignment graded $-$ for a second time.

Part II

Linear Models

Intermezzo: Multivariate Normal distribution

Introductory example from RG

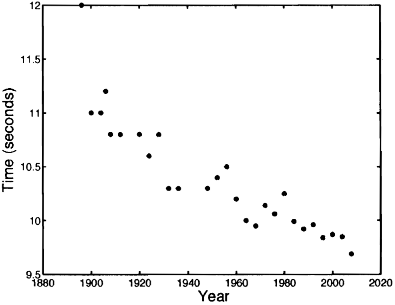


FIGURE 1.1: Winning men's 100 m times at the Summer Olympics since 1896. Note that the two world wars interrupted the games in 1914, 1940 and 1944.

Goal: learn a model of the functional dependence between Olympics year and 100m winning time and use this model to make predictions about winning times in future games

Linear models

The linear model is defined by

$$y = X\theta + \epsilon,$$

where we assume $X \in \mathbb{R}^{n \times p}$ and ε models “noise”.

Example: fit a parabola through a cloud of points

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Least squares solution

“Forget about the noise, forget about probability and statistics...”

The system

$$y = X\theta$$

has no solution.

Least squares:

- find θ minimising $\theta \mapsto \|y - X\theta\|^2$.
- Linear algebra course: solution satisfies

$$X^T X\theta = X^T y.$$

Why $X^T X\theta = X^T y$? Recap from Linear Algebra

Best approximation theorem: If W is a subspace of $\mathbb{R}^n$, then the closest point to y within W is given by $\text{proj}_W y$.

- For any $\theta \in \mathbb{R}^p$, $X\theta \in \text{Col}X$,
  so we look for the closest point to y in $\text{Col}X$.
- Hence, $\hat{\theta}$ satisfies
  $$\text{proj}_{\text{Col}X} y = X\hat{\theta}.$$
- But then $y - X\hat{\theta} \perp \text{Col}X$. That is, for *all* column vectors X_j of X we have
  $$X_j^T(y - X\hat{\theta}) = 0.$$
- Equivalently, $X^T(y - X\hat{\theta}) = 0$.

When is the solution unique?

If X has full column-rank (equivalently, $\mathsf{N}(X) = \{0\}$), then the LS-solution is unique and given by

$$\hat{\theta} = (X^T X)^{-1} X^T y.$$

If $p > n$ then uniqueness fails!

Linearity is about the parameter

We can take

$$y_i = \sum_{j=1}^{3} \theta_j h_j(x_i)$$

with

$$\begin{aligned} h_1(x) &= 1 \\ h_2(x) &= x \\ h_3(x) &= \sin\left(\frac{x-a}{b}\right), \end{aligned}$$

for known a and b.

Still gives a linear model.

A statistical model

Informally:

- A **statistical experiment** is an experiment with uncertain outcome.
- A **random quantity** is the outcome within an experiment with uncertainty.
  1. Scalar valued outcome: random variable.
  2. Vector valued outcome: random vector.
  3. Function valued outcome: random process.

  Instead of “random”, the word “stochastic” is often used.
- Probability theory provides a language to describe stochasticity (uncertainty).
- **Stochastic modelling**: describing how we believe the data that we have could be generated (RG, sections 2.1.1 and 2.7 “Thinking generatively”).

Turning LS to a statistical model

Instead of

$$y = X\theta$$

consider

$$y = X\theta + \varepsilon$$

Idea: not all points satisfy the equation exactly, so to generate data like the data we have, we add some noise.

Most common choice for noise: **Multivariate Normal (Gaussian) distribution.**

Intermezzo: notation

For a random variable X, we write f_X to denote either its probability mass or density function.

When evaluated in u, we write $f_X(u)$.

Hence, the density of X^2 evaluated in u is $f_{X^2}(u)$.

We abbreviate

$$p(x) = f_X(x)$$

This is sometimes called **Bayesian notation**.

Note that then $p(x^2) = f_{X^2}(x^2)$, and not $f_X(x^2)$ or $f_{X^2}(x)$.

If there is risk of confusion: don't use Bayesian notation.

Intermezzo: random vectors

Simply stack random variables into a vector!

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbb{E}Y = \begin{bmatrix} \mathbb{E}Y_1 \\ \vdots \\ \mathbb{E}Y_n \end{bmatrix}$$

The covariance matrix is defined by

$$\mathsf{Cov}Y = \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T].$$

Hence, for a compatible matrix A

$$\mathbb{E}[AY] = A\mathbb{E}Y \qquad \text{and} \qquad \mathsf{Cov}(AY) = A\left(\mathsf{Cov}Y\right)A^T.$$

Multivariate Normal (Gaussian) distribution

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent $N(0,1)$-distributed random variables. Let

$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \cdots & \varepsilon_n \end{bmatrix}^T .$$

Then

$$\begin{aligned} p(\varepsilon_1, \ldots, \varepsilon_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\varepsilon_i^2\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|\varepsilon\|^2\right). \end{aligned}$$

We write $\varepsilon \sim N(0, I)$.

Normal distribution in $\mathbb{R}^2$: the basic idea

- Start with independent random variables $U \sim N(0,1)$ and $V \sim N(0,1)$.
- Take $\rho \in [-1,1]$ and put

$$\begin{aligned} X_1 &= \sqrt{1-\rho^2}U + \rho V \\ X_2 &= \phantom{\sqrt{1-\rho^2}U + \rho} V \end{aligned}$$

Then

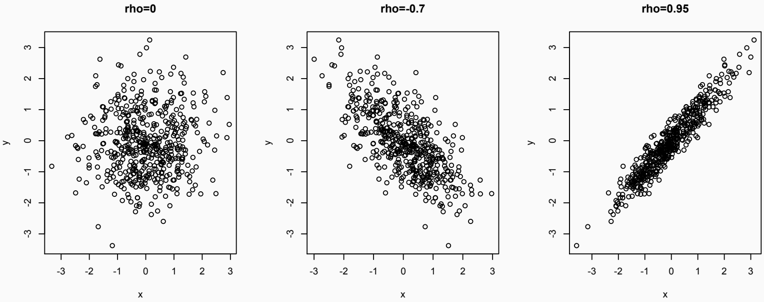
$$X_1 \sim N(0,1) \qquad \text{and} \qquad X_2 \sim N(0,1)$$

and

$$\rho(X_1, X_2) = \rho.$$

Bivariate normal distribution (simple version)

Simulations in which we sample 500 times from the distribution of the example for various values of ρ.



The figures illustrates that ρ measures *linear dependence*.

Multivariate Normal (Gaussian) distribution

If $\varepsilon \sim N(0, I)$, define

$$Z = \mu + L\varepsilon.$$

We write $Z \sim N(\mu, \Sigma)$, where $\Sigma = LL^T$.

If Σ is nonsingular, then

$$p(z) = (2\pi)^{-n/2} |\det \Sigma|^{-1/2} \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu) \right).$$

where $z = \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}^T$.

- μ: mean vector
- Σ: covariance matrix
- Σ^{-1}: precision matrix

Canonical Multivariate Normal distribution

The distribution can also be parametrised by (v, P) with

1. precision $P = \Sigma^{-1}$
2. potential $v = \Sigma^{-1}\mu$

We write $N^{\mathrm{can}}(v, P)$.

Density in terms of (μ, Σ):

$$p(z) = (2\pi)^{-n/2}|\det\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right).$$

Density in terms of (v, P):

$$p(z) = (2\pi)^{-n/2}|\det P|^{1/2}\exp\left(-\frac{1}{2}z^TPz + z^Tv - \frac{1}{2}v^TP^{-1}v\right).$$

The linear model

Consider

$$y = X\theta + \epsilon$$

with $\epsilon \sim N(0, \Sigma)$.

Why the multivariate normal? Suppose $Z \sim N_n(\mu, \Sigma)$

- Central Limit Theorem suggests it is good for modelling accumulated noise effect.
- Tractability, so many nice properties:
  1. Σ is diagonal if and only if the components $Z_1, \ldots, Z_n$ are statistically independent and Normally distributed.
  2. $Z \sim N_n(\mu, \Sigma)$ if and only if for any $a \in \mathbb{R}^n$ the random variable $a^T Z$ has distribution $N(a^T \mu, a^T \Sigma a)$.
  3. If $A \in \mathbb{R}^{k \times n}$, then $AZ \sim N_k(A\mu, A\Sigma A^T)$.
  4. If $(X, Y) \sim N_n(\mu, \Sigma)$, then $X \mid Y = y$ also has a multivariate normal distribution.

Parameter estimation for the linear model

Consider

$$y = X\theta + \epsilon$$

with $\epsilon \sim N(0, \Sigma)$.

The **likelihood** is defined as the probability density of the data:[1]

$$L(\theta, \Sigma; y) \overset{\text{def}}{=} p(y; \theta, \Sigma).$$

Maximum Likelihood Estimator (MLE) is defined by

$$(\hat{\theta}, \hat{\Sigma}) = \underset{\theta, \Sigma}{\operatorname{argmax}} L(\theta, \Sigma; y)$$

(if it exists..., Σ should be symmetric and positive-definite...)

[1] For the linear model this becomes $L(\theta, \Sigma; y) = \varphi(y; X\theta, \Sigma)$, where $\varphi(x; \mu, \Upsilon)$ denotes the density of the $N(\mu, \Upsilon)$-distribution, evaluated at x.

Loglikelihood

$$\ell(\theta, \Sigma; y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\det\Sigma| - \frac{1}{2}(y - X\theta)^T\Sigma^{-1}(y - X\theta).$$

Convenient assumption: $\Sigma = \sigma^2 I$. Then

- $\hat{\theta}$ minimises $\theta \mapsto \|y - X\theta\|^2$ (so equals the LS solution).
- $\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\theta}\|^2$.
- The Hessian matrix of ℓ, containing all 2nd order derivatives with respect to θ, equals $-\sigma^{-2}X^TX$. In notation from RG:

$$\frac{\partial^2 \ell}{\partial\theta\partial\theta^T} = -\sigma^{-2}X^TX.$$

Maximum likelihood favours complex models

If

$$\ell(\theta, \Sigma; y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\det\Sigma| - \frac{1}{2}(y - X\theta)^T\Sigma^{-1}(y - X\theta)$$

then

$$\begin{aligned} \ell(\hat{\theta}, \hat{\Sigma}; y) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2 \\ &= -\frac{n}{2}(1 + \log(2\pi) + \log\hat{\sigma}^2). \end{aligned}$$

So if $\hat{\sigma}^2$ can be decreased, the loglikelihood is increased.
This can be accomplished with a more complex model.

Sampling distribution of the MLE for θ

Assuming X has full column rank, we have $\hat{\theta} = (X^T X)^{-1} X^T y$.

As $y \sim N_n(X\theta, \sigma^2 I)$, we get

$$\hat{\theta} \sim N_p(\theta, \sigma^2 (X^T X)^{-1}).$$

As

$$I(\theta) := -\frac{\partial^2 \ell}{\partial \theta \partial \theta^T} = \sigma^{-2} X^T X,$$

we get

$$\hat{\theta} \sim N_p\left(\theta, I(\theta)^{-1}\right).$$

1. $\hat{\theta}$ is unbiased for θ
2. the variance can get very large if $X^T X$ is close to singularity (as explained on later slides).

Is the MLE a good estimator?

“Good” estimators

What makes a good estimator $\hat{\theta}$ for θ?

1. Unbiasedness is not a sensible criterion!
2. Most common approach: define a **loss function**. Suppose $\theta \in \Omega$, the parameter set. Let
$$\mathcal{L} \, : \, \Omega \times \Omega \to \mathbb{R}.$$
3. Good estimator: estimator for which its **risk** at θ
$$R_{\hat{\theta}}(\theta) = \mathbb{E}_\theta \, \mathcal{L}(\theta, \hat{\theta})$$
is small.
4. Ideally, we would have one estimator $\hat{\theta}$ for which $R_{\hat{\theta}}(\theta)$ is minimal for all θ. This does not exist.

Bias-variance trade-off

Most common choice $\mathcal{L}(\theta, \bar{\theta}) = \|\theta - \bar{\theta}\|^2$. In this case

$$R_{\hat{\theta}}(\theta) = \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 =: \mathsf{MSE}_{\hat{\theta}}(\theta).$$

Mean Squared Error for $\hat{\theta}$ at θ.

We can write [2]

$$\mathsf{MSE}_{\hat{\theta}}(\theta) = \sum_{i=1}^{p} \left(\mathbb{E}_\theta\left[\hat{\theta}_i\right] - \theta_i\right)^2 + \sum_{i=1}^{p} \mathrm{Var}\left(\hat{\theta}_i\right).$$

To remember:

$$\mathsf{MSE} = \mathsf{Bias}^2 + \mathsf{Variance}.$$

[2] Here, we have used that for a random vector Z in $\mathbb{R}^n$ we have $\mathrm{E}\left[\|Z\|^2\right] = \mathsf{tr}(\mathsf{Cov}(Z)) + \mathrm{E}[Z]^T\,\mathrm{E}[Z]$.

Bias-variance trade-off for the MLE in the linear model: the variance term

1. Assume X has full column rank so that $X^T X$ has p strictly positive eigenvalues λ_k (nonsingular, posdef)
2. $(X^T X)^{-1}$ has eigenvalues λ_k^{-1}

This implies

$$\sum_{i=1}^{p} \mathrm{Var}\left(\hat{\theta}_i\right) = \mathsf{tr}\left(\mathsf{Cov}\,\hat{\theta}\right) = \mathsf{tr}\left(\sigma^2 (X^T X)^{-1}\right) = \sigma^2 \sum_{k=1}^{p} \frac{1}{\lambda_k}.$$

If $\min_k \lambda_k \downarrow 0$, the variance term tends to ∞ and hence the risk blows up.

Prediction in the linear model

As

$$y_i = \theta^T x_i + \epsilon_i$$

for given x_{new} we predict

$$y_{\text{new}} = \hat{\theta}^T x_{\text{new}}.$$

Properties of y_{new}:

$$\mathbb{E} y_{\text{new}} = \theta^T x_{\text{new}}$$

$$\begin{aligned} \operatorname{Var}(y_{\text{new}}) &= \operatorname{Var}\left(x_{\text{new}}^T \hat{\theta}\right) = x_{\text{new}}^T \left(\operatorname{Cov} \hat{\theta}\right) x_{\text{new}} \\ &= \sigma^2 x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}} \end{aligned}$$

Then plug in $\hat{\sigma}$ for σ to obtain $\widehat{\operatorname{Var}(y_{\text{new}})}$.

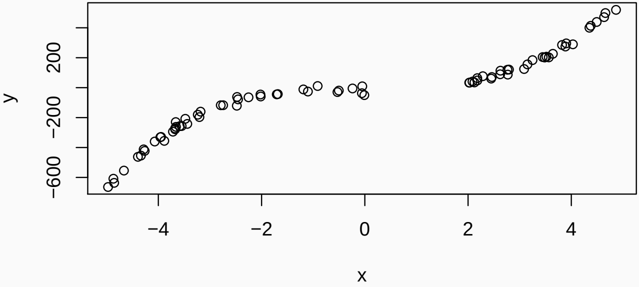
Example (see section 2.11 in RG)

Consider data generated as follows: first generate $\{x_i\}$ independently $Unif\,(-5, 5)$; then set

$$y_i = 5x_i^3 - x_i^2 + x_i + \varepsilon_i$$

with $\{\varepsilon_i\}$ independent $N\,(0, 300)$.

Remove all $x_i \in [0, 2]$.



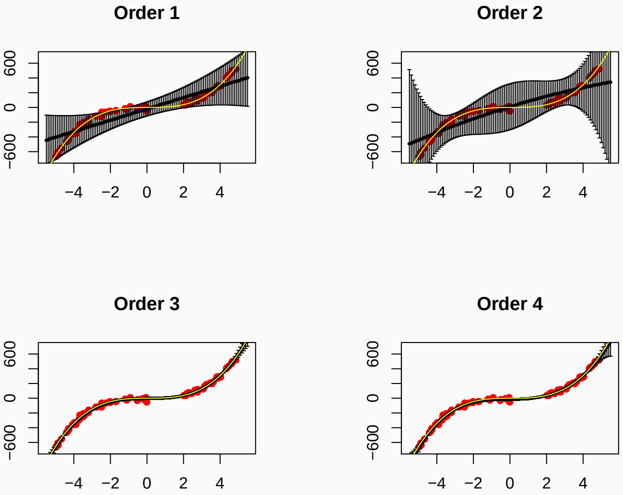
Example continued

- Fit the model while assuming a polynomial of degree k, with $k \in \{1, \dots, 8\}$.
- Evaluate
$$\left(y_{\text{new}}, \widehat{\operatorname{Var}(y_{\text{new}})}\right)$$
for x_{new} a fine grid of evenly spaced values in $[-5.5, 5.5]$.
- Plot the data, along with
$$x_{\text{new}} \mapsto y_{\text{new}}$$
and
$$x_{\text{new}} \mapsto y_{\text{new}} \pm \widehat{\operatorname{Var}(y_{\text{new}})}.$$

Note: it is more customary to plot

$$x_{\text{new}} \mapsto y_{\text{new}} \pm 2\sqrt{\widehat{\operatorname{Var}(y_{\text{new}})}}.$$

Example continued

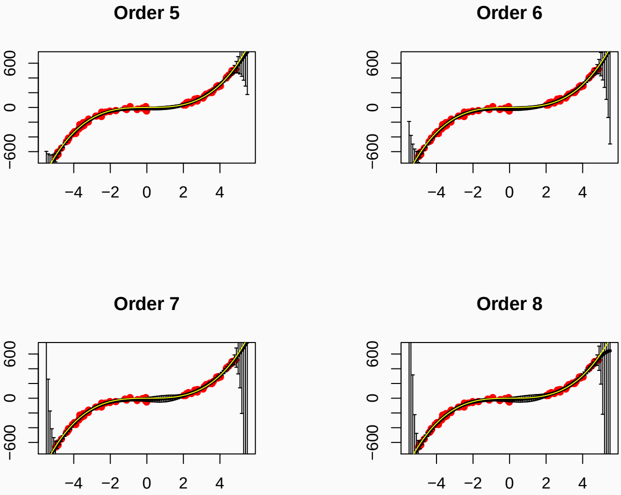

Order 1

Order 2

Order 3

Order 4

Example continued



Prediction in the linear model

Important observation: prediction is a problematic concept in classical statistics.

Why?

- Inconsistency: we assume all y_i and y_{new} are statistically independent. Then how can $y_1, \ldots, y_n$ help in predicting y_{new}?
- We don't get a truly predictive distribution.
- For $\mathrm{Var}(y_{\mathrm{new}})$ we plug in an estimate, but ignore uncertainty in this estimate.

Bayesian approach to statistics does not suffer from such issues.

Model tuning

Most statistical learning methods have tuning parameters:

Key example: polynomial regression with degree parameter. The higher degree model: the higher the likelihood.

How to obtain good risk (i.e. bias variance trade-off)?

1. Use estimator on each model (for example MLE) and evaluate performance by **predictive performance**.
2. Build in **regularisation** to penalise model-complexity.
3. Use **Bayesian** approach to statistics with a **prior distribution** that induces penalisation on model-complexity.

Evaluating predictive performance using cross-validation

Basic idea is data-splitting: split data into train and validation data.

- "Fit" model using training data.
- Evaluate predictions on validation data.

Suppose data $\{(x_i, y_i),\, 1 \le i \le n\}$.

1. Let $1 \le K < n$ and make a partition of $\{1, \dots, n\}$ into K groups:
$$\{1, \dots, n\} = \bigcup_{k=1}^{K} I_k.$$
2. For each $k \in \{1, \dots, K\}$
    - 2.1 Fit model using data $\{(x_i, y_i),\, i \in \{1, \dots, n\} \setminus I_k\}$, yielding estimate θ_{-k}.
    - 2.2 Compute $e_k = \sum_{\ell \in I_k} (y_\ell - \theta^T_{-k} x_\ell)^2$.
    - 2.3 Compute CV $= \sum_k e_k$.

If there is a tuning parameter η, then repeat step (2) over its range to obtain $\eta \mapsto \mathrm{CV}(\eta)$.

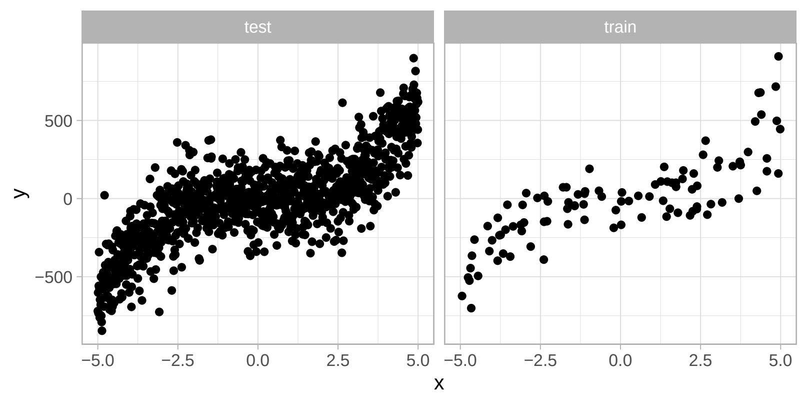
Cross-validation example

- Generate data, with $\{\varepsilon_i\} \overset{\text{iid}}{\sim} N(0, 1)$.

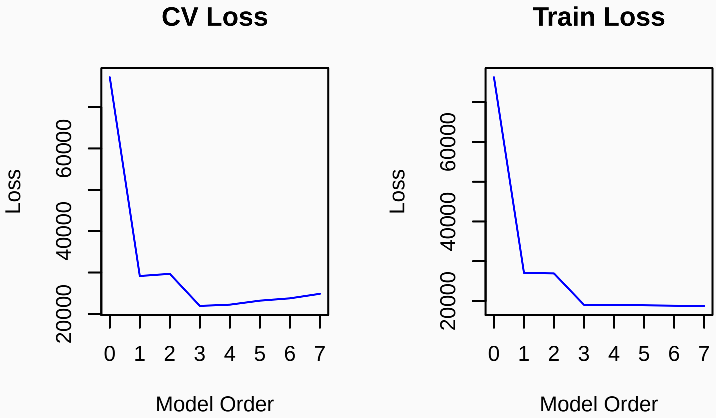
$$y_i = 5x_i^3 - x_i^2 + x_i + 150\varepsilon_i$$

- Use 10-fold CV to determine polynomial degree.
- Evaluate
  - CV loss
  - Training loss

Train and independent test data



Results from 10-fold cross-validation



CV Loss does the right thing here.