LECTURE 10: Validation

Modeling and Simulation 2

Daniel Georgiev

Winter 2014

OUTLINE

- Parameter Fitting
  - regression analysis
  - cross validation
- Hypothesis testing
  - p-value
  - t-test
  - Wilcoxon signed rank test
- Distribution metrics
  - functional norms
  - Wasserstein pseudometric

REGRESSION ANALYSIS

(linear model)

STEP 1: MODEL

dependent variable

independent variable

$$Y = \alpha + \beta X + \epsilon$$

residual

regressors

STEP 2: RESIDUALS

$$\{(x_1, y_1), ..., (x_n, y_n)\}, \epsilon_i = y_i - \alpha - \beta x_i$$

set of measurements

model residuals

STEP 3: PARAMETER FIT

$$\min_{\alpha,\beta} SSR, \text{ where } SSR = \sum_{i=1}^{n} \epsilon_i^2$$

$$\frac{\partial SSR}{\partial \alpha} = 0, \frac{\partial SSR}{\partial \beta} = 0 \implies \alpha^*, \beta^*$$

STEP 4: EVALUATE MODEL QUALITY

standard error

$$\sigma^* = \sqrt{\frac{SSR}{n-2}}$$

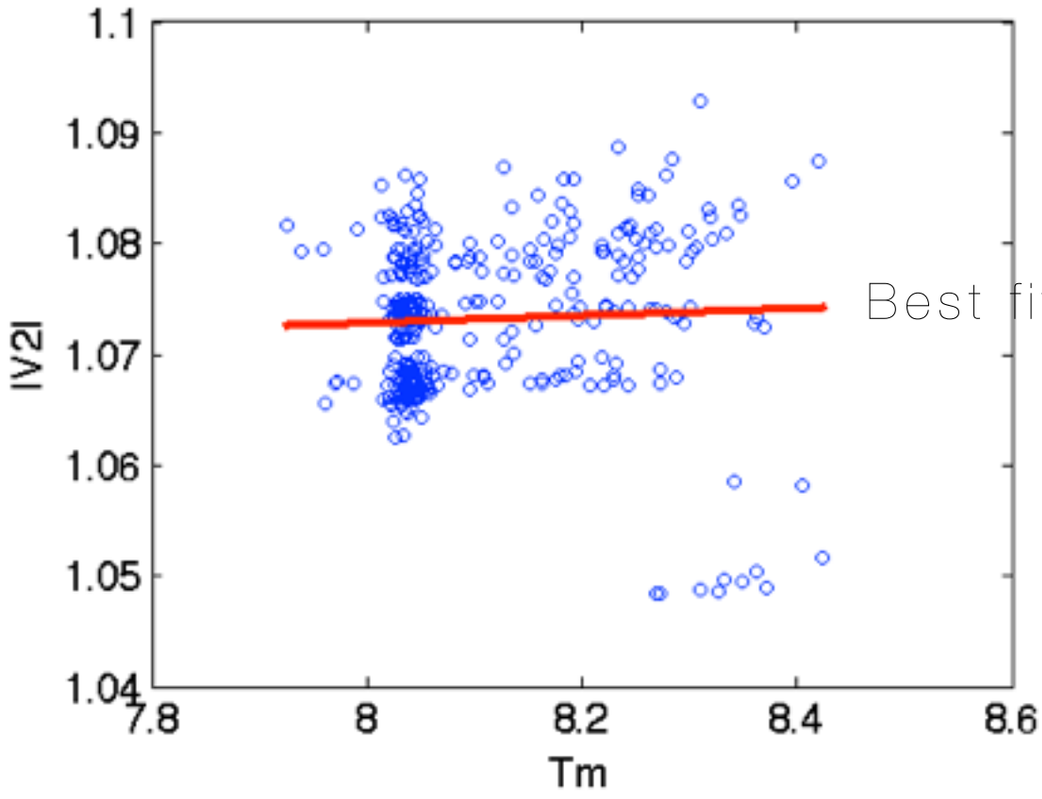
increases with noise

coefficient of determination

$$R^2 = 1 - \frac{SSR}{\sum_{i=1}^{n}(y_i - \bar{y})}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

REGRESSION ANALYSIS

(network example)



Best fitting linear model.

Linear model is clearly not a good fit for the network case study. If it was we wouldn't have bothered with all that modelling and coding and debugging.

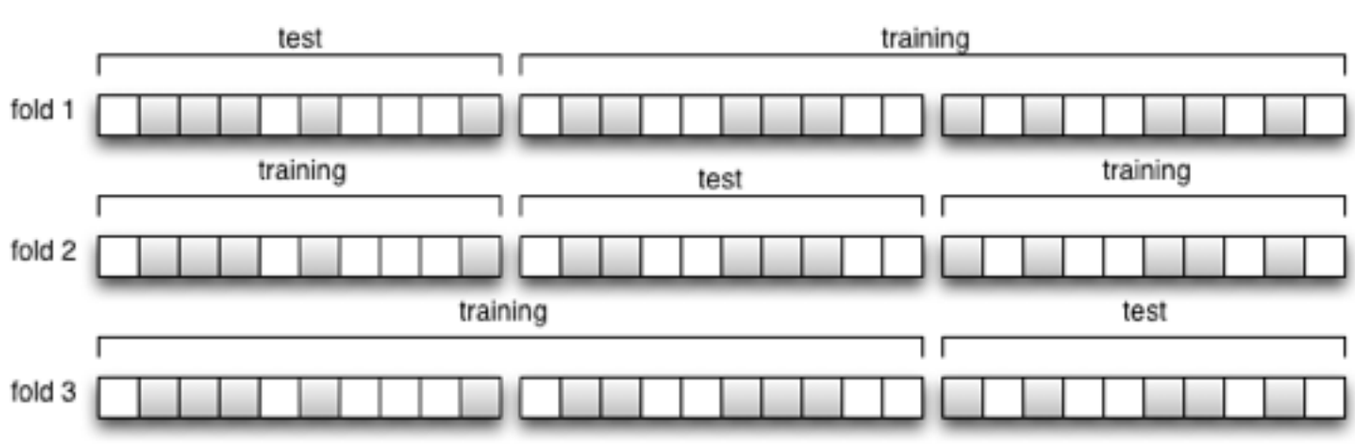
For DAE systems, can compute sensitivity functions to find good parameter fits.

$$\begin{aligned} \dot{x} &= f(x,y,p) \\ 0 &= g(x,y,p) \end{aligned} \Longrightarrow \begin{cases} \frac{\partial \dot{x}}{\partial p} = \frac{\partial f(x,y,p)}{\partial x}\frac{\partial x}{\partial p} + \frac{\partial f(x,y,p)}{\partial y}\frac{\partial y}{\partial p} + \frac{\partial f(x,y,p)}{\partial p} \\ 0 = \frac{\partial g(x,y,p)}{\partial x}\frac{\partial x}{\partial p} + \frac{\partial g(x,y,p)}{\partial y}\frac{\partial y}{\partial p} + \frac{\partial g(x,y,p)}{\partial p} \end{cases}$$

CROSS VALIDATION
(predictive models)

Regression analysis yields models that fit the measured data. These models may be poor predictors of future data. To test prediction ability, cross validation is used.

Algorithm (K-fold cross validation):
STEP 1. separate data into bins



STEP 2: select K-1 bins for model training (e.g., regression analysis)

STEP 3: test the regression model on the unused bin.

STEP 4: select different K-1 bins and repeat STEPS 2-3 until all bins have been used for validation.

HYPOTHESIS TESTING
(model invalidation)

Hypothesis tests determine whether a hypothesis (or a model) is invalidated by the data.

DEFINITION (p-value): Given a null hypothesis Ho, the p-value is the probability of obtaining a result equal to or more extreme than what was actually observed.

$$P(X \geq x|Ho) \vee P(X \leq x|Ho) \vee 2\min\{P(X \leq x|Ho), P(X \geq x|Ho)\}$$

right tail event left tail event double tail event

EXAMPLE (coin toss):

Ho = the coin is fair, i.e., P(H) = P(T) = 0.5
observation = 16 out of 20 Heads
p-value = P(16 H or more out of 20 | fair coin)

$$\text{p-value} = \frac{1}{2^{20}}\left(\binom{20}{16} + \binom{20}{17} + \binom{20}{18} + \binom{20}{19} + \binom{20}{20}\right) \approx 0.059$$

Hence, the fair coin model is invalidated with confidence 0.9 but it is not invalidated with confidence above 0.95.

T-TEST
(normal distribution case)

p-value is not computable in general because we do not know the conditional probability P(X=x|Ho).

Assumptions (t-test):
1. the sample mean is normally distributed
2. the sample variance is chi-squared distributed
3. the sample mean and variance are independent random variables

Computation:
1. Compute the t-score:the t-score is a ratio of two random variables. Under the assumptions, this ratio is described by the t-distribution with dof degrees of freedom.

$$\text{t-score} = \frac{\mu_N - \mu}{\sigma_N / \sqrt{N}}$$

the number of degrees of freedom equals the number of regressors. If we're only approximating the mean, dof = N-1.

2. Table lookup:

t-score for a sample mean, for 10 measurements, the dof = 9
for 5% confidence, the critical value for the t-score = 2.262

T-TEST

(applied to regression analysis)

$Ho : \beta = \beta_0$ the null hypothesis is some true value for the linear regressor

STEP 1: consider sample values and the optimal residuals for these values

$$\{(x_1, y_1), ..., (x_n, y_n)\}, \epsilon_i = y_i - \alpha^* - \beta^* x_i$$

STEP 2: the t-score is the probability that the optimal regressors for a sample have the given values or worse. For two regressors, the t-score has two degrees of freedom and the form

$$\text{t-score} = \frac{(\beta^* - \beta_0)\sqrt{n-2}}{\sqrt{\left(\sum_{i=1}^{n} \epsilon_i^2\right) / \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)}}$$

Note the similarity to the coefficient of determination.

DISTRIBUTION METRICS

Why not just compare the empirical distribution of the sample with the simulated distribution of the model?

Recall the functional norms:

$$f : \Omega \to \mathbb{R}, ||f||_p = \left(\int_\Omega |f|^p \, d\mu\right)^{1/p}$$

Hence, if we estimate the sample probability distribution P1 and the model probability distribution P2, we can take the norm of their difference to measure the distance between the model and the sample data.

$$P_1 : \Omega \to [0\ 1], P_2 : \Omega \to [0\ 1], d_p(P1, P2) = ||P_1 - P_2||_p$$

ADVANTAGE: no need to match measurements or to know independent variable values corresponding to measurements

DISADVANTAGE: introduce further error by approximating the empirical distribution first. The estimation may be especially poor if not enough measurements are given.

WASSERSTEIN PSEUDO METRIC

Wasserstein pseudo metric resolves the issues with distribution norms for the 1-dimensional case. It is defined as:

$$W_p(P_1, P_2) = \left(\inf_{Q \in J(P_1, P_2)} \int_{\Omega \times \Omega} d(\omega, \eta)^p \, dQ(\omega, \eta) \right)^{1/p}$$

where J(P1,P2) is the set of joint probability distributions with marginals equal to P1 and P2.

Intuitively, the Wasserstein pseudo metric can be interpreted as a solution to the minimum transportation problem.

Minimise the distance traveled in transforming a pile of sand into another pile of sand with equal mass.

WASSERSTEIN 1D

Consider two probability distributions

$$P_1 : \Omega \to [0\ 1], P_2 : \Omega \to [0\ 1]$$

and values sampled from these distributions

$$\{X_{1,1}, ..., X_{1,n}\}, \{X_{2,1}, ..., X_{2,\ell n}\}$$

Suppose the individual data sets are in increasing order, then the Wasserstein pseudo metric satisfies

$$W_p\left(P_1, P_2\right) = \lim_{n\to\infty} \left(\frac{1}{\ell n} \sum_{i=1}^{\ell n} \left| X_{1,\lceil i/\ell \rceil} - X_{2,i} \right|^p \right)^{1/p}$$

ADVANTAGES:
no need to compute empirical distributions
distance is defined in physical units
no need to match data