

Lecture 5: Linear regression and hypothesis testing

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 1/21

Announcements

- Homework 1 returned today – general issues:
 - linear regression problem
 - proving versus examples
 - showing results
 - verifying results
 - grading policy
- Homework 2 assigned today, due February 3 – more work than previous assignment

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 2/21

Questions?

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 3/21

Chapter 3: Linear methods for regression

- 3.1 – Introduction
- 3.2 – Linear regression models and least squares
- 3.3 – Multiple regression from simple univariate regression
- 3.4 – Subset selection and coefficient shrinkage
- 3.5 – Computational considerations

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 4/21

Linear regression (3.2)

Linear model:

$$\begin{aligned}f(X) &= \beta_0 + \sum_{j=1}^d \beta_j X_j \\ &= \beta_0 + \langle \beta, X \rangle \\ &= \beta_0 + \beta^T X\end{aligned}$$

Or if X has a column of ones (a dummy variable):

$$f(X) = \beta^T X$$

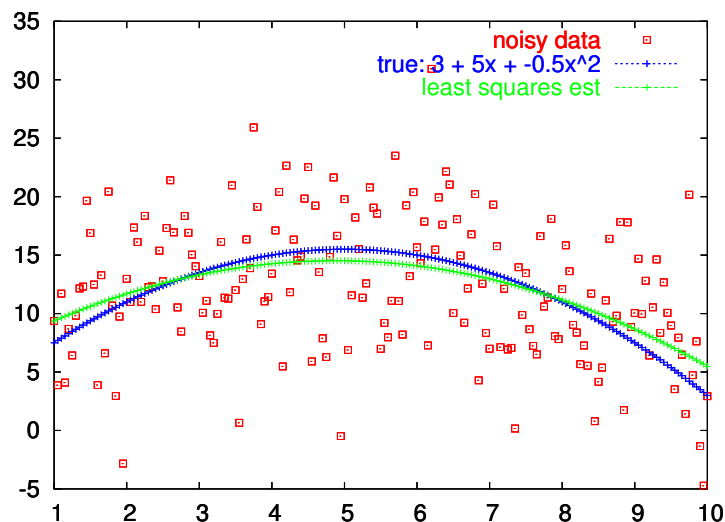
A linear model is linear in the parameters.

Basis expansions allow flexible models

This is (still) a linear model in terms of β :

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

This model can produce curved lines, but in β it is still linear:



Least squares criterion and derivatives

$$\begin{aligned}RSS(\beta) &= \sum_{i=1}^n (y_i - \beta^T x_i)^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = -2\mathbf{X}^T \mathbf{X}$$

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 7/21

Finding the least-squares solution

Take the partial derivative, set to zero:

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

...gives ...

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

...and ...

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

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 8/21

Why does least squares make sense?

Reasonable as long as y_i are independent given x_i (x_i need not be independent).

Criterion measures average lack of fit.

Potential problems

If \mathbf{X} is not full-rank, then $(\mathbf{X}^T \mathbf{X})^{-1}$ is singular, and there is not a unique solution for $\hat{\beta}$.

- This occurs when $d > n$ (can you think of an example of this?)
- This also occurs when two dimensions are perfectly correlated.

In practice, we can ignore the multiple solutions with some common-sense decisions.

Properties of the estimate $\hat{\beta}$

Assumptions:

- y_i are uncorrelated and have constant variance σ^2
- x_i are fixed (non-random)

$$\text{Var}[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

Why? Derivation on the board. . .

Note that if we spread out the x_i values, this will reduce the variance of β !

Statistically eliminating variables

Do we need every variable (aka feature) in the training set?

Variables that have larger coefficients are generally more predictive (but this depends on the spread of the inputs!)

Eliminating multiple variables

If we assume that the linear model is the *correct* model, we can do some manipulations to discover if any variables are not significant for predictive purposes.

Assuming the correct model, three things are true:

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$$

Where $N(\mu, \Sigma)$ is the multivariate Gaussian (aka normal) distribution with mean μ and covariance Σ . Second,

$$(N - d - 1) \hat{\sigma}^2 \sim \sigma^2 \chi_{N-d-1}^2$$

Where χ_{α}^2 is the chi-squared distribution with α degrees of freedom.

Third, $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

Eliminating multiple variables

Using these three properties:

- $\hat{\beta} \sim N(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$
- $(N - d - 1) \hat{\sigma}^2 \sim \sigma^2 \chi_{N-d-1}^2$
- $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent

We can form a statistical hypothesis test for significance of β_j .

Statistical hypothesis testing

A statistical hypothesis test has four parts:

- H_0 , the “null hypothesis”
- H_1 , the “alternative hypothesis”, which is the complement of H_0
- A test *statistic* and *distribution of that statistic* that we will apply.
- A desired significance level, α , such as $\alpha = 5\%$.

We either *reject the null and accept the alternative*, or we *accept the null and reject the alternative*. However, we can never do either with 100% certainty.

Errors in hypothesis testing:

- Type I error: rejecting H_0 when it is true. This has probability α (which we choose).
- Type II error: accepting H_0 when it is false. This has probability β .

The hypothesis test for $\beta_j = 0$

Hypotheses:

- H_0 (null hypothesis): $\beta_j = 0$
- H_1 (alternative hypothesis): $\beta_j \neq 0$

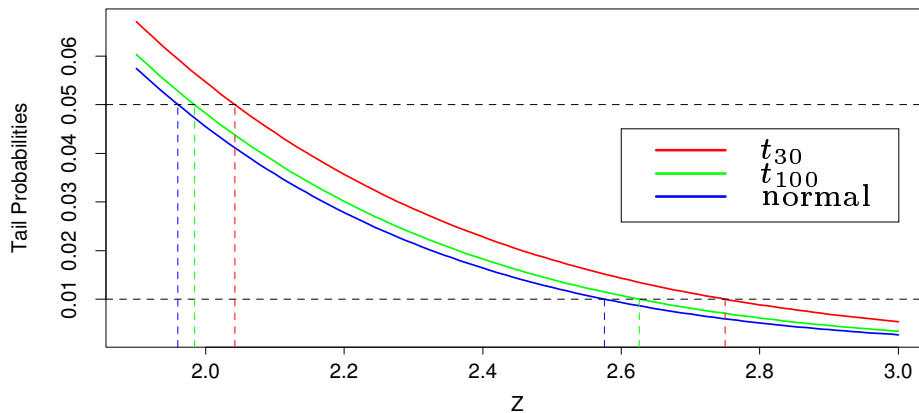
The test statistic is:

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_j}}$$

Where v_j is the j th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$.

Under H_0 , $z_j \sim t_{N-d-1}$. The t distribution is like the Gaussian distribution, but with fatter tails.

The t and Gaussian distributions



The two are very related, but the t distribution arises because we don't know the true σ^2 of the data, only an estimate of it. Note that t requires the number of samples, Gaussian does not.

The test is then if the z_j score is large enough that it falls outside of the acceptance region, and lies in the rejection region.

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 17/21

Example: Testing the hypothesis

- True model: $f(X) = 30 + 50X - 10X^2$
- Assumed model: $g(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \sqrt{X} + \epsilon$
- Generate noisy data: $y_i = f(x_i) + \epsilon$
- Compute $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ (under assumed model $g(X)$)

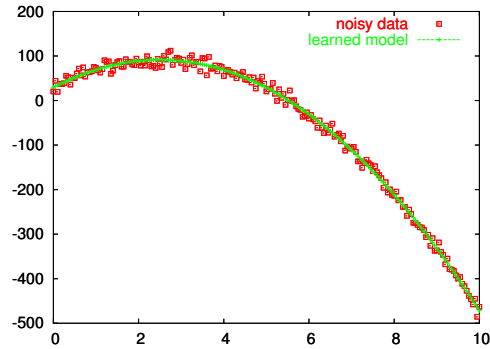
Question: Is $\hat{\beta}_3$ significant?

Applying the test:

- $H_0: \hat{\beta}_3 = 0$
- $H_1: \hat{\beta}_3 \neq 0$
- Compute $z_3 = \hat{\beta}_3 / (\hat{\sigma} \sqrt{v_3})$
- Under H_0 , $z_3 \sim t_{N-d-1}$
- Set $\alpha = 0.05$
- If $Pr(Z > z_3) < \alpha$, then reject H_0

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 18/21

Matlab example



- True model: $f(X) = 30 + 50X - 10X^2$
- Assumed model: $g(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \sqrt{X} + \epsilon$
- $\hat{\beta} = [38.5 \quad 57.8 \quad -10.3 \quad -16.8]$
- $z = [6.4 \quad 12.3 \quad -47.8 \quad -1.6]$
- $Pr(|Z| > 1.9721) = 1 - 0.05 = 0.95$
- Therefore, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are all significant (by themselves), and $\hat{\beta}_3$ is NOT significant (by itself).

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 19/21

Eliminating multiple variables

Note that this z -test for significance only applies to one variable, and not multiple variables at once.

To eliminate multiple variables, eliminate one variable at a time, re-running the test each time.

Next time we will look at eliminating multiple variables at once using the F test.

CSI 5v93: Introduction to machine learning, Lecture 5 – p. 20/21

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification