

Lecture 6: 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 6 – p. 1/22

Announcements

- Homework 2 due February 8th – extension

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 2/22

Questions?

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 3/22

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 6 – p. 4/22

$\hat{\beta}$ is a minimum-error point (BOOK CORRECTION)

In your book, equation 3.4 *incorrectly* states:

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

Taking the partial derivative with respect to β of the first equation gives:

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

So the second derivative is positive along all axes (the diagonal of $\mathbf{X}^T \mathbf{X}$ is positive), so $\hat{\beta}$ is a minimum point. Since $\hat{\beta}$ is the only local minimum, it is the global minimum.

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 5/22

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 $\hat{\beta}$!

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 6/22

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!)

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): $\beta_j \neq 0$

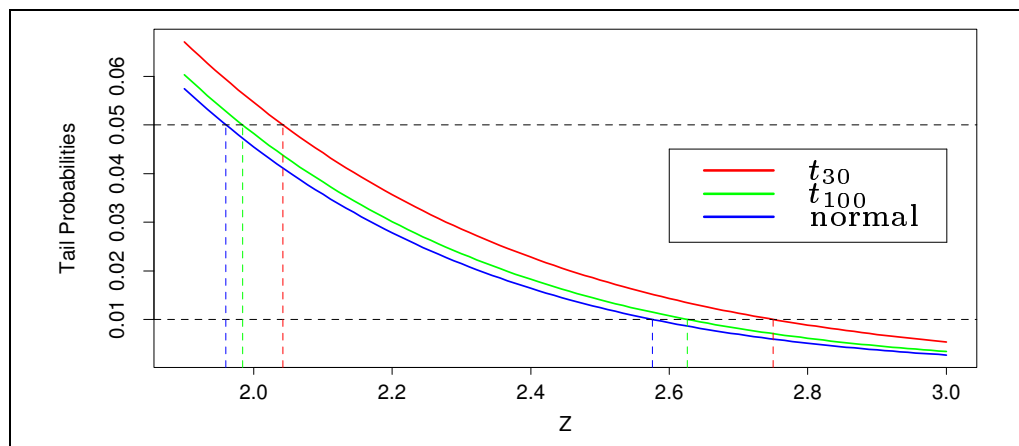
The test statistic is:

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

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.

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 $g(X)$)

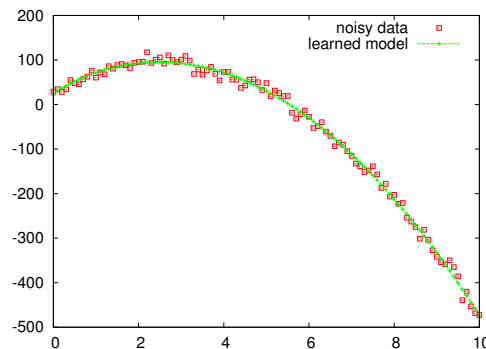
Question: Is $\hat{\beta}_3$ significant?

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 6 – p. 11/22

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 6 – p. 12/22

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.

Eliminating multiple variables

The F-test allows multiple variables to be tested for significance.

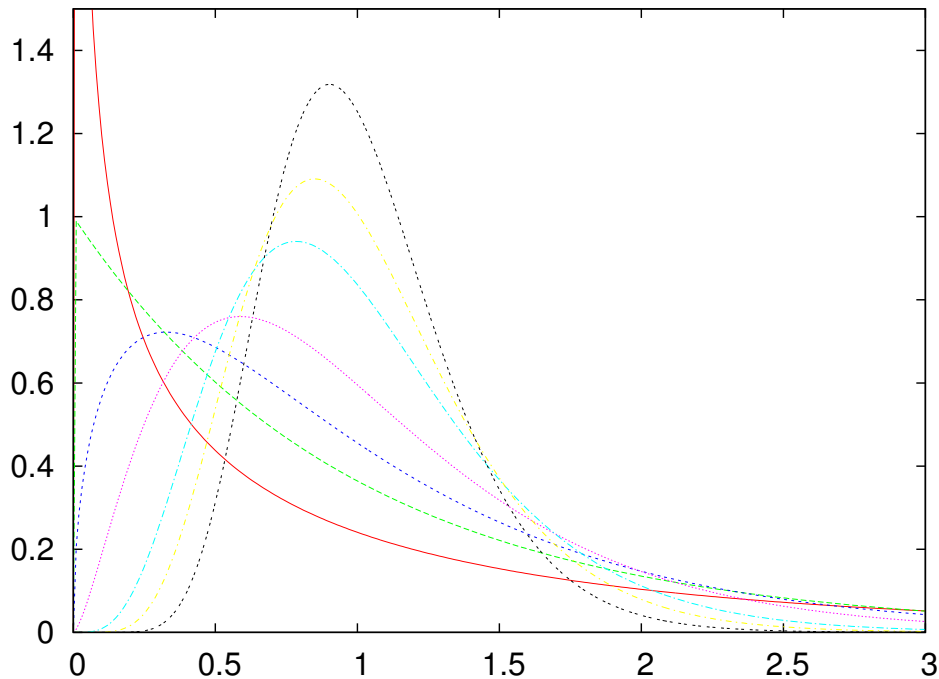
$$F = \frac{(RSS_0 - RSS_1)/(d_1 - d_0)}{RSS_1/(n - d_1 - 1)}$$

- RSS_0 and d_0 refer to the smaller model (with $d_0 + 1$ parameters)
- RSS_1 and d_1 refer to the larger model (with $d_1 + 1$ parameters)

Then we know that F is distributed as $F_{d_1 - d_0, n - d_1 - 1}$ distribution.

The $F_{\alpha,\beta}$ distribution

$$\beta = 100; \alpha = 1, 2, 3, 5, 10, 15, 25$$



CSI 5v93: Introduction to machine learning, Lecture 6 – p. 15/22

Practical issues: statistics in Matlab

Your version of Matlab does not have the statistics toolbox, which is essential for these hypothesis tests. You will need to do these hypothesis tests for your homework.

You have several options:

- use Octave (<http://www.octave.org/>), a matlab-like open-source program (it is installed on earth/wind/fire)
- use the Octave functions in Matlab (I haven't tried this yet)
- build your own functions from tables of values (obtained from Octave, or elsewhere). This may be the easiest method, if you only need a few values of the f_{inv} function, for example.

This issue may (probably will) come up again, and I apologize for the inconvenience.

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 16/22

Subset selection and coefficient shrinkage (3.4)

There are two reasons why least-squares estimates are not totally satisfactory:

- prediction accuracy – many parameters means that the model may have *high variance*; by reducing the number of parameters, we may reduce the variance (but increase the bias). This may help the model to do better on unseen data.
- interpretability – a model with many parameters is not as easy to understand as a model with few parameters.

Your book introduces two different methods of reducing the number of parameters:

- subset selection – choosing only a subset of the variables
- coefficient shrinkage – causing the parameters of less important variables to go to zero

Subset selection

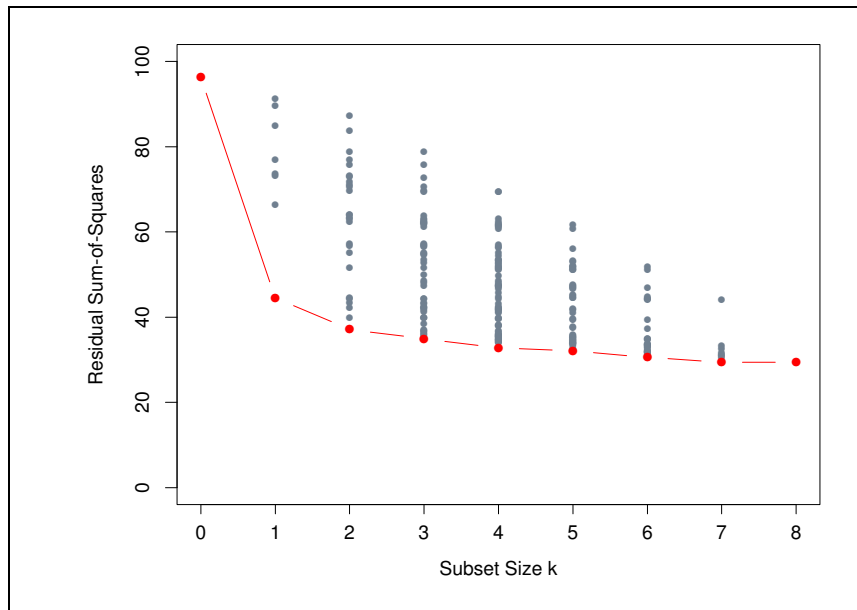
Consider subsets of size 1, 2, 3, ..., d of d variables, and choose the subset that gives a small error.

Considering all possible subsets is computationally infeasible for large d ; the number of subsets goes up exponentially.

Choosing the subset

The RSS will always be smallest for the model with the most parameters (all d parameters), so this is not very useful.

RSS for different subsets for the cancer-prediction problem:



CSI 5v93: Introduction to machine learning, Lecture 6 – p. 19/22

Stepwise selection

Rather than consider every possible subset, stepwise selection adds one variable at a time.

Forward stepwise selection:

- start with one parameter (the intercept)
- add the next best parameter (using the F statistic to determine the best)
- repeat until the change in the F statistic is not significant (with some probability, e.g. 95%)

Backwards stepwise selection is similar, but works in reverse.

CSI 5v93: Introduction to machine learning, Lecture 6 – p. 20/22

Issues with stepwise selection

What are some drawbacks with stepwise selection? Is forward or backward easier? Do you expect they will produce the same results?

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