The Lasso: a brief review and a new significance test

Robert Tibshirani, Stanford University

PIMS-UBC Statistics Constance Van Eeden Lecture
2014
Is Statistics boring?

The Statistician and the Biologist

They are both being executed, and are each granted one last request.

The Statistician asks that he be allowed to give one last lecture on his grand theory of statistics.

The Biologist asks that he be executed first.
Is Statistics boring?

The Statistician and the Biologist

They are both being executed, and are each granted one last request.

The Statistician asks that he be allowed to give one last lecture on his grand theory of statistics.

The Biologist asks that he be executed first.
Is Statistics boring?

*The Statistician and the Biologist*

They are both being executed, and are each granted one last request.

The Statistician asks that he be allowed to give one last lecture on his grand theory of statistics.

The Biologist asks that he be executed first.
Is Statistics boring?

The Statistician and the Biologist

They are both being executed, and are each granted one last request.

The Statistician asks that he be allowed to give one last lecture on his grand theory of statistics.

The Biologist asks that he be executed first.
Statistics is a hot field!

- our enrollments (Masters and PhD) are way up
- graduates find jobs in academia, drug companies, finance, and now: Google, FaceBook, Twitter etc.
- PhD graduates can step into Asst Professor positions right away, often without doing a postdoc
- Other fields—math, physics, biology etc— are not nearly as well off
Outline of this talk

- Sparsity and the Lasso – motivated by a problem in cancer surgery
- A new significance test for the Lasso
Sparsity and the Lasso

For motivation, a real problem:

- I am currently working in a cancer diagnosis project with co-workers at Stanford;
- They have collected samples of tissue from a number of patients undergoing surgery for stomach cancer.
- We are working to build a classifier than can distinguish three kinds of tissue: normal epithelial, normal stromal and cancer.
- Such a classifier could be used to assist surgeons in determining, in real time, whether they had successfully removed all of the tumor. Current pathologist error rate for this call is about 20%.
Left in patient

Extracted part

Normal margin

Cancer

Stromal

Epithelial
Details

• 20 patients, each contributing a sample of epithelial, stromal and cancer tissue. At each pixel in the image, the intensity of metabolites is measured by a new type of mass spectometry. Peaks in the spectrum representing different metabolites.

• The spectrum has been finely sampled, with the intensity measured at about 11,000 $m/z$ sites across the spectrum, and 8000 pixels.
The data for one patient

Spectrum sampled at 11,000 m/z values

Epithelial

Cancer

Stromal
The overall data

Patient 1

Patient 2

Patient 3

Patient 4

... Patient 20

--- 11,000 m/z sites ---

Label of pixel:
1 = Epithelial, 2 = Stromal
3 = Cancer

Robert Tibshirani, Stanford University
The Lasso: a brief review and a new significance test.
What we need to solve this problem

- A statistical classifier (algorithm) that sorts through the large number of features, and finds the most informative ones. This is called a *sparse* set of features.
- Then it must use these features in combination to accurately classify pixels.
The Lasso

• Regression problem: We observe \( n \) feature-response pairs \((x_i, y_i)\), where \( x_i \) is a \( p \)-vector and \( y_i \) is real-valued.

• Let \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \)

• Consider a linear regression model:

\[
y_i = \beta_0 + \sum_j x_{ij} \beta_j + \epsilon_i
\]

\( y_i \) = class of pixel \( i \), (1, 2, or 3); \( x_{ij} \) = height of spectrum at site \( j \) for pixel \( i \). \( \epsilon_i \) is an error term. \( \beta_j \) is the weight given to height of spectrum at site \( j \) (not quite appropriate since \( y_i \) is actually unordered)

• Least squares fitting is defined by

\[
\text{minimize } \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2
\]

• This won’t work when \( p > n \) (why not)?
The Lasso—continued

The **Lasso** is an estimator defined by the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2 \\
\text{subject to} & \quad \sum |\beta_j| \leq s
\end{align*}
\]

- Penalty \(\implies\) sparsity (feature selection)
- Convex problem (good for computation and theory)
- *Ridge regression* uses penalty \(\sum_j \beta_j^2 \leq s\) and does not yield sparsity
Why does the lasso give a sparse solution?

Lasso: \[ \sum_j |\beta_j| \leq s \]

Ridge: \[ \sum_j \beta_j^2 \leq s \]
Prostate cancer example

Back to our problem

• $K = 3$ classes (epithelial vs stromal vs normal): multinomial model

$$\log \frac{Pr(Y_i = k|x)}{1 - Pr(Y_i = K|x)} = \beta_{0k} + \sum_j x_{ij} \beta_{jk}, \ k = 1, 2, \ldots K$$

Here $x_{ij}$ is height of spectrum for sample $i$ at $j$th $m/z$ position

• We replace the least squares objective function by one that is appropriate for a 3-class outcome (no details here)

• Add lasso penalty $\sum |\beta_j| \leq s$; optimize, using cross-validation to estimate best value for budget $s$.

• yields a pixel classifier, and also reveals which $m/z$ sites are informative.
Fast computation is essential

- Our lab has written an open-source R language package called `glmnet` for fitting lasso models.
- It is very fast - can solve the current problem in a few minutes on a PC.
- Not “off-the-shelf” : Many clever computational tricks were used to achieve the impressive speed.

Jerry Friedman

Trevor Hastie
## Results

*Cross-validation* - min at 129 peaks; overall error rate = 4.2%

<table>
<thead>
<tr>
<th>true</th>
<th>Epi</th>
<th>Canc</th>
<th>Strom</th>
<th>Prop correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epi</td>
<td>3277.00</td>
<td>80.00</td>
<td>145.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Canc</td>
<td>73.00</td>
<td>5106.00</td>
<td>13.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Strom</td>
<td>79.00</td>
<td>86.00</td>
<td>2409.00</td>
<td>0.94</td>
</tr>
</tbody>
</table>

*Test set:* overall error rate = 5.7%

<table>
<thead>
<tr>
<th>true</th>
<th>Epi</th>
<th>Canc</th>
<th>Strom</th>
<th>Prop correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epi</td>
<td>1606.00</td>
<td>149.00</td>
<td>19.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Canc</td>
<td>23.00</td>
<td>1622.00</td>
<td>5.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Strom</td>
<td>67.00</td>
<td>5.00</td>
<td>1222.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Peak #</td>
<td>m/z value</td>
<td>Epi</td>
<td>Canc</td>
<td>Strom</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-----</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>101.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>107.5</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>110.5</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>123.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>132.5</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>134.5</td>
<td>0.10</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>135.5</td>
<td>0.13</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>137.5</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>145.5</td>
<td>-0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>146.5</td>
<td>-0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>151.5</td>
<td>-0.15</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>12</td>
<td>157.5</td>
<td>-0.07</td>
<td>-0.07</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>170.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>171.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>174.5</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>175.5</td>
<td>0.14</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>179.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>188.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>212.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>214.5</td>
<td>-0.13</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>215.5</td>
<td>-1.17</td>
<td>-1.17</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>222.5</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>224.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>225.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>231.5</td>
<td>-0.01</td>
<td>-0.31</td>
<td>-0.31</td>
</tr>
<tr>
<td>26</td>
<td>244.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>247.5</td>
<td>-0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>258.5</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>270.5</td>
<td>-0.14</td>
<td>-0.14</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>278.5</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>279.5</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>280.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>285.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>289.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>293.5</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>297.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>299.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>301.5</td>
<td>-0.45</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>39</td>
<td>312.5</td>
<td>0.10</td>
<td>-0.44</td>
<td>-0.44</td>
</tr>
<tr>
<td>40</td>
<td>322.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>324.5</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>325.5</td>
<td>0.03</td>
<td>-0.00</td>
<td>-0.00</td>
</tr>
<tr>
<td>43</td>
<td>333.5</td>
<td>0.82</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>340.5</td>
<td>-0.02</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>341.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>347.5</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>349.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>353.5</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>355.5</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>363.5</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>364.5</td>
<td>-0.10</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>52</td>
<td>365.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>391.5</td>
<td>0.14</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>392.5</td>
<td>0.14</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>393.5</td>
<td>0.26</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>394.5</td>
<td>-0.12</td>
<td>-0.00</td>
<td>-0.00</td>
</tr>
<tr>
<td>57</td>
<td>412.5</td>
<td>-0.00</td>
<td>-0.00</td>
<td>-0.00</td>
</tr>
</tbody>
</table>

Robert Tibshirani, Stanford University

The Lasso: a brief review and a new significance test.
patient = 167 at m/z = 788.6

true

predicted

epithelial
cancer
stromal
A Significance test for the lasso

Lockhart, Taylor, Ryan Tibshirani, Rob Tibshirani
To appear, Annals of Statistics 2014 with discussion
Robert Tibshirani, Stanford University

The Lasso: a brief review and a new significance test.
Setup and basic question

- Given an outcome vector $y \in \mathbb{R}^n$ and a predictor matrix $X \in \mathbb{R}^{n \times p}$, we consider the usual linear regression setup:

$$y = X\beta^* + \sigma \epsilon,$$

where $\beta^* \in \mathbb{R}^p$ are unknown coefficients to be estimated, and the components of the noise vector $\epsilon \in \mathbb{R}^n$ are i.i.d. $N(0, 1)$.

- What we show today: how to provide a p-value for each variable as it is added to lasso model via the LAR algorithm.
Forward stepwise regression

- This procedure enters predictors one at a time, choosing the predictor that most decreases the residual sum of squares at each stage.
- Defining $\text{RSS}$ to be the residual sum of squares for the model containing $k$ predictors, and $\text{RSS}_{\text{null}}$ the residual sum of squares before the $k$th predictor was added, we can form the usual statistic

\[
R_k = \frac{(\text{RSS}_{\text{null}} - \text{RSS})}{\sigma^2}
\]  

(2)

(with $\sigma$ assumed known for now), and compare it to a $\chi^2_1$ distribution.
Simulated example- Forward stepwise- F statistic

$N = 100, p = 10$, true model null

Test is too liberal: for nominal size 5%, actual type I error is 39%. Can get proper p-values by sample splitting: but messy, loss of power
Quick review of least angle regression

Least angle regression is a method for constructing the path of lasso solutions. A more democratic version of forward stepwise regression.

- find the predictor most correlated with the outcome,
- move the parameter vector in the least squares direction until some other predictor has as much correlation with the current residual.
- this new predictor is added to the active set, and the procedure is repeated.
- If a non-zero coefficient hits zero, that predictor is dropped from the active set, and the process is restarted. [This is “lasso” mode, which is what we consider here.]
Least angle regression
The covariance test statistic

Suppose that we want a p-value for predictor 2, entering at the 3rd step.
Compute covariance at $\lambda_4$: $\langle y, X\hat{\beta}(\lambda_4) \rangle$
Drop $x_2$, yielding active yet $A$; refit at $\lambda_4$, and compute resulting covariance at $\lambda_4$, giving

$$T = \left( \langle y, X\hat{\beta}(\lambda_4) \rangle - \langle y, X_A\hat{\beta}_A(\lambda_4) \rangle \right) / \sigma^2$$
The covariance test statistic: formal definition

- Suppose that we wish to test significance of predictor that enters LARS at $\lambda_j$.
- Let $A$ be the active set before this predictor added.
- Let the estimates at the end of this step be $\hat{\beta}(\lambda_{j+1})$.
- We refit the lasso, keeping $\lambda = \lambda_{j+1}$ but using just the variables in $A$. This yields estimates $\hat{\beta}_A(\lambda_{j+1})$. The proposed covariance test statistic is defined by
\[
T_j = \frac{1}{\sigma^2} \cdot \left( \langle y, X \hat{\beta}(\lambda_{j+1}) \rangle - \langle y, X_A \hat{\beta}_A(\lambda_{j+1}) \rangle \right).
\]

- measures how much of the covariance between the outcome and the fitted model can be attributed to the predictor which has just entered the model.
Main result

Under the null hypothesis that all signal variables are in the model:

\[ T_j = \frac{1}{\sigma^2} \cdot \left( \langle y, x \hat{\beta}(\lambda_{j+1}) \rangle - \langle y, x_A \hat{\beta}_A(\lambda_{j+1}) \rangle \right) \rightarrow \text{Exp}(1) \]

as \( p, n \rightarrow \infty \).

More details to come
Comments on the covariance test

\[ T_j = \frac{1}{\sigma^2} \cdot \left( \langle y, X \hat{\beta} (\lambda_{j+1}) \rangle - \langle y, X_A \hat{\beta}_A (\lambda_{j+1}) \rangle \right). \] (4)

- Generalization of standard \( \chi^2 \) or \( F \) test, designed for fixed linear regression, to adaptive regression setting.
- \( \text{Exp}(1) \) is the same as \( \chi^2_2/2 \); its mean is 1, like \( \chi^2_1 \): overfitting due to adaptive selection is offset by \textit{shrinkage} of coefficients
- Form of the statistic is very specific- uses covariance rather than residual sum of squares (they are equivalent for projections)
- Covariance must be evaluated at specific knot \( \lambda_{j+1} \)
- Applies when \( p > n \) or \( p \leq n \): although asymptotic in \( p \), \( \text{Exp}(1) \) seem to be a very good approximation even for small \( p \)
Simulated example- Lasso- Covariance statistic

$N = 100, p = 10$, true model null
Example: Prostate cancer data

<table>
<thead>
<tr>
<th></th>
<th>Stepwise</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>lcavol</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lweight</td>
<td>0.000</td>
<td>0.052</td>
</tr>
<tr>
<td>svi</td>
<td>0.041</td>
<td>0.174</td>
</tr>
<tr>
<td>lbph</td>
<td>0.045</td>
<td>0.929</td>
</tr>
<tr>
<td>pgg45</td>
<td>0.226</td>
<td>0.353</td>
</tr>
<tr>
<td>age</td>
<td>0.191</td>
<td>0.650</td>
</tr>
<tr>
<td>lcp</td>
<td>0.065</td>
<td>0.051</td>
</tr>
<tr>
<td>gleason</td>
<td>0.883</td>
<td>0.978</td>
</tr>
</tbody>
</table>
Simplifications

• For any design, the first covariance test $T_1$ can be shown to equal $\lambda_1(\lambda_1 - \lambda_2)$.

• For orthonormal design, $T_j = \lambda_j(\lambda_j - \lambda_{j+1})$ for all $j$; for general designs, $T_j = c_j\lambda_j(\lambda_j - \lambda_{j+1})$

• For orthonormal design, $\lambda_j = |\hat{\beta}_{(j)}|$, the $j$th largest univariate coefficient in absolute value. Hence

$$T_j = (|\hat{\beta}_{(j)}||\hat{\beta}_{(j)}| - |\hat{\beta}_{(j+1)}|).$$ (5)
Rough summary of theoretical results

Under somewhat general conditions, after all signal variables are in the model, distribution of $T$ for $k$th null predictor $\rightarrow \text{Exp}(1/k)$
Theory for orthogonal case

Global null case: first predictor to enter

Recall that in this setting,

\[ T_j = \lambda_j (\lambda_j - \lambda_{j+1}) \]

and \( \lambda_j = |\hat{\beta}_{(j)}|, \hat{\beta}_j \sim N(0, 1) \)

So the question is:

Suppose \( V_1 > V_2 \ldots > V_n \) are the order statistics from a \( \chi_1 \) distribution (absolute value of a standard Gaussian).

What is the asymptotic distribution of \( V_1 (V_1 - V_2) \)?

[Ask Richard Lockhart!]
Theory for orthogonal case

Global null case: first predictor to enter

Lemma

Lemma 1: Top two order statistics: Suppose $V_1 > V_2 \ldots > V_p$ are the order statistics from a $\chi_1$ distribution (absolute value of a standard Gaussian) with cumulative distribution function $F(x) = (2\Phi(x) - 1)1(x > 0)$, where $\Phi(x)$ is standard normal cumulative distribution function. Then

$$V_1(V_1 - V_2) \rightarrow \text{Exp}(1).$$  
(6)

Lemma

Lemma 2: All predictors. Under the same conditions as Lemma 1,

$$(V_1(V_1 - V_2), \ldots, V_k(V_k - V_{k+1})) \rightarrow (\text{Exp}(1), \text{Exp}(1/2), \ldots \text{Exp}(1/k))$$

Proof uses a theorem from de Haan & Ferreira (2006). We were unable to find these remarkably simple results in the literature.
Heuristically, the Exp(1) limiting distribution for $T_1$ can be seen as follows:

- The spacings $|\hat{\beta}_{(1)}| - |\hat{\beta}_{(2)}|$ have an exponential distribution asymptotically, while $|\hat{\beta}_{(1)}|$ has an extreme value distribution with relatively small variance.
- It turns out that $|\hat{\beta}_{(1)}|$ is just the right (stochastic) scale factor to give the spacings an Exp(1) distribution.
General $\mathbf{X}$ results

Under appropriate condition on $\mathbf{X}$, as $p, N \to \infty$,

1. **Global null case:** $T_1 = \lambda_1(\lambda_1 - \lambda_2) \to \text{Exp}(1)$.

2. **Non-null case:** After the $k$ strong signal variables have entered, under the null hypothesis that the rest are weak,

   $$T_{k+1}^{n,p \to \infty} \leq \text{Exp}(1)$$

Jon Taylor: “Something magical happens in the math”
Conditions on $X$

- The main condition is that for each $(j, s_j)$ the random variable $M_{j,s}(g)$ grows sufficiently fast.
- A sufficient condition: for any $j$, we require the existence of a subset $S$ not containing $j$ such that the variables $U_i, i \in S$ are not too correlated, in the sense that the conditional variance of any one on all the others is bounded below. This subset $S$ has to be of size at least $\log p$. 
HIV mutation data

\( N = 1057 \) samples

\( p = 217 \) mutation sites \((x_{ij} = 0 \text{ or } 1)\)

\( y = \) a measure of drug resistance

The data were randomly divided 50 times into training and test sets of size \((150, 907)\).

Top row shows the training set p-values for forward stage regression and the lasso. The bottom panels show the test set error for the models of each size.
Case of Unknown $\sigma$

Let

$$W_k = \left( \langle y, X\hat{\beta}(\lambda_{k+1}) \rangle - \langle y, X_A\hat{\beta}_A(\lambda_{k+1}) \rangle \right).$$

(7)

and assuming $n > p$, let $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_{\text{full}})^2 / (n - p)$. Then asymptotically

$$F_k = \frac{W_k}{\hat{\sigma}^2} \sim F_{2,n-p}$$

(8)

$[W_j/\sigma^2$ is asymptotically Exp(1) which is the same as $\chi^2_2/2$, $(n - p) \cdot \hat{\sigma}^2 / \sigma^2$ is asymptotically $\chi^2_{n-p}$ and the two are independent.]

When $p > n$, $\sigma^2$ must be estimated with more care.
Extensions

- **Elastic Net:** $T_j$ is simply scaled by $(1 + \lambda_2)$, where $\lambda_2$ multiplies the $\ell_2$ penalty.

- **Generalized likelihood models:**

  $$T_j = \frac{S_0 l_0^{-1/2} X \hat{\beta}(\lambda_{j+1}) - S_0^T l_0^{-1/2} X A \hat{\beta}_A(\lambda_{j+1})}{2}$$

  where $S_0, l_0$ are null score and information matrices, respectively. Works e.g. for generalized linear models and Cox model.
More recent developments

- Exact post-selection inference with the lasso (2013) (Lee et al) in arXiv
- Post-selection adaptive inference for Least Angle Regression and the Lasso (2014) (Taylor et al) in arXiv (provides exact p-values for finite \( n, p \))

\[
\frac{1 - \Phi(\lambda_1/\sigma)}{1 - \Phi(\lambda_2/\sigma)} \sim U(0, 1)
\]

- False Discovery Rate Control for Sequential Selection Procedures, with Application to the Lasso (2013). (G’Sell et al) in arXiv
- In the works: extensions to PCA, group lasso, forward stepwise regression

Robert Tibshirani, Stanford University