



[< Previous](#)

✓

✓

✓

✓

✓

✓

[Next >](#)

7. Model selection and regularization

🔖 Bookmark this page

Exercises due Mar 1, 2021 18:59 EST
Model selection and regularization

References

- D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007. Part III
- D. Freedman. *Statistical Models – Theory and Practice*. 2009. Chapters 2–4.

and we can fit, now, a lot of features with this.

We can use a lot of features with this prediction,

but we can also do model selection to choose only some of them that are relevant.

And if you want to read up more about the parts

of this lecture, I recommend two books to you.

One is the statistics book by Freedman, Pisani, and Purves,

which is a bit more detailed, or *Statistical Models* by Freedman.

And that is going to be chapter 2 and 4.

18:24 / 18:24

1.50x

Video

[Download video file](#)

Transcripts

[Download SubRip \(.srt\) file](#)
[Download Text \(.txt\) file](#)

Video note: At 12:50 The expectation value over ϵ_i ($\mathbb{E} [\epsilon_i]$) should be an expectation value over ϵ_i squared: $\mathbb{E} [\epsilon_i^2] = \sigma^2$, as $\mathbb{E} [\epsilon_i] = 0$.

Video note: After 13:06, the equation for the variance of ϵ_i is missing a square in the equation for the linear model. Since the variance of ϵ_i is σ^2 , the corresponding equation should be

$$\mathbf{E}[(\epsilon_i)^2] = \sigma^2.$$

In this unit, we will

- Create a statistical test for the significance of a predictor in multiple linear regression.
- Derive the statistical distribution for this test.
- Apply the significance test to the exoplanetary data in order to assess the strength of the predictors.

For a model

$$y = X\beta + \epsilon$$

Detailed derivation

The least squares estimator is conditionally unbiased, so $\mathbb{E}[\hat{\beta}|X] = \beta$. As for the covariance matrix of $\hat{\beta}$, the deviation from the mean is

$$\begin{aligned}\hat{\beta} - \mathbb{E}[\hat{\beta}|X] &= \hat{\beta} - \beta \\ &= (X^T X)^{-1} X^T y - \beta \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) - \beta \\ &= (X^T X)^{-1} (X^T X) \beta - \beta + (X^T X)^{-1} X^T \epsilon \\ &= (X^T X)^{-1} X^T \epsilon\end{aligned}$$

Now note that the covariance matrix for the noise is $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I$. So the covariance matrix for $\hat{\beta}$ is

$$\begin{aligned}\mathbb{E}[(X^T X)^{-1} X^T \epsilon ((X^T X)^{-1} X^T \epsilon)^T | X] &= (X^T X)^{-1} X^T \mathbb{E}[\epsilon\epsilon^T | X] X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

It remains to find σ . We have $\mathbb{E}[\epsilon] = 0$, so

$$\begin{aligned}\mathbb{E}[y|X] &= \mathbb{E}[X\beta|X] \\ &= X\beta.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}[(y - \mathbb{E}[y|X]) (y - \mathbb{E}[y|X])^T | X] &= \mathbb{E}[\epsilon\epsilon^T | X] \\ &= \sigma^2 I,\end{aligned}$$

so the variance of y could be used to estimate σ :

$$\sigma^2 = \frac{1}{N-1} \sum_i^N (Y_i - x_i^T \beta)^2$$

However, we don't have access to β directly, just its estimate, so we can't just compute the variance directly as it requires knowing the mean. Instead we must find

$$\begin{aligned}S &= \sum_i^N (Y_i - x_i^T \hat{\beta})^2 \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= (y - X(X^T X)^{-1} X^T y)^T (y - X(X^T X)^{-1} X^T y) \\ &= (X\beta + \epsilon)^T (I - X(X^T X)^{-1} X^T)^2 (X\beta + \epsilon) \\ &= \epsilon^T (I - H)^2 \epsilon\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

We can verify that $\mathbb{E}[(\mathbf{I} - \mathbf{H}) \boldsymbol{\epsilon} | \mathbf{X}] = \mathbf{0}$, and thus S is a sum of squares each with mean 0.

Now note that \mathbf{H} is idempotent, and so $(\mathbf{I} - \mathbf{H})$ is also idempotent. In addition, the rank of an idempotent matrix is equal to the trace of the matrix. So $\text{rank}(\mathbf{I} - \mathbf{H}) = N - \text{rank}(\mathbf{H})$, and the rank of \mathbf{H} is equal to the number of columns in \mathbf{X} , denoted by p . (Note that p includes any intercept column, in the slides $p - 1$ is used so that p is just the number of model parameters.) Thus $\text{rank}(\mathbf{I} - \mathbf{H}) = N - p$.

As the variance of ϵ is σ^2 , it follows that ϵ/σ is a standard normally distributed random variable. Then, from Cochran's theorem, we have that

$$\frac{S}{\sigma^2} = \frac{\boldsymbol{\epsilon}^\top (\mathbf{I} - \mathbf{H}) \boldsymbol{\epsilon}}{\sigma^2},$$

is χ^2 distributed with number of degrees of freedom equal to $\text{rank}(\mathbf{I} - \mathbf{H})$.

Therefore

$$\mathbb{E}\left[\frac{S}{\sigma^2} | \mathbf{X}\right] = N - p.$$

We can now use this as an estimator of σ^2 :

$$\hat{\sigma}^2 = \frac{S}{N - p}$$

[Hide](#)

Recall $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ is the covariance matrix for $\hat{\boldsymbol{\beta}}$, and let

$$\Sigma_j^2 = (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}$$

be the j th diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$. It follows that

$$\frac{\hat{\beta}_j - 0}{\sigma \Sigma_j} \sim \mathcal{N}(0, 1)$$

is a normally distributed variable that can be used in a z -test to test the null hypothesis that $\beta_j = 0$.

Now

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{N - p}$$

is an estimator for σ^2 , and

$$(N - p) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-p}^2$$

where p is the number of columns in $\hat{\mathbf{X}}$.

If $Z \sim \mathcal{N}(0, 1)$ and $\omega \sim \chi_n^2$ then

$$\frac{Z}{\sqrt{\frac{\omega}{n}}} \sim t_n$$

is t -distributed with n degrees of freedom.

Therefore

$$T_j = \frac{\frac{\hat{\beta}_j - 0}{\sigma \Sigma_j}}{\sqrt{\frac{(N-p) \frac{\hat{\sigma}^2}{\sigma^2}}{N-p}}} = \frac{\hat{\beta}_j}{\hat{\sigma} \Sigma_j}$$

is t distributed with $N - p$ degrees of freedom and can be used as a t -test to test the hypothesis that $\beta_j = 0$.

Exoplanet test

3 points possible (graded)
Compute T_j for each $\hat{\beta}_j$ in the exoplanet dataset.

What is the most significant predictor?

☐ LogPlanetRadius

☐ LogPlanetOrbit

☐ StarMetallicity

☐ LogStarMass

☐ LogStarAge

What is the second most significant predictor?

☐ LogPlanetRadius

☐ LogPlanetOrbit

☐ StarMetallicity

☐ LogStarMass

☐ LogStarAge

Which predictor should you remove first from this model?

☐ LogPlanetRadius

☐ LogPlanetOrbit

☐ StarMetallicity

☐ LogStarMass

☐ LogStarAge

Submit
















You have used 0 of 3 attempts

Discussion

Hide Discussion

Topic: Module 1. Review: Statistics, Correlation, Regression, Gradient Descent:Correlation and Least Squares Regression / 7. Model selection and regularization

Add a Post

Show all posts	by recent activity
 Checking my answer	4
 About the notation...	4
 Are p-values the appropriate way to do feature selection?	2
 Precisions of Slide 6 : is beta a random variable ? The slide says that β is a random variable, but at the same time that $\mathbb{E}(\hat{\beta} X) = \beta$, how can an expectation be a random variable ? Is...	1
 What is the null hypothesis for each beta?	5
 Are the test-statistics and p-values found in most standard regression/ANOVA tables exactly equivalent to the hand-constructed hypothesis test on the betas we did here? Quick question, I appreciate doing the hypothesis test by hand for the betas. However, I'm curious, from the standard regression tabl...	1
 Didn't get why we need two sided p-values	2
 Predictor that's removed first... I got the part right but only when i chose second worst predictor. In the worst one i got LogStarMass having highest p value. Dont w...	3
 Any potential "bias" in backward model selection? Maybe I'm not getting the concept right but please let me ask a question. In backward model selection, isn't it possible that having a ...	6
 [STAFF] What is Sigma j ? It is not clear what Sigma j is. Can we have more explicit directions ?	14
 [STAFF] Bias term regularization	4
 Video notes clarification	1
 Linear Regression vs. Machine Learning	5
 help? I got all of the p-value by t test given T value, but I do not know how to recognize the significant parameters I got all of the p-value by t test given T value, but I do not know how to recognize the significant parameters. in the video, the profes...	3
 Probably missed word in the text description	

< Previous

Next >



- About
- Affiliates
- edX for Business
- Open edX
- Careers
- News

Legal

- Terms of Service & Honor Code
- Privacy Policy
- Accessibility Policy
- Trademark Policy
- Sitemap

Connect

- Blog
- Contact Us
- Help Center
- Media Kit
- Donate



© 2021 edX Inc. All rights reserved.
深圳市恒宇博科技有限公司 [粤ICP备17044299号-2](#)