L15: Cross-Validation & p-values

Jeff M. Phillips

March 4, 2019
Model: Ridge Regression

\[ L_2^\text{reg}(x, y, \alpha) = \| x\alpha - y \|_2^2 + \lambda \| \alpha \|_2^2 \]

Cross-Validation

- Choose Parameter (e.g., \( \lambda \))
- Measure Generalization
\( \text{Cost}(x_{\text{test}}, s) \)

\( \| X_{\text{test}} x^*(s) - y_{\text{train}} \| ? \)

\( x^*(s) \) = \text{soln}(x_{\text{train}}, s) \)

over-fitting

train on \( x_{\text{train}} \)

test on \( x_{\text{test}} \)
Generalization

How well will Model $a$ do on new data?

$$\mathbb{E}_{x \sim \text{Data}^\text{train}} \left[ (y - M(x))^2 \right]$$

- $x$: input data
- $y$: target output
- $M(x)$: model output

$\text{generalization}$

$\text{learned}$ vs $\text{new data}$
How to Split Data

- 70% train 30% test
- 90% train 10% test

More → allows more complex model

12-fold Cross-Validation

\[ X \rightarrow X_1, X_2, \ldots, X_{12} \]

for \( j = 1 \) to 12

train on \((X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{12})\)

test on \(X_j\)

leave-one-out

\[ b = n = 1 \times |X| \]
Important:

\[ Pr(\text{observation} \mid \text{hypothesis}) \neq Pr(\text{hypothesis} \mid \text{observation}) \]

The probability of observing a result given that some hypothesis is true is not equivalent to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error: the transposed conditional fallacy.

A p-value (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.
1. Multiple Hypothesis Testing

https://xkcd.com/882/
1. Multiple Hypothesis Testing

https://xkcd.com/882/
1. Multiple Hypothesis Testing

![Comic strip illustration of multiple hypothesis testing](https://xkcd.com/882/)

In the first panel, someone argues that jelly beans cause acne. Scientists investigate and find no link between jelly beans and acne ($P > 0.05$). In the second panel, scientists are asked to investigate again, but they are playing Minecraft. In the third panel, the conclusion is that jelly beans cause acne because a certain color causes it. In the fourth panel, scientists are asked to investigate purple jelly beans and acne, and in the fifth panel, they are asked to investigate brown jelly beans and acne.
1. Multiple Hypothesis Testing

https://xkcd.com/882/
1. Multiple Hypothesis Testing

https://xkcd.com/882/
1. Multiple Hypothesis Testing

https://xkcd.com/882/
1. Multiple Hypothesis Testing

- enormous noisy data
- data squashing
- working data
- data mining
- draw conclusion
1. Multiple Hypothesis Testing

- enormous noisy data
- data squashing
- working data
- data mining
- draw conclusion
1. Multiple Hypothesis Testing

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis PLOS 2:8, 2005

Summary
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologists have pointed out that the framework is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among many null relationships.
1. Multiple Hypothesis Testing

Why Most Published Research Findings Are False

John P.A. Ioannidis  PLOS 2:8, 2005

Summary
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologists have pointed out (e.g., Begg, 1994; Ioannidis, 1998) that the framework for false positive findings is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among many potential candidates.

Bonferroni Correction?
1. Multiple Hypothesis Testing

- enormous noisy data
- data squashing
- working data
- data mining
- draw conclusion

Essay

Why Most Published Research Findings Are False
John P.A. Ioannidis  PLOS 2:8, 2005

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologists have pointed out [8,9] that there is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among the hundreds or thousands of relationships tested.

Bonferroni Correction?
Garden of Forking Paths [Gelman + Loken 2013]

1. Simple classical test based on a unique test statistic, $T$, which when applied to the observed data yields $T(y)$. 

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \cdot)$, with preregistered. For example, might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \cdot(y))$, where the function $\cdot(y)$ is observed in the observed case.

4. “Fishing”: computing $T(y; j)$ for $j = 1, \ldots, J$: that is, performing $J$ tests and then reporting the best result given the data, thus $T(y; \text{best}(y))$. 

1. Simple classical test based on a unique test statistic, $T$, which when applied to the observed data yields $T(y)$.

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered $\phi$. For example, $\phi$ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.
1. Simple classical test based on a unique test statistic, $T$, which when applied to the observed data yields $T(y)$.

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered $\phi$. For example, $\phi$ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.
1. Simple classical test based on a unique test statistic, $T$, which when applied to the observed data yields $T(y)$.

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered $\phi$. For example, $\phi$ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.

4. "Fishing": computing $T(y; \phi_j)$ for $j = 1, \ldots, J$: that is, performing $J$ tests and then reporting the best result given the data, thus $T(y; \phi^{\text{best}}(y))$. 

---

Garden of Forking Paths [Gelman + Loken 2013]