About this class

We'll talk about the concepts of mean squared error, bias, and variance, and discuss the tradeoffs

We'll discuss linear regression and show how to estimate the parameters of a linear model

Evaluating Estimators

Statistical evaluation – ways of choosing without access to test data

Mean Squared Error (MSE): The MSE of an estimator W of a parameter θ is the function of θ defined by $E_\theta(W-\theta)^2$

Alternatives? (Any increasing function of $|W-\theta|$ could work...)

Bias/Variance decomposition:

$$E(W-\theta)^2 =$$

$$E[W^2] + \theta^2 - 2\theta E[W] + (E[W])^2 - (E[W])^2$$

$$= (\text{Bias } W)^2 + E[W^2] - (E[W])^2$$

$$= (\text{Var } W) + (\text{Bias } W)^2$$

where

$$\text{Bias } W = E_\theta W - \theta$$

Unbiased estimators ($E_\theta W = \theta$ for all θ) are good at controlling bias! An unbiased estimator has MSE equal to its variance

Estimators for the Normal Distribution

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$

Unbiased estimator for mean is sample mean

Unbiased estimator for variance is the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Proof:

$$E[S^2] = E[\frac{1}{n-1}(\sum_{i=1}^{n} (X_i - \overline{X})^2)$$

$$= \frac{1}{n-1}[E(\sum_{i=1}^{n} X_i^2) + n\overline{X}^2 - 2\overline{X} \sum_{i=1}^{n} X_i]$$

$$= \frac{1}{n-1} E(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$$

$$= \frac{1}{n-1}(nEX_1^2 - nE\overline{X}^2)$$

Now we need to use a couple of additional facts:

$$EX_1^2 - (EX_1)^2 = \sigma^2$$

and

$$E\overline{X}^2 - (E\overline{X})^2 = \sigma^2/n$$

(This second is basically the definition of standard error)

To show the second, here's a lemma:

$$\text{Var} \sum_{i=1}^{n} g(X_i) = n\text{Var}g(X_1)$$

(where $Eg(X_i)$) and Var$g(X_i)$ exist)

Proof:

$$\text{Var} \sum_{i=1}^{n} g(X_i) = E[\sum_{i=1}^{n} g(X_i) - E(\sum_{i=1}^{n} g(X_i))]^2$$

$$= E[\sum_{i=1}^{n} (g(X_i) - Eg(X_i))]^2$$

If we expand this, there are n terms of the form

$$(g(X_i) - Eg(X_i))^2$$

The expectation of this term is Var $g(X_i)$. Therefore, for n of them we get nVar $g(X_1)$.

What about the other terms? They are all of the form:

$$(g(X_i) - Eg(X_i))(g(X_j) - Eg(X_j))$$

with $i \neq j$ The expectation of this is the covariance of X_i and X_j, which is 0 from independence.

Now we plug back into the expression for $E[S^2]$ and find:

$$E[S^2] = \frac{1}{n-1}(nEX_1^2 - nE\overline{X}^2)$$

$$= \frac{1}{n-1}(n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2))$$

$$= \sigma^2$$

MSEs for Estimators for the Normal Distribution

Unbiased estimator for the mean μ is $\overline{X}$ Unbiased estimator for the variance σ^2 is S^2

MSEs for these estimators are:

$$E(\overline{X} - \mu)^2 = \text{Var } \overline{X} = \frac{\sigma^2}{n}$$

$$E(S^2 - \sigma^2)^2 = \text{Var } S^2 = \frac{2\sigma^4}{n-1}$$

MLE for the variance is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$

$$E\hat{\sigma}^2 = E(\frac{n-1}{n}S^2) = (\frac{n-1}{n})\sigma^2$$

$$\text{Var } \hat{\sigma}^2 = \text{Var } (\frac{n-1}{n}S^2)$$

$$= (\frac{n-1}{n})^2 \text{Var } S^2$$

$$= (\frac{n-1}{n})^2 \frac{2\sigma^4}{n-1}$$

$$= \frac{2(n-1)\sigma^4}{n^2}$$

MSE, using the bias/variance decomposition

$$E(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + (\frac{n-1}{n}\sigma^2 - \sigma^2)^2$$

$$= \frac{2n-1}{n^2}\sigma^4$$

Which is less than

$$\frac{2\sigma^4}{n-1}$$

Bias/Variance Tradeoff in General

Keep in mind: MSE is not the last word. Should we be comfortable using biased estimators? Why are they biased?

Is MSE reasonable for scale parameters (as opposed to location ones?) – forgives underestimation...

Hypothesis space too simple? High bias, low variance

Hypothesis space too complex? Low bias, high variance

Regression

Statistics: describing data, inferring conclusions

Machine learning: predicting future data (out-of-sample)

What would be a reasonable thing to do in the following case (diagram of point cloud)?

Assumption for linear regression: data can be modeled by

$$y_i = \alpha + \beta x_i + \epsilon_i$$

First algorithmic question for us: how to find α and β ?

Least Squares

Define $\overline{x}$ and $\overline{y}$ as usual from our sample data. Now define:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Let's fit a line to the data as best as we can. How do we define this? Residual sum of squares (RSS)

$$\sum_{i=1}^{n} (y_i - (c + dx_i))^2$$

Now, find a and b, estimators of α and β, such that:

$$\min_{c,d} \sum_{i=1}^{n} (y_i - (c + dx_i))^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

For any fixed value of d, the minimizing value of c can be found as follows.

$$\sum_{i=1}^{n} (y_i - (c + dx_i))^2 = \sum_{i=1}^{n} ((y_i - dx_i) - c)^2$$

Turns out the right side is minimized at

$$c = \frac{1}{n} \sum_{i=1}^{n} (y_i - dx_i)$$

$$= \overline{y} - d\overline{x}$$

Why?

$$\min_a \sum_{i=1}^{n} (x_i - a)^2 = \min_a \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - a)^2$$

$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + 2 \sum_{i=1}^{n} (x_i - \overline{x})(\overline{x} - a) + \sum_{i=1}^{n} (\overline{x} - a)^2$$

Second term drops out, basically giving us our result

For a given value of d, the minimum value of RSS is then

$$\sum_{i=1}^{n} ((y_i - dx_i) - (\overline{y} - d\overline{x}))^2$$

$$= \sum_{i=1}^{n} ((y_i - \overline{y}) - d(x_i - \overline{x}))^2$$

$$= S_{yy} - 2dS_{xy} + d^2 S_{xx}$$

Take the derivative with respect to d and set to 0

$$-2S_{xy} + 2dS_{xx} = 0$$

$$\Rightarrow d = \frac{S_{xy}}{S_{xx}}$$

We'll get different lines if we regress x on y! (exercise)

A Statistical Method: BLUE

Assumptions:

$EY_i = \alpha + \beta x_i$

Var $Y_i = \sigma^2$

Second one implies that variance is the same for all data points No assumption needed on the distribution of the Y_i

BLUE: Best Linear Unbiased Estimator

Linear: estimator of the form $\sum_{i=1}^{n} d_i Y_i$

Unbiased: estimator must satisfy $E \sum_{i=1}^{n} d_i Y_i = \beta$

Therefore $\beta = \sum_{i=1}^{n} d_i E[Y_i]$

$$= \sum_{i=1}^{n} d_i(\alpha + \beta x_i)$$

$$= \alpha \sum_{i=1}^{n} d_i + \beta \sum_{i=1}^{n} d_i x_i$$

Must hold for *all* α and β. This is true iff $\sum_{i=1}^{n} d_i = 0$ and $\sum_{i=1}^{n} d_i x_i = 1$

Best: Smallest variance (Equal to MSE for unbiased estimators)

$$\text{Var } \sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} d_i^2 \text{Var } Y_i$$

$$= \sum_{i=1}^{n} d_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} d_i^2$$

The BLUE is then defined by constants d_i that minimize $\sum_{i=1}^{n} d_i^2$ while satisfying the constraints derived above.

It turns out that the choices $d_i = \frac{x_i - \overline{x}}{S_{xx}}$ are the choices that do this, which gives us $b = \frac{S_{xy}}{S_{xx}}$

The advantage of working under statistically explicit assumptions is we also get statistical knowledge about our estimator

$$\text{Var } b = \sigma^2 \sum_{i=1}^{n} d_i^2 = \frac{\sigma^2}{S_{xx}}$$

If you can choose the x_i, you can design the experiment to try and minimize the variance!

Similar analysis shows that the BLUE of α is the same a as in least squares