STAT 88: Lecture 35

Contents

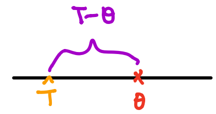
Warm up:

German tanks were numbered $1, 2, 3, ..., N$, with N unknown, during World War 2 and the Allies needed to estimate N. They captured 5 tanks numbered $20, 31, 43, 78$ and 92. Can you find an unbiased estimate of N?

Last time

Bias and Variance

We score how good an estimator T of a parameter θ is by



$$\text{MSE}_\theta(T) = E_\theta((T-\theta)^2).$$

And we showed

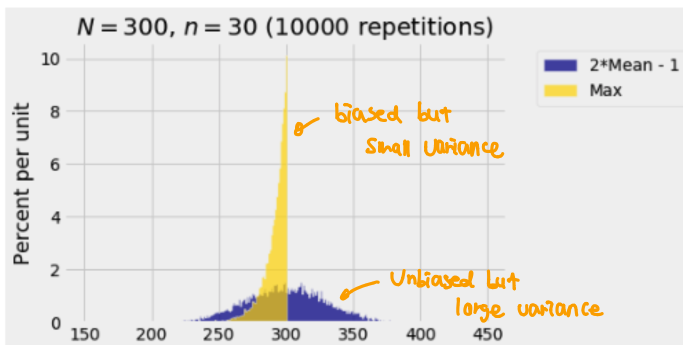
$$\text{MSE}_\theta(T) = B_\theta^2(T) + \text{Var}_\theta(T),$$

where

$$B_\theta(T) = E_\theta(T) - \theta \text{ and } \text{Var}_\theta(T) = E_\theta((T - E_\theta(T))^2).$$

bias   variance

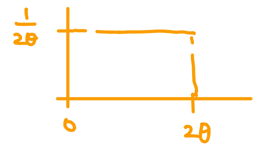
The best estimator is *not* always unbiased.



To find an unbiased estimator, start with a statistic whose expectation is a linear function of the parameter.

11.2. The German Tank Problem

Practice for finding an unbiased estimator



Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, 2\theta)$. Let

$$M = \max\{X_1, \ldots, X_n\}.$$

Is M a biased estimator?

Find $E(M)$.

Find an unbiased estimator for 2θ.

11.3. Least Squares Linear Regression

Let (X, Y) be a random pair of father and son heights from the population:

$$X: \text{ father height, and } Y: \text{ son height.}$$

We want to estimate Y, call this $\widehat{Y}$, by the function

$$\widehat{Y} = aX + b,$$

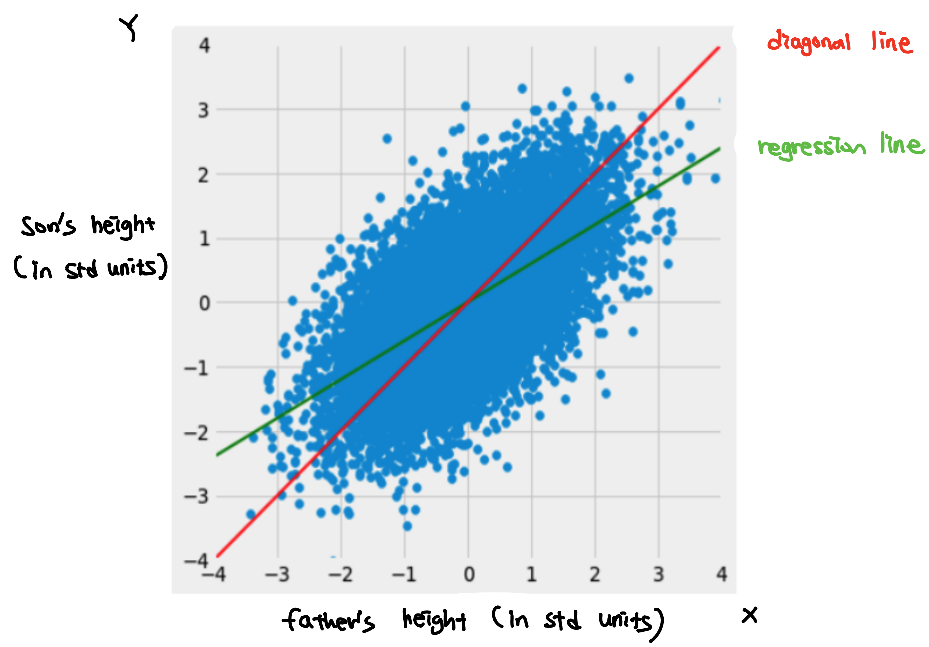
for some slope a and intercept b.

You plug in X into $\widehat{Y} = aX + b$ to predict Y. To find a and b, in Data 8, you collected n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and made a scatter plot. The regression line is the "best" fitting line $\widehat{Y} = aX + b$ through your scatter plot. The formulas are:

$$\text{slope of the regression line} = r\frac{\text{SD of } Y}{\text{SD of } X},$$

and

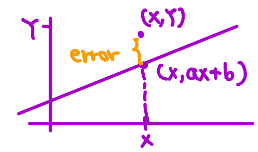
$$\text{intercept of the regression line} = (\text{average of } Y) - \text{slope} \times (\text{average of } X).$$



We will now derive the formulas mathematically using calculus and properties of expectation and variance.

Mean Squared Error For the random point (X, Y), the mean squared error of a linear predictor of Y based on X depends on the slope a and intercept b of the line used. So let us define $\text{MSE}(a, b)$ to be the mean squared error when we use the line $aX + b$ to predict Y. That is,

$$\text{MSE}(a, b) = E((Y - (aX + b))^2).$$



Note that we average over all random (X, Y) pairs in the population. We have to find the values of a and b that minimize this function.

<u>Notation</u>

- $E(X) = \mu_X$, $\text{SD}(X) = \sigma_X$.
- $E(Y) = \mu_Y$, $\text{SD}(Y) = \sigma_Y$.

Best Intercept for a Fixed Slope Fix slope a, and solve $\frac{\partial \text{MSE}(a,b)}{\partial b} = 0$. Since

$$\begin{aligned} \text{MSE}(a, b) &= E((Y - (aX + b))^2) \\ &= E(((Y - aX) - b)^2) \\ &= E((Y - aX)^2 - 2b(Y - aX) + b^2) \\ &= E((Y - aX)^2) - 2b \cdot E(Y - aX) + b^2. \end{aligned}$$

Solve $\frac{\partial \text{MSE}(a,b)}{\partial b} = 0$ for b:

Best Slope For each fixed slope a, we first plug in the best intercept we just found. The the error becomes

$$\begin{aligned} Y - (aX + \widehat{b}_a) &= Y - (aX + \mu_Y - a\mu_X) \\ &= Y - aX - \mu_Y + a\mu_X \\ &= Y - \mu_Y - a(X - \mu_X) \\ &= D_Y - aD_X. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{MSE}(a, \widehat{b}_a) &= E((D_Y - aD_X)^2) \\ &= E(D_Y^2) - 2aE(D_X D_Y) + a^2 E(D_X^2) \\ &= \sigma_Y^2 - 2aE(D_X D_Y) + a^2 \sigma_X^2. \end{aligned}$$

Solve $\frac{d\mathrm{MSE}(a,\widehat{b}_a)}{da} = 0$ for a:

So the regression line is

$$\widehat{Y} = \widehat{a}X + \widehat{b},$$

where

$$\widehat{a} = \frac{E(D_X D_Y)}{\sigma_X^2} \text{ and } \widehat{b} = \mu_Y - \widehat{a} \cdot \mu_X.$$

Correlation $E(D_X D_Y)$ is called the covariance of X and Y. If X is father's height (ft) and Y is son's height (ft), then $E(D_X D_Y)$ has unit ft^2.

If we divide it by $\sigma_X \sigma_Y$,

$$r = \frac{E(D_X D_Y)}{\sigma_X \sigma_Y}$$

is unitless and called the correlation coefficient of X and Y. This tells you

$$\text{Covariance } E(D_X D_Y) = r\sigma_X \sigma_Y,$$

so

$$\widehat{a} = \frac{E(D_X D_Y)}{\sigma_X^2} = \frac{r\sigma_X \sigma_Y}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X}.$$

Appendix

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, 2\theta)$. Let $M = \max\{X_1, \ldots, X_n\}$. Calculate the density of M by first calculating the CDF of M.

$$\begin{aligned} F(m) &= P(M \leq m) \\ &= P(X_1 \leq m, \ldots, X_n \leq m) \\ &= P(X_1 \leq m) \cdots P(X_n \leq m) \\ &= P(X_1 \leq m)^n = \left(\frac{m}{2\theta}\right)^n. \end{aligned}$$

So,

$$f(m) = \frac{dF(m)}{dm} = nm^{n-1} \cdot \frac{1}{(2\theta)^n}.$$

Now we calculate

$$\begin{aligned} E(M) &= \int_0^{2\theta} mf(m)dm \\ &= \frac{n}{(2\theta)^n} \int_0^{2\theta} m^n dm \\ &= \frac{n}{(2\theta)^n} \frac{m^{n+1}}{n+1} |_0^{2\theta} \\ &= (2\theta) \frac{n}{n+1}. \end{aligned}$$