Homework 4

ESE 402/542

Due November 12, 2020 at 11:59pm

Type or scan your answers as a single PDF file and submit on Canvas.

Problem 1. Suppose we fit n data points with a line by minimizing RSS (least squares), and that we want to estimate the line at a new point, x_0. Denoting its value on the line by μ_0, the estimate is:

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

(a) Find variance of $\hat{\mu}_0$.

(b) The standard deviation of $\hat{\mu}_0$ can be expressed as a function of $(x_0 - \bar{x})$. Find this function and briefly explain its shape.

(c) Find 95% confidence interval for $\mu_0 = \beta_0 + \beta_1 x_0$ under assumption of normality.

Sloution:

(a) The line we are observing is given by:

$$y = \beta_0 + \beta_1 x$$

where β_0 and β_1 are:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$
$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

We know that the line goes through point $(x_0, \hat{\mu}_0)$.

$$\begin{aligned}\hat{\mu}_0 &= \beta_0 + \beta_1 x_0 \\ &= \bar{y} - \beta_1 \bar{x} + \beta_1 x_0 \\ &= \bar{y} + \beta_1 (x_0 - \bar{x})\end{aligned}$$

Now we can find the variance of $\hat{\mu}_0$:

$$\begin{aligned}
\operatorname{Var}(\hat{\mu}_0) &= \operatorname{Var}(\bar{y} + \beta_1 (x_0 - \bar{x})) \\
&= \operatorname{Var}(\bar{y}) + (x_0 - \bar{x})^2 \operatorname{Var}(\beta_1) \\
&= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + (x_0 - \bar{x})^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)
\end{aligned}$$

And according to theorem:

$$\operatorname{Var}\left(\hat{\beta}_1\right) = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

We have:

$$\operatorname{Var}(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

(b) Variance of $\hat{\mu}_0$ is given by:

$$\operatorname{Var}(\hat{\mu}_0) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Standard deviation of $\hat{\mu}_0$ is given by:

$$s(\hat{\mu}_0) = \sqrt{\frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Writing standard deviation as a function of $x_0 - \bar{x}$. Now we have:

$$f(x_0 - \bar{x}) = \sqrt{\frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Or simply:

$$f(z) = \sqrt{\frac{\sigma^2}{n} + z^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Since n, σ^2 and $\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ are constants, we can define them as:

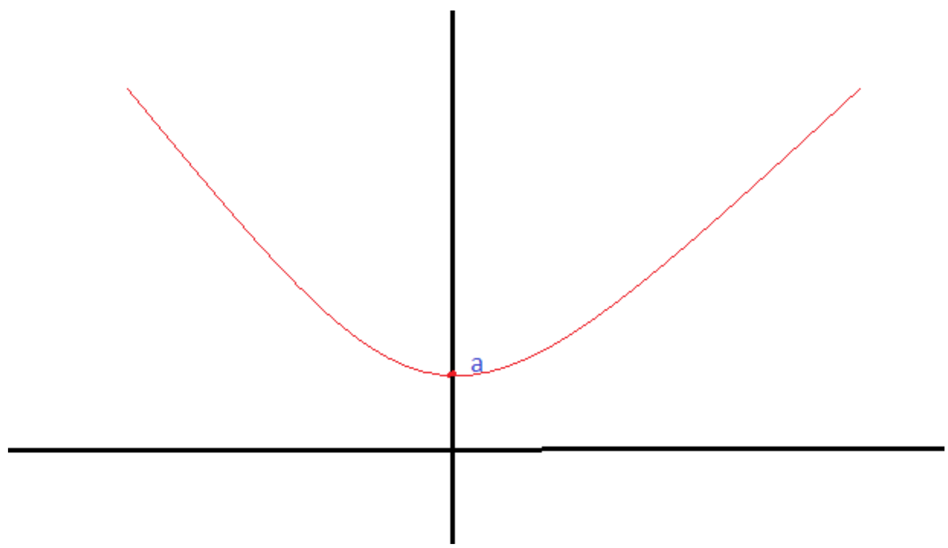
$$\frac{\sigma^2}{n} := a$$

$$\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} := b$$

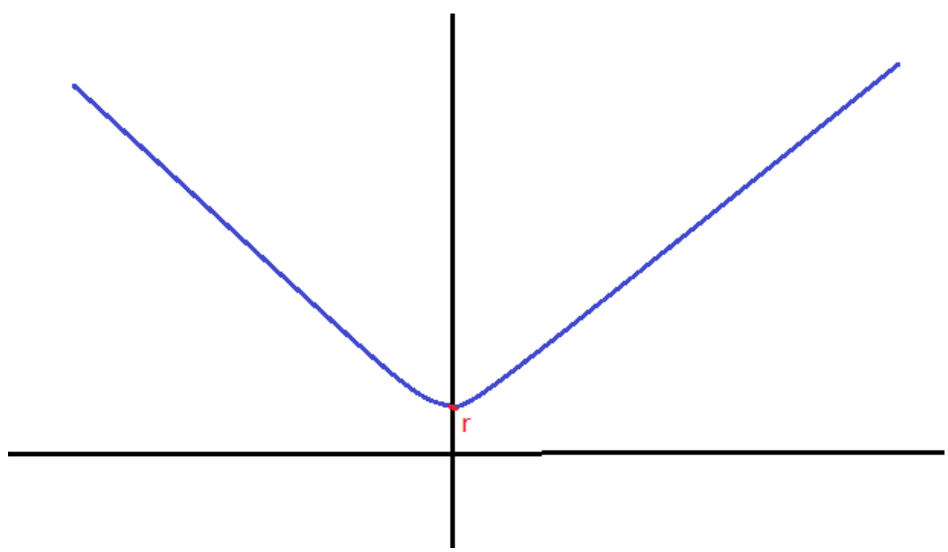
Notice that both a and b are positive constants. Now, function f can be written as:

$$f(z) = \sqrt{a + z^2 b}$$

First, we can draw function $g(z) = a + z^2 b$. The result is parabola:



Now we can draw function f, the square root of function g. Point r represents the square root of a:



(c) Standard deviation of $\hat{\mu}_0$ is given by:

$$s\left(\hat{\mu}_0\right) = \sqrt{\frac{\sigma^2}{n} + \left(x_0 - \bar{x}\right)^2 \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

95% confidence interval is given by:

$$\hat{\mu}_0 \pm s_{\hat{\mu}_0} z \left(\frac{\alpha}{2}\right)$$

Problem 2. $X \sim N(0,1)$, $E \sim N(0,1)$, X and E are independent, and $Y = X + \beta E$. Show that:

$$r_{XY} = \frac{1}{\sqrt{\beta^2 + 1}}$$

Note that r_{XY} is defined as $r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Sloution:

Variables X and E have a standard normal distribution and variable Y is defined as $X + \beta E$. Variables X and E are independent, thus the covariance between these two is 0.
We know that correlation between two variables A and B is:

$$\rho_{A,B} = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B}$$

Applying this formula to variables X and Y, we get:

$$\begin{aligned}
\rho_{X,Y} &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{\mathrm{Cov}(X, X+\beta E)}{\sigma_X \sigma_{X+\beta E}} \\
&= \frac{\mathrm{Cov}(X,X) + \mathrm{Cov}(X, \beta E)}{\sigma_X \sqrt{\mathrm{Var}(X+\beta E)}} \\
&= \frac{\mathrm{Cov}(X,X) + \beta\,\mathrm{Cov}(X,E)}{1 \cdot \sqrt{\mathrm{Var}(X+\beta E)}} \\
&= \frac{\mathrm{Var}(X)}{\sqrt{1+\beta^2 \cdot 1}} \\
&= \frac{1}{\sqrt{1+\beta^2}}
\end{aligned}$$

Problem 3. Suppose there are n data points. We fit a line $y = a + bx$ with least squares, and fit a line $x = c + dy$ with least squares. Show that $bd \leq 1$, and briefly explain when $bd = 1$ and what it means.

Solution:
The first line $y = a + bx$ has:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The second line $x = c + dy$ has:

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

So we have:

$$\begin{aligned} bd &= \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &= \frac{\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &= \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \end{aligned}$$

According to Cauchy-Schwarz inequality, $(\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$, so $bd \leq 1$.

And two sides are equal iff $X - \bar{X} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})$ and $Y - \bar{Y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})$ are linearly dependent, which means you can fit a perfect line on n data points with zero error.

Problem 4. A student wants to predict a variable, Y, from two other variables, $X1$ and $X2$, using multiple regression. He defines a new variable $X3 = X1 + X2$ and uses multiple regression to predict Y from $X1$, $X2$, $X3$. Show that this method is problematic.
Hint 1: $A_{n\times n}$ is invertible $\Leftrightarrow$ $\mathrm{Rank}(A) = n$.
Hint 2: $\mathrm{Rank}(AB) \leq \min(\mathrm{Rank}(A), \mathrm{Rank}(B))$.

Solution:

We are given varaibles X_1, X_2 and Y. Also, the sum of variables X_1 and X_2 is variable X_3.
We want to predict Y from three X variables using multiple regression.
The model we will be analysing is

$$Y = \beta X + E$$

where matrices Y, X,β and E are

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{bmatrix}$$

Since $X_1 + X_2 = X_3$, columns of X are connceted and the column rank of matrix X is 3. Since matrix X doesn't have a full column rank, it doesn't have an inverse.
First, let's notice that the rank of matrix X is the same as the rank of matrix $X^\intercal$.
Now we can remember that if we have two matrices, A and B, the nex inequality is valid for the rank of AB:

$$r(AB) \leq min\{r(A), r(B)\}$$

Applying this to our matrices X and $X^\intercal$, we have:

$$r(X^\intercal X) \leq min\{r(X), r(X^\intercal)\} = min\{3, 3\} = 3$$

Since matrix $X^\intercal X$ comes from the set of matrices with dimension 4×4 and its rank is 3 or less, we can conclude that it doesn't have an inverse.

Now we have a problem, because we have to find the least square estimator for β which is given by

$$\beta = (X^\intercal X)^{-1} X^\intercal Y$$

So, we cannot find β because we cannot find the inverse of $X^\intercal X$.