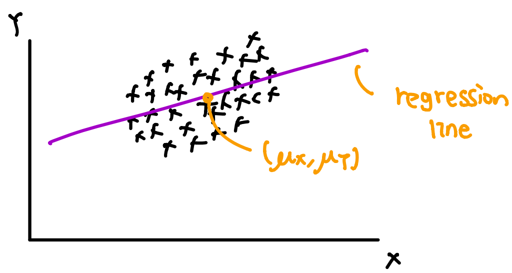
STAT 88: Lecture 37

Contents

Warm up:

Let (X, Y) be a random pair. The average of this pair is (μ_X, μ_Y), called the point of averages. Show that the point of averages lies on the regression line.



Last time

Least squares regression

Let (X, Y) be a random pair. We write

- $E(X) = \mu_X$, $\mathrm{SD}(X) = \sigma_X$.
- $E(Y) = \mu_Y$, $\mathrm{SD}(Y) = \sigma_Y$.
- Correlation
$$r = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}.$$

The regression line $\widehat{Y} = \widehat{a}X + \widehat{b}$ is the best fitting line, in the sense that it minimizes the mean squared error

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

We showed that

$$\widehat{a} = r\frac{\sigma_Y}{\sigma_X} \text{ and } \widehat{b} = \mu_Y - \widehat{a} \cdot \mu_X.$$

Correlation

Let $X^\star$ be X in standard units and $Y^\star$ be Y in standard units:

$$X^\star = \frac{X - \mu_X}{\sigma_X} \text{ and } Y^\star = \frac{Y - \mu_Y}{\sigma_Y}.$$

Then

$$r = r(X, Y) = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = E\left(\frac{X-\mu_X}{\sigma_X} \cdot \frac{Y-\mu_Y}{\sigma_Y}\right) = E(X^\star Y^\star).$$

The correlation satisfies the following properties:

- $-1 \le r(X, Y) \le 1$.
- $r(X, Y) = r(Y, X)$.
- $r(aX + b, cY + d) = \begin{cases} r(X, Y) & \text{if } ac > 0 \\ -r(X, Y) & \text{if } ac < 0 \end{cases}$

11.5. The Error in Regression

r As a Measure of Linear Association The error in the regression estimate is called the residual and is defined as

$$D = Y - \widehat{Y}.$$

We showed

$$E(D) = 0 \text{ and } \mathrm{SD}(D) = \sqrt{1 - r^2}\sigma_Y.$$

So if r is close to ± 1, $\mathrm{SD}(D)$ is close to 0, which implies that Y is close to $\widehat{Y}$. In other words, Y is close to being a linear function of X. ← r measures strength of linear relationship between X and Y

In the extreme case $r = \pm 1$, $\mathrm{SD}(D) = 0$ and Y is a perfectly linear function of X.

The Residual is Uncorrelated with X We will show that the correlation between X and residual D is zero. Note that

$$r(D, X) = \frac{E((D - \mu_D)(X - \mu_X))}{\sigma_D \sigma_X} = \frac{1}{\sigma_D \sigma_X} E(D D_X),$$

because $\mu_D = 0$. We thus show $E(D D_X) = 0$:

$$\begin{aligned}
E(D D_X) &= E((D_Y - \widehat{a} D_X) D_X) \\
&= E(D_X D_Y) - \widehat{a} E(D_X^2) \\
&= r \sigma_X \sigma_Y - r \frac{\sigma_Y}{\sigma_X} \sigma_X^2 \\
&= 0.
\end{aligned}$$

Example: (Exercise 11.6.11) Let X have the uniform distribution on the three points -1, 0, and 1. Let $Y = X^2$.

(a) Show that X and Y are uncorrelated.

(b) Are X and Y independent?

12.1. The Simple Linear Regression Model

The model involves a variable called the response and another called a predictor variable or feature.

$$Y = \text{response} \quad \text{and} \quad x = \text{predictor variable/covariate/feature}$$

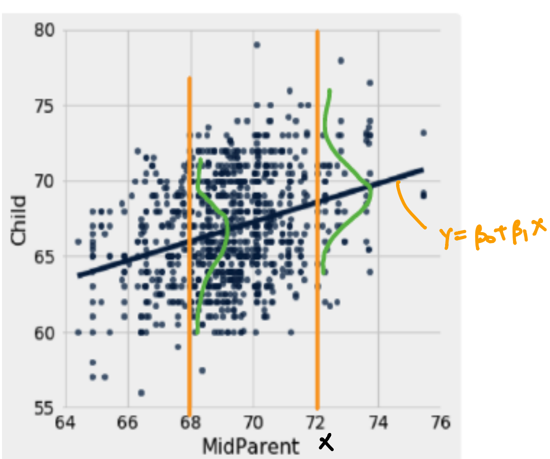
big Y (random) Small x (fixed / not random)

We observe the response and the predictor variable of n individuals, i.e. $(x_1, Y_1), \ldots, (x_n, Y_n)$. Our assumption is

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{noise}},$$

where

- β_0 and β_1 are unobservable constant parameters.
- x_i is the value of the predictor variable for individual i and is assumed to be constant (that is, not random).
- The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d. normal $\mathcal{N}(0, \sigma^2)$ random variables.
- The error variance σ^2 is an unobservable constant parameter, and is assumed to be the same for all individuals i.



Data generation process

Pop of X → Generate $Y = \beta_0 + \beta_1 X + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$

↓ Draw n samples

$X_1, X_2, \ldots, X_n$

↳ $x_1, x_2, \ldots, x_n$ → Generate $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \underset{iid}{\sim} N(0, \sigma^2)$

Fix

Individual Responses Fix x_i, then

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim \mathcal{N}(\qquad\qquad , \qquad).$$

$Y_1, Y_2, \ldots, Y_n$ are independent because $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent.

Average Response Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}$ denote the average response of the n individuals.

What distribution does $\bar{Y}$ follow?

Find $E(\bar{Y})$.

Find $\text{Var}(\bar{Y})$.