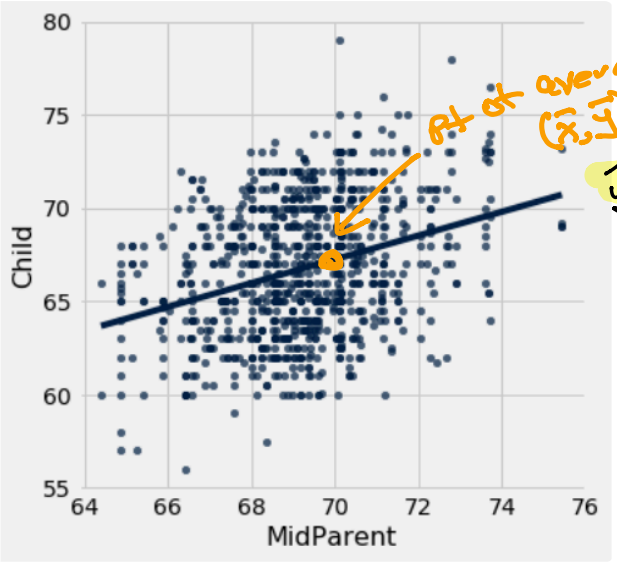
# Stat 88 lec 37

warm up 2:00-2:10

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be random pairs. The average of these pairs $(\bar{x}, \bar{y})$ is called the **point of averages** in the scatter diagram. Show that the point of averages lies on the regression line.




$$\hat{y} = r\frac{SD(y)}{SD(x)}x + \underbrace{\bar{y} - r\frac{SD(y)}{SD(x)}\bar{x}}_{\hat{a}} \qquad \hat{b}$$

Hint: plug $\bar{x}$ into $x$ in the regression line.

$$\hat{y} = r\frac{SD(y)}{SD(x)}\bar{x} + \bar{y} - r\frac{SD(y)}{SD(x)}\bar{x} = \bar{y}$$

Last time

Sec 11.3 and 11.4 linear regression.

For a sample of pairs $(x_1,y_1), (x_2,y_2), \dots, (x_n,y_n)$ the regression line is the best fitting line through your scatter plot:

$$\hat{y} = r\frac{SD(y)}{SD(x)}x + \bar{y} - r\frac{SD(y)}{SD(x)}\bar{x}$$

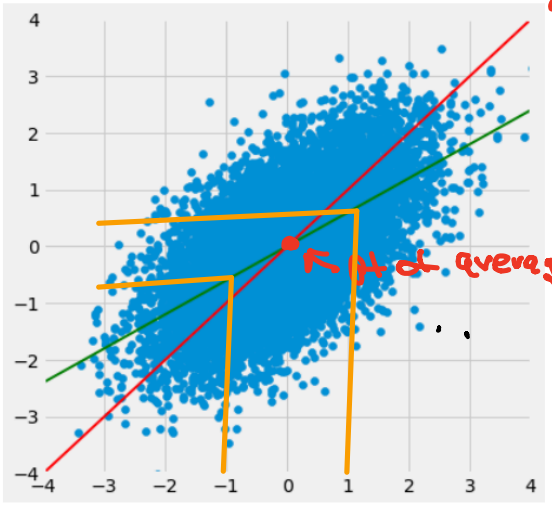
$\hat{a}$ = slope, $\hat{b}$ = intercept

where $r = E(x^* y^*) = \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^*$ where $x^*$ is $x$ in std units.

The regression line for $(x^*, y^*)$ is much simpler

$\hat{y}^* = r x^*$ so some people prefer to work in std units.

$y^*$ = final




diagonal line

$\hat{y}^* = r x^*$

R at of averages

$x^*$ = midterm

regression effect: if you are above the median on the midterm ($x$), you can expect to do relatively worse on the final,

If you are below the median on the midterm you can expect to do better on the final,

Today ① Sec 11.5 The error in regression

② Sec 12.1 The simple linear regression model

## (2) Sec 11.5 The error in regression

The error in the regression estimate is called the **residual** and is defined as

$$D = y - \hat{y}$$

It is useful to write this in terms of the deviations $D_x = x - \mu_x$ and $D_y = y - \mu_y$

$$\hat{y} = \hat{a}x + \hat{b} = \hat{a}x + \mu_y - \hat{a}\mu_x$$
$$= \hat{a}(x - \mu_x) + \mu_y$$

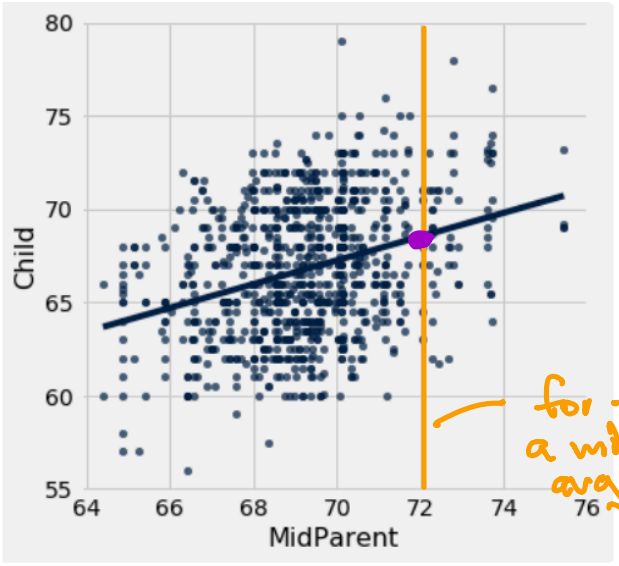
$$D = y - \hat{y} = y - \left[\hat{a}(x-\mu_x) + \mu_y\right]$$
$$= y - \mu_y - \hat{a}(x-\mu_x)$$
$$= D_y - \hat{a}D_x \leftarrow$$

What is $E(D)$? $E(D) = E(D_y) - \hat{a}E(D_x) = 0$

($E(D_y) = 0$, $E(D_x) = 0$)

$E(y - \mu_y) = E(y) - \mu_y = 0$

This says for a fixed $X = x$, the average deviation of $y$ from $\hat{y}$ is 0.




for those parents with a midparent ht = 72, the avg child ht is the pt on the regression line at x = 72.

what is $E(D_X^2)$ ? $E(D_X^2) = Var(D_X) + (E(D_X))^2 = \sigma_X^2$

$E(D_X D_Y)$ ?

$Var(D_X) = Var(X - \mu_X) = Var(X) = \sigma_X^2$ ; $(E(D_X))^2 = 0$

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

$$Var(D) = E(D^2) - (E(D))^2$$

$D^2 = (D_Y - \hat{a} D_X)^2$

$$= E(D_Y^2) - 2\hat{a}E(D_X D_Y) + \hat{a}^2 E(D_X^2)$$

$$= \sigma_Y^2 - 2r\frac{\sigma_Y}{\sigma_X} r\sigma_X\sigma_Y + r^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2$$

$= ?$

$= \sigma_y^2 - 2r^2\sigma_y^2 + r^2\sigma_y^2$

$= \sigma_y^2 - r^2\sigma_y^2 = (1 - r^2)\sigma_y^2$

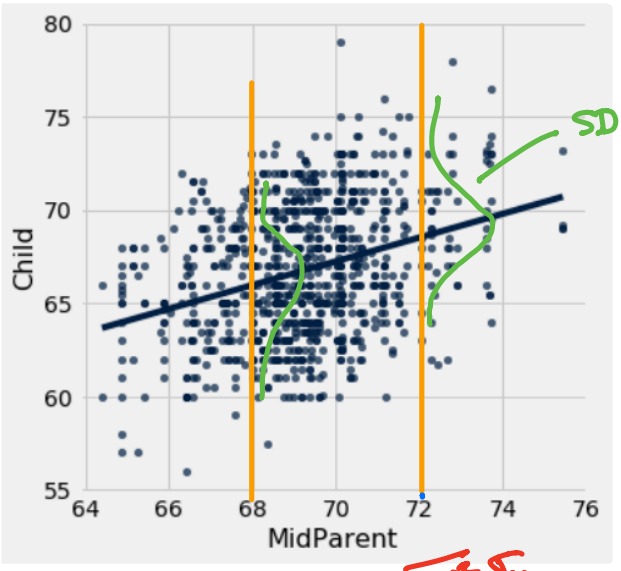
$\Rightarrow SD(D) = \sqrt{1-r^2}\,\sigma_y$

The residuals are normally distributed

so

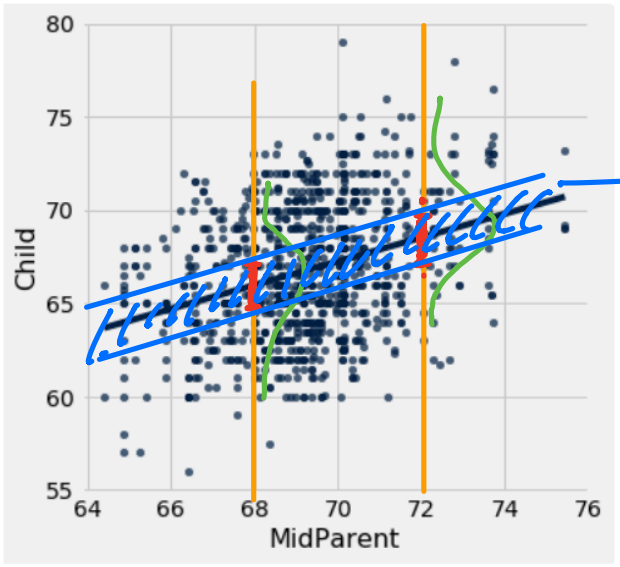
$$D \sim N\left(0, (1-r^2)\sigma_y^2\right)$$

Picture




I have marked 1 SD $= \sqrt{1-r^2}\,\sigma_y$ on either side of $\hat{y}$ below.

Shade the middle 2/3 of your data.



ex

. For about 1/3 of the students, the prediction for the final score was off by more than how many points?

x average midterm score $\approx 50$, SD $\approx 25$
y average final score $\approx 55$, SD $\approx 15$, $r \approx 0.60$

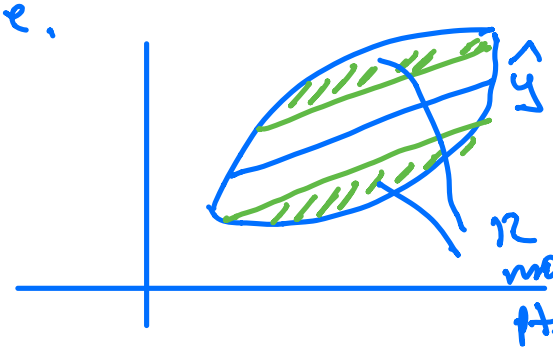
**a** 6 pts
**b** 9 pts
**c** 12 pts
**d** 15 pts

$\frac{1}{3}$ of your data is at least 1 SD away from regression line.

$SD(D) = \sqrt{1-r^2}\,\sigma_y$

$= \sqrt{1-.6^2}\,(15)$

$= (.8)(15) = 12$



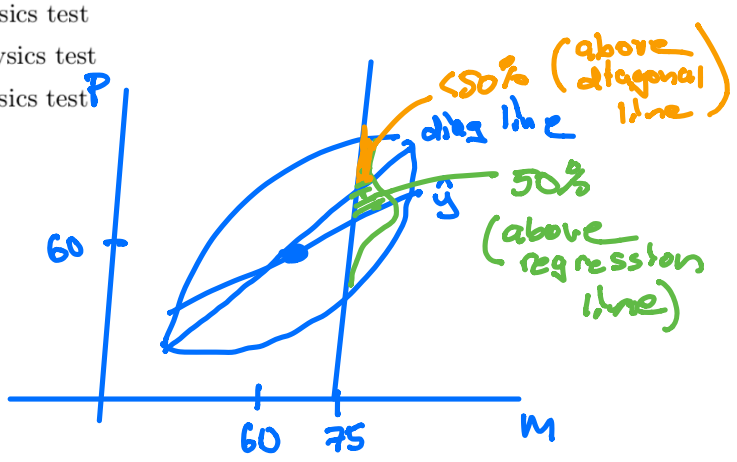

12 or more pts away from regression line.

ex

At Cal's engineering school, students are required to take aptitude tests in various subjects, including mathematics and physics. Students who do wellonthe mathematics test also tend to score high on the Physics tests. On both tests, the average is 60, and the spreads are about the same for both tests. The scatter diagram is football shaped Of the students scoring about 75 on the Mathematics test:

**i** just about half scored over 75 on the Physics test
**ii** more than half scored over 75 on the Physics test
**iii** less than half scored over 75 on the Physics test

Choose one option and explain.



## ② Sec 12.1 The Simple regression model

small x (not random)

$x$ = predictor variable

$Y$ = response variable

big Y (random)

assumption:

Formally, for individuals $i = 1, 2, 3, \ldots, n$, the response is assumed to be

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$Y_i$: response; $\beta_0 + \beta_1 x_i$: signal; $\epsilon_i$: random error

where:

- $\beta_0$ and $\beta_1$ 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 $(0, \sigma^2)$ random variables.
- The error variance $\sigma^2$ is an unobservable constant parameter, and is assumed to be the same for all individuals $i$.

we see the responses $Y_1, Y_2, \vdots, Y_n$

and we want to pick out the signals

$Y_1 = \beta_0 + \beta_1 x_1$,

$Y_2 = \beta_0 + \beta_1 x_2$

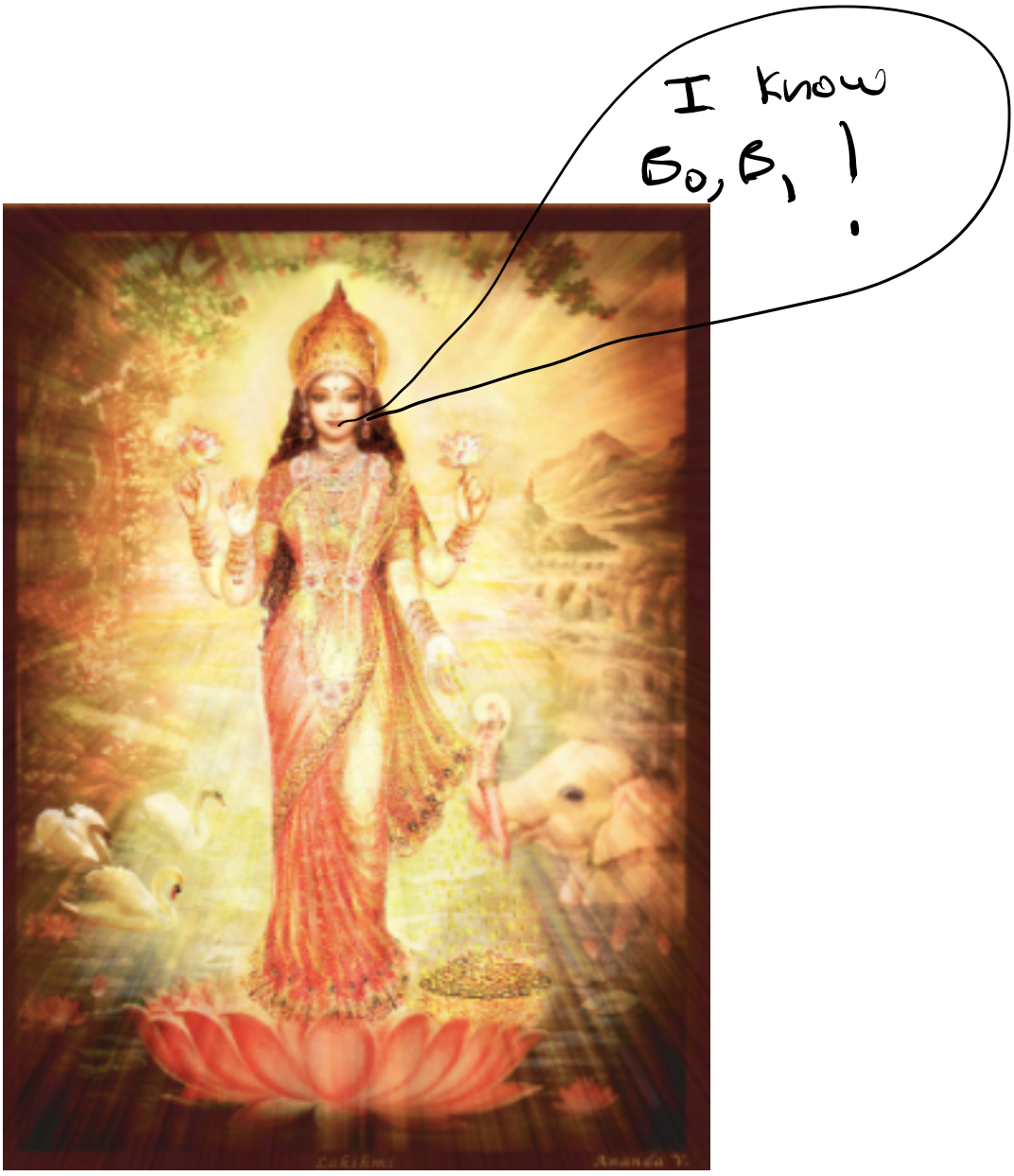
$\vdots$

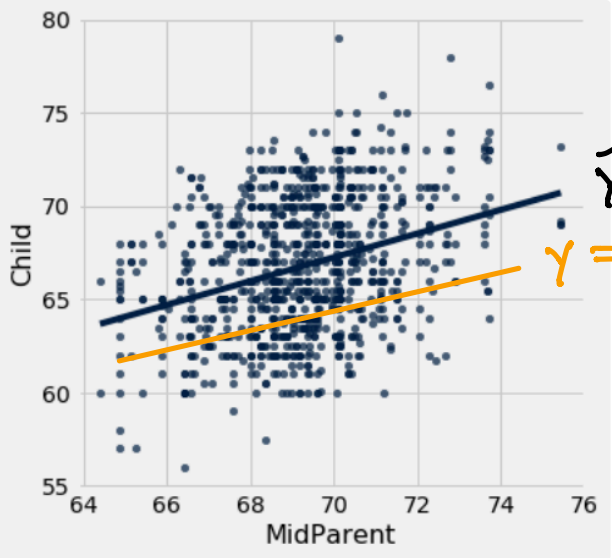
$Y_n = \beta_0 + \beta_1 x_n$

I call the line $Y = \beta_0 + \beta_1 X$ the Tyche line.

The Tyche line is signal part of the response

Image of Tyche the goddess of fortune.
Only Tyche knows $B_0, B_1$.






$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ regression line

$Y = \beta_0 + \beta_1 X$ Tyche line

The regression line estimates the Tyche line.

using $\hat{\beta}_1$ we can hypothesize whether $\beta_1$ is zero (i.e whether Y and x are independent)

For example is the height of the child dependent on the height of the parents?