Stat 88 lec 36

warm up 2:00-2:10

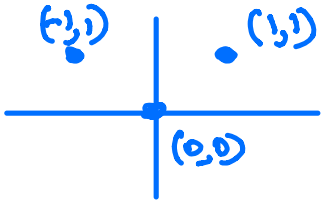
exercise 11.6.11

11. Let X have the uniform distribution on the three points $-1, 0$, and 1. Let $Y = X^2$.

$$r = \frac{E(D_x D_y)}{\sigma_x \sigma_y}$$

a) Show that X and Y are uncorrelated.

Picture



we show that $r = 0$

i.e $E(D_x D_y) = 0$.

$X \sim \text{Unif}(-1,0,1) \Rightarrow \mu_X = E(X) = \frac{1+(-1)}{2} = 0$

$\mu_Y = E(Y) = E(X^2) = \frac{1}{3}\left((-1)^2 + (0)^2 + (1)^2\right) = 2/3$

$E(D_x D_y) = E\left((X-\mu_X)(Y-\mu_Y)\right) = E(X(Y-2/3))$

$= \frac{1}{3}\left(-1(1-2/3) + 0(0-2/3) + 1(1-2/3)\right)$

$= \frac{1}{3}\left(-1/3 + 0 + 1/3\right) = 0$

$\Rightarrow r = \frac{E(D_x D_y)}{\sigma_x \sigma_y} = \frac{0}{\sigma_x \sigma_y} = 0 \Rightarrow X, Y$ are uncorrelated

$Y = X^2$ so Y is dependent on X

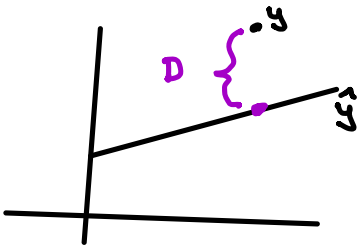
So uncorrelated $\not\Rightarrow$ indep.

But if you scatter plot is football shaped then uncorrelated $\Leftrightarrow$ independent.

Last time

Sec 11.5 The error in regression

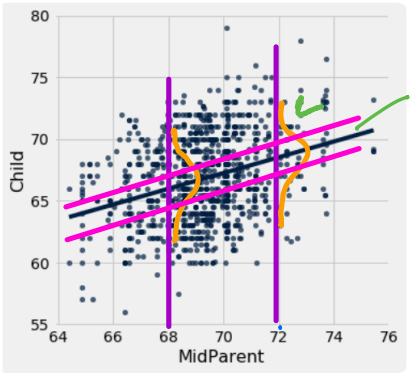
let $D = \hat{y} - y$ be the residual error



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

$E(D)$ $\quad$ $Var(D)$

2/3 of your data are within two lines parallel to $\hat{y}$ $1\,SD = \sqrt{1-r^2}\,\sigma_y$ away.



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

Sec 12.1 the simple regression model

small x (not random)

x = Predictor variable

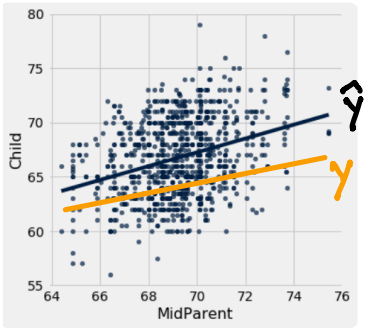
Y = response variable

big Y (random)

we assume for each of n observations

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

- $x_1, x_2, \ldots, x_n$ are fixed
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$
- only Tyche knows β_0, β_1, and σ^2
- $Y = \beta_0 + \beta_1 x$ is called the Tyche line.



regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

Tyche line is the signal part of the response

The regression line estimates the Tyche line.

Today

1. Sec 12.1 The simple linear regression model
2. Sec 12.2 The distribution of the estimated slope.

Sec 12.1 The simple linear regression model

Is the response variable dependent on the predictor variable(s)? In other words is the typical line flat? This is the question we answer in Chapter 12.

Individual responses

Fix β_0, β_1, x_i:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal (constant)}} + \underbrace{\varepsilon_i}_{\text{error (random) } N(0,\sigma^2)} \qquad \text{is } N(\beta_0+\beta_1 x_i, \sigma^2)$$

Y_i: response (random)

$Y_1, \ldots, Y_n$ are independent since $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

Average responses

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be average response of indep normals is normal

Show $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$? $\quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(\beta_0+\beta_1 x_i + \varepsilon_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} (\beta_0+\beta_1 x_i + \underbrace{E(\varepsilon_i)}_{0})$$

$$= \beta_0 + \beta_1 \bar{x} \quad \checkmark$$

Furthermore,

$$Var(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

So $\bar{Y} \sim N\left(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n}\right)$

(2) Sec 12.2 The distribution of the estimated slope.

To make a hypothesis about whether $\beta_1 = 0$ we need to know the distribution of $\hat{\beta}_1$

$\hat{\beta}_1$ replaces $\hat{a}$ from chapter 11.

recall $\hat{a} = r\frac{\sigma_y}{\sigma_x} = \frac{E(D_x D_y)}{\sigma_x^2} = \frac{E((x-\mu_x)(y-\mu_y))}{\sigma_x^2}$

Estimated Slope

For our sample,

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

($\bar{x}$ ← μ_x, $\bar{Y}$ ← μ_y, denominator ← σ_x^2)

$\hat{\beta}_1$ is a linear combination of independent normals $Y_1, \ldots, Y_n$ and hence $\hat{\beta}_1$ is normal

Expectation of the estimated slope

We have $E(Y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i)$, where $E(\varepsilon_i) = 0$

$$E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$$

So $E(Y_i - \bar{Y}) = \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x}$

$$= \beta_1 (x_i - \bar{x})$$

Show $E(\hat{\beta}_1) = \beta_1$?

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\beta_1 (x_i - \bar{x})$

$$E(\hat{\beta}_1) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,E(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 \frac{\frac{1}{n}\sum(x_i - \bar{x})^2}{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

$$= \beta_1$$ ✓

Hence $\hat{\beta}_1$ is an unbiased estimate of β_1

We need $\text{Var}(\hat{\beta}_1)$ to know the distribution and can make a 95% CI for β_1.

Variance of the estimated slope :

error variance, $Var(\varepsilon_i)$, is an unknown parameter

FACT $Var(\hat{\beta}_1) = \frac{\sigma^2}{n\,var(x)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Note as $n \to \infty$, $\sum_{i=1}^{n}(x_i-\bar{x})^2$ gets very large and $Var(\hat{\beta}_1) \to 0$ so the difference between $\hat{\beta}_1$ and β_1 becomes very small with high probability.

Hence $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n\,var(x)}\right)$

exercise 12.4.1

1. Recall that the intercept of the regression line is given by "the average of Y minus the slope times the average of x. That is, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$. Is $\hat{\beta}_0$ an unbiased estimator of β_0?

$E(\bar{y}) = \beta_0 + \beta_1\bar{x}$

$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$

($E(\bar{y}) = \beta_0 + \beta_1\bar{x}$, $E(\hat{\beta}_1) = \beta_1$)

Standard error of the estimated slope :

We have $SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{n\,var(x)}}$ (unknown) where σ is SD of the errors.

We estimate σ with the SD of the residuals. This is calculated in Python.

When the SD of an estimator is approximated by the data it is called the SE (standard error).

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$ is approximately normal for large n.

Ex (Pulse Rates)

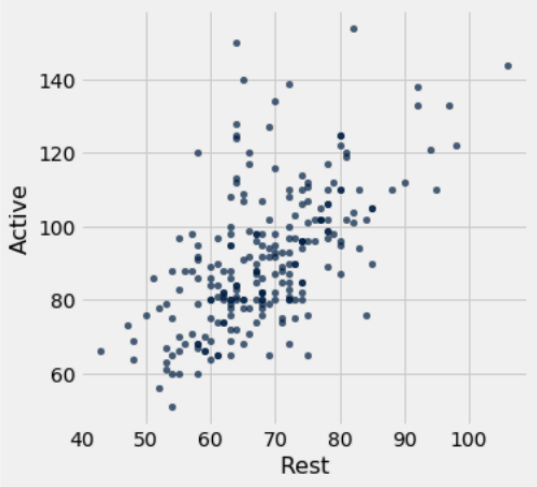
We wish to predict active pulse rates from resting pulse rates.

```
pulse
```

Active	Rest	Smoke	Sex	Exercise	Hgt	Wgt
97	78	0	1	1	63	119
82	68	1	0	3	70	225
88	62	0	0	3	72	175
106	74	0	0	3	72	170
78	63	0	1	3	67	125
109	65	0	0	3	74	188
66	43	0	1	3	67	140
68	65	0	0	3	70	200
100	63	0	0	1	70	165
70	59	0	1	2	65	115

... (222 rows omitted) (n=232)

```
pulse.scatter('Rest', 'Active')
```




```
active = pulse.column(0)
resting = pulse.column(1)
```

```
stats.linregress(x=resting, y=active)
```

Output:

$\hat{\beta}_1$	(1.142879681904831,
$\hat{\beta}_0$	13.182572776013345,
r	0.6041870881060092,
P-val	1.7861044071652305e-24,
$SE(\hat{\beta}_1)$	0.0993888436389145)

$n=232$ is large so $T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$

A 95% CI for β_1 is

$\hat{\beta}_1 \pm 2\,SE(\hat{\beta}_1) = (0.944, 1.342)$

this doesn't contain 0 so we reject the null $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

under the null

$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = 11.5$ (0 ← β_0)