STAT 88: Lecture 38

Contents

Warm up:

Let (X, Y) be a random pair and we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$ from the linear regression model

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

(a) We defined the mean squared error of a linear function of X as

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

How would you estimate $\text{MSE}(a, b)$ from $(x_1, Y_1), \ldots, (x_n, Y_n)$?

(b) How can you estimate the best regression line, $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$?

(c) Compare the regression line in (b) with the population regression line $\widehat{Y} = \widehat{a}X + \widehat{b}$.

Last time

The simple regression model

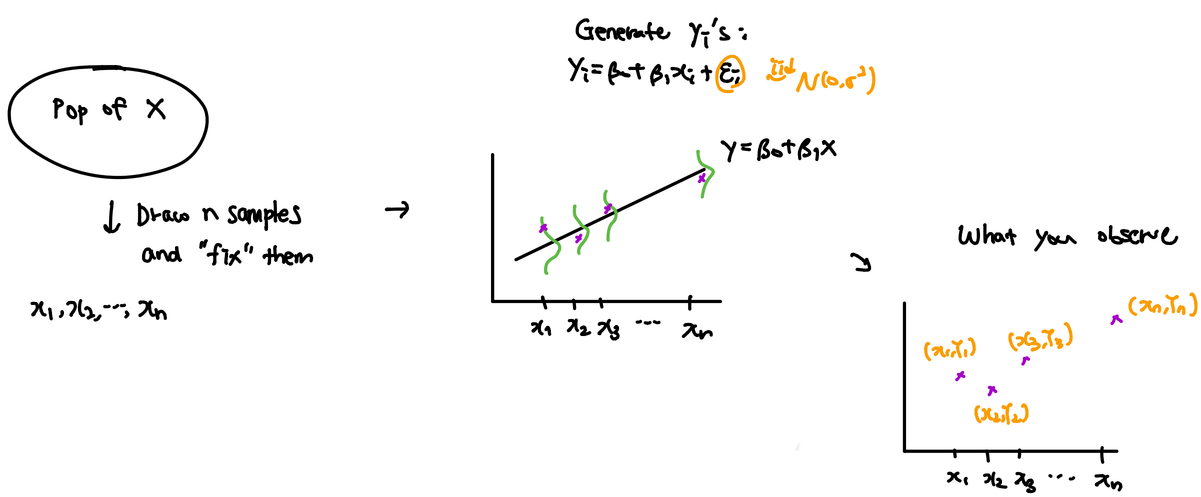
$$Y = \text{response} \quad \text{and} \quad x = \text{predictor variable/covariate/feature}$$

We assume for each of n observations

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

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.



Goal: Estimate β_0 and β_1 from $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$

12.2. The Distribution of the Estimated Slope

Estimated Slope

The least-squares estimate of the true slope β_1 is the slope of the regression line, given by

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

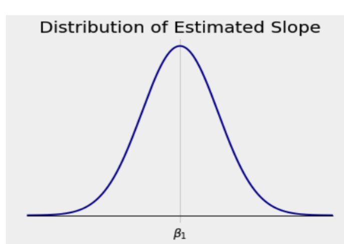
Is $\widehat{\beta}_1$ random or constant? What distribution does $\widehat{\beta}_1$ follow?

Expectation of the Estimated Slope

Let's find $E(\widehat{\beta}_1)$.

First, what is $E(Y_i)$ and $E(\bar{Y})$?

Hence $\widehat{\beta}_1$ is an unbiased estimator of β_1.



Variance of the Estimated Slope

FACT:

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

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

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Example: (Exercise 12.4.1) 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, $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$. Is $\widehat{\beta}_0$ an unbiased estimate of β_0?

Standard Error of the Estimated Slope

We have

$$\mathrm{SD}(\widehat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

σ is unknown so we have to estimate it. Recall σ is the SD of the error, $\mathrm{SD}(\epsilon_1) = \sigma$. So we estimate σ with the SD of the residuals. If

$$D_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i,$$

then

$$\widehat{\sigma} = \mathrm{SD}(D_1, \ldots, D_n) = \frac{1}{n} \sum_{i=1}^n (D_i - \bar{D})^2.$$

This can be calculated in Python.

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

$$\mathrm{SE}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

When n is large, it can be shown $\mathrm{SE}(\widehat{\beta}_1)$ converges to $\mathrm{SD}(\widehat{\beta}_1)$ so

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)}$$

is approximately $\mathcal{N}(0,1)$ for large n.

Pulse Rates

Y X

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

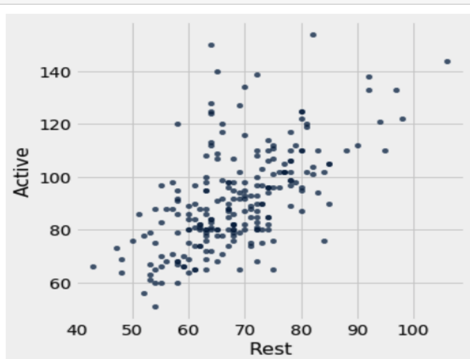
```
pulse
```

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)

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

Assumptions of simple linear regression model?



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

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

Output:

```
(1.142879681904831,
 13.182572776013345,
 0.6041870881060092,
 1.7861044071652305e-24,
 0.09938884436389145)
```

$\widehat{\beta}_1$

$\widehat{\beta}_0$

r

p-val

$SE(\widehat{\beta}_1)$

$n = 232$ is large so

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \sim \mathcal{N}(0,1).$$

A 95% CI for β_1 is

$$(\widehat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\widehat{\beta}_1)) = (0.944, 1.342).$$

A fundamentally important question is whether the true slope β_1 is 0. If it is 0, then the resting pulse rate isn't involved in the prediction of the active pulse rate, according to the regression model. Our testing problem is

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0.$$

T is our test statistic. Under H_0,

$$T = \frac{\widehat{\beta}_1}{\mathrm{SE}(\widehat{\beta}_1)} \sim \mathcal{N}(0,1).$$

The observed value of the test statistic is 11.5. So the p-value is

$$\text{p-value} = P(T \geq 11.5) + P(T \leq -11.5) \approx 0.$$

We reject H_0 at 5% level.