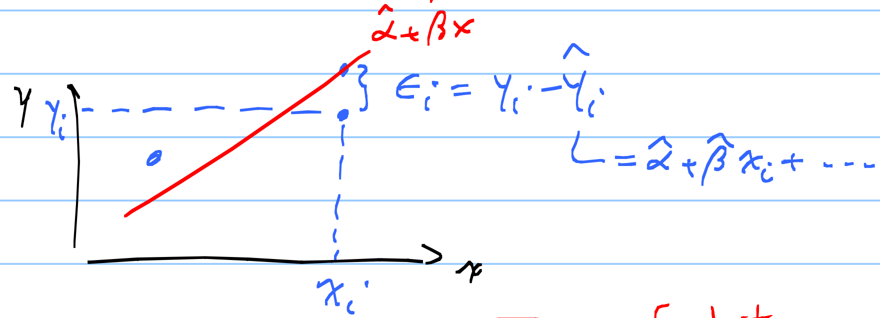
Lecture 24 (Ch. 11)

We did regression $y_i = \alpha + \beta x_i + \cdots + \epsilon_i$ — Ch. 3.

We did inference on $\mu, \pi, \mu_1 - \mu_2, \pi_1 - \pi_2, \pi_i, \ldots$ — Ch 7, 8.

Now we do inference on β (and α), y, ... — Ch. 11.

Review:



For a sample we write $y_i = \alpha + \beta x_i + \epsilon_i$ — arbitrary params to be estimated by OLS, ie. $\frac{\partial}{\partial \alpha} SSE$, etc.

and $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

where $\underbrace{\hat{\alpha}, \hat{\beta}}$ (a, b in book) are the OLS estimates of α, β, ie.

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}.$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Recall that Sample var. $= s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{S_{xx}}{n-1}$

For a population, there exists an OLS fit as well! Just because we can fit a line to the whole pop., it does not follow that there are no errors when predicting y from x!

So, now, in regression we need notation that can distinguish between pop. and sample quantities; like $\bar{x}$ and μ_x, but for regression. I will use the following notation for the predictions:

$\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$ (for sample) $\qquad y(x) = \alpha + \beta x$ (for population)

But, in Ch. 11, you have to keep in mind that this α, β are <u>NOT</u> free params that we can do things like $\frac{\partial}{\partial \alpha} SSE$, etc.; they <u>are</u> fixed quantities obtained by "fitting" a line to the whole population.

Then There is The Analysis of Variance:

$$SST = \sum (y_i - \bar{y})^2 = \underbrace{SS_{explained}}_{\hat{\beta} S_{xy}} + \underbrace{SS_{unexpl.}}_{SSE}$$

df: $n-1 \quad = \quad k \quad + \quad n-(k+1)$

$R^2 = \dfrac{SS_{expl.}}{SST}$

percent of var. in y explained by x ... (Goodness of fit)

of β's (excluding α)

of predictors

$y = \alpha + \beta_1 x_1 + \cdots \beta_k x_k$

$S_e = \sqrt{\dfrac{SSE}{n-(k+1)}}$ ~RMSE

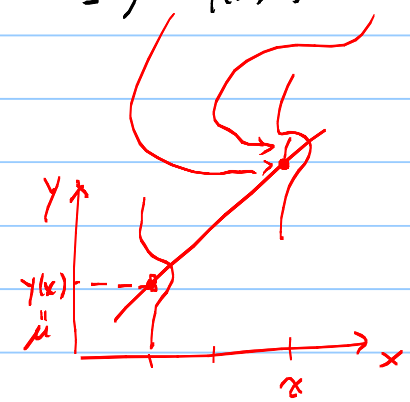
std. dev. of errors ~ Typical error or spread about fit.

Now, to do inference we need a Probability model (for regression):

Assume y's are Normally distr. at each x, with params $\mu = y(x)$, $\sigma = \sigma_\epsilon$

e.g. $\mu = y(x) = \alpha + \beta x + \cdots$ $\quad$ $\sigma = \sigma_\epsilon$ = fixed

estimate α, β with $\hat{\alpha}, \hat{\beta}$ $\quad$ estimate, with S_e



Note: $y \sim N(y(x), \sigma_\epsilon)$

$\epsilon = y - y(x) \sim N(0, \sigma_\epsilon^2)$

This allows us to say things like:

1) $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$ = estimates mean of y, given x

2) In about 95% of The cases, we expect to have y-values within $y(x) \pm 1.96\, \sigma_\epsilon$, for a given x

like 95% of cases are within $\mu \pm 1.96\sigma$ (Ch. 1)

3) other probs, e.g. $\text{prob}(a < y < b \mid x) =$

True prediction = True mean | x.

$$\text{prob}\left(\frac{a - y(x)}{\sigma_\epsilon} < \boxed{\frac{y - y(x)}{\sigma_\epsilon}} < \frac{b - y(x)}{\sigma_\epsilon} \right) = \text{Table I}$$

$z \sim N(0,1)$

like $pr(a < x < b) =$ (Ch. 1)

$$pr\left(\frac{a-\mu}{\sigma} < \boxed{\frac{x-\mu}{\sigma}} < \frac{b-\mu}{\sigma} \right)$$

$z \sim N(0,1)$

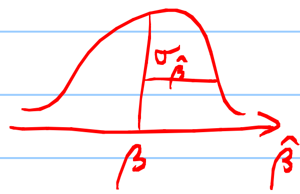
Let's build a CI (and hyp. test) for ONE β : $Y_i = \alpha + \beta x_i + \epsilon_i$

Theorem: If $\epsilon \sim N(0, \sigma_\epsilon^2)$, Then $\hat\beta$ is normal with params:

Expected value (or mean) of The sampling dist. of $\hat\beta$

$E[\hat\beta] \equiv \mu_{\hat\beta} = \beta$ ← pop. slope

$\sqrt{V[\hat\beta]} \equiv \sigma_{\hat\beta} = \dfrac{\sigma_\epsilon}{\sqrt{S_{xx}}} = \dfrac{\sigma_\epsilon}{\sqrt{n-1}\, s_x}$



Ch. 7
If $x \sim N(\mu_x, \sigma_x^2)$, Then $\bar x$ is Normal with params
$E[\bar x] = \mu_{\bar x} = \mu_x$
$\sqrt{V[\bar x]} = \sigma_{\bar x} = \sigma_x / \sqrt{n}$

$S_{xx} = \sum_i^n (x_i - \bar x)^2 = (n-1)\, s_x^2$
Defn. of sample var.

Recall σ_ϵ is const. and s_x does not vary as $\frac{1}{n-1}$ because of $\sum$ in The numerator of s_x.
So, $\sigma_{\hat\beta}$ falls off as $1/\sqrt{n}$

Q1: What is The quantity That has a std. Normal dist?

A) $\hat\beta$ B) $\dfrac{\hat\beta - \mu_Y}{\sigma_Y/\sqrt{n}}$ C) $\dfrac{\hat\beta - \beta}{\sigma_\beta/\sqrt{n}}$ D) $\dfrac{\hat\beta - \beta}{\sigma_{\hat\beta}}$

If $w \sim N(\mu_w, \sigma_w)$, Then $z = \dfrac{w - \mu_w}{\sigma_w} \sim N(0,1)$.

$z = \dfrac{\hat\beta - \beta}{\sigma_{\hat\beta}} = \dfrac{\hat\beta - \beta}{\sigma_\epsilon / \sqrt{S_{xx}}} \sim N(0,1)$

$t = \dfrac{\hat\beta - \beta}{s_e / \sqrt{S_{xx}}} \sim$ t-dist. $df = n - 2$ ($k+1$)

Ch. 7
$z = \dfrac{\bar x - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$t = \dfrac{\bar x - \mu}{s/\sqrt{n}} \sim$ t-dist. $df = n-1$

Then, self-evident fact gives:

C.I. for β : $\hat\beta \pm t^* \dfrac{s_e}{\sqrt{S_{xx}}}$ $df = n-2$ (Table VI or IV) ($k+1$)

$H_0 : \beta \;\square\; \beta_0$

$H_1 : \beta \;\square\; \beta_0$

$t_{obs} = \dfrac{\hat\beta_{obs} - \beta_0}{s_e / \sqrt{S_{xx}}}$

$df = n-2$ ($k+1$)

p-value $= (1,2) \cdot pr(\hat\beta \;\square\; \hat\beta_{obs}) = pr(t \;\square\; t_{obs}) =$ Table VI

1 or 2-sided.

problem 11.17 [Revised; remove the word "positive", ie. do 2-sided]

$n = 13$ $\quad x$ = nickel content, $\quad y$ = percentage austentite.

Data: $\quad \Sigma(x_i - \bar{x})^2 = 1.183 \quad = S_{xx}$

$\Sigma(y_i - \bar{y})^2 = 0.0508 \quad = S_{yy}$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 0.2073 \quad = S_{xy}$

Question: Is There a statistically significant ($\alpha = 0.05$) relationship between x and y?

1) C.I. β: $\quad \hat{\beta} \pm t^* \, S_e / \sqrt{S_{xx}}$

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{.2073}{1.183} = .1752$

$SSE = SST - \hat{\beta} S_{xy}$

$= .0508 - (.1752)(.2073)$

$= .014$

$S_e = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{.014}{13-2}} = 0.0357$

$\therefore$ 95% CI for β: $\quad .1752 \pm \underset{df = 13-2}{2.201} \left(\dfrac{.0357}{\sqrt{1.183}} \right) = 0.0328 \quad = (0.10, 0.24)$

We are 95% Confident That The pop. β is in here.

Also, zero is not included $\Rightarrow$ Relationship is statistically significant

2) $H_0: \beta = 0$ $\qquad t_{obs} = \dfrac{.1752 - 0}{.0328} = 5.31$,

$H_1: \beta \neq 0$

p-value $= 2\,pr(\hat{\beta} > \hat{\beta}_{obs}) = 2\,pr(t > t_{obs})$

$= 2\,pr(t > 5.31) < 0.001$

Table VI

df = 13-2

p-value $< \alpha$

$\therefore$ Evidence That $\beta \neq 0$. (same conclusion as above).

In Summary, we have 2 ways of testing an association between x, y.

(A Third way, next FYI)

FYI

Note that the test of $\beta = 0$ is equivalent to testing if there is a linear relationship between x and y. But if a linear relationship is all that you are testing, then we can test the population correlation coeff

$H_0: \rho = 0$

$H_1: \rho \neq 0$

The test statistic for this test is a bit weird:

$$\boxed{t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}}}$$ has a t distr. with $df = n-2$.

Recall $r = S_{xy} / \sqrt{S_{xx} S_{yy}}$

This way, you take your data (x_i, y_i), compute the sample correl. coeff (r), then t_{obs}, and then p-value, all without any fitting.

3) For the above example:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \cdots = .8456$$

$$t_{obs} = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = \cdots = 5.3$$ ← same value as t_{obs} we got above when testing β.

p-value $= 2\,\text{prob}(t > t_{obs})$ = same as above.

∴ same conclusion.

In summary: We have 3 ways of testing if there is a useful relation between x & y:

1) C.I. for β 2) Testing $H_0: \beta = 0$ 3) $H_0: \rho = 0$

hw-lect 24-1 (by R)

The very beginning of section 3.3 in lab4 shows how to make/simulate data on x and y that are linearly associated. The x data consists of 100 cases from a uniform distribution, and the TRUE/population relationship between x and y is given by y = 10 + 2x.

a) What is the value of sigma_epsilon in that simulation?

b) Using the same settings used in section 3.3, write code to build the (empirical) sampling distribution of beta_hat based on 5000 trials. This code should produce a histogram.

c) According to the lecture, the mean of the histogram is supposed to be equal (or close) to what quantity? Is it?

d) According to the lecture, the standard deviation of that histogram is supposed to be equal (or close) to what quantity? Is it?

e) According the lecture, the distribution of the beta_hat is supposed to be normal with certain parameters. Use qqnorm() and abline () to confirm that.

hw-lect 24-2

In a problem dealing with flow rate (y) and pressure drop (x) acrsss filters, it is known That $y = -0.12 + 0.095x$. I.e. This is The "fit" to The population. Suppose it is also known That $\sigma_\epsilon = 0.025$. Now, IF we were to make repeated observations of y when $x = 10$, what's The prob. of a flow rate exceeding 0.835?

Hint: $\dfrac{y - (\text{true mean of } y \text{ at some } x)}{\sigma_\epsilon} \sim N(0,1)$.