Lecture 25 (Ch. 11)

We are back in the realm of regression, and so far we have made inferences about the slope regression coefficient β (α is in 11.11).

Q What about the true (pop.) prediction itself? $y(x) = \alpha + \beta x + \ldots$

Unfortunately, the prediction $\hat{y}(x)$ has 2 different meanings:

- (point estimate of) The true/pop. conditional mean of y, given x. ← discussed last time
- (point) prediction of a single y, given x

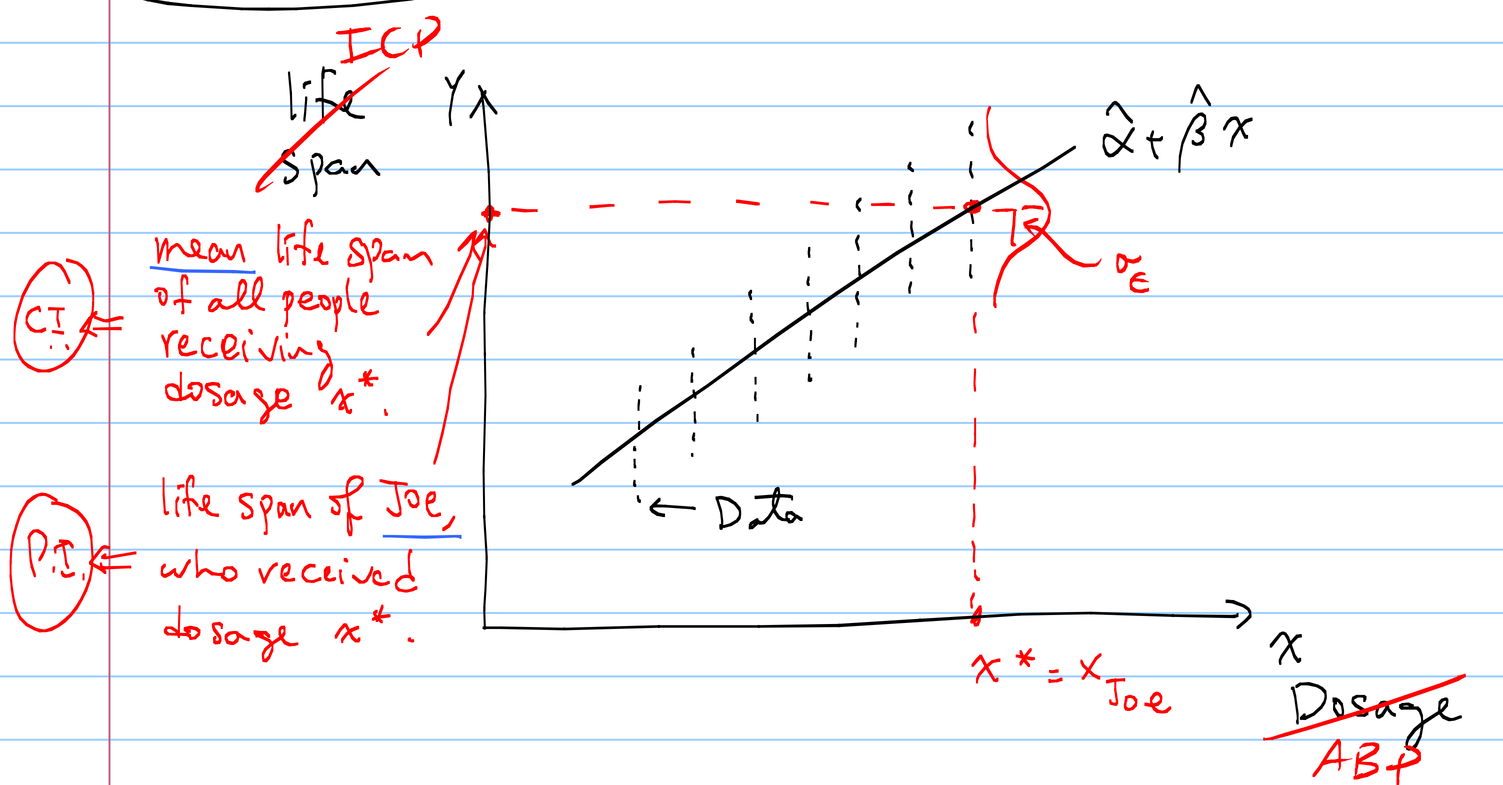
Note: The prediction $\hat{y}(x)$ is the same in both cases.
But the interpretation is different ⟹ different intervals & tests.

The two intervals/tests answer 2 diff. questions:

→ What's the true cond'l mean of y for all cases, given $x = x^*$?

→ What's the predicted y for an individual case at $x = x^*$?

Example:



The first interval is just a confidence interval because it pertains to a pop. param (ie. true mean of y, given x).

The 2^{nd} interval is not a conf. interval at all!
It is called a Prediction Interval (P.I.).

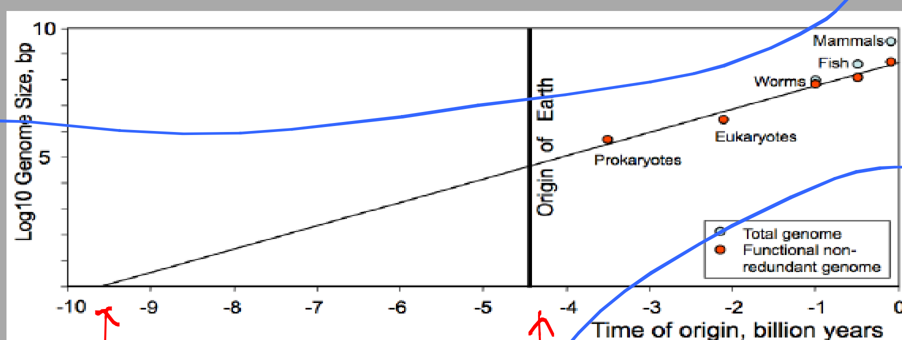
The "levels" of the two intervals are often called
confidence level and prediction level

C.I/P.I. are important, because they allow us to make uncertainty "bands". Without them, wrong conclusions may follow. E.g.

Sharov & Gordon (2013) "Life Before Earth":

What is most interesting in this relationship is that it can be extrapolated back to the origin of life. Genome complexity reaches zero, which corresponds to just one base pair, at time ca. 9.7 billion years ago (Fig. 1). A sensitivity analysis gives a range for the extrapolation of ±2.5 billion years (Sharov, 2006). Because the age of Earth is only 4.5 billion years, life could not have originated on Earth even in the most favorable scenario (Fig. 2). Another complexity measure yielded an estimate for the origin of life date about 5 to 6 billion years ago, which is similarly not compatible with the origin of life on Earth (Jørgensen, 2007). Can we take these estimates as an approximate age of life in the universe? Answering this question is not easy because several other problems have to be addressed. First, why the increase of genome complexity follows an exponential law instead of fluctuating erratically? Second, is it reasonable to expect that biological evolution had started from something equivalent in complexity to one nucleotide? And third, if life is older than the Earth and the Solar System, then how can organisms survive interstellar or even intergalactic transfer? These problems as well as consequences of the exponential increase of genome complexity are discussed below.

Uncertainty Bands



origin of life

Earth's age

From this they conclude that life predates Earth, and that life must have been formed on some other planet, then transported to Earth.

In a follow-up paper (Marzban et al. (2014): "Earth Before Life", Biology Direct 9:1) we showed that there are (at least) 2 problems with that analysis

1) Extrapolation is bad!
2) Uncertainty Bands must be considered.

1) C.I. for the population mean, $y(x)$, given x:

We need the sampling distr. of $\hat{y}(x)$.

Analogous to the sampl. dist. of $\bar{x}$.

$\bar{x} \sim N(\mu, \sigma^2/n)$

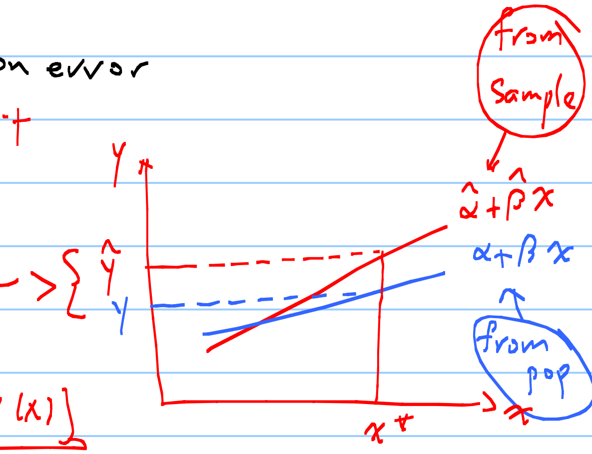
The sampl. distr. of $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$ is Normal with params:

$\mu = y(x) = \alpha + \beta x, \quad \sigma^2 = \sigma^2_{\text{estimation error}}$

where

sample fit   pop. fit

estimation error $= \underset{\underset{\hat{\alpha}+\hat{\beta}x}{\parallel}}{\hat{y}(x)} - \underset{\underset{\alpha+\beta x}{\parallel}}{y(x)}$



$\sigma^2_{est.err} = V[est.\ err] = \underbrace{V[\hat{y}(x)]}_{} + \underbrace{V[y(x)]}_{=0}$

$\sigma^2_{est.err} = \sigma^2_{\hat{y}} + 0$

because $y(x) =$ pop fit.

Approximate the σ's by sample estimates:

$$S^2_{est.\ err.} = S^2_{\hat{y}} + 0 \implies \boxed{S^2_{est.err} = S^2_{\hat{y}}}$$

where $\boxed{S^2_{\hat{y}} = S^2_e \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}$ No proof!

It follows that

estimation error

$$z = \frac{\boxed{\hat{y}(x) - y(x)}}{\sigma_{est.err}} \sim N(0,1),$$

$$t = \frac{\hat{y}(x) - y(x)}{S_{est.err}} \sim t\text{-dist.} \quad df = n-2$$

$k+1$

I.e. C.I. for mean $y(x)$, given x:

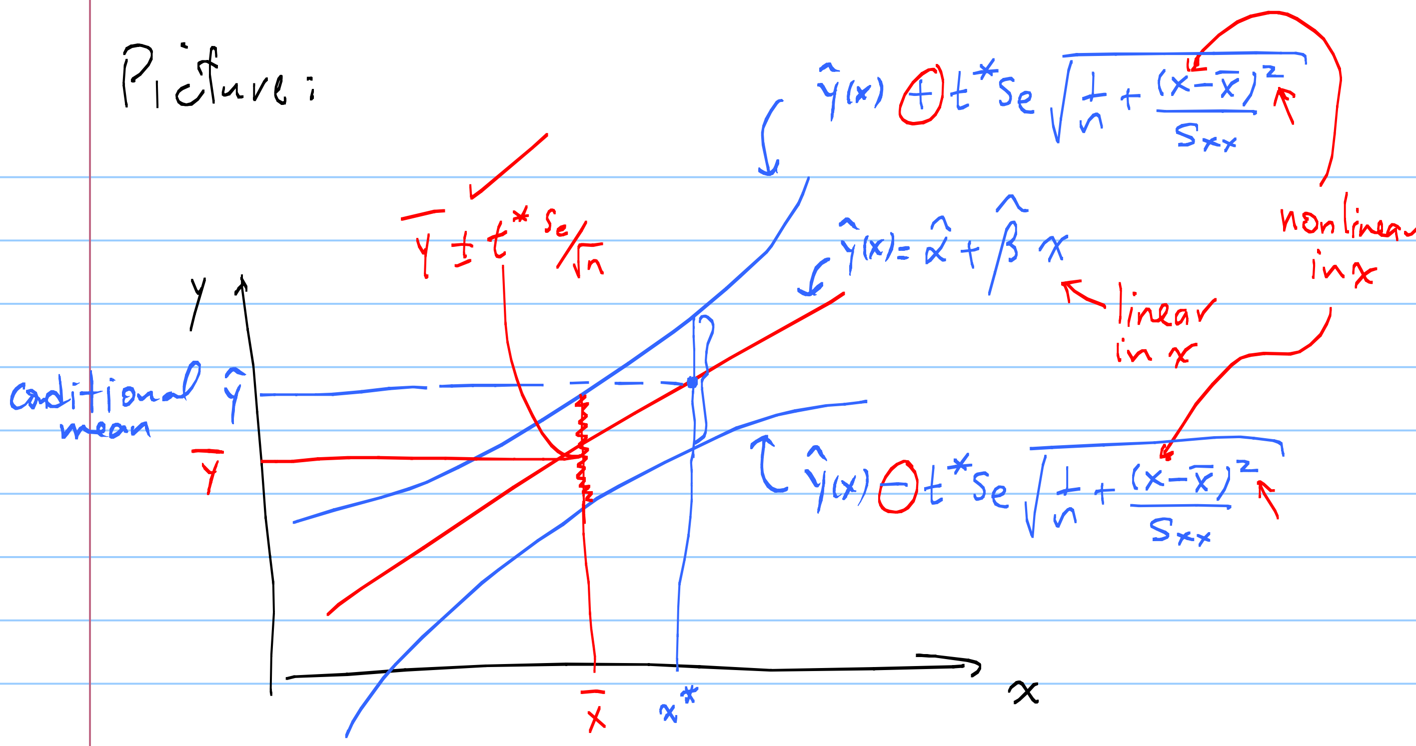
Table IV

$df = n-2$.

$k+1$

$$\hat{y}(x) \pm t^* S_{est.\ err} = \hat{y}(x) \pm t^* S_{\hat{y}} = \hat{y}(x) \pm t^* S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

Picture:



Note: The C.I. gets wider the farther x gets from $\bar{x}$. Why?

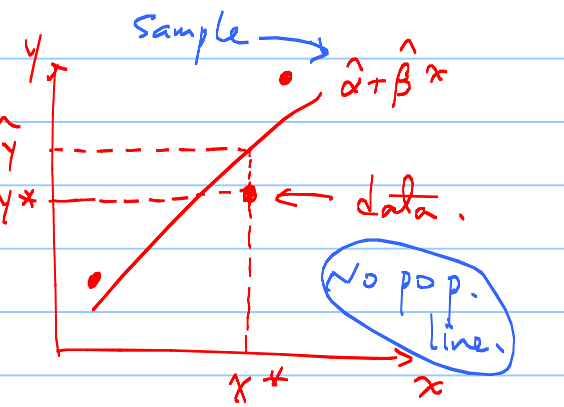
Regression has the property where the fit must go through the point $(x, y) = (\bar{x}, \bar{y})$. So, now, imagine a line that is fixed at that point. Any uncertainty in the slope will then cause the line the sweep a larger vertical direction in regions far away from $x = \bar{x}$.

2) Prediction Interval (P.I.) for a single y.

Suppose y^* is Joe's y value corresponding to his x-value, x^*.

A theorem states that $\overbrace{(\hat{y}(x) - y^*)}^{\text{prediction error}}$ has a normal distr. with params

$\mu = 0$, $\sigma^2 = \sigma^2_{\text{prediction error}}$



where prediction error $= \hat{y}(x) - y^*$

$$\sigma^2_{\text{pred. err}} = V[\text{pred. err}] = \underbrace{V[\hat{y}(x)]}_{\sigma^2_{\hat{y}}} + \underbrace{V[y^*]}_{\sigma^2_\epsilon}$$

$$\sigma^2_{\text{pred. err}} = \sigma^2_{\hat{y}} + \sigma^2_\epsilon$$

Think! The variance of all y values at a given x is σ^2_ϵ.

Approximate the σ's with sample estimates:

$$\therefore \; s^2_{\text{pred. err}} = s^2_{\hat{y}} + s^2_e$$

above

$SSE / (n-2)$ k+1

$$z = \frac{\boxed{\hat{y}(x) - y^*}}{\sigma_{\text{pred. err}}} \sim N(0,1)$$

= prediction error

$$t = \frac{\hat{y}(x) - y^*}{s_{\text{pred. err}}} \sim t\text{-distr. } df = n - \underset{k+1}{2}$$

$\therefore$ P.I. for a single y: $\hat{y} \pm t^* s_{\text{pred. err}} = \hat{y} \pm t^* \sqrt{s^2_{\hat{y}} + s^2_e}$

Compare with C.I for y (the cond'l mean): $\hat{y} \pm t^* s_{\hat{y}}$

Q which one is bigger? P.I. makes sense?

Don't forget what these intervals mean:

⇒ 2 interpretations for C.I:

1) We are 95% confident that the true (conditional) mean of y, given x, is in the observed C.I.

2) About 95% of random CIs will cover the true cond'l mean of y, given x.

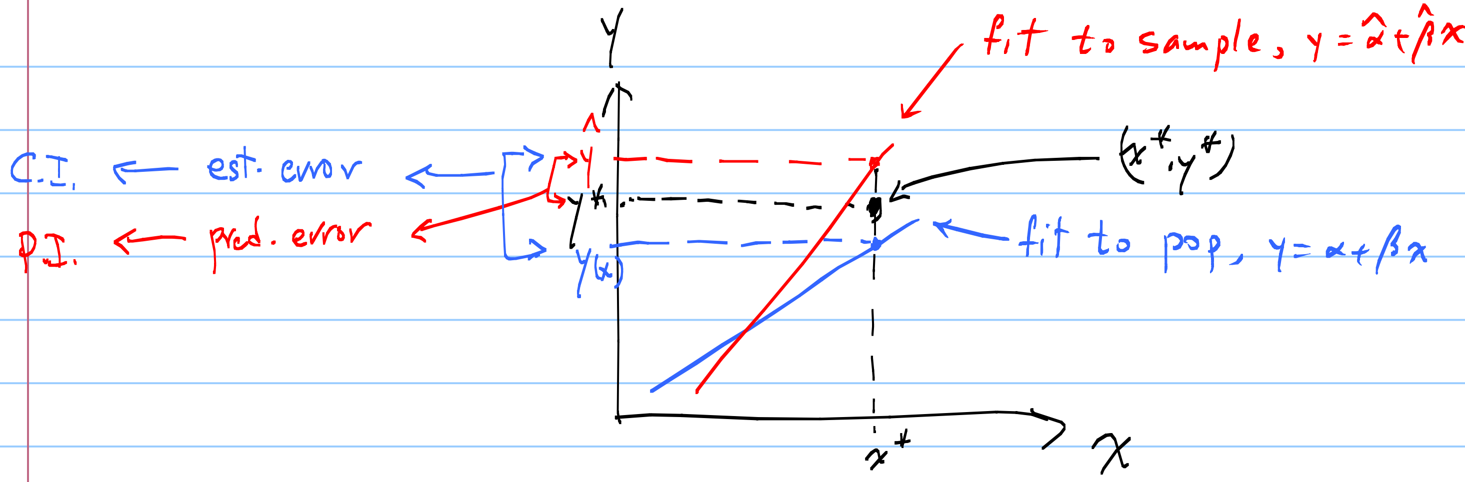
⇒ For P.I. the most straightforward interpretation is

1) About 95% of random PIs will cover a single y, at a given x.

1') After we are more comfortable with interpretations we will allow ourselves to also say things like "plausible y values, at a given x, are in the observed PI, at the 95% prediction level."

(See example, below)

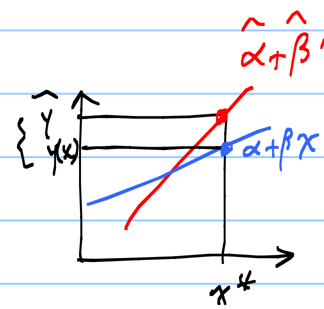
CI, PI on top of each other



CI, PI side-by-side

$\boxed{\text{est. error}}$

$= \hat{y} - y(x)$



$$\sigma^2_{est.err} = \sigma^2_{\hat{y}} + \sigma^2_{y(x)}$$

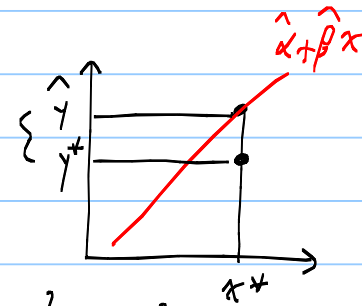
Recall that $\sigma^2_{y(x)}$ means the variance of $y(x)$ under resampling. But $y(x)$ is the fit to the pop. So, $\sigma^2_{y(x)} = 0$.

$$\therefore \sigma^2_{est.err.} = \sigma^2_{\hat{y}}$$

$$\therefore s^2_{est.err.} = s^2_{\hat{y}}$$

$\boxed{\text{pred. err.}}$

$= \hat{y} - y^*$



$$\sigma^2_{pred.err} = \sigma^2_{\hat{y}} + \sigma^2_{y^*}$$

Again, $\sigma^2_{y^*}$ means the var. of y^* under resampling. But y^* is the y for a given x, and so, its variance under resampling is just σ^2_ϵ.

$$\therefore \sigma^2_{pred.err} = \sigma^2_{\hat{y}} + \sigma^2_\epsilon$$

$$\therefore s^2_{pred.err} = s^2_{\hat{y}} + s^2_e$$

C.I.: $\hat{y} \pm t^* \underbrace{s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}_{s_{\hat{y}}}$

P.I.: $\hat{y} \pm t^* \sqrt{s^2_{\hat{y}} + s^2_e}$

Example 11.20 (re-worded and revised, for clarity)

x = temperature y = oxygen diffusivity.

$n = 9$, $\sum x = 12.6$ $\sum y = 27.68$

$\sum x^2 = 18.24$ $\sum y^2 = 93.3448$ $\sum xy = 40.968$

predict oxyg. diffusivity when temperature is 1.5 (in 1000 F°) in a way that conveys info about reliability & precision.

11.5

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 18.24 - 9\left(\frac{12.6}{9}\right)^2 = 0.6$$

$$SST = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = 93.3448 - 9\left(\frac{27.68}{9}\right)^2 = 8.213$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} = 40.968 - 9\left(\frac{12.6}{9}\right)\left(\frac{27.68}{9}\right) = 2.216$$

$$\hat{y} = -2.095 + 3.6933\,x$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{SST - \hat{\beta}(S_{xy})}{n-2}} = \sqrt{\frac{8.2134 - 3.6933(2.216)}{9-2}} = 0.0644$$

When temp = 1.5 in (1000 F), what is the prediction for the mean of diffusivity at that temp.?

A point estimate for that mean is given by the OLS line:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -2.095 + 3.6933\,x$$

i.e. $\hat{y} = -2.095 + 3.6933(1.5) = 3.445$.

A C.I. for the true mean at that temp. gives an interval estimate of that mean:

$$\hat{y} \pm t^* s_{est.\,err} = \hat{y} \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

$$= 3.445 \pm \underset{\substack{\uparrow \\ df = 9-2}}{2.365}\ \underbrace{(0.0644)\sqrt{\frac{1}{9} + \frac{\left(1.5 - \frac{12.6}{9}\right)^2}{0.6}}}_{0.02302\ =\ s_{est.\,err}\ =\ s_{\hat{y}}}$$

∴ C.I. for mean, y(x), at temp = 1.5 : 3.445 ± 0.0544

(3.39, 3.50)

Interpretation?

---

predict oxyg. diffusivity (for a single case) when temperature is 1.5 K°F in a way that conveys info about reliability & precision.

Interval estimate.

This is asking for a prediction interval:

$$\hat{y} \pm t^* \sqrt{s_{\hat{y}}^2 + s_e^2}$$

$$= 3.445 \pm 2.365 \sqrt{(0.02302)^2 + (0.0644)^2}$$

$$= 3.445 \pm 0.1617 = (3.28,\ 3.61)$$

1) 95% of such PI's will cover single values of y, at x = 1.5.

2) At 95% prediction level, plausible values for a single y value, at x = 1.5, are between 3.28 and 3.61.

hw-lect 25-1 Revised 11.21

Mist (airborne droplets or aerosols) is generated when metal-removing fluids are used in machining operations to cool and lubricate the tool and work-piece. Mist generation is a concern to OSHA, which has recently lowered substantially the workplace standard. The article "Variables Affecting Mist Generation from Metal Removal Fluids" (Lubrication Engr., 2002: 10-17) gave the accompanying data on x = fluid flow velocity for a 5% soluble oil (cm/sec) and y = the extent of mist droplets having diameters smaller than some value:

x:	89	177	189	354	362	442	965
y:	.40	.60	.48	.66	.61	.69	.99

a. Make a scatterplot of the data. By R.

b. What is the point estimate of the beta coefficient? (By R.) Interpret it.

c. What is s_e? (By R) Interpret it.

d. Estimate the true average change in mist associated with a 1 cm/sec increase in velocity, and do so in a way that conveys information about precision and reliability.
Hint: This question is asking for a CI for beta. Compute it AND interpret it.
By hand; i.e. you must use the basic formulas for the CI. E.g. for beta:
beta_hat +- t* s_e/sqrt(S_xx) ,
but you may use R to compute the various terms in the formula.
Use 95% confidence level.

e. Suppose the fluid velocity is 250 cm/sec. Compute an interval estimate of the corresponding mean y value. Use 95% confidence level. Interpret the resulting interval.
By hand, as in part d.

f. Suppose the fluid velocity for a specific fluid is 250 cm/sec. Predict the y for that specific fluid in a way that conveys information about precision and reliability. Use 95% prediction level. Interpret the resulting interval. By hand, as in part d.

hw_lect25-2

Consider the defining formulas for C.I and P.I:

C.I. $\hat{y}(x) \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$

P.I $\hat{y}(x) \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$ where $\hat{y}(x) = \hat{\alpha} + \hat{\beta}x$

a) As n becomes large (but <u>not</u> quite ∞) what does each of the following approach? For example $\hat{\alpha} \rightarrow \alpha$.

$\hat{\alpha} \rightarrow \alpha$

$\hat{\beta} \rightarrow$

$\hat{y}(x) \rightarrow$

$t^* \rightarrow$

$s_e \rightarrow$

$\bar{x} \rightarrow$

$S_{xx} \rightarrow$

b) As $n \rightarrow \infty$, what does CI converge to?

c) As $n \rightarrow \infty$, what does PI converge to?