What you have learned in this class.

Dealing with ambiguity
Random variable
histograms
Comparative boxplots
quantiles
distributions
probability (e.g. from Poisson)
sample mean and variance
distr mean and variance
qqplots

scatterplots
correlation
regression (multiple, polynomial, ...)
ANOVA (R^2, $s_e \sim$ RMSE)
overfitting, collinearity, interaction

sampling distribution
1-sample Confidence Interval for ...
2-sample CI for ...
t-distribution

Hypothesis testing with p-values
1-sample, 2-sample, paired, ... tests
tests for means and proportions
chi-squared test of multiple proportions in 1 pop
chi-squared test of indepedence of two categorical variables ← skipped.

1-way ANOVA F-test for the equality of multiple pop means.

t-test of regression coefficients
Confidence and Prediction Intervals
F-test of model utility

Model selection via bootstrapping (and cross-validation)
Neural networks (as a regression model).) skipped.

Lecture 26 (ch. 11)

problem 11.11

Last time, we did inference for 1 β (and α) t-interval & t-test
and for many β's p-value " & F-test.

Q: What about the true (pop.) prediction itself? $y(x)=\alpha+\beta_1 x_1+\beta_2 x_2+\dots$

Unfortunately, the (sample) prediction $\hat{y}(x)$ has 2 diff. interpretations:

- (point estimate of) the true/pop. conditional mean of y, given x, i.e. $y(x)$.
- (point) prediction of a single y, given x, call it y^*.

Note: The prediction $\hat{y}(x)$ is the same in both cases.

But the interpretation is different $\Rightarrow$ different intervals & tests.

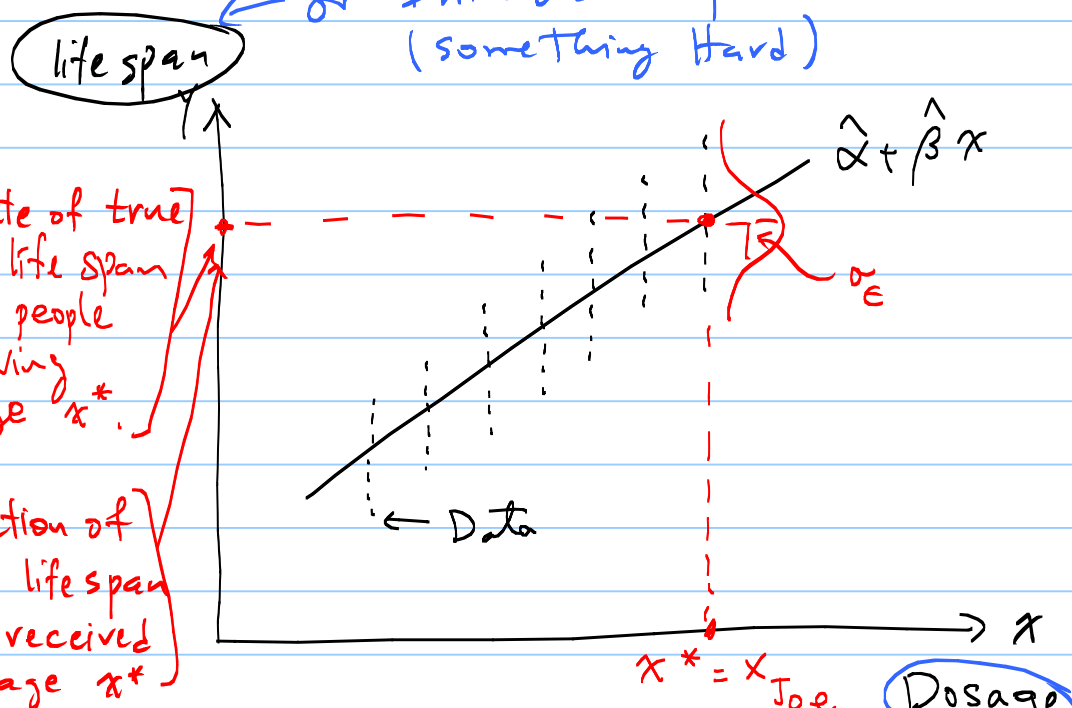
The two intervals/tests answer 2 diff. questions:

→ 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:

← or Intracranial pressure. (something Hard)



Estimate of true mean life span of all people receiving dosage x^*.

Conf. Int. for true mean, $y(x)$

Today's Topics

Prediction of Joe's life span who received dosage x^*

Prediction Interval (PI) for Joe's true life span, y^*

The "level" of this interval is called prediction level (e.g. 95%)

or Blood pressure (something Easy).

(CI for $y(x)$)

1) For C.I. of the population mean, $y(x)$, given x: we need the sampling distr. of $\hat{y}(x)$.

Analogous to: For C.I of μ_x, we need sampl. dist. of $\bar{x}$. $\bar{x} \sim N(\mu, \sigma^2/n)$

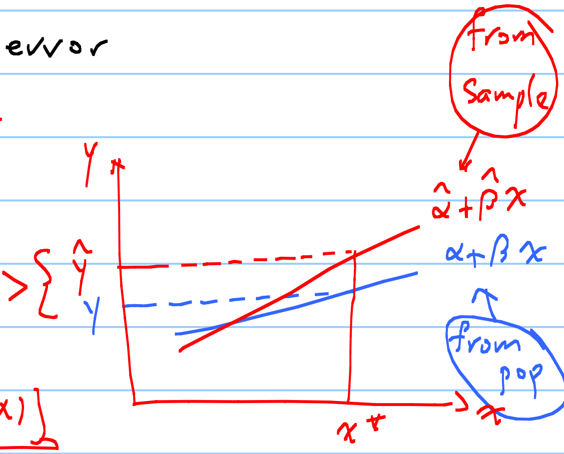
$\boxed{\text{Theorem}}$ The sampl. distr. of $\hat{y}(x) = \hat{\alpha} + \hat{\beta}x$ is Normal with params:
$\mu = y(x) = \alpha + \beta x$, $\sigma^2 = \sigma^2_{\text{estimation error}}$

where

sample fit — pop. fit

$$\boxed{\text{estimation error}} = \hat{y}(x) - y(x)$$

$$\hat{y}(x) = \hat{\alpha} + \hat{\beta}x, \quad y(x) = \alpha + \beta x$$



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

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

because $y(x)$ = pop fit.

Approximate/Estimate the σ's with their sample analog:

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

where

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

It follows that

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

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

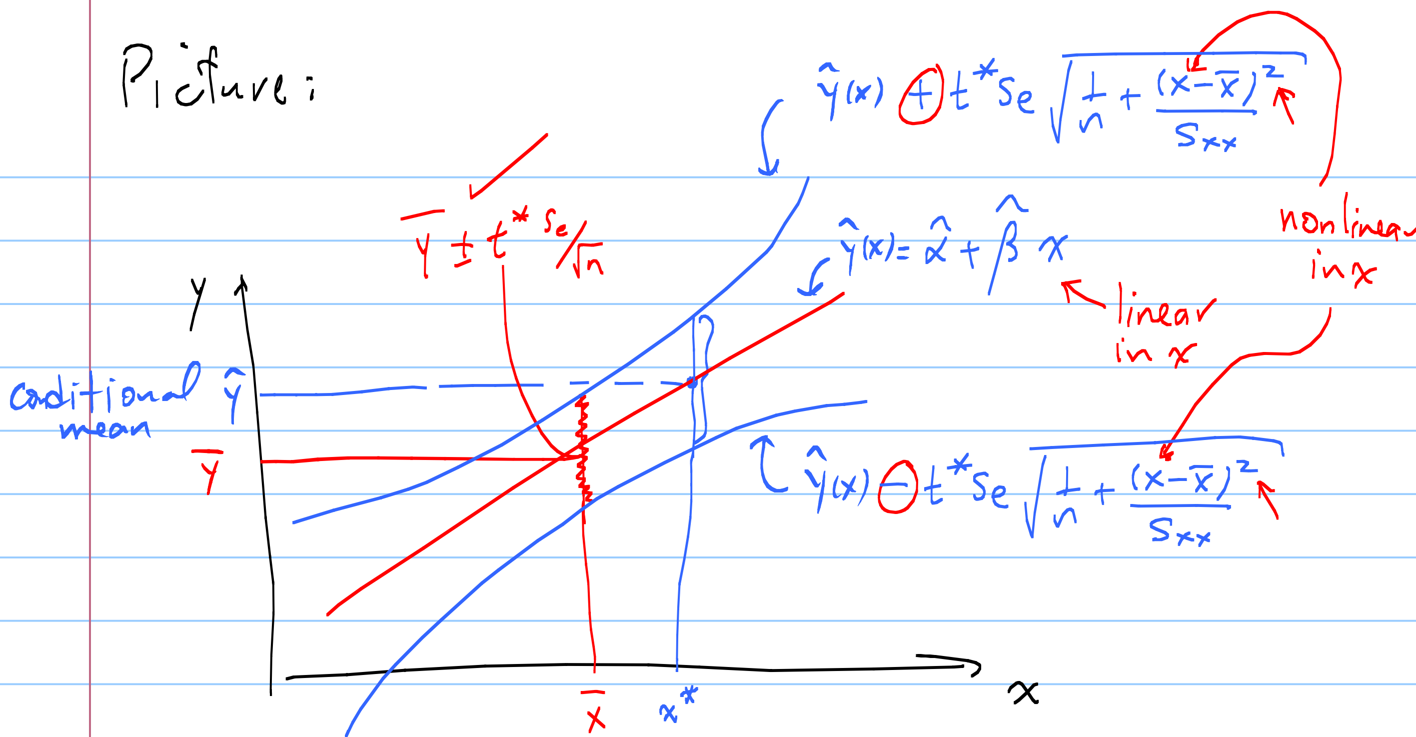
($2 = k+1$)

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

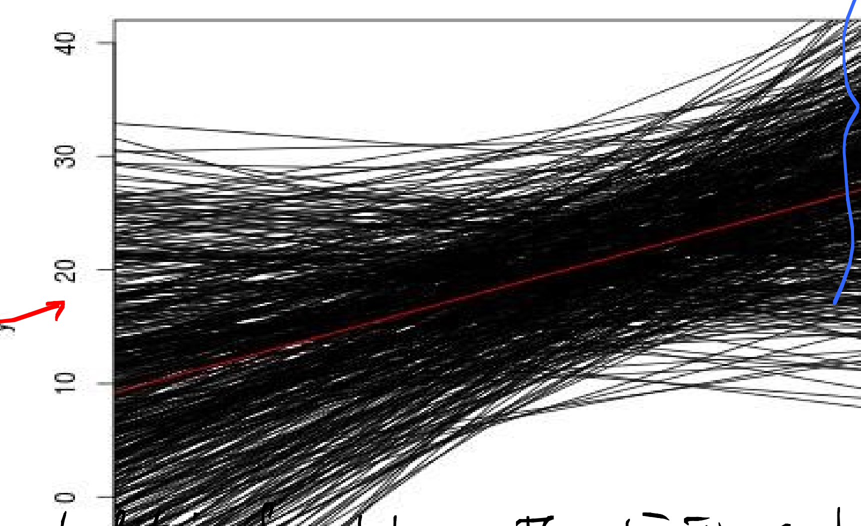
Table IV, $df = n-2$ ($k+1$)

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

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.
Recall, in "our" regression, $\bar{x}$ = fixed (ie. x's have no uncertainty).
But $\bar{y}$ is a r.v. (ie. every sample will have a different $\bar{y}$. So, that will shift the sample fit up/down across trials. BUT the $\hat{\beta}$ will also change across trials. So, now imagine what all the fits will look like:



These are all the sample fits (in blue) from last lect's big fig.

The red line is the pop fit

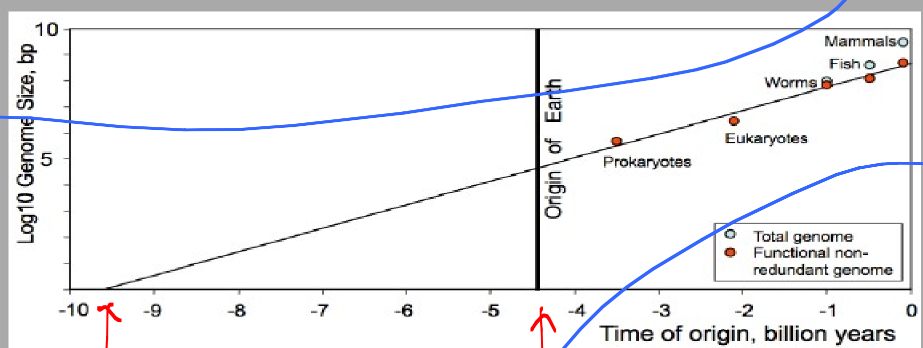
Each sample fit is forced to go thru $(\bar{x},\bar{y})$, and the slope changes.
Do you see how a change in slope sweeps a wider region at larger x's?
That's the reason why the CI is wider for larger x's.

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.

Confidence 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) Confidence intervals must be considered.

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

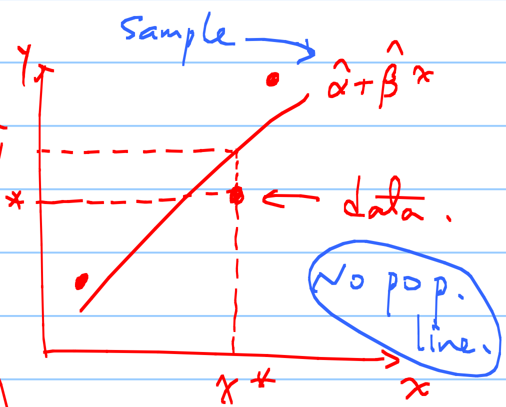
Suppose y^* is Joe's y-value, corresponding to his x-value (x^*), but y^* is not observed. So, y^* is just a random y, at a given x.

The * simply emphasizes That it has not been observed, or more importantly, That it has not been used in estimating $\hat{\alpha}, \hat{\beta}$.

Data That are not used in The development ("training") of The model are often called "out of sample."

$\boxed{\text{Theorem}}$: Denote prediction error $= (\hat{y}(x) - y^*)$

Then pred. err $\sim N(\mu = 0, \sigma^2 = \sigma^2_{\text{pred err}})$



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

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

As explained above, y^* is just a random y at $x = x^*$.

Approximate The σ's with sample estimates:

above

$$\therefore \; s^2_{\text{pred. err}} = s^2_{\hat{y}} + s^2_e \leftarrow SSE/(n-2)$$

$k+1$

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

prediction error

$k+1$

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

$\therefore$ P.I. for 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 wider? P.I. Makes sense?

Don't forget what these intervals mean:

⇒ 2 interpretations for C.I :

1) About 95% of random CI's cover the true (Conditional) mean, $y(x)$, at given x.

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

⇒ For P.I. the most straightforward interpretation is

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

2) We are 95% "confident" that y^*, at a given x, is in observed PI.

(Technically, we should not use the word "confident" because that word is reserved for pop. params (μ, π, $y(x)$, ...). So, people often say something like "plausible y values, at a given x, are in the observed PI, at 95% prediction level."

(See example, below)

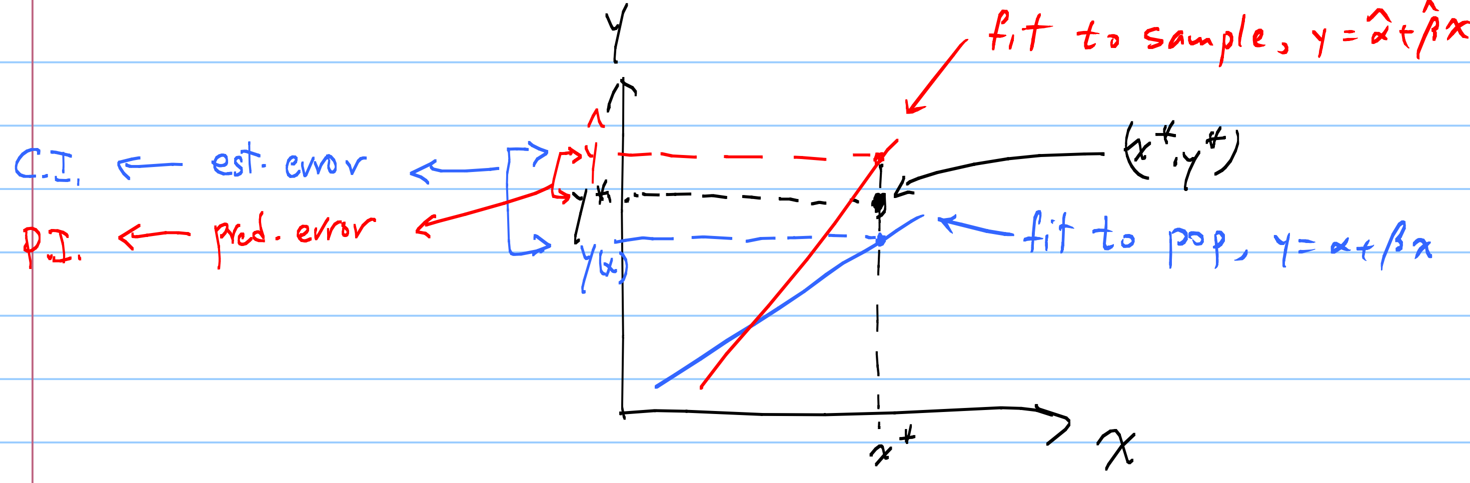
Note: The $\underline{\underline{3}}$ "errors" in regression:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

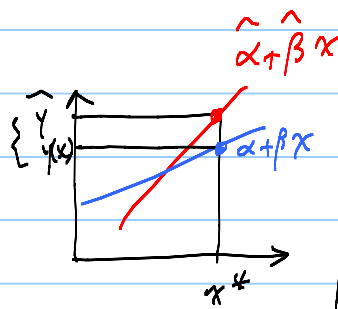
$$\begin{cases} y - y(x) = \epsilon = \text{observation error} \Rightarrow \sigma_\epsilon \\ \hat{y}(x) - y(x) = \text{est. err} \Rightarrow \sigma_{\text{est. err}} \\ \hat{y}(x) - y^* = \text{pred. err.} \Rightarrow \sigma_{\text{pred. err}} \end{cases}$$

[This is just a random y, at $x = x^*$.
You can even denote it as just y (without the *) as long as you remember that it's not observed.

CI, PI on top of each other



CI, PI side-by-side



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

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

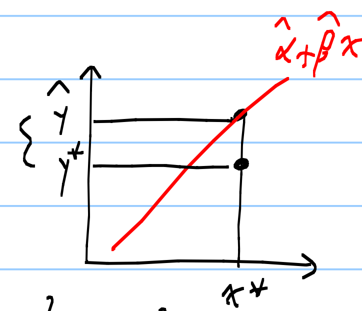
$$\sigma^2_{est.err} = \sigma^2_{\hat{y}} + \sigma^2_{y(x)}$$

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 $\qquad$ y = oxygen diffusivity.

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

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

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

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

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

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

∴ (obs) CI for $y(x)$, ie. mean of y at temp = 1.5 : 3.445 ± 0.0544

(3.39, 3.50)

Interpretations: 1) With 95% confidence, the true mean of y, at $x = 1.5$, is between 3.39 and 3.50.

2) There is 95% prob that a random C.I will cover the true mean of y at $x = 1.5$.

---

for a single case

predict oxyg. diffusivity 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_lect26-1

In a simple regression problem, we have $n = 16$, $\bar{x} = 10$, $S_x = \frac{1}{\sqrt{8}}$, and $S_e = 4$. At $x = 11$, what is the value of T_0 such that

prob(prediction error > T_0) = 0.01 ?

(Hint: how do we standardize prediction error?)

hw-lect24-2

Here is the regression version of a problem we have seen many times before.

a) Show that, at a given x, the prob that a random Y would fall into the obs CI for $Y(x)$ is

$$pr\left(t_{obs} - t^* \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} < t < t_{obs} + t^* \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

where $t_{obs} = \dfrac{\hat{Y}_{obs}(x) - Y(x)}{S_e}$

(Hint: How do you standardize observation error?)

b) Show that, at a given x, the prob that a random $\hat{Y}(x)$ would fall into the obs CI for $Y(x)$ is $pr(t_{obs} - t^* < t < t_{obs} + t^*)$

where $t_{obs} = \dfrac{\hat{Y}_{obs}(x) - Y(x)}{S_{est.\,err.}}$.

(Hint: how do you standardize estimation error?)

hw-optional

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

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

where $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$

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

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

$\hat{\alpha} \to \alpha$ As n increases, $\hat{\alpha}$ approaches the population y-intercept α.

$\hat{\beta} \to$

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

$t^* \to$

$s_e \to$

$\bar{x} \to$

$S_{xx} = (n-1) s_x^2 \to$

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

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

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

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

hw_optional (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.