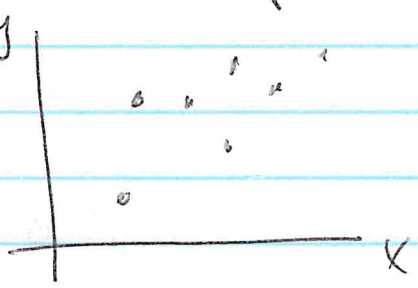
Ch 11 Regression + Least-Squares

Intro: best-fit lines (two-ways)

Consider a sample (data) $S = \{(x_i, y_i)\}$



with some underlying joint pdf $f(x,y)$.

Assumptions: ① y is dependent upon X (not indep).
② in general, $Y = mX + b$ is the relationship btwn X and Y.

A) a probabalistic construction.

Given all possible lines $y = mx + b$ that could describe the data, the most likely line will intersect the point (μ_x, μ_y).
(This is the fundamental idea of "regression analysis".)

That is $y - \mu_y = m(x - \mu_x)$

To find the "best" line, we choose to minimize the vertical distance (error in the y-direction) btwn the points and the line.

for a point (X_k, Y_k) in S, the vertical distance is

$$\text{dist} = |\underbrace{Y_k}_{\text{actual}} - \underbrace{(b(X_k - \mu_X) + \mu_Y)}_{\text{predicted by model}}|$$

Minimizing will be a calculus problem. Old trick from Calc I, min dist^2 instead,

$$\text{dist}^2 = (Y_k - (b(X_k - \mu_X) + \mu_Y))^2$$
$$= ((Y_k - \mu_Y) - b(X_k - \mu_X))^2$$

As X, Y distributed via a PDF, best way to compute this problem is by minimizing the expectation of the parameter b.

Define $K(b) = E(\text{dist}^2(b))$
$= E[((Y_k - \mu_Y) - b(X_k - \mu_X))^2]$

the problem of finding the "b" that minimizes expectation is called a <u>least-squares</u> problem.

$$K(b) = E[(Y_k - \mu_Y)^2 - 2b(Y_k - \mu_Y)(X_k - \mu_X) + b^2(X_k - \mu_X)^2]$$
$$= E[(Y_k - \mu_Y)^2] - 2bE[(X_k - \mu_X)(Y_k - \mu_Y)] + b^2 E[(X_k - \mu_X)^2]$$
$$= S_Y^2 - 2b\,\text{Cov}(X, Y) + b^2 S_X^2.$$

Minimize $\Rightarrow \frac{d}{db}$.

$K'(b) = -2Cov(X,Y) + 2b S_x^2$

$K'(b) = 0$ when $b = \frac{Cov(X,Y)}{S_x^2}$.

Note that this is a minimum as $K''(b) = 2S_x^2 > 0$.

In AP Stats, the slope is moved to a "prettier" form:

$$b = \frac{Cov(X,Y)}{S_x S_y} \cdot \frac{S_y}{S_x} = \rho \frac{S_y}{S_x}$$

where ρ is the correlation coef $\frac{Cov(X,Y)}{S_x S_y}$

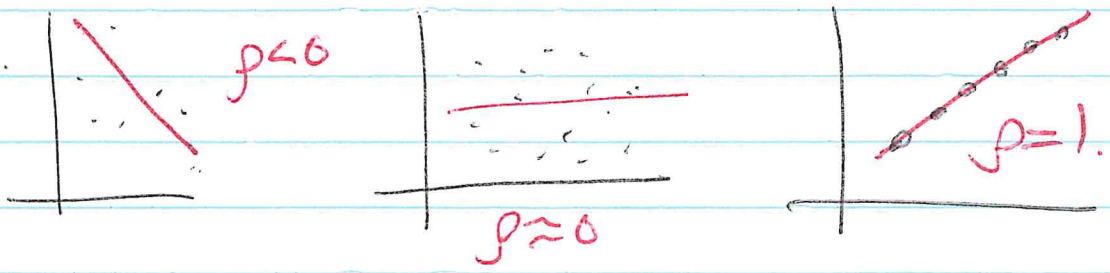
From last semester,

FACT #1: The sign of the slope depends entirely upon ρ, whose sign is dependent upon $Cov(X,Y)$

FACT #2: $-1 \le \rho \le 1$.

We proved this last semester in HW. Here, much easier to prove simply by evaluating $K(b)$ at our critical point.

FACT #3: Closer $|\rho|$ is to 1, the better the line $y = mx + b$ "fits" the data.



The AP Stats defn of the best fit line

is $\hat{y} = \hat{b}_0 + \hat{b}_1 x$

where $\hat{b}_1 = \rho \frac{S_y}{S_x}$ and $\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$

B) the linear algebra construction.

We want $y = mx + b$.
Using $S = \{(x_i, y_i)\}$ yields the over determined and always inconsistent system

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ &\vdots \\ y_n &= mx_n + b \end{aligned}$$

as a vector eqn ⟹

$$\vec{y} = m\vec{x} + b\vec{1}$$

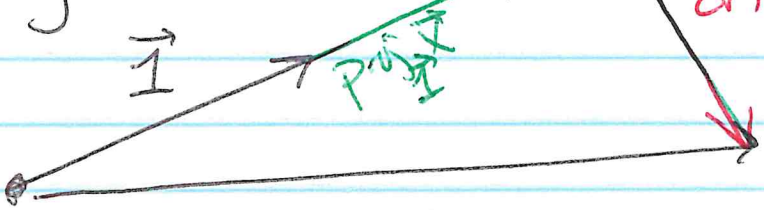
where $\vec{y} = \langle y_1, \ldots, y_n \rangle$, $\vec{x} = \langle x_1, \ldots, x_n \rangle$ and $\vec{1} = \langle 1, \ldots, 1 \rangle$.

Consider the plane $P = \text{span}\{\vec{x}, \vec{1}\}$.

Inconsistent $\rightarrow$ $\vec{y} \notin P$
But the vector in P that is closest to $\vec{y}$ is $\text{proj}_P \vec{y}$. (closest under Euclidean distance $\text{dist}^2 = \sum (y_i - v_i)^2$ ... i.e. least squares!)

Easiest way to project onto P is to have an orthogonal basis of P.

$\text{orth}_{\vec{1}} \vec{x} = \vec{x} - \text{proj}_{\vec{1}} \vec{x}$.



$$P = \text{span}\{\vec{1}, \text{orth}_{\vec{1}} \vec{x}\}$$
$$= \text{span}\{\vec{1}, \vec{v}\},$$
where $\vec{v} = \vec{x} - \frac{\vec{x}\cdot\vec{1}}{\vec{1}\cdot\vec{1}}\vec{1} = \vec{x} - \frac{\vec{x}\cdot\vec{1}}{n}\vec{1}$.

Then $\text{proj}_P \vec{y} = \frac{\vec{y}\cdot\vec{1}}{\vec{1}\cdot\vec{1}}\vec{1} + \frac{\vec{y}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\vec{v}$

$$= \frac{\vec{y}\cdot\vec{1}}{n}\vec{1} + \frac{\vec{y}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\vec{x} - \frac{\vec{y}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\left(\frac{\vec{x}\cdot\vec{1}}{n}\right)\vec{1}$$

$$= \underbrace{\frac{\vec{y}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}}_{\text{this is slope } m}\vec{x} + \underbrace{\left(\frac{\vec{y}\cdot\vec{1}}{n} - \frac{\vec{y}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\cdot\frac{\vec{x}\cdot\vec{1}}{n}\right)}_{\text{intercept } b.}\vec{1}$$

$$\vec{V}\cdot\vec{V} = \left(\vec{x} - \frac{\vec{x}\cdot\vec{1}}{n}\vec{1}\right)\cdot\left(\vec{x} - \frac{\vec{x}\cdot\vec{1}}{n}\vec{1}\right)$$

$$= \vec{x}\cdot\vec{x} - 2\frac{(\vec{x}\cdot\vec{1})^2}{n} + \left(\frac{\vec{x}\cdot\vec{1}}{n}\right)^2 \vec{1}\cdot\vec{1}$$

$$= \vec{x}\cdot\vec{x} - \frac{2(\vec{x}\cdot\vec{1})^2}{n} + \frac{(\vec{x}\cdot\vec{1})^2}{n}$$

$$= \frac{n(\vec{x}\cdot\vec{x}) - (\vec{x}\cdot\vec{1})^2}{n}$$

$$\vec{V}\cdot\vec{Y} = \vec{x}\cdot\vec{y} - \frac{(\vec{x}\cdot\vec{1})(\vec{y}\cdot\vec{1})}{n}$$

$$= \frac{n(\vec{x}\cdot\vec{y}) - (\vec{x}\cdot\vec{1})(\vec{y}\cdot\vec{1})}{n}$$

Thm $\quad m = \dfrac{\vec{V}\cdot\vec{Y}}{\vec{V}\cdot\vec{V}} = \dfrac{n(\vec{x}\cdot\vec{y}) - (\vec{x}\cdot\vec{1})(\vec{y}\cdot\vec{1})}{n(\vec{x}\cdot\vec{x}) - (\vec{x}\cdot\vec{1})^2}$

Also note $\quad b = \dfrac{\vec{y}\cdot\vec{1}}{n} - m\cdot\dfrac{\vec{x}\cdot\vec{1}}{n}$

$$= \bar{y} - m\bar{x}$$

Claim: These are the same set of formulas.

$$m = \frac{Cov(X,Y)}{S_X^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Useful formula: $\sum (x_i - \bar{x})(y_i - \bar{y})$

$$= \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$
$$= \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{x}\bar{y}$$
$$= \sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$
$$= \sum x_i y_i - n\bar{x}\bar{y}.$$

So $$m = \frac{Cov(X,Y)}{S_X^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$= \frac{\vec{x} \cdot \vec{y} - n \frac{\vec{x} \cdot \vec{1}}{n} \cdot \frac{\vec{y} \cdot \vec{1}}{n}}{\vec{x} \cdot \vec{x} - n \left(\frac{\vec{x} \cdot \vec{1}}{n}\right)^2}$$

$$= \frac{n\,\vec{x} \cdot \vec{y} - (\vec{x} \cdot \vec{1})(\vec{y} \cdot \vec{1})}{n\,\vec{x} \cdot \vec{x} - (\vec{x} \cdot \vec{1})^2}$$

save this!

§ 11.4 – the coefficients as estimators.

We have shown the best-fit line to be

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{Y})$

$S_{xx} = \sum (x_i - \bar{x})^2$.

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

But we recognize $\hat{\beta}_0$ and $\hat{\beta}_1$ as stats dependent upon $\bar{x}$, S_x^2 and $\bar{Y}$. But what exactly are they estimating (Ch 9)

In theory $X \times Y$ distributed via $f(x,y)$ and there is some "best" linear relationship over the entire probability space.

$$Y = \beta_0 + \beta_1 X$$

That is, the @ x, $E[Y] = \beta_0 + \beta_1 x$.

A modern approach to the distribution of Y @ x is to add an error parameter ϵ

That is $Y = \underbrace{\beta_0 + \beta_1 X}_{\text{deterministic component of } Y} + \epsilon$ the random component.

Still want $E[Y] = \beta_0 + \beta_1 x$, hence $E[\epsilon] = 0$ is required.
We make the additional assumption that the variance of ϵ is independent of x.
That is $V(Y) = V(\epsilon) = \sigma^2$
i.e. y depends on x, but the spread of Y does not

Pop: $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators of β_0, β_1 where $Y = \beta_0 + \beta_1 X + \epsilon$.

Reason: $E(\hat{\beta}_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) = E\left(\frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}}\right)$

$= E\left(\frac{\sum (x_i - \bar{x}) Y_i - \bar{Y} \sum (x_i - \bar{x})}{S_{xx}}\right)$ $\to = 0$ deft.

$= \frac{\sum (x_i - \bar{x}) E(Y_i)}{S_{xx}}$

$= \frac{\sum (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)}{S_{xx}}$

$= \frac{\beta_0 \sum (x_i - \bar{x})}{S_{xx}} + \frac{\beta_1 \sum (x_i - \bar{x}) x_i}{S_{xx}}$ $\to 0$

* Lastly $\sum (x_i - \bar{x}) x_i = \sum x_i^2 - n\bar{x}^2 = S_{xx}$.

So $E(\hat{\beta}_1) = \beta_1$. unbiased

for $E(\hat{\beta}_0) = E[\bar{Y} - \hat{\beta}_1 \bar{x}]$

$= E[\bar{Y}] - \bar{x} E[\hat{\beta}_1]$

$= E[\bar{Y}] - \beta_1 \bar{x}$.

But $\bar{Y} = \frac{1}{n}\sum Y_i = \frac{1}{n}\sum (\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$

and $E[\bar{Y}] = \beta_0 + \beta_1 \bar{x}$.

and $E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$ ∎.

Cor: $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, $V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$.

Reason: $V(\hat{\beta}_1) = V\left[\frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}}\right] = \frac{1}{S_{xx}^2} V\left[\sum (x_i - \bar{x}) Y_i\right)$

$= \frac{\sum (x_i - \bar{x})^2 V(Y_i)}{S_{xx}^2} = \frac{\sigma^2 \sum (x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$.

$V(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1 \bar{x})$

note: in our ϵ set-up, $\bar{Y}$ and $\hat{\beta}_1$ both functions of Y_i.

$= V(\bar{Y}) + V(\hat{\beta}_1 \bar{x}) - 2\,Cov(\bar{Y}, \hat{\beta}_1 \bar{x})$

$= V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}\,Cov(\bar{Y}, \hat{\beta}_1)$

i) $V(\bar{Y}) = V(\bar{\epsilon}) = \frac{1}{n} V(\epsilon) = \frac{\sigma^2}{n}$

ii) $Cov(\bar{Y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum Y_i, \sum \frac{(x_i - \bar{x})}{S_{xx}} Y_i\right)$

$$= \sum_i^n \frac{(x_i-\bar{x})}{n S_{xx}} V(Y_i) + \sum\sum_{i \neq j} \frac{(x_i-\bar{x})}{n S_{xx}} \underbrace{Cov(Y_i, Y_j)}_{=0 \text{ by ind}}$$

$$= \frac{\sigma^2}{n S_{xx}} \sum^n (x_i - \bar{x}) \quad \rightarrow = 0$$

$$= 0.$$

$$\therefore V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} + 0$$

$$= \frac{\sigma^2 (S_{xx} + n\bar{x}^2)}{n}$$

$$= \frac{\sigma^2 \sum x_i^2}{n S_{xx}} \quad \text{since } S_{xx} = \sum x_i^2 - n\bar{x}^2. \; \square$$

Cor: $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{xx}}$

Reason: Exercise

Note: $\hat{\beta}_0, \hat{\beta}_1$ are dependent unless $\bar{x} = 0$.

~~Disc:~~

topic disc: estimating σ^2.

We have been working w/ assumption $V(Y) = V(\epsilon) = \sigma^2$, but this is also usually unknown.

In the past, we used $V(Y) = \frac{1}{n-1}\sum (y_i - \bar{Y})^2$.

However, in this setting, our estimator of $E[Y_i]$ is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and not $\bar{Y}$.

We want to define S^2 via $SSE = \sum_1^n (y_i - \hat{y}_i)^2$

$$S^2 := K \sum_1^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

where K is chosen to make S^2 an unbiased estimator of σ^2.

(I don't want to do this again pg 580-581)
As we did in Ch 9, we can show

$$E[SSE] = (n-2)\sigma^2.$$

So we will use

$$S^2 := \frac{1}{n-2} SSE = \frac{1}{n-2} \sum_1^n (y_i - \hat{y}_i)^2$$

Prop: $E[S^2] = \sigma^2$.

Cor: (The computation cor.) $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$
where $S_{yy} = \sum_1^n (y_i - \bar{y})^2$.

ex: (11.16)

potency of antibiotic after storage (y)	38, 43, 29	32, 26, 33	19, 27, 23	14, 19, 21
temp stored at (x)°F	30°	50°	70°	90°

note $n = 12$

$\bar{y} = 27$, $\bar{x} = 60$

$S_{xy} = -1900 = \sum (x_i - \bar{x})(y_i - \bar{y})$

$S_{xx} = 6000 = \sum (x_i - \bar{x})^2$

$S_{yy} = 792 = \sum (y_i - \bar{y})^2$.

Then $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-19}{60}$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 27 + \frac{19}{60} \cdot 60 = 46$

So the best-fit line is $\hat{y} = 46 - \frac{19}{60} x$.

For $S^2 = \frac{1}{n-2} SSE = \frac{1}{n-2} (S_{yy} - \hat{\beta}_1 S_{xy})$

$= \frac{1}{10} \left[792 - \left(\frac{-19}{60} \right) (-1900) \right]$

$= 571/30 \approx 19.03$

disc: do we know how S^2 is distributed?

This depends on our assumptions on how ϵ is distributed.

We have $Y = \beta_0 + \beta_1 X + \epsilon$ w/ $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$, and these are independent of X.

Note the distns of the point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are σ^2 dependent (which we can estimate w/ S^2)!

However, it is a common assumption that things are normally distributed.

If we also assume $\epsilon \sim N(0, \sigma^2)$

then we can prove a Fisher's Thm - like result for S^2.

Namely,

Thm: $\frac{(n-2)S^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2(n-2)$

and S^2 is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Moreover, $\hat{\beta}_0$ and $\hat{\beta}_1$ are also normally distributed.

Pf: beyond our scope.

§ 11.5 Inferences concerning the point estimators.

Under the assumption $\epsilon \sim N(0, \sigma^2)$, we have $\hat{\beta}_0, \hat{\beta}_1$ are Normal $N(\beta_i, V(\beta_i))$

Thus, we can do confidence intervals for the true β_0, β_1 ($Y = \beta_0 + \beta_1 x + \epsilon$) and hypothesis tests.

Use the same Z / T Rules as before.
- If $n \leq 30$, use t-dist'n
- $n > 30$, use Z.

ex: Given $z_{\alpha/2}$, $\hat{\beta}_i \pm z_{\alpha/2} \sqrt{V(\hat{\beta}_i)}$ $n > 30$ is an α-level confidence interval for the true β_i.

Ex: Find a 90% C.I. for the slope $\hat{\beta}_1$ of the putrescency example.

We have $\hat{\beta}_1 = \frac{-19}{60} \approx -0.31667$

$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \approx \frac{S^2}{S_{xx}} \approx \frac{19.63}{60000} = 0.00317\ldots$

$\Rightarrow S = 0.0563\ldots$

* Because we are using S^2, we need to use S^2 df. *

$n = 12$

$S^2 \sim \chi^2(12-2) = \chi^2(10)$

Need $t_{0.\alpha}(10) = 1.812$.

And $\hat{\beta}_1 \pm t_{0.\alpha}(10) S$

is $-0.31667 \pm (0.0522)(1.812)$

-0.31667 ± 0.102856

or $(-0.4187, -0.2146)$

ex: (potency example again)

Let us assume that it is normally known that penicillin derivatives are considered "good" if when stored in a deep freezer (0°F) the potency is 50 (or better)

We wish to test if our drug is considered "good".

We have $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 46 - \frac{4}{60} x$.

Then $\hat{y}(0) = \hat{\beta}_0 = 46$

We want to test if the true $\beta_0 = 50$

Given our sample observation $\hat{\beta}_0 = 46$.

$H_0: \beta_0 = 50$
$H_a: \beta_0 < 50$.

Assn $S^2 \sim \chi^2(10)$ ... need to use t-test

p-value $p = P(\hat{\beta}_0 \leq 46 \mid \beta_0 = 50)$

We assume $\hat{\beta}_0 \sim N(\beta_0, V(\hat{\beta}_0))$

$$p = P\left(\frac{\hat{\beta}_0 - \beta_0}{\sqrt{V(\hat{\beta}_0)}} \leq \frac{46 - 50}{\sqrt{V(\hat{\beta}_0)}} \right)$$

didn't compute $V(\hat{\beta}_0)$ earlier.

Recall $V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$

Have $\sigma^2 \approx S^2 \approx 19.03$, $n = 10$, $S_{xx} = 600$

$\sum x_i^2 = 30^2 + 43^2 + \dots + 80^2 = 5740$ ~~16,400~~ $\cdot 3$
~~4900~~ ~~16,400~~ $= 49200$

So $V(\hat{\beta}_0) = \frac{19.03 \cdot 5740}{10 \cdot 600} = 2.274$ ~~13.003~~

$\sqrt{V(\hat{\beta}_0)} = 3.606$

Then
$$p = P\left(T(10) \leq \frac{46 - 50}{\sqrt{2.274}\ 3.606} = -2.482 \right) \quad -1.812$$

$= P(T \geq 2.482)$ ~~< 0.025~~ > 0.01

~~Small~~ p-value $\Rightarrow$ ~~can~~ reject H_0. Our drug is ~~not~~ "good"!

§ 11.6 : predictions via $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Our focus so far has been on the coefs $\hat{\beta}_i$ and their distribution.
Of course, the original goal was to understand y as a fcn of x

ex: (potency example again)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 46 - \frac{19}{60} x.$$

If we started the antibiotic at $x = 20°F$, what would we expect the potency to be?

$$\hat{y}(20) = 46 - \frac{19}{60}(20) = 119/3 = 39.\overline{6}.$$

But what does this really mean?
This is the expected value of y when $x = 20$.
If we store a bunch of samples at 20°F, we'd expect the sample mean $\approx 39.\overline{6}$.

In other words $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is an estimator for the average value of y @ x.

$\hat{y}$ is a statistic of its own.

And where there are point estimates, there are interval estimates.

We have that $\hat{\beta}_i \sim N(\beta_i, V(\beta_i))$.
So $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is also Normal at any fixed x.
To avoid formula confusion, we use x^* for the fixed x.

Note $E[\hat{Y}(x^*)] = E(\hat{\beta}_0 + \hat{\beta}_1 x^*)$
$= \beta_0 + \beta_1 x^*$

We need $V(\hat{Y})$.

$$V(\hat{Y}) = V(\hat{\beta}_0 + \hat{\beta}_1 x^*)$$
$$= V(\hat{\beta}_0) + (x^*)^2 V(\hat{\beta}_1) + 2x^* Cov(\hat{\beta}_0, \hat{\beta}_1)$$
$$= \frac{\sigma^2 \sum x_i^2}{n S_{xx}} + \frac{(x^*)^2 \sigma^2}{S_{xx}} + 2x^* \left(\frac{-\bar{x}\sigma^2}{S_{xx}}\right)$$
$$= \frac{\sigma^2}{S_{xx}} \left(\frac{\sum x_i^2 + n(x^*)^2 - 2x^* \bar{x} n}{n} \right)$$

(using $S_{xx} = \sum x_i^2 - n\bar{x}^2$)

$$= \frac{\sigma^2}{S_{xx}} \left(\frac{S_{xx} + n\bar{x}^2 + n(x^*)^2 - 2nx^*\bar{x}}{n} \right)$$
$$= \frac{\sigma^2}{S_{xx}} \left(\frac{S_{xx} + n((x^*)^2 - 2x^*\bar{x} + \bar{x}^2)}{n} \right)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$$

11/20

When estimating σ^2, make sure to use the new one $S^2 = \frac{1}{n-2} SSE$.

As in all things, we standardize $\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}}}$.

Here $\frac{\hat{Y}(x^*) - E(Y|x^*))}{\sqrt{V(\hat{Y})}} = \frac{\hat{Y}(x^*) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$

This is a t-statistic with $n-2$ df!

An $1-\alpha$ level confidence interval comes from

$$P\left(\left| \frac{\hat{Y}(x^*) - E(\hat{Y}(x^*))}{\sqrt{V(\hat{Y})}} \right| < t_{\alpha/2}(n-2) \right) = 1-\alpha.$$

$$\Rightarrow \boxed{(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2}(n-2) S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

ex. Find a 90% confidence interval for the average potency of our antibiotic when stored at 20°.

Here $x^* = 20$

$\hat{Y} = 46 - \frac{19}{60}(20) = 39.\overline{6}$

need $V(\hat{Y}) = S^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$

data set had $n = 12 \Rightarrow df = 10$

$\bar{x} = 60$

$S^2 = \frac{571}{30}$, $S_{xx} = 6000$

For 90%, $t_{0.05}(10) = 1.812$

$$39.6 \pm 1.812\sqrt{\frac{571}{30}\left(\frac{1}{12} + \frac{(100-60)^2}{6000}\right)}$$

$\sqrt{2.5810}$

$\Rightarrow 39.6 \pm 4.677$

Rmk: Of course, we could now do hypothesis testing if we'd like.

Something like.. $H_0: \hat{Y}(x^*) = y_0$

$H_a: \hat{Y}(x^*) \neq y_0$ etc

Our t-stat $$T = \frac{y_0 - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

w/ $n-2$ df

§11.7 : predictions on Y.

The difference btwn this section and the previous is perspective.
In 11.6, we focused on the spread of the average of y, $E[\hat{Y}]$
What if instead we wanted to focus on the distribution of the values that Y can take when $x = x^*$.

Recall $Y = \underbrace{\beta_0 + \beta_1 x}_{\text{deterministic part}} + \underbrace{\epsilon}_{\text{the "spread" of Y is here}}$

$V(Y) = V(\epsilon) = \sigma^2$.

Our estimate for Y at x^* is defined

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* + \epsilon$$

Since $\epsilon \sim N(0, \sigma^2)$, $E[\hat{Y}^*] = E[\hat{\beta}_0 + \hat{\beta}_1 x^*]$

$$= \hat{Y}(x^*)$$
$$= \beta_0 + \beta_1 x^*.$$

Computing variance is surprisingly easy.
As before, $\hat{Y}^*$ is the sum of Normal r.v.
Hence it is Normal.

Moreover, the assumption on ϵ is that it is independent of x.
We also have that S^2 is indy of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\text{So } V(\hat{Y}^*) = V(\hat{\beta}_0 + \hat{\beta}_1 x^*) + V(\epsilon) \text{ by indy.}$$
$$= \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) + \sigma^2$$
$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$$

this is the only difference from $V(\hat{Y})$.

When σ^2 is unknown...

$$V(\hat{Y}^*) = S^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$$

For the $1-\alpha$ level confidence interval

$$P(|T| < t_{\alpha/2}(df)) \geq 1-\alpha.$$

$$Z = \frac{\hat{Y}^* - E[\hat{Y}^*]}{\sigma_{\hat{Y}^*}} \quad \sim T = \frac{Y^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

t-distributed w $n-2$ df.

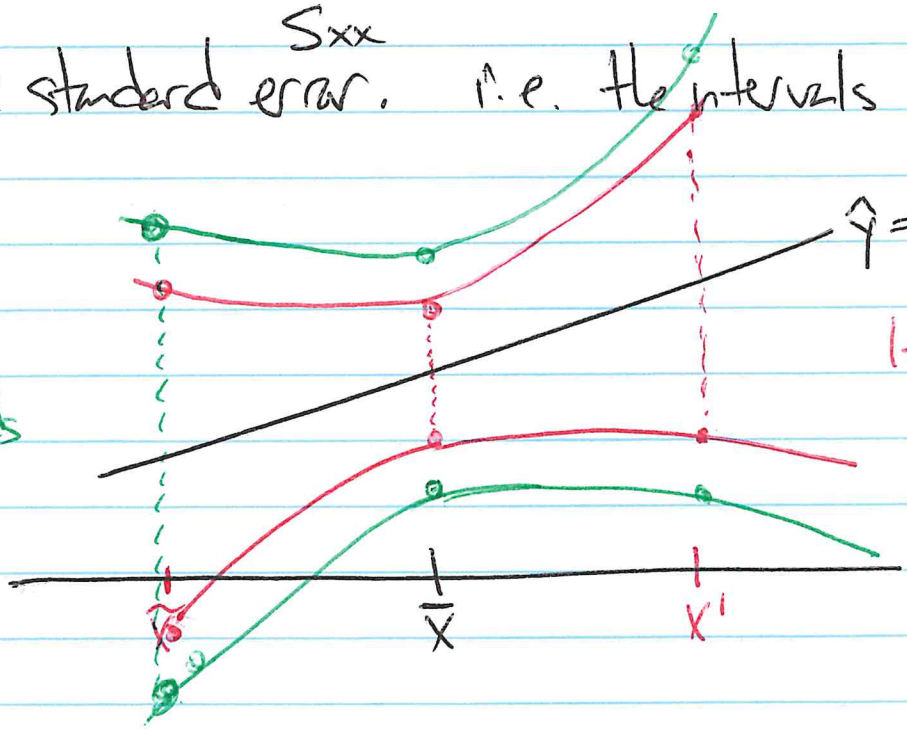
and $1-\alpha$ level C.I. is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

disc: confidence + prediction bands.

In §11.6, we had C.I. for the mean of $\hat{Y}(x^*)$
In §11.7, we have C.I. for the spread of Y at x^*.

Note that in either case, as x^* moves away from $\bar{x}$, the $\frac{(x^*-\bar{x})^2}{S_{xx}}$ increases and so does the standard error. i.e. the intervals get larger.



ex: potency of antibiotic when stored at 20°F

Last section we showed $E[\hat{Y}(20)]$ had the 90% C.I.

39.6 ± 4.677 (red band above)

What spread of potency would we expect to see with 90% confidence when stored at 20°F

11pdf

(the same calculation but w/ the new standard error)

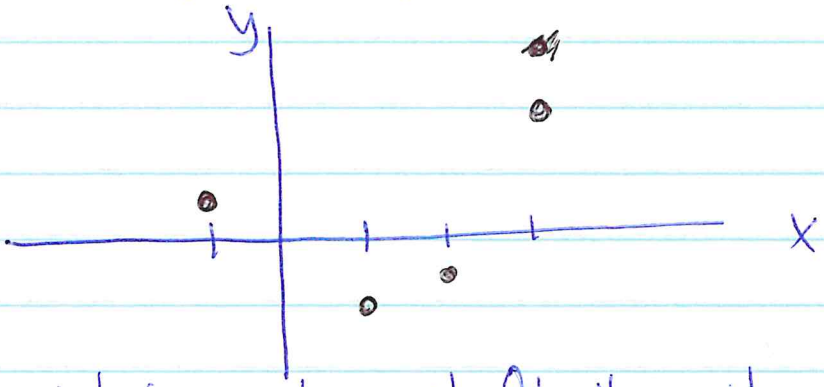
$$39.\bar{6} \pm 1.812\sqrt{1+\frac{1}{10}+\frac{1600}{6000}}$$

39.6 ± 5.069 (green band above)

§ 11.10 Multiple Linear Regression.

ex: Consider the data
$(-1, \frac{1}{2}), (1, -1), (2, -\frac{1}{2}), (3, 2)$

Plot



Seems unlikely a line will fit this well.
What about a parabola?
Want $y = b_0 + b_1 x + b_2 x^2$

(of course, in the probabalistic set up
$y = \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{\text{deterministic}} + \underbrace{\epsilon}_{\text{error}}$)

Using the data

$$\begin{aligned} \tfrac{1}{2} &= \beta_0 + \beta_1(-1) + \beta_2(-1)^2 \\ -1 &= \beta_0 + \beta_1 + \beta_2 \\ -\tfrac{1}{2} &= \beta_0 + 2\beta_1 + 4\beta_2 \\ 2 &= \beta_0 + 3\beta_1 + 4\beta_2 \end{aligned}$$

but this is the same idea as before, just a higher-dimensional problem

11/p27

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x} + \beta_2 \vec{x^2}$$

$$= \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} -1 \\ 1 \\ 2 \\ 3 \end{pmatrix} + \beta_2 \begin{pmatrix} 1 \\ 1 \\ 4 \\ 9 \end{pmatrix}$$

or, as a matrix-eqn

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{bmatrix}$$

Rmk: This is why it is called "linear" regression. The system for the coefs β_i will be a linear system.

But just as before, over determined $\Rightarrow$ no soln. The best we can do is find the projection.

We need a way to solve this projection problem in general.

topic: The Normal Eqns.

$$Y = \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x) + \cdots + \beta_n f_n(x) + \epsilon$$

Let there be m data pairs (x_i, y_i):
This will yield a matrix eqn

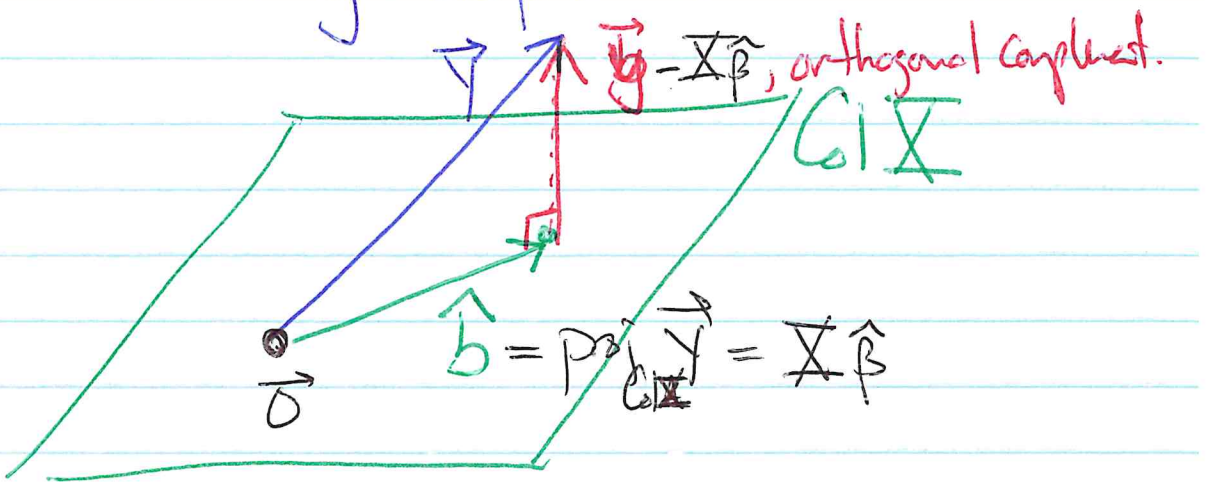
$$\begin{bmatrix} 1 & f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & & & \vdots \\ 1 & f_1(x_m) & & f_n(x_m) \end{bmatrix} \vec{\beta} = \vec{Y}$$

Call this $m \times n$ matrix X

$$X\vec{\beta} = \vec{Y}$$

Note, this setup only makes sense when system is overdetermined. That is $m > n$.
(more rows than columns)

We seek the coefficient vector $\hat{\beta} = \langle \hat{\beta}_0, \ldots, \hat{\beta}_n \rangle$
such that $\| \vec{y} - X\hat{\beta} \|$ is minimized



In the general case, there a 2 new issues to contend with:

① no reason the column vectors of X are actually a basis for $\mathrm{Col}\, X$ (might just be a spanning set).

② certainly no reason columns are an orthogonal set.

The implication of ① is that a least squares soln $\hat{\beta}$ need not be unique.
As for ②, ends up we don't need this.

By defn $\vec{y} - X\hat{\beta}$ is orthogonal to $\mathrm{Col}\, X$.
In other words, $\vec{y} - X\hat{\beta}$ must lie in the null space of X^T.
(Recall $(\mathrm{Col}\, A)^\perp = \mathrm{Nul}\, A^T$.)

So $X^T(\vec{y} - X\hat{\beta}) = \vec{0}$

or

$$X^T\vec{y} - X^TX\hat{\beta} = \vec{0}$$

which yields...

defn: The Normal Eqns $X^TX\hat{\beta} = X^T\vec{y}$.

FACTS:

① $\underline{X}^T\underline{X}$ is a square $n \times n$ matrix.

② The normal eqns is always a consistent system,
That is, there is a sol'n to the least squares problem.

③ If the columns of $\underline{X}$ are linearly independent, then $\underline{X}^T\underline{X}$ is invertible and the unique sol'n is

$$\hat{\beta} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\vec{y}.$$

ex: our ~~best~~-fit parabola.

$$\underline{X} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \qquad \underline{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 2 & 3 \\ 1 & 1 & 4 & 9 \end{bmatrix}$$

$\hat{\beta} = \langle \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_0 \rangle$, $\vec{y} = \langle 1/2, -1, -1/2, 2 \rangle$.

$$\underline{X}^T\underline{X} = \begin{bmatrix} 4 & 5 & 15 \\ 5 & 15 & 35 \\ 15 & 35 & 99 \end{bmatrix}$$

(nice $\det(\underline{X}^T\underline{X}) = 440$ ... invertible)

So $\hat{\beta} = (X^T X)^{-1} X^T \vec{Y} = \begin{bmatrix} -41/44 \\ -379/440 \\ 53/88 \end{bmatrix}$

So the best-fit parabola is

$$y = \frac{-41}{44} - \frac{379}{440}x + \frac{53}{88}x^2$$
$$\approx -0.932 - 0.861x + 0.602x^2$$

(show picture)

disc: revisit the best-fit line.

given n (x_i, y_i)'s, $y = \beta_0 + \beta_1 x + \epsilon$

$\Rightarrow [\vec{1}\ \vec{x}]\hat{\beta} = \vec{Y}$, $\hat{\beta} = \langle \hat{\beta}_0, \hat{\beta}_1 \rangle$

Here $X = [\vec{1}\ \vec{x}]$ is $n \times 2$.

$X^T = \begin{bmatrix} \vec{1}^T \\ \vec{x}^T \end{bmatrix}$ is $2 \times n$.

Then $X^T X = \begin{bmatrix} \vec{1}^T \\ \vec{x}^T \end{bmatrix} [\vec{1}\ \vec{x}]$ is 2×2

$$= \begin{bmatrix} \vec{1}\cdot\vec{1} & \vec{1}\cdot\vec{x} \\ \vec{x}\cdot\vec{1} & \vec{x}\cdot\vec{x} \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

Note $\det(X^T X) = n\sum x_i^2 - (\sum x_i)^2 = \Delta$

$$= n\sum x_i^2 - n^2\bar{x}^2$$
$$= n\left(\sum x_i^2 - n\bar{x}\right)$$
$$= nS_{xx} > 0 \quad \ldots \text{ invertible!}$$

So $\hat{\beta} = (X^TX)^{-1}X^T\vec{y}$ unique LSR sol'n!

$$X^T\vec{y} = \begin{bmatrix}\vec{1}^T\\ \vec{x}^T\end{bmatrix}\vec{y} = \begin{bmatrix}\vec{1}\cdot\vec{y}\\ \vec{x}\cdot\vec{y}\end{bmatrix} = \begin{bmatrix}\sum y_i\\ \sum x_i y_i\end{bmatrix}$$

Recall $\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$

So $(X^TX)^{-1} = \frac{1}{nS_{xx}}\begin{bmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{bmatrix}$

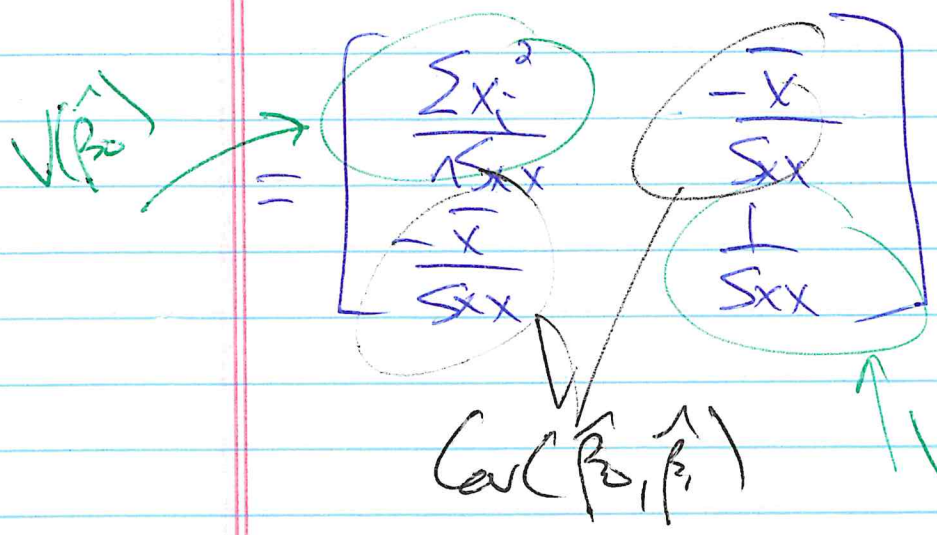
and
$$\hat{\beta} = \frac{1}{nS_{xx}}\begin{bmatrix}\sum x_i^2 & -\sum x_i\\ -\sum x_i & n\end{bmatrix}\begin{bmatrix}\sum y_i\\ \sum x_i y_i\end{bmatrix}$$

$$= \frac{1}{nS_{xx}}\begin{bmatrix}(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)\\ (-\sum x_i)(\sum y_i) + n\sum x_i y_i\end{bmatrix}$$

(these are the same as before)

But wait! There is more.

Look at $(X^TX)^{-1} = \frac{1}{nS_{xx}}\begin{bmatrix}\sum x_i^2 & -n\bar{x}\\ -n\bar{x} & n\end{bmatrix}$



We have seen these before! They are the coef. of all the $Cov(\,,\,)$

This actually always happens!

FACT: $\sigma^2 (X^T X)^{-1}$ is the table of covariances.

i.e. $\sigma^2 a_{ij} = Cov(\hat{\beta}_i, \hat{\beta}_j)$ where $[a_{ij}] = (X^T X)^{-1}$.

Lastly, $s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ ...

FACT: $SSE = \vec{Y}^T \vec{Y} - \hat{\beta}^T X^T \vec{Y}$

Reason: $SSE = \sum (y_i - \hat{y}_i)^2$

$= \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

can be interpreted as the dot product

$= (\vec{Y} - X\hat{\beta})^T (\vec{Y} - X\hat{\beta})$

$= \vec{Y}^T \vec{Y} - 2\hat{\beta}^T X^T \vec{Y} + \hat{\beta}^T X^T (X\hat{\beta})$

Since $\hat{\beta} = (X^T X)^{-1} X^T \vec{Y}$, we have

$= \vec{Y}^T \vec{Y} - 2\hat{\beta}^T X^T \vec{Y} + \hat{\beta}^T X^T X (X^T X)^{-1} X^T \vec{Y}$

but $(X^TX)(X^TX)^{-1} = I$ so

$$SSE = \vec{y}^T\vec{y} - 2\hat{\beta}^T X^T\vec{y} + \hat{\beta}^T X^T\vec{y}$$

$$= \vec{y}^T\vec{y} - \hat{\beta}^T X^T\vec{y}.$$

ex: (anthbiotic, yet again)

temp $\vec{x} = \langle 30,30,30,50,50,50,70,70,70,90,90,90\rangle$

potency $\vec{y} = \langle 38,43,29,32,26,33,19,27,23,14,19,21\rangle$

$n=12$, $X^TX = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x^2 \end{bmatrix} = \begin{bmatrix} 12 & 720 \\ 720 & 49200 \end{bmatrix}$

$\det(X^TX) = 72000$

$(X^TX)^{-1} = \begin{bmatrix} 41/60 & -1/40 \\ -1/40 & 1/6000 \end{bmatrix}$

Note $V(\hat{\beta}_0) = \frac{41}{60}\sigma^2$,

$V(\hat{\beta}_1) = \frac{\sigma^2}{6000}$

$$X^T\vec{y} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix} = \begin{bmatrix} 324 \\ 17540 \end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{bmatrix} 46 \\ -19/60 \end{bmatrix}$$

same as before.

$\hat{y} = 46 - \frac{19}{60}x$.

$SSE = \vec{y}^T\vec{y} - \hat{\beta}^T X^T \vec{y} = \frac{571}{3}$

$\sigma^2 \simeq s^2 = \frac{SSE}{n-2} = \frac{573}{30} = 19.1$

§ 11.11 a big ole thm

Thm Let $Y_i = \beta_0 + \beta_1 f_1(x_i) + \beta_2 f_2(x_i) + \cdots + \beta_k f_k(x_i) + \epsilon_i$

where $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2 \ \forall i$.

Then the least-squares estimators are given by

$\hat{\beta} = (X^T X)^{-1} X^T \vec{y}$

provided $(X^T X)^{-1}$ exists and

① $E(\hat{\beta}_i) = \beta_i$ (unbiased estimators)

② $V(\hat{\beta}_i) = c_{ii}\sigma^2$ where $(X^T X)^{-1} = [c_{ij}]$

③ $Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$

④ $S^2 = \frac{SSE}{n-(k+1)}$ where $SSE = \vec{y}^T\vec{y} - \hat{\beta}^T X^T \vec{y}$

and $E[S^2] = \sigma^2$.

if additionally $\varepsilon_i \sim N(0, \sigma^2)$, then

(5) $\hat{\beta}_i$ is normally distributed

(6) $\frac{(n-(k+1))\,s^2}{\sigma^2} \sim \chi^2(n-(k+1))$

(7) s^2 and $\hat{\beta}_i$ are independent for all i.

§ 11.1.2 Hypothesis Testing + C.I. multiple regression!

We have $Y = \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x) + \cdots + \beta_k f_k(x)$

w/ LSR soln $\hat{\beta} = (X^T X)^{-1} X^T \vec{Y}$
$= \langle \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \rangle$

Motivational example: how do we test a specific $\hat{\beta}_i$?

To "pick off" a specific $\hat{\beta}_i$, we dot $\hat{\beta}$ with the $(i+1)^{th}$ standard basis vector $\vec{e}_{i+1} = \langle 0, 0, \ldots, 0, 1, 0, \ldots, 0 \rangle$ (i^{th} position $= +1$)

Then $\hat{\beta}_i = \vec{e}_{i+1} \cdot \hat{\beta} = \vec{e}_{i+1}^T \hat{\beta}$

Note $E(\vec{e}_{i+1}^T \hat{\beta}) = E(1 \cdot \hat{\beta}_i) = E(\hat{\beta}_i) = \beta_i$
and $V(\vec{e}_{i+1}^T \hat{\beta}) = V(\hat{\beta}_i)$

So a test of the form $H_0: \beta_i = (\beta_i)_0$
$H_a: \beta_i > (\beta_i)_0$
$<$
$=$

yields the statistic

$$Z = \frac{\hat{\beta}_i - (\beta_i)_0}{\sqrt{V(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - (\beta_i)_0}{\sqrt{\sigma^2 c_{ii}}}$$

$[c_{ii}] = (X^T X)^{-1}$

or $T = \dfrac{\hat{\beta}_i - (\beta_i)_0}{S\sqrt{c_{ii}}}$, a t-stat of $n-(k+1)$ df.

ex: We fit $(-1, 1/2)$, $(1,-1)$, $(2,-1/2)$, $(3,2)$

w/ $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ because it "looked" non-linear. Is there evidence that the data is in fact non-linear? That is, are we certain $\beta_2 \neq 0$?

$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

We had $\hat{\beta} = (X^T X)^{-1} X^T y$
$= \left\langle \frac{-41}{44}, \frac{-379}{440}, \frac{53}{88} \right\rangle$.

To pick off $\hat{\beta}_2$, we dot w/ $\vec{e}_3 = \langle 0\,0\,1 \rangle$.

$\vec{e}_3 \cdot \hat{\beta} = 53/88$

We had $X^T X = \begin{bmatrix} 4 & 5 & 15 \\ 5 & 15 & 35 \\ 15 & 35 & 99 \end{bmatrix}$, $\det(X^T X) = 440$

$(X^T X)^{-1} = \frac{1}{440}\begin{bmatrix} 260 & 30 & -50 \\ 30 & 171 & -65 \\ -50 & -65 & 35 \end{bmatrix}$

$V(\hat{\beta}_2) = c_{22} \cdot \sigma^2 = \frac{35}{440}\sigma^2$

For σ^2, need $S^2 = \frac{SSE}{n-3}$ ← 3 $\hat{\beta}_i$'s.

$SSE = \vec{Y}^T\vec{Y} - \hat{\beta}^T X^T \vec{Y} = \frac{4791}{440}$

$S^2 = \frac{4791}{1320} \approx 3.6295$

and $V(\hat{\beta}_0) \approx \frac{35}{440} \cdot S^2 \approx 0.317585$

~~w/ $\sqrt{V(\hat{\beta}_0)}$ w/ $\sqrt{V(\hat{\beta}_1)}$ = 1.0687~~

w/ $\sqrt{V(\hat{\beta}_0)} \approx 0.563547$

Using $T = \frac{\vec{e}_3^T\hat{\beta} - (\beta_0)_0}{\sqrt{V(\hat{\beta}_0)}} = \frac{53/88 - 0}{0.563547}$

$= 1.06872$ @ $df = 4 - 3 = 1$

$P(|T| > 1.06872 \mid df = 1) = 0.478760$

Surprising! Supports H_0. What went wrong?

Probably the small # of data points ⇒ $df = 1$.

This is why statisticians use multiple tests. For example, the SSE of the quadratic fit will be much smaller than the SSE of the linear fit.

ex: What about the potency example? Would quadratic have been a better fit?

i.e. have line of best fit $y = 46 - \frac{19}{60}x$

w/ SSE line $= \frac{571}{3} = 190.\overline{3}$

Instead try $y = \beta_0 + \beta_1 x + \beta_2 x^2$

$\Rightarrow \underline{X} = \left[\vec{1} \;\; \vec{x} \;\; \vec{x^2}\right]$

$$(\underline{X}^T\underline{X})^{-1} = \frac{1}{1.92 \times 10^6}\begin{bmatrix} 1.0944 \times 10^6 & -391200 & 3100 \\ -391200 & 147200 & -1200 \\ 3100 & -1200 & 1 \end{bmatrix}$$

$$\hat{\beta} = \left\langle \frac{582}{12}, \frac{-5}{12}, \frac{1}{1200} \right\rangle$$

any $y = \frac{582}{12} - \frac{5}{12}x + \frac{1}{1200}x^2$

$\approx 48.58 - 0.4166x + 0.00083x^2$

note $\hookrightarrow \sim 46$ $\hookrightarrow \sim -\frac{19}{60} = -0.31666$

pretty small!

To test hypo that model is better as a linear model consider $H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$.

Need $V(\hat{\beta}_2) = \frac{1}{1920000}\sigma^2 \simeq \frac{S^2}{1920000}$

For $S^2 = \frac{SSE}{n-3} = \frac{SSE}{9}$,

$SSE = \vec{Y}^T\vec{Y} + \hat{\beta}^T X^T \vec{Y} = 189$

and $S^2 = \frac{189}{9} = 21$.

So $V(\hat{\beta}_2) \simeq \frac{21}{1920000} = 1.09375 \times 10^{-5}$

and $\sqrt{V(\hat{\beta}_2)} \simeq 0.0033071$.

for T-stat, $T = \frac{\hat{\beta}_2 - (\beta_2)_0}{\sqrt{V(\hat{\beta}_2)}}$ @ df = 9

$= \frac{1/1200 - 0}{0.0033071}$

$= 0.251976$.

Well... this is no more conclusive than the last

$P(|T| > 0.251876 \mid df = 9)$

$= 0.8067194$.

Again does not support H_0.

disc: generalization.

Of course, this is math and we love to generalize. Don't have to use an SRV $\vec{e}_i$.

Let $\vec{a} = \langle a_0, \ldots, a_k \rangle$ be constant.

Then $\vec{a}^T \hat{\beta} = a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1 + \cdots + a_k \hat{\beta}_k$

a weighted linear combo of the $\hat{\beta}_i$'s.

Note $E(\vec{a}^T \hat{\beta}) = \sum_1^k a_i E(\hat{\beta}_i)$ by linearity

$= \sum_1^k a_i \beta_i$ by unbiased estimators.

$= \vec{a}^T \cdot \hat{\beta}$

Variance is more complicated.

$V(\vec{a}^T \hat{\beta}) = V(a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1 + \cdots + a_k \hat{\beta}_k)$

$= \sum_{i=1}^k \sum_{j=1}^k a_i a_j \operatorname{Cov}(a_i \hat{\beta}_i, a_j \hat{\beta}_j)$ by Chapter 5.

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \underbrace{\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j)}_{\text{these are just } c_{ij}\sigma^2 \text{ like } [c_{ij}] = (X^T X)^{-1}}$$

In fact, using matrix algebra, this is the same as

$$V(\vec{a}^T \hat{\beta}) = \vec{a}^T (X^T X)^{-1} \vec{a} \cdot \sigma^2$$

yielding $Z = \dfrac{\vec{a}^T\hat{\beta} - E(\vec{a}^T\hat{\beta})}{\sqrt{V(\vec{a}^T\hat{\beta})}} = \dfrac{\vec{a}^T\hat{\beta} - \vec{a}\beta}{\sigma\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}}$

(used if σ^2 known)

For a test, $H_0: \vec{a}^T\beta = (\vec{a}^T\beta)_0$

$H_a: \vec{a}^T\beta \ (\neq, >, <) \ (\vec{a}^T\beta)_0$

we use

$$T = \frac{\vec{a}^T\hat{\beta} - (\vec{a}^T\beta)_0}{S\sqrt{\vec{a}^T(X^TX)^{-1}\vec{a}}} \quad \text{on } n-(k+1) \text{ df}$$

where $S^2 = \dfrac{SSE}{n-(k+1)}$

Aside: Dr. Anand says a common usage of this is to test collections of $\hat{\beta}_i$ to be zero. In other words, to test the necessity of components in your model (like we did in previous examples)