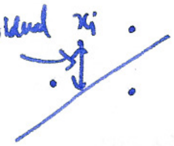
2.5 Warnings about correlation and regression

residuals: residual x_i  residual plot: 

- correlation does not imply causation.

10.1 Inference for linear regression

step 0 : plot data, check it looks linear.

model for simple linear regression: $(x_1, y_1), \dots (x_n, y_n)$ $\quad y_i = \beta_0 + \beta_1 x + \epsilon_i$

ϵ_i independent dist as $N(0, \sigma)$ parameters for the model are β_0, β_1, σ.

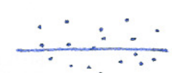
recall, best fit line has slope $b_1 = r\frac{s_y}{s_x}$ and intercept $b_0 = \bar{y} - b_1\bar{x}$.

(these are our estimates for β_1 and β_0). what about σ? look at residuals.

$e_i =$ observation $-$ predicted $= y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$

Fact estimate for σ^2 is $s^2 = \dfrac{\sum e_i^2}{n-2} = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}$ $\quad s = \sqrt{s^2}$ is model standard deviation.

model assumptions

- sample is an SRS
- linear relation in data (plot it)
- common standard deviation of residuals:  good  bad.
- which is normal: use qqplot.

Confidence intervals for slopes choose level C: $b_1 \pm t^* SE_{b_1}$

t^* critical C-value from t_{n-2}, $\quad SE_{b_1} = \dfrac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$ $\quad SE_{b_0} = s\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$

Hypothesis tests for slopes

$H_0: \beta_1 = 0$ compute test statistic $t = \dfrac{b_1}{SE_{b_1}}$ dist as t_{n-2}.

Estimated mean value at x^* : ~~true~~ mean $\hat{\mu}_y = b_0 + b_1 x^*$

Level C confidence interval for the mean response: $\hat{\mu}_y \pm t^* SE_{\hat{\mu}} = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

Estimated value at x^* : $\hat{y} = b_0 + b_1 x^*$ confidence interval $\hat{y} \pm t^* SE_{\hat{y}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

§12.1 One way analysis of variance (ANOVA)

recall §7: comparing two means. aim: compare means of many groups.

Example 12.2

Store	$\bar{x}_i$	s_i	n_i	pop	
A	38	8	50	μ_1	σ
B	48	13	50	μ_2	σ
C	42	11	50	μ_3	σ
D	28	7	50	μ_4	σ
E	35	10	50	μ_5	σ

data points.

x_{ij} i-th pop, j-th person in sample.

H_0: the means are all equal

H_a: the means are not all equal.

model: each population has distribution $N(\mu_i, \sigma)$ (here $\mu_1 \cdots \mu_5$ ~ $\mu_A \cdots \mu_E$).

note: s.d. assumed to be same in each pop!

estimates for parameters: $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$. estimate for μ_i

standard deviation of each sample: $s_i = \sqrt{\frac{\sum_{j=1}^{n_i} (x_i - \bar{x}_i)}{n_i - 1}}$.

note: ANOVA not particularly sensitive to differing σ_i, rough guide: if s_i vary by less than a factor of 2 usually ok.

pooled estimate for σ: $s_p^2 = \frac{(n_1-1)s_1^2 + \cdots + (n_I - 1)s_I^2}{(n_1-1) + (n_2-1) + \cdots + (n_1-1)}$. $s_p = \sqrt{s_p^2}$

note: s_p not average of s_i, its weighted by size of sample.

```
stores.aov <- aov(age ~ store, data = stores)
```

numeric for mean; categorical groups

summary(stores.aov):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
store	4	11154	2788.5	27.39	10^{-16}
residuals	245	24954	101.8		

sum sq/Df

if H_0 is true all means are the same, so $F = \frac{MSG}{MSE} \approx 1$ and dist as F-dist w/ $df = I - 1$

$MSE = s_p^2 \leftarrow$ estimate for ~~within~~ total group variance

$MSG = \leftarrow$ estimate for variance between groups.

use stores.csv data for this