STA 302 / 1001

Note Title 10/20/2011

Test has both interpretation / response to SAS output + theoretical / conceptual / derivation questions

Page 1 is posted.

Adjusted assignment 1 marks on Blackboard.

## Normal quantile plots

- assess whether data (assumed to be i.i.d. realizations of a r.v.) come from a normal distribution

## Order statistics of $r_1, \dots, r_n$

are $r_{(1)} = \min\{r_1, \dots, r_n\}$

$r_{(2)}$ = second smallest

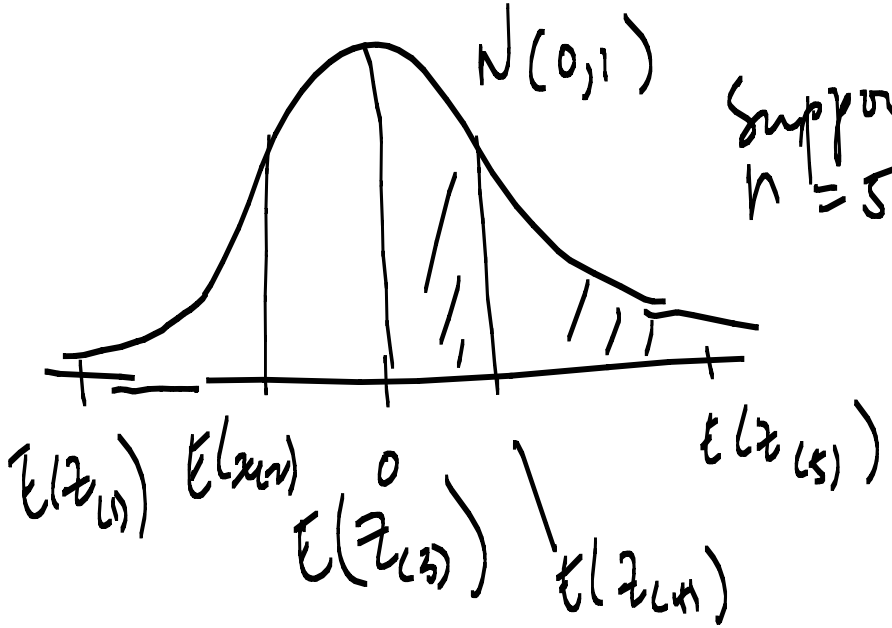
$\vdots$

$r_{(n)} = \max\{r_1, \dots, r_n\}$

Normal scores for samples of size n

$$E(Z_{(1)}), \ldots, E(Z_{(n)})$$

Where $Z_1, \ldots, Z_n$ are a sample of size n from $N(0,1)$ distribution




Suppose $n = 5$

equal probability between each normal score

If $r_1, \ldots, r_n$ are random samples from $N(\mu, \sigma^2)$

$z_i = \frac{r_i - \mu}{\sigma} \Rightarrow$ Samples from $N(0,1)$

$r_{(i)} = r_{(i)} - \mu$

$z_{(i)} = \frac{r_{(i)} - \mu}{\sigma}$

So if $r_i$ are $N(\mu, \sigma^2)$ distributed

$r_{(i)} \approx \sigma E(z_{(i)}) + \mu$

or $r_{(i)} = \sigma E(z_{(i)}) + \mu$

So plot $r_{(i)}$ vs $E(z_{(i)})$ should be approx. straight.

How to get $E(z_{(i)})$

Idea $E(z_{(i)}) = \Phi^{-1}\left(\frac{i}{n}\right)$

$\Phi^{-1}$: inverse standard Normal cdf

$\rightarrow$ value s.t. prob $\leq$ it is $i/n$

Have problem when $i = n$, $\Phi^{-1}\left(\frac{n}{n}\right) = \infty$

Most common adjust:

Blom $\quad E(z_{(i)}) \doteq \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right)$

- has been shown to work well in simulations

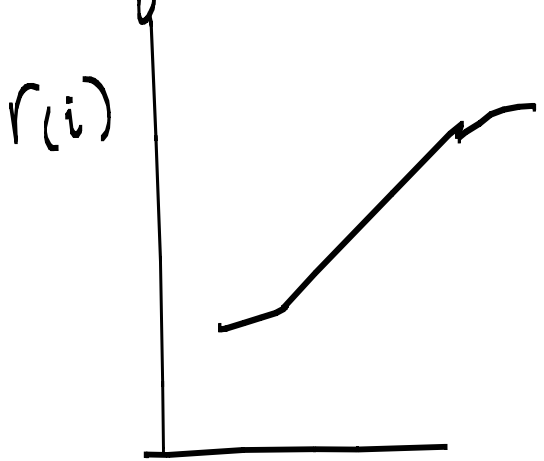
If assumption of normality is relaxed slightly to allow for heavier tails but still symmetric,

F- and t-tests and CIs for parameters are still approximately correct (p-value close, CI coverage % close)

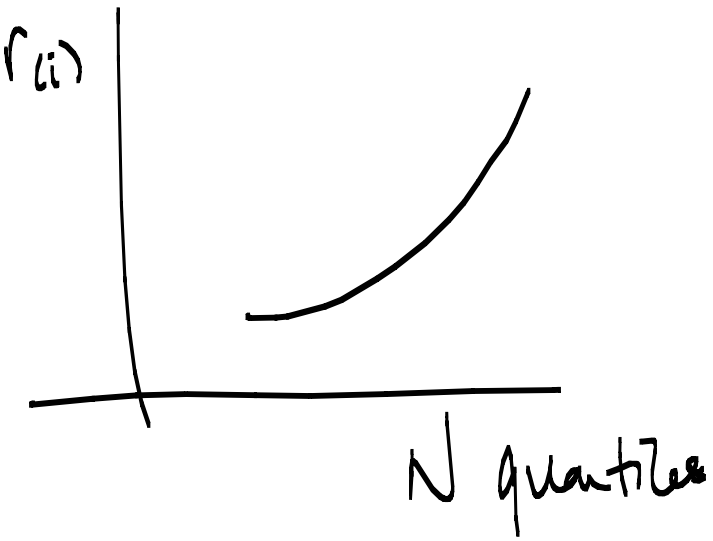
Why? estimators of parameters are linear combinations of r.v.'s and CLT

Heavy tails:


Light tails



Skewed

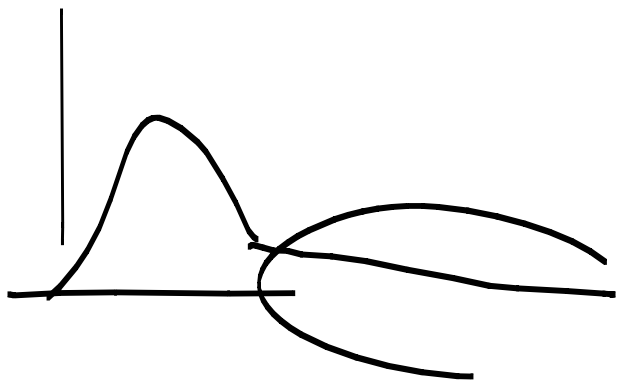



Mauna Loa CO2 example

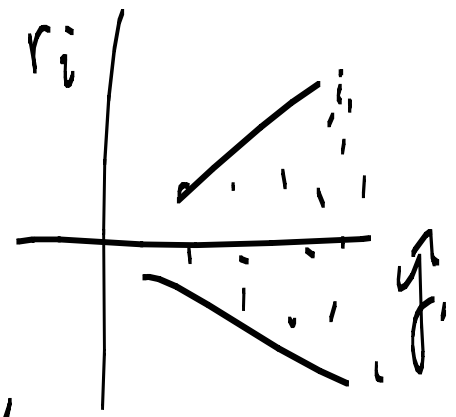
- residual plots (vs time or vs $\hat{y}$) show curvature
- this might indicate that a polynomial model would be better

- but in this case, that doesn't fix the problem
- The real problem: data collected over time, violation of G-M condition of uncorrelated error.

(to fix this: STA 457)

# TRANSFORMATIONS TO STABILIZE THE VARIANCE




(Theoretical result: "Delta Method")

If $Y$ has a distribution with mean $\mu$ and variance $\sigma_Y^2$

Then $Z = f(Y)$ has mean and variance approximately equal to

$$E(Z) \approx f(\mu)$$

("1st order linear

$$\mathrm{Var}(Z) \simeq \sigma_Y^2 \left[f'(\mu)\right]^2$$ ("approximately")

Taylor series expansion about $\mu$

"Proof" $$Z = f(Y) = f(\mu) + (Y-\mu) f'(\mu) + \text{remainder}$$

$$E(Z) = f(\mu) + E(Y-\mu) f'(\mu) + E(\text{remainder})$$

$$\simeq f(\mu)$$

$$\mathrm{Var}(Z) \simeq E\left[(Z - f(\mu))^2\right] + \text{different remainder term}$$

$$= E\left[\left((Y-\mu) f'(\mu)\right)^2\right] + \cdots$$

$$= [f'(\mu)]^2 \, E[(Y-\mu)^2]$$

$$= \sigma_Y^2 \, [f'(\mu)]^2 \quad //$$

This result gets used to derive variance stabilizing transformations

In SLR, $E(Y_i) = \beta_0 + \beta_1 x_i = \mu_i$

and want $\text{Var}(Y_i) = \sigma^2$ doesn't depend on $i$, can't depend on $\mu_i$,

Suppose $Y_i$ has mean $\mu_i$ and variance proportional to a function of $\mu_i$

i.e. $\text{Var}(Y_i) \propto V(\mu_i)$, say

Want to find a transformation $Z = f(Y)$ so

$$\text{Var}(Z) \simeq \text{const}$$

Want $f$ such that $\text{Var}(Z) \propto [f'(\mu)]^2 V(\mu)$

Need $[f'(\mu)]^2 = c / V(\mu)$, $c$ a constant

$$f'(\mu) \propto \frac{1}{\sqrt{V(\mu)}}$$

$$f(\mu) \propto \int \frac{1}{\sqrt{V(\mu)}} d\mu$$

For example (Poisson counts)

Suppose $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i$

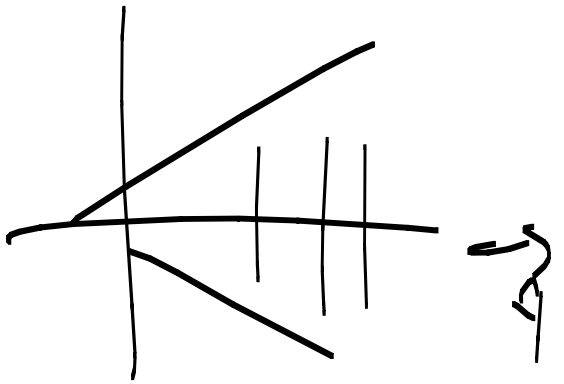
for

$$V(\mu) = \mu$$

$$f'(\mu) \propto 1/\sqrt{\mu}$$

$$f(\mu) \propto \sqrt{\mu}$$

If $\mathrm{Var}(Y)$ is linearly proportional to $E(Y)$, then $Z = \sqrt{Y}$ has variance that's approximately constant.



Another example:

Exponential, Gamma

$E(Y) = \mu$

$Var(Y) \propto \mu^2$

$V(\mu) = \mu^2$

$f'(\mu) \propto 1/\mu$

$f(\mu) \propto \log(\mu)$

NOTE: "log" = "ln" (natural logarithm)

ALWAYS for me (+ SAS).

If Var(Y) ↑ faster than linearly in E(Y),
log transformation often stabilizes the variance

In regression, these are the most useful transformations
- log stronger than $\sqrt{\ }$
- reciprocal: $f(Y) = \frac{1}{Y}$ is even stronger

Exercise: How is Var(Y) related to E(Y) if reciprocal transformation is appropriate.

What if data includes 0's or negative numbers?
and log transformation is appropriate

Use $\log(Y + k)$, $k$ a constant of your choice