Stat 411/511

ANOVA ASSUMPTIONS

Nov 9 2015

Charlotte Wickham

stat511.cwick.co.nz

DA #1 followup

Population inference to all opposite sex married couple households in Oregon.

(Response rates for ACS ~ 97.9%)

No causal inference.

Gender was not randomly assigned to members of a couple by a researcher.

Households were not randomly assigned to older or newer houses by a researcher.

…households with in older houses have higher (?) electricity costs on average…

NOT…having an old house will increase (?) your electricity costs…

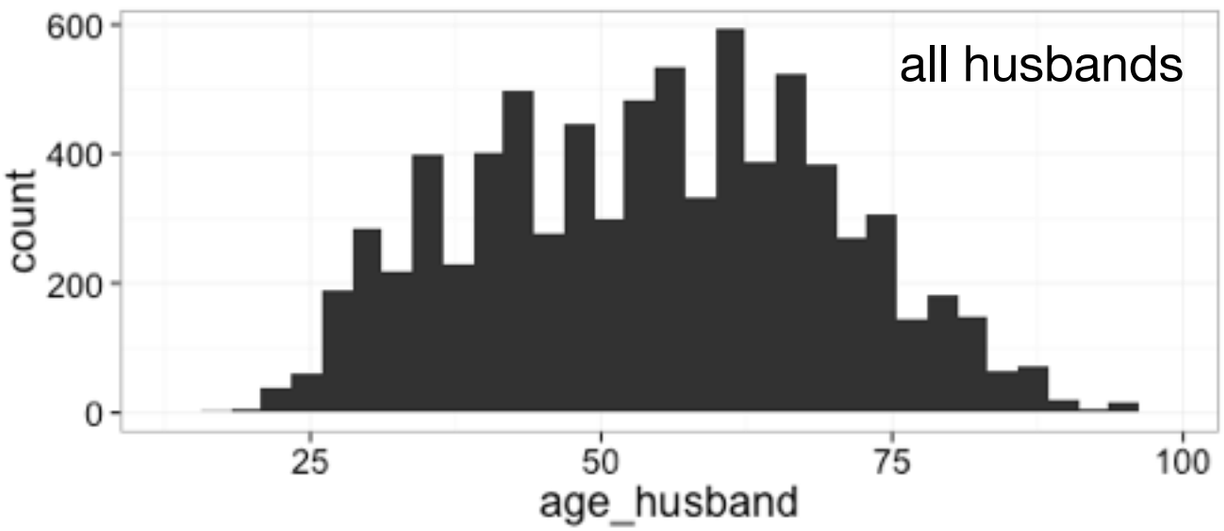
Is there dependence between the husband's age and the wife's age?

"If I **don't** know a wife's age, what's a good guess for the husband's age?"

"If I **do** know a wife's age, what's a good guess for the husband's age?"

If the answer to these questions is different, there is dependence between wife's age and husband's age.

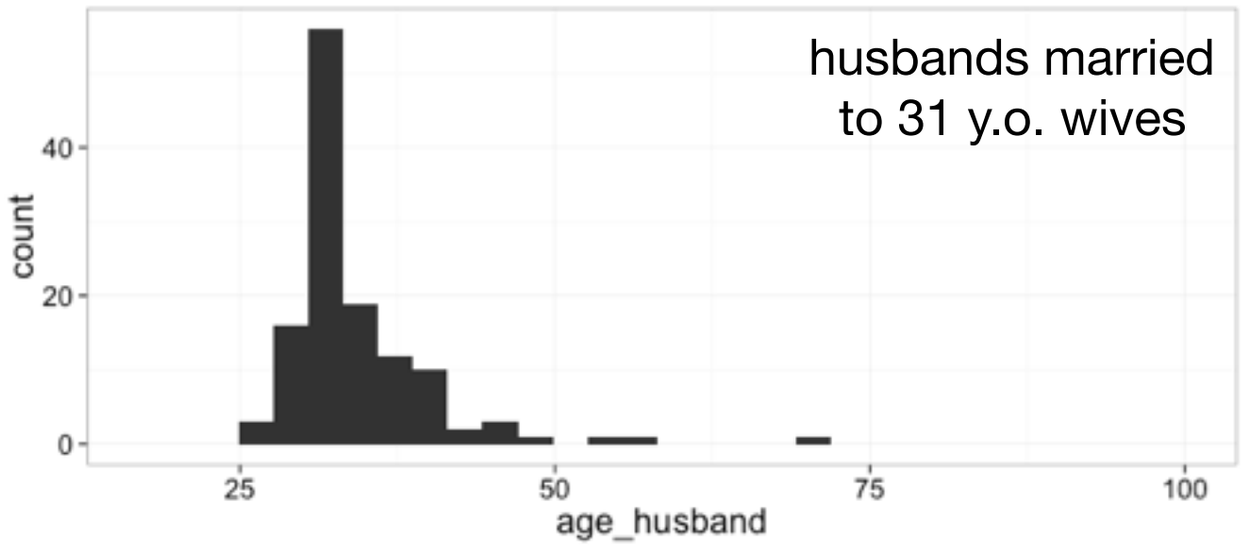
“If I **don’t** know a wife’s age, what’s a good guess for the husband’s age?”



On average,
husbands are 54 years old

50% of husbands are between
42 y.o. and 66 y.o.

“If I **do** know a wife’s age, what’s a good guess for the husband’s age?” Let’s say the wife is 31



On average,
husbands are 34 years old

50% of husbands are between
31 y.o. and 35 y.o.

different answers, dependent!

If I know a wife's age, I **can** make a better guess of their husband's age.

If I know a wife's age, I **cannot** make a better guess of a husband in a different household.

If I know a wife's age, I **cannot** make a better guess of a wife in a different household.

Dependence only comes from the pairing of wives and husbands into couples, do a **paired t-test**.

One way ANOVA practice

See practice problems posted under this lecture on the website.

Get the solutions by replacing “worksheet” with “solution” in the url.

Assumptions

of ANOVA F-test,
two group t comparisons &
linear combinations

1. Normally distributed populations.

2. Equal population standard deviations,

   $\sigma_1 = \sigma_2 = \ldots = \sigma_I = \sigma$.

3. Independence of subjects between and within groups.

Same as the two-sample t-test,
but now many samples.

Robustness

1. Normality

Like the t-tools, the ANOVA is robust to the normal population assumption with large sample sizes.

I.e. normality becomes less important for larger samples, because of the Central Limit Theorem

2. Equal population standard deviations.

Unlike the t-tools, the ANOVA is **not robust** to the equal standard deviation assumption.

Display 5.13 **p. 131**

Success rates for 95% confidence intervals for μ_1-μ_2 from samples simulated from normal populations with possibly different SDs

			$\sigma_2=\sigma_1$			$\sigma_2=\mathbf{2}\sigma_1$		
n_1	n_2	n_3	$\sigma_3=\sigma_1$	$\sigma_3=\mathbf{2}\sigma_1$	$\sigma_3=\mathbf{4}\sigma_1$	$\sigma_3=\sigma_1$	$\sigma_3=\mathbf{2}\sigma_1$	$\sigma_3=\mathbf{4}\sigma_1$
10	10	10	95.4	98.9	99.9	91.9	96.8	99.6
20	10	10	95.5	98.7	99.8	84.8	91.7	98.9
10	20	10	94.1	98.7	99.9	97.0	98.8	99.8
10	10	20	95.6	99.6	99.9	90.4	97.5	99.9

3. Independence of subjects between and within groups.

As always, this assumption is crucial. The ANOVA is not robust to this assumption.

Using residuals to check assumptions

It's often easier to verify the assumptions using the **residuals** from the full model, rather than the observations.

Plot:

Residuals by group

Residuals against fitted means

Residuals against other variables (like time)

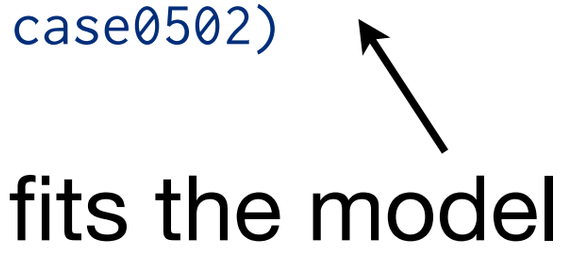
Should be centered around zero (in y direction) and roughly equal spread.

Probability plot of residuals

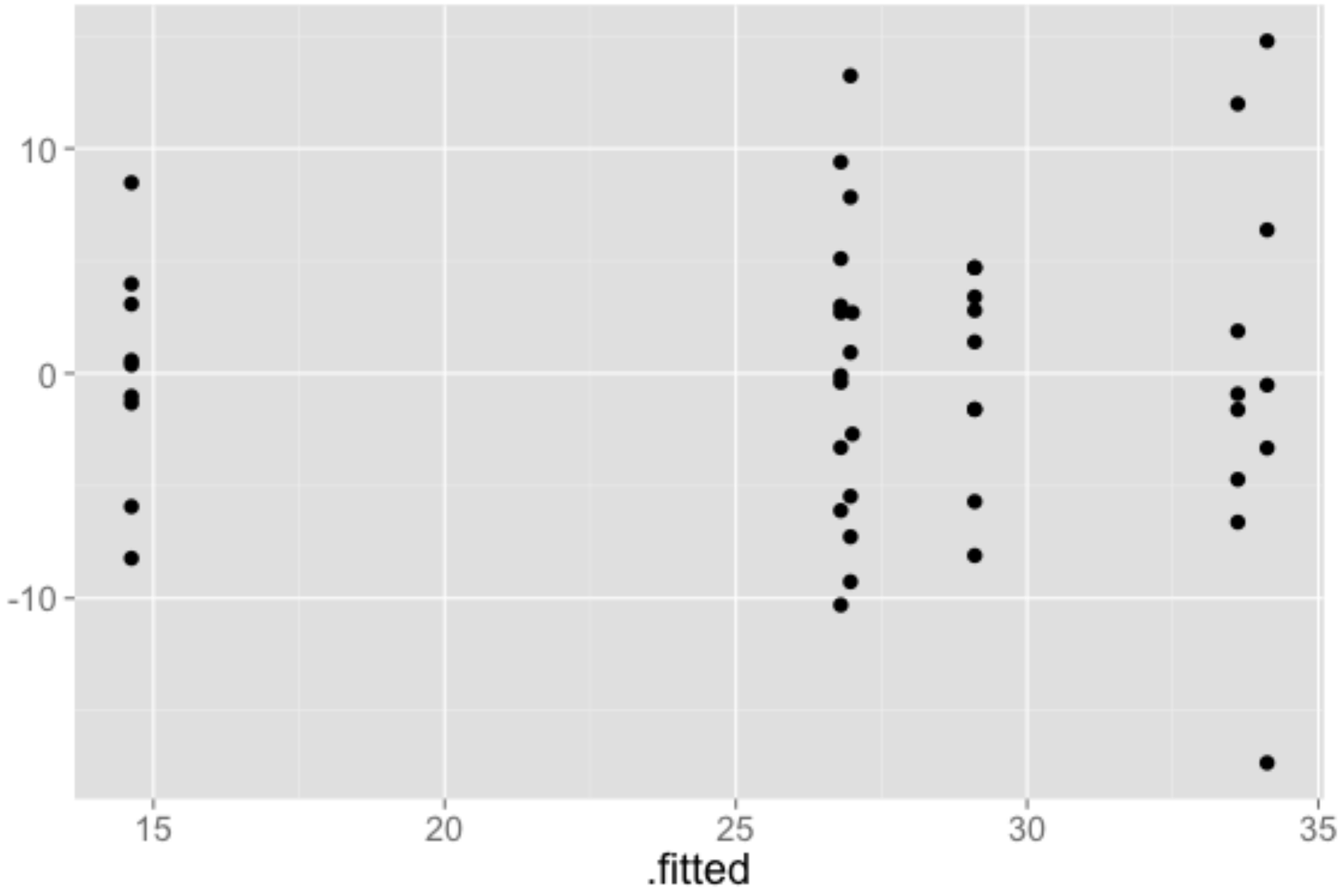
```
library(Sleuth3)
library(ggplot2)

separate_means_model <- lm(Percent ~ Judge, data = case0502)

str(separate_means_model)
summary(separate_means_model)
```



```
qplot(.fitted, .resid,
      data = separate_means_model)
```

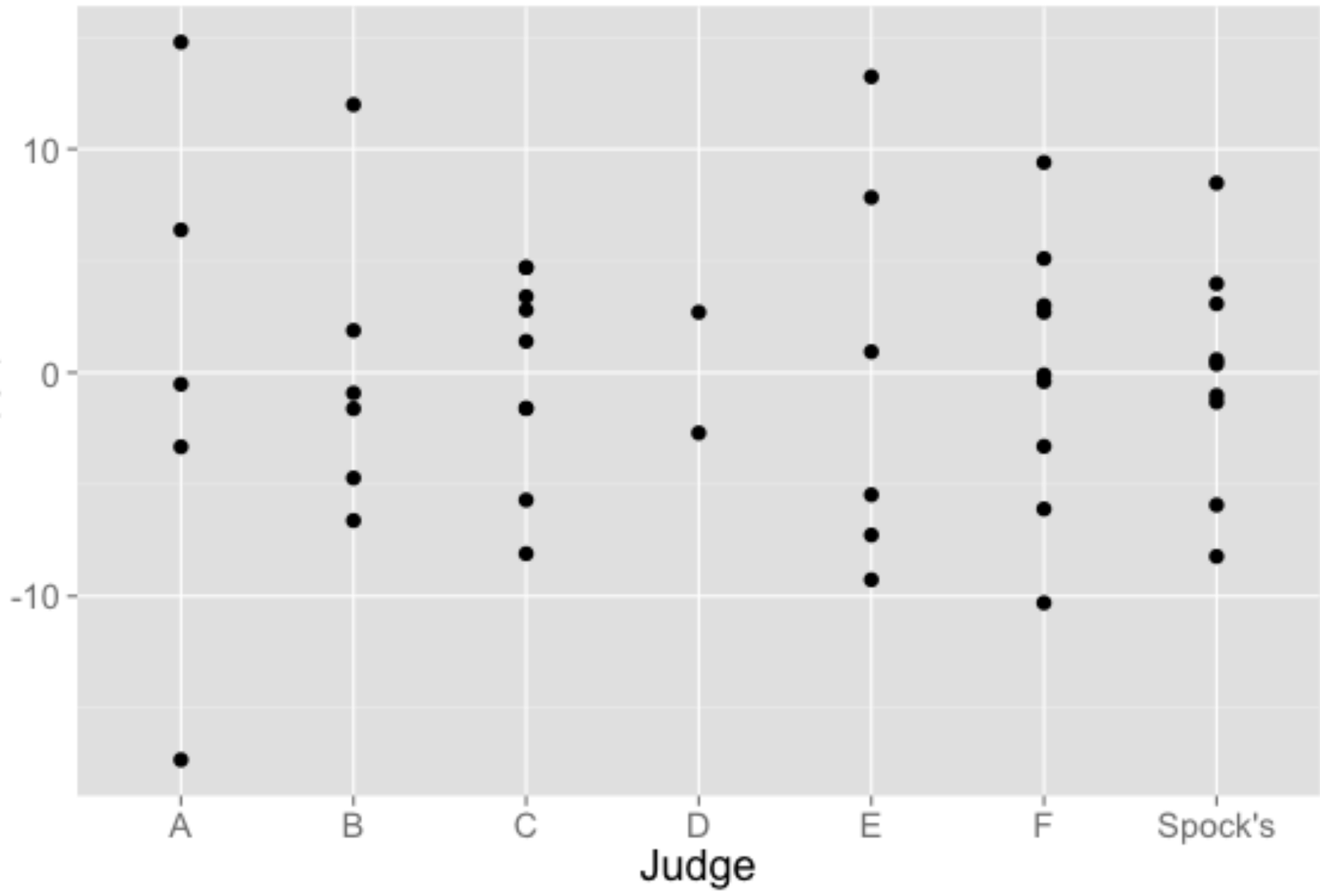


Residuals against fitted (estimated) group means

a funnel shape is a sign a log transform might be appropriate

```
qplot(Judge, .resid,
      data = separate_means_model)
```

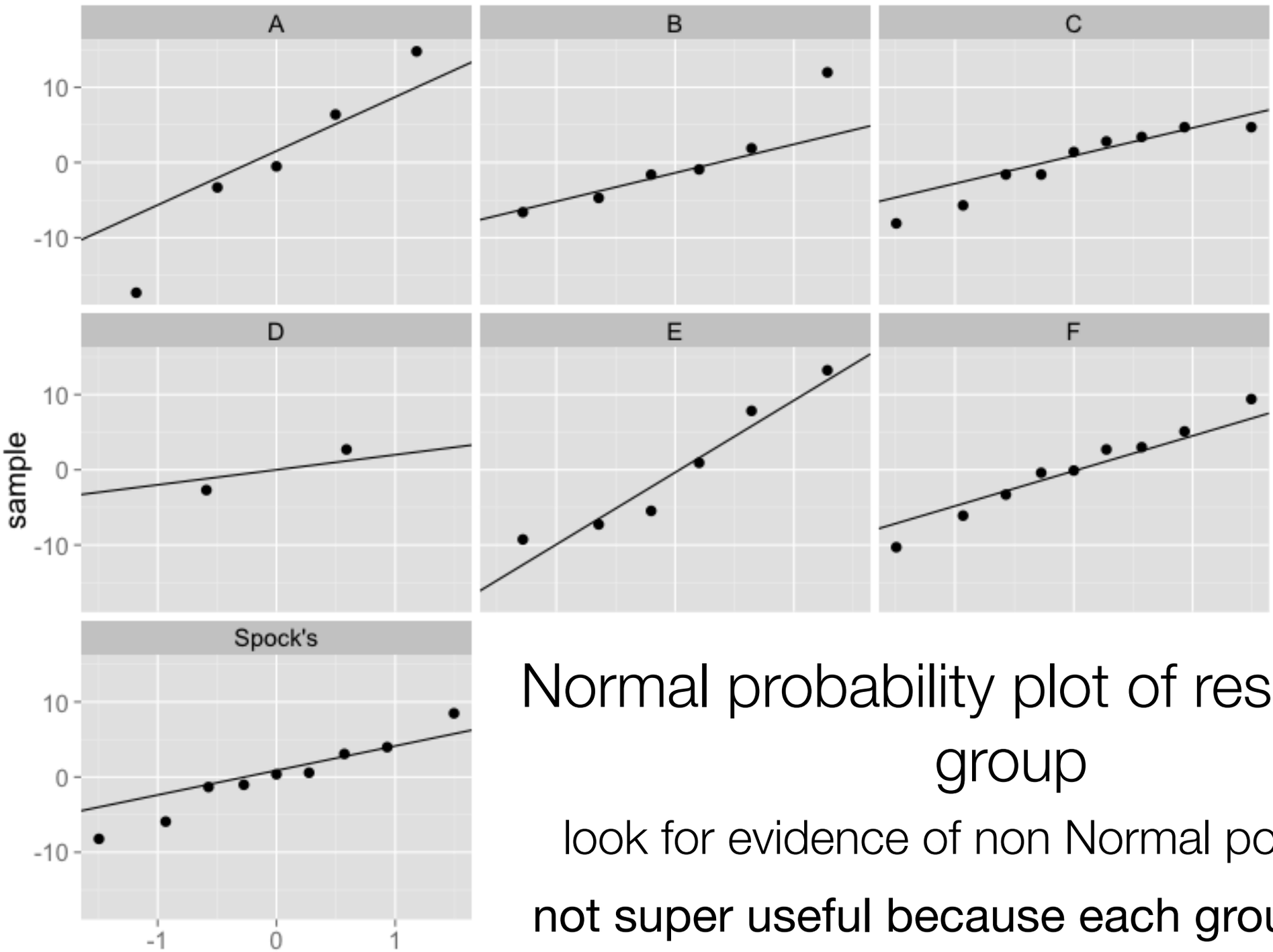
could also do histograms



Residuals against groups

look for evidence of unequal population standard deviations

```
qplot(sample =  .resid,
      data = separate_means_model) +
   stat_qqline() +
   facet_wrap(~ Judge)
```



Normal probability plot of residuals by group

look for evidence of non Normal populations

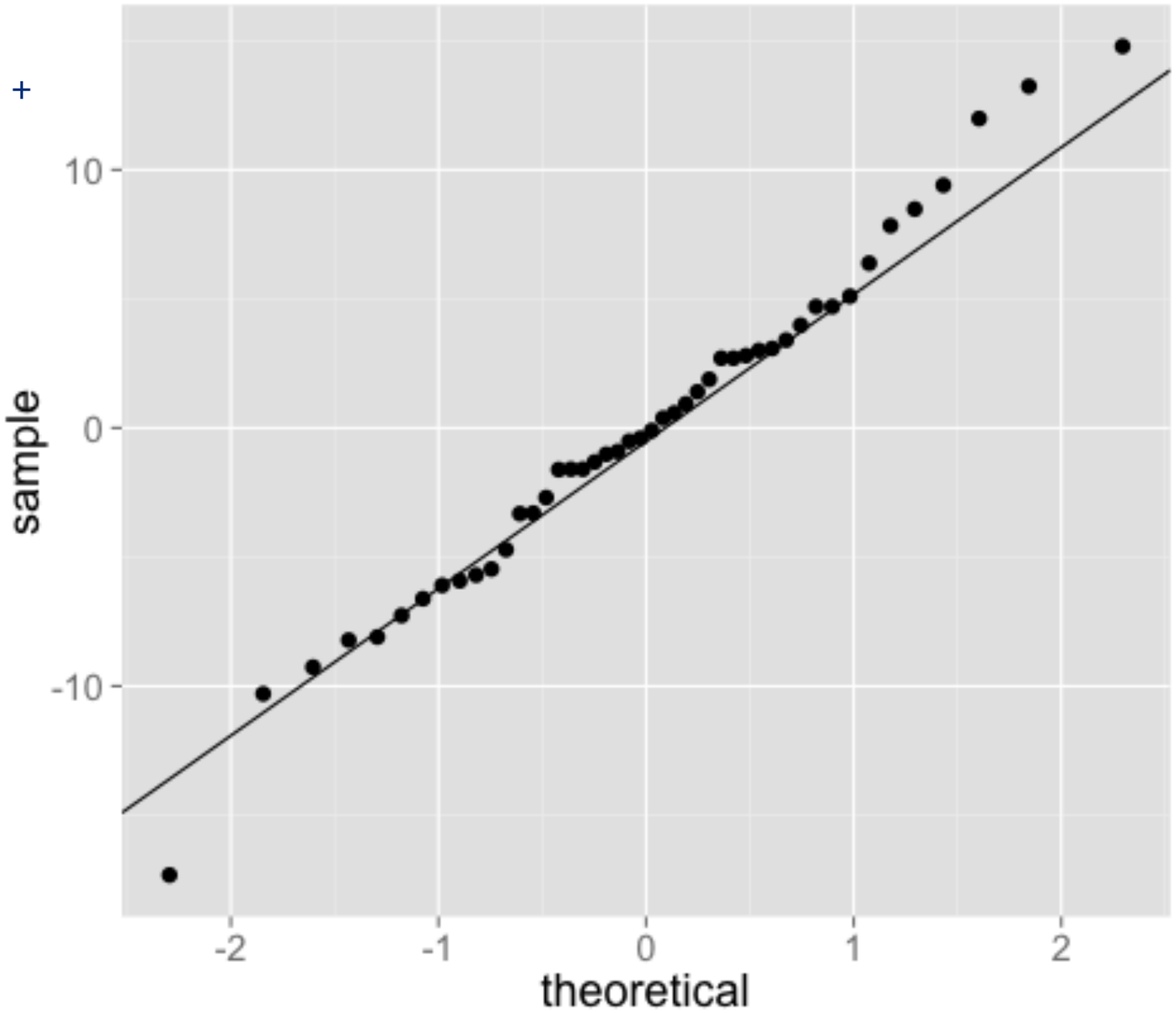
not super useful because each group is small

```
qplot(sample =  .resid,
      data = separate_means_model) +
   stat_qqline()
```

Normal probability plot of all residuals

look for evidence of non Normal populations

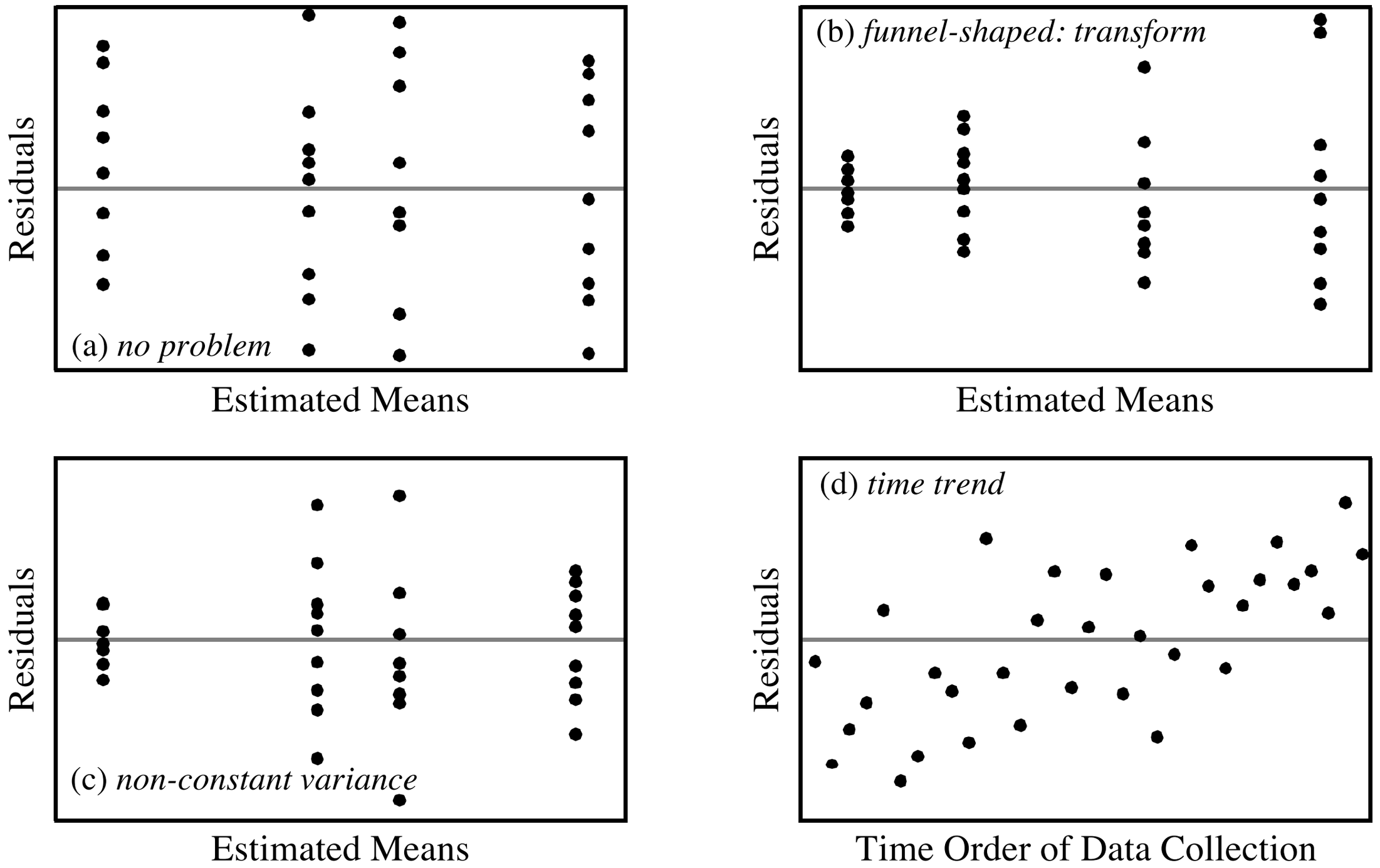
more useful



Why can we do this?

The populations have different means, but by using residuals we have subtracted off these different means (or at least a good guess). We also assume equal population standard deviations, so these residuals should all come from the same Normal distribution with mean zero regardless of the group.

Some important patterns in residual plots



Some extensions to the one-way ANOVA

Levene's test

Null: All groups have the same population standard deviation.

Procedure: Do a one-way ANOVA on the absolute deviations from the group medians.

Kruskal-Wallis ANOVA

The one-way ANOVA **is not resistant** to outliers.

The Kruskal-Wallis ANOVA removes the assumption of normal populations and is resistant to outliers.

Null: All groups have the same **median**.

Procedure: Convert response to ranks (ignoring groups, just like Wilcoxon Rank Sum), then do ANOVA on ranks (almost).

```
library(coin)
kruskal_test(Percent ~ Judge, data = case0502)
```

Section 5.6.2 in Sleuth

The one-way ANOVA is a common starting point for a multiple groups analysis, but it rarely directly answers the question of interest.

In some studies we know beforehand that at least one group will be different, and the one-way ANOVA isn't necessary.