Stat 412/512

Model Checking

Jan 30 2015

Charlotte Wickham

stat512.cwick.co.nz

Another note on indicator variables

You may have noticed it's difficult to write summaries about slopes relative to a baseline category.

A different **parameterization**, has an indicator variable for every category, but you have to drop some terms

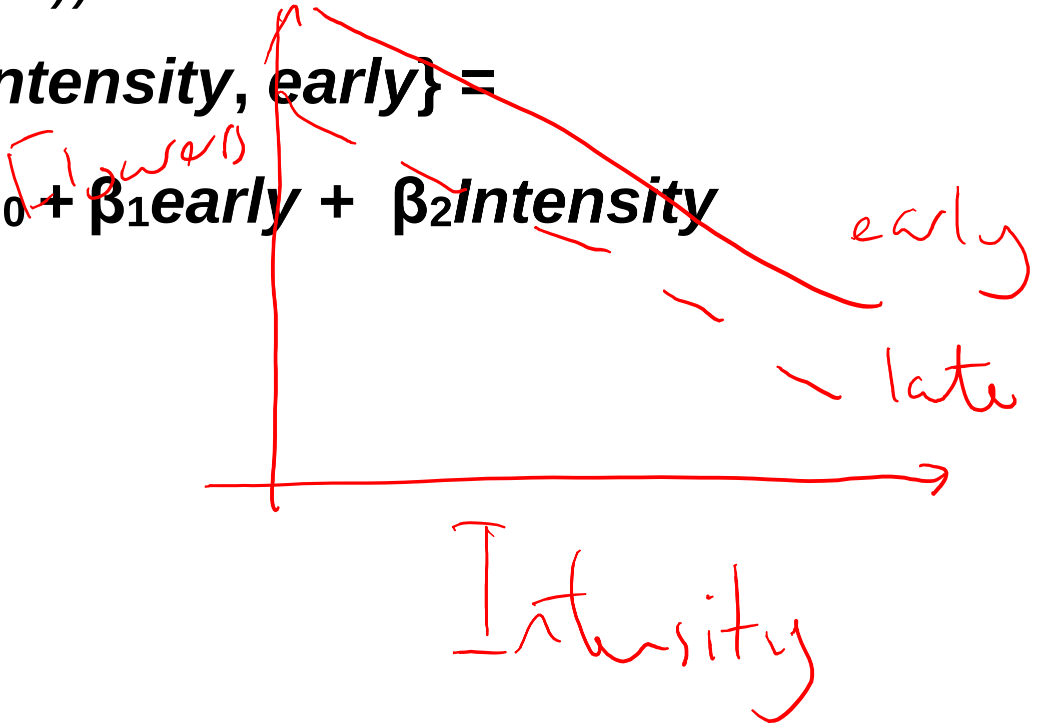
different parameterization: same model, but the parameters mean different things

```
> summary(lm(Flowers ~ Intens + Time, data = case0901))
...

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.305834   3.273772  21.781 6.77e-16 ***
Intens      -0.040471   0.005132  -7.886 1.04e-07 ***
TimeEarly   12.158333   2.629557   4.624 0.000146 ***
```

$\mu\{$ ***flowers* | *Intensity*, *early*** $\} =$

$\beta_0 + \beta_1$***early*** $+\ \beta_2$***Intensity***



drop the intercept

```
> summary(lm(Flowers ~ Intens + Time - 1, data = case0901))
...
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
Intens    -0.040471   0.005132  -7.886 1.04e-07 ***
TimeLate  71.305834   3.273772  21.781 6.77e-16 ***
TimeEarly 83.464167   3.273772  25.495  < 2e-16 ***
```

$\mu\{$ ***flowers* | *Intensity*, *early*** $\} =$

β_0***early*** $+\ \beta_1$ ***late*** $+\ \beta_2$***Intensity***

The models are equivalent, but we move from parameters that describe intercepts relative to the baseline, to absolute intercepts for each category.

$\mu\{\textit{flowers} \mid \textit{Intensity}, \textit{early}\} = \beta_0 + \beta_1 \textit{early} + \beta_2 \textit{Intensity} + \beta_3 \textit{early} \times \textit{Intensity}$

sep lines

```
> summary(lm(Flowers ~ Time + Intens + Intens:Time, data = case0901))
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      71.623333   4.343305  16.491 4.14e-13 ***
TimeEarly        11.523333   6.142361   1.876   0.0753 .
Intens          -0.041076   0.007435  -5.525 2.08e-05 ***
TimeEarly:Intens  0.001210   0.010515   0.115   0.9096
---
```

$\mu\{\textit{flowers} \mid \textit{Intensity}, \textit{early}\} = \beta_0 \textit{early} + \beta_1 \textit{late} + \beta_2 \textit{early} \times \textit{Intensity} + \beta_3 \textit{late} \times \textit{Intensity}$

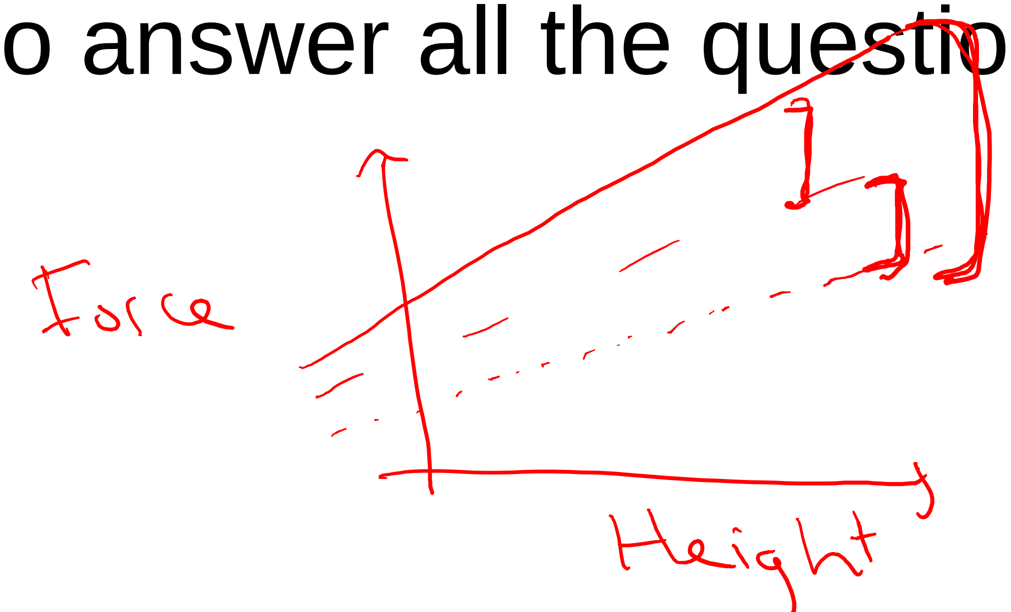
```
> summary(lm(Flowers ~ Time - 1 + Intens:Time, data = case0901))
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
TimeLate          71.623333   4.343305  16.491 4.14e-13 ***
TimeEarly         83.146667   4.343305  19.144 2.49e-14 ***
TimeLate:Intens  -0.041076   0.007435  -5.525 2.08e-05 ***
TimeEarly:Intens -0.039867   0.007435  -5.362 3.01e-05 ***
```

The models are equivalent, but we move from parameters that describe intercepts and slopes relative to the baseline, to absolute intercepts and slopes for each category.

It's a lot easier to picture the model for each group with this parameterization, but we lose the easy access to p-values that tell us whether there is evidence the groups have different lines.

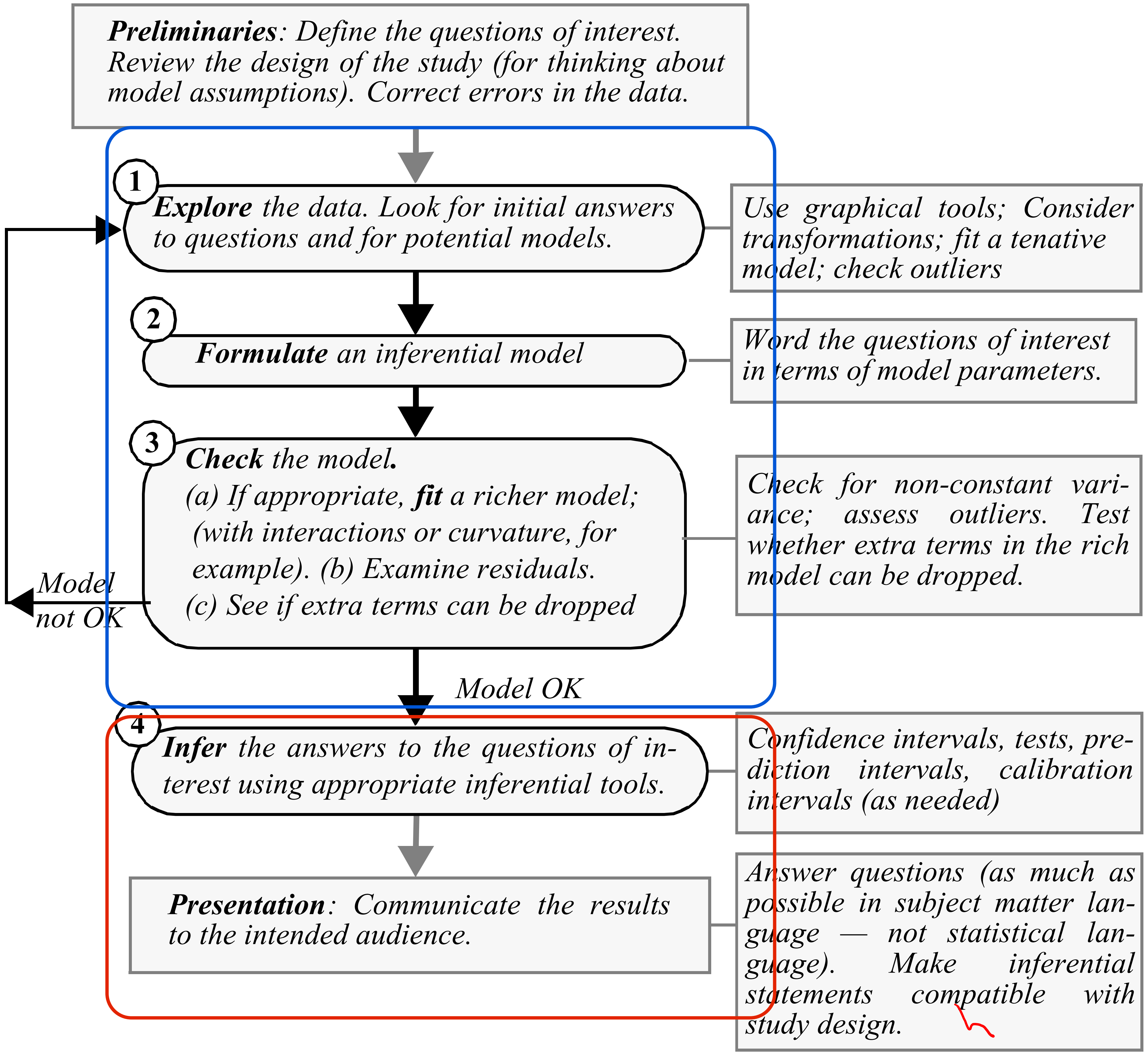
Convenience is generally the driver of a particular parameterization.

And often multiple parameterizations of the same model will be used to answer all the questions on interest.



Display 9.9 **p. 251**

A strategy for data analysis using statistical models



Preliminaries: *Define the questions of interest. Review the design of the study (for thinking about model assumptions). Correct errors in the data.*

1 ***Explore*** *the data. Look for initial answers to questions and for potential models.*

Use graphical tools; Consider transformations; fit a tenative model; check outliers

2 ***Formulate*** *an inferential model*

Word the questions of interest in terms of model parameters.

3 ***Check*** ***the model.*** *(a) If appropriate,* ***fit*** *a richer model; (with interactions or curvature, for example). (b) Examine residuals. (c) See if extra terms can be dropped*

Check for non-constant variance; assess outliers. Test whether extra terms in the rich model can be dropped.

Model not OK

Model OK

4 ***Infer*** *the answers to the questions of interest using appropriate inferential tools.*

Confidence intervals, tests, prediction intervals, calibration intervals (as needed)

Presentation: *Communicate the results to the intended audience.*

Answer questions (as much as possible in subject matter language — not statistical language). Make inferential statements compatible with study design.

Model Checking and Refinement

The best way to check the model is with residual plots, but you to have a model to fit.

Generally, you want to start with a model that:

- can answer your questions of interest
- includes confounding variables
- captures important relationships

and be willing to make adjustments as you go

Case 11.01 Alcohol Metabolism

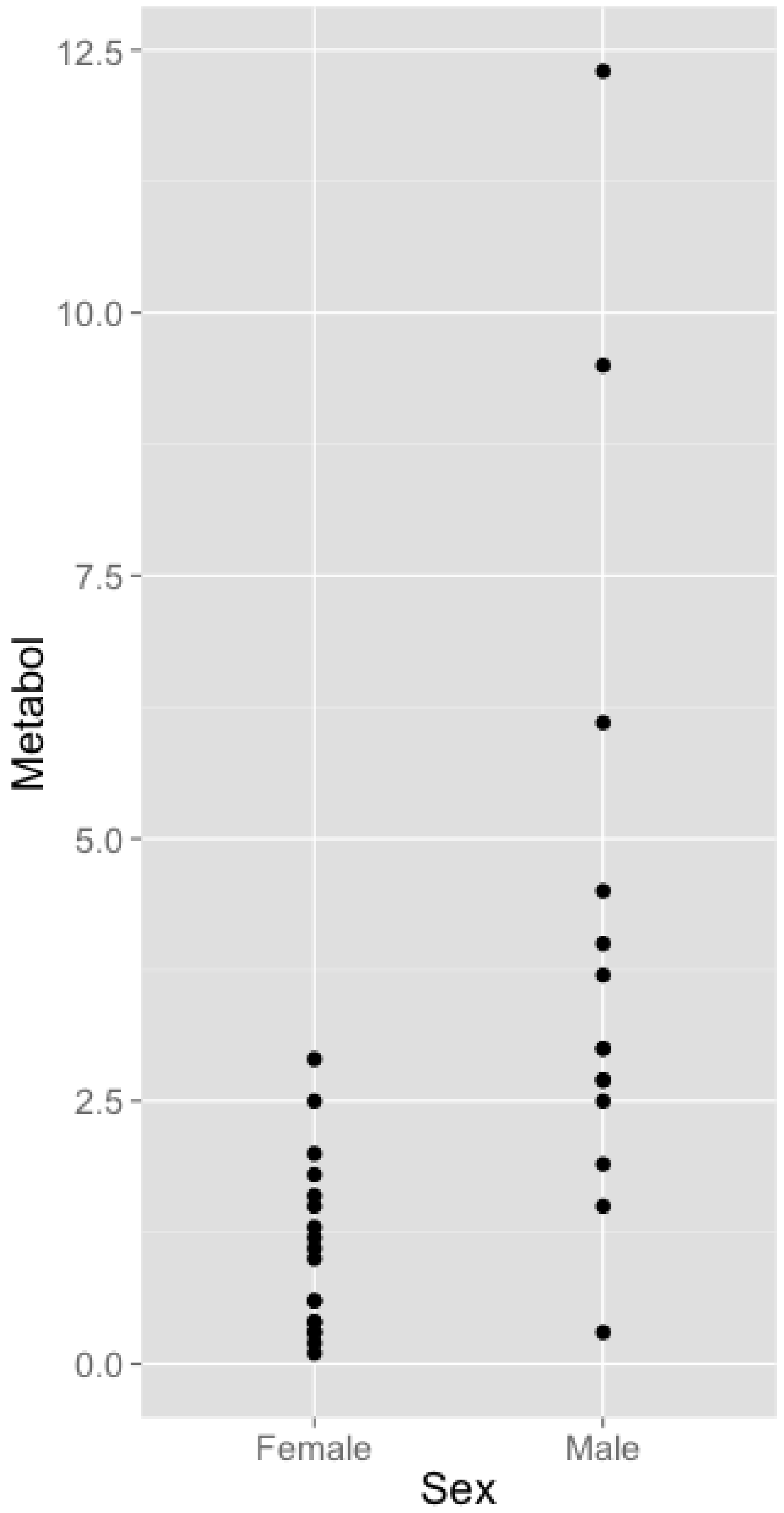
Women get drunk quicker than men. Women also develop alcohol related liver disease more readily.

Theory: a particular enzyme responsible for alcohol metabolism in the stomach is more active in men.

"first pass metabolism" = alcohol metabolized in the stomach so it doesn't reach the bloodstream

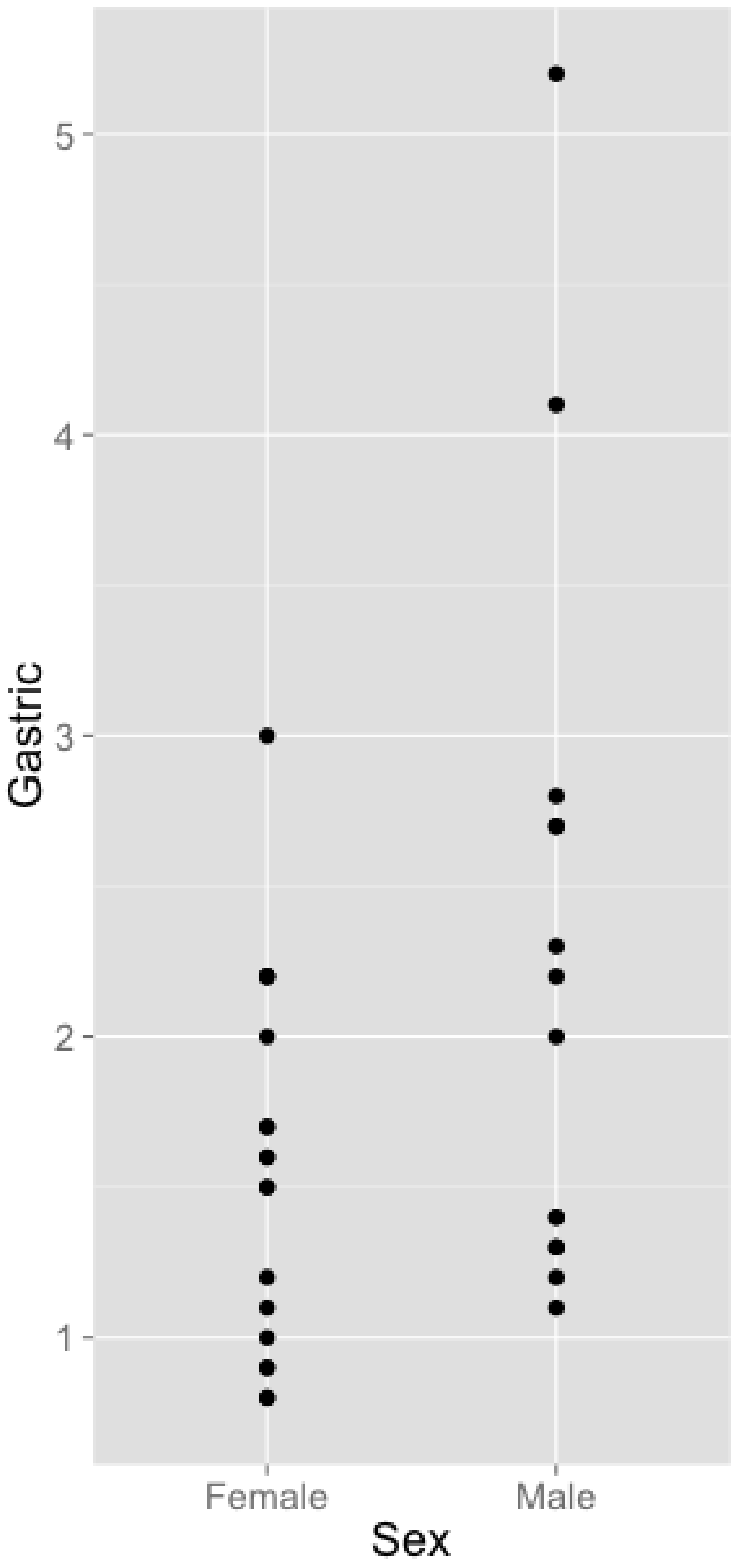
To determine first pass metabolism, compare blood alcohol levels after drinking to after intravenous alcohol.

Also measure enzyme activity.



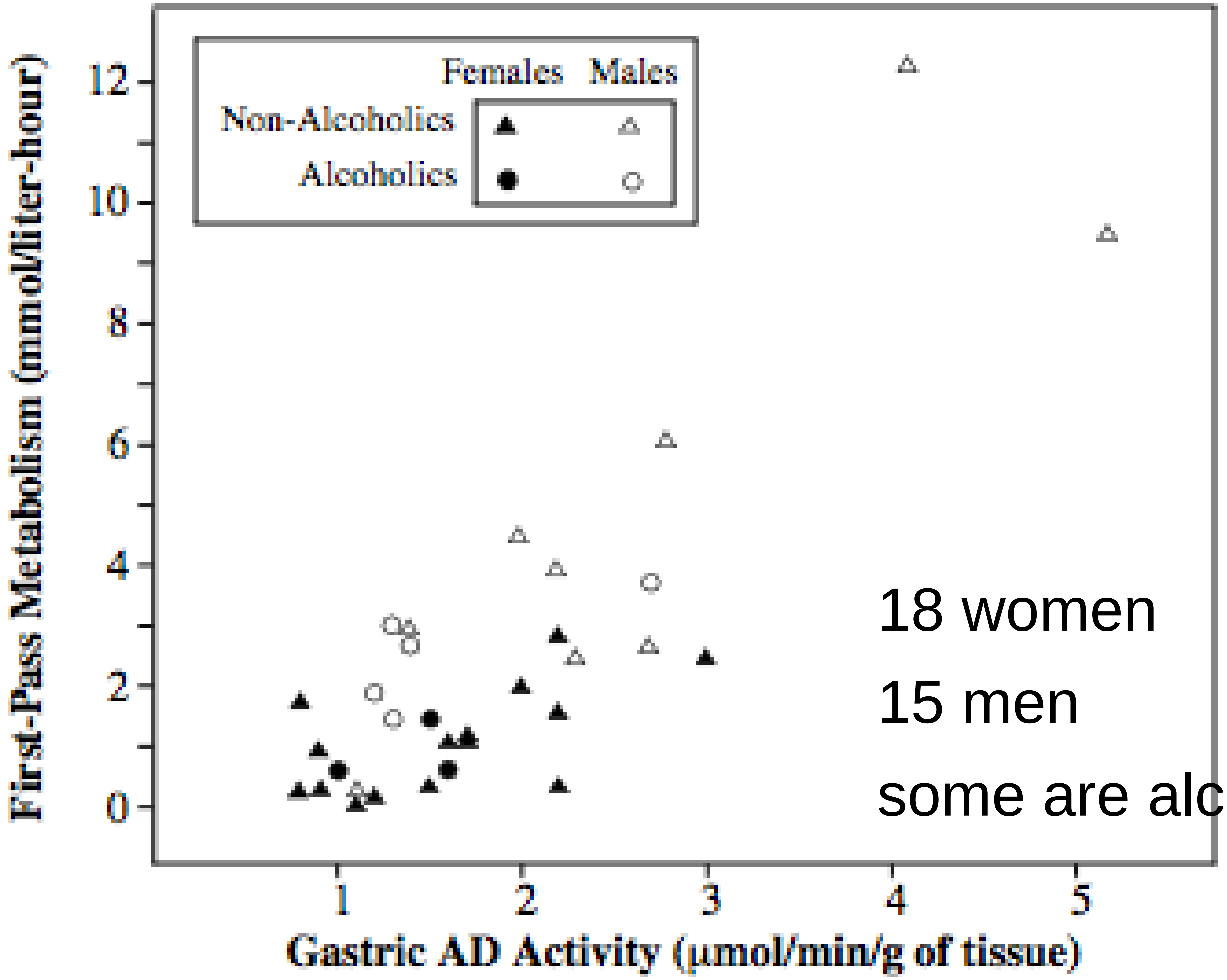
Alcohol Metabolism is greater for men

...but so is Gastric AD activity



Display 11.2 **p. 306**

First-pass metabolism and gastric alcohol dehydrogenase activity in alcoholic and non-alcoholic men and women



18 women

15 men

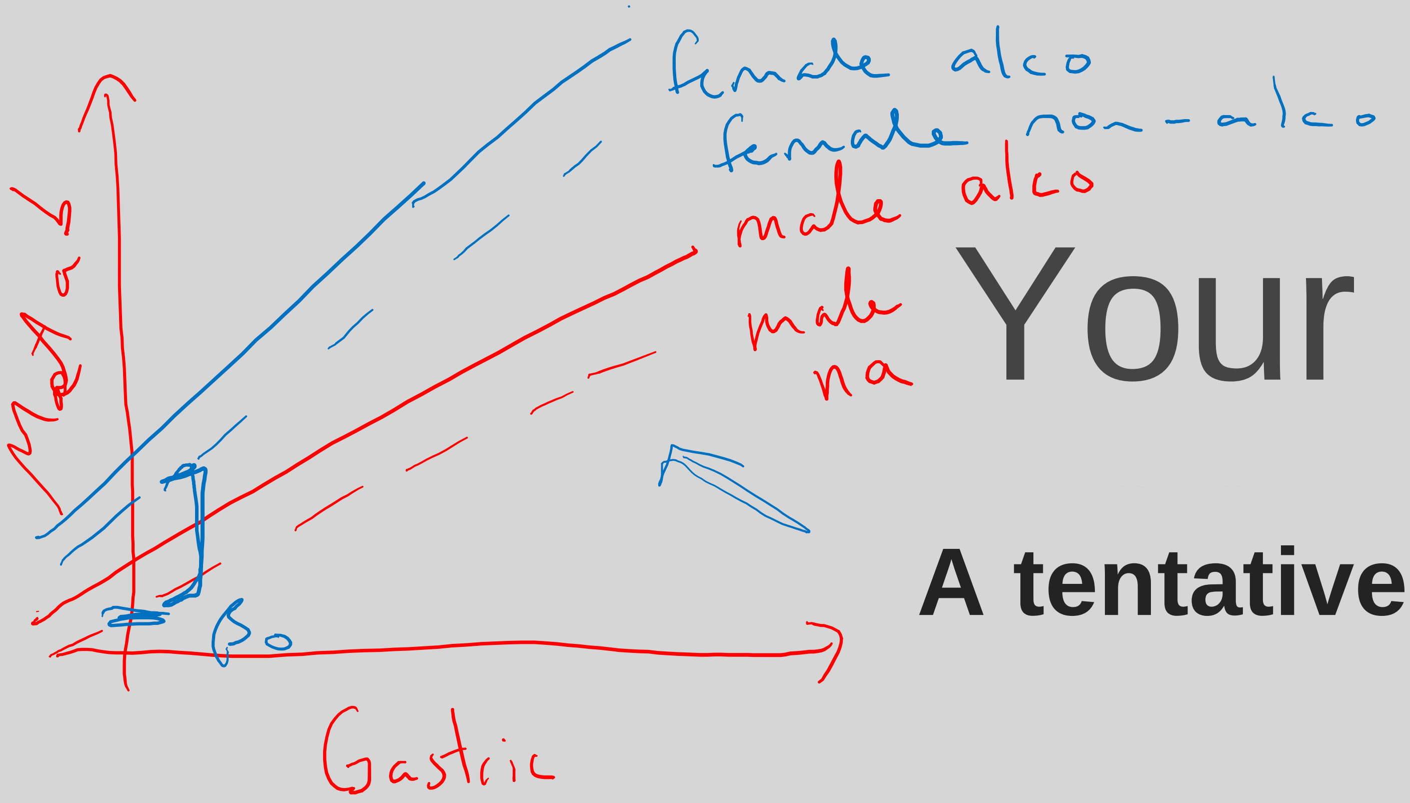
some are alcoholics

Questions of interest

Do levels of first pass metabolism differ between men and women?

Can the difference be explained by postulating that men have more enzyme activity in their stomachs?

Are the answers to these questions complicated by an alcoholism effect?



Your turn

baseline

male

non-alcoholic

A tentative model?

μ{ *First pass metabolism* | *gast*, female, *alcoholic*} =

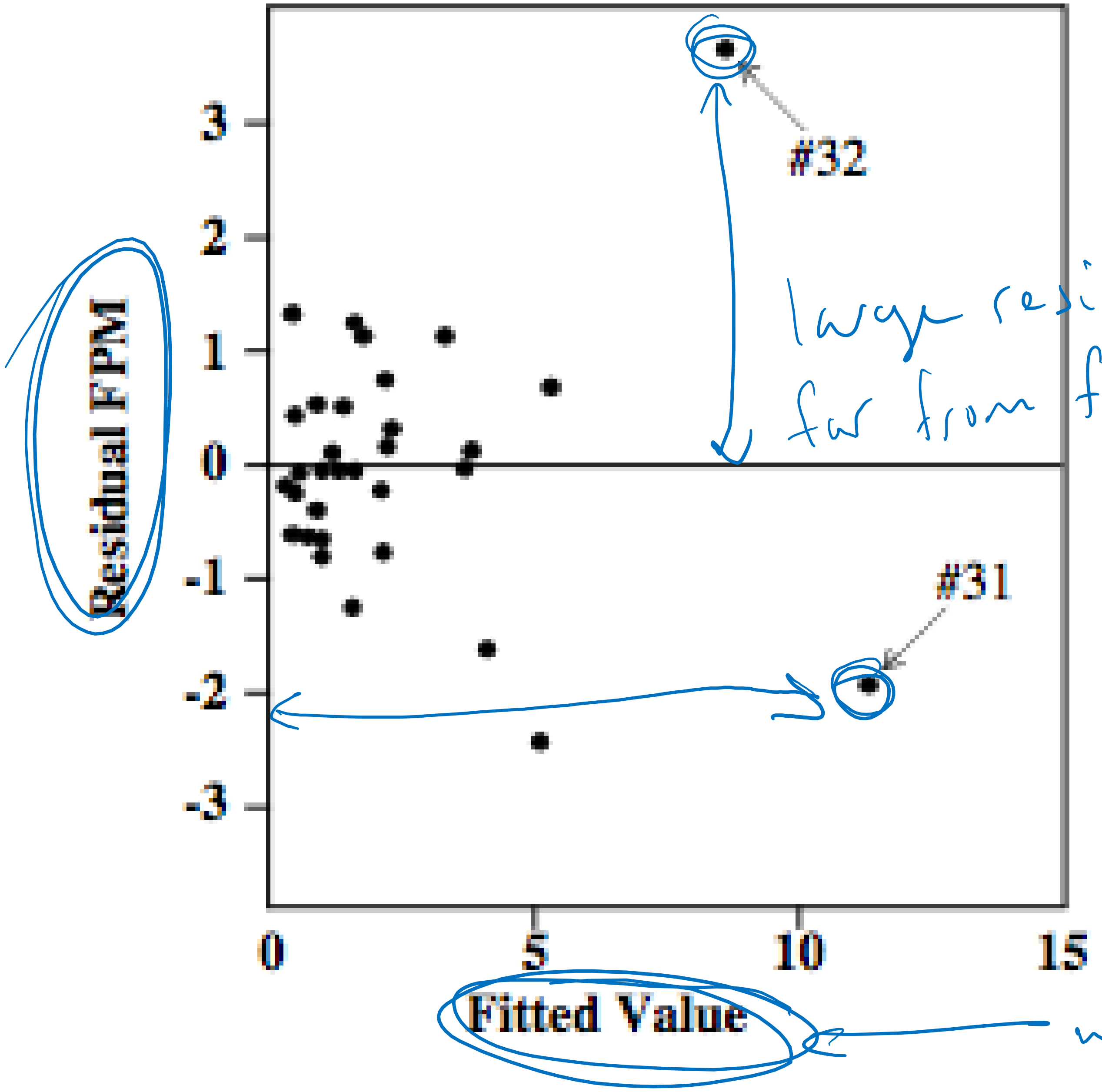
$$\beta_0 + \beta_1 \, gast + \beta_2 \, female + \beta_3 \, alcoholic$$

$$+ \beta_4 \, gast \times alcoholic + \beta_5 \, female \times alcoholic$$

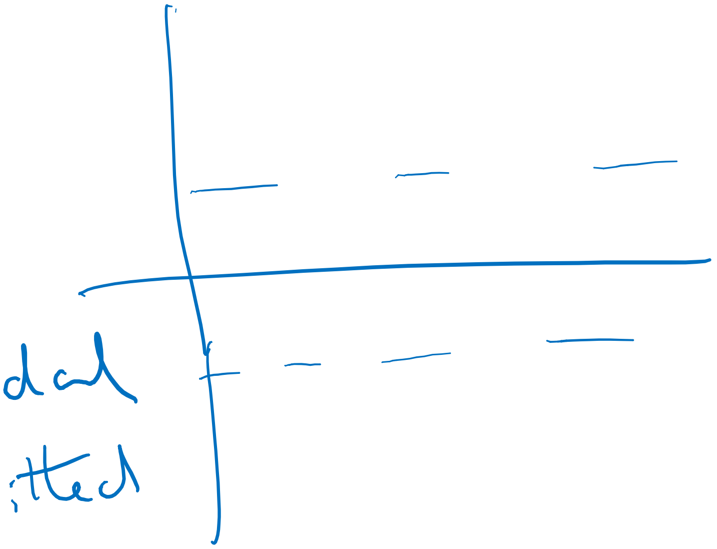
1 when female & alcoholic

$$\beta_6 \, female \times gast + \beta_7 \, female \times alcoholic \times gast$$

Residual plot from the regression of first-pass metabolism on gastric activity, sex indicator, alcoholism indicator, and all 2nd and 3rd-order interactions



large residual far from fitted line



unusual explanatory unusual

Outliers

Least squares estimates are not **robust** to outliers.

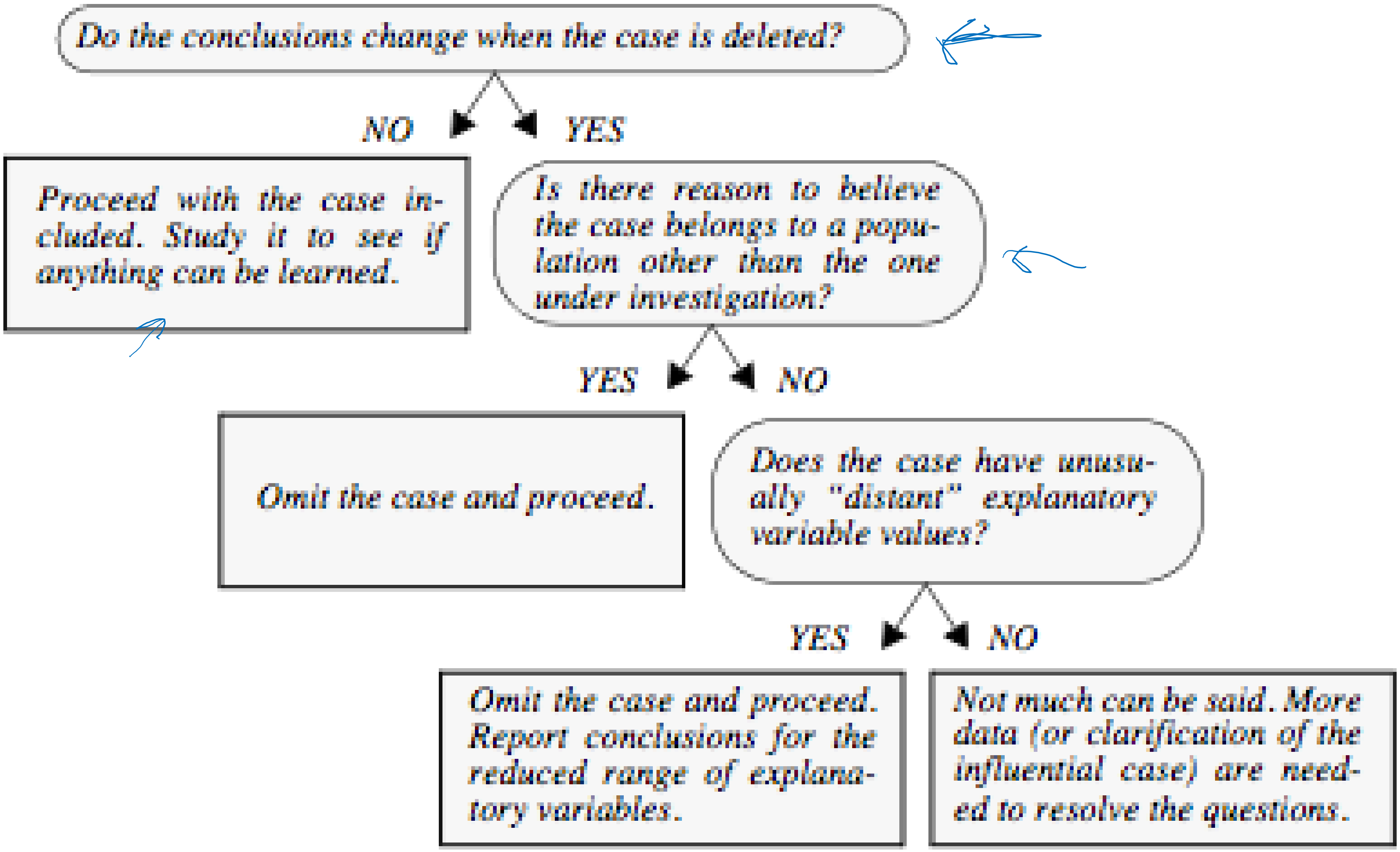
Identify outliers early on, so you don't end up tailoring the model to to fit a few unusual observations.

An observation is said to be **influential** if the fitted model depends unduly on its value.

For example, removing it: changes the estimate of parameters greatly, changes conclusions, or changes which terms are included in the model.

Display 11.8 p. 314

A strategy for dealing with suspected influential cases



Do the conclusions change when the case is deleted?

NO

Proceed with the case included. Study it to see if anything can be learned.

YES

Is there reason to believe the case belongs to a population other than the one under investigation?

YES

Omit the case and proceed.

NO

Does the case have unusually "distant" explanatory variable values?

YES

Omit the case and proceed. Report conclusions for the reduced range of explanatory variables.

NO

Not much can be said. More data (or clarification of the influential case) are needed to resolve the questions.

Case influence statistics

Case influence statistics help identify **observations** that may be influential.