Inference: Case study

Patrick Breheny

February 8

Alcohol metabolism data

- Our data set for today involves a study of 18 women and 14 men which investigated one aspect of why women exhibit a lower tolerance for alcohol and develop alcohol-related liver disease more readily than men
- The data set contains the following variables:
  - `Metabol`: First-pass metabolism of alcohol in the stomach (mmol/liter-hour); this is the outcome variable
  - `Gastric`: Gastric alcohol dehydrogenase activity in the stomach (μmol/min/g of tissue)
  - `Sex`: Sex of the subject
  - `Alcohol`: Whether the subject is alcoholic or not
- We will be asking a wide variety of inferential questions about this data set to illustrate the rich set of quantities we can estimate and hypotheses we can test with the tools we have learned so far

Sex as the explanatory variable

- Let's begin by fitting a simple model, with Sex as the only explanatory predictor:

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male},$$

  where Male is an indicator variable for being male; you could use Female instead, but recall that you can't use both
- In R, you can fit the above model with `lm(Metabol~Sex)`, but in PROC REG, you'll have to manually set up the indicator variable Male
- Alternatively in SAS, you could use PROC GLM with a CLASS SEX statement, but you would have to specify SOLUTION to get the parameter estimates

Thinking about the coefficients

- Now let's think about the model and what the coefficients mean:

$$\mathrm{E}(\texttt{Metabol}) = \begin{cases} \beta_0 & \text{if } \texttt{Sex==``Female''} \\ \beta_0 + \beta_1 & \text{if } \texttt{Sex==``Male''} \end{cases}$$

- Thus, β_0 is the expected alcohol metabolism for women, while β_1 is the difference in expected alcohol metabolism between men and women (in this situation, women are sometimes said to the *reference category*)
- Standard output reports the estimate, $\widehat{\mathrm{SE}}$, the test statistic, and the p-value for each coefficient in the model

Thinking about the coefficients (cont'd)

- The tests in the standard output may or may not be meaningful
- For example, $H_0 : \beta_1 = 0$ is a hypothesis that alcohol metabolism is the same in males as it is in female
- This is an interesting hypothesis to test
- The test of $H_0 : \beta_0 = 0$, testing whether alcohol metabolism in females is 0, on the other hand, is probably meaningless

Confidence intervals

- Recall where all of the items in the standard output come from: the estimate comes from $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, $\widehat{\mathrm{SE}}$ comes from the square root of the diagonal of $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$, the test statistic is $\hat{\beta}/\widehat{\mathrm{SE}}$, and the p-value comes from the t distribution with $n-p$ degrees of freedom
- Now suppose we want confidence intervals, which are not reported by default; we could compute them by hand:

$$\left[\hat{\beta} - t_{\alpha/2,n-p}\widehat{\mathrm{SE}}, \hat{\beta} + t_{\alpha/2,n-p}\widehat{\mathrm{SE}}\right]$$

- In SAS, we could add a `CLB` option to the `MODEL` statement
- In `R`, we could use the `confint` function, as in

```
fit <- lm(Metabol~Sex)
confint(fit)
```

The t test is a special case of linear regression

- As I alluded to in our first lecture, t tests are a special case of the linear regression model
- Note that the two-sample Student's t test comparing men and women in terms of alcohol metabolism is the same test as $H_0 : \beta_1 = 0$; the two have the same test statistic, degrees of freedom, null distribution and p-value
- Note that it is only equivalent to Student's t test, the "equal variance" t test; our model assumptions did not allow σ^2 to vary depending on the covariates

Reparameterizing our model

- Finally, let's explore what happens when we reparameterize our model as

$$\mathrm{E}(\texttt{Metabol}) = \beta_1 \texttt{Female} + \beta_2 \texttt{Male},$$

- This model is equivalent to the first one, in the sense that all of its residuals and fitted values are the same, but the coefficients (and thus, the default hypothesis tests) don't have the same meaning
- The coefficients are perhaps now easier to interpret, but neither default hypothesis test is meaningful
- In order to test for or estimate the difference between the two groups, we would need to set up the linear combination $\boldsymbol{\lambda}^T \boldsymbol{\beta}$, where $\boldsymbol{\lambda}^T = (-1, 1)$

R^2 for models without an intercept

- One additional note about the reparameterization: note that the R^2 for the first parameterization was only 33%, while that for the second is 64%
- It is clearly illogical, however, to conclude that the second model fits the data better (since their fitted values are exactly the same)
- The reason for this apparent paradox is that our decomposition of the variances depended on the presence of an intercept; without it, R^2 is of questionable meaning
- So, beware of interpreting R^2 for models without an intercept

Gastric as the explanatory variable

- We can also fit one-variable regression models involving Gastric and Alcohol
- A point which may be obvious, but I want to make sure we're all clear on: for a numeric variable like Gastric, the coefficient refers to the effect of a one-unit increase
- Because of linearity, it doesn't matter where that one-unit increase occurs; the effect of moving from 1 to 2 is exactly the same as going from 4 to 5
- So, for example, if we were to center Gastric by subtracting off its mean, we would change the value of the intercept, but $\hat{\beta}_1$ (and associated inferences) would remain exactly the same

Recap

To briefly recap our univariate findings:

- Alcohol metabolism much higher in men than women
- Alcohol metabolism highly positively correlated with gastric alcohol dehydrogenase
- Alcohol metabolism lower in alcoholics, but this was not significant

Male + Alcoholic

- Let's move on to two-variable models, such as:

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Alcoholic},$$

- We are now estimating the effect of alcoholism while controlling for sex, and vice versa
- Note that alcoholics have an alcohol metabolism 0.65 units lower than non-alcoholics, but this estimate jumps to 1.5 if we compare alcoholics to non-alcoholics while holding sex constant
- This is because alcoholics are more likely to be men, which confounded the univariate analysis, resulting a biased underestimate
- What does the intercept mean now?

Male + Gastric

- Another two-variable model:

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Gastric},$$

- Note that males had a metabolism 3 units higher than females, but that this estimate drops to 1.6 if we compare males to females who have the same levels of gastric alcohol dehydrogenase
- This is because men tend to have higher alcohol dehydrogenase activity, which again confounded the univariate analysis and led to systematic bias, although in this case, it led to overestimation
- What does the intercept mean now?

A key assumption

- Now, both of these models make a rather important kind of assumption
- Our first model assumed that the effect of alcoholism is the same for men as it is for women
- Our second model assumed that the effect of gastric dehydrogenase was the same for men as it was for women
- These assumptions might be true, but obviously we have no guarantee of that

Interactions

- Suppose we wanted a more flexible model – one that allowed alcoholism to have one effect in men and a different effect in women
- We can achieve that by introducing a new variable called, say, `AlcoholicMale`, and then fitting the model

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Alcoholic} + \beta_3 \texttt{AlcoholicMale}$$

- In the lingo of regression modeling, this is called introducing an *interaction* between sex and alcoholism, with $\beta_3\texttt{AlcoholicMale}$ said to be the *interaction term*

Interaction terms in R/SAS

- Note that we don't really need a new variable to represent `AlcoholicMale`
- If we have the indicator variables `Male` and `Alcoholic`, then

$$\texttt{AlcoholicMale} = \texttt{Alcoholic} \cdot \texttt{Male}$$

- Note that `R` and `PROC GLM` will allow you to include mathematical operations (logarithms, multiplication, etc.) in the model statement, whereas `PROC REG` will not

Interaction terms in R/SAS (cont'd)

- Thus, if you're using `PROC GLM`, you can specify the model with

  `MODEL Metabol = Male Alcohol Male*Alcohol;`

  or alternatively, `Male|Alcohol`, which includes both the interaction and the *main effects*
- If you're using `PROC REG`, you have to specify everything manually
- In `R`, `Male*Alcohol` includes both interactions and main effects; to specifically request only the interaction, use `Male:Alcohol`

- Let's think about the model and what the coefficients mean:

$$\mathrm{E}(\texttt{Metabol}) = \begin{cases} \beta_0 & \text{Female non-alcoholics} \\ \beta_0 + \beta_1 & \text{Male non-alcoholics} \\ \beta_0 + \beta_2 & \text{Female alcoholics} \\ \beta_0 + \beta_1 + \beta_2 + \beta_3 & \text{Male alcoholics} \end{cases}$$

- Thus,
  - β_0 is the expected alcohol metabolism for female non-alcoholics
  - β_1 is the difference in expected alcohol metabolism between male and female non-alcoholics
  - β_2 is the difference in expected alcohol metabolism between alcoholic and non-alcoholic females
  - β_3 is how much higher the difference in expected alcohol metabolism between alcoholic and non-alcoholic males is than the same difference for females

Default tests: interpretation

- From the default output, we can see that $H_0 : \beta_1 = 0$ looks pretty doubtful, but that $H_0 : \beta_2 = 0$ and $H_0 : \beta_3 = 0$ look plausible
- This model, then, supports the conclusion that male non-alcoholics have a significantly higher alcohol metabolism than female non-alcoholics, but finds no evidence that the metabolism of alcoholic females differs from that of non-alcoholic females, nor evidence of a bigger difference between alcoholic and non-alcoholic males than there is between alcoholic and non-alcoholic females
- However, it's important to note that there are additional interesting tests here

Additional tests

- For example, do male alcoholics have a significantly higher alcohol metabolism than female alcoholics?
- Mathematically, we want to test $H_0 : \beta_1 + \beta_3 = 0$
- Thus, $\boldsymbol{\lambda}^T = (0, 1, 0, 1)$, and we obtain $p = .27$
- Do alcoholic males have a significantly lower metabolism than non-alcoholic males?
- Perhaps $(p = .051)$

The easy way out

- The lazy statistician's way of getting those same estimates and tests is to change the reference category and refit the model
- Note that if we fit the model with `Female` and `Nonalcoholc` as explanatory variables, we get our $p = .051$ and $p = .27$ by default

Male*Gastric

- What about our other two-variable model, with `Sex` and `Gastric`?
- Do interaction terms mean anything here?

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Gastric} + \beta_3 \texttt{MaleGastric}$$

- Indeed they do: they allow the effect of gastric alcohol dehydrogenase to differ by sex

Interpreting the parameters

- Thinking about the coefficients:

$$\mathrm{E}(\texttt{Metabol}) = \begin{cases} \beta_0 + \beta_2 \texttt{Gastric} & \text{Females} \\ \beta_0 + \beta_1 + (\beta_2 + \beta_3)\texttt{Gastric} & \text{Males} \end{cases}$$

- In other words, we are fitting different lines, with separate slopes and intercepts, for each sex
- What would this model mean?

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_2 \texttt{Gastric} + \beta_3 \texttt{MaleGastric}$$

- This one?

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_3 \texttt{MaleGastric}$$

Trellis plots - R

- It is often useful to visualize these separate regression lines; one way to do this is with what are called *trellis plots*
- In R, the `lattice` package provides these sorts of plots
- The basic function in the `lattice` package is `xyplot`, whose syntax works like `y~x|z`, meaning plot `y` versus `x`, conditioning on `z`
- So for example:

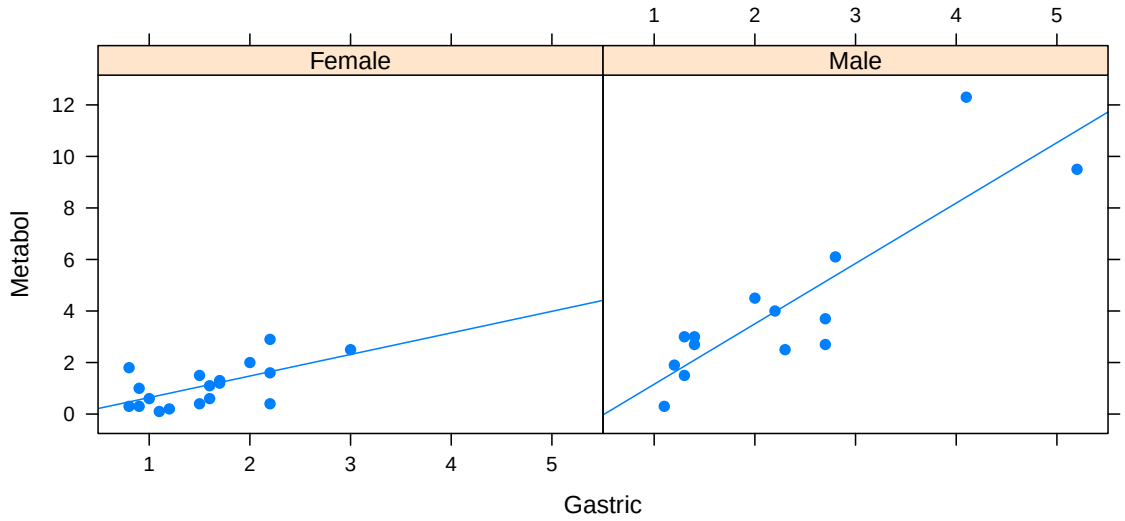
```
require(lattice)
xyplot(Metabol~Gastric|Sex)
```

Trellis plots - R (cont'd)

- The `type` option controls what is plotted in each panel, and accepts multiple arguments
- So, for example,

  `xyplot(Metabol~Gastric|Sex,type=c("p","r"))`

  plots both points and a regression line



Trellis plots - SAS

- In SAS, one can obtain such plots with `PROC SGPANEL`, which requires a plotting statement like you would find in `PROC SGPLOT` and also a `PANELBY` statement, which sets up the panels
- So for example:

```
PROC SGPANEL DATA=alcohol;
  PANELBY Sex;
  REG Y=Metabol X=Gastric;
RUN;
```

Interpreting main effects

- The `Gastric-Sex` interaction model has a significant interaction term, which suggests that our earlier models may have been too simplistic
- If the effect of alcohol dehydrogenase depends on sex, then it's impossible to consider the effect of one without the other
- For example, what conclusions should we draw about the fact that the main effect of `Sex` is no longer significant in our model?
- Can we conclude that there is no difference in alcohol metabolism between males and females once we have adjusted for the effect of alcohol dehydrogenase?

Interpreting main effects (cont'd)

- No; this test only compares the two at the specific alcohol dehydrogenase activity level of 0
- If we compared males and females at a different level of alcohol dehydrogenase – still holding it constant across the comparison, just not constant at 0 – we might get a significant result
- For example, if we center `Gastric` in our model, the main effect of `Sex` is now highly significant ($p = .0006$)

Two-variable recap

To recap, some salient conclusions from our two-variable models are:

- Alcoholism is associated with sex, so any conclusion about one, if the other one hasn't been adjusted for, is subject to confounding
- The effect of alcohol dehydrogenase on alcohol metabolism seems to depend on sex

Male+Alcoholic+Gastric

- Moving on to three-variable models, the simplest one is

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Alcoholic} + \beta_3 \texttt{Gastric}$$

- This model would indicate that males have higher alcohol metabolism than females, as do people with higher levels of alcohol dehydrogenase
- Furthermore, it suggests that there is very little effect of alcoholism on alcohol metabolism after adjusting for sex and enzymatic activity

Including an interaction

- However, we have reason to be dissatisfied with this model, as our earlier modeling indicated that the relationship between `Gastric` and `Metabol` depends on sex
- Thus, it would seem prudent to investigate this model:

$$\begin{aligned}\mathrm{E}(\texttt{Metabol}) =& \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Alcoholic} + \beta_3 \texttt{Gastric} \\ &+ \beta_4 \texttt{MaleGastric}\end{aligned}$$

- This allows a different `Metabol`-`Gastric` regression line for each sex, while controlling for alcoholism

Interpretation

This model indicates that

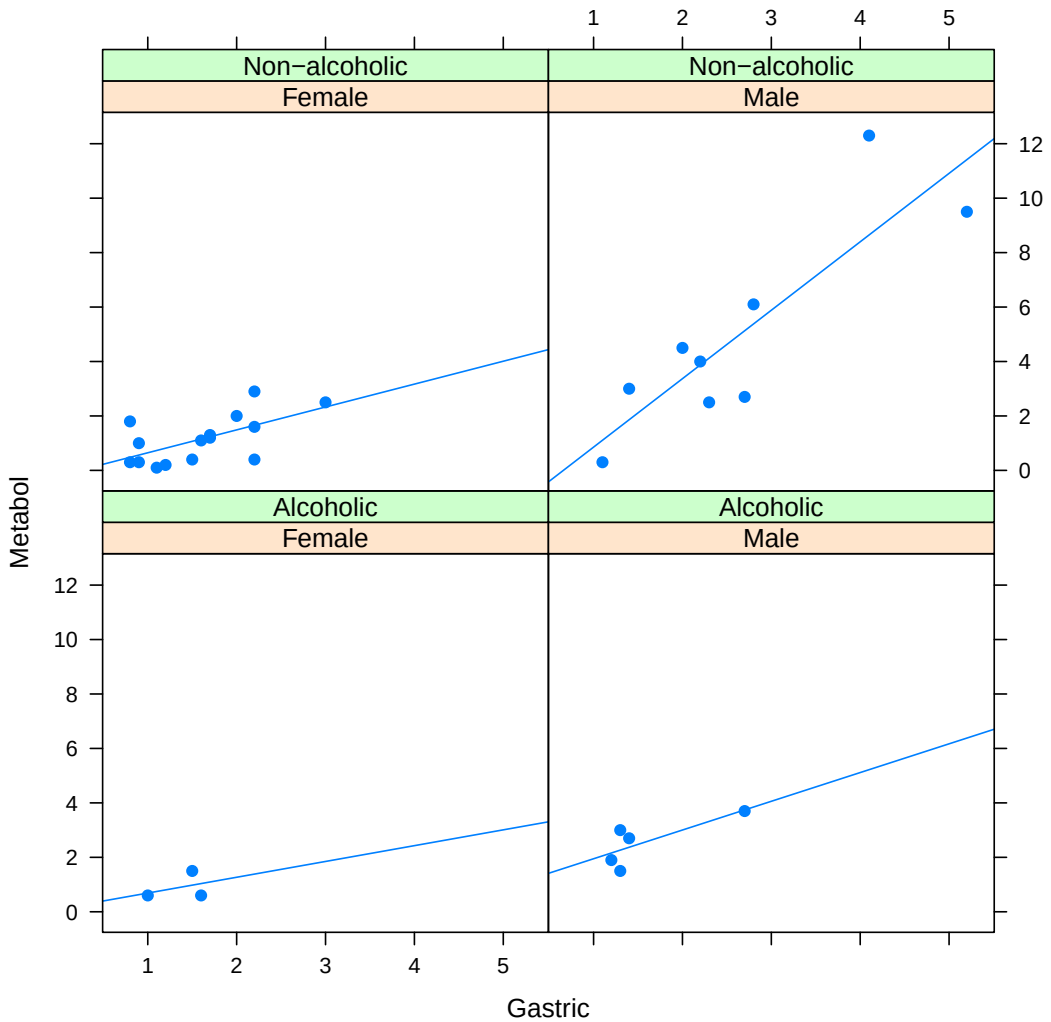
- There is a very strong effect of alcohol dehydrogenase in males
- There is a much less strong effect of alcohol dehydrogenase in females – indeed, we cannot even rule out that alcohol dehydrogenase has no effect in females
- At low levels of alcohol dehydrogenase, men and women have similar alcohol metabolism; at higher levels, however, men have much higher alcohol metabolism
- There is no evidence that alcoholism plays an important role

Assumptions

- However, let's think about what assumptions and restrictions are imposed by the preceding model
- It allows different slopes for each sex, but the slope for male alcoholics is assumed to be the same as the slope for male non-alcoholics
- Is this true?

Plotting the four lines

A two-way trellis plot suggests that this is not necessarily true:



Male*Alcoholic*Gastric

- So let's try fitting a model that allows for a different regression line for each combination of sex and alcoholism status:

$$\begin{aligned}\mathrm{E}(\texttt{Metabol}) =& \beta_0 + \beta_1\texttt{Male} + \beta_2\texttt{Alcoholic} + \beta_3\texttt{Gastric} \\ &+ \beta_4\texttt{AlcoholicMale} + \beta_5\texttt{AlcoholicGastric} \\ &+ \beta_6\texttt{MaleGastric} + \beta_7\texttt{AlcoholicMaleGastric}\end{aligned}$$

- Note that this model involves a term that is the product of three variables; this is a so-called *three-way interaction* model

Remarks

- Note that the number of parameters add up to what it should: we need an intercept and a slope for every combination of sex and alcoholism, so we need $2 \times 4 = 8$ parameters, and that's exactly how many coefficients we have
- You can specify this model manually, with `Alcoholic*Male*Gastric` in `R` or with `Alcoholic|Male|Gastric` in `PROC GLM`
- Note that the above constructions automatically include all possible lower-order interactions (in this case, all the two-way interactions)

Interpreting the parameters

Let's think about what the regression lines look like in the four different groups:

Group	Intercept	Slope
Female non-alcoholics	β_0	β_3
Male non-alcoholics	$\beta_0 + \beta_2$	$\beta_3 + \beta_6$
Female alcoholics	$\beta_0 + \beta_1$	$\beta_3 + \beta_5$
Male alcoholics	$\beta_0 + \beta_1 + \beta_2 + \beta_4$	$\beta_3 + \beta_5 + \beta_6 + \beta_7$

Inferences

- If we carry out comparisons of slopes between groups, we find that:
  - There is significant evidence of a difference in slopes between male and female non-alcoholics
  - There is no evidence of any difference in slope between alcoholic and non-alcoholic females, between male and female alcoholics, or between alcoholic and non-alcoholic males (although there is a fairly large estimated difference in this last comparison)
- However, the test of the three-way interaction was not significant, suggesting that we might not need different slopes depending on alcoholism status

Concluding remarks

- We've ran through a spectrum today from very simple models that were equivalent to the two-sample t-test, to rather complicated multiparameter models with a zoo of interaction terms
- Two concluding remarks:
  - Linear models are very flexible, and accommodate a wide range of complexity all in a single framework
  - How do we decide where in this spectrum the most appropriate model lies? This is where we will direct our attention next, as we start discussing the art of model selection and diagnostics