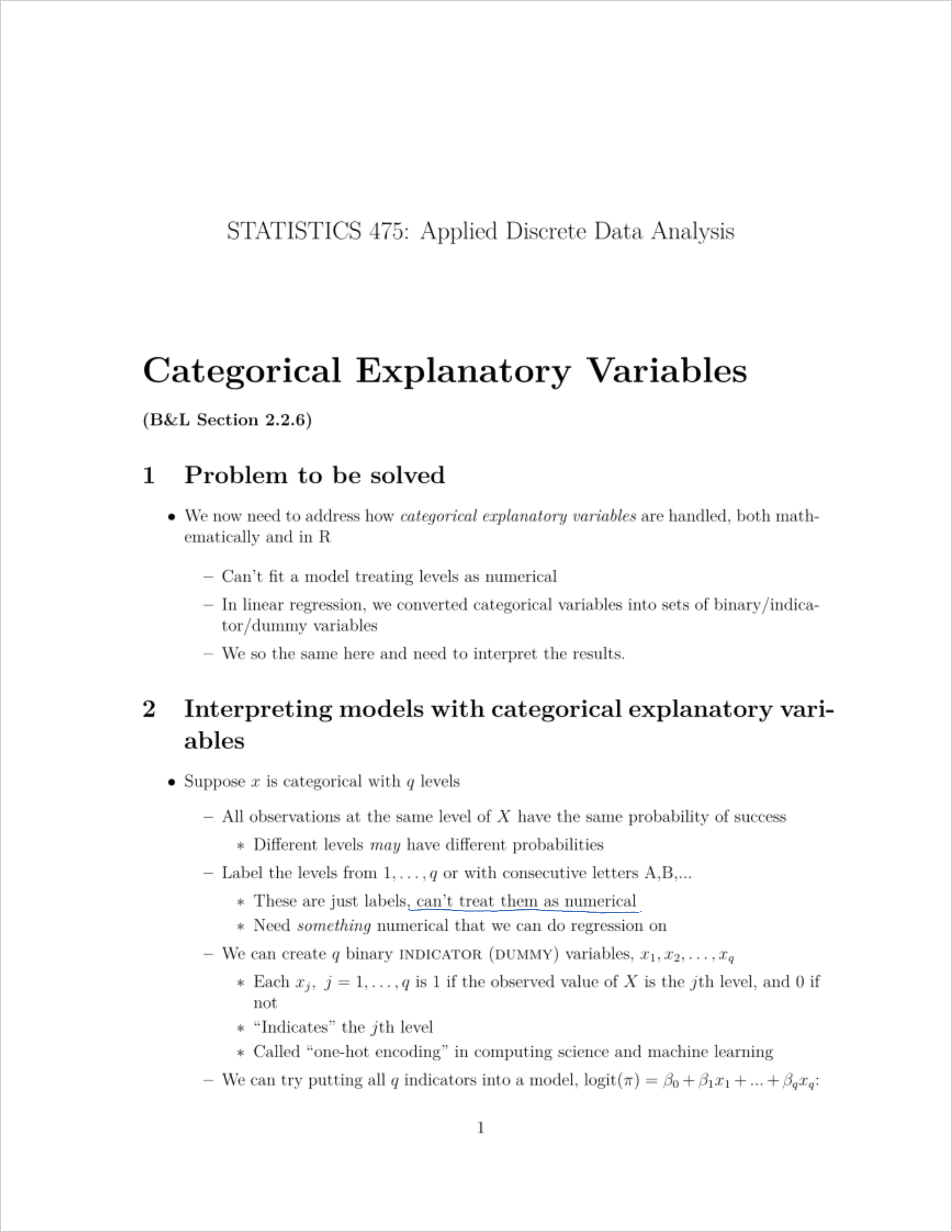
Lecture 11

February 5, 2019 10:57 AM



STATISTICS 475: Applied Discrete Data Analysis

Categorical Explanatory Variables

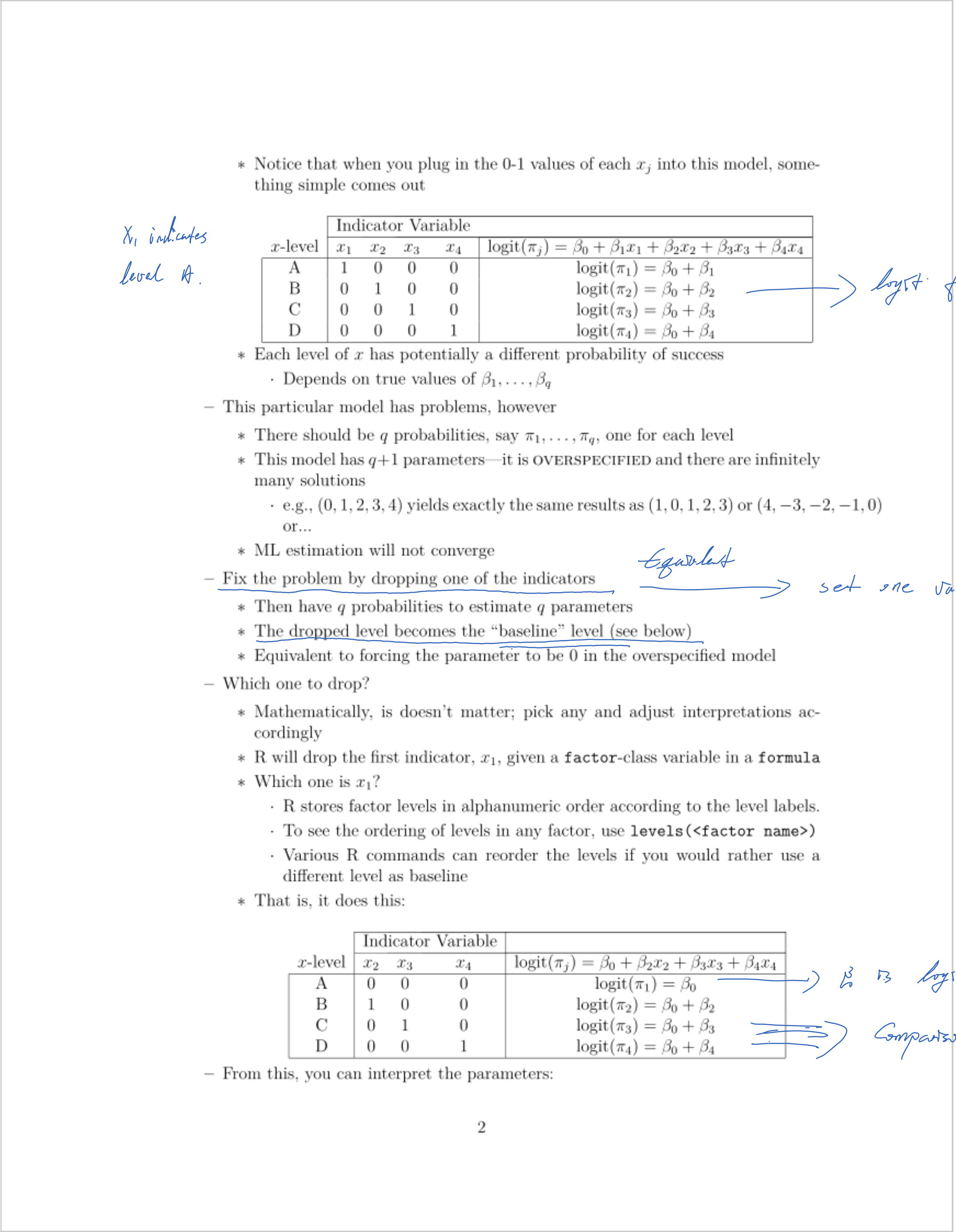
(B&L Section 2.2.6)

1 Problem to be solved

- We now need to address how *categorical explanatory variables* are handled, both mathematically and in R
  - Can't fit a model treating levels as numerical
  - In linear regression, we converted categorical variables into sets of binary/indicator/dummy variables
  - We so the same here and need to interpret the results.

2 Interpreting models with categorical explanatory variables

- Suppose x is categorical with q levels
  - All observations at the same level of X have the same probability of success
    - Different levels *may* have different probabilities
  - Label the levels from $1, \ldots, q$ or with consecutive letters A,B,...
    - These are just labels, can't treat them as numerical
    - Need *something* numerical that we can do regression on
  - We can create q binary INDICATOR (DUMMY) variables, $x_1, x_2, \ldots, x_q$
    - Each $x_j,\ j = 1, \ldots, q$ is 1 if the observed value of X is the jth level, and 0 if not
    - "Indicates" the jth level
    - Called "one-hot encoding" in computing science and machine learning
  - We can try putting all q indicators into a model, $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$:



- Notice that when you plug in the 0-1 values of each x_j into this model, something simple comes out

x_1 indicates level A.

	Indicator Variable				
x-level	x_1	x_2	x_3	x_4	$\text{logit}(\pi_j) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	0	$\text{logit}(\pi_1) = \beta_0 + \beta_1$
B	0	1	0	0	$\text{logit}(\pi_2) = \beta_0 + \beta_2$
C	0	0	1	0	$\text{logit}(\pi_3) = \beta_0 + \beta_3$
D	0	0	0	1	$\text{logit}(\pi_4) = \beta_0 + \beta_4$

logit functions are constants.

- Each level of x has potentially a different probability of success
  - Depends on true values of $\beta_1, \ldots, \beta_q$

– This particular model has problems, however

- There should be q probabilities, say $\pi_1, \ldots, \pi_q$, one for each level
- This model has $q+1$ parameters—it is OVERSPECIFIED and there are infinitely many solutions
  - e.g., $(0, 1, 2, 3, 4)$ yields exactly the same results as $(1, 0, 1, 2, 3)$ or $(4, -3, -2, -1, 0)$ or...
- ML estimation will not converge

– Fix the problem by dropping one of the indicators

Equivalent → set one variable to 0

- Then have q probabilities to estimate q parameters
- The dropped level becomes the "baseline" level (see below)
- Equivalent to forcing the parameter to be 0 in the overspecified model

– Which one to drop?

- Mathematically, is doesn't matter; pick any and adjust interpretations accordingly
- R will drop the first indicator, x_1, given a `factor`-class variable in a `formula`
- Which one is x_1?
  - R stores factor levels in alphanumeric order according to the level labels.
  - To see the ordering of levels in any factor, use `levels(<factor name>)`
  - Various R commands can reorder the levels if you would rather use a different level as baseline
- That is, it does this:

	Indicator Variable			
x-level	x_2	x_3	x_4	$\text{logit}(\pi_j) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	0	0	0	$\text{logit}(\pi_1) = \beta_0$
B	1	0	0	$\text{logit}(\pi_2) = \beta_0 + \beta_2$
C	0	1	0	$\text{logit}(\pi_3) = \beta_0 + \beta_3$
D	0	0	1	$\text{logit}(\pi_4) = \beta_0 + \beta_4$

β_0 is logit for π_1

Comparisons to the baseline

– From this, you can interpret the parameters:

    - $\ast$ β_0 is the log-odds of success at the first level of X
      - · This is the "baseline" level against which others are compared.
    - $\ast$ For $j = 2, \dots, q$, $\beta_j = \text{logit}(\pi_j) - \text{logit}(\pi_1)$,
      - · This is the log(OR) between levels j and 1!
  - For convenience of notation, we can sometimes pretend that we are using the original model including a $\beta_1 x_1$ term, and that $\beta_1 = 0$.
  - Note that the book uses $x_1, \dots, x_{q-1}$ instead of $x_2, \dots, x_q$
    - $\ast$ It doesn't matter what you call them, as long as you remember that they represent

2.1 Odds Ratios

- We can form odds ratios *within* a categorical explanatory by comparing the odds of success at level a to the odds of success at level b:

$$\frac{\text{Odds}_a}{\text{Odds}_b} = \frac{\exp(\beta_0 + \beta_a)}{\exp(\beta_0 + \beta_b)} = \exp(\beta_a - \beta_b)$$

  where β_1 is taken to be 0 when level 1 is involved in an OR

  - *So odds ratios are just exponentiated differences between the numerator- and denominator-level parameter combinations*
  - (We have seen this before.)

- We already know how to do Wald and LR inferences on parameters of this form!
  - Work in the logit scale, taking differences of log-odds as a log(OR)
  - Estimate with $\hat{\beta}_a - \hat{\beta}_b$, which uses a normal approximation in large samples
  - Wald or LR test for $H_0 : OR = 1$ tests $H_0 : \beta_a - \beta_b = 0$
  - Get $\sqrt{\widehat{Var}(\hat{\beta}_a - \hat{\beta}_b)}$ mathematically and computationally for Wald
  - LR CI through `mcprofile` uses a coefficient matrix where a row contains a 1 in the ath position, a -1 in the bth position, and a 0 everywhere else
    - $\ast$ Unless level 1 is one of a or b. In that case, there is no coefficient for that level, because R has removed it from the model
    - $\ast$ e.g.,

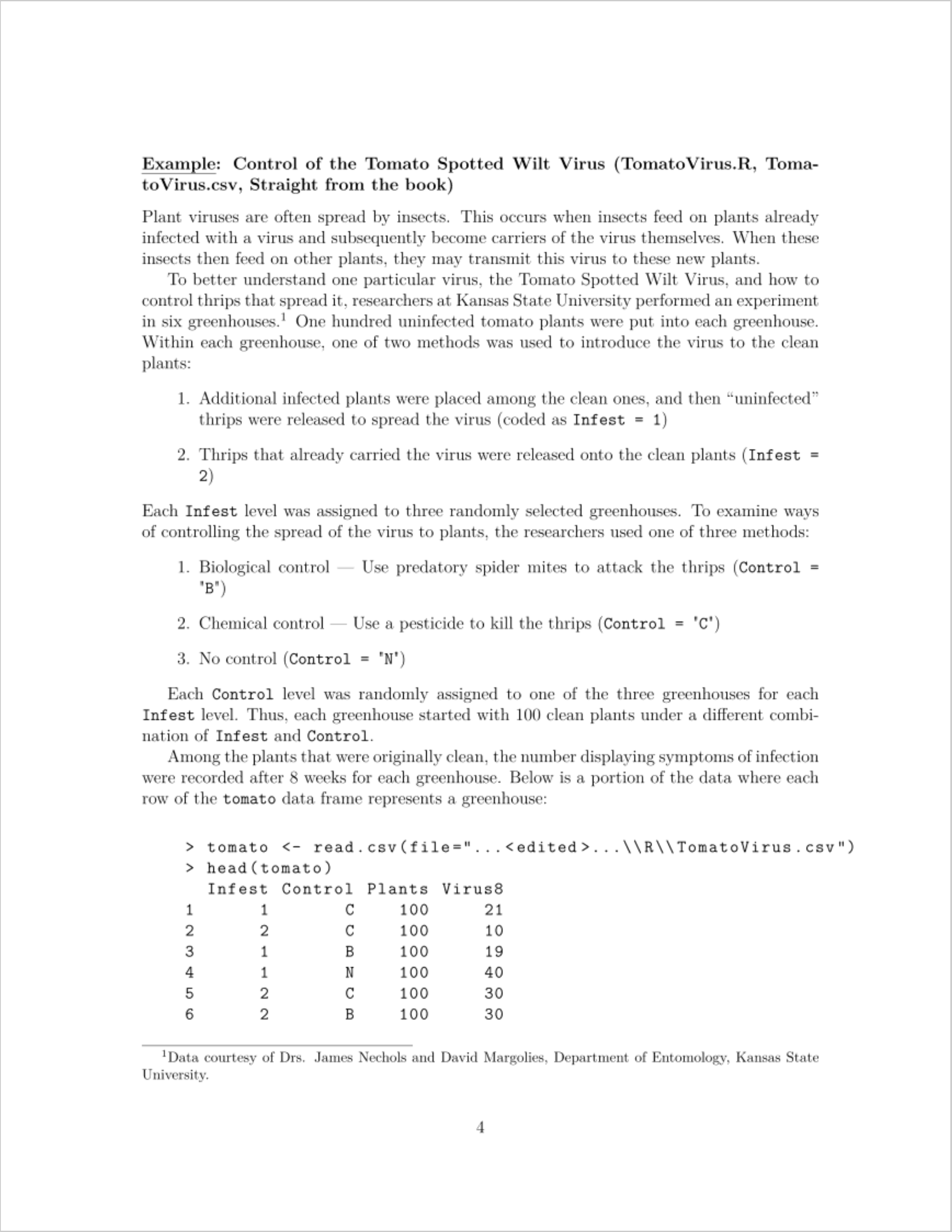
$$\frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0 + \beta_2)} = \frac{\exp(\beta_0)}{\exp(\beta_0 + \beta_2)} = \exp(-\beta_{2})$$

  - Exponentiate endpoints of the resulting confidence intervals

- Very often test for a "variable X effect": do any of the levels of X have different probabilities of success?
  - $H_0 : (\beta_1 =)\beta_2 = \beta_3 = \dots = \beta_q = 0$ versus $H_a :$ Not all $\beta_j = 0$

is the baseline, already 0.

  - LR model comparison test
    - $\ast$ How many df? $q - 1$



Example: Control of the Tomato Spotted Wilt Virus (TomatoVirus.R, TomatoVirus.csv, Straight from the book)

Plant viruses are often spread by insects. This occurs when insects feed on plants already infected with a virus and subsequently become carriers of the virus themselves. When these insects then feed on other plants, they may transmit this virus to these new plants.

To better understand one particular virus, the Tomato Spotted Wilt Virus, and how to control thrips that spread it, researchers at Kansas State University performed an experiment in six greenhouses.[1] One hundred uninfected tomato plants were put into each greenhouse. Within each greenhouse, one of two methods was used to introduce the virus to the clean plants:

1. Additional infected plants were placed among the clean ones, and then "uninfected" thrips were released to spread the virus (coded as `Infest = 1`)

2. Thrips that already carried the virus were released onto the clean plants (`Infest = 2`)

Each `Infest` level was assigned to three randomly selected greenhouses. To examine ways of controlling the spread of the virus to plants, the researchers used one of three methods:

1. Biological control — Use predatory spider mites to attack the thrips (`Control = "B"`)

2. Chemical control — Use a pesticide to kill the thrips (`Control = 'C'`)

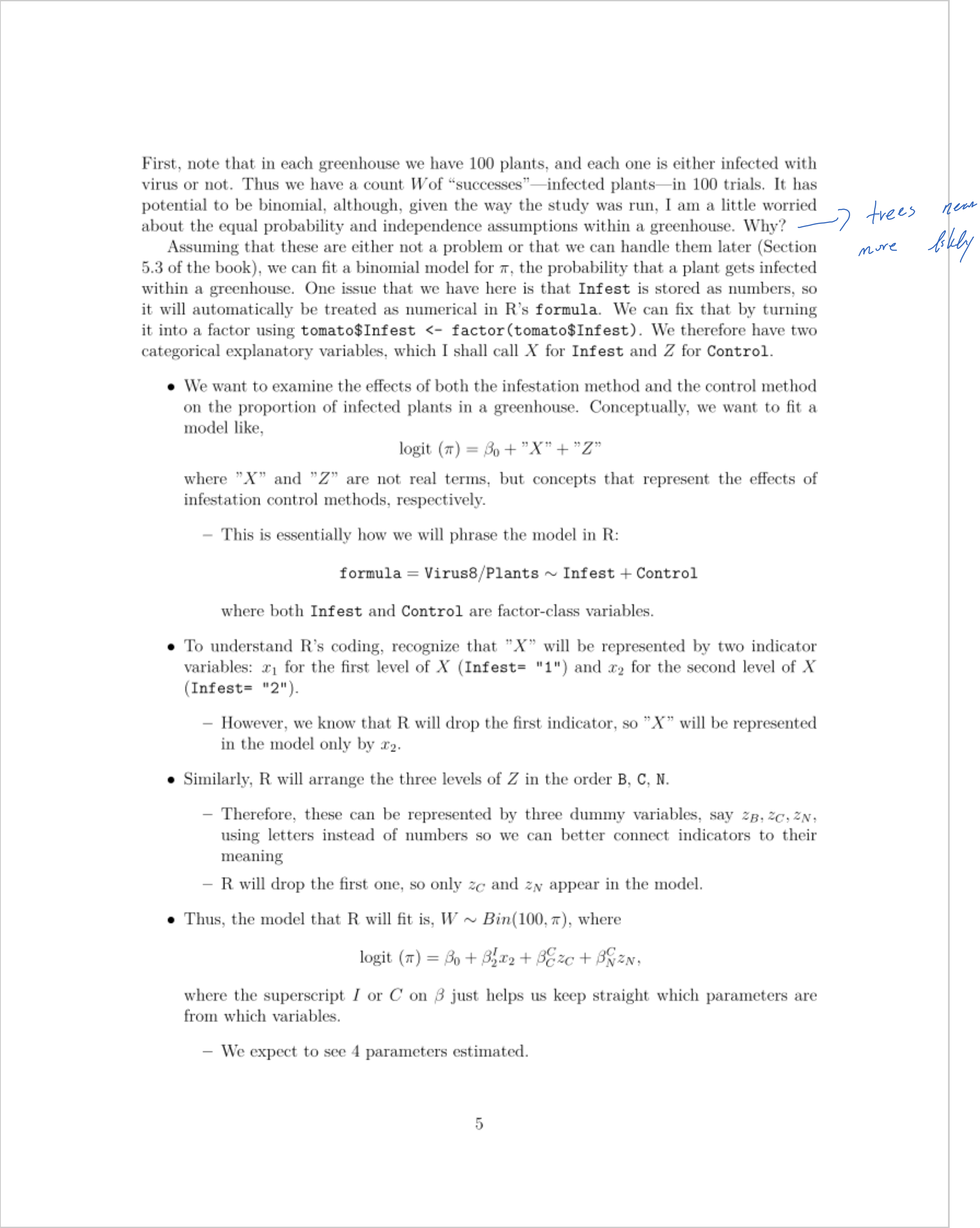
3. No control (`Control = 'N'`)

Each `Control` level was randomly assigned to one of the three greenhouses for each `Infest` level. Thus, each greenhouse started with 100 clean plants under a different combination of `Infest` and `Control`.

Among the plants that were originally clean, the number displaying symptoms of infection were recorded after 8 weeks for each greenhouse. Below is a portion of the data where each row of the `tomato` data frame represents a greenhouse:

```
> tomato <- read.csv(file="...<edited>...\\R\\TomatoVirus.csv")
> head(tomato)
  Infest Control Plants Virus8
1      1       C    100     21
2      2       C    100     10
3      1       B    100     19
4      1       N    100     40
5      2       C    100     30
6      2       B    100     30
```

[1] Data courtesy of Drs. James Nechols and David Margolies, Department of Entomology, Kansas State University.



First, note that in each greenhouse we have 100 plants, and each one is either infected with virus or not. Thus we have a count $W$of "successes"—infected plants—in 100 trials. It has potential to be binomial, although, given the way the study was run, I am a little worried about the equal probability and independence assumptions within a greenhouse. Why?

→ trees near the Infected trees are more likely to be infected.

Assuming that these are either not a problem or that we can handle them later (Section 5.3 of the book), we can fit a binomial model for π, the probability that a plant gets infected within a greenhouse. One issue that we have here is that `Infest` is stored as numbers, so it will automatically be treated as numerical in R's `formula`. We can fix that by turning it into a factor using `tomato$Infest <- factor(tomato$Infest)`. We therefore have two categorical explanatory variables, which I shall call X for `Infest` and Z for `Control`.

- We want to examine the effects of both the infestation method and the control method on the proportion of infected plants in a greenhouse. Conceptually, we want to fit a model like,

$$\text{logit }(\pi) = \beta_0 + "X" + "Z"$$

  where $"X"$ and $"Z"$ are not real terms, but concepts that represent the effects of infestation control methods, respectively.

  - This is essentially how we will phrase the model in R:

$$\texttt{formula} = \texttt{Virus8/Plants} \sim \texttt{Infest} + \texttt{Control}$$

    where both `Infest` and `Control` are factor-class variables.

- To understand R's coding, recognize that $"X"$ will be represented by two indicator variables: x_1 for the first level of X (`Infest= "1"`) and x_2 for the second level of X (`Infest= "2"`).

  - However, we know that R will drop the first indicator, so $"X"$ will be represented in the model only by x_2.

- Similarly, R will arrange the three levels of Z in the order `B`, `C`, `N`.

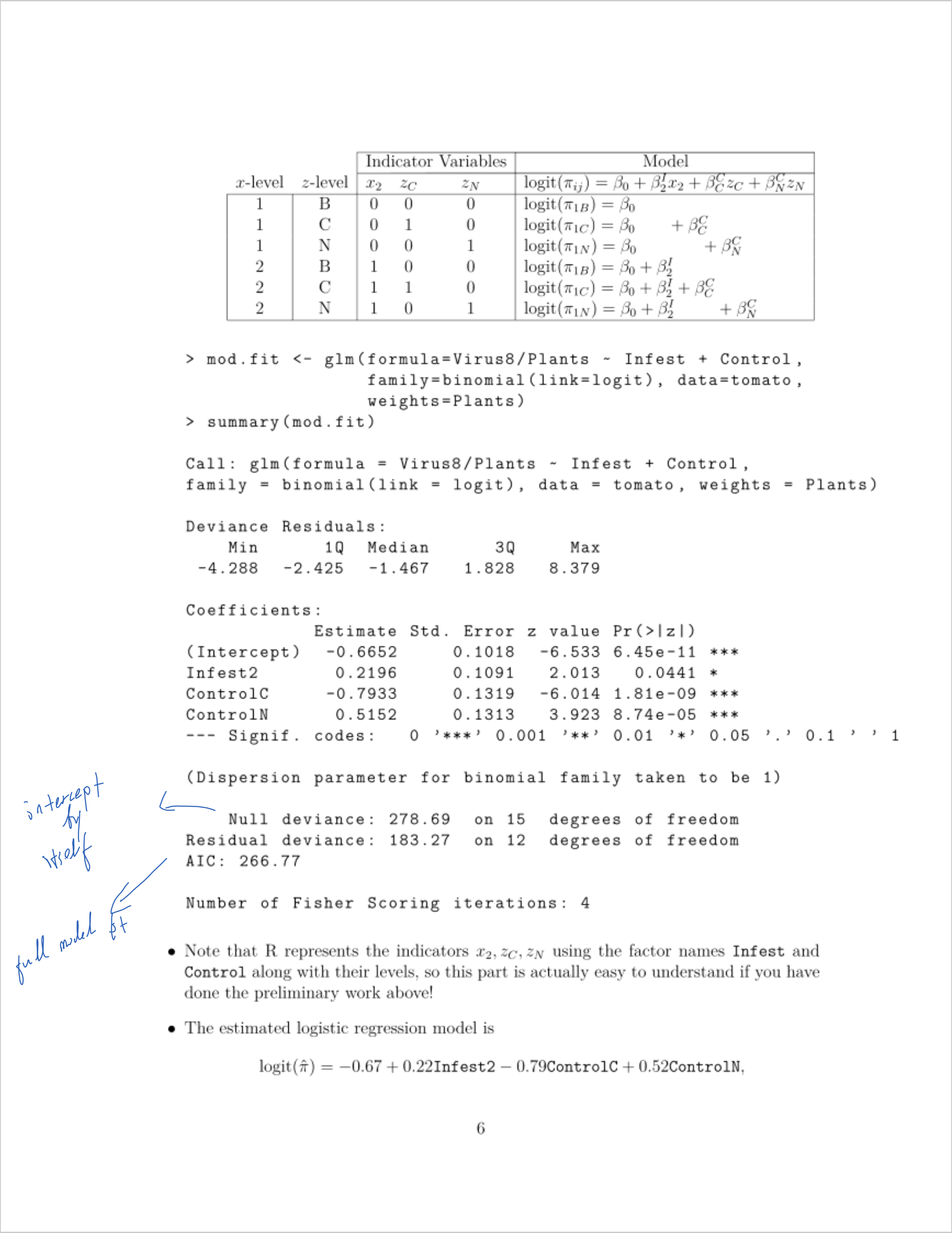
  - Therefore, these can be represented by three dummy variables, say z_B, z_C, z_N, using letters instead of numbers so we can better connect indicators to their meaning
  - R will drop the first one, so only z_C and z_N appear in the model.

- Thus, the model that R will fit is, $W \sim Bin(100, \pi)$, where

$$\text{logit }(\pi) = \beta_0 + \beta_2^I x_2 + \beta_C^C z_C + \beta_N^C z_N,$$

  where the superscript I or C on β just helps us keep straight which parameters are from which variables.

  - We expect to see 4 parameters estimated.



		Indicator Variables			Model
x-level	z-level	x_2	z_C	z_N	$\text{logit}(\pi_{ij}) = \beta_0 + \beta_2^I x_2 + \beta_C^C z_C + \beta_N^C z_N$
1	B	0	0	0	$\text{logit}(\pi_{1B}) = \beta_0$
1	C	0	1	0	$\text{logit}(\pi_{1C}) = \beta_0 \quad + \beta_C^C$
1	N	0	0	1	$\text{logit}(\pi_{1N}) = \beta_0 \quad\quad + \beta_N^C$
2	B	1	0	0	$\text{logit}(\pi_{1B}) = \beta_0 + \beta_2^I$
2	C	1	1	0	$\text{logit}(\pi_{1C}) = \beta_0 + \beta_2^I + \beta_C^C$
2	N	1	0	1	$\text{logit}(\pi_{1N}) = \beta_0 + \beta_2^I \quad + \beta_N^C$

```
> mod.fit <- glm(formula=Virus8/Plants ~ Infest + Control,
                 family=binomial(link=logit), data=tomato,
                 weights=Plants)
> summary(mod.fit)

Call: glm(formula = Virus8/Plants ~ Infest + Control,
family = binomial(link = logit), data = tomato, weights = Plants)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
 -4.288   -2.425   -1.467    1.828    8.379

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.6652     0.1018  -6.533 6.45e-11 ***
Infest2        0.2196     0.1091   2.013   0.0441 *
ControlC      -0.7933     0.1319  -6.014 1.81e-09 ***
ControlN       0.5152     0.1313   3.923 8.74e-05 ***
--- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 278.69  on 15  degrees of freedom
Residual deviance: 183.27  on 12  degrees of freedom
AIC: 266.77

Number of Fisher Scoring iterations: 4
```

intercept by itself

full model fit

- Note that R represents the indicators x_2, z_C, z_N using the factor names `Infest` and `Control` along with their levels, so this part is actually easy to understand if you have done the preliminary work above!

- The estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = -0.67 + 0.22\texttt{Infest2} - 0.79\texttt{ControlC} + 0.52\texttt{ControlN},$$

- $\hat{\beta}_0 = -0.67$, so the log odds of a virus infection with `Infest=1` (release uninfected thrips and scatter infected plants) and `Control="B"` (biological control with spider mites) are -0.67
  - ∗ Correspondingly, the probability of virus infection is $\exp(-0.67)/(1+\exp(-0.67)) = 0.34$ from `plogis(coef(mod.fit)[1])`
- $\hat{\beta}_2^I = 0.22$, so the log odds of virus infection are estimated to be 0.22 higher with `Infest` level 2 (release infected thrips on uninfected plants) than with level 1 (release uninfected thrips and scatter infected plants)
- $\hat{\beta}_2^C = -0.79$, so the log odds of virus infection with chemical control (level `C` of `Control`) are estimated to be 0.79 lower than with biological control (level `B` of `Control`).
- What about the interpretation of $\hat{\beta}_3^C$?

- Although I don't show them here, tests and confidence intervals could be computed in the usual way
  - Test for the effect of different methods of infestation or the effect of different types of control would be tested using model-comparison LR tests from `Anova()` in `car`.

---

3 Interactions with Categorical Variables

- Interactions among categorical variables are common (more so than with numerical variables)
  - e.g., does the comparison of biological control to chemical control change depending on the method used to create the infestation?
- Model interactions as cross-products among *all* indicators for the two variables involved
  - e.g., X with 3 levels and Z with 4 levels
    - ∗ Cross-products of x_2, x_3 with z_2, z_3, z_4
    - ∗ Model becomes

$$\begin{aligned}\text{logit}(\pi) &= \beta_0 + \beta_2^X x_2 + \beta_3^X x_3 + \beta_2^Z z_2 + \beta_3^Z z_3 + \beta_4^Z z_4 \\ &\quad + \beta_{22}^{XZ} x_2 z_2 + \beta_{23}^{XZ} x_2 z_3 + \beta_{24}^{XZ} x_2 z_4 + \beta_{32}^{XZ} x_3 z_2 + \beta_{33}^{XZ} x_3 z_3 + \beta_{34}^{XZ} x_3 z_4\end{aligned}$$

    - ∗ β^X and β^Z parameters are called MAIN EFFECT PARAMETERS
    - ∗ β^{XZ}parameters are INTERACTION PARAMETERS
- *To understand meaning of parameters, make a table* (here shown with 2 × 2 levels)

		Indicator Variables		Model
x-level	z-level	x_2	z_2	$\text{logit}(\pi_{ij}) = \beta_0 + \beta_2^X x_2 + \beta_2^Z z_2 + \beta_{22}^{XZ} x_2 z_2$
A	C	0	0	$\text{logit}(\pi_{00}) = \beta_0$
A	D	0	1	$\text{logit}(\pi_{01}) = \beta_0 + \beta_2^Z$
B	C	1	0	$\text{logit}(\pi_{10}) = \beta_0 + \beta_2^X$
B	D	1	1	$\text{logit}(\pi_{11}) = \beta_0 + \beta_2^X + \beta_2^Z + \beta_{22}^{XZ}$

→ compute single logit by taking difference

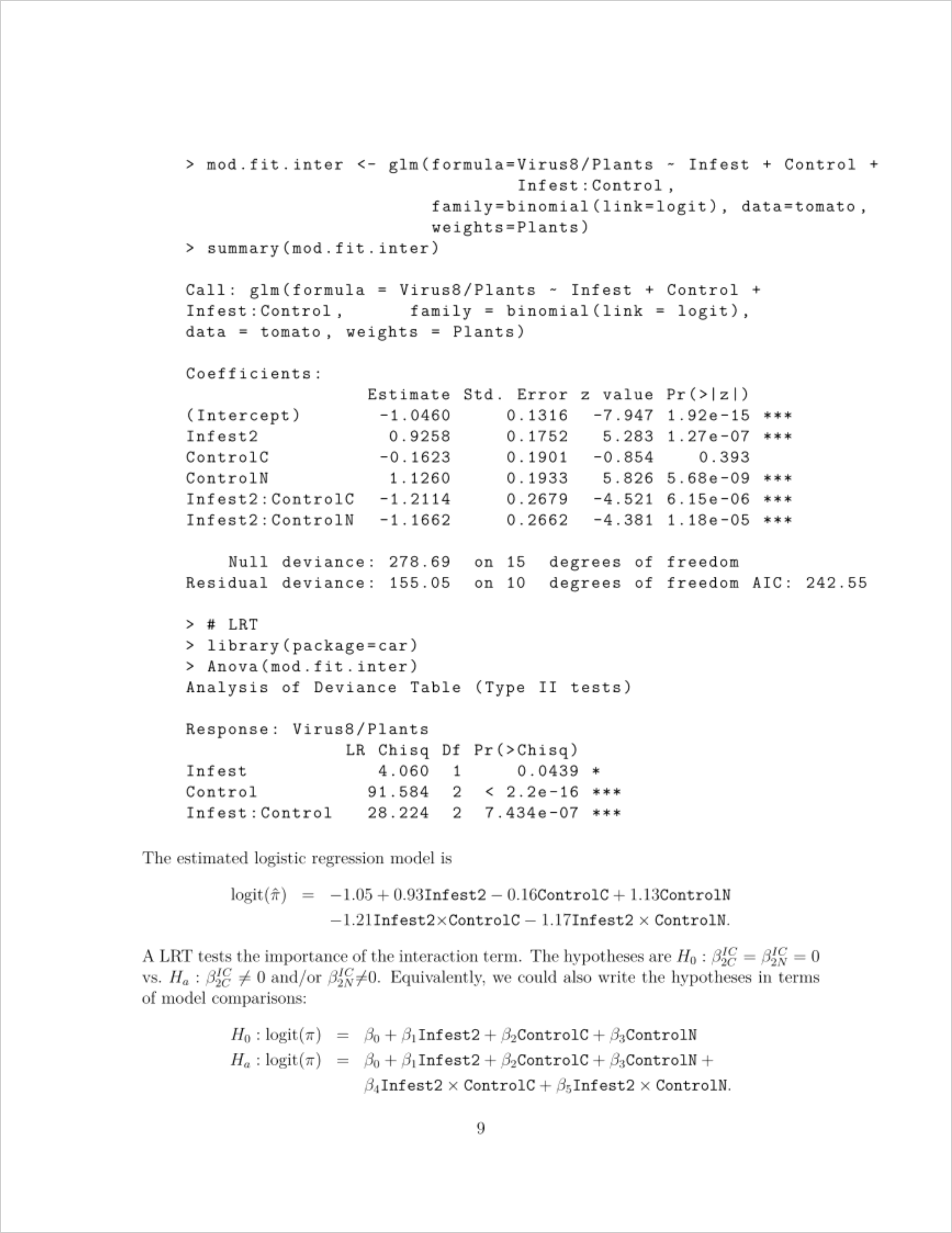
- From this, you can interpret the parameters:
  - β_0 is the log-odds of success at the first level of *both* x and z
  - The main-effect parameter, $\beta_2^X = \text{logit}(\pi_{10}) - \text{logit}(\pi_{00})$, is the log(OR) between levels B and A of x, *fixing z at its first level, C*
  - The main-effect parameter, $\beta_2^Z = \text{logit}(\pi_{01}) - \text{logit}(\pi_{00})$, is the log(OR) between levels D and C of z, *fixing x at its first level, A*
  - The interaction parameter, β_{22}^{XZ}, can be interpreted in a few equivalent ways:
    - *Always*, an interaction between two variables means that the effect of one variable changes depending on the level of the other variable
    - In the model, "effect" is measured by the log(OR) for two comparing two levels of one variable
    - Then an interaction term measures the difference between log(OR) (the "effect") of one variable at the two levels of the other variable
    - You can freely reverse which variable is which in this interpretation
    - $\beta_{22}^{XZ} = [\text{logit}(\pi_{11}) - \text{logit}(\pi_{10})] - [\text{logit}(\pi_{01}) - \text{logit}(\pi_{00})]$ is the difference between the log(OR) for Z when $X = B$ and the log(OR) for Z when $X = A$
    - $\beta_{22}^{XZ} = [\text{logit}(\pi_{11}) - \text{logit}(\pi_{01})] - [\text{logit}(\pi_{10}) - \text{logit}(\pi_{00})]$ is the difference between the log(OR) for X when $Z = D$ and the log(OR) for X when $Z = C$
- Higher order interactions (3 variables, 4 variables, etc) are possible, but become cumbersome to interpret

Example: Control of the Tomato Spotted Wilt Virus (TomatoVirus.R, TomatoVirus.csv)

I add interaction term to the model we fit previously to see whether the effect of the control method is associated with the infestation method.

- We have 2 new terms, $x_2 z_C$ and $x_2 z_N$
- This adds two parameters: β_{2C}^{IC} and β_{2N}^{IC} (show this in a table!).

The model fit is below (abridged output):



```
> mod.fit.inter <- glm(formula=Virus8/Plants ~ Infest + Control +
                                  Infest:Control,
                       family=binomial(link=logit), data=tomato,
                       weights=Plants)
> summary(mod.fit.inter)

Call: glm(formula = Virus8/Plants ~ Infest + Control +
Infest:Control,      family = binomial(link = logit),
data = tomato, weights = Plants)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.0460     0.1316  -7.947 1.92e-15 ***
Infest2            0.9258     0.1752   5.283 1.27e-07 ***
ControlC          -0.1623     0.1901  -0.854    0.393
ControlN           1.1260     0.1933   5.826 5.68e-09 ***
Infest2:ControlC  -1.2114     0.2679  -4.521 6.15e-06 ***
Infest2:ControlN  -1.1662     0.2662  -4.381 1.18e-05 ***

    Null deviance: 278.69  on 15  degrees of freedom
Residual deviance: 155.05  on 10  degrees of freedom AIC: 242.55

> # LRT
> library(package=car)
> Anova(mod.fit.inter)
Analysis of Deviance Table (Type II tests)

Response: Virus8/Plants
               LR Chisq Df Pr(>Chisq)
Infest            4.060  1     0.0439 *
Control          91.584  2  < 2.2e-16 ***
Infest:Control   28.224  2  7.434e-07 ***
```

The estimated logistic regression model is

$$\begin{aligned}\text{logit}(\hat{\pi}) &= -1.05 + 0.93\texttt{Infest2} - 0.16\texttt{ControlC} + 1.13\texttt{ControlN} \\ &\quad -1.21\texttt{Infest2}\times\texttt{ControlC} - 1.17\texttt{Infest2}\times\texttt{ControlN}.\end{aligned}$$

A LRT tests the importance of the interaction term. The hypotheses are $H_0 : \beta_{2C}^{IC} = \beta_{2N}^{IC} = 0$ vs. $H_a : \beta_{2C}^{IC} \neq 0$ and/or $\beta_{2N}^{IC}\neq 0$. Equivalently, we could also write the hypotheses in terms of model comparisons:

$$\begin{aligned}H_0 : \text{logit}(\pi) &= \beta_0 + \beta_1\texttt{Infest2} + \beta_2\texttt{ControlC} + \beta_3\texttt{ControlN} \\ H_a : \text{logit}(\pi) &= \beta_0 + \beta_1\texttt{Infest2} + \beta_2\texttt{ControlC} + \beta_3\texttt{ControlN} + \\ &\quad \beta_4\texttt{Infest2} \times \texttt{ControlC} + \beta_5\texttt{Infest2} \times \texttt{ControlN}.\end{aligned}$$

The test statistic is $-2\log(\Lambda) = 28.224$, and the p-value is 7.4×10^{-7} using a χ^2_2 approximation (why 2 df?). Thus, there is strong evidence of an interaction between the infestation and control methods.

3.1 Odds ratios with interactions

- An interaction in the model implies that the odds ratios that compare two levels of a given variable are not the same for all levels of the other variable
- Therefore, need to estimate odds ratios for two levels of X separately for each level of Z and vice-versa.
  - Write it out in terms of logits using π_{ij} to represent the probability of success with level i_1 of X and level i_2 of X at level j of Z:

$$\log(OR) = \log\left(\frac{odds_1}{odds_2}\right) = logit(\pi_1) - logit(\pi_2)$$

$$\begin{aligned}\text{Odds}_{x=i_1 z=j}/\text{Odds}_{x=i_2,z=j} &= \exp[\text{logit}(\pi_{i_1 j}) - \text{logit}(\pi_{i_2 j})] \\ = \exp[& (\cancel{\beta_0} + \beta^X_{i_1} + \cancel{\beta^Z_j} + \beta^{XZ}_{i_1 j}) - (\cancel{\beta_0} + \beta^X_{i_2} + \cancel{\beta^Z_j} + \beta^{XZ}_{i_2 j})] \\ &= \exp[\beta^X_{i_1} - \beta^X_{i_2} + \beta^{XZ}_{i_1 j} - \beta^{XZ}_{i_2 j}]\end{aligned}$$

  - If any of i_1, i_2, or j are 1, then the corresponding parameter is 0.
    - Otherwise, this tells you the position and values of coefficients needed to create the coefficient matrix for `mcprofile()`to estimate ORs and to find confidence intervals:
      - +1 at the numerator level of X
      - −1 at the denominator level X
      - +1 on the interaction term for the numerator level of X at the fixed level of Z
      - −1 on the interaction term for the denominator level of X at the fixed level of Z
  - Similar work tells you how to compute the OR for two levels of Z at a given level of X

Example: Control of the Tomato Spotted Wilt Virus (TomatoVirus.R, TomatoVirus.csv)

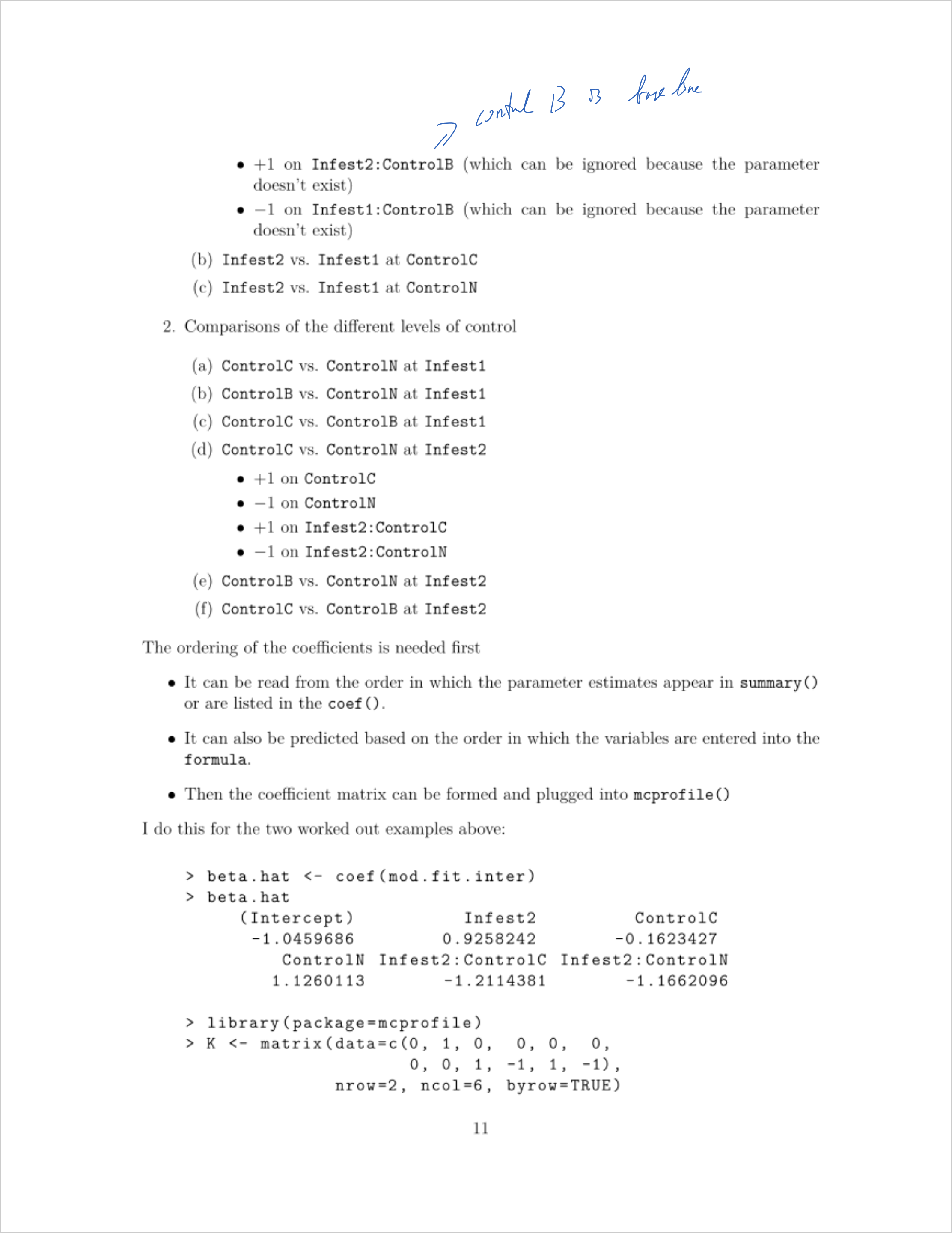
The ORs we can compute here are:

1. Comparisons of the different levels of infestation

   (a) `Infest2` vs. `Infest1` at `ControlB`

   - +1 on `Infest2`
   - −1 on `Infest1` (which can be ignored because the parameter doesn't exist)

baseline



control B is base line

- +1 on Infest2:ControlB (which can be ignored because the parameter doesn't exist)
- −1 on Infest1:ControlB (which can be ignored because the parameter doesn't exist)

(b) Infest2 vs. Infest1 at ControlC

(c) Infest2 vs. Infest1 at ControlN

2. Comparisons of the different levels of control

(a) ControlC vs. ControlN at Infest1

(b) ControlB vs. ControlN at Infest1

(c) ControlC vs. ControlB at Infest1

(d) ControlC vs. ControlN at Infest2

- +1 on ControlC
- −1 on ControlN
- +1 on Infest2:ControlC
- −1 on Infest2:ControlN

(e) ControlB vs. ControlN at Infest2

(f) ControlC vs. ControlB at Infest2

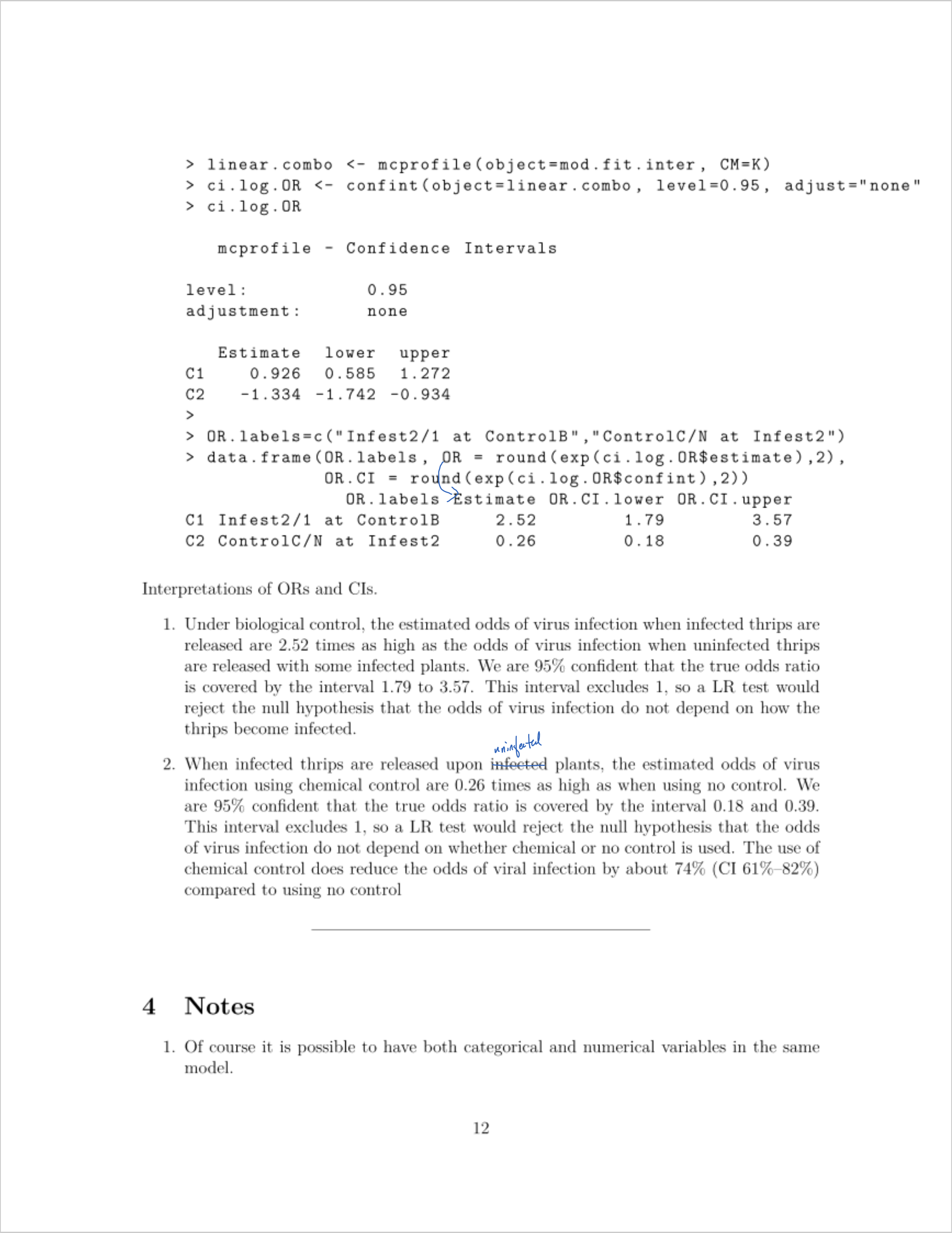
The ordering of the coefficients is needed first

- It can be read from the order in which the parameter estimates appear in summary() or are listed in the coef().
- It can also be predicted based on the order in which the variables are entered into the formula.
- Then the coefficient matrix can be formed and plugged into mcprofile()

I do this for the two worked out examples above:

```
> beta.hat <- coef(mod.fit.inter)
> beta.hat
     (Intercept)           Infest2          ControlC
      -1.0459686         0.9258242        -0.1623427
         ControlN  Infest2:ControlC  Infest2:ControlN
        1.1260113        -1.2114381        -1.1662096

> library(package=mcprofile)
> K <- matrix(data=c(0, 1, 0,  0, 0,  0,
                     0, 0, 1, -1, 1, -1),
               nrow=2, ncol=6, byrow=TRUE)
```



```
> linear.combo <- mcprofile(object=mod.fit.inter, CM=K)
> ci.log.OR <- confint(object=linear.combo, level=0.95, adjust="none"
> ci.log.OR

   mcprofile - Confidence Intervals

level:            0.95
adjustment:       none

   Estimate  lower  upper
C1    0.926  0.585  1.272
C2   -1.334 -1.742 -0.934
>
> OR.labels=c("Infest2/1 at ControlB","ControlC/N at Infest2")
> data.frame(OR.labels, OR = round(exp(ci.log.OR$estimate),2),
             OR.CI = round(exp(ci.log.OR$confint),2))
               OR.labels Estimate OR.CI.lower OR.CI.upper
C1 Infest2/1 at ControlB     2.52        1.79        3.57
C2 ControlC/N at Infest2     0.26        0.18        0.39
```

Interpretations of ORs and CIs.

1. Under biological control, the estimated odds of virus infection when infected thrips are released are 2.52 times as high as the odds of virus infection when uninfected thrips are released with some infected plants. We are 95% confident that the true odds ratio is covered by the interval 1.79 to 3.57. This interval excludes 1, so a LR test would reject the null hypothesis that the odds of virus infection do not depend on how the thrips become infected.

2. When infected thrips are released upon uninfected ~~infected~~ plants, the estimated odds of virus infection using chemical control are 0.26 times as high as when using no control. We are 95% confident that the true odds ratio is covered by the interval 0.18 and 0.39. This interval excludes 1, so a LR test would reject the null hypothesis that the odds of virus infection do not depend on whether chemical or no control is used. The use of chemical control does reduce the odds of viral infection by about 74% (CI 61%–82%) compared to using no control

4 Notes

1. Of course it is possible to have both categorical and numerical variables in the same model.