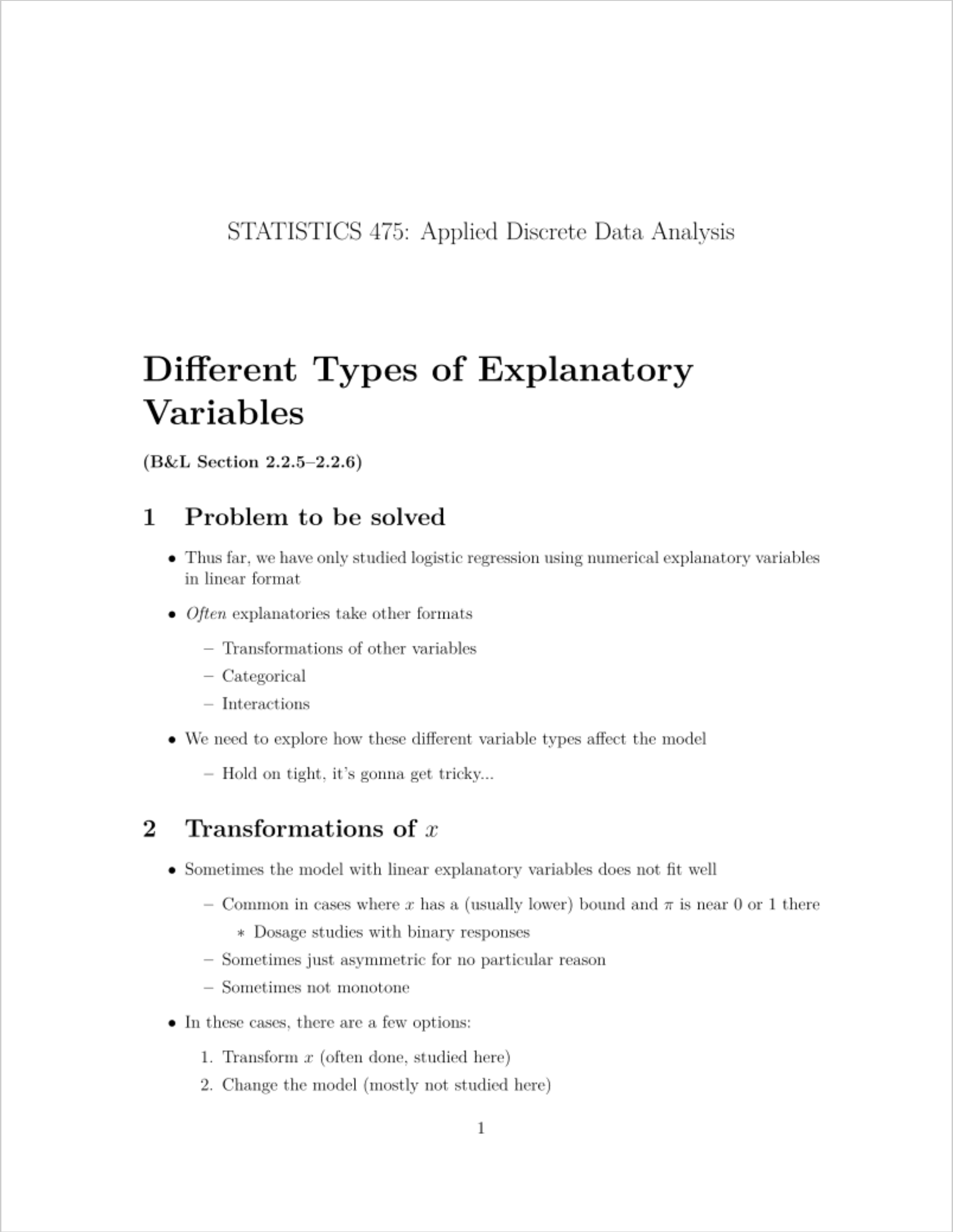
Lecture 10

January 31, 2019 10:40 AM



STATISTICS 475: Applied Discrete Data Analysis

Different Types of Explanatory Variables

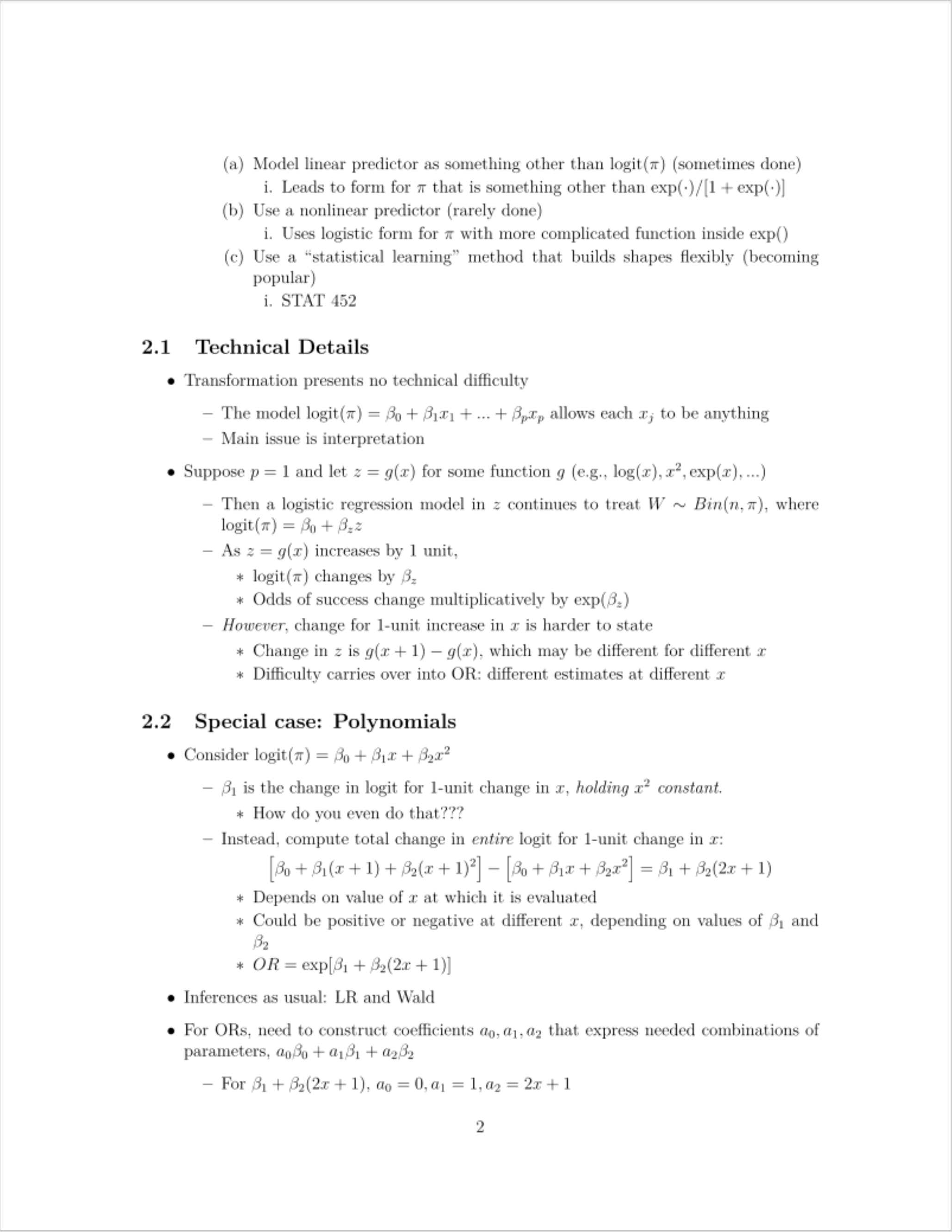
(B&L Section 2.2.5–2.2.6)

1 Problem to be solved

- Thus far, we have only studied logistic regression using numerical explanatory variables in linear format
- *Often* explanatories take other formats
  - Transformations of other variables
  - Categorical
  - Interactions
- We need to explore how these different variable types affect the model
  - Hold on tight, it's gonna get tricky...

2 Transformations of x

- Sometimes the model with linear explanatory variables does not fit well
  - Common in cases where x has a (usually lower) bound and π is near 0 or 1 there
    - Dosage studies with binary responses
  - Sometimes just asymmetric for no particular reason
  - Sometimes not monotone
- In these cases, there are a few options:
  1. Transform x (often done, studied here)
  2. Change the model (mostly not studied here)



(a) Model linear predictor as something other than logit(π) (sometimes done)
   i. Leads to form for π that is something other than $\exp(\cdot)/[1+\exp(\cdot)]$

(b) Use a nonlinear predictor (rarely done)
   i. Uses logistic form for π with more complicated function inside exp()

(c) Use a "statistical learning" method that builds shapes flexibly (becoming popular)
   i. STAT 452

2.1 Technical Details

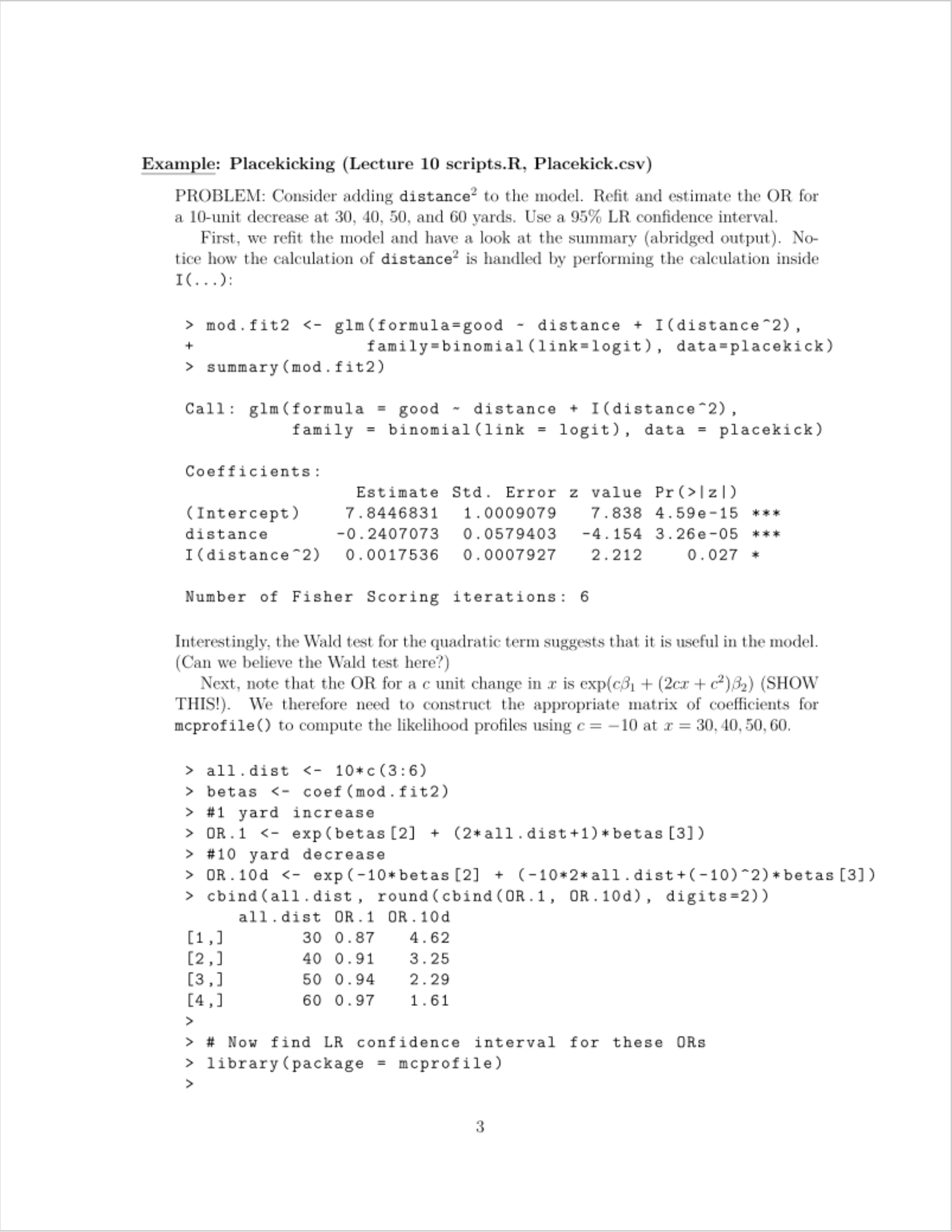
- Transformation presents no technical difficulty
  - The model $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ allows each x_j to be anything
  - Main issue is interpretation
- Suppose $p = 1$ and let $z = g(x)$ for some function g (e.g., $\log(x), x^2, \exp(x), ...$)
  - Then a logistic regression model in z continues to treat $W \sim Bin(n, \pi)$, where $\text{logit}(\pi) = \beta_0 + \beta_z z$
  - As $z = g(x)$ increases by 1 unit,
    - $\text{logit}(\pi)$ changes by β_z
    - Odds of success change multiplicatively by $\exp(\beta_z)$
  - *However*, change for 1-unit increase in x is harder to state
    - Change in z is $g(x+1) - g(x)$, which may be different for different x
    - Difficulty carries over into OR: different estimates at different x

2.2 Special case: Polynomials

- Consider $\text{logit}(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2$
  - β_1 is the change in logit for 1-unit change in x, *holding x^2 constant.*
    - How do you even do that???
  - Instead, compute total change in *entire* logit for 1-unit change in x:

$$\left[\beta_0 + \beta_1(x+1) + \beta_2(x+1)^2\right] - \left[\beta_0 + \beta_1 x + \beta_2 x^2\right] = \beta_1 + \beta_2(2x+1)$$

    - Depends on value of x at which it is evaluated
    - Could be positive or negative at different x, depending on values of β_1 and β_2
    - $OR = \exp[\beta_1 + \beta_2(2x+1)]$
- Inferences as usual: LR and Wald
- For ORs, need to construct coefficients a_0, a_1, a_2 that express needed combinations of parameters, $a_0\beta_0 + a_1\beta_1 + a_2\beta_2$
  - For $\beta_1 + \beta_2(2x+1)$, $a_0 = 0, a_1 = 1, a_2 = 2x+1$



Example: Placekicking (Lecture 10 scripts.R, Placekick.csv)

PROBLEM: Consider adding `distance`^{2} to the model. Refit and estimate the OR for a 10-unit decrease at 30, 40, 50, and 60 yards. Use a 95% LR confidence interval.

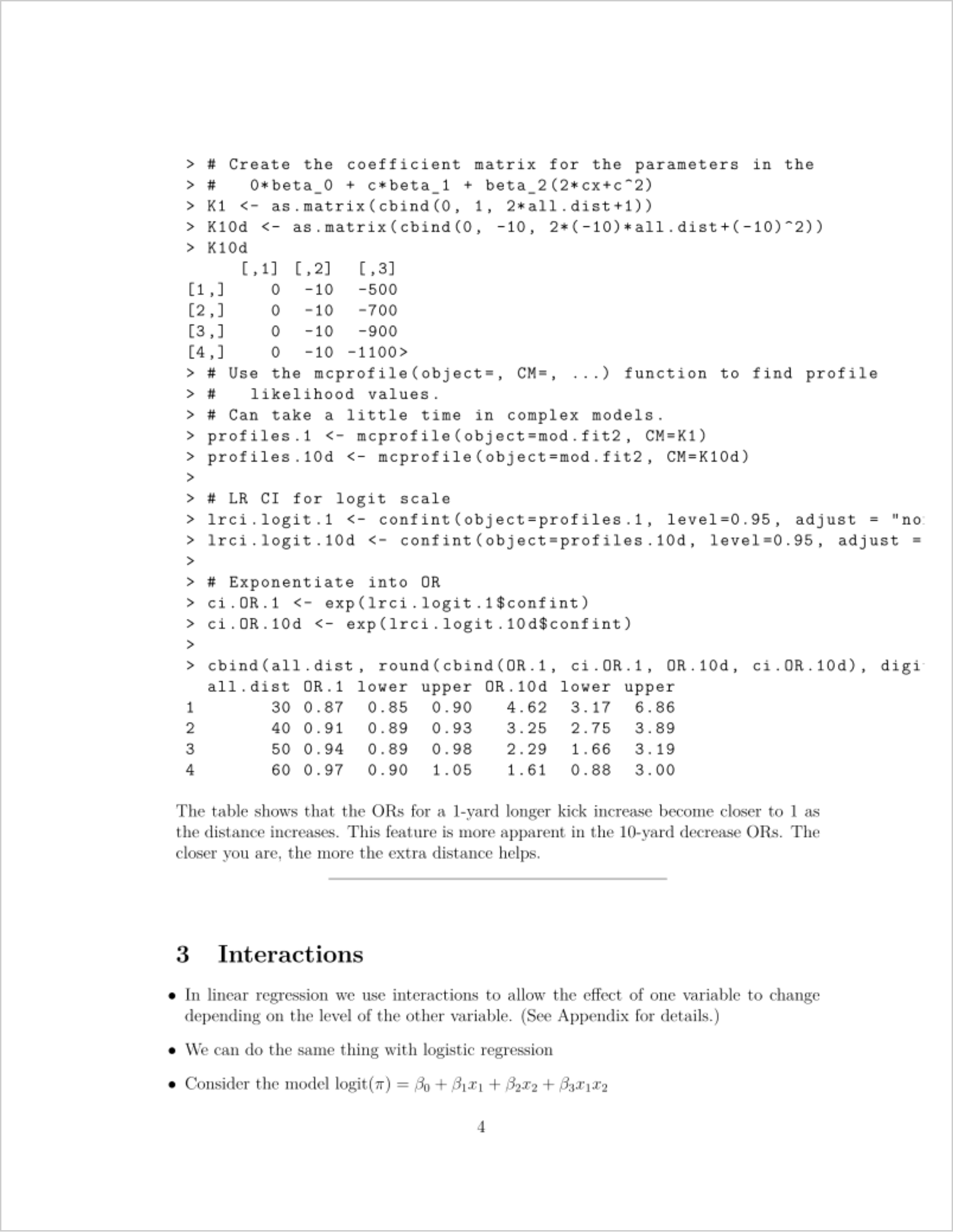
First, we refit the model and have a look at the summary (abridged output). Notice how the calculation of `distance`^{2} is handled by performing the calculation inside `I(...)`:

```
> mod.fit2 <- glm(formula=good ~ distance + I(distance^2),
+                  family=binomial(link=logit), data=placekick)
> summary(mod.fit2)

Call: glm(formula = good ~ distance + I(distance^2),
          family = binomial(link = logit), data = placekick)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    7.8446831  1.0009079   7.838 4.59e-15 ***
distance      -0.2407073  0.0579403  -4.154 3.26e-05 ***
I(distance^2)  0.0017536  0.0007927   2.212    0.027 *

Number of Fisher Scoring iterations: 6
```

Interestingly, the Wald test for the quadratic term suggests that it is useful in the model. (Can we believe the Wald test here?)

Next, note that the OR for a c unit change in x is $\exp(c\beta_1 + (2cx + c^2)\beta_2)$ (SHOW THIS!). We therefore need to construct the appropriate matrix of coefficients for `mcprofile()` to compute the likelihood profiles using $c = -10$ at $x = 30, 40, 50, 60$.

```
> all.dist <- 10*c(3:6)
> betas <- coef(mod.fit2)
> #1 yard increase
> OR.1 <- exp(betas[2] + (2*all.dist+1)*betas[3])
> #10 yard decrease
> OR.10d <- exp(-10*betas[2] + (-10*2*all.dist+(-10)^2)*betas[3])
> cbind(all.dist, round(cbind(OR.1, OR.10d), digits=2))
     all.dist OR.1 OR.10d
[1,]       30 0.87   4.62
[2,]       40 0.91   3.25
[3,]       50 0.94   2.29
[4,]       60 0.97   1.61
>
> # Now find LR confidence interval for these ORs
> library(package = mcprofile)
>
```

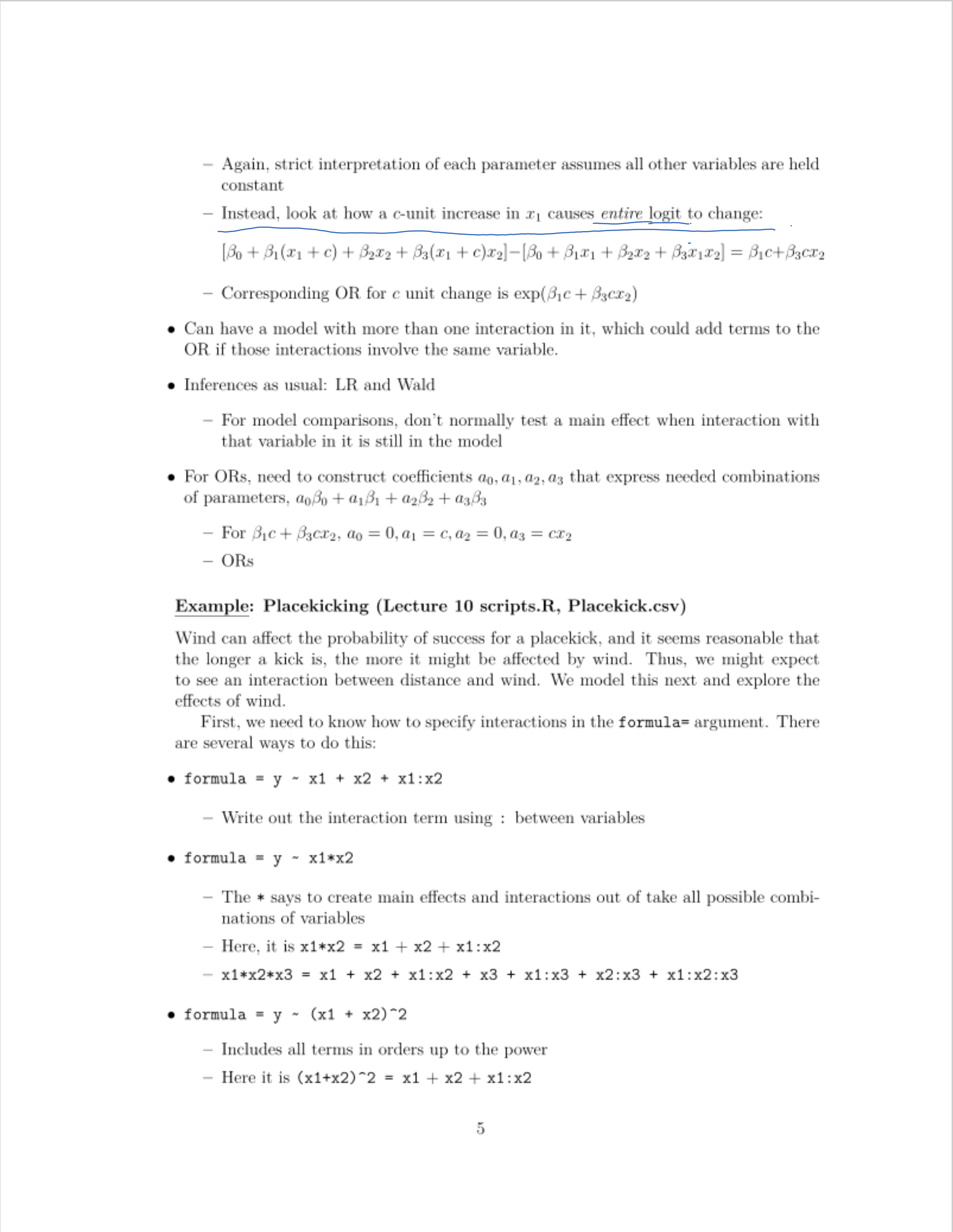


```
> # Create the coefficient matrix for the parameters in the
> #   0*beta_0 + c*beta_1 + beta_2(2*cx+c^2)
> K1 <- as.matrix(cbind(0, 1, 2*all.dist+1))
> K10d <- as.matrix(cbind(0, -10, 2*(-10)*all.dist+(-10)^2))
> K10d
     [,1] [,2]  [,3]
[1,]    0  -10  -500
[2,]    0  -10  -700
[3,]    0  -10  -900
[4,]    0  -10 -1100>
> # Use the mcprofile(object=, CM=, ...) function to find profile
> #   likelihood values.
> # Can take a little time in complex models.
> profiles.1 <- mcprofile(object=mod.fit2, CM=K1)
> profiles.10d <- mcprofile(object=mod.fit2, CM=K10d)
>
> # LR CI for logit scale
> lrci.logit.1 <- confint(object=profiles.1, level=0.95, adjust = "no
> lrci.logit.10d <- confint(object=profiles.10d, level=0.95, adjust =
>
> # Exponentiate into OR
> ci.OR.1 <- exp(lrci.logit.1$confint)
> ci.OR.10d <- exp(lrci.logit.10d$confint)
>
> cbind(all.dist, round(cbind(OR.1, ci.OR.1, OR.10d, ci.OR.10d), digi
  all.dist OR.1 lower upper OR.10d lower upper
1       30 0.87  0.85  0.90   4.62  3.17  6.86
2       40 0.91  0.89  0.93   3.25  2.75  3.89
3       50 0.94  0.89  0.98   2.29  1.66  3.19
4       60 0.97  0.90  1.05   1.61  0.88  3.00
```

The table shows that the ORs for a 1-yard longer kick increase become closer to 1 as the distance increases. This feature is more apparent in the 10-yard decrease ORs. The closer you are, the more the extra distance helps.

3 Interactions

- In linear regression we use interactions to allow the effect of one variable to change depending on the level of the other variable. (See Appendix for details.)
- We can do the same thing with logistic regression
- Consider the model $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$



- Again, strict interpretation of each parameter assumes all other variables are held constant
- Instead, look at how a c-unit increase in x_1 causes *entire* logit to change:

  $$[\beta_0 + \beta_1(x_1 + c) + \beta_2 x_2 + \beta_3(x_1 + c)x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2] = \beta_1 c + \beta_3 c x_2$$

- Corresponding OR for c unit change is $\exp(\beta_1 c + \beta_3 c x_2)$

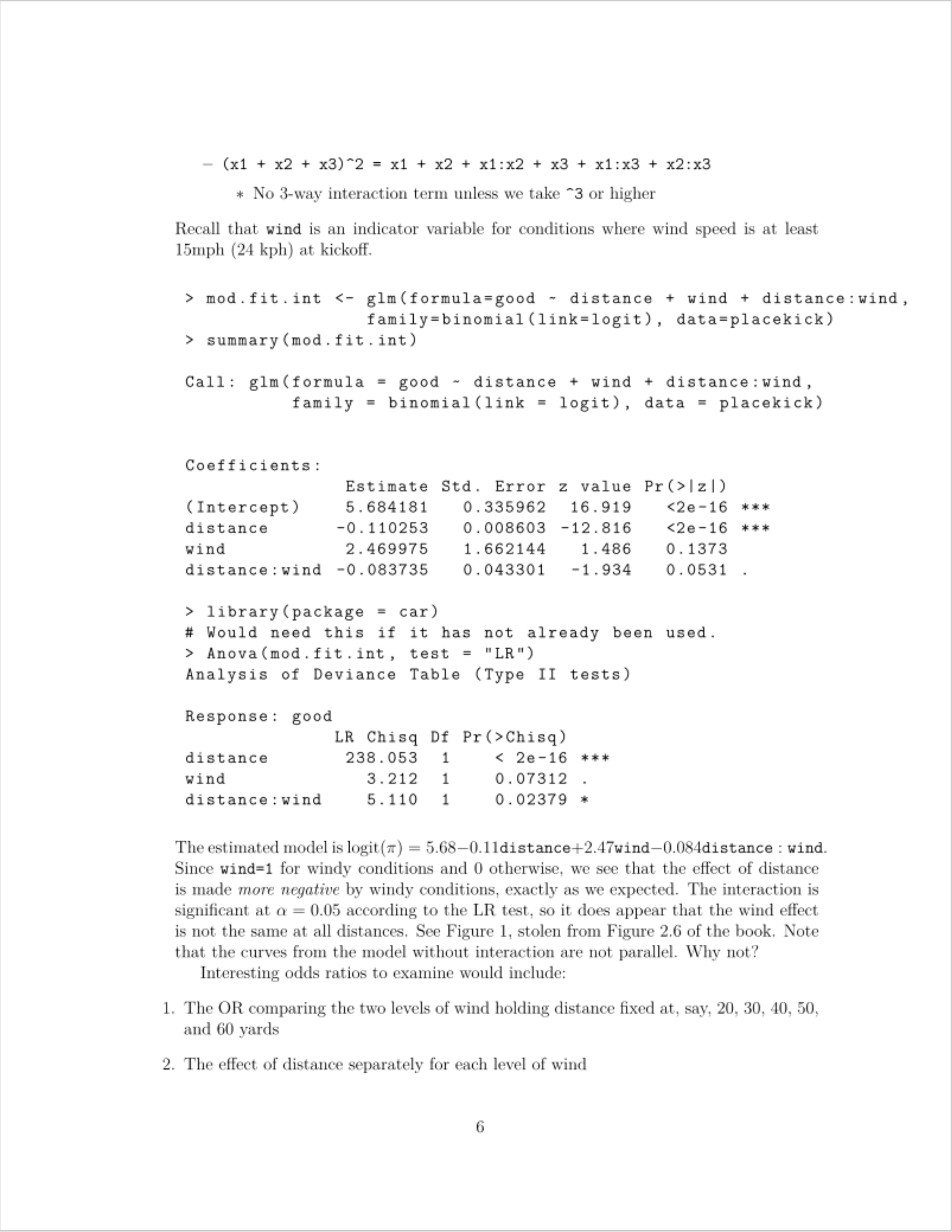
- Can have a model with more than one interaction in it, which could add terms to the OR if those interactions involve the same variable.
- Inferences as usual: LR and Wald
  - For model comparisons, don't normally test a main effect when interaction with that variable in it is still in the model
- For ORs, need to construct coefficients a_0, a_1, a_2, a_3 that express needed combinations of parameters, $a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + a_3\beta_3$
  - For $\beta_1 c + \beta_3 c x_2$, $a_0 = 0, a_1 = c, a_2 = 0, a_3 = cx_2$
  - ORs

Example: Placekicking (Lecture 10 scripts.R, Placekick.csv)

Wind can affect the probability of success for a placekick, and it seems reasonable that the longer a kick is, the more it might be affected by wind. Thus, we might expect to see an interaction between distance and wind. We model this next and explore the effects of wind.

First, we need to know how to specify interactions in the `formula=` argument. There are several ways to do this:

- `formula = y ~ x1 + x2 + x1:x2`
  - Write out the interaction term using `:` between variables
- `formula = y ~ x1*x2`
  - The `*` says to create main effects and interactions out of take all possible combinations of variables
  - Here, it is `x1*x2 = x1 + x2 + x1:x2`
  - `x1*x2*x3 = x1 + x2 + x1:x2 + x3 + x1:x3 + x2:x3 + x1:x2:x3`
- `formula = y ~ (x1 + x2)^2`
  - Includes all terms in orders up to the power
  - Here it is `(x1+x2)^2 = x1 + x2 + x1:x2`



- `(x1 + x2 + x3)^2 = x1 + x2 + x1:x2 + x3 + x1:x3 + x2:x3`
  - No 3-way interaction term unless we take `^3` or higher

Recall that `wind` is an indicator variable for conditions where wind speed is at least 15mph (24 kph) at kickoff.

```
> mod.fit.int <- glm(formula=good ~ distance + wind + distance:wind,
                   family=binomial(link=logit), data=placekick)
> summary(mod.fit.int)

Call: glm(formula = good ~ distance + wind + distance:wind,
          family = binomial(link = logit), data = placekick)


Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.684181   0.335962  16.919   <2e-16 ***
distance      -0.110253   0.008603 -12.816   <2e-16 ***
wind           2.469975   1.662144   1.486   0.1373
distance:wind -0.083735   0.043301  -1.934   0.0531 .

> library(package = car)
# Would need this if it has not already been used.
> Anova(mod.fit.int, test = "LR")
Analysis of Deviance Table (Type II tests)

Response: good
              LR Chisq Df Pr(>Chisq)
distance       238.053  1    < 2e-16 ***
wind             3.212  1    0.07312 .
distance:wind    5.110  1    0.02379 *
```

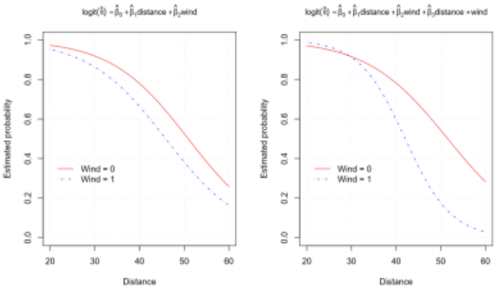
The estimated model is $\text{logit}(\pi) = 5.68 - 0.11\texttt{distance} + 2.47\texttt{wind} - 0.084\texttt{distance} : \texttt{wind}$. Since `wind=1` for windy conditions and 0 otherwise, we see that the effect of distance is made *more negative* by windy conditions, exactly as we expected. The interaction is significant at $\alpha = 0.05$ according to the LR test, so it does appear that the wind effect is not the same at all distances. See Figure 1, stolen from Figure 2.6 of the book. Note that the curves from the model without interaction are not parallel. Why not?

Interesting odds ratios to examine would include:

1. The OR comparing the two levels of wind holding distance fixed at, say, 20, 30, 40, 50, and 60 yards

2. The effect of distance separately for each level of wind

Figure 1: Plot of probability of successful placekick vs. `distance` separately for each level of `wind`. On left, from model without interaction. On right, from model with interaction. Code is in the program **placekick.R** on the book's website.



Each of these requires only knowing what coefficients to use when we form the linear combinations of parameters that are exponentiated into ORs. The rest is exactly as we have done before. I show the math work here to find the needed coefficients, and leave the numerical calculations to the program.

1. The OR for comparing the odds of success with windy conditions to the odds without windy conditions at fixed distance d:

   - $OR = \text{Odds}_{d,1}/\text{Odds}_{d,0} =$

     $$\exp(\beta_0 + \beta_1 d + \beta_2 1 + \beta_3 d * 1)/\exp(\beta_0 + \beta_1 d + \beta_2 0 + \beta_3 d * 0)$$

   - Focusing on the logits:

     $$\log(OR) = [\beta_0 + \beta_1 d + \beta_2 1 + \beta_3 d * 1] - [\beta_0 + \beta_1 d + \beta_2 0 + \beta_3 d * 0] = \beta_2 + d\beta_3$$

     - So the coefficients on the 4 parameters are $(0, 0, 1, d)$

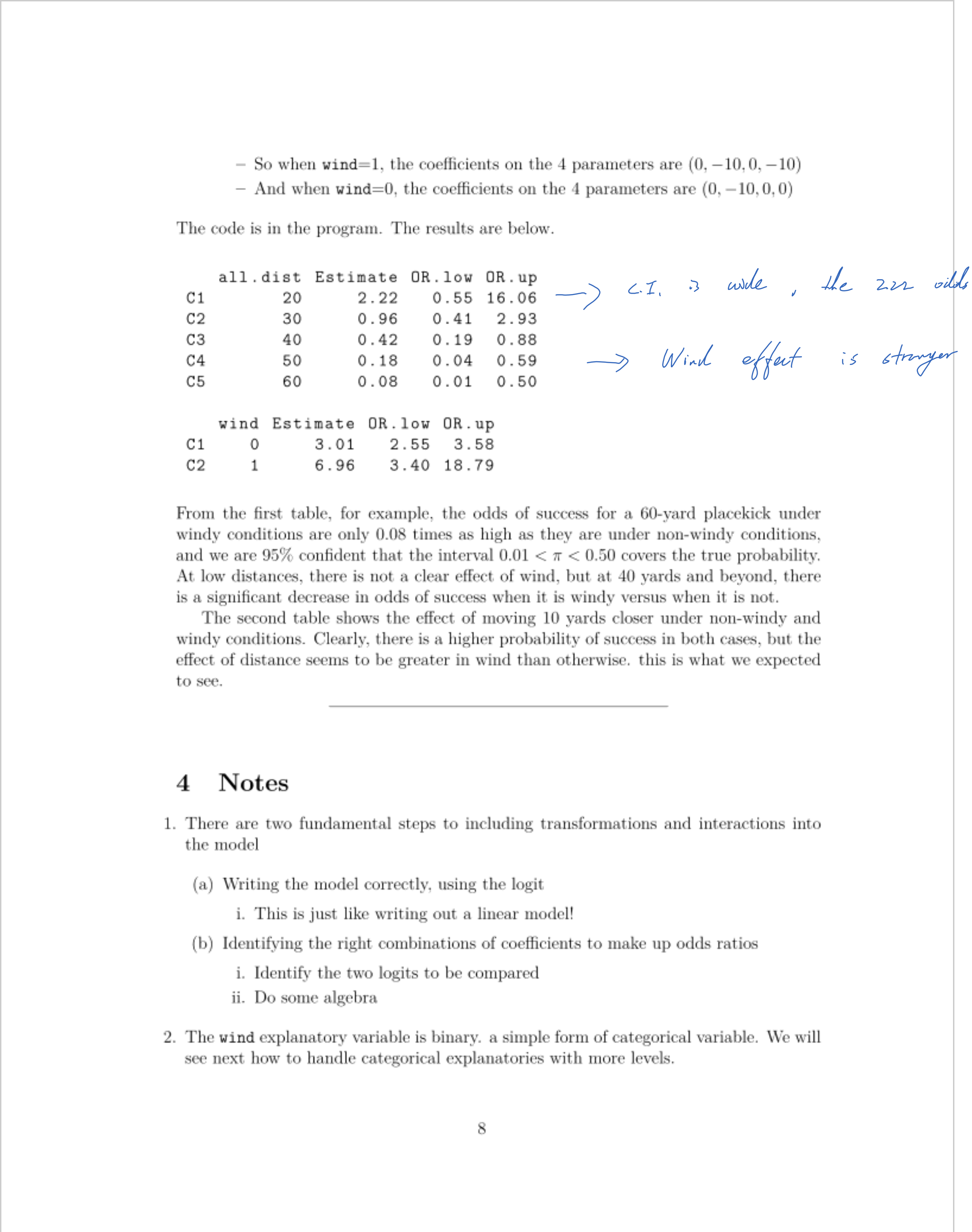
2. The OR comparing the odds of success at distance $d - 10$ to the odds of success at distance d at fixed wind level w:

   - ~~$OR = \text{Odds}_{d,w}/\text{Odds}_{d-10,w}$~~ $=$ $OR = \frac{\text{Odds}_{d-w,w}}{\text{Odds}_{d,w}}$

     $$\exp(\beta_0 + \beta_1(d-10) + \beta_2 w + \beta_3(d-10) * w)/\exp(\beta_0 + \beta_1 d + \beta_2 w + \beta_3 d * w)$$

   - Focusing on the logits:

     $$\log(OR) = [\beta_0+\beta_1(d-10)+\beta_2 w+\beta_3(d-10)*w]-[\beta_0+\beta_1 d+\beta_2 w+\beta_3 d*w] = -10\beta_1+(-10)w\beta_3$$



- So when `wind`=1, the coefficients on the 4 parameters are $(0, -10, 0, -10)$
- And when `wind`=0, the coefficients on the 4 parameters are $(0, -10, 0, 0)$

The code is in the program. The results are below.

```
   all.dist Estimate OR.low OR.up
C1       20     2.22   0.55 16.06
C2       30     0.96   0.41  2.93
C3       40     0.42   0.19  0.88
C4       50     0.18   0.04  0.59
C5       60     0.08   0.01  0.50

   wind Estimate OR.low OR.up
C1    0     3.01   2.55  3.58
C2    1     6.96   3.40 18.79
```

→ C.I. is wide, the 222 odds may not be accurate.

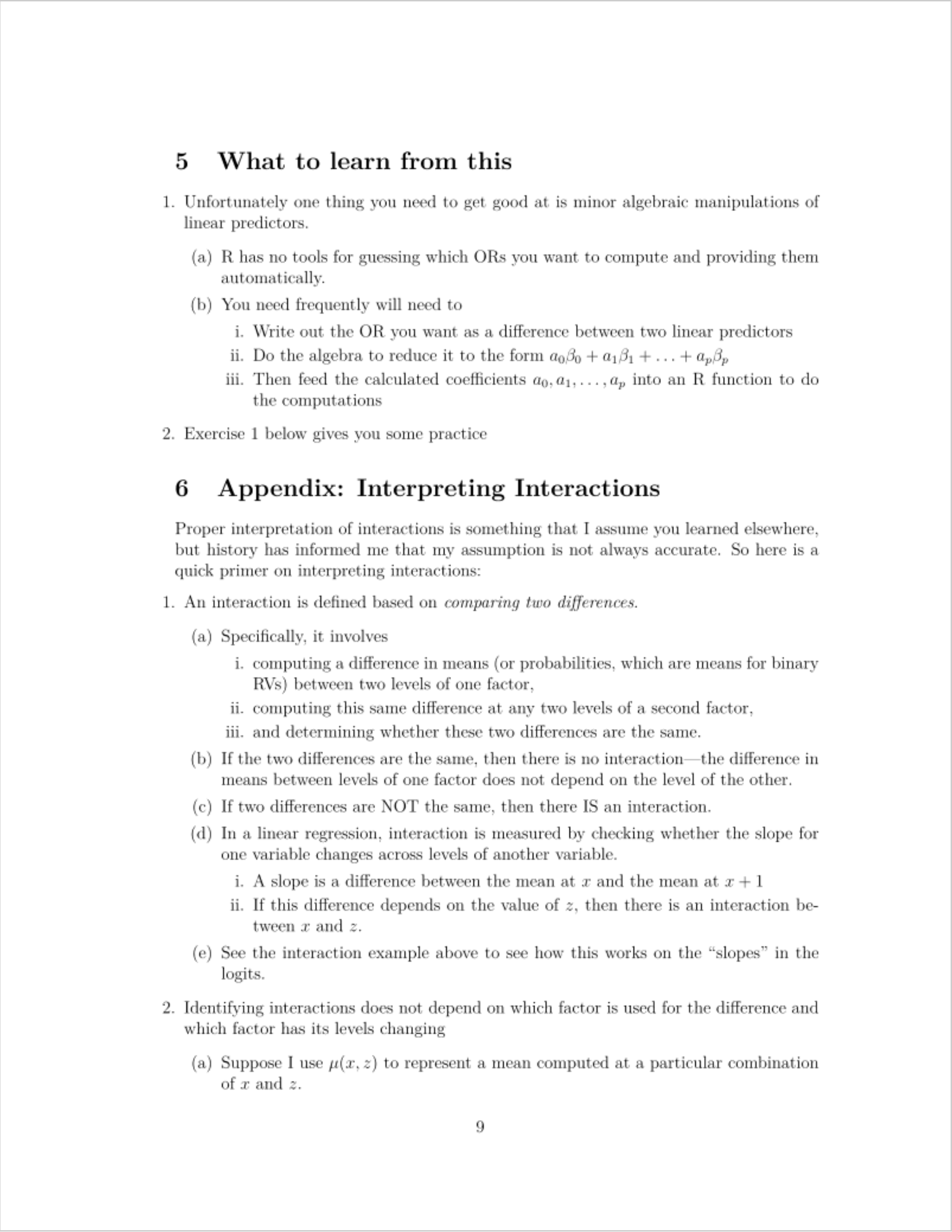
→ Wind effect is stronger when d ↑, less odds of success.

From the first table, for example, the odds of success for a 60-yard placekick under windy conditions are only 0.08 times as high as they are under non-windy conditions, and we are 95% confident that the interval $0.01 < \pi < 0.50$ covers the true probability. At low distances, there is not a clear effect of wind, but at 40 yards and beyond, there is a significant decrease in odds of success when it is windy versus when it is not.

The second table shows the effect of moving 10 yards closer under non-windy and windy conditions. Clearly, there is a higher probability of success in both cases, but the effect of distance seems to be greater in wind than otherwise. this is what we expected to see.

4 Notes

1. There are two fundamental steps to including transformations and interactions into the model
   (a) Writing the model correctly, using the logit
      i. This is just like writing out a linear model!
   (b) Identifying the right combinations of coefficients to make up odds ratios
      i. Identify the two logits to be compared
      ii. Do some algebra
2. The `wind` explanatory variable is binary. a simple form of categorical variable. We will see next how to handle categorical explanatories with more levels.



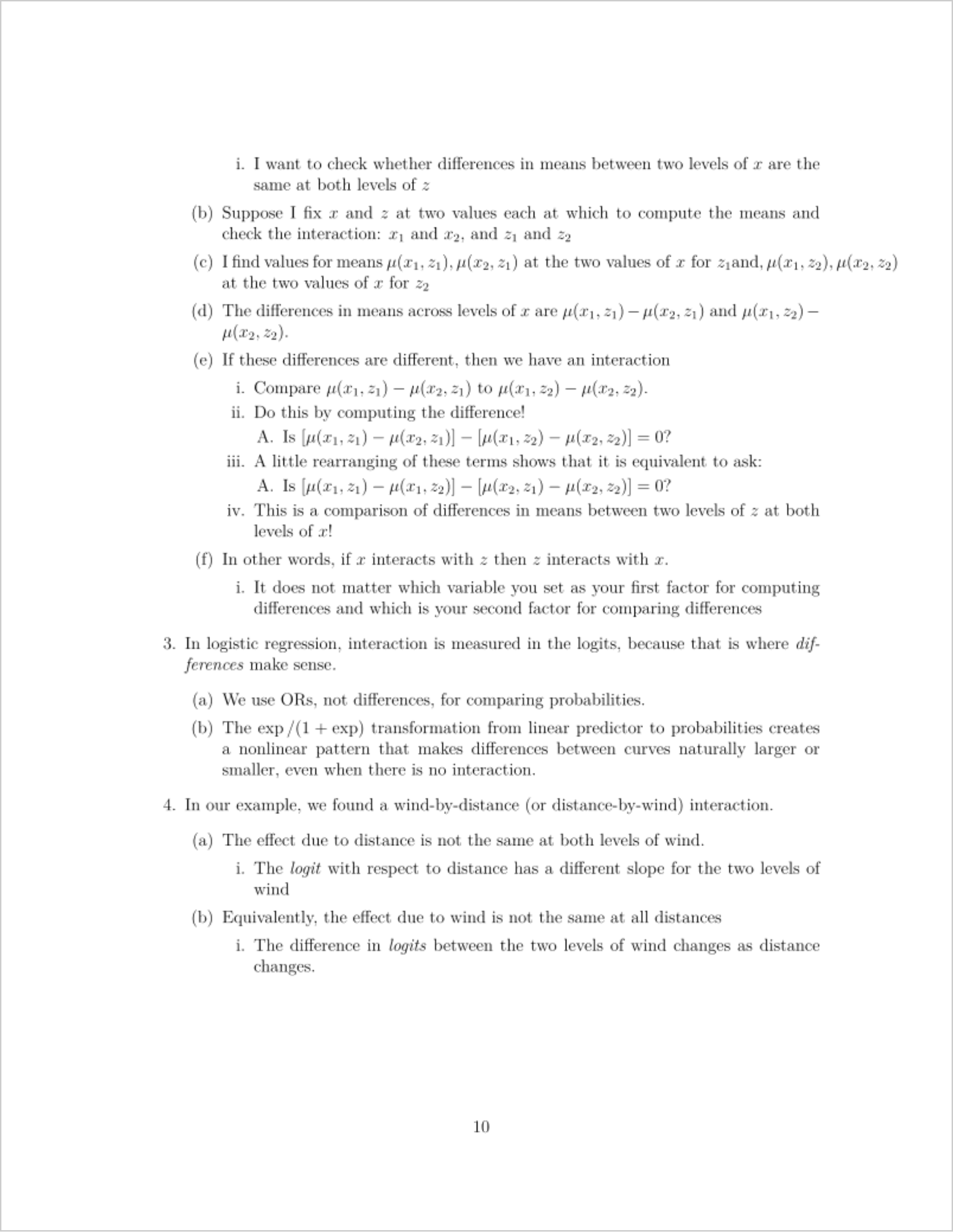
5 What to learn from this

1. Unfortunately one thing you need to get good at is minor algebraic manipulations of linear predictors.
   (a) R has no tools for guessing which ORs you want to compute and providing them automatically.
   (b) You need frequently will need to
      i. Write out the OR you want as a difference between two linear predictors
      ii. Do the algebra to reduce it to the form $a_0\beta_0 + a_1\beta_1 + \ldots + a_p\beta_p$
      iii. Then feed the calculated coefficients $a_0, a_1, \ldots, a_p$ into an R function to do the computations
2. Exercise 1 below gives you some practice

6 Appendix: Interpreting Interactions

Proper interpretation of interactions is something that I assume you learned elsewhere, but history has informed me that my assumption is not always accurate. So here is a quick primer on interpreting interactions:

1. An interaction is defined based on *comparing two differences.*
   (a) Specifically, it involves
      i. computing a difference in means (or probabilities, which are means for binary RVs) between two levels of one factor,
      ii. computing this same difference at any two levels of a second factor,
      iii. and determining whether these two differences are the same.
   (b) If the two differences are the same, then there is no interaction—the difference in means between levels of one factor does not depend on the level of the other.
   (c) If two differences are NOT the same, then there IS an interaction.
   (d) In a linear regression, interaction is measured by checking whether the slope for one variable changes across levels of another variable.
      i. A slope is a difference between the mean at x and the mean at $x + 1$
      ii. If this difference depends on the value of z, then there is an interaction between x and z.
   (e) See the interaction example above to see how this works on the "slopes" in the logits.
2. Identifying interactions does not depend on which factor is used for the difference and which factor has its levels changing
   (a) Suppose I use $\mu(x, z)$ to represent a mean computed at a particular combination of x and z.



i. I want to check whether differences in means between two levels of x are the same at both levels of z

(b) Suppose I fix x and z at two values each at which to compute the means and check the interaction: x_1 and x_2, and z_1 and z_2

(c) I find values for means $\mu(x_1, z_1), \mu(x_2, z_1)$ at the two values of x for $z_1$and, $\mu(x_1, z_2), \mu(x_2, z_2)$ at the two values of x for z_2

(d) The differences in means across levels of x are $\mu(x_1, z_1) - \mu(x_2, z_1)$ and $\mu(x_1, z_2) - \mu(x_2, z_2)$.

(e) If these differences are different, then we have an interaction

i. Compare $\mu(x_1, z_1) - \mu(x_2, z_1)$ to $\mu(x_1, z_2) - \mu(x_2, z_2)$.

ii. Do this by computing the difference!

A. Is $[\mu(x_1, z_1) - \mu(x_2, z_1)] - [\mu(x_1, z_2) - \mu(x_2, z_2)] = 0$?

iii. A little rearranging of these terms shows that it is equivalent to ask:

A. Is $[\mu(x_1, z_1) - \mu(x_1, z_2)] - [\mu(x_2, z_1) - \mu(x_2, z_2)] = 0$?

iv. This is a comparison of differences in means between two levels of z at both levels of x!

(f) In other words, if x interacts with z then z interacts with x.

i. It does not matter which variable you set as your first factor for computing differences and which is your second factor for comparing differences

3. In logistic regression, interaction is measured in the logits, because that is where *differences* make sense.

(a) We use ORs, not differences, for comparing probabilities.

(b) The $\exp/(1 + \exp)$ transformation from linear predictor to probabilities creates a nonlinear pattern that makes differences between curves naturally larger or smaller, even when there is no interaction.

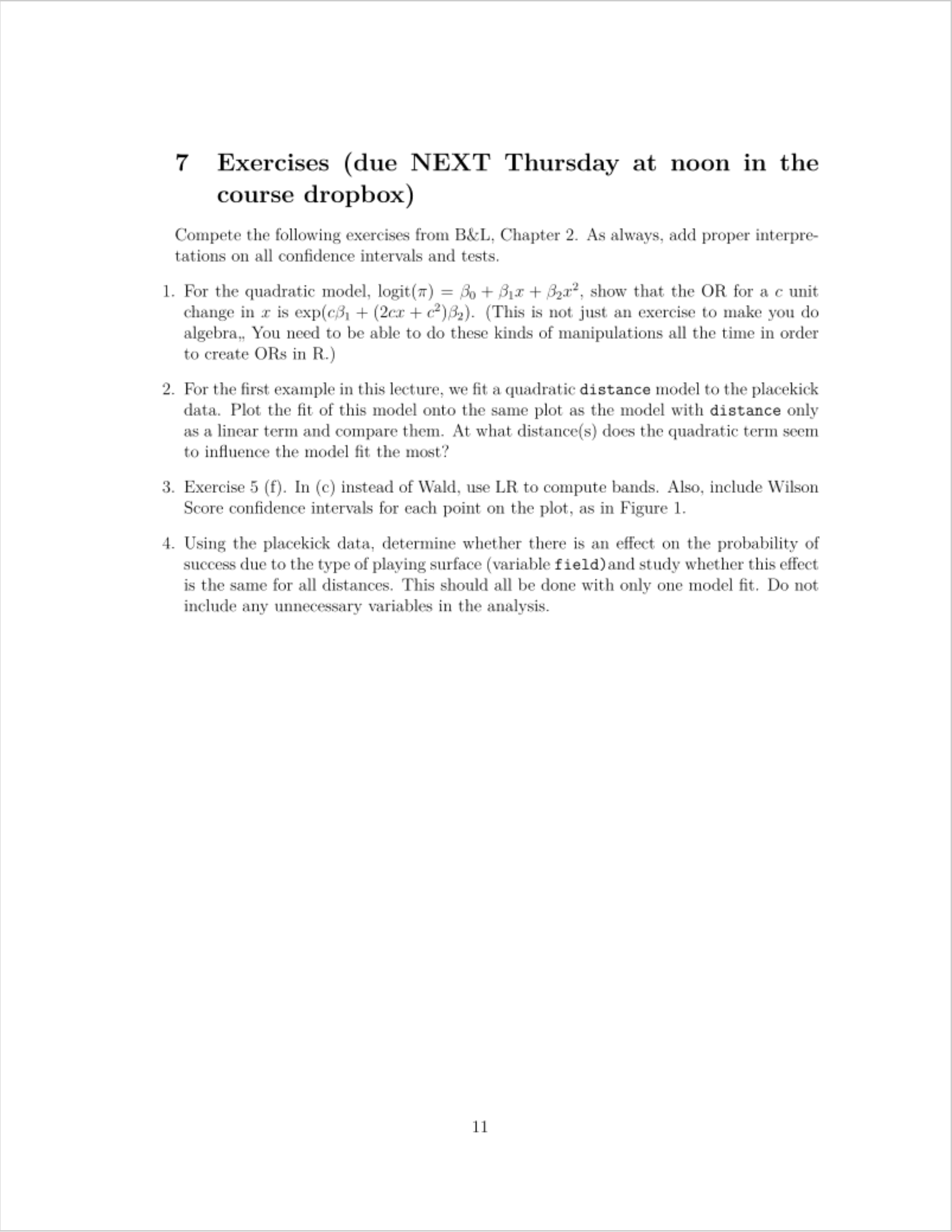
4. In our example, we found a wind-by-distance (or distance-by-wind) interaction.

(a) The effect due to distance is not the same at both levels of wind.

i. The *logit* with respect to distance has a different slope for the two levels of wind

(b) Equivalently, the effect due to wind is not the same at all distances

i. The difference in *logits* between the two levels of wind changes as distance changes.



7 Exercises (due NEXT Thursday at noon in the course dropbox)

Compete the following exercises from B&L, Chapter 2. As always, add proper interpretations on all confidence intervals and tests.

1. For the quadratic model, $\text{logit}(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2$, show that the OR for a c unit change in x is $\exp(c\beta_1 + (2cx + c^2)\beta_2)$. (This is not just an exercise to make you do algebra,, You need to be able to do these kinds of manipulations all the time in order to create ORs in R.)

2. For the first example in this lecture, we fit a quadratic `distance` model to the placekick data. Plot the fit of this model onto the same plot as the model with `distance` only as a linear term and compare them. At what distance(s) does the quadratic term seem to influence the model fit the most?

3. Exercise 5 (f). In (c) instead of Wald, use LR to compute bands. Also, include Wilson Score confidence intervals for each point on the plot, as in Figure 1.

4. Using the placekick data, determine whether there is an effect on the probability of success due to the type of playing surface (variable `field`)and study whether this effect is the same for all distances. This should all be done with only one model fit. Do not include any unnecessary variables in the analysis.