Logistic Regression

Max Turgeon

SCI 2000–Introduction to Data Science

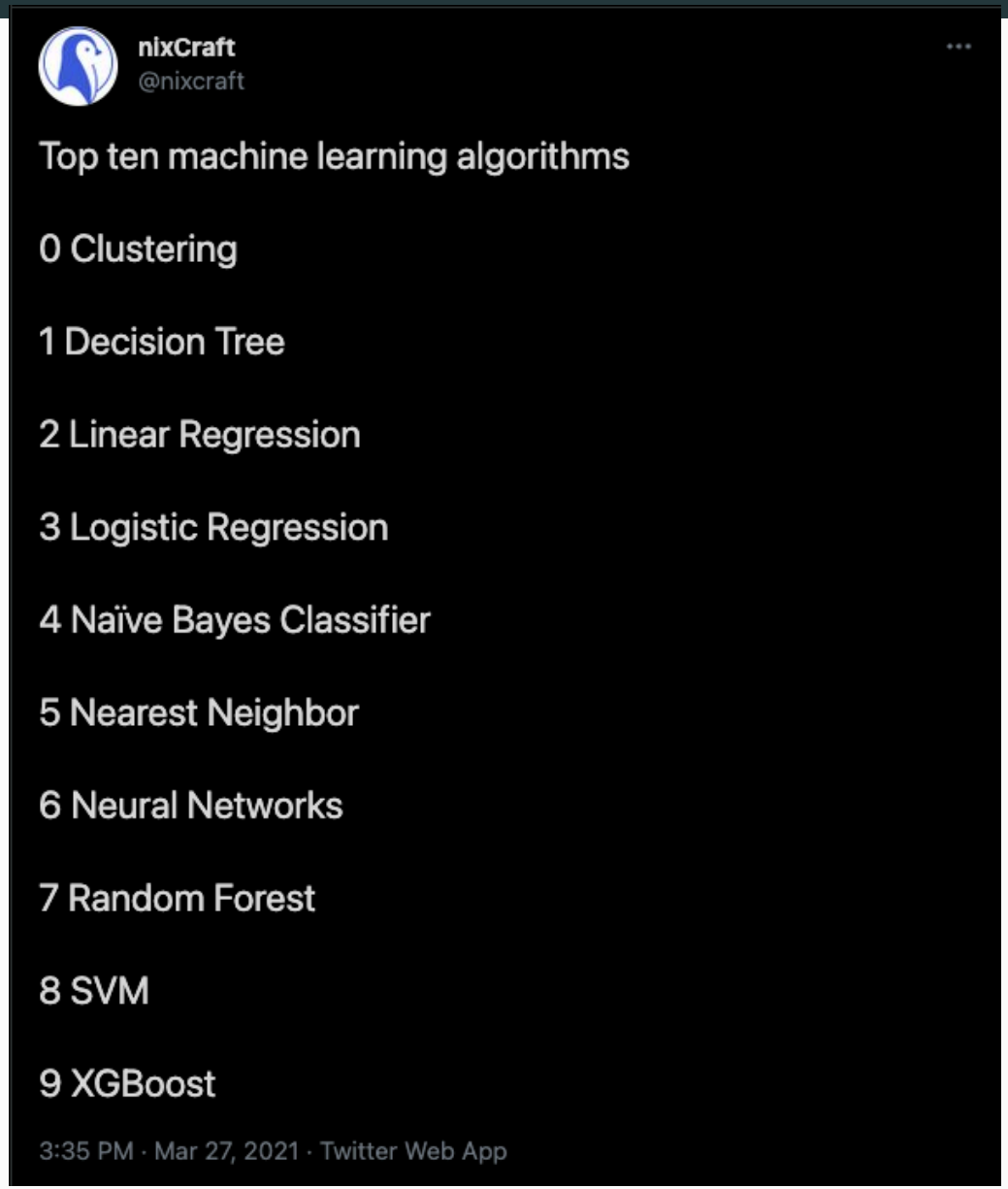
Lecture Objectives

- Fit logistic regression models using R.
- Understand the output and interpret the coefficients.
- Evaluate the goodness of fit.

Motivation i

- Earlier in the semester, we discussed linear regression.
  - Measure differences in **averages** between different subgroups.
  - For continuous outcome variables.
- **Logistic regression** is a way to model the relationship between a binary outcome variable and a set of covariates.
  - It's also used as a basis for prediction modeling in machine learning.

Motivation ii



Main definitions

- Y is a binary outcome variable (i.e. $Y = 0$ or $Y = 1$).

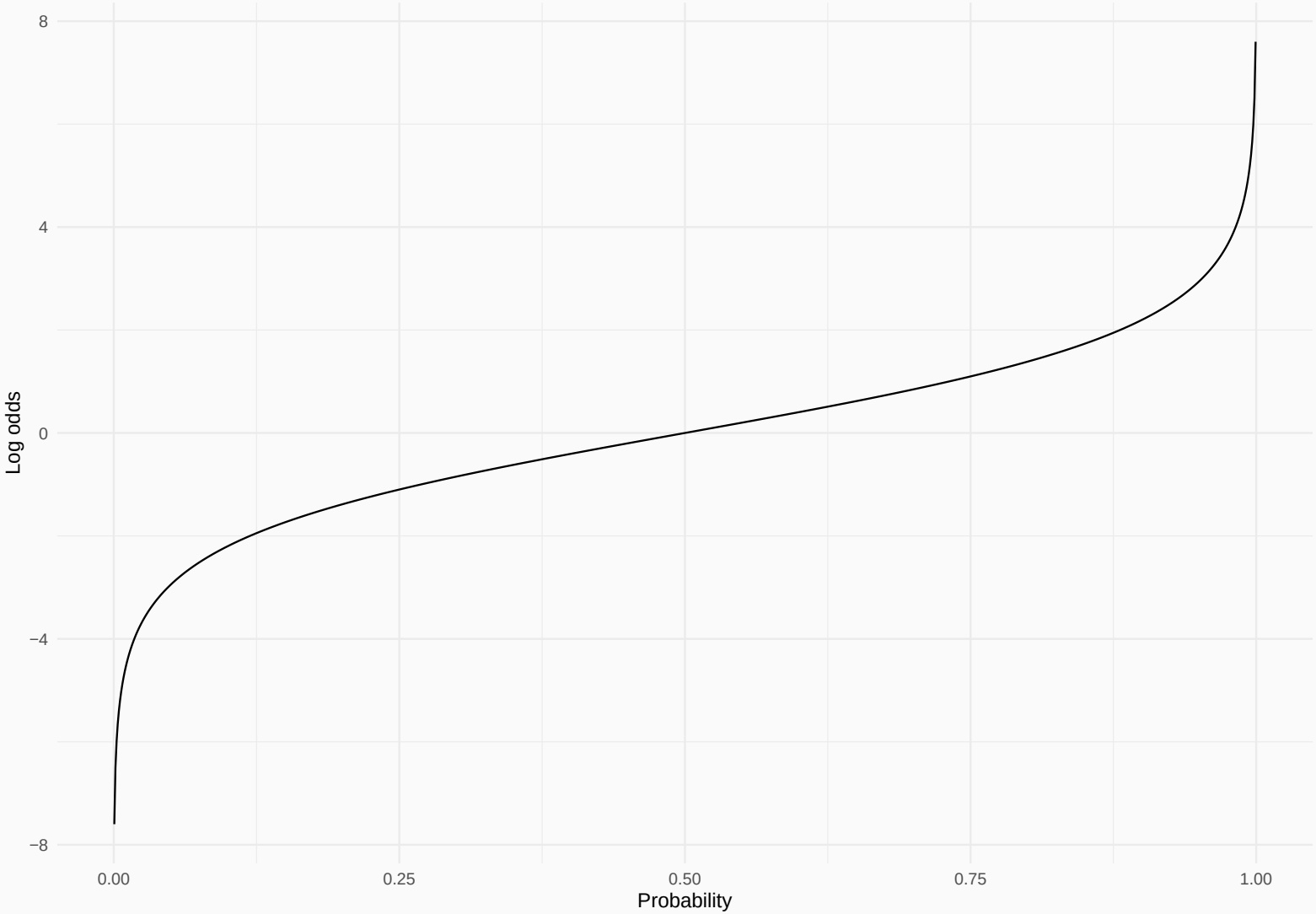
$$\operatorname{logit}\left(E(Y \mid X_1, \ldots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Note: $\operatorname{logit}(t) = \log(t/(1-t))$.
- The coefficients β_i represent comparisons of **log odds** for different values of the covariates (i.e. for different subgroups).

Comments i

- If Y is a binary random variable, then $E(Y) = P(Y = 1)$.
- The **odds** is the ratio $P(Y = 1)/P(Y = 0)$.
  - E.g. if the odds is 2, then $Y = 1$ is twice as likely than $Y = 0$.
  - In other words, $P(Y = 1) = 0.66$.
- The **logit function** takes probabilities (which are between 0 and 1) and transforms them to a real number (from $-\infty$ to ∞)

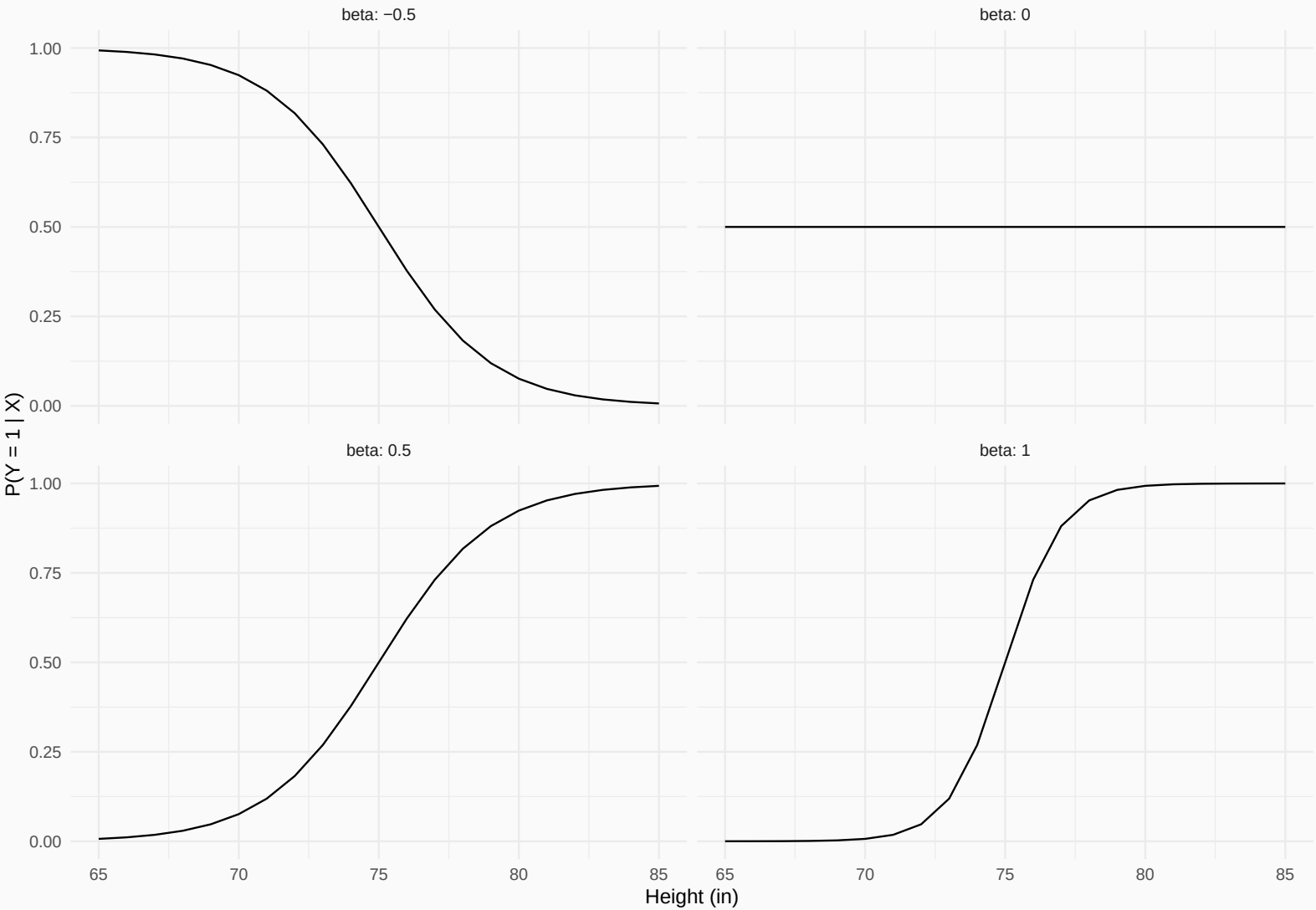
Comments ii



Example i

- Assume we have one covariate X: height in inches.
- The covariate Y: whether someone is a good basketball player (or not).
- Let's look at the effect of β on the relationship between X and $P(Y = 1 \mid X)$.

Example ii


beta: −0.5
beta: 0
beta: 0.5
beta: 1
P(Y = 1 | X)
1.00
0.75
0.50
0.25
0.00
65
70
75
80
85
Height (in)

Example i

- Consider the following 2x2 table:

	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- Let Y be handedness, and let X be sex.
- **Note**: The odds for female is $(44/48)/(4/48) = 11$; the odds for male is $(43/52)/(9/52) = 4.78$.

Example ii

```
library(tidyverse)
# Create dataset
dataset <- bind_rows(
  data.frame(Y = rep("right", 43),
             X = rep("male", 43)),
  data.frame(Y = rep("right", 44),
             X = rep("female", 44)),
  data.frame(Y = rep("left", 9),
             X = rep("male", 9)),
  data.frame(Y = rep("left", 4),
             X = rep("female", 4)))
```

Example iii

```
glimpse(dataset)
```

```
## Rows: 100
## Columns: 2
## $ Y <chr> "right", "right", "right", "right",
"right", "right", "right", "righ~
## $ X <chr> "male", "male", "male", "male",
"male", "male", "male", "male", "mal~
```

Example iv

```
# Outcome must be 0 or 1
dataset <- mutate(dataset, Y = as.numeric(Y=="right"))

glm(Y ~ X, data = dataset,
    family = "binomial")
```

```
##
## Call: glm(formula = Y ~ X, family =
"binomial", data = dataset)
##
## Coefficients:
```

Example v

```
## (Intercept) Xmale
## 2.3979 -0.8339
##
## Degrees of Freedom: 99 Total (i.e. Null); 98
Residual
## Null Deviance: 77.28
## Residual Deviance: 75.45 AIC: 79.45
```

```
# Relationship with odds?
log(11)
```

```
## [1] 2.397895
```

Example vi

```
log(4.78/11)
```

```
## [1] -0.8334547
```

Interpreting coeffficients i

- The regression coefficients in logistic regression measure differences in **log odds**.
    - Or put another way: they measure ratios of odds on the log scale.
    - Very common to take the exponential of coefficients (and confidence intervals).
- Let's start with the example of a single binary covariate X.

Interpreting coeffficients ii

- If $X = 0$, we have

$$\log \frac{P(Y = 1 \mid X = 0)}{P(Y = 0 \mid X = 0)} = \beta_0.$$

- In other words, the intercept term β_0 corresponds to the log-odds when all covariates are equal to zero.

Interpreting coeffficients iii

- Now, let's look at $X = 1$

$$\log \frac{P(Y = 1 \mid X = 1)}{P(Y = 0 \mid X = 1)} = \beta_0 + \beta_1.$$

- Therefore, β_1 is the difference in log-odds between $X = 1$ and $X = 0$.
- Using logarithm rules, the difference in log-odds is the same as the log of the odds ratio.

Exercise

The dataset `case2001` from the `Sleuth3` package contains information about members of the Donner party who got trapped by snow on their way to California.

Using logistic regression, investigate the relationship between age and survival. Carefully interpret the regression coefficient estimates. Is the association statistically significant?

Solution i

```
library(Sleuth3)
library(tidyverse)

# First transform outcome to 0/1
dataset <- mutate(case2001,
                  Y = as.numeric(Status == "Died"))

fit <- glm(Y ~ Age, data = dataset,
           family = "binomial")
```

Solution ii

```
coef(fit)
```

```
## (Intercept)         Age
## -1.81851831  0.06647028
```

- We can't interpret the intercept, as it would correspond to age 0.
- The coefficient for age is 0.07, which means for two groups whose age differ by 1 year, the log odds differ by 0.07.
- Alternatively, the odds ratio is $\exp(0.07) = 1.07$.
    - Sometimes you'll see "odds increased by 7%".

Solution iii

```
confint(fit)
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.99016010 0.005987258
## Age          0.01016096 0.139737905
```

```
exp(confint(fit))
```

```
##                   2.5 %   97.5 %
## (Intercept) 0.01849675 1.006005
## Age         1.01021276 1.149972
```

Assumptions

Logistic regression has less assumptions than linear regression.

1. Validity (with respect to the research question).
2. Representativeness (of the data with respect to the population).
3. Additivity and linearity.
4. (Conditional) Independence of the outcomes.

Note: There is only one possible distribution for binary outcomes, i.e. Bernoulli. As a consequence, we **always** have heteroscedasticity.

Diagnostic plots i

- Diagnostic plots are trickier with logistic regression because the data is *discrete*.
    - And therefore the *residuals* are also discrete.
- One useful solution: *bin the outcomes/residuals*.
    - Bin observations with similar fitted values.
    - Take the average of residuals and fitted values.
    - Plot the averages against one another.
- As residual plots in linear regression, we are looking for random pattern around horizontal line.
- **Note**: There is a balance between enough bins to see patterns and enough observations by bins to have stable averages.

Example i

- We will use data on Duchenne Muscular Dystrophy (DMD).
  - Can be downloaded from `https://biostat.app.vumc.org/wiki/Main/DataSets`
- Goal of the study was to develop a screening program for female relatives of boys with DMD.
- **Outcome**: Carrier status
- **Covariates**: serum markers; creatine kinase (`ck`), hemopexin (`h`), pyruvate kinase (`pk`) and lactate dehydroginase (`ld`).

Example ii

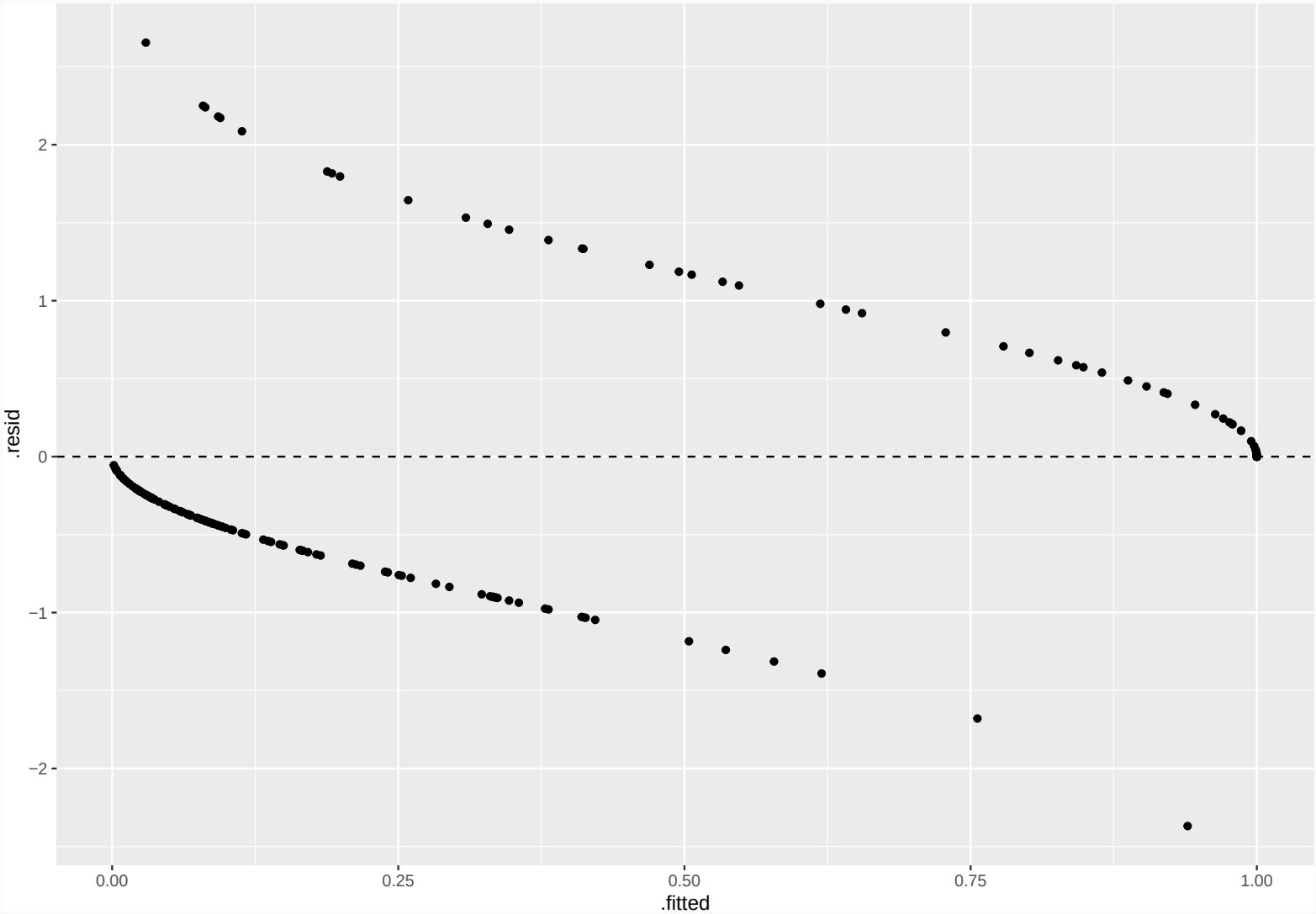
```
library(tidyverse)
# Import dataset into R
data_dmd <- read_csv("dmd.csv")

## Warning: Missing column names filled in: 'X1' [1]

# Remove rows with missing values
data_dmd <- na.omit(data_dmd)

model <- glm(carrier ~ ck + h, data = data_dmd,
             family = "binomial")
confint(model)
```

Example iii

```
##                    2.5 %       97.5 %
## (Intercept) -20.76823776 -10.43024757
## ck            0.04058575   0.08519017
## h             0.07813791   0.17837069
```

```
library(broom)
# Plot residuals and probabilities (no binning)
augment(model, type.predict = "response") %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dashed")
```

Example iv

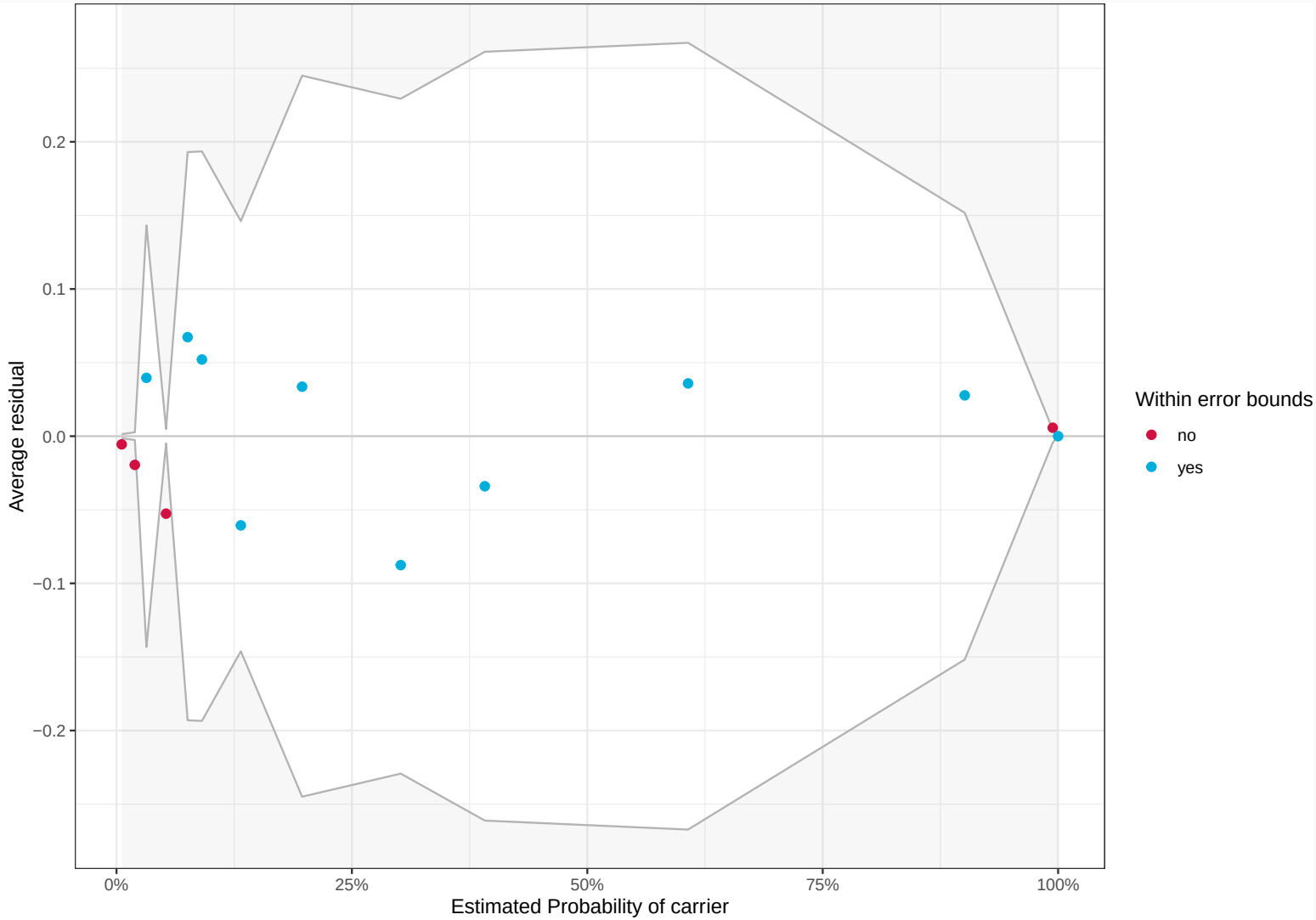


Example v

```
# We will use the performance package
library(performance)

# By default: residuals vs fitted probs
#             sqrt(n) bins (~14 bins)
binned_residuals(model)
```

```
## Warning: Probably bad model fit. Only about
71% of the residuals are inside the error bounds.
```

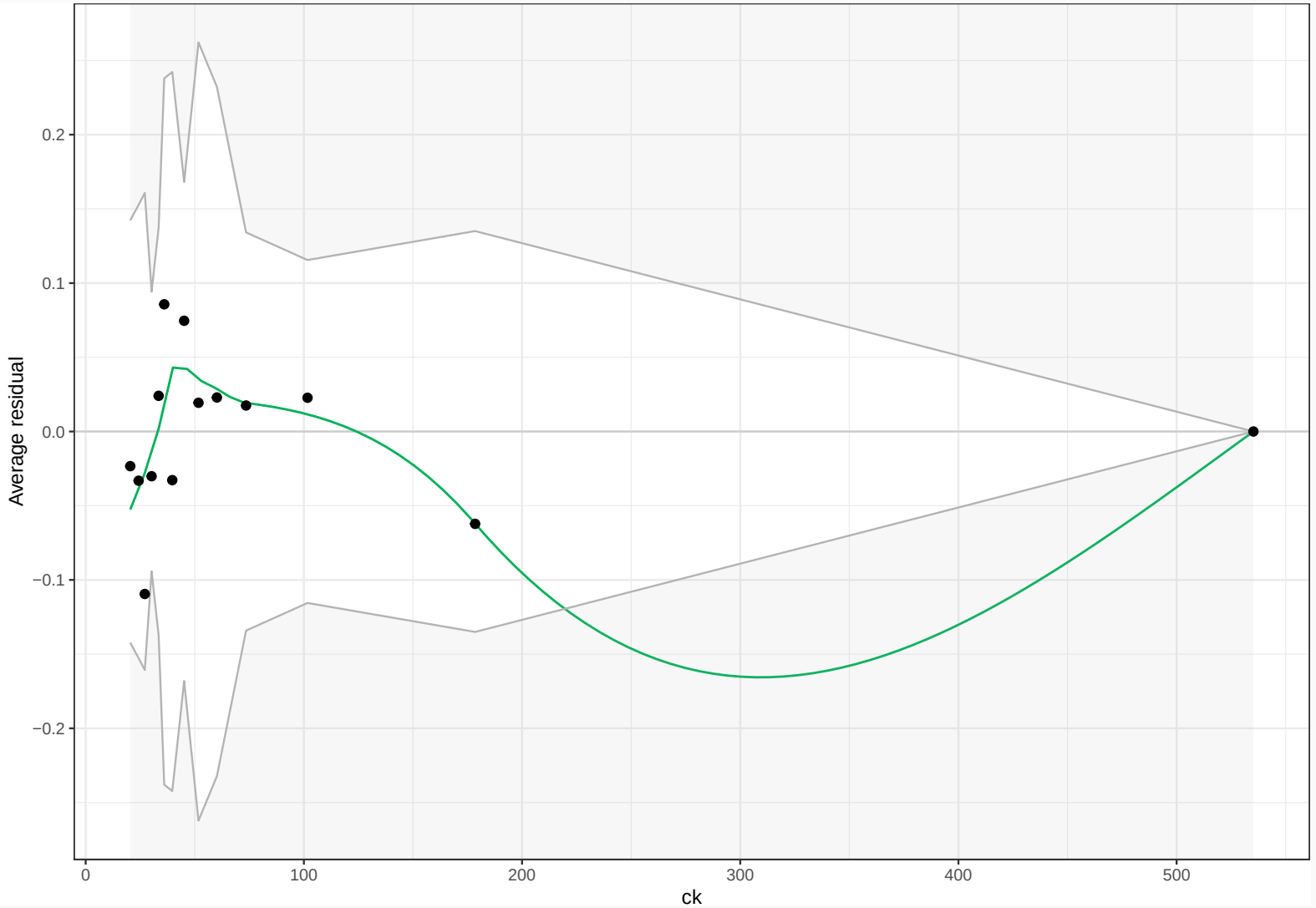
Example vi


0.2
0.1
0.0
−0.1
−0.2
Average residual
0%
25%
50%
75%
100%
Estimated Probability of carrier
Within error bounds
no
yes

Example vii

```
# Use 'term' to plot against covariate
binned_residuals(model, term = "ck")


## Ok: About 100% of the residuals are inside the
error bounds.
```

Example viii

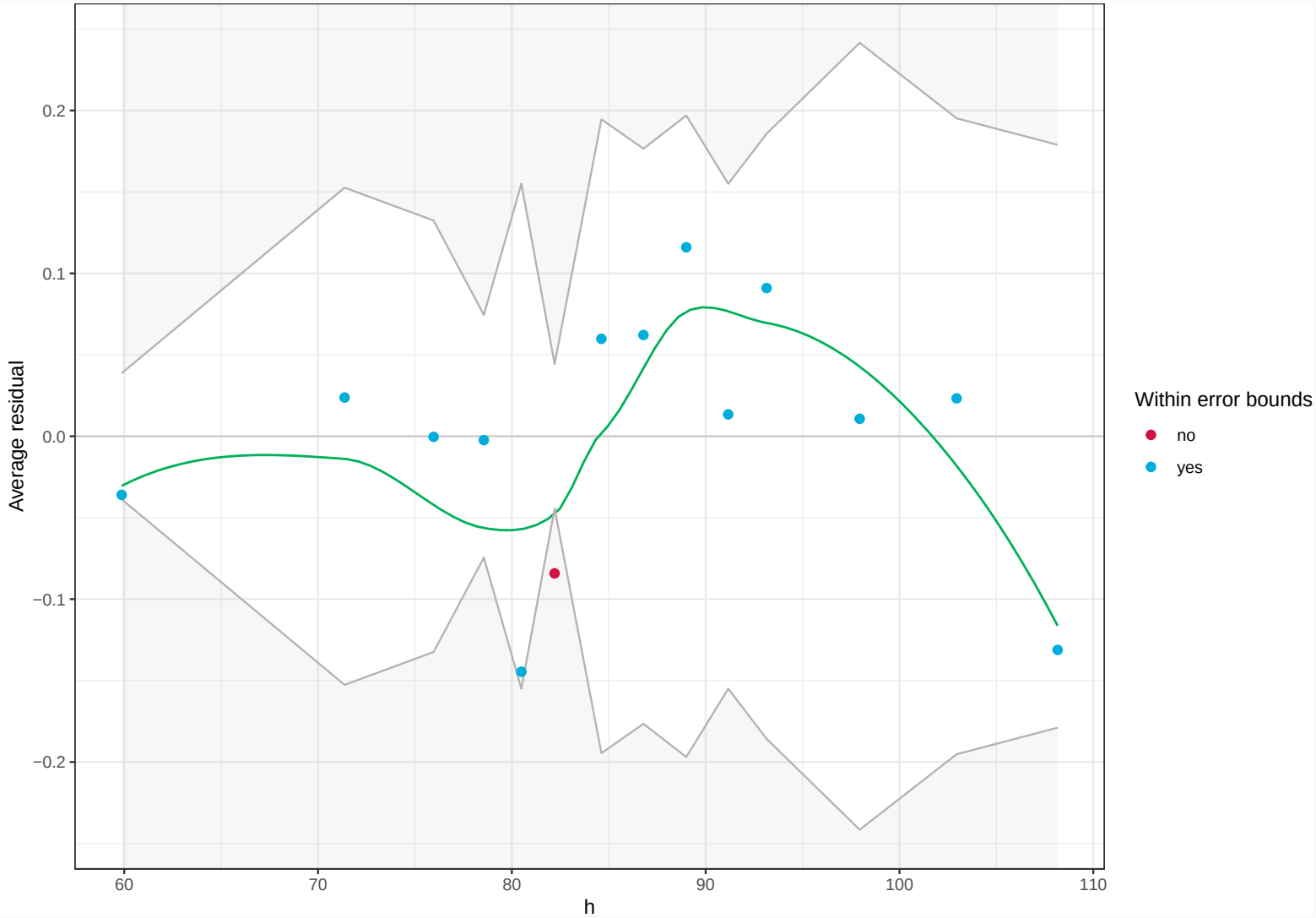


Example ix

```
binned_residuals(model, term = "h")
```

```
## Warning: About 93% of the residuals are inside
the error bounds (~95% or higher would be good).
```

Example x


0.2
0.1
0.0
−0.1
−0.2
Average residual
60
70
80
90
100
110
h
Within error bounds
no
yes

Example xi

- We have evidence of poor model fit (from binned residuals vs fitted probabilities).
    - But the evidence is weak.
- It may be driven by non-linearity of the effect of h on the log-odds.
    - Or it could be driven by a missing covariate.

Other considerations i

- **Calibration**: Are the estimated probabilities close to empirical probabilities?
    - Hosmer-Lemeshow, Brier score
- **Discrimination**: Are cases more likely to be given large scores (or large probabilities) than non-cases?
    - Area under the ROC curve (AUC), Percentage of Correct Predictions (PCP)
    - **Note**: the AUC is not a very sensitive measure of model performance.

Other considerations ii

```
performance_hosmer(model)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 3.305
##            df: 8
##       p-value: 0.914
```

Other considerations iii

```
performance_score(model) # Quadratic = Brier
```

```
## # Proper Scoring Rules
##
## logarithmic:   -Inf
##   quadratic: 8.1783
##   spherical: 0.0280
```

```
performance_roc(model)
```

```
## AUC: 92.73%
```

```r
performance_pcp(model)
```

```
## # Percentage of Correct Predictions from
Logistic Regression Model
##
## Full model: 81.53% [76.07% - 86.99%]
## Null model: 54.78% [47.78% - 61.79%]
##
## # Likelihood-Ratio-Test
##
## Chi-squared: 133.685
## p-value: 0.000
```

Summary

- Logistic regression is an extension of linear regression for binary outcomes.
    - Easily extended to any binomial outcome.
- Instead of measuring differences in means, regression coefficients measure differences in log-odds.
    - But $\beta = 0$ still corresponds to no association!
- Residual analysis is more complicated.
    - **Key**: Binned residuals.
- As a prediction model, logistic regression is surprisingly powerful.
    - Neural networks can be seen as a generalization.