Simple Logistic Regression

Biostat 515/518
Discussion – Week 4

David Clausen

Adapted from notes by Anu Mishra
University of Washington

PSA Study

- Goal of study was to assess if PSA can be used to identify those patients in whom cancer is progressing
- Prospective cohort study of men who have received hormonal therapy for prostate cancer
- Followed for at least 24 months
- Lowest PSA and cancer severity measured

Scientific Question

- **Is PSA nadir (the lowest value observed post therapy) highly associated with time to relapse?**

PSA Data

- What are the relevant variables for this scientific question?

PSA Data

- What are the relevant variables for this scientific question?
  - PSA Nadir (continuous, uncensored)
  - Time in remission / time to relapse (continuous, possibly censored)
  - Indicator of relapse status (binary)

- What are valid analysis approaches?

Analysis Approaches

- Approach 1:
  Logistic regression – binary predictor
  - Response: Indicator of relapse within 24 months
  - Predictor: Dichotomized PSA nadir
  - Statistical question: Are the odds of relapse within 24 months different for those with high PSA nadir compared to those with low PSA nadir.
- Drawbacks?
  - Cut-off may be arbitrary (what is “high” or “low” PSA?)

Analysis Approaches

- Approach 2:
  Logistic regression – continuous predictor
  - Response: Indicator of relapse status at 24 months
  - Predictor: PSA nadir
  - Statistical question: Are the odds of relapse with 24 months different for those with different PSA levels?

- Drawbacks?
  - Slightly harder to interpret (but we'll go over this!)

Review of Terms

- Probability of event occurring (remission at last followup)

$$P(Y_i = 1) = p_i$$

- Odds of event occurring: Ratio of probabilities

$$\text{odds} = \frac{p_i}{1 - p_i}$$

- Odds ratio: Ratio of odds of event occurring to odds of event not occurring

$$OR = \frac{\text{odds event in group 1}}{\text{odds event in group 2}}$$

Logistic Regression Review

- Uses the model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_i$$

- Parameter interpretations
  - log odds for X = 0 : β_0
  - log odds for X = x : $\beta_0 + \beta_1 * x$
  - log odds for X = x+ 1 : $\beta_0 + \beta_1 * (x + 1)$

Logistic Regression Review

- Parameter interpretation (cont.)
  - Odds of event for X=x: $\exp(\beta_0 + \beta_1 * x)$
  - Odds of event for X=x+1: $\exp(\beta_0 + \beta_1 * (x+1))$
  - Odds ratio comparing groups:

$$\frac{\text{odds of event for X=x+1}}{\text{odds of event for X=x}} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)}$$

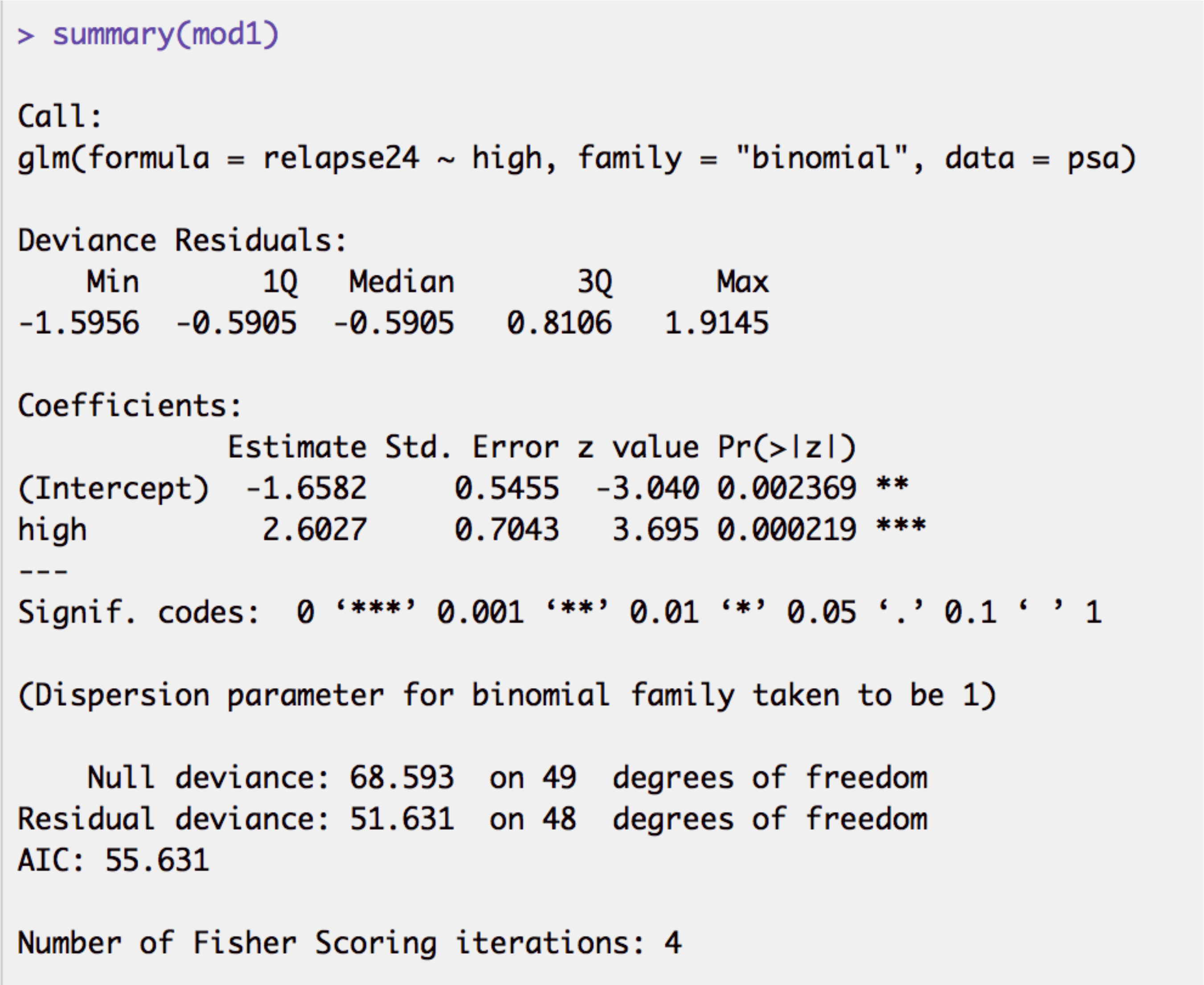
$$= \frac{\exp(\beta_0 + \beta_1 x + \beta_1)}{\exp(\beta_0 + \beta_1 x)}$$

$$= \exp(\beta_1)$$

Approach 1: Application

```
#clear objects from workspace
rm(list=ls())

#set working directory
setwd("/Users/davidclausen/Dropbox/BIOST 515/Discussion")

#read in data
psa <- read.table('psa.txt',header=T)

#create indicator of relapse within 24 months
psa$relapse24 <- ifelse(psa$inrem=="no"&psa$obstime<=24,1,0)

#create dichotomized PSA variable
psa$high <- ifelse(psa$nadirpsa>=median(psa$nadirpsa),1,0)

#logistic regression of relapse status on dichotomized PSA nadir
mod1 <- glm(relapse24~high,family='binomial',data=psa)summary(mod1)
```

Approach 1: Application



```
> summary(mod1)

Call:
glm(formula = relapse24 ~ high, family = "binomial", data = psa)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5956  -0.5905  -0.5905   0.8106   1.9145

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6582     0.5455  -3.040 0.002369 **
high          2.6027     0.7043   3.695 0.000219 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.593  on 49  degrees of freedom
Residual deviance: 51.631  on 48  degrees of freedom
AIC: 55.631

Number of Fisher Scoring iterations: 4
```

Approach 1: Results

- Results
  - OR : exp(2.60) = 13.5
  - 95% CI : [exp(1.22) , exp(3.98)]

    = [3.40,53.7]
  - P value: 0.000219

- Note: Above analysis does not use robust standard errors, but could use them here.

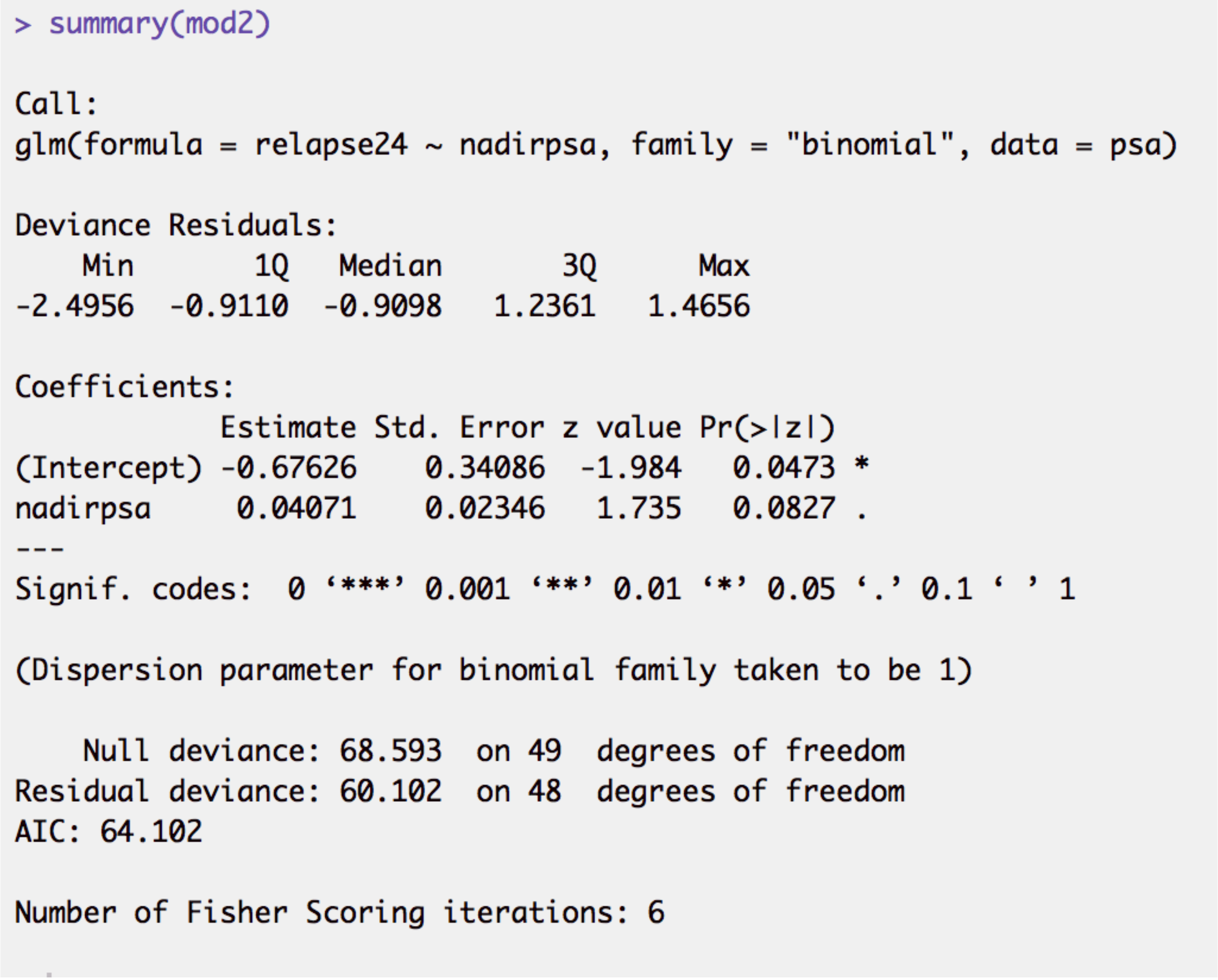
Approach 1: Results

- The estimated odds of relapse within 24 months among prostate cancer patients with above-median PSA nadir level are 13.5 times higher relative to a group of prostate cancer patients with below-median PSA nadir level. Based on a 95% CI it would not be unusual to observe an OR between 3.40 and 53.7. With a p-value of 0.000219 we find this result significant at the 0.05 level.

Approach 2: Application

```
#logistic regression of relapse status on (continuous) PSA
nadirmod2 <- glm(relapse24~nadirpsa,family='binomial',data=psa)

#extract point estimate and compute 95% CI for PSA effect
mod2.pointest <- exp(summary(mod2)$coefficients["nadirpsa","Estimate"])

mod2.95ci <- exp(summary(mod2)$coefficients["nadirpsa","Estimate"] +
        c(-1,1)*qnorm(.975)*summary(mod2)$coefficients["nadirpsa","Std. Error"])
```

Approach 2: Application



```
> summary(mod2)

Call:
glm(formula = relapse24 ~ nadirpsa, family = "binomial", data = psa)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4956  -0.9110  -0.9098   1.2361   1.4656

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.67626    0.34086  -1.984   0.0473 *
nadirpsa     0.04071    0.02346   1.735   0.0827 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.593  on 49  degrees of freedom
Residual deviance: 60.102  on 48  degrees of freedom
AIC: 64.102

Number of Fisher Scoring iterations: 6
```

Approach 2: Results

- Results
  - OR : exp(0.041) = 1.04
  - 95% CI : [exp(-0.0052) , exp(0.087)]

    = [0.995, 1.09]
  - P value: 0.0827
- Note: Above analysis does not use robust standard errors, but could use them here.

Approach 2: Results

The estimated odds of relapse within 24 months in a group of prostate cancer patients are 4% higher relative to a group of prostate cancer patients with a 1 ng/ml lower PSA nadir level. Based on a 95% CI it would not be unusual to observe an OR between 0.995 and 1.09. With a p-value of 0.08 we find this result is not significant at the 0.05 level.

Approach 3: Using $\log_2$(PSA)

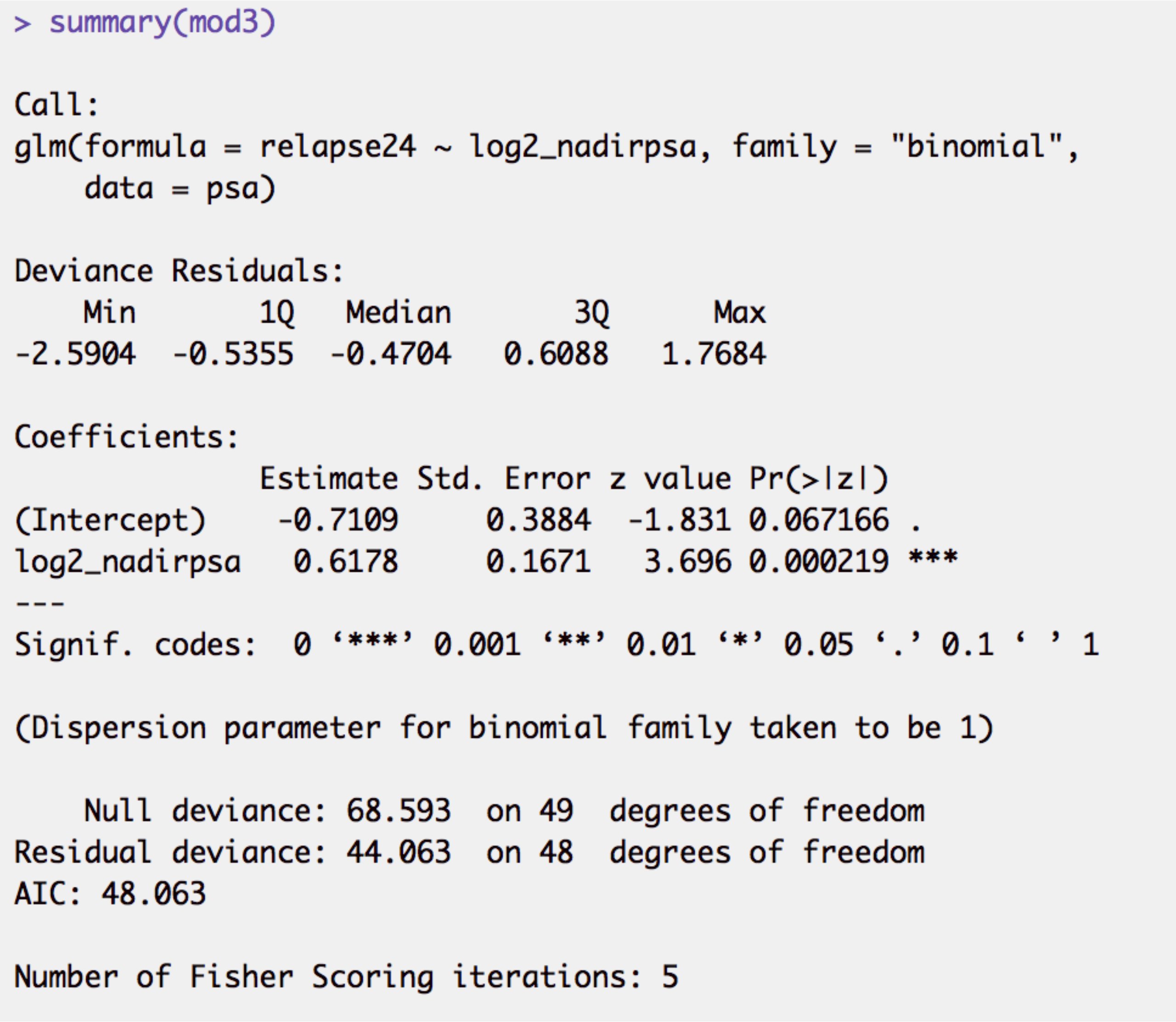
- In the previous example we compared groups on an additive scale (1 unit different in PSA)

- If we want wanted to compare groups on a multiplicative scale we can use a log-transformed predictor

Approach 3: Application

```
#create log_2 PSA nadir variable
psa$log2_nadirpsa <- log(psa$nadirpsa)/log(2)

#logistic regression of relapse status on log PSA nadir
mod3 <- glm(relapse24~log2_nadirpsa,family="binomial",data=psa)

#extract point estimate and compute 95% CI for PSA nadir effect
mod3.pointest <- exp(summary(mod3)$coefficients["log2_nadirpsa","Estimate"])

mod3.95ci <- exp(summary(mod3)$coefficients["log2_nadirpsa","Estimate"]  +
        c(-1,1)*qnorm(.975)*summary(mod2)$coefficients["log2_nadirpsa","Std. Error"])
```

Approach 3: Application



```
> summary(mod3)

Call:
glm(formula = relapse24 ~ log2_nadirpsa, family = "binomial",
    data = psa)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5904  -0.5355  -0.4704   0.6088   1.7684

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.7109     0.3884  -1.831 0.067166 .
log2_nadirpsa   0.6178     0.1671   3.696 0.000219 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.593  on 49  degrees of freedom
Residual deviance: 44.063  on 48  degrees of freedom
AIC: 48.063

Number of Fisher Scoring iterations: 5
```

Approach 3: Results

The estimated odds of relapse within 24 months in a group of prostate cancer patients are 1.85 times the odds of relapse for group of prostate cancer patients with a PSA nadir twice as low (two-fold decrease). Based on a 95% CI it would not be unusual to observe an OR between 1.34 and 2.57. With a p-value less than 0.001 we find this result to be significant and reject the null hypothesis.

Summary

- Logistic regression requires a binary dependent/response variable

- Without a good scientific reason, dichotomization of continuous predictors is not recommended.

- Choice of transformation of independent variable depends the scientific question (additive or multiplicative change).