Chapter 2. Binary Response

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Contents

§2.1 Heart Disease Example (1)

- What might affect the chance of getting heart disease?
- The Western Collaborative Group Study (WCGS) started in 1960 and used 3154 healthy men, aged from 39 to 59, from the San Francisco area. At the beginning, all were free of heart disease.
- 8.5 years later, the study recorded whether these men now suffered from heart disease along with many other variables that might be related to the chance of developing this disease.
- A subset of the WCGS data is used.

```
> data(wcgs, package="faraway")
> help(wcgs)
```

- *Reference:* Coronary Heart Disease in the WCGS Final Follow-up Experience of 8 1/2 Years. Rosenman et al; JAMA. 1975 **233**(8):872-877. **doi**:10.1001/jama.1975.03260080034016.

§2.1 Description of the wcgs data

A data frame with 3154 observations on the following 13 variables:

age:	age in years
height:	height in inches
weight:	weight in pounds
sdp:	systolic blood pressure in mm Hg
dbp:	diastolic blood pressure in mm Hg
chol:	Fasting serum cholesterol in mm %
behave:	behavior type, a factor with levels A1 A2 B3 B4
cigs:	number of cigarettes smoked per day
dibep:	behavior type, a factor with levels A (Agressive) B (Passive)
chd:	coronary heart disease developed, a factor with levels no yes
typechd:	CHD type, with levels angina infdeath none silent
timechd:	Time of CHD event or end of follow-up
arcus:	arcus senilis is a factor with levels absent present

§2.1 Initial analysis of the wcgs data (1)

```
> dim(wcgs); head(wcgs)
```

```
[1] 3154   13
      age height weight sdp dbp chol behave cigs dibep chd  typechd timechd   arcus
2001   49     73    150 110  76  225     A2   25     B  no     none    1664  absent
2002   42     70    160 154  84  177     A2   20     B  no     none    3071 present
2003   42     69    160 110  78  181     B3    0     A  no     none    3071  absent
2004   41     68    152 124  78  132     B4   20     A  no     none    3064  absent
2005   59     70    150 144  86  255     B3   20     A yes infdeath    1885 present
2006   44     72    204 150  90  182     B4    0     A  no     none    3102  absent
```

```
> summary(wcgs)
```

```
      age            height          weight          sdp             dbp             chol      
 Min.   :39.00   Min.   :60.00   Min.   : 78   Min.   : 98.0   Min.   : 58.00   Min.   :103.0  
 1st Qu.:42.00   1st Qu.:68.00   1st Qu.:155   1st Qu.:120.0   1st Qu.: 76.00   1st Qu.:197.2  
 Median :45.00   Median :70.00   Median :170   Median :126.0   Median : 80.00   Median :223.0  
 Mean   :46.28   Mean   :69.78   Mean   :170   Mean   :128.6   Mean   : 82.02   Mean   :226.4  
 3rd Qu.:50.00   3rd Qu.:72.00   3rd Qu.:182   3rd Qu.:136.0   3rd Qu.: 86.00   3rd Qu.:253.0  
 Max.   :59.00   Max.   :78.00   Max.   :320   Max.   :230.0   Max.   :150.00   Max.   :645.0  
                                                                                NA's   :12     
 behave       cigs       dibep     chd          typechd       timechd          arcus      
 A1: 264   Min.   : 0.0   A:1565   no :2897   angina  :  51   Min.   :  18   absent :2211  
 A2:1325   1st Qu.: 0.0   B:1589   yes: 257   infdeath: 135   1st Qu.:2842   present: 941  
 B3:1216   Median : 0.0                       none    :2897   Median :2942   NA's   :   2  
 B4: 349   Mean   :11.6                       silent  :  71   Mean   :2684                 
           3rd Qu.:20.0                                       3rd Qu.:3037                 
           Max.   :99.0                                       Max.   :3430                 
```

§2.1 Initial analysis of the wcgs data (2)

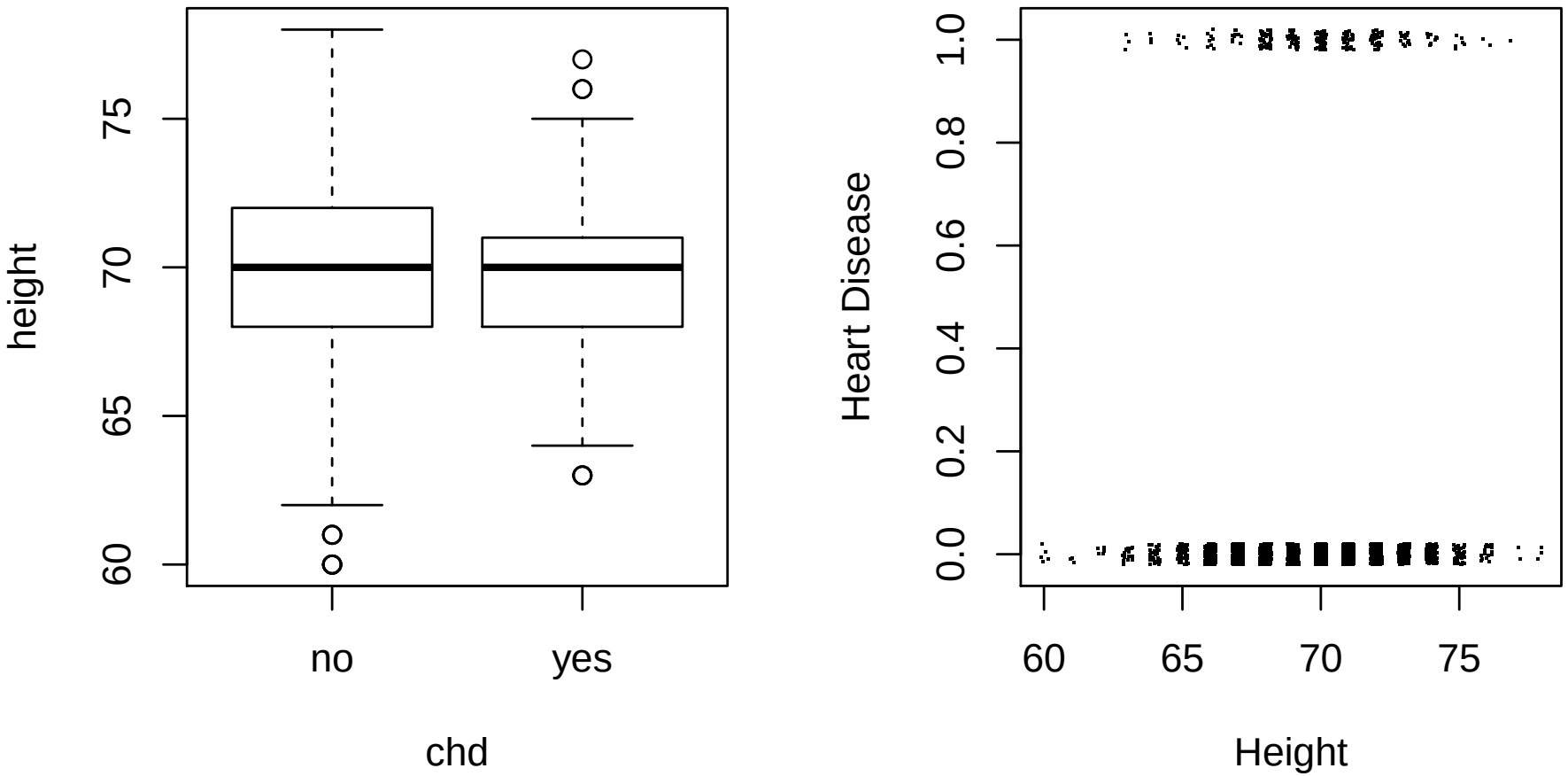

Figure 2.1: Plots of the presence/absence of CHD according to height in inches.

```
> plot(height ~ chd, wcgs); wcgs$y <- ifelse(wcgs$chd == "no",0,1)
> plot(jitter(y,0.1)~jitter(height),wcgs,xlab="Height",ylab="Heart Disease",pch=".")
```

§2.1 Initial analysis of the wcgs data (3)

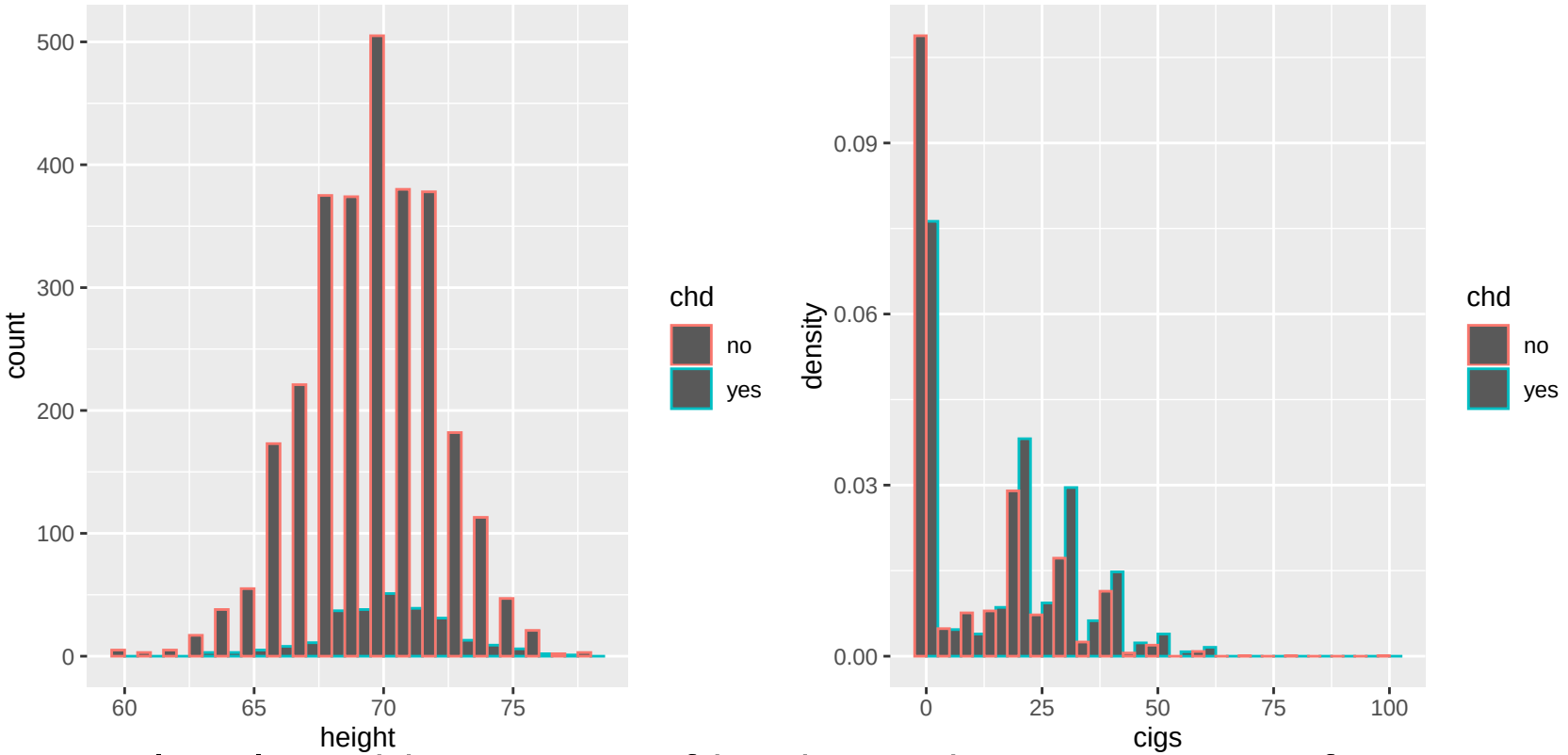

Figure 2.2: Interleaved histograms of heights and cigarette usage for men with and without chd.

```
> library(ggplot2)
> ggplot(wcgs,aes(x=height,color=chd))+geom_histogram(position="dodge",binwidth=1)
> ggplot(wcgs, aes(x=cigs, color=chd)) + geom_histogram(position="dodge",
                        binwidth=5, aes(y=..density..))
```

§2.1 Initial analysis of the wcgs data (4)

```
> ggplot(wcgs, aes(x=height,y=cigs)) + geom_point(alpha=0.2, position
            =position_jitter()) + facet_grid(~ chd)
```

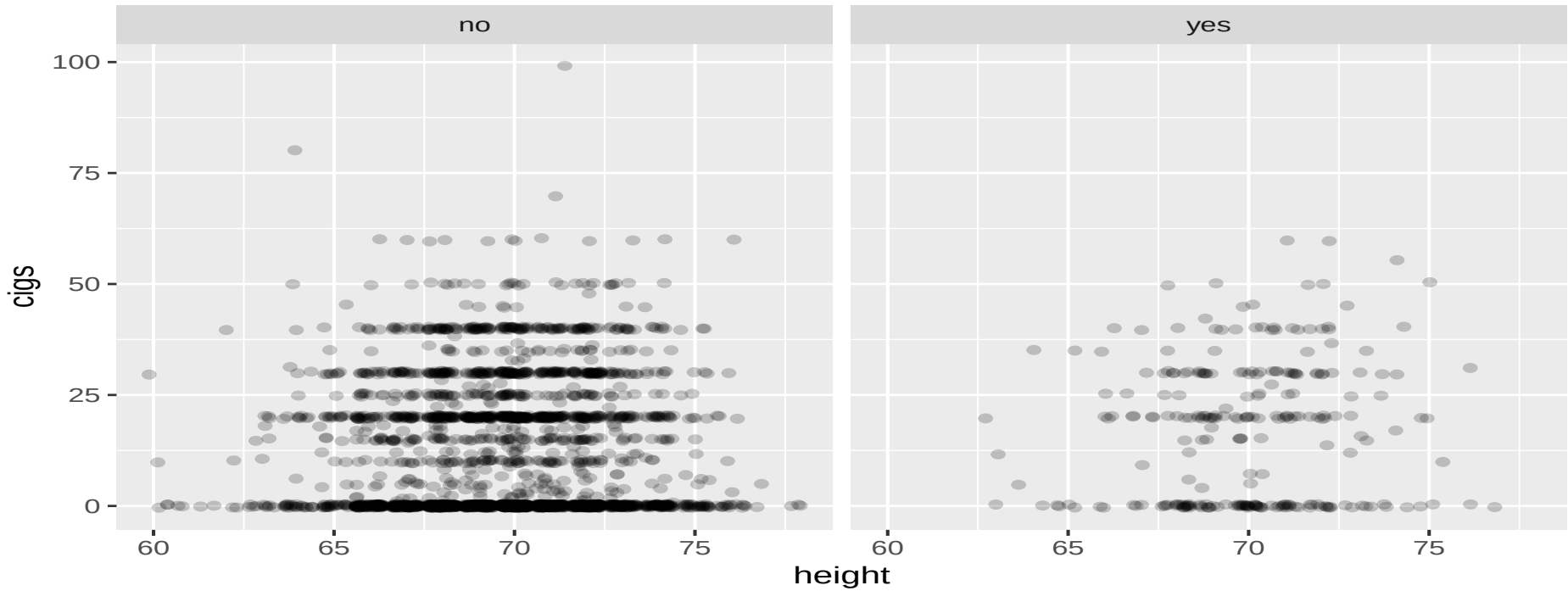

Figure 2.3: Height and cigarette consumption for men without CHD on the left and with CHD on the right. Some jittering and transparency have been used to reduce overplotting problems.

§2.1 Initial analysis of the wcgs data (5)

- In the first panel of Figure 2.2, the two histograms show the distributions of `height` for both those with and without CHD are similar. Therefore, `height` does not seem to have significant effect on changing the risk of CHD.
- In the second panel of Figure 2.2, the two density histograms show the distributions of `cigs` for both those with and without CHD are not similar. This suggests men smoking more cigarettes are more likely to have CHD.
- The above findings can be confirmed in Figure 2.3.

§2.1 Initial analysis of the `wcgs` data (6)

- One problem of our interest is to **predict** the heart disease outcome for a given individual and also to **explain the relationship** between height, cigarette usage and heart disease.
- We observe that, for the same height and cigarette consumption, both outcomes occur. This occurs quite regularly. Hence it makes better sense to model the probability of getting heart disease rather than the outcome itself.
- This model, however, *cannot be obtained by fitting a linear model* to the data where the response variable (`chd`) is binary, because the fitted value from a linear model cannot be binary.

§2.2 Logistic regression (1)

- Suppose the response variable Y is binary with

$$P(Y = 1) = p \quad \text{and} \quad P(Y = 0) = 1 - p.$$

  Thus Y follows a Bernoulli (a special binomial) distribution.
- There are q predictors $x_1, \cdots, x_q$, and we want to see how p is related to these predictors.
- There are n independent sample observations of $(Y, x_1, \cdots, x_q)$: $(Y_i, x_{i1}, \cdots, x_{iq})$, $i = 1, \cdots, n$.
- A **logistic regression** model consists of the following 3 components:
  1. **Distribution** Y_i's have independent $\text{Bin}(1, p_i)$ distributions.
  2. **Linear predictor** $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$
  3. **Link** $\eta_i = \log \frac{p_i}{1-p_i}$ or equivalently $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

  In short

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

§2.2 Logistic regression (2)

- Note the logistic regression model is **conceptually different** from the linear model

$$\log \frac{Y_i}{1-Y_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

  which is not properly defined for binary Y_i.
- A fitted logistic regression model returns a fitted linear predictor

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$$

  then a fitted probability value $\hat{p}_i = \dfrac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$, from which a fitted response $\hat{Y}_i$ can be randomly generated from Bin(1, $\hat{p}_i$).
- The link function $\eta = \log \dfrac{p}{1-p}$ is called the **logistic link**, and can be computed using `logit` or `ilogit` in `faraway`.

```
> curve(ilogit(x),-6,6, xlab=expression(eta), ylab="p")
```

§2.2 Logistic regression (3)

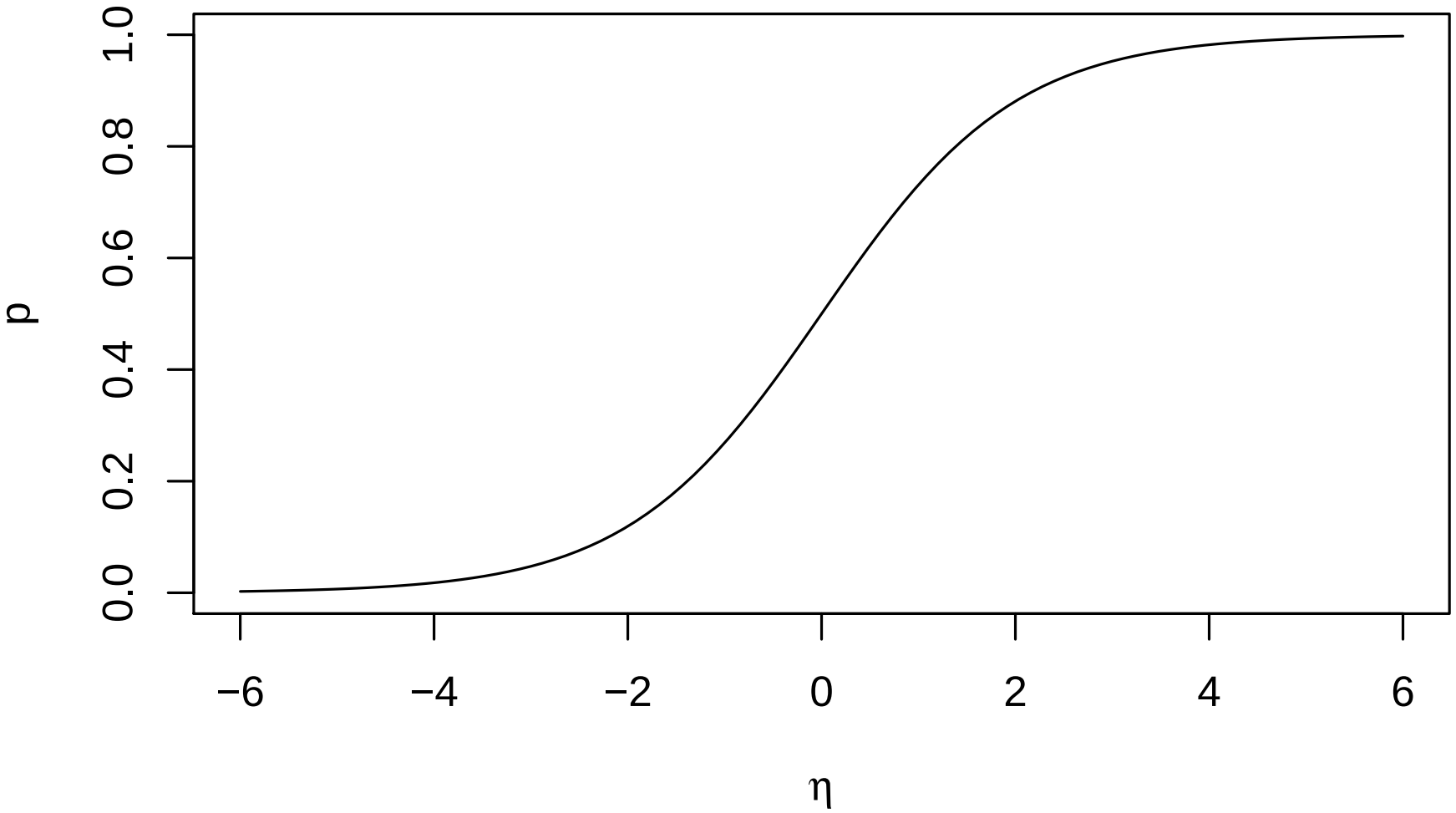

Figure 2.4: A logistic relationship between the probability of the response, p, and the linear predictor, η.

§2.2 Logistic regression: Parameter estimation (1)

- The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)^\top$ in the logistic regression model are estimated by the method of **maximum likelihood**.
- The log-likelihood function is

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [Y_i \log p_i + (1 - Y_i) \log(1 - p_i)] = \sum_{i=1}^{n} [Y_i \eta_i - \log(1 + e^{\eta_i})] \\
&= \sum_{i=1}^{n} \left[Y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}) - \log\left(1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}}\right) \right] \\
&= \sum_{i=1}^{n} \left[Y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log\left(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right) \right], \quad \text{where} \quad \mathbf{x}_i = (1, x_{i1}, \cdots, x_{iq})^\top
\end{aligned}$$

- **Score function**

$$\mathbf{u}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right) \mathbf{x}_i = \sum_{i=1}^{n} (Y_i - p_i) \mathbf{x}_i = X^\top (\mathbf{y} - \mathbf{p})$$

where $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$, $\mathbf{y} = (Y_1, \cdots, Y_n)^\top$ and $\mathbf{p} = (p_1, \cdots, p_n)^\top$.

§2.2 Logistic regression: Parameter estimation (2)

- **Hessian function**

$$\begin{aligned} H(\beta) &= \frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} \frac{e^{\mathbf{x}_i^\top \beta}}{\left(1+e^{\mathbf{x}_i^\top \beta}\right)^2} \mathbf{x}_i \mathbf{x}_i^\top = -\sum_{i=1}^{n} p_i(1-p_i)\mathbf{x}_i \mathbf{x}_i^\top \\ &= -X^\top \text{diag}\left\{(p_1(1-p_1), \cdots, p_n(1-p_n)\right\} X \end{aligned}$$

- **Observed information** $J(\beta) = -H(\beta)$.
- **Fisher information** $I(\beta) = -E\left[H(\beta)\right] = \text{Var}(\mathbf{u}(\beta))$.
- For logistic regression, $I(\beta) = J(\beta) = -H(\beta)$.
- **MLE** $\hat{\beta}$ is solved by *Newton-Raphson* or *Fisher scoring*

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left[H(\hat{\beta}^{(k)})\right]^{-1} \mathbf{u}(\hat{\beta}^{(k)}) \quad \text{or} \quad \hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left[I(\hat{\beta}^{(k)})\right]^{-1} \mathbf{u}(\hat{\beta}^{(k)})$$

- **Estimated variance** $\widehat{\text{Var}}(\hat{\beta}) = \left[I(\hat{\beta})\right]^{-1}$. Also $\hat{\beta} \overset{d}{\approx} N\left(\beta,\ [I(\beta)]^{-1}\right)$.

§2.2 Logistic regression model fit (1)

- We fit a logistic regression model to study the relationship between chd and height and cigarette usuage based on data wcgs

```
> lmod <- glm(chd ~ height + cigs, family = binomial, wcgs)
> beta <- coef(lmod); summary(lmod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0041  -0.4425  -0.3630  -0.3499   2.4357
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
---
    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1749.0  on 3151  degrees of freedom
AIC: 1755;     Number of Fisher Scoring iterations: 5

> attributes(summary(lmod))

$`names`
 [1] "call"           "terms"          "family"         "deviance"
 [5] "aic"            "contrasts"      "df.residual"    "null.deviance"
 [9] "df.null"        "iter"           "deviance.resid" "coefficients"
[13] "aliased"        "dispersion"     "df"             "cov.unscaled"
[17] "cov.scaled"
```

§2.2 Logistic regression model fit (2)

```
> sumary(lmod)

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -4.5016140  1.8418627 -2.4441   0.01452
height       0.0252078  0.0263274  0.9575   0.33833
cigs         0.0231274  0.0040402  5.7243 1.038e-08

n = 3154 p = 3
Deviance = 1749.04923 Null Deviance = 1781.24374 (Difference = 32.19451)

> plot(jitter(y,0.1) ~ jitter(height), wcgs, xlab="Height", ylab="Heart Disease",
                    pch=".")
> curve(ilogit(beta[1] + beta[2]*x + beta[3]*0),add=TRUE)
> curve(ilogit(beta[1] + beta[2]*x + beta[3]*20),add=TRUE,lty=2)
> plot(jitter(y,0.1) ~ jitter(cigs), wcgs, xlab="Cigarette Use",
                      ylab="Heart Disease",pch=".")
> curve(ilogit(beta[1] + beta[2]*60 + beta[3]*x),add=TRUE)
> curve(ilogit(beta[1] + beta[2]*78 + beta[3]*x),add=TRUE,lty=2)
```

§2.2 Logistic regression model fit (3)

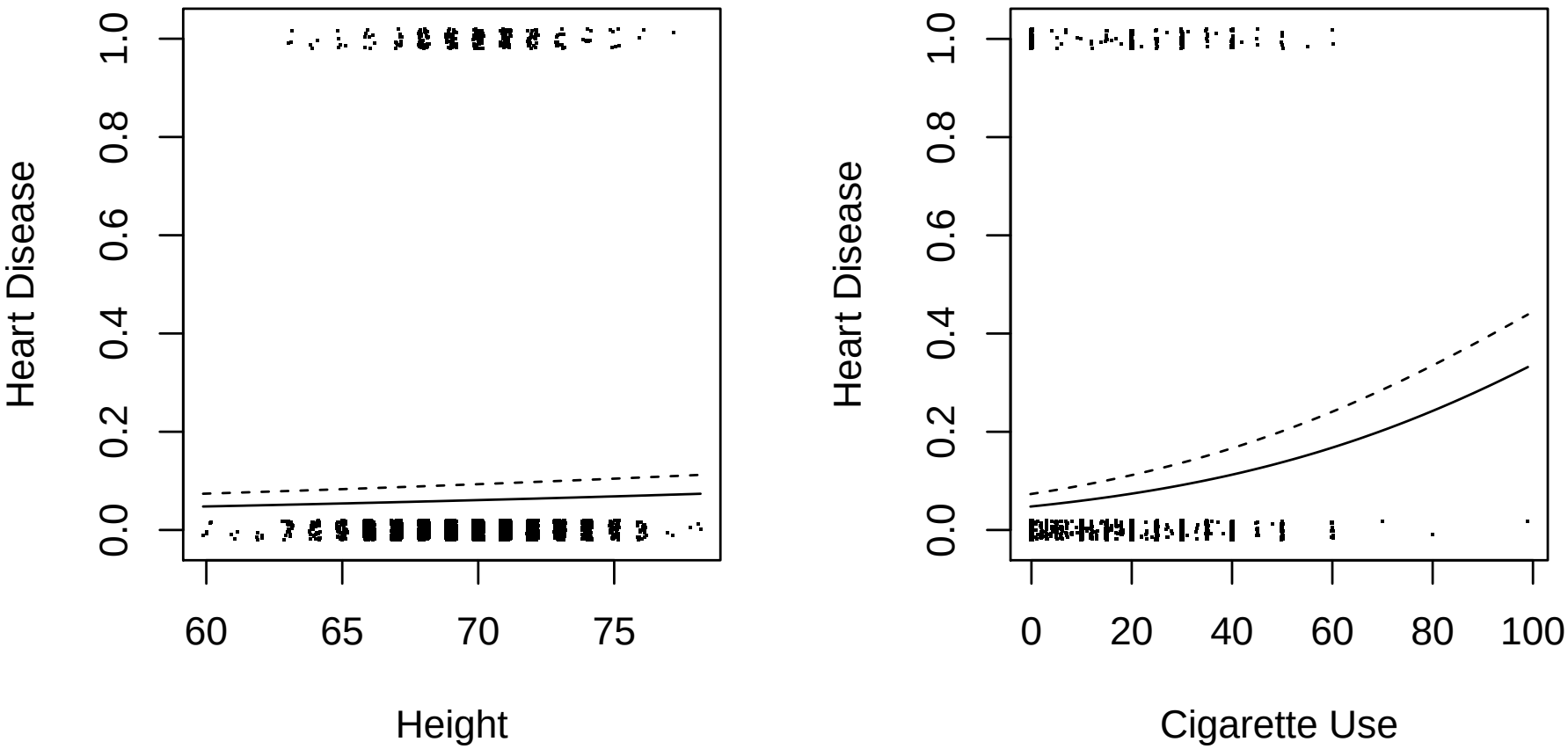

Figure 2.5: Predicted probability of heart disease as height and cigarette consumption vary. In the first panel, the solid line represents a nonsmoker, while the dashed line is a pack-a-day smoker. In the second panel, the solid line represents a very short man (60 in. tall) while the dashed line represents a very tall man (78 in. tall.)

§2.2 Logistic regression model interpretation

- The mathematical form of the model `lmod` is

$$\begin{aligned}\log \widehat{\text{odds}} = \log \frac{\hat{p}}{1-\hat{p}} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot \texttt{height} + \hat{\beta}_2 \cdot \texttt{cigs} \\ &= -4.502 + 0.025 \cdot \texttt{height} + 0.023 \cdot \texttt{cigs}\end{aligned}$$

Equivalently,

$$\begin{aligned}\widehat{\text{odds}} = \frac{\hat{p}}{1-\hat{p}} &= e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 \cdot \texttt{height}} \cdot e^{\hat{\beta}_2 \cdot \texttt{cigs}} \\ &= e^{-4.502} \cdot e^{0.025 \cdot \texttt{height}} \cdot e^{0.023 \cdot \texttt{cigs}} \\ &= 0.011 \times 1.026^{\texttt{height}} \times 1.023^{\texttt{cigs}}\end{aligned}$$

- $\hat{\beta}_1$ is interpreted as the increase of the estimated log-odds when `height` increases by one unit (i.e. 1 inch). $\hat{\beta}_2$ can be similarly interpreted w.r.t. `cigs`.
- e^{β_1} and e^{β_2} are relevant **odds ratio** values.
- The estimated odds ratio of heart disease for a man against another man 1 inch shorter is 1.026. The estimated odds of `chd` increase by 2.3% with each additional cigarette smoked per day.

§2.3 Logistic regression: Hypothesis testing (1)

- Possible significance of any set of the predictors on the response in a logistic regression model may be assessed by testing a *linear hypothesis* about the relevant β parameters in the model.

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad \text{vs.} \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}$$

  where C is a known matrix of full row rank, and $\boldsymbol{\xi}$ is a known vector.

- As a special case of the above, if we want to know whether a model Ω can be replaced by one of its sub-models, ω, meaning the predictors not belonging to ω have no effect on the response, we can test

$$H_0 : \boldsymbol{\beta}_{\Omega-\omega} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_{\Omega-\omega} \neq \mathbf{0}$$

  where $\boldsymbol{\beta}_{\Omega-\omega}$ is the subset of $\boldsymbol{\beta}$ in Ω but not in ω.

§2.3 Logistic regression: Hypothesis testing (2)

- Three types of tests are available for testing $H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi}$ vs. $H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}$:

  Likelihood ratio (LR) test, **Wald test**, and **score test**.

- **LR test statistic**: $\lambda = -2\left\{\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})\right\}$, where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0, and $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_1; also $\hat{\boldsymbol{\beta}}$ is the unrestricted MLE of $\boldsymbol{\beta}$.
- **Wald test statistic**: $W = \left(C\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right)^T \left[CI^{-1}(\hat{\boldsymbol{\beta}})C^T\right]^{-1} \left(C\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right)$, where $\hat{\boldsymbol{\beta}}$ is the unrestricted MLE of $\boldsymbol{\beta}$.
- **Score test statistic**: $U = \mathbf{u}^T(\tilde{\boldsymbol{\beta}})I^{-1}(\tilde{\boldsymbol{\beta}})\mathbf{u}(\tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0.
- Under H_0, all 3 test statistics asymptotically follow a $\chi^2(s)$ distribution with degrees of freedom $s = \text{rank}(C)$.

§2.3 Logistic regression: Hypothesis testing (3)

Remarks

1. In testing $H_0 : \beta_j = \xi$ vs. $H_1 : \beta_j \neq \xi$, Wald test is equivalent to

$$\sqrt{W} = \frac{\hat{\beta}_j - \xi}{\text{s.e.}(\hat{\beta}_j)} \overset{d}{\approx} N(0,1) \text{ under } H_0,$$

the p-value of which can be read off from the R output when $\xi = 0$.

2. In testing $H_0 : \boldsymbol{\beta}_{\Omega-\omega} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta}_{\Omega-\omega} \neq \mathbf{0}$, the LR test becomes

$$\lambda = D_\omega - D_\Omega \overset{d}{\approx} \chi^2(s) \text{ under } H_0 \text{ with } s = \dim(\boldsymbol{\beta}_{\Omega-\omega}),$$

where D_Ω= **deviance** of model Ω, and D_ω = **deviance** of model ω.

3. $D_\omega = -2\sum_{i=1}^{n} \left[Y_i \log \hat{p}_{i\omega} + (1 - Y_i) \log(1 - \hat{p}_{i\omega}) \right]$, and

$$D_\Omega = -2\sum_{i=1}^{n} \left[Y_i \log \hat{p}_{i_\Omega} + (1 - Y_i) \log(1 - \hat{p}_{i_\Omega}) \right], \text{ where } \hat{p}_{i\omega} \text{ (or } \hat{p}_{i_\Omega})$$

is computed from $\hat{\boldsymbol{\beta}}_\omega$ (or $\hat{\boldsymbol{\beta}}_\Omega$) based on model ω (or Ω).

§2.3 Logistic regression: Hypothesis testing (4)

Remarks (continued)

- In other examples of GLMs, the deviance is a measure of how well the model fit the data. But in the case where the response is binary, the deviance cannot be used for measuring goodness of fit, because the deviance is just a function of $\hat{p}_i$'s.
- `summary(lmod)` returns the `Residual deviance` (i.e. the deviance of model `lmod`) $D_{\text{lmod}} = 1749.0$ and the `Null deviance` (i.e. the deviance of the model having only an intercept term) $D_o = 1781.2$.
- $D_o - D_{\text{lmod}} = 32.2$ with p-value 1.0183×10^{-7}.

```
> 1-pchisq(32.2,2)

[1] 1.01826e-07
```

- This concludes that there is significant relationship between the predictors `height` and `cigs` and the response `chd`.
- Are both `height` and `cigs` significant to `chd`, or one of them is significant? Need further tests.

§2.3 Logistic regression: Hypothesis testing (5)

```
> lmodc <- glm(chd ~ cigs, family = binomial, wcgs)
> anova(lmodc,lmod, test="Chi")

Analysis of Deviance Table

Model 1: chd ~ cigs
Model 2: chd ~ height + cigs
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3152       1750
2      3151       1749  1  0.92025   0.3374
```

- The analysis of deviance table displays all the information of the LR test. We see that height is not significant in a model that already includes cigarette consumption, with p-value=0.3374.

§2.3 Logistic regression: Hypothesis testing (6)

- Can test all the predictors in the model using the `drop1` function:

```
> drop1(lmod,test="Chi")

Single term deletions

Model:
chd ~ height + cigs
       Df Deviance    AIC     LRT Pr(>Chi)
<none>      1749.0 1755.0
height  1   1750.0 1754.0  0.9202   0.3374
cigs    1   1780.1 1784.1 31.0695 2.49e-08 ***
```

- An alternative to the above LR test is the Wald test with the $\sqrt{W}$- or z-value, which is $\hat{\beta}/\text{se}(\hat{\beta})$, which is approximately follows $N(0,1)$:

```
> sumary(lmod)

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -4.5016140  1.8418627 -2.4441   0.01452
height       0.0252078  0.0263274  0.9575   0.33833
cigs         0.0231274  0.0040402  5.7243 1.038e-08
```

- Both tests show that `cigs` is significant. Often the deviance-based tests are preferred, especially with sparse data (Hauck-Donner effect).

§2.3 Logistic regression: confidence intervals for β_i

- Using normal approximation, a $100(1-\alpha)\%$ confidence interval for β_i would be

$$\hat{\beta}_i \pm z^{(\alpha/2)}\text{se}(\hat{\beta}_i), \quad \text{with } z^{(\alpha/2)} \text{ the upper } \tfrac{\alpha}{2}\text{th } N(0,1) \text{ quantile.}$$

- A normal approximation based 95% C.I. for β_1 (`height`) is

$$0.0252078 \pm 1.96 \cdot 0.0263274 = (-0.0263939,\ \ 0.0768095).$$

- This is very close to the 95% C.I. obtained from a *profile likelihood method*:

```
> confint(lmod)

Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -8.13475465 -0.91297018
height      -0.02619902  0.07702835
cigs         0.01514949  0.03100534
```

- Profile likelihood method is preferable due to Hauck–Donner effect.

§2.4 Diagnostics in binary regression (1)

- Regression diagnostics are for checking the model assumptions and identifying unusual data points.
- Residuals are the most important means for doing this.
- **Raw residuals**: $r_i^{(\text{raw})} = Y_i - \hat{p}_i$, $i = 1, \cdots, n$, where the probability fitted values $\hat{p}_i = \text{logit}^{-1}(\hat{\eta}_i) = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$ with $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$ being the fitted linear predictor values.
- Calculate $\hat{\eta}_i$ by `predict()`, and $\hat{p}_i$ by `predict( , type="response")` in `R`.
- **Deviance residuals** (the default ones in GLM) are defined as

$$r_i = \text{sign}(y_i - \hat{p}_i)\sqrt{-2\left[Y_i \log \hat{p}_i + (1-Y_i)\log(1-\hat{p}_i)\right]}, \quad i = 1, \cdots, n.$$

  Thus, $\sum_{i=1}^n r_i^2$ equals the deviance of the model.
- **Pearson residuals**: $r_i^{(P)} = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$, $i = 1, \cdots, n$.

§2.4 Diagnostics in binary regression (2)

```
> linpred <- predict(lmod)
> predprob <- predict(lmod, type="response")
> head(linpred)

     2001      2002      2003      2004      2005      2006
-2.083261 -2.274521 -2.762277 -2.324936 -2.274521 -2.686653

> head(predprob)

0.11073449 0.09325523 0.05939705 0.08907868 0.09325523 0.06376553

> head(ilogit(linpred))

     2001      2002      2003      2004      2005      2006
0.11073449 0.09325523 0.05939705 0.08907868 0.09325523 0.06376553

> rawres <- wcgs$y - predprob
> head(rawres)

      2001       2002       2003       2004       2005       2006
-0.11073449 -0.09325523 -0.05939705 -0.08907868  0.90674477 -0.06376553

> head(residuals(lmod, type="response"))

      2001       2002       2003       2004       2005       2006
-0.11073449 -0.09325523 -0.05939705 -0.08907868  0.90674477 -0.06376553
```

§2.4 Diagnostics in binary regression (3)

Calculate the deviance residuals and Pearson residuals.

```
> head(residuals(lmod, type="deviance"))

      2001       2002       2003       2004       2005       2006
-0.4844779 -0.4424800 -0.3499548 -0.4319693  2.1782631 -0.3630133

> head(residuals(lmod))

      2001       2002       2003       2004       2005       2006
-0.4844779 -0.4424800 -0.3499548 -0.4319693  2.1782631 -0.3630133

> head(residuals(lmod, type="pearson"))

2001       2002       2003       2004       2005       2006
-0.3528789 -0.3206964 -0.2512923 -0.3127134  3.1182141 -0.2609761
```

§2.4 Diagnostics in binary regression (4)

Create residual plots and binned residual plots

```
> ####Create Figure 2.6
> par(mfrow=c(1,2))
> plot(rawres ~ linpred, xlab="linear predictor", ylab="raw residuals")
> library(dplyr) ###Creat the binned residual plot
> wcgs <- mutate(wcgs, residuals=residuals(lmod), linpred=predict(lmod))
> gdf <- group_by(wcgs, cut(linpred, breaks=unique(quantile(linpred,(1:100)/101))))
> diagdf <- summarise(gdf, residuals=mean(residuals), linpred=mean(linpred))
> plot(residuals ~ linpred, diagdf, xlab="linear predictor")

> ####Create Figure 2.7 left panel
> gdf <- group_by(wcgs, height)
> diagdf <- summarise(gdf, residuals=mean(residuals))
> ggplot(diagdf, aes(x=height,y=residuals)) + geom_point()
> filter(wcgs, height==77) %>% select(height, cigs, chd, residuals)

  height cigs chd  residuals
1     77    0  no -0.3857933
2     77    0 yes  2.2956622
3     77    5  no -0.4078515

> ####Create Figure 2.7 right panel
> group_by(wcgs, cigs) %>% summarise(residuals=mean(residuals), count=n())
        %>% ggplot(aes(x=cigs, y=residuals, size=sqrt(count)))+geom_point()
```

§2.4 Diagnostics in binary regression (5)

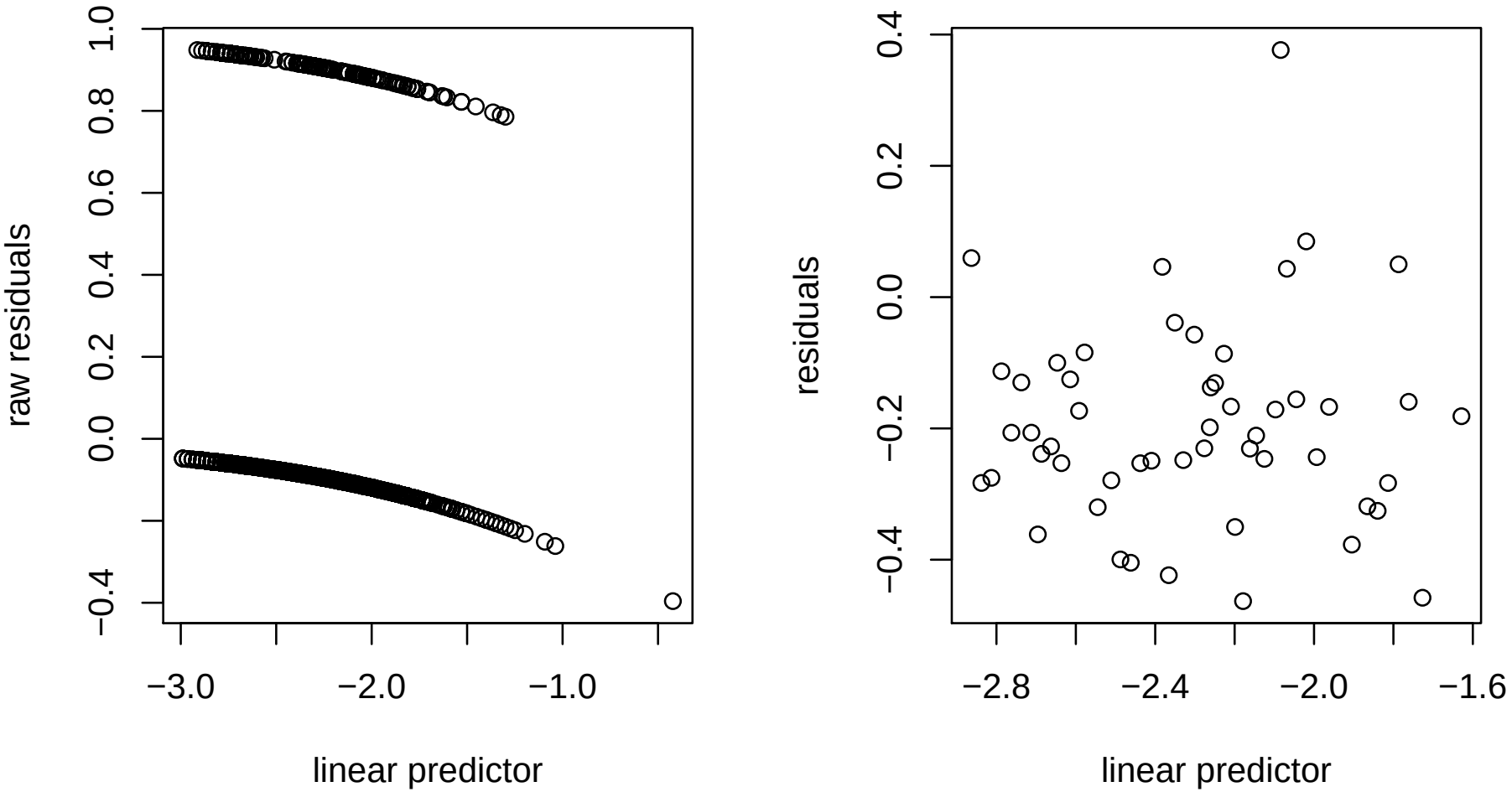

Figure 2.6: The panel on the left shows the raw residuals and linear predictor. The two lines are due to the binary response. The panel on the right shows the binned version of the plot, which reveals **no inadequacy of the model**.

§2.4 Diagnostics in binary regression (6)

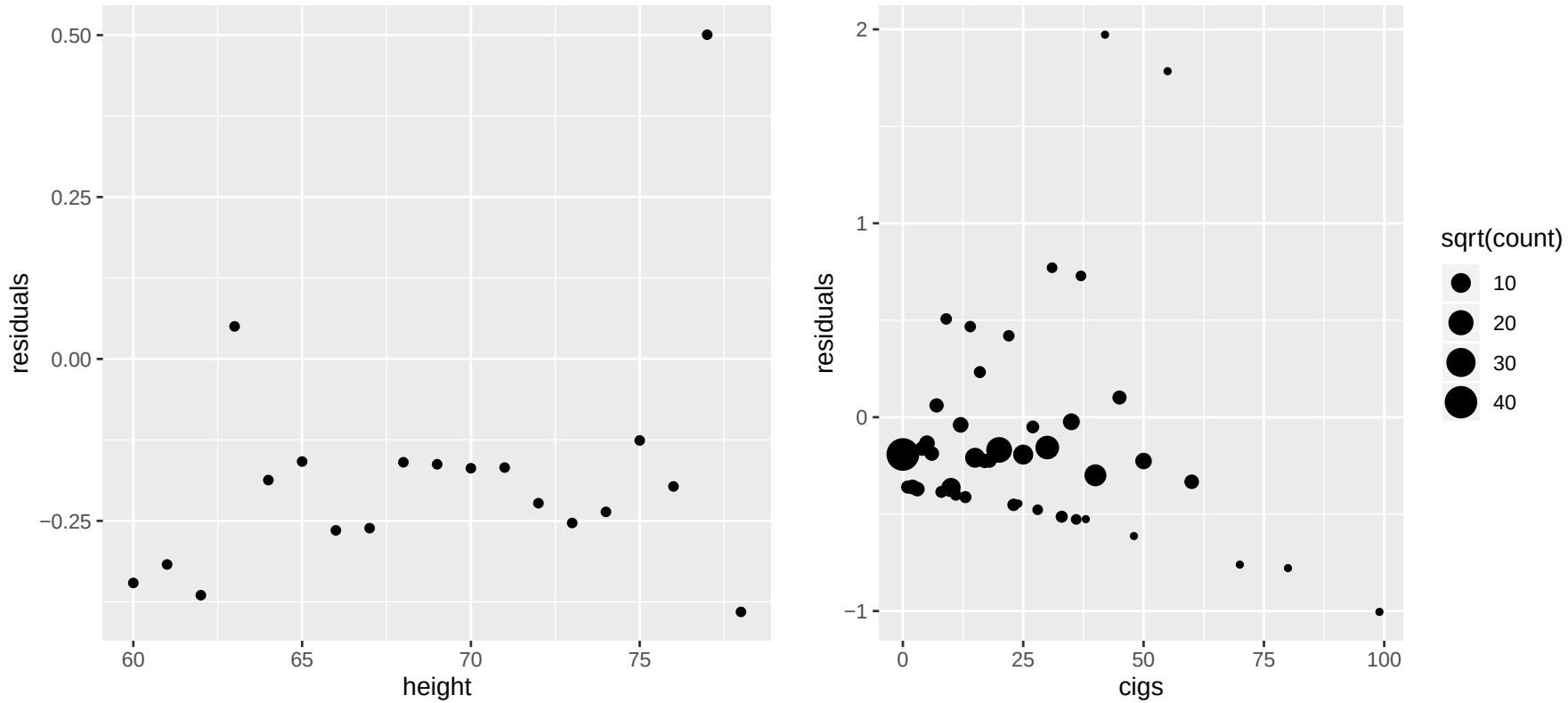

Figure 2.7: Binned residuals plots for the predictors. Left plot reveals nothing remarkable except for a large residual at height=77in. Right plot reveals a few points with large residuals but their corresponding bin sizes are small, so not of major concern.

§2.4 Diagnostics in binary regression (7)

- QQ plot of the residuals is for checking their normality or not .
  `> qqnorm(residuals(lmod))`
- Half-normal plot of hat values is for finding large leverage points.
  `>halfnorm(hatvalues(lmod))`
  `> filter(wcgs, hatvalues(lmod) > 0.015) %>% select(height, cigs, chd)`

```
  height cigs chd
1     71   99  no
2     64   80  no
```

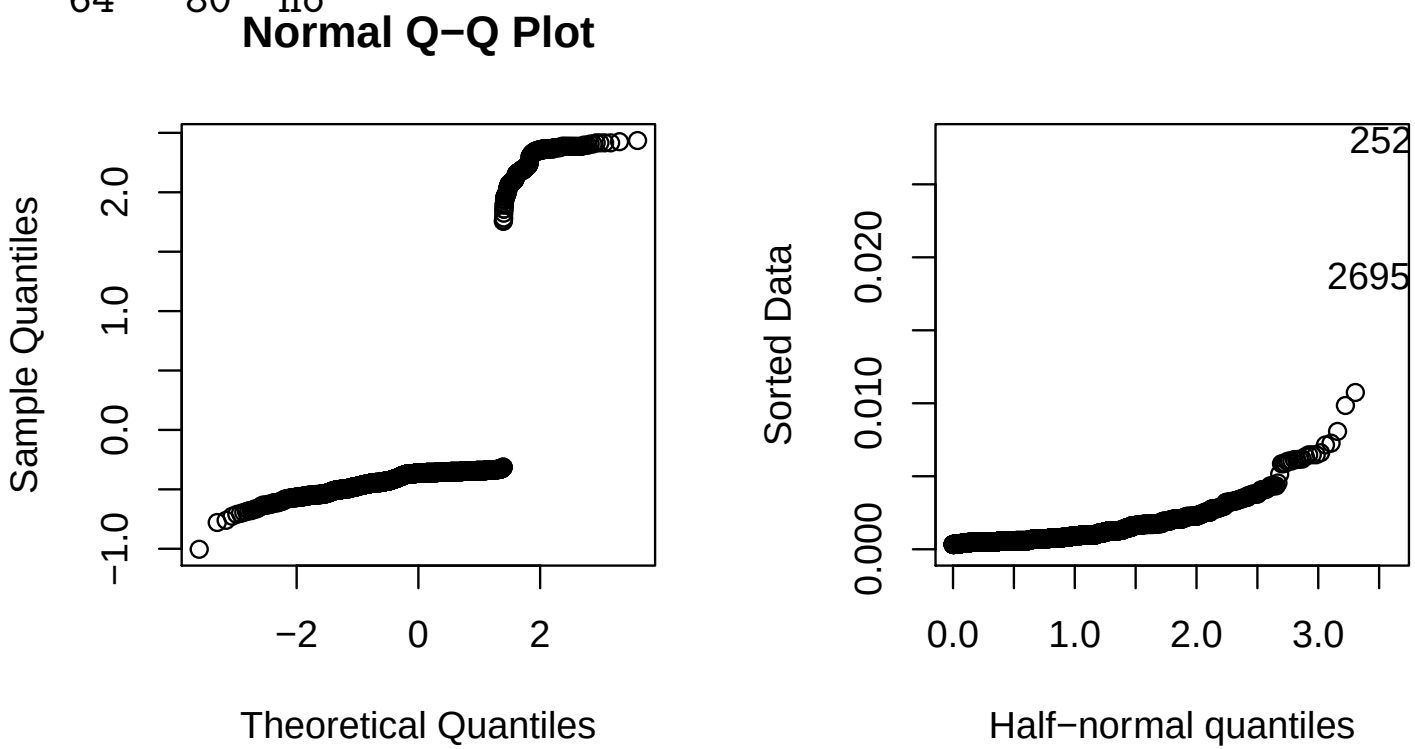

Figure 2.8: Left: QQ plot of the deviance residuals. Right:half-normal plot of the leverages.

§2.5 Model selection in logistic regression (1)

- Model selection or variable selection may be undertaken by either hypothesis testing or a numerical information criterion.
- In the hypothesis testing approach, **backward elimination** is a common procedure. The `drop1` function in `R` can be used to implement.
- Commonly used information criteria are AIC, BIC and their variants.
- A **subset selection** procedure may be used to implement these criteria. The `step` function in `R` is for this purpose.
- The subset selection procedure becomes computationally infeasible when the candidate model space is of exponential order. New methods such as LASSO and Gibbs-BIC etc., have been developed.

§2.5 Model selection in logistic regression (2)

The **backward elimination** method proceeds sequentially:

1. Start with the full model including all the available predictors. The full model may include derived predictors formed from transformations or interactions between two or more predictors.
2. Compare this model with all the models consisting of one less predictor. Compute the p-value corresponding to each dropped predictor.
3. Eliminate the term with largest p-value that is greater than some preset critical value, say 0.05. Return to the previous step. If no such term meets this criterion, stop and use the current model.

This is an inferior procedure for variable selection, because the hypothesis testing error involved in the procedure cannot be controlled. This procedure is often used for explaining the effect of some predictors on the response.

§2.5 Model selection in logistic regression (3)

- The Akaike information criterion (AIC) for a model with the maximum likelihood L and number of parameters q is defined by
  $\text{AIC} = -2\log L + 2q \iff \text{AIC} = \text{deviance} + 2q$, (use $q \log n$ not $2q$ if BIC)
- The model having the smallest AIC value among all candidate models is selected. Although unsatisfactory, `step` function is still in use.

```
> wcgs$bmi <- with(wcgs, 703*wcgs$weight/(wcgs$height^2))   #dim(wcgs)=c(3154,15)
> wcgsm <- na.omit(wcgs) #3140x15 dataframe; sum(is.na(wcgs)); 12 chol & 2 arcus NAs
> lmod <- glm(chd ~ age + height + weight +bmi + sdp + dbp + chol + dibep + cigs
            +arcus, family=binomial, wcgsm)
> lmodr <- step(lmod, trace=0, k=2); sumary(lmodr)  #k=2 or log(n) for AIC or BIC
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-15.9575989	2.2860760	-6.9803	2.945e-12
age	0.0615904	0.0123968	4.9683	6.756e-07
height	0.0501608	0.0278236	1.8028	0.07142
bmi	0.0603846	0.0265986	2.2702	0.02319
sdp	0.0177284	0.0041547	4.2671	1.981e-05
chol	0.0107089	0.0015285	7.0062	2.450e-12
dibepB	0.6576159	0.1458984	4.5074	6.564e-06
cigs	0.0210406	0.0042625	4.9363	7.963e-07
arcuspresent	0.2109985	0.1437175	1.4681	0.14206

```
n = 3140 p = 9; Deviance = 1569.33 Null Deviance = 1769.17 (Difference = 199.846)
```

§2.5 Model selection in logistic regression (4)

- A different model is selected by repeatedly using `drop1`.

```
> drop1(lmod, test="Chi")

Single term deletions

Model:
chd ~ age + height + weight + bmi + sdp + dbp + chol + dibep +  cigs + arcus
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1569.2 1591.2
age     1   1593.8 1613.8 24.618 6.989e-07 ***
height  1   1569.5 1589.5  0.285  0.593689
weight  1   1569.3 1589.3  0.099  0.753181
bmi     1   1569.5 1589.5  0.258  0.611578
sdp     1   1577.0 1597.0  7.826  0.005151 **
dbp     1   1569.2 1589.2  0.011  0.916620   #largest p-value
chol    1   1620.0 1640.0 50.735 1.057e-12 ***
dibep   1   1590.5 1610.5 21.333 3.860e-06 ***
cigs    1   1592.2 1612.2 23.013 1.609e-06 ***
arcus   1   1571.3 1591.3  2.098  0.147446
```

§2.5 Model selection in logistic regression (5)

```
> drop1(glm(chd ~ age + height + weight + bmi + sdp + chol + dibep +
        cigs + arcus, family=binomial, wcgsm), test="Chi")

Single term deletions

Model:
chd ~ age + height + weight + bmi + sdp + chol + dibep + cigs +  arcus
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1569.2 1589.2
age     1   1593.8 1611.8 24.609 7.024e-07 ***
height  1   1569.5 1587.5  0.287    0.5921
weight  1   1569.3 1587.3  0.101    0.7511  #largest p-value
bmi     1   1569.5 1587.5  0.259    0.6111
sdp     1   1586.5 1604.5 17.254 3.271e-05 ***
chol    1   1620.0 1638.0 50.755 1.046e-12 ***
dibep   1   1590.5 1608.5 21.323 3.881e-06 ***
cigs    1   1592.7 1610.7 23.448 1.283e-06 ***
arcus   1   1571.3 1589.3  2.116    0.1458
```

§2.5 Model selection in logistic regression (6)

```
> drop1(glm(chd ~ age + height + bmi + sdp + chol + dibep +
      cigs + arcus, family=binomial, wcgsm), test="Chi")

Single term deletions

Model:
chd ~ age + height + bmi + sdp + chol + dibep + cigs + arcus
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1569.3 1587.3
age     1   1593.9 1609.9 24.598 7.062e-07 ***
height  1   1572.6 1588.6  3.277   0.07028 .
bmi     1   1574.4 1590.4  5.098   0.02396 *
sdp     1   1586.6 1602.6 17.285 3.217e-05 ***
chol    1   1620.0 1636.0 50.658 1.099e-12 ***
dibep   1   1590.7 1606.7 21.320 3.886e-06 ***
cigs    1   1592.8 1608.8 23.453 1.280e-06 ***
arcus   1   1571.5 1587.5  2.130   0.14441 #largest p-value
```

§2.5 Model selection in logistic regression (7)

```
> drop1(glm(chd ~ age + height + bmi + sdp + chol + dibep + cigs,
                  family=binomial, wcgsm), test="Chi")

Single term deletions

Model: chd ~ age + height + bmi + sdp + chol + dibep + cigs
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1571.5 1587.5
age     1   1600.2 1614.2 28.748 8.243e-08 ***
height  1   1575.0 1589.0  3.515   0.06080 .  #p-value > 0.05
bmi     1   1576.3 1590.3  4.860   0.02749 *
sdp     1   1588.2 1602.2 16.722 4.328e-05 ***
chol    1   1624.7 1638.7 53.207 3.002e-13 ***
dibep   1   1592.8 1606.8 21.388 3.752e-06 ***
cigs    1   1595.6 1609.6 24.129 9.009e-07 ***
```

§2.5 Model selection in logistic regression (8)

```
> drop1(glm(chd~age+bmi+sdp+chol+dibep+cigs,family=binomial, wcgsm), test="Chi")

Single term deletions
Model:  chd ~ age + bmi + sdp + chol + dibep + cigs
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1575.0 1589.0
age     1   1601.9 1613.9 26.946 2.093e-07 ***
bmi     1   1579.5 1591.5  4.483   0.03424 *
sdp     1   1592.2 1604.2 17.219 3.331e-05 ***
chol    1   1626.2 1638.2 51.268 8.056e-13 ***
dibep   1   1597.1 1609.1 22.094 2.596e-06 ***
cigs    1   1599.8 1611.8 24.839 6.233e-07 ***

> lmodBE <- glm(formula = chd ~ age + bmi + sdp + chol + dibep + cigs, family =
                binomial, data = wcgsm)  #model selected by Backward Elimination.
> lmodAIC <- glm(formula = chd ~ age + height + bmi + sdp + chol + dibep + cigs +
              arcus, family = binomial, data = wcgsm) #model selected by AIC.
> anova(lmodBE,lmodAIC, test="Chi")

Analysis of Deviance Table
Model 1: chd ~ age + bmi + sdp + chol + dibep + cigs
Model 2: chd ~ age + height + bmi + sdp + chol + dibep + cigs + arcus
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3133     1575.0
2      3131     1569.3  2   5.6458  0.05943 .  #Two models not significantly differ
```

§2.6 Goodness of fit (1)

- As seen before, deviance cannot be used as a measure of fit for binary response GLM.
- *Hosmer-Lemeshow statistic*, constructed based on the binned residuals, can be used as a goodness of fit measure.
- *Receiver operating characteristic* (ROC) curve gives a graphic description of the binary response GLM's goodness of fit.

§2.6 Goodness of fit (2)

- Divide the observations up into J bins based on the linear predictor $\hat{\eta}_i$ values. Let the sum of responses in the j-th bin be y_j and the mean predicted probability $\bar{\hat{p}}_j = m_i^{-1} \sum_{i \in \text{bin } j} \hat{p}_i$, with m_j observations within the j-th bin.
- The **Hosmer-Lemeshow statistic** is defined as

$$X^2_{HL} = \sum_{j=1}^{J} \frac{(y_j - m_j \bar{\hat{p}}_j)^2}{m_j \bar{\hat{p}}_j (1 - \bar{\hat{p}}_j)}.$$

- This statistic has an approximate χ^2 distribution with df. $J - 1$.
- We have some freedom to decide on the binning. We need sufficient observations per bin to ensure the accuracy of the χ^2 approximation yet not so few bins that the fit can hardly be tested.

§2.6 Goodness of fit (3)

- A plot of the binned predicted probabilities obtained from the model `lmodAIC` vs. the observed proportions of `chd` is given by the code:

```
> wcgsm <- na.omit(wcgs)
> wcgsm <- mutate(wcgsm, predprob=predict(lmodAIC,type="response"),
                  linpred=predict(lmodAIC))
> gdf <- group_by(wcgsm, ntile(linpred,100))
> hldf <- summarise(gdf, y=sum(y), ppred=mean(predprob), count=n())
> hldf <- mutate(hldf, se.fit=sqrt(ppred*(1-ppred)/count))
> ggplot(hldf,aes(x=ppred,y=y/count,ymin=y/count-2*se.fit,ymax=y/count+2*se.fit))+
    geom_point()+geom_linerange(color=grey(0.75))+geom_abline(intercept=0,slope=1)
    +xlab("Predicted Probability")+ylab("Observed Proportion")
```

- The Hosmer-Lemeshow statistic for `lmodAIC` is given in the following:

```
> hlstat <- with(hldf, sum( (y-count*ppred)^2/(count*ppred*(1-ppred))))
> c(hlstat, nrow(hldf))
[1]  94.28042 100.00000
> 1-pchisq(94.28042, 100-1)
[1] 0.6153527
```

Since the p-value is large, we detect no lack of fit in `lmodAIC`.

§2.6 Goodness of fit (4)

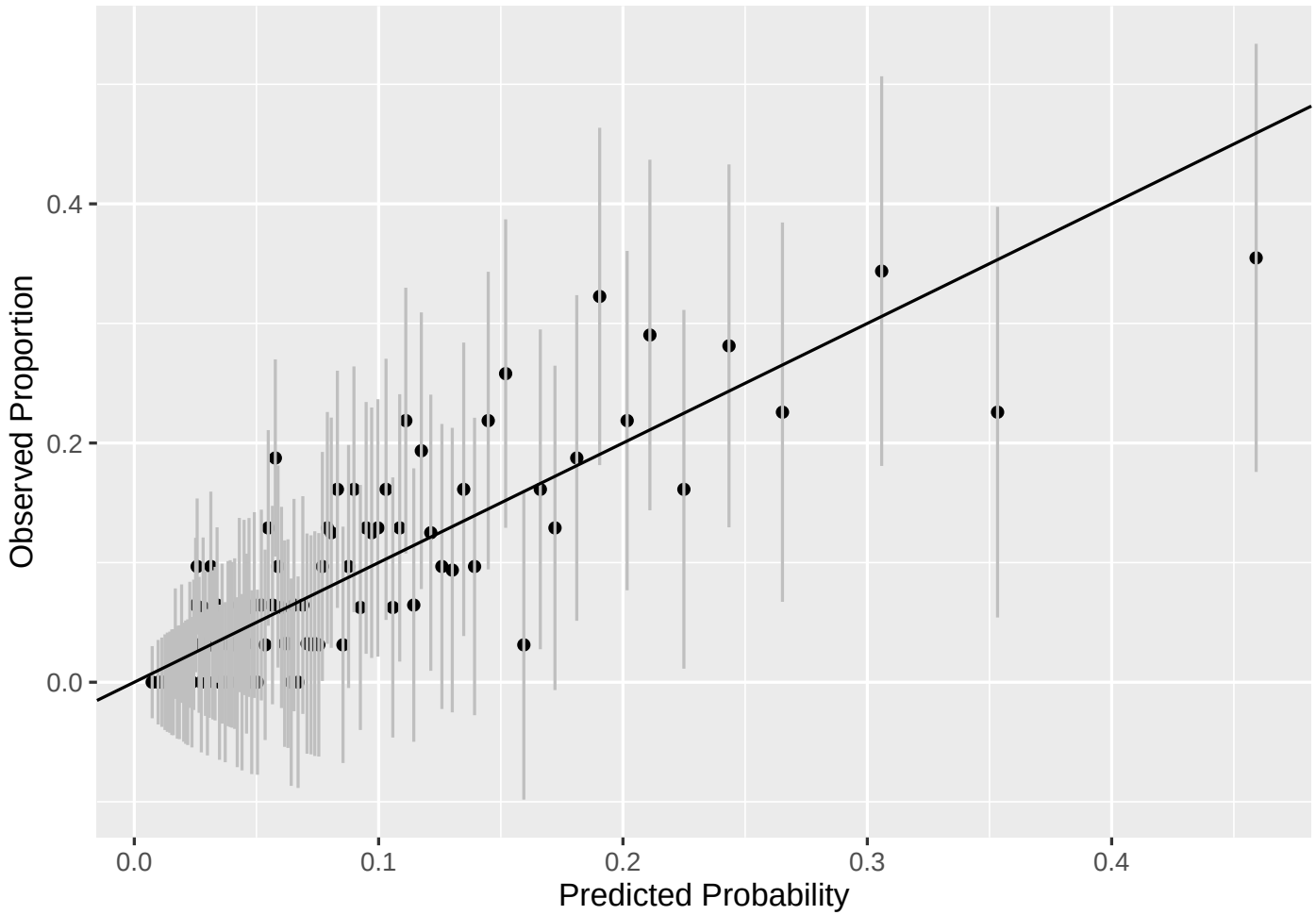

Figure 2.9: Binned predicted probabilities and observed proportions for the heart disease model.

§2.6 Goodness of fit: ROC curve (1)

- A logistic regression model gives the fitted/predicted probability $\hat{p}_i$ automatically. To further obtain a fitted/predicted value of Y_i we need to specify a *threshold value* p^* so that Y_i can be fitted/predicted according to the rule $\hat{Y}_i = I(\hat{p}_i \geq p^*)$.
- The model `lmodAIC` with $p^* = 0.5$ gives the following classification **confusion matrix**

```
> wcgsm <- mutate(wcgsm, predout=ifelse(predprob < 0.5, "no", "yes"))
> xtabs( ~ chd + predout, wcgsm)
```

```
     predout
chd     no  yes
  no  2882    3
  yes  253    2
```

- The model has a small **mis-classification rate** $(253 + 3)/(2882 + 3 + 253 + 2) = 0.0815$.
- The model has a high **specificity rate** $2882/(2882 + 3) = 0.999$, but a very small (poor) **sensitivity rate** $2/(253 + 2) = 0.00784$.

§2.6 Goodness of fit: ROC curve (2)

- When the threshold value p^* increases, the specificity will rise but the sensitivity will fall.

```
thresh <- seq(0.01,0.5,0.01)
Sensitivity <- numeric(length(thresh))
Specificity <- numeric(length(thresh))
for(j in seq(along=thresh)){
pp <- ifelse(wcgsm$predprob < thresh[j],"no","yes")
xx <- xtabs( ~ chd + pp, wcgsm)
Specificity[j] <- xx[1,1]/(xx[1,1]+xx[1,2])
Sensitivity[j] <- xx[2,2]/(xx[2,1]+xx[2,2])
}

matplot(thresh,cbind(Sensitivity,Specificity),type="l",xlab="Threshold",
                    ylab="Proportion",lty=1:2)
```

- The plot of *sensitivity* vs. *1 – specificity* (also called *false positive rate*) is named the **ROC curve**.

```
> plot(1-Specificity,Sensitivity,type="l");  abline(0,1,lty=2)
> AUCv <- numeric(length(thresh)-1)
> for(i in 1:(length(thresh)-1)){ #Trapezoid area =0.5*(Left+Right)*Width
    AUCv[i]<-0.5*sum(Sensitivity[i:(i+1)])*(Specificity[i+1]-Specificity[i])}
> AUC <- sum(AUCv)  # Area under the curve (AUC)=0.7365725
```

§2.6 Goodness of fit: ROC curve (3)

- A model is better when its ROC curve stretches more towards top-left corner. Area under the curve (AUC)=0.737 here.

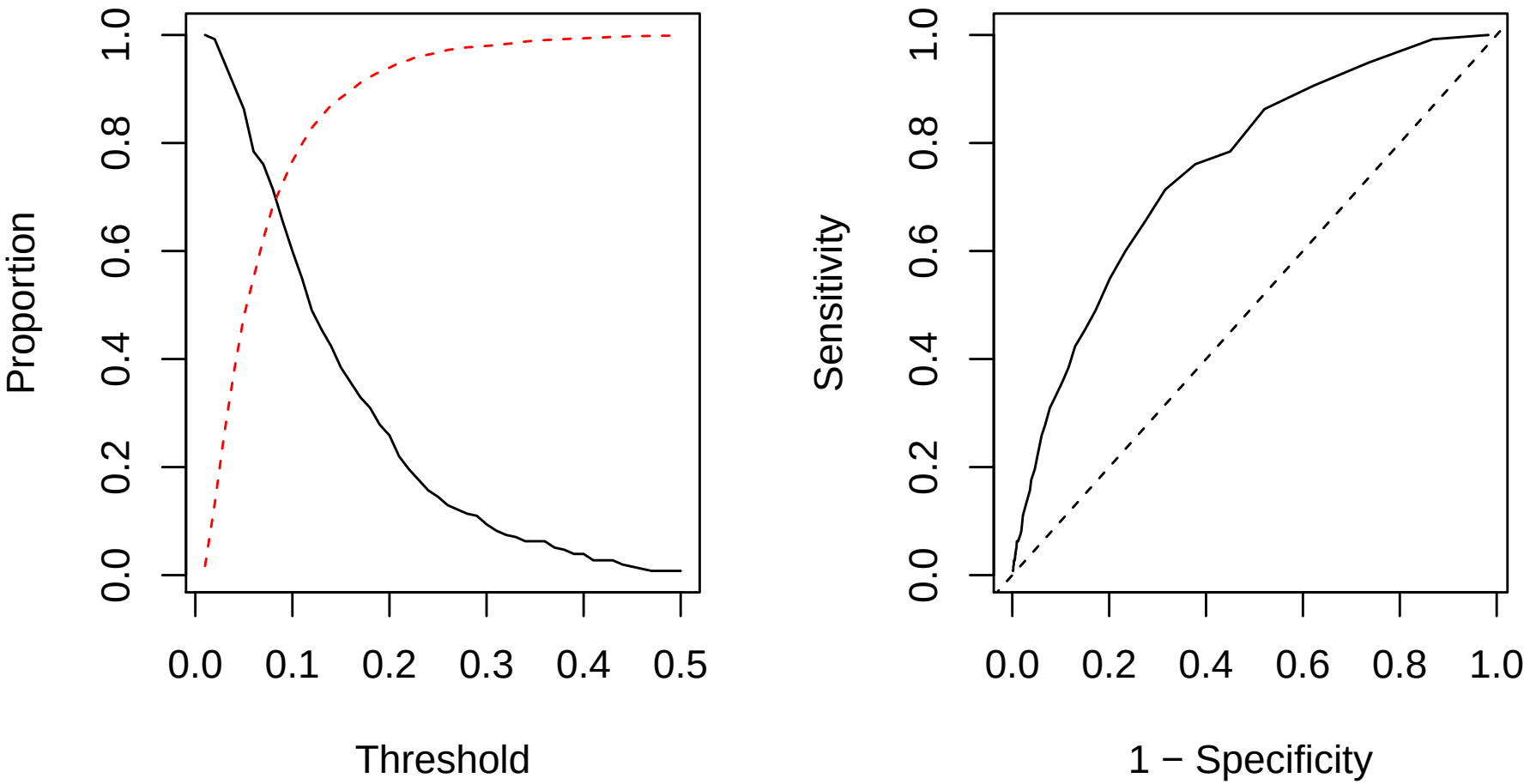

Figure 2.10: Sensitivity and specificity for the heart disease model plotted as a function of the probability threshold (left plot) and as the ROC curve (right plot).