STAT 426

Logistic Regression

Spring 2018

Suppose we observe 0/1 (Bernoulli) response Y and quantitative explanatory variable X.

Let

$$\pi(x) \;=\; \mathrm{P}(Y = 1 \mid X = x) \quad \in (0, 1) \text{ for all } x$$

The (simple) logistic regression of Y on X uses

$$\pi(x) \;=\; \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

i.e.

$$\mathrm{logit}\big(\pi(x)\big) \;=\; \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \;=\; \alpha + \beta x$$

Interpretation

β determines nature of X-Y relationship:

- $\beta > 0$: increasing X increases prob. $Y = 1$
- $\beta = 0$: no relationship
- $\beta < 0$: increasing X decreases prob. $Y = 1$

Can show

$$\frac{d}{dx}\pi(x) \quad = \quad \beta\,\pi(x)\big(1 - \pi(x)\big)$$

(For what $\pi(x)$ is this steepest?)

Note:

$$\text{odds of } Y = 1 \quad = \quad \frac{\pi(x)}{1 - \pi(x)} \quad = \quad e^{\alpha + \beta x}$$

Odds ratio for $Y = 1$ at $x + 1$ versus at x:

$$\frac{e^{\alpha + \beta(x+1)}}{e^{\alpha + \beta x}} \quad = \quad e^{\beta}$$

Note: Doesn't depend on x.

So β is the log-odds ratio for the effect of increasing x by one unit (which depends on the units of x).

Eg: $\alpha = 1$, $\beta = 2$

The **median effective level** is the x value at which $\pi(x) = 1/2$, which is generally

$$-\alpha/\beta \qquad \text{if } \beta \neq 0$$

Other summaries:

- change in $\pi(x)$ over observed range of x
- change in $\pi(x)$ between observed upper and lower quartiles of x
- x distance between $\pi(x) = 1/4$ and $\pi(x) = 3/4$

Note: The first two do not depend on the units of x.

Interpreting α can be more difficult — it's the log-odds when $x = 0$.

If a mean-centered version of X is used, α is the log-odds at the sample mean of the original X.

Remark: Logistic regression remains valid for retrospective (e.g. case-control) studies (Y fixed, X random).

The same β is estimated as in a prospective study (although α may be different). (Agresti, Sec. 5.1.4)

Inference

Assume independently-sampled data pairs

$$(x_i, y_i) \qquad i = 1, \ldots, N$$

Let $\hat{\alpha}$, $\hat{\beta}$ be the MLEs.

Test and form CIs using

- Wald
- Likelihood ratio
- Score ("Rao")

Eg: Wald z-statistic for

$$H_0 : \beta = 0 \qquad\qquad H_a : \beta \neq 0$$

is

$$\frac{\hat{\beta}}{SE(\hat{\beta})} \quad \underset{H_0}{\dot{\sim}} \quad N(0,1)$$

where the square of SE comes from the estimated asymptotic covariance matrix (estimated inverse information matrix).

Similarly, can form a Wald CI for β.

Odds ratio e^{β} for increasing x by one unit is estimated by

$$e^{\hat{\beta}}$$

and has CI

$$\left(e^{L}, e^{U}\right)$$

where (L, U) is a CI for β.

Estimated logistic curve:

$$\hat{\pi}(x) \;=\; \text{logit}^{-1}\big(\hat{\alpha} + \hat{\beta}x\big)$$

with estimated slope

$$\hat{\beta}\,\hat{\pi}(x)\big(1 - \hat{\pi}(x)\big)$$

So the estimated median effective level (x such that $\hat{\pi}(x) = 1/2$) is

$$-\hat{\alpha}/\hat{\beta} \qquad\qquad \text{if } \hat{\beta} \neq 0$$

The "fitted values" are

$$\hat{\pi}_i \quad = \quad \hat{\pi}(x_i) \qquad\qquad\qquad i = 1, \ldots, N$$

More generally, for possible X value x_0,

estimate $\pi(x_0) = \mathrm{P}(Y = 1 \mid X = x_0)$ with $\hat{\pi}(x_0)$.

Note: $SE\big(\operatorname{logit}\hat{\pi}(x_0)\big)$ is the square root of

$$\begin{aligned}\widehat{\operatorname{var}}\big(\operatorname{logit}\hat{\pi}(x_0)\big) &= \widehat{\operatorname{var}}(\hat{\alpha}+\hat{\beta}x_0) \\ &= \widehat{\operatorname{var}}(\hat{\alpha}) + x_0^2\,\widehat{\operatorname{var}}(\hat{\beta}) + 2x_0\,\widehat{\operatorname{cov}}(\hat{\alpha},\hat{\beta})\end{aligned}$$

where the estimated variances and covariances are from the estimated asymptotic covariance matrix.

The Wald CI

$$\operatorname{logit}\hat{\pi}(x_0) \pm z_{\alpha/2}\,SE\big(\operatorname{logit}\hat{\pi}(x_0)\big)$$

can be transformed (inverse logit) to a Wald CI for $\pi(x_0)$.

R Example: Horseshoe Crabs — Satellite Presence

```
> horseshoe <- read.table("horseshoe.txt", header=TRUE)

> head(horseshoe)
  color spine width satell weight y
1     3     3  28.3      8   3050 1
2     4     3  22.5      0   1550 0
3     2     1  26.0      9   2300 1
4     4     3  24.8      0   2100 0
5     4     3  26.0      4   2600 1
6     3     3  23.8      0   2100 0
```

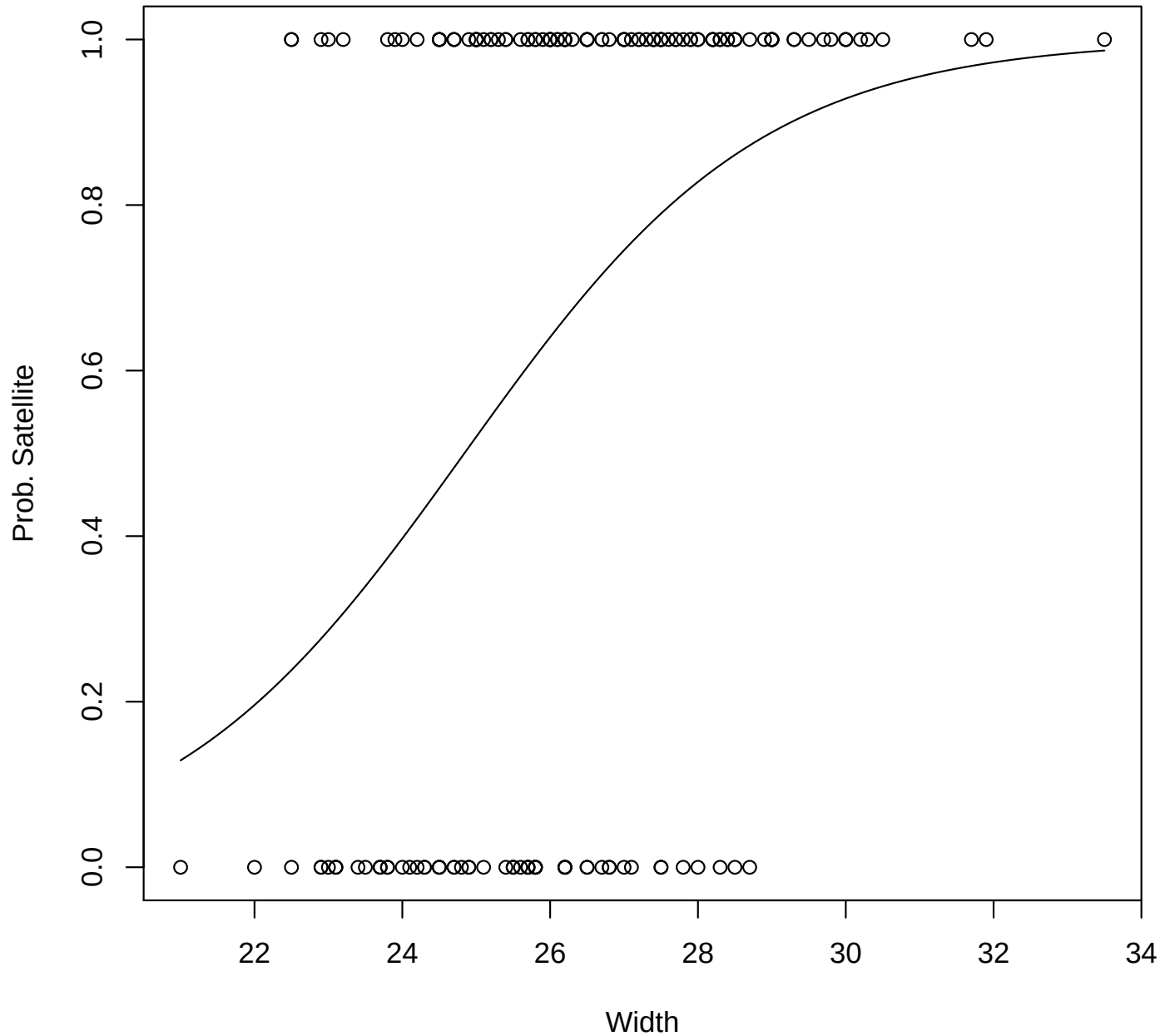
y indicates if there are any satellites (0=no, 1=yes)

```
> hsfit <- glm(y ~ width, family=binomial, data=horseshoe)

> summary(hsfit)

...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45

Number of Fisher Scoring iterations: 4
```

Plot observations and estimated simple logistic regression curve:

```
> plot(y ~ width, data=horseshoe, xlab="Width", ylab="Prob. Satellite")
```

```
> curve(predict(hsfit, data.frame(width=x), type="response"), add=TRUE)
```


Prob. Satellite
0.0
0.2
0.4
0.6
0.8
1.0
22
24
26
28
30
32
34
Width

Likelihood ratio test (for $\beta = 0$):

```
> drop1(hsfit, test="Chisq")
Single term deletions

Model:
y ~ width
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      194.45 198.45
width   1   225.76 227.76 31.306 2.204e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Profile likelihood CIs:

```
> confint(hsfit)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -17.8100090 -7.4572470
width         0.3083806  0.7090167

> exp(confint(hsfit)[2,])  # for odds ratio when increasing width 1cm
Waiting for profiling to be done...
   2.5 %   97.5 %
1.361219 2.031992
```

Estimation of median effective level (width with prob = 0.5):

```
> coef(hsfit)
(Intercept)       width
-12.3508177   0.4972306

> -coef(hsfit)[1]/coef(hsfit)[2]
(Intercept)
   24.83922
```

Estimated logit and probability at width $= 26.5$:

```
> predict(hsfit, data.frame(width=26.5))  # estimated logit
        1
0.8257928

> predict(hsfit, data.frame(width=26.5), type="response")  # estimated prob
        1
0.6954646
```

Wald 95% CIs for logit and probability at width = 26.5:

```
> predict(hsfit, data.frame(width=26.5), se.fit=TRUE)
$fit
        1
0.8257928

$se.fit
[1] 0.1886957

$residual.scale
[1] 1


> logit.Wald.CI <- 0.8257928 + c(-1,1) * 1.96 * 0.1886957

> logit.Wald.CI  # for logit
[1] 0.4559492 1.1956364

> exp(logit.Wald.CI) / (1 + exp(logit.Wald.CI))  # for probability
[1] 0.6120528 0.7677476
```

First 10 fitted values (estimated probs. for first 10 obs):

```
> fitted(hsfit)[1:10]
        1         2         3         4         5         6         7         8
0.8482329 0.2380991 0.6404177 0.4951254 0.6404177 0.3736172 0.6954646 0.4827014
        9        10
0.3620557 0.5934595
```

Note: Wald inference, though common, is often inferior to likelihood or score — see Agresti, Sec. 5.2.6

Goodness of Fit Tests

Various strategies:

- Test an added higher-order term, e.g. test $\beta_2 = 0$ in the quadratic
$$\alpha + \beta_1 x + \beta_2 x^2$$
- If there are **replicates** (repeated x values), use **grouped** (binomial) data to test using the deviance
$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$$
- Otherwise, consider Hosmer-Lemeshow Test (Sec. 5.2.5)

Generalization: Binomial (Grouped) Data

$$Y_i \sim \text{indep. binomial}\big(n_i, \pi(x_i)\big)$$

Data must have both Y_i and n_i (or $n_i - Y_i$).

Ungrouped (binary Y) data with replicates can be converted to grouped data:

Sum Y over each replicate group,

and let n_i be the size of replicate group i.

Eg: Simple Logistic Regression

Ungrouped (binary)

Y	X
1	1
0	1
1	1
0	2
0	3
1	3

$N = 6$

$\hat{\alpha} \approx 0.7690$

$\hat{\beta} \approx -0.4203$

Deviance ≈ 8.109

Grouped (binomial)

Y	n	X
2	3	1
0	1	2
1	2	3

$N = 3$

$\hat{\alpha} \approx 0.7690$

$\hat{\beta} \approx -0.4203$

Deviance ≈ 1.518

Remarks on grouping:

- MLEs and likelihood-based inference remain the same, since grouping doesn't change the kernel of the likelihood.
- The deviance changes, since the saturated model for the grouped data is different.

  But deviance-based comparisons between nested sub-models do not change, since the saturated model log-likelihood cancels out.
- A deviance-based goodness of fit test may be valid for the grouped data, provided the values of $E(Y_i)$ and $n_i - E(Y_i)$ are not too small.
- "Fitted values" from R are still the probabilities $\hat{\pi}(x_i)$ (rather than estimated means $n_i\hat{\pi}(x_i)$).

Categorical Predictors

Consider categorical X with I categories.

After grouping, we obtain an $I \times 2$ table:

X category	success	failure
1	Y_1	$n_1 - Y_1$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
I	Y_I	$n_I - Y_I$

Then

$$Y_i \sim \text{indep. binomial}(n_i, \pi_i) \qquad i = 1, \ldots, I$$

Code X using $I - 1$ indicator variables:

$$\tilde{X}_i = \begin{cases} 1 & \text{if row } i \\ 0 & \text{otherwise} \end{cases} \qquad i = 2, \ldots, I$$

Note: No indicator variable for row 1 (to avoid redundancy).

Differs from coding used in Agresti (which omits the *last* level, not the first), but consistent with behavior of R.

For logistic regression,

$$\begin{aligned} \text{logit}(\pi_i) &= \alpha + \beta_2\tilde{x}_2 + \cdots + \beta_I\tilde{x}_I \\ &= \alpha + \beta_i \end{aligned}$$

if we define $\beta_1 \equiv 0$. Then

$$\alpha = \text{logit}(\pi_1)$$

$$\begin{aligned} \beta_i &= \text{logit}(\pi_i) - \text{logit}(\pi_1) = \ln\left(\frac{\pi_i/(1-\pi_i)}{\pi_1/(1-\pi_1)}\right) \\ &= \text{log-odds ratio from category } i \text{ to category 1} \end{aligned}$$

Similarly,

$$\beta_i - \beta_{i'} = \text{log-odds ratio from category } i \text{ to category } i'$$

MLEs of the π_is are

$$\hat{\pi}_i \;=\; p_i \;=\; y_i/n_i \qquad\qquad \text{(why?)}$$

and MLEs of α and β_i are obtained from the **empirical** (or **sample**) **logits**

$$\text{logit}(\hat{\pi}_i)$$

Letting $\beta_1 \equiv 0 \equiv \hat{\beta}_1$, we find

$$\hat{\alpha} \;=\; \text{logit}(\hat{\pi}_1) \qquad\qquad \hat{\beta}_i \;=\; \text{logit}(\hat{\pi}_i) \;-\; \hat{\alpha}$$

For $i < i'$,

$$\hat{\beta}_i \;-\; \hat{\beta}_{i'}$$

is the empirical log-odds ratio for the sub-table with only rows i and i'.

Note:

- If $y_i = 0$ or $y_i = n_i$, then $\text{logit}(\hat{\pi}_i)$ does not exist (why?), and thus neither does the MLE for the logistic regression.

- This model is saturated, so a deviance-based goodness of fit test is not available.

Testing for any X effect is testing

$$H_0 : \beta_2 = \cdots = \beta_I = 0 \quad (\equiv \beta_1)$$

$$H_a : \text{not all } \beta_i\text{s equal (including } \beta_1)$$

A LRT based on

$$M_0 : \ \text{logit}(\pi_i) \ = \ \alpha \qquad M_1 : \ \text{logit}(\pi_i) \ = \ \alpha + \beta_i$$

uses $G^2(M_0 \mid M_1)$, which equals the G^2 statistic for testing independence/homogeneity in the $I \times 2$ table.

Alternative: The Pearson X^2 statistic.

Remark: When X is ordinal, can sometimes improve power by assigning numerical scores x_i^* to the X levels and then using the linear logit model

$$\text{logit}(\pi_i) \quad = \quad \alpha + \beta x_i^*$$

R Example: Alcohol & Infant Malformation

```
> malform <- data.frame(Present=c(48, 38, 5, 1, 1),
+                       Absent=c(17066, 14464, 788, 126, 37),
+                       Drinks=factor(c("0", "<1", "1-2", "3-5", ">=6")))

> malform
  Present Absent Drinks
1      48  17066      0
2      38  14464     <1
3       5    788    1-2
4       1    126    3-5
5       1     37    >=6
```

```
> mffit <- glm(cbind(Present,Absent) ~ Drinks, family=binomial, data=malform)

> summary(mffit)

...

Deviance Residuals:
[1]  0  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.87364    0.14454 -40.637   <2e-16 ***
Drinks<1    -0.06819    0.21743  -0.314   0.7538
Drinks1-2    0.81358    0.47134   1.726   0.0843 .
Drinks3-5    1.03736    1.01431   1.023   0.3064
Drinks>=6    2.26272    1.02368   2.210   0.0271 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  6.2020e+00  on 4  degrees of freedom
Residual deviance: -1.7897e-13  on 0  degrees of freedom
AIC: 28.627

Number of Fisher Scoring iterations: 4
```

```
> model.matrix(mffit)  # X matrix
  (Intercept) Drinks<1 Drinks1-2 Drinks3-5 Drinks>=6
1           1        0         0         0         0
2           1        1         0         0         0
3           1        0         1         0         0
4           1        0         0         1         0
5           1        0         0         0         1
attr(,"assign")
[1] 0 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$Drinks
[1] "contr.treatment"


> predict(mffit)  # empirical logits
        1         2         3         4         5
-5.873642 -5.941832 -5.060060 -4.836282 -3.610918

> fitted(mffit)  # MLEs of probabilities
          1           2           3           4           5
0.002804721 0.002620328 0.006305170 0.007874016 0.026315789
```

LRT for any relationship:

```
> drop1(mffit, test="Chisq")  # LRT (same as G^2 test)
Single term deletions

Model:
cbind(Present, Absent) ~ Drinks
       Df Deviance    AIC   LRT Pr(>Chi)
<none>       0.000 28.627
Drinks  4    6.202 26.829 6.202   0.1846
```

(Test is perhaps questionable because of small expected counts in the contingency table.)

Pearson X^2 test — two different ways:

```
> drop1(mffit, test="Rao")  # Pearson test
Single term deletions

Model:
cbind(Present, Absent) ~ Drinks
       Df Deviance    AIC Rao score Pr(>Chi)
<none>       0.000 28.627
Drinks  4    6.202 26.829    12.082  0.01675 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> chisq.test(cbind(malform$Present, malform$Absent), correct=FALSE)

Pearson's Chi-squared test

data:  cbind(malform$Present, malform$Absent)
X-squared = 12.082, df = 4, p-value = 0.01675

Warning message:
In chisq.test(cbind(malform$Present, malform$Absent), correct = FALSE) :
  Chi-squared approximation may be incorrect
```

Now use scores for Drink as an ordinal variable:

```
> drink.score <- c(0, 0.5, 1.5, 4, 7)[as.integer(malform$Drinks)]

> mffit2 <- glm(cbind(Present,Absent) ~ drink.score, family=binomial,
+               data=malform)

> summary(mffit2)

...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
drink.score   0.3166     0.1254   2.523   0.0116 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
AIC: 24.576

Number of Fisher Scoring iterations: 4
```

LRT using scores:

```
> drop1(mffit2, test="Chisq")  # LRT
Single term deletions

Model:
cbind(Present, Absent) ~ drink.score
            Df Deviance    AIC    LRT Pr(>Chi)
<none>           1.9487 24.576
drink.score  1   6.2020 26.829 4.2533  0.03917 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Check goodness of fit (of score model):

```
> 1 - pchisq(deviance(mffit2), df.residual(mffit2))
[1] 0.5831178
```

Agresti, Sec. 5.3 (later subsections), gives arguments for using scores when dealing with ordinal predictors.

Multiple Logistic Regression

Allow p explanatory variables:

$$\text{logit}\big(\pi(\boldsymbol{x})\big) \;=\; \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

Increasing just x_j by 1 increases the logit by β_j, so

$$e^{\beta_j} \;=\; \text{odds ratio at } x_j + 1 \text{ relative to } x_j$$

(all other variables held fixed)

Application: $2 \times 2 \times 2$ Contingency Tables

$Z = 1$:

	Y	$n - Y$
$X = 1$		
$X = 0$		

$Z = 0$:

	Y	$n - Y$
$X = 1$		
$X = 0$		

Recall: Z could be a stratification variable, for which we want to adjust our analysis.

Note: X and Z are indicator variables.

We consider three nested models, in order from largest to smallest ...

- The saturated model:

$$\text{logit}\big(\pi(x,z)\big) \;=\; \alpha \,+\, \beta_1 x \,+\, \beta_2 z \,+\, \beta_3\, xz$$

  Note: Uses 4 parameters to model the 4 possible success probabilities.

- The additive (main-effects) model:

$$\text{logit}\big(\pi(x,z)\big) \;=\; \alpha \,+\, \beta_1 x \,+\, \beta_2 z$$

Note:

log-odds ratio b/w X and Y

$$\begin{aligned} &= \text{logit}\big(\pi(1,z)\big) \,-\, \text{logit}\big(\pi(0,z)\big) \\ &= \alpha \,+\, \beta_1 \,+\, \beta_2 z \,-\, (\alpha \,+\, \beta_2 z) \\ &= \beta_1 \end{aligned}$$

does not depend on Z.

Thus, there is homogeneous XY association:

$$\theta_{XY(0)} \;=\; \theta_{XY(1)} \;=\; e^{\beta_1}$$

- The Z-only model:

$$\text{logit}\big(\pi(x,z)\big) \;=\; \alpha \,+\, \beta_2 z \qquad\qquad (\text{so } \beta_1 = 0)$$

corresponds to

conditional independence between X and Y given Z.

(Why?)

- The Z-only model:

$$\text{logit}\big(\pi(x,z)\big) \;=\; \alpha + \beta_2 z \qquad\qquad (\text{so } \beta_1 = 0)$$

corresponds to

conditional independence between X and Y given Z.

(Why?)

Test for homogeneous association by comparing the additive and saturated models.

Test for conditional independence by comparing the Z-only and additive (or saturated) models.

R Example: Death Penalty & Race

Recall the Florida death penalty data:

```
> deathpenalty <- read.table("deathpenalty.txt")

> deathpenalty
  DeathPenalty Defendant Victim Freq
1          Yes     White  White   53
2           No     White  White  414
3          Yes     Black  White   11
4           No     Black  White   37
5          Yes     White  Black    0
6           No     White  Black   16
7          Yes     Black  Black    4
8           No     Black  Black  139
```

We need separate columns for the `Yes` and `No` values of `DeathPenalty`.

Here is one possible approach:

```
> dp <- reshape(deathpenalty, v.names="Freq", timevar="DeathPenalty",
+               idvar=c("Defendant","Victim"), direction="wide")

> dp
  Defendant Victim Freq.Yes Freq.No
1     White  White       53     414
3     Black  White       11      37
5     White  Black        0      16
7     Black  Black        4     139
```

Fit the saturated model:

```
> sat.mod <- glm(cbind(Freq.Yes,Freq.No) ~ Defendant*Victim, family=binomial,
+                data=dp)

> summary(sat.mod)

...

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.5482     0.5071  -6.996 2.63e-12 ***
DefendantWhite             -21.9957 53403.2302   0.000 0.999671
VictimWhite                  2.3352     0.6125   3.813 0.000137 ***
DefendantWhite:VictimWhite  21.1531 53403.2302   0.000 0.999684
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.2266e+01  on 3  degrees of freedom
Residual deviance: 2.5801e-10  on 0  degrees of freedom
AIC: 20.92

Number of Fisher Scoring iterations: 22
```

Note: The MLE doesn't actually exist for this logistic regression (why?).

That explains the unusually large standard errors and P-values.

It also explains why so many iterations were needed.

Fit the additive model:

```
> add.mod <- glm(cbind(Freq.Yes,Freq.No) ~ Defendant + Victim, family=binomial,
+                data=dp)

> summary(add.mod)

...

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.5961     0.5069  -7.094 1.30e-12 ***
DefendantWhite  -0.8678     0.3671  -2.364   0.0181 *
VictimWhite      2.4044     0.6006   4.003 6.25e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.26591  on 3  degrees of freedom
Residual deviance:  0.37984  on 1  degrees of freedom
AIC: 19.3

Number of Fisher Scoring iterations: 4
```

Estimated odds ratio of receiving the death penalty for a white versus a black defendant (after controlling for victim race):

$$e^{-0.8678} \approx 0.42$$

Associated Wald 95% CI:

$$e^{-0.8678 \pm 1.96 \cdot 0.3671} \approx (0.20, 0.86)$$

But is the additive model adequate? ...

Test for homogeneous association (LRT):

```
> anova(add.mod, sat.mod, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Freq.Yes, Freq.No) ~ Defendant + Victim
Model 2: cbind(Freq.Yes, Freq.No) ~ Defendant * Victim
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1    0.37984
2         0    0.00000  1  0.37984   0.5377
```

(Or we could have just noticed the small `Residual deviance` from the `summary`.)

This is the same as the G^2 goodness of fit test. However, it may be questionable for this data because of some small expected cell counts.

Test the main effects (LRTs):

```
> drop1(add.mod, test="Chisq")
Single term deletions

Model:
cbind(Freq.Yes, Freq.No) ~ Defendant + Victim
          Df Deviance    AIC     LRT Pr(>Chi)
<none>         0.3798 19.300
Defendant  1   5.3940 22.314  5.0142  0.02514 *
Victim     1  20.7298 37.650 20.3499 6.45e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The test for (conditional) defendant race effect implicitly uses the additive model as the full model.

Alternatively, we can test for (conditional) defendant race effect with the saturated model as the full model (LRT):

```
> vic.mod <- glm(cbind(Freq.Yes,Freq.No) ~ Victim, family=binomial, data=dp)

> anova(vic.mod, sat.mod, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Freq.Yes, Freq.No) ~ Victim
Model 2: cbind(Freq.Yes, Freq.No) ~ Defendant * Victim
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2      5.394
2         0      0.000  2    5.394  0.06741 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

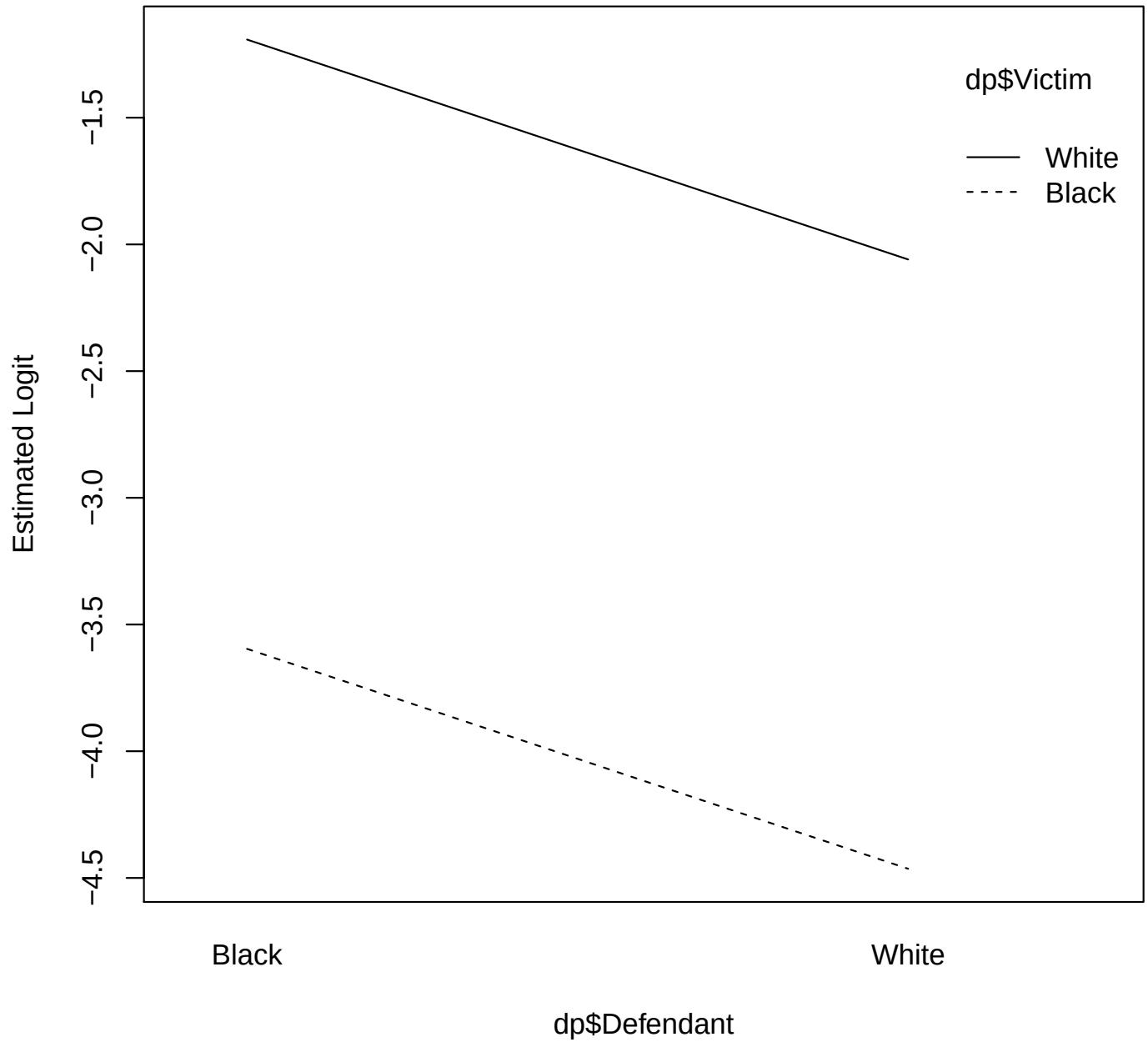
This is the same as the G^2 goodness of fit test. Again, it may be questionable for this data because of some small expected cell counts.

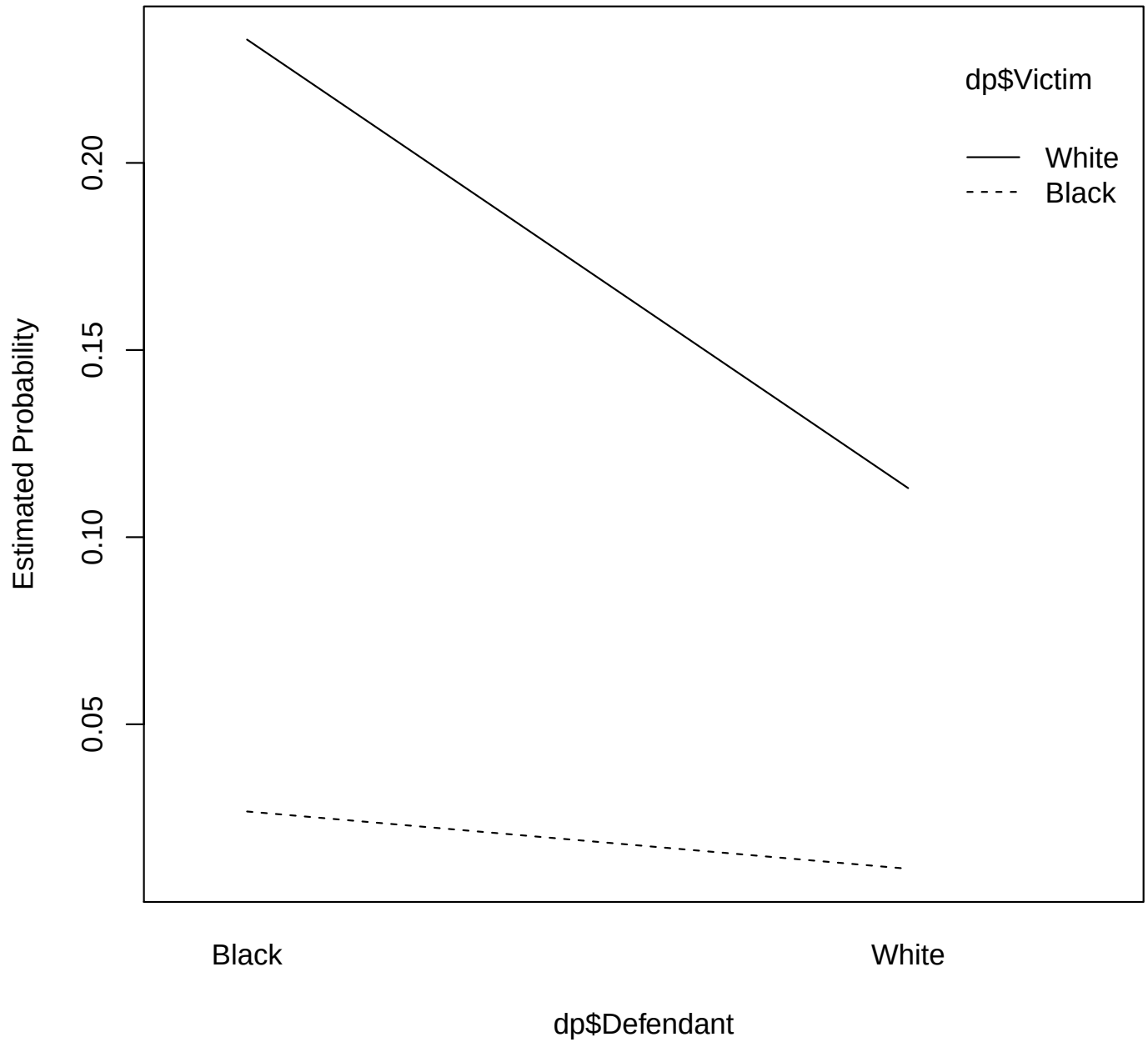
An "interaction plot" for the additive model is parallel on the logit scale:

```
interaction.plot(dp$Defendant, dp$Victim, predict(add.mod),
                 ylab="Estimated Logit")
```

... but not on the probability scale:

```
interaction.plot(dp$Defendant, dp$Victim, fitted(add.mod),
                 ylab="Estimated Probability")
```


dp$Victim
White
Black
Estimated Logit
−1.5
−2.0
−2.5
−3.0
−3.5
−4.0
−4.5
Black
White
dp$Defendant


dp$Victim
White
Black
Estimated Probability
0.05
0.10
0.15
0.20
Black
White
dp$Defendant

See Agresti, Sec. 5.4, for a different example: AIDS & AZT use.

Note: Can extend to situations where X has $I > 2$ categories and Z has $K > 2$ categories — just use more indicator variables.

Application: Categorical & Continuous Predictors

Use indicator variables for the categorical predictors, and linear terms for the continuous predictors.

Alternatively, consider using scores for an ordinal categorical variable.

R Example: Horseshoe Crab Satellites

```
> horseshoe <- read.table("horseshoe.txt", header=TRUE)

> head(horseshoe)
  color spine width satell weight y
1     3     3  28.3      8   3050 1
2     4     3  22.5      0   1550 0
3     2     1  26.0      9   2300 1
4     4     3  24.8      0   2100 0
5     4     3  26.0      4   2600 1
6     3     3  23.8      0   2100 0
```

Consider logistic regression of satellite presence on color and width, but now treat color as categorical (and not ordinal):

```
> hsfit <- glm(y ~ factor(color) + width, family=binomial, data=horseshoe)

> summary(hsfit)

...

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
factor(color)3   0.07242    0.73989   0.098    0.922
factor(color)4  -0.22380    0.77708  -0.288    0.773
factor(color)5  -1.32992    0.85252  -1.560    0.119
width            0.46796    0.10554   4.434 9.26e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46

Number of Fisher Scoring iterations: 4
```

Estimated odds ratio for a 1cm width increase, w/ color unchanged:

$$e^{0.46796} \approx 1.60$$

Estimated odds ratio for darkest (5) versus next darkest (4), w/ same width:

$$e^{-1.32992 - (-0.22380)} \approx 0.33$$

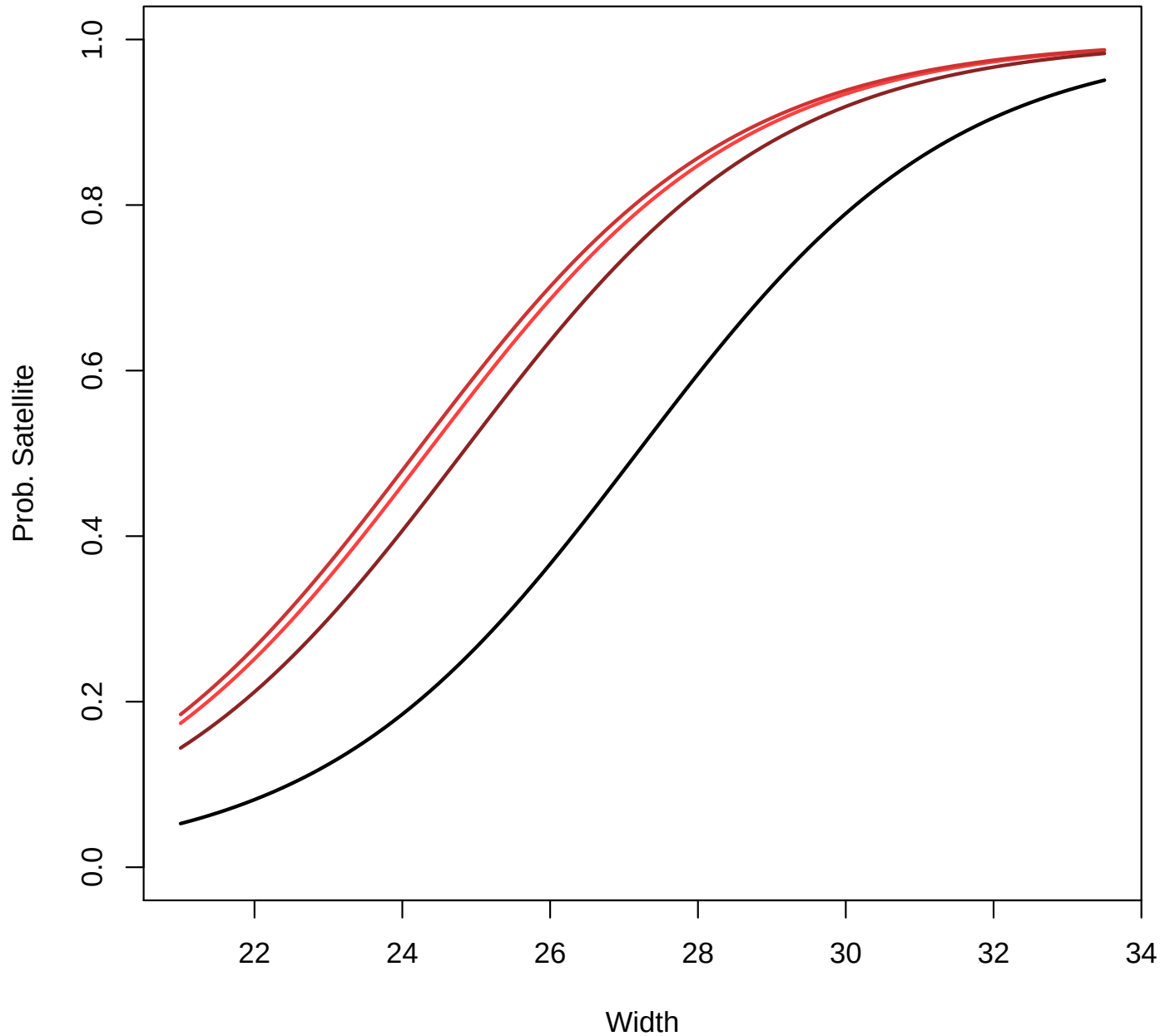
```
> drop1(hsfit, test="Chisq")
Single term deletions

Model:
y ~ factor(color) + width
              Df Deviance    AIC     LRT  Pr(>Chi)
<none>             187.46 197.46
factor(color)  3   194.45 198.45  6.9956   0.07204 .
width          1   212.06 220.06 24.6038 7.041e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

So color is not quite significant.

Plot estimated probability versus width, separately for each color:

```
> plot(y ~ width, data=horseshoe, xlab="Width", ylab="Prob. Satellite",
+      type="n")
> curve(predict(hsfit, data.frame(color=2,width=x), type="response"),
+       col="brown1", add=TRUE, lwd=2)
> curve(predict(hsfit, data.frame(color=3,width=x), type="response"),
+       col="brown3", add=TRUE, lwd=2)
> curve(predict(hsfit, data.frame(color=4,width=x), type="response"),
+       col="brown4", add=TRUE, lwd=2)
> curve(predict(hsfit, data.frame(color=5,width=x), type="response"),
+       col="black", add=TRUE, lwd=2)
```

Curves have same shape, but different shift ...


Prob. Satellite
0.0
0.2
0.4
0.6
0.8
1.0
22
24
26
28
30
32
34
Width

For comparison, try scores (2,3,4,5) for color:

```
> hsfit2 <- glm(y ~ color + width, family=binomial, data=horseshoe)

> summary(hsfit2)

...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5618     2.8828  -3.317  0.00091 ***
color        -0.5090     0.2237  -2.276  0.02286 *
width         0.4583     0.1040   4.406 1.05e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 189.12  on 170  degrees of freedom
AIC: 195.12

Number of Fisher Scoring iterations: 4
```

```
> drop1(hsfit2, test="Chisq")
Single term deletions

Model:
y ~ color + width
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      189.12 195.12
color   1   194.45 198.45  5.3315   0.02094 *
width   1   213.30 217.30 24.1767 8.789e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Color is now significant.

Matrix-Vector Formulation

Let

$$\boldsymbol{x}_i^T \;=\; i\text{th row of } \boldsymbol{X}$$

i.e. the explanatory variable values for observation i, and (typically) a 1 for the intercept.

Then

$$\text{logit}\big(\pi(\boldsymbol{x}_i)\big) \;=\; \eta_i \;=\; \boldsymbol{x}_i^T\boldsymbol{\beta} \qquad\qquad \boldsymbol{\eta} \;=\; \boldsymbol{X}\boldsymbol{\beta}$$

The log-likelihood (binomial response) is

$$L(\boldsymbol{\beta}) \;=\; \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} \;-\; \sum_i n_i \ln(1+e^{\eta_i})$$

The **score vector** is

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{\beta}) &= \nabla L(\boldsymbol{\beta}) &= \boldsymbol{X}^T\boldsymbol{y} - \sum_i n_i \, \boldsymbol{x}_i \, \pi(\boldsymbol{x}_i) \\ &&= \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\mu} \end{aligned}$$

where $\boldsymbol{\mu} = E(\boldsymbol{Y})$.

The **score vector** is

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{\beta}) &= \nabla L(\boldsymbol{\beta}) &= \boldsymbol{X}^T\boldsymbol{y} - \sum_i n_i \, \boldsymbol{x}_i \, \pi(\boldsymbol{x}_i) \\ &&= \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\mu} \end{aligned}$$

where $\boldsymbol{\mu} = E(\boldsymbol{Y})$.

This equals $\mathbf{0}$ at the MLE, giving the **likelihood equations**

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}$$

(Note: $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$ is orthogonal to the columns of $\boldsymbol{X}$, just like in linear regression.)

The second derivative matrix is

$$\begin{aligned} \nabla^2 L(\boldsymbol{\beta}) &= -\sum_i n_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \, \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big) \\ &= -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \end{aligned}$$

where

$$\boldsymbol{W} = \operatorname{diag}\Big(n_i \, \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big)\Big) = \operatorname{diag}\big(\operatorname{var}(Y_i)\big)$$

The second derivative matrix is

$$\begin{aligned} \nabla^2 L(\boldsymbol{\beta}) &= -\sum_i n_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \, \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big) \\ &= -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \end{aligned}$$

where

$$\boldsymbol{W} = \operatorname{diag}\Big(n_i \, \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big)\Big) = \operatorname{diag}\big(\operatorname{var}(Y_i)\big)$$

Since this is not random (doesn't depend on $\boldsymbol{Y}$), the information matrix is

$$\boldsymbol{\mathcal{J}} = -\nabla^2 L(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

and

$$\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$$

where $\hat{\boldsymbol{W}} = \operatorname{diag}\big(n_i \, \hat{\pi}(\boldsymbol{x}_i)\big(1 - \hat{\pi}(\boldsymbol{x}_i)\big)\big)$.

The standard error of

$$\operatorname{logit}\big(\hat{\pi}(\boldsymbol{x})\big) \;\; = \;\; \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$$

is then

$$\sqrt{\boldsymbol{x}^T\,\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}})\,\boldsymbol{x}}$$

and a Wald CI for $\operatorname{logit}\big(\pi(\boldsymbol{x})\big)$ is

$$\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \;\; \pm \;\; z_{\alpha/2}\sqrt{\boldsymbol{x}^T\,\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}})\,\boldsymbol{x}}$$

which can be transformed to a CI for $\pi(\boldsymbol{x})$.

The derivatives of $L(\boldsymbol{\beta})$ are also useful in the Newton-Raphson algorithm for finding $\hat{\boldsymbol{\beta}}$ — Agresti, Sec. 5.5.4.

Recall also the estimated hat matrix

$$\hat{\boldsymbol{H}}_{at} = \hat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}^{1/2}$$