

Logistic Regression Case Study

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 21, 2008

- We are studying the *runoff* data set.
- Data was collected over a four-year period from a Madison home.
- The outcome variable is an indicator if a rain storm produces runoff.
- There are multiple inputs.
- Graphical examinations show that the *total amount of precipitation* and various measures of *storm intensity* are good predictors.
- Storm duration and time since the previous storm are less predictive.
- We first study a model with storm total precipitation as a single input.

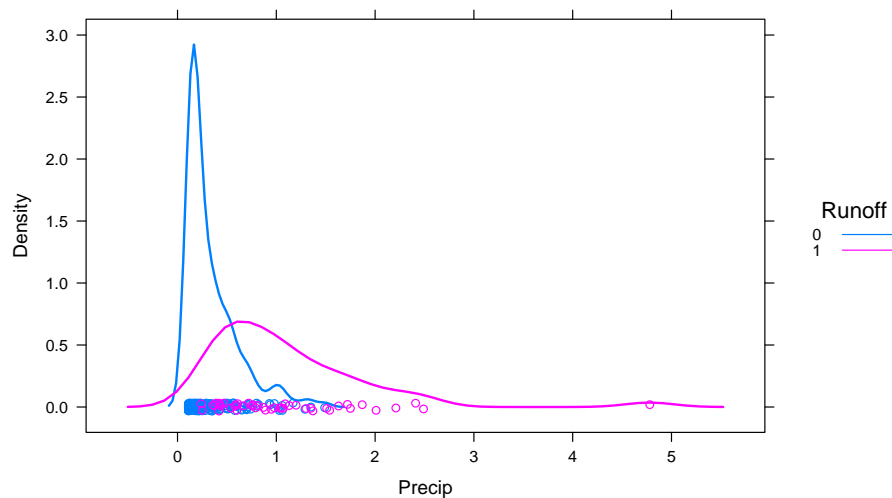
1 / 14

Runoff

2 / 14

Graph

Precipitation (inches)

Runoff
0 ———
1 ———

Model in R

```
> library(arm)
> fit1 = glm(RunoffEvent ~ Precip, data = runoff,
+   family = binomial)
> display(fit1)

glm(formula = RunoffEvent ~ Precip, family = binomial, data = runoff)
      coef.est coef.se
(Intercept) -3.64    0.42
Precip       3.81    0.58
---
n = 231, k = 2
residual deviance = 148.1, null deviance = 227.8 (difference = 79.7)
```

Runoff

3 / 14

Runoff

4 / 14

Fitted Model

- The variable Precip measures precipitation in inches
- The general logistic regression formula is

$$P\{y = 1 | X_i\} = \frac{1}{1 + \exp(-\eta)}$$

where $\eta = X_i \hat{\beta}$.

- The probability of runoff in this model is:

$$P\{\text{runoff}\} = \frac{1}{1 + \exp(-(-3.64 + 3.81(\text{Precip})))}$$

Interpretations

- The intercept is related to predictions when the predictor has value 0.
- The slope is an approximation of the change in inverse logistic probability per unit change in the predictor when the prediction is near the middle of the curve, which is hard to interpret directly.
- We can use the derivative

$$\frac{\hat{\beta}_2 \exp(\hat{\beta}_1 + \hat{\beta}_2 x)}{(1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x))^2}$$

- Because of the curve, we may need to consider smaller changes than one unit.
- When the precipitation is near one inch, $\exp(-3.64 + 3.81(1)) = 1.18$ and an increase of 0.1 inches increases the probability by about

$$0.1 \times \frac{3.81(1.18)}{(1 + 1.18)^2} = 0.09$$

Finding the 50/50 point

-

$$\begin{aligned} p &= \frac{1}{1 + \exp(-\eta)} \\ 1 + \exp(-\eta) &= \frac{1}{p} \\ \exp(-\eta) &= \frac{1}{p} - 1 = \frac{1-p}{p} \\ \eta &= \log\left(\frac{p}{1-p}\right) \end{aligned}$$

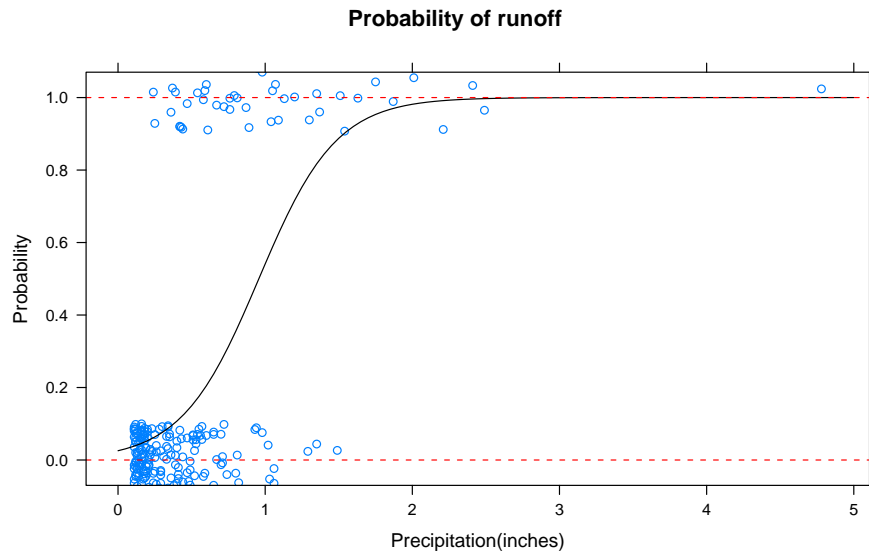
- With $p = 0.5$ and one predictor (plus an intercept),
 $\hat{\eta} = \hat{\beta}_1 + \hat{\beta}_2(\text{Precip}) = \log(1) = 0$

- so

$$\text{Precip} = -\frac{\hat{\beta}_1}{\hat{\beta}_2} = -\frac{-3.64}{3.81} = 0.96$$

Predictions

```
> invlogit = function(x) {
+   1/(1 + exp(-x))
+ }
> print(invlogit(predict(fit1, data.frame(Precip = c(1,
+   1.1))))))
      1      2
0.54 0.63
> values = c(0, 0.25, 0.5, 0.75, 1, 1.25, 1.5,
+   1.75, 2, 4)
> prob = round(invlogit(predict(fit1, data.frame(Precip = values))),
+   2)
> rbind(values, prob)
      1      2      3      4      5      6      7      8      9     10
values 0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00  4
prob   0.03 0.06 0.15 0.31 0.54 0.75 0.89 0.95 0.98  1
```



- Roughly speaking, a coefficient will be *statistically significant* if it is *at least two standard errors away from zero*.
- It rarely makes sense to test the intercept.
- Here, Precip is a useful predictor.

```
> display(fit1)

glm(formula = RunoffEvent ~ Precip, family = binomial, data = runoff)
      coef.est coef.se
(Intercept) -3.64   0.42
Precip       3.81   0.58
---
n = 231, k = 2
residual deviance = 148.1, null deviance = 227.8 (difference = 79.7)
```

Adding another predictor

```
> fit2 = glm(RunoffEvent ~ Precip + MaxIntensity10,
+ data = runoff, family = binomial)
> display(fit2)

glm(formula = RunoffEvent ~ Precip + MaxIntensity10, family = binomial,
data = runoff)
      coef.est coef.se
(Intercept)  -4.90   0.62
Precip       2.81   0.68
MaxIntensity10 1.84   0.38
---
n = 231, k = 3
residual deviance = 116.1, null deviance = 227.8 (difference = 111.7)
```

Including an interaction

```
> fit3 = glm(RunoffEvent ~ Precip * MaxIntensity10,
+ data = runoff, family = binomial)
> display(fit3)

glm(formula = RunoffEvent ~ Precip * MaxIntensity10, family = binomial,
data = runoff)
      coef.est coef.se
(Intercept)  -5.43   0.86
Precip       3.59   1.04
MaxIntensity10 2.42   0.69
Precip:MaxIntensity10 -0.84   0.77
---
n = 231, k = 4
residual deviance = 115.3, null deviance = 227.8 (difference = 112.5)
```

```
> anova(fit1, fit2, fit3)
```

Analysis of Deviance Table

Model 1: RunoffEvent ~ Precip

Model 2: RunoffEvent ~ Precip + MaxIntensity10

Model 3: RunoffEvent ~ Precip * MaxIntensity10

	Resid. Df	Resid. Dev	Df	Deviance
1	229	148.1		
2	228	116.1	1	32.0
3	227	115.3	1	0.8

