

# STAT 213

## Multiple Logistic Regression I

Colin Reimer Dawson

Oberlin College

April 13, 2018

# Outline

Multiple Predictors

Nested Model Tests

# Outline

Multiple Predictors

Nested Model Tests

# Logistic Regression With Multiple Predictors

We are combining logistic regression (Ch. 9) with multiple regression (Chs 3-4). Nothing really fundamentally new.

All of the “usual” options for predictors:

- Quantitative variables
- Powers of variables (e.g., second-order models)
- Other transformations of variables (e.g., log)
- Interactions of variables (modeling “change of slope”)
- Indicator variables for binary predictors
- Collections of  $M - 1$  indicators for categorical predictors w/  $M$  levels

# Two Equivalent Forms of (Multiple) Logistic Regression

Probability Form

$$\pi = \frac{e^{\beta_0 + \beta_1 X + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X + \dots + \beta_k X_k}}$$

Logit Form

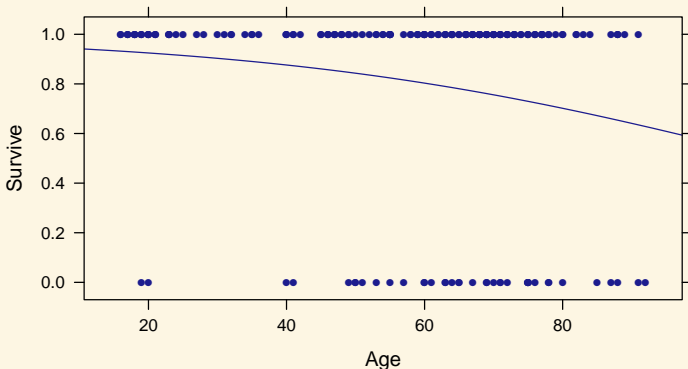
$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

# Example: Survival in ICU

- Response: Survive =  $\begin{cases} 0 & \text{Died} \\ 1 & \text{Lived} \end{cases}$
- Predictors:
  - Age
  - SysBP (Systolic Blood Pressure)
  - Pulse

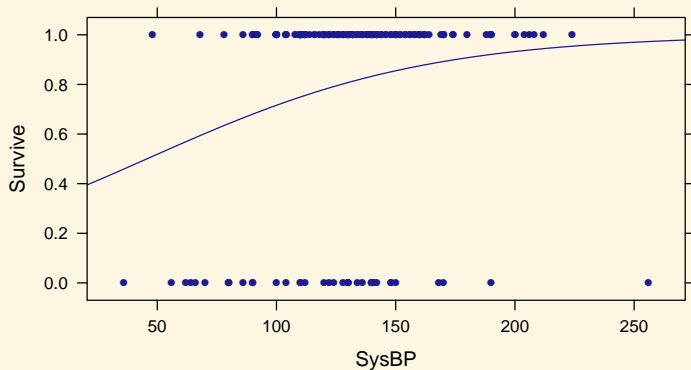
# Simple Logistic Models

```
library("Stat2Data"); data("ICU")  
m1 <- glm(Survive ~ Age, family = "binomial", data = ICU)  
plotModel(m1)
```



# Simple Logistic Models

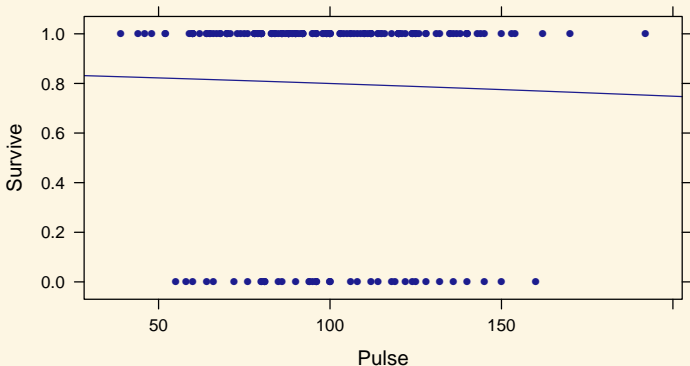
```
m2 <- glm(Survive ~ SysBP, family = "binomial", data = ICU)
plotModel(m2)
```





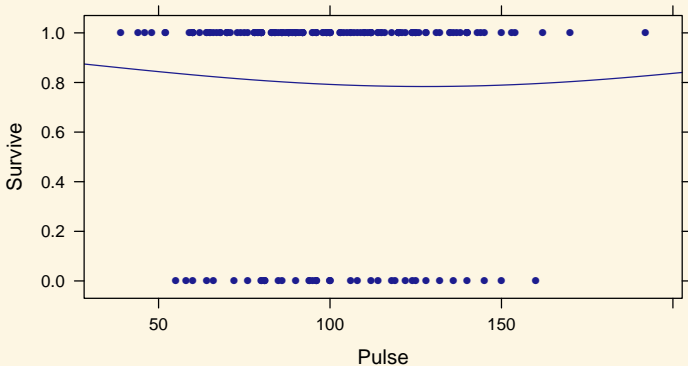
# Simple Logistic Models

```
m3 <- glm(Survive ~ Pulse, family = "binomial", data = ICU)
plotModel(m3)
```



# Quadratic Model

```
m3 <- glm(Survive ~ Pulse + I(Pulse^2),  
          family = "binomial", data = ICU)  
plotModel(m3)
```



Log odds of survival modeled as a quadratic function of Pulse

# Multiple Predictor Model

```
full.model <- glm(Survive ~ Age + SysBP + Pulse + I(Pulse^2),  
                 family = "binomial", data = ICU)  
summary(full.model) %>% coef() %>% round(digits = 3)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2.625	2.341	1.121	0.262
Age	-0.029	0.011	-2.662	0.008
SysBP	0.017	0.006	2.899	0.004
Pulse	-0.034	0.043	-0.795	0.426
I(Pulse^2)	0.000	0.000	0.780	0.436

How to interpret tests of individual coefficients? Just as in linear regression: is there evidence that including the predictor should affect predictions, given everything else already in the model?

# Checking For Multicollinearity

Multicollinearity is about relationships among predictors, so no difference from linear regression!

```
dplyr::select(ICU, Age, SysBP, Pulse) %>% cor() %>% round(digits = 2)
```

```
      Age SysBP Pulse
Age    1.00  0.04  0.04
SysBP  0.04  1.00 -0.06
Pulse  0.04 -0.06  1.00
```

```
vif(full.model)
```

```
      Age      SysBP      Pulse I(Pulse^2)
1.006822 1.010579 33.678411 33.754245
```

Not surprising about pulse since we have a polynomial.

# Checking For Multicollinearity

Check VIFs without the quadratic term

```
full.minus.quadratic <-  
  glm(Survive ~ Age + SysBP + Pulse, family = "binomial", data = ICU)  
vif(full.minus.quadratic)
```

Age	SysBP	Pulse
1.003413	1.005230	1.004810

Looks OK, so high VIFs in full model are *only* due to the polynomial.

# Outline

Multiple Predictors

Nested Model Tests

# Overall and Nested LR Tests

```
full.model <-  
  glm(Survive ~ Age + SysBP + Pulse + I(Pulse^2),  
      family = "binomial", data = ICU)  
no.pulse.model <-  
  glm(Survive ~ Age + SysBP, family = "binomial", data = ICU)  
anova(no.pulse.model, full.model, test = "LRT")
```

## Analysis of Deviance Table

Model 1: Survive ~ Age + SysBP

Model 2: Survive ~ Age + SysBP + Pulse + I(Pulse^2)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	197	183.25			
2	195	182.57	2	0.68431	0.7102

Test statistic:  $G :=$  Difference in (residual) deviance

# Nested Comparison: One vs. Two Curves

## Is Sex an important predictor, controlling for BP?

```
two.curves.model <- glm(Survive ~ SysBP + factor(Sex) + SysBP:factor(Sex),
                        family = 'binomial', data = ICU)
summary(two.curves.model) %>% coef() %>% round(3)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.439	1.021	-1.410	0.159
SysBP	0.023	0.008	2.762	0.006
factor(Sex)1	1.455	1.526	0.954	0.340
SysBP:factor(Sex)1	-0.013	0.012	-1.088	0.277

```
one.curve.model <- glm(Survive ~ SysBP, family = 'binomial', data = ICU)
anova(one.curve.model, two.curves.model, test = "LRT")
```

### Analysis of Deviance Table

Model 1: Survive ~ SysBP

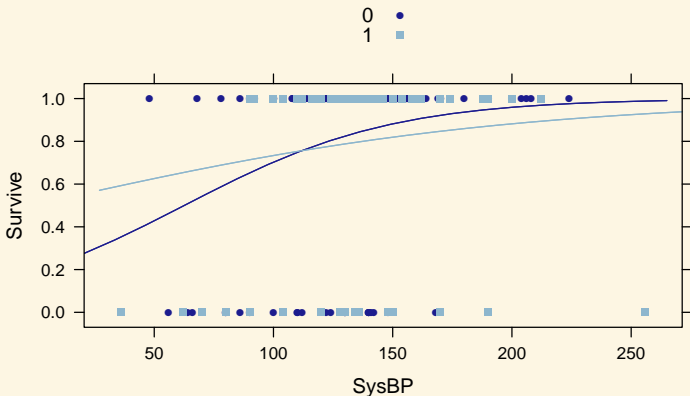
Model 2: Survive ~ SysBP + factor(Sex) + SysBP:factor(Sex)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	198	191.34			
2	196	189.99	2	1.3421	0.5112



# One vs. Two Curves

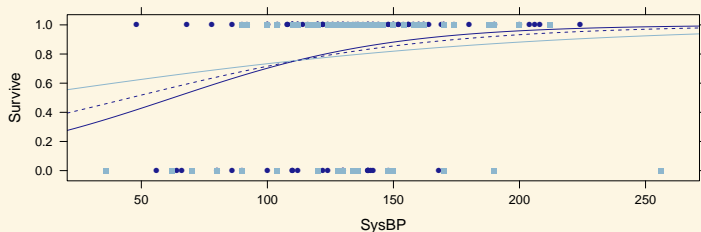
```
# This only works if binary variable is a factor  
plotModel(two.curves.model, auto.key = TRUE)
```



Little evidence that curves are different (in population)

# One vs. Two Curves

```
f.hat.two <- makeFun(two.curves.model)
f.hat.one <- makeFun(one.curve.model)
xyplot(Survive ~ SysBP, data = ICU, groups = factor(Sex))
plotFun(f.hat.two(SysBP, Sex) ~ SysBP, Sex = 0,
        xlim = c(0,300), col = 1, add = TRUE)
plotFun(f.hat.two(SysBP, Sex) ~ SysBP, Sex = 1, add = TRUE, col = 2)
plotFun(f.hat.one(SysBP) ~ SysBP, add = TRUE, lty = 2)
```



Little evidence that curves are different (in population)

# Parallel vs. Non-Parallel log odds lines

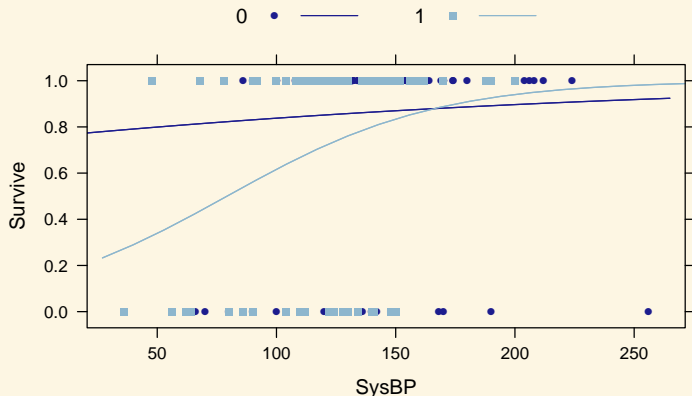
```
nonparallel.model <-  
  glm(Survive ~ SysBP + factor(Infection) + SysBP:factor(Infection),  
      family = 'binomial', data = ICU)  
summary(nonparallel.model)$coefficients %>% round(digits = 2)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.12	1.20	0.94	0.35
SysBP	0.01	0.01	0.60	0.55
factor(Infection)1	-2.93	1.59	-1.85	0.06
SysBP:factor(Infection)1	0.02	0.01	1.44	0.15

Weak evidence for different slopes (in log odds space)

# One vs. Two Curves

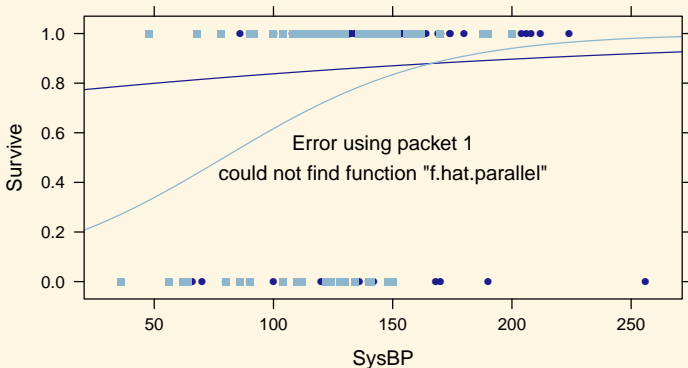
```
plotModel(nonparallel.model)
```



Weak evidence for different "slopes" (based on  $z$  test of interaction term)

# Parallel vs. Non-parallel logit lines

```
Error in makeFun(parallel.model): object 'parallel.model' not found
```



Weak evidence for different “slopes”

# One Line?

```
parallel.model <-  
  glm(Survive ~ SysBP + factor(Infection), family = 'binomial', data = ICU)  
summary(parallel.model) %>% coef() %>% round(3)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.124	0.836	-0.148	0.882
SysBP	0.015	0.006	2.384	0.017
factor(Infection)1	-0.729	0.373	-1.953	0.051

Borderline evidence for different “intercepts”