STAT 215
Multiple Logistic Regression

Colin Reimer Dawson

Oberlin College

November 16, 2017

Outline

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 (products) of variables
- Indicator variables for binary predictors
- Collections of $k-1$ indicators for categorical predictors w/ k levels

Two Equivalent Forms of (Multiple) Logistic Regression

Probability Form

$$\pi = \frac{e^{\beta_0+\beta_1 X+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}$$

Logit Form

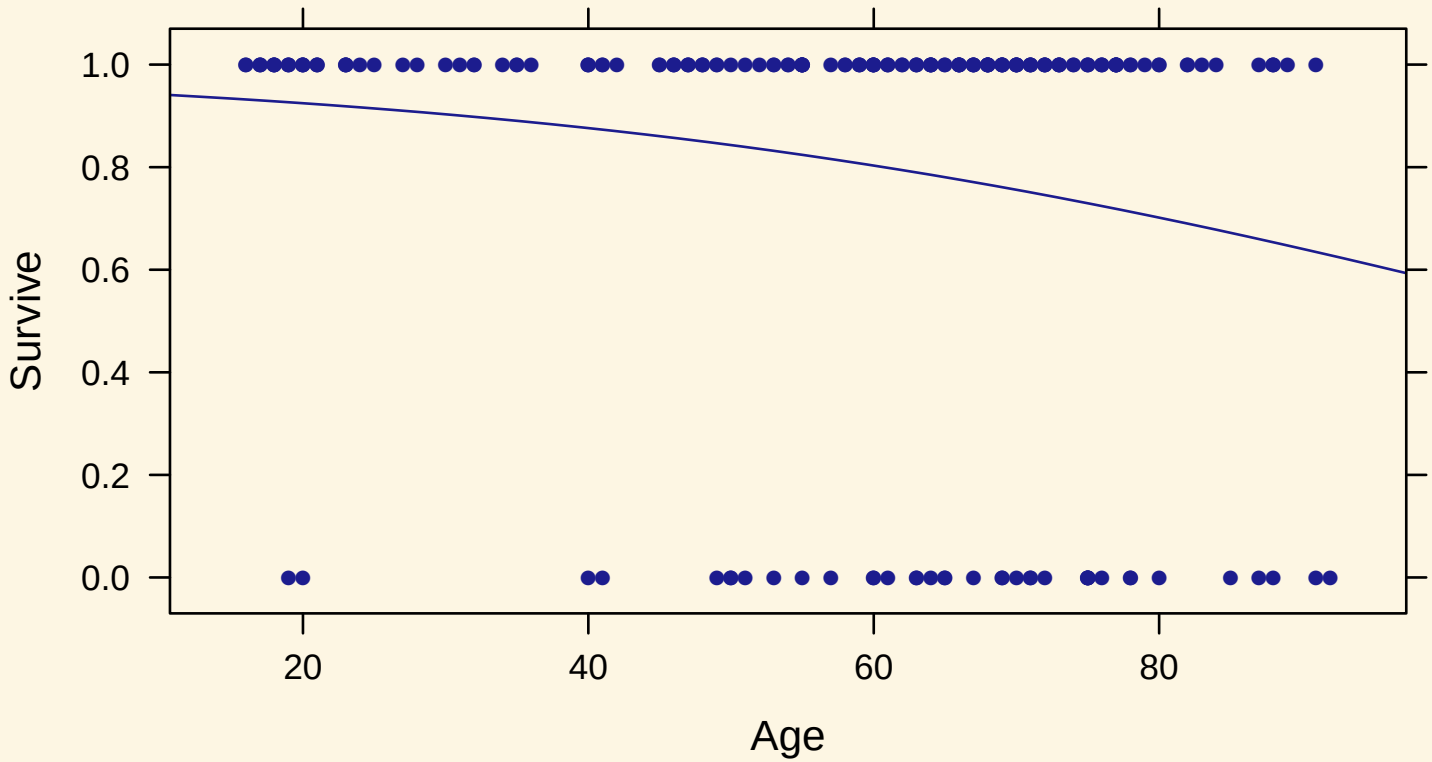
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k$$

Example: Survival in ICU

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

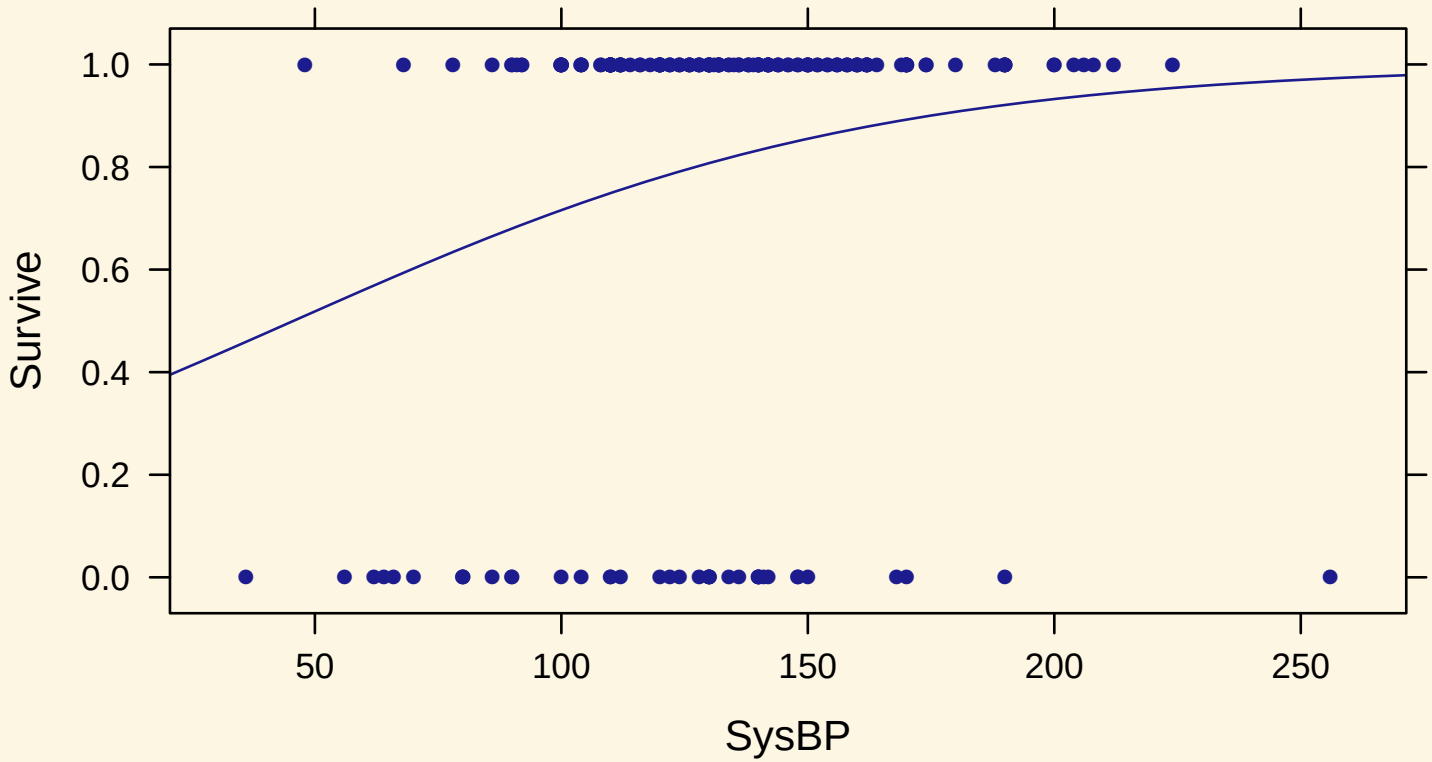
Simple Logistic Models

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



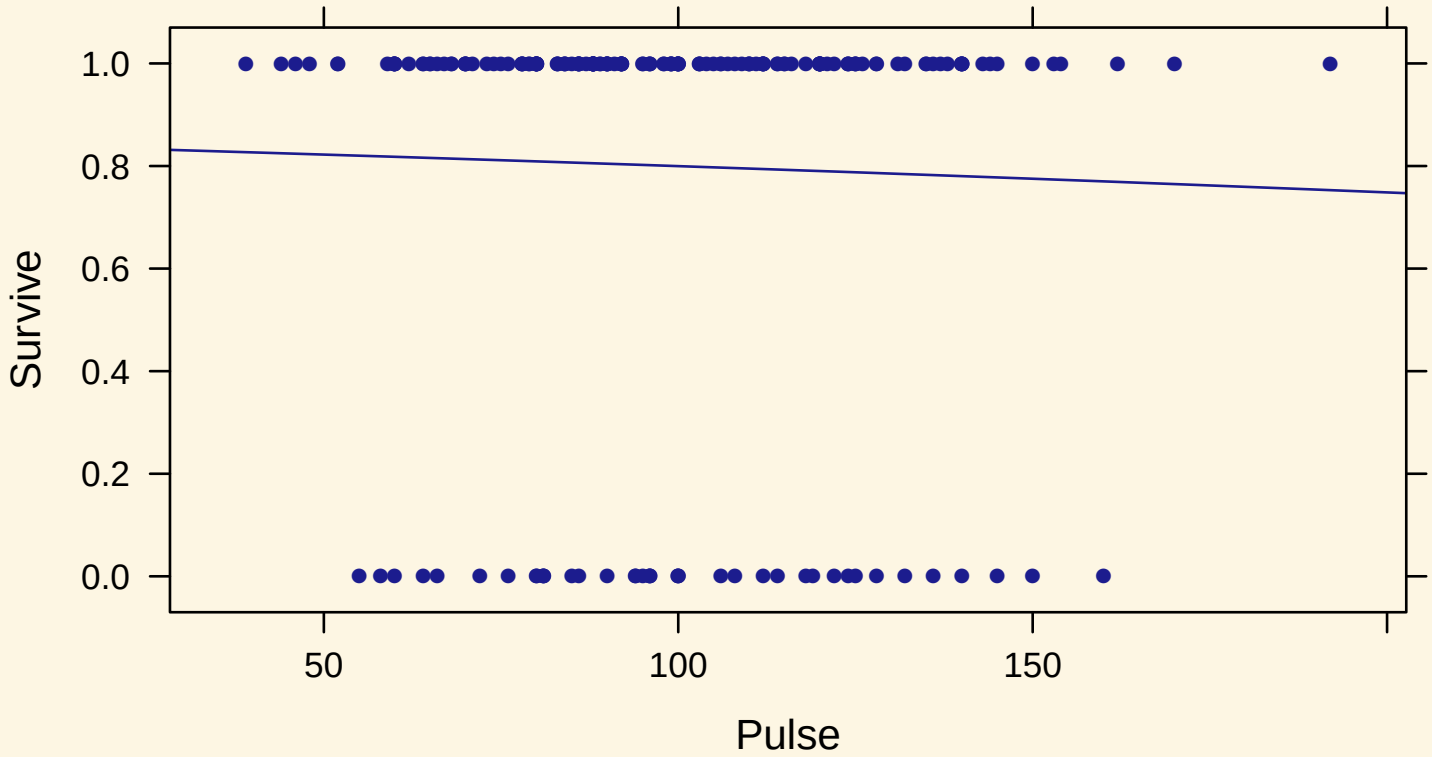
Simple Logistic Models

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



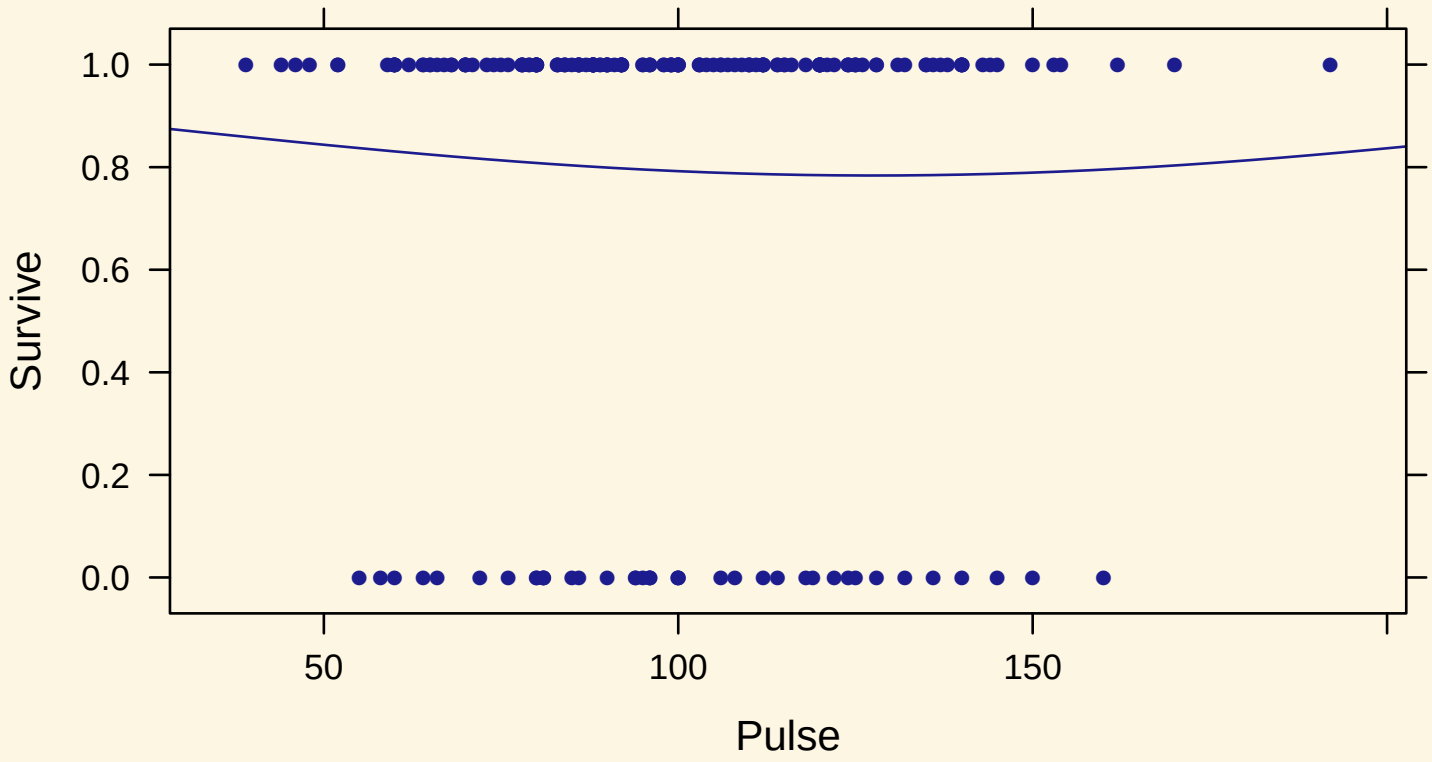
Simple Logistic Models

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



Simple Logistic Models

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



Multiple Predictor Model

```
full.model <- glm(Survive ~ Age + SysBP,
                  family = "binomial", data = ICU)
summary(full.model)$coefficients %>% round(digits = 3)
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.962      1.000   0.962    0.336
Age            -0.028      0.011  -2.637    0.008
SysBP           0.017      0.006   2.873    0.004
```

How to interpret tests of individual coefficients? Just as in linear regression: is the predictor adding something over the others?

Checking For Multicollinearity

Same issues with multicollinearity can arise!

```
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
1.001818 1.001818
```

But no worries in this case

Overall and Nested LR Tests

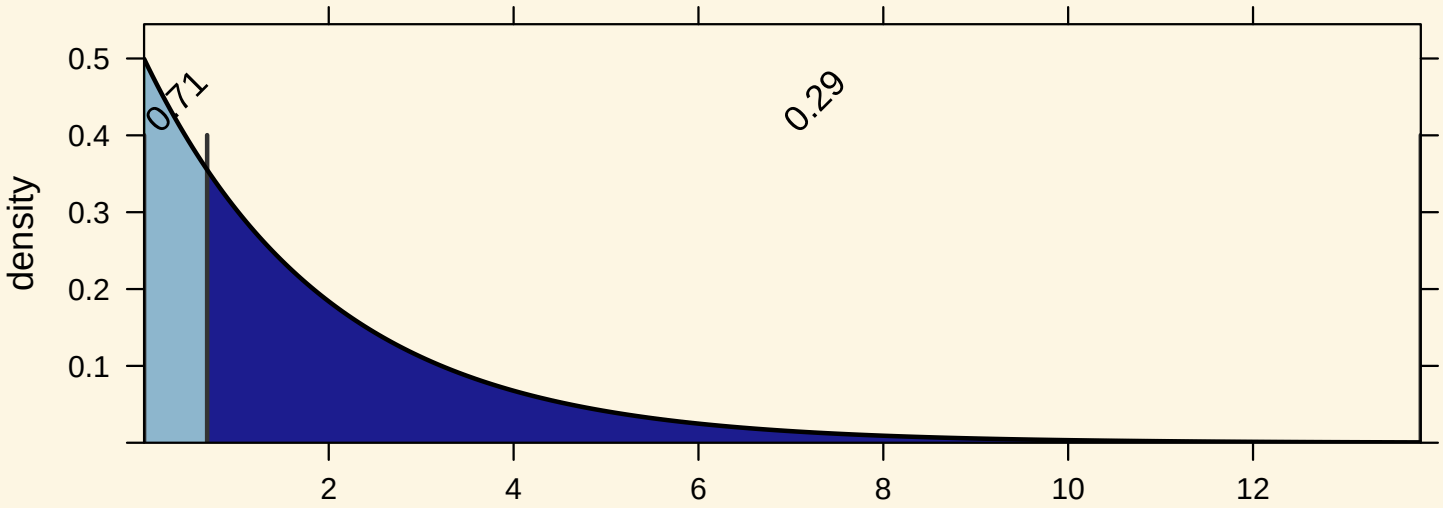
```
pulse.quad.model <-
    glm(Survive ~ Age + SysBP + Pulse + I(Pulse^2),
        family = "binomial", data = ICU)
no.pulse.model <-
    glm(Survive ~ Age + SysBP,
        family = "binomial", data = ICU)
anova(no.pulse.model, pulse.quad.model, test = "LRT")


     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:

$$G = -2(\log P(\text{Data} \mid \text{Full}) - \log P(\text{Data} \mid \text{Reduced}))$$

Overall and Nested LR Tests

```
xpchisq(0.68431, df = 2, lower.tail = FALSE)
```




```
[1] 0.7102381
```

One vs. Two Curves

Is Sex an important predictor, controlling for BP?

```
full.model <- glm(Survive ~ SysBP + factor(Sex) + SysBP:factor(Sex),
                  family = 'binomial', data = ICU)
summary(full.model)$coefficients
```

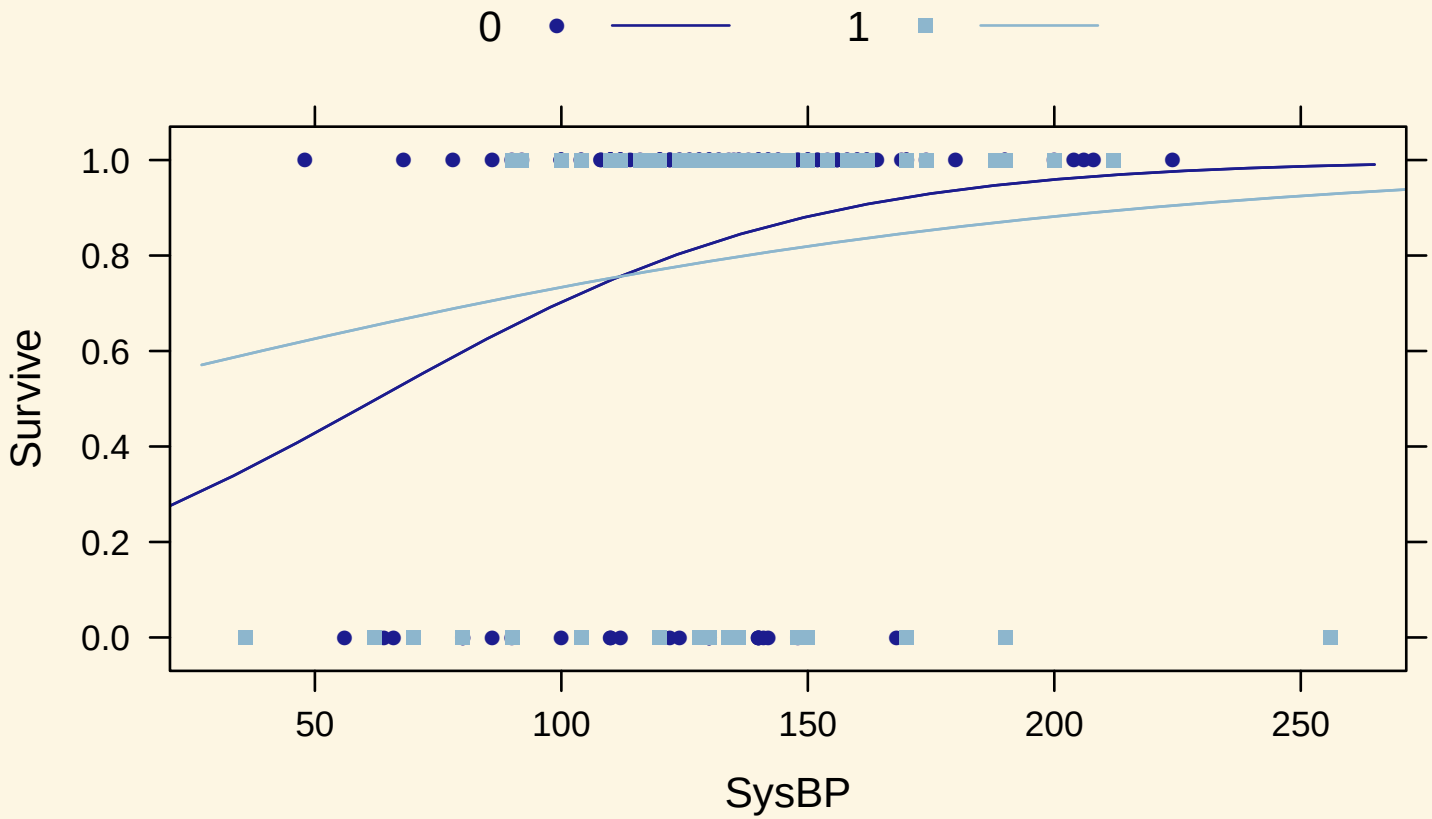
```
                          Estimate  Std. Error   z value    Pr(>|z|)
    (Intercept)        -1.43930431 1.021041657 -1.409643 0.158645099
    SysBP               0.02299392 0.008325432  2.761889 0.005746799
    factor(Sex)1        1.45516591 1.525558283  0.953858 0.340155546
    SysBP:factor(Sex)1 -0.01301957 0.011964883 -1.088148 0.276529569
```

```
reduced.model <- glm(Survive ~ SysBP, family = 'binomial', data = ICU)
anova(reduced.model, full.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

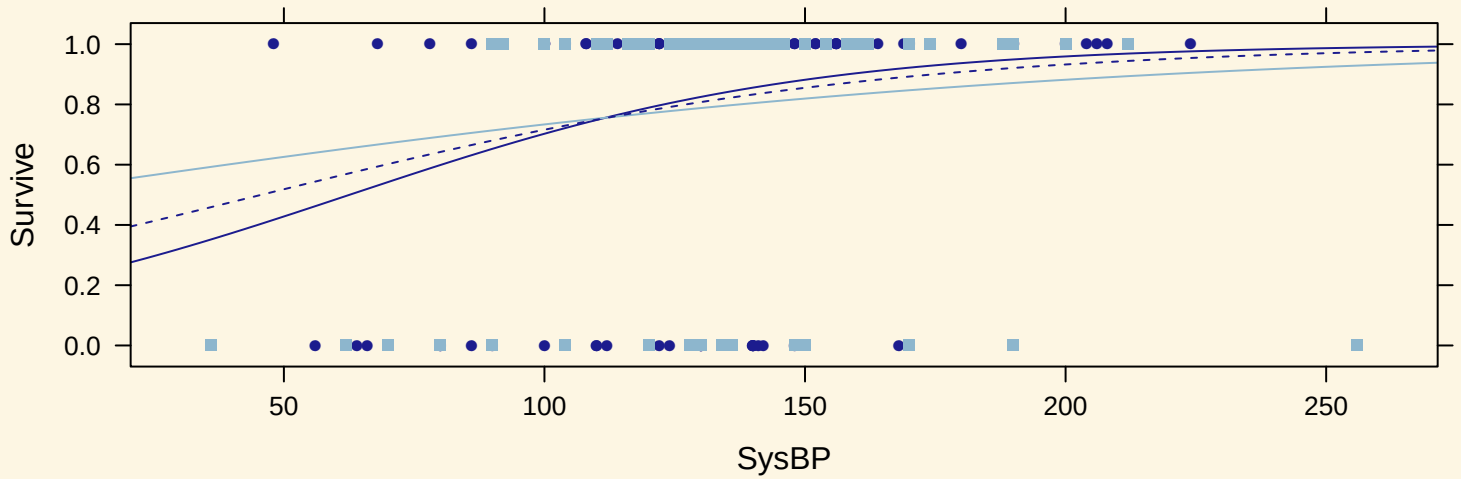
```
plotModel(full.model)
```



Curves are not significantly different

One vs. Two Curves

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



Curves are not signifcantly different

Parallel vs. Non-Parallel logit lines

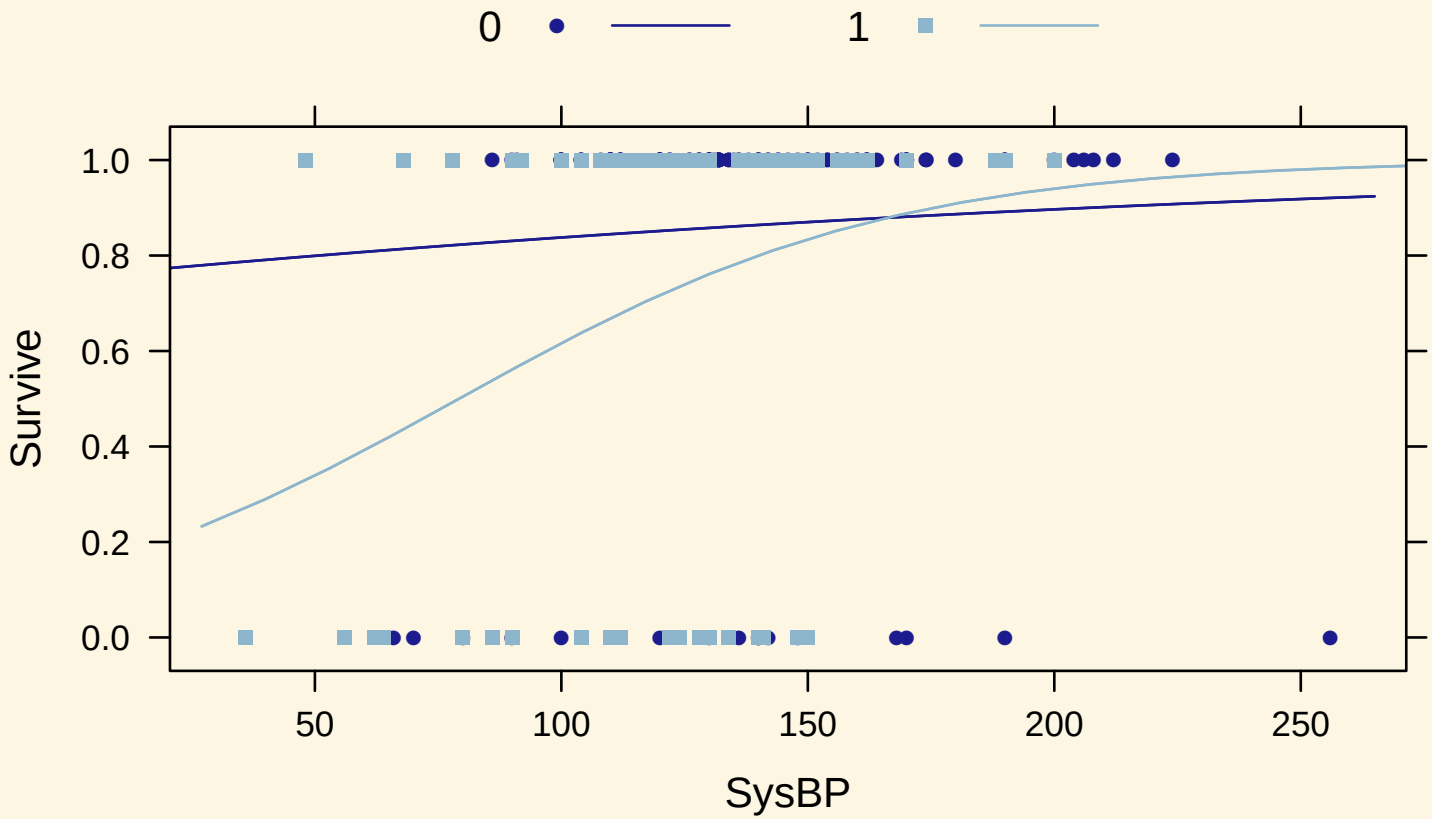
```
full.model <- glm(Survive ~ SysBP + factor(Infection) + SysBP:factor(Infection),
                  family = 'binomial', data = ICU)
summary(full.model)$coefficients %>% round(digits = 3)
```

```
                            Estimate Std. Error z value Pr(>|z|)
    (Intercept)                1.123      1.195   0.940    0.347
    SysBP                      0.005      0.009   0.601    0.548
    factor(Infection)1        -2.934      1.589  -1.846    0.065
    SysBP:factor(Infection)1   0.018      0.012   1.436    0.151
```

```
reduced.model <- glm(Survive ~ SysBP + factor(Infection), family = 'binomial',
```

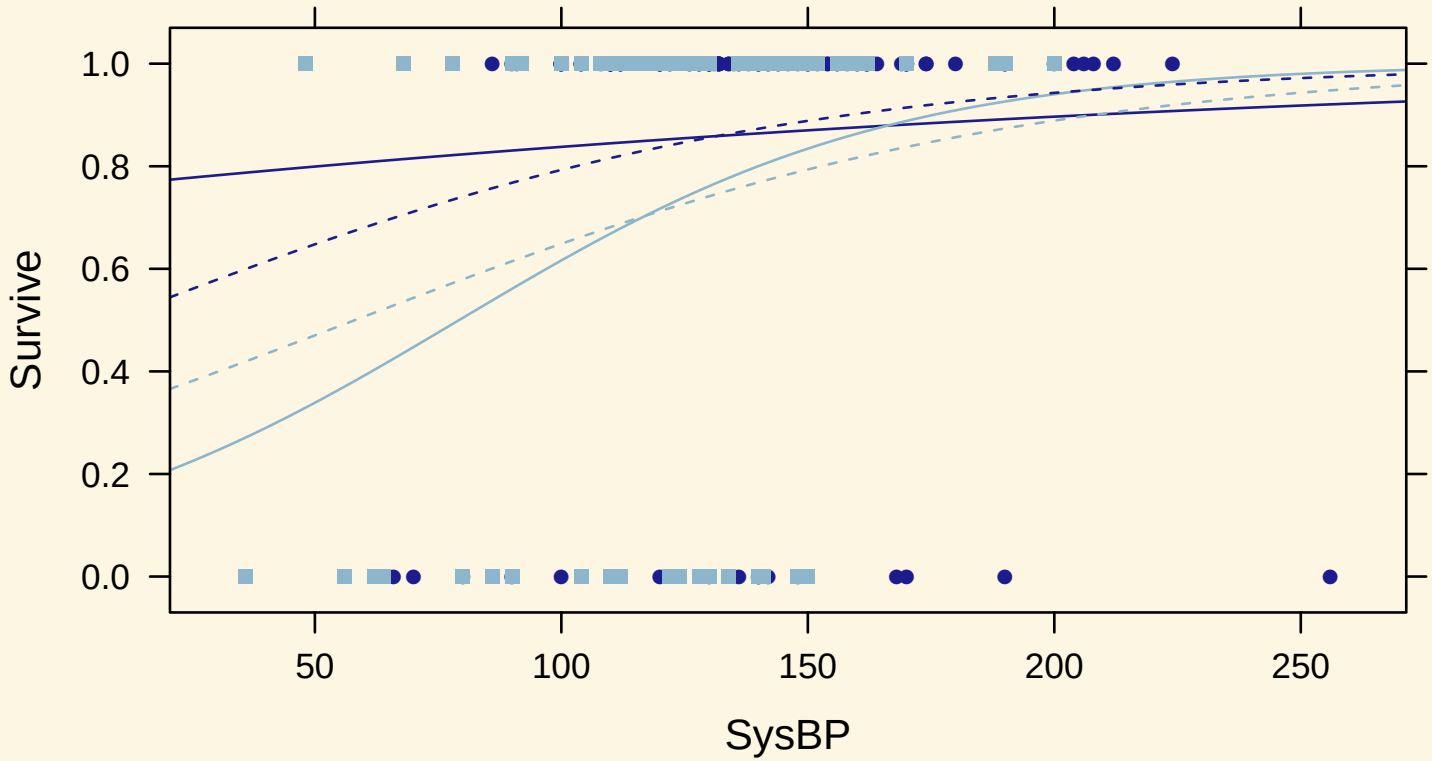
One vs. Two Curves

```
plotModel(full.model)
```



Curves do not have significantly different "slopes"

Parallel vs. Non-parallel logit lines



Lines are not signifcantly non-parallel

Model Selection Criteria

The usual metrics no longer apply:

- adj. R^2
- Mallow's C_p

Instead:

- **Akaike Information Criterion** (AIC): Deviance $+2p$ (lower is better)
- (Hard or Soft) Prediction Error (only evaluate out-of-sample)
  - Hard: How many cases did the model yield $\hat{\pi}$ on the wrong side of 1/2?
  - Soft: Sum absolute difference between $\hat{\pi}_i$ to Y_i

Stepwise Regression and Cross-Validation

Demo