Lecture 4 - Introduction to Logistic Regression

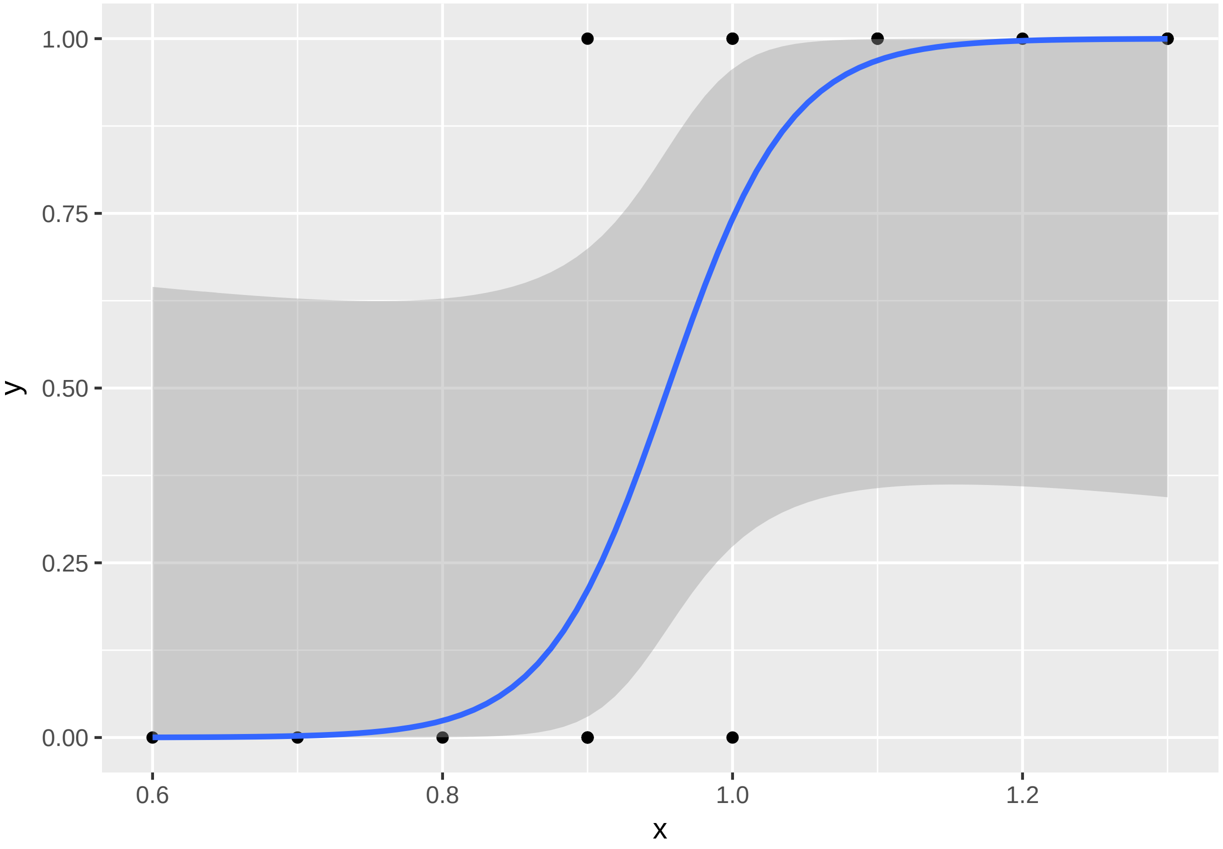
WILD 502- Jay Rotella

Logistic Regression using *glm* in R

Here, we'll work with the small dataset used in lecture to see how to implement logistic regression in **R**. This is a very small dataset that's used so that you can work through the likelihood calculations more easily than you could with a larger dataset. As you can see below, there is a lot of uncertainty in the estimates from this model.

```
library(ggplot2)
dat <- data.frame(
x=c(0.6, 0.7, 0.8, 0.9, 0.9, 0.9, 0.9, 1.0, 1.0, 1.0, 1.1, 1.1, 1.1, 1.2, 1.3, 1.3, 1.3),
y=c(0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1))

ggplot(dat, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "glm",
              method.args = list(family = "binomial"))
```



Above, we used `ggplot2` to run the simple model and view graphical output. Now, let's formally run the model in `glm` and view some of the numerical output.

Model summary, lnL, and AIC score

```
mod <- glm(y ~ x, data = dat, family = binomial)
summary(mod)
```

```
## 
## Call:
## glm(formula = y ~ x, family = binomial, data = dat)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.64751  -0.22083   0.02342   0.25122   1.76499  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)   -22.75      12.26  -1.855   0.0636 .
## x              23.81      12.82   1.857   0.0633 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 23.0348  on 16  degrees of freedom
## Residual deviance:  8.6897  on 15  degrees of freedom
## AIC: 12.69
## 
## Number of Fisher Scoring iterations: 7
```

```
logLik(mod)
```

```
## 'log Lik.' -4.344853 (df=2)
```

```
AIC(mod)
```

```
## [1] 12.68971
```

Covariance of Parameter Estimates

Here, we have very strong negative covariation between the intercept and slope. If you look back at the plot above, you can see that a lot of lines fit within the confidence bands. With negative covariation between the 2 estimates, as the estimate of the slope increases, the estimate of the intercept decreases and vice versa.

```
vcov(mod)
```

```
##             (Intercept)         x
## (Intercept)    150.3953 -156.8291
## x             -156.8291  164.3425
```

```
cov2cor(vcov(mod))
```

```
##             (Intercept)          x
## (Intercept)   1.0000000 -0.9975499
## x            -0.9975499  1.0000000
```

Deviance Residuals

You can ask **R** to print out the residuals, but because there are a variety of types of residuals that can be calculated, you have to specify that you want the deviance residuals. Notice that the sum of the squared deviance residuals is $-2 \cdot logLik$ and that $-2 \cdot logLik + 2 \cdot k = AIC$. Here, $k = 2$ as we have $\hat{\beta}_0$ and $\hat{\beta}_1$ and there is no $\hat{\sigma}$ in logistic regression because the errors are assumed to be binomially distributed.

```
(dev.resids <- residuals(mod, type = c("deviance")))
```

```
##           1           2           3           4           5           6 
## -0.02054956 -0.06753365 -0.22082974 -0.68780438 -0.68780438 -0.68780438 
##           7           8           9          10          11          12 
##  1.76498575 -1.64751167  0.77147863  0.77147863  0.25122184  0.25122184 
##          13          14          15          16          17 
##  0.25122184  0.07695355  0.02341952  0.02341952  0.02341952
```

```
sum(dev.resids ^ 2)
```

```
## [1] 8.689706
```

```
mod$deviance
```

```
## [1] 8.689706
```

```
AIC(mod)
```

```
## [1] 12.68971
```