Discrete Response Model
Lecture 2

datascience@berkeley

Visual Assessment of the Logistic Regression Model

Visual Assessment of the Logistic Regression Model With One Explanatory Variable

Because the response variable in logistic regression is binary, constructing a scatterplot with the raw binary response variable is not very informative because all plotted points would be at y = 0 or 1 on the y-axis.

Instead, we can plot the observed proportion of successes at each x instead to obtain a general understanding of how well the model fits the data.

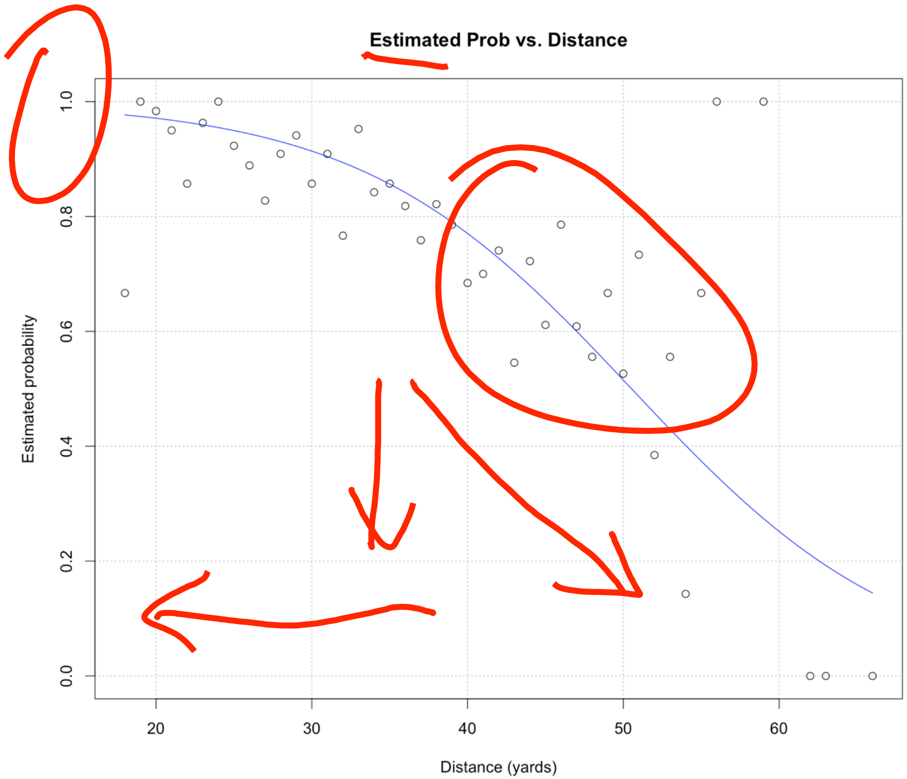
Note: This works well only when the number of observations at each possible x is not small. For truly continuous x, the number of observations would be 1 and this plot would not be very useful. An alternative for this situation include grouping observations by x and finding the observed proportion of successes for each group; however, this leads to potentially different results depending on how the grouping is done.

Example

```
> w<-aggregate(formula = good ~ distance, data = placekick, FUN = sum)
> n<-aggregate(formula = good ~ distance, data = placekick, FUN = length)
> w.n<-data.frame(distance = w$distance, success = w$good,
+     trials = n$good, proportion = round(w$good/n$good,4))
> head(w.n)
  distance success trials proportion
1       18       2      3     0.6667
2       19       7      7     1.0000
3       20     776    789     0.9835
4       21      19     20     0.9500
5       22      12     14     0.8571
6       23      26     27     0.9630
```

This was used to estimate a logistic regression model using a binomial response form of the data. Instead, we can plot the observed proportion of successes at each distance and overlay the estimated logistic regression model.

Aggregated Scatterplot of Estimated Probability vs. Distance

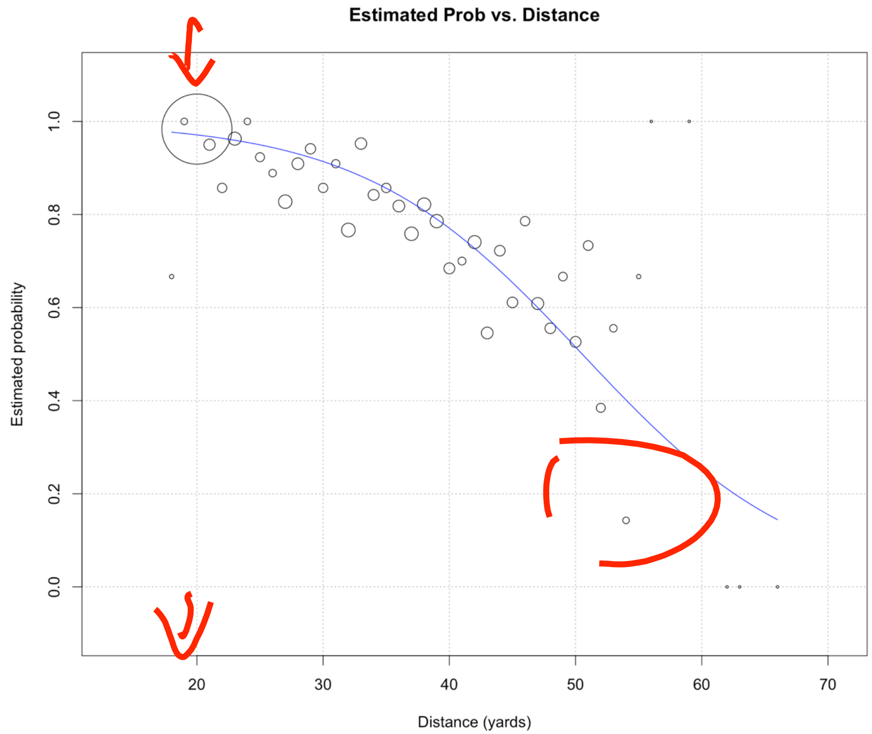



```
w<-aggregate(formula = good ~ distance, data = placekick, FUN = sum)
n<-aggregate(formula = good ~ distance, data = placekick, FUN = length)
w.n<-data.frame(distance = w$distance, success = w$good,
    trials = n$good, proportion = round(w$good/n$good,4))
head(w.n)

win.graph(width = 7, height = 6, pointsize = 12)
plot(x = w$distance, y = w$good/n$good, main="Estimated Prob vs. Distance",
    xlab="Distance (yards)", ylab="Estimated probability",
    panel.first = grid(col = "gray", lty = "dotted"))
```

Bubble Plot

To include a measure of how many observations are at each distance, we can use a bubble plot. For this plot, we make the plotting point size proportional to the observed number of observations at each unique distance.




```
win.graph(width = 7, height = 6, pointsize = 12)
symbols(x = w$distance, y = w$good/n$good, circles =
    sqrt(n$good), inches = 0.5,  main="Estimated Prob vs. Distance",
    xlab="Distance (yards)",
    ylab="Estimated probability", panel.first = grid(col =
    "gray", lty = "dotted"))
curve(expr = predict(object = mod.fit, newdata =
    data.frame(distance = x), type = "response"), col =
    "blue", add = TRUE, xlim = c(18, 66))
```

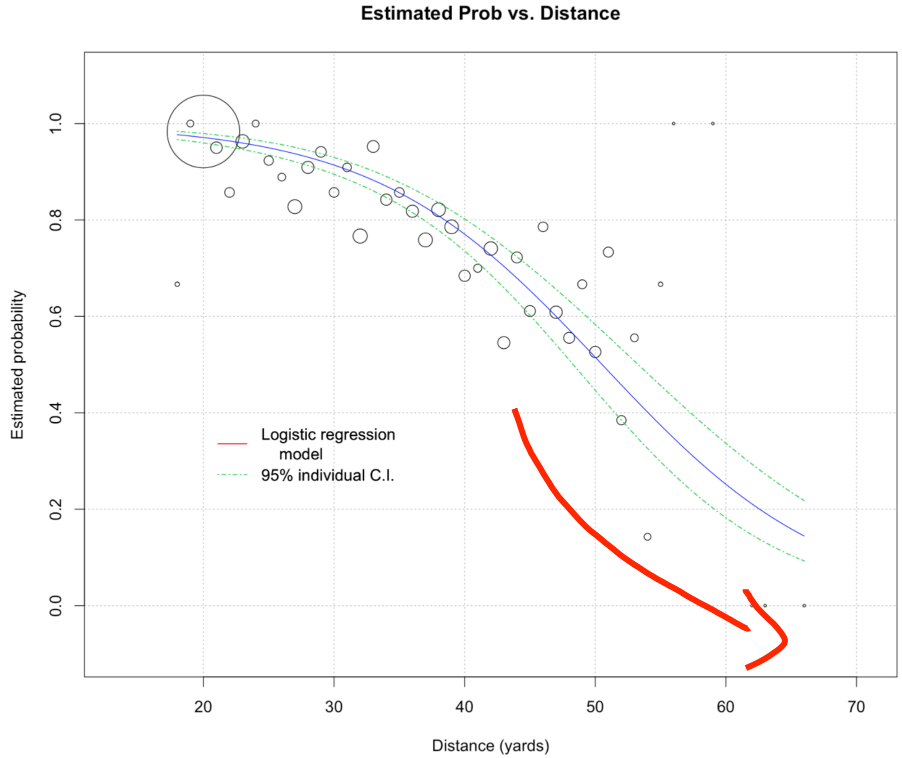
Further Remarks

Points that may have caused us concern before, now generally do not because we see they represent a small number of observations. For example,

- 18-yard placekicks: Only 3 observations occurred and there were 2 successes.
- The largest in distance placekicks: Many of these correspond to 1 observation only.

However, there are a few large plotting points, such as at 32 yards ($\hat{\pi}$ = 0.89, observed proportion = 23/30 = 0.77) and 51 yards ($\hat{\pi}$ = 0.49, observed proportion = 11/15 = 0.73), that may not be fit well by the model. How to more formally assess these observations and others will be an important subject of Week 5 when we examine model diagnostic measures.

Adding Confidence Bands to the Plot




```
# Add confidence bands to the previous plot
curve(expr = ci.pi(newdata = data.frame(distance = x),
    mod.fit.obj = mod.fit, alpha = 0.05)$lower, col =
    "green", lty = "dotdash", add = TRUE, xlim = c(18, 66))
curve(expr = ci.pi(newdata = data.frame(distance = x),
    mod.fit.obj = mod.fit, alpha = 0.05)$upper, col =
    "green", lty = "dotdash", add = TRUE, xlim = c(18, 66))
legend(x = 20, y = 0.4, legend = c("Logistic regression
    model", "95% individual C.I."), lty = c("solid",
    "dotdash"), col = c("red", "green"), bty = "n")
```

```
ci.pi<-function(newdata, mod.fit.obj, alpha){
  linear.pred<-predict(object = mod.fit.obj, newdata =
    newdata, type = "link", se = TRUE)
  CI.lin.pred.lower<-linear.pred$fit - qnorm(p = 1-
    alpha/2)*linear.pred$se
  CI.lin.pred.upper<-linear.pred$fit + qnorm(p = 1-
    alpha/2)*linear.pred$se
  CI.pi.lower<-exp(CI.lin.pred.lower) / (1 +
    exp(CI.lin.pred.lower))
  CI.pi.upper<-exp(CI.lin.pred.upper) / (1 +
    exp(CI.lin.pred.upper))
  list(lower = CI.pi.lower, upper = CI.pi.upper)
}
```

Berkeley SCHOOL OF INFORMATION