**Section 5.2 – Tools to assess model fit (continued, part 2)**

The purpose of this set of notes is to judge whether an observation is poorly fit by a model. This is not straightforward!

Example scenarios:

* Suppose a normal distribution approximation does actually work and M = 100. Approximately how many standardized Pearson residuals would you expect outside of ±2?
* Generally, if the model is o.k., one should expect to see a random scattering of the points in a plot of the standardized Pearson residuals vs. an important component of the model, such as an explanatory variable or . If instead there is some type of trend, this suggests there is a problem with the model, such as perhaps a transformation is needed of an explanatory variable. One just needs to be careful with looking for trend, and more will be discussed about this in the examples.

Example: Plots with about 50 points plotted on each

Random scattering of points, everything within ±2, assuming normal approximation works





Random scattering of points, everything within ±3, assuming normal approximation works





Random scattering of points, a few are outside of ±3, assuming normal approximation works





Random scattering of points, many more points than expected outside of ±2, assuming normal approximation works





Trend in points (could also plot an explanatory variable on the x-axis) assuming normal approximation works





Chapter 2 models: Points outside of ±3 when  is close to 0 and 1; normal approximation may not work

 

Chapter 2 models: Purely binary responses (M = n); notice separation between points (due to positive and negative responses)



Example: Placekicking (Placekick\_NotInBook.R, Placekick.csv)

Consider the logistic regression model with distance as the only explanatory variable.

> placekick <- read.csv(file = "C:\\data\\placekick.csv")

> w <- aggregate(x = good ~ distance, data = placekick, FUN = sum)

> n <- aggregate(x = good ~ distance, data = placekick, FUN = length)

> w.n <- data.frame(distance = w$distance, success = w$good, trials = n$good, prop = round(w$good/n$good,4))

> head(w.n)

 distance success trials prop

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

> mod.fit.bin <- glm(formula = success/trials ~ distance,

 weights = trials, family = binomial(link = logit), data

 = w.n)

Next, let’s find the standardized Pearson residuals along with probabilities from a binomial distribution to assess how unusual an observation is relative to the model.

Note: dbinomial() finds P(W = w) and pbinomial() finds P(W ≤ w)

> pi.hat <- predict(mod.fit.bin, type = "response")

> s.res <- rstandard(mod.fit.bin, type = "pearson")

> # If want Pearson residuals

> # p.res <- residuals(mod.fit.bin, type = "pearson")

> # P(W\_m <= w\_m)

> prob.smaller <- pbinom(q = w.n$success, size = w.n$trials, prob = pi.hat, lower.tail = TRUE)

> # P(W\_m >= w\_m)

> prob.higher <- pbinom(q = w.n$success, size = w.n$trials, prob = pi.hat, lower.tail = FALSE) + dbinom(x = w.n$success, size = w.n$trials, prob = pi.hat)

> # Mininum of P(W <= w\_m) and P(W >= w\_m)

> tail.prob <- apply(X = cbind(prob.smaller, prob.higher), MARGIN = 1, FUN = min)

> lin.pred <- mod.fit.bin$linear.predictors

> w.n <- data.frame(w.n, pi.hat, s.res, tail.prob, lin.pred)

> round(head(w.n), digits = 3)

 distance success trials prop pi.hat s.res tail.prob

1 18 2 3 0.667 0.977 -3.575 0.068

2 19 7 7 1.000 0.974 0.433 0.832

3 20 776 789 0.984 0.971 3.628 0.017

4 21 19 20 0.950 0.968 -0.448 0.483

5 22 12 14 0.857 0.964 -2.149 0.089

6 23 26 27 0.963 0.960 0.091 0.672

 lin.pred

1 3.742

2 3.627

3 3.512

4 3.397

5 3.281

6 3.166

> round(w.n[abs(w.n$s.res) > 2, ], digits = 3)

 distance success trials prop pi.hat s.res tail.prob

1 18 2 3 0.667 0.977 -3.575 0.068

3 20 776 789 0.984 0.971 3.628 0.017

5 22 12 14 0.857 0.964 -2.149 0.089

10 27 24 29 0.828 0.937 -2.476 0.032

15 32 23 30 0.767 0.894 -2.299 0.034

 lin.pred

1 3.742

3 3.512

5 3.281

10 2.706

15 2.131

> round(w.n[w.n$tail.prob < 0.025, ], digits = 3)

 distance success trials prop pi.hat s.res tail.prob

3 20 776 789 0.984 0.971 3.628 0.017

 lin.pred

3 3.512

The tail.prob variable is the minimum of P(Wm ≤ wm) and P(Wm ≥ wm) where Wm has a binomial distribution with trials nm and probability of success . This provides a measurement of how extreme the wm is relative to this estimated probability distribution. It can be used as an alternative to the rough ±2 or ±3 boundaries. Examine this value for the first 6 EVPs and compare to s.res relative to the ±3 boundary.

> # Standardized Pearson residual vs x plot

> plot(x = w.n$distance, y = w.n$s.res, xlab = "Distance",

 ylab = "Standardized Pearson residuals", main =

 "Standardized residuals vs. \n X")

> abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")

> # Add loess model to help visualize trend

> smooth.stand <- loess(formula = s.res ~ distance, data =

 w.n, weights = trials)

> # Make sure that loess estimates are ordered by "x" for

 the plots, so that they are displayed properly

> order.dist <- order(w.n$distance)

> lines(x = w.n$distance[order.dist], y =

 predict(smooth.stand)[order.dist], lty = 3, col =

 "red", lwd = 3)

See the first plot on the left corresponding to the code above.



Additional code in the program produces the other two plots. Because distance is the only variable in the model and it has a negative coefficient, the plot of the residuals against the linear predictor is just a mirror image of the plot against distance.

Comments:

* The residual at 18 yards is outside of the ±3 lines.
	+ I am not concerned about this because there are only 3 placekicks attempted from this distance (2 out of 3 were successes). Thus, a normal approximation would not be accurate. In fact, the only way this residual would not be outside of the ±3 lines is if all placekicks were successes!
	+ The tail.prob value is 0.068, suggesting it is definitely not as unusual as what would be suggested by a standard normal approximation (probability of 0.00018 for being this extreme).
	+ With binary outcomes, there are simply going to be examples like this where a failure occurs despite a high probability of success.
* The residual at 20 yards is also outside of the ±3 lines. In this case, there are 776 successes out of 789 trials (776/789 = 0.984) when the estimated probability of success is 0.971. The tail.prob value is 0.017. This is calculated as



where I used

> pbinom(q = 776, size = 789, prob = 0.971)

[1] 0.9910736

> pbinom(q = 776, size = 789, prob = 0.971, lower.tail = FALSE) + dbinom(x = 776, size = 789, prob = 0.971)

[1] 0.01720304

in R. Due to the small probability, I am somewhat concerned about this residual. Taking into my knowledge about football, I am even more concerned because there are two different types of placekicks typically observed at 20 yards – field goals and PATs. Perhaps another variable needs to be added to the model to account for this? More on this in a later section!

* There are a few other standardized Pearson residuals outside of ±2, but within ±3. Are these observations of concern?
* If you are not familiar with loess regression models, they provide a way to determine trend in a scatter plot. My normal linear regression course gives further details about these models. One important aspect of their application is that predicted values from loess models can have a large amount of variability where data are sparse or near the extreme values of the variable on the x-axis. For our plots here, I focus on the middle of the plots and see some upward trend, suggesting there may be some problems.
* If there were not potential problems with a model, where would you expect the loess curve to be on a plot?