STATISTICS 608 Linear Models - Final Exam
May 7, 2015

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **14** pages including this cover page.
2. You have exactly 2 hours to complete the exam.
3. There may be more than one correct answer; choose the best answer.
4. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
5. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions next week.
6. You may use one 8.5" X 11" sheet of notes and a calculator.
7. At the end of the exam, leave your sheet of notes with your proctor along with the exam.

I attest that I spent no more than 2 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature: _____

INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

1. I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. I certify that the student has left the exam and sheet of notes with me, to be returned to the student no less than one week after the exam or shredded.
5. I certify that the copy of the exam that is uploaded is readable and in PDF format.

Proctor's Signature: _____

Part I: Multiple choice - Circle the best answer for each of the following.

1. Suppose the probability of an event occurring is between 0 and 0.5. What does that tell you about the odds of the event occurring? Be as specific as possible.

   (a) The odds are smaller than 0.

   (b) The odds are smaller than 0.5.

   (c) The odds are smaller than 1. 0.5 / (1 - 0.5) = 1

   (d) The odds are greater than 0.5.

   (e) The odds are greater than 1.

2. If we were going to use a weighted least squares regression model when $\text{Var}(e_i|x_i) = x_i^2\sigma^2$, which of the following would we use for the weights w_i in the model?

   (a) x_i

   (b) $1/x_i$

   (c) x_i^2

   (d) $1/x_i^2$

   (e) $\sqrt{x_i}$

   (f) $1/\sqrt{x_i}$

3. Suppose that a predictor variable x is Poisson distributed. For a logistic regression model shown below, which of the following is true?

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + e,$$

   where $\theta(x) = P(Y = 1|X = x)$.

   (a) The log odds are always a linear function of x. Notice x is POISSON.

   (b) The log odds are a linear function of x if the means of x are equal when $Y = 1$ and when $Y = 0$.

   (c) The log odds are a linear function of x if the variances of x are equal when $Y = 1$ and when $Y = 0$.

   (d) The log odds are a linear function of x if the means and variances of x are equal when $Y = 1$ and when $Y = 0$.

4. Suppose we fit a simple linear regression model $y = \beta_0 + \beta_1 x + e$, but the errors, while being normally distributed with mean 0, are autocorrelated. Which of the following statistics will be estimated incorrectly if we assume that the errors are independent?

   (a) sample slope

   (b) sample intercept

   (c) standard errror of the slope, thus t-values and p-values as well

   (d) predicted value of y for given x

5. Suppose we have two logistic regression models, $\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_{0i} + \beta_{1i}\, x, i = 1, 2$, where $\theta(x) = P(Y = 1|X = x)$ that have the same intercept: $\beta_{01} = \beta_{02}$, but different slopes: $\beta_{11} \neq \beta_{12}$. Which of the following statements is always true?

   (a) The graphs of log(odds) versus x will cross the y-axis at different values of y. They have the same intercept.

   (b) The graphs of log(odds) versus x will cross the horizontal line log(odds) $= 0.5$ at the same value of x.The two lines intersect at the intercept, which is at $x = 0$, so they can't intersect anywhere else. Unless log(odds) $= 0.5$ when $x = 0$, this can't be true.

   (c) The graphs of $\theta(x)$ versus x will cross the horizontal line $\theta(x) = 0.5$ at the same value of x. The graphs cross that horizontal line at the point $-\beta_1/4$, which is different for the two models.

   (d) The graphs of $\theta(x)$ versus x will cross the vertical line $x = 0$ at the same value of $\theta(x)$.

6. A real estate agent wanted to predict $y =$ price of a house in California using $x_1 =$ house size and x_2 is an indicator variable, takinig the value 1 if the house has an ocean-front view and 0 otherwise. What does the parameter β_2 represent in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$?

   (a) β_2 is the intercept for ocean-front houses.

   (b) β_2 is the intercept for houses without an ocean-front view.

   (c) β_2 is the difference between the the intercepts for ocean-front houses and houses without an ocean-front view.

   (d) β_2 is the slope for ocean-front houses.

   (e) β_2 is the difference between the slopes for ocean-front houses and houses without an ocean-front view.

7. Suppose we have three simple linear regression models (involving only one predictor variable). In Model 1, all of the standard assumptions are met. In Model 2, all the assumptions except normality of the residuals are met. In Model 3, all the assumptions except constant variance are met. For which models are prediction intervals calculated in the usual way appropriate? **Circle all that apply.**

   (a) Model 1

(b) Model 2 Remember that if the the residuals aren't normal, prediction intervals aren't appropriate: prediction intervals are for individuals. We need those individuals to be normally distributed.

(c) Model 3 If variance is not constant, you need both prediction and confidence intervals to be different widths for different values of x.

Part II: Long Answer

8. The results from a model predicting y = graduation rate (proportion of incoming freshmen who graduate in four years) using x = median SAT score for incoming students at $n = 19$ public universities in the Big 12 and SEC conferences are output below. Use this output to answer the following questions.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0697439  0.3644412  -2.935 0.009244 **
MedSAT       0.0012662  0.0003111   4.070 0.000796 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.06708 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4936,Adjusted R-squared:  0.4638
F-statistic: 16.57 on 1 and 17 DF,  p-value: 0.000796
```

(a) Assuming appropriate assumptions are met, calculate a 95% confidence interval for the slope of the model. The t-critical value for a 95% confidence interval with 17 degrees of freedom is 2.110. Show your work.

$$0.0012662 \pm 2.110(0.0003111)$$
$$(0.0006, 0.002)$$

Remember that the standard error is the estimate of the standard deviation of a statistic. You don't need to take its square root.

(b) A 95% confidence interval for the predicted four-year graduation rate when the median SAT score is 1210 is (0.42, 0.50). Interpret this confidence interval in context. (I picked 1210 because that's TAMU's median SAT score.)

We are 95% confident that the **average** four-year graduation rate among Big 12 and SEC universities with median SAT score 1210 is between 42% and 50%.

9. For the logistic regression model $\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0+\beta_1 x_1+\beta_2 x_2+e$, show that the predicted odds ratio of a success for $x_2 = 12$ to $x_2 = 10$ is $\exp(2\hat{\beta}_2)$, holding x_1 constant.

$$\begin{aligned}
\log\left(\frac{\hat{\theta}(12)}{1-\hat{\theta}(12)}\right) - \log\left(\frac{\hat{\theta}(10)}{1-\hat{\theta}(10)}\right) &= [\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 12] - [\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 10] \\
\log\left(\frac{\hat{\theta}(12)/(1-\hat{\theta}(12))}{\hat{\theta}(10)/(1-\hat{\theta}(10))}\right) &= \hat{\beta}_2(12-10) \\
\text{Odds Ratio} &= \exp(2\hat{\beta}_2)
\end{aligned}$$

10. Suppose that Y is a binary random variable (taking values 0 and 1), and that we use a single predictor variable X to predict Y. When X is discrete, we note that the log odds that $Y = 1$ can be expressed as:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right)$$

Suppose that X is a dummy variable, taking the value 1 with probability $\pi_j, j = 0, 1$ conditional on $Y = 0, 1$. Show that the log odds that $Y = 1$ are a linear function of x. (Recall that the probability distribution function for a Bernoulli random variable Z with success probability π is $P(Z = z) = \pi^z(1-\pi)^{1-z}, z = 0, 1.$)

To prove this from the beginning (rather than using the results above), let's first recall the definition of conditional probability: $P(A|B) = P(A \cap B)/P(B)$. Thus we also have $P(A \cap B) = P(A|B)P(B)$. Thus we have:

$$\begin{aligned}
\log\left(\frac{\theta(x)}{1-\theta(x)}\right) &= \log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) \\
&= \log\left(\frac{P(Y=1\cap X=x)/P(X=x)}{P(Y=0\cap X=x)/P(X=x)}\right) \\
&= \log\left(\frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{\pi_1^x(1-\pi_1)^{1-x}}{\pi_0^x(1-\pi_0)^{1-x}}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + x\log\left(\frac{\pi_1}{\pi_0}\right) + (1-x)\log\left(\frac{1-\pi_1}{1-\pi_0}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{1-\pi_1}{1-\pi_0}\right) + x\left[\log\left(\frac{\pi_1}{\pi_0}\right) - \log\left(\frac{1-\pi_1}{1-\pi_0}\right)\right]
\end{aligned}$$

So we have that the log odds are linear in x, with an intercept of $\log\left(\frac{P(Y=1)(1-\pi_1)}{P(Y=0)(1-\pi_0)}\right)$ and a slope of $\log\left(\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_0)}\right)$.

11. We are interested in predicting $Y =$ monthly sales (in hundreds of dollars) from a bookstore using predictors advertising spent in previous month and indicator (dummy) variables for November and December; we've taken the logs of Sales and Advertising:

$$\log(Sales) = \beta_0 + \beta_1 \log(Advertising) + \beta_2 iNov + \beta_3 iDec + e$$

Output is labeled "Bookstore Sales" in the appendix. (We've added 1 to Advertising before taking the log in the actual model to prevent taking log(0).)

(a) Is this a valid model? Give the most important reason why or why not.

No, the model is not valid; the data are collected monthly, and so are time series data. Notice the residuals have significant autocorrelation at lags 8, 11, and 12 (even after accounting for differences in November and December).

(b) A suggestion is made that we consider adding an interaction term between $iDec$ and $\log(Advertising)$. What would the interaction term imply in context?

It would mean that advertising had a different effect on sales in December from months January through October. (Notice November has its own indicator variable.)

(c) Assume the model is valid. Is it necessary to include the indicator for November in this model? Conduct a hypothesis test using $\alpha = 0.05$. Be sure to include hypotheses, appropriate statistics, and a conclusion in context.

$H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$
$t = 0.335772/0.07412261 = 4.53$, p-value ≈ 0
We have evidence that the indicator for November is necessary; our model predicts that November has significantly different sales from months January through October.

12. Data from the 2000 through 2008 seasons of the National Footbal League have been used to predict whether a player makes a field goal from kicks attempted at varying distances (in yards) away from the field goal. (The minimum distance was 18 yards because the shortest possible kick is from the opponent's 1-yard line, with the kicker standing 7 yards further away from that, plus the 10-yard depth of the end zone.) In these nine years, multiple attempted field goals are available at almost every distance, so we will use a binomial, rather than a binary, model.

    Output is labeled "NFL" in the appendix.

    (a) First, consider Model 1. Regardless of whether it fits well, interpret the parameter estimate for β_1 in context.

$$\text{Model 1: } \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + e$$

    If a kicker moves one more yard further from the field goal, our model predicts his odds to be multiplied by 0.9; that is, his odds are decreased by about 10%.

    (b) A second model is also being considered to model the relationship between $x =$ distance from the field goal and whether the kicker makes the field goal:

$$\text{Model 2: } \log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

    Regardless of whether the second model fits well, use the second model to estimate the probability of an NFL player making a field goal from a distance of 40 yards. Show your work.

    $6.774 - 0.173(40) + 0.0009(40^2) = 1.249675$

$$\hat{\theta}(x) = \frac{1}{1+\exp(-1.249675)} = 0.78$$

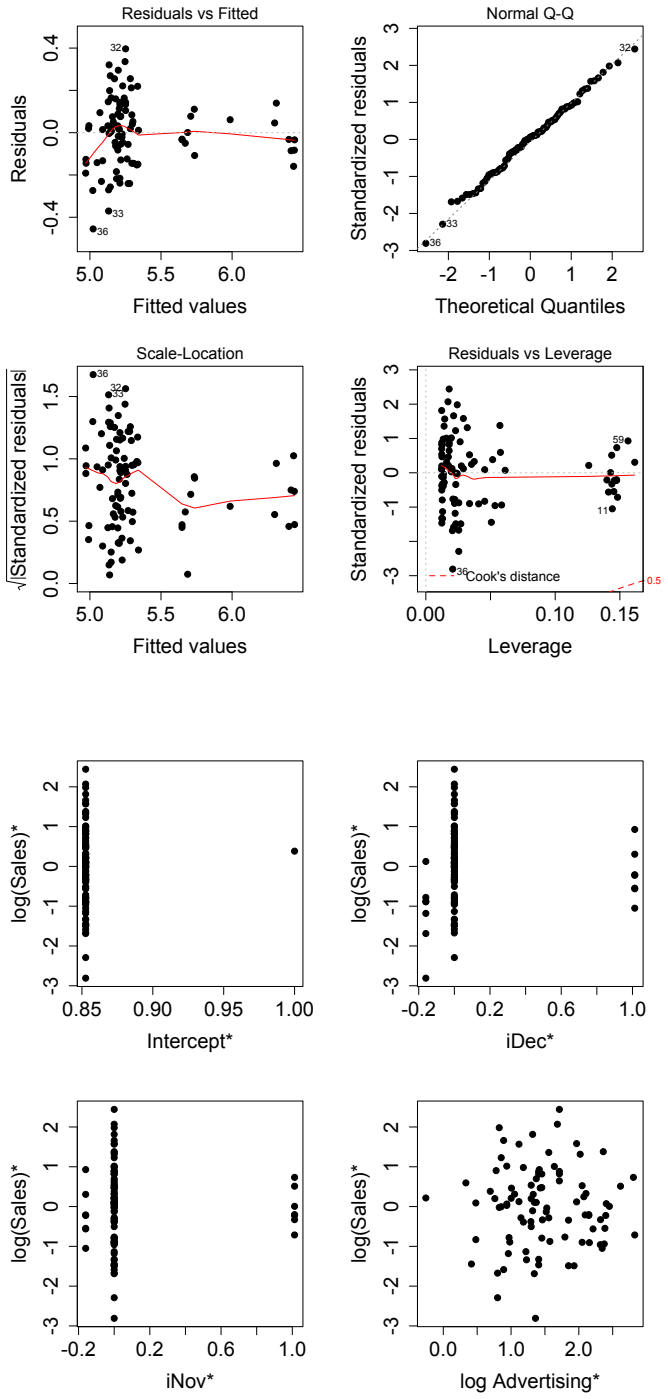
(c) Use the output in the appendix to carefully explain which model, Model 1 or Model 2, is preferable. Give three reasons why that model is preferable. At least one reason should not involve any hypothesis tests. Use a significance level of $\alpha = 0.05$ for any tests you conduct. (Areas to the right of three chi-square statistics are included.)

The second model is preferable for the following reasons:

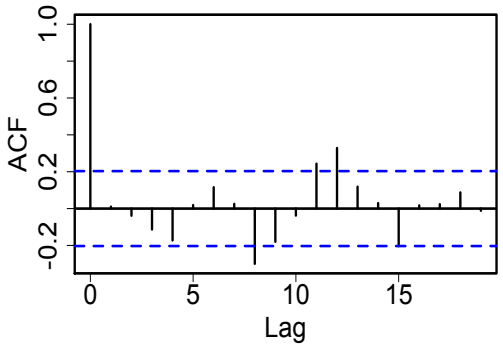
- First, notice that the variances in distributions of distance from the field goal are different depending on the value of Y (whether the player made the kick or not). If x is normally distributed, we take this as a sign that the squared term may be necessary in the model.
- The plot of the logistic model against a nonparametric smooth appears to match the second model slightly better than the first, except for the very largest distances (which don't happen very often). (There weren't quite enough field goals at these distances for the binomial model to be appropriate, anyway. We really need more attempted field goals at these distances.)
- AIC is lower for the second model than the first.
- In a test of $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$ (that is, that model 2 is significantly better than model 1), we reject the null hypothesis. We can use the Wald test to do this, in which case we have $z = 2.479$ and p-value $= 0.0132$, or preferably we can use the deviance test, in which case we have a chi-squared test statistic of 66.863 - 60.625 $= 6.238$ with 46 - 45 $= 1$ df, and a p-value of 0.0125. The Wald test and the deviance test p-values match pretty well in this case. Either way, we conclude that we have evidence that the squared term is necessary in the model.
- We can also conduct Goodness of Fit tests for both models. Notice that we cannot simply say "Model 2 fits well" without also discussing the first model; to compare two models, we need to say that one model fits *better* than the other. For both goodness of fit tests, the null hypothesis is that the logistic regression model predicts the probability of making a kick as well as the saturated model, and the alternative hypothesis is that the saturated model is better (implying the logistic regression model does not fit well). For the first model, we have a chi-squared test statistic of 66.863 with 46 df, yielding a p-value of 0.0238, and for the second the test statistic is 60.625 with 45 df, yielding a p-value of 0.0598. This means we have evidence the saturated model is better than the first logistic model, meaning the first logistic model doesn't fit well, while we don't in the second model, so we don't have evidence the saturated model fits better than the second model. That is, the first model doesn't fit while the second does, which is a reason to prefer the second model over the first.

Appendix

Bookstore Sales



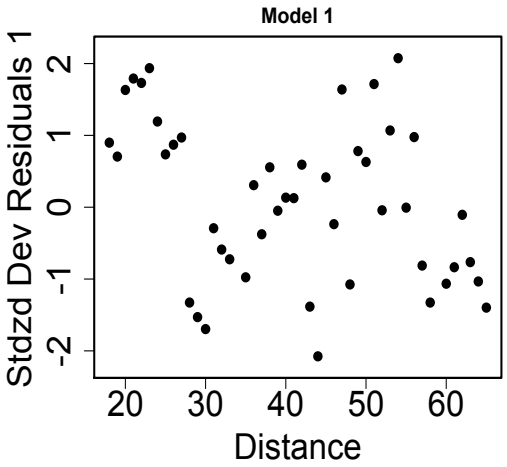
Bookstore Sales

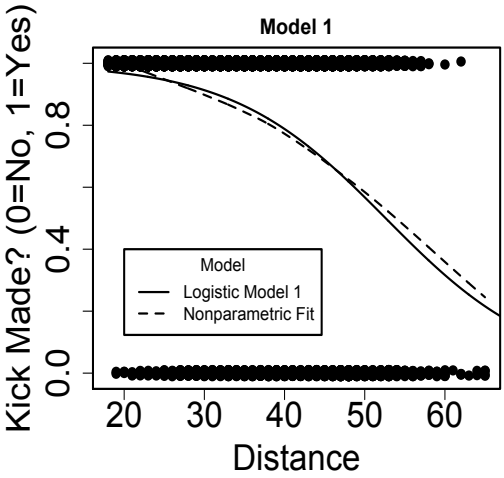



```
Generalized least squares fit by maximum likelihood
  Model: log(Sales) ~ Month_12 + Month_11 + log(Advert + 1)
  Data: books
       AIC      BIC   logLik
  -66.9709 -51.7753 39.48545

Correlation Structure: AR(1)
 Formula: ~Time
 Parameter estimate(s):
      Phi
0.1580911

Coefficients:
                   Value  Std.Error   t-value p-value
(Intercept)     5.895485 0.05839944 100.95106   <2e-16
Month_12        1.138222 0.06988855  16.28625   <2e-16
Month_11        0.335772 0.07412261   4.52996   <2e-16
log(Advert + 1) 0.130892 0.03437437   3.80782   3e-04
```

NFL

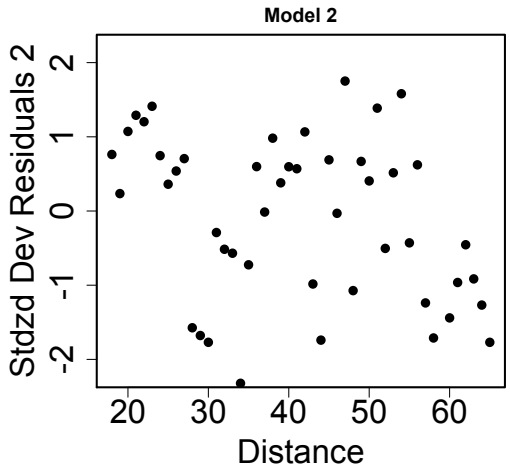
Model 1:

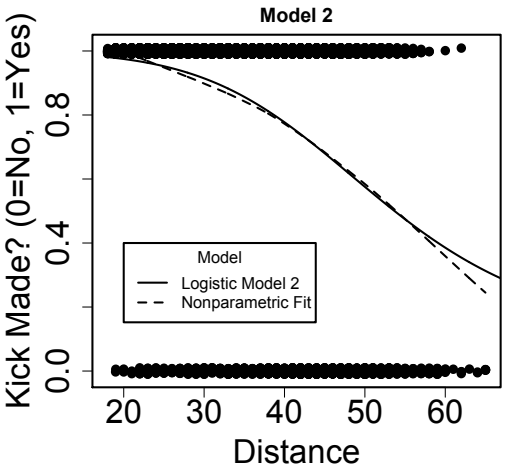





```
Call:
glm(formula = cbind(Makes, not.makes) ~ Dist, family = binomial())

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.73219  -0.97376  -0.02361   0.86066   2.02072

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.477986   0.147745   37.08   <2e-16 ***
Dist        -0.104129   0.003495  -29.79   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1179.206  on 47  degrees of freedom
Residual deviance:   66.863  on 46  degrees of freedom
AIC: 267.26
```

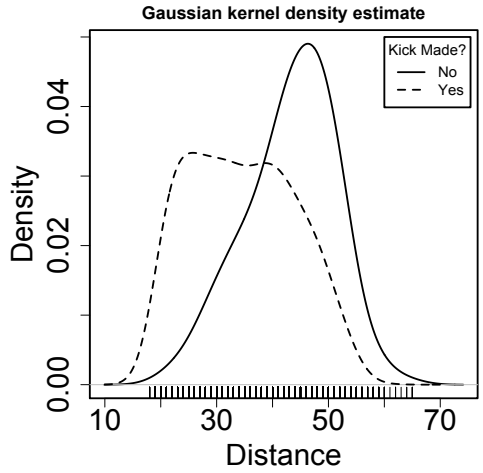
NFL Model 2:

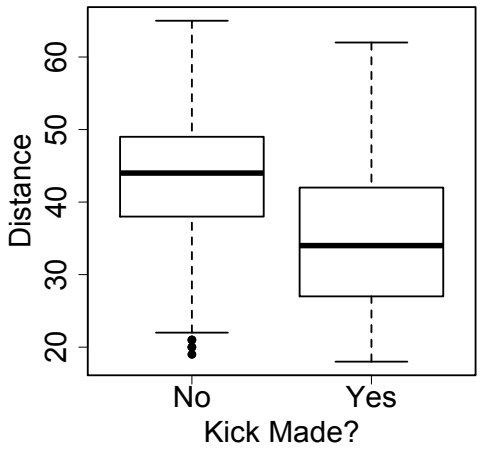





```
Call:
glm(formula = cbind(Makes, not.makes) ~ Dist + I(Dist^2), family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0421  -0.9760   0.1073   0.6714   1.6861

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.7736510  0.5497692  12.321  < 2e-16 ***
Dist        -0.1729634  0.0281231  -6.150  7.74e-10 ***
I(Dist^2)    0.0008716  0.0003516   2.479   0.0132 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1179.206  on 47  degrees of freedom
Residual deviance:   60.625  on 45  degrees of freedom
AIC: 263.02
```

NFL





> **pchisq(66.863, 46, lower.tail=FALSE)**
0.0238

> **pchisq(60.625, 45, lower.tail=FALSE)**
0.0598

> **pchisq(6.238, 1, lower.tail=FALSE)**
0.0125