STATISTICS 408 / 608 Linear Models - Final Exam
May 4, 2017

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **15** pages including this cover page.
2. You have exactly 2 hours to complete the exam.
3. Complete the exam on this form.
4. Pen is preferred as it usually scans more clearly, though blue pen, black pen, or pencil are allowed.
5. There may be more than one correct answer; choose the **best** answer.
6. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
7. You may use **three** 8.5" X 11" sheets of notes and a calculator. (Clear memory on a TI-83/84 with 2nd +, 7, All, 1.)
8. At the end of the exam, leave your sheet of notes with your proctor along with the exam.
9. You may choose not to scan in the appendix if you made no notes on it.
10. 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.

I attest that I spent no more than 2 hours minutes 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. I certify that:

1. The time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. The student has followed the instructions listed above.
3. The exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. 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.

Proctor's Signature: _____

Part I: Multiple Choice (5 points each)

1. A statistician is interested in modeling the response variable number of doctors in a city using the predictor number of hospitals. The first four entries for the design matrix for the model are shown below (the first four cities had 2, 5, 3, and 12 hospitals); which model does this design matrix match?

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 5 & 25 \\ 1 & 3 & 9 \\ 1 & 12 & 144 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

   (a) $y = \beta_0 + \beta_1 NumHospitals + e$

   (b) $y = \beta_1 NumHospitals + \beta_2 NumHospitals^2 + e$

   (c) $y = \beta_1 NumHospitals + \beta_2 \log(NumHospitals) + e$

   (d) $**y = \beta_0 + \beta_1 NumHospitals + \beta_2 NumHospitals^2 + e$

   (e) $y = \beta_0 + \beta_1 NumHospitals + \beta_2 \log(NumHospitals) + e$

2. Suppose the odds of an event occuring is 1/3. What is the probability that the event occurs?

   (a) **1/4

   (b) 1/3

   (c) 1/2

   (d) 2/3

   (e) 3/4

3. A researcher fits a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, where y and x_1 are quantitative variables, and x_2 is an indicator variable. Which of the following is the **best** interpretation of β_2?

   (a) β_2 is the mean change in y when x_2 increases by 1 unit.

   (b) β_2 is the mean change in y when x_1 increases by 1 unit.

   (c) β_2 is the mean of y when both x_1 and x_2 equal 0.

   (d) β_2 is the difference between the mean of y when $x_2 = 1$ and when $x_2 = 0$.

   (e) $**\beta_2$ is the difference between the mean of y when $x_2 = 1$ and when $x_2 = 0$, for a fixed value of x_1.

4. For the usual logistic regression model $\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + e$, which of the following is true?

   (a) **The relationship between log odds and x is assumed to be linear.

   (b) The relationship between probability and x is assumed to be linear.

   (c) The relationship between odds and x is assumed to be linear.

   (d) The relationship between y and x is assumed to be linear.

5. In which of the following situations are we most likely to fit a model with an error covariance matrix that has non-zero off-diagonals?

   (a) **We model a single book's price, y, across two years of weekly sales, using x = number of weeks since release.

   (b) We model the relationship between y = average book price for each of 50 sub-genres and x = number of pages, where each genre has a different number of books contributing to the average.

   (c) We model the relationship between y = book price and x = major genre, where 50 books are randomly selected from a book store's inventory.

   (d) We model the relationship between y = book price and x = number of pages, where 50 books are randomly selected from a book store's inventory.

6. A statistician wants to model the relationship between students' final grades in a class (y) and their opinion score on their major (x) for capstone courses across the university. Capstone course grades are calculated as the average of scores on 2-6 major projects throughout the semester, but no exams. What adjustment to the model might you suggest?

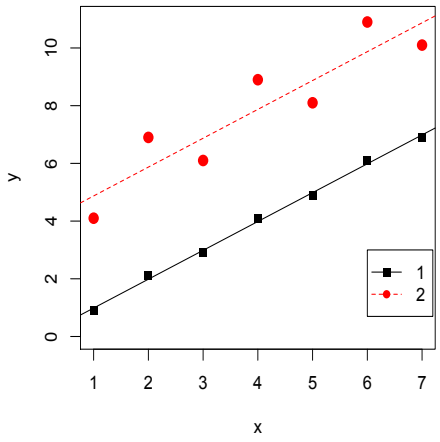
   (a) A square root transformation for the students' final grades

   (b) A log transformation for the students' final grades

   (c) **A weighted least squares model, with the number of projects as the weights

   (d) A weighted least squares model, with the inverse number of projects as the weights

7. Below are two simple linear regression models with the usual assumptions from two data sets, everything drawn to scale. Both data sets have the same number of points, 7, and result in the same slope, 1. The correlation for model 1, however, is larger than the correlation for model 2. Which of the following is also true of the two models?



   (a) Residual SS for Model 1 > Residual SS for Model 2.

   (b) **Residual SS for Model 1 < Residual SS for Model 2.

   (c) Regression SS for Model 1 > Regression SS for Model 2.

   (d) Regression SS for Model 1 < Regression SS for Model 2.

Part II: Long Answer

8. Statisticians working with a smaller film studio are interested in predicting whether or not a movie will make any profit using its budget (in millions of dollars), how much money it makes (in millions of dollars) opening weekend, and the number of theaters it is shown in.

   (a) A preliminary model fit to the data set was Model 1, below. Parameter θ is the probability of making a profit, and depends on the values of the predictor variables. For a movie with a \$50 million budget, making \$25 million opening weekend, and shown in 2000 theaters, what is the predicted probability of profit? Be sure to show your work.

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = \beta_0 + \beta_1 Budget + \beta_2 Opening + \beta_3 Theaters + e$$

$$-3.8716 - 0.1591(50) + 0.3464(25) + 0.002066(2000) = 0.9647$$
$$\hat{\theta}(\mathbf{x}) = \frac{1}{1+e^{-0.9647}}$$
$$\hat{\theta}(\mathbf{x}) = 0.724$$

   (b) Do the marginal model plots tell us our model is valid? Briefly explain why or why not.

No, the difference between our model and a nonparametric fit is too large. For example, for opening weekend around \$40 million, our model predicts a 50% larger probability of making a profit than the nonparametric smoothed model (Probability 0.6 vs. probability 0.4.).

(c) Box plots for the distributions of Budget, Opening, and Theaters are shown; what adjustments do you suggest to the above model based on the box plots? Explain.

Because the variance for number of theaters is not equal for the two groups, we may want to use a quadratic term in number of theaters.
Because of the right-skewness in budget and especially opening, I may want to use a transformation such as log to both help prevent outliers from being quite as influential in the model and because the relationship between log odds and a normal predictor is guaranteed to be at least quadratic (we may need interaction terms; see the next page).

(d) Box-Cox output is provided for the predictor variables. What specific transformations would you suggest for the three predictor variables?

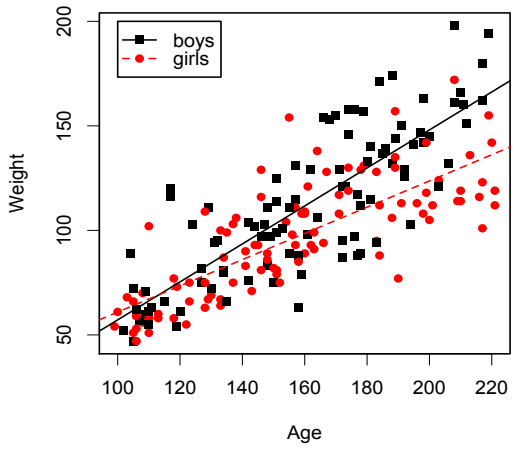
The square root, cube root, and squared transformations are suggested for budget, opening, and theaters, respectively. We won't consider log transformations for all three since the p-value for (0, 0, 0) is small, and 0 does not fall in any of the confidence intervals.

(e) If you were to choose two interaction terms to include in the model, which interactions would be most important to include? Why?

I would choose the Budget and Opening interaction and the Theaters and Opening interaction: we notice that their plots indicate very different relationships for those pairs of predictors when profit = no and when profit = yes. The relationship between Budget and Theaters is more similar when profit=no and when profit=yes. The relationship between budget and profit will thus be different for different values of opening, and the relationship between theaters and profit will also be different for different values of opening.

9. A doctor studying children's growth has data on the height in inches, weight in pounds, and age in months of 198 children ages approximately 8-18. Boys are expected to increase in weight faster than girls, since eight-year-old girls and boys are approximately the same size; the variable "Sex" takes the value 0 for boys and 1 for girls (that is, it is an indicator for girls). A model with separate slopes was thus fit (and separate intercepts, as our data comes nowhere near the intercept), with error terms independent with mean 0 and constant variance:

$$Weight_i = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Age_i \times Sex_i + e_i$$



(a) Name an advantage of keeping boys and girls in the same data set and fitting the model above rather than putting them into two separate data sets. (Name something that has nothing to do with the hypothesis test below.)

If we keep girls and boys in the same data set, we have a larger sample size to use to examine the distribution of the residuals, and thus we'll have narrower confidence intervals for slopes and more power: smaller p-values for slopes. That is, we'll essentially have a larger sample size.

A couple people also mentioned that we can test for whether $\mu_{Y|Age=x,Sex=M} = \mu_{Y|Age=x,Sex=F}$, that is, whether boys and girls have the same average weight at the same age. That's slightly different from testing whether an interaction exists, so I mostly accepted it.

(b) Test the hypothesis that boys grow faster than girls, using a significance level of 0.05. Assume the appropriate assumptions are met. Be sure to state your hypotheses, test statistic, and p-value, and conclusion in context. Output from the model is below:

```
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.69254   10.00727  -3.367 0.000917 ***
Age           0.90871    0.06106  14.882  < 2e-16 ***
Sex          31.85057   13.24269   2.405 0.017106 *
Age:Sex      -0.28122    0.08164  -3.445 0.000700 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 19.19 on 194 degrees of freedom
Multiple R-squared:  0.6683,Adjusted R-squared:  0.6631
F-statistic: 130.3 on 3 and 194 DF,  p-value: < 2.2e-16
```

Be careful: this is a one-sided hypothesis test. Since the indicator variable Sex is an indicator for girls, and boys grow faster, that means girls grow more slowly, so β_3 is negative if the alternative hypothesis is true.
$H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 < 0$
$t = -0.28122/0.08164 = -3.445$, p-value $= 0.00035$
Because my p-value is so small, I have very strong evidence that boys do indeed grow faster than girls, assuming the appropriate assumptions are met.

(c) A 95% prediction interval for someone with age=180, sex=0, and weight=133lbs is (91.8, 168.0). Interpret this interval in context.

For reference, an 8-year-old is at least 96 months old; a 15-year-old is at least 180 months old; and an 18-year-old is at least 216 months old.

My model predicts that 95% of 15-year-old boys will weigh between 91.8 and 168.0 lbs. (This particular girl does fall in the prediction interval, but the prediction interval actually only applies to future observations.)

10. Babe Ruth is argued to be the best baseball player of all time. One of his records that still stands today is OPS, the on base percentage plus slugging average, which correlates well with team run scoring (don't mix this up with the On Base Percentage from last year's exam). Interest centers in predicting OPS using PA, the number of plate appearances per year; H, the number of hits per year; and Team, the team Babe Ruth played on that year. Babe Ruth played for the Boston Red Sox the first six years of his career, the Boston Braves the last year, and the New York Yankees the remaining years. We have data from the 22 years Babe Ruth played for the major leagues.

    (a) First, do you suspect that autocorrelation may be a problem for fitting this model? Give at least two reasons from the plots why or why not.

    Yes: first, the ACF plots show significant correlation at lag one for OPS (barely), hits, and plate appearances. Second, if you plot this year vs. last year, you can notice an increasing trend for all three variables. Finally, if you plot any of the three variables across time, they tend to all be lower and then higher and then lower values, so if this year's value was higher, next year's value tends to also be high.

    (b) A bit of code for a model transformation is shown in the appendix. What is the (1,2) entry of the matrix Sigma (the estimated covariance matrix for the errors) in this code? Explain, as if to someone with no experience in statistics, what this number means.
    The (1,2) entry will simply be the estimate of the correlation, rho, -0.367. This means that the relationship between the residuals from time t and time t-1 is weakly negatively correlated: if this year's OPS for Babe Ruth is above what our model predicts, we're saying next year's is predicted to be slightly below the model's prediction.
    On the one hand, we're surprised that the correlation is negative; on the other hand, the fact that it is so close to zero makes us suspicious that it is no longer significant at all, now that we have taken team, plate appearances, and number of hits into account.

(c) Finally, the generalized least squares model with errors e_t autocorrelated AR(1) was fit. Does this model appear valid? Why or why not?

The very first observation will have higher leverage than the other points due to the nature of the transformation, so we check its residual to make sure it is not a "bad" leverage point: since its residual is indeed less than -2, it is an influential point, and we probably want to consider re-fitting the model with that point removed and notice the difference in predictions made by the model. If the difference is too large, one remedy is to fit an indicator for that observation to reduce the amount of influence it has on the model. (We might also suggest fitting a quadratic through time: it appears that Ruth gets better through the middle of his career and then declines toward the end.)

We also see non-constant variability in the residuals through the second half of the plot; this also indicates our model is not valid. (It may be caused by any number of things; all that we know is that the model is not valid.) We also note normality of the residuals is problematic; this is somewhat important since our sample size is only 22; if the goal of the model is to make any kind of inference, we should bootstrap.

(d) (2 bonus points): When I ran the generalized least squares model below, R printed a warning message, "not plotting observations with leverage one: 22" (that is, observation 22 in the data set, the last observation, was not plotted). Why did this happen? (Hint: No, I don't think it has anything to do with the fact that this is autocorrelated data.)

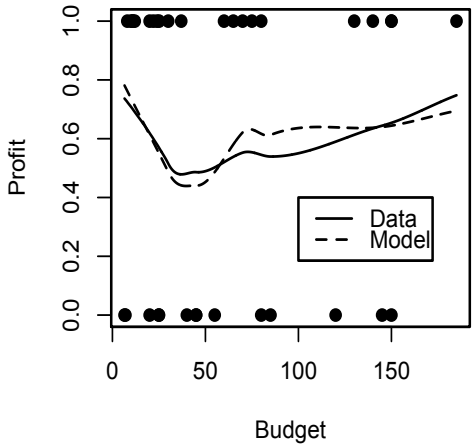
$$OPS_t = \beta_0 + \beta_1 PA_t + \beta_2 H_t + \beta_3 iBSN_t + \beta_4 iNYY_t + e_t$$

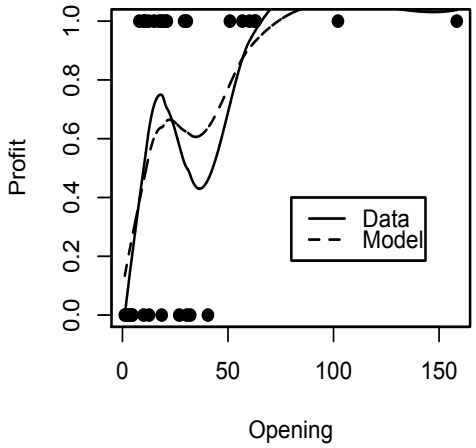
(where $iBSN$ is an indicator variable for playing for the Boston Braves, $iNYY$ is an indicator for playing for the New York Yankees, and the error term was autocorrelated AR(1): $e_t = \rho e_{t-1} + \nu_t$, where the ν_t were independent and identically normally distributed.

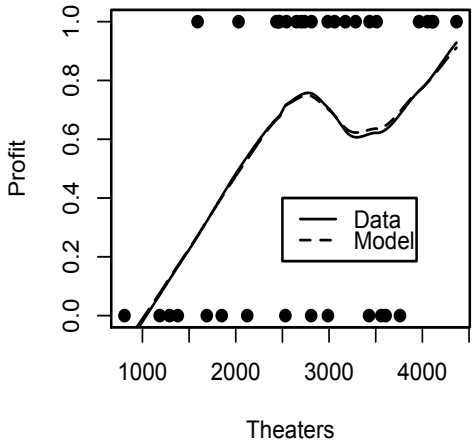
The problem is that we only have one year of data when Ruth was playing for the Boston Braves. We have no measure of variability of the sample mean OPS when Ruth was playing for the Boston Braves since he was only there one year.

Appendix

Movie Profits
Model 1: Marginal Model Plots



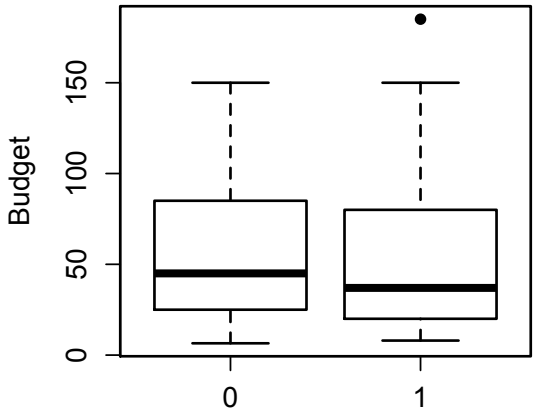


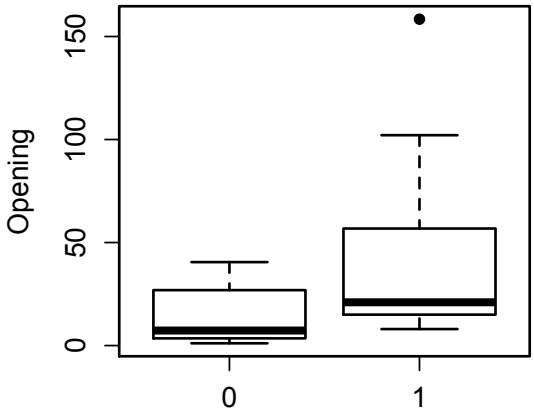


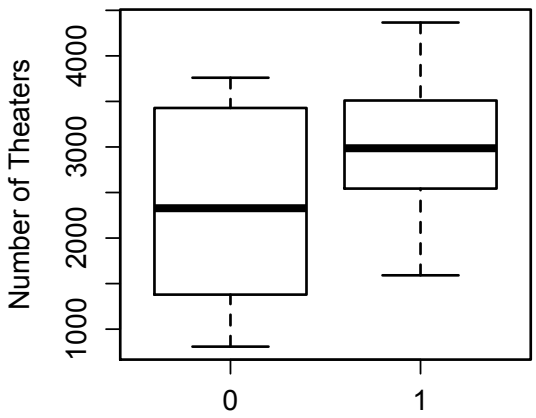
Movie Profits
Model 1 Output

```
Call:
glm(formula = Profit ~ Budget + Opening + Theaters, family = binomial(),
    data = movies)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.871589   3.351796  -1.155   0.2481
Budget      -0.159113   0.067706  -2.350   0.0188 *
Opening      0.346399   0.144127   2.403   0.0162 *
Theaters     0.002066   0.001734   1.192   0.2334
```

Boxplots

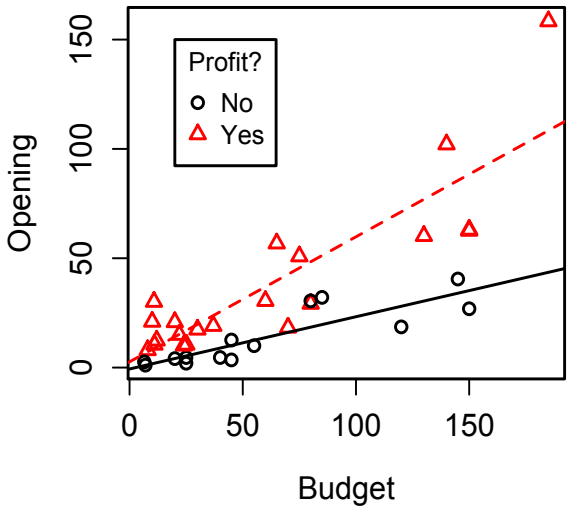


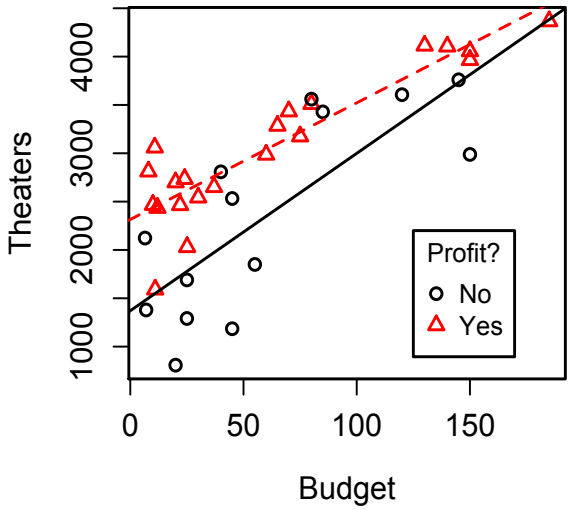


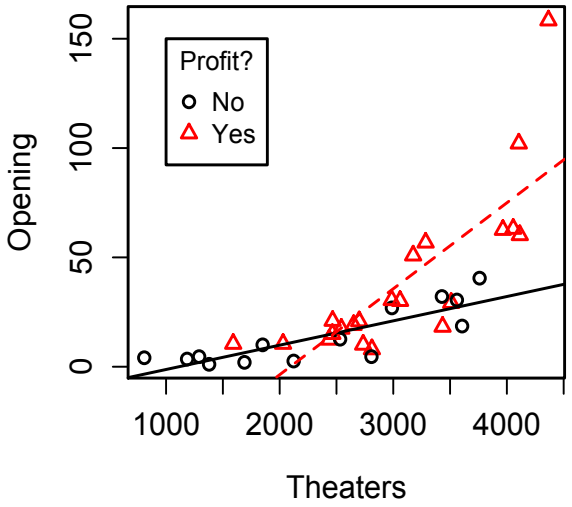


Movie Profits
Box-Cox Output

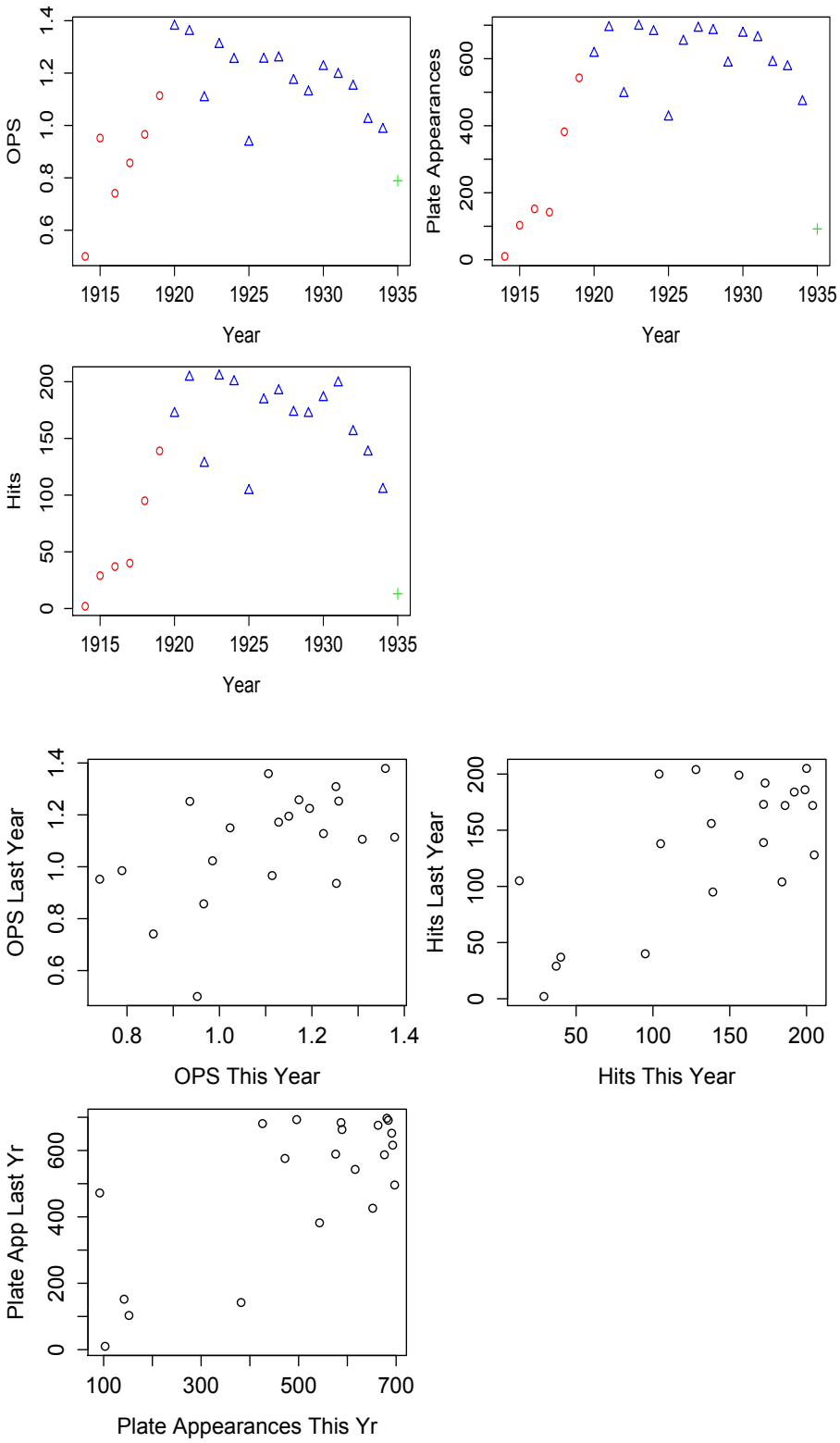
```
bcPower Transformations to Multinormality
         Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Budget      0.4320   0.1694           0.0999           0.7640
Opening     0.2476   0.0943           0.0628           0.4324
Theaters    1.9811   0.3765           1.2432           2.7191

Likelihood ratio tests about transformation parameters
                                      LRT df         pval
LR test, lambda = (0 0 0)      33.6043702  3 2.401223e-07
LR test, lambda = (1 1 1)      74.9983507  3 3.330669e-16
LR test, lambda = (0.5 0.33 2)  0.9645658  3 8.098251e-01
```



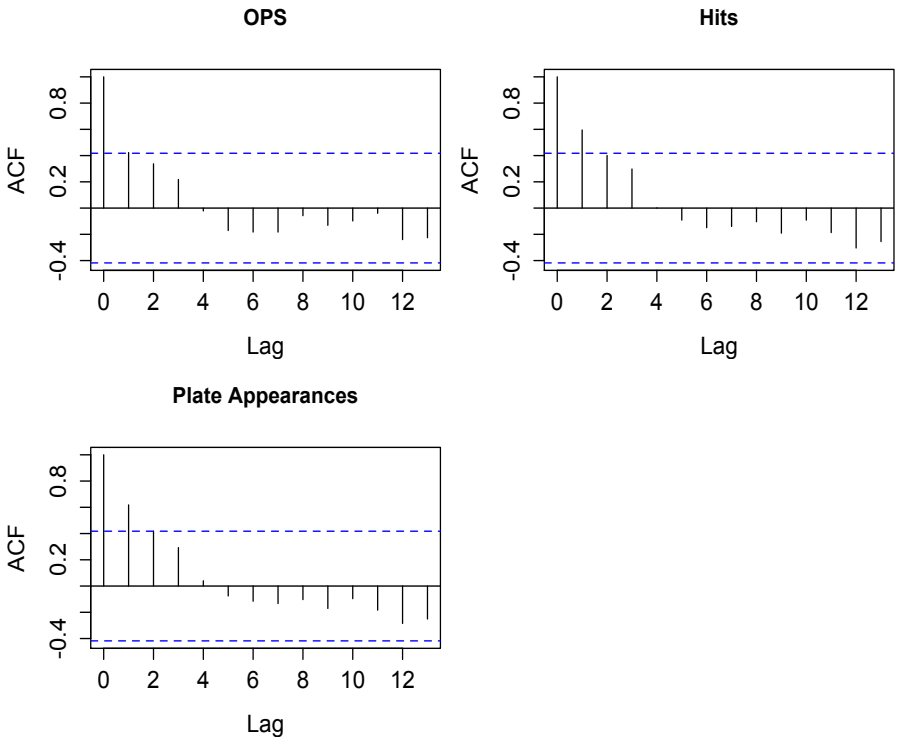




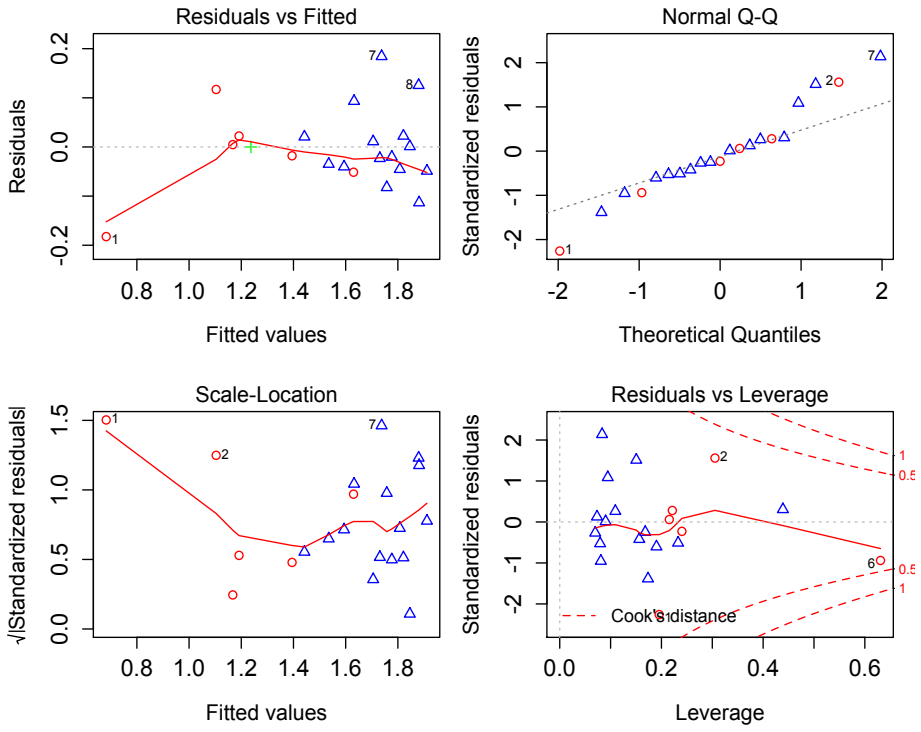
Babe Ruth


OPS
0.6
0.8
1.0
1.2
1.4
1915
1920
1925
1930
1935
Year
Plate Appearances
0
200
400
600
Hits
0
50
100
150
200
OPS Last Year
OPS This Year
Hits Last Year
Hits This Year
Plate App Last Yr
100
300
500
700
Plate Appearances This Yr

Babe Ruth



Generalized Least Squares Model with errors AR(1):



Babe Ruth
Model transformation code:

```
m2g <- gls(OPS ~ PA + Tm + H, correlation=corAR1(form = ~Year), method="ML")

rho <- -0.367068
x <- model.matrix(m1)
iden <- diag(n)
Sigma <- rho^abs(row(iden)-col(iden))
sm <- chol(Sigma)
smi <- solve(t(sm))
xstar <- smi %*% x
ystar <- smi %*% OPS
m1tls <- lm(ystar ~ xstar - 1)
```