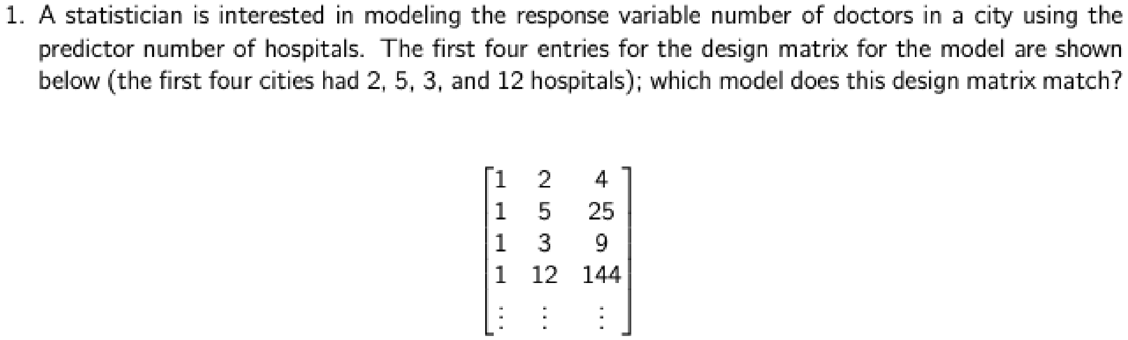
Part I: Multiple Choice



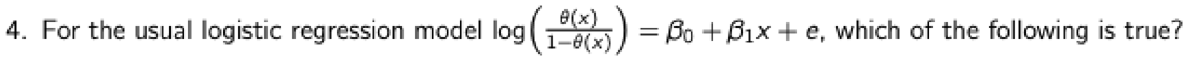
1. A statistician is interested in modeling the response variable number of doctors in a city using the predictor number of hospitals. The first four entries for the design matrix for the model are shown below (the first four cities had 2, 5, 3, and 12 hospitals); which model does this design matrix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

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

(a) y = β0 + β1NumHospi tals + e
(b) y = β1NumHospi tals + β2NumHospi tals2 + e
(c) y = β1NumHospi tals + β2 log(NumHospitals) + e
(d) y = β0 + β1NumHospi tals + β2NumHospi tals2 + e
(e) y = β0 + β1NumHospi tals + β2 log(NumHospitals) + e

2. Suppose the odds of an event occurring 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 = β0 + β1x1 + β2x2 + e, where y and x1 are quantitative variables, and x2 is an indicator variable. Which of the following is the best interpretation of β2?
(a) β2 is the mean change in y when x2 increases by 1 unit.
(b) β2 is the mean change in y when x1 increases by 1 unit.
(c) β2 is the mean of y when both x1 and x2 equal 0.
(d) β2 is the difference between the mean of y when x2 = 1 and when x2 = 0.
(e) β2 is the difference between the mean of y when x2 = 1 and when x2 = 0, for a fixed value of x1.



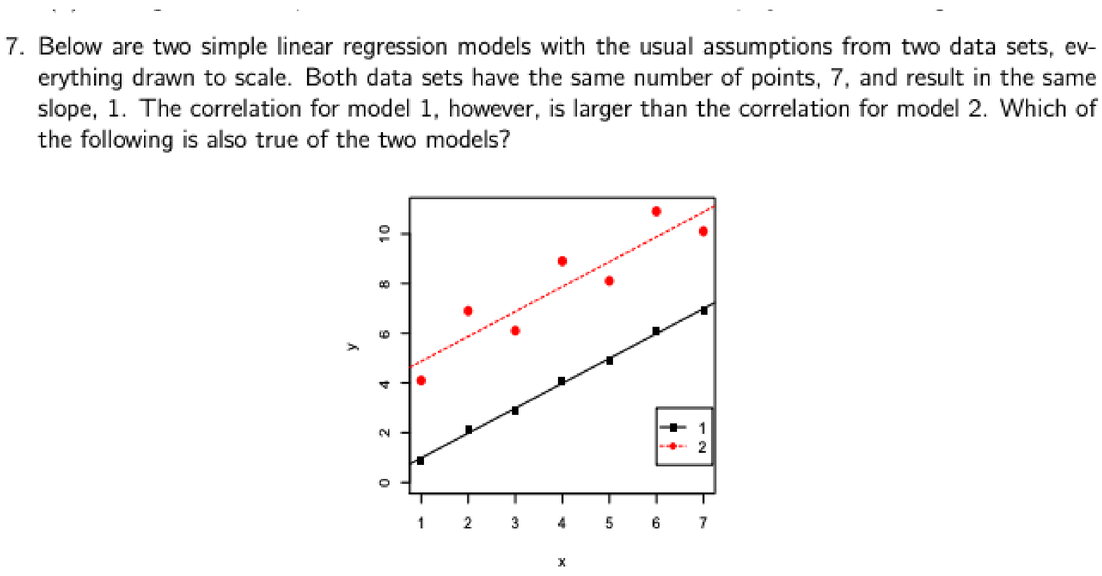
4. For the usual logistic regression model $\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + e$, which of the following is true?

(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.

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

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

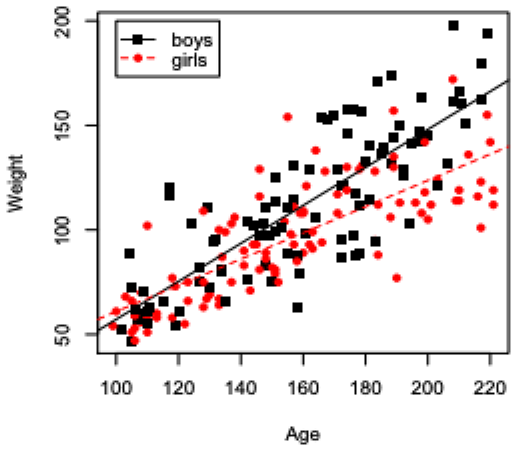
(c) Box plots for the distributions of Budget, Opening, and Theaters are shown; what adjustments do you suggest to the above model based on the box plots? Explain.

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

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

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.)

(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
```

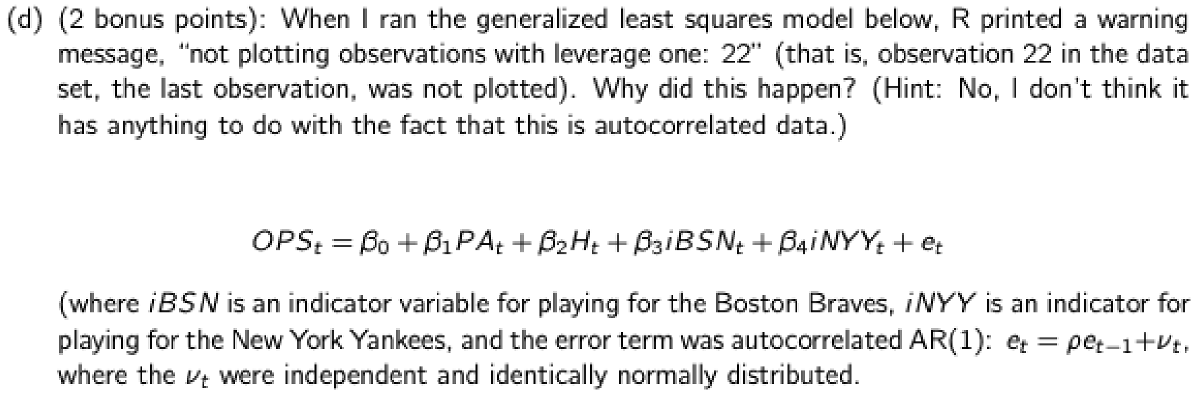
(c) A 95% prediction interval for someone with age=180, sex=0, and weight=133lbs is (91.8, 168.0). Interpret this interval in context. For reference, an 8-year-old is at least 96 months old; a 15-year-old is at least 180 months old; and an 18-year-old is at least 216 months old.

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.

(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.

(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?



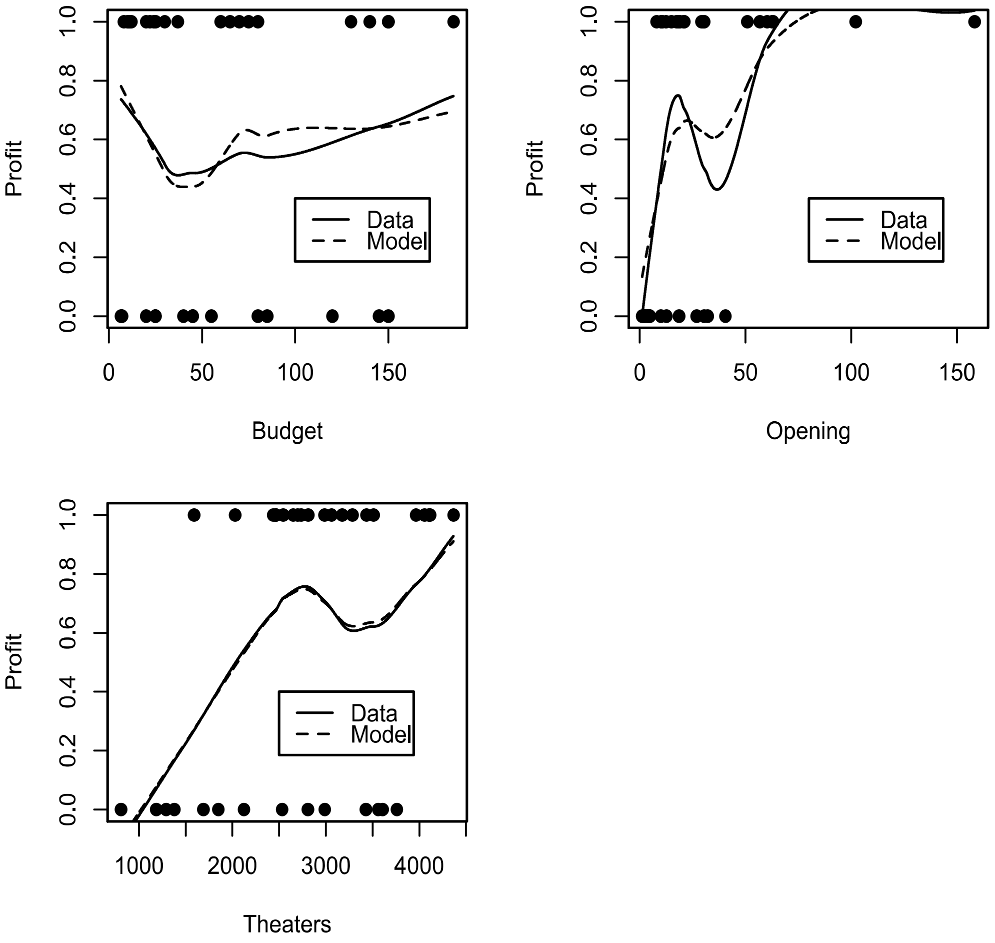
(d) (2 bonus points): When I ran the generalized least squares model below, R printed a warning message, "not plotting observations with leverage one: 22" (that is, observation 22 in the data set, the last observation, was not plotted). Why did this happen? (Hint: No, I don't think it has anything to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

Appendix

Movie Profits
Model 1: Marginal Model Plots

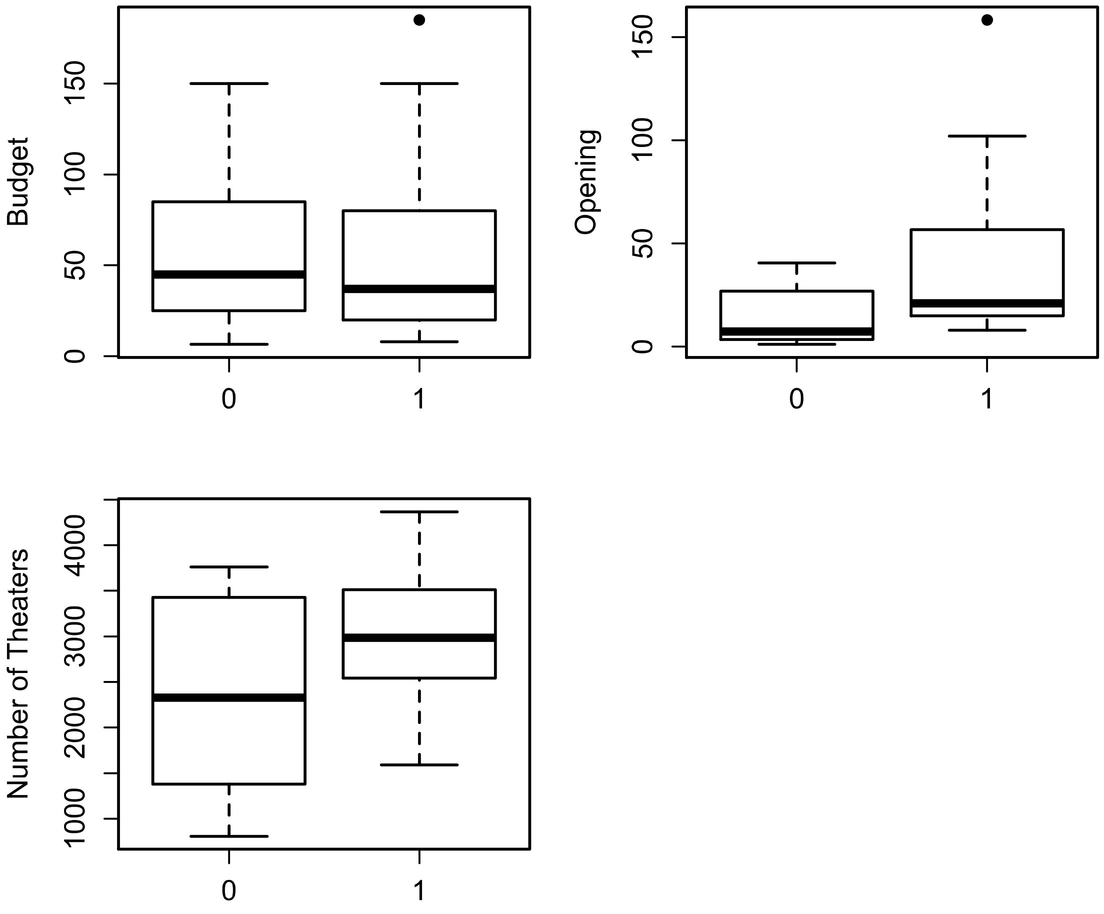


Movie Profits
Model 1 Output

```
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
```

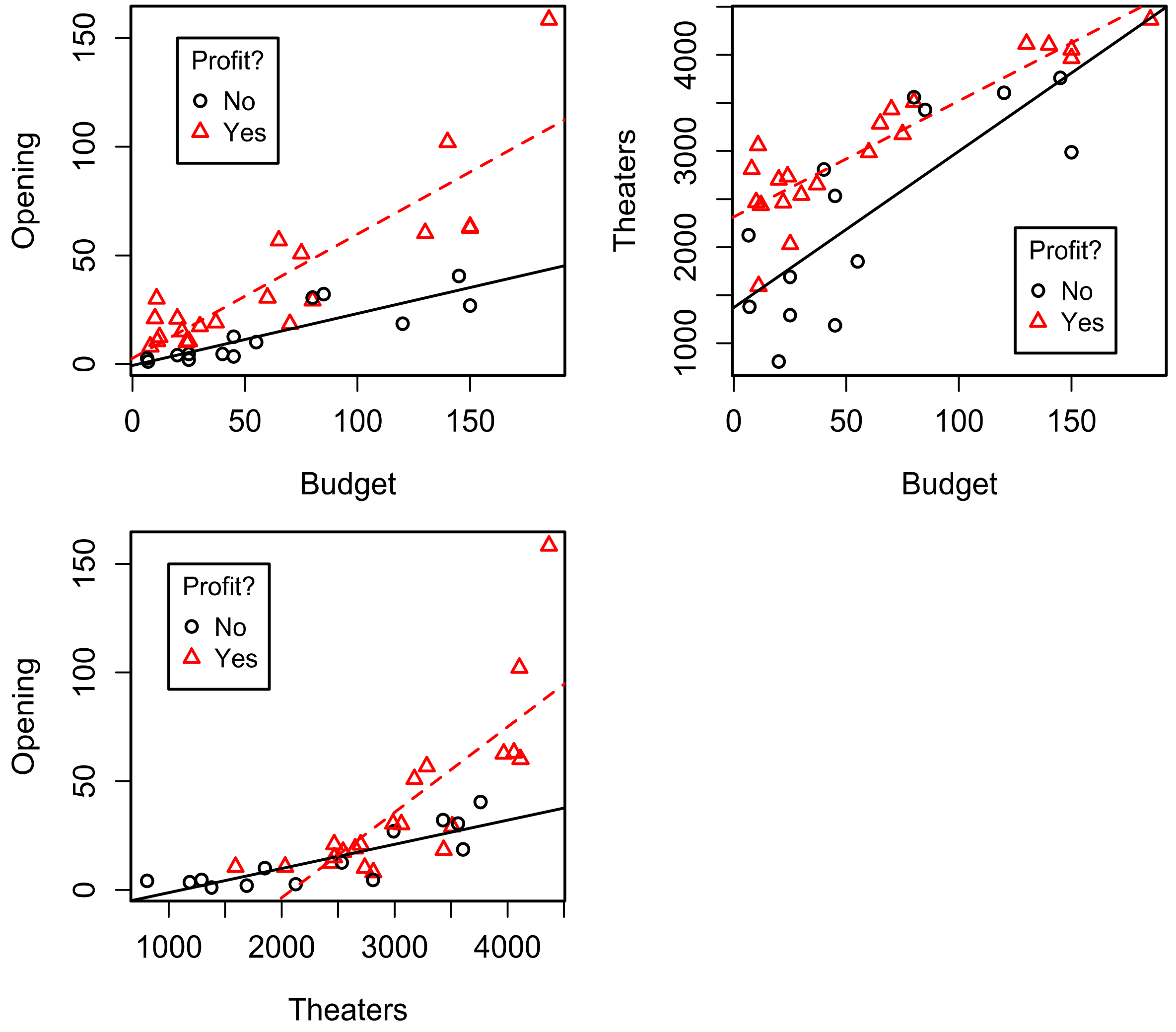
Boxplots



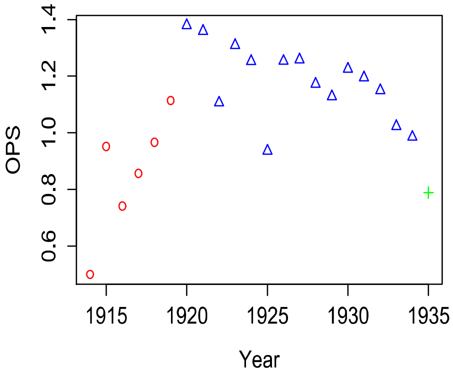
Movie Profits
Box-Cox Output

```
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

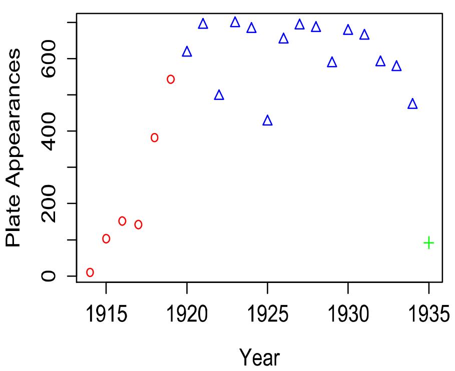
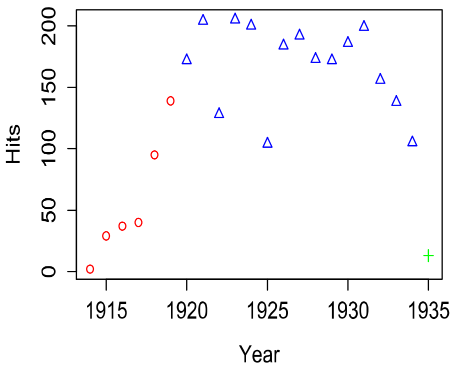
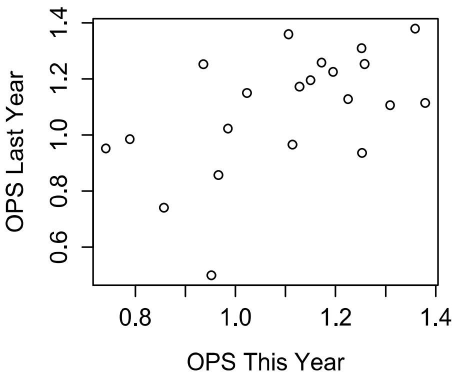
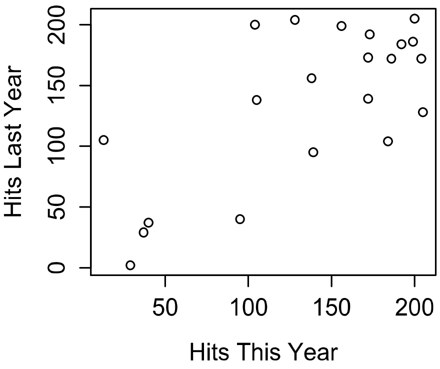


Plate Appearances
0
200
400
600
1915
1920
1925
1930
1935
Year


Hits
0
50
100
150
200
1915
1920
1925
1930
1935
Year


OPS Last Year
0.6
0.8
1.0
1.2
1.4
0.8
1.0
1.2
1.4
OPS This Year


Hits Last Year
0
50
100
150
200
50
100
150
200
Hits This Year

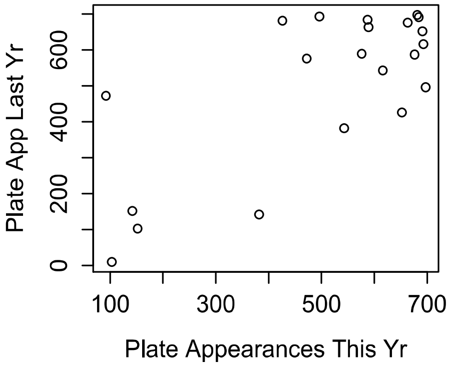
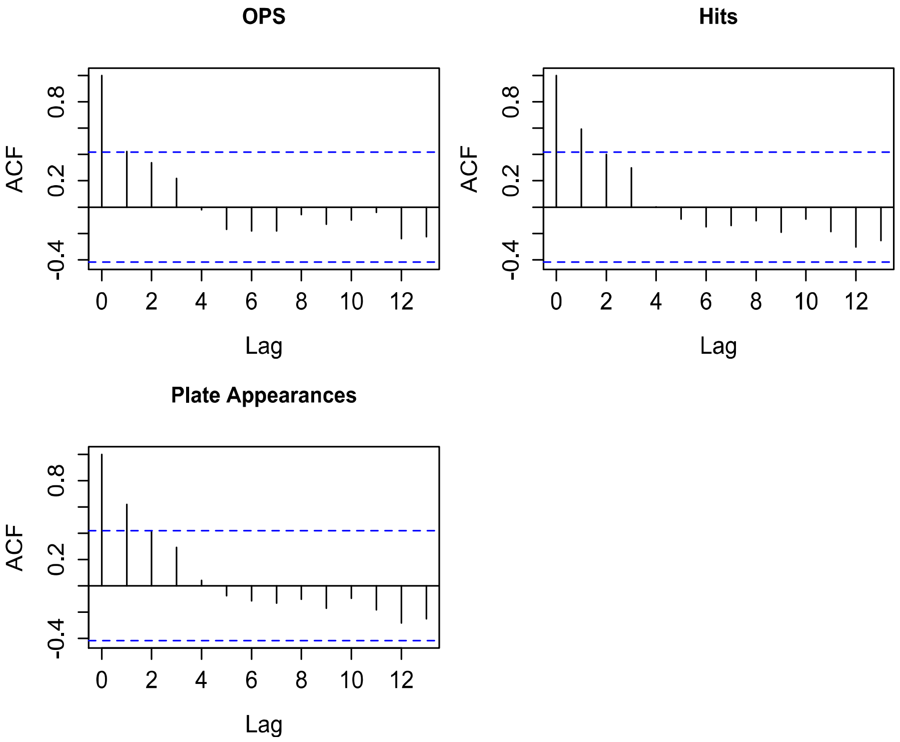
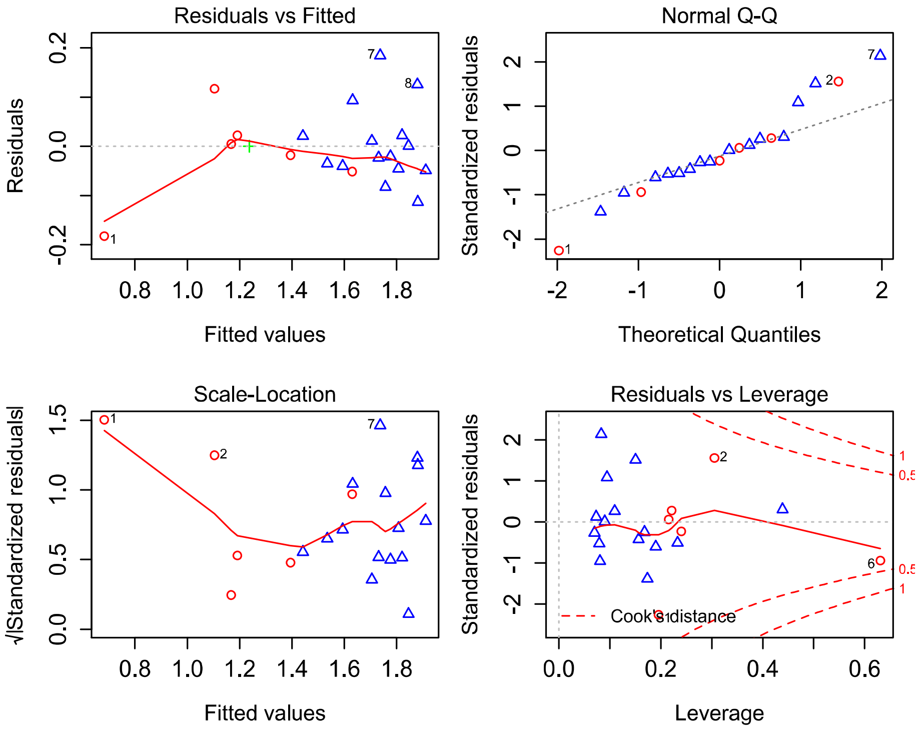


Plate App Last Yr
0
200
400
600
100
300
500
700
Plate Appearances This Yr

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)
```