

## STATISTICS 415, Homework Solution for Logistic Regression

1. Below are data on the relationship between the proportion of male turtles and incubation temperature for turtle eggs from New Mexico. The turtles are the same species as those from Illinois. The New Mexico data are given below.

| Temp | male | female | % male |
|------|------|--------|--------|
| 27.2 | 0    | 10     | 0%     |
| 28.3 | 8    | 4      | 67%    |
| 29.9 | 8    | 2      | 80%    |

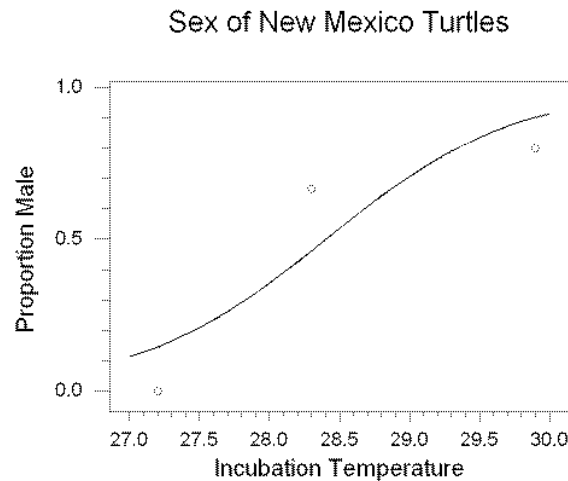
- (a) Use logistic regression to analyze these data. Turn in the summary of the logistic regression fit. Give the formula for the estimated curve and test to see if temperature has a significant relationship with the sex of New Mexico turtles.

**Predicted log odds =  $-41.742 + 1.4696 \cdot \text{Temperature}$**

**Change in Deviance = 12.9983, P-value = 0.0003, Temperature is a statistically significant predictor of the proportion of male turtles. Or the parameter estimate ChiSquare for Temp is 8.22 with a P-value = 0.0041, this leads to the same conclusion.**

**Residual Deviance = 6.0787, P-value = 0.0137, there is significant lack of fit.**

- (b) Turn in a plot of the data with the logistic regression curve superimposed.



- (c) What is the temperature at which you would get a 50:50 split of males to females? How does this compare to the temperature for a 50:50 split for Illinois turtles?

**For New Mexico turtles, the temperature for a 50:50 split is  $28.404^\circ \text{C}$ . For the Illinois turtles, the temperature for a 50:50 split was  $27.733^\circ \text{C}$ .**

- (d) Estimate the probability that a male turtle hatches from a New Mexico egg incubated at  $27^\circ \text{C}$ . Construct a 95% confidence interval for this probability.

**The estimated probability of male is 0.1127 with a 95% confidence interval from 0.0256 to 0.3909.**

- (e) Estimate the probability that a male turtle hatches from a New Mexico egg incubated at  $30^\circ \text{C}$ . Construct a 95% confidence interval for this probability.

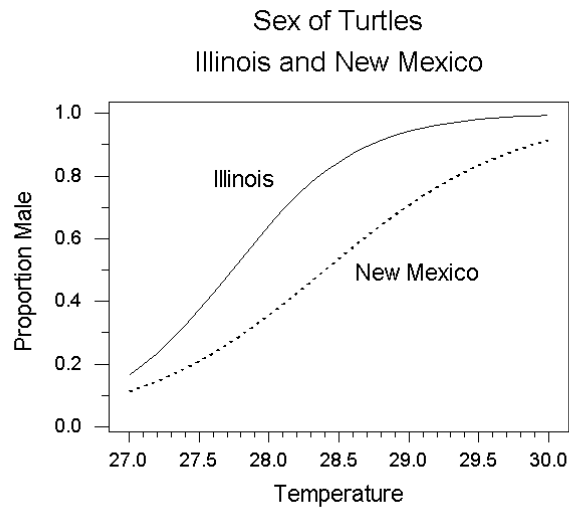
**The estimated probability of male is 0.9126 with a 95% confidence interval from 0.6094 to 0.9859.**

- (f) Compare the estimates and confidence intervals in (d) and (e) to the estimates and confidence intervals for the same temperatures for Illinois turtles.

| Temp | Illinois    |                  | New Mexico  |                  |
|------|-------------|------------------|-------------|------------------|
|      | Probability | 95% CI           | Probability | 95% CI           |
| 27.0 | 0.1651      | 0.0762 to 0.3216 | 0.1127      | 0.0256 to 0.3809 |
| 30.0 | 0.9934      | 0.9603 to 0.9989 | 0.9126      | 0.6094 to 0.9859 |

The estimated probabilities of a male turtle for the Illinois turtles are higher than those for the New Mexico turtles. The confidence intervals are much wider for the New Mexico turtles.

- (g) To visually compare the the fits for New Mexico turtles and the Illinois turtles, construct a plot that has both logistic regression curves on it. See the S-plus code for the turtle data on the course website for some hints.



2. A study was conducted to see the effect of coupons on purchasing habits of potential customers. In the study, 1000 homes were selected and a coupon and advertising material for a particular product was sent to each home. The advertising material was the same but the amount of the discount on the coupon varied from 5% to 30%. The number of coupons redeemed was counted. Below are the data.

| Price Reduction | Number of Coupons | Number Redeemed | Proportion Redeemed |
|-----------------|-------------------|-----------------|---------------------|
| $X_i$           | $n_i$             | $Y_i$           | $p_i$               |
| 5               | 200               | 32              | 0.160               |
| 10              | 200               | 51              | 0.255               |
| 15              | 200               | 70              | 0.350               |
| 20              | 200               | 103             | 0.515               |
| 30              | 200               | 148             | 0.740               |

- (a) Fit a simple linear regression to the observed proportions. Use this regression to estimate the proportion redeemed. Is there a significant linear relationship between proportion redeemed and price reduction? According to this regression at what price reduction will you get a 25% redemption rate?

**Simple linear regression of observed proportions redeemed on the price reduction. The simple linear regression equation is**

$$\hat{\pi}_i = 0.0243 + 0.0237\text{PriceReduce}$$

The slope coefficient for Price Reduction has a t-value of 20.26 with an associated P-value of 0.0003. Such a small P-value indicates that such a large slope coefficient could not have happened by chance alone and therefore there is a significant linear relationship between the proportion redeemed and price reduction. A note of caution, we are violating the assumptions for the t-test. However, the result is so extreme the violation is comparatively unimportant. If we want a 25% redemption rate we need to offer a 9.5% price reduction.

- (b) Fit a simple linear regression of the logit transformed proportions on the price reduction. Is there a significant linear relationship between the logit and the price reduction? Use this regression to estimate the proportion redeemed for each price reduction. According to this regression at what price reduction will you get a 25% redemption rate?

**The simple linear regression of the logit of the proportion redeemed on the price reduction is:**

$$\hat{\pi}_i' = -2.18602 + 0.10858\text{PriceReduce}$$

See the table below for the estimates of the proportions redeemed.

| Price Reduction | Number of Coupons | Number Redeemed | Proportion Redeemed | Predicted Proportions |
|-----------------|-------------------|-----------------|---------------------|-----------------------|
| $X_i$           | $n_i$             | $Y_i$           | $p_i$               | $\hat{p}_i$           |
| 5               | 200               | 32              | 0.160               | 0.162                 |
| 10              | 200               | 51              | 0.255               | 0.250                 |
| 15              | 200               | 70              | 0.350               | 0.364                 |
| 20              | 200               | 103             | 0.515               | 0.496                 |
| 30              | 200               | 148             | 0.740               | 0.745                 |

The t-value is 34.47 with a P-value of essentially zero. Such a small P-value indicates that such a large slope coefficient could not have happened by chance alone and therefore there is a significant linear relationship between the logit and price reduction. A note of caution, we are violating the assumptions for the t-test. However, the result is so extreme the violation is comparatively unimportant. If we want a 25% redemption rate we need to offer a 10% price reduction.

- (c) Fit a logistic regression of the proportion redeemed on the price reduction. Comment on the adequacy of the fit of the logistic model. Support your answer statistically. Is price reduction a significant predictor in this logistic regression model? Support your answer statistically. Use the logistic regression to estimate the proportion redeemed for each price reduction. According to this regression at what price reduction will you get a 25% redemption rate?

The logistic regression of the of the proportion redeemed on the price reduction is:

$$\hat{\pi}_i' = -2.18551 + 0.10872 \text{PriceReduce}$$

See the table below for the estimates of the proportions redeemed.

| Price Reduction | Number of Coupons | Number Redeemed | Proportion Redeemed | Predicted Proportions |
|-----------------|-------------------|-----------------|---------------------|-----------------------|
| $X_i$           | $n_i$             | $Y_i$           | $p_i$               | $\hat{p}_i$           |
| 5               | 200               | 32              | 0.160               | 0.162                 |
| 10              | 200               | 51              | 0.255               | 0.250                 |
| 15              | 200               | 70              | 0.350               | 0.365                 |
| 20              | 200               | 103             | 0.515               | 0.497                 |
| 30              | 200               | 148             | 0.740               | 0.746                 |

The residual deviance is 0.5102 on 3 df. Such a small residual deviance indicates that the logistic regression model is a very good fit. The change in deviance is 180.54 on 1 df. The associated P-value would be virtually zero. This indicates that the predictor variable, Price Reduction is statistically significant.

If we want a 25% redemption rate we need to offer a 10% price reduction.

- (d) Compare the three regression equations and price reductions to get a 25% redemption rate.

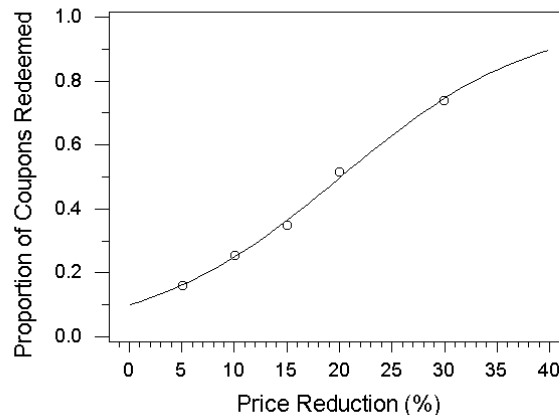
Comparison of equations, predicted proportions and predicted price reduction necessary for a 25% redemption.

The two equations using the logit transform (logitslr and logistic) are virtually the same in terms of the estimates. The standard errors are somewhat different. The predicted proportions are very similar, especially for the logitslr and the logistic regression. The slr model indicated a 9.5% price reduction while the other two models say to offer a 10% price reduction to get a 25% redemption rate.

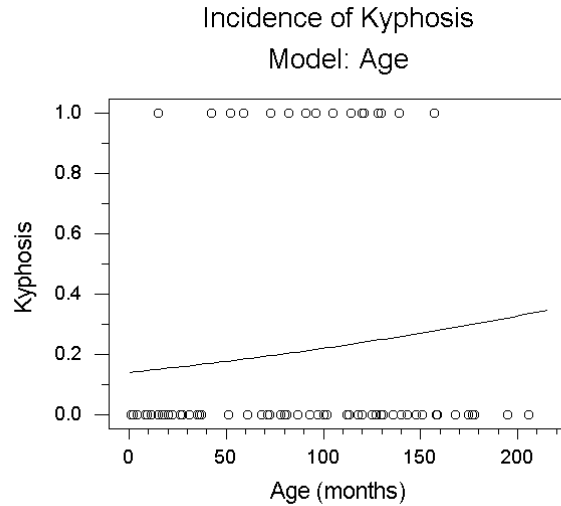
- (e) Create plots that show the data and each of the fits.

The plot below is the plot of the logistic regression fit.

Coupons and Purchasing Habits



3. Kyphosis is a spinal deformity found in young children who have corrective spinal surgery. The incidence of spinal deformities following corrective spinal surgery ( $kyp=1$  if deformity is present,  $kyp=0$  if there is no deformity present) is thought to be related to the Age (in months) at the time of surgery, Start (the starting vertebra for the surgery) and Num (the number of vertebrae involved in the surgery).
- (a) Plot the binary response for the incidence of Kyphosis versus the age of the child. Fit a simple logistic regression of incidence of Kyphosis on Age. Examine the fit and significance of Age.



The logistic regression of kyphosis on Age is

$$\hat{\pi}'_i = -1.8091 + 0.00544 * \text{Age}$$

The residual deviance is large but there a lot of degrees of freedom so there is no indication of significant lack of fit. However the change in deviance is only 1.302 on 1 df which has a P-value of 0.25. So Age is not a significant predictor of Kyphosis. Looking at the graph, although there is a slight increasing trend in Kyphosis with Age, what is more apparent is that Kyphosis is not as common for very young or older children.

- (b) Fit a quadratic logistic regression model in Age. You will need to create a new variable AgeSq = Age\*Age. Examine the fit and significance of Age and AgeSq.

The quadratic logistic regression of Kyphosis on Age is

$$\hat{\pi}'_i = -3.7703 + 0.07003 * \text{Age} - 0.00365 * \text{AgeSq}$$

The residual deviance is large but there are a lot of degrees of freedom which indicates no significant lack of fit. However the change in deviance from the model with only Age in it is 9.19 on 1 df, this is highly significant. Note also that the test statistic for AgeSq is  $\chi^2=6.10$  (JMP) and  $t=-2.47$  (S-plus) which also indicates the significance of the addition of AgeSq. What is somewhat unusual is that the test for Age (indicating the significance of Age once AgeSq is in the model) is  $\chi^2=6.74$  (JMP) and  $t=2.60$  (S-Plus) which is also significant. So together both Age and AgeSq are significant predictors of Kyphosis.

- (c) Repeat part (a) with the explanatory variable Number.

The logistic regression of Kyphosis on the Number of vertebrae is

$$\hat{\pi}'_i = -3.6510 + 0.5317 * \text{Num}$$

Both the change in deviance of 9.88 on 1 df and the test statistic for number  $\chi^2=8.26$  (JMP) and  $t=2.88$  (S-Plus), indicate that Number is a significant predictor of Kyphosis.

- (d) Fit the following logistic regression models to examine the effects of Age and Num. For each model, comment on the adequacy of the fit and the significance of each of the terms.

- Regress on Age and Num

$$\hat{\pi}'_i = -4.4185 + 0.0070885 * \text{Age} + 0.5644841 * \text{Num}$$

The model is statistically significant (change in deviance = 11.61 on 2 df, P-value=0.0030). Only Num (give Age is in the model) is statistically significant ( $\chi^2=8.09$  P-value=0.0045). Age is not statistically significant ( $\chi^2=1.67$ , P-value=0.1969).

- Regress on Age, Num, AgeSq

$$\hat{\pi}'_i = -6.5269 + 0.0758597 * \text{Age} + 0.5440377 * \text{Num} - 0.00038429\text{AgeSq}$$

The model is statistically significant (change in deviance = 19.37 on 3 df, P-value=0.0002). All terms in the model are statistically significant.

| Term  | Chi Square | P-value | Significant?        |
|-------|------------|---------|---------------------|
| Age   | 5.85       | 0.0156  | Yes, P-value < 0.05 |
| Num   | 6.98       | 0.0082  | Yes, P-value < 0.05 |
| AgeSq | 5.11       | 0.0238  | Yes, P-value < 0.05 |

- Regress on Age, Num, AgeSq, NumSq

$$\hat{\pi}'_i = -8.0943 + 0.0756414 * \text{Age} + 1.1713864 * \text{Num} - 0.00037599\text{AgeSq} - 0.05866968 * \text{NumSq}$$

The model is statistically significant (change in deviance = 19.89 on 4 df, P-value=0.0005). Only Age and AgeSq are statistically significant in this model, given the other variables in the model.

| Term  | Chi Square | P-value | Significant?        |
|-------|------------|---------|---------------------|
| Age   | 5.81       | 0.0160  | Yes, P-value < 0.05 |
| Num   | 1.68       | 0.1943  | No, P-value > 0.05  |
| AgeSq | 4.97       | 0.0258  | Yes, P-value < 0.05 |
| NumSq | 0.54       | 0.4634  | No, P-value > 0.05  |

- Regress on Age, Num, AgeSq, NumSq, Age\*Num

$$\hat{\pi}'_i = -17.1010 + 0.194447 * \text{Age} + 2.2015456 * \text{Num} - 0.00062148\text{AgeSq} - 0.00378401 * \text{NumSq} - 0.01383399 * \text{Age} * \text{Num}$$

The model is statistically significant (change in deviance = 24.56 on 5 df, P-value=0.0002). Only Age and AgeSq are statistically significant in this model, given the other variables in the model.

| Term    | Chi Square | P-value | Significant?        |
|---------|------------|---------|---------------------|
| Age     | 4.57       | 0.0326  | Yes, P-value < 0.05 |
| Num     | 3.55       | 0.0594  | No, P-value > 0.05  |
| AgeSq   | 4.79       | 0.0268  | Yes, P-value < 0.05 |
| NumSq   | 0.00       | 0.9625  | No, P-value > 0.05  |
| Age*Num | 3.14       | 0.0763  | No, P-value > 0.05  |

- (e) Give a final model that includes only those explanatory that you think are important enough to include in the model. What can you conclude about the relationships of Age, Num, and Start on the incidence of spinal deformities known as Kyphosis?

Using a backward elimination procedure you can eliminate Age\*Start, NumSq, Num\*Start and Start.

$$\hat{\pi}'_i = -20.42059 + 0.26982 * \text{Age} + 2.77622 * \text{Num} - 0.00081076\text{AgeSq} - 0.020305\text{Age} * \text{Num} - 0.0163363 * \text{StarSq}$$

The model is statistically significant (change in deviance = 38.91 on 5 df, P-value<0.0001). All variables in the model are statistically significant.

| Term    | Chi Square | P-value | Significant?        |
|---------|------------|---------|---------------------|
| Age     | 5.79       | 0.0161  | Yes, P-value < 0.05 |
| Num     | 5.49       | 0.0191  | Yes, P-value < 0.05 |
| AgeSq   | 5.05       | 0.0268  | Yes, P-value < 0.05 |
| AgeNum  | 4.92       | 0.0266  | Yes, P-value < 0.05 |
| StartSq | 10.42      | 0.0012  | Yes, P-value < 0.05 |