

## STATISTICS 415, Homework 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.
  - (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?
  - (d) Estimate the probability that a male turtle hatches from a New Mexico egg incubated at 27° C. Construct a 95% confidence interval for this probability.
  - (e) Estimate the probability that a male turtle hatches from a New Mexico egg incubated at 30° C. Construct a 95% confidence interval for this probability.
  - (f) Compare the estimates and confidence intervals in (d) and (e) to the estimates and confidence intervals for the same temperatures for Illinois turtles.
  - (g) To visually compare 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?
- (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?
- (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?

- (d) Compare the three regression equations and price reductions to get a 25% redemption rate.
- (e) Create plots that show the data and each of the fits.
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).

Age	Start	Num	Kyp	Age	Start	Num	Kyp	Age	Start	Num	Kyp
71	5	3	0	100	14	3	0	140	15	4	0
158	14	3	0	4	16	3	0	72	15	5	0
128	5	4	1	151	16	2	0	2	13	3	0
2	1	5	0	31	11	3	0	120	8	5	1
1	15	4	0	125	11	2	0	51	9	7	0
1	16	2	0	130	13	5	0	102	13	3	0
61	17	2	0	112	16	3	0	130	1	4	1
37	16	3	0	140	11	5	0	114	8	7	1
113	16	2	0	93	16	3	0	81	1	4	0
59	12	6	1	1	9	3	0	118	16	3	0
82	14	5	1	52	6	5	1	118	16	4	0
148	16	3	0	20	9	6	0	17	10	4	0
18	2	5	0	91	12	5	1	195	17	2	0
1	12	4	0	73	1	5	1	159	13	4	0
168	18	3	0	35	13	3	0	18	11	4	0
1	16	3	0	143	3	9	0	15	16	5	0
78	15	6	0	61	1	4	0	158	14	5	0
175	13	5	0	97	16	3	0	127	12	4	0
80	16	5	0	139	10	3	1	87	16	4	0
27	9	4	0	136	15	4	0	206	10	4	0
22	16	2	0	131	13	5	0	11	15	3	0
105	5	6	1	121	3	3	1	178	15	4	0
96	12	3	1	177	14	2	0	157	13	3	1
131	3	2	0	68	10	5	0	26	13	7	0
15	2	7	1	9	17	2	0	120	13	2	0
9	13	5	0	139	6	10	1	42	6	7	1
8	6	3	0	2	17	2	0	36	13	4	0

- (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.
- (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.
- (c) Repeat part (a) with the explanatory variable Number.
- (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
  - Regress on Age, Num, AgeSq
  - Regress on Age, Num, AgeSq, NumSq
  - Regress on Age, Num, AgeSq, NumSq, Age\*Num
- (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?