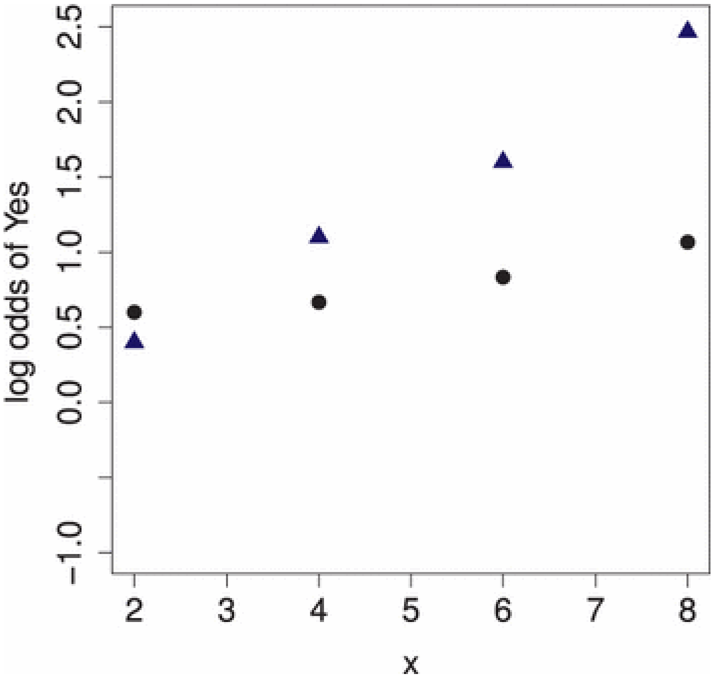
Assignment Name: Homework 12

Student Name: Mo Pei

10.4 ***Empirical logits again.*** Here is a plot of empirical logits for a dataset with two predictors: *x*, a continuous variable, and *Group*, a categorical variable with two levels, 1 and 2. The circles are for*Group* = 1 and the triangles are for *Group* = 2. What model is suggested by this plot?



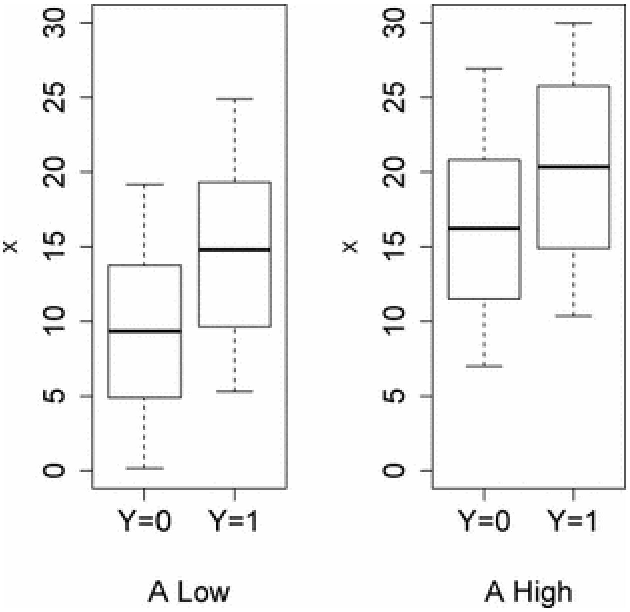
Because the two lines are not parallel and so it involves interaction term.

Log(PI/(1-PI)) = B0 + B1*Group1 + B2*Group2 + B3*Group1*Group2

PI is the probability of yes

B1 and B2 are indicators variables for group. Group1*Group2 are interaction coefficient.

10.6 ***Model building.*** Here are parallel boxplots. Suppose we want to model *Y* as depending on the levels of *A* and on the continuous variable *X*. What logistic model is suggested by this plot?



Log(pi/(1-pi)) = B0 + B1*levelofA+B2*X

B1 is the indicator variable of level A, B2 is coefficient of value of x

10.8 ***CAFE.*** Consider the CAFE data presented in this chapter, where we considered the model of logit(*Vote*) depending on *logContr* and on party (through the indicator variable *Dem*). Now we want to ask whether the slope between logit(*Vote*) and *logContr* is the same for the two parties.

a. Fit the model in which logit(*Vote*) depends on *logContr* and on party, allowing for different coefficients of *logContr* for Democrats (*Dem*) and for Republicans. Use a Wald z-test to check for a difference. What is the p-value?

H0: B3=0

H1: B3 is not zero

Since Wald statistics is 0.516 and p value is 0.472 > 0.05 we cannot reject null hypothesis. This indicates that there is no interaction between LogContr and Party.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LogContr	3.002	1.357	4.891	1	.027	20.116
	Dem	2.544	5.974	.181	1	.670	12.726
	LogContr_Dem	-1.088	1.515	.516	1	.472	.337
	Constant	-10.164	5.401	3.541	1	.060	.000

a. Variable(s) entered on step 1: LogContr, Dem, LogContr_Dem.

b. Repeat part (a) but this time use nested models and the drop-in-deviance test (the nested likelihood ratio test). What is the p-value?

H0: B3=0

H1: B3 is not zero

Test statistic: 87.336 – 86.781 = 0.555

DF=1

Since X^2 = 0.555 and p value is 0.456 we can reject null hypothesis This indicates that there interaction term of dem and logContr are not useful to predict log odds of whether some vote, after accounting their logContr and party(dem)

Full model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	86.781[a]	.369	.502

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Nested model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	87.336[a]	.365	.497

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

c. How do the p-values from parts (a) and (b) compare? If they are the same, why is that? If they are different, why are they different?

No, p value is different. Z test p value is 0.472 and LRT p value is 0.456

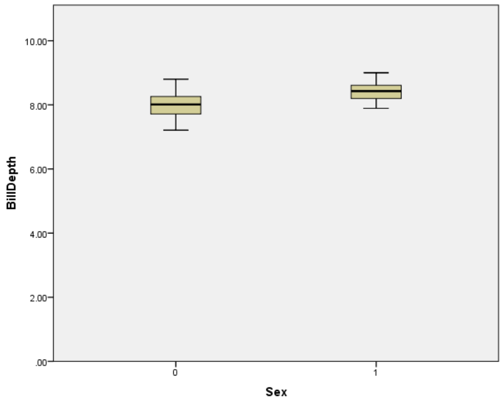
The z-test and the LRT are large–sample approximations even if all the required conditions are satisfied, the distributions for the z-statistics and for the drop-in-deviance are not exactly equal to the normal and chi–square distributions we use to get p-values. The p-values are only approximations to the exact p-values. The approximations tend to be better for larger samples.

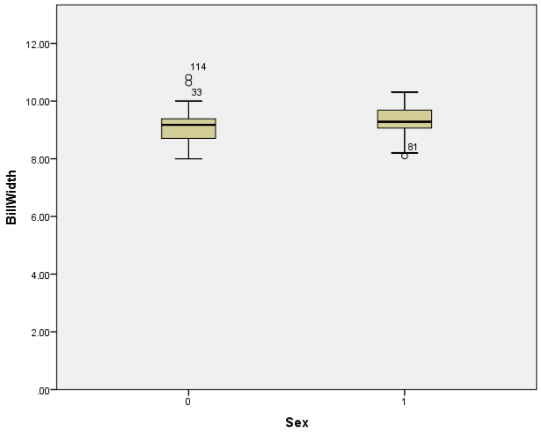
10.9 ***Blue jay morphology.*** Biologists took several measurements of the heads of blue jays. Among the variables measured were *BillLength*, *BillWidth*, and *BillDepth* (where *BillDepth* is the distance between the upper surface of the upper bill and the lower surface of the lower bill, measured at the nostril). All measurements were in millimeters. The data are in the file **Blue-Jays.**[13] We want to study the relationship between sex (coded as M/F in the variable KnownSex and as 1/0 in the variable Sex) and measurements of the blue jays.

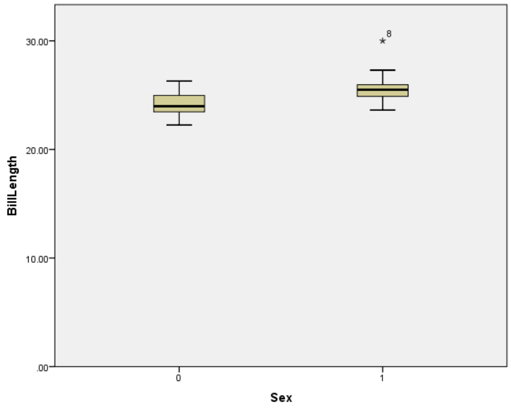
a. Make parallel boxplots of *BillLength* by *KnownSex*, *BillWidth* by *KnownSex*, and *BillDepth* by*KnownSex*. Which of these three predictors has the weakest relationship with *KnownSex*? Which has the strongest relationship?

   Based on boxplots below, the relationship between BillWidth and KnownSex is the weakest because for the different sex there is tiny change about BillWidth, compared with Knowsex and BillWidth, the same scale but more obvious difference between sex code 1 and sex code 0.

   BillLength and KnownSex have the strongest relationship. Compared knownsex with Billwidth and Billdepth, BillLengh has bigger scale and more obvious difference.







b. Fit a multiple logistic regression model on Sex depending on *BillLength*, *BillWidth*, and*BillDepth*. Which predictor has the largest p-value?

Billwidth has largest p value of 0.609 with wald statistics of .262

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BillWidth	-.240	.470	.262	1	.609	.786
	BillDepth	2.937	.831	12.485	1	.000	18.854
	BillLength	1.054	.288	13.393	1	.000	2.868
	Constant	-48.111	8.700	30.583	1	.000	.000

a. Variable(s) entered on step 1: BillWidth, BillDepth, BillLength.

c. Fit a simple logistic regression model of Sex depending on the predictor from part (b) that has the weakest relationship with Sex. What is the p-value for the coefficient of that predictor?

B1 represents coefficient of BillWidth

H0: B1=0

H1: B1 not equal to 0

Since wald statistics is 3.986 and p value is 0.046. So we can reject null hypothesis. This indicates that there is log linear relationship between BillWidth and odds of which sex.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BillWidth	.715	.358	3.986	1	.046	2.044

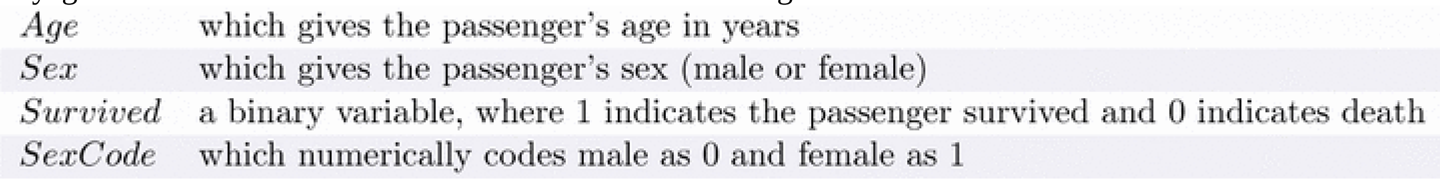
	Constant	-6.539	3.304	3.916	1	.048	.001

a. Variable(s) entered on step 1: BillWidth.

d. Comment on the results of parts (b) and (c). Why are the p-values so different between parts (b) and (c)?

The reason why p value are so different for part b and c is that BillWith highly correlate with Billdepth and Billlength and so it will be insignificant when we use them all in one model but it will be significant when BillWith works alone as one predictor.

10.12 ***Sinking of the Titanic (continued).*** In Exercises 9.17–9.20, we considered data on the passengers who survived and those who died when the oceanliner *Titanic* sank on its maiden voyage in 1912. The dataset in **Titanic** includes the following variables:



Age	which gives the passenger's age in years
Sex	which gives the passenger's sex (male or female)
Survived	a binary variable, where 1 indicates the passenger survived and 0 indicates death
SexCode	which numerically codes male as 0 and female as 1

a. In Exercises 9.17–9.20, you fit separate logistic regression models for the binary response*Survived* using *Age* and then *SexCode*. Now fit a multiple logistic model using these two predictors. Write down both the logit and probability forms for the fitted model.

Logit form: Log(odds survived) = B0 + B1*Age + B2* SexCode

Probability form: Probability of Survive = e^(B0 + B1*Age + B2* SexCode)/(1+e^(B0 + B1*Age + B2* SexCode))

b. Comment on the effectiveness of each of the predictors in the two-predictor model.

H0: B1=0

H1: B1 is not zero

Since Wald statistics is 190.952 and p value is smaller than 0.05 we reject null hypothesis. This indicates that there is a log linear relationship between sex and odds of whether survived, after accounting their ages

H0: B2=0

H1: B2 is not zero

Since Wald statistics is 1.054 and p value is greater than 0.05 we cannot reject null hypothesis. This indicates that there is no a log linear relationship between age and odds of whether survived, after accounting their sex

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	2.466	.178	190.952	1	.000	11.775
	Age	-.006	.006	1.054	1	.305	.994
	Constant	-1.160	.220	27.882	1	.000	.314

a. Variable(s) entered on step 1: Sex, Age.

c. According to the fitted model, estimate the probability and odds that an 18-year-old man would survive the *Titanic* sinking.

Probability form: Probability of Survive = e^(B0 + B1*Age + B2* SexCode)/(1+e^(B0 + B1*Age + B2* SexCode))

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	2.466	.178	190.952	1	.000	11.775
	Age	-.006	.006	1.054	1	.305	.994
	Constant	-1.160	.220	27.882	1	.000	.314

a. Variable(s) entered on step 1: Sex, Age.

Probability of Survive^ = e^(-1.160 + -.006*18 +2.466* 0)/(1+ e^(-1.160 + -.006*18 +2.466* 0)) = 0.22

d. Repeat the calculations for an 18-year-old woman and find the odds ratio compared to a man of the same age.

Probability form: Probability of Survive = e^(B0 + B1*Age + B2* SexCode)/(1+e^(B0 + B1*Age + B2* SexCode))

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	2.466	.178	190.952	1	.000	11.775
	Age	-.006	.006	1.054	1	.305	.994

	Constant	-1.160	.220	27.882	1	.000	.314

a. Variable(s) entered on step 1: Sex, Age.

Probability of Survive^ = e^(-1.160 + -.006*18 +2.466* 1)/(1+ e^(-1.160 + -.006*18 +2.466* 1)) = 0.77

Interpretation of odds ratio of sex: going from 18 year old male to 18 year old female, the odds of surviving increases by a factor of 11.775.

10.16 ***Leukemia treatments (continued).*** Refer to **Exercise 9.21** that describes data in **Leukemia** that arose from a study of 51 patients treated for a form of leukemia. The first six variables in that dataset all measure pretreatment variables: *Age, Smear, Infil, Index, Blasts*, and *Temp*. Fit a multiple logistic regression model using all six variables to predict *Resp*, which is 1 if a patient responded to treatment and 0 otherwise.

a. Based on values from a summary of your model, which of the six pretreatment variables appear to add to the predictive power of the model, given that other variables are in the model?

Temp and Age

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-.062	.027	5.149	1	.023	.940
	Smear	-.005	.040	.014	1	.907	.995
	Infil	.031	.038	.671	1	.413	1.032
	Index	.373	.132	7.920	1	.005	1.452
	Blasts	.033	.046	.503	1	.478	1.033
	Temp	-.112	.043	6.856	1	.009	.894
	Constant	108.331	41.844	6.703	1	.010	1.116E+47

a. Variable(s) entered on step 1: Age, Smear, Infil, Index, Blasts, Temp.

b. Specifically, interpret the relationship (if any) between *Age* and *Resp* and also between *Temp*and *Resp* indicated in the multiple model.

If a H0: B1=0

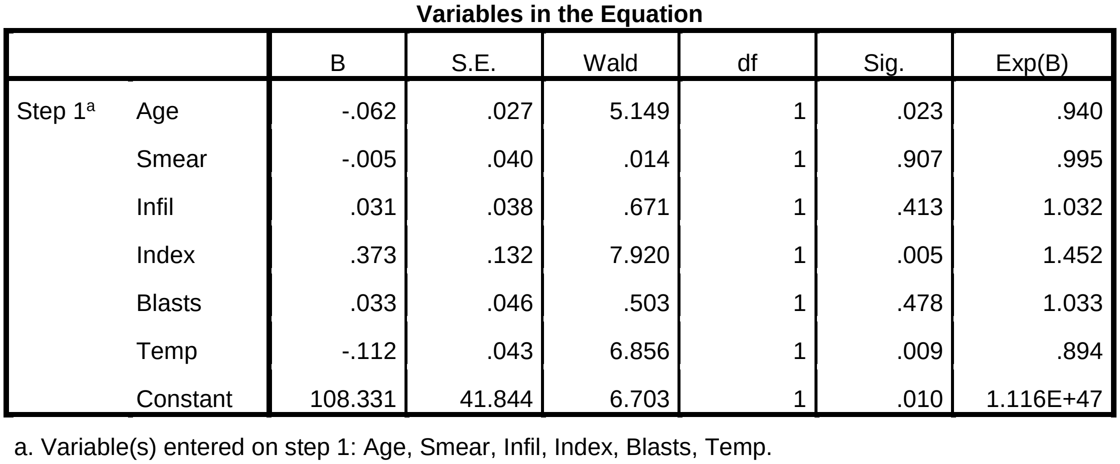
H1: B1 is not zero

Since Wald statistics is 5.149 and p value is greater than 0.05 we cannot reject null hypothesis. This indicates that there is no a log linear relationship between age and odds of whether responded, after accounting their Smear, Infil, Index, Blasts, and Temp

H0: B6=0

H1: B6 is not zero

Since Wald statistics is 6.856 and p value is greater than 0.05 we can reject null hypothesis. This indicates that there is a log linear relationship between Temp and odds of whether responded, after accounting their Age, Smear, Infil, Index, and Blasts.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-.062	.027	5.149	1	.023	.940
	Smear	-.005	.040	.014	1	.907	.995
	Infil	.031	.038	.671	1	.413	1.032
	Index	.373	.132	7.920	1	.005	1.452
	Blasts	.033	.046	.503	1	.478	1.033
	Temp	-.112	.043	6.856	1	.009	.894
	Constant	108.331	41.844	6.703	1	.010	1.116E+47

a. Variable(s) entered on step 1: Age, Smear, Infil, Index, Blasts, Temp.

c. predictor variable is nonsignificant in the fitted model here, might it still be possible that it should be included in a final model? Explain why or why not.

Yes, some predictors are nonsignificant but might be significant when we take out other predictors. So we need to do the further analysis to see whether it is significant or not.

d. Despite your answer above, sometimes one gets lucky, and a final model is, simply, the model that includes all "significant" variables from the full additive model output. Use a nested likelihood ratio (drop-in-deviance) test to see if the model that excludes precisely the non significant variables seen in (a) is a reasonable choice for a final model. Also, comment on the stability of the estimated coefficients between the full model from (a) and the reduced model without the “nonsignificant” terms.

H0: B2=B3=B4=B5

H1: at least one is not zero

Test statistic: 60.778 – 39.275 = 21.503

DF=4

P value is 0.0002. Since X^2 = 21.503and p value is 0.0002 we can reject null hypothesis This indicates that Smear, Infil, and Index, Blasts are useful to predict odds of whether someone responded, after accounting their age and temp.

Full model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	39.275[a]	.458	.612

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Nested model

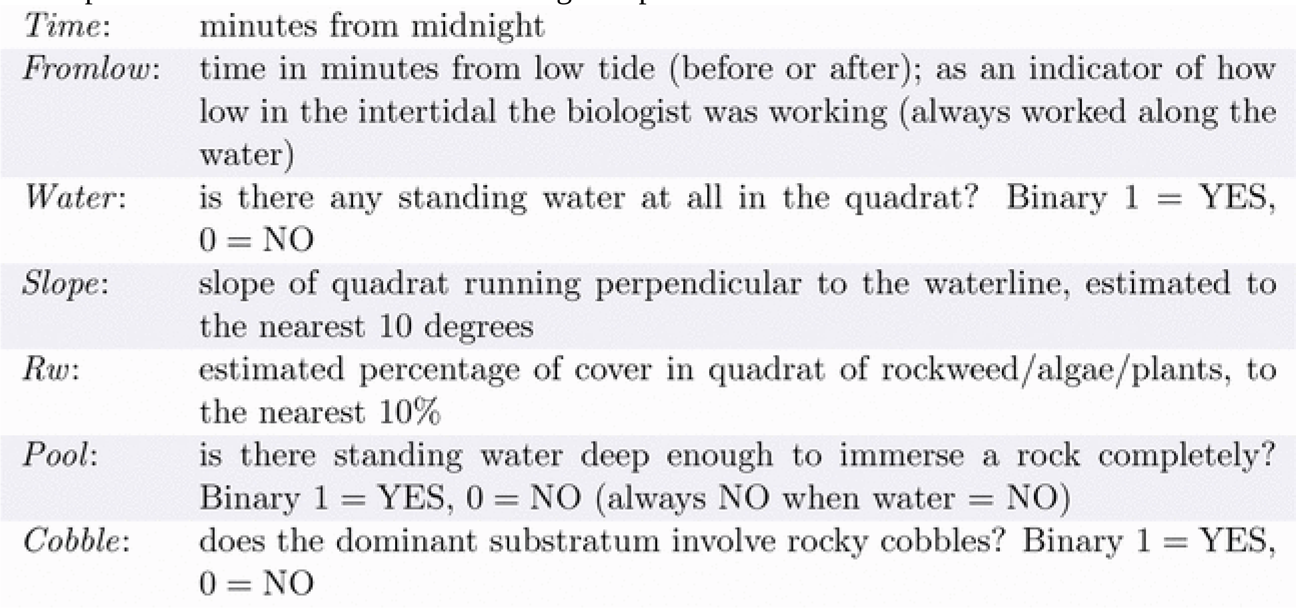
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	60.778[a]	.174	.232

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

10.26 ***Gunnels (continued).*** Construct a model that predicts the presence of gunnels using as predictors seven variables that relate to the geology and timing of the observation: *Time*, *Fromlow*, *Water*, *Slope*, *Rw*, *Pool*, and *Cobble*.

These predictor variables have the following interpretations:



Time: minutes from midnight

Fromlow: time in minutes from low tide (before or after); as an indicator of how low in the intertidal the biologist was working (always worked along the water)

Water: is there any standing water at all in the quadrat? Binary 1 = YES, 0 = NO

Slope: slope of quadrat running perpendicular to the waterline, estimated to the nearest 10 degrees

Rw: estimated percentage of cover in quadrat of rockweed/algae/plants, to the nearest 10%

Pool: is there standing water deep enough to immerse a rock completely? Binary 1 = YES, 0 = NO (always NO when water = NO)

Cobble: does the dominant substratum involve rocky cobbles? Binary 1 = YES, 0 = NO

a. Give a summary table of the model that lists the statistical significance or lack thereof for the seven predictors. Using the 0.05 significance level, which variables are statistically significant?

   Based on the table below, the significant predictors include Fromlow with p value of <0.001, Time p value of 0.003, Rw p value of 0.016, and cobble with p value of <0.001

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Fromlow	-.031	.005	32.407	1	.000	.970
	Water	.460	.443	1.076	1	.299	1.584

	Slope	.026	.014	3.431	1	.064	1.026
	Time	-.004	.001	8.698	1	.003	.996
	Rw	1.431	.592	5.837	1	.016	4.183
	Pool	.443	.411	1.166	1	.280	1.558
	Cobble	2.683	.395	46.221	1	.000	14.624
	Constant	-1.137	.781	2.119	1	.146	.321

a. Variable(s) entered on step 1: Fromlow, Water, Slope, Time, Rw, Pool, Cobble.

b. Explain in plain language why we might not trust a model that simply eliminates from the model the nonsignificant predictors.

We need to further analysis to compare efficiency of nested model and full model. It is possible that full model is better than nested model.

c. Use a nested likelihood ratio test to ascertain whether the subset of predictors deemed non significant in (a) can be eliminated, as a group, from the model. Give a summary of the calculations in the test: a statement of the full and reduced models, the calculation of the test statistic, the distribution used to compute the p-value, the p-value, and a conclusion based on the p-value.

H0: B2=B3=B6

H1: at least one is not zero

G-Test statistic: 253.701 – 247.875 = 5.826

DF=3

P value is 0.1203. Since X^2 = 5.826 and we cannot reject null hypothesis This indicates that water, slope, and pool are not useful to predict odds of whether presence of gunnel, after accounting their fromlow, rw, time, and cobble.

Full model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square

1	247.875[a]	.108	.458

a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

Nested model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	253.701[a]	.105	.444

a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

d. Would you recommend using a series of nested LRT tests to find a final model that is more than the reduced model and less than the full model? Explain why or why not?

No, because LRT's result is insignificant. That means nested LRT test already decides the model nested model is better one. So there is no need to find a reduced or less the full model.

e. Using the reduced model from (c), state the qualitative nature of the relationship between each of the predictors and the response variable, *Gunnel*; that is, state whether each has a positive or negative relationship, controlling for all other predictors.

For predictor fromlow

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Fromlow	-.032	.005	34.084	1	.000	.969

	Time	-.004	.001	8.327	1	.004	.996
	Rw	1.552	.577	7.237	1	.007	4.719
	Cobble	2.593	.371	48.809	1	.000	13.371
	Constant	-.631	.693	.830	1	.362	.532

a. Variable(s) entered on step 1: Fromlow, Time, Rw, Cobble.