STAT 659 Spring 2016
Homework 7 Solution

5.1

(a) The prediction equation is $\text{logit}\pi = -9.355 + 0.834\text{weight} + 0.307\text{width}$.

(b) The likelihood ratio test statistic is 32.867 which has a chi-squared distribution with degree of freedom 2. The P-value is 7.295×10^{-8} which is smaller than 0.05, so we reject the null hypothesis. There is evidence that at least one of weight and width have effect on the response.

(c) The Wald test statistic for the weight variable is 2.8430 with P-value 0.0918 and the Wald test statistic for the width variable is 1.5411 with P-value 0.2145. So we can see neither test show evidence of an effect. This is because of the multi-collinearity of the variables. Since one predictor can be explained by the other, so the test on one variable is not significant.

5.2

(a) Stepwise Procedure
At the beginning, we fit the intercept only model. Then in Step 1, we compute the LR statistic for each variable. Notice that the variable weight is with the smallest P-value corresponding to the LR statistics, but its P-value is less than $\pi_E = 0.15$. Thus the variable weight is selected into the model.
Then in step 2 we compute the LR statistics for the remaining variables. The variable color is with the smallest P-value corresponding to the LR statistics, but its P-value is less than $\pi_E = 0.15$. Thus the variable color is selected into the model. Also, use a LR test to obtain the P-value for deletion of the variable weight and the P-value is not greater than $\pi_R = 0.15$. We retain the variable weight.
Finally, no effects met the 0.15 significance level for entry into the model, and all variables in the model have P-values to remove that are less than $\pi_R = 0.15$ and P-values to enter that are greater than $\pi_E = 0.15$. Thus the final model consists of weight and color.

(b) Forward Selection
At the beginning, we fit the "intercept only model". Then in step 1, we compute the LR statistic for each variable. The variable weight has the smallest p-value which is less than $\pi_E = 0.15$.Thus the variable weight is selected in the model.
Then in stpe 2 we comput the LR statistic for remaining variables. The variable color

has the smallest p-value and it is less than $\pi_E = 0.15$. The variable color is selected in the model.
Finally, no additiona variables met the 0.15 significance level for entering to the model. Thus, final model has variables weight and color.

(c) Backward elimination
At the beginning, all predictors are entered to the model. Then in stpe 1, the wald statistic is computed for each variable. The variable spine has the greatest p-value and it is greater than $\pi_R = 0.15$. Thus, the variable spine is deleted.
Finally, no additional effects met the removal criterion. Thus, the final model contains variables weight and color.
Three selection procedures select the same model.

5.4

(a) According to the output, the Pearson chi-square test statistic is 10.9756 with P-value $0.4453 > 0.05$. So the model is adequate.

(b) The predictor JP can be removed because it has the smallest chi-square value which implies the highest P-value.

(c) The likelihood ratio test statistic is $11.1491 - 3.74 = 7.4091$ which has a Chi-square distribution with degree of freedom 6. The P-value is $0.2846 > 0.05$ which means the null hypothesis is not violated and none of the interaction terms are significant.

5.5

The model with only four binary main effect terms is preferred, since it has the lowest AIC score.

5.6

(a) Let Y be 1 if the people frequently drinks and be 0 if the people does not frequently drink.Then $P(Y = 1) = 0.092, P(\hat{Y} = 1|Y = 1) = 0.53, P(\hat{Y} = 0|Y = 0) = 0.66$. We can see the sensitivity gives the probability of positive diagnosis given the people frequently drinks and the specificity gives the probability of negative diagnosis given the people does not frequently drink.

(b) $P(\text{Correct Classification}) = P(\hat{Y} = 1|Y = 1)P(Y = 1) + P(\hat{Y} = 0|X = 0)P(Y = 0) = 0.092 \times 0.53 + (1 - 0.092) \times 0.66 = 0.648$.

(c) (i). We will choose the first one with highest concordance index. (ii). We will choose the second one, which has close prediction power to the first one but fewer number of predictors.

5.7

(a) The model with only the intercept is logit$\pi = \alpha$ with one parameter; The model with five parameters is logit$\pi = \alpha + \beta_1 EI + \beta_2 SN + \beta_3 TF + \beta_4 JP$; the model with 11 parameters is logit$\pi = \alpha + \beta_1 EI + \cdots, \beta_4 JP + \gamma_1 EI \times SN + \cdots + \gamma_6 TF \times JP$; the model with 15 parameters is logit$\pi = \alpha + \beta_1 EI + \cdots, \beta_4 JP + \gamma_1 EI \times SN + \cdots + \gamma_6 TF \times JP + \gamma_7 EI \times SN \times TF + \cdots, + \gamma_{10} SN \times TF \times JP$.(With 4 additional three-interaction terms).

(b) According the definition of AIC, we have for the 1 parameter model, $AIC = 1130.23{+}2 = 1132.23$; for the 5 parameters model, $AIC = 1124.86 + 2 \times 5 = 1134.86$; for the 11 parameters model, $AIC = 1119.87 + 2 \times 11 = 1141.87$; for the 15 parameters model, $AIC = 1116.47 + 2 \times 15 = 1146.47$. Since the model with only the intercept has the smallest AIC score, it is preferred.

(c) Since the prediction power is around 0.5, so the result will be close to that of the random guess. So it will not help us to predict well whether someone is a frequent smoker.

5.10

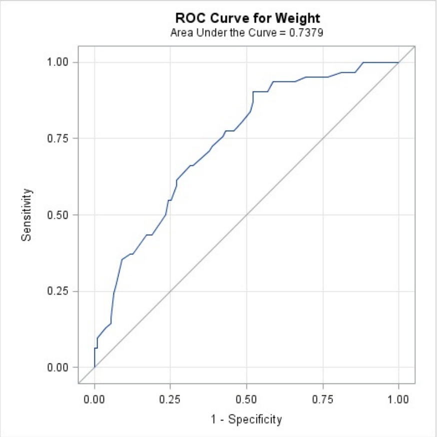
(a) The classification table is as follow:

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	68	43
$y = 0$	18	44

The sensitivity is $P(\hat{Y} = 1|Y = 1) = \frac{68}{68+43} = 0.613$ and the specificity is $P(\hat{Y} = 0|Y = 0) = \frac{44}{44+18} = 0.710$. So given the crab has satellites, the probability of correct classification is 0.613; given the crab does not have satellites, the probability of correct classification is 0.710.

(b) ROC curve is as follows:

Area under the curve is 0.7379, which is the concordance index. In the other words, the



probability that the predictions and outcomes are concordant is 0.7379.
(d) The linear model is $\text{logit}\pi = -3.6947 + 1.8151\text{weight}$ and the quadratic model is $\text{logit}\pi = -1.8877 + 0.2182\text{weight} + 0.3393\text{weight}^2$. The likelihood ratio test statistic for testing sgnificance of the quadratic term is 0.27669 with P-value 0.59888. So the linear model is adequate.
(e) The AIC score for the linear model is 199.7371 and the AIC score for the quadratic model is 201.4605. So the linear model is better according to the AIC score.

5.13

Through backward and forward selections, the variables chosen are laufkont, credit duration, moral, verw, and famges. Then we conduct a Hosmer and Lemeshow test, the test statistic is $6.7503 \sim \chi^2_8$. Since the P-value is $0.5638 > 0.05$, the adequacy of the model is not violated.

Variable Selection Problem:

First define variables race and loc to be nominal variables. For the remaining part of the problem, variables race_i is a dummy variable with one when race have value of i. Variables loc_i are defined by similar way. Other variables are either dummy variables with zero and one values or numeric variables

(a) Set SLENTRY=0.15 and SLSTAY=0.15. For forward selection, the final selected model has variables loc_2, typ, age, can and sys. For backward and stepwise selections, the final selected model has variables: age, can, sys, typ, ph, pco, loc_2.

(b) First, univariate analyses is performed using STA as the response and each of the other variables as explanatory variables. The results appear in the table I:
The Pvalues for sex, race, can, hra, pre, fra, po2, ph, pco, bic are large, and we choose to omit them from our main effects model. The main effects model containing the 9 predictors was fit to the data. The results appear in the table II:
Only the variables age, sys, typ, and loc_2 appear to be statistically. Thus, final main effect model is sta=age sys typ loc.

(c) By using methods of part (a) and (b), the final model commonly contains variables age,sys,typ,and loc_2. But variables ph,can,and pco are only contained in the final model which is selected by backward elimination, and the variable can is contained in the final models which are selected by forward method and backward elimination.

Table I: Table I

Parameter	Estimate	Std. Error	G^2	DF	P-value
age	0.0275	0.0106	7.8546	1	0.0051
sex	0.1054	0.3617	0.0845	1	0.7713
race_2	-1.3227	1.0515	2.2599	2	0.3231
race_3	-0.0700	0.8120			
ser	-0.9469	0.3682	6.9190	1	0.0085
can	0	0.5893	0.0000	1	1.0000
crn	1.2198	0.5039	5.4242	1	0.0199
inf	0.9163	0.3617	6.5756	1	0.0103
cpr	1.6945	0.5885	7.9324	1	0.0049
sys	-0.0170	0.00600	8.8258	1	0.0030
hra	0.00294	0.00655	0.2006	1	0.6543
pre	0.2339	0.4732	0.2374	1	0.6261
typ	2.1846	0.7450	15.1120	1	0.0001
fra	0	0.6712	0.0000	1	1.0000
po2	0.6601	0.5711	1.2402	1	0.2654
ph	0.6228	0.6289	0.9099	1	0.3401
pco	0	0.5893	0.0000	1	1.0000
bic	0.7621	0.5790	1.5988	1	0.2061
cre	1.4881	0.6596	4.7650	1	0.0290
loc_1	16.4226	680.8	36.3765	2	<.0001
loc_2	3.1531	0.8175			

Table II: Table II

Parameter	DF	Estimate	Std. Error	Wald Chi-Square	Pr ¿ ChiSq
Intercept	1	-3.7237	1.6973	4.8133	0.0282
age	1	0.0321	0.0128	6.3159	0.0120
ser	1	-0.0630	0.5071	0.0154	0.9011
crn	1	0.5859	0.7104	0.6802	0.4095
inf	1	0.0990	0.4784	0.0429	0.8360
cpr	1	0.8722	0.7910	1.2157	0.2702
sys	1	-0.0171	0.00790	4.7037	0.0301
typ	1	2.3793	1.0918	4.7493	0.0293
cre	1	0.5280	0.8538	0.3824	0.5363
loc_1	1	17.6618	546.5	0.0010	0.9742
loc_2	1	1.9378	0.9402	4.2477	0.0393