Homework 07
Joseph Blubaugh
jblubau1@tamu.edu
STAT 659-700

5.1

a. Model: $log(\frac{\pi(x)}{1-\pi(x)}) = -3.6201 + .1594(weight) + .1043(width)$

```
crabs = read.csv("crabs.csv")

mdl = glm(cbind(crabs$y, crabs$n) ~ weight + width, family = binomial(),
          data = crabs)
summary(mdl)
```

```
Call:
glm(formula = cbind(crabs$y, crabs$n) ~ weight + width, family = binomial(),
    data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1281  -0.8639   0.1161   0.3554   0.7188

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6201     2.5230  -1.435    0.151
weight        0.1594     0.4841   0.329    0.742
width         0.1043     0.1335   0.781    0.435

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 72.305  on 172  degrees of freedom
Residual deviance: 66.114  on 170  degrees of freedom
AIC: 225.99

Number of Fisher Scoring iterations: 4
```

b. Likelihood Ratio Test: 1 - pchisq(72.305 - 66.114, 2) = .045. Conclude that there is some evidence that the betas are not equal to 0 so we reject $Ho : \beta_1 = \beta_2 = 0$.

c. $Ho : Weight = 0, Ha : Weight > 0, .329 < 1.96$, fail to reject
$Ho : Width = 0, Ha : Width > 0, .781 < 1.96$, fail to reject
The individual coefficients may be 0 but the linear combination of the 2 plus the intercept is significant.

5.2

Backward selection and forward selection differ

Using Backward Selection

Iteration 1: $log(\frac{\pi(x)}{1-\pi(x)}) = weight + spine + color$

- Color Removed: AIC = 226.5
- Spine Removed: AIC = 228.4
- Weight Removed: AIC = 232.3
- None Removed: AIC = 230.0

Iteration 2: $log(\frac{\pi(x)}{1-\pi(x)}) = weight + spine$

- Spine Removed: AIC = 224.6
- Weight Removed: AIC = 230.1
- None Removed: AIC = 226.5

Iteration 3: $log(\frac{\pi(x)}{1-\pi(x)}) = weight$

- Weight Removed: AIC = 228.18
- None Remove: AIC = 224.6

Final Model: $log(\frac{\pi(x)}{1-\pi(x)}) = weight$

Using Forward Selection

Iteration 1:

- Color Added: AIC = 230.4
- Width Added: AIC = 224.1
- Weight Added: AIC = 224.6
- Intercept Only: AIC = 228.1

Iteration 2: $log(\frac{\pi(x)}{1-\pi(x)}) = width$

- Color Added: AIC = 228.0
- Weight Added: AIC = 225.9
- None Added: AIC = 224.1

Final Model: $log(\frac{\pi(x)}{1-\pi(x)}) = width$

5.4

a. $G^2 = 30.48 - 11.14 = 19.34 \rightarrow 1 - pchisq(19.34, 4) = .009$ There is enough evidence to reject the null hypothesis that at least one variable is significant.
b. I would remove JP, the coefficient is the closest to 0, the 95% confidence interval includes 0, and the Chi-squared statistic is very low.
c. $G^2 = 10.97 - 3.74 = 7.23 \rightarrow 1 - pchisq(7.23, 6) = .30$ Fail to reject the null hypothesis that the interaction model is an improvement of the simpler model

5.5

The best model is the model with the lowest AIC = 637.5. AIC tends to pick more complicated models, but in this case the interaction terms are not significant so adding the terms does not improve the model enough for the degrees of freedom that are taken away.

5.6

a. Sensitivity is the conditional probability of getting a true reponse, given that the true answer is true. Specificity is the conditional probability pf gettomg a false response, given that the true answer is false. The higher the sensitivity and specificity, the more powerful the test. With these results, there is a 53% probability of getting a true positive and a 66% probability of getting a false negative.

b. If n = 1000, $(48.8 + 642.5)/1000 = .65$

	Predicted.Drinking	Predicted.NoDrinking	Total
Drinking	48.8	43.2	92
No Drinking	308.7	599.3	908
Total	357.5	642.5	1000

c. i. If you want the model with the best predictive power then you choose the model with the highest concordance index that has the 4 main effects and interaction terms.

ii. If you are concerned with statistical parsimony you want the simplest model which is the model with only the T/F indicator, however I would choose the main effects model over the T/F model because it explains a lot more without the added complexity of interpreting the interactions.

5.7

a.

- Model 1: $log(\frac{\pi(x)}{1-\pi(x)}) = -1.214$
- Model 2: $log(\frac{\pi(x)}{1-\pi(x)}) = -1.214 - .0915(TF) - .287(JP) - .146(EI) - .17(SN)$
- Model 3: $log(\frac{\pi(x)}{1-\pi(x)}) = -1.214 - .0915(TF) - .287(JP) - .146(EI) - .17(SN) + .195(TF * JP) + .054(TF * EI) + .078(TF * SN) - .141(JP * EI) - .117(JP * SN) + .024(EI * SN)$

- Model 4: $log(\frac{\pi(x)}{1-\pi(x)}) = -1.214 - .0915(TF) - .287(JP) - .146(EI) - .17(SN) + .195(TF * JP) + .054(TF * EI) + .078(TF * SN) - .141(JP * EI) - .117(JP * SN) + .024(EI * SN) + .341(TF * JP * EI) + .365(TF * JP * SN) + .224(TF * EI * SN) + .029(JP * EI * SN)$

b. Under AIC where the best model has the lowest AIC, model 4 is the best.

- Model 1: AIC = -2*(1 + 32) + 1130.23 = 1064.2
- Model 2: AIC = -2*(5 + 32) + 1124.86 = 1050.86
- Model 3: AIC = -2*(11 + 32) + 1119.87 = 1033.87
- Model 4: AIC = -2*(15 + 32) + 1116.47 = 1022.47

c. 1 - Specificity = .45, at that point sensitivity is only slightly better at .48 which is slightly better than a 50/50 guess so personality does help the prediction.

5.10

a. Given that the true answer is > 0 satellites, the probability of correctly predicting a crab has a satellite is .61. Given that the true answer is satellites = 0, the probability of correctly predicting a crab has 0 satellites is .72

```
pandoc.table(Class.Tbl, caption = "Classification Table")
```

Table 2: Classification Table

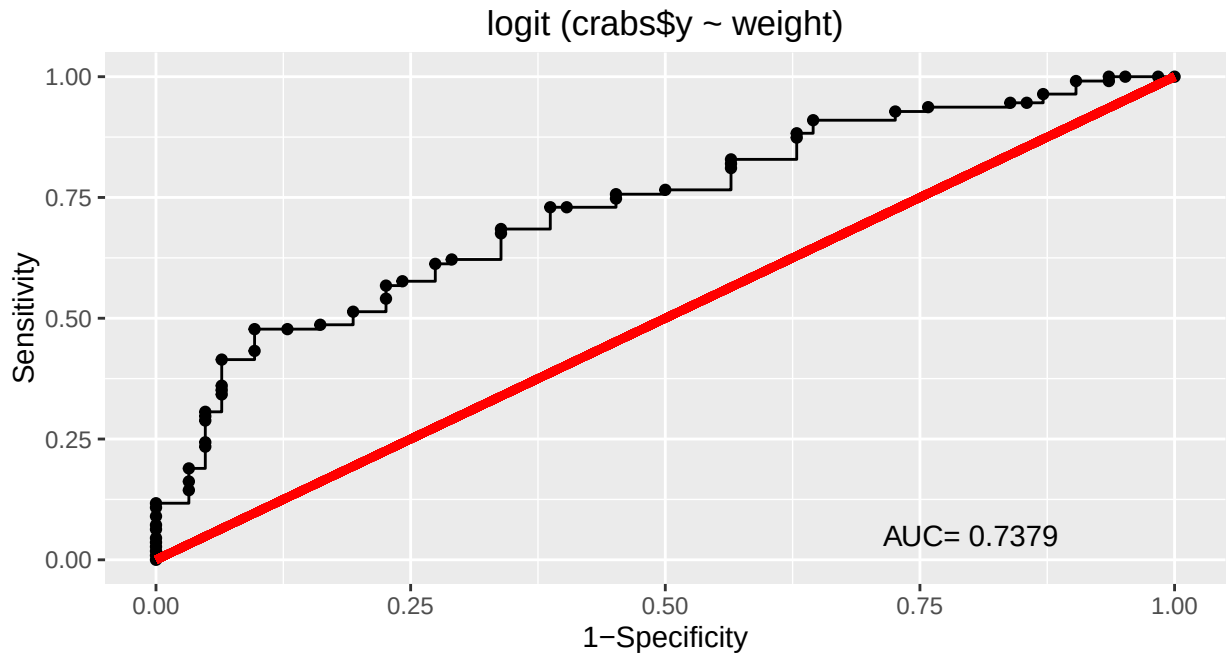
	yhat.1	yhat.0
y1	68	43
y0	17	43

```
(sensitivity = 68/111); (specificity = 43/60)
```

[1] 0.6126126 [1] 0.7166667

b. Sensitivity is relatively high compared to 1-specificity. As specificity decreases, sensitivity increases meaning we are more likely to correctly predict the correct answer, but also more likely to incorrectly predict the wrong answer. The higher the AUC, the better the predictive power of the model.

```
library(Deducer)
rocplot(mdl)
```



d. The liklihood ratio test shows that the polynomial model is not an improvement over the original model. $G^2 = 195.74 - 195.46 = .28 < 3.84 = X^2_{1,.05} \rightarrow 1 - pchisq(.28, 1) = .59$

```
## Model used earlier
(mdl)
```

```

Call:  glm(formula = crabs$y ~ weight, family = binomial(), data = crabs)

Coefficients:
(Intercept)       weight
     -3.695        1.815

Degrees of Freedom: 172 Total (i.e. Null);  171 Residual
Null Deviance:      225.8
Residual Deviance: 195.7    AIC: 199.7
```

```
(mdl2 = glm(crabs$y ~ poly(weight, 2), family = binomial(), data = crabs))
```

```

Call:  glm(formula = crabs$y ~ poly(weight, 2), family = binomial(),
    data = crabs)

Coefficients:
     (Intercept)  poly(weight, 2)1  poly(weight, 2)2
          0.7717           15.1986            2.7228
```

```
Degrees of Freedom: 172 Total (i.e. Null);  170 Residual
Null Deviance:      225.8
Residual Deviance: 195.5    AIC: 201.5
```

e. Model AIC: 199.7, Model 2 AIC: 201.4. The simpler model has the lower AIC and is the better model which is confirmed by the likelihood ratio test in part d.

5.13

Using stepwise backward selection we need only one iteration to get to the final model. Only verw is insignificant and with it removed the model has the lowest AIC. On the second iteration the AIC does not decrease when any other variable is removed so we take the original model minus verw as the final model.

```
credit = read.csv("credit.csv")

mdl = glm(kredit ~ laufkont + laufzeit + moral + verw + famges,
          family = binomial(), data = credit)

stepAIC(mdl, direction = "backward")
```

```
Start:  AIC=1032.55
kredit ~ laufkont + laufzeit + moral + verw + famges

           Df Deviance    AIC
- verw      1   1020.9 1030.9
<none>          1020.5 1032.5
- famges    1   1026.2 1036.2
- moral     1   1046.6 1056.6
- laufzeit  1   1056.0 1066.0
- laufkont  1   1120.4 1130.4

Step:  AIC=1030.86
kredit ~ laufkont + laufzeit + moral + famges

           Df Deviance    AIC
<none>          1020.9 1030.9
- famges    1   1026.5 1034.5
- moral     1   1046.6 1054.6
- laufzeit  1   1056.2 1064.2
- laufkont  1   1121.6 1129.6


Call:  glm(formula = kredit ~ laufkont + laufzeit + moral + famges,
    family = binomial(), data = credit)

Coefficients:
```

```
(Intercept)     laufkont     laufzeit        moral       famges
   -1.41552      0.62695     -0.03632      0.36878      0.25382

Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
Null Deviance:      1222
Residual Deviance: 1021     AIC: 1031
```

Additional 1

```
icu = read.csv("icu.csv"); icu = icu[, -1]

mdl.null = glm(sta ~ 1, family = binomial(), data = icu)
mdl.full = glm(sta ~ ., family = binomial(), data = icu)

forward = step(mdl.null, scope = list(lower = formula(mdl.null), upper = formula(mdl.full)),
               direction = "forward", trace = 0)
backward = step(mdl.full, scope = list(lower = formula(mdl.null), upper = formula(mdl.full)),
               direction = "backward", trace = 0)
stepwise = step(mdl.null, scope = list(lower = formula(mdl.null), upper = formula(mdl.full)),
                direction = "both", trace = 0)

forward; backward; stepwise
```

```
Call:  glm(formula = sta ~ loc + typ + age + can + pco + ph + sys, family = binomial(),
    data = icu)

Coefficients:
(Intercept)          loc          typ          age          can
   -5.27888      2.34391      2.75305      0.04043      2.16474
        pco           ph          sys
   -2.29744      1.80960     -0.01099

Degrees of Freedom: 199 Total (i.e. Null);  192 Residual
Null Deviance:      200.2
Residual Deviance: 136.2    AIC: 152.2
```

```
Call:  glm(formula = sta ~ age + can + sys + typ + ph + pco + loc, family = binomial(),
    data = icu)

Coefficients:
(Intercept)          age          can          sys          typ
   -5.27888      0.04043      2.16474     -0.01099      2.75305
         ph          pco          loc
    1.80960     -2.29744      2.34391
```

```
Degrees of Freedom: 199 Total (i.e. Null);  192 Residual
Null Deviance:      200.2
Residual Deviance: 136.2    AIC: 152.2
```

```
Call:  glm(formula = sta ~ loc + typ + age + can + pco + ph + sys, family = binomial(),
    data = icu)

Coefficients:
(Intercept)          loc          typ          age          can
   -5.27888      2.34391      2.75305      0.04043      2.16474
        pco           ph          sys
   -2.29744      1.80960     -0.01099

Degrees of Freedom: 199 Total (i.e. Null);  192 Residual
Null Deviance:      200.2
Residual Deviance: 136.2    AIC: 152.2
```

Additional 2

Forward, Backward, and Stepwise all converge on the same model while the manual method of pulling out insignificant vairiables based on an alpha of .05 results in a model with the only difference being that sys is not in the manual model.

```
## removing all non significant variables based on the highest pvalue > .05
mdl = glm(sta ~ ., family = binomial(), data = icu)
mdl = update(mdl, ~ . - race) # race pvalue = .993
mdl = update(mdl, ~ . - inf)  #  inf pvalue = .916
mdl = update(mdl, ~ . - crn)  #  crn pvalue = .880
mdl = update(mdl, ~ . - cre)  #  cre pvalue = .837
mdl = update(mdl, ~ . - hra)  #  hra pvalue = .697
mdl = update(mdl, ~ . - po2)  #  po2 pvalue = .679
mdl = update(mdl, ~ . - bic)  #  bic pvalue = .417
mdl = update(mdl, ~ . - ser)  #  ser pvalue = .435
mdl = update(mdl, ~ . - fra)  #  fra pvalue = .280
mdl = update(mdl, ~ . - cpr)  #  cpr pvalue = .314
mdl = update(mdl, ~ . - sex)  #  sex pvalue = .252
mdl = update(mdl, ~ . - pre)  #  pre pvalue = .205
mdl = update(mdl, ~ . - sys)  #  sys pvalue = .103

sort(forward$coefficients); sort(mdl$coefficients)
```

```
(Intercept)         pco         sys         age          ph         can
-5.27888001 -2.29744394 -0.01099413  0.04042549  1.80960204  2.16473984
        loc         typ
 2.34390530  2.75305027
```

```
(Intercept)         pco         age          ph         can         loc
-6.75127609 -2.13253582  0.04018236  1.76829880  2.14667503  2.30889684
        typ
 2.81591561
```

Additional 3

AIC selection shows that the best model includes sys and even though its not significant, it is not very insignificant either with a pvalue of .1 and so AIC determines that the added model complexity results in an overall better model than letting sys drop out. The models are the same except for the sys variable that dropped out.