Lab 2

K.C. Tobin, Weixing Sun, Winston Lin

July 2, 2017

The Lab

1. Even though it is not included in these questions, conduct a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course.

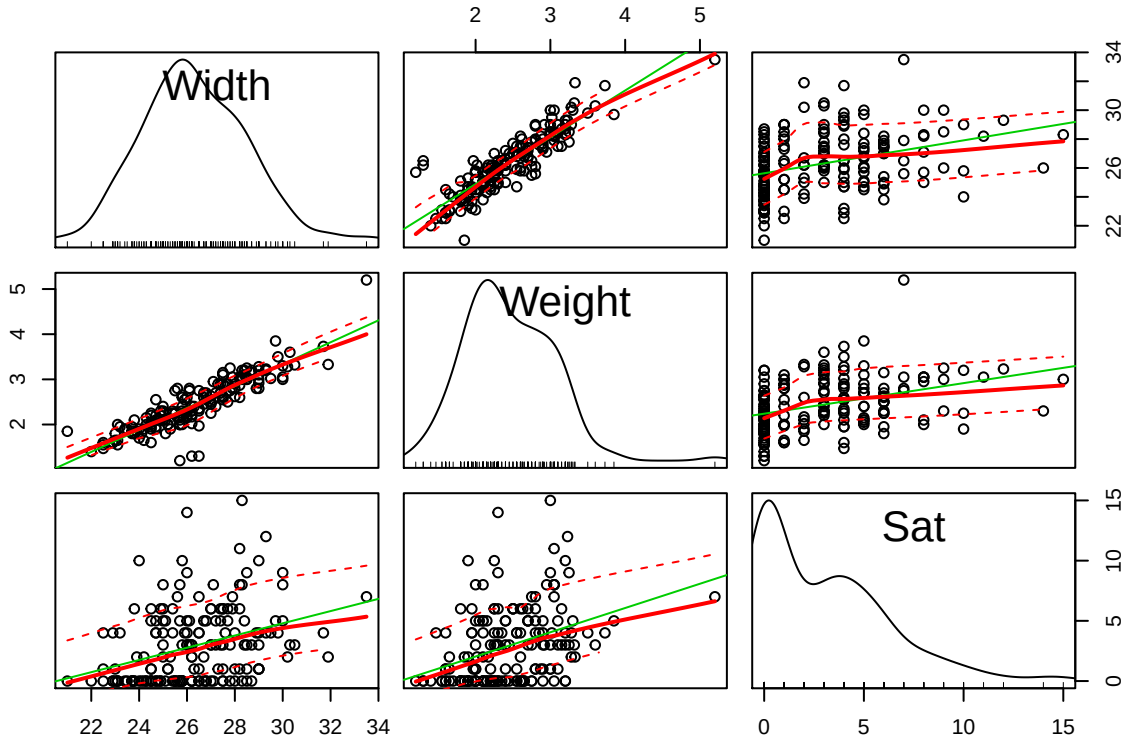
According to online information of this study (http://isites.harvard.edu/fs/docs/icb.topic79671.files/glmhand1.pdf):

- Color = female crab's color (1,2,3,4; 1=lightest to 4=darkest)
- Spine = female crab's spine condition (1 = both good, 2= one worn or broken, 3=both worn or broken)
- Width = female crab's carapace width (cm)
- Weight = female crab's weight (kg)

Insights from scatterplots:

- Width and weight have a clear positive linear correlation, which makes sense since a heavier horse shoe crab should be wider
- Sat shows a frequency distribution that might be fitted by Poisson model.

```
hs <- read.csv("HorseshoeCrabs.csv")
hs$Color.cat <- factor(hs$Color, labels = c("very light", "light", "dark", "very dark"))
hs$Spine.cat <- factor(hs$Spine, labels = c("2 good", "1 good", "0 good"))
scatterplotMatrix(hs[c(3,4,5)])
```



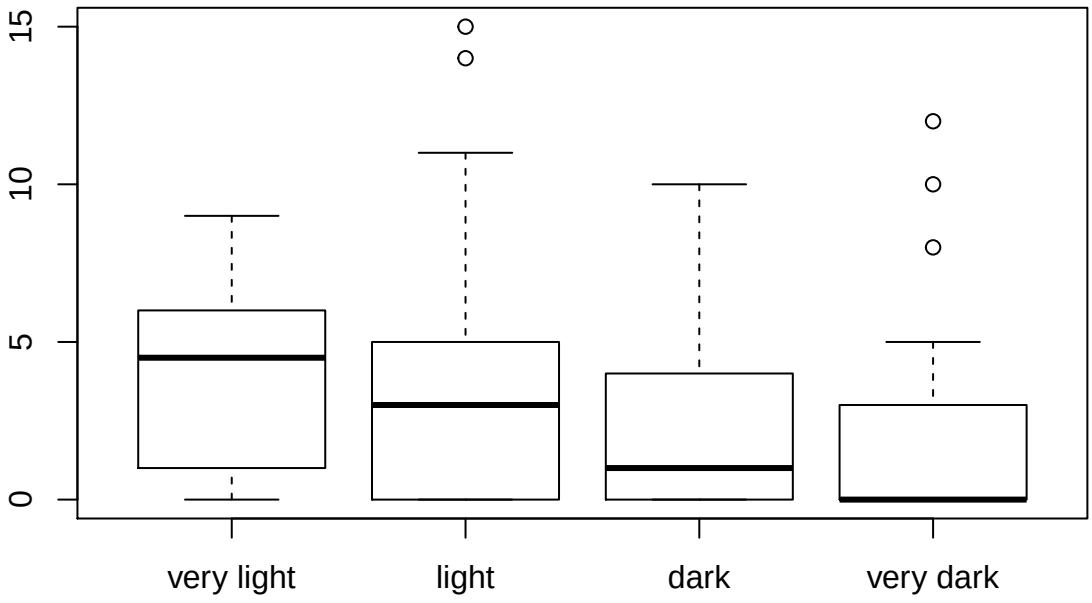
Insights from the contingency table and boxplots:

- Darker color and worse spine health ("0 good") appear to be associated with a lower number of satellite crabs. However, the strength of the correlation between spine health and satellite crabs is debatable given the low number of observations in the "1 good" category.
- For Color, only a few data points locate in the level "very light", which may give rise to a large uncertainty of fitting
- For Spine, only a few data points locate in the level "1 good", which may also give rise to a large uncertainty of fitting.

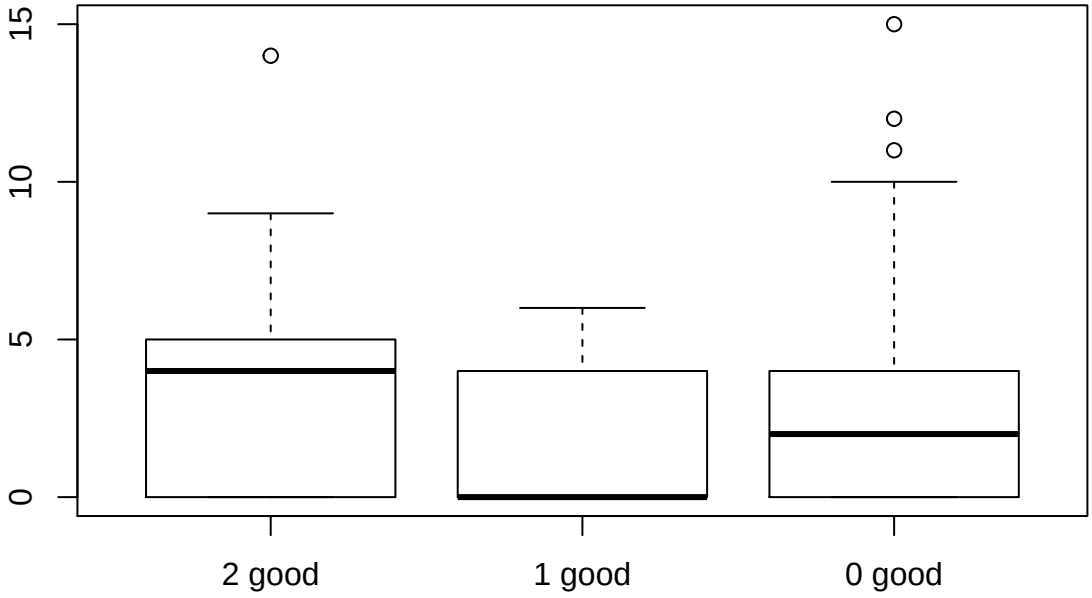
```
xtabs(~Color.cat + Spine.cat, data = hs)
```

```
##             Spine.cat
## Color.cat    2 good 1 good 0 good
##   very light      9      2      1
##   light          24      8     63
##   dark            3      4     37
##   very dark       1      1     20
```

```
boxplot(hs$Sat~hs$Color.cat)
```




```
boxplot(hs$Sat~hs$Spine.cat)
```



2. Do question 23, 24, 25 in C Chapter 5 (page 351 and 352) of Bilder and Loughin's *Analysis of Categorical Data with R.*

5.23. Agresti (2007) provides data on the social behavior of horseshoe crabs. These data are contained in the HorseshoeCrabs.csv file available on our website. Each observation corresponds to one female crab. The response variable is Sat, the number of "satellite" males in her vicinity. Physical measurements of the female—Color (4-level ordinal), Spine (3-level ordinal), Width (cm), and Weight (kg)—are explanatory variables.

a. Fit a Poisson regression model with a log link using all four explanatory variables in a linear form. Test their significance and summarize results.

- Results show that only Color and Weight parameters are statistically significant. LRT also shows strong evidence that the effect of color and weight are statistically significant (p-value < 0.05).

```r
# Poisson with ordinal variables
mod1 <- glm(Sat ~ Color + Spine + Width + Weight, family = poisson(link="log"), data=hs)
summary(mod1)
```

```
##
## Call:
## glm(formula = Sat ~ Color + Spine + Width + Weight, family = poisson(link = "log"),
##     data = hs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0127  -1.8844  -0.5401   0.9449   4.9605
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.52381    0.94909  -0.552  0.58101
## Color       -0.18503    0.06652  -2.781  0.00541 **
## Spine        0.04007    0.05681   0.705  0.48061
## Width        0.02728    0.04796   0.569  0.56954
## Weight       0.47319    0.16493   2.869  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 551.83  on 168  degrees of freedom
## AIC: 917.13
##
## Number of Fisher Scoring iterations: 6
```

```r
Anova(mod1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Sat
##        LR Chisq Df Pr(>Chisq)
## Color    7.9690  1   0.004759 **
## Spine    0.5009  1   0.479111
## Width    0.3221  1   0.570374
## Weight   8.3329  1   0.003893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Next, we transform Color and Spine to categorical variables to run Poisson fitting again. It shows similar results with mod1. However, it is noticable that only dark colors (3 & 4) show a significant effect on satellites.

```r
# Poisson with categorical variables
mod2 <- glm(Sat ~ Color.cat + Spine.cat + Width + Weight, family = poisson(link="log"), data=hs)
summary(mod2)
```

```
##
## Call:
## glm(formula = Sat ~ Color.cat + Spine.cat + Width + Weight, family = poisson(link = "log"),
##     data = hs)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0291  -1.8632  -0.5991   0.9331   4.9449
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.35722    0.96700  -0.369  0.71182
## Color.catlight     -0.26491    0.16811  -1.576  0.11507
## Color.catdark      -0.51374    0.19536  -2.630  0.00855 **
## Color.catvery dark -0.53126    0.22692  -2.341  0.01922 *
## Spine.cat1 good    -0.15044    0.21358  -0.704  0.48119
## Spine.cat0 good     0.08742    0.11993   0.729  0.46604
## Width               0.01651    0.04894   0.337  0.73582
## Weight              0.49712    0.16628   2.990  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 549.56  on 165  degrees of freedom
## AIC: 920.86
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(mod2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Sat
##           LR Chisq Df Pr(>Chisq)
## Color.cat   9.2463  3   0.026190 *
## Spine.cat   1.7984  2   0.406896
## Width       0.1135  1   0.736183
## Weight      9.0654  1   0.002605 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Compute deviance/df and interpret its value.

- Deviance/df, the ratio of residual deviance to residual degrees of freedom, is 3.285 (551.83/168). The ratio is greater than 1.181, the heuristics indicating that the model is a poor fit (Bilder, p. 293).

```
# null model
mod0 = glm(formula = Sat ~ 1, family = poisson(link = 'log'), data = hs)

# LRT to calculate deviance and df
lrt = anova(mod0, mod1, test = 'Chisq')
lrt
```

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ 1
## Model 2: Sat ~ Color + Spine + Width + Weight
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       172     632.79
```

```
## 2       168     551.83  4   80.962 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

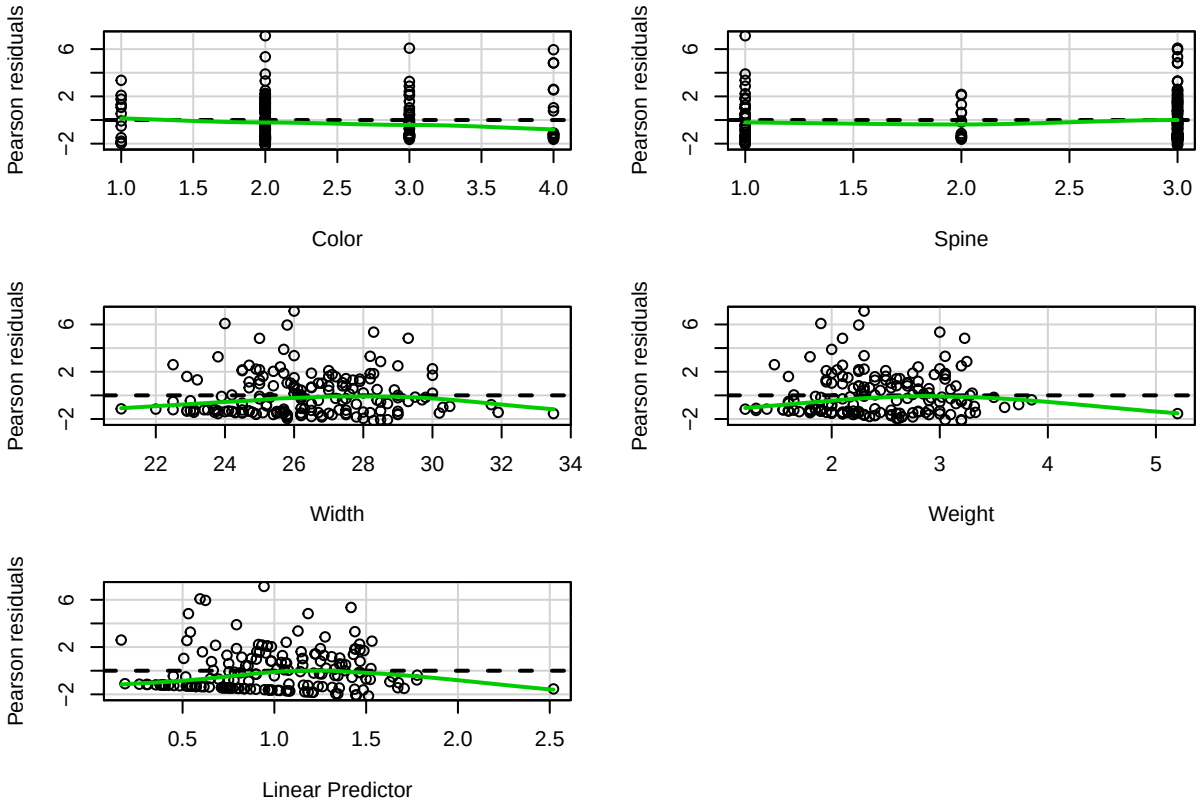
```
dev.df = lrt$`Resid. Dev`/lrt$`Resid. Df`

# heuristics for concern with model and poor fit
t1 = 1 + 2*sqrt(2/lrt$`Resid. Dev`[2])
t2 = 1 + 3*sqrt(2/lrt$`Resid. Dev`[2])

# results
round(c(devdf = dev.df[2], concern = t1, poor = t2), 3)
```

```
##   devdf concern    poor
##   3.285   1.120   1.181
```

c. Examine residual diagnostics and identify any potential problems with the model.

- Residual plots show that there is no systematic patterns for residual distribution, except for a few outliners. E(u|x) is slightly away from 0, while the homoskedasticity is highly preserved. Again, there are a few outliers that deviate expected residuals from 0. The presence of large residuals may indicate overdispersion and imply that an important explanatory variable has been omitted by the model or the distribution used to estimate the model is not appropriate. More influence of outliners will be discussed in 24.

```
# plot(allEffects(mod2))
residualPlots(mod1)
```




```
##        Test stat Pr(>|t|)
## Color      0.910    0.340
## Spine      1.156    0.282
```

```
## Width     14.855    0.000
## Weight    10.879    0.001
```

```
# residualPlots(mod2)

# # standardized pearson residuals - appear to be random
# std.resid = rstandard(mod2, type = c('pearson'))
# plot(std.resid)
```

d. Carry out the GOF test described on page 296 using the PostFitGOFTest() function available in the PostFitGOFTest.R program of the same name from our website. Use the default number of groups. (M/5 when M $<$ 100)

e. State the hypotheses, test statistic and p-value, and interpret the results.

- The hypothesis is the observed residuals of our model predictions for the 20 quantiles is not significantly different from the predictions thus indicating a good model fit. The GOF test gives us a test statistic of 776.2152 which is the total Pearson residuals at our 20 quantiles. This test statistic gives a p-value of effectively 0 on chi-squared distribution with 18 degrees of freedom indicating there is some evidence that our model is not well fit. This evidence support previous graphical analysis already conducted above.

ii. Plot the Pearson residuals for the groups against the interval centers (availabe in the pear.res and centers components, respectively, of the list returned by the function). Use this plot and the resiudal plots from part (c) to explain the results.

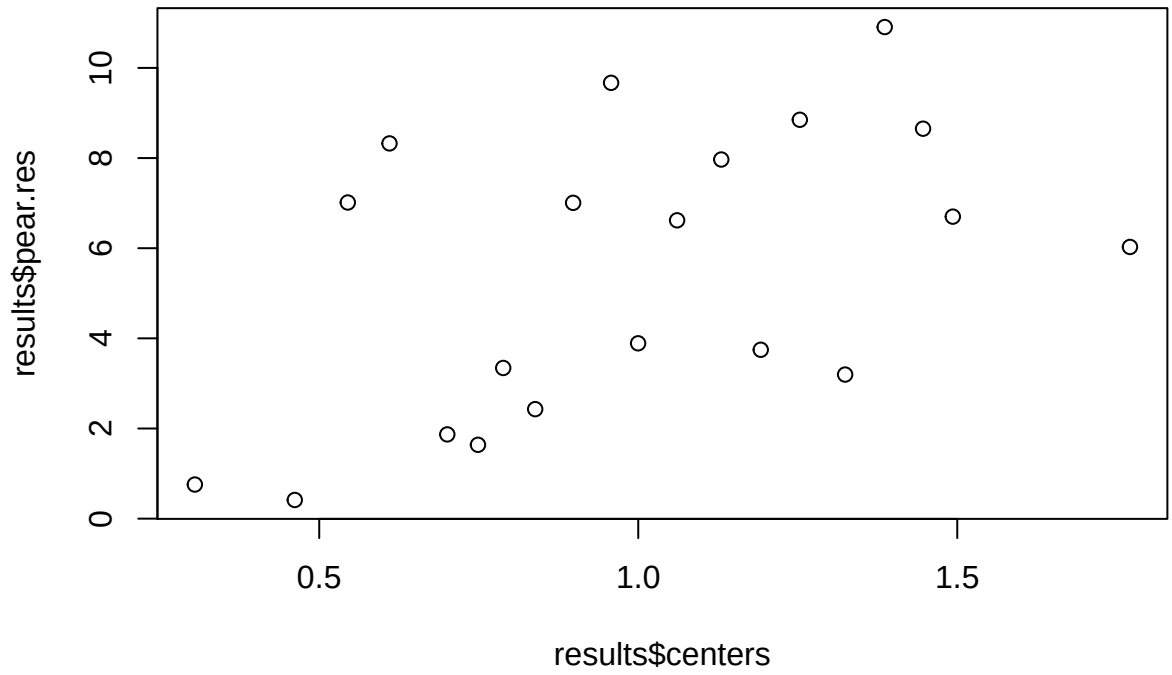
- As you can see in the plot below the Pearson residuals for out 20 quantiles range quite drastically from just below 0 to over 8. These residuals are much higher than we would expect if our model was an accurate representation of the data. In addition you can see the model is not consistent and the residuals vary based on predicted values.

```
PostFitGOFTest = function(obs, pred, g = 0) {
  if(g == 0) g = round(min(length(obs)/5,20))
  ord <- order(pred)
  obs.o <- obs[ord]
  pred.o <- pred[ord]
  interval = cut(pred.o, quantile(pred.o, 0:g/g), include.lowest = TRUE)
  # Creates factor with levels 1,2,...,g
  counts = xtabs(formula = cbind(obs.o, pred.o) ~ interval)
  centers <- aggregate(formula = pred.o ~ interval, FUN = "mean")
  pear.res <- rep(NA,g)
  for(gg in (1:g)) pear.res[gg] <- (counts[gg] - counts[g+gg])/sqrt(counts[g+gg])
  pearson <- sum(pear.res^2)
  if (any(counts[((g+1):(2*g))] < 5))
  warning("Some expected counts are less than 5. Use smaller number of groups")
  P = 1 - pchisq(pearson, g - 2)
  cat("Post-Fit Goodness-of-Fit test with", g, "bins", "\n",
      "Pearson Stat = ", pearson, "\n", "p = ", P, "\n")
  return(list(pearson = pearson, pval = P, centers = centers$pred.o, observed = counts[1:g],
              expected = counts[(g+1):(2*g)], pear.res = pear.res))
}

preds <- predict(mod1, hs)
obs <- hs$Sat
results <- PostFitGOFTest(obs,preds)
```

```
## Warning in PostFitGOFTest(obs, preds): Some expected counts are less than
## 5. Use smaller number of groups
```

```
## Post-Fit Goodness-of-Fit test with 20 bins
##  Pearson Stat =  785.136
##  p =  0
```

```
# results

plot(results$pear.res~results$centers)
```



24. Conduct an influence analysis. Interpret the results.

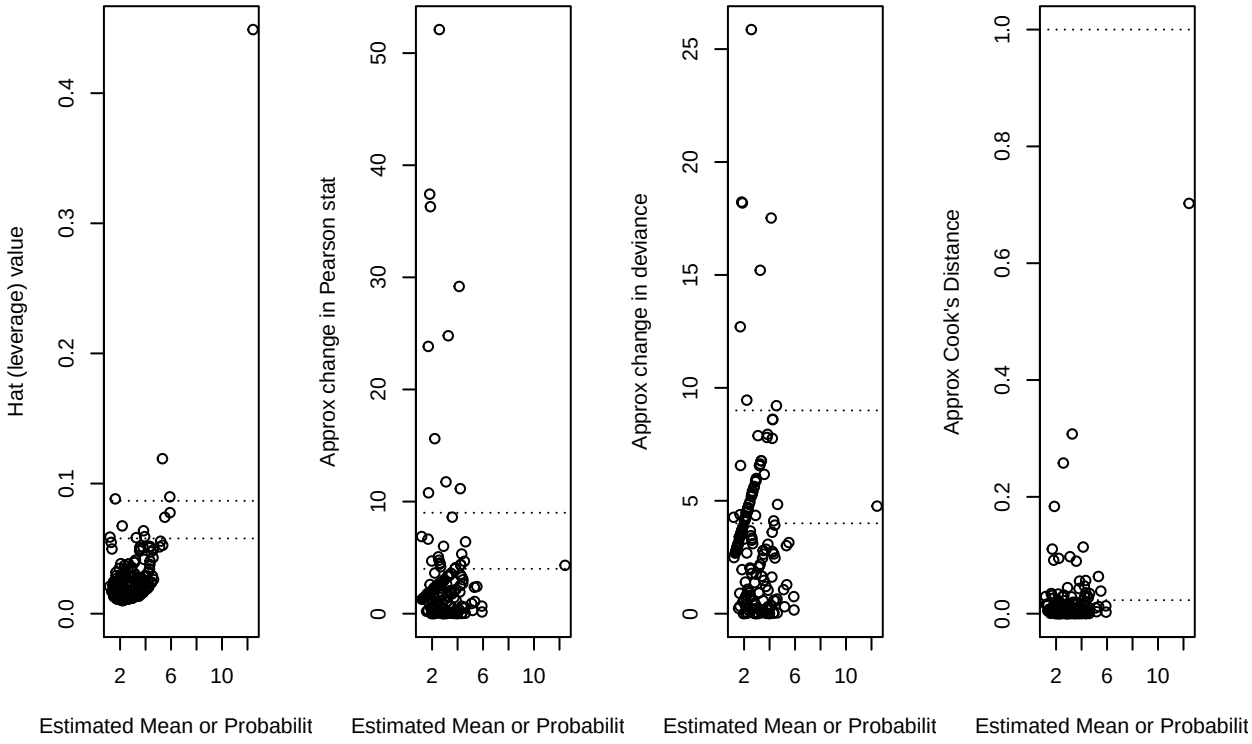
- The residuals-fitted plot shows that the there is some heteroskedasticity as spread of the residuals increases with the predicted values. Overall, it appears that the residuals are random and normal, as seen in the QQ plot. The residuals-leverage plot shows observation 165 to be a highly influential point, which we may remove in order to re-estimate the model.

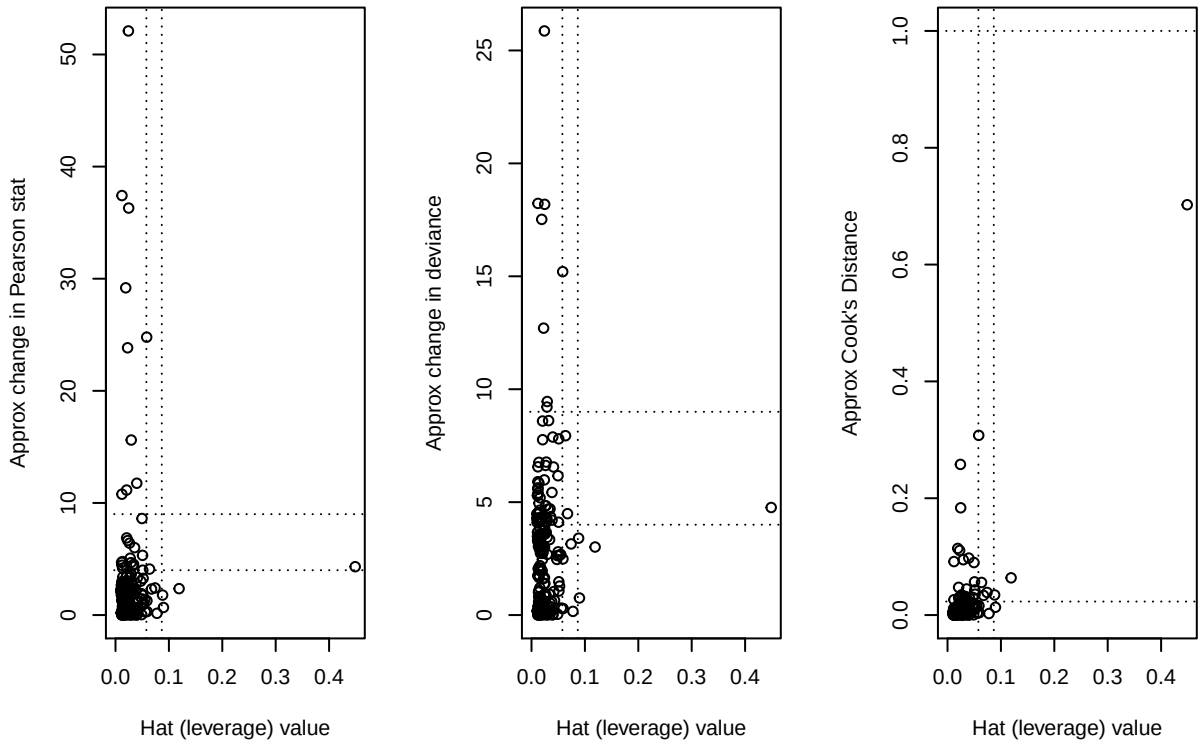
- In mod1 (ordinal), only one data point (No.165) shows a Cook's distance above 0.5, meaning a high influence on the fitting. In mod2 (categorical), this outliner shows a Cook's distance slightly below 0.5. The influence of this outliner is reduced by mod2, contributing to a more accurate fitting.

```
# plot(mod1)[4]
# plot(mod2)[4]

glmInflDiag <- function(mod.fit, print.output = TRUE, which.plots = c(1,2)){

 # Which set of plots to show
 show <- rep(FALSE, 2)  # Idea from plot.lm()
 show[which.plots] <- TRUE

 # Main quantities: Pearson and deviance residual, model Pearson and deviance stats
 pear <- residuals(mod.fit, type = "pearson")
 dres <- residuals(mod.fit, type = "deviance")
 x2 <- sum(pear^2)
 N <- length(pear)
 P <- length(coef(mod.fit))
 # Hat values (leverages)
```

```
 hii <- hatvalues(mod.fit)
 # Computed quantities: Standardized Pearson residual, Delta-beta, Delta-deviance
 sres <- pear/sqrt(1-hii)
# D.beta <- (pear^2*hii/(1-hii)^2)
# cookD <- D.beta / (P * summary(mod.fit)$dispersion)
 cookD <- pear^2 * hii / ((1-hii)^2 * (P) * summary(mod.fit)$dispersion)
 D.dev2 <- dres^2 + hii*sres^2
 D.X2 <- sres^2

 yhat <- fitted(mod.fit)

 # Plots against fitted values
 if(show[1] == TRUE) {
  # x11(height = 7,width = 15, pointsize = 15)
  par(mfrow = c(1,4), lty = "dotted")
  plot(x = yhat, y = hii, xlab = "Estimated Mean or Probability", ylab = "Hat (leverage) value",
   ylim = c(0, max(hii,3*P/N)))
  abline(h = c(2*P/N,3*P/N))

  plot(x = yhat, y = D.X2, xlab = "Estimated Mean or Probability", ylab = "Approx change in Pearson sta
   ylim = c(0, max(D.X2,9)))
  abline(h = c(4,9), lty = "dotted")

  plot(x = yhat, y = D.dev2, xlab = "Estimated Mean or Probability", ylab = "Approx change in deviance"
   ylim = c(0, max(D.dev2,9)))
  abline(h = c(4,9), lty = "dotted")

  plot(x = yhat, y = cookD, xlab = "Estimated Mean or Probability", ylab = "Approx Cook's Distance",
   ylim = c(0, max(cookD, 1)))
  abline(h = c(4/N,1), lty = "dotted")
 }

 # Plots against hat values
 if(show[2] == TRUE) {
  # x11(height = 6, width = 12, pointsize = 20)
  par(mfrow = c(1,3))
  plot(x = hii, y = D.X2, xlab = "Hat (leverage) value", ylab = "Approx change in Pearson stat",
   ylim = c(0, max(D.X2, 9)), xlim = c(0, max(hii,3*P/N)))
  abline(h = c(4,9), lty = "dotted")
  abline(v = c(2*P/N,3*P/N), lty = "dotted")

  plot(x = hii, y = D.dev2, xlab = "Hat (leverage) value", ylab = "Approx change in deviance",
   ylim = c(0, max(D.dev2,9)), xlim = c(0, max(hii,3*P/N)))
  abline(h = c(4,9), lty = "dotted")
  abline(v = c(2*P/N,3*P/N), lty = "dotted")

  plot(x = hii, y = cookD, xlab = "Hat (leverage) value", ylab = "Approx Cook's Distance",
   ylim = c(0, max(cookD, 1)), xlim = c(0, max(hii,3*P/N)))
  abline(h = c(4/N,1), lty = "dotted")
  abline(v = c(2*P/N,3*P/N), lty = "dotted")
 }

 # Listing of values to check
```

```
 # Create flags to identify high values in listing
 hflag <- ifelse(test = hii > 3*P/N, yes = "**", no =
          ifelse(test = hii > 2*P/N, yes = "*", no = ""))
 xflag <- ifelse(test = D.X2 > 9, yes = "**", no =
          ifelse(test = D.X2 > 4, yes = "*", no = ""))
 dflag <- ifelse(test = D.dev2 > 9, yes = "**", no =
          ifelse(test = D.dev2 > 4, yes = "*", no = ""))
 cflag <- ifelse(test = cookD > 1, yes = "**", no =
          ifelse(test = cookD > 4/N, yes = "*", no = ""))

 chk.hii2 <- which(hii > 3*P/N)
 chk.DX22 <- which(D.X2 > 9 | (D.X2 > 4 & hii > 2*P/N))
 chk.Ddev2 <- which(D.dev2 > 9 | (D.dev2 > 4 & hii > 2*P/N))
 chk.cook2 <- which(cookD > 4/N)

 all.meas <- data.frame(h = round(hii,2), hflag, Del.X2 = round(D.X2,2), xflag,
             Del.dev = round(D.dev2,2), dflag, Cooks.D = round(cookD,3), cflag)

 if(print.output == TRUE) {
  cat("Potentially influential observations by any measures","\n")
  print(all.meas[sort(unique(c(chk.hii2, chk.DX22, chk.Ddev2, chk.cook2))),])
  cat("\n","Data for potentially influential observations","\n")
  print(cbind(mod.fit$data, yhat = round(yhat, 3))[sort(unique(c(chk.hii2, chk.DX22, chk.Ddev2, chk.cook
 }

 data.frame(hat  =  hii, CD = cookD, delta.Xsq = D.X2, delta.D = D.dev2)
}

influence1 = glmInflDiag(mod1, print.output = FALSE, which.plots = 1:2)
```






```
# influence2 = glmInflDiag(mod2, print.output = FALSE, which.plots = 1:2)
```

25. Remove the influential crab from the analysis and repeat the steps in Question 23. Has this fixed the problems with the model? Are there any other problems with the model, and what could be done to solve these problems?

- In both (ordinal and categorical) models, the removal of the outlier mildly decreases the significance of Color and mildly increases the significance of Weight. The influence of the outlier is limited.
- Excluding observation 165 from the model estimation does not substantially change the residual deviance and deviance/df (from 3.285 to 3.275 = 547.09/167). The goodness-of-fit test gives a higher Pearson statistic than the original model (from 785 to 850) and a similar p-value of 0. Thus, the model is still not well fit to the data based on the heuristic of deviance/df > 1 and goodness-of-fit test.
- There are 4 additional points above the top line in the leverage plot (6, 31, 95, 127) that we could remove before restimating a model for better fit.
- Random sampling: From the Q-Q plot above in Problem 24, we can observe that the residual is not well normally distributed. From other residual plots, we observed a slight heteroscedasticity. We have no information about sampling in this study. But a better randomly selected sample and independently and identically distributed observations will help on a better model fitting.
- Interactions between parameter. It is possible that weight and color, or spine and color, have interactions. Though it is not studied here, the interaction term may be statistically significant and improve model fitting.

```
hs2 <- hs[-165,]

# compare mod1 and mod1.3(without the outlier)
mod1.3 <- glm(Sat~Color+Spine+Width+Weight, family = poisson(link="log"), data=hs2)
summary(mod1.3)
```

```
##
## Call:
```

```
## glm(formula = Sat ~ Color + Spine + Width + Weight, family = poisson(link = "log"),
##     data = hs2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1306  -1.8654  -0.5758   0.9632   4.9814
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.230599   0.961641  -0.240 0.810487
## Color       -0.171880   0.066418  -2.588 0.009657 **
## Spine        0.026684   0.056431   0.473 0.636306
## Width       -0.002688   0.050374  -0.053 0.957444
## Weight       0.674908   0.193091   3.495 0.000474 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 628.68  on 171  degrees of freedom
## Residual deviance: 547.09  on 167  degrees of freedom
## AIC: 908.58
##
## Number of Fisher Scoring iterations: 6
```

```
#Anova(mod1.3)
stargazer(mod1,mod1.3,type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                 Sat
##                           (1)            (2)
## -----------------------------------------------
## Color                  -0.185***      -0.172***
##                         (0.067)        (0.066)
##
## Spine                   0.040          0.027
##                         (0.057)        (0.056)
##
## Width                   0.027         -0.003
##                         (0.048)        (0.050)
##
## Weight                 0.473***       0.675***
##                         (0.165)        (0.193)
##
## Constant                -0.524         -0.231
##                         (0.949)        (0.962)
##
## -----------------------------------------------
## Observations             173            172
## Log Likelihood         -453.564       -449.288
## Akaike Inf. Crit.      917.128         908.575
## ===============================================
```

```
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

```
# compare mod2 and mod2.3(without the outlier)
mod2.3 <- glm(Sat ~ Color.cat + Spine.cat + Width + Weight, family = poisson(link="log"), data=hs2)
summary(mod2.3)
```

```
##
## Call:
## glm(formula = Sat ~ Color.cat + Spine.cat + Width + Weight, family = poisson(link = "log"),
##     data = hs2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1525  -1.8517  -0.5667   0.9336   4.9439
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.08024    0.97742  -0.082 0.934573
## Color.catlight      -0.22340    0.16853  -1.326 0.184979
## Color.catdark       -0.47060    0.19560  -2.406 0.016131 *
## Color.catvery dark -0.46914    0.22791  -2.058 0.039550 *
## Spine.cat1 good     -0.16432    0.21301  -0.771 0.440456
## Spine.cat0 good      0.05544    0.11901   0.466 0.641329
## Width               -0.01345    0.05130  -0.262 0.793207
## Weight               0.69752    0.19410   3.594 0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 628.68  on 171  degrees of freedom
## Residual deviance: 544.86  on 164  degrees of freedom
## AIC: 912.35
##
## Number of Fisher Scoring iterations: 6
```

```
#Anova(mod2.3)
stargazer(mod2,mod2.3,type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Sat
##                           (1)            (2)
## ------------------------------------------------
## Color.catlight          -0.265         -0.223
##                        (0.168)        (0.169)
##
## Color.catdark         -0.514***       -0.471**
##                        (0.195)        (0.196)
##
## Color.catvery dark     -0.531**       -0.469**
##                        (0.227)        (0.228)
##
```

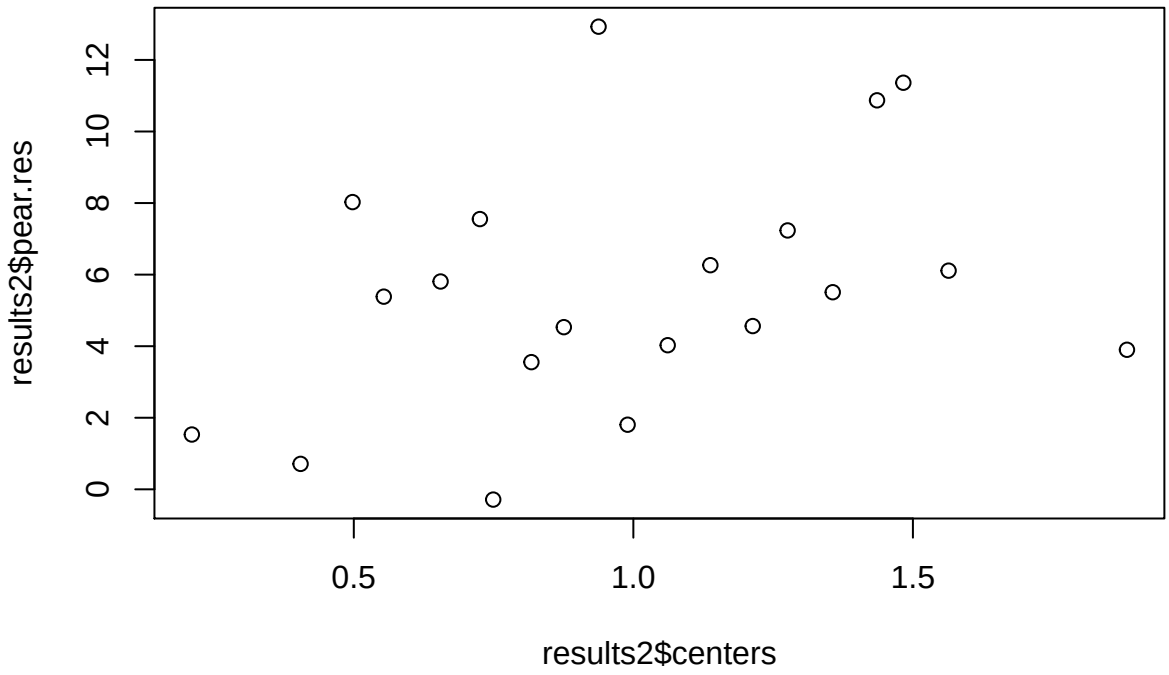
```
## Spine.cat1 good        -0.150        -0.164
##                       (0.214)       (0.213)
##
## Spine.cat0 good        0.087         0.055
##                       (0.120)       (0.119)
##
## Width                  0.017        -0.013
##                       (0.049)       (0.051)
##
## Weight                0.497***      0.698***
##                       (0.166)       (0.194)
##
## Constant               -0.357        -0.080
##                       (0.967)       (0.977)
##
## ------------------------------------------------
## Observations             173           172
## Log Likelihood        -452.431      -448.173
## Akaike Inf. Crit.     920.862       912.346
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# mod2 <- glm(Sat~Color+Spine+Width+Weight, family = poisson(link="log"), data=hs2)
# summary(mod2)
# Anova(mod2)
# stargazer(mod1,mod2,type="text")
#
preds2 <- predict(mod1.3, hs)
obs2 <- hs$Sat
results2 <- PostFitGOFTest(obs2,preds2)
```

```
## Warning in PostFitGOFTest(obs2, preds2): Some expected counts are less than
## 5. Use smaller number of groups
```

```
## Post-Fit Goodness-of-Fit test with 20 bins
##  Pearson Stat =  849.5179
##  p =  0
```

```
# results2
plot(results2$pear.res~results2$centers)
```


results2$pear.res
0
2
4
6
8
10
12
0.5
1.0
1.5
results2$centers