```
Log for R-bases Case-Study of Horseshoe Crab Analysis
======================================== 4/24/03

> Crabs <- read.table("Crabs", skip=1)
> names(Crabs) <- c("Color","Spine","Width","Satell","Wt")
> Crabs[,5] <- Crabs[,5]/1000
 Crabs[,1] <- Crabs[,1]-1
> Crabs[1:5,]
  Color Spine Width Satell   Wt
1     2     3  28.3      8 3.05
2     3     3  22.5      0 1.55
3     1     1  26.0      9 2.30
4     3     3  24.8      0 2.10
5     3     3  26.0      4 2.60
> attach(Crabs)
> table(Color)
Color
 1  2  3  4
12 95 44 22
> table(Spine)
Spine
  1   2   3
 37  15 121
> Crabs[,"Color"] <- factor(Color)
 Crabs[,"Spine"] <- factor(Spine)

> plot(Width, Satell, type="n", xlab="Width, cm",
  ylab="Number of satellites", main=
  "Plot of # Satell by Width and Color")
> for (i in 1:4) { inds <- (1:173)[Color==i]
    points(Width[inds],Satell[inds], pch=as.character(i)) }

> plot(Width, Satell, type="n", xlab="Width, cm",
  ylab="Number of satellites", main=
  "Plot of # Satell by Width and Spine")
  for (i in 1:3) { inds <- (1:173)[Spine==i]
    points(Width[inds],Satell[inds], pch=as.character(i)) }
```

```
### Not so clear from the pictures ...

### Now try grouping and look at mean versus width and
###   variance versus mean

> wdthgp <- split(Satell, cut(Width, breaks=seq(20,34,2)))
>  wdmat <- cbind(seq(21,33,2), unlist(lapply(wdthgp, mean)),
         unlist(lapply(wdthgp, var)), unlist(lapply(wdthgp,
         function(wid) mean(wid>0))))
        [,1]     [,2]      [,3]      [,4]
(20,22]   21 0.000000  0.000000 0.0000000
(22,24]   23 1.391304  6.612648 0.3478261
(24,26]   25 2.600000 10.006780 0.5500000
(26,28]   27 2.846154  5.975867 0.7115385
(28,30]   29 4.933333 15.374713 0.9000000
(30,32]   31 2.800000  0.700000 1.0000000
(32,34]   33 7.000000        NA 1.0000000

>  wtgp <- split(Satell, cut(Wt, breaks=
         c(0,.5,1,1.5,2,2.5,3,3.5,6)))
> cbind((3:8)/2, unlist(lapply(wtgp, mean)),
         unlist(lapply(wtgp, var)))
        [,1]     [,2]      [,3]
(1,1.5]  1.5 0.800000  3.200000
(1.5,2]  2.0 1.512195  6.056098
(2,2.5]  2.5 2.867925  9.847605
(2.5,3]  3.0 3.250000  9.866279
(3,3.5]  3.5 4.807692 11.601538
(3.5,6]  4.0 4.750000  2.916667

### By either criterion, we find variance is a large multiple
### of mean decreasing to less than mean for highest category

> anova(lm(Satell ~ . , data=Crabs))
Analysis of Variance Table

Response: Satell
           Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
Color       3   67.52   22.51  2.5659 0.0563798 .
Spine       2   19.25    9.63  1.0975 0.3361279
Width       1  134.12  134.12 15.2907 0.0001344 ***
Wt          1   36.68   36.68  4.1822 0.0424389 *
Residuals 165 1447.29    8.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>  anova(lm(Satell ~ Width + Wt + Color + Spine, data=Crabs))
Analysis of Variance Table

Response: Satell
           Df  Sum Sq Mean Sq F value    Pr(>F)
Width       1  196.96  196.96 22.4542 4.619e-06 ***
Wt          1   36.72   36.72  4.1867   0.04233 *
Color       3   20.67    6.89  0.7856   0.50352
Spine       2    3.23    1.61  0.1840   0.83214
Residuals 165 1447.29    8.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

### This suggests that Color and Spine are not such valuable
###    predictors

## Next try a model with Satell > 0
> table(Crabs[,"Satell"]>0)
FALSE  TRUE
   62   111

> Crabglm1 <- glm(I(Satell>0) ~ Width + Wt, family=binomial,
        data=Crabs)
> Crabglm1
Call:  glm(formula = I(Satell > 0) ~ Width + Wt, family =
        binomial, data = Crabs)
...
Coefficients:
(Intercept)        Width           Wt
    -9.3545       0.3068       0.8338
```

```
Degrees of Freedom: 172 Total (i.e. Null);  170 Residual
Null Deviance:      225.8
Residual Deviance: 192.9        AIC: 198.9
> anova(Crabglm1)
Analysis of Deviance Table,   Model: binomial, link: logit

Response: I(Satell > 0), Terms added sequentially

       Df Deviance Resid. Df Resid. Dev
NULL                     172    225.759
Width  1   31.306      171    194.453
Wt      1    1.561       170    192.892

> sum(111*log(111/173)+62*log(62/173))
                ### max loglik, NULL model
[1] -112.8793
### So saturated logLik = -112.88 + 0.5*(225.8) = 0,
###    and  Crabglm1  logLik = -192.9/2 = -96.4

> sum(log(Crabglm1$fit[Crabs[,"Satell"]>0])) +
    sum(log(1-Crabglm1$fit[ Crabs[,"Satell"]==0]))
[1] -96.44594

> plot(Crabglm1$fitted,as.numeric(Crabs[,"Satell"]>0)-
    Crabglm1$fitted,type="n", xlab="Fitted prob", ylab=
    "Residual", main="Residuals Plot by Color")
> for(i in 1:4) { inds <- Crabs[,"Color"]==i
    points(Crabglm1$fitted[inds], (as.numeric(Crabs[,
      "Satell"]>0)-Crabglm1$fitted)[inds],
      pch=as.character(i)) } ### Fairly useless picture

> plot(Crabs$Color, as.numeric(Crabs[,"Satell"]>0)-
      Crabglm1$fitted, xlab="Color", ylab="Residuals",
      main="Boxplot of Residuals, by Color")
### Saved as "ResidColr.ps" : shows Color=4 may be different !
> plot(Crabs$Spine, as.numeric(Crabs[,"Satell"]>0)-
      Crabglm1$fitted, xlab="Spine", ylab="Residuals", main=
```

```
      "Boxplot of Residuals, by Spine")
### Saved as "ResidSpine.ps" : Spine=2 looks different,
###   maybe not significant !

> plot(Crabs$Width, as.numeric(Crabs[,"Satell"]>0)-
      Crabglm1$fitted, xlab="Spine", ylab="Residuals", main=
      "Boxplot of Residuals, by Width")

> anova(glm(I(Satell>0) ~ Width + Wt + I(Color==4),
        family=binomial, data=Crabs))
               Df Deviance Resid. Df Resid. Dev
NULL                              172    225.759
Width           1   31.306       171    194.453
Wt              1    1.561       170    192.892
I(Color == 4)   1    6.197       169    186.695

> tmpglm <- glm(I(Satell>0) ~ Width + I(Color==4), family=
    binomial(link=probit), data=Crabs)
...
Residual Deviance: 187.7       ### hardly any different

### Overall, MANY possible functions could fit the plot
>    plot(wdmat[,1], wdmat[,4], xlab="Width", ylab=
           "Prob(Satell > 0)", main=
          "Frac Pos Satell vs Width, Crab Data")
>  tmpglm <- glm(I(Satell>0) ~ Width , family=
          binomial(link=probit), data=Crabs)
> tmpglm$coef
(Intercept)       Width
 -7.4999984   0.3019387
> glm(I(Satell>0) ~ Width , family=binomial, data=Crabs)$coef
(Intercept)       Width
 -12.350677    0.497225
> lines(wdmat[,1], plogis(-12.3507+0.497225*wdmat[,1]), lty=2)
> lines(wdmat[,1], pnorm(-7.50+0.301939*wdmat[,1]), lty=5)
> legend(locator(), legend=c("Logit link fit","Probit link fit"),
     lty=c(2,5))  ### Saved as Linkfits.ps
```

```
> anova(glm(I(Satell>0) ~ Width + Wt + I(Spine==2),
        family=binomial, data=Crabs))
               Df Deviance Resid. Df Resid. Dev
NULL                              172    225.759
Width           1   31.306       171    194.453
Wt              1    1.561       170    192.892
I(Spine == 2)   1    0.066       169    192.826

### Best available model seems to be: omit Wt, include Color=4

> Crabglm1B <- glm(I(Satell>0) ~ Width + I(Color==4),
        family=binomial, data=Crabs)
NULL                              172    225.759
Width           1   31.306       171    194.453
I(Color == 4)   1    6.495       170    187.958

### But there was additional structure in Number of Satellites.

> table(Crabs$Satell)
 0  1  2  3  4  5  6  7  8  9 10 11 12 14 15
62 16  9 19 19 15 13  4  6  3  3  1  1  1  1

> Crabglm2 <- glm(Satell ~ Width + Wt + I(Color==4), family=
       poisson, data=Crabs)
> summary(Crabglm2)$coef
                 Estimate Std. Error    z value   Pr(>|z|)
(Intercept)   -1.16271018 0.90881936 -1.2793634 0.20076913
Width          0.04207901 0.04708027  0.8937716 0.37144411
Wt             0.44464802 0.15944996  2.7886367 0.00529304
I(Color == 4) -0.18707508 0.15872500 -1.1786113 0.23855299
> anova(Crabglm2)
Analysis of Deviance Table,       Model: poisson, link: log

Response: Satell

               Df Deviance Resid. Df Resid. Dev
NULL                              172     632.79
```

```
Width           1    64.91       171     567.88
Wt              1     7.98       170     559.90
I(Color == 4)   1     1.45       169     558.45

### Both Width and Wt  seem worth retaining in model.
### Next try residual plots ...

> plot (Crabs$Width, Crabs$Satell - Crabglm2$fit)
> plot (Crabs$Wt, Crabs$Satell - Crabglm2$fit)
> plot(Crabs$Color, Crabs$Satell - Crabglm2$fit)

> Crabglm2B <- glm(Satell ~ Wt + Width + Wt:Width , family=
      poisson, data=Crabs)
> Crabglm2B$dev
[1] 545.29   ### quite a bit better than before
> anova(Crabglm2B)
NULL                         172     632.79
Wt          1    71.93       171     560.87
Width       1     0.97       170     559.90
Wt:Width    1    14.61       169     545.29
> Crabglm2B$coef
(Intercept)          Wt       Width    Wt:Width
-6.91097413  3.06646729  0.23034746 -0.08660654
> plot(Crabs$Width, Crabs$Satell - Crabglm2B$fit)
> plot(Crabs$Wt, Crabs$Satell - Crabglm2B$fit)
       ### Better than before
> sum((Crabs$Satell - Crabglm2B$fit)^2/Crabglm2B$fit)/171
[1] 3.11824     #### Scale parameter !! High overdispersion !!
### Now try to visualize alternative links ?!
> plot(Crabs$Wt, Crabs$Satell)
> points(Crabs$Wt, Crabglm2B$fitted, pch=3)
> points(Crabs$Wt,  glm(Satell ~ Wt + Width + Wt:Width , family=
      poisson(link=sqrt), data=Crabs)$fitted, pch=6)
> par(mfrow=c(2,1))
   plot(Crabs$Wt, Crabs$Satell)
   points(Crabs$Wt, Crabglm2B$fitted, pch=3)
   par(mfrow=c(2,1))
   plot(Crabs$Wt, Crabs$Satell)
```

```
   points(Crabs$Wt,  glm(Satell ~ Wt + Width + Wt:Width , family=
      poisson(link=sqrt), data=Crabs)$fitted, pch=6)

> Crabglm2C <-  glm(Satell ~ Wt + Width + Wt:Width , family=
+       poisson(link=sqrt), data=Crabs)
> c( Crabglm2C$dev, Crabglm2B$dev)
[1] 543.4073   545.29
> sum((Crabs$Satell - Crabglm2C$fit)^2/Crabglm2C$fit)/171
[1] 3.114793

### Could also perform critical tests of Poisson assumption by
###  calculating expected numbers of obs with number of Satell
###  equal to 1, 2, ...
> c(sum(Crabglm2C$fit exp(-Crabglm2C$fit)), sum(Crabs$Satell==1))
[1] 31.08952 16.00000

### But we knew this assumption would fail because of the great
overdispersion !!

### Other goodness-of-fit approaches involve GROUPING
    (chi-sq tests) and LR test (eg for quadratic and
    interaction terms in the fit).
```

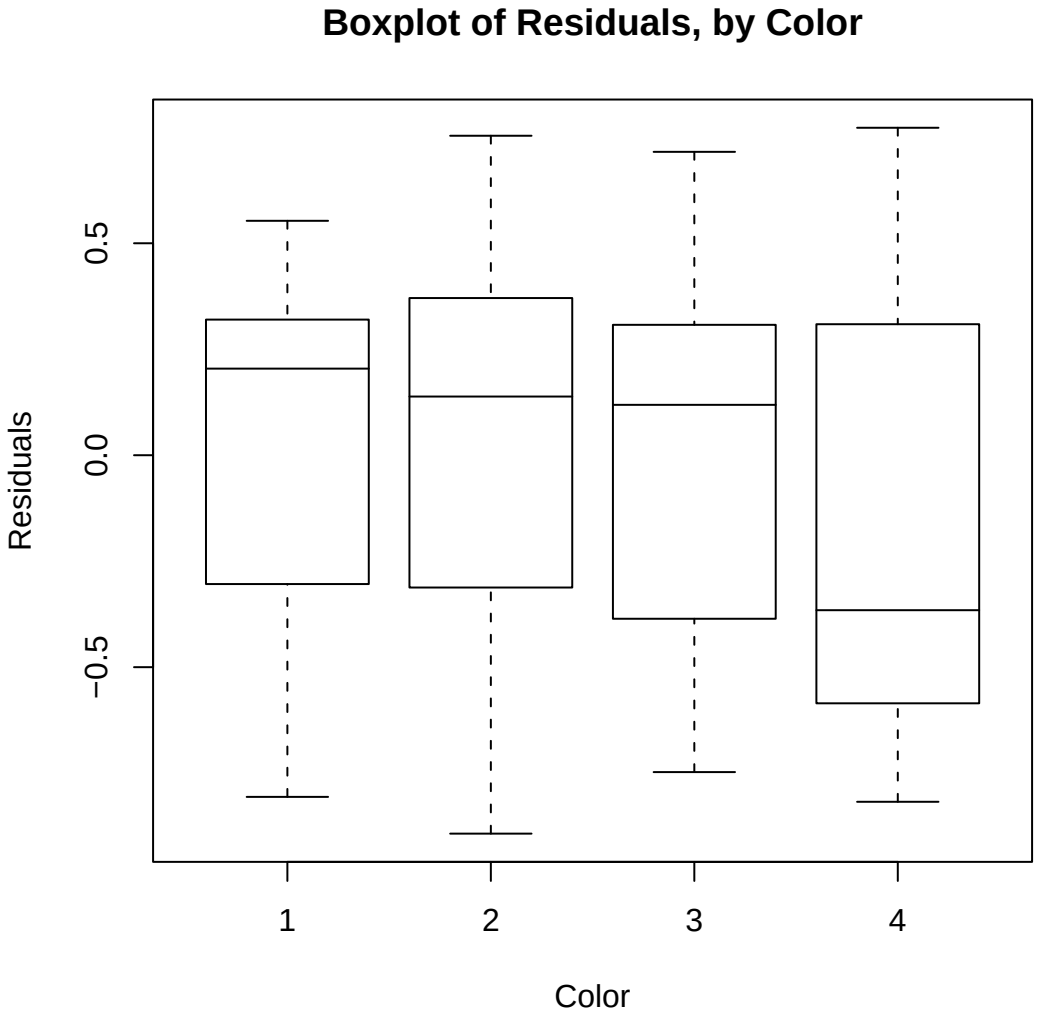

Figure 1: R-Language boxplot showing residuals from logistic-regression fit to Crab Data versus Color category. Note that the residual median and distribution appear quite different for Color-category 4 (the darkest) than for the other colors.

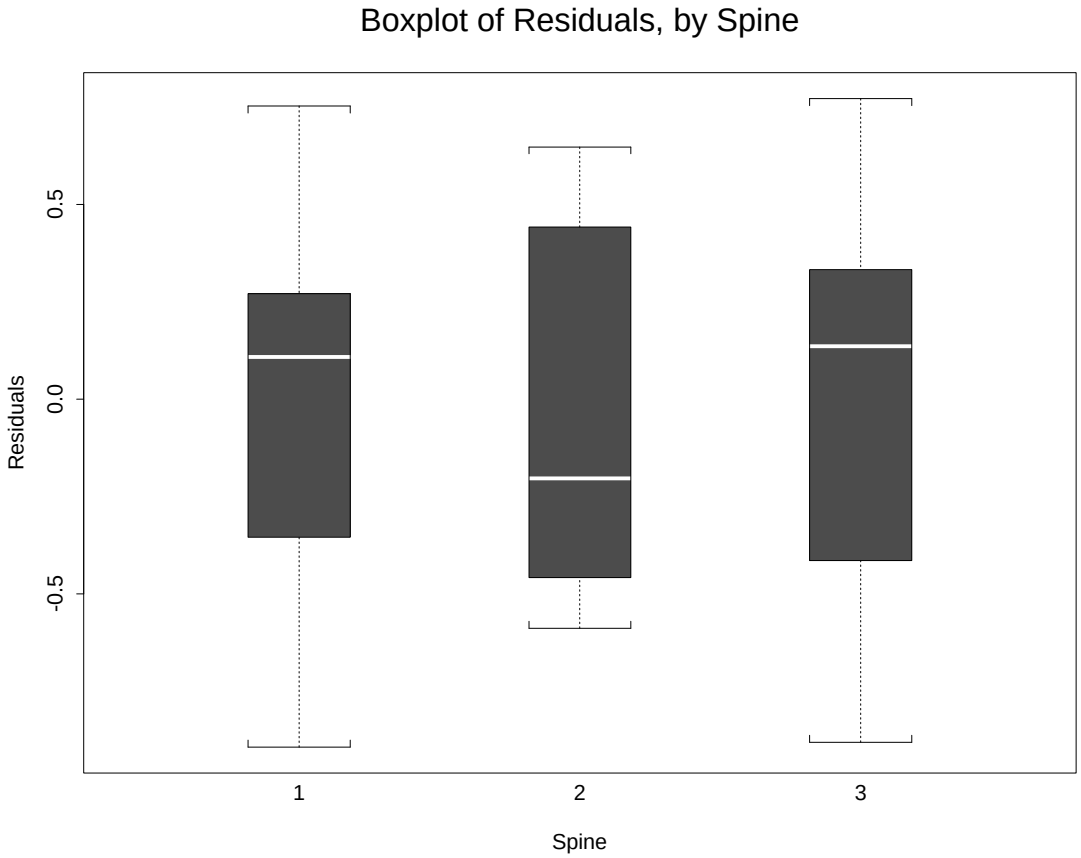

Figure 2: Splus3.4 boxplot showing residuals from logistic-regression fit to Crab Data versus Spine category. Here the Spine=2 category has residual distribution *looking* a little different from the other two, but as one might guess abd as it turns out, not significantly so.

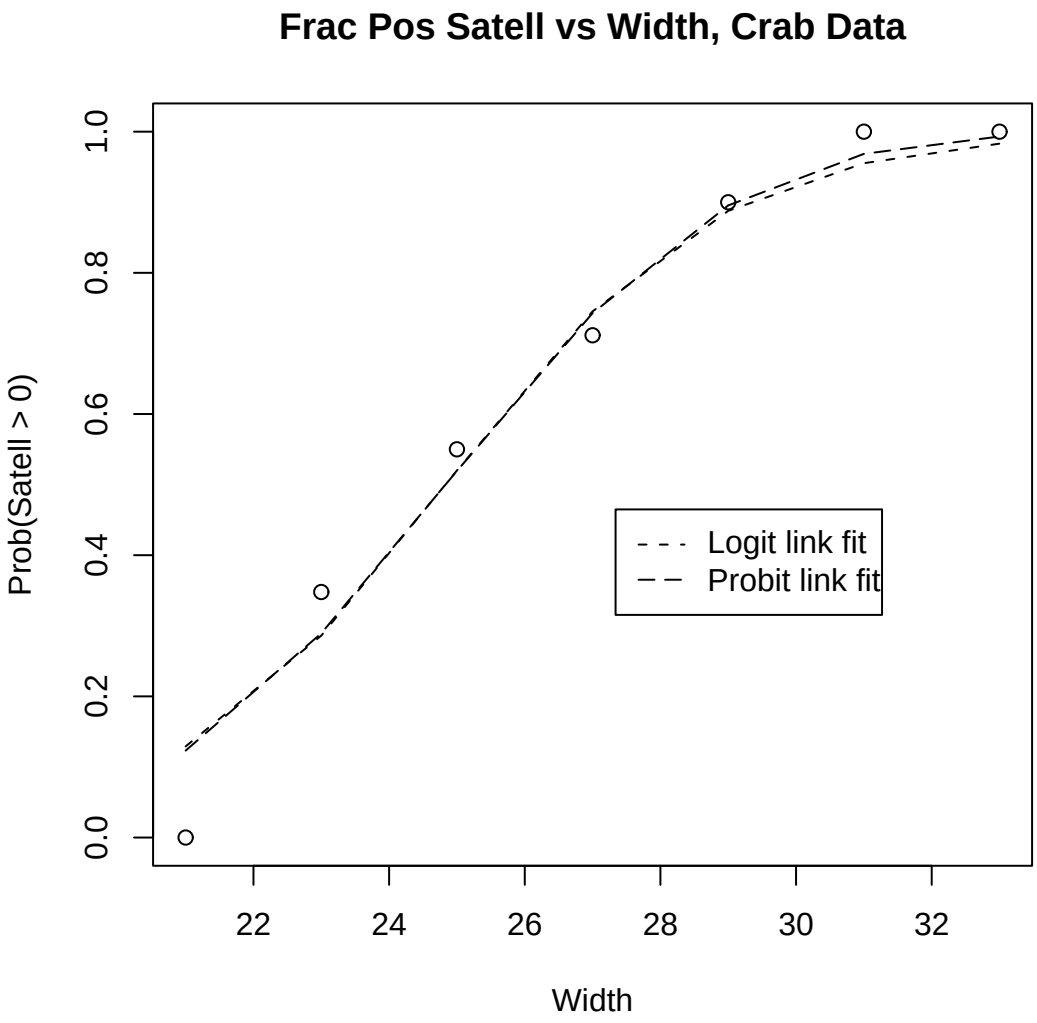

Figure 3: R-Language plot showing mean proportion of (female) crabs with positive number of satellites (males) for 7 ordered categories by Width (of shell or carapace) along with fitted means for midpt of category-interval in logistic and probit regressions with Width as only predictor.