Question 3

Question 3. R excercise.

Load the data set crabs available on MyCourses under Content -> Datasets, viz.This data set contains the number of satelites (males attached to the female's nest) for 173 female horseshoe crabs. The weight (in g), carapace width (in cm), spine condition and color of the crab are also recorded; weight and width are continuous whereas color (1=medium light, 2=medium, 3=medium dark, 4=dark) and spine condition (1= both good, 2 = one worn or broken, 3 = both worn or broken) are categorical. It is of interest to understand how the width, weight, color and spine condition affect the number of satelites.

```
crabs <- read.table("crabs.txt",header=TRUE)
attach(crabs)
head(crabs)
```

```
##   color spine width satell weight
## 1     3     3  28.3      8   3050
## 2     4     3  22.5      0   1550
## 3     2     1  26.0      9   2300
## 4     4     3  24.8      0   2100
## 5     4     3  26.0      4   2600
## 6     3     3  23.8      0   2100
```

```
library(faraway)
```

(a). Fit a linear regression model to the data with width and spine condition as main effects, using lm. Plot the data along with the fitted regression line(s) and interpret the model.

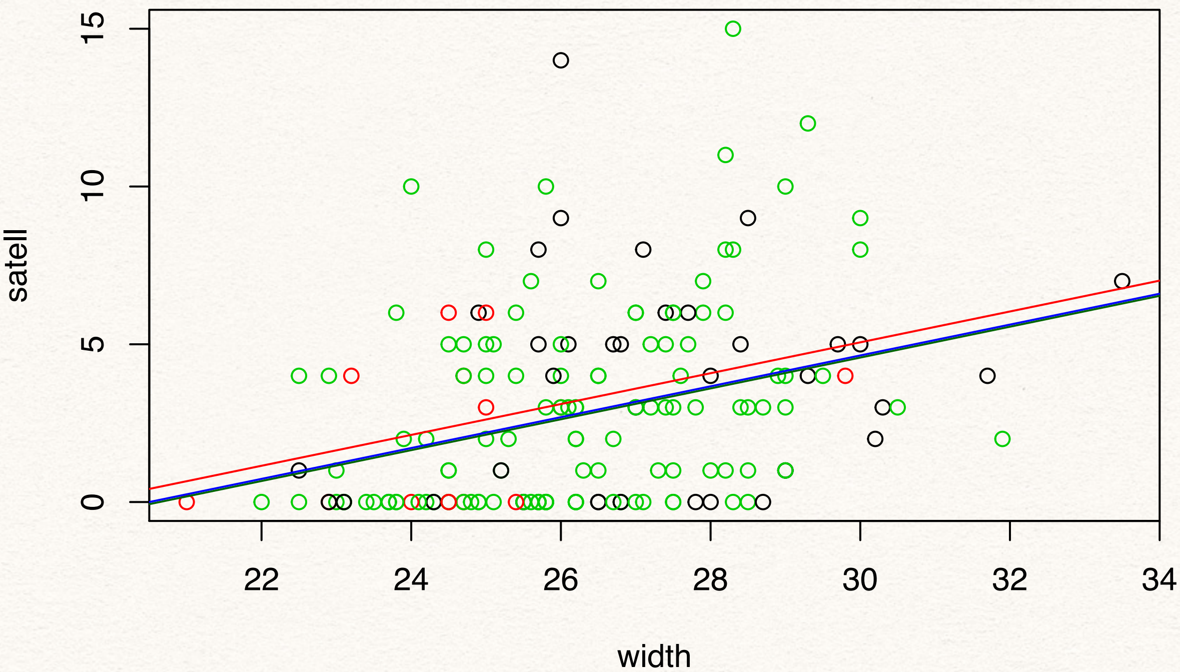
Fit the linear regression model using lm

```
width<-crabs$width
spine<-crabs$spine
# creasting the factor variable
satell<-crabs$satell
fit.crabs<-lm(satell~factor(spine)+width)
summary(fit.crabs)
```

```
##
## Call:
## lm(formula = satell ~ factor(spine) + width)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.426 -2.250 -0.734  1.846 11.185
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -9.6081     3.0878  -3.112  0.00218 **
## factor(spine)2  -0.4828     0.9514  -0.508  0.61245
## factor(spine)3  -0.4156     0.5686  -0.731  0.46586
## width            0.4890     0.1125   4.348 2.36e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.982 on 169 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.1028
## F-statistic:  7.57 on 3 and 169 DF,  p-value: 8.816e-05
```

```
#plot the data
plot(width,satell,col=factor(spine))
# Line for width =1
# factor value of spine(1,2,3) just specifying the category so it will be decribed by indicator
# X2=0,X3=0 represents spine=1
# a represents the intercept and b represent the slope
abline(a=coef(fit.crabs)[1],b=coef(fit.crabs)[4],col="red")
#line for width=2
# X2=1,X3=0 represents spine=2
abline(a=coef(fit.crabs)[1]+coef(fit.crabs)[2],b=coef(fit.crabs)[4],col="darkgreen")
#line for width=3
# X2=0,X3=1 represent spine=3
abline(a=coef(fit.crabs)[1]+coef(fit.crabs)[3],b=coef(fit.crabs)[4],col="blue")
```



The above figures show that a better spins condition can cause a larger number of satelites. According to the line trend, we can conclude that higher width can can cause larger number of satellites.

(b). Fit the same model as in part (b), but using glm with family = gaussian. If you leave the link unspecified, what link is being used? Using the function summary, compare the estimated coefficients and standard errors to those obtained in part (a). What do you think the quantity given in the output line (Dispersion parameter for gaussian family taken to be . . .) estimates? How could you alternatively estimate the same quantity from the output in part (a)?

Fit the model in part(a) with unspecified link, then the canonical link is being used.

```r
gaussianmod<-glm(satell~factor(spine)+width,family=gaussian)
summary(gaussianmod)
```

```
##
## Call:
## glm(formula = satell ~ factor(spine) + width, family = gaussian)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.426  -2.250  -0.734   1.846  11.185
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -9.6081     3.0878  -3.112  0.00218 **
## factor(spine)2  -0.4828     0.9514  -0.508  0.61245
## factor(spine)3  -0.4156     0.5686  -0.731  0.46586
## width            0.4890     0.1125   4.348 2.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.892932)
##
##     Null deviance: 1704.9  on 172  degrees of freedom
## Residual deviance: 1502.9  on 169  degrees of freedom
## AIC: 874.96
##
## Number of Fisher Scoring iterations: 2
```

The coefficient and standard error obtained in part(a) are the same as thoese in part(b). As the the disperssion parameter ϕ is taken to be 8.892932 in part(b) and $\phi = \sigma^2$ in Gaussian family. The estimated value of σ is equal to 2.982 in part(a), so both of these 2 parts gain the same value for the standard errors and dispersion paramter.

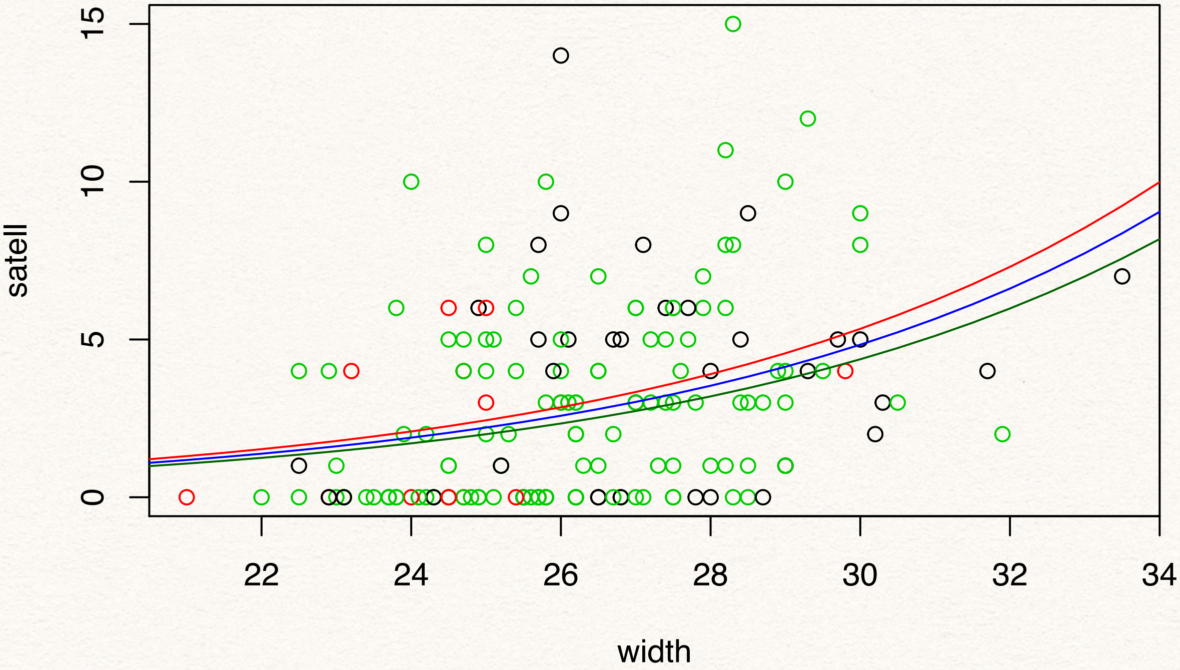
(c). Fit a Poisson glm to the data, again with width and spine condition as main effects, using the canonical link. Plot the data along with the fitted regression curve(s) and compare the plot to the one in part (a).

```r
# Fit the model with poisson family
poissonmod<-glm(satell~factor(spine)+width,family=poisson)
summary(poissonmod)
```

```
##
## Call:
## glm(formula = satell ~ factor(spine) + width, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9509  -1.9740  -0.4963   1.0832   4.7173
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.02570    0.59421  -5.092 3.54e-07 ***
## factor(spine)2 -0.19932    0.20983  -0.950    0.342
```

```
## factor(spine)3 -0.09899    0.10490  -0.944    0.345
## width           0.15668    0.02096   7.476 7.69e-14 ***
## ---
## 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: 566.60  on 169  degrees of freedom
## AIC: 929.9
##
## Number of Fisher Scoring iterations: 6
```

```r
# plot both graph
plot(width,satell,col=factor(spine))
# g(x)=e^x
x<- seq(from=20,to=34,by=0.5)
lines(x,exp(coef(poissonmod)[1]+coef(poissonmod)[4]*x),col="red")
lines(x,exp(coef(poissonmod)[1]+coef(poissonmod)[2]+coef(poissonmod)[4]*x),col="darkgreen")
lines(x,exp(coef(poissonmod)[1]+coef(poissonmod)[3]+coef(poissonmod)[4]*x),col="blue")
```



Comparing to the graph in part(a), they both have the increasing trend as width raising where part(a) are growing linearly while lines in this grapth are growing exponentially. Better spines condition can still have a larger number of satell.

(d). Calculate the estimated number of satellites for a female crab with carapace width 28cm, and one worn or broken spine using (i) the linear model in part (a) and (ii) the Poisson glm in part (c).

Since the female crab has one worn or broken spin, then the spin condition is 2.

```r
# Calculate the linar model
coef(fit.crabs)[1]+coef(fit.crabs)[2]+coef(fit.crabs)[4]*28
```

```
## (Intercept)
##    3.600611
```

```r
# Calculate the poisson GLM model
exp(coef(poissonmod)[1]+coef(poissonmod)[2]+coef(poissonmod)[4]*28)
```

```
## (Intercept)
##     3.19653
```

The above result show that the estimated value for Linear model in part(a) will be $3.601 \approx 4$ satellites and the estimated result for poisson glm model will be $3.197 \approx 3$ satellites.

(e). Fit a glm to the data with only spine condition as a predictor,and (i) family=gaussian and (ii) family =poisson(link=identity). Compare the estimated coefficients, and explain why they are the same (you need to provide a proof by carrying out appropriate calculation(s)). Compare the estimated standard errors. Why do you think they differ (an explanation in words suffices)? Which standard errors do you think are more trustworthy?

```r
canomod <- glm(satell ~ factor(spine),family=gaussian)
summary(canomod)
```

```
##
## Call:
## glm(formula = satell ~ factor(spine), family = gaussian)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6486  -2.8099  -0.8099   2.0000  12.1901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.6486     0.5154   7.079 3.66e-11 ***
## factor(spine)2  -1.6486     0.9597  -1.718   0.0876 .
## factor(spine)3  -0.8387     0.5890  -1.424   0.1563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.829768)
##
##     Null deviance: 1704.9  on 172  degrees of freedom
## Residual deviance: 1671.1  on 170  degrees of freedom
## AIC: 891.3
##
## Number of Fisher Scoring iterations: 2
```

```r
identmod<- glm(satell ~ factor(spine),family=poisson(link=identity))
summary(identmod)
```

```
##
## Call:
## glm(formula = satell ~ factor(spine), family = poisson(link = identity))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7014  -2.3706  -0.5097   1.1252   5.0859
```

```
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.6486     0.3140  11.619  < 2e-16 ***
## factor(spine)2  -1.6486     0.4816  -3.423 0.000619 ***
## factor(spine)3  -0.8387     0.3490  -2.403 0.016265 *
## ---
## 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: 621.16  on 170  degrees of freedom
## AIC: 982.46
##
## Number of Fisher Scoring iterations: 3
```

Proofing the same estimated coefficients:
For(i) family = Gaussian that is $g^{-1}(x) = b^{'}(x) = x$:

$$g(E[Y|X]) = \beta X$$

$$b^{'-1}(E[Y|X]) = \beta X$$

$$E[Y|X] = \beta X$$

For(ii) family =poisson(link=identity) that is $g(x) = x$:

$$g(E[Y|X]) = \beta X$$

$$E[Y|X] = \beta X$$

Also β can be calculated through the score function. First consider the Gaussian with identity link:

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{y_i - b'(\theta)}{a(\phi) b''(\theta)} * \frac{1}{b'(\theta_i)} * x_{ij}$$

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma^2} * \frac{1}{\mu_i} * x_{ij}$$

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij}$$

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} X_i \beta_1 x_{ij}$$

$$X^T Y = X^T X \beta_1$$

$$\beta_1 = (X^T X)^{-1} X^T Y$$

Consider the poisson with identity link, then $g'(x) = 1$:

$$\frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{y_i - b'(\theta)}{a(\phi) b''(\theta)} * x_{ij}$$

$$\sum_{i=1}^{n} \frac{y_i - \lambda_i}{\lambda_i} * x_{ij} = 0$$

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \lambda_i x_{ij}$$

$$\beta_2 = (X^T X)^{-1} X^T Y$$

Then both of these 2 model will give the same estimation on β. Since both models are fitting the data by $E[Y|X] = X\beta$, so they obatain the same estimated coefficients.
For (i), we have that the dispersion parameter $\phi = \sigma^2 = 9.830$ then the estimated standard error is 3.135 while (ii)'s dispersion parameter is 1 and $var[Y|X] = b''(\theta)a(\phi) = e^{\theta} = E[Y|X]$. Since they have different distriburion , so they have different estimated standard error.I think the standard error of (ii) is more trustworthy.