Horseshoe crab example:
There are 173 female crabs for which we wish to model the presence or absence of male "satellites" dependant upon characteristics of the female horseshoe crabs.

$$Y_i = \begin{cases} 1 & \text{satellite present} \\ 0 & \text{otherwise} \end{cases}$$

Explanatory variables are: weight, width of shell, color (medium light, medium, medium dark, dark), and condition of spine.

```
> library(splines)
> library(gam)
> crabs=read.table("http://www.stat.ufl.edu/~dathien/STA6505/crabdata.txt", header=TRUE)
> attach(crabs)
> crabs[1:5,]
```

```
  color spine width satellite 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
```

```
> y=ifelse(satellite>0, 1, 0) # Y = a binary indicator of satellites
> weight=weight/1000 # weight in kilograms rather than grams
```

Let us fist start with a simple model where we model the presence of satellites based upon "weight" in kg.

```
> fit=glm(y ~ weight, family=binomial(link=logit))
> summary(fit)
```

```
Call:
glm(formula = y ~ weight, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1108  -1.0749   0.5426   0.9122   1.6285

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4
```

The maximum likelihood fit is then $\text{logit}[\hat{\pi}(x)] = -3.6947 + 1.8151x$. Note that β is positive, implying that $\hat{\pi}(x) \uparrow$ as $x \uparrow$.

$$\hat{\pi}(x) = \frac{\exp(-3.6947 + 1.8151x)}{1 + \exp(-3.6947 + 1.8151x)}$$

At $x = \bar{x} = 2.44, \hat{\pi}(2.44) = 0.676$. Also, the rate of change at $x = 2.44$ is $\hat{\beta}\hat{\pi}(1-\hat{\pi}) = 1.8151(0.676)(0.324) = 0.398$. Consequently, the estimated change in π per 0.1kg increase is about 0.0398. Also, for a 0.1 kg increase in weight, the estimated odds of the presence of a satellite are multiplied by $\exp(0.1(1.8151)) = 1.2$, i.e. the odds increase by 20%.

The Wald statistic of $z = 4.819$ (or $z^2 = 23.2$) provides strong evidence that $H_0 : \beta = 0$ can be rejected. Similarly, the likelihood ratio test (which is the change in deviances between the model and the null model) $G^2 = 225.76 - 195.74 = 30.02$ compared to a χ^2_1 provides extremely strong evidence that $\beta \neq 0$.

```
> beta_h=coef(fit)[2]
> se.beta_h=sqrt(vcov(fit)[2,2])
> beta_h+c(-1.96,1.96)*se.beta_h
```

```
[1] 1.076821 2.553469
```

```
> exp(0.1*(beta_h+c(-1.96,1.96)*se.beta_h))
```

```
[1] 1.113694 1.290909
```

The 95% Wald CI for β is $1.8151 \pm 1.96(0.3767) \Rightarrow (1.08, 2.55)$ and therefore for the odds ratio per 0.1 kg increase in x is $(\exp(0.1(1.08)), \exp(0.1(2.55)))$ or $(1.11, 1.29)$. Similarly,

```
> confint(fit,"weight")
```

```
   2.5 %   97.5 %
1.113790 2.597305
```

is the likelihood-ratio CI.

Now that the model has been found to be significant, we can use it to estimate
$P(Y = 1|x = 2.4) = \pi(2.4)$.

```
> # number of females with 2.4kg weight (6)
> length(which(crabs$weight==2400))
```

```
[1] 6
```

```
> # number of those 6 females with satellites (4)#$
> crabs$satellite[which(crabs$weight==2400)]
```

```
[1] 0 3 1 5 5 0
```

There are 6 female crabs with a weight of 2.4 kg (or 2400 g), of whom only 4 have at least one satellite. Therefore, a naive estimate is 4/6 with 95% CI (without the use of the model)

```
> 4/6+c(-1.96,1.96)*sqrt((4/6*(1-4/6))/6)
```

```
[1] 0.2894645 1.0438688
```

However, using the model we construct a 95% CI for $\text{logit}[\hat{\pi}(2.4)]$ and by Equation (**??**) the CI for $\hat{\pi}(2.4)$

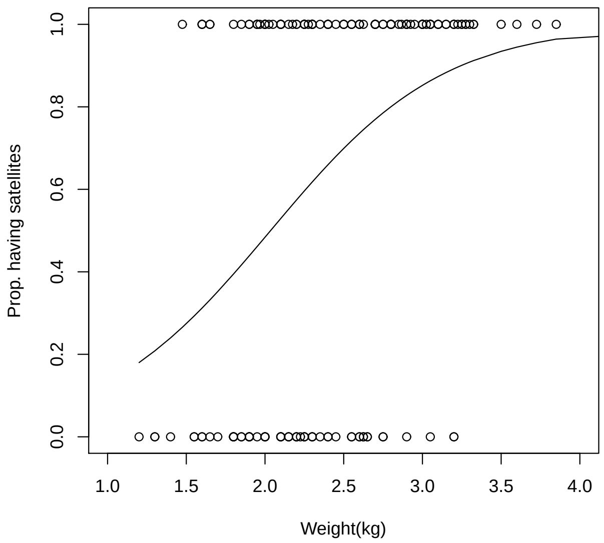
```
> eta=predict(fit,newdata=data.frame(weight=2.4),type="link",se.fit=TRUE)
> eta
```

```
$fit
        1
0.6616206

$se.fit
[1] 0.1780615

$residual.scale
[1] 1
```

```
> sqrt(vcov(fit)[1,1]+2.4^2*vcov(fit)[2,2]+2*2.4*vcov(fit)[1,2])

[1] 0.1780615

> eta.ci=eta$fit+c(-1,1)*qnorm(0.975)*eta$se.fit
> eta.ci # This is (l,u) interval

[1] 0.3126265 1.0106148

> plogis(eta.ci) # This is (exp(l)/(1+exp(l)),exp(u)/(1+exp(u)))

[1] 0.5775262 0.7331404
```

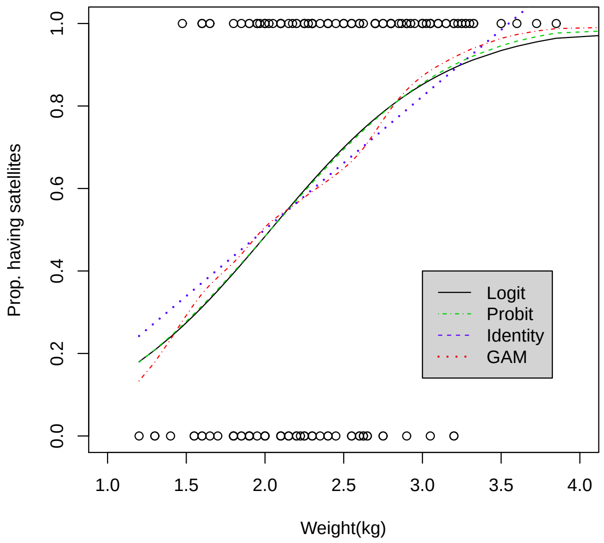
The Delta method standard error of the model based $\hat{\pi}(2.4) = 0.66$ is 0.04. The logistic regression model is

```
> plot(c(1,4),c(0,1),type="n",xlab="Weight(kg)",ylab="Prop. having satellites")
> ind=order(weight)
> lines(weight[ind], fit$fitted.values[ind],type="l",lty=1)
> lines(y ~ weight,type="p")
```



```
> #I2
> gam.fit=gam(y ~ s(weight), family=binomial(link=logit))
> lines(weight[ind], gam.fit$fitted.values[ind],type="l",lty=4,col=2)
> #I3
> fit.probit=glm(y ~ weight, family=binomial(link=probit))
> summary(fit.probit)

Call:
glm(formula = y ~ weight, family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1436  -1.0774   0.5336   0.9196   1.6216
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2383     0.5116  -4.375 1.22e-05 ***
weight        1.0990     0.2151   5.108 3.25e-07 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.46  on 171  degrees of freedom
AIC: 199.46

Number of Fisher Scoring iterations: 4

> lines(weight[ind], fit.probit$fitted.values[ind],type="l",lty=2,col=3)
> #I4
> fit.linear=glm(y ~ weight, family=gaussian())
> #summary(fit.linear)
> lines(weight[ind], fit.linear$fitted.values[ind],type="l",lty=3,lwd=2,col=4)
> legend(3,0.4,c("Logit","Probit","Identity","GAM"),col=c(1,3,4,2),lty=c(1,4,2,3),
+ lwd=c(1,1,1,2), bg = "light gray")
```



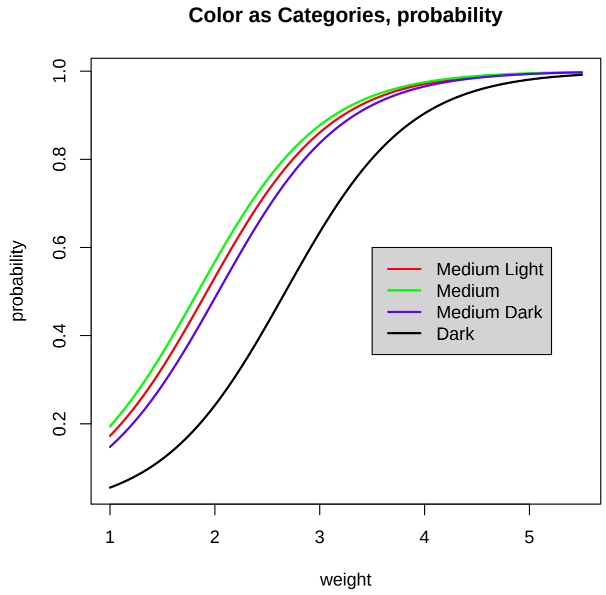
Next we introduce the color variable into the model by creating 3 indicator variables for the 4 levels of color. Let,

$$c_1 = \begin{cases} 1 & \text{medium light} \\ 0 & \text{o/w} \end{cases} \quad c_2 = \begin{cases} 1 & \text{medium} \\ 0 & \text{o/w} \end{cases} \quad c_3 = \begin{cases} 1 & \text{medium dark} \\ 0 & \text{o/w} \end{cases}$$

and hence $c_1 = c_2 = c_3 = c_4 = 0$ indicate whether a female crab is dark (i.e. base group). The model is then

$$\text{logit}[\hat{\pi}(x)] = \alpha + \beta_1 x + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_3$$

```
> color=color - 1 # color now takes values 1,2,3,4
> color=factor(color) # treat color as a factor
> fit2=glm(y ~ weight + color, family=binomial(link=logit),
+ contrasts=list(color=contr.treatment(4,base=4,contrasts=TRUE)))
> summary(fit2)

Call:
glm(formula = y ~ weight + color, family = binomial(link = logit),
    contrasts = list(color = contr.treatment(4, base = 4, contrasts = TRUE)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1908  -1.0144   0.5101   0.8683   2.0751

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5266     1.0038  -4.510 6.50e-06 ***
weight        1.6928     0.3888   4.354 1.34e-05 ***
color1        1.2694     0.8488   1.495  0.13479
color2        1.4143     0.5449   2.595  0.00945 **
color3        1.0833     0.5884   1.841  0.06561 .
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 188.54  on 168  degrees of freedom
AIC: 198.54

Number of Fisher Scoring iterations: 4

> ## Transformed for plot of probabilities
> cols=rainbow(3)
> curve(plogis(fit2$coefficients[1]+fit2$coefficients[2]*x), from=1,to=5.5,lwd=2,
+       ylab="probability",xlab="weight",main="Color as Categories, probability")
> for (i in 1:3)
+     curve(plogis(fit2$coefficients[1]+fit2$coefficients[2+i]+fit2$coefficients[2]*x),
+            from=1,to=5.5,lwd=2,col=cols[i],add=TRUE)
> legend(3.5,.6,col=c(cols,"black"),lwd=2,
+       legend=c("Medium Light","Medium","Medium Dark","Dark"),bg="light gray")
```



At $x = \bar{x} = 2.44$,

$$\hat{\pi} = \begin{cases} 0.704 & \text{medium light} \\ 0.401 & \text{dark} \end{cases}$$

The estimated odds ratio comparing these colors is $\exp 1.27 = 3.6$ at any fixed level of weight. This is equivalent to $(0.704/0.296)/(0.401/0.599)$.

To test the significance of color, controlling for weight we must test $\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$. The likelihood-ratio statistic is

$$\begin{aligned} G^2 &= -2(L_0 - L_1) \\ &= D_0 - D_1 \\ &= 195.7 - 188.5 = 7.2 \end{aligned}$$

which when compared to a χ^2_3 produces a p-value of 0.07.

Let us use color as a continuous variable

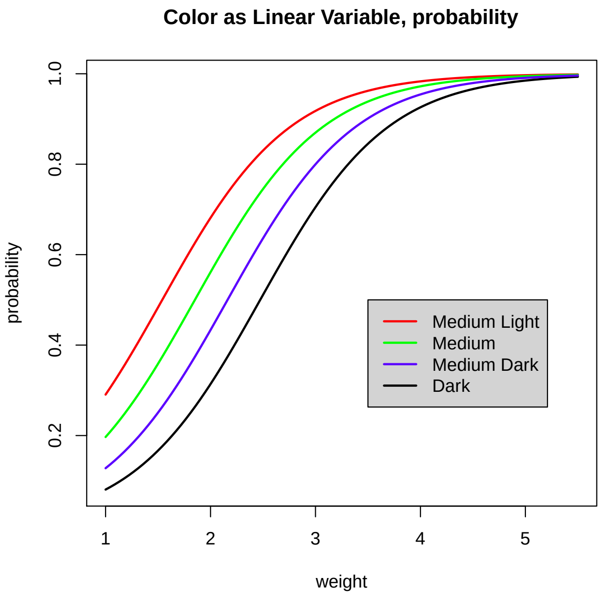
```
> linear=unclass(color)  #  convert back to integer levels
> fit2.1=glm(y ~ weight + linear, family=binomial(link=logit))
> summary(fit2.1)

Call:
glm(formula = y ~ weight + linear, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1596  -0.9998   0.5237   0.8825   1.9109

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0316     1.1161  -1.820   0.0687 .
weight        1.6531     0.3825   4.322 1.55e-05 ***
linear       -0.5142     0.2234  -2.302   0.0213 *
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 190.27  on 170  degrees of freedom
AIC: 196.27

Number of Fisher Scoring iterations: 4
```

```
> # TEST WHETHER COLOR IS SIGNIFICANT
> 1-pchisq(fit$deviance-fit2.1$deviance,fit$df.residual-fit2.1$df.residual)
```

```
[1] 0.0193637
```

```
> ## Plot of probabilities
> curve(plogis(fit2.1$coefficients[1]+fit2.1$coefficients[2]*x+fit2.1$coefficients[3]*4),
+      from=1,to=5.5,lwd=2,ylab="probability",xlab="weight",main="Color as Linear Variable")
> for (i in 1:3)
+ curve(plogis(fit2.1$coefficients[1]+fit2.1$coefficients[2]*x+fit2.1$coefficients[3]*i),
         from=1,to=5.5,lwd=2,col=cols[i],add=TRUE)
> legend(3.5,.5,col=c(cols,"black"),lwd=2,
        legend=c("Medium Light", "Medium", "Medium Dark","Dark"),bg="light gray")
```



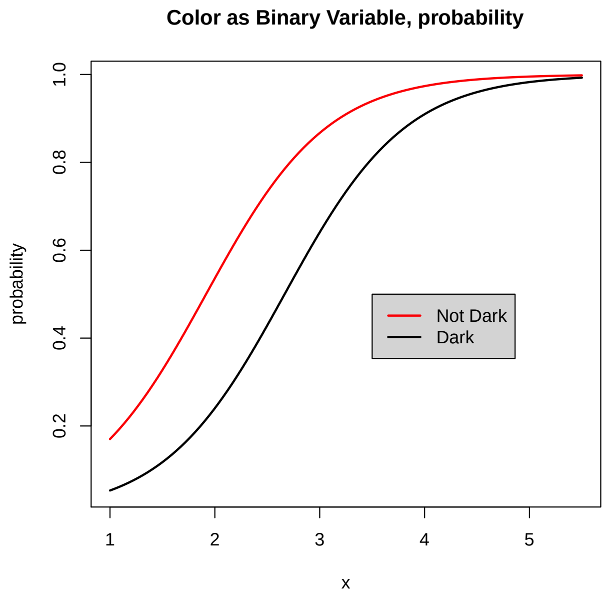
We notice that in fact there may not be 4 color categories but in fact only 2, dark and non-dark.

```
> dark=ifelse(unclass(color)<4,1,0)
> fit2.2=glm(y ~ weight + dark, family=binomial(link=logit))
> summary(fit2.2)
```

```
Call:
glm(formula = y ~ weight + dark, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-2.1555  -1.0233   0.5132   0.8484   2.0873

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6088     0.9922  -4.645 3.40e-06 ***
weight        1.7292     0.3825   4.520 6.18e-06 ***
dark          1.2954     0.5222   2.481   0.0131 *
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 189.17  on 170  degrees of freedom
AIC: 195.17

Number of Fisher Scoring iterations: 4
```

```
> ## Plot of probability
> curve(plogis(fit2.2$coefficients[1]+fit2.2$coefficients[2]*x), from=1,to=5.5,lwd=2,
+       col=c("black"),ylab="probability",main="Color as Binary Variable, probability")
> curve(plogis(fit2.2$coefficients[1]+fit2.2$coefficients[2]*x+fit2.2$coefficients[3]),
+       from=1,to=5.5,lwd=2,col=c("red"),add=TRUE)
> legend(3.5,.5,col=c("red","black"),lwd=2,legend=c("Not Dark","Dark"),bg = "light gray")
```



The likelihood ratio statistic indicates that this model is adequate and in fact note that the AIC is smaller than when 4 color categories were used.

Next we can add weight as a third predictor.

```
> fit4=glm(y ~ weight + linear + width, family=binomial(link=logit))
> summary(fit4)
```

```
Call:
glm(formula = y ~ weight + linear + width, family = binomial(link = logit))
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1605  -0.9650   0.5094   0.9012   1.9855

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.4065     3.7125  -1.995    0.046 *
weight        0.7450     0.6865   1.085    0.278
linear       -0.4937     0.2247  -2.197    0.028 *
width         0.2872     0.1873   1.533    0.125
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.92  on 169  degrees of freedom
AIC: 195.92

Number of Fisher Scoring iterations: 4
```

```
> # TEST WHETHER WIDTH IS SIGNIFICANT
> 1-pchisq(fit2.1$deviance-fit4$deviance,fit2.1$df.residual-fit4$df.residual)
```

```
[1] 0.1257814
```

```
> # TEST WHETHER COLOR AND WIDTH ARE SIGNIFICANT
> 1-pchisq(fit$deviance-fit4$deviance,fit$df.residual-fit4$df.residual)
```

```
[1] 0.02011901
```

A Poisson loglinear can also be fit since we have counts. Will be covered in later chapters.

```
> fit.poi=glm(satellite ~ weight, family=poisson(link=log))
> summary(fit.poi)
```

```
Call:
glm(formula = satellite ~ weight, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9307  -1.9981  -0.5627   0.9298   4.9992

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.42841    0.17893  -2.394   0.0167 *
weight       0.58930    0.06502   9.064   <2e-16 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.87  on 171  degrees of freedom
AIC: 920.16

Number of Fisher Scoring iterations: 5
```