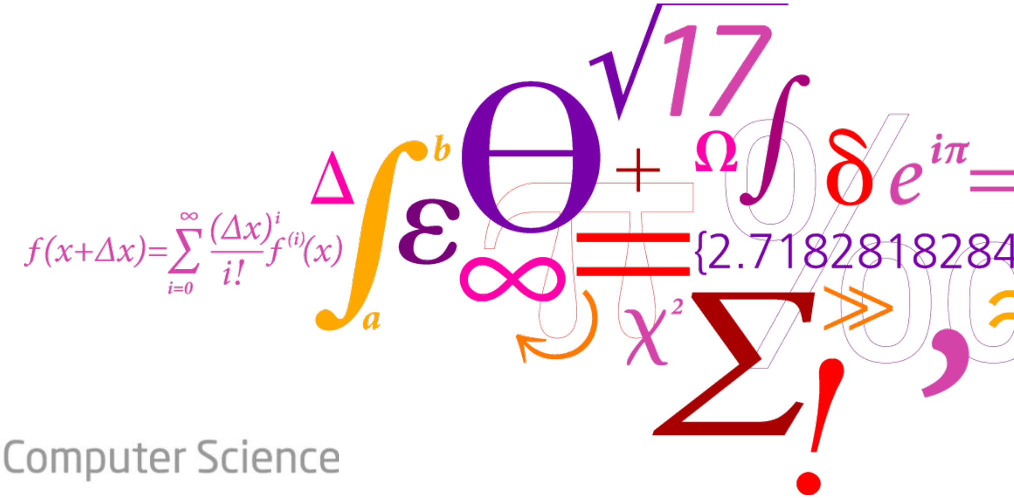
Logistic Regression

January 9th, 2025

Anders Stockmarr
Section for Statistics and Data Analysis, DTU
anst@dtu.dk



DTU Compute
Department of Applied Mathematics and Computer Science

Programme

- Monday : Statistical Inference, the t-test
- Tuesday : Simple and Multiple Regression
- Wednesday : ANOVA, ANCOVA, and Linear Models
- Thursday: Categorical Data, Writing Statistical Reports, **Logistic regression**
- Friday: Repeated Measurements, Principal Component Analysis

Contents:

1. Introduction.

2. Main example: Sperm competition among horseshoe crabs.

3. Exercise.

3. Logistic regression for frequency data.

5. Exercises.

6. Logistic regression for ordinal data.

What you should be able to do after the lecture:

a) Identify data suitable for logistic regression.

b) Carry out simple logistic regression analyses, and estimate the parameters.

c) Perform standard model control of logistic regression models.

Introduction

Logistic Regression

- Applies to:

- Binary data
  - Yes/No
  - Dead/Alive

- Frequency data
  - Percentage of sick people
  - Ratio of bycatch for fishing trawlers

- [Nominal data]

- Ordinal data

Color Blind Example

X=Number of events (colour blind children) out of N. With p the probability of event, it holds that

$$P(X = x) = \binom{N}{x} p^x (1-p)^{N-x};$$

X is binomially distributed (N, p).

The optimal estimator for p is the observed proportion of colour blind:

$$\hat{p} = X/N$$

In the example from the Categorical Data Session,

$$\hat{p} = 7/270 \approx 0.026$$

Color Blind Example

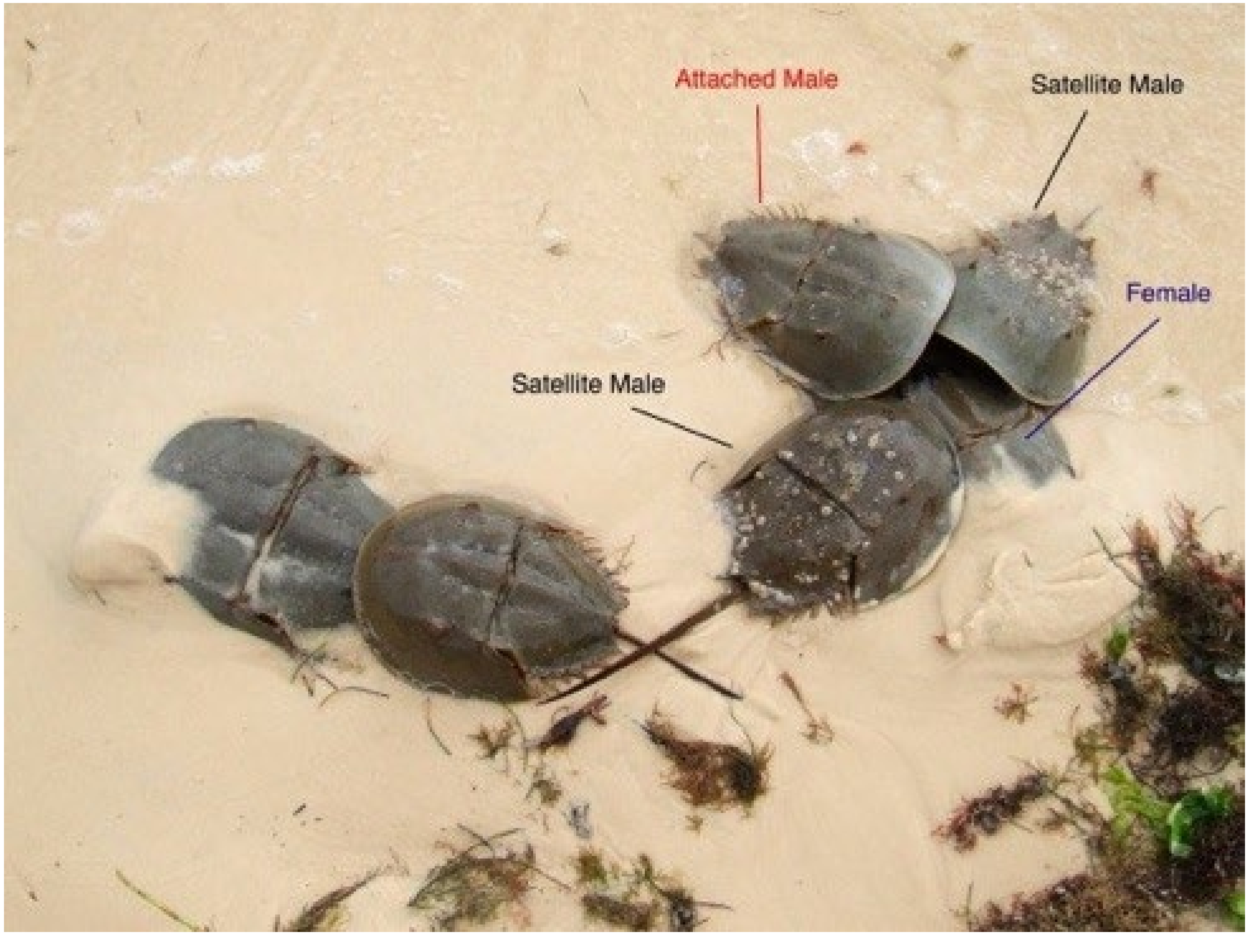
$$\hat{p} = X/N$$

Requires the observations to be *repetitions*;

Ie. each person investigated is assumed to have the **same** probability p of being colorblind.

If this probability *varies from person to person*; depending on t.ex. gender, but perhaps also quantitative genetic information (such as t.ex. the number of alleles at a locus associated with color blindness), a different type of analysis is required.

Data Example:
Sperm Competition in Horseshoe Crabs



http://people.biology.ufl.edu/dsasson

Crab Data

```
setwd("C:/<your data directory>")
crab.data<-read.table("crab.data.txt")

head(crab.data)

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

Nominal variables; categories

Horseshoe Crab Data Analysis

Central Question for sperm competition analysis:

What is the probability that a female has a satellite?

”y” in the crab dataset denotes the presence/absence of satellites.

If satelites attach themselves to females completely at random, $\mathbb{Y}$ will be either 0 or 1 with the same probability for all individuals:

$$P(Y = 1) = p; \quad P(Y = 0) = 1 - p,$$

Where p is the probability of having a satellite.

Horseshoe Crab Data Analysis

- Let $\mathbb{X}$ be the number of females with a satellite attached; then

$$P(\mathbb{X} = x) = \binom{N}{x} p^{x}(1-p)^{N-x}, \qquad \hat{p} = \frac{\mathbb{X}}{N}.$$

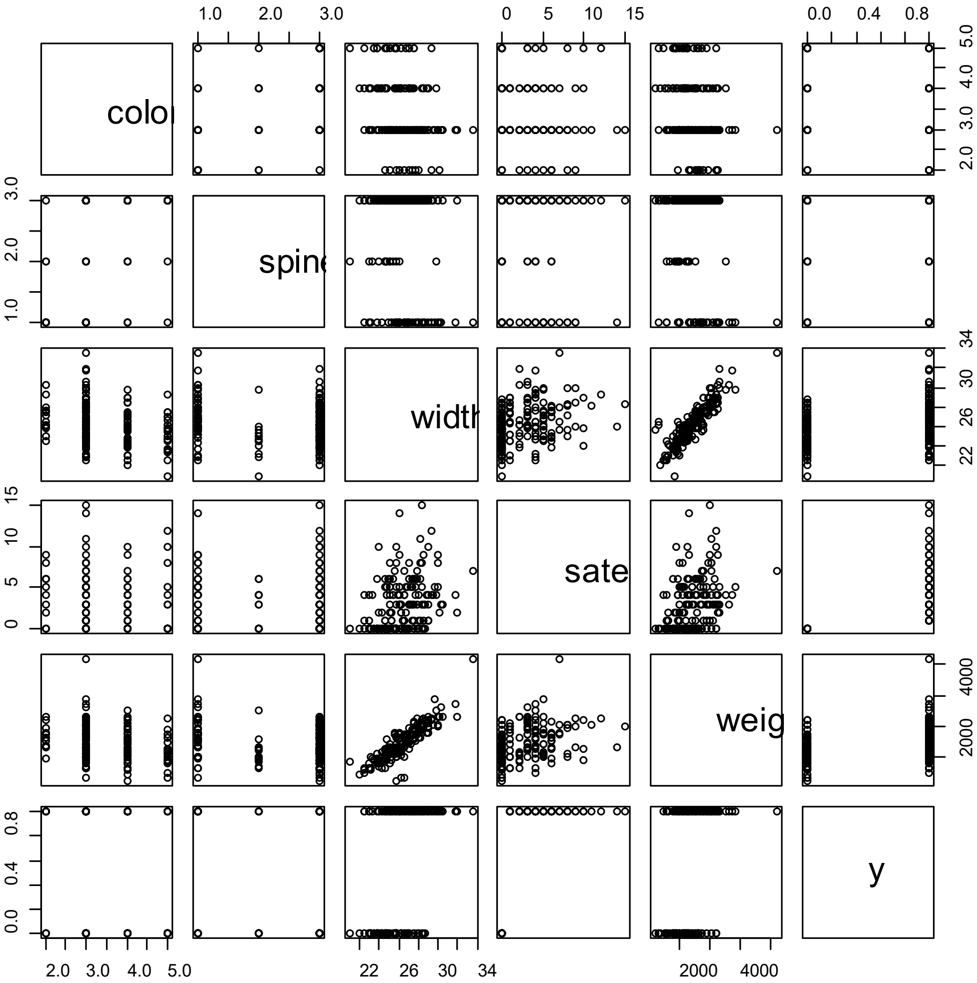
.

Finding N and $\hat{p}$:

```
N<-length(crab.data$y)
N
[1] 173

sum(y)/N
[1] 0.6416185
```

Horseshoe Crab Data Analysis



Horseshoe Crab Data Analysis

H: *p depends on the width of the crab*

Linear regression is one bid on how to model the effect. But there isn't really much hope, as the data hardly satify the normality assumption.

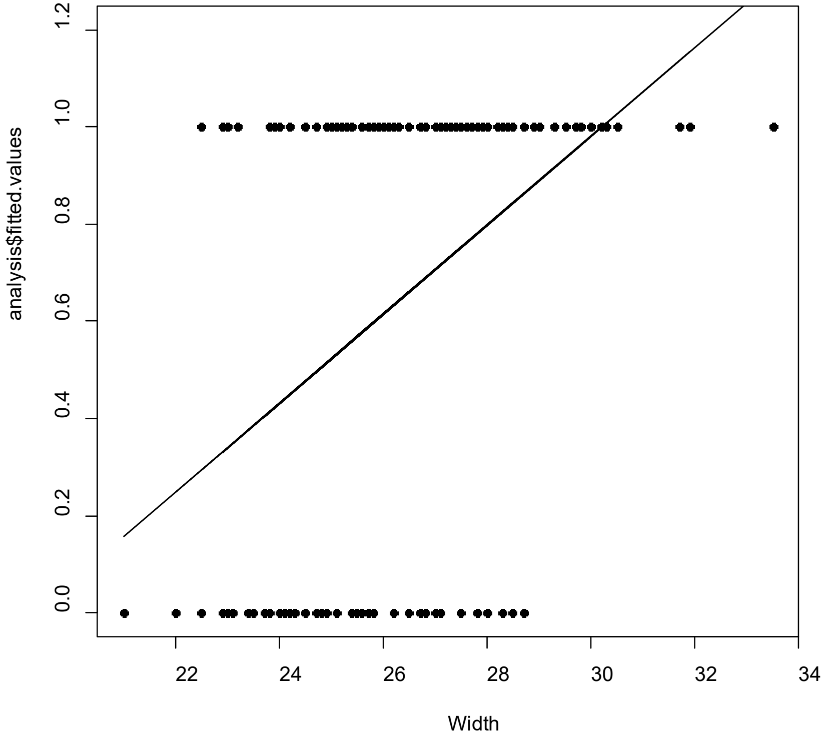
Lets see what happens….

Horseshoe Crab Data Analysis

- Linear regression:

```
analysis<-lm(y~width, data=crab.data)
analysis
Call:
lm(formula = y ~ width,
   data = crab.data)

Coefficients:
(Intercept)        width
   -1.76553      0.09153
```



Dependency of Width: Logistic Regression

Odds of a satellite (similar to Categorical Data session):

$$\frac{p}{1-p}$$

Log(odds):

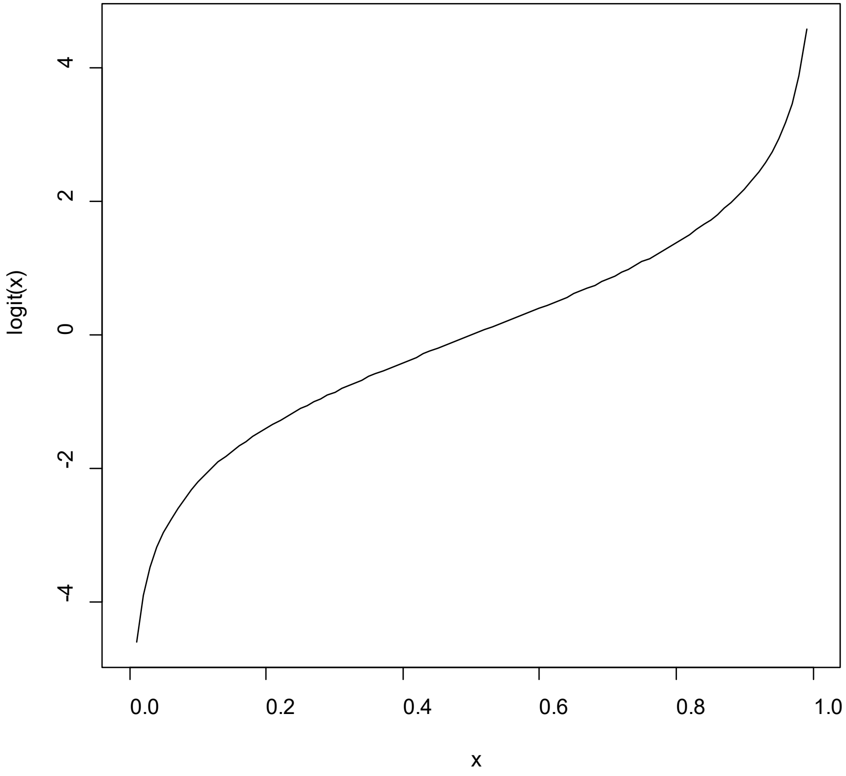
$$log\left(\frac{p}{1-p}\right)$$

This is the logit function:

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

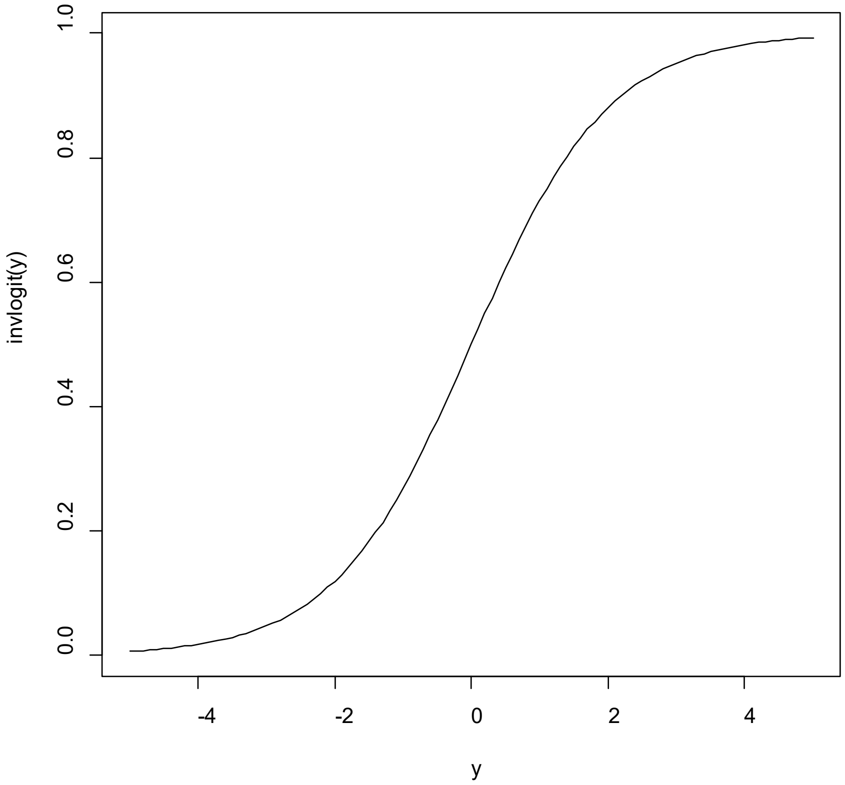
The Logit Function

$$logit(p) = log\left(\frac{p}{1-p}\right)$$



The Inverse Logit Function

$$invlogit(y) = \frac{e^y}{1 + e^y}$$



Dependency of Width: Logistic Regression

Model:

$$logit(p) = \alpha + \beta \cdot width$$

R: Use the glm function with option `family=binomial(link=logit)` :

```
analysis<-glm(y~width,family=binomial(link=logit),data=crab.data)
analysis
```

```
Call:  glm(formula = y ~ width, family = binomial(link = logit), data =
crab.data)

Coefficients:
(Intercept)        width
   -12.3508       0.4972

Degrees of Freedom: 172 Total (i.e. Null);  171 Residual
Null Deviance:      225.8
Residual Deviance: 194.5        AIC: 198.5
```

Dependency of Width: Logistic Regression

Model:

$$logit(p) = \alpha + \beta \cdot width$$

```
summary(analysis)
Call:
glm(formula = y ~ width, family = binomial(link = logit), data = crab.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942
```

$\widehat{\alpha}$!

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
```

$\widehat{\beta}$!

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45

Number of Fisher Scoring iterations: 4
```

Dependency of Width: Logistic Regression

Model:

$$P(\mathbb{Y}_i = 1) = p_i, \qquad logit(p_i) = \alpha + \beta \cdot width_i, \qquad i = 1, \dots 173.$$

Test:

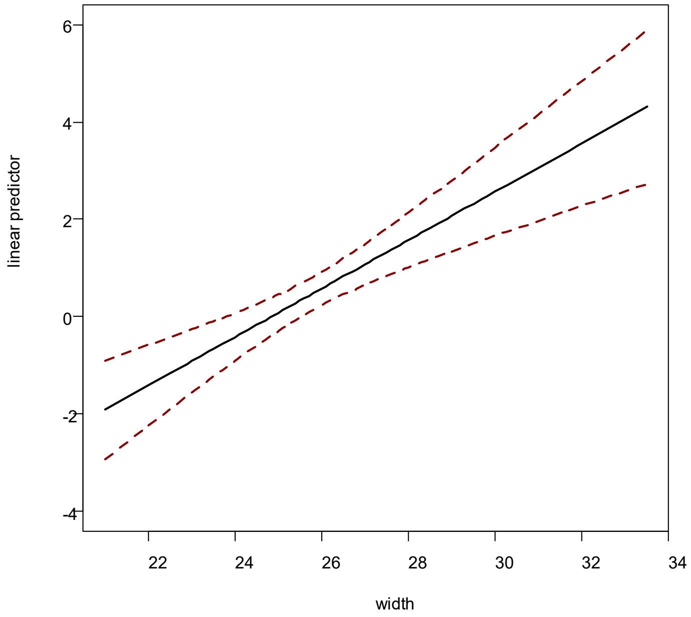
```
drop1(analysis,test=”Chisq”)

Single term deletions

Model:
y ~ width
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      194.45 198.45
width   1   225.76 227.76 31.306 2.204e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Control I:

```
prediction.temp<-as.data.frame(predict(analysis,se.fit=T))
prediction.data<-data.frame(pred=prediction.temp$fit,
                            upper=prediction.temp$fit+
                                  1.96*prediction.temp$se.fit,
                            lower=prediction.temp$fit-
                                  1.96*prediction.temp$se.fit)
```

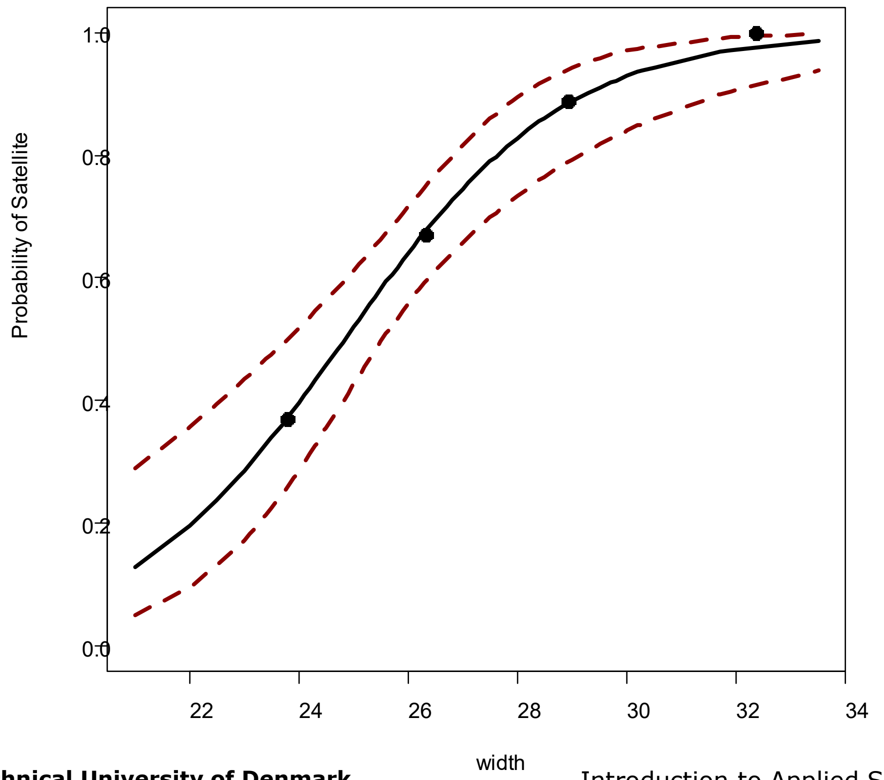


Model Control II:

```
prediction.data.original<-invlogit(prediction.data)
```

Plot with original data grouped frequencies:

P(satellite) as a Function of Width



Model Control III:

Polynomial regression:

$$logit(p_i) = \alpha + \beta \cdot width_i + \gamma \cdot width_i^2$$

```
analysis2<-update(analysis,~.+I(width^2))
drop1(analysis2,test=”Chisq”)

Single term deletions

Model:
y ~ width + I(width^2)
           Df Deviance    AIC     LRT Pr(>Chi)
<none>          193.63 199.63
width       1   194.10 198.10 0.47378   0.4913
I(width^2)  1   194.45 198.45 0.82542   0.3636
```

Model control IV:

- Summary:

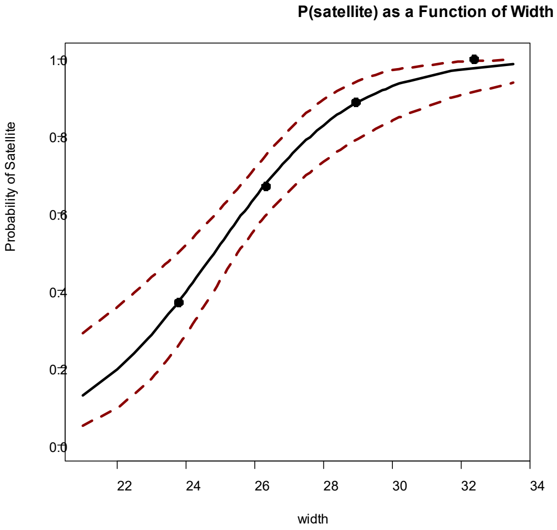
- Plot the predictors and the confidence intervals, group the original data, and check if they fall into the confidence area.

- Polynomial regression; if multiple covariates apply consider interaction terms (ie. the product of the covariates).

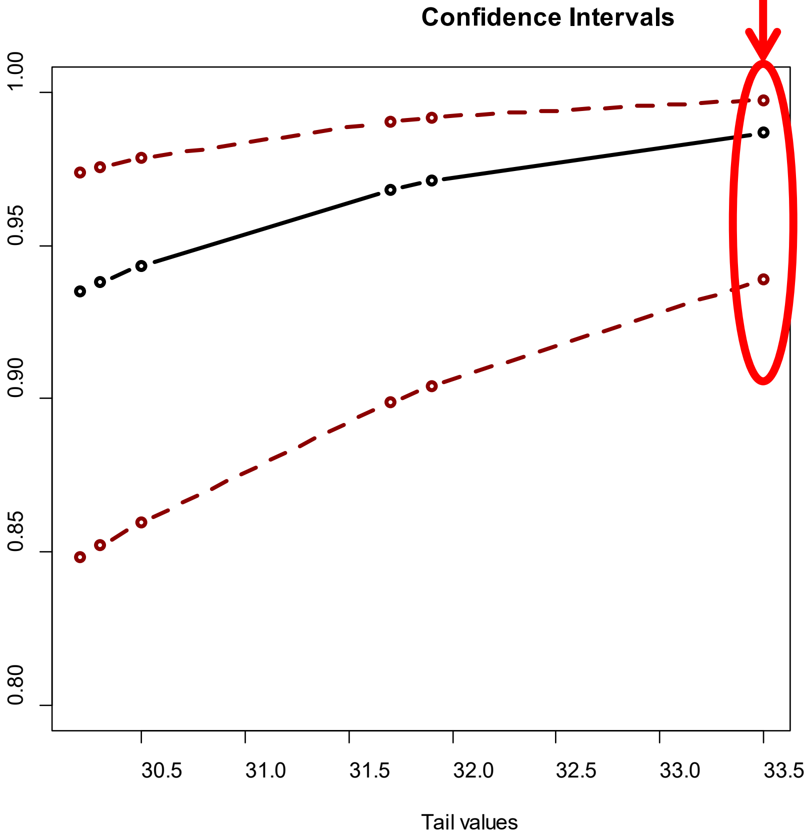
It is concluded that the model is a fair description of the data.

Crab Data Analysis

```
tail(prediction.data.original)
 pred      upper      lower
168 0.9349627 0.9736655 0.8482453
169 0.9379216 0.9753623 0.8522055
170 0.9434658 0.9784454 0.8598511
171 0.9680587 0.9904320 0.8987182
172 0.9709946 0.9916535 0.9041445
173 0.9866974 0.9972205 0.9387802
```



Asymmetric!



Crab Data Analysis Revisited

```
summary(analysis)

glm(formula = y ~ width, family = binomial(link = logit), data
= crab.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\hat{\alpha} = -12.3508, \qquad v(\hat{\alpha}) = 2.6287^2 = 6.910$$
$$\hat{\beta} = 0.4972, \qquad v\left(\hat{\beta}\right) = 0.1017^2 = 0.01035$$

Crab Data Analysis Revisited

Estimates are correlated. Covariance between estimators:

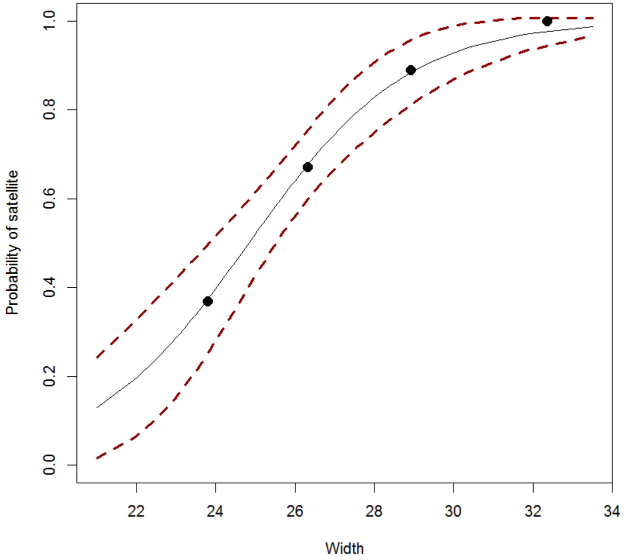
```
summary(analysis)$cov.scaled
             (Intercept)       width
(Intercept)    6.9101576 -0.26684761
width         -0.2668476  0.01035012
```

$cov(\hat{\alpha},\hat{\beta})$

Prediction Intervals – Brute Force (not recommended)

- You can predict directly on the original scale with predict:

```
predict(analysis,type="response",se.fit=T)
```

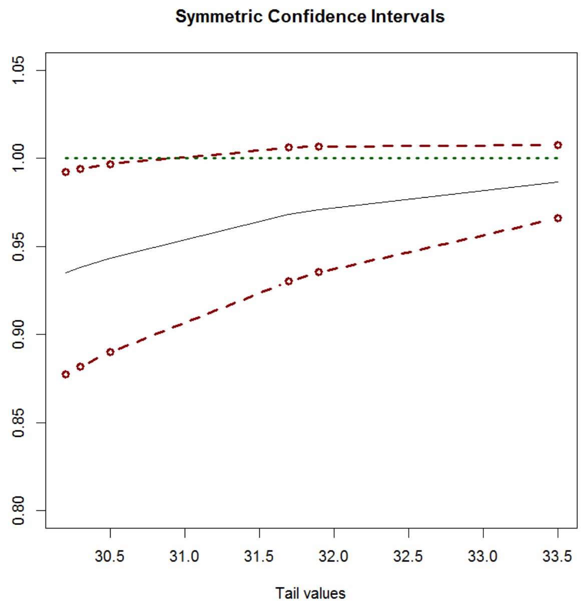


- Symmetric intervals; do not reflect that the link scale is the appropriate for that.

Prediction Intervals – Brute Force (not recommended)

- You can predict directly on the original scale with predict:

```
predict(analysis,type="response",se.fit=T)
```



Hard to interpret intervals - t. ex. values above 1; not super for a probability. **Assigning ±1.96sd should be done on the link scale, not the original/response scale**.

Prediction Intervals

Suppose we have an additional crab with width w.

What is a 95% confidence interval for this crab to have satellites?

$$logit(\hat{p}) = \hat{\alpha} + \hat{\beta}w$$

$$sd = sd\big(logit(\hat{p})\big)$$

$$\boxed{\begin{aligned} \hat{p} &= \frac{e^{\hat{\alpha}+\hat{\beta}w}}{1+e^{\hat{\alpha}+\hat{\beta}w}} = invlogit\big(\hat{\alpha} + \hat{\beta}w\ \big) \\ upper &= invlogit\big(\hat{\alpha} + \hat{\beta}w\ \ + 1.96 * sd\big) \\ lower &= invlogit\big(\hat{\alpha} + \hat{\beta}w\ \ - 1.96 * sd\big) \end{aligned}}$$

Prediction Intervals

- Assume w=15. In R, use the `newdata` option in `predict()`:

```
new.data<-data.frame(width=15)
new.prediction<-predict(analysis,newdata=new.data,se.fit=TRUE)
new.prediction.2<-data.frame(fit=new.prediction$fit,
                           upper=new.prediction$fit+1.96*
                                new.prediction$se.fit,
                           lower=new.prediction$fit-1.96*
                                new.prediction$se.fit)
invlogit(new.prediction.2)
  fit          upper      lower
1 0.007447815 0.06206388 0.00085019
```

Syntax: `?predict.glm`

Exercise:

recall that we have defined the main model as

```
analysis<-glm(y~width,family=binomial(link=logit),data=crab.data)
```

1) Plot the crab data again: plot(crab.data)

2) Deduce from the graph that another possible predictor for a satellite is the crab weight. Use the update() function to add weight to the model as on slide 24. How does that alter the model? If you should choose between width and weight, which one would you choose?

3) A third possible predictor for satellites is the color of the female. The color is a nominal covariate where higher value indicates darker skin, so it is added to the model as a factor:

```
analysis2<-update(analysis,~.+ as.factor(color))
```

Check that color does not add significantly to the model. Which color label stands out the most?

4) Create a new dataset that included an indicator for darkskinned females:

```
crab.data.2<-data.frame(crab.data,dark=1*(crab.data$color==5))
```

Add ‘dark’ to the model with the command

```
analysis2<-update(analysis,~.+dark,data=crab.data.2)
```

Do satellite males prefer light-skinned or dark-skinned females, or are they indifferent?

Real Interest:

Does Horseshoe Crabs recognize high fertility?

- Light Skin of horseshoe female crabs is associated with increased fertility!
- To investigate this, we are not really interested in the effect of the width;
- But we have to model the effect of the width, as it is a **confounder** for the color preference; an attachment could be either because the female is wide, or because it has a light colored skin.
- To model the width effect, the logistic regression model is obvious.

What if we just used the t-test?

```
t.test(crab.data.3$y[crab.data.3$dark==0],
       crab.data.3$y[crab.data.3$dark==1])
```

yields p=0.002.

```
BUT:
mean(crab.data.3$width[crab.data.3$dark==1])
[1] 25.28182
mean(crab.data.3$width[crab.data.3$dark==0])
[1] 26.44702
```

```
t.test(crab.data.3$width[crab.data.3$dark==0],
       crab.data.3$width[crab.data.3$dark==1])
```

yields p=0.01: In this dataset, light.skinned crabs are significantly wider than dark-skinned crabs.

We cannot know if the conclusion from the t-test is because of the color or the width.

Logistic Regression For Frequency Data

Smoking, Obesity, Snoring (SOS)

Effect of Smoking, Obesity and Snoring on **Hypertension** (Altman (1991, page 353)):

```
sosdata<-read.table(”sosdata.txt”)
sosdata
```

```
  smoking obesity snoring n.tot n.hyp
1      No      No      No    60     5
2     Yes      No      No    17     2
3      No     Yes      No     8     1
4     Yes     Yes      No     2     0
5      No      No     Yes   187    35
6     Yes      No     Yes    85    13
7      No     Yes     Yes    51    15
8     Yes     Yes     Yes    23     8
```

Smoking, Obesity, Snoring (SOS)

Model: Let *p* be the probability of hypertension. Then

$$logit(p) = \alpha + \beta_{smoking} + \beta_{obese} + \beta_{snoring}$$

Thus: The odds ratio of hypertension for a smoker vs. a non-smoker, with the same snoring and obesity status, is given by

$$exp(\beta_{smoking})$$

Coding in R:

```
analysis.sos<-glm(n.hyp/n.tot~smoking + obesity + snoring,
family=binomial(link=logit), weights = n.tot)
```

Compactified table on slide 37! Requires weights.

Smoking, Obesity, Snoring (SOS)

```
analysis.sos<-glm(n.hyp/n.tot~smoking + obesity + snoring,
family=binomial(link=logit), weights = n.tot)

summary(analysis.sos)
Call:
glm(formula = n.hyp/n.tot ~ smoking + obesity + snoring, family
= binomial(link = logit),
    data = sosdata, weights = n.tot)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Smoking, Obesity, Snoring (SOS)

Odds ratios for smoking, obesity and snoring:

```
exp(cbind (OR = coef(analysis.sos), confint(analysis.sos)))
Waiting for profiling to be done...
                    OR      2.5 %   97.5 %
(Intercept) 0.09276726 0.04063914 0.183823
smokingYes  0.93447081 0.53379700 1.594628
obesityYes  2.00432951 1.13345994 3.478922
snoringYes  2.39154432 1.15660384 5.605594
```

Note that the interval for smoking contains 1; smoking is insignificant.

Exercises:

1) Load the data set surgery as

```
surgery<-read.table(”surgery.txt”,header=T)
```

The dataset shows the results of a study about Y= whether a patient having surgery with general anesthesia experienced a sore throat on waking up (0=no, 1=yes), as a function of the D= duration of the surgery in minutes; and the T= type of device used to secure the airway (0= laryngeal mask airway, 1= tracheal tube). Fit a logistic regression model using these predictors, interpret parameter estimates, and conduct inference about the effects.

Source: D. Collett, in Encyclopedia of Biostatistics (Wiley, New York 1998), pp.350-358.

2) Alternative formulation for frequency data: Access the internal R dataset menarche (proportion of female children that have reached menarche/first menstruation), by typing

```
library(MASS); data(menarche); attach(menarche)
```

Model the matrix `cbind(Menarche,Total-Menarche)` as a function of Age, and make a plot with the data and the fitted logistic regression curve.

ORDINAL REGRESSION

Ordinal Regression

- In the lecture on logistic regression, we modeled the probability of a satellite for a Female Horseshoe Crab as

$$logit(p_i) = \alpha + \beta w_i$$

where w was the width of the crab.

The logit function is the log of the odds:

$$logit(p_i) = log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right)$$

However, this relies on if data are binary. What if there were more response groups than two?

Ordinal Regression

Let’s activate the `ordinal` package, and look at the wine data:

```
library(ordinal)
summary(wine)
 response        rating   temp      contact       bottle        judge   
 Min.   :12.00   1: 5   cold:36   no :36   1      : 9   1      : 8  
 1st Qu.:32.00   2:22   warm:36   yes:36   2      : 9   2      : 8  
 Median :46.00   3:26                      3      : 9   3      : 8  
 Mean   :47.22   4:12                      4      : 9   4      : 8  
 3rd Qu.:60.00   5: 7                      5      : 9   5      : 8  
 Max.   :90.00                             6      : 9   6      : 8  
                                           (Other):18   (Other):24  
```

`temp` is the temperature when chrushing the grapes, while `contact` indicates contact between juice and skin during the crushing process. The `rating` of the wines have 5 ordinal levels.

Ordinal Regression

- With the response `Y=rating`, the response is no longer binary, but ordinal. Instead of the success/failure events (satellite/no satellite), we organise the data into the 4 consecutive success events:

$$Y \leq 1, Y \leq 2, Y \leq 3, Y \leq 4;$$
$$Y \leq j, j = 1{:}\,4$$

None of these events happen if $Y = 5$; so $Y = 5$ is the ‘failure’ here. We thus have 4 versions of ‘success’; lets model the odds of success:

$$logit(P(Y_i \leq j)) = \alpha_j + temp_i + contact_i + temp_i{:}\,contact_i, j = 1, \ldots 4.$$

Remember that the logit function is the log of the odds. The odds of the event $\{Y_i \leq j\}$ are

$$\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} = \frac{P(Y_i \leq j)}{P(Y_i > j)}, j = 1, \ldots 4$$

Ordinal Regression

In the ordinal package, we can use the `clm` function to do ordinal regression:

```
analysis <- clm(rating ~ temp * contact, data = wine)
drop1(analysis,test=”Chisq”)

Single term deletions

Model:
rating ~ temp * contact
             Df    AIC     LRT Pr(>Chi)
<none>          186.83
temp:contact  1 184.98 0.15145   0.6972
```

Ordinal Regression

No interaction:

```
analysis2 <-update(analysis,~.-temp:contact)
drop1(analysis2,test="Chisq")

Single term deletions

Model:
rating ~ temp + contact
        Df    AIC    LRT  Pr(>Chi)
<none>     184.98
temp     1 209.91 26.928 2.112e-07 ***
contact  1 194.03 11.043 0.0008902 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ordinal Regression

- Parameter estimates:

```
summary(analysis2)$coefficients
             Estimate Std. Error   z value     Pr(>|z|)
1|2         -1.344383  0.5171020 -2.599842 9.326680e-03
2|3          1.250809  0.4378802  2.856509 4.283277e-03
3|4          3.466887  0.5977604  5.799793 6.639670e-09
4|5          5.006404  0.7309063  6.849584 7.406519e-12
tempwarm     2.503102  0.5286801  4.734625 2.194605e-06
contactyes   1.527798  0.4766226  3.205466 1.348440e-03
```

- We need to get back to the probabilities of the response groups. We know how to get from the (log) odds to probabilities from the categorical data lecture.

Ordinal Regression

- Consider the reference group, the situation where `temp=”cold”` and `contact=”no”`. In this case, the linear predictor for the probabilities $P(Y \leq j), j = 1, \ldots, 4$ is exactly the first four parameter estimates in the table below.

```
summary(analysis2)$coefficients
             Estimate Std. Error   z value     Pr(>|z|)
1|2         -1.344383  0.5171020 -2.599842 9.326680e-03
2|3          1.250809  0.4378802  2.856509 4.283277e-03
3|4          3.466887  0.5977604  5.799793 6.639670e-09
4|5          5.006404  0.7309063  6.849584 7.406519e-12
tempwarm     2.503102  0.5286801  4.734625 2.194605e-06
contactyes   1.527798  0.4766226  3.205466 1.348440e-03
```

- Lets extract them:

```
my.linear.predictor<- analysis2$alpha
```

Ordinal Regression

- From (log) odds to probabilities:

```
temp<-exp(my.linear.predictor)/(1+exp(my.linear.predictor))
```

From cumulated probabilities to category probabilities:

```
my.probabilities<-c(temp[1],diff(temp),1-temp[4])
my.probabilities
       1|2        2|3        3|4        4|5        4|5
0.20679013 0.57064970 0.19229094 0.02361882 0.00665041
```

When the temperature is cold and there is no contact, the wine is most often rated 2.

Ordinal Regression

Cold and no contact:

```
0.20679013 0.57064970 0.19229094 0.02361882 0.00665041
```

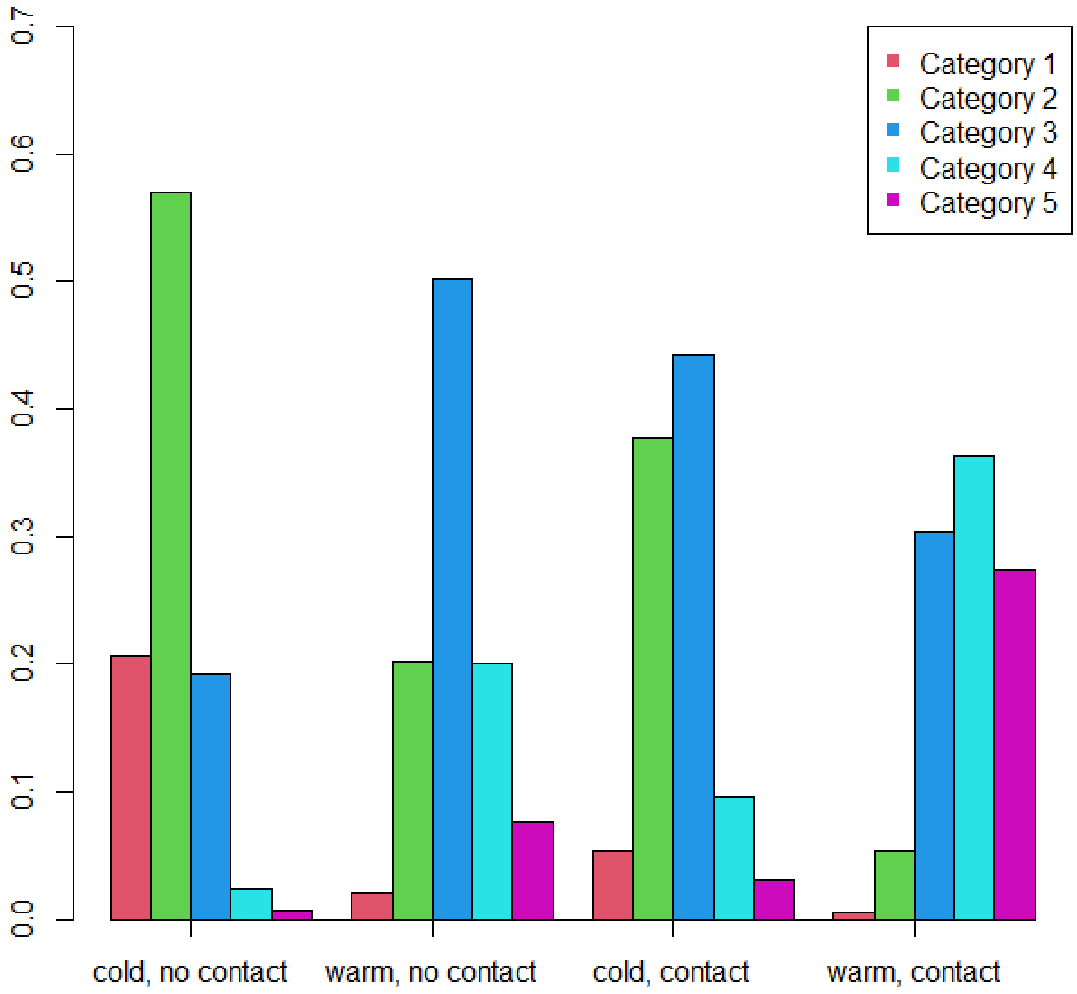
- All estimates in one go:

```
my.predictdata<-expand.grid(temp=c("cold","warm"),contact=c("no","yes"))

cbind(my.predictdata,predict(analysis2,newdata=my.predictdata)$fit)
  temp contact           1          2         3          4          5
1 cold      no 0.206790132 0.57064970 0.1922909 0.02361882 0.00665041
2 warm      no 0.020887709 0.20141572 0.5015755 0.20049402 0.07562701
3 cold     yes 0.053546010 0.37764614 0.4430599 0.09582084 0.02992711
4 warm     yes 0.004608274 0.05380128 0.3042099 0.36359581 0.27378469
```

We recognize the first line as the values we calculated. The wines are generally rated highest when the temperature is warm and there is contact between the juice and skin.

Ordinal Regression



Ordinal Regression the Proportional Odds assumption and more

- In the formula

$$logit(P(Y_i \leq j)) = \alpha_j + temp_i + contact_i + temp_i\!:contact_i, j = 1, \ldots 4$$

the impact of `temp` and `contact` does not depend on the response level j at the logit scale.

This assumption is standard, but one may wish to be able to model a changing effect of temperature and contact when the category level changes.

This can be done in R using the `vglm` function from the `VGAM` library. This is referred to as self-study.

General nominal regression can be done in R with the `multinom` function from the `nnet` library. This too is referred to as self-study.

Confidence intervals on the original scale: One can use similar methods as in the lecture on logistic regression, to obtain uncertainties about probabilities.