# Binomial Data

## Ex 1, Almost Factorial Experiment

- **ncases, ncontrols** = y-variable (cancer cases, vs. no cancer cases)
- **agegp** = Age group (six levels)
- **alcgp** = Alcohol level (four levels)
- **tpbgp = Tobacco consumption (four levels)**

**<< R printout >>**

```
> data(esoph)
> dim(esoph)
[1] 88  5
> head(esoph)
  agegp     alcgp    tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day      0        40
2 25-34 0-39g/day    10-19      0        10
3 25-34 0-39g/day    20-29      0         6
4 25-34 0-39g/day      30+      0         5
5 25-34     40-79 0-9g/day      0        27
6 25-34     40-79    10-19      0         7
> attach(esoph)
> table(agegp)
agegp
25-34 35-44 45-54 55-64 65-74   75+
   15    15    16    16    15    11
> table(alcgp,tobgp)
           tobgp
alcgp       0-9g/day 10-19 20-29 30+
  0-39g/day        6     6     5   6
  40-79            6     6     6   5
  80-119           6     6     4   5
  120+             6     6     5   4
> model1 <-
glm(cbind(ncases,ncontrols)~agegp+alcgp*tobgp, binomial)
> summary(model1)

Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp * tobgp,
    family = binomial)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.75985    0.19822  -8.878  < 2e-16 ***
agegp.L          2.99646    0.65386   4.583 4.59e-06 ***
agegp.Q         -1.35008    0.59197  -2.281   0.0226 *
agegp.C          0.13436    0.45056   0.298   0.7655
agegp^4          0.07098    0.30974   0.229   0.8187
agegp^5         -0.21347    0.19627  -1.088   0.2768
alcgp.L          1.37077    0.21136   6.485 8.85e-11 ***
alcgp.Q         -0.14913    0.19645  -0.759   0.4478
alcgp.C          0.22823    0.18203   1.254   0.2099
tobgp.L          0.63846    0.19710   3.239   0.0012 **
tobgp.Q          0.02922    0.19617   0.149   0.8816
tobgp.C          0.15607    0.19796   0.788   0.4304
alcgp.L:tobgp.L -0.70426    0.41128  -1.712   0.0868 .
alcgp.Q:tobgp.L  0.12948    0.38889   0.333   0.7392
alcgp.C:tobgp.L -0.16118    0.36697  -0.439   0.6605
alcgp.L:tobgp.Q  0.12225    0.42044   0.291   0.7712
alcgp.Q:tobgp.Q -0.44527    0.39224  -1.135   0.2563
alcgp.C:tobgp.Q  0.04843    0.36211   0.134   0.8936
alcgp.L:tobgp.C -0.29187    0.42939  -0.680   0.4967
alcgp.Q:tobgp.C -0.05205    0.39538  -0.132   0.8953
alcgp.C:tobgp.C -0.13905    0.35754  -0.389   0.6973
---
 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  47.484  on 67  degrees of freedom
AIC: 236.96
```

```
Number of Fisher Scoring iterations: 6
```

# # Simplify model

```
> model2 <-
glm(cbind(ncases,ncontrols)~agegp+alcgp+tobgp, binomial)
> anova(model1, model2)
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + alcgp * tobgp
Model 2: cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
  Resid. Df Resid. Dev Df Deviance
1        67     47.484
2        76     53.973 -9  -6.4895
> anova(model1, model2, test="F")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + alcgp * tobgp
Model 2: cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        67     47.484
2        76     53.973 -9  -6.4895 0.7211 0.6901
Warning message:
using F test with a 'binomial' family is inappropriate
> anova(model1, model2, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + alcgp * tobgp
Model 2: cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        67     47.484
2        76     53.973 -9  -6.4895   0.6901
```

# # Simpler model is better!

```
> summary(model2)

Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
    family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.77997    0.19796  -8.992  < 2e-16 ***
agegp.L      3.00534    0.65215   4.608 4.06e-06 ***
agegp.Q     -1.33787    0.59111  -2.263  0.02362 *
agegp.C      0.15307    0.44854   0.341  0.73291
agegp^4      0.06410    0.30881   0.208  0.83556
agegp^5     -0.19363    0.19537  -0.991  0.32164
alcgp.L      1.49185    0.19935   7.484 7.23e-14 ***
alcgp.Q     -0.22663    0.17952  -1.262  0.20680
alcgp.C      0.25463    0.15906   1.601  0.10942
tobgp.L      0.59448    0.19422   3.061  0.00221 **
tobgp.Q      0.06537    0.18811   0.347  0.72823
tobgp.C      0.15679    0.18658   0.840  0.40071
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  53.973  on 76  degrees of freedom
AIC: 225.45

Number of Fisher Scoring iterations: 6
```

# # Simplify factor-levels

```
> tobgp2 <- tobgp
> levels(tobgp2)[2:3] <- "10-30"
> table(tobgp2)
tobgp2
0-9g/day    10-30      30+
      24       44       20
> agegp2 <- agegp
> levels(agegp2)[4:6] <- "55+"
> levels(agegp2)[1:2] <- "under45"
> table(agegp2)
agegp2
under45   45-54     55+
     30      16      42
> model3 <- glm(cbind(ncases,ncontrols)~agegp2*alcgp*tobgp2, binomial)
> model4 <- step(model3)
Start:  AIC=241.11
cbind(ncases, ncontrols) ~ agegp2 * alcgp * tobgp2

                      Df Deviance    AIC
- agegp2:alcgp:tobgp2 12   30.807 226.29
<none>                     21.631 241.11

Step:  AIC=226.29
cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2 + agegp2:alcgp +
    agegp2:tobgp2 + alcgp:tobgp2

                Df Deviance    AIC
- alcgp:tobgp2   6   38.225 221.71
- agegp2:tobgp2  4   38.021 225.50
<none>               30.807 226.29
- agegp2:alcgp   6   45.443 228.92

Step:  AIC=221.71
cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2 + agegp2:alcgp +
    agegp2:tobgp2

                Df Deviance    AIC
- agegp2:tobgp2  4   44.677 220.16
<none>               38.225 221.71
- agegp2:alcgp   6   52.295 223.78

Step:  AIC=220.16
cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2 + agegp2:alcgp

               Df Deviance    AIC
<none>              44.677 220.16
- agegp2:alcgp  6   60.388 223.87
- tobgp2        2   52.991 224.47
Warning messages:
1: glm.fit: fitted probabilities numerically 0 or 1 occurred
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
3: glm.fit: fitted probabilities numerically 0 or 1 occurred
4: glm.fit: fitted probabilities numerically 0 or 1 occurred

> model5 <- update(model4,~.-agegp2:alcgp)
> anova(model4, model5, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2 + agegp2:alcgp
Model 2: cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        74     44.677
2        80     60.388 -6  -15.711  0.01539 *
---
> summary(model4)

Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp2 + alcgp + tobgp2 +
    agegp2:alcgp, family = binomial)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)       -3.20524  159.13069  -0.020  0.98393
agegp2.L           4.41490  337.56710   0.013  0.98957
agegp2.Q          -1.88022  194.89460  -0.010  0.99230
alcgp.L            0.95484  142.33121   0.007  0.99465
alcgp.Q            2.35990  318.26137   0.007  0.99408
alcgp.C            4.11708  426.99230   0.010  0.99231
tobgp2.L           0.53876    0.19258   2.798  0.00515 **
tobgp2.Q           0.05769    0.15042   0.383  0.70135
agegp2.L:alcgp.L   1.48668  301.92965   0.005  0.99607
agegp2.Q:alcgp.L  -2.19843  174.32018  -0.013  0.98994
agegp2.L:alcgp.Q  -6.11440  675.13419  -0.009  0.99277
agegp2.Q:alcgp.Q   4.07646  389.78919   0.010  0.99166
agegp2.L:alcgp.C  -8.07675  905.78741  -0.009  0.99289
agegp2.Q:alcgp.C   4.10633  522.95669   0.008  0.99373
---
```

# # L, Q, C … is because factor levels were "ordered"

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

    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  44.677  on 74  degrees of freedom
AIC: 220.16

Number of Fisher Scoring iterations: 17
```

# # One more simplification of factor-levels & using "unordered" factor levels

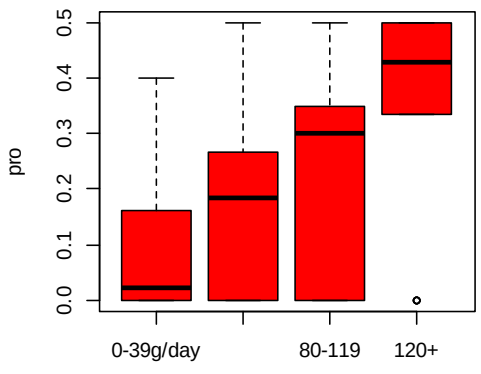
```
> alcgp3 <- alcgp
> levels(alcgp3)[2:3] <- "40-119"
> table(alcgp3)
alcgp3
0-39g/day    40-119      120+
       23        44        21
> is.factor(alcgp3)
[1] TRUE
> is.ordered(alcgp3)
[1] TRUE

> alcgp3 <- factor(alcgp3,ordered=F)
> is.ordered(alcgp3)
[1] FALSE
> agegp3 <- factor(agegp2,ordered=F)
> tobgp3 <- factor(tobgp2,ordered=F)
> is.ordered(agegp2)
[1] TRUE
> is.ordered(agegp3)
[1] FALSE
> is.ordered(tobgp2)
[1] TRUE
> is.ordered(tobgp3)
[1] FALSE
> model6 <- glm(cbind(ncases,ncontrols)~agegp3*alcgp3*tobgp3, binomial)
> model7 <- step(model6)
Start:  AIC=227.78
cbind(ncases, ncontrols) ~ agegp3 * alcgp3 * tobgp3

                      Df Deviance    AIC
- agegp3:alcgp3:tobgp3  8   34.972 220.45
<none>                      26.299 227.78

Step:  AIC=220.45
cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3 + agegp3:alcgp3 +
    agegp3:tobgp3 + alcgp3:tobgp3

                Df Deviance    AIC
- agegp3:tobgp3  4   42.052 219.53
- alcgp3:tobgp3  4   42.100 219.58
<none>               34.972 220.45
- agegp3:alcgp3  4   47.862 225.34

Step:  AIC=219.53
```

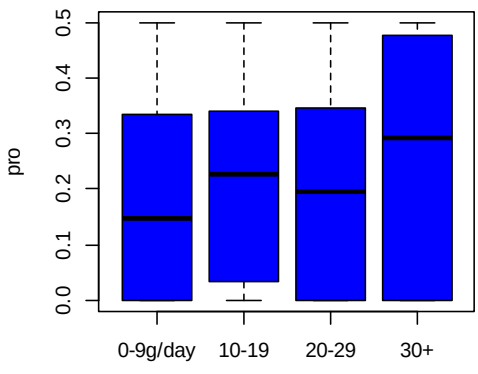
```
cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3 + agegp3:alcgp3 +
    alcgp3:tobgp3

                Df Deviance    AIC
- alcgp3:tobgp3  4   48.559 218.04
<none>               42.052 219.53
- agegp3:alcgp3  4   56.310 225.79

Step:  AIC=218.04
cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3 + agegp3:alcgp3

                Df Deviance    AIC
<none>               48.559 218.04
- tobgp3         2   57.532 223.01
- agegp3:alcgp3  4   62.468 223.95
> summary(model7)

Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3 +
    agegp3:alcgp3, family = binomial)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.1929     1.0066  -5.159 2.49e-07 ***
agegp345-54                0.6559     1.4217   0.461 0.644562
agegp355+                  3.0628     1.0248   2.989 0.002801 **
alcgp340-119               1.3133     1.1248   1.168 0.242985
alcgp3120+                 3.8060     1.1313   3.364 0.000767 ***
tobgp310-30                0.3293     0.1798   1.832 0.066980 .
tobgp330+                  0.7837     0.2715   2.887 0.003893 **
agegp345-54:alcgp340-119   1.6374     1.5225   1.075 0.282189
agegp355+:alcgp340-119    -0.2974     1.1504  -0.259 0.795984
agegp345-54:alcgp3120+     0.2203     1.5625   0.141 0.887895
agegp355+:alcgp3120+      -2.2843     1.1780  -1.939 0.052494 .
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  48.559  on 77  degrees of freedom
AIC: 218.04

Number of Fisher Scoring iterations: 6

> model8 <- update(model7,~.-agegp3:alcgp3)
> anova(model7, model8, test="Chi")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3 + agegp3:alcgp3
Model 2: cbind(ncases, ncontrols) ~ agegp3 + alcgp3 + tobgp3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        77     48.559
2        81     62.468 -4  -13.909  0.00759 **
---
```

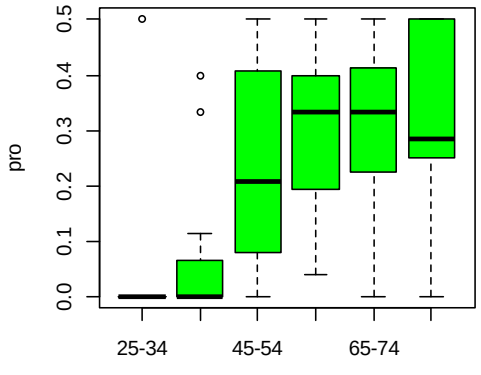
# # Model 7 is the best!

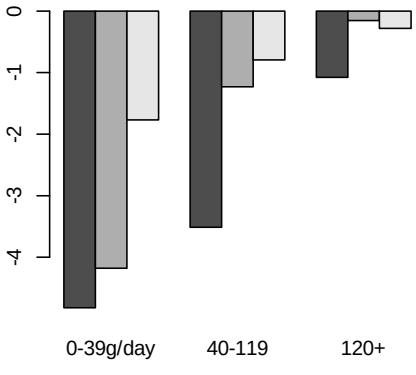
# # Plots

```
> pro <- ncases/(ncases+ncontrols)
> par(mfrow=c(2,2))
> plot(pro~alcgp,col="red"); plot(pro~tobgp,col="blue");plot(pro~agegp,col="green")
> barplot(tapply(predict(model7),list(agegp3,alcgp3),mean),beside=T)
```




# Ex 2, Binomial vs. x

- **`males, females`** = y-variable (number of males & females)
- **`density`** = x variable

**<< R printout >>**

```
data1 <- read.table("D:\\STAT999\\RBOOK\\sexratio.txt",header=T)
dim(data1)
[1] 8 3
data1
  density females males
1       1       1     0
2       4       3     1
3      10       7     3
4      22      18     4
5      55      22    33
6     121      41    80
7     210      52   158
8     444      79   365
attach(data1)
pro <- males/(males+females)
y <- cbind(males, females)
model1 <- glm(y~density, binomial)
summary(model1)

Call:
glm(formula = y ~ density, family = binomial)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0807368  0.1550376   0.521    0.603
density     0.0035101  0.0005116   6.862 6.81e-12 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.159  on 7  degrees of freedom
Residual deviance: 22.091  on 6  degrees of freedom
AIC: 54.618

Number of Fisher Scoring iterations: 4
```

# # Overdispersion!

```
model1 <- glm(y~log(density), binomial)
summary(model1)

Call:
glm(formula = y ~ log(density), family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.65927    0.48758  -5.454 4.92e-08 ***
log(density)  0.69410    0.09056   7.665 1.80e-14 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.1593  on 7  degrees of freedom
Residual deviance:  5.6739  on 6  degrees of freedom
AIC: 38.201

Number of Fisher Scoring iterations: 4
```

# # No overdispersion!

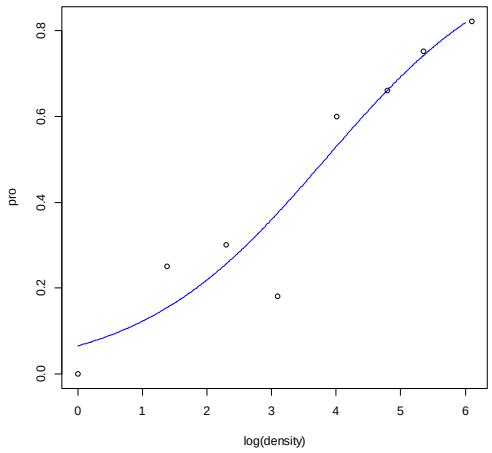
```
(-2.65927+(0.6941*4.5))
[1] 0.46418
(predict(model1, list(density=exp(4.5))))
        1
0.4641847
(predict(model1, list(density=exp(4.5)),type="response"))
        1
0.6140064
```

# # OR use “quasibinomial”. Then p-values become very different.

```
> model11 <- glm(y~density, quasibinomial)
> summary(model11)

Call:
glm(formula = y ~ density, family = quasibinomial)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0807368  0.2888702   0.279   0.7893
density     0.0035101  0.0009532   3.683   0.0103 *
---
(Dispersion parameter for quasibinomial family taken to be 3.471613)

    Null deviance: 71.159  on 7  degrees of freedom
Residual deviance: 22.091  on 6  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

# # Plot

```
xx <- seq(0,6,0.01)
yy <- predict(model1, list(density=exp(xx)),type="response")
plot(log(density),pro)
lines(xx,yy,col="blue")
```




# Ex 3, Binomial vs. x. If "quasibinomial" is used (due to OVERDISPERSION), use F test instead of $\chi^2$ test.

- **count, sample**= y-variable (number germinated out of total sample)
- **Orobanche** = x variable (two kinds of genotypes)
- **extract =** x variable (two kinds of host plants)

**<< R printout >>**

```
data1 <- read.table("D:\\STAT999\\RBOOK\\germination.txt",header=T)
dim(data1)
[1] 21  4
data1
   count sample Orobanche  extract
1     10     39       a75     bean
2     23     62       a75     bean
3     23     81       a75     bean
4     26     51       a75     bean
5     17     39       a75     bean
6      5      6       a75 cucumber
7     53     74       a75 cucumber
8     55     72       a75 cucumber
9     32     51       a75 cucumber
10    46     79       a75 cucumber
11    10     13       a75 cucumber
12     8     16       a73     bean
13    10     30       a73     bean
14     8     28       a73     bean
15    23     45       a73     bean
16     0      4       a73     bean
17     3     12       a73 cucumber
18    22     41       a73 cucumber
19    15     30       a73 cucumber
20    32     51       a73 cucumber
21     3      7       a73 cucumber
attach(data1)
y <- cbind(count, sample-count)
xtabs(~Orobanche+extract)
         extract
Orobanche bean cucumber
      a73    5        5
      a75    5        6
model1 <- glm(y~Orobanche*extract, binomial)
summary(model1)

Call:
glm(formula = y ~ Orobanche * extract, family = binomial)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.4122     0.1842  -2.238   0.0252 *
Orobanchea75                 -0.1459     0.2232  -0.654   0.5132
extractcucumber               0.5401     0.2498   2.162   0.0306 *
Orobanchea75:extractcucumber  0.7781     0.3064   2.539   0.0111 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

Number of Fisher Scoring iterations: 4
```

# Overdispersion!

```
model1 <- glm(y~Orobanche*extract, quasibinomial)
model2 <- update(model1, ~. -Orobanche:extract, quasibinomial)
anova(model1, model2, test="Chi") # WRONG

Analysis of Deviance Table

Model 1: y ~ Orobanche * extract
Model 2: y ~ Orobanche + extract
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        17     33.278
2        18     39.686 -1  -6.4081  0.06357 .
---
anova(model1, model2, test="F") # CORRECT!

Analysis of Deviance Table

Model 1: y ~ Orobanche * extract
Model 2: y ~ Orobanche + extract
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1        17     33.278
2        18     39.686 -1  -6.4081 3.4418 0.08099 .
---
```

# Proceed to model simplification!

```
anova(model2)
Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: y

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                         20     98.719
Orobanche  1    2.544        19     96.175
extract    1   56.489        18     39.686
anova(model2, test="F")
Analysis of Deviance Table

Model: quasibinomial, link: logit
```

```
Response: y

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                         20     98.719
Orobanche  1    2.544        19     96.175  1.1954    0.2887
extract    1   56.489        18     39.686 26.5412 6.692e-05 ***
---
```

```
model3 <- update(model2, ~. -Orobanche)
anova(model2, model3, test="F")
Analysis of Deviance Table

Model 1: y ~ Orobanche + extract
Model 2: y ~ extract
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        18     39.686
2        19     42.751 -1   -3.065 1.4401 0.2457
```

```
summary(model3)

Call:
glm(formula = y ~ extract, family = quasibinomial)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.5122     0.1531  -3.345   0.0034 **
extractcucumber   1.0574     0.2118   4.992 8.09e-05 ***
---

(Dispersion parameter for quasibinomial family taken to be 2.169821)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 42.751  on 19  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
1/(1+exp(-(-0.5122)))
[1] 0.3746779                          # Mean germination rate for BEAN = 37.46%
1/(1+exp(-(-0.5122+1.0574)))
[1] 0.6330212                          # Mean germination rate for CUCUMBER = 63.30%
```

# Ex 4, ANCOVA with Binomial data

- **flowered, number** = y-variable (how many were flowering out of total number)
- **variety** = categorical x variable (A,B,C,D,E,F)
- **dose** = continuous x variable (0,1,4,8,16,32)

**<< R printout >>**

```
data1 <- read.table("C:\\STAT999\\RBOOK\\flowering.txt",header=T)
dim(data1)
[1] 30  4
data1
   flowered number dose variety
1         0     12    1       A
2         0     17    4       A
3         4     10    8       A
4         9     11   16       A
5        10     10   32       A
6         0     17    1       B
7         3     15    4       B
8         6     12    8       B
9         9     10   16       B
10        9     18   32       B
11        2     14    1       C
12        1     15    4       C
13        3     17    8       C
14        5     20   16       C
15       15     15   32       C
16        2     18    1       D
17        3     19    4       D
18       15     28    8       D
19       19     26   16       D
```

```
20       21     27   32       D
21        0     13    1       E
22        0     15    4       E
23        3     19    8       E
24       15     20   16       E
25       17     17   32       E
26        0     11    0       A
27        1     12    0       B
28        0     17    0       C
29        1     15    0       D
30        0     10    0       E
attach(data1)
y <- cbind(flowered, number-flowered)

model1 <- glm(y~dose*variety,binomial)
summary(model1)

Call:
glm(formula = y ~ dose * variety, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.59165    1.03215  -4.449 8.64e-06 ***
dose           0.41262    0.10033   4.113 3.91e-05 ***
varietyB       3.06197    1.09317   2.801 0.005094 **
varietyC       1.23248    1.18812   1.037 0.299576
varietyD       3.17506    1.07516   2.953 0.003146 **
varietyE      -0.71466    1.54849  -0.462 0.644426
dose:varietyB -0.34282    0.10239  -3.348 0.000813 ***
dose:varietyC -0.23039    0.10698  -2.154 0.031274 *
dose:varietyD -0.30481    0.10257  -2.972 0.002961 **
dose:varietyE -0.00649    0.13292  -0.049 0.961057
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 303.350  on 29  degrees of freedom
Residual deviance:  51.083  on 20  degrees of freedom
AIC: 123.55

Number of Fisher Scoring iterations: 5

> model11 <- glm(y~dose*variety,quasibinomial)
> summary(model11)

Call:
glm(formula = y ~ dose * variety, family = quasibinomial)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.59165    1.56314  -2.937  0.00814 **
dose           0.41262    0.15195   2.716  0.01332 *
varietyB       3.06197    1.65555   1.850  0.07922 .
varietyC       1.23248    1.79934   0.685  0.50123
varietyD       3.17506    1.62828   1.950  0.06534 .
varietyE      -0.71466    2.34511  -0.305  0.76371
dose:varietyB -0.34282    0.15506  -2.211  0.03886 *
dose:varietyC -0.23039    0.16201  -1.422  0.17043
dose:varietyD -0.30481    0.15534  -1.962  0.06380 .
dose:varietyE -0.00649    0.20130  -0.032  0.97460
---
```

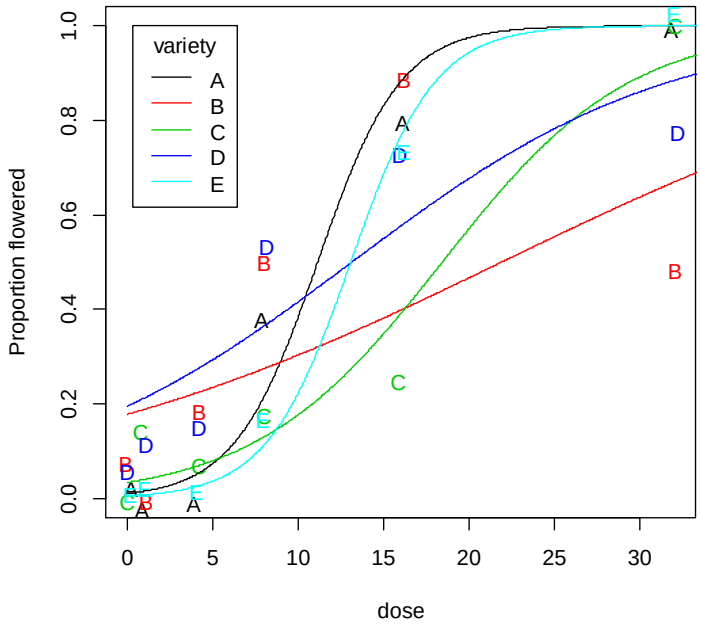
# # When “quasibinomial” is used, p-values become very different.

```
(Dispersion parameter for quasibinomial family taken to be 2.293557)

    Null deviance: 303.350  on 29  degrees of freedom
```

```
Residual deviance:  51.083  on 20  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

# # Plot

```
p_f <- flowered/number
sp_f <- split(p_f, variety)
sdose <- split(dose,variety)
plot(dose,p_f,type="n",ylab="Proportion flowered")
points(jitter(sdose[[1]]),jitter(sp_f[[1]]),pch="A",col=1)
points(jitter(sdose[[2]]),jitter(sp_f[[2]]),pch="B",col=2)
points(jitter(sdose[[3]]),jitter(sp_f[[3]]),pch="C",col=3)
points(jitter(sdose[[4]]),jitter(sp_f[[4]]),pch="D",col=4)
points(jitter(sdose[[5]]),jitter(sp_f[[5]]),pch="E",col=5)

xx <- seq(0,35,0.05)
vv1 <- rep("A",length(xx))
yy1 <- predict(model1,list(variety=vv1,dose=xx),type="response")
lines(xx,yy1,col=1)
vv2 <- rep("B",length(xx))
yy2 <- predict(model1,list(variety=vv2,dose=xx),type="response")
lines(xx,yy2,col=2)
vv3 <- rep("C",length(xx))
yy3 <- predict(model1,list(variety=vv3,dose=xx),type="response")
lines(xx,yy3,col=3)
vv4 <- rep("D",length(xx))
yy4 <- predict(model1,list(variety=vv4,dose=xx),type="response")
lines(xx,yy4,col=4)
vv5 <- rep("E",length(xx))
yy5 <- predict(model1,list(variety=vv5,dose=xx),type="response")
lines(xx,yy5,col=5)
legend(locator(1),legend=c("A","B","C","D","E"),title="variety",lty=rep(1,5),col=1:5)
```




```
> tapply(p_f,list(dose,variety),mean)
           A          B          C          D         E
0  0.0000000 0.08333333 0.00000000 0.06666667 0.0000000
1  0.0000000 0.00000000 0.14285714 0.11111111 0.0000000
4  0.0000000 0.20000000 0.06666667 0.15789474 0.0000000
8  0.4000000 0.50000000 0.17647059 0.53571429 0.1578947
16 0.8181818 0.90000000 0.25000000 0.73076923 0.7500000
32 1.0000000 0.50000000 1.00000000 0.77777778 1.0000000
```

**# The reason why B & D look weird.**

# Ex 5, Binomial data with many categorical variables

- **species** = y-variable (Ao vs. Ag)
- **sun, height, perch, time** = Four categorical x variables\

**<< R printout >>**

```
data1 <- read.table("C:\\STAT999\\RBOOK\\lizards.txt",header=T)
dim(data1)
[1] 48  6
head(data1)                                          # case #1~24=Ao, case #25~48=Ag
    n   sun height  perch      time  species
1  20 Shade   High  Broad   Morning opalinus
2  13 Shade    Low  Broad   Morning opalinus
3   8 Shade   High Narrow   Morning opalinus
4   6 Shade    Low Narrow   Morning opalinus
5  34   Sun   High  Broad   Morning opalinus
6  31   Sun    Low  Broad   Morning opalinus
attach(data1)
xtabs(~species+perch)
          perch
species    Broad Narrow
  grahamii    12     12
  opalinus    12     12
sorted <- data1[order(species,sun,height,perch,time),]
head(sorted)
   n   sun height  perch      time  species
41 4 Shade   High  Broad Afternoon grahamii
33 1 Shade   High  Broad   Mid.day grahamii
25 2 Shade   High  Broad   Morning grahamii
43 3 Shade   High Narrow Afternoon grahamii
35 1 Shade   High Narrow   Mid.day grahamii
27 3 Shade   High Narrow   Morning grahamii
first <- sorted[1:24,]
first
    n   sun height  perch      time  species
41  4 Shade   High  Broad Afternoon grahamii
33  1 Shade   High  Broad   Mid.day grahamii
25  2 Shade   High  Broad   Morning grahamii
43  3 Shade   High Narrow Afternoon grahamii
35  1 Shade   High Narrow   Mid.day grahamii
27  3 Shade   High Narrow   Morning grahamii
42  0 Shade    Low  Broad Afternoon grahamii
34  0 Shade    Low  Broad   Mid.day grahamii
26  0 Shade    Low  Broad   Morning grahamii
44  1 Shade    Low Narrow Afternoon grahamii
36  0 Shade    Low Narrow   Mid.day grahamii
28  0 Shade    Low Narrow   Morning grahamii
45 10   Sun   High  Broad Afternoon grahamii
37 20   Sun   High  Broad   Mid.day grahamii
29 11   Sun   High  Broad   Morning grahamii
47  8   Sun   High Narrow Afternoon grahamii
39 32   Sun   High Narrow   Mid.day grahamii
31 15   Sun   High Narrow   Morning grahamii
46  3   Sun    Low  Broad Afternoon grahamii
38  4   Sun    Low  Broad   Mid.day grahamii
30  5   Sun    Low  Broad   Morning grahamii
48  4   Sun    Low Narrow Afternoon grahamii
40  5   Sun    Low Narrow   Mid.day grahamii
32  1   Sun    Low Narrow   Morning grahamii
names(first)[1] <- "Ag"
names(first)
[1] "Ag"      "sun"     "height"  "perch"   "time"    "species"
first <- first[,-6]
names(first)
[1] "Ag"     "sun"    "height" "perch"  "time"
data2 <- data.frame(sorted$n[25:48],first)
head(data2)
   sorted.n.25.48. Ag   sun height  perch      time
41               4  4 Shade   High  Broad Afternoon
33               8  1 Shade   High  Broad   Mid.day
25              20  2 Shade   High  Broad   Morning
43               5  3 Shade   High Narrow Afternoon
35               4  1 Shade   High Narrow   Mid.day
27               8  3 Shade   High Narrow   Morning
names(data2)[1] <- "Ao"
data2
   Ao Ag   sun height  perch      time
41  4  4 Shade   High  Broad Afternoon
```

```
33  8  1 Shade   High  Broad   Mid.day
25 20  2 Shade   High  Broad   Morning
43  5  3 Shade   High Narrow Afternoon
35  4  1 Shade   High Narrow   Mid.day
27  8  3 Shade   High Narrow   Morning
42 12  0 Shade    Low  Broad Afternoon
34  8  0 Shade    Low  Broad   Mid.day
26 13  0 Shade    Low  Broad   Morning
44  1  1 Shade    Low Narrow Afternoon
36  0  0 Shade    Low Narrow   Mid.day
28  6  0 Shade    Low Narrow   Morning
45 18 10   Sun   High  Broad Afternoon
37 69 20   Sun   High  Broad   Mid.day
29 34 11   Sun   High  Broad   Morning
47  8  8   Sun   High Narrow Afternoon
39 60 32   Sun   High Narrow   Mid.day
31 17 15   Sun   High Narrow   Morning
46 13  3   Sun    Low  Broad Afternoon
38 55  4   Sun    Low  Broad   Mid.day
30 31  5   Sun    Low  Broad   Morning
48  4  4   Sun    Low Narrow Afternoon
40 21  5   Sun    Low Narrow   Mid.day
32 12  1   Sun    Low Narrow   Morning

detach(data1)
rm(first,sorted)
attach(data2)
names(data2)
[1] "Ao"     "Ag"     "sun"    "height" "perch"  "time"
```

# # Ready now!

```
y <- cbind(Ao,Ag)
model1 <- glm(y~sun*height*perch*time,binomial)
model2 <- step(model1)
Start:  AIC=102.82
y ~ sun * height * perch * time

                         Df   Deviance    AIC
- sun:height:perch:time  1 2.1801e-10 100.82
<none                     3.5823e-10 102.82

Step:  AIC=100.82
y ~ sun + height + perch + time + sun:height + sun:perch + height:perch +
    sun:time + height:time + perch:time + sun:height:perch +
    sun:height:time + sun:perch:time + height:perch:time

                    Df Deviance     AIC
- sun:height:time    2   0.4416  97.266
- sun:perch:time     2   0.8101  97.634
- height:perch:time  2   3.2217 100.046
<none                  0.0000 100.824
- sun:height:perch   1   2.7088 101.533

Step:  AIC=97.27
y ~ sun + height + perch + time + sun:height + sun:perch + height:perch +
    sun:time + height:time + perch:time + sun:height:perch +
    sun:perch:time + height:perch:time

                    Df Deviance    AIC
- sun:perch:time     2   1.0713 93.896
<none                   0.4416 97.266
- height:perch:time  2   4.6476 97.472
- sun:height:perch   1   3.1113 97.936

Step:  AIC=93.9
y ~ sun + height + perch + time + sun:height + sun:perch + height:perch +
    sun:time + height:time + perch:time + sun:height:perch +
    height:perch:time

                    Df Deviance    AIC
- sun:time           2   3.3403 92.165
<none                   1.0713 93.896
- sun:height:perch   1   3.3016 94.126
- height:perch:time  2   5.7906 94.615
```

```
Step:  AIC=92.16
y ~ sun + height + perch + time + sun:height + sun:perch + height:perch +
    height:time + perch:time + sun:height:perch + height:perch:time

                    Df Deviance    AIC
<none                  3.3403 92.165
- sun:height:perch   1   5.8273 92.651
- height:perch:time  2   8.5418 93.366

> model9 <- glm(y~sun+height+perch+time, binomial)
> summary(model9)

Call:
glm(formula = y ~ sun + height + perch + time, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2079     0.3536   3.416 0.000634 ***
sunSun       -0.8473     0.3224  -2.628 0.008585 **
heightLow     1.1300     0.2571   4.395 1.11e-05 ***
perchNarrow  -0.7626     0.2113  -3.610 0.000306 ***
timeMid.day   0.9639     0.2816   3.423 0.000619 ***
timeMorning   0.7368     0.2990   2.464 0.013730 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.102  on 22  degrees of freedom
Residual deviance: 14.205  on 17  degrees of freedom
AIC: 83.029

Number of Fisher Scoring iterations: 4

> table(time)
time
Afternoon   Mid.day   Morning
        8         8         8
> time2 <- time
> levels(time2)[c(2,3)] <- "Other"
> table(time2)
time2
Afternoon     Other
        8        16
> model10 <- glm(y~sun+height+perch+time2, binomial)
> anova(model9,model10,test="Chi")
Analysis of Deviance Table

Model 1: y ~ sun + height + perch + time
Model 2: y ~ sun + height + perch + time2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        17     14.205
2        18     15.023 -1 -0.81863   0.3656
```

# # It's OK to reduce the levels of TIME

```
> summary(model10)

Call:
glm(formula = y ~ sun + height + perch + time2, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1595     0.3484   3.328 0.000874 ***
sunSun       -0.7872     0.3159  -2.491 0.012722 *
heightLow     1.1188     0.2566   4.360  1.3e-05 ***
perchNarrow  -0.7485     0.2104  -3.557 0.000375 ***
time2Other    0.8717     0.2611   3.338 0.000844 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.102  on 22  degrees of freedom
Residual deviance: 15.023  on 18  degrees of freedom
AIC: 81.847

Number of Fisher Scoring iterations: 4
```

# # Two species differ significantly in all the variables. There are no significant interactions.

```
> ftable(tapply(n,list(species,sun,height,perch,time),sum))
                          Afternoon Mid.day Morning

grahamii Shade High Broad           4       1       2
                    Narrow          3       1       3
                Low  Broad          0       0       0
                    Narrow          1       0       0
         Sun    High Broad         10      20      11
                    Narrow          8      32      15
                Low  Broad          3       4       5
                    Narrow          4       5       1
opalinus Shade High Broad           4       8      20
                    Narrow          5       4       8
                Low  Broad         12       8      13
                    Narrow          1       0       6
         Sun    High Broad         18      69      34
                    Narrow          8      60      17
                Low  Broad         13      55      31
                    Narrow          4      21      12
```

# # Doing it with "poisson"

```
> model1 <-
glm(n~sun+height+perch+time+species,poisson)
> summary(model1)

Call:
glm(formula = n ~ sun + height + perch + time + species, family = poisson)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.52293    0.16089   3.250  0.00115 **
sunSun         1.48684    0.10858  13.694  < 2e-16 ***
heightLow     -0.60659    0.08812  -6.884 5.83e-12 ***
perchNarrow   -0.45447    0.08640  -5.260 1.44e-07 ***
timeMid.day    1.07799    0.11695   9.218  < 2e-16 ***
timeMorning    0.59682    0.12579   4.745 2.09e-06 ***
speciesopalinus  1.17576    0.09919  11.853  < 2e-16 ***
---

 (Dispersion parameter for poisson family taken to be 1)

    Null deviance: 737.56  on 47  degrees of freedom
Residual deviance: 152.61  on 41  degrees of freedom
AIC: 329.86

Number of Fisher Scoring iterations: 5
```

# # Overdispersion!

```
> model1 <- glm(n~sun+height+perch+time+species,quasipoisson)
> plot(model1)
> summary(model1)

Call:
glm(formula = n ~ sun + height + perch + time + species, family = quasipoisson)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.5229     0.3157   1.656  0.10528
sunSun          1.4868     0.2131   6.978 1.77e-08 ***
heightLow      -0.6066     0.1729  -3.508  0.00111 **
perchNarrow    -0.4545     0.1695  -2.681  0.01054 *
timeMid.day     1.0780     0.2295   4.697 2.95e-05 ***
timeMorning     0.5968     0.2468   2.418  0.02014 *
speciesopalinus  1.1758     0.1946   6.041 3.79e-07 ***
---
```

```
(Dispersion parameter for quasipoisson family taken to be 3.85066)

    Null deviance: 737.56  on 47  degrees of freedom
Residual deviance: 152.61  on 41  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

# p-values become different.

```
> model3 <- update(model1,~.+sun:species)
```

```
> anova(model1,model3,test="F")
```

# CORRECT

```
Analysis of Deviance Table

Model 1: n ~ sun + height + perch + time + species
Model 2: n ~ sun + height + perch + time + species + sun:species
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        41     152.61
2        40     146.14  1   6.4724 1.6901  0.201
```

```
> anova(model1,model3,test="Chi")
```

# WRONG

```
Analysis of Deviance Table

Model 1: n ~ sun + height + perch + time + species
Model 2: n ~ sun + height + perch + time + species + sun:species
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        41     152.61
2        40     146.14  1   6.4724   0.1936
```

# Ex 6, Back to Basic Binomial vs. x

- **damage, no damage** = binomial y-variable (out of 6 trials)
- **temp** = x variable

**<< R printout >>**

```
> library(faraway)
> data(orings)
> dim(orings)
[1] 23  2
> orings
   temp damage
1    53      5
2    57      1
3    58      1
4    63      1
5    66      0
6    67      0
7    67      0
8    67      0
9    68      0
10   69      0
11   70      1
12   70      0
13   70      1
14   70      0
15   72      0
16   73      0
17   75      0
18   75      1
19   76      0
20   76      0
21   78      0
22   79      0
23   81      0
> attach(orings)
```

```
> model1 <- glm(cbind(damage, 6-damage)~temp, binomial)
```

# # Uses default "logit" link *log*{p/(1–p)}

```
> summary(model1)
```

```
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom # Model fits well
AIC: 33.675

Number of Fisher Scoring iterations: 6

plot(damage/6~temp,xlim=c(25,85),ylim=c(0,1),ylab="Prob of damage")
xx <- seq(25,85,0.1)
yy1 <- predict(model1,list(temp=xx))
yy2 <- 1/(1+exp(-yy1))
lines(xx,yy2,lty=1,col=1)
# OR
lines(xx,ilogit(model1$coef[1]+model1$coef[2]*xx),lty=1,col=1)
# For example, to predict prob at temp=31
> 1/(1+exp(-predict(model1, list(temp=31))))
        1
0.9930342
# OR
> ilogit(predict(model1, list(temp=31)))
        1
0.9930342
> model2 <- glm(cbind(damage, 6-damage)~temp, binomial(link=probit))
# Uses optional "probit" link, Φ-1(p)
> summary(model2)

Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(link = probit))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.59145    1.71055   3.269  0.00108 **
temp        -0.10580    0.02656  -3.984 6.79e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 18.131  on 21  degrees of freedom
AIC: 34.893

Number of Fisher Scoring iterations: 6

> lines(xx,pnorm(model2$coef[1]+model2$coef[2]*xx),lty=2,col=2)
# For example, to predict prob at temp=31
> pnorm(predict(model2, list(temp=31)))
        1
0.9895983

# Uses optional "cloglog" link, log{-log(1–p)}
> model3 <- glm(cbind(damage, 6-damage)~temp, binomial(link=cloglog))
```

```
> summary(model3)

Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(link = cloglog))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.86388    2.73668   3.970 7.20e-05 ***
temp        -0.20552    0.04561  -4.506 6.59e-06 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.029  on 21  degrees of freedom
AIC: 32.791

Number of Fisher Scoring iterations: 7

> predict(model3, list(temp=31))
       1
4.492617
```

# # For example, to predict prob at temp=31

```
> 1-exp(-exp(predict(model3, list(temp=31))))
1
1

> lines(xx, 1-exp(-exp(model3$coef[1]+model3$coef[2]*xx)),lty=4,col=4)
```