MLR by LM and GLM

```r
library(gamlss.data)
fitLM = lm(rent ~ area + location + bath + kitchen + cheating, data = rent99)
summary(fitLM)
fitGLM = glm(rent ~ area + location + bath + kitchen + cheating, data = rent99)
summary(fitGLM)
```

$Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$

```
##
## Call:
## lm(formula = rent ~ area + location + bath + kitchen + cheating,
##     data = rent99)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633.41  -89.17   -6.26   82.96 1000.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.9733    11.6549  -1.885   0.0595 .
## area          4.5788     0.1143  40.055  < 2e-16 ***
## location2    39.2602     5.4471   7.208 7.14e-13 ***
## location3   126.0575    16.8747   7.470 1.04e-13 ***
## bath1        74.0538    11.2087   6.607 4.61e-11 ***
## kitchen1    120.4349    13.0192   9.251  < 2e-16 ***
## cheating1   161.4138     8.6632  18.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.2 on 3075 degrees of freedom
## Multiple R-squared:  0.4504, Adjusted R-squared:  0.4494
## F-statistic:   420 on 6 and 3075 DF,  p-value: < 2.2e-16
##
##
```

$\hat{\sigma}^2 = \frac{SSE}{n-p}$

$\hat{\sigma}$

$n-p$

$R^2 = 1 - \frac{SSE}{SST}$

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

H_1: at least one $\neq 0$

```
## Call:
## glm(formula = rent ~ area + location + bath + kitchen + cheating,
##     data = rent99)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -633.41   -89.17    -6.26    82.96  1000.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.9733    11.6549  -1.885   0.0595 .
## area          4.5788     0.1143  40.055  < 2e-16 ***
## location2    39.2602     5.4471   7.208 7.14e-13 ***
## location3   126.0575    16.8747   7.470 1.04e-13 ***
## bath1        74.0538    11.2087   6.607 4.61e-11 ***
## kitchen1    120.4349    13.0192   9.251  < 2e-16 ***
## cheating1   161.4138     8.6632  18.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 21079.53)
##
```

sqrt(diag($(X^T X)^{-1} \sigma^2$))

$\hat{\beta}$ $\quad$ $\widehat{SD}(\hat{\beta})$

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$t_j = \frac{\hat{\beta}_j - 0}{\widehat{SD}(\hat{\beta}_j)}$ ← for LM $\sim t_{n-1}$

also called Z and $Z \approx N(0,1)$

asymptotically

$\hat{\sigma}^2$

Wald teststatistic: $Z^2 \approx \chi^2_1$ here

SSE SST $n-1$ $n-p$

```
##     Null deviance: 117945363  on 3081  degrees of freedom
## Residual deviance:  64819547  on 3075  degrees of freedom
## AIC: 39440
##
## Number of Fisher Scoring iterations: 2
```

$= -2 \ln L(\hat\beta) + 2p$

GLM - Binomial regression with logit-link

$G = 8$ → df for saturated

8 covariate patterns

```r
library(investr)
fitgrouped = glm(cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)
summary(fitgrouped)
```

```
##
## Call:
## glm(formula = cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

$\text{sqrt}(\text{diag}(F^{-1}(\hat\beta)))$

$\hat\beta$ $\quad \widehat{SD}(\hat\beta)$

$Z = \frac{\hat\beta_j - 0}{\widehat{SD}(\hat\beta_j)} \approx N(0,1)$ under H_0

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$(P(\text{success})/P(\text{failure}))$

odds change with e^{β_j} when x_j increase to $x_j + 1$ while all else is kept unchanged x's

ϕ

Compare saturated to model with intercept only

$-2(l_{\text{candidate}} - l_{\text{saturated}}) \approx \chi^2_6$

G # param in candidate

$6 = 8 - 2$

GLM - Poisson regression with log-link

```r
crab = read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/crab.txt")
colnames(crab) = c("Obs", "C", "S", "W", "Wt", "Sa")
crab = crab[, -1]  #remove column with Obs
crab$C = as.factor(crab$C)
model3 = glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum"))
summary(model3)
```

```
##
## Call:
## glm(formula = Sa ~ W + C, family = poisson(link = log), data = crab,
##     contrasts = list(C = "contr.sum"))
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0415  -1.9581  -0.5575   0.9830   4.7523
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.92089    0.56010  -5.215 1.84e-07 ***
## W            0.14934    0.02084   7.166 7.73e-13 ***
## C1           0.27085    0.11784   2.298   0.0215 *
## C2           0.07117    0.07296   0.975   0.3294
## C3          -0.16551    0.09316  -1.777   0.0756 .
## ---
## 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: 559.34  on 168  degrees of freedom
## AIC: 924.64
##
## Number of Fisher Scoring iterations: 6
```

$\hat{\mu}$ change with factor $\exp(\hat{\beta}_j)$
when x_j go to $x_j + 1$

$\hat{\beta}_{C4} = -(\hat{\beta}_{C1} + \hat{\beta}_{C2} + \hat{\beta}_{C3})$

ϕ

Categorical regression, nominal model

```r
# data from Agresti (2015), section 6, with use of the VGAM packages
data = "http://www.stat.ufl.edu/~aa/glm/data/Alligators.dat"
ali = read.table(data, header = T)
attach(ali)
y.data = cbind(y2, y3, y4, y5, y1)
x.data = model.matrix(~size + factor(lake), data = ali)
library(VGAM)
# We fit a multinomial logit model with fish (y1) as the reference category:
fit.main = vglm(cbind(y2, y3, y4, y5, y1) ~ size + factor(lake), family = multinomial,
    data = ali)
summary(fit.main)
pchisq(deviance(fit.main), df.residual(fit.main), lower.tail = FALSE)
```

$$\pi_{ir} = \frac{\exp(\eta_{ir})}{1 + \sum_{s=1}^{c} \exp(\eta_{is})}$$

```
##
## Call:
## vglm(formula = cbind(y2, y3, y4, y5, y1) ~ size + factor(lake),
##     family = multinomial, data = ali)
##
##
## Pearson residuals:
##   log(mu[,1]/mu[,5]) log(mu[,2]/mu[,5]) log(mu[,3]/mu[,5])
## 1             0.0953           0.028205           -0.54130
## 2            -0.5082           0.003228            0.66646
## 3            -0.3693          -0.461102           -0.42005
## 4             0.4125           0.249983            0.19772
## 5            -0.5526          -0.191149            0.07215
## 6             0.6500           0.110694           -0.02784
```

$\exp(1.458)$ = factor increase in OR if size change from 0 to 1

```
## 7              0.6757           0.827737            0.79863
## 8             -1.3051          -0.802694           -0.69525
##   log(mu[,4]/mu[,5])
## 1            -0.7268
## 2             1.2589
## 3             1.8347
## 4            -1.3779
## 5             0.2790
## 6            -0.2828
## 7            -0.3081
## 8             0.4629
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept):1      -3.2074     0.6387  -5.021 5.13e-07 ***
## (Intercept):2      -2.0718     0.7067  -2.931 0.003373 **
## (Intercept):3      -1.3980     0.6085  -2.297 0.021601 *
## (Intercept):4      -1.0781     0.4709  -2.289 0.022061 *
## size:1              1.4582     0.3959   3.683 0.000231 ***
## size:2             -0.3513     0.5800  -0.606 0.544786
## size:3             -0.6307     0.6425  -0.982 0.326296
## size:4              0.3316     0.4482   0.740 0.459506
## factor(lake)2:1     2.5956     0.6597   3.934 8.34e-05 ***
## factor(lake)2:2     1.2161     0.7860   1.547 0.121824
## factor(lake)2:3    -1.3483     1.1635  -1.159 0.246529
## factor(lake)2:4    -0.8205     0.7296  -1.125 0.260713
## factor(lake)3:1     2.7803     0.6712   4.142 3.44e-05 ***
## factor(lake)3:2     1.6925     0.7804   2.169 0.030113 *
## factor(lake)3:3     0.3926     0.7818   0.502 0.615487
## factor(lake)3:4     0.6902     0.5597   1.233 0.217511
## factor(lake)4:1     1.6584     0.6129   2.706 0.006813 **
## factor(lake)4:2    -1.2428     1.1854  -1.048 0.294466
## factor(lake)4:3    -0.6951     0.7813  -0.890 0.373608
## factor(lake)4:4    -0.8262     0.5575  -1.482 0.138378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 17.0798 on 12 degrees of freedom
##
## Log-likelihood: -47.5138 on 12 degrees of freedom
##
## Number of iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1'
##
## Reference group is level  5  of the response
## [1] 0.1466189
```

$$\left[\frac{\dfrac{P(\text{invertebrate} \mid \text{size } 1, \text{lake}=a)}{P(\text{fish} \mid \cdots)}}{\dfrac{P(\text{inv} \mid \text{size}=0, \text{lake}=a)}{P(\text{fish} \mid \cdots)}}\right]$$

as before

$\hat\beta$

β_{01}

β_{04}

β_{11}

β_{14}

$$\hat\pi_{ir} = \frac{\exp(x_i^T\hat\beta_r)}{1+\sum_{s=1}^{c}\exp(x_i^T\hat\beta_s)}$$

$$\pi_{i\,c+1} = 1-\hat\pi_{i1}-\dots-\hat\pi_{ic} = \frac{1}{1+\sum_{s=1}^{c}\exp(x_i^T\hat\beta_s)}$$

$$\left(\ln\left(\frac{\pi_{ia}}{\pi_{ib}}\right) = x_i^T(\beta_a-\beta_b)\right)$$

η's

$-2(\ln L_{cen} - \ln L_{sat})$ as before

$c \cdot p = 4 \cdot 5 = 20$

$G = 32$

4 lakes × 2 sizes = 8

× 5 foods: 4 → 32

$\sum$ prob = 1

Not converged

p-value from deviance test → not reject H_0

so ok model

Easiest to interpret by calc. $\hat{P}(y=1)$ for diff val of x_1, x_2 and compare

life

ses

40 obs, 2 covs (continuous)

Categorical regresion, ordinal model

```
# Read mental health data from the web:
library(knitr)
data = "http://www.stat.ufl.edu/~aa/glm/data/Mental.dat"
mental = read.table(data, header = T)
library(VGAM)
# We fit a cumulative logit model with main effects of 'ses' and 'life':
fit.imp = vglm(impair ~ life + ses, family = cumulative(parallel = T), data = mental)
# parallell=T gives proportional odds structure - only intercepts differ
summary(fit.imp)
```

$\pi_{i1} = F(\eta_{i1})$, $\pi_{ir} = F(\eta_{ir}) - F(\eta_{i\,r-1})$ where $F(\eta) = \frac{e^\eta}{1+e^\eta}$ ← logistic cdf

$\eta_{ir} = \theta_r + x_i^T\beta$

$\text{logit}(\hat{P}(y_i \le j)) = \hat{\theta}_j - 0.32\, x_{i1} + 1.11\, x_{i2}$

```
##
## Call:
## vglm(formula = impair ~ life + ses, family = cumulative(parallel = T),
##     data = mental)
##
##
## Pearson residuals:
##                     Min      1Q  Median     3Q   Max
## logit(P[Y<=1]) -1.568 -0.7048 -0.2102 0.8070 2.713
## logit(P[Y<=2]) -2.328 -0.4666  0.2657 0.6904 1.615
## logit(P[Y<=3]) -3.688  0.1198  0.2039 0.4194 1.892
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.2819     0.6231  -0.452  0.65096
## (Intercept):2   1.2128     0.6511   1.863  0.06251 .
## (Intercept):3   2.2094     0.7171   3.081  0.00206 **
## life           -0.3189     0.1194  -2.670  0.00759 **
## ses             1.1112     0.6143   1.809  0.07045 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Residual deviance: 99.0979 on 115 degrees of freedom
##
## Log-likelihood: -49.5489 on 115 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##      life       ses
## 0.7269742 3.0380707
```

replace β_0's

θ_1 θ_2 θ_3

common β's

as before

$c+1 = 4$ → $c + k = 3 + 2 = 5$ params

$\eta_{ir} = \theta_r + x_i^T\beta = \underbrace{\text{logit}(P(Y_i \le r))}_{\text{"our } g(\mu_i) \text{ here"}}$

40 obs & 3 responses = 120, 5 param estimated } df = 115

ses is low or high

exp(1.11)

[Given any life score, at a high SES level the est odds of mental impairment below any fixed level is 3 times the estimated odds at the low SES level

Now it is rather complex to interpret!

LMM - random intercept and slope

$Y_{ij} = \beta_0 + \beta_1 x_{i1} + \gamma_{0i} + \gamma_{1i} \cdot x_{i1} + \varepsilon_{ij}$

```r
library(lme4)
fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
summary(fm1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.09   24.740
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.825  36.838
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

$\gamma_{0i} + \gamma_{1i} x_{i1}$

$0.07 \cdot \tau_0 \cdot \tau_1$

$Q = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$

$\hat{\sigma}^2$

$\hat{\beta}$

$\approx N(0,1)$ under H_0

$\leftarrow Corr(\hat{\beta}_0, \hat{\beta}_1)$

$\tau_{01} = Cov(\gamma_{0i}, \gamma_{1i})$

$Corr = \frac{\tau_{01}}{\tau_0 \cdot \tau_1}$

GLMM - random intercept and slope Poisson

```r
library("AED")
data(RIKZ)
library(lme4)
fitRI = glmer(Richness ~ NAP + (1 + NAP | Beach), data = RIKZ, family = poisson(link = log))
summary(fitRI)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Richness ~ NAP + (1 + NAP | Beach)
##    Data: RIKZ
##
##      AIC      BIC   logLik deviance df.resid
##    218.7    227.8   -104.4    208.7       40
```

= 45 obs − 5 par. est

$\beta_0, \beta_1, \tau_0^2, \tau_1^2, \tau_{01}$

$-2 \log Lik + 2p$

$-2 \log Lik + \ln n \cdot p$

```
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.35846 -0.51129 -0.21846  0.09802  2.45384
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  Beach  (Intercept) 0.2630   0.5128
##         NAP         0.0891   0.2985   0.18
## Number of obs: 45, groups:  Beach, 9
```

as for LMM

```
##
## Fixed effects:
```

as for Poisson GLM

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6942     0.1868   9.071  < 2e-16 ***
## NAP          -0.6074     0.1374  -4.421 9.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##     (Intr)
## NAP 0.121
```

$\text{Corr}(\hat\beta_0, \hat\beta_1) \leftarrow (F^{-1}(\hat\beta))_{\text{scaled}}$

Exam and exam preparation

We take look at the information posted at Blackboard, and the relevant exams are found on the bottom of each module page.

Dates for supervision are also found on Bb.

After TMA4315

What is next in the spring semester?

For the 4th year student

- TMA4250 Spatial statistics
- TMA4268 Statistical learning
- TMA4275 Survival analysis
- TMA4300 Computational statistics
- KLMED8005 Analysis of repeated measurements
- SMED8002 Epidemiology 2
- TDT4300 Datavarehus og datagruvedrift
- TDT4173 Machine learning and case-based reasoning (Big overlap with TMA4268)
- NEVR3004 Neural networks (in the brain)

For the 5th year student

- MA8701 General statistical models Phd course with selected topics relevant for statistical learning and inference.

Also, for the autumn of 2019 the Deep learning course at IDI which up to now was 3.75STP is planned to be an ordinary 7.5STP course.