The linear model and extensions

The linear model is arguably one of the most important ideas in statistics, both in its own right and as a component of other models.

$$
\begin{aligned}
Y &\sim \text{Normal}(\mu, \sigma^2) \\
\mu &= X\beta
\end{aligned}
$$

- assumes:
  - independence
  - linearity
  - constant variance
  - normality
  - input variables measured without error
- even when assumptions are violated, they're often OK *asymptotically* and/or the results may still be *unbiased* (just less efficient)
- transformations can help (Box-Cox) (but see O'Hara and Kotze (2010), Warton and Hui (2011))
- random-variable format ($\sim$) and matrix format (with the *design matrix* X) may be unfamiliar, but are extremely useful
- the model can be solved computationally in a few very standard linear algebra steps. This means it can be solved very efficiently by standard libraries (e.g. optimized BLAS); it can also be decomposed and solved out-of-memory (e.g. the biglm package)
- expected responses are a linear function of the *predictor variables* (which may or may not be the same as the *input variables*)
- you can use `model.matrix()` to create the design matrix by specifying the formula and the data

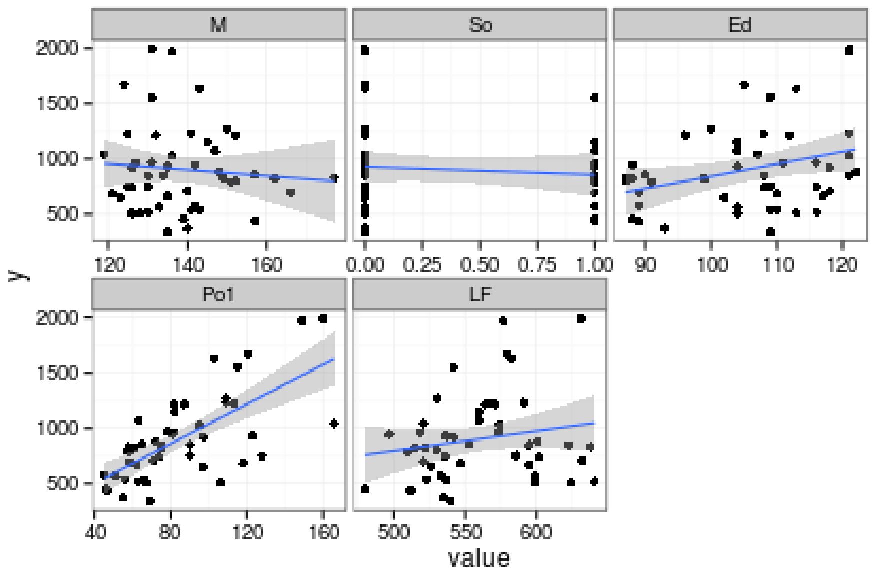
Linear regression

- $\mu = \beta_0 + \beta_1 x$
- formula: `y~x` (or `y~1+x`)
- Design matrix:

$$
\begin{pmatrix}
1 & x_1 \\
1 & x_2 \\
1 & x_3 \\
\dots & \dots
\end{pmatrix}
$$

US crime data: "The variables seem to have been rescaled to convenient numbers." (M=percentage of males, So=indicator for southern state, Ed=mean years of schooling, Po1=police expenditure in 1960, LF=labour force participation rate)

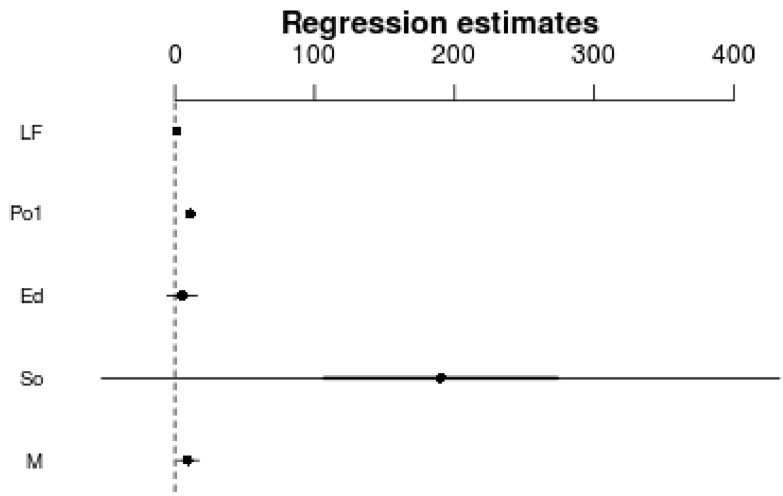
```
library(reshape2)
data(UScrime,package="MASS")
USsub <- subset(UScrime,select=c(y,M,So,Ed,Po1,LF))
mUS <- melt(USsub,id.var="y")
g0 <- ggplot(mUS,aes(x=value,y=y))+
  facet_wrap(~variable,scale="free_x")+
      geom_point()
g0 + geom_smooth(method="lm")
```



(Note limitation on `ggplot` linear models)

(Note difference between marginal [univariate] and multivariate models)

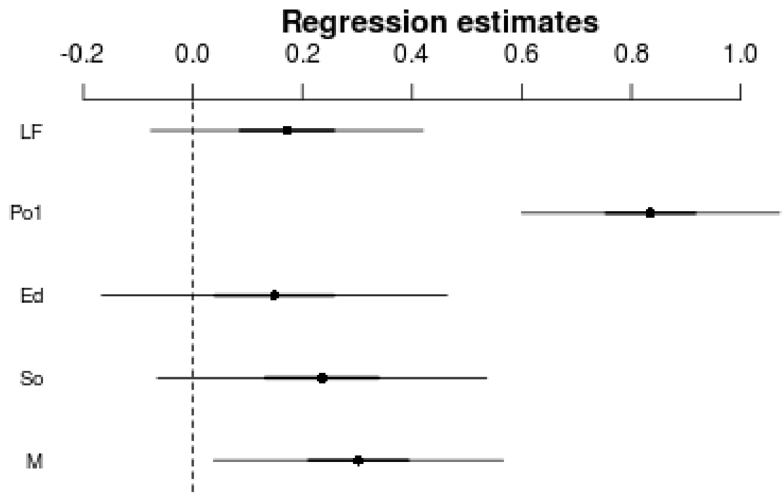
```
lm1 <- lm(y~.,data=USsub)
library(coefplot2)
coefplot2(lm1)
```




```r
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ ., data = USsub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -498   -161     37    125    550
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2835.10     899.00   -3.15    0.003 **
## M               9.29       4.15    2.24    0.031 *
## So            190.09     123.76    1.54    0.132
## Ed              5.14       5.57    0.92    0.361
## Po1            10.87       1.57    6.95    2e-08 ***
## LF              1.64       1.21    1.35    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257 on 41 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.558
## F-statistic: 12.6 on 5 and 41 DF,  p-value: 1.88e-07
```

```
lm2 <- update(lm1,data=as.data.frame(scale(USsub)))
coefplot2(lm2)
```



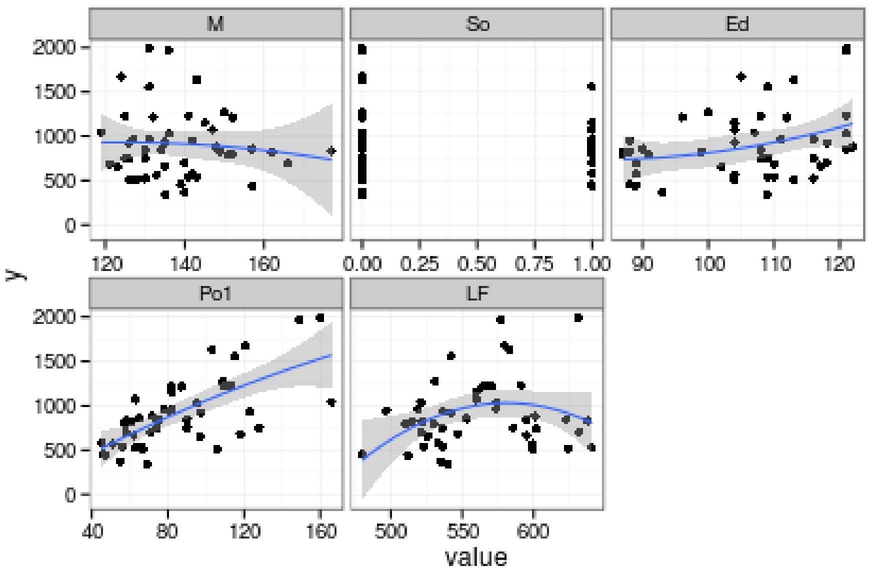
Polynomial regression:

- $\mu = \sum_{i=0}^{n} \beta_i x_i^n$
- `y~x+I(x^2)` (or `~poly(x,2)` [orthogonal polynomial] or `~poly(x,2,raw=TRUE)`)
- Design matrix:

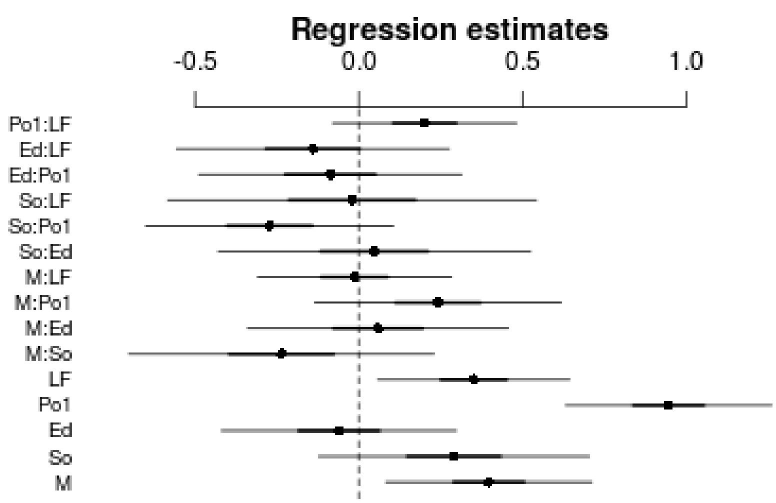
$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots \\ 1 & x_2 & x_2^2 & \dots \\ 1 & x_3 & x_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

- Here (especially if we specify the model with `poly()` we have a single *input* variable x, but multiple *predictor* variables (x, x^2, ...))
- Polynomial models beyond quadratics are probably a bad idea (unstable). Consider splines/GAMs instead (see below).

```
g0 + geom_smooth(method="lm",formula=y~poly(x,2))
```




```
lm2P <- update(lm2,.~.^2)
coefplot2(lm2P)
```



ANOVA

- Treatment separately from linear regression is really a historical accident
- $\mu = \beta_0 + \beta_1 I(x=2) + \beta_2 I(x=3) + \ldots$
- `y~f`
- Design matrix (if first observations are in level 1, 2, 2, 3 respectively)

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \ldots & \ldots & \end{pmatrix}$$

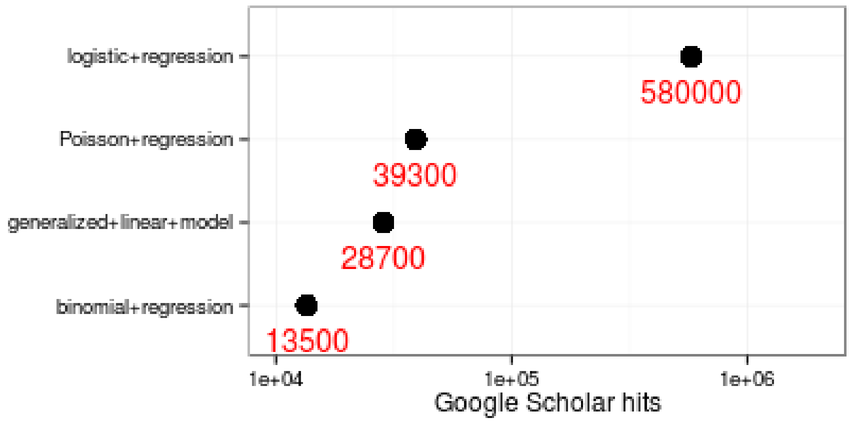
- **Contrasts** determine the translation of input variables into dummy (0/1) predictor variables, e.g. treatment (default: differences from baseline) vs. sum-to-zero (differences from mean in a balanced design)
- Interactions are easy to set up (but possibly hard to understand)
  - 'differences in differences'; e.g. the difference among regions in effects of government spending on phosphorus trends of ver time is a (region $\times$ money $\times$ time) interaction. In a before-after-control-impact treatment we are looking at the difference between (after-before) between control and impact sites
  - interpretation of main effects **depends on presence of interactions** ("principle of marginality": (Venables 1998)); where is the zero/baseline level? What are the contrasts? As a general rule, avoid "type III sums of squares" issues by centering data (Schielzeth 2010)

Generalized linear models

$$Y \sim F(g^{-1}(\mu), \phi)$$
$$\mu = X\beta$$

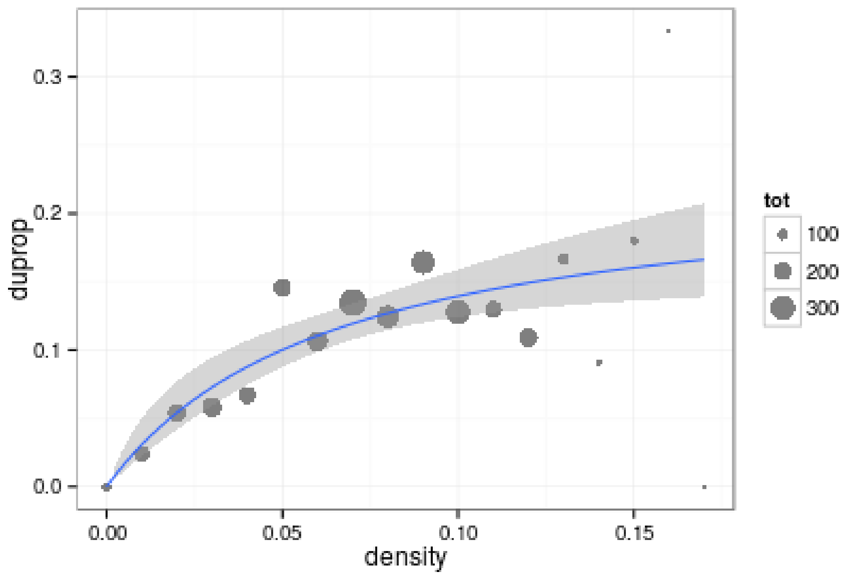
where F is an exponential family probability distribution (e.g. binomial, Poisson, Gamma) with a known mean-variance relationship; g is a *link function* (log, logit, probit ...) * logistic regression is (by far) the most common, followed by Poisson regression

These data were scraped from Google Scholar hits on the relevant search terms.



- iteratively reweighted least squares
- extensions: bias-reduced, Tweedie, negative binomial, zero-inflated/hurdle
- example: from [@Tiwari+2005]

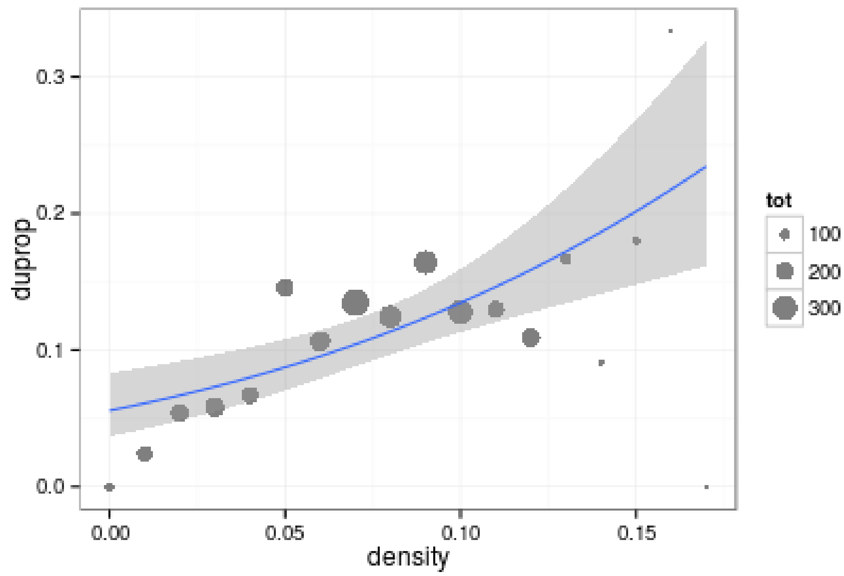
```
dat <- read.csv("data/dufemalepers.csv")
dat <- transform(dat,
                 tot=du+notdu,
                 duprop=du/(du+notdu))
ggplot(dat,aes(x=density,y=duprop))+geom_point(aes(size=tot),alpha=0.5)+
    geom_smooth(data=subset(dat,duprop>0),
                method="glm",formula=y~I(1/x),aes(weight=tot),
                family=quasibinomial(link="inverse"),
                fullrange=TRUE)
```



Generalized additive models

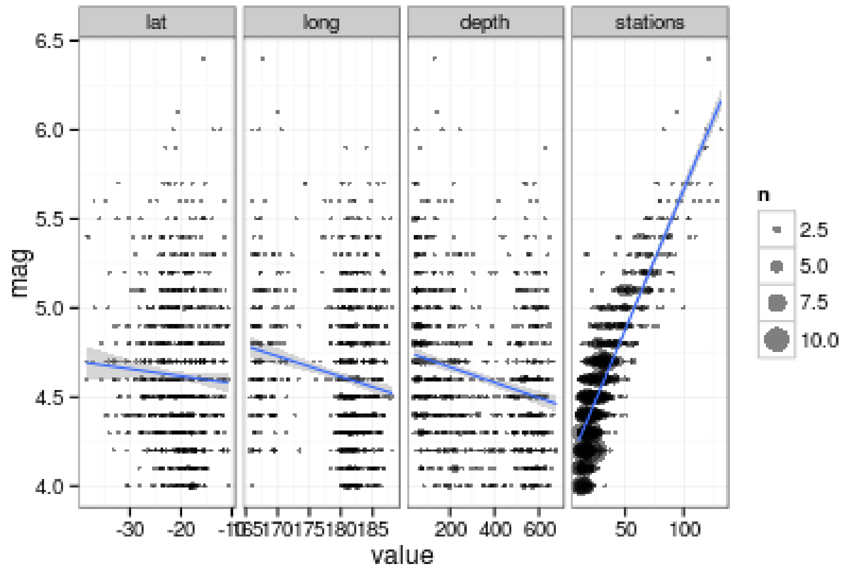
- allow splines: generalized additive models
- really still “just” linear models
- model complexity (number of knots); can be chosen by AIC
- `splines` package: `ns`, `bs`, `periodicSpline`; specify input variable and number of knots (knot placement is done automatically)
- or *smoothing splines* (`mgcv` package); use lots of knots, shrink via penalization (generalized cross-validation)
- multidimensional splines, e.g. tensor product ...
- highly efficient – can model small-scale spatial variation (vs correlation, see below)
- see Wood (2006)

```
ggplot(dat,aes(x=density,y=duprop))+geom_point(aes(size=tot),alpha=0.5)+
    geom_smooth(method="gam",
                aes(weight=tot),
                family=quasibinomial,
                fullrange=TRUE)
```

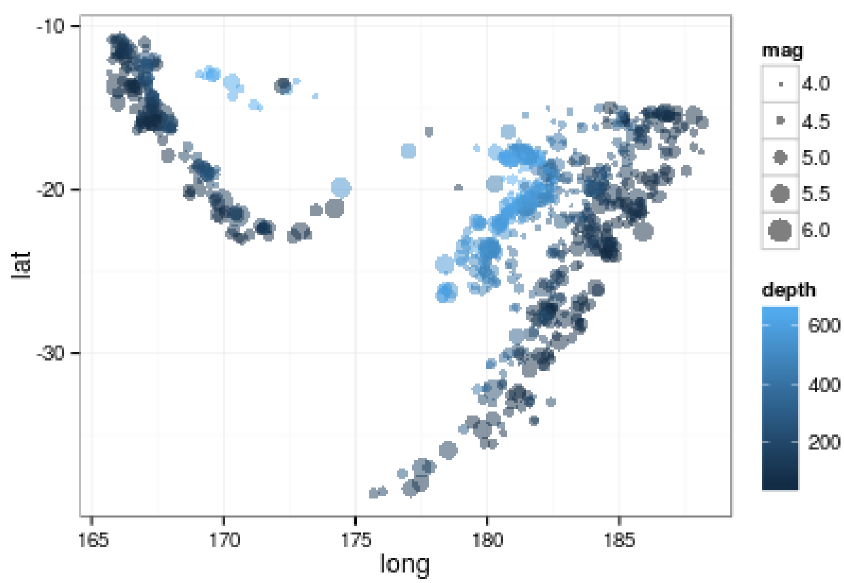


(GAM doesn't actually do very well here - might possibly be able to pin it at zero, but that would be difficult . . .)

```
mquake <- melt(quakes,id.var="mag")
ggplot(mquake,aes(x=value,y=mag))+
    facet_grid(.~variable,scale="free_x")+
  stat_sum(alpha=0.5,aes(size=..n..))+
  geom_smooth(method="gam")
```




```
ggplot(quakes,aes(x=long,y=lat))+
  geom_point(aes(size=mag,colour=depth),alpha=0.5)
```



Generalized least squares (correlation and heteroscedasticity)

$$
\begin{aligned}
Y &\sim MVN(\mu, \Sigma) \\
\mu &= X\beta \\
\Sigma &= f(\theta)
\end{aligned}
$$

Σ is the variance-covariance matrix of the residuals.

- heteroscedasticity structures: power, exponential, differing by stratum …
  - generalized least squares
  - Σ is diagonal; $\sigma_i^2 = f(x_i, \theta)$
  - `f()` can be anything, but chosen to be positive
- correlation structures: temporal, spatial, phylogenetic
  - Σ is no longer diagonal: in particular, specify *correlation* structure in terms of θ, e.g. $\rho_{ij} = (t_i - t_j)^{-\theta}$ (AR1 structure)
  - temporal, evenly spaced (ARMA)
  - temporal, uneven sampling (`corCAR1`=exponential decay)
  - spatial: linear, Gaussian, exponential, etc.
  - may measure spatial distance according to great-circle distance (`ramps` package)
- phylogenetic: Brownian, Ornstein-Uhlenbeck …
- R: `gls`

Nonlinear least squares

- relax linearity: nonlinear least squares
- lose almost all of the computational advantages
- need to know gory details of optimization algorithms
- starting values!

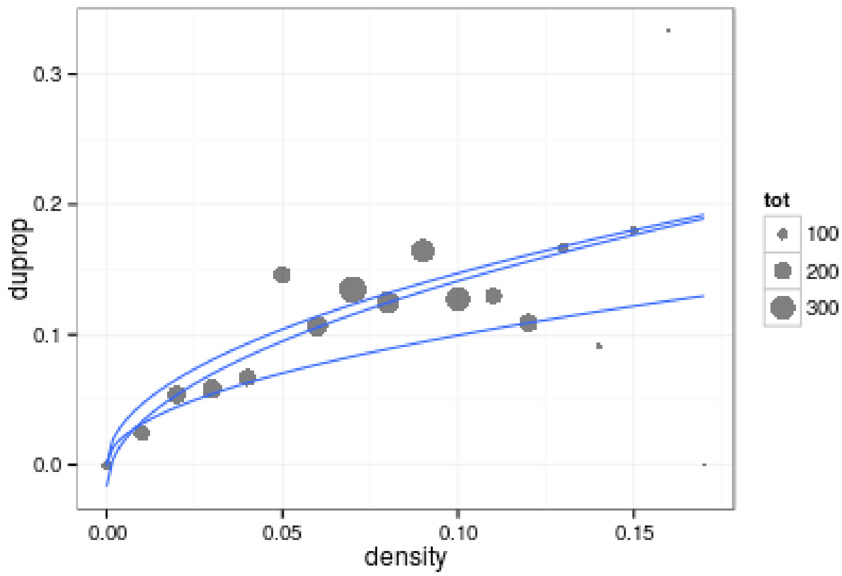
Mixed models

- relax independence (groups): mixed models
- random effects defined by group membership ("G-side")
- design matrix for random effects: $X\beta + Zu$
- penalization on u, with automatically determined penalty

Quantile regression

*

```r
ggplot(dat,aes(x=density,y=duprop))+geom_point(aes(size=tot),alpha=0.5)+
stat_quantile(formula=y~sqrt(x))
```



Penalized regression

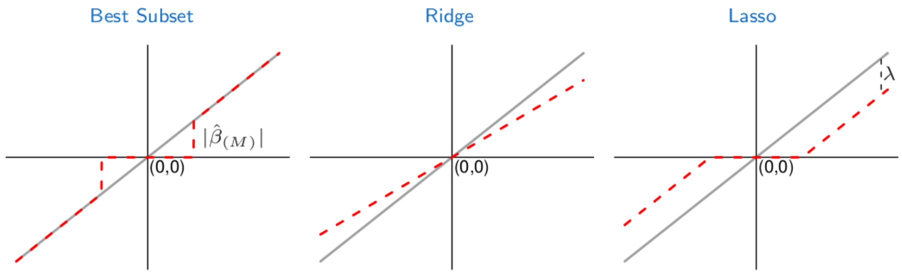
- ridge regression: penalty of the form $\alpha \sum \beta_i^2$
- lasso: penalty of the form $\alpha \sum |\beta_i|$ (reduces some variables to zero)

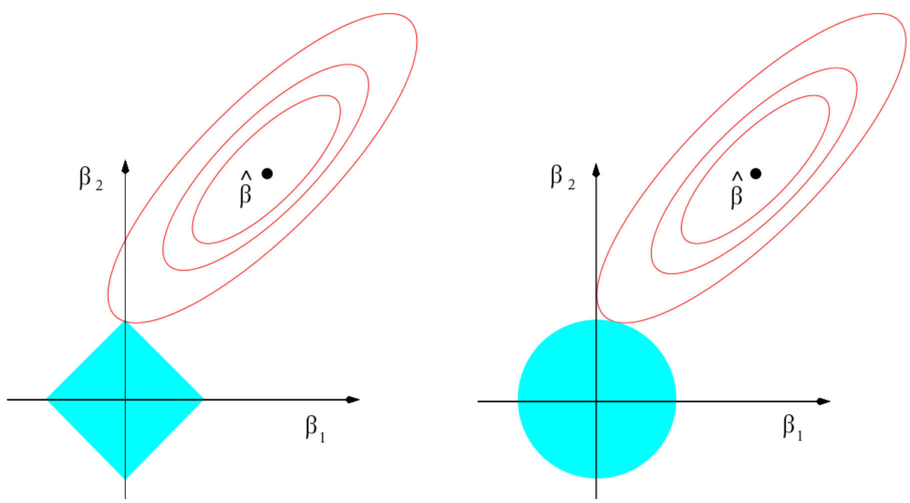
Even more: mix and match

- GAMs include both linear and generalized linear models; can also use GAMMs (generalized additive mixed models)
- spatial (or temporal) GLMMs: put a Poisson (or whatever) layer on top of a correlated MVN

$$
\begin{aligned}
Y &\sim Distrib(\mu) \\
\mu &\sim g^{-1}(MVN(X\beta, \Sigma)) \\
\Sigma &= f(\theta)
\end{aligned}
$$

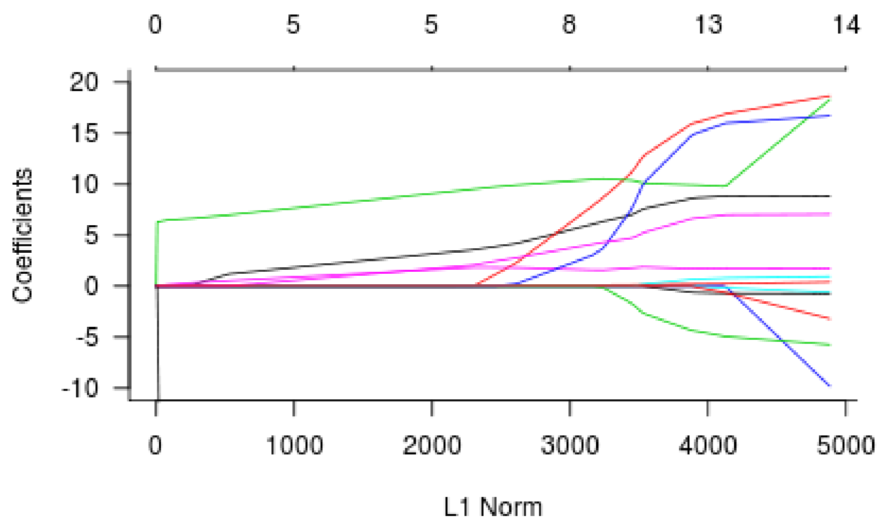
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$





(Hastie, Tibshirani, and Friedman 2009)

```r
library(glmnet)
resp <- as.matrix(subset(UScrime,select=-c(y,So)))
g1 <- glmnet(resp,UScrime$y,alpha=1)
par(las=1,bty="l")
plot(g1,ylim=c(-10,20))
```



(this is a bad example: penalized regression actually doesn't like correlated predictors!)

- more complex conditional distributions (negative binomial, Tweedie, zero-inflation); may allow linear models for the dispersion parameters as well as the mean

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning data mining, inference, and prediction.* New York: Springer. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=437866.

O'Hara, Robert B., and D. Johan Kotze. 2010. "Do not log-transform count data." *Methods in Ecology and Evolution* 1 (2) (jun): 118–122. doi:10.1111/j.2041-210X.2010.00021.x. http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2010.00021.x/abstract.

Schielzeth, Holger. 2010. "Simple means to improve the interpretability of regression coefficients." *Methods in Ecology and Evolution* 1: 103–113. doi:10.1111/j.2041-210X.2010.00012.x. http://dx.doi.org/10.1111/j.2041-210X.2010.00012.x.

Venables, W. N. 1998. "Exegeses on Linear Models." In . Washington, DC. http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf.

Warton, David I., and Francis K. C. Hui. 2011. "The arcsine is asinine: the analysis of proportions in ecology." *Ecology* 92 (jan): 3–10. doi:10.1890/10-0340.1. http://www.esajournals.org/doi/full/10.1890/10-0340.1.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.