Insurance sum models

Recall the two linear models of log claim size as a function of log insurance sum and trade group.

```
formula(claimsLm)

## log(claims) ~ log(sum)

formula(claimsLmAdd)

## log(claims) ~ log(sum) + grp
```

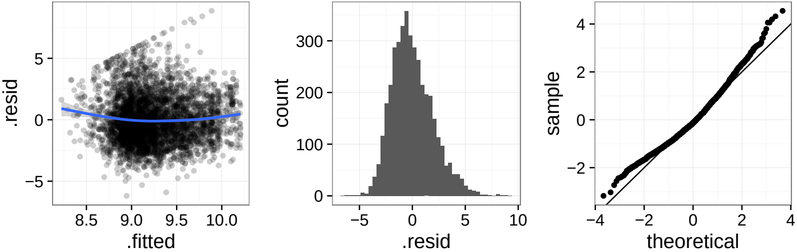
Diagnostics

Once a model is fitted we can investigate the model assumptions via the residuals,

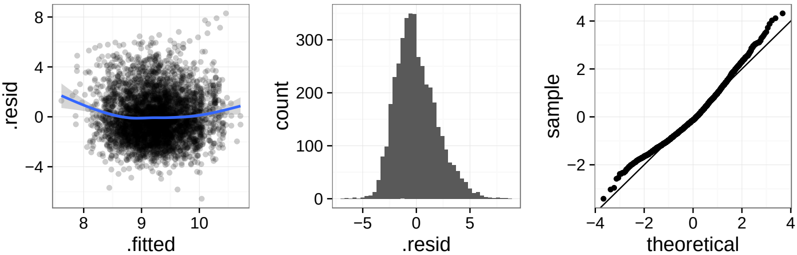
$$\hat{\epsilon}_i = Y_i - X_i^T \hat{\beta}.$$

```
claimsDiag <- fortify(claimsLm) ## Residuals etc.
grid.arrange(
  qplot(.fitted, .resid, data = claimsDiag, alpha = I(0.2)) +
    geom_smooth(),
  qplot(.resid, data = claimsDiag, bins = I(40)),
  qplot(sample = .stdresid, data = claimsDiag, geom = "qq") +
    geom_abline(),
  ncol = 3
)
```

Diagnostics for the first model



Diagnostics for the additive model



Interactions

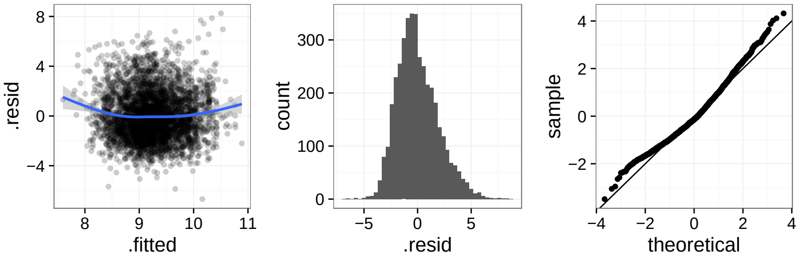
The additive model has the same slope for all groups but different intercepts. An interaction model gives individual slopes and intercepts for each group.

$$
\begin{aligned}
E(\log(Y_i) \mid X_{i,\mathtt{sum}}, \mathtt{grp}_i) = {} & \beta_0 + \beta_{\mathtt{sum}} \log(X_{i,\mathtt{sum}}) \\
& + (\beta_{\mathtt{grp2}} + \beta_{\mathtt{sum,grp2}} \log(X_{i,\mathtt{sum}}))X_{i,\mathtt{grp2}} \\
& + (\beta_{\mathtt{grp3}} + \beta_{\mathtt{sum,grp3}} \log(X_{i,\mathtt{sum}}))X_{i,\mathtt{grp3}} \\
& + (\beta_{\mathtt{grp4}} + \beta_{\mathtt{sum,grp4}} \log(X_{i,\mathtt{sum}}))X_{i,\mathtt{grp4}}.
\end{aligned}
$$

Model matrix

```
model.matrix(claimsLmAdd)[781:784, ]

##     (Intercept) log(sum) grp2 grp3 grp4
## 781           1    19.42    0    0    0
## 782           1    14.63    1    0    0
## 783           1    14.91    0    0    1
## 784           1    15.25    1    0    0

model.matrix(claimsLmInt)[781:784, ]

##     (Intercept) log(sum) grp2 grp3 grp4 log(sum):grp2
## 781           1    19.42    0    0    0          0.00
## 782           1    14.63    1    0    0         14.63
## 783           1    14.91    0    0    1          0.00
## 784           1    15.25    1    0    0         15.25
##     log(sum):grp3 log(sum):grp4
## 781             0          0.00
## 782             0          0.00
## 783             0         14.91
## 784             0          0.00
```

Diagnostics for the interaction model



Distributional results ($\mathbf{W} = \mathbf{I}$, $\mathbf{\Omega} = 0$)

Under assumptions A3 and A5:

$$\hat{\beta} \mid \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

$$(n-p)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p}.$$

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p}.$$

The F-test statistic for testing the hypothesis

$$H_0 : E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}'\beta', \quad \mathbf{X}' = \mathbf{X}C$$

with C a $p \times p_0$ matrix, $p_0 < p$, is

$$F = \frac{||\mathbf{X}\hat{\beta} - \mathbf{X}'\hat{\beta}'||^2/(p-p_0)}{||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2/(n-p)} \sim F(p-p_0, n-p).$$

Coefficient of determination

With

$$\mathrm{RSS} = ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$$

the residual sum of squares the coefficient of determination is

$$R^2 = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}_0} = 1 - \frac{\mathrm{RSS}}{\mathrm{RSS}_0}.$$

We can interpret $1 - R^2$ as a ratio of variance estimates,

$$1 - R^2 = \frac{\mathrm{RSS}/n}{\mathrm{RSS}_0/n} = \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}.$$

The adjusted R^2 is

$$\overline{R}^2 = 1 - \frac{\mathrm{RSS}/(n-p)}{\mathrm{RSS}_0/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p}.$$

Summary of simple model

```
summary(claimsLm)

##
## Call:
## lm(formula = log(claims) ~ log(sum), data = claims)
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.8410     0.2949    19.8   <2e-16
## log(sum)      0.2115     0.0182    11.6   <2e-16
##
## Residual standard error: 1.95 on 4034 degrees of freedom
## Multiple R-squared:  0.0322,Adjusted R-squared:  0.032
## F-statistic:  134 on 1 and 4034 DF,  p-value: <2e-16
```

Summary of additive model

```
summary(claimsLmAdd)

##
## Call:
## lm(formula = log(claims) ~ log(sum) + grp, data = claims)
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5974     0.3615    9.95  < 2e-16
## log(sum)      0.3300     0.0212   15.55  < 2e-16
## grp2          0.5473     0.0908    6.03  1.8e-09
## grp3          0.4013     0.1194    3.36  0.00078
## grp4          0.9143     0.0868   10.53  < 2e-16
##
## Residual standard error: 1.92 on 4031 degrees of freedom
## Multiple R-squared:  0.0591,Adjusted R-squared:  0.0582
## F-statistic: 63.3 on 4 and 4031 DF,  p-value: <2e-16
```

Summary of interaction model

```
summary(claimsLmInt)

##
## Call:
## lm(formula = log(claims) ~ log(sum) * grp, data = claims)
...
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.52566    0.43610    8.08  8.2e-16
## log(sum)        0.33429    0.02567   13.02  < 2e-16
## grp2            0.43426    0.97624    0.44    0.656
## grp3            3.66649    1.43647    2.55    0.011
## grp4            0.04078    1.02473    0.04    0.968
## log(sum):grp2   0.00772    0.06191    0.12    0.901
## log(sum):grp3 -0.20999    0.09158   -2.29    0.022
## log(sum):grp4   0.05934    0.06732    0.88    0.378
##
## Residual standard error: 1.92 on 4028 degrees of freedom
## Multiple R-squared:  0.0606,Adjusted R-squared:  0.059
## F-statistic: 37.1 on 7 and 4028 DF,  p-value: <2e-16
```

ANOVA tests

```
anova(claimsLm, claimsLmAdd, claimsLmInt)

## Analysis of Variance Table
##
## Model 1: log(claims) ~ log(sum)
## Model 2: log(claims) ~ log(sum) + grp
## Model 3: log(claims) ~ log(sum) * grp
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   4034 15328
## 2   4031 14903  3       425 38.4 <2e-16
## 3   4028 14878  3        24  2.2  0.086
```

Transformations and expansions

We used the log-transform. We could also try other transformations, e.g.

$$E(\log(Y_i) \mid X_{i,\mathtt{sum}}) = \beta_0 + \beta_{\mathrm{sum}} \sqrt{X_{i,\mathtt{sum}}}$$

or polynomial expansions of the log-transformed insurance sum

$$\begin{aligned} E(\log(Y_i) \mid X_{i,\mathtt{sum}}) = {} & \beta_0 + \beta_{\mathrm{sum},1} \log(X_{i,\mathtt{sum}}) \\ & + \beta_{\mathrm{sum},2} \log(X_{i,\mathtt{sum}})^2 \\ & + \beta_{\mathrm{sum},3} \log(X_{i,\mathtt{sum}})^3 \end{aligned}$$

The latter could be done in R using the formula

```
log(claims) ~ log(sum) + I(log(sum)^2) + I(log(sum)^3)
```

Basis expansions

Using the formula

```
log(claims) ~ ns(log(sum), knots = c(13, 15, 17, 19))
```

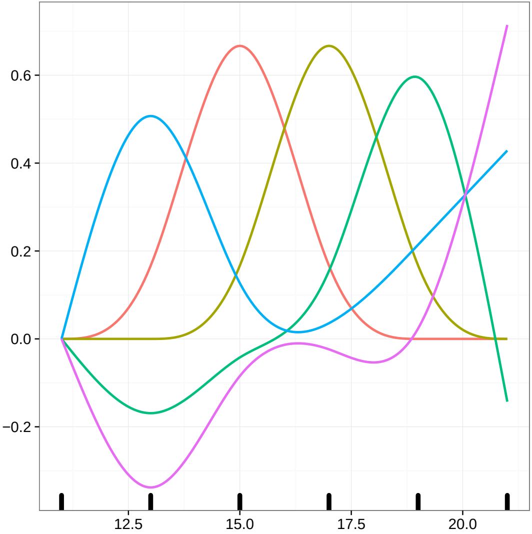
in `lm` results in a basis expansion using natural cubic splines. Here with 5 basis functions.

This means that we model the conditional mean of the response as the function

$$\beta_0 + \beta_1 h_1 + \beta_2 h_2 + \beta_3 h_3 + \beta_4 h_4 + \beta_5 h_5$$

of log-insurance sum, where h_1, h_2, h_3, h_4, h_5 are the five spline basis functions.

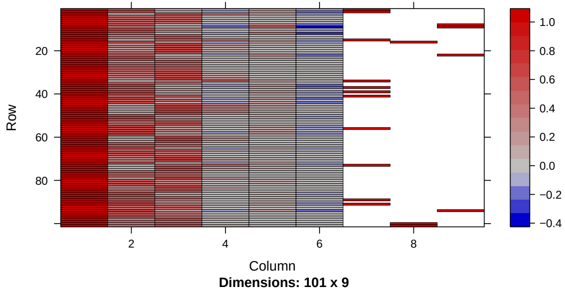
Natural cubic splines basis



Spline expansions

```
claimsLmSplineAdd <- lm(
  log(claims) ~ ns(log(sum), knots = c(13, 15, 17, 19)) + grp,
  data = claims)
```

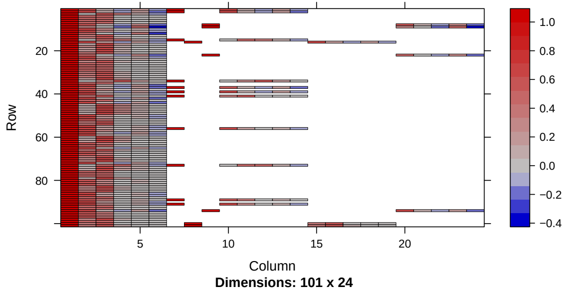
Model matrix:



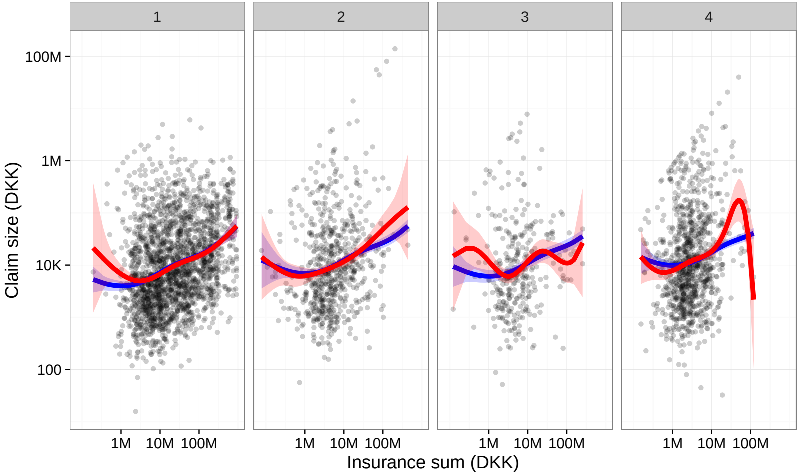
Spline expansions, interaction with trade group

```
claimsLmSplineInt <- lm(
  log(claims) ~ ns(log(sum), knots = c(13, 15, 17, 19)) * grp,
  data = claims)
```

Model matrix:



Spline based model fits



A formal comparison using F-tests

```
anova(claimsLmAdd, claimsLmSplineAdd, claimsLmSplineInt)

## Analysis of Variance Table
##
## Model 1: log(claims) ~ log(sum) + grp
## Model 2: log(claims) ~ ns(log(sum), knots = c(13, 15, 17, 19)) + grp
## Model 3: log(claims) ~ ns(log(sum), knots = c(13, 15, 17, 19)) * grp
##   Res.Df   RSS Df Sum of Sq    F  Pr(>F)
## 1   4031 14903
## 2   4027 14822  4      80.3 5.48 0.00021
## 3   4012 14683 15     139.0 2.53 0.00094
```

The least squares solution

The normal equation

$$\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{\Omega}\right)\beta = \mathbf{X}^T\mathbf{W}\mathbf{Y}$$

is usually solved without computing the matrix inverse of $\mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{\Omega}$.

The solution can be computed using

- Gaussian elimination (LU-decomposition).
- Using the sweep operator (symmetric matrices).
- Using the Cholesky decomposition of $\mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{\Omega}$.
- Using a QR decomposition as implemented in `lm`, `lm.fit` and `lm.wfit`.

The least squares solution

The R function `solve` calls the Fortran routine DGESV from the LAPACK library for solution of linear equations using LU decomposition with partial pivoting (Gaussian elimination with row permutations).

```
X <- model.matrix(claimsLmSplineInt)
y <- model.response(model.frame(claimsLmSplineInt))
XtX <- crossprod(X)
Xty <- crossprod(X, y)
coefHat <- solve(XtX, Xty)
```

In R, `crossprod(X)` computes $X^T X$ and `crossprod(X, y)` computes $X^T y$.

```
cbind(coefHat, coefficients(claimsLmSplineInt))
```

The least squares solution

```
##                                              [,1]     [,2]
## (Intercept)                               10.7533  10.7533
## ns(log(sum), knots = c(13, 15, 17, 19))1  -2.6252  -2.6252
## ns(log(sum), knots = c(13, 15, 17, 19))2  -1.4026  -1.4026
## ns(log(sum), knots = c(13, 15, 17, 19))3  -0.7424  -0.7424
## ns(log(sum), knots = c(13, 15, 17, 19))4  -1.5648  -1.5648
## ns(log(sum), knots = c(13, 15, 17, 19))5   1.1076   1.1076
## grp2                                      -1.1906  -1.1906
## grp3                                      -1.4600  -1.4600
## grp4                                      -0.5076  -0.5076
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp2   2.0113   2.0113
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp2   1.4743   1.4743
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp2   2.6491   2.6491
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp2   2.9986   2.9986
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp2   2.1465   2.1465
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp3   1.0077   1.0077
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp3   2.9934   2.9934
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp3  -0.9487  -0.9487
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp3   5.5283   5.5283
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp3   2.9652   2.9652
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp4   2.0440   2.0440
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp4   0.7091   0.7091
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp4   8.1703   8.1703
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp4 -42.4868 -42.4868
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp4 -64.8214 -64.8214
```

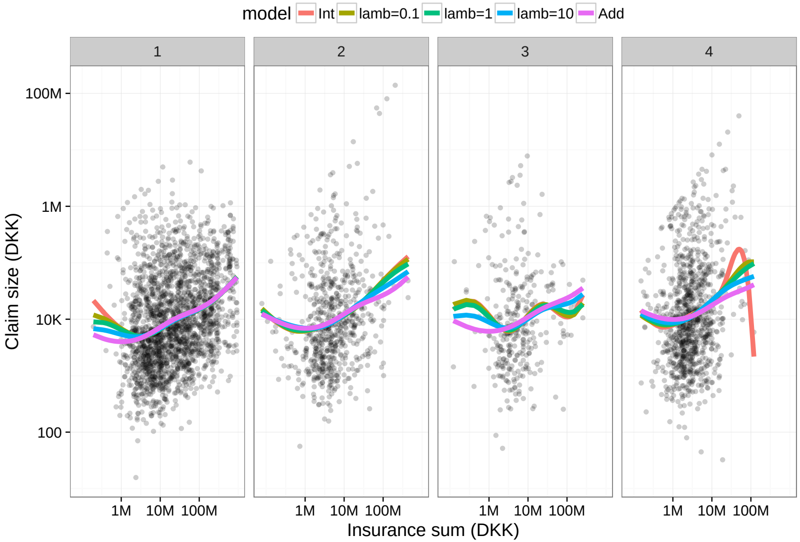
Penalized regression

```
Omega <- diag(c(rep(0, 9), rep(1, 15)))
coefHat <- cbind(solve(XtX + 0.1 * Omega, Xty),
                 solve(XtX + Omega, Xty),
                 solve(XtX + 10 * Omega, Xty))
cbind(coefHat,
      coefficients(claimsLmSplineInt),
      c(coefficients(claimsLmSplineAdd), rep(0, 15))
)
```

Penalized regression

```
##                                             Int lamb=0.1  lamb=1  lamb=10     Add
## (Intercept)                             10.7533  9.52365  9.0408  8.85931  8.8815
## ns(log(sum), knots = c(13, 15, 17, 19))1       -2.6252 -1.42990 -0.9227 -0.63479 -0.6004
## ns(log(sum), knots = c(13, 15, 17, 19))2       -1.4026 -0.15621  0.3132  0.45553  0.4772
## ns(log(sum), knots = c(13, 15, 17, 19))3       -0.7424  0.12625  0.5142  0.79648  0.9362
## ns(log(sum), knots = c(13, 15, 17, 19))4       -1.5648  0.79947  1.6315  1.62326  1.1017
## ns(log(sum), knots = c(13, 15, 17, 19))5        1.1076  1.55845  1.7969  2.08478  2.4040
## grp2                                    -1.1906  0.11804  0.5314  0.55277  0.5450
## grp3                                    -1.4600  0.09223  0.3790  0.39915  0.4237
## grp4                                    -0.5076  0.21078  0.7477  0.85908  0.9052
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp2   2.0113  0.74481  0.2485 -0.05161  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp2   1.4743  0.14018 -0.1946  0.03943  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp2   2.6491  1.74961  1.2956  0.60048  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp2   2.9986  0.34410 -0.4918 -0.24206  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp2   2.1465  1.47349  0.9410  0.33003  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp3   1.0077 -0.40670 -0.5498 -0.20548  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp3   2.9934  1.27736  0.7484  0.17380  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp3  -0.9487 -1.66247 -1.3683 -0.47788  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp3   5.5283  1.46333  0.3036  0.16021  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp3   2.9652  0.89131 -0.1442 -0.17439  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp4   2.0440  1.00593  0.3683  0.02885  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp4   0.7091  0.77982  0.4209  0.51195  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp4   8.1703  3.05257  1.8591  0.62697  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp4 -42.4868 -2.15256 -0.8488 -0.51179  0.0000
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp4 -64.8214 -2.99525  0.2050  0.33261  0.0000
```

Penalized regression



Penalized regression, QR-decomposition

The function `lm.fit` can be used for computing the penalized fit by augmenting the model matrix and the response data.

```
fitLamb1 <- lm.fit(rbind(X, sqrt(0.1) * Omega),
                   c(y, rep(0, ncol(Omega))))
fitLamb2 <- lm.fit(rbind(X, sqrt(10) * Omega),
                   c(y, rep(0, ncol(Omega))))
```

The result is (numerically) the same as using `solve` (see below).

If several penalized fits using matrices $\lambda\Omega$ for different λ are to be fitted, a third option using a diagonalization of $X^T X$ is beneficial, see the `lm.ridge` function in the package MASS.

Penalized regression

```
##                                                LU (0.1) QR (0.1)   LU (10)  QR (10)
## (Intercept)                                     9.52365  9.52365  8.85931  8.85931
## ns(log(sum), knots = c(13, 15, 17, 19))1      -1.42990 -1.42990 -0.63479 -0.63479
## ns(log(sum), knots = c(13, 15, 17, 19))2      -0.15621 -0.15621  0.45553  0.45553
## ns(log(sum), knots = c(13, 15, 17, 19))3       0.12625  0.12625  0.79648  0.79648
## ns(log(sum), knots = c(13, 15, 17, 19))4       0.79947  0.79947  1.62326  1.62326
## ns(log(sum), knots = c(13, 15, 17, 19))5       1.55845  1.55845  2.08478  2.08478
## grp2                                            0.11804  0.11804  0.55277  0.55277
## grp3                                            0.09223  0.09223  0.39915  0.39915
## grp4                                            0.21078  0.21078  0.85908  0.85908
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp2  0.74481  0.74481 -0.05161 -0.05161
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp2  0.14018  0.14018  0.03943  0.03943
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp2  1.74961  1.74961  0.60048  0.60048
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp2  0.34410  0.34410 -0.24206 -0.24206
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp2  1.47349  1.47349  0.33003  0.33003
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp3 -0.40670 -0.40670 -0.20548 -0.20548
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp3  1.27736  1.27736  0.17380  0.17380
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp3 -1.66247 -1.66247 -0.47788 -0.47788
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp3  1.46333  1.46333  0.16021  0.16021
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp3  0.89131  0.89131 -0.17439 -0.17439
## ns(log(sum), knots = c(13, 15, 17, 19))1:grp4  1.00593  1.00593  0.02885  0.02885
## ns(log(sum), knots = c(13, 15, 17, 19))2:grp4  0.77982  0.77982  0.51195  0.51195
## ns(log(sum), knots = c(13, 15, 17, 19))3:grp4  3.05257  3.05257  0.62697  0.62697
## ns(log(sum), knots = c(13, 15, 17, 19))4:grp4 -2.15256 -2.15256 -0.51179 -0.51179
## ns(log(sum), knots = c(13, 15, 17, 19))5:grp4 -2.99525 -2.99525  0.33261  0.33261
```

Distributional results ($\mathbf{W} = \mathbf{I}$, $\mathbf{\Omega} = 0$)

Under assumption A1 and A4 the least squares estimator is unbiased,

$$E(\hat{\beta} \mid \mathbf{X}) = \beta.$$

Assuming also A2

$$V(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Moreover, the estimator

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2 = \frac{1}{n-p}||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$$

is an unbiased estimator of the variance σ^2.

More distributional results

Assuming A3 and A5 recall that for the standardized Z-score

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p}.$$

More generally, for any $a \in \mathbb{R}^p$

$$\frac{a^T\hat{\beta} - a^T\beta}{\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} \sim t_{n-p}.$$

A 95% confidence interval for $a^T\beta$ is obtained as

$$a^T\hat{\beta} \pm z_{n-p}\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a} \tag{1}$$

where $\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}$ is the estimated standard error of $a^T\hat{\beta}$ and z_{n-p} is the 97.5% quantile in the t_{n-p}-distribution.

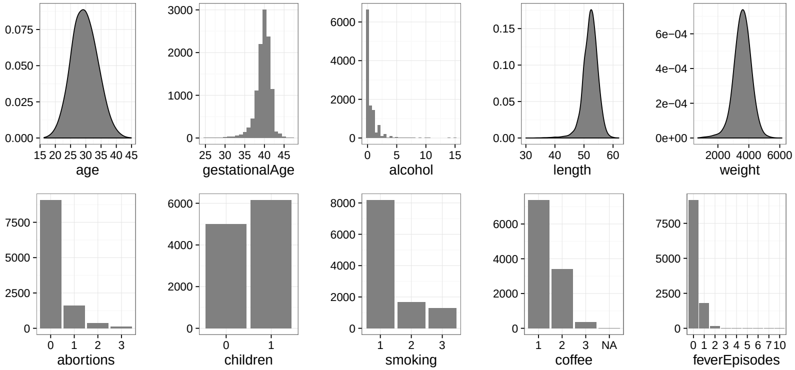
Birth weight case

```
pregnant <- read.table(
  "http://www.math.ku.dk/~richard/regression/data/pregnant.txt",
  header = TRUE,
  colClasses = c("factor", "factor", "numeric", "factor", "factor",
                 "integer", "factor", "numeric", "factor", "numeric",
                 "numeric", "integer")
)
```

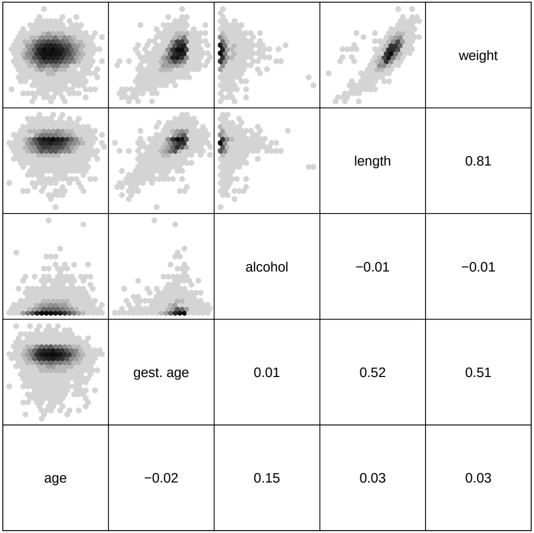
Birth weight case

`interviewWeek`: Pregnancy week at interview.
`fetalDeath`: Indicator of fetal death ($1 =$ death).
`age`: Mother's age at conception in years.
`abortions`: Number of previous spontaneous abortions (0, 1, 2, $3+$).
`children`: Indicator of previous children ($1 =$ previous children).
`gestationalAge`: Gestational age in weeks at end of pregnancy.
`smoking`: Smoking status; 0, 1–10 or $11+$ cigs/day encoded as 1, 2, 3.
`alcohol`: Number of weekly drinks during pregnancy.
`coffee`: Coffee consumption; 0, 1–7 or $8+$ cups/day encoded as 1, 2, 3.
`length`: Birth length in cm.
`weight`: Birth weight in gram.
`feverEpisodes`: Number of mother's fever episodes before interview.

Marginal distributions



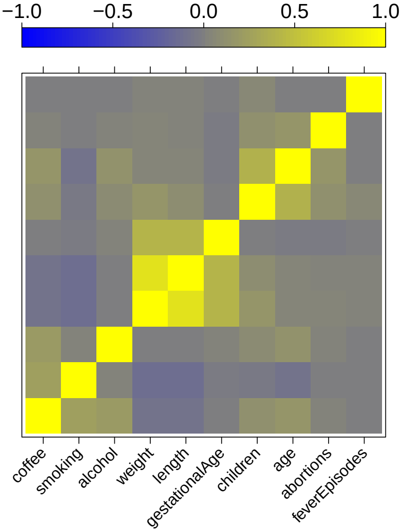
Scatter plot matrix



Spearman correlations

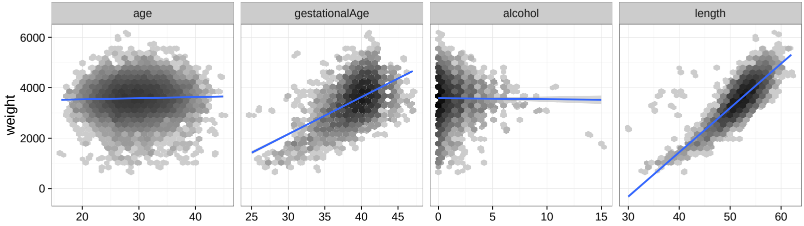
```
cp <- cor(data.matrix(na.omit(pregnant)), method = "spearman")
ord <- rev(hclust(as.dist(1 - abs(cp)))$order)
colPal <- colorRampPalette(c("blue", "yellow"), space = "rgb")(100)

levelplot(cp[ord, ord],
          xlab = "",
          ylab = "",
          col.regions = colPal,
          at = seq(-1, 1, length.out = 100),
          colorkey = list(space = "top", labels = list(cex = 1.5)),
          scales = list(x = list(rot = 45),
                        y = list(draw = FALSE),
                        cex = 1.2)
)
```

Spearman correlations



Marginal association – scatter plots revisited

```
mPregnant <- melt(pregnant[, contVar], id.vars = "weight")
binScale <- scale_fill_continuous(breaks = c(1, 10, 100, 1000),
                                  low = "gray80", high = "black",
                                  trans = "log", guide = "none")
qplot(value, weight, data = mPregnant, xlab = "", geom = "hex") +
  stat_binhex(bins = 25) + binScale +
  facet_wrap(~ variable, scales = "free_x", ncol =  4) +
  geom_smooth(size = 1, method = "lm")
```



Marginal association tests

```
form <- weight ~ gestationalAge + length + age + children +
  coffee + alcohol + smoking + abortions + feverEpisodes
pregnant <- na.omit(pregnant)
nulModel <- lm(weight ~ 1, data = pregnant)
add1(nulModel, form, test = "F")

## Single term additions
##
## Model:
## weight ~ 1
##                Df Sum of Sq      RSS    AIC  F value  Pr(>F)
## <none>                      3.61e+09 141506
## gestationalAge  1  9.32e+08 2.68e+09 138181  3876.54 < 2e-16
## length          1  2.35e+09 1.26e+09 129777 20774.87 < 2e-16
## age             1  3.95e+06 3.61e+09 141496    12.21 0.00048
## children        1  9.76e+07 3.52e+09 141203   309.55 < 2e-16
## coffee          2  2.20e+07 3.59e+09 141442    34.13 1.7e-15
## alcohol         1  1.70e+05 3.61e+09 141508     0.52 0.46898
## smoking         2  5.33e+07 3.56e+09 141344    83.48 < 2e-16
## abortions       3  6.27e+06 3.61e+09 141493     6.46 0.00023
## feverEpisodes   1  1.09e+06 3.61e+09 141505     3.35 0.06717
```

A linear additive model

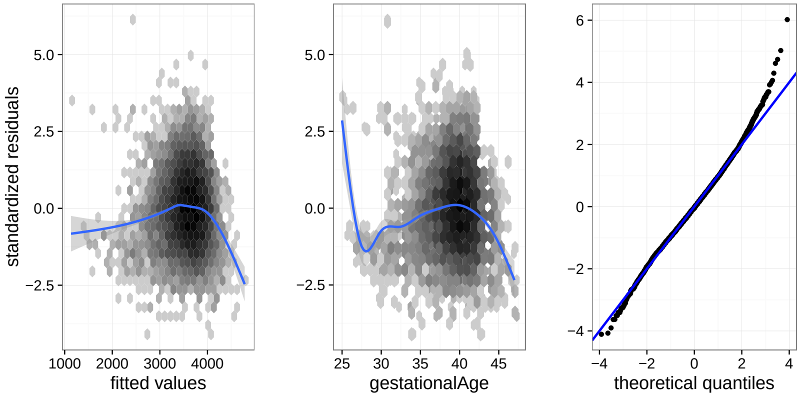
```
form <- update(form, . ~ . - length)
pregnantLm <- lm(form, data = pregnant)
summary(pregnantLm)
```

```
...
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2169.44      98.60  -22.00  < 2e-16
## gestationalAge   145.16       2.30   63.01  < 2e-16
## age               -2.00       1.20   -1.66    0.097
## children1        185.95       9.90   18.79  < 2e-16
## coffee2          -65.54      10.39   -6.31  2.9e-10
## coffee3         -141.78      27.24   -5.20  2.0e-07
## alcohol           -2.75       5.09   -0.54    0.589
## smoking2        -101.95      13.05   -7.81  6.1e-15
## smoking3        -131.19      14.91   -8.80  < 2e-16
## abortions1        27.84      13.09    2.13    0.033
## abortions2        48.76      25.45    1.92    0.055
## abortions3       -50.03      45.80   -1.09    0.275
## feverEpisodes      6.36       9.39    0.68    0.498
##
## Residual standard error: 477 on 11139 degrees of freedom
## Multiple R-squared:  0.298,Adjusted R-squared:  0.297
## F-statistic:  394 on 12 and 11139 DF,  p-value: <2e-16
```

F-test of individual terms

```
drop1(pregnantLm, test = "F")

## Single term deletions
##
## Model:
## weight ~ gestationalAge + age + children + coffee + alcohol +
##     smoking + abortions + feverEpisodes
##                Df Sum of Sq      RSS    AIC F value  Pr(>F)
## <none>                      2.54e+09 137587
## gestationalAge  1  9.04e+08 3.44e+09 140985 3970.33 < 2e-16
## age             1  6.29e+05 2.54e+09 137588    2.76   0.097
## children        1  8.04e+07 2.62e+09 137933  353.03 < 2e-16
## coffee          2  1.29e+07 2.55e+09 137640   28.35 5.2e-13
## alcohol         1  6.65e+04 2.54e+09 137586    0.29   0.589
## smoking         2  2.66e+07 2.56e+09 137700   58.44 < 2e-16
## abortions       3  2.07e+06 2.54e+09 137590    3.03   0.028
## feverEpisodes   1  1.05e+05 2.54e+09 137586    0.46   0.498
```

Residuals and diagnostics



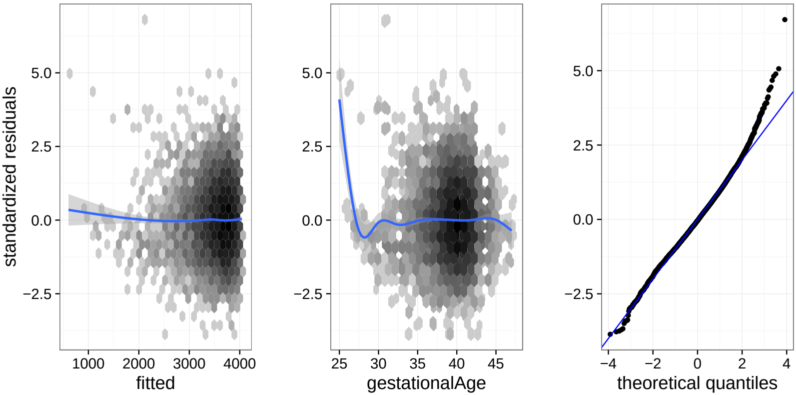
Including nonlinearity

```
nsg <- function(x)
  ns(x, knots = c(38, 40, 42), Boundary.knots = c(25, 47))
form <- weight ~ nsg(gestationalAge) + ns(age, df = 3) + children +
  coffee + alcohol + smoking + abortions + feverEpisodes
pregnantLm3 <- lm(form, data = pregnant)
anova(pregnantLm, pregnantLm3)


...
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1  11139 2.54e+09
## 2  11134 2.47e+09  5  71212926 64.3 <2e-16
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11139	2.5376e+09				
2	11134	2.4664e+09	5	7.1213e+07	64.3	2.254e−66

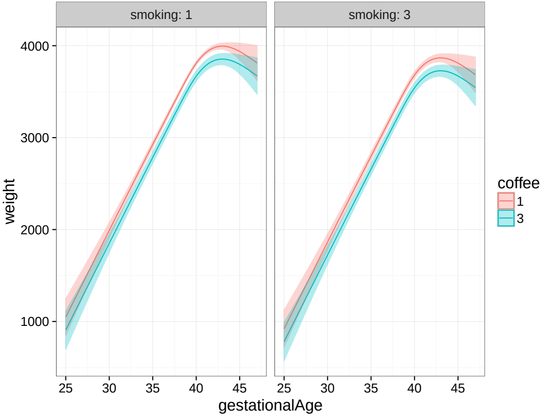
The models are nested because the basis expansion includes the identity function in its span.

Residuals and diagnostics



Reporting the model

Prefer visual summaries (predictions) of models over tables of parameter estimates.



Reporting the model

For selected parameters, estimates and/or confidence intervals may be reported.

	2.5 %	97.5 %
children1	155.47	194.09
coffee2	−82.61	−42.43
coffee3	−193.02	−87.63
smoking2	−125.76	−75.26
smoking3	−155.76	−97.98

Correct interpretation: Descriptive subpopulation differences and not causal effects.

Statistics with basis expansions

Expanding the effect of a variable using K basis functions $h_1, \ldots, h_K$ should generally be understood and analyzed as follows:

- The estimated function $\hat{h} = \sum_k \hat{\beta}_k h_k$ is interpretable and informative whereas the individual parameters $\hat{\beta}_k$ are typically not.
- One should consider combined F-tests that h is 0 or h is linear against the model with a fully expanded h and not parallel or successive tests of individual parameters.
- Pointwise confidence intervals for $\hat{h}(x)$ are easily computed by observing that

$$\hat{h}(x) = a^T \hat{\beta}, \quad a_k = h_k(x)$$

cf. (1).

Reporting the model