Regression 2016

- Practical and theoretical regression
- Linear, generalized linear and survival models
- One larger practical assignment and 27-hours take-home exam
- R, writing reports, using knitr and ggplot2

Insurance claims

```
claims <- read.table(
  "http://www.math.ku.dk/~richard/regression/data/claims.txt",
  sep = ";",
  colClasses = c("character", "numeric", "numeric", "factor")
  )
head(claims)
```

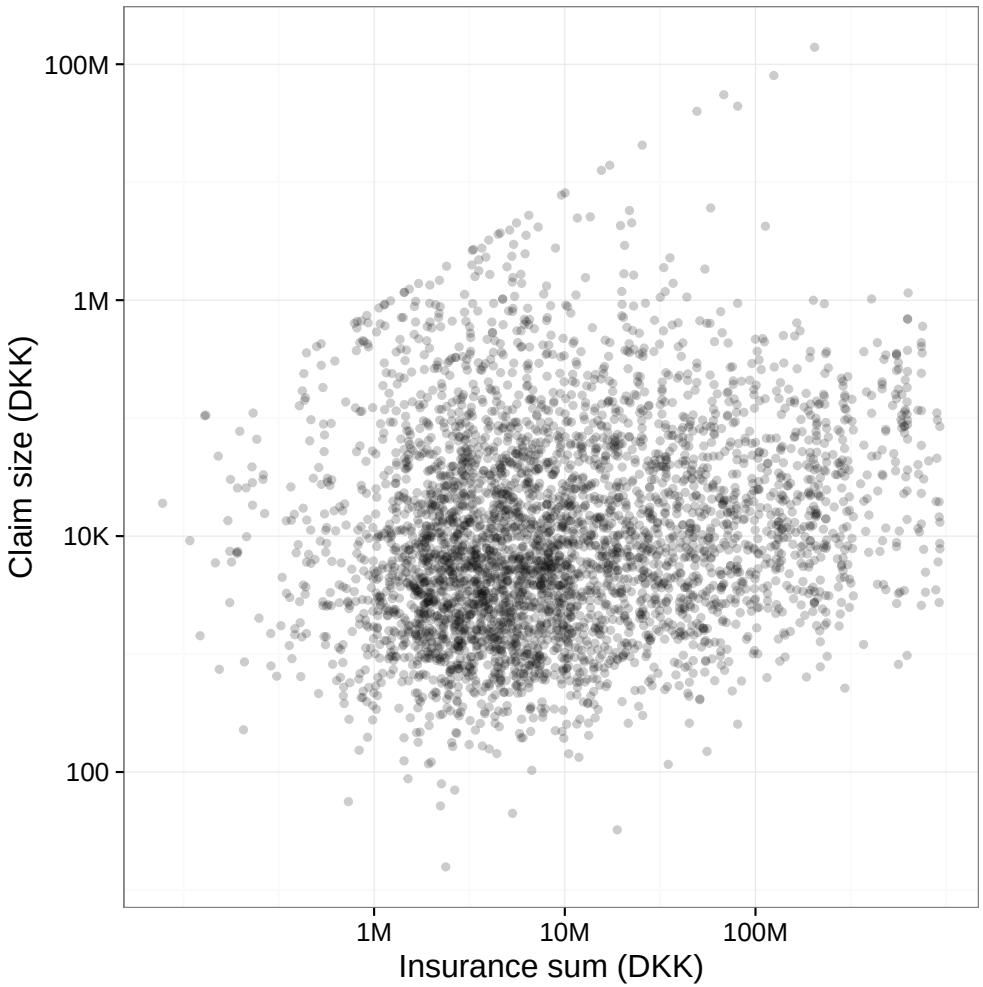
```
##   claims       sum grp
## 1   3853  34570294   1
## 2   1194   7776469   1
## 3   3917  26343305   1
## 4   1259   5502915   1
## 5  11594 711453346   1
## 6  33535 716609368   1
```

Insurance claims

```
p0 <- qplot(sum, claims, data = claims, alpha = I(0.2)) +
  scale_x_log10("Insurance sum (DKK)",
                breaks = 10^c(6, 7, 8),
                labels = c("1M", "10M", "100M")) +
  scale_y_log10("Claim size (DKK)",
                breaks = 10^c(2, 4, 6, 8),
                labels = c("100", "10K", "1M", "100M"))
```

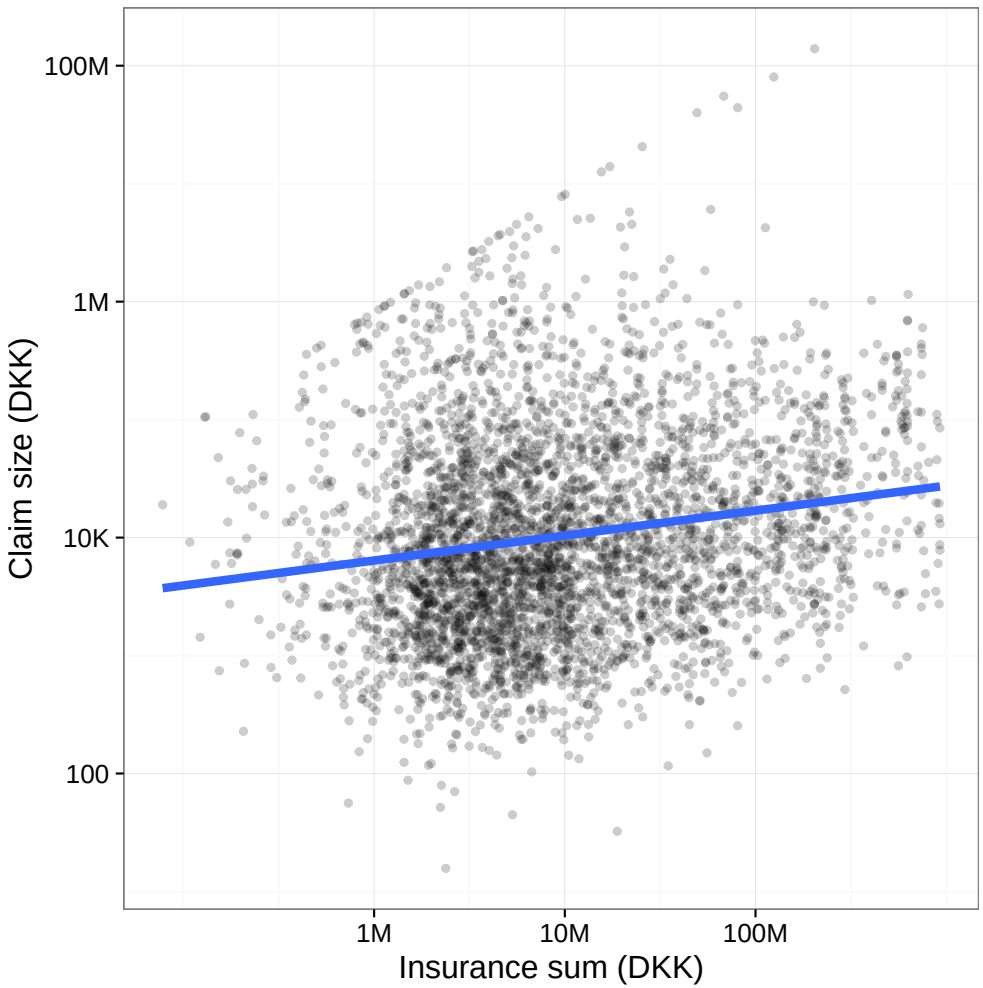
Insurance claims

```
p0
```



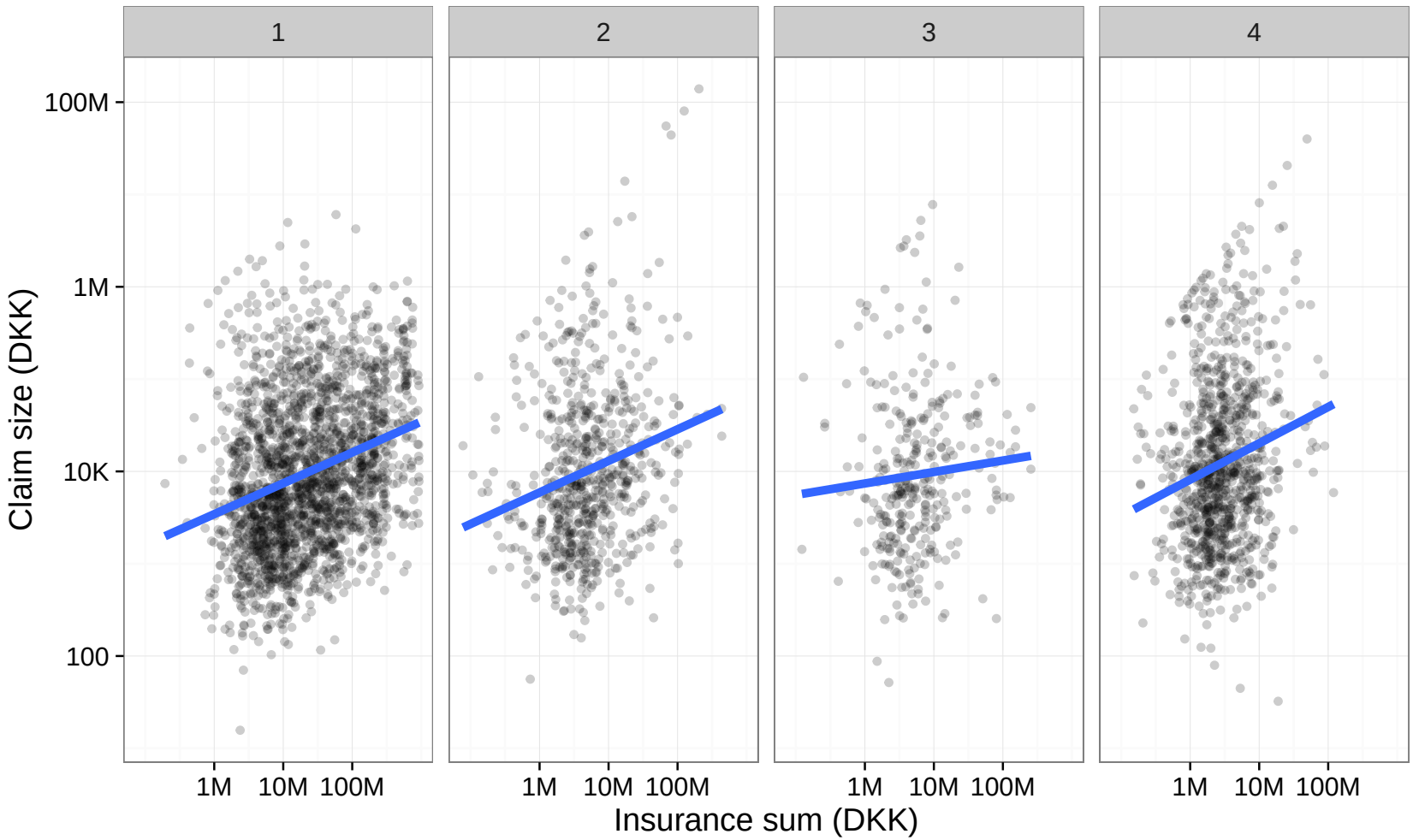
Insurance claims

```
p <- p0 + geom_smooth(method = "lm", size = 2, se = FALSE); p
```



Insurance claims

```
p + facet_wrap(~ grp, ncol = 4)
```



Model Assumptions

We consider models of the response $Y \in \mathbb{R}$ given predictors $X \in \mathbb{R}^p$.

A1: The conditional expectation of Y given X is

$$E(Y \mid X) = X^T \beta.$$

A2: The conditional variance of Y given X does not depend upon X,

$$V(Y \mid X) = \sigma^2.$$

A3: The conditional distribution of Y given X is a normal distribution,

$$Y \mid X \sim \mathcal{N}(X^T \beta, \sigma^2).$$

Model Assumptions

The reponses $Y_1, \ldots, Y_n$ are organized as a column vector $\mathbf{Y}$, and the predictors $X_1, \ldots, X_n$ as an $n \times p$ model matrix $\mathbf{X}$ with the i'th row of $\mathbf{X}$ being X_i^T.

A4: The conditional distribution of Y_i given $\mathbf{X}$ depends upon X_i only, and Y_i and Y_j are conditionally uncorrelated given $\mathbf{X}$,

$$\text{cov}(Y_i, Y_j \mid \mathbf{X}) = 0.$$

A5: The conditional distribution of Y_i given $\mathbf{X}$ depends upon X_i only, and $Y_1, \ldots, Y_n$ are conditionally independent given $\mathbf{X}$.

Model Assumptions

The model assumptions A1 + A2 + A4 are weak distributional assumptions about moments (mean, variance and covariance).

The model assumptions A3 + A5 are strong distributional assumptions (normality, independence).

Precise assumptions are necessary for mathematical deduction. Some results require strong assumptions, some require weaker assumptions.

Regression modeling relies on mathematical results, but applying the results is not mathematics! You cannot prove the model assumptions, but model assumptions can be critically investigated.

Insurance claims simple model

We will first fit the linear model

$$E(\log(Y_i) \mid X_{i,\texttt{sum}}) = \beta_0 + \beta_{\texttt{sum}} \log(X_{i,\texttt{sum}}).$$

```
claimsLm <- lm(log(claims) ~ log(sum), data = claims)
coefficients(claimsLm)

## (Intercept)    log(sum)
##      5.8410      0.2115
```

The parameter β_0 is the `(intercept)` coefficient and $\beta_{\texttt{sum}}$ is the `log(sum)` coefficient.

Insurance claims additive model

We can add an additive effect of trade group to the model:

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

```
claimsLmAdd <- lm(log(claims) ~ log(sum) + grp, data = claims)
coefficients(claimsLmAdd)

## (Intercept)    log(sum)        grp2        grp3        grp4
##      3.5974      0.3300      0.5473      0.4013      0.9143
```

Model matrices

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

##     (Intercept) log(sum)
## 781           1    19.42
## 782           1    14.63
## 783           1    14.91
## 784           1    15.25

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
```

Dummy variable encoding and contrasts

The dummy variable encoding of the categorical variable grp was

$$X_{i,\texttt{grp}k} = 1(\texttt{grp}_i = k).$$

Overparametrization was avoided by leaving out the first dummy variable and with the group parameters being contrast parameters.

An alternative parametrization is

$$\begin{aligned} E(\log(Y_i) \mid X_{i,\texttt{sum}}, \texttt{grp}_i) = \beta_0 + \beta_{\texttt{sum}} \log(X_{i,\texttt{sum}}) \\ + \beta_{\texttt{grp1}} X_{i,\texttt{grp1}} + \beta_{\texttt{grp2}} X_{i,\texttt{grp2}} \\ + \beta_{\texttt{grp3}} X_{i,\texttt{grp3}} + \beta_{\texttt{grp4}} X_{i,\texttt{grp4}}. \end{aligned}$$

with the constraint $\beta_{\texttt{grp1}} + \beta_{\texttt{grp2}} + \beta_{\texttt{grp3}} + \beta_{\texttt{grp4}} = 0$.

Additive model with different contrasts

Alternative parametrizations can be obtained using the `contrasts` argument in `lm`.

```
claimsLmAdd2 <- lm(log(claims) ~ log(sum) + grp, data = claims,
                   contrasts = list(grp = "contr.sum"))
coefficients(claimsLmAdd2)

## (Intercept)    log(sum)        grp1        grp2        grp3
##     4.06314     0.33005    -0.46573     0.08154    -0.06438
```

The `"contr.sum"` gives the sum contrast parametrization as above. The corresponding model matrix is automatically computed to avoid overparametrization.

Model matrices

```
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(claimsLmAdd2)[781:784, ]

##     (Intercept) log(sum) grp1 grp2 grp3
## 781           1    19.42    1    0    0
## 782           1    14.63    0    1    0
## 783           1    14.91   -1   -1   -1
## 784           1    15.25    0    1    0
```

Model fitting criteria

The squared error

$$(Y_i - X_i^T \beta)^2$$

quantifies the deviation of the predicted mean value $X_i^T \beta$ for the parameter β from the observed value Y_i.
The (residual) sum of squares

$$\sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

quantifies the total deviation. It is natural to minimize the sum of squares to obtain an estimate of β.

Aggregated data

Suppose that the observations come in groups, $Y_{i,j}$ for $j = 1, \ldots, m_i$ fulfilling, A1 + A2 + A4 with $X_{i,j} = X_i$. That is, for fixed i the variables in the group all share the same predictor X_i.

With

$$Y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{i,j}$$

the Y_i's fulfill A1 and A4, and

$$V(Y_i \mid X_i) = \frac{\sigma^2}{m_i}.$$

The constant variance assumption is fulfilled if the groups are all equally large.

If we only have the aggregated data they should generally be weighted according to the group size.

Weighted sum of squares

Suppose that A2 is not fulfilled but that

$$V(Y_i) = \frac{\sigma^2}{w_i}$$

for a known weight $w_i > 0$.

Then it is more reasonable to minimize the weighted sum of squares

$$\sum_{i=1}^{n} w_i (Y_i - X_i^T \beta)^2.$$

Weighted sum of squares

In matrix notation the weighted sum of squares equals

$$\ell(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta)$$

with

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix}.$$

Weighted sum of squares

For $\mathbf{W}$ any positive definite matrix there is an inner product on $\mathbb{R}^n$ defined by

$$(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{y}^T \mathbf{W} \mathbf{x}$$

and a corresponding norm $\| \cdot \|_{\mathbf{W}}$

Then

$$\ell(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2_{\mathbf{W}}.$$

Penalized sum of squares

If we believe that several of the parameters are small it can be sensible to replace the sum of squares by

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{i=1}^{p}\beta_i^2.$$

Here $\lambda > 0$ is a parameter that controls the tradeoff between $\sum_{i=1}^{p}\beta_i^2$, which is minimized for $\beta = 0$, and the model fit as measured by the sum of squares.

In vector notation the penalized sum of squares is

$$(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

Or

$$\|\mathbf{Y} - \mathbf{X}\beta^T\|^2 + \lambda\|\beta\|^2.$$

The least squares solution

We will consider estimators that are minimizers of the penalized, weighted squared error loss

$$\begin{aligned}\ell_{\mathbf{\Omega}}(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta) + \beta^T\mathbf{\Omega}\beta \\ &= \|\mathbf{Y} - \mathbf{X}\beta\|^2_{\mathbf{W}} + \|\beta\|^2_{\mathbf{\Omega}}\end{aligned}$$

for positive definite $\mathbf{W}$ and positive semidefinite $\mathbf{\Omega}$.

Theorem (Thm. 2.1)

A solution of the equation

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

is a minimizer of $\ell_{\mathbf{\Omega}}$. There is a unique minimizer if either $\mathbf{X}$ has full column rank p or if $\mathbf{\Omega}$ is positive definite.