MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 10, spring 2019
Generalized linear models - Quantile regression

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

13/5-19

Generalized linear models (GLMs)

Linear regression

The model is $Y_i \sim N(\mu_i, \sigma^2)$ where $E(Y_i) = \mu_i = \mathbf{x}_i\boldsymbol{\beta}$.

Logistic regression

The model is $Y_i \sim Bin(1, p_i)$ where $E(Y_i) = \mu_i = p_i = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}$.
We thus have $\ln \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i\boldsymbol{\beta}$.

- In both cases either $E(Y_i)$ or a function of $E(Y_i)$ is a linear model.
- Can we extend this to other models?

Generalized linear models (GLMs)

In a generalized linear model we have:

- Y_i all independently distributed from the same member of the *exponential family* (see later).
- $E(Y_i) = \mu_i$
- $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ is the *linear predictor*.
- $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta} = \eta_i$ where $g(\cdot)$ is some monotonous and differentiable function called a *link function*.

Examples:

- Linear regression: $g(\mu_i) = \eta_i = \mu_i$,
  $g(x) = x$ is the identity function.
- Logistic regression: $g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$,
  $g(x) = \ln \frac{x}{1-x}$ is the *logit* transformation.

The Exponential Family

The *exponential family* (EF) is a large class of probability distributions (both continuous and discrete). There are several definitions, here follows a specific one:

One-parameter (or "Natural") EF

- Let Y_i be a continuous (discrete) random variable with density function (probability mass function) $f(\cdot)$ as given below.
- (one-parameter) EF:

$$f(Y_i;\theta) = h(Y_i) \cdot \exp\left(\eta_i(\theta) \cdot T(Y_i) - A(\theta)\right)$$

where $h(y)$, $\eta_i(\theta)$, $T(y)$ and $A(\theta)$ are known functions.

- The Gaussian distribution is a member of the exponential family. So are exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Binomial, Poisson, Negative binomial, Wishart, Inverse Wishart and many others.

Example: Poisson distribution

If $Y_i \sim Po(\mu_i)$ we can write the probability mass function as

$$f(Y_i;\boldsymbol{\beta}) = e^{-\mu_i}\frac{\mu_i^{Y_i}}{Y_i!} = \underbrace{\frac{1}{Y_i!}}_{h(Y_i)} \cdot \exp(\underbrace{\ln \mu_i}_{\eta_i(\boldsymbol{\beta})} \cdot \underbrace{Y_i}_{T(Y_i)} - \underbrace{\mu_i}_{A(\boldsymbol{\beta})})$$

where $\eta_i = \ln \mu_i = \mathbf{x}_i\boldsymbol{\beta}$ and thus $\mu_i = e^{\mathbf{x}_i\boldsymbol{\beta}}$ is a natural parametrization.

Advantage of working with GLMs: the generality of methods. For whatever distribution from the EF:

- we can write the likelihood function as $\prod_i f(Y_i; \boldsymbol{\beta})$ (for independent Y_i).
- maximize the (log)likelihood by Newton-Raphson.
- invoke asymptotic normality of maximum likelihood estimates for inferences.
- use likelihood-ratios and deviances for testing.
- Because of the generality of GLMs+EF we do not need to reintroduce specific calculations for each possible distribution for our responses.

Re-define deviance as distance from the saturated model, $\hat{\boldsymbol{\mu}}_S = \mathbf{Y}$:

$$D = -2 \ln L(\hat{\boldsymbol{\beta}}) - (-2 \ln L(\hat{\boldsymbol{\mu}}_S))$$

In logistic regression $L(\hat{\boldsymbol{\mu}}_S) = 1$ so it did not matter.

- As mentioned in Lecture 8, everything is based on the construction of the likelihood function and consequent inferences.

$$\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \mathbf{H}^{-1}), \qquad (n \to \infty)$$

where $\mathbf{H}$ is the Hessian matrix of $-\ln L(\boldsymbol{\beta})$ evaluated at $\hat{\boldsymbol{\beta}}$.

We are going to look in detail into 2 additional members (Gaussian and Binomial have been considered already):

- Poisson distribution $\to$ Poisson regression;
- Negative Binomial distribution $\to$ Negative Binomial regression;
- Poisson regression: just need to change the `family` argument into `glm()` accordingly.

Poisson regression

We want to investigate the relationship between a variable taking non-negative integer values and some covariates.

This type of response often represents a **count** (though not exclusively).

Examples:

- Response: The number of people in line at a certain time in the grocery store. Predictors: the number of items currently offered at a special discounted price and whether a special event (e.g., a holiday, a big sporting event) is incoming.
- Response: The number of awards earned by students at a high school. Predictors: the type of program in which the students were enrolled (e.g., vocational, general or academic) and the score on their final exam in math.
- Response: the number of customers calling some technical support by phone during an hour. Predictors: time of the day; day of the week (Monday morning should be especially busy).

Poisson regression or loglinear models?

As previously mentioned, Y does not have to represent a count. However this is most often the case of interest in practical applications.

Convention:

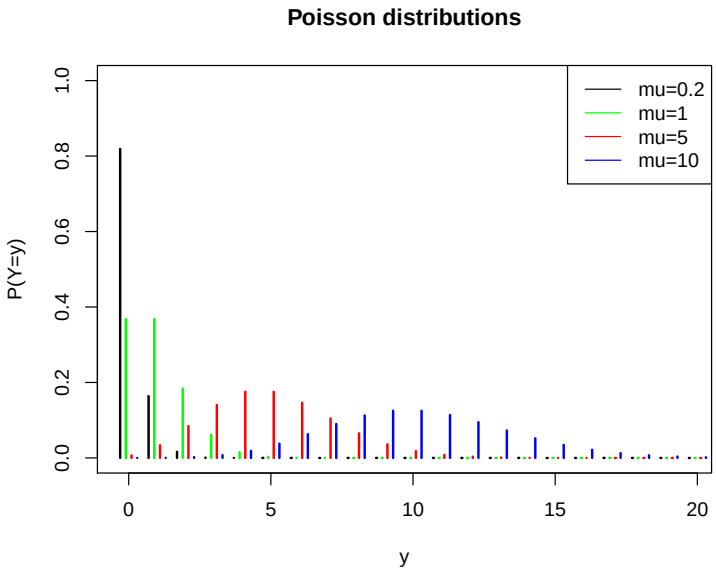
- All covariates categorical: responses can be grouped in a (contingency) table with counts in the cells. In literature, these type of models are called **loglinear models**.
- Numerical/continuous covariates: in literature convention is to call them **Poisson regression** models.

For the rest of this lecture we use the term "Poisson regression" for all cases.

Poisson regression

We observe $Y_i =$ "number of events in experiment i" $\sim Po(\mu_i)$ with $E(Y_i) = V(Y_i) = \mu_i$. Since μ_i must be positive a suitable function could be $\mu_i = e^{\mathbf{x}_i\beta}$ and the link $\ln \mu_i = \mathbf{x}_i\beta$ (log-link). Thus, the probabilities are given by

$$P(Y_i = y_i) = \frac{e^{-\mu_i} \cdot \mu_i^{y_i}}{y_i!} = \frac{e^{-e^{\mathbf{x}_i\beta}}(e^{\mathbf{x}_i\beta})^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$



As previously remarked: with GLMs we need to derive the likelihood function.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} Pr(Y_i = y_i) = \prod_{i=1}^{n} \frac{e^{-e^{\mathbf{x}_i\boldsymbol{\beta}}}(e^{\mathbf{x}_i\boldsymbol{\beta}})^{y_i}}{y_i!},$$

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}(-e^{\mathbf{x}_i\boldsymbol{\beta}}) + \sum_{i=1}^{n} y_i\mathbf{x}_i\boldsymbol{\beta} - \sum_{i=1}^{n}\ln(y_i!)$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_0} = -\sum_{i=1}^{n} e^{\mathbf{x}_i\boldsymbol{\beta}} + \sum_{i=1}^{n} y_i \qquad = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_1} = -\sum_{i=1}^{n} X_{1i} \cdot e^{\mathbf{x}_i\boldsymbol{\beta}} + \sum_{i=1}^{n} X_{1i}y_i \qquad = 0$$

$$\vdots$$

The non-linear system $\mathbf{X}'\mathbf{M} = \mathbf{X}'\mathbf{Y}$ is solved by Newton-Raphson.

- The rate ratio e^{β_j} is the relative increase in the expected value when X_j is increased by 1 (*and all other predictors are kept fixed*): $\frac{\mu(X_j+1)}{\mu(X_j)} = e^{\beta_j}, \qquad j = 1, ..., p.$
- Model comparison: as we know, for GLM models we can test hypotheses about several β_j (i.e. larger vs smaller models) using likelihood ratio (deviance) tests.
- Use $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, Var(\hat{\boldsymbol{\beta}}))$ and $\ln \hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \approx N(\mathbf{x}_0\boldsymbol{\beta}, \mathbf{x}_0 Var(\hat{\boldsymbol{\beta}})\mathbf{x}_0')$ for large n to construct intervals for β_j and $\ln \mu_0$.
- Here, and for all GLMs, $V(\hat{\beta}_j)$ is obtained from the diagonal elements of the inverted Hessian matrix at convergence of Newton-Raphson.
- A confidence interval for μ_0 is then given by $I_{\mu_0} = e^{I_{\mathbf{x}_0\boldsymbol{\beta}}}$.

Example: number of awards

Response: the number of awards earned by students at a high school. $\bar{Y} = 0.63 < s^2 = 1.11$. Larger variance explained by covariates?

Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

Model: $Y_i \sim Po(\mu_i)$ where

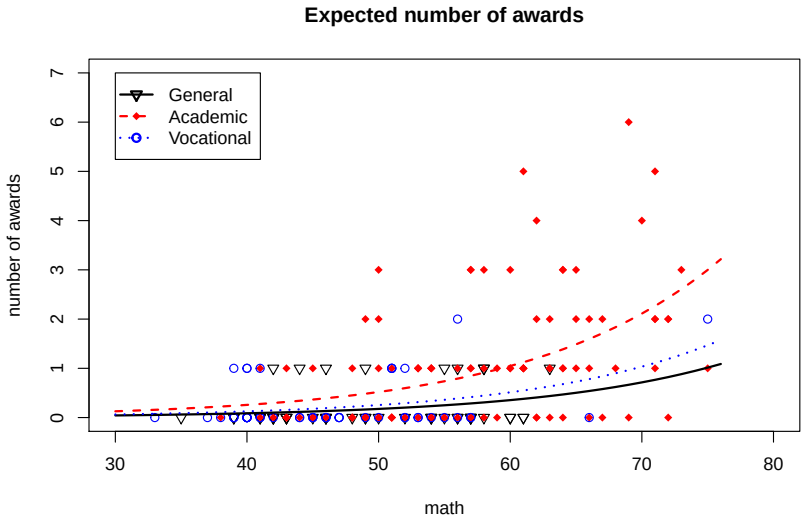
$\ln \mu_i = \beta_0 + \beta_1 X_{\text{Academic},i} + \beta_2 X_{\text{Vocational},i} + \beta_3 X_{\text{math},i} = \mathbf{x}_i\boldsymbol{\beta}$.

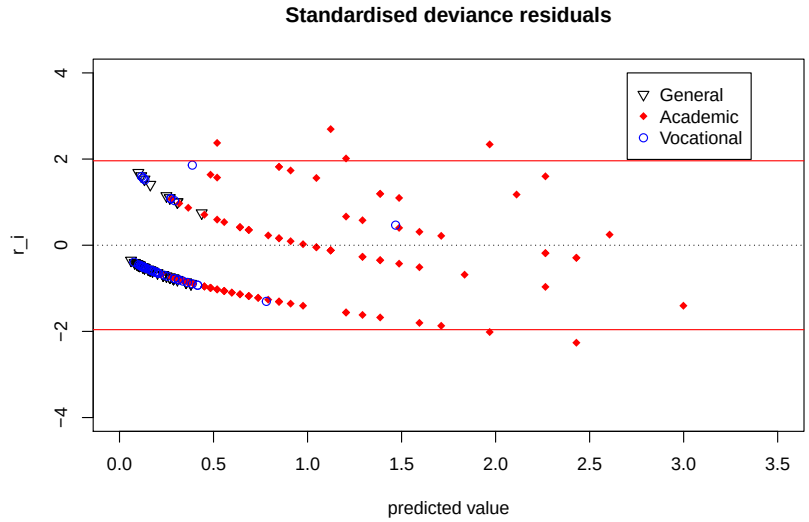
Estimated model:

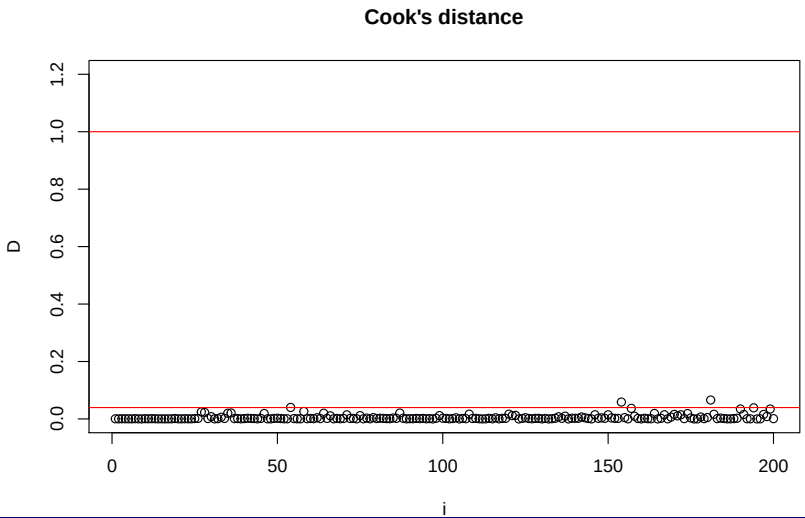
$$\hat{\mu}_i = \begin{cases} e^{-5.25+0.07\cdot X_{\text{math},i}} & = 0.005 \cdot 1.07^{X_i}, \quad \text{General} \\ e^{-5.25+1.08+0.07\cdot X_{\text{math},i}} & = 0.016 \cdot 1.07^{X_i}, \quad \text{Academic} \\ e^{-5.25+0.37+0.07\cdot X_{\text{math},i}} & = 0.008 \cdot 1.07^{X_i}, \quad \text{Vocational} \end{cases}$$

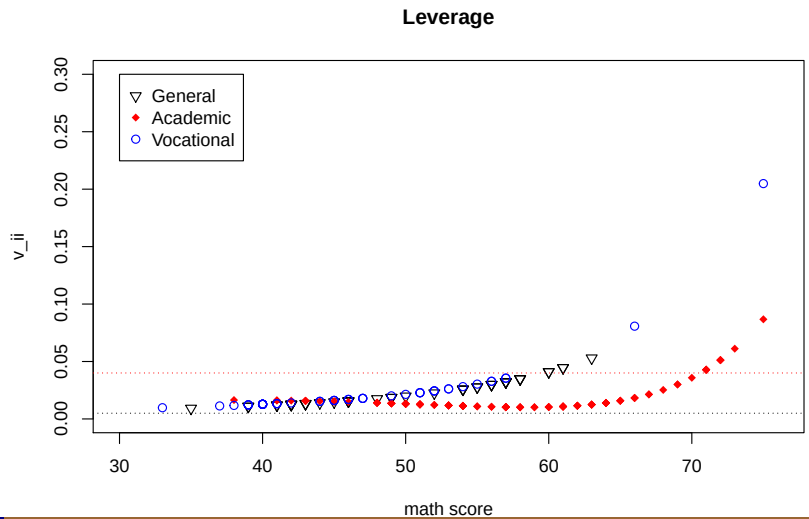
The expected number of awards increases by 7 % for each extra math score and Academic students get $e^{1.08} = 2.96 \approx 3$ times more awards than General, for the same math score.

A few outliers. However no observation has a strong influence on the estimates. Note that the residuals have constant variance, even though the original observations have increasing variance.









Negative binomial regression

Count data often vary more than the Poisson distribution allows. That is the fact that for Poisson variables the mean equals the variability is a rather strong assumptions that in many situations does not hold[1].

- Let each individual have their own poisson mean, $z_i \cdot \mu_i$, randomly distributed around some common value, $\mu_i = e^{\mathbf{x}_i\boldsymbol{\beta}}$.
- Then let Y_i, $i = 1, \ldots, n$ be independent observations with

$$(Y_i \mid Z_i = z_i) \sim Po(z_i\mu_i)$$

and

$$Z_i \sim \Gamma(\theta, 1/\theta) \quad \text{with } E(Z_i) = 1 \text{ and } V(Z_i) = 1/\theta.$$

We then get the distribution of Y_i as (Law of Total Probability)

$$Pr(Y_i = y_i) = \int_0^\infty Pr(Y_i = y_i \mid Z_i = z) \cdot f_{Z_i}(z)\, dz$$

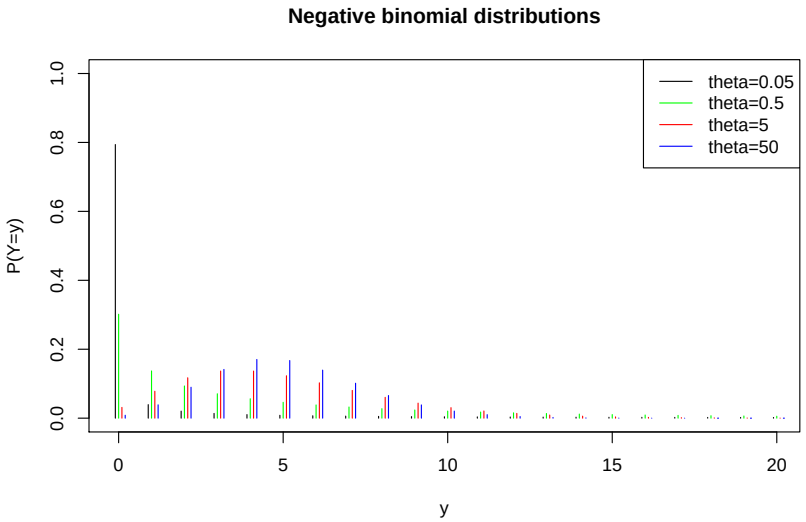
[1] See the nice example at
http://stats.idre.ucla.edu/r/dae/negative-binomial-regression/

Now Y_i, $i = 1, \ldots, n$ are independent observations from the "Negative Binomial" distribution below

$$
\begin{aligned}
Pr(Y_i = y_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \cdot \frac{(\mu_i/\theta)^{y_i}}{(1 + \mu_i/\theta)^{\theta + y_i}} \\
&= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \cdot \frac{(e^{\mathbf{x}_i\boldsymbol{\beta}}/\theta)^{y_i}}{(1 + e^{\mathbf{x}_i\boldsymbol{\beta}}/\theta)^{\theta + y_i}}, \qquad y_i = 0, 1, 2 \ldots \\
E(Y_i) &= \mu_i \\
V(Y_i) &= \mu_i + \mu_i^2/\theta > \mu_i \qquad (\theta > 0)
\end{aligned}
$$

($\Gamma(x) = (x-1)!$ if x is an integer, else $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$.)

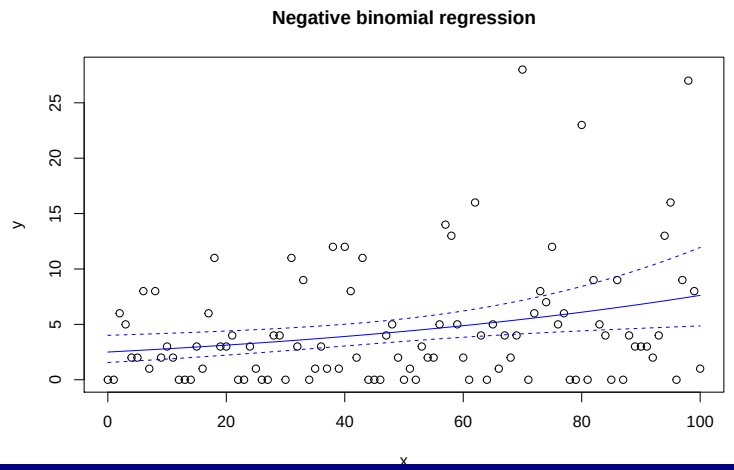
Variances: $5 + 5^2/0.05 = 505$; $5 + 5^2/0.5 = 55$; $5 + 5^2/5 = 10$; $5 + 5^2/50 = 5.5$:



Estimate $\boldsymbol{\beta}$ and θ by Maximum Likelihood. Solved by Newton-Raphson.

R: negative-binomial regression is not implemented in the basic R library. Need to load the `MASS` package via `library(MASS)`

- notice again, maximization of the likelihood will also return a $\hat{\theta}$, see the bottom of the `summary(model)`.
- Test hypotheses about several β_j using likelihood ratio (deviance) tests.
- Use $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, Var(\hat{\boldsymbol{\beta}}))$ and $\ln \hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \approx N(\mathbf{x}_0\boldsymbol{\beta}, \mathbf{x}_0 Var(\hat{\boldsymbol{\beta}})\mathbf{x}_0')$ for large n to construct intervals for β_j and $\ln \mu_0$.
- A confidence interval for μ_0 is then given by $I_{\mu_0} = e^{I_{\mathbf{x}_0\boldsymbol{\beta}}}$.
- Test if it is worth to use a negative-binomial instead of the (simpler) Poisson using a likelihood ratio test: $-2 \ln L_{\text{poisson}} - (-2 \ln L_{\text{negbin}}) > \chi^2_{1-\alpha}(1)$.



Warning

A little remark: in order to execute a likelihood ratio test to compare a Poisson vs Negative-Binomial model do NOT use `anova()`.

`anova()` is built to compare (nested) models where observations follow the same distribution.

Compare "by hand" using

```
-2*(logLik(model.pois)[1]-logLik(model.nb)[1])
```

For a nice additional example see `http://stats.idre.ucla.edu/r/dae/negative-binomial-regression/`

Distribution free

For the entire course we assumed that our data were samples from some specific distribution:

- Gaussian responses $\rightarrow$ linear regression;
- binary responses $\rightarrow$ Bernoulli/binomial distributions $\rightarrow$ logistic regression;
- ... and other members from the exponential family $\rightarrow$ GLMs.

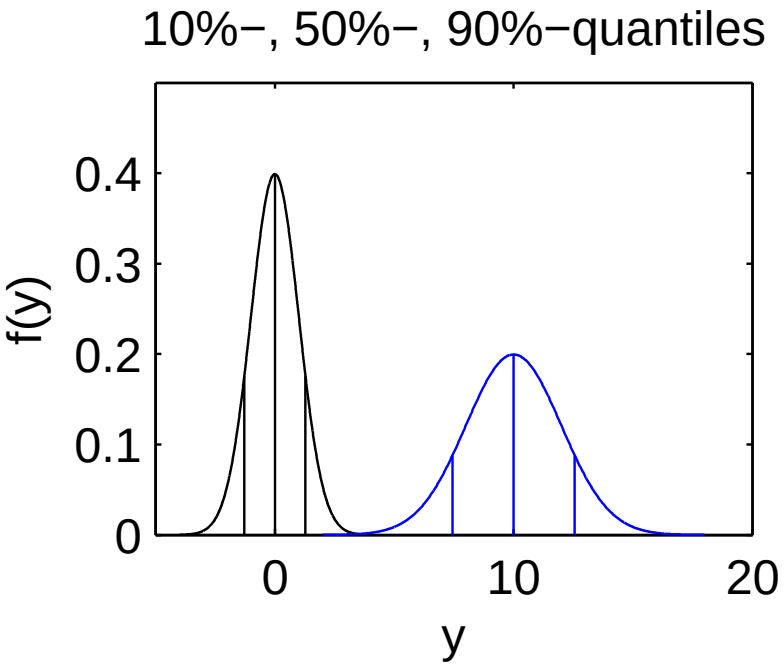
It is implicit that when we say "we assume a certain distribution" we should use diagnostic methods to verify that the assumption holds.

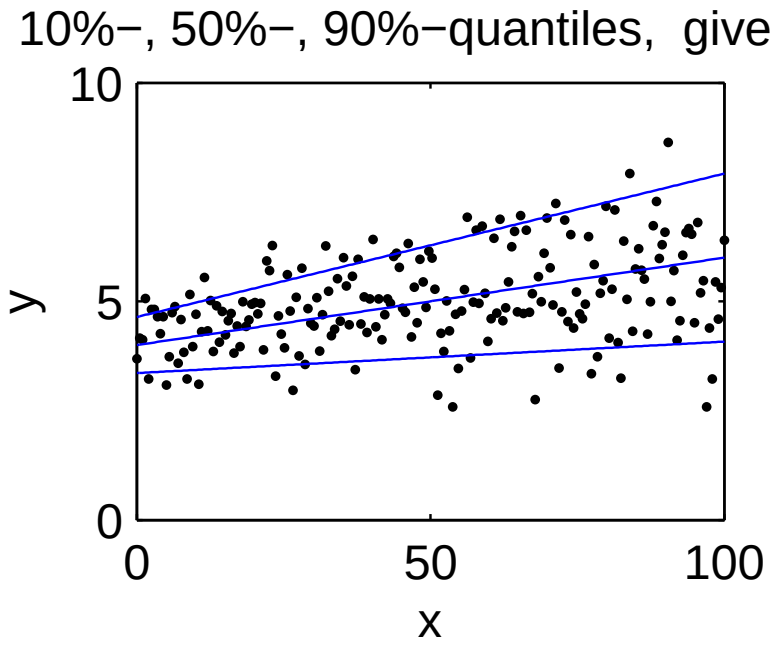
For example by using qq-plots, to compare sample quantiles with the exact quantiles from an hypothesized distribution.

Now we consider a methodology which is "distribution free".

Quantiles

General case: The α-quantile y_α is defined as $P(Y > y_\alpha) = \alpha$.
Regression: The α-quantile $y_{i\alpha}$ is defined as $P(Y_i > y_{i\alpha} \mid \mathbf{x}_i) = \alpha$.



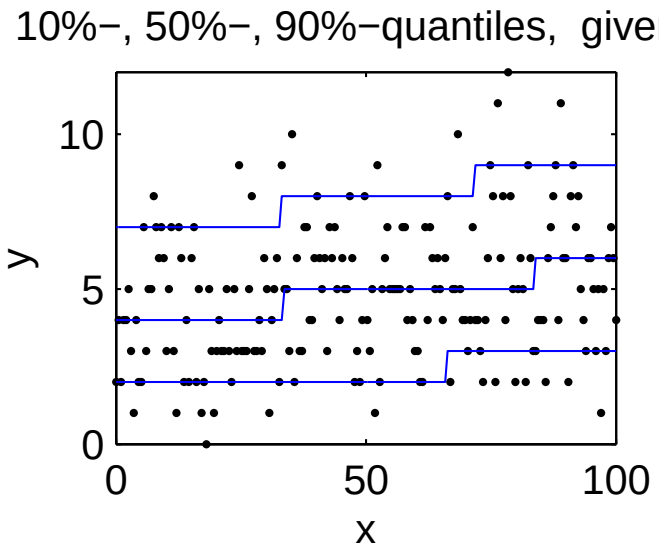


Linear regression

With $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ we have $y_{i\alpha} = \mathbf{x}_i\boldsymbol{\beta} + \lambda_\alpha \cdot \sigma$.
Estimated by the prediction interval. Requires that ϵ_i really are $N(0, \sigma^2)$.
[here λ_α is the α-quantile from N(0,1)]

Poisson regression

With $Y_i \sim Po(e^{\mathbf{x}_i\beta})$ we can use the quantiles in the (estimated) Poisson distribution:

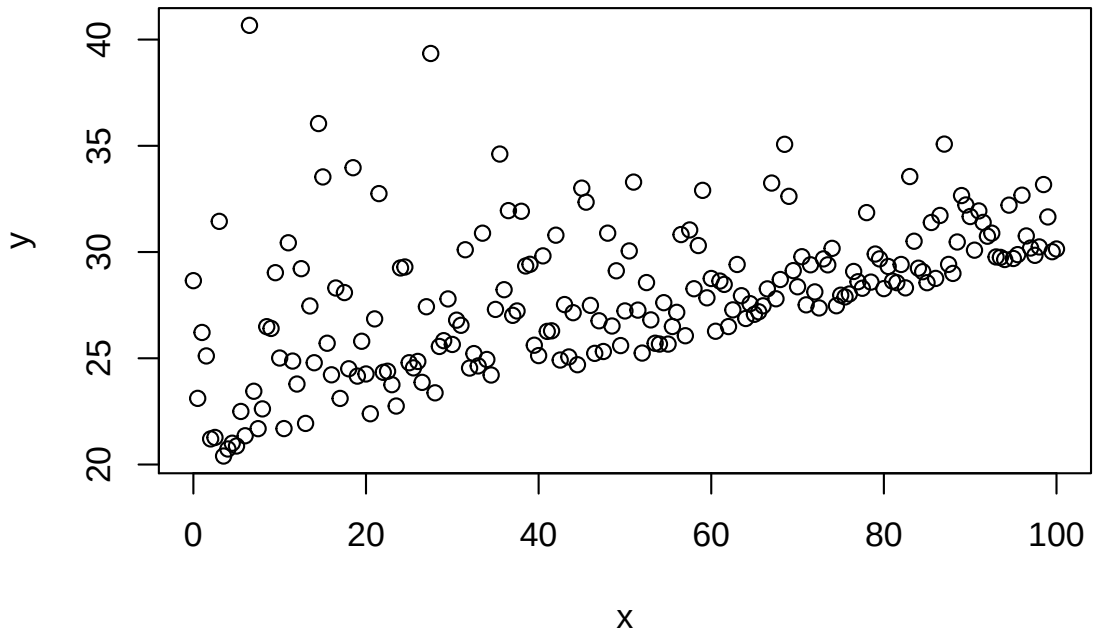


Other distributions

As long as we know (read "pretend to know") the distribution type and can estimate all its parameters we can use its quantiles.

What if we don't know the distribution type?

Example:



Skewed distribution with larger variability for lower expected values.
Difficult to find a transformation.

Quantile as solution to minimization problem

- Sample mean as a solution to minimization: $\hat{\mu} = \bar{y}$ solves

$$\min_{\mu} \sum (y_i - \mu)^2$$

- Median (i.e. 50% quantile) $\hat{m} = median(y_1, ..., y_n)$ solves:

$$\min_{m} \sum |y_i - m| \qquad \text{(robust to outliers!)}$$

- generic empirical quantile y_α corresponding to a probability α solves:

$$\min_{y_\alpha} \left\{ (1-\alpha) \sum_{y_i > y_\alpha} |y_i - y_\alpha| + \alpha \sum_{y_i \leq y_\alpha} |y_i - y_\alpha| \right\}$$

and is *robust to outliers*.

Quantile regression

Set $y_{i\alpha} = \mathbf{x}_i\boldsymbol{\beta}_\alpha$ where $P(Y_i > y_{i\alpha} \,|\, \mathbf{x}_i) = \alpha$.
Replace Least-squares $(Y_i - \mu_i)^2$ by

$$\rho_\alpha(Y_i - y_{i\alpha}) = \begin{cases} (1-\alpha)\cdot|Y_i - y_{i\alpha}| & \text{if } Y_i > y_{i\alpha}, \\ \alpha\cdot|Y_i - y_{i\alpha}| & \text{if } Y_i \leq y_{i\alpha}. \end{cases}$$

and minimize $\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i\boldsymbol{\beta}_\alpha)$ with respect to $\boldsymbol{\beta}_\alpha = \begin{pmatrix} \beta_{0\alpha} \\ \beta_{1\alpha} \\ \vdots \\ \beta_{p\alpha} \end{pmatrix}$

Symbol $\boldsymbol{\beta}_\alpha$ is meant to emphasize that it is not an estimate based on least squares or maximum likelihood (in the latter case it would not be possible as we do not specify the distribution for observed data).

Features

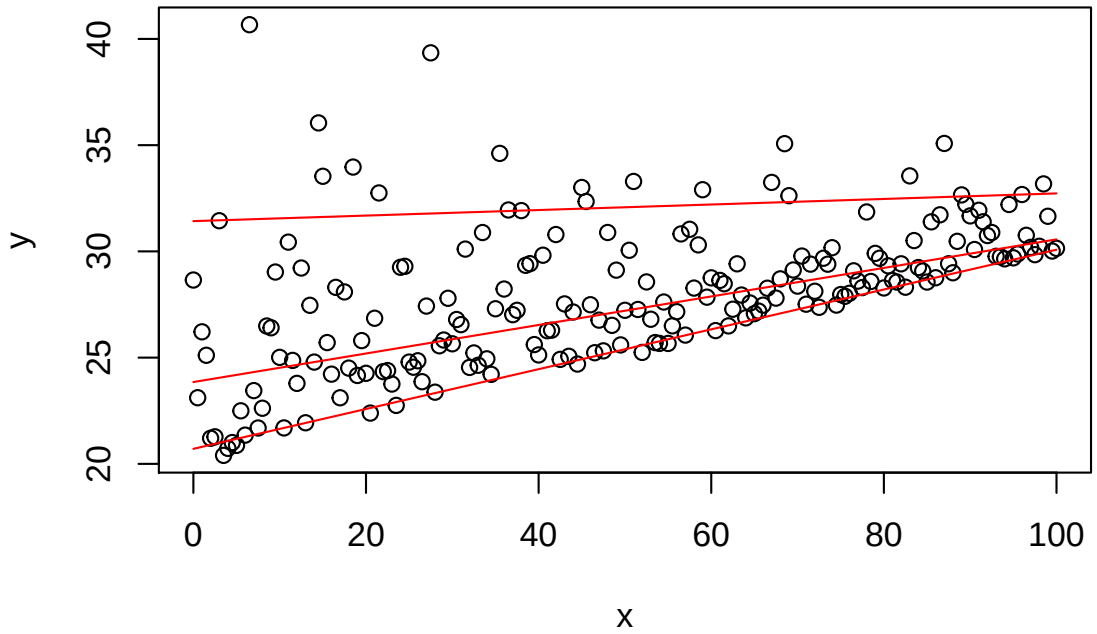
- the resulting α-quantile regression line is such that, for given covariates $\mathbf{x}_i$, a proportion of approximately α data points lies above the fitted value

$$y_{i\alpha} = \beta_{0\alpha} + \beta_{1\alpha} X_{i1} + \cdots + \beta_{p\alpha} X_{ip}$$

  and a proportion $1 - \alpha$ lies below.
- For example, when we estimate the coefficients for the .10th quantile regression line, 10% of the data points will lie above the fitted value leading to positive residuals, and 90% lie below the fitted value and thus have negative residuals.
- Conversely, to estimate the coefficients for the .90th quantile regression, 10% of observations have negative residuals and the remaining 90% have positive residuals.

Example (cont)

In R: install the package `quantreg`, then `library(quantreg)` and `rq(y ~ x, data = data, tau = c(0.1, 0.5, 0.9))`



Quantile regressions of 10, 50 and 90 %-quantiles.

- Advantages of QR: (i) does not require assumptions on the distribution of Y. (ii) more robust to outliers than least squares (quantiles do not change that much in presence of outliers).
- Disadvantages of QR: because of (i) estimated asymptotic variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ does **not** attain minimal variance (Cramer-Rao bound), unlike maximum likelihood estimates (MLE).

Cramer-Rao*

Under (mild) conditions an unbiased estimator $\hat{\beta}$ of a scalar parameter β has variance

$$Var(\hat{\beta}) \geq \frac{1}{I_{fish}(\beta)} \qquad \text{(Cramer-Rao bound)}$$

$I_{fish}(\beta)$ is the "Fisher information", given by
$I_{fish}(\beta) = -E(\frac{\partial^2 \log L(y_1,..,y_n;\beta)}{\partial^2 \beta})$
As $n \to \infty$ a MLE reach the minimal variance I_{fish}.

Even though for most models I_{fish} is unavailable analytically, an approximate is given by $I_{fish} \approx H$, the Hessian matrix at the MLE $\hat{\beta}$.