MASM22/FMSN3/FMSN30F0: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 11, spring 2026
Multinomial and Ordinal logistic regression
Quantile regression

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

13/5-26

Multinomial distribution

The multinomial distribution is an extension of the binomial distribution with more than two possible outcomes.

- Assume that an experiment can result in $q+1$ mutually exclusive outcomes, each with probabilities $p_0, \ldots, p_q$ where $0 \leq p_k \leq 1$, $k = 0, \ldots, q$, and $\sum_{k=0}^{q} p_k = 1$.
- Perform the experiment n independent times and calculate the total number of times each outcome occured as $\mathbf{Y} = (Y_0, \ldots, Y_q)$ where $\sum_{k=0}^{q} Y_k = n$.
- Then $\mathbf{Y} \sim \text{Multinomial}(n, p_0, \ldots, p_q)$ and

$$\Pr(\mathbf{Y} = \mathbf{y}) = \Pr(Y_0 = y_0, \ldots, Y_q = y_q) = \\ = \frac{n!}{y_0! \cdot \ldots \cdot y_q!} \cdot p_0^{y_0} \cdot \ldots \cdot p_q^{y_q}$$

for $y_0, \ldots, y_q = 0, 1, \ldots, n$ and $\sum_{k=0}^{q} y_k = n$

Unordered or ordered categories?

Unordered categories: multinomial logistic regression

If there is no logical order between the categories

Ordered categories: ordinal logistic regression

If there is a logical order between the categories

Multinomial logistic regression

Assume that Y_i can take any of the values $k = 0, 1, \ldots, q$ with probabilities $p_{0,i}, \ldots, p_{q,i}$ where $\sum_{k=0}^{q} p_{k,i} = 1$ giving $p_{0,i} = 1 - \sum_{k=1}^{q} p_{k,i}$.
Model:

$$Y_i \sim \text{Multinomial}(1, p_{0,i}, \ldots, p_{q,i}), \qquad i = 1, \ldots, n$$

Using category 0 as reference, we assume that the log-odds for being in category k instead of in category 0 can be modeled as

Multinomial

$$\ln \frac{p_{1,i}}{p_{0,i}} = \beta_{1,0} + \beta_{1,1} x_{i1} + \ldots + \beta_{1,p} x_{ip} = \mathbf{x}_i \boldsymbol{\beta}_1$$

$$\vdots$$

$$\ln \frac{p_{q,i}}{p_{0,i}} = \beta_{q,0} + \beta_{q,1} x_{i1} + \ldots + \beta_{q,p} x_{ip} = \mathbf{x}_i \boldsymbol{\beta}_q$$

Binomial

$$\ln \frac{p_i}{1 - p_i} = \mathbf{x}_i \boldsymbol{\beta}$$

Rewriting gives the probabilities

Multinomial

$$p_{0,i} = \frac{1}{1 + \sum_{k=1}^{q} \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}_k}}$$

$$p_{1,i} = \frac{\mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}_1}}{1 + \sum_{k=1}^{q} \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}_k}}$$

$$\vdots$$

$$p_{q,i} = \frac{\mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}_q}}{1 + \sum_{k=1}^{q} \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}_k}}$$

Binomial

$$p_{0,i} = 1 - p_i = \frac{1}{1 + \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}$$

$$p_{1,i} = p_i = \frac{\mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}}$$

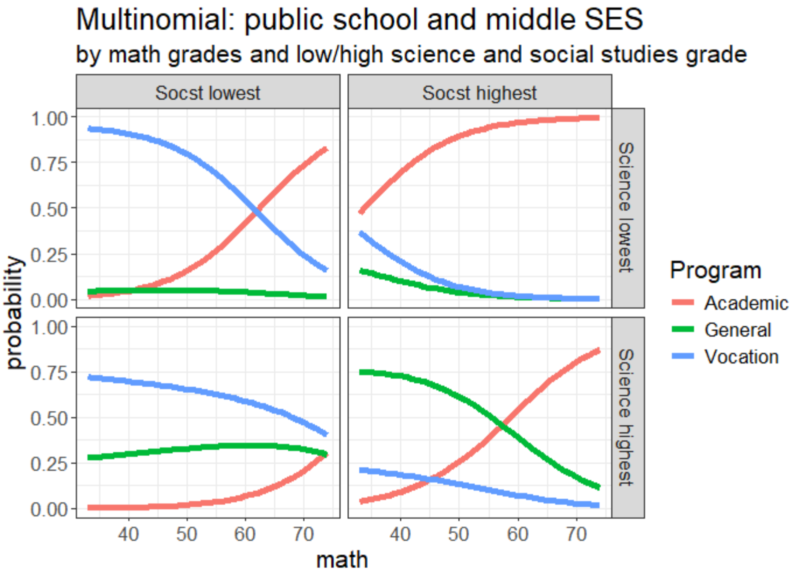
- Estimate all the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ jointly using Maximum-Likelihood.
- The $\hat{\boldsymbol{\beta}}_k$ are again asymptotically multivariate normal distributed and we can use the Wald test for a specific $\beta_{k,j}$ and construct confidence intervals for both $\beta_{k,j}$ and the linear predictors $\mathbf{x}_i \boldsymbol{\beta}_k$ as before.
- Compare nested models with Likelihood Ratio (Deviance) tests.

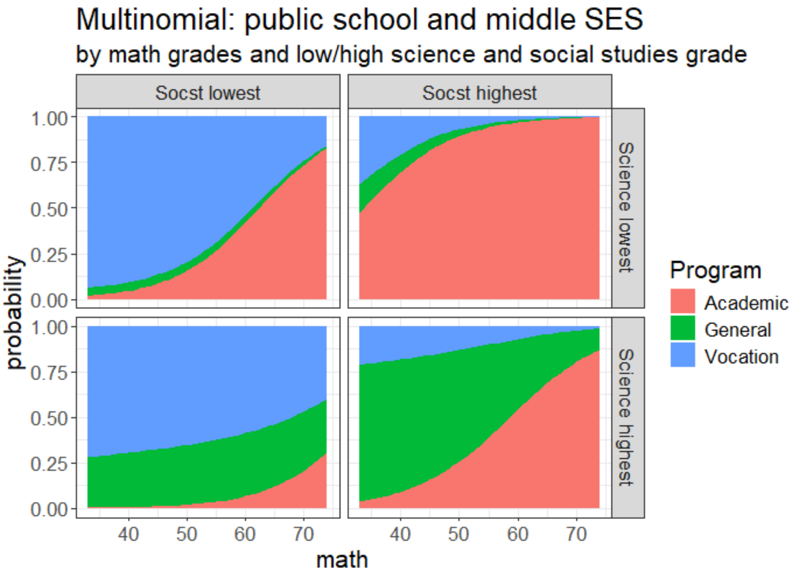
Example: which type of program does a student apply to

- `prog`: type of program student applies to (Academic, General, Vocation). Dependent categorical variable

	General				Academic				
	$\hat{\beta}_{1,j}$	OR	lwr	upr	$\hat{\beta}_{2,j}$	OR	lwr	upr	LR
(Intercept)	−5.96 ***	0.00	0.00	0.08	−9.89 ***	0.00	0.00	0.00	
math	0.02 n.s.	1.02	0.95	1.09	0.14 ***	1.15	1.07	1.23	***
socst	0.05 *	1.05	1.00	1.10	0.09 ***	1.10	1.04	1.16	***
science	0.04 n.s.	1.04	0.99	1.10	−0.04 n.s.	0.96	0.91	1.02	***
schtyppublic	0	1	-	-	0	1	-	-	*
schtypprivate	1.35 n.s.	3.85	0.70	21.11	2.00 **	7.36	1.50	36.19	
seslow	1.40 **	4.06	1.39	11.82	1.22 *	3.39	1.13	10.19	
sesmiddle	0	1	-	-	0	1	-	-	*
seshigh	0.51 n.s.	1.66	0.50	5.54	1.21 *	3.34	1.12	9.96	

- LR test for differences in $\beta_{\cdot,j}$ between any of the three programs.
- The odds of applying to Academic, compared to Vocation, increases by 15 % for each 1 unit higher math grade.
- The math grade has no significant effect on the odds of applying to General, compared to Vocation.
- The science grade has no significant effect on the odds when comparing General or Academic to Vocation, but it has a significant effect when comparing General and Academic with each other.


Multinomial: public school and middle SES
by math grades and low/high science and social studies grade
Socst lowest
Socst highest
Science lowest
Science highest
probability
1.00
0.75
0.50
0.25
0.00
40
50
60
70
math
Program
Academic
General
Vocation


Multinomial: public school and middle SES
by math grades and low/high science and social studies grade
Socst lowest
Socst highest
Science lowest
Science highest
probability
1.00
0.75
0.50
0.25
0.00
40
50
60
70
math
Program
Academic
General
Vocation

Model	AIC	BIC	R^2_{D}	$R^2_{\text{D,adj}}$
Null	412.2	418.8	0.000	0.000
Math	364.2	377.4	0.127	0.122
Grades	350.7	377.1	0.180	0.165
Final	343.6	389.7	0.227	0.198
Full	356.7	435.9	0.244	0.190

Goodness-of-fit

	True				
Predicted	Vocation	Academic	General	Total	PPV
Vocation	29	11	13	53	$29/53 = 54.7\,\%$
Academic	17	87	22	126	$87/126 = 69.0\,\%$
General	4	7	10	21	$10/21 = 47.6\,\%$
Total	50	105	45	200	
Sens.	$29/50 = 58.0\,\%$	$87/105 = 82.9\,\%$	$10/45 = 22.2\,\%$		Accuracy $= \frac{29+87+10}{200} = 63.0\,\%$
Spec.	$1 - \frac{53-29}{200-50} = 84.0\,\%$	$1 - \frac{126-87}{200-105} = 58.9\,\%$	$1 - \frac{21-10}{200-45} = 92.9\,\%$		

McNemar-Bowker test of symmetry

An extension of the McNemar test which tests H_0: $p_{ij} = p_{ji}$, $\forall i \neq j$ against H_1: $p_{ij} \neq p_{ji}$ for some $i \neq j$, i.e., if the confusion matrix is symmetrical.

Ordinal logistic regression

If Y_i is an ordinal categorical variable, where the categories, $k = 1, \ldots, q$ have a logical order, we should do an ordinal logistic regression instead. We then model the log-odds of being in category k or lower instead of in category $k+1$ or higher.

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i > 1)} = \zeta_{1,0} - \beta_1 x_{i1} - \ldots - \beta_p x_{ip} = \zeta_{1,0} - \mathbf{x}_i\boldsymbol{\beta},$$

$$\ln \frac{\Pr(Y_i \leq 2)}{\Pr(Y_i > 2)} = \zeta_{2,0} - \mathbf{x}_i\boldsymbol{\beta},$$

$$\vdots$$

$$\ln \frac{\Pr(Y_i \leq q-1)}{\Pr(Y_i = q)} = \zeta_{q-1,0} - \mathbf{x}_i\boldsymbol{\beta}.$$

The linear part $\mathbf{x}_i\boldsymbol{\beta}$ is the same for all categories, only the intercept $\zeta_{k,0}$ changes!

Note: in $\mathbf{x}_i\boldsymbol{\beta}$ we redefine $\boldsymbol{\beta}$ as $(0, \beta_1, \ldots, \beta_p)$ so we can re-use the notation $\mathbf{x}_i$, including 1 in the first column.

- The parameters, $\boldsymbol{\zeta}_0' = (\zeta_{1,0}, \ldots, \zeta_{q-1,0})$ and $\boldsymbol{\beta}$ are estimated using Maximum-likelihood and are asymptotically normally distributed.
- The probabilities for each of the categories have to be calculated recursively:

$$\ln \frac{1-\hat{p}_{q,i}}{\hat{p}_{q,i}} = \hat{\zeta}_{q-1,0} - \mathbf{x}_i\hat{\boldsymbol{\beta}} \Rightarrow \hat{p}_{q,i} = \frac{1}{1+\mathrm{e}^{\hat{\zeta}_{q-1,0}-\mathbf{x}_i\hat{\boldsymbol{\beta}}}}$$

$$\frac{\hat{\Pr}(Y_i \leq k-1)}{\hat{\Pr}(Y_i > k-1)} = \frac{\sum_{j=1}^{k-1}\hat{p}_{j,i}}{\sum_{j=k}^{q}\hat{p}_{j,i}} = \frac{1-\sum_{j=k}^{q}\hat{p}_{j,i}}{\sum_{j=k}^{q}\hat{p}_{j,i}} = \mathrm{e}^{\hat{\zeta}_{k-1,0}-\mathbf{x}_i\hat{\beta}} \Rightarrow$$

$$\sum_{j=k}^{q}\hat{p}_{j,i} = \frac{1}{1+\mathrm{e}^{\hat{\zeta}_{k-1,0}-\mathbf{x}_i\hat{\boldsymbol{\beta}}}} \Rightarrow$$

$$\hat{p}_{k,i} = \frac{1}{1+\mathrm{e}^{\hat{\zeta}_{k-1,0}-\mathbf{x}_i\hat{\boldsymbol{\beta}}}} - \sum_{j=k+1}^{q}\hat{p}_{j,i}, \text{ for } k = 1, \ldots, q-1$$

- LR test for nested models. AIC, BIC and Confusion matrices for non-nested models.

Example: What grade will a student get on the first exam?

Data: a random sample of $n = 200$ students out of the 690 first-time registered F, I, Pi-students on FMSF80 Mathematical Statistics 9hp in HT22, HT23 and HT24.

We want to model the probabilities for the different results on the main exam in January for the corresponding course registrations. The Fail category includes both those who did not show up at the exam and those who did, but did not pass.

- `exam`: Y_i with $q = 4$ ordered categories; "Fail", "3:a", "4:a", "5:a".
- `tests`: passed the compulory in-course tests on time (No/Yes),
- `mathnbr`: number (0, 1, 2, 3) of passed basic mathematics courses, in addition to the one required to be admitted to the course. Calculus in one variable B1, B2*, in several* variables and Linear algebra ($*$ = one of these is required),
- `mathavg`: average excess grade on the passed mathematics courses (0–2 where 0 = 3:a on all courses and 2 = 5:a on all courses),

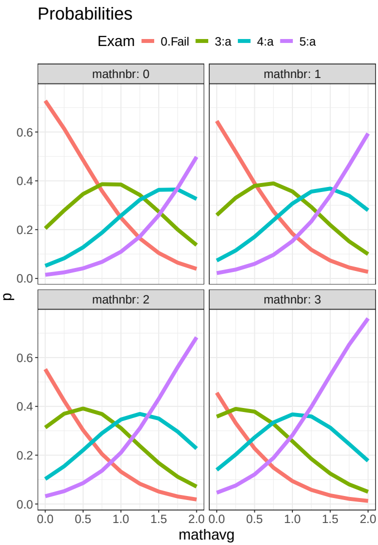
Model

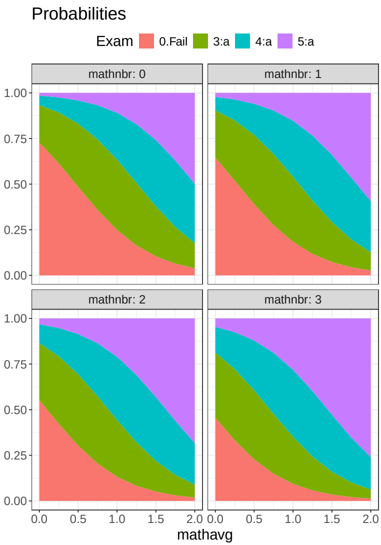
$$\ln \frac{\Pr(Y_i = \text{Fail})}{\Pr(Y_i = \text{3:a, 4:a or 5:a})} = \zeta_{1,0} - \mathbf{x}_i\boldsymbol{\beta},$$
$$\ln \frac{\Pr(Y_i = \text{Fail or 3:a})}{\Pr(Y_i = \text{4:a or 5:a})} = \zeta_{2,0} - \mathbf{x}_i\boldsymbol{\beta},$$
$$\ln \frac{\Pr(Y_i = \text{Fail, 3:a or 4:a})}{\Pr(Y_i = \text{5:a})} = \zeta_{3,0} - \mathbf{x}_i\boldsymbol{\beta}$$

where $\mathbf{x}_i\boldsymbol{\beta} = \beta_1 \cdot \text{testsYes} + \beta_2 \cdot \text{mathnbr} + \beta_3 \cdot \text{mathavg}$.

Variable	Param.	$\hat{\beta}_j$	$e^{\hat{\beta}_j}$	95 % CI(e^{β_j})	
testsYes	β_1	1.56	4.7	2.0	12.0
mathnbr	β_2	0.39	1.47	0.81	2.87
mathavg	β_3	2.09	8.1	4.9	13.7
Fail vs 3:a,4:a,5:a	$\zeta_{1,0}$	2.5	12.7		
Fail,3:a vs 4:a,5:a	$\zeta_{2,0}$	4.2	66.4		
Fail,3:a,4:a vs 5:a	$\zeta_{3,0}$	5.7	312.6		

- A student who has only passed one maths course, got a 3:a on that course and did not pass the tests on time, has an odds of failing $e^{\zeta_{1,0}} = 12.7$, giving a probability of failing as $\frac{12.7}{1+12.7} = 92.7\,\%$.
- If the student had passed the tests on time, the odds of failure would instead be $e^{\zeta_{1,0}-\beta_1} = \frac{e^{\zeta_{1,0}}}{e^{\beta_1}} = \frac{12.7}{4.7} = 2.7$, giving a probability of failure as $\frac{2.7}{1+2.7} = 72.8\,\%$.
- If the student had passad all four maths courses, got one 4:a and three 5:a on them, and passed the tests on time, the odds of getting less than grade 5:a would be
  $e^{\zeta_{3,0}-\beta_1-3\beta_2-1.75\beta_3} = \frac{e^{\zeta_{3,0}}}{e^{\beta_1+3\beta_2+1.75\beta_3}} = \frac{312.6}{12.7\cdot1.47^3\cdot8.1^{1.75}} = 0.53$.
  The probability of getting less than 5:a is then $\frac{0.53}{1+0.53} = 34.7\,\%$ and the probability of getting 5:a is $\frac{1}{1+0.53} = 65.3\,\%$.
- If $\beta_j > 0$, increasing the value of x_j increases the probability of getting a higher grade.


Probabilities
Exam 0.Fail 3:a 4:a 5:a
mathnbr: 0
mathnbr: 1
mathnbr: 2
mathnbr: 3
p
mathavg


Probabilities
Exam 0.Fail 3:a 4:a 5:a
mathnbr: 0
mathnbr: 1
mathnbr: 2
mathnbr: 3
mathavg

Model	df	AIC	BIC	R^2_{D}	$R^2_{\text{D,adj}}$
Null	3	558.2	568.1	0.000	0.000
Final	6	455.3	475.1	0.197	0.192

Confusion matrix

Final model	True					
Predicted	Fail	3:a	4:a	5:a	total	PPV
Fail	32	16	5	3	56	57.1 %
3:a	16	29	9	6	60	48.3 %
4:a	6	6	13	9	34	38.2 %
5:a	3	2	16	29	50	58.0 %
total	57	53	43	47	200	
Sens.	56.2 %	54.7 %	30.2 %	61.7 %		Acc.: 51.5 %

- No information rate: 28.5 % and Acc $>$ NIR is very significant
- McNemar(et al): P-value $= 0.59$ not significant. No systematic under- or over-prediction of grades.
- Grade 4:a is the most difficult to predict correctly.

Distribution free methods

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

- Normally distributed 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.

Quantiles

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

Linear regression

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

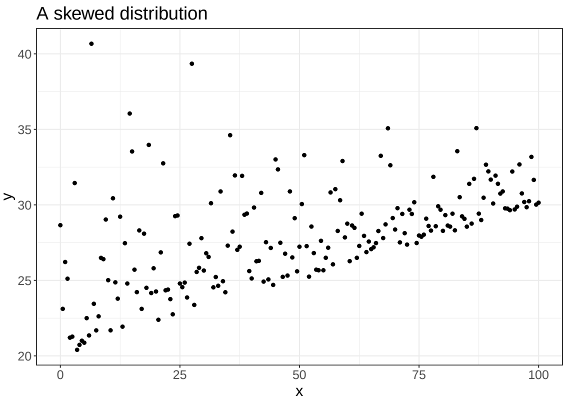
Poisson regression

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

Other distributions

As long as we know ("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.

Least Absolute Deviations

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

$$\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 \qquad \text{(Ordinary least squares)}$$

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

$$\min_{m} \sum_{i=1}^{n} |y_i - m| \qquad \text{(robust to outliers!)}$$

- 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 $\Pr(Y_i > y_{i\alpha} \mid \mathbf{x}_i) = \alpha$.
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 $\displaystyle\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i \boldsymbol{\beta}_\alpha)$ with respect to $\boldsymbol{\beta}_\alpha = \begin{bmatrix} \beta_{0\alpha} \\ \beta_{1\alpha} \\ \vdots \\ \beta_{p\alpha} \end{bmatrix}$

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).

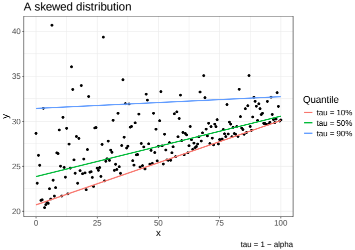
Features

- 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

$$y_{i\alpha} = \hat{\beta}_{0\alpha} + \hat{\beta}_{1\alpha} x_{i1} + \cdots + \hat{\beta}_{p\alpha} x_{ip}$$

  and a proportion $1 - \alpha$ lies below.
- In R: install the package `quantreg`, then `library(quantreg)`
- `rq(y ~ x, data = data, tau = c(0.1, 0.5, 0.9))` where $\tau = 1 - \alpha$ is a list of quantiles.
- `anova(model)` tests if the lines are parallel, i.e. H$_0$: $\beta_{j,\alpha_1} = \ldots = \beta_{j,\alpha_m}$ for $j = 1, \ldots, p$.
- Use $D_0 - D =$ `-2*(logLik(model.red) - logLik(model.full))` $> \chi^2_{0.05}(k)$ to compare models for a specific τ.

Example

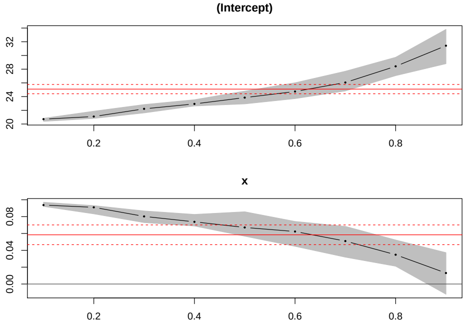


- Test for parallel lines, H_0: $\beta_{1,0.1} = \beta_{1,0.5} = \beta_{1,0.9}$, gives F-value $= 14.64 > F_{0.05,2,601} = 3.01$ (P-value $= 6 \cdot 10^{-7} < 0.05$). The lines are not parallel.
- H_0: $\beta_{1,0.1} = 0$ cannot be rejected and its AIC is very similar to the null model, i.e., the line is horizontal.

	Null model		
	$\tau = 0.1, \alpha = 0.9$	$\tau = \alpha = 0.5$	$\tau = 0.9, \alpha = 0.1$
Intercept ($\hat{\beta}_{0,\alpha}$)	23.9	28.1	32.2
AIC	1156.729	1071.145	1174.435
	Model		
	$\tau = 0.1, \alpha = 0.9$	$\tau = \alpha = 0.5$	$\tau = 0.9, \alpha = 0.1$
Intercept ($\hat{\beta}_{0,\alpha}$)	20.7 (20.4, 20.8)	23.9 (22.9, 24.8)	31.4 (28.8, 33.8)
Slope ($\hat{\beta}_{1,\alpha}$)	0.094 (0.092, 0.097)	0.067 (0.057, 0.086)	0.013 (-0.012, 0.037)
AIC	860.4221	962.9234	1174.0644
LR-test Model vs Null model	$D_0 - D = 298.3$	$D_0 - D = 110.2$	$D_0 - D = 2.37$
Reject if $> \chi^2_{0.05}(1) = 3.84$	sign.	sign.	n.s.

Parameters changing with α

```
plot(summary(rq(..., tau = seq(0.1, 0.9, 0.1))))
```



- Black dots show increasing intercepts and decreasing slopes for varying $\tau = 1 - \alpha$ levels. Grey areas are the corresponding 95 % confidence bands.
- Horizontal red lines are the least squares estimates with 95 % confidence intervals, assuming all quantiles have the same slope.

- You cannot compare nested models using `anova()`.
- You cannot use `step` for forward/backward/stepwise selection.
- `AIC()` can only handle one model at a time and will give AIC for each separate tau. `BIC()` does not work.
- No R^2 or pseudo R^2.
- Advantages of QR:
  - does not require assumptions on the distribution of Y,
  - more robust to outliers than least squares (quantiles do not change that much in presence of outliers).
- Disadvantages of QR: due to lack of distribution assumptions the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ does **not** attain minimal variance (Cramer-Rao bound), unlike maximum likelihood estimates (MLE).