

MASM22/FMSN30: Linear and Logistic
Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp
Lecture 11, spring 2024
Multinomial and ordinal logistic regression
Quantile regression

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

8/5-24

Categorical response variable

Multinomial distribution

Multinomial logistic
regression

Ordinal

Quantile regression

Quantiles

Example

Absolute deviation

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, \dots, p_q where $\sum_{k=0}^q p_k = 1$.
- ▶ Perform the experiment n independent times and calculate the total number of times each outcome occurred as $\mathbf{Y} = (Y_0, \dots, Y_q)$ where $\sum_{k=0}^q Y_k = n$.
- ▶ Then $\mathbf{Y} \sim \text{Multinomial}(n, p_0, \dots, p_q)$ and

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}) &= P(Y_0 = y_0, \dots, Y_q = y_q) = \\ &= \frac{n!}{y_0! \cdot \dots \cdot y_q!} p_0^{y_0} \cdot \dots \cdot p_q^{y_q} \end{aligned}$$

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

Multinomial logistic regression

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

$$p_{0,i} = 1 - \sum_{k=1}^q p_{k,i}$$

Model:

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

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

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

$$\vdots$$

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

Rewriting gives the probabilities

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

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

$$\vdots$$

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

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

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

- ▶ Estimate all the parameter vectors $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q$ 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.

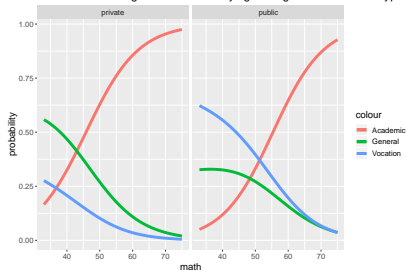
Example: which type of program does a student apply to

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

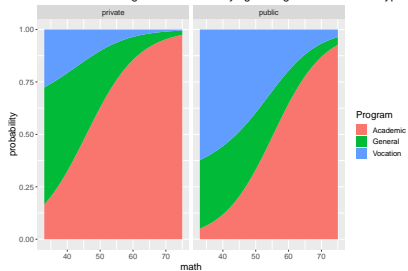
	$\hat{\beta}_j$	General			$\hat{\beta}_j$	Academic			LR
		OR	lwr	upr		OR	lwr	upr	
(Intercept)	-5.96 ***	0.00	0.00	0.08	-9.89 ***	0.00	0.00	0.00	***
math	0.02 n.s.	1.02	0.95	1.09	0.14 ***	1.15	1.07	1.23	***
socst	0.05 *	1.05	1.00	1.10	0.09 ***	1.10	1.04	1.16	***
science	0.04 n.s.	1.04	0.99	1.10	-0.04 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 (deviance) test for differences in β 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: average student with varying math grades and school type



Multinomial: average student with varying math grades and school type



Model	AIC	BIC	R_D^2	$R_{D,adj}^2$
Null	412.2	418.8	0.0 %	0.0 %
Math	364.2	377.4	12.7 %	12.2 %
Grades	350.7	377.1	18.0 %	16.5 %
Final	343.6	389.7	22.7 %	19.8 %
Full	356.7	435.9	24.4 %	19.0 %

Goodness-of-fit

Predicted	True			Total	PPV
	Vocation	Academic	General		
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.	$1 - \frac{53-29}{200-50} = 84.0\%$	$1 - \frac{126-87}{200-105} = 58.9\%$	$1 - \frac{21-10}{200-45} = 92.9\%$		

Ordinal logistic regression

If Y_i is an ordinal categorical variable, where the categories, $k = 1, \dots, q$ 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.

$$\begin{aligned}\ln \frac{Pr(Y_i = 1)}{Pr(Y_i > 1)} &= \zeta_{1,0} - \beta_1 x_{i1} - \dots - \beta_p x_{ip} = \zeta_{1,0} - \mathbf{x}_i \boldsymbol{\beta}, \\ \ln \frac{Pr(Y_i \leq 2)}{Pr(Y_i > 2)} &= \zeta_{2,0} - \mathbf{x}_i \boldsymbol{\beta}, \\ &\vdots \\ \ln \frac{Pr(Y_i \leq q-1)}{Pr(Y_i = q)} &= \zeta_{q-1,0} - \mathbf{x}_i \boldsymbol{\beta}.\end{aligned}$$

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 $(0, \beta_1, \dots, \beta_p)$ so we can re-use the notation \mathbf{x}_i , including 1 in the first column.

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

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

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

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

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

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

Example: How likely is a student to apply to graduate school?

We have Y_i with $q = 3$ ordered categories: "unlikely", "somewhat likely", "very likely". The potential x -variables are

- ▶ pared = at least one parent has a graduate degree (no/yes),
- ▶ gpa = student's grade point average (continuous).

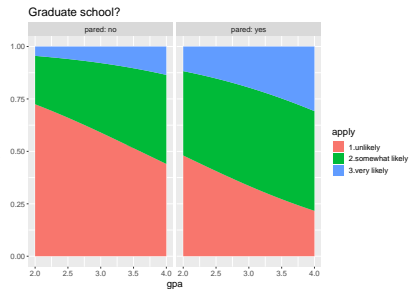
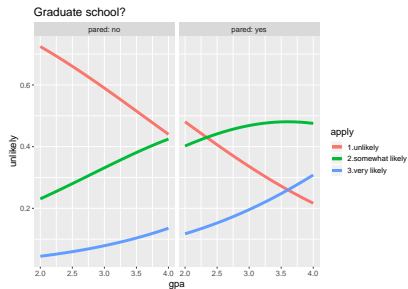
Model

$$\ln \frac{\Pr(Y_i = \text{unlikely})}{\Pr(Y_i = \text{somewhat or very likely})} = \zeta_{1,0} - \mathbf{x}_i \boldsymbol{\beta},$$
$$\ln \frac{\Pr(Y_i = \text{unlikely or somewhat likely})}{\Pr(Y_i = \text{very likely})} = \zeta_{2,0} - \mathbf{x}_i \boldsymbol{\beta}$$

where $\mathbf{x}_i \boldsymbol{\beta} = \beta_1 \cdot \text{pared} + \beta_2 \cdot \text{gpa}$.

Variable	Param.	Est,	e^{est}	95 % CI(e^{β_j})	
pared	β_1	1.046	2.85	1.69	4.81
gpa	β_2	0.604	1.83	1.12	3.02
unlikely vs somewhat or very	$\zeta_{1,0}$	2.176	8.81		
unlikely or somewhat vs very	$\zeta_{2,0}$	4.272	71.64		

- ▶ A student where neither parent has a graduate degree and who had 0 as grade point average, has a probability to answer "unlikely" that is $e^{\zeta_{1,0}} = 8.81$ times as large as the probability to answer something more positive, i.e, the odds is 8.81.
- ▶ If at least one parent has a graduate degree, the probability to answer "unlikely" is only $e^{\zeta_{1,0}-\beta_1} = \frac{e^{\zeta_{1,0}}}{e^{\beta_1}} = \frac{8.81}{2.85} = 3.10$ times as large.
- ▶ If he also has a grade point average of 3, the probability to answer "unlikely" is $e^{\zeta_{1,0}-\beta_1-3\beta_2} = \frac{e^{\zeta_{1,0}}}{e^{\beta_1+3\beta_2}} = \frac{8.81}{2.85 \cdot 1.84^3} = 0.51$ times as large.
- ▶ If $\beta_j > 0$, increasing the value of x_j increases the probability of giving a more positive answer.



Model	AIC	BIC	R_D^2	$R_{D,adj}^2$
Null	745.2	753.2	0.0 %	0.0 %
Final	725.1	741.0	3.3 %	3.0 %
Full	727.0	747.0	3.3 %	2.9 %

Confusion matrix

Final model Predikted	True			total	PPV
	unlikely	somewhat	very		
unlikely	201	110	27	338	59.5 %
somewhat	19	30	13	62	48.8 %
very	0	0	0	0	—
total	220	140	40	400	
Sens.	91.4 %	21.4 %	0.0 %	57.8 %	Acc.: 57.75 %

Categorical response variable

Multinomial distribution

Multinomial logistic
regression

Ordinal

Quantile regression

Quantiles

Example

Absolute deviation

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.

Now we consider a methodology which is “distribution free”.

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 $Y_i = \mathbf{x}_i\beta + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ we have $y_{i\alpha} = \mathbf{x}_i\beta + \lambda_\alpha \cdot \sigma$.

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

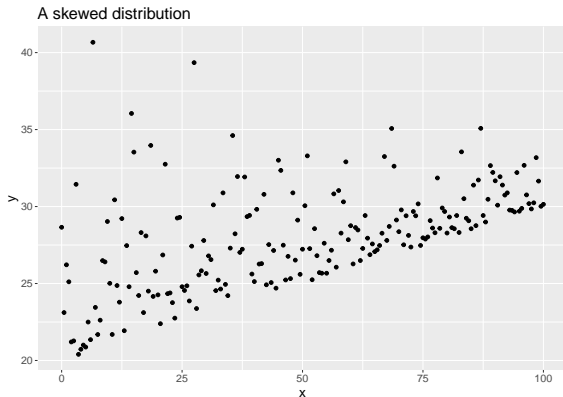
With $Y_i \sim Po(e^{\mathbf{x}_i\beta})$ we can use the quantiles in the (estimated) Poisson distribution.

Other distributions

As long as we know ("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

$$\min_{\mu} \sum_{i=1}^n (y_i - \mu)^2$$

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

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

- ▶ 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 $Pr(Y_i > y_{i\alpha} | \mathbf{x}_i) = \alpha$.

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

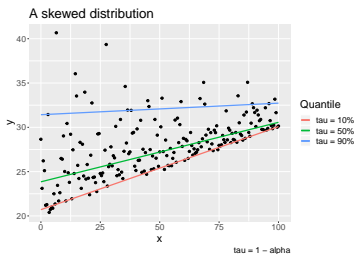
- ▶ the resulting α -quantile regression line is such that, for given covariates \mathbf{x}_i , a proportion of approximately α data points lies above the fitted value

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

and a proportion $1 - \alpha$ lies below.

- ▶ 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.
 $H_0: \beta_{j,\alpha_1} = \dots = \beta_{j,\alpha_m}$ for $j = 1, \dots, p$.
- ▶ Use $D_0 - D = -2 * (\log\text{Lik}(\text{model.red}) - \log\text{Lik}(\text{model.full})) > \chi^2_{0.05}(k)$ to compare models for a specific τ .

Example (cont)

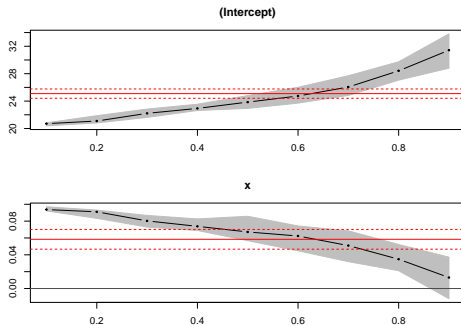


- ▶ Test for parallel lines,
 $H_0: \beta_{1,0.1} = \beta_{1,0.5} = \beta_{1,0.9}$, gives
 $F\text{-value} = 14.64 > F_{0.05,2,601} = 3.01$
(P-value = $6 \cdot 10^{-7} < 0.05$). The lines are not parallel.
- ▶ $H_0: \beta_{1,0.1} = 0$ cannot be rejected and its AIC is very similar to the null model, 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 $\hat{\beta}_\alpha$ does **not** attain minimal variance (Cramer-Rao bound), unlike maximum likelihood estimates (MLE).