Models for Count Data and Categorical Response Data

Christopher F Baum

Boston College and DIW Berlin

June 2010

Poisson regression

In statistical analyses, dependent variables may be limited by being *count data*, only taking on nonnegative (or only positive) integer values. This is a natural form for data such as the number of children per family, the number of jobs an individual has held or the number of countries in which a company operates manufacturing facilities.

Just as with the other limited dependent variable models we have discussed, linear regression is not an appropriate estimation technique for count data, as it fails to take into account the limited number of possible values of the response variable.

The most common technique employed to model count data is *Poisson regression*, so named because the error process is assumed to follow the Poisson distribution. As an aside, you may notice that the insignia (colophon) of Stata Press appears to be a soldier with a horse. The Poisson distribution was first applied to data on the number of Prussian cavalrymen who died after being kicked by a horse, and the colophon refers to that historical detail.

The technique is implemented in Stata by the `poisson` command, which has the same format as other estimation commands, where the *depvar* is a nonnegative count variable; that is, it may be zero. It is a maximum likelihood estimation technique.

In some contexts, the Poisson distribution describes the number of events that occur in a given time period where its mean μ is the average number of events per period. It has the unusual feature that its mean equals its variance. Its probability density function is $\Pr(Y = y) = \left(\frac{e^{-\mu}\mu^{y}}{y!}\right)$, y=0,1,2,... where *e* is the base of the natural logarithms and *y*! is the factorial of *y*.

The skewness of the Poisson distribution is $(1/\sqrt{\mu})$ and the kurtosis is $(3 + 1/\mu)$, so that for large μ, the distribution approaches the Normal $N(\mu, \mu)$ with skewness of zero and kurtosis of three.

We illustrate count data techniques using a dataset from the U.S. Medical Expenditure Panel Survey (MEPS) containing information on the number of doctor visits in 2003 (`docvis`) for a number of elderly patients as well as a number of patient characteristics.

`private` is an indicator of private insurance coverage, supplemental to Medicare. `medicaid` indicates the patient is eligible for low-income Medicaid coverage. `actlim` indicates the presence of activity limitations, while `totchr` is the number of chronic conditions. `educyr` indicates the number of years of education attained.

```
. summarize docvis private medicaid age age2 educyr actlim totchr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
docvis	3677	6.822682	7.394937	0	144
private	3677	.4966005	.5000564	0	1
medicaid	3677	.166712	.3727692	0	1
age	3677	74.24476	6.376638	65	90
age2	3677	5552.936	958.9996	4225	8100
educyr	3677	11.18031	3.827676	0	17
actlim	3677	.333152	.4714045	0	1
totchr	3677	1.843351	1.350026	0	8

The default parameterization of the Poisson model, in which the conditional mean of observation *i* depends on a number of covariates, is the exponential mean:

$$\mu_i = \exp(x_i'\beta), \ i = 1, ..., N$$

This model may be estimated by maximum likelihood (ML), where the parameter estimates are the solutions to the first order conditions

$$\sum_{i=1}^{N}(y_i - \exp(x_i'\beta))x_i = 0$$

The likelihood function is globally concave and the estimation converges rapidly.

```
. poisson docvis private medicaid age age2 educyr actlim totchr, nolog
Poisson regression                              Number of obs   =       3677
                                                LR chi2(7)      =    4477.98
                                                Prob > chi2     =     0.0000
Log likelihood =  -15019.64                     Pseudo R2       =     0.1297
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.1422324	.0143311	9.92	0.000	.114144	.1703208
medicaid	.0970005	.0189307	5.12	0.000	.0598969	.134104
age	.2936722	.0259563	11.31	0.000	.2427988	.3445457
age2	-.0019311	.0001724	-11.20	0.000	-.0022691	-.0015931
educyr	.0295562	.001882	15.70	0.000	.0258676	.0332449
actlim	.1864213	.014566	12.80	0.000	.1578726	.2149701
totchr	.2483898	.0046447	53.48	0.000	.2392864	.2574933
_cons	-10.18221	.9720115	-10.48	0.000	-12.08732	-8.277101

If the model is correctly specified, but the distribution of errors is not Poisson (as we will discuss next) one approach is to estimate the model with pseudo-ML, generating robust standard errors:

```
. poisson docvis private medicaid age age2 educyr actlim totchr, ///
>         vce(robust) nolog
Poisson regression                              Number of obs   =       3677
                                                Wald chi2(7)    =     720.43
                                                Prob > chi2     =     0.0000
Log pseudolikelihood =  -15019.64               Pseudo R2       =     0.1297
```

docvis	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.1422324	.036356	3.91	0.000	.070976	.2134889
medicaid	.0970005	.0568264	1.71	0.088	-.0143773	.2083783
age	.2936722	.0629776	4.66	0.000	.1702383	.4171061
age2	-.0019311	.0004166	-4.64	0.000	-.0027475	-.0011147
educyr	.0295562	.0048454	6.10	0.000	.0200594	.039053
actlim	.1864213	.0396569	4.70	0.000	.1086953	.2641474
totchr	.2483898	.0125786	19.75	0.000	.2237361	.2730435
_cons	-10.18221	2.369212	-4.30	0.000	-14.82578	-5.538638

Although all parameters (except `medicaid`) are still highly significant, the standard errors and z-statistics are much smaller, indicating that the errors may not be distributed as Poisson.

The coefficients may be interpreted as semielasticities. A coefficient of 0.029 on `educyr` indicates that a patient with one more year of education is expected to have 2.9% more doctor visits.

The average marginal effects (AMEs) may be calculated with `margins`:

```
. margins, dydx(_all)
Average marginal effects                        Number of obs   =       3677
Model VCE    : Robust
Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.private 1.medicaid age age2 educyr 1.actlim totchr
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.private	.9701906	.2473149	3.92	0.000	.4854622	1.454919
1.medicaid	.6830664	.4153252	1.64	0.100	-.130956	1.497089
age	2.003632	.4303207	4.66	0.000	1.160219	2.847045
age2	-.0131753	.0028473	-4.63	0.000	-.0187559	-.0075947
educyr	.2016526	.0337805	5.97	0.000	.1354441	.2678612
1.actlim	1.295942	.2850588	4.55	0.000	.7372367	1.854647
totchr	1.694685	.0908883	18.65	0.000	1.516547	1.872823

Note: dy/dx for factor levels is the discrete change from the base level.

An individual with one more year of education is predicted to have 0.2 more visits, other things equal.

Negative binomial regression

A limitation of the Poisson distribution is the equality of its mean and variance. We may often observe count data processes where this equality is not reasonable: in particular, where the conditional variance is larger than the conditional mean. This is termed *overdispersion*, and its presence renders the assumption of a Poisson distribution for the error process untenable. It is particularly likely to occur in the case of unobserved heterogeneity.

In this circumstance, a reasonable alternative is *negative binomial regression*. This model allows the variance to differ from the mean. In its Stata implementation as `nbreg`, a Poisson model is also estimated and a test of overdispersion is provided. If the dispersion parameter is zero, it is appropriate to fit a Poisson regression model.

The negative binomial (NB) distribution is a two-parameter distribution. For positive integer *n*, it is the distribution of the number of failures that occur in a sequence of trials before *n* successes have occurred, where the probability of success in each trial is *p*. The distribution is defined for any positive *n*. The negative binomial distribution is a mixture of the Poisson distribution and the Gamma distribution, or generalized factorial function.

Unlike the Poisson, which is fully characterized by its mean μ, the NB distribution is a function of both μ and α. Its mean is still μ, but its conditional variance is $\mu(1 + \alpha\mu)$. As is evident, as $\alpha \rightarrow 0$, the distribution becomes the Poisson distribution.

We reestimate the model with Stata's `nbreg`:

```
. nbreg docvis private medicaid age age2 educyr actlim totchr, nolog
Negative binomial regression                    Number of obs   =       3677
                                                LR chi2(7)      =     773.44
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -10589.339                     Pseudo R2       =     0.0352
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.1640928	.0332186	4.94	0.000	.0989856	.2292001
medicaid	.100337	.0454209	2.21	0.027	.0113137	.1893603
age	.2941294	.0601588	4.89	0.000	.1762203	.4120384
age2	-.0019282	.0004004	-4.82	0.000	-.0027129	-.0011434
educyr	.0286947	.0042241	6.79	0.000	.0204157	.0369737
actlim	.1895376	.0347601	5.45	0.000	.121409	.2576662
totchr	.2776441	.0121463	22.86	0.000	.2538378	.3014505
_cons	-10.29749	2.247436	-4.58	0.000	-14.70238	-5.892595
/lnalpha	-.4452773	.0306758			-.5054007	-.3851539
alpha	.6406466	.0196523			.6032638	.6803459

```
Likelihood-ratio test of alpha=0:  chibar2(01) = 8860.60 Prob>=chibar2 = 0.000
```

The likelihood ratio test of $\alpha = 0$ strongly rejects the null hypothesis that the errors do not exhibit overdispersion. Thus, the Poisson regression model is rejected in favor of its generalized version, the NB regression model. The coefficients are similar between the two models, and the NB estimates are comparable to those from `poisson` with robust standard errors.

Extended count data models

In many social science datasets, count data may include a large number of zero values. If the data were dichotomized into zero and non-zero subsets so that a probit or logit model could be fit, the unconditional probability of zero would be sizable: larger than that arising in a Poisson or negative binomial distribution. For instance, we might have a random sample from the population in which the number of postgraduate degrees is recorded. For many individuals, the count will be zero. For many professionals, it will be one, and for most academics, it will be two (or more).

To model data with these characteristics, we may employ the *zero-inflated* variants of Poisson regression (`zip`) or negative binomial regression (`zinb`). In these commands, there is an auxiliary logit model specified in the `inflate()`) option that determines whether the observed count is zero. This model could contain only a constant or additional covariates.

With the `vuong` option, a test of the ZIP versus standard Poisson regression model is computed. For `zinb`, the `zip` option computes a test of the `zinb` model versus the zero-inflated Poisson model which is nested within.

To illustrate these models, we consider a different dependent variable: the number of emergency room (ER) visits, which is small for most patients, with 80% of the sample recording no ER visits in 2003. The sample mean of `er` is 0.2774. We could choose to ignore the high prevalence of zeros and fit a `nbreg` model:

```
. nbreg er age actlim totchr, nolog
Negative binomial regression                    Number of obs   =       3677
                                                LR chi2(3)      =     225.15
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -2314.4927                     Pseudo R2       =     0.0464
```

er	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0088528	.0061341	1.44	0.149	-.0031697	.0208754
actlim	.6859572	.0848127	8.09	0.000	.5197274	.8521869
totchr	.2514885	.0292559	8.60	0.000	.1941481	.308829
_cons	-2.799848	.4593974	-6.09	0.000	-3.700251	-1.899446
/lnalpha	.4464685	.1091535			.2325315	.6604055
alpha	1.562783	.1705834			1.26179	1.935577

```
Likelihood-ratio test of alpha=0:  chibar2(01) =  237.98 Prob>=chibar2 = 0.000
```

The likelihood ratio test rejects the Poisson distribution. But should these data be treated as zero-inflated? To use `zinb`, we must specify the `inflate()` option, listing the variable or variables that are expected to influence whether the count is zero or not.

```
. zinb er age actlim totchr, inflate(totchr) vuong nolog
Zero-inflated negative binomial regression      Number of obs   =      3677
                                                Nonzero obs     =       710
                                                Zero obs        =      2967
Inflation model = logit                         LR chi2(3)      =     98.06
Log likelihood  =  -2310.65                     Prob > chi2     =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
er						
age	.0076908	.006134	1.25	0.210	-.0043317	.0197133
actlim	.6761249	.0849168	7.96	0.000	.509691	.8425588
totchr	.1600338	.0461155	3.47	0.001	.0696492	.2504185
_cons	-2.333669	.501506	-4.65	0.000	-3.316603	-1.350736
inflate						
totchr	-.8182987	.3673752	-2.23	0.026	-1.538341	-.0982565
_cons	-.3149276	.4843635	-0.65	0.516	-1.264263	.6344074
/lnalpha	.2305631	.2038915	1.13	0.258	-.169057	.6301832
alpha	1.259309	.2567625			.8444608	1.877955

```
Vuong test of zinb vs. standard negative binomial: z =     1.35  Pr>z = 0.0885
```

The results show that `totchr` is significant in the logit estimation of `er` as zero or nonzero. It is also significant, along with `actlim`, in the estimated equation.

The Vuong test weakly rejects the standard negative binomial model in favor of the zero-inflated NB model.

A second variant of the count data model appears when we only record positive integer values for the response variable, although zero values appear in the population. This is a form of *truncation*, as discussed earlier. This could occur, for example, if we collected data from an elementary school from each pupil on how many children under 18 were in their family. This would only capture information from households containing children: a subset of households in the population.

To make appropriate inferences on the population, we must take into account the truncated nature of the data. In Stata, this technique is implemented as *zero-truncated* Poisson regression (`ztp`) or negative binomial regression (`ztnb`).

We illustrate by fitting a zero-truncated Poisson regression model on the nonzero observations of `er`:

```
. ztp er age actlim totchr if er>0, nolog
Zero-truncated Poisson regression               Number of obs   =        710
                                                LR chi2(3)      =     196.31
                                                Prob > chi2     =     0.0000
Log likelihood = -642.72434                     Pseudo R2       =     0.1325

------------------------------------------------------------------------------
          er |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0013535   .0082979     0.16   0.870      -.01491    .0176171
      actlim |   .2402127   .1218004     1.97   0.049     .0014884    .4789371
      totchr |   .1370198   .0384868     3.56   0.000      .061587    .2124525
       _cons |  -.8600034   .6309487    -1.36   0.173     -2.09664    .3766333
------------------------------------------------------------------------------
```

Multinomial logit models

Categorical data often fall into one of several mutually exclusive categories: e.g., different ways of commuting to work, or different categories of self-assessed health status. In the latter case, where the categories are ordered, we may utilize ordered probit or ordered logit techniques, as we have discussed. But in the case where the choices are underordered, we have *multinomial* data, with the most common technique being that of *multinomial logit*.

The outcome, y_i, is one of m alternatives. We set $y_i = j$ if the outcome is the j^{th} alternative. The probability that individual i chooses alternative j, conditional on regressors x_i, is:

$$p_{ij} = \Pr(y_i = j) = F_j(X_i, \theta), \; j = 1, \ldots, m, \; i = 1, \ldots, N$$

with different functional forms $F_j(\cdot)$ corresponding to different multinomial models.

Only $m-1$ of the probabilities can be freely specified, as they must sum to unity. Multinomial models require a normalization. Their parameters are generally not directly interpretable: for instance, a positive coefficient on x_k does not imply that an increase in x_k increases the probability that the alternative is selected.

Instead, marginal effects are computed for individual i, alternative j, and regressor k:

$$ME_{ijk} = \frac{\partial \Pr(y_i = j)}{\partial x_{ik}} = \frac{\partial F_j(x_i, \theta)}{\partial x_{ik}}$$

For each regressor, there will be m marginal effects corresponding to the m probabilities, and the marginal effects must sum to zero. As with other nonlinear models, the marginal effects vary with the point at which they are evaluated, x_i.

Some regressors, such as gender, do not vary across alternatives and are termed *case-specific* regressors. Other regressors, such as price or time, may vary across alternatives and are termed *alternative-specific* regressors. For instance, we may record the price charged by different vendors from which an individual could buy the good, or the time required for each commuting mode.

The Stata commands used to estimate multinomial logit models vary according to the form of regressors. In the simplest case, all regressors are case-specific, and we may use the `mlogit` command. In more complicated specifications, some or all of the regressors are alternative-specific, and we could use the `asclogit` command. Other choices exist, including *nested logit* (`nlogit`) and *stereotype logit* (`slogit`).

We illustrate with a dataset on individuals choosing one of four fishing modes: from the beach, the pier, a private boat or a charter boat. Selected characteristics of the dataset are:

```
. summarize mode price crate d* income, sep(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mode	1182	3.005076	.9936162	1	4
price	1182	52.08197	53.82997	1.29	666.11
crate	1182	.3893684	.5605964	.0002	2.3101
dbeach	1182	.1133672	.3171753	0	1
dpier	1182	.1505922	.3578023	0	1
dprivate	1182	.3536379	.4783008	0	1
dcharter	1182	.3824027	.4861799	0	1
income	1182	4.099337	2.461964	.4166667	12.5

`mode` is the choice of fishing mode; the `d` variables are indicators of each choice. `price` and `crate` are the price and catch rate for the chosen mode. `income` is monthly income in thousands of USD.

We may examine how income varies across fishing mode:

```
. table mode, contents(N income mean income sd income)
```

Fishing mode	N(income)	mean(income)	sd(income)
beach	134	4.051617	2.50542
pier	178	3.387172	2.340324
private	418	4.654107	2.777898
charter	452	3.880899	2.050029

We see that the highest-income anglers use a private boat, and that the lowest-income individuals fish from the pier.

We now fit a multinomial logit, using the only case-specific regressor, and mode 1 (beach) as the base outcome:

```
. mlogit mode income, baseoutcome(1) nolog
Multinomial logistic regression                 Number of obs   =       1182
                                                LR chi2(3)      =      41.14
                                                Prob > chi2     =     0.0000
Log likelihood = -1477.1506                     Pseudo R2       =     0.0137
```

mode	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
beach	(base outcome)					
pier						
income	-.1434029	.0532884	-2.69	0.007	-.2478463	-.0389595
_cons	.8141503	.228632	3.56	0.000	.3660399	1.262261
private						
income	.0919064	.0406637	2.26	0.024	.0122069	.1716058
_cons	.7389208	.1967309	3.76	0.000	.3533352	1.124506
charter						
income	-.0316399	.0418463	-0.76	0.450	-.1136571	.0503774
_cons	1.341291	.1945167	6.90	0.000	.9600457	1.722537

Although the overall fit (as judged by `Pseudo R2`) is poor, the χ^2 test against the null model is highly significant. We test whether `income` is an important determinant with a joint (Wald) test on the three coefficients:

```
. test income
 ( 1)  [beach]income = 0
 ( 2)  [pier]income = 0
 ( 3)  [private]income = 0
 ( 4)  [charter]income = 0
       Constraint 1 dropped
           chi2(  3) =   37.70
         Prob > chi2 =    0.0000
```

The multinomial logit model is equivalent to a series of pairwise logit models comparing each category with the base category. A positive coefficient thus indicates that as x_k increases, we are more likely to choose alternative j than the base category, number 1. The coefficients may also be expressed as *proportional odds* or *relative-risk* ratios,

$$\frac{\Pr(y_i = j)}{\Pr(y_i = 1)} = \exp(x_i \beta_j)$$

```
. mlogit mode income, rr baseoutcome(1) nolog
Multinomial logistic regression                 Number of obs   =       1182
                                                LR chi2(3)      =      41.14
                                                Prob > chi2     =     0.0000
Log likelihood = -1477.1506                     Pseudo R2       =     0.0137
```

mode	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
beach	(base outcome)					
pier						
income	.8664049	.0461693	-2.69	0.007	.7804799	.9617896
private						
income	1.096262	.0445781	2.26	0.024	1.012282	1.18721
charter						
income	.9688554	.040543	-0.76	0.450	.8925639	1.051668

A one thousand dollar increase in income leads to relative odds of choosing to fish from a pier (rather than the beach) of 0.866 times what they were at the original level of income, so the relative odds (pier vs. beach) have declined.

We may also create predictions for each alternative and individual.

```
. predict pml1 pml2 pml3 pml4, pr
. summarize pml*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pml1	1182	.1133672	.0036716	.0947395	.1153659
pml2	1182	.1505922	.0444575	.0356142	.2342903
pml3	1182	.3536379	.0797714	.2396973	.625706
pml4	1182	.3824027	.0346281	.2439403	.4158273

These predicted values have the same means as the observed data, by construction. As the predicted values for beach fishing (`pml1`) vary only between 0.094 and 0.115, the model using only income performs very poorly. Ideally, it would produce predictions of 1.0 for the 134 individuals that chose beach fishing, and 0 for the rest.

We may calculate marginal effects:

$$\frac{\partial p_{ij}}{\partial x_i} = p_{ij}(\beta_j - \bar{\beta}_i)$$

where $\bar{\beta}_i$ is a probability-weighted average of the estimated β coefficients. The marginal effects vary with the point in regressor space as p_{ij} varies with x_i.

The signs of the coefficients do not give the signs of the marginal effects, as the sign of the marginal effect is positive if $\beta_j > \bar{\beta}_i$.

```
. margins, predict(pr outcome(3)) dydx(income)
Average marginal effects                        Number of obs   =       1182
Model VCE    : OIM
Expression   : Pr(mode==private), predict(pr outcome(3))
dy/dx w.r.t. : income
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
income	.0317562	.0052589	6.04	0.000	.021449	.0420633

A one-unit change (thousand-dollar increase) in `income` increases the probability of choosing to fish from a private boat by 0.032, or 3.2%.

Alternative-specific multinomial logit

When alternative-specific data are available, they must be transformed into the *long form* by `reshape` so that each individual has one record per alternative. The `asclogit` command can then be employed. The `case()` option is used to identify the individual, `alternatives()` specifies the choices and `casevars()` may be used to give a *varlist* of case-specific regressors.

In the long form fishing data, `d` indicates the mode choice, `p` indicates the choice-specific price and `q` gives the choice-specific catch rate. We also use the case-specific variable `income`.

```
. asclogit d p q, case(id) alternatives(fishmode) ///
>   casevars(income) basealternative(beach) nolog
```

Alternative-specific conditional logit — Number of obs = 4728
Case variable: id — Number of cases = 1182
Alternative variable: fishmode — Alts per case: min = 4
avg = 4.0
max = 4
Wald chi2(5) = 252.98
Log likelihood = -1215.1376 — Prob > chi2 = 0.0000

d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fishmode						
p	-.0251166	.0017317	-14.50	0.000	-.0285106	-.0217225
q	.357782	.1097733	3.26	0.001	.1426302	.5729337

beach	(base alternative)					
charter						
income	-.0332917	.0503409	-0.66	0.508	-.131958	.0653745
_cons	1.694366	.2240506	7.56	0.000	1.255235	2.133497
pier						
income	-.1275771	.0506395	-2.52	0.012	-.2268288	-.0283255
_cons	.7779593	.2204939	3.53	0.000	.3457992	1.210119
private						
income	.0894398	.0500671	1.79	0.074	-.0086898	.1875694
_cons	.5272788	.2227927	2.37	0.018	.0906132	.9639444

In the alternative-specific regression, we may readily interpret the coefficients for the r^{th} regressor:

$$\frac{\partial p_{ij}}{\partial x_{rik}} = \begin{cases} p_{ij}(1 - p_{ij})\beta_r & j = k \\ -p_{ij}p_{ik}\beta_r & j \neq k \end{cases}$$

If $\beta_r > 0$, then the own-effect is positive, and the cross-effect is negative. A positive coefficent indicates that category *j* is chosen more frequently and other categories are chosen less frequently, and vice versa.

In our example, the negative price coefficient (p) indicates that an increase in the price of choice *j* causes it to be chosen less often. The positive coefficient on the catch rate, q, indicates that an increase causes choice *j* to be choisen more often. An increase in income reduces the probability of charter boat fishing and pier fishing, and increases the probability of private boat fishing, relative to beach fishing.

Another alternative specification of this model could be made employing the *nested logit* (in Stata, `nlogit`) technique. In this framework, we assume that individuals make a sequence of choices. For instance, they choose a fishing mode of *shore* or *boat*, perhaps depending how much they like being out on the water. After choosing a mode, they then choose among the alternatives in that branch of the decision tree. For instance, for fishing from shore, they then choose *beach* or *pier*. This model may be relevant for a number of outcomes that can be considered as sequential choices. We do not discuss it further.

DIscriminant analysis

Limited-dependent-variable techniques such as binomial logit or probit may be used to model decisions, such as a lender's willingness to extend credit to an applicant or a consumer's willingness to purchase a product. Another body of statistical methodology that may be used to analyze data of that nature is *discriminant analysis*, also known in some contexts as *classification*.

Discriminant analysis describes the difference between groups in order to exploit those differences in allocating of classifying observations of unknown group membership to the groups.

These techniques, as implemented in Stata by subcommands of the `discrim` command, include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discriminant analysis and k^{th}-nearest-neighbor discriminant analysis (KNN). These techniques may be both *predictive* and *descriptive*, depending on whether the researcher is seeking to classify unknown observations or to merely analyze the determinants of group membership.

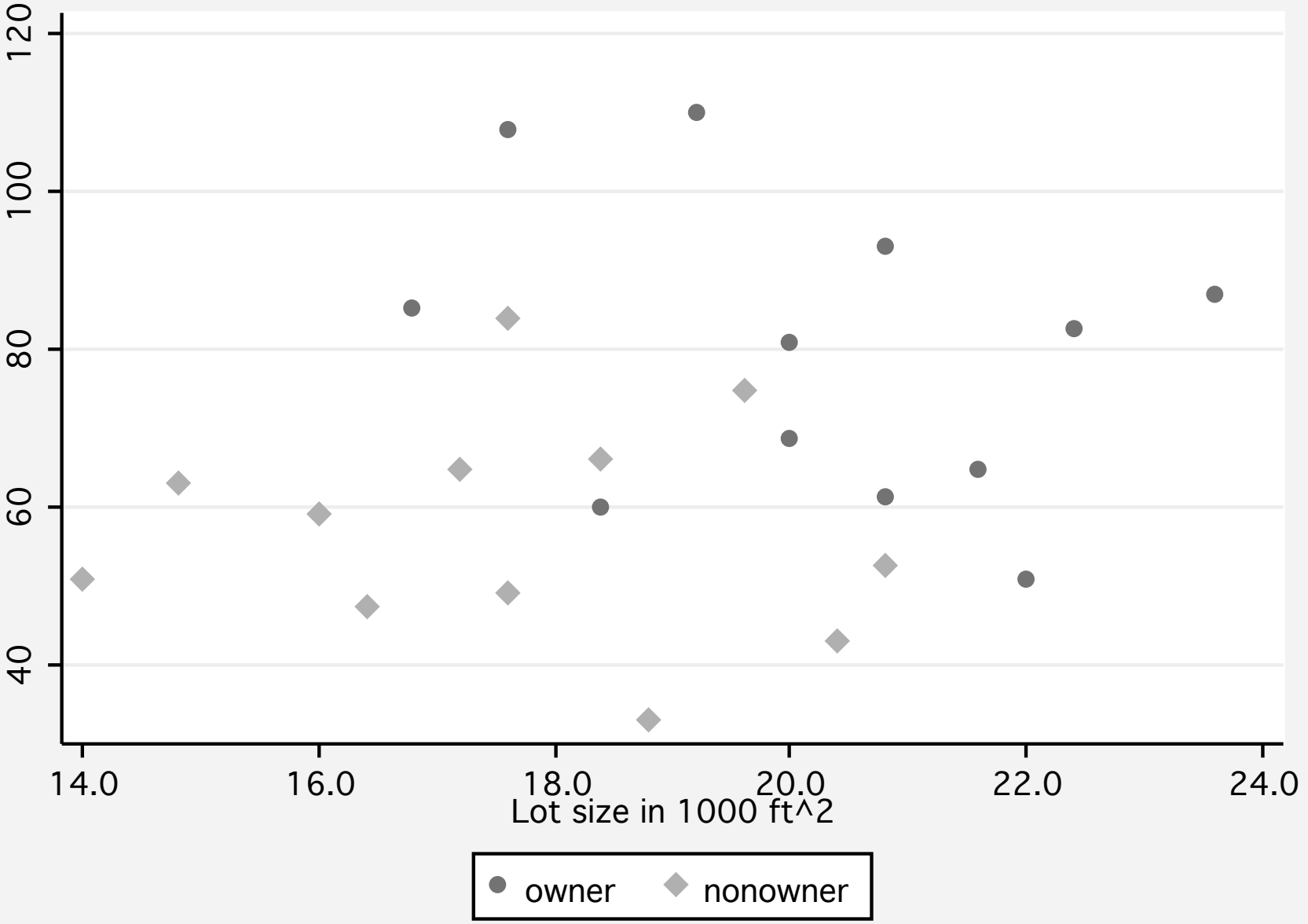
As an example, consider a dataset in which 12 riding-lawnmower owners and 12 nonowners appear, with their family income and lot size. Using predictive analysis, do these variables adequately classify observations into owner/nonowner status? We apply linear discriminant analysis (LDA) with `discrim lda`:

```
. discrim lda lotsize income, group(owner)
Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True owner	Classified nonowner	owner	Total
nonowner	10 83.33	2 16.67	12 100.00
owner	1 8.33	11 91.67	12 100.00
Total	11 45.83	13 54.17	24 100.00
Priors	0.5000	0.5000	

This classification table shows that 10 of the nonowners and 11 of the owners are correctly classified, with three being misclassified. A *leave-one-out* analysis provides a more robust approach, using a sort of jackknife strategy to build the LDA model, and using it to classify each omitted observation in turn.


120
100
80
60
40
14.0
16.0
18.0
20.0
22.0
24.0
Lot size in 1000 ft^2
owner
nonowner

```
. estat classtable, loo nopriors
Leave-one-out classification table
```

Key
Number Percent

True owner	LOO Classified nonowner	owner	Total
nonowner	9 75.00	3 25.00	12 100.00
owner	2 16.67	10 83.33	12 100.00
Total	11 45.83	13 54.17	24 100.00

With leave-one-out (`loo`) classification, we see that 5 (rather than 3) of the 24 observations are misclassified.

We may use predictive discriminant analysis to explore how the groups are separated. The postestimation command `estat loadings` allows us to view the discriminant function coefficients, or loadings.

```
. estat loadings, unstandardized
Canonical discriminant function coefficients
```

	function1
lotsize	.3795228
income	.0484468
_cons	-10.50754

These coefficients may be expressed as the equation

$$lotsize = -0.1277 income + 27.6862$$

which provides the separating line between riding-lawnmower owners and nonowners.

The difference between the `discrim` techniques involves the choice of density function for each group. The LDA technique assumes that the groups are multivariate normal with equal covariance matrices. The QDA technique assumes that they are multivariate normal with potentially unequal covariance matrices. The k^{th} nearest neighbor (KNN) technique is a nonparametric alternative, similar to kernel density estimation.

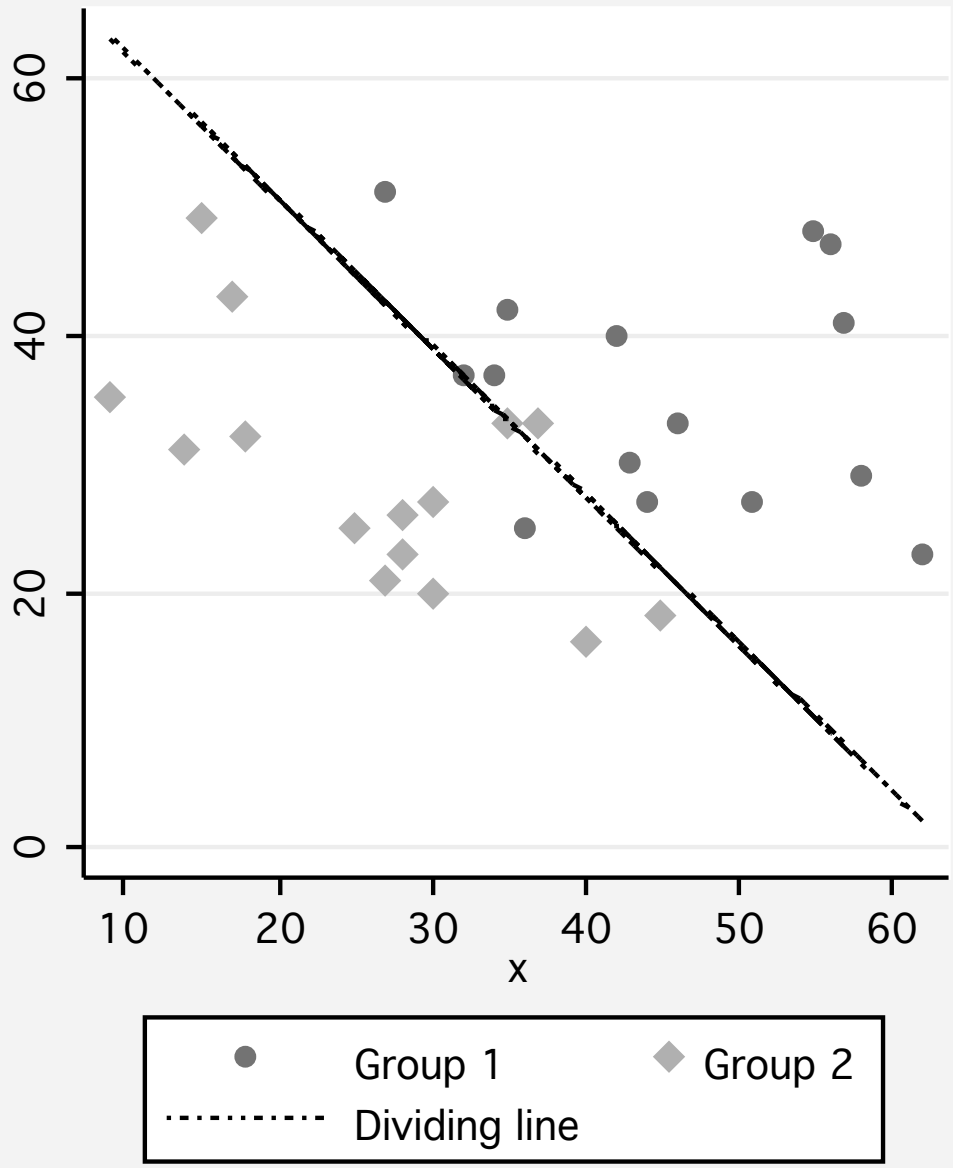
Linear discriminant analysis

Linear discriminant analysis (LDA) is based on seeking the linear combination of the discriminating variables that provides maximal separation between the groups. It is based on an eigensystem analysis of matrices formed from the between-group and within-group matrices of sums of squares and cross products. The first linear discriminant function is the eigenvector associated with the largest eigenvalue.

We illustrate with the dataset `twogroup` from the Stata website which contains 30 observations on $\{x, y\}$ pairs. We fit the LDA and retrieve the unstandardized coefficients, which may then be expressed in standard $y = mx + b$ form, as illustrated in the following figure.

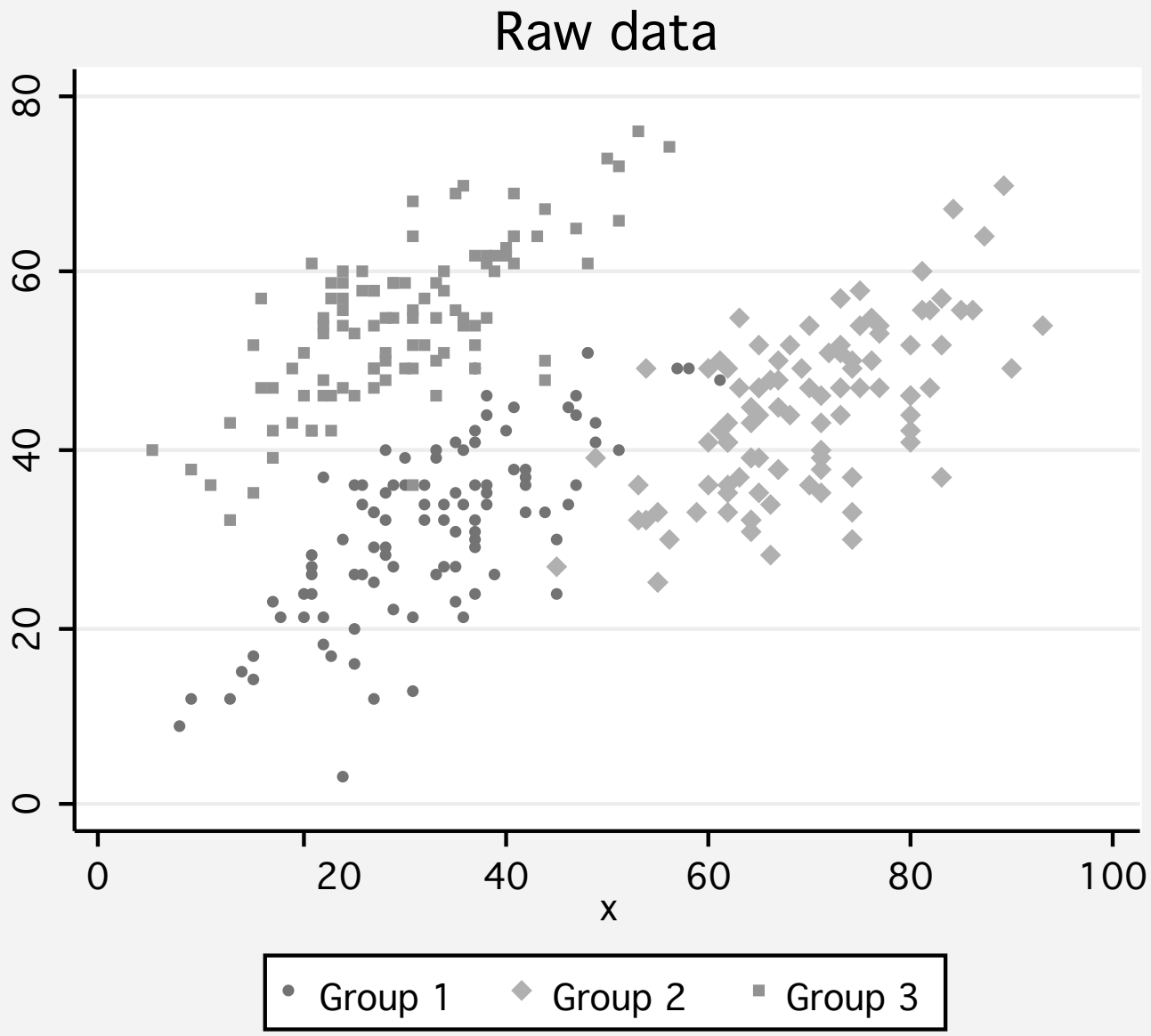
```
. discrim lda y x, group(group) notable
. estat loadings, unstandardized
Canonical discriminant function coefficients
```

	function1
y	.0862145
x	.0994392
_cons	-6.35128

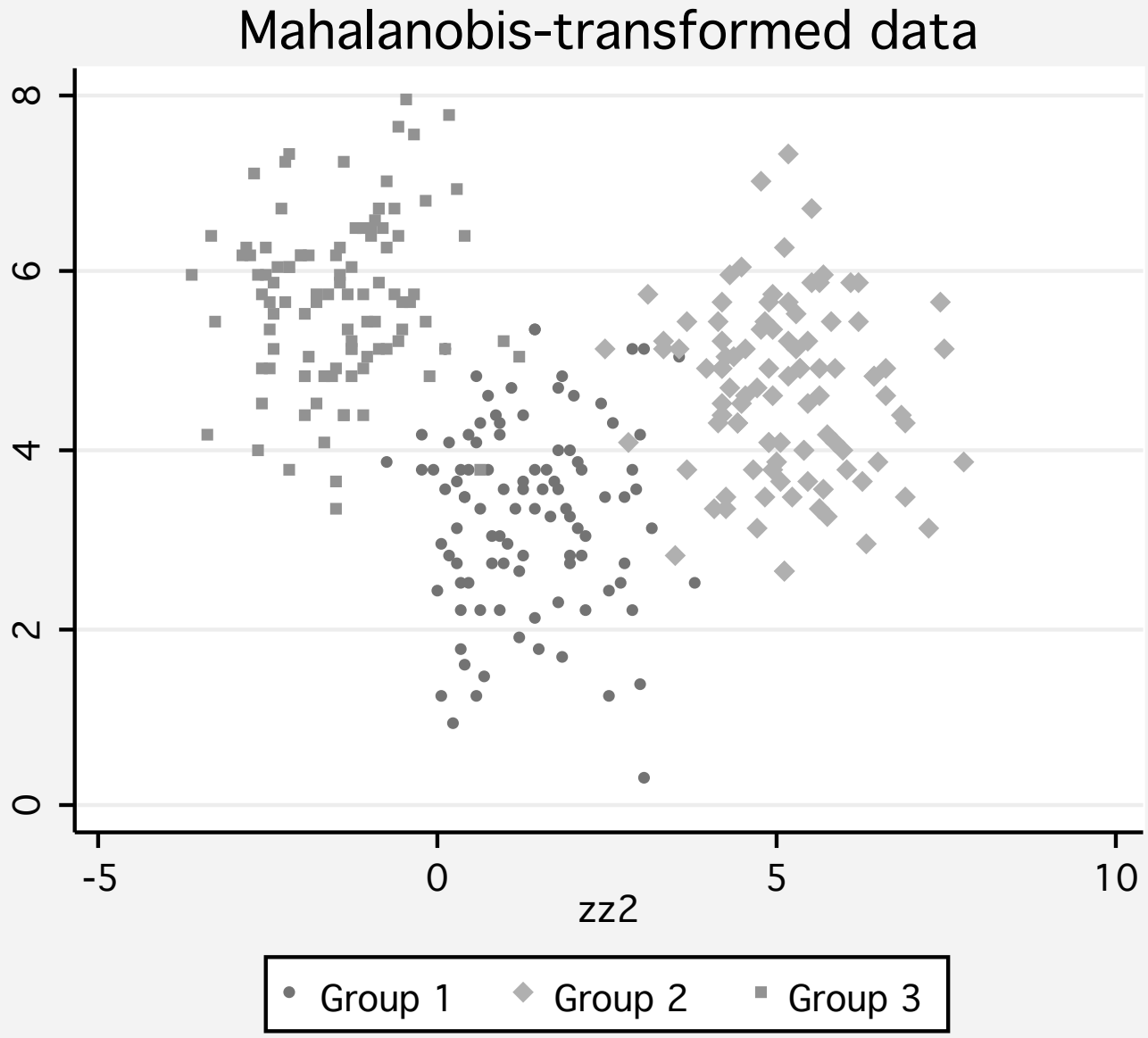

60
40
20
0
10
20
30
40
50
60
x
Group 1
Group 2
Dividing line

Another approach, *predictive LDA*, is based on the assumption that the observations are multivariate normal with equal covariance matrices, but different locations, or means, for different groups. LDA then uses the *Mahalanobis distance* for classification, grouping by observations by their smallest Mahalanobis distance from the group mean. This approach can be viewed as a transformation of the data, and then calculation of Euclidian distance measures. Group membership is based on Euclidian distance in the transformed space.

To illustrate, we use dataset `threegroup` from the Stata website. This dataset contains 300 $\{y, x\}$ pairs, 100 from each of three groups. A scatterplot of the raw data shows significant overlap between the groups' observations.


Raw data
0
20
40
60
80
0
20
40
60
80
100
x
Group 1
Group 2
Group 3

Predictive LDA transforms the data into Mahalanobis distances:



In the transformed space, the groups are more distinct.

```
. discrim lda y x, group(group)
Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True group	Classified 1	2	3	Total
1	93 93.00	4 4.00	3 3.00	100 100.00
2	3 3.00	97 97.00	0 0.00	100 100.00
3	3 3.00	0 0.00	97 97.00	100 100.00
Total	99 33.00	101 33.67	100 33.33	300 100.00
Priors	0.3333	0.3333	0.3333	

Classification was quite successful, with 93, 97 and 97 observations correctly classified into groups 1, 2 and 3, respectively. We could also examine the misclassified observations' characteristics with `estat list, varlist misclassifed` and use several options of the `predict` command to generate additional insight into the results of this predictive LDA analysis.

k^{th} nearest neighbor discriminant analysis

k^{th} nearest neighbor (KNN) discriminant analysis, unlike LDA or QDA, is a nonparametric technique that is based on the k nearest neighbors of each observation. We illustrate with a dataset, `head`, from the Stata website produced to study a possible link between American football helmet design and neck injuries. The three groups of 30 observations include high school football players, college football players, and nonfootball players. The discriminating variables we employ include `wdim`, head width; `circum`, head circumference; and `fbeye`, front-to-back measurement at eye level.

We first produce a LDA for these variables:

```
. discrim lda wdim circum fbeye, group(group)
Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True group	Classified high school	college	nonplayer	Total
high school	17 56.67	6 20.00	7 23.33	30 100.00
college	6 20.00	17 56.67	7 23.33	30 100.00
nonplayer	4 13.33	12 40.00	14 46.67	30 100.00
Total	27 30.00	35 38.89	28 31.11	90 100.00
Priors	0.3333	0.3333	0.3333	

We now produce a KNN analysis, using three nearest neighbors in the `k()` option:

```
. discrim knn wdim circum fbeye, group(group) k(3) mahalanobis
Kth-nearest-neighbor discriminant analysis
Resubstitution classification summary
```

Key
Number
Percent

	Classified				
True group	high school	college	nonplayer	Unclassified	Total
high school	17	4	3	6	30
	56.67	13.33	10.00	20.00	100.00
college	3	13	7	7	30
	10.00	43.33	23.33	23.33	100.00
nonplayer	4	5	19	2	30
	13.33	16.67	63.33	6.67	100.00
Total	24	22	29	15	90
	26.67	24.44	32.22	16.67	100.00
Priors	0.3333	0.3333	0.3333		

The results will be sensitive to the choice of k, the number of nearest neighbors to be considered. The “Unclassified" observations are those for which the method resulted in ties. The `ties()` option may be used to break ties by one of several methods. Using the `ties(nearest)` option results in all observations being classified.

```
. discrim knn wdim circum fbeye, group(group) k(3) mahalanobis ties(nearest)
Kth-nearest-neighbor discriminant analysis
Resubstitution classification summary
```

Key
Number
Percent

True group	Classified high school	college	nonplayer	Total
high school	23	4	3	30
	76.67	13.33	10.00	100.00
college	3	20	7	30
	10.00	66.67	23.33	100.00
nonplayer	4	5	21	30
	13.33	16.67	70.00	100.00
Total	30	29	31	90
	33.33	32.22	34.44	100.00
Priors	0.3333	0.3333	0.3333	

We see that the KNN classification with this option of handling tied scores is considerably more successful than LDA. LDA correctly classified 17, 17 and 14 of the observations in each 30-person group. The KNN analysis correctly classified 23, 20 and 21 observations.

This discussion of discriminant analysis only scratches the surface of Stata's capabilities in multivariate statistics. Other available techniques include correspondence analysis (`help ca`), cluster analysis (`help cluster`), factor analysis (`help factor`), multivariate analysis of variance (`help manova`), multiple classification analysis (`help mca`), multidimensional scaling (`help mds`), and principal component analysis (`help pca`).

Case study: Analyzing health status

- use and describe the mus18dataH.dta dataset:

```
use mus18dataH, clear
describe
```

- tabulate the health status variable:

```
tabulate hlthstat
```

- summarize the explanatory variables:

```
summarize hlthstat age linc ndisease num
```

- evaluate how (log) income differs across health status: `table hlthstat, contents(N linc mean linc p50 linc)`
  Are individuals from wealthier families more healthy?
- evaluate how age differs across health status:

```
table hlthstat, contents(N age mean age p50 age)
```

  Are younger individuals more healthy?

Case study: Analyzing health status

- analyze as a multinomial logit, using poor_fair as the base outcome:

  `mlogit hlthstat age linc ndisease num, nolog base(1)`
- perform Wald tests for each of the explanatory variables:

  `test age` (etc.)
- estimate the model in terms of proportional odds or relative risk ratios:

  `mlogit hlthstat age linc ndisease num, nolog base(1) rr`
- calculate marginal effects for each outcome:

  `margins, predict(pr outcome(1)) dydx(_all)` (etc.)

  Which of the explanatory factors have the greatest effect for each outcome?

Case study: Analyzing health status

- fit the model as an ordered logistic regression, taking the ordered nature of `hlthstat` into account:
  `ologit hlthstat age linc ndisease num, nolog`
- calculate marginal effects for each outcome from the `ologit`:
  `margins, predict(pr outcome(1)) dydx(_all)` (etc.)
  Which of the explanatory factors have the greatest effect for each outcome?
- Compare and contrast the marginal effects from the `mlogit` and `ologit` forms of the model. Which do you prefer? Why?