

Index of Lecture 3a

Page	Title
1	Introduction to logistic regression
2	Example dataset <code>mice</code>
3	Why not linear regression?
4	Logit transformation
5	Logistic regression model
6	Logistic regression for mice data
7	2×2 -table analysis
8	2×2 -table and logistic regression
9	Goodness-of-fit tests
10	Linear vs. logistic modelling
11	Computing predictions in linear models
12	Stata: the <code>margins</code> command
13	Simplest examples: 1 predictor
14	Examples with 2 categorical predictors
15	Examples with 2 predictors (cont.)
16	Predictions in multivariable models
17	Predictions for VER Example 14.16
18	Predictions in logistic regression
19	Scale-dependence of predictions with averaging/weighting

INTRODUCTION TO LOGISTIC REGRESSION

Logistic regression:

- binary (0/1, dichotomous) outcomes, possibly grouped to binomial outcomes (e.g., 3 positive out of 10 animals),
- first of several regression-type models not relying on normal distribution assumptions,
 - * sometimes called *generalised linear models* (glm's),
 - * model building from the predictors similar to linear regression,
 - * some common features in their analysis that distinguish them from linear regression analysis.

Today's session:

- review of simple logistic regression (one predictor) and its relation to known analyses, in particular 2×2 -table analysis,
- predictions, in particular using `margins` command (Stata),
- computer-assisted using Stata (good facilities available, superior to many simpler programs).

VER/MER textbooks:

- today: 16.1–5, 8 with some omissions (in next lecture).
- maybe also check Chapter 28 of VHM 801 textbook.

Homework for Tuesday: Last linear reg. exercise (VER 14.3).

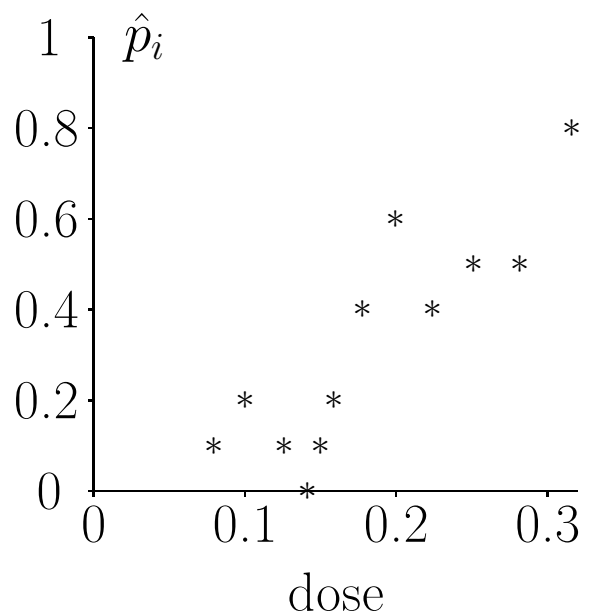
EXAMPLE DATASET MICE

Toxicity study of dose-response curve (Woodward 1941):

- lethality of different doses of chloracetic acid, measured as the mortality among 10 mice subjected to each dose,

group	dose	# mice died	# mice total	prop. died	
i	x_i	r_i	N_i	$\hat{p}_i = r_i/N_i$	$\text{logit}(\hat{p}_i)$
1	0.0794	1	10	0.1	-2.197
2	0.1000	2	10	0.2	-1.386
3	0.1259	1	10	0.1	-2.197
4	0.1413	0	10	0.0	undef.
5	0.1500	1	10	0.1	-2.197
6	0.1588	2	10	0.2	-1.386
7	0.1778	4	10	0.4	-0.405
8	0.1995	6	10	0.6	0.405
9	0.2239	4	10	0.4	-0.405
10	0.2512	5	10	0.5	0
11	0.2818	5	10	0.5	0
12	0.3162	8	10	0.8	1.386

- grouped binary data,
- statistical model:
 $r_i \sim \text{Bin}(N_i, p_i)$, and
 r_1, \dots, r_{12} independent,
- parameters: p_1, \dots, p_{12}
 (prob. of death in groups),
- question:
 how to use the doses?



WHY NOT LINEAR REGRESSION?

Regression for binary outcomes ($Y_i = 0$ or $Y_i = 1$)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i,$$

conflicts with the model assumptions:

- (1) errors ε_i are far from normally distributed (can only take two possible values¹),
- (2) with $p_i = P(Y_i = 1)$, we have $\text{Var}(Y_i) = p_i(1 - p_i)$, which is not constant when p_i is modelled by predictors,
- (3) both Y_i and p_i are bounded (do not go beyond 0 and 1) but linear predictions by the x -variables can easily give predictions outside the interval.

Regression for grouped binary outcomes (proportions r_i/N_i)

- same problems (1)–(3), although (1) is less severe,
- transformation is a possibility, usually with *variance-stabilising* transformation (slide L1a–12):

$$Y_i = \arcsin(\sqrt{r_i/N_i}), \quad \arcsin = \text{inverse sine function},$$

— however, not recommended unless

- * the denominators N_i are all “large” and approximately the same,
 - * the prop.’s r_i/N_i are not too extreme (close to 0 or 1),
- and usually offers no advantages over logistic regression.

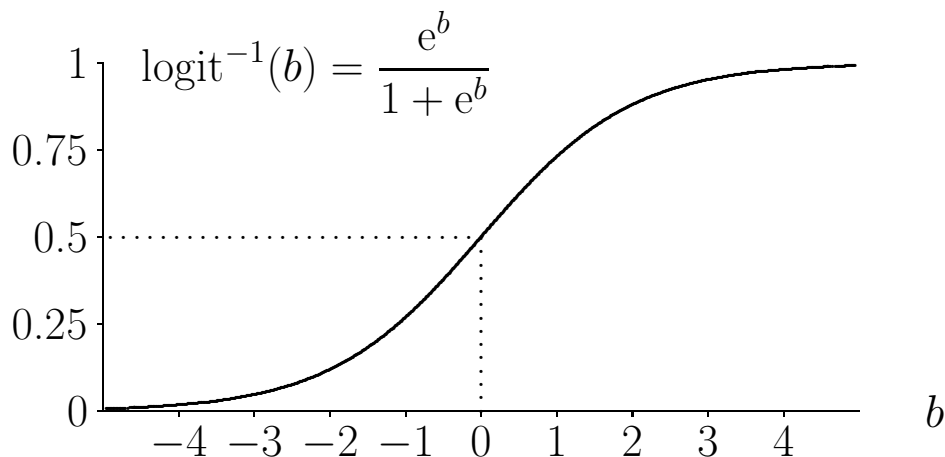
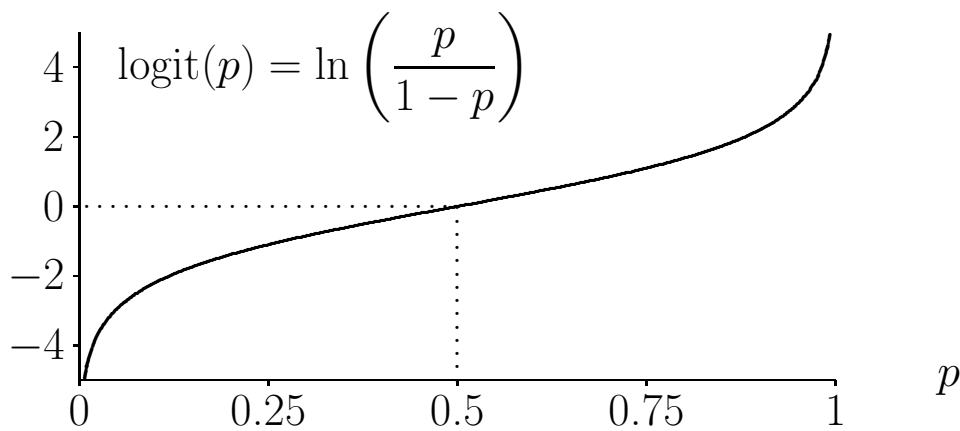
¹ Possible values for the error of obs. i are $\varepsilon_i = 1 - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$, and $\varepsilon_i = -(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$.

LOGIT TRANSFORMATION

Define for $0 < p < 1$ and any b ,

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right) \quad \text{and} \quad \text{logit}^{-1}(b) = \frac{e^b}{1+e^b} = \frac{1}{1+e^{-b}},$$

- the logit function stretches the interval $(0,1)$, excl. endpoints!, onto the entire real axis (from $-\infty$ to ∞),
- $\text{logit}(\frac{1}{2})=0$, and $\text{logit}(p)$ is increasing in p ,
- with $\text{odds}(p) = \frac{p}{1-p}$, we have $\text{logit}(p) = \ln(\text{odds}(p))$,
- logit and inverse logit functions:



LOGISTIC REGRESSION MODEL

= a different transformation approach:

- keep observations (binary/grouped binary) untransformed,
- transform probability parameter p by logit function to logit scale where linear modelling takes place, e.g.

$$\ln \left(\frac{p_i}{1 - p_i} \right) = \text{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad (1)$$

where

$$p_i = P(Y_i = 1),$$

$$Y_i = \begin{cases} 1 & \text{“success”} \\ 0 & \text{“failure”} \end{cases} \quad \text{for } i = 1, \dots, n,$$

$$x_i = \text{predictor variable for observation } i.$$

Model assumptions:

- independence of all the observations (Y_i 's),
- linearity of relation (1) on logit scale.²

Grouped binary data (with N_i repl. in i th group)

- equation (1) for $p_i = \text{prob. of “success” in group } i$,
- same model as if set up as binary data (with $n = \sum_i N_i$).

Multiple logistic regression model for predictors x_1, \dots, x_k :

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}.$$

² For a single predictor model, the linearity assumption applies only to the case where x_1 is continuous.

LOGISTIC REGRESSION FOR MICE DATA

Estimates (with SE):

$$\hat{\beta}_0 = -3.57 (0.71),$$

$$\hat{\beta}_1 = 14.64 (3.33).$$

Test of $H_0: \beta_1 = 0$:

$$z = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1) = 4.39$$

very sign. in $N(0,1)$

\Rightarrow strong effect of dose.

Estimated line:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{dose}$$

– on logit-scale.

Estimated curve $\hat{p}(x)$:

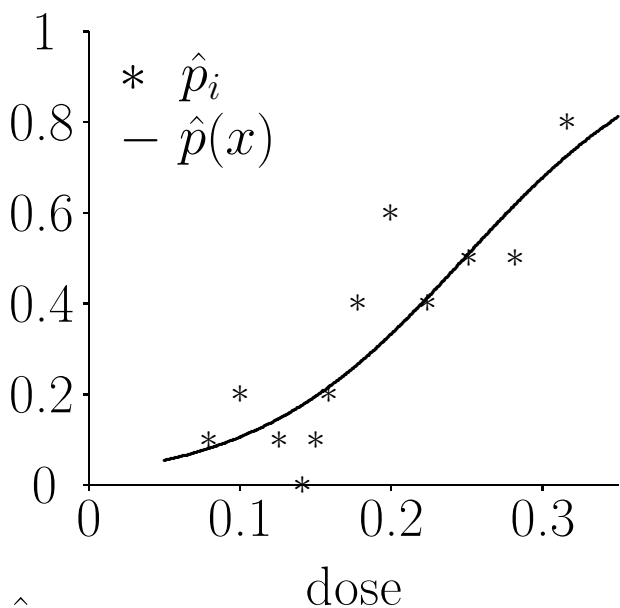
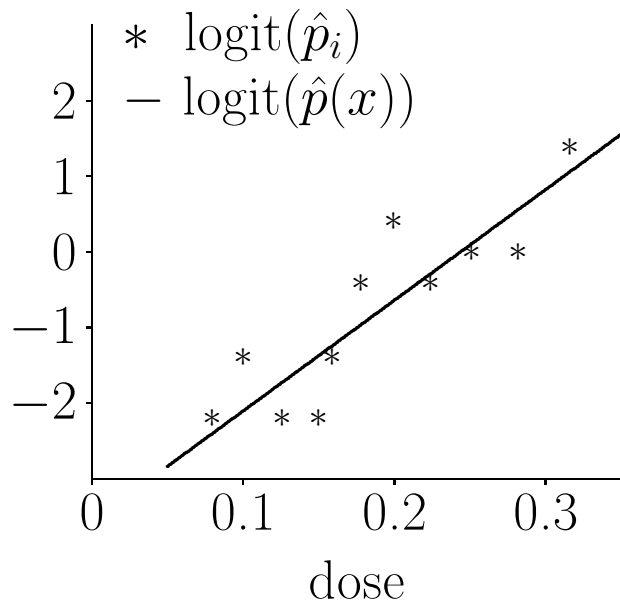
$$\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{dose})$$

– on probability-scale.

Interpretation of $\hat{\beta}_1$:

change in dose of a units \Rightarrow

- change in $\text{logit}(p)$ of $a\hat{\beta}_1$ units,
- change in $\text{odds}(p)$ by factor $\exp(a\hat{\beta}_1)$ (= *odds-ratio*), where $\text{odds}(p) = p/(1-p)$.



Test of model: (Pearson goodness-of-fit test)

$$X^2 = 8.74, P = 0.56 \Rightarrow \text{no lack of fit (model ok).}$$

2×2 -TABLE ANALYSIS

Mice data: outcome = mortality, explanatory = dichotomous version of dose (for illustration only):

	dose > 0.16		
dead	1	0	Total
1	32	7	39
0	28	53	81
Total	60	60	120

Statistical model (with binary dose predictor):

two binomial distributions $\text{Bin}(60, p_1)$ and $\text{Bin}(60, p_0)$,

– analyzed in VHM 801 by computing:

$$\begin{aligned} \hat{p}_1 &= 32/60 = 0.533, & \text{SE}(\hat{p}_1) &= \sqrt{\hat{p}_1(1-\hat{p}_1)/60} = 0.0644, \\ \hat{p}_0 &= 7/60 = 0.117, & \text{SE}(\hat{p}_0) &= \sqrt{\hat{p}_0(1-\hat{p}_0)/60} = 0.0414, \\ \hat{p}_1 - \hat{p}_0 &= 0.533 - 0.117 = 0.417, & \text{SE} &= \sqrt{0.0644^2 + 0.0414^2} = 0.077. \end{aligned}$$

Alternative ways of comparing the probabilities \hat{p}_1 and \hat{p}_0 :

relative risk : $\text{RR} = \hat{p}_1/\hat{p}_0 = 0.533/0.117 = 4.57,$

$$\begin{aligned} \text{odds-ratio : OR} &= \text{odds}(\hat{p}_1)/\text{odds}(\hat{p}_0) = [\hat{p}_1/(1-\hat{p}_1)] / [\hat{p}_0/(1-\hat{p}_0)] \\ &= [0.533/(1-0.533)] / [0.117/(1-0.117)] \\ &= 1.143/0.132 = 8.653 = (32 \cdot 53)/(28 \cdot 7). \end{aligned}$$

Advantages of these statistics (over the simple $\hat{p}_1 - \hat{p}_0$):

- multiplicative effects are more meaningful than additive effects for proportions bounded by 0 and 1,
- when both probabilities “close” to zero: OR \approx RR (clearly not the case in the example),
- both statistics more useful than $(\hat{p}_1 - \hat{p}_0)$ when multiple factors studied simultaneously.

2 × 2–TABLE AND LOGISTIC REGRESSION

Mice data: (same as on previous slide):

outcome = mortality, predictor = **dose2** (dichotomous version of dose):

	dose2 = (dose > 0.16)		
dead	1	0	Total
1	32	7	39
0	28	53	81
Total	60	60	120

Logistic regression model with **dose2** as predictor:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{dose2}_i,$$

gives the estimates

$$\hat{\beta}_1 = 2.158 = \ln(8.653) = \ln(\text{OR}),$$

$$\hat{\beta}_0 = -2.024 = \text{logit}(0.117) = \text{logit}(\hat{p}_0).$$

Interpretations (valid also in multiple logistic regression):

- odds-ratio for effect of **dose2** = $e^{\hat{\beta}_1} = e^{2.158} = 8.653$,
- baseline prob. = $\text{logit}^{-1}(\hat{\beta}_0) = \text{logit}^{-1}(-2.024) = 0.117$.

Summary:

The 2 × 2-table analysis and the logistic regression analysis are equivalent (also, the *P*-values are similar³).

³ The likelihood-ratio tests (next lecture) of the two models are identical.

GOODNESS-OF-FIT TESTS

New feature compared to normal distribution models:

- formal statistic test for whether agreement between observed and expected values is “acceptable”,
 - * in the sense of corresponding to the distribution assumed (e.g. binomial), because the variance of that distribution determines bounds for “acceptable” differences between observed and expected,⁴
 - * note: this does not work for a normal distribution because of its separate dispersion parameter (σ),
- different tests exist (discussed later in course), but the most intuitive one is the Pearson goodness-of-fit test.

For grouped binary (binomial) data, the Pearson test⁵ takes the form,

$$X^2 = \sum_i \frac{(Y_i - N_i \hat{p}_i)^2}{N_i \hat{p}_i (1 - \hat{p}_i)} \approx \chi^2(I - 1 - k),$$

where $\hat{p}_i = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki})$ is the predicted probability for the i th group, and $i = 1, \dots, I \sim$ groups. The approximate reference chi-square distribution is valid⁶ under the assumed logistic regression model (\sim null hypothesis), and large values of X^2 indicate departures from the model.

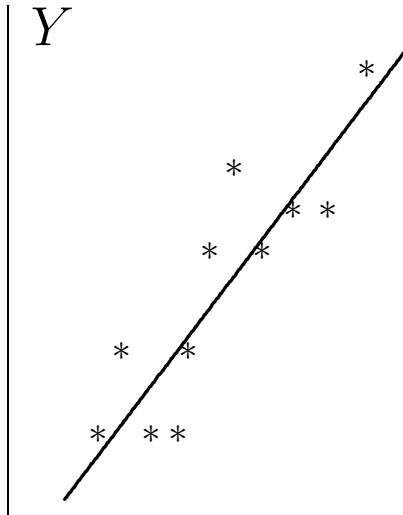
⁴ E.g., in a binomial distribution (N, p) the standard deviation is $\sqrt{Np(1-p)}$.

⁵ The test coincides with the Pearson chi-square test in two-way tables.

⁶ The approximation may be poor if the expected counts $N_i \hat{p}_i$ are too low.

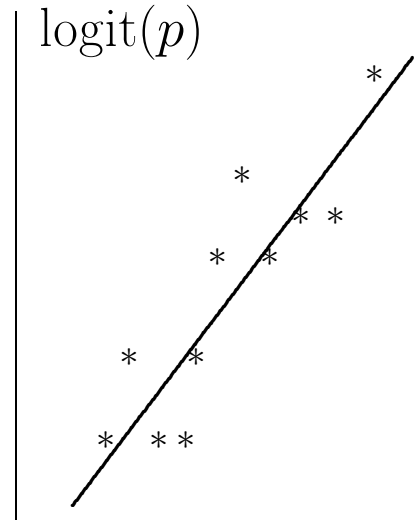
LINEAR VS. LOGISTIC MODELLING

Linear regression



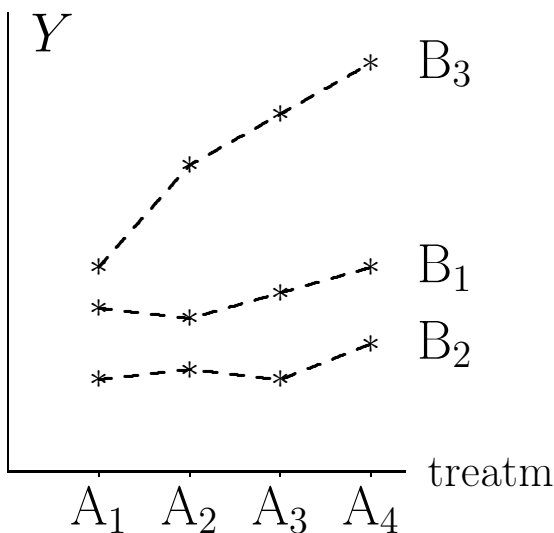
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
 where ε_i 's $\sim N(0, \sigma^2)$

Logistic regression



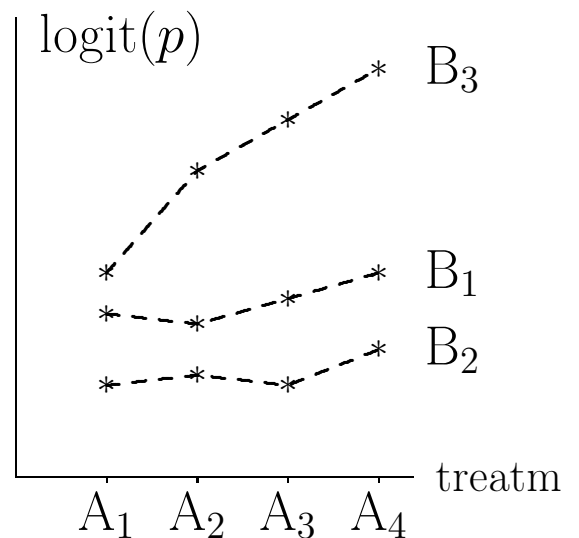
Model: $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$
 where $p_i = P(Y_i = 1)$

Factorial design



Model:
 $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
 where ε_{ij} 's $\sim N(0, \sigma^2)$

Logistic factorial design



Model:
 $\text{logit}(p_{ij}) = \mu + \alpha_i + \beta_j$
 where $p_{ij} = P(Y_{ij} = 1)$

COMPUTING PREDICTIONS IN LINEAR MODELS

We distinguish between two types of predictions (or purposes):

- i*) for individual observations: “real” prediction,
- ii*) for purposely selected combinations of predictor values: “illustrative” prediction.

All software packages for linear models offer predictions of type *i*), directly for observed predictor patterns and for new predictor patterns:

- Stata/SAS: add extra observations to data with missing outcome,
- Minitab/R: specify new observations in separate columns/dataset,
- Stata/SAS: special commands (`lincom/estimate`) give estimates for linear combinations of regression coefficients.

Fully specified predictions of type *ii*): may be done using methods for type *i*) (perhaps tedious).

Some software offer both fully and partially specified predictions of type *ii*):

- Stata: the `margins` command,
- Minitab/SAS: least squares means⁷; i.e., all predictors not included in prediction are set at their average value.

⁷ “Least squares means” originate from experimental designed studies/data where factors are often balanced by design.

STATA: THE MARGINS COMMAND

- very flexible command (from version 12) with a wide range of options and setups; so flexible that caution is needed to not use it wrongly. . . .
- strongly recommended to always check your predictions with simpler methods (in a few examples),
- linkage to the `marginsplot` command allows easy plotting of predicted values; in particular, this is the easiest way to generate an interaction plot,
- mainly intended for “illustrative” predictions, and uses predictions from the `predict` command behind the scenes to come up with the requested predictions,
- the online help is pretty confusing \Rightarrow recommended to work from well-established examples, and to avoid use of numerous extra “fancy” options.

Coverage in course: worked examples (from simple to more complex) illustrating the basic features of command:

- 1-predictor settings (categorical and continuous),
- 2-predictor settings, and the questions arising from omitting a predictor from a prediction,
- VER 14.16 worked example,
- plots and transformations as needed.

SIMPLEST EXAMPLES: 1 PREDICTOR

(1a) categorical: simple means, with model-based SE,

```
regress wpc i.herd
margins herd
lincom _cons+2.herd
```

(1b) continuous: predictions at specified set of values, with subsequent plot by `marginsplot`:

```
regress wpc milk120
margins , at( milk120=(1200(1000)5600) )
marginsplot
lincom _cons+milk120*3200
```

note: flexible format of “atspec”; e.g., `1200(1000)5600` = 1200, 2200, ..., 5200, but list can also include statistics (e.g., mean and percentiles) and the special names “asobserved” and “asbalanced”,

(1c) continuous with quadratic effect: predictions as above, but need to use factor notation,

```
regress wpc c.milk120##c.milk120
margins , at( milk120=(1200(1000)5600) )
marginsplot
```

(1d) backtransformation from transformed scale: can be specified by formula, but note CI problems,⁸

```
regress lnwpc milk120
margins , at( milk120=(1200(1000)5600))
margins , at( milk120=(1200(1000)5600)) expression(exp(predict(xb)))
marginsplot
```

⁸ The correct CIs are obtained by backtransformation, but the method used by `margins` command is based on an approximate SE on original scale.

EXAMPLES WITH 2 CATEGORICAL PREDICTORS

(2a) additive model: predictions require decision about how to weight contributions from other predictor, e.g.:

- * equally/balanced (standard in least squares means),
- * total data weights (default choice),
- * choices corresponding to specific prediction settings,

```
regress wpc i.rp i.vag_disch
margins rp vag_disch
table rp vag_disch, row col
lincom _cons+1.rp+1.vag_disch*82/1574 /* rp=1 */
margins rp vag_disch, asbalanced
lincom _cons+1.rp+1.vag_disch*0.5 /* rp=1 */
margins rp, over(vag_disch) /* same as: at(vag_disch=(0 1)) */
lincom _cons+1.rp+1.vag_disch /* rp=1, vag_disch=1 */
```

(2b) model with interaction: combined effect \sim simple means, separate effects require decision about how to weight contributions from other predictor (as above for additive model),

```
regress wpc rp##vag_disch
margins rp#vag_disch
marginsplot, noci /* this is the interaction plot! */
marginsplot, noci x(vag_disch) /* x() to control variable on x */
marginsplot, noci x(rp) /* same as default */
margins rp
lincom _cons+1.rp+(1.vag_disch+1.rp#1.vag_disch)*82/1574 /* rp=1 */
margins rp, asbalanced
lincom _cons+1.rp+(1.vag_disch+1.rp#1.vag_disch)*0.5 /* rp=1 */
```

EXAMPLES WITH 2 PREDICTORS (CONT.)

Categorical + continuous predictor:

(2c) similar to single continuous predictor, with multiple groups (intercepts and lines),

```
regress wpc i.dyst milk120
margins dyst, at( milk120=(1200 2200 3200 4300 5500))
marginsplot, noci
lincom _cons+1.dyst+milk120*3200 /* dyst=1, milk120=3200 */
margin dyst, atmeans
lincom _cons+1.dyst+milk120*3215.096
* interaction model
regress wpc dyst##c.milk120
margins dyst, at( milk120=(1200 2200 3200 4300 5500))
marginsplot, noci /* this is the interaction plot! */
lincom _cons+1.dyst+(c.milk120+1.dyst#c.milk120)*3200
/* dyst=1, milk120=3200 */
```

Two continuous predictors:

(2d) need values (possibly lists) for both predictors \Rightarrow predictions usually fully specified (no averaging/weighting),

```
regress wpc parity milk120
margins , at( parity=(1(1)6) milk120=(1200 2200 3200 4300 5500))
marginsplot, noci
margins , at( milk120=(1200 2200 3200 4300 5500) parity=(1(1)6) )
marginsplot, noci /* changing roles in plot */
margins , at( milk120=(1200 2200 3200 4300 5500) (median)parity)
marginsplot
lincom _cons+milk120*1200+parity*2 /* milk120=1200, parity=2 */
margins, atmeans
lincom _cons+milk120*3215.096+parity*2.73628 /* both at means */
```

PREDICTION IN MULTIVARIABLE MODELS

Main challenge/thing to remember: predictions need values or weights for all predictor terms in model

⇒ no software can do this automatically (so that it always makes sense)!

Some issues to consider when setting up predictions:

- the purpose (e.g., “real” versus “illustrative”),
- should the prediction correspond to an average or a real situation? (e.g., when using weights for categorical predictors, the predictions will not correspond to real situations),⁹
- is the predictor distribution in the observed data representative for the population or the targeted setting?⁹
- are the predictor distributions independent enough to set the values for different predictors independently?⁹
- for categorical predictors, are predictions intended to facilitate pairwise comparisons beyond comparisons with baseline? (perhaps the main motivation of least squares means),
- if modelling is carried out on transformed scale, should any weighting take place on transformed or original scale? (as they will lead to different results).

⁹ Using margins with its default settings implies that your answer to this question is “yes”.

PREDICTIONS FOR VER EXAMPLE 14.16

Model summary:

- outcome: wpc, on square-root transformed scale,
- categorical predictors: aut_calv, twin, dyst, rp##vag_disch,
- continuous predictors: parity, herd_size with quadratic term.

Some possible prediction aims:

- 1) illustrate combined effect of diseases (rp,dyst,vag_disch) on wpc,
- 2) illustrate interaction rp#vag_disch (effectively included under 1),
- 3) illustrate effect of herd_size on wpc.

1): Prediction/Estimates for combinations of disease, with backtransformed (squared) means \sim median wpc-values:

Estimates*		\sqrt{wpc} (mean)		wpc (median)	
rp	vag_d	dyst=0	dyst=1	dyst=0	dyst=1
0	0	7.517	8.059	56.50	64.95
0	1	7.503	8.046	56.30	64.73
1	0	7.906	8.448	62.51	71.38
1	1	9.384	9.926	88.06	98.53

* at: parity=1, twin=0, herd_size=251, aut_calv=0
 (~ the mean herd size, and the most frequent categories)

3): Prediction/Estimates for the 7 observed herd sizes:

Estimates*	herd sizes						
scale	125	185	201	235	263	294	333
\sqrt{wpc}	7.092	7.013	7.079	7.448	7.674	8.177	9.002
wpc	50.29	49.19	50.11	53.85	58.90	66.86	81.03

* at: parity=1, twin=0, aut_calv=0, all diseases=0

PREDICTIONS IN LOGISTIC REGRESSION

Predictions/presentation of effects on probability scale:

- easier to understand probabilities than OR's,
- not additive \Rightarrow more complicated (care is needed).

Illustration for Nocardia data and effect of `dcpxt`:

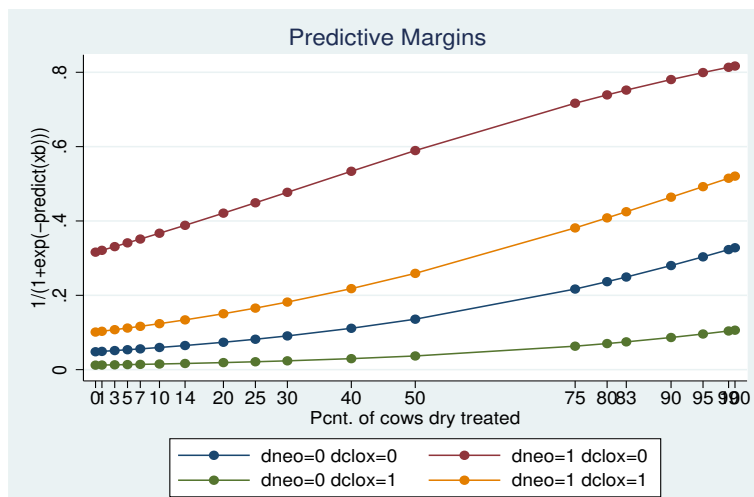
$$\text{logit}(\hat{p}) = -2.984 + 0.023 \text{ dcpxt} + 2.212 \text{ dneo} - 1.412 \text{ dclox},$$

- OR for 1% “change” in `dcpxt` = $e^{0.023} = 1.023$,
- OR for 10% “change” in `dcpxt` = $e^{0.023 \cdot 10} = e^{0.23} = 1.25$.

or on probability scale¹⁰

$$\hat{p} = \text{logit}^{-1}(\text{logit}(\hat{p})) = 1/(1 + e^{-\text{logit}(\hat{p})})$$

can be plotted against `dcpxt` values in a suitable range, for *fixed* values of all other predictors (here `dneo` and `dclox`):



note: non-linear relation!
non-parallel curves!

How to compute predictions — same options as for linear models:

- predicted values for actual/new observations,
- Stata: `margins` and `marginsplot` commands.¹¹

¹⁰ For demonstration purposes only; in a case-control design predicted probabilities *do not make sense* because the proportion of cases and controls is controlled.

¹¹ See next page for discussion of averaging or weighting.

SCALE-DEPENDENCE OF PREDICTIONS

WITH AVERAGING/WEIGHTING

Main message: in models involving transformations (e.g. logistic regression), any averaging of predictions involves a choice of scale:

- different results (even after transformation to same scale),
- different interpretations.

Consider again the logistic “predictive” equation

$$\text{logit}(\hat{p}) = -2.984 + 0.023 \text{ dcpct} + 2.212 \text{ dneo} - 1.412 \text{ dclox},$$

— Table of predictions for dcpct = 0:

dneo	dclox	$\text{logit}(\hat{p})$	$\hat{p} = \text{logit}^{-1}(\hat{p})$
0	0	-2.984	0.0481
0	1	-4.397	0.0123
1	0	-0.772	0.316
1	1	-2.184	0.101
weighted*		-1.821	0.198
(transformed)		0.139	-1.399

* based on data counts for the 4 categories of (dneo,dclox): 22, 12, 59, 15

Interpretations:

- \hat{p} (averaged): \sim averaged probabilities, might correspond to a population (if data representative),
- $\text{logit}(\hat{p})$ (averaged and then transformed): simpler, because backtransformation of a “natural” value.