

## Index of Lecture 3b: Introduction to logistic regression & Prediction for presentation

Page	Title
1	Practical information
2	Example dataset mice
3	Why not linear regression?
4	Logit transformation
5	Logistic regression model
6	Logistic regression for mice data
7	$2 \times 2$ -table analysis
8	$2 \times 2$ -table and logistic regression
9	Goodness-of-fit tests
10	Linear versus logistic modelling
11	Computing predictions in linear models
12	Stata: the margins 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

## PRACTICAL INFORMATION

### 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 without normal distribution assumptions,
  - \* sometimes called **generalised linear models** (glm's),
  - \* model building from the predictors is similar to linear regression,
  - \* but some common features in their analysis that distinguish them from linear regression analysis.

### This lecture:

- possibly brief review/discussion of VER Exercise 14.3,
- review of simple logistic regression (one predictor) and its relation to known analyses, in particular  $2 \times 2$ -table analysis; **using Stata** to illustrate,
- more on predictions, in particular using margins command (**Stata**).

### VER/MER textbooks:

- **today**: 16.1–5, 8 with some omissions (included in next lecture),
- maybe also check PSLS Chapter 28 (the VHM 801 textbook; at Moodle).

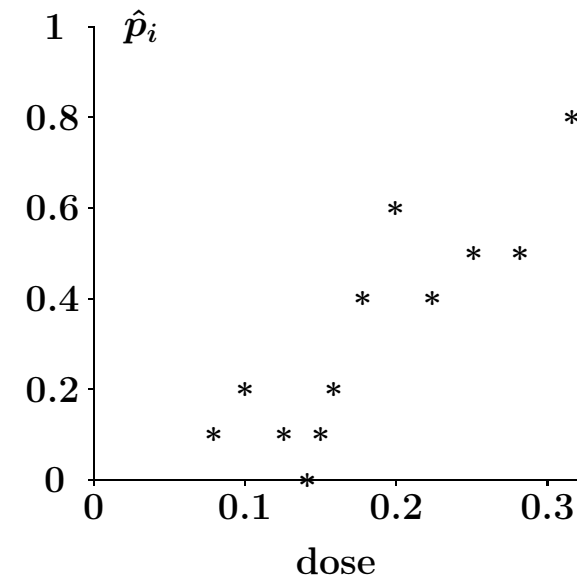
**Preparation for Friday:** the Model-building exercise (VER 15).

## EXAMPLE DATASET MICE

**Toxicity study** of dose-response curve (Woodward 1941, from RC textbook):

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

group $i$	dose $x_i$	# died $r_i$	# total $N_i$	prop. died $\hat{p}_i = r_i/N_i$	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**:  $p_1, \dots, p_{12}$  (probability of death in dose 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) the errors  $\varepsilon_i$  are far from normally distributed (can only take two possible values<sup>1</sup>),
- (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 (0,1).

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 the **variance-stabilising** transformation:  $Y_i = \arcsin(\sqrt{r_i/N_i})$  (slide 1aL–12) — however, **not recommended** unless
  - \* the denominators  $N_i$  are all “large” and approximately the same,
  - \* the proportions  $r_i/N_i$  are not too extreme (close to 0 or 1),and **usually offers no advantages** over logistic regression.

---

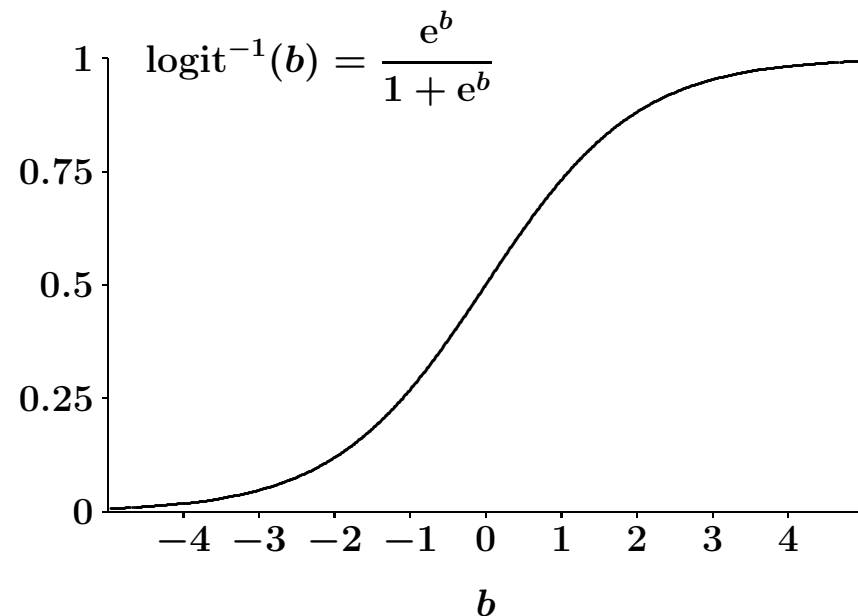
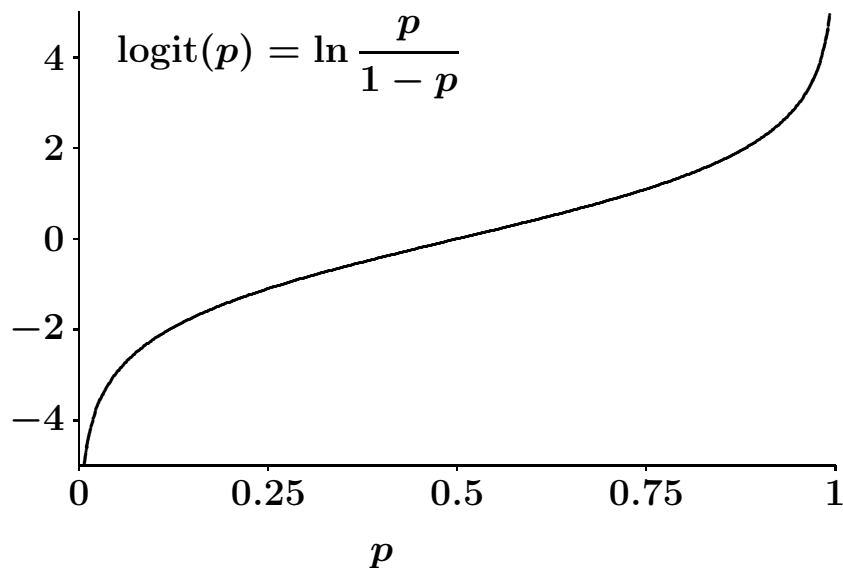
<sup>1</sup> Possible values for the error of obs.  $i$  are:  $\varepsilon_i = 1 - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$ , and  $\varepsilon_i = -(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$ .

## 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  $(0,1)$ , **excluding endpoints!**, onto the entire real axis (from  $-\infty$  to  $\infty$ ),
- $\text{logit}(\frac{1}{2}) = 0$ , 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.

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

where  $Y_i = \begin{cases} 1 & \text{“success”} \\ 0 & \text{“failure”} \end{cases}$ ,  $p_i = P(Y_i = 1)$ , and  $x_i =$  predictor value for obs.  $i$ .

**Model assumptions:**

- **independence** of all the observations ( $Y_i$ 's),
- **linearity** in modelling equation (above) on logit scale.<sup>2</sup>

**Grouped binary data** (with  $N_i$  replicates in  $i$ th group):

- same modelling equation for  $p_i =$  probability 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  $x_1, \dots, x_k$ :

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

---

<sup>2</sup> 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):

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

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

**Estimated line:**

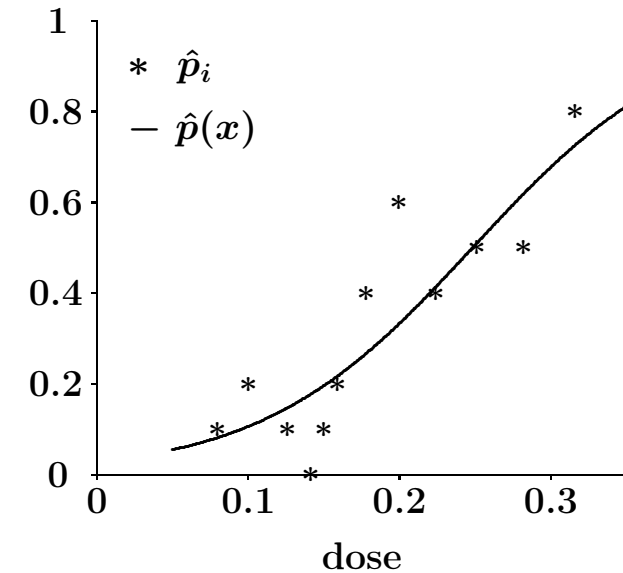
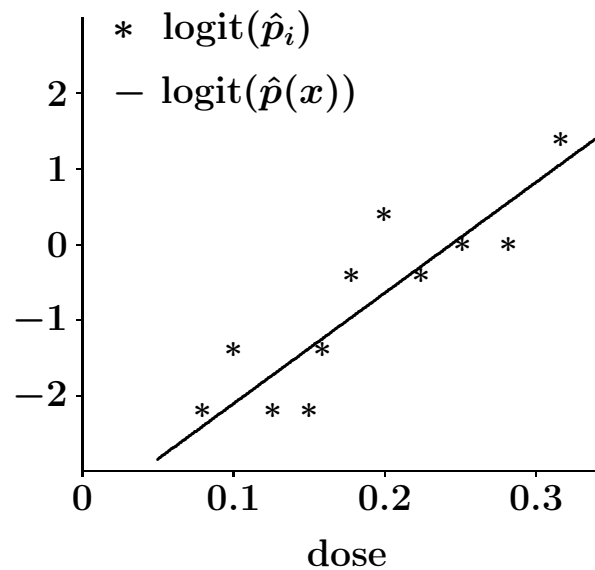
$$\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{dose}$$

– on logit-scale.

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

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

– on probability-scale.



**Test** of  $H_0: \beta_1 = 0$ :

$$z = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1) = 4.39 \sim \text{very significant in } N(0, 1) \Rightarrow \text{strong effect of dose.}$$

**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):

	dose2 = (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:

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

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

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

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

**Advantages** of the ratio 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 are “close” to zero (clearly not the case here): **OR  $\approx$  RR**,
- both statistics more useful than  $(\hat{p}_1 - \hat{p}_0)$  when multiple factors studied together.

## 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

$$\begin{aligned}\hat{\beta}_1 &= 2.158 = \ln(8.653) = \ln(\text{OR}), \\ \hat{\beta}_0 &= -2.024 = \text{logit}(0.117) = \text{logit}(\hat{p}_0).\end{aligned}$$

**Interpretations** (valid also in multiple logistic regression):

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

**Summary:** the 2 × 2-table analysis and the logistic regression analysis are equivalent (also, the *P*-values are similar<sup>3</sup>).

<sup>3</sup> The likelihood-ratio tests (next lecture) for the effect of dose2 are identical in the two models.

## GOODNESS-OF-FIT TESTS

New feature compared to normal distribution models<sup>4</sup>:

- o formal statistic test for whether the 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,<sup>5</sup>
- o 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 test 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 valid<sup>6</sup> under the assumed logistic regression model ( $\sim$  null hypothesis), and large values of  $X^2$  indicate departures from the model.

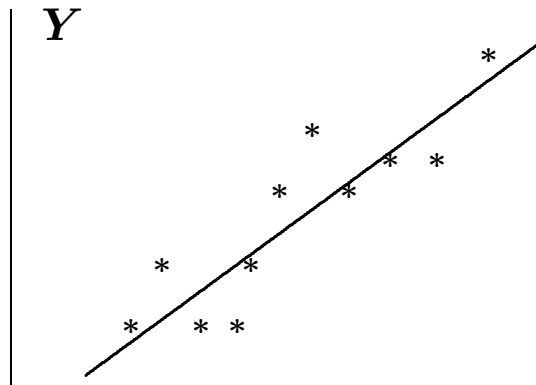
<sup>4</sup> Such tests do not work for a normal distribution because of its separate dispersion parameter ( $\sigma$ ).

<sup>5</sup> For example, in a binomial distribution  $(N, p)$  the standard deviation equals  $\sqrt{Np(1-p)}$ .

<sup>6</sup> The approximation may be poor if the denominators  $N_i \hat{p}_i (1 - \hat{p}_i)$  are too low.

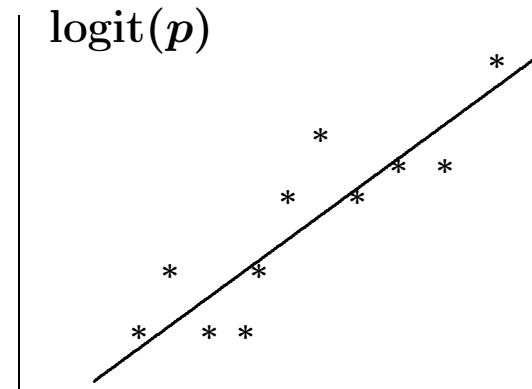
# LINEAR VERSUS LOGISTIC MODELLING

## Linear regression



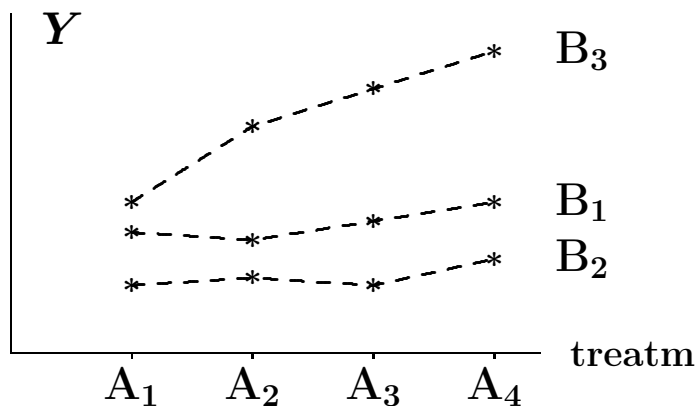
**Model:**  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$   
 where  $\varepsilon_i$ 's  $\sim N(0, \sigma^2)$

## Logistic regression



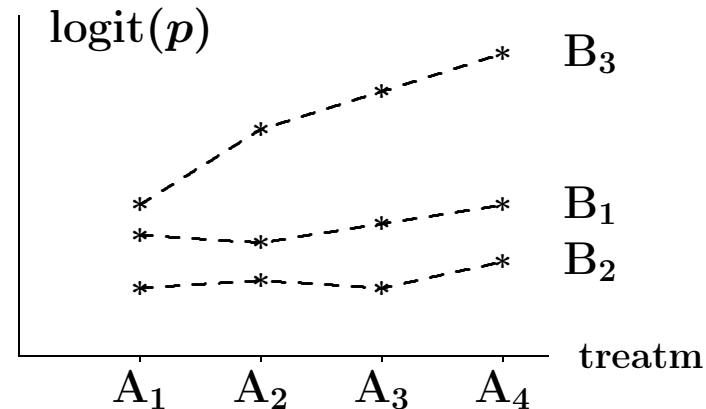
**Model:**  $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$   
 where  $p_i = P(Y_i = 1)$

## Factorial design



**Model:**  $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$   
 where  $\varepsilon_{ij}$ '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 also for **new predictor patterns**:

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

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

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

- o Stata: the `margins` command (R: `emmeans` library),
- o Minitab/SAS: **least squares means**<sup>7</sup>; i.e., all predictors not included in the prediction are set at their average value or as equally distributed across categories.

---

<sup>7</sup> The term “least squares means” originates 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) — if you don't know what is computed, how can others know...
- 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=(1400(1000)5400) )
marginsplot
lincom _cons+milk120*3400
```

**note**: flexible format of “atspec”; e.g., 1400(1000)5400 = 1400, 2400, . . . , 5400, 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 for quadratic term:

```
regress wpc c.milk120##c.milk120
margins , at( milk120=(1400(1000)5400) )
marginsplot
```

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

```
regress lnwpc milk120
margins , at( milk120=(1400(1000)5400))
margins , at( milk120=(1400(1000)5400)) expression(exp(predict(xb)))
marginsplot
```

---

<sup>8</sup> The correct CIs are obtained by backtransformation, but the method used by the margins command is based on an approximate SE (from the so-called “delta method”) on original scale.

## EXAMPLES WITH 2 CATEGORICAL PREDICTORS

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

- \* equally/balanced ( $\sim$  standard of least squares means),
- \* total data weights (default choice of margins),
- \* choices corresponding to specific prediction settings,

```
regress wpc i.rp i.vag_disch
margins rp vag_disch
table rp vag_disch
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 and the **interaction plot**; separate effects require decision about how to weight contributions from the other predictor (as above for the 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 (CONTINUED)

### Categorical + continuous predictor:

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

```
regress wpc i.dyst milk120
margins dyst, at( milk120=(1400(1000)5400))
marginsplot, noci
lincom _cons+1.dyst+milk120*3400 /* dyst=1, milk120=3400 */
margin dyst, atmeans
lincom _cons+1.dyst+milk120*3215.096
* interaction model
regress wpc dyst##c.milk120
margins dyst, at( milk120=(1400(1000)5400))
marginsplot, noci /* this is the interaction plot! */
lincom _cons+1.dyst+(c.milk120+1.dyst#c.milk120)*3400 /* dyst=1, milk120=3400 */
```

### 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=(1400(1000)5400))
marginsplot, noci
margins , at( milk120=(1400(1000)5400) parity=(1(1)6) )
marginsplot, noci /* changing roles in plot */
margins , at( milk120=(1400(1000)5400) (median)parity)
marginsplot
lincom _cons+milk120*1400+parity*2 /* milk120=1400, 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  $\Rightarrow$  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** instead of a **real situation**? (e.g., when using weights for categorical predictors, the predictions will not correspond to real situations),<sup>9</sup>
- is the predictor distribution in the observed data **representative** for the population or the targeted setting?<sup>9</sup>
- are the predictor distributions **independent** enough to set the values for different predictors independently?<sup>9</sup>
- 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).

---

<sup>9</sup> **Beware** that 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 and 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 means ~ **median** wpc-values:

Estimates*		$\sqrt{wpc}$ (mean)		wpc (median)	
rp	vag_d	dyst=0	dyst=1	dyst=0	dyst=1
0	0	7.328	7.870	53.70	61.94
0	1	7.315	7.857	53.51	61.73
1	0	7.717	8.260	59.56	68.22
1	1	9.195	9.737	84.55	94.82

\* at: parity=1, twin=0, herd\_size=234, aut\_calv=0 (~ mean herd size, 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