PhD course 2023

Epidemiological methods in medical research

Interaction, quantitative covariates

Clayton & Hills, Ch. 24.3-7, 25-26

23 February 2023

Per Kragh Andersen

Diet data, additive model for the log(Rate)

Table 24.1. Program output for the diet data

Parameter	Estimate (M)	SD (S)	W
Corner	-5.4180	0.4420	
Exposure (1)	0.8697	0.3080	7.97
Age (1)	0.1290	0.4753	0.07
Age (2)	0.6920	0.4614	2.25

Max. log-likelihood is -247.03

Interaction

We have assumed that the **effect of exposure** is **constant over age bands** (and vice versa).

Is that reasonable?

Or is there **interaction** between age and exposure?

$$\log(\text{Rate}) = \text{Corner} + \text{Exposure} + \text{Age} + \text{Exposure}\cdot\text{Age}$$

Note the relationship with the Breslow-Day test for homogeneity over age strata. However, we now:

- get a quantification of heterogeneity
- are able to adjust for other explanatory variables when examining interaction

Table 24.5. Estimates of parameters in the model with interaction

Parameter	Estimate	SD
Corner	-5.0237	0.500
Exposure(1)	-0.0258	0.866
Age(1)	-0.5153	0.671
Age(2)	0.3132	0.612
Age(1) · Exposure(1)	1.2720	1.020
Age(2) · Exposure(1)	0.8719	0.973

Test for no interaction: Max. log likelihood for

$$\text{Corner} + \text{Age} + \text{Exposure} + \text{Age.Exposure}$$

is -246.19 leading to the LR test 1.67 (2 d.f.)

Illustrative example **without interaction Table 22.4**

	Exposure	
Age	0	1
0	5.0	15.0
1	12.0	36.0
2	30.0	90.0
0	5.0	5.0×3.0
1	12.0	12.0×3.0
2	30.0	30.0×3.0
0	5.0	5.0×3.0
1	5.0×2.4	$5.0 \times 2.4 \times 3.0$
2	5.0×6.0	$5.0 \times 6.0 \times 3.0$

Corner = 5.0

Age(1) = 2.4

Exposure (1) = 3.0

Age(2) = 6.0

Example: Illustrative values of rates with interaction

Table 24.2. Definition of interactions in **terms of exposure**

	Exposure	
Age	0	1
0	5.0	15.0
1	12.0	42.0
2	30.0	135.0
0	5.0	5.0×3.0
1	12.0	12.0×3.5
2	30.0	30.0×4.5
0	5.0	5.0×3.0
1	12.0	$12.0 \times 3.0 \times 1.167$
2	30.0	$30.0 \times 3.0 \times 1.5$

interaction parameters

Example: Illustrative values of rates with interaction

Table 24.3. Definition of interactions in **terms of age**

	Exposure		
Age	0	1	
0	5.0	15.0	
1	12.0	42.0	
2	30.0	135.0	
0	5.0	15.0	
1	5.0×2.4	15.0×2.8	
2	5.0×6.0	15.0×9.0	
0	5.0	15.0	
1	5.0×2.4	$15.0 \times 2.4 \times 1.167$	interaction
2	5.0×6.0	$15.0 \times 6.0 \times 1.5$	parameters

Table 24.4. Definition of interactions in **terms of exposure and age**

	Exposure	
Age	0	1
0	5.0	5.0×3.0
1	5.0×2.4	$5.0 \times 3.0 \times 2.4 \times 1.167$
2	5.0×6.0	$5.0 \times 3.0 \times 6.0 \times 1.5$

Exercise 24.4, p. 242

Table 22.6. Energy intake and IHD incidence per 1000 person-years

Current	Exposed (< 2750 kcal)			Unexposed (≥ 2750 kcal)			
age	Cases	P-yrs.	Rate	Cases	P-yrs.	Rate	RR
40–49	2	311.9	6.41	4	607.9	6.58	0.97
50–59	12	878.1	13.67	5	1271.1	3.93	3.48
60–69	14	667.5	20.97	8	888.9	9.00	2.33
Total	28	1857.5	15.07	17	2768.9	6.14	2.45

Verify that, in the model with interaction, the Corner is the log(observed rate) for the youngest unexposed, and Exposure(1) is the log(observed rate ratio) for exposure among the youngest.

Exercise 24.4: solution

Parameter	Estimate	SD
Corner	-5.0237	0.500
Exposure(1)	-0.0258	0.866
Age(1)	-0.5153	0.671
Age(2)	0.3132	0.612
Age(1) · Exposure(1)	1.2720	1.020
Age(2) · Exposure(1)	0.8719	0.973

$$\log \frac{4}{607.9} = -5.0237, \log \frac{\frac{2}{311.9}}{\frac{4}{607.9}} = -0.0258$$

(except for rounding errors)

SAS and R code

```
proc genmod data=ihd;
class exposure (ref='0') age (ref='0') ;
model cases=exposure age exposure*age/dist=poisson offset=lpyrs type3;
run;
```

```
# Fit Poisson regression model with interaction
fit <- glm(cases ~ factor(exposure) + factor(age) +
    factor(age):factor(exposure) + offset(log(pyrs)), ihd, family = "poisson")
summary(fit)

# and compare with model without interaction
fit0 <- glm(cases ~ factor(exposure) + factor(age) +
     offset(log(pyrs)), ihd, family = "poisson")


anova(fit0,fit,test="Chisq")
```

Table 24.5. Reporting estimates from the model with interaction:

Reparametrize into separate effects of Exposure within each Age band.

Parameter	Estimate	SD	RR
Corner	-5.0237	0.500	
Exposure(1)·Age(0)	-0.0258	0.866	0.97
Exposure(1)·Age(1)	1.2461	0.532	3.48
Exposure(1)·Age(2)	0.8461	0.443	2.33
Age(1)	-0.5153	0.671	0.60
Age(2)	0.3132	0.612	1.37

This parametrization may be obtained by excluding the 'main effect' of exposure, e.g. in SAS:

```
MODEL cases = age exposure*age/ ...
```

Interactions: which to study?

When the model contains p covariates there are $p(p-1)/2$ possible two-factor interactions (e.g., 45 for $p = 10$).

It is out of the question to study them all, so a general recommendation is to restrict attention to those that were pre-specified in the research protocol:

"Don't ask a question if you are not interested in the reply!"

There will also be a type I error problem: "if you ask too many questions you will get too many wrong answers".

Interaction is scale dependent (Sect. 26.6)

Table of disease rates (say, per 1000 years):

Factor A	Factor B	
	Absent	Present
Absent	0.1	0.2
Present	0.3	λ

If $\lambda = 0.6$ then the *rate ratio* associated with the presence of factor A is 3 both when factor B is absent or present; and the *rate ratio* associated with the presence of factor B is 2 both when factor A is absent or present.

However, the *rate difference* associated with the presence of factor A is 0.2 when factor B is absent and 0.4 if it is present and the *rate difference* associated with the presence of factor B is 0.1 when factor A is absent and 0.3 if it is present

Factor A	Factor B	
	Absent	Present
Absent	0.1	0.2
Present	0.3	λ

If $\lambda = 0.4$ then the *rate difference* associated with the presence of factor A is 0.2 both when factor B is absent or present; the *rate difference* associated with the presence of factor B is 0.1 both when factor A is absent or present.

However, the *rate ratio* associated with the presence of factor A is 3 when factor B is absent and 2 if it is present and the *rate ratio* associated with the presence of factor B is 2 when factor A is absent and 1.33 if it is present

Interaction is scale dependent

On which scale should we study interaction?

Items to consider:

- Interpretation: absolute vs. relative effects
- Goodness of fit
- 'Biological' interaction

Potential interaction between 2 exposures

Table 24.6. Cases (controls) for **oral cancer** study.

	Alcohol (oz/day, 1 drink $\sim$ 0.3 oz/day).							
Tobacco	0		1		2		3	
(cigs/day)	0		0.1-0.3		0.4-1.5		1.6+	
0 (0)	10	(38)	7	(27)	4	(12)	5	(8)
1 (1-19)	11	(26)	16	(35)	18	(16)	21	(20)
2 (20-39)	13	(36)	50	(60)	60	(49)	125	(52)
3 (40+)	9	(8)	16	(19)	27	(14)	91	(27)

Table 24.7. Case/control ratios for the **oral cancer** data.

	Alcohol			
Tobacco	0	1	2	3
0	0.26	0.26	0.33	0.63
1	0.42	0.46	1.13	1.05
2	0.36	0.83	1.22	2.40
3	1.12	0.84	1.93	3.37

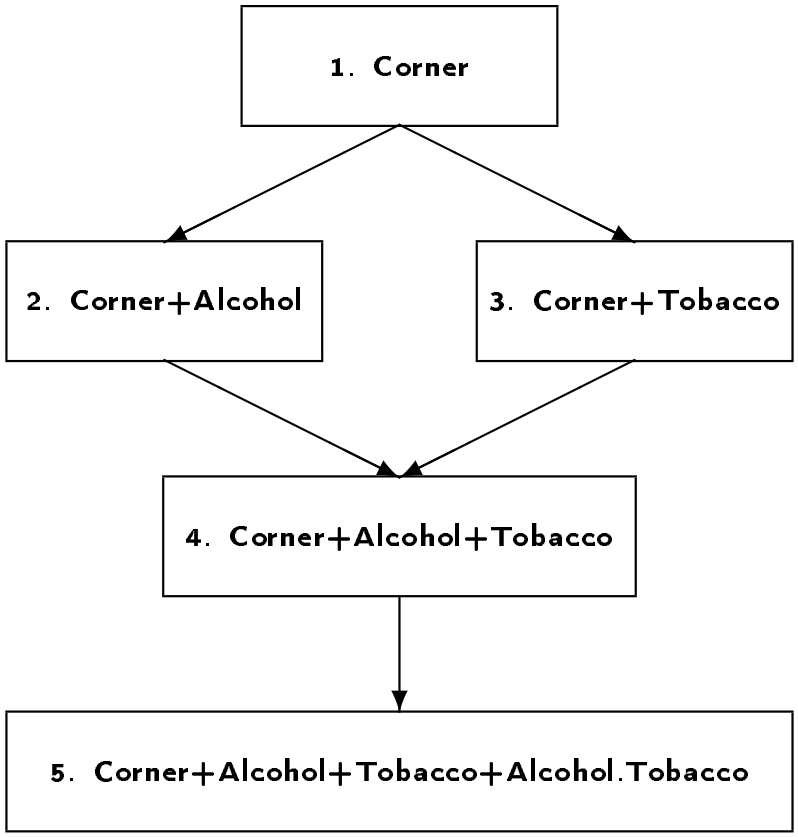
Is the effect of tobacco the same for all levels of alcohol consumption?

SYNERGY?

= INTERACTION

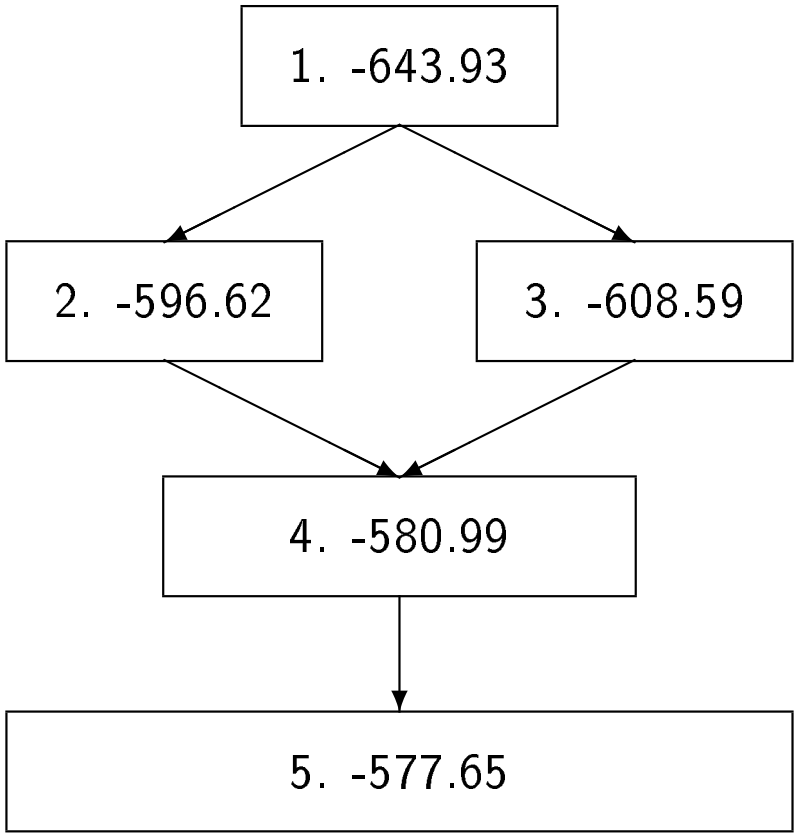
Note that **CORRELATION** is something completely different

Fig. 24.2. Nesting of models



Exercise 24.6, p. 246: Calculate the likelihood ratio test statistics between successive nested models.

Exercise 24.6: Log-likelihoods



Exercise 24.6: solution

Likelihood ratio tests:

Hypothesis	Test statistic	Degrees of freedom
Model 4. vs. 5.	$2 \cdot (580.99 - 577.65) = 6.68$	9=16-7 =(4-1)(4-1)
Model 2. vs. 4.	$2 \cdot (596.62 - 580.99) = 31.26$	3=4-1
Model 3. vs. 4.	$2 \cdot (608.59 - 580.99) = 51.20$	3=4-1

Quantitative covariates (dose-response models)

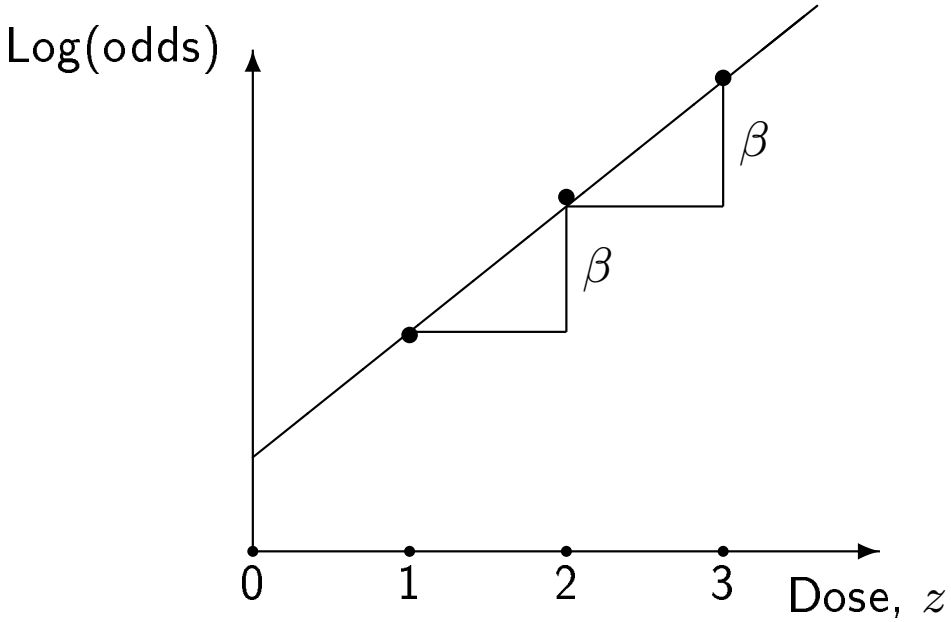
Explanatory variables with **ordered categories**.

Table 25.1. Alcohol and tobacco treated as categorical variables

Parameter	Estimate	SD
Corner	-1.6090	0.2654
Alcohol(1)	0.2897	0.2327
Alcohol(2)	0.8437	0.2383
Alcohol(3)	1.3780	0.2256
Tobacco(1)	0.5887	0.2844
Tobacco(2)	1.0260	0.2544
Tobacco(3)	1.4090	0.2823

Alternative: monotone effect of tobacco

Fig. 20.1. Log-linear trend



Look at **successive differences** between effects:

Tobacco(1), Tobacco(2)-Tobacco(1), Tobacco(3)-Tobacco(2)

Exercise 25.1, p. 249: Calculate the values of these new parameters.

Introduce a variable

taking values 0, 1, 2 or 3 and denote its effect by

$$[\text{Tobacco}]$$

Model: $\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}]$

Exercise 25.1: solution

Table 25.1. Alcohol and tobacco treated as categorical variables

Parameter	Estimate	SD	Succ. diff.
Corner	-1.6090	0.2654	
Alcohol(1)	0.2897	0.2327	0.2897
Alcohol(2)	0.8437	0.2383	0.5540
Alcohol(3)	1.3780	0.2256	0.5543
Tobacco(1)	0.5887	0.2844	0.5887
Tobacco(2)	1.0260	0.2544	0.4373
Tobacco(3)	1.4090	0.2823	0.3830

Model: log(Odds) = Corner + Alcohol + [Tobacco]

Table 25.2. The **linear effect of tobacco** consumption

Alcohol	Tobacco	log(Odds)=Corner + ...
0	0	-
0	1	1×[Tobacco]
0	2	2×[Tobacco]
0	3	3×[Tobacco]
1	0	Alcohol(1)
1	1	Alcohol(1)+1×[Tobacco]
1	2	Alcohol(1)+2×[Tobacco]
1	3	Alcohol(1)+3×[Tobacco]
2	0	Alcohol(2)
2	1	Alcohol(2)+1×[Tobacco]
2	2	Alcohol(2)+2×[Tobacco]
2	3	Alcohol(2)+3×[Tobacco]
3	0	Alcohol(3)
3	1	Alcohol(3)+1×[Tobacco]
3	2	Alcohol(3)+2×[Tobacco]
3	3	Alcohol(3)+3×[Tobacco]

Table 25.3. Linear effect of tobacco per level

Parameter	Estimate	SD
Corner	−1.5250	0.219
Alcohol(1)	0.3020	0.232
Alcohol(2)	0.8579	0.237
Alcohol(3)	1.3880	0.225
[Tobacco]	0.4541	0.083

Both in SAS and R, treating a variable as quantitative is the *default*.

That is, to treat `x` as quantitative, you should *not* declare it as `CLASS` in SAS, and you should write `x` instead of `factor(x)` in R.

Similarly with **alcohol consumption:**
introduce variable with values=0, 1, 2 or 3
and denote its effect [Alcohol]

Table 25.4. Linear effects of alcohol and tobacco per level

Parameter	Estimate	SD
Corner	–1.6290	0.1860
[Alcohol]	0.4901	0.0676
[Tobacco]	0.4517	0.0833

Exercise 25.3, p. 251: Estimate the log(odds ratio) between (Alc 3, Tob 3) and (Alc 0, Tob 0) based on the models in Tables 25.1 and 25.4 (i.e., either treating both as categorical or treating both as quantitative).

Exercise 25.3: solution

Tobacco(3)+Alcohol(3)=2.7870

$3 \times [\text{Tobacco}] + 3 \times [\text{Alcohol}] = 2.8254$

Alternative ways of scoring

Tobacco: cigarettes/day (0 : 0, 1-19 : 10, 20-39 : 30, 40+ : 50)

Alcohol: ounces/day (0.0 : 0, 0.1-0.3 : 0.2, 0.4-1.5 : 1.0, 1.6+ : 2.0)

Table 25.5. Alcohol in **ounces/day** and tobacco in **cigarettes/day**

Parameter	Estimate	SD
Corner	−1.2657	0.1539
[Alcohol]	0.6484	0.0881
[Tobacco]	0.0253	0.0046

Test for linearity

1. Compare the "nested" models:

   $$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco}$$

   and

   $$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}],$$

   here: LR test=0.38, 2. d.f.,

2. Eliminate [Tobsq] (=0, 1, 4, 9) from

   $$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}] + [\text{Tobsq}],$$

   here LR test=0.02, 1 d.f.

Trend test

Always 1. d.f.!

Eliminate [Tobacco] from the model:

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}],$$

here LR test=30.88 (Wald test: similar).

Using individual levels of the quantitative covariate

Why not use individual levels, that is, a truly quantitative covariate and no categorization at all?

Pros and cons

- Information is lost by categorization
- Categories may be more robust (e.g., smoking)
- Few outliers may have large influence ("Casanova effect"!)
- Model with a linear effect is no longer "nested" in categorical model $\Rightarrow$ alternative alternatives are needed when testing linearity

Indicator ('dummy') variables

The way in which the categorical covariates are entered into the regression model.

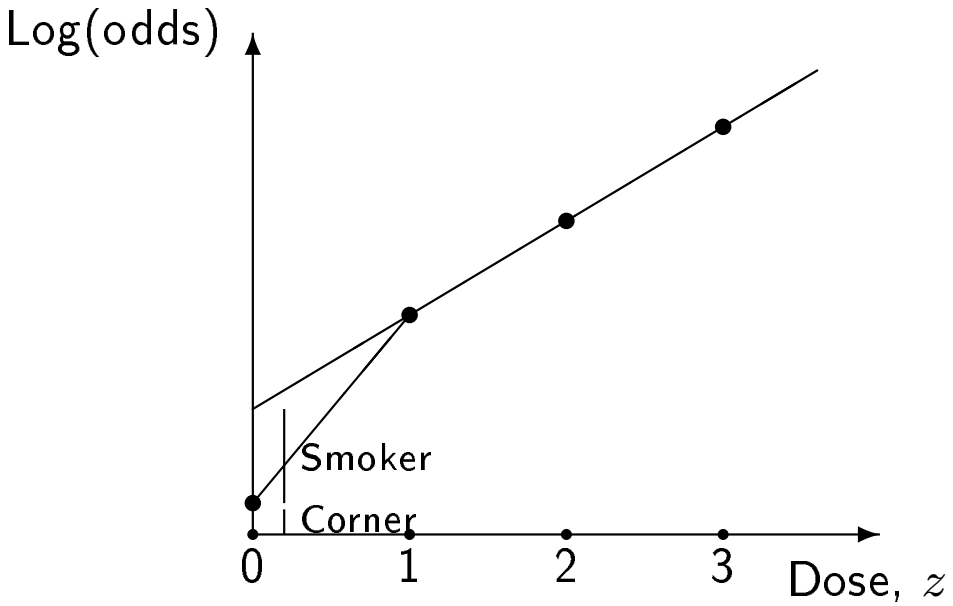
Table 25.8. Indicator variables for the four alcohol levels - include A_1, A_2, A_3:

A_0	A_1	A_2	A_3	Level	log(Odds) = Corner + $\cdots$
1	0	0	0	0	–
0	1	0	0	1	Alcohol(1)
0	0	1	0	2	Alcohol(2)
0	0	0	1	3	Alcohol(3)

The use of indicator variables enables the programmer to choose his/her preferred **reference level** by *excluding the corresponding indicator* (here: level 0).

Treating the zero level differently

Fig. 25.1. Separating zero exposure from the dose-response.



Corresponds to adding a **new variable** [Smoker]

'**Table 25.11.**' Separating zero exposure from the dose-response

Tobacco	Smoker	log(Odds) = Corner + $\cdots$
0	0	-
1	1	[Smoker] + 1 $\times$ [Tobacco]
2	1	[Smoker] + 2 $\times$ [Tobacco]
3	1	[Smoker] + 3 $\times$ [Tobacco]

Truly quantitative covariates, x

In a model like

$$\text{log(Rate)=Corner + Exposure + } [x]$$

the effect of x is assumed to be linear, i.e. $[x]$ expresses the change in log(Rate) per 1 unit change of x.

To test for linearity, one may add [xsq] to the model where xsq$= x^2$.

An alternative alternative is a *linear spline*.

Linear splines

An alternative to a straight line could be a broken line.

Introduce break points for x, e.g., a_1, a_2, a_3 and add the three linear splines

$$I_1 \times [x - a_1], I_2 \times [x - a_2], I_3 \times [x - a_3]$$

to $[x]$:

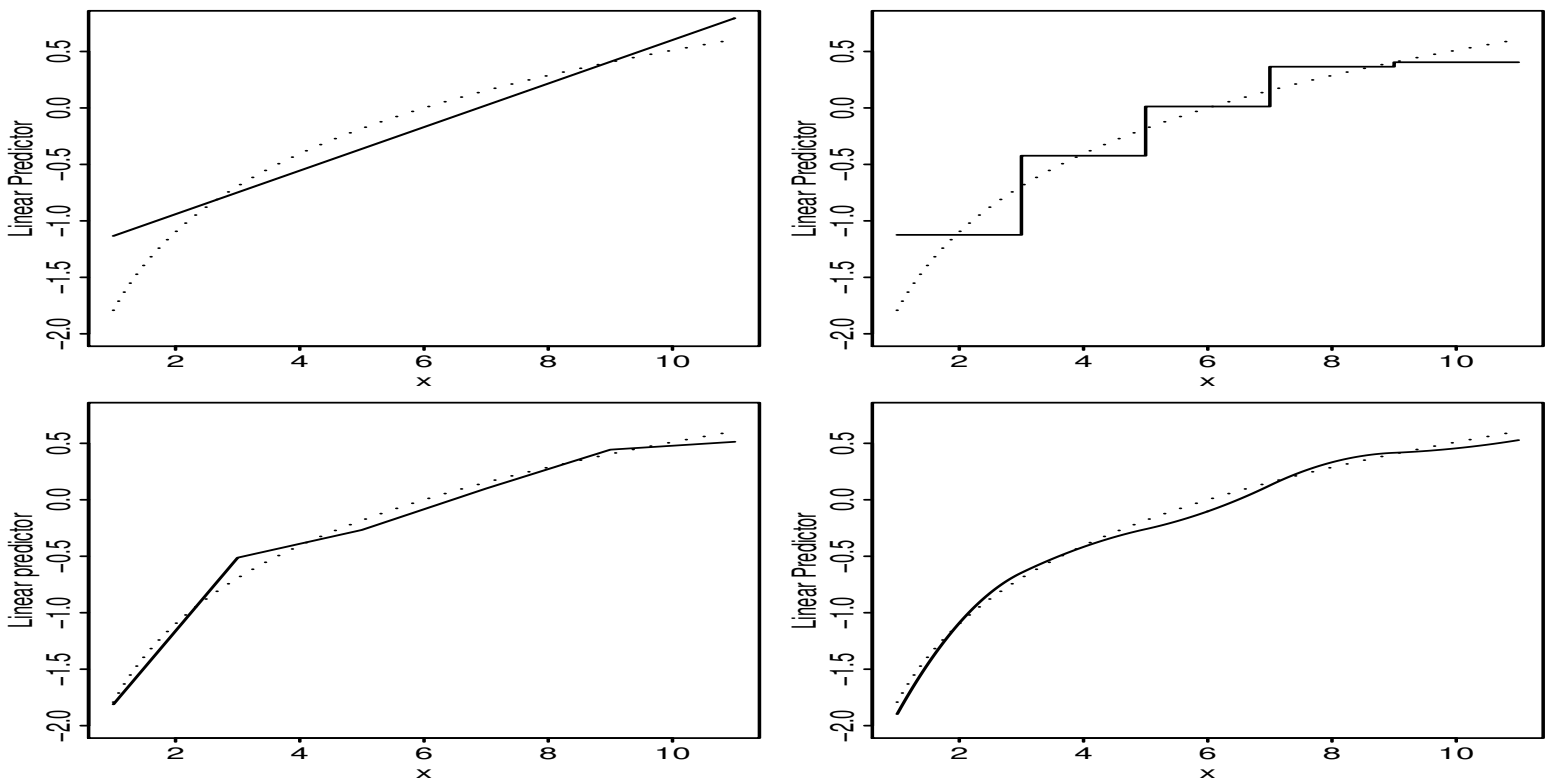
Here, $I_1 = $ indicator for $x \geq a_1$

$I_2 = $ indicator for $x \geq a_2$

$I_3 = $ indicator for $x \geq a_3$

The parameter for the spline $I_1 \times [x - a_1]$ gives the *change in slope* at the break point a_1. Similarly for a_2, a_3.

Linear splines are easy to program and parameters are easier to interpret than for quadratic terms (quadratic and cubic splines also exist - but then the nice interpretation is lost).


Linear Predictor
x
Linear Predictor
x
Linear predictor
x
Linear Predictor
x

Code for indicator variables and splines (1)

Indicator variables (Z0, Z1) in SAS may be created in the obvious way from a binary variable Z:

```
if Z=0 then Z0=1; if Z=1 then Z0=0;
if Z=0 then Z1=0; if Z=1 then Z1=1;
```

A shorter, but less transparent code uses 'logical expressions':

```
Z0=(Z=0); Z1=(Z=1);
```

Then include *either* Z0 *or* Z1 in the model (depending on the preferred reference group).

Code for indicator variables and splines (2)

The last way of coding makes creation of splines easy.

Suppose X is quantitative and we want a linear spline with break points at A1 and A2:

```
X1=(X-A1)*(X>A1); X2=(X-A2)*(X>A2);
```

Then include *both* X, X1 *and* X2 in the model to obtain a piecewise linear effect of X.

The test for linearity corresponds to eliminating *both* X1 *and* X2 from the model.

Completely analogously in R:

```
X1<-(X-A1)*(X>A1)
X2<-(X-A2)*(X>A2)
```

Higher order ('smoothing') splines

Linear splines are easy to compute and they provide parameters with a simple interpretation.

Higher order ('quadratic' or 'cubic') splines are often available in regression software. They are smooth and more flexible than linear splines, however, they are purely descriptive and provide estimated dose-response curves, but no interpretable parameters.

Last curve on the figure is a quadratic spline.