

# Epidemiology workshop

Cate Dewey

## Coding variables

Example: Temperature on loss during transport of pigs

1. linear (not very logical increase of 1 degree at 10 °C will not have the same influence as an increase at 28°C)
2. linear + quadratic + cubic (curve with a high loss at a temp lower than e.g. -25°C and above e.g. 28 °C)
3. Categorical: loss related to  $T > 27$  °C or  $T$  between 24 – 27 °C or below 24 °C
4. Categorical by season or month

## Hierarchical dummy variables

Normal dummies

Temperature	T11	T12	T13	T14
11	1	0	0	0
12	0	1	0	0
13	0	0	1	0
14	0	0	0	1

Hierarchical dummies

Temperature	T11	T12	T13	T14
11	1	0	0	0
12	1	1	0	0
13	1	1	1	0
14	1	1	1	1

The interpretation is that if T12 is significant it will mean that temperatures above 12 degrees will have the effect. If there is another dummy variable significant the effect is related to the interval.

## Cross-over study on farm

An example was given on a cross-over study on the effect of food withdrawal on in-transit loss. Although the study was too small the example shows that a veterinarian can measure simple things on farms to be able to convince the farmer.

## 2<sup>nd</sup> source data, existing data

Validation

- Paper records
- no paper record
- ? outlier ?

How do you validate the data if you have no paper records?

- descriptive statistics
  - o e.g. different ways of recording (fraction or percentage of haematocrit)
  - o look at maximum and minimum and the numbers

- are the data biologically possible

Outliers are those values that are biologically possible but extreme and influence the outcome of the model (e.g. previous lactation length of 42 days is in the model previous lactation length is significant in the model and without the record with 42 days lactation length it is not significant).

In this case you should probably describe both the outcomes in the paper.

Validation of data is sometimes possible if the same animal has multiple record. One can e.g. calculate the farrowing to service interval from different records, and drop the interval what is not correct e.g. farrowing after service.

## Data from 2 sources

- farm management
- research data
- slaughter data

Risk of overinflation Validation of in each data sources, but also after merging. E.g descriptive statistics on original farm management dataset by farm, and also after merging by doing descriptive statistics per farm.

## Understanding limitations of our data

- Representative of the population
- Covariate representative for the situation (e.g. max temp as data is that representative for the overall temperature or would the maximum be important)
- Can only draw conclusions within the range you measured (advice to restrict graphs between the 10 and 90 percentile of observed observations)
- Drivers (for collection (often not collected for the research, so beware of selection bias)
- Differences between slaughter plants e.g. in average pneumonia and a farmer send pigs to both slaughter slaughterhouses. In that case the slaughterhouse should be in the model