EPSE 596: Correlational Designs & Analysis

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Experimental vs. Observational Paradigm

- Quantitative study design, and so statistics, were traditionally split into two traditions:
  - Experimental: analysis of variance (ANOVA)
  - Observational: regression
- Modern days, no need to separate these two traditions, as the mathematics is ultimately the same.
- However, some distinctive hallmarks remain:

Experimental vs. Observational Paradigm

- Experimental: direct manipulation of variable(s) of interest: researcher control of *treatments*, existence of *"no treatment"* group.
- Observational: no or imprecise control.

- Experimental: categorical predictors only.
- Observational: continuous and categorical predictors.

- Experimental: direct causal claims *sometimes* possible.
- Observational: direct causal claims never* possible.

Experimental vs. Observational Paradigm

- Mathematically, statistical analysis of the experimental (ANOVA) tradition is essentially subsumed as a special case of the observational (regression) tradition, but...
- ...But *interpretations* (i.e. causal vs. correlational) remain different.
- Classical ANOVA and (parametric) regression are special cases of the *general linear model* (very general).
- *ALL* modern techniques propose some kind of *model* for the phenomena of interest, usually written as:

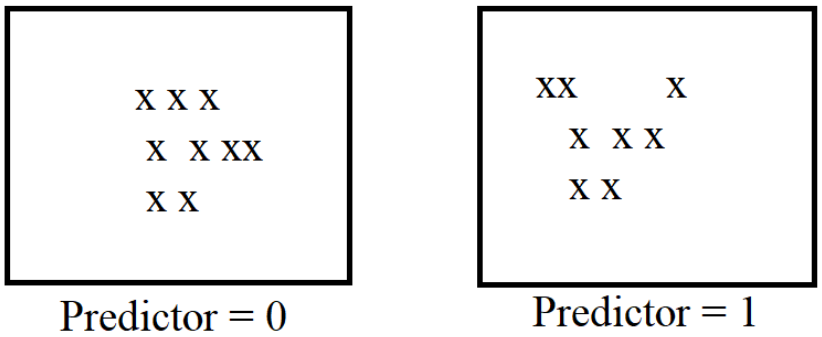
$$\text{response} = \text{predictors} + \text{error}$$

or

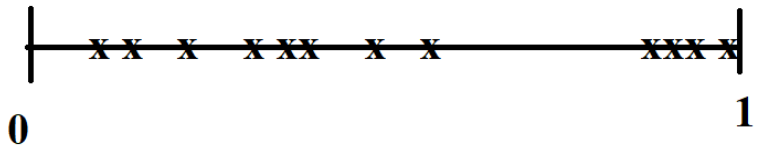
$$g(\textit{response}) = f(\textit{predictors}) + \textit{error}$$

Difficulties of continuous predictors

- If you are studying how some random *response* phenomenon *varies* over different levels of some *predictor* variable...
  - If your predictor is *discrete* (usually finitely many categories), then you just need to collect multiple observations within *each category* to understand how variation in the predictor explains variation in the response.



  - If your predictor is *continuous* (over a continuum), then you can *never* study how the response varies over all possible predictor values.



Difficulties of continuous predictors

- This means we must be extremely careful about *interpolating* or *extrapolating*.
  - Interpolation: filling in the blanks between two numbers; e.g., observe effect of blood pressure medication outcomes for patients aged $<$50 or $>$65, so is it okay to *interpolate* what we learned for ages *between 50 and 65*?
  - Extrapolation: filling in the blanks when you only have information to *one side*; e.g., observe effect of blood pressure medication outcomes on patients aged $<$50, so is it okay to *extrapolate* what we learned for ages *above 50*?
- General scientific best-practice guidelines to follow:
  - NEVER extrapolate.
  - Only interpolate when you have good reason to suspect that the blank spaces in your predictors are not hiding important information.

Difficulties of continuous predictors

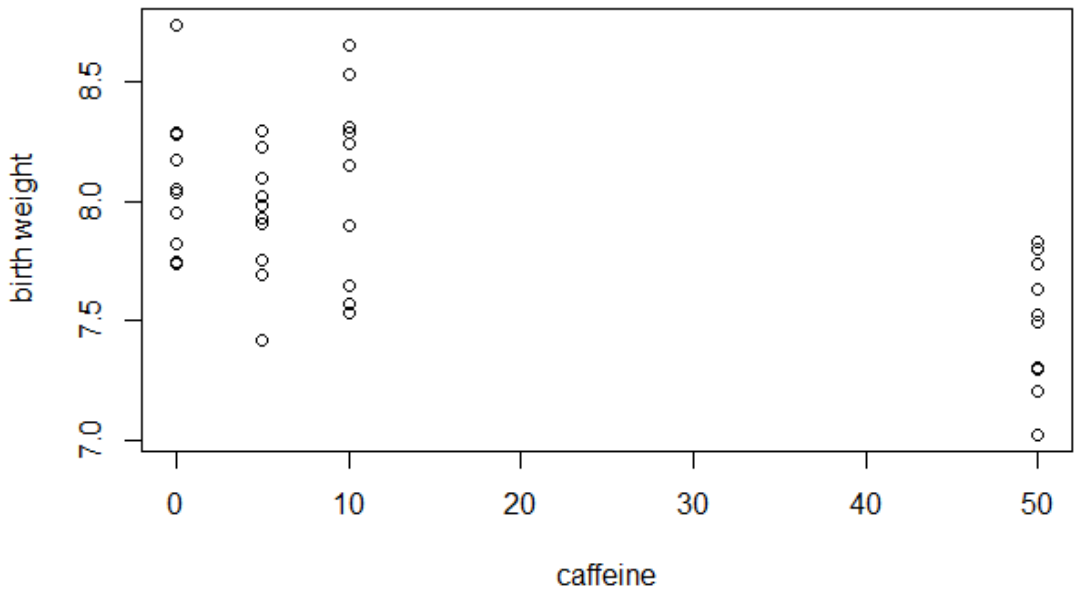
Another big problem of working with continuous predictors is that there are (infinitely) many different ways a response can be a function of such a predictor.

- Consider a scenario (1) where we are trying to model/understand birth weight as a function of mother's daily caffeine intake.
  - Scenario (1a): We experimentally manipulate/control caffeine intake in sample subjects as: 10 mothers consume exactly 0 mg/day, 10 mothers consume exactly 5 mg/day, 10 mothers consume exactly 10 mg/day, 10 mothers consume exactly 50 mg/day.
  - Scenario (1b): We forgo experimental manipulation/control and just observe outcomes on sample subjects who set their own caffeine-intake levels, say, ranging from 0 mg/day to 50 mg/day.

Difficulties of continuous predictors

Scenario (1a):

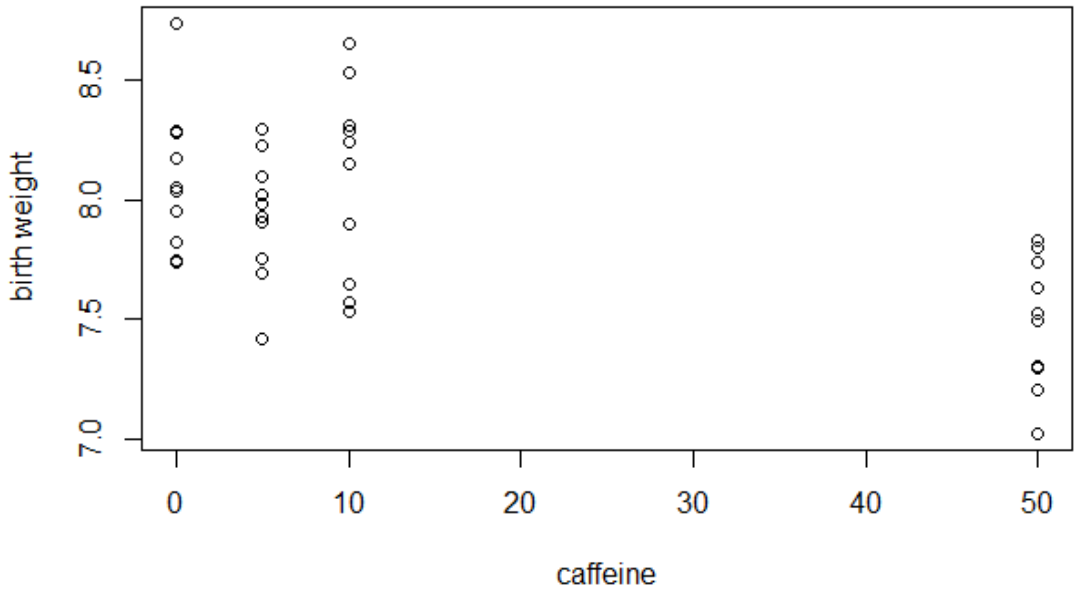
- Here, we are interested in comparing the 4 different treatments/diets directly.
- Every sample subject falls into exactly one of 4 different categorical intake categories; within each category, sample subjects are *exchangeable* (at least with respect to caffeine intake).



Difficulties of continuous predictors

Scenario (1a):

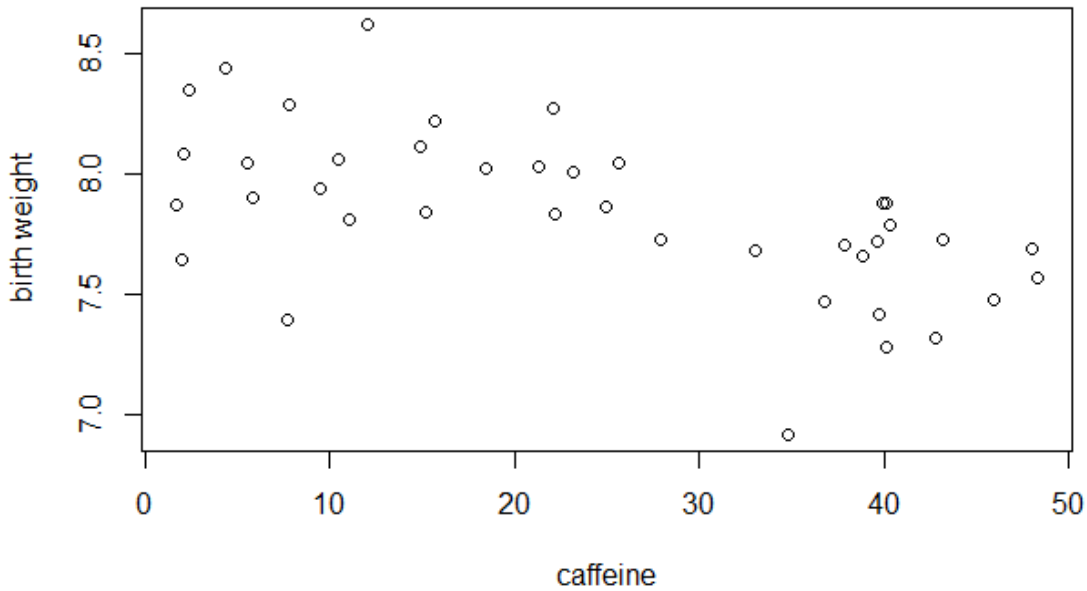
- An ANOVA (or regression) would compare pairwise mean differences between the *discrete* number of treatment groups.
- Note though, that any justification for *interpolating* the observed effects would be based upon previously established domain knowledge (probably reasonable in this context) rather than direct experimental evidence.



Difficulties of continuous predictors

Scenario (1b):

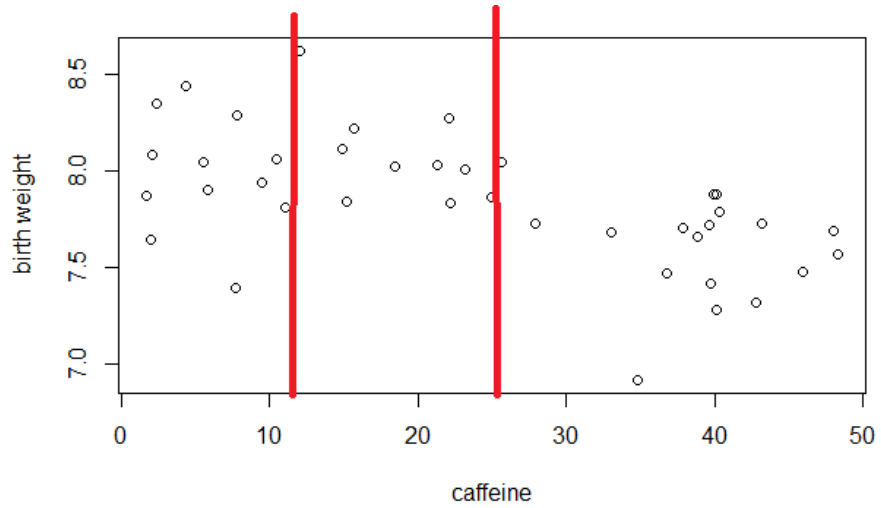
- Here, we observe many different "treatment levels," as the predictor of caffeine intake can apriori vary *anywhere* between 0 and 50 mg/day.
- Very likely that every sample subject has a slightly different daily caffeine intake; just how *exchangeable* are any of them?



Difficulties of continuous predictors

Scenario (1b):

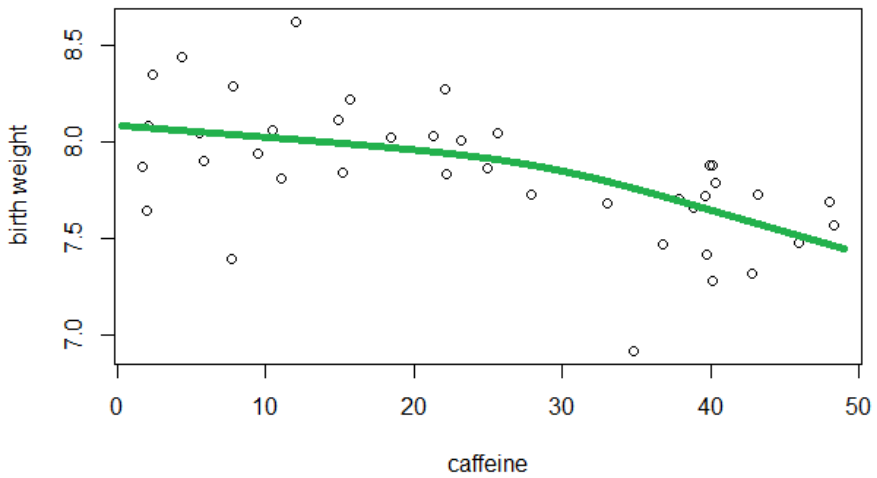
- WARNING: do NOT discretize an observed continuum (very common, but very bad) because:
  - Will rely on arbitrary cutoffs/thresholds
  - Implicitly assumes that all sample units in one category are *exchangeable* (often unjustified/untenable)
  - Observations close to thresholds can create artificial effects that appear as (spurious) associations/effects in a formal analysis.



Difficulties of continuous predictors

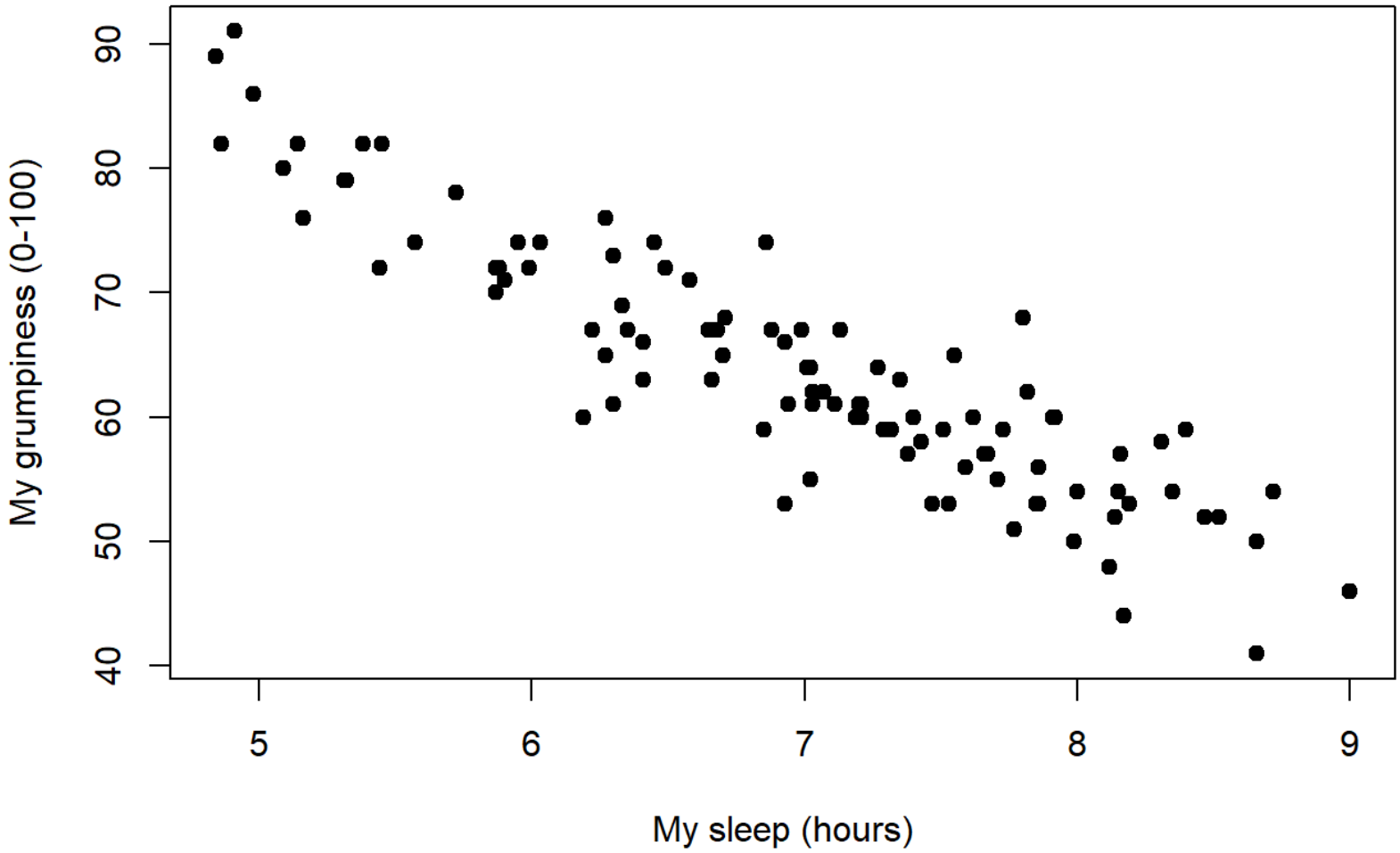
Scenario (1a):

- A proper regression will aim to find the "best" fitting *function* through the data points; here, that looks like it may be slightly curvilinear.
- Note that now *interpolation* of observed effects (on 0 to 50) is more directly evidenced by the raw data themselves, although we have less power to estimate pairwise differences between any select pair of caffeine intakes.



Visualizing/modelling data as predictor(s) vs. response

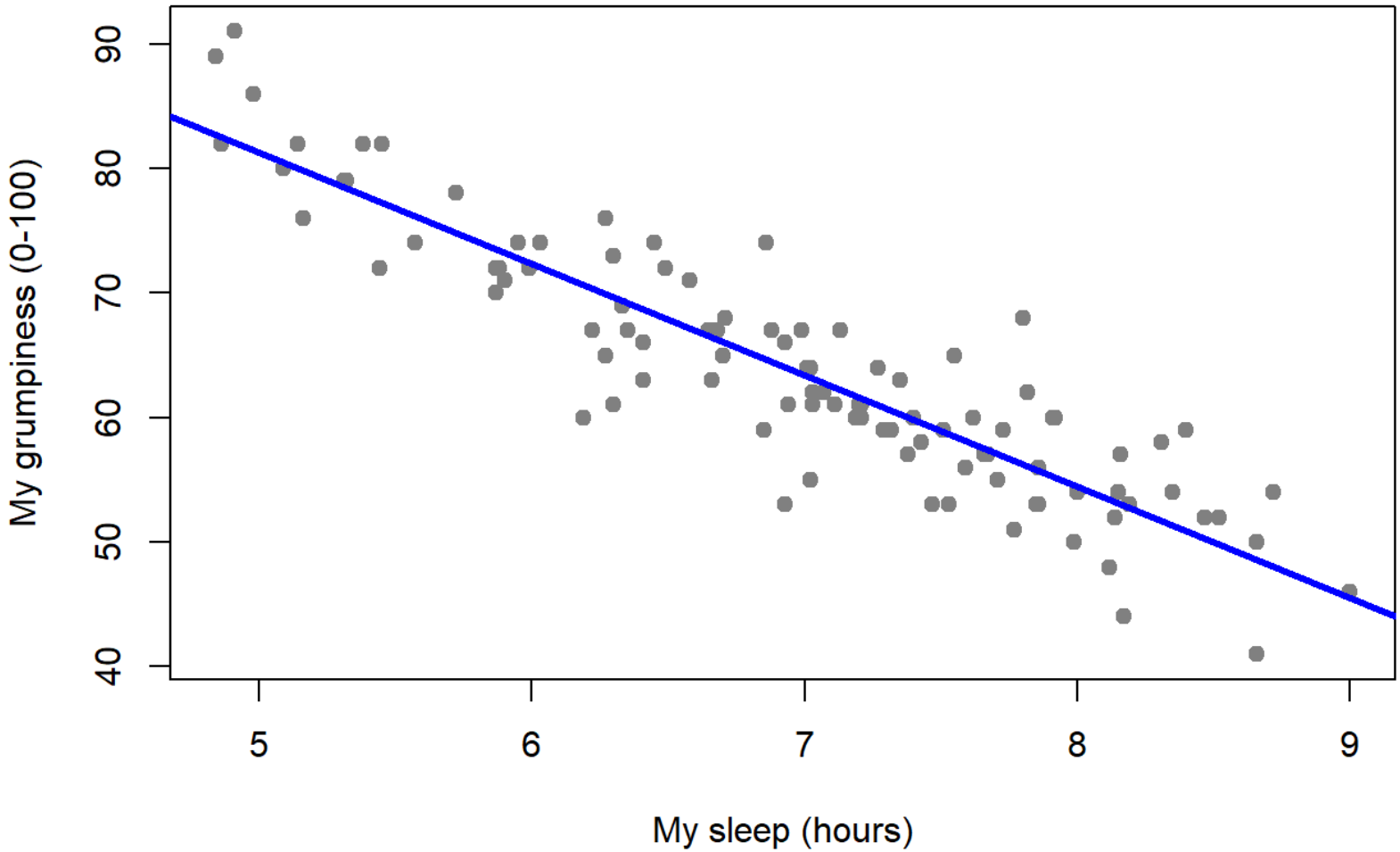
One predictor (horizontal axis, X) vs. one response (vertical axis, Y):



Example from *Learning Statistics with R* by D. Navarro

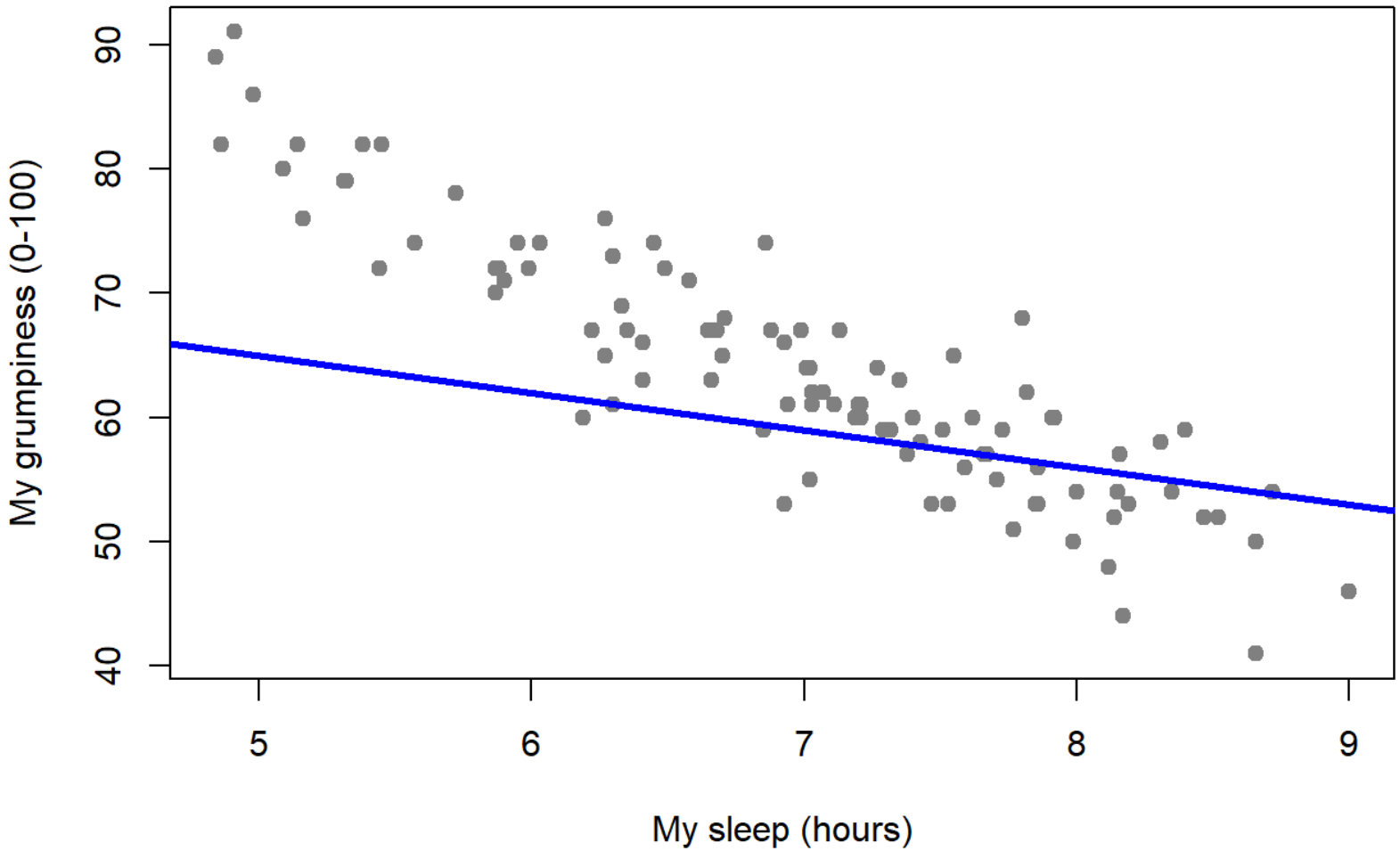
Visualizing/modelling data as predictor(s) vs. response

Could model the *typical* response of Y for any *given* X as a linear relationship: $typical(Y) = a + bX$, where a is the Y-intercept and b is the slope.



Visualizing/modelling data as predictor(s) vs. response

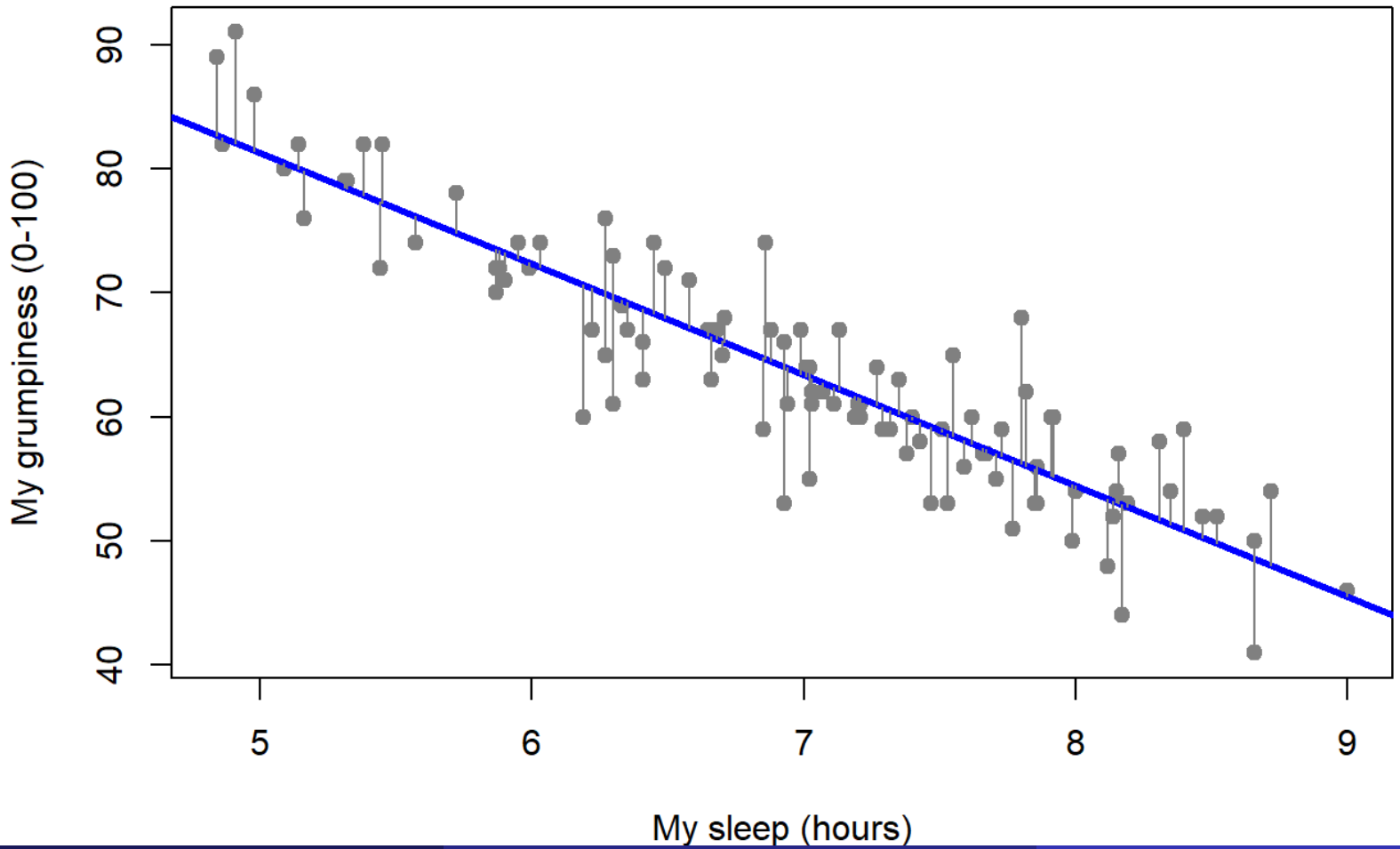
Many different linear relationships possible, but not all are equally plausible: $typical(Y) = c + dX$.



Visualizing/modelling data as predictor(s) vs. response

Many ways to define which relationship is "best", but probably should depend on the amount of *error* that is generated by a proposed model:

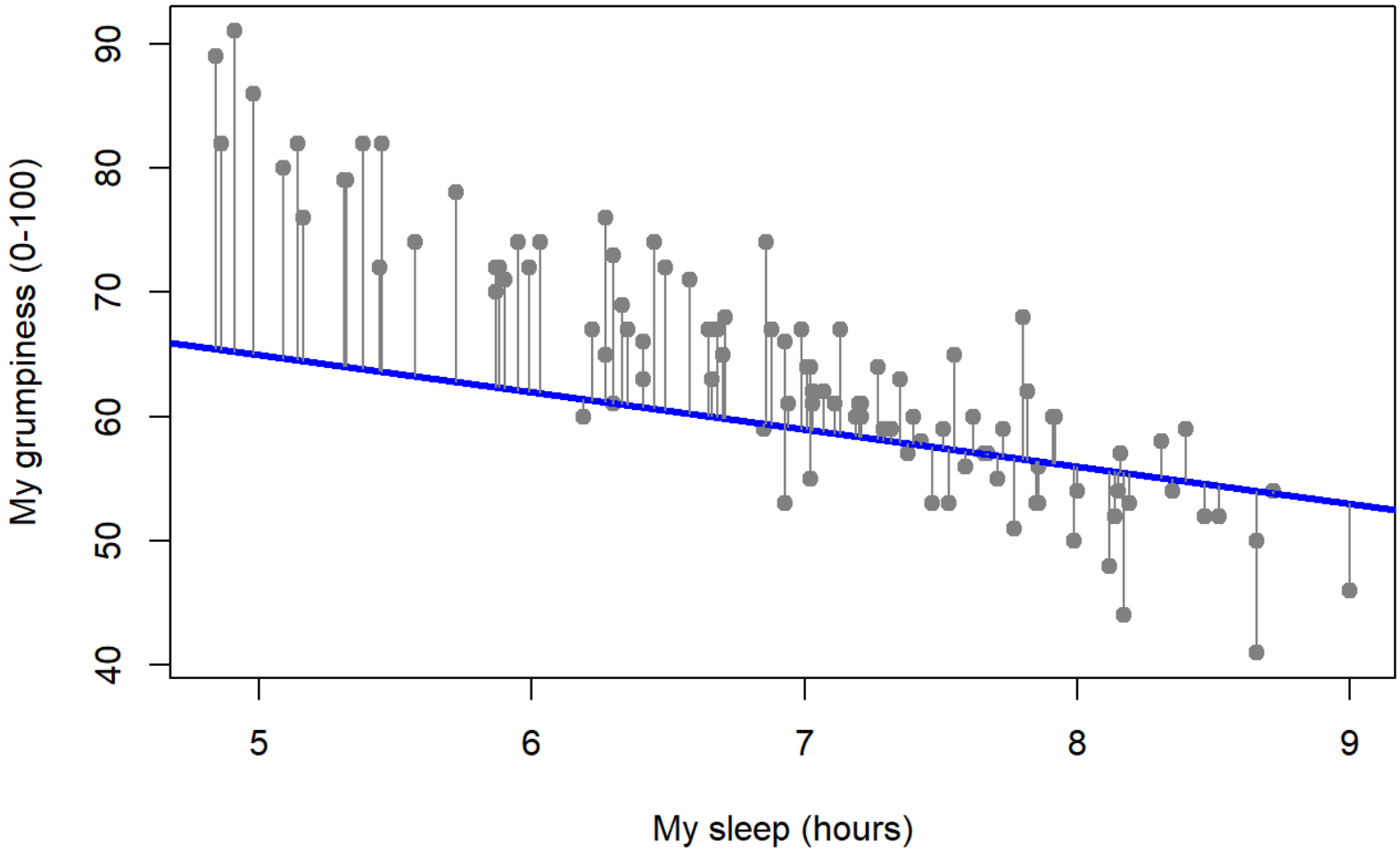
$$Y = a + bX + \varepsilon$$



Visualizing/modelling data as predictor(s) vs. response

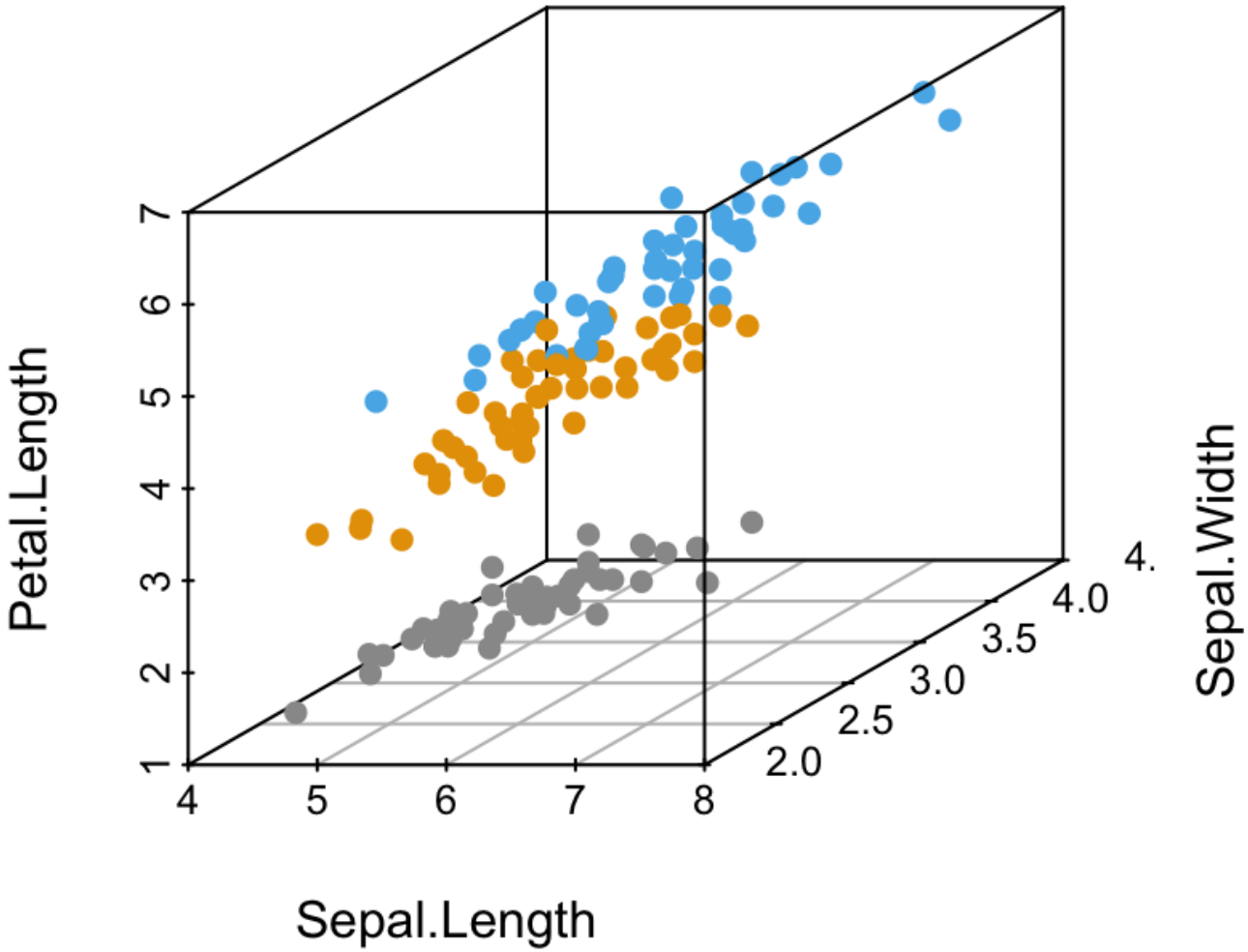
Different hypothesized relationships will generate different *errors*:

$$Y = c + dX + \delta$$



Visualizing/modelling data as predictor(s) vs. response

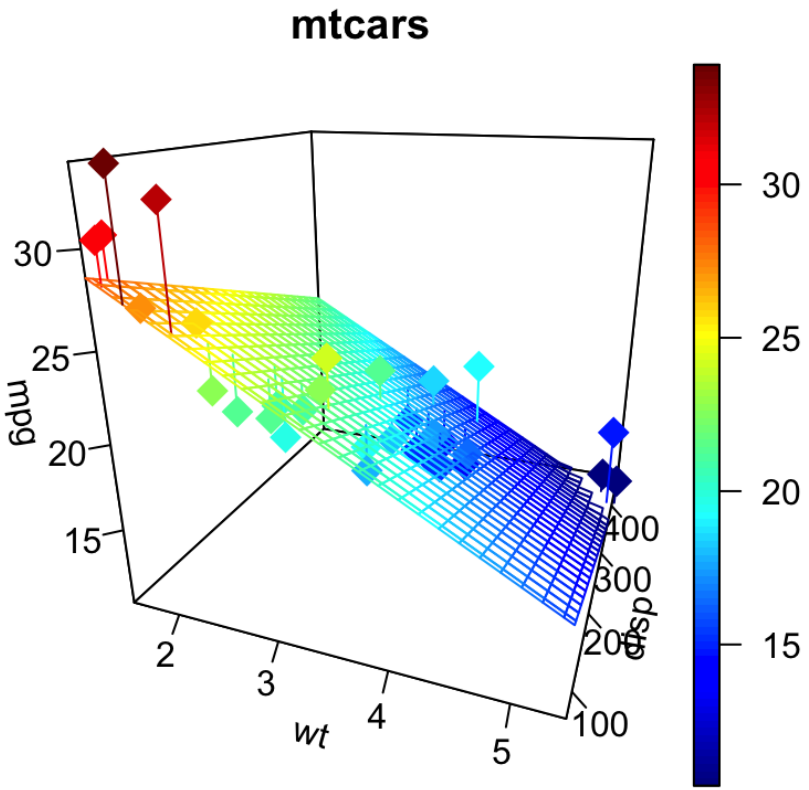
When we have more than one predictor, our data/model will occupy more than two dimensions.



Visualizing/modelling data as predictor(s) vs. response

Can still hypothesize linear relationships between predictors and response to generate *(hyper)planes* of typical responses (plus errors):

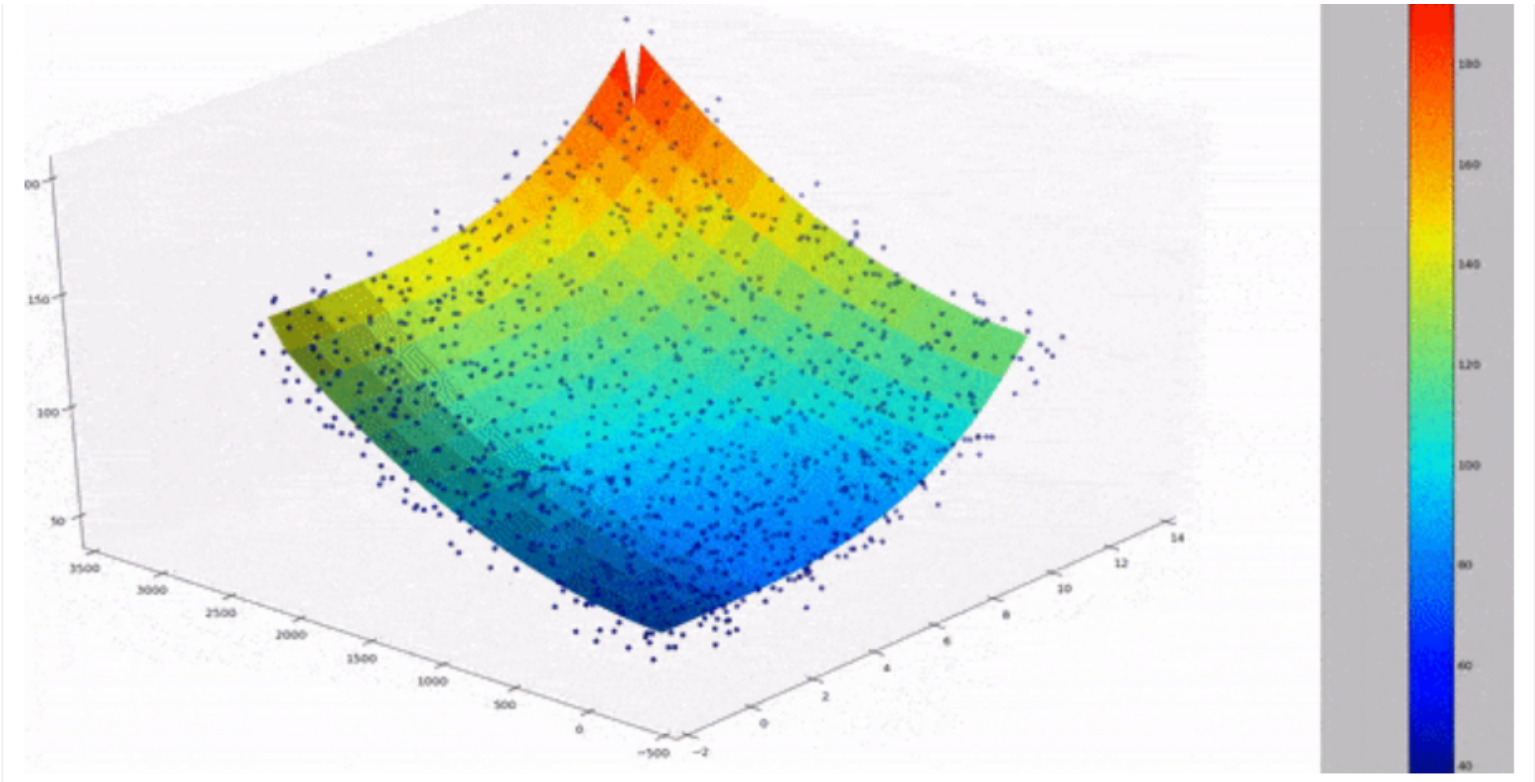
$$Y_{mpg} = a + bX_{wt} + cX_{disp} + \varepsilon$$



Visualizing/modelling data as predictor(s) vs. response

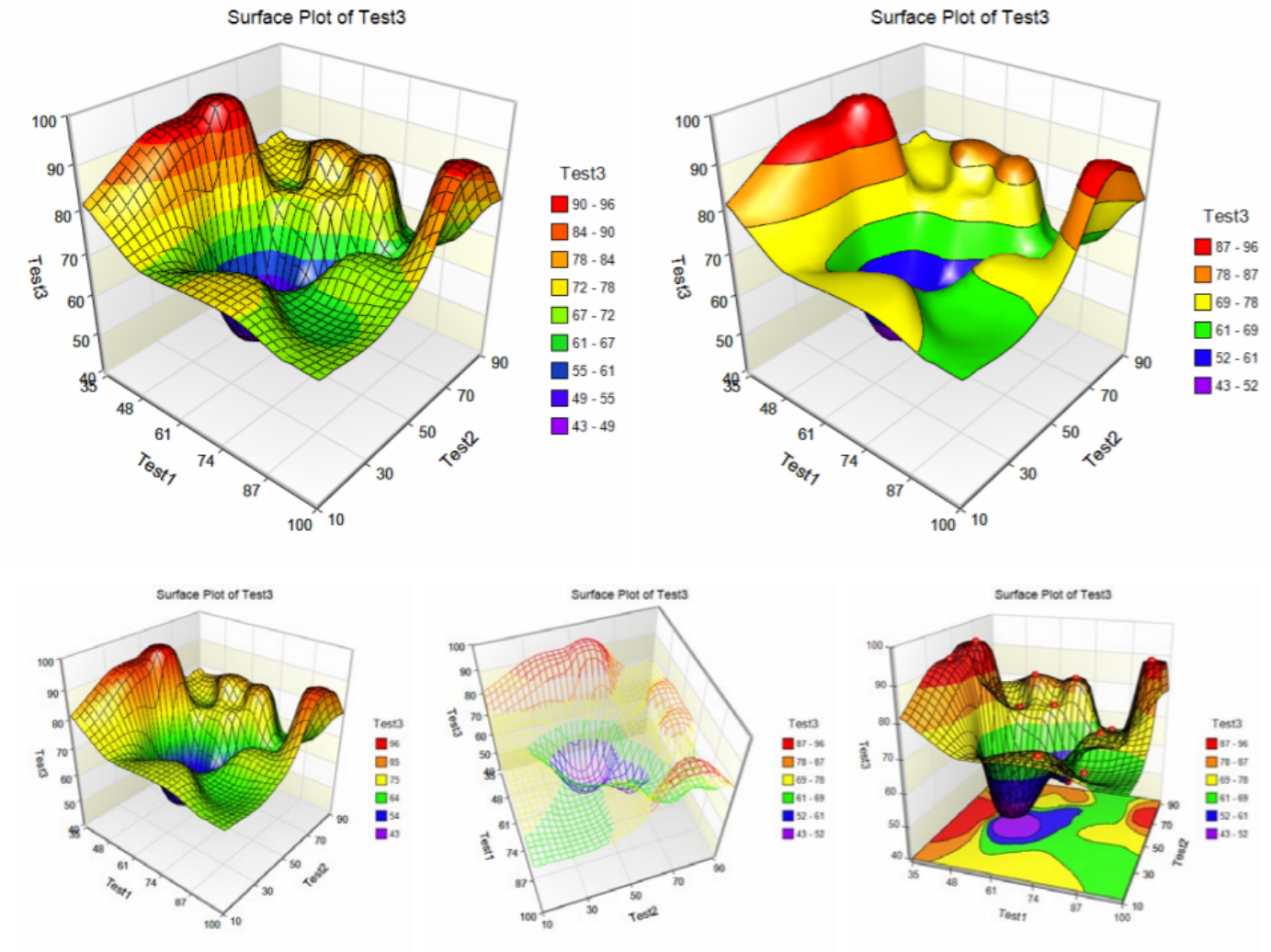
Moreover, relationships need not be strictly linear in the geometric sense:

- `http://www.semspirit.com/artificial-intelligence/machine-learning/regression/polynomial-regression/polynomial-regression-in-python/`



Visualizing/modelling data as predictor(s) vs. response

All kinds of complicated looking relationships can fall under the umbrella of *linear regression modelling*:



Algebraic linearity vs. geometric linearity

- A super-common point of confusion: *Linear regression allows for the modelling of very non-linear relationships!*
- The reason for this confusion is that we use the same words, *linear/linearity*, to mean two different things depending on the mathematical context.
- Geometrically, pretty much everyone understands a *linear relationship* between two variables is one given by a *straight line*. Between more than two variables, the relationship is given by a *flat plane* (or hyperplane).
- Algebraically, on the other hand, a *linear function/transformation/map* is a function f that obeys the following algebraic structure:

$$f(\lambda(x+y)) = \lambda f(x) + \lambda f(y), \text{ for any constant } \lambda \in \mathbb{R}.$$

Algebraic linearity vs. geometric linearity

Examples of algebraically linear and nonlinear maps:

- $f(x) = x$ is algebraically linear, because:

$$f(\lambda(x + y)) = \lambda(x + y) = \lambda x + \lambda y.$$

- However, $f(x) = x + 1$ is NOT algebraically linear:

$$f(\lambda(x + y)) = \lambda(x + y) + 1 \neq \lambda f(x) + \lambda f(y).$$

- So, confusingly, a geometric line through the origin is an algebraically linear function, but a geometric line that does *not* pass through the origin is *not* algebraically linear; instead, both our instances of an *affine function/transformation/map*.

Algebraic linearity vs. geometric linearity

This rears its ugly head in regression all the time because we can think of regression equations as either *functions of their predictor variables* or *functions of their unknown (to be estimated) parameters*:

- For instance, when one considers the following regression equation

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

  as a function of the predictor variable X, then the function is *not* algebraically (nor geometrically) linear.

- However, that *same equation is an example of linear regression* because we consider it to be a function of *the unknown parameters* $\beta_0, \beta_1, \beta_2$; i.e., the function is *algebraically linear in the parameters* even though it is *geometrically nonlinear in the predictors*.

- The term *linear regression* ALWAYS refers to algebraic linearity of the predictors.