EMAT30007 Applied Statistics Lecture 8: Model Building

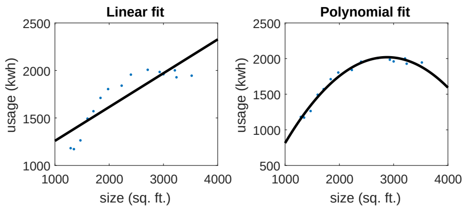
Ksenia Shalonova & Nikolai Bode

Recap and plan

- So far, we have looked at linear models of the form: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} + \epsilon_i$
- In this lecture, we extend this concept:
  - We look at how we can use different types of predictors to capture a wide range of relationships (e.g. capture non-linear relationship with linear model)
  - We'll think about how to formulate models, depending on their use.
  - We'll consider some common pitfalls in the use of linear models.

The importance of model building

- Example: consider energy usage in houses, based on their size.

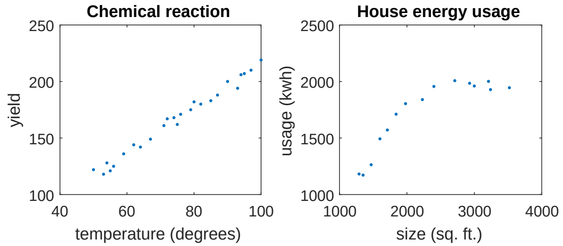


- ... a simple linear model might tell us that a predictor (e.g. size) is important, but it's no good for prediction.
- *Model building* is a process that includes:
  - *Formulating a model* — model structure, we'll look at different types of predictors today.
  - *Model fitting* (last two lectures).
  - *Model evaluation.* Check models assumptions hold, avoid common pitfalls (we'll look at those today).

Good practice before model formulation

Before formulating a statistical model, it is good practice to explore the data.

- Look at scatterplots of predictors against the response (linear/non-linear relationship) and predictors against each other (multicollinearity). Can help to develop an intuition for model structure. Examples:



- Look at distributions of response and predictors (e.g. look for outliers, can you capture the distribution with your model?)
- Think about what the model will be used for (e.g. prediction, or simply to find relevant explanatory variables).

Types of predictors: qualitative vs quantitative

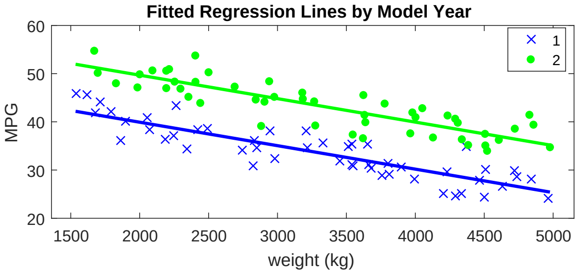
- So far, we have looked at *quantitative predictors* (numerical variables), e.g. temperature, energy usage, waiting time before computer processes data...
- But we can have *qualitative predictors* as well (categorical variables), e.g. type of engine, type of fuel used, type of computer processor used...
- These are included in models using *dummy variables*.
- Example: consider the performance Y_i of diesel engines for three different fuel types A, B and C. We use the model $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$, where $x_{1,i}$ and $x_{2,i}$ are dummy variables such that

$$x_{1,i} = \begin{cases} 1 & \text{if fuel B is used in engine } i \\ 0 & \text{if not} \end{cases} \qquad x_{2,i} = \begin{cases} 1 & \text{if fuel C is used} \\ 0 & \text{if not} \end{cases}$$

... the perfomance for fuel type A is captured in β_0.

Types of predictors: qualitative vs quantitative

- Qualitative and quantitative predictors can be combined in models.
- Example: consider fuel efficiency of car models from the 1950s (type 1) and from the 1960s (type 2), depending on their weight.
- For car models $i = 1, ..., n$, we might consider the model: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \epsilon_i$, where Y_i is the efficieny in miles per gallon (MPG), x_i is a dummy variable for the car type (time period) and w_i is the weight of a car model.
- We might find:

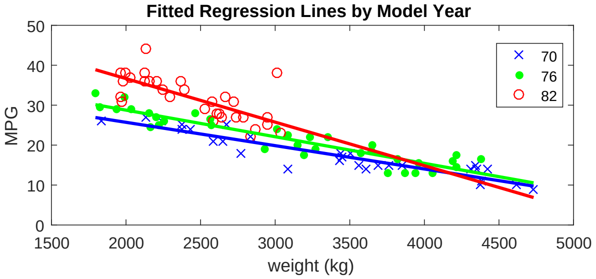


Types of predictors: interaction terms

- So far, we have assumed that the effects of all explanatory variables are additive, e.g. as in $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$.
- What if the relationship between Y_i and $x_{1,i}$ depends on the value of $x_{2,i}$?
- Then we need to consider *interaction terms* in our model, e.g. $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \epsilon_i$.
- Interpretation of model parameters:
  - $(\beta_1 + \beta_3 x_2)$ represents the change in Y for every unit increase in x_1, holding x_2 fixed.
  - $(\beta_2 + \beta_3 x_1)$ represents the change in Y for every unit increase in x_2, holding x_1 fixed.
  - In model checking: if interaction term is important (e.g. T-test), then the interacting explanatory variables must be important and T-tests on them are meaningless.
- This is still a linear model — the effects captured by parameters are additive.

Types of predictors: interaction terms

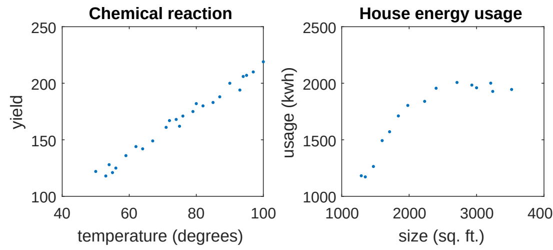
- Example: fuel efficieny Y_i of car models from 1970, '76 and '82 depending on their weight, w_i. Consider interaction between year built and weight.
- $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 w_i + \beta_4 x_{1,i} w_i + \beta_5 x_{2,i} w_i + \epsilon_i$.
- This model has *main effects* for year and weight and *interaction terms*.
- Model fitting:



- To test if interactions are important, could use the Likelihood-ratio test $(H_0 : \beta_4 = \beta_5 = 0)$.

Types of predictors: polynomials of predictors

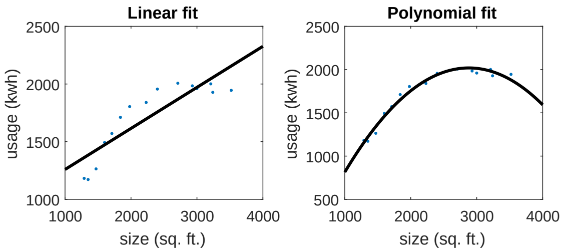
- How to deal with non-linear relationships between variables?
- Example: curvature in relationship between variables.



- Can account for this in models using polynomials of predictors, e.g. $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + \epsilon_i$, where β_0 is the intercept, β_1 is a shift parameter and β_2 is the rate and direction of curvature.
- Higher-order polynomials (e.g. x^3, x^4) are also possible.
- These are still linear models — parameters effects are additive.

Types of predictors: polynomials of predictors

- Example: consider data on energy usage in houses depending on their size and second-order model: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + \epsilon_i$.



- Interpreting model parameters:
  - β_0 can only be interpreteted directly if range of data includes $x = 0$.
  - β_1 no longer represents a slope and cannot be interpeted in isolation.
  - The sign of β_2 indicates the direction of the curvature (concave upward or downward).
- **Warning:** polynomials of predictors lead to correlations between predictors by design (see below: multicollinearity).

Types of predictors: data transformations

Data transformations can help to address problems in model fit (e.g. when relationship is not linear, but residuals are normally distributed). There are many data transformations and there are two examples below, but in general I'd recommend to be cautious about this.

- *Log-transform* predictors, response or both (we'll look at an example in the lab).
- *Code* predictors, so that their range is similar. E.g. for temperature T, code (transform) this as $x = \frac{T-100}{50}$. Can reduce computational rounding errors in model fit and can help to address problems with multicollinearity in polynomial regression models.

Warning: data transformations do not make residuals more normal and statistical tests performed on transformed data are not necessarily relevant for the original data (Feng et al. (2014) *Shanghai Arch Psychiatry.* 26: 105–109).

Pitfalls

- Often there is not one correct way to build a model, espcially for large data sets with many predictors.

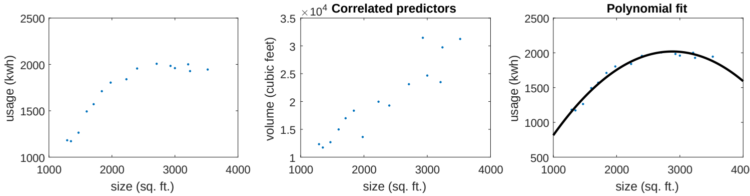
Fact (...*sort of...* — after work by George Box, 1919-2013)

"Essentially, all models are wrong, but some are useful."

- Whether a model is useful or not often depends on how it is used. Things to consider are: prediction or data analysis, exploratory or for decision-making.
- There are many wrong ways of doing things. Common pitfalls are listed in the following.

Pitfalls: multicollinearity

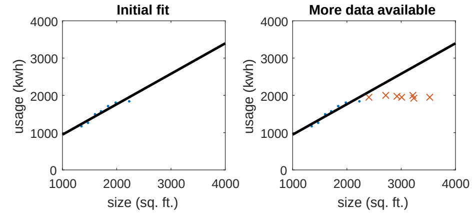
- *Multicollinearity* arises when predictors are correlated.
- This can be *data-based* (i.e. inherent in the data collected) or *structural* when new predictors are created from existing ones (e.g. polynomial regression).
- Example: consider predicting energy usage in house using area and volume (ceiling heights do not vary that much) using polynomial regression:



- This causes problems: parameter estimates cannot be interpreted sensibly and statistical tests on them are meaningless.
- However, prediction from models within the range covered by the data is not affected by multicollinearity.

Pitfalls

- *Model assumptions are violated*. See previous lectures for model assumptions (e.g. normality and independence of errors).
- *Extrapolation beyond the scope of the model*. Trends identified in data by a model do not necessarily hold beyond the range of the data. Example:



If she loves you more each and every day,
by linear regression she hated you before you met.

- *Excluding important predictors*. Can lead to models that contain misleading associations between variables. Avoid by data exploration and considering background information on data.

Pitfalls

- *Parameter interpretation*. A common misconception is that parameter estimates always measure the effect of a predictor on the response independent from other predictors (e.g. not the case in models with interactions). Another misinterpretation is that significant p-values for a parameter indicate a cause-and-effect relationship. Unless we control for all other effects, they do not.
- *Overfitting*. Recall Occam's Razor (parsimony). An extreme case of overfitting is to use as many model parameters as there are data points. In less extreme cases, including too many predictors makes interpretation difficult.
- *Power and sample size*. Small data sets can lead to poorly fitted models with large standard errors for parameter estimates. The more data, the better. General guidelines, such as a number of data points per predictor, are not possible, as they depend on the context (e.g. effect size, variability in the data).

Summary

- Model building is the process of formulating a model, fitting the model and evaluating it (e.g. check assumptions hold).
- There is a lot of flexibility in constructing linear models (this lecture).
- The key challenge is how to select the right model for the questions to be addressed (last week and this week).
- One way to approach this is to think very carefully about how the data are collected, e.g. in experiments (next lecture).