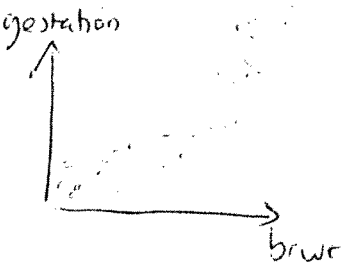
# Polynomial Regression



$$y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3}_{\text{polynomial in } x_1} + \beta_4 x_{i2} + \ldots + \varepsilon_i$$

Example




**Note** - poly-reg is (not) the same as transform models

$$y_i = \beta_0 + \beta_1 (x_{i1})^2 + \beta_2 x_{i4} + \beta_3 (\log(x_{i5}) + z_i$$

transform x's prior to model fitting

(+) a transform model suffices → you save some parameters (esp. poly-reg and avoid possible collinearities between $(x, x^2, x^3 \ldots)$

**Note** - poly-reg is (not) a nonlinear model, since $E(y)$ still linear in the parameters $\underset{\sim}{\beta}$. (and that makes life a lot easier)

Nonlinear models: $y_i = \beta_0 + \beta_1 \exp(\beta_2 x_{i1}) + \varepsilon_i$

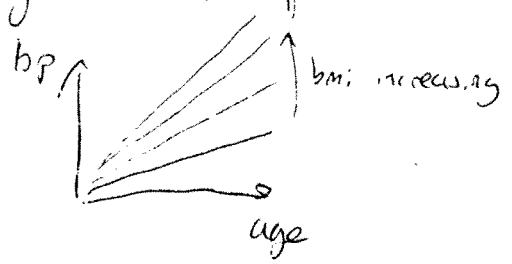
( even $y_i = \beta_1 \exp(\beta_2 x_{i1}) + \varepsilon_i$ since error additive
o/w $E(y) = \beta_1 \exp(\beta_2 x_i)$ $\log E(y_i) = \log(\beta_1) + \beta_2 x_{i1}$ ⇒ Generalized Linear Models )

## Interactions

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{1,2} x_{i,1} x_{i,2} + \beta_3 x_{i,1}^2 + \beta_4 x_{i,2}^2 + \varepsilon_i$$

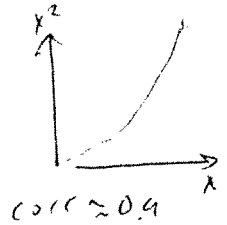
Interaction term.

Example: age * bmi ⇒ effect on blood pressure (bp)




## Problems?

$(x, x^2, \ldots x^p)$ are often highly correlated

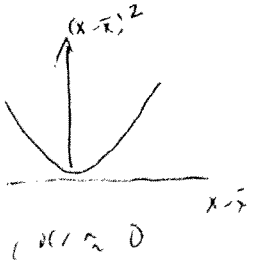



⇒ $(X'X)^{-1}$ unstable, near singular

Is there any way around this?

(centering $x$ before fitting the polynomial!

$\tilde{x}_{i,1} = (x_{i,1} - \bar{x}_1)$ $\quad \tilde{x}_{i,1}^2 = (x_{i,1} - \bar{x}_1)^2$ ...




The correlation between $\tilde{x}, \tilde{x}^2, \ldots \tilde{x}^p$ much reduced compared to corr between $x, x^2, \ldots x^p$.

Fit $y_i = \beta_0 + \beta_1 \tilde{x}_{i,1} + \beta_2 \tilde{x}_{i,1}^2 + \ldots + \varepsilon_i$

Note → $\beta$'s have slightly different interpretation now.

A word of warning

→ it's best to keep the order of the polynomial as low as possible

→ Always try transform models first.

→ Keep it reasonable → limit the order & number of interaction terms.

The idea behind centering, i.e. preprocessing x's & higher order terms to reduce the correlation can be extended to other "bases" → e.g. splines, wavelets (more later)

Practical

- To fit poly-reg → incorporate $\tilde{x}_{i1}, \tilde{x}_{i1}^2 \ldots$ in $\underline{\tilde{X}}$ (design matrix)
- $\hat{\beta} = (\underline{\tilde{X}}'\underline{\tilde{X}})^{-1}\underline{\tilde{X}}'y$ as before
- You can use the t-test, F-test, but be aware of remaining collinearities
- (Carefully examine residual plots & rely on data context to determine the order of the polynomial regression model.

## Detecting interactions

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$

→ if this is the true model then

(1) if $x_2$ is held constant & $x_1$ increases by 1

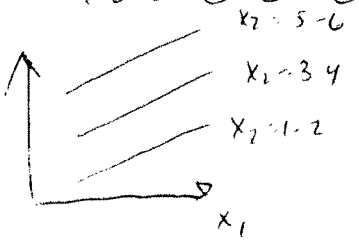
→ expect $y$ to increase by $\beta_1 + \beta_3 x_2$

(2) if $x_1$ is held constant & $x_2$ increases by 1

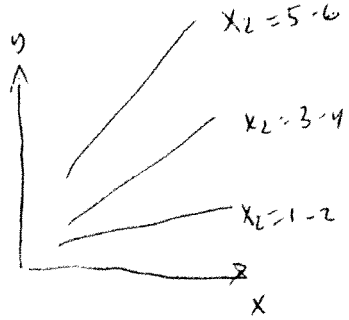
→ expect $y$ to increase by $\beta_1 + \beta_2 x_1$

→ so effect of $x_1$ on $y$ depends on $x_2$ & v.v.

Use conditional effects plots to detect

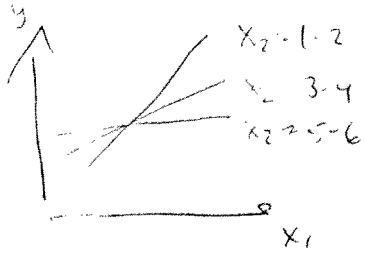
- 

→ "Additive model", no interaction

- 

⇒ "Reinforcement model" → effect of $x_1$ on $y$ enhanced by $x_2$ & v.v.

Ex: Salary ~ yr employment * yr of schooling

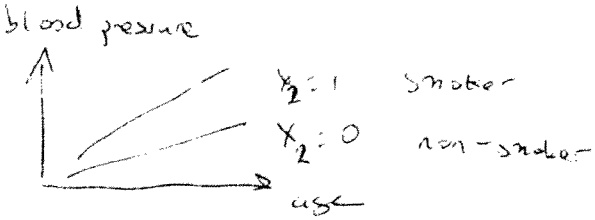
- 

⇒ "Interference model" → effect of $x_1$ on $y$ is counteracted by $x_2$ & v.v.

# Qualitative variables

E.g. smoker, non-smoker

$X_2 = 0, 1$




It's easy to incorporate qualitative variables in regression using <u>indicator variables</u>

$X_2 = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{Bmatrix}$

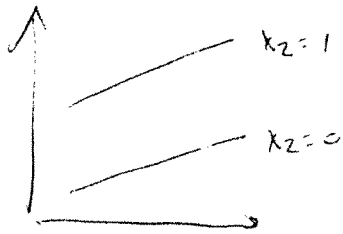
$\Rightarrow$ add to $\underline{X}$ (the design matrix)

$\Downarrow$

an additive model

$E(y) = \underbrace{\beta_0 + \beta_1 x_1}_{\beta_0^{NS}}$ if $x_2 = 0$

$E(y) = \underbrace{\beta_0 + \beta_2 + \beta_1 x_1}_{\beta_0^{S}}$ if $x_2 = 1$




"Additive model"

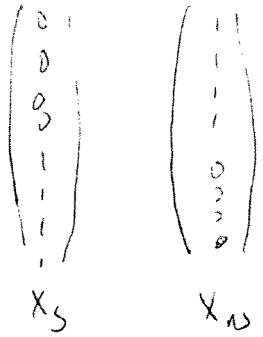
So including an indicator in $\underline{X}$ allows for a different intercept - but what if being a smoker / non-smoker changes the relationship between age and blood pressure?

$\Rightarrow$ Interactions

$\Rightarrow$ Create a new $x_1$ variable $x_1 \cdot$ indicator variable and add to $\underline{X}$.

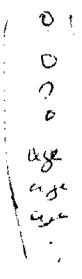
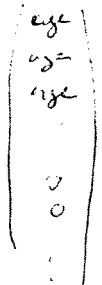
# A few alternatives

(1) 2 separate regression models

- use 2 indicators  , for smokers & nonsmokers

$X_S$ $X_{NS}$

- create 2 new age variables

$age * X_S =$  , $age * X_{NS} =$ 

- add the new variables to the design matrix $X$

- run regression → get $E(y) = \beta_0^S + \beta_1^S \cdot age$ if smoker

$E(y) = \beta_0^{NS} + \beta_1^{NS} \cdot age$ if nonsmoker

(But) we usually want to make inference about the contrasts between smokers & non-smokers . . .

(2) One indicator

- baseline model

$X_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$ , $age \cdot X_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ age \\ \vdots \\ age \end{pmatrix}$

→ add to $\underline{X}$

- run regression & get

$$\begin{cases} E(y) = \beta_0 + \beta_1 \cdot age & \text{if } X_2 = 0 \\ E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot age & \text{if } X_2 = 1 \end{cases}$$

- Now we can test

a) $\beta_3 = 0$ → additive model

b) $\beta_2 = \beta_3 = 0$ → no effect due to smoking.

- Note, this formulation implicitly treats non-smokers as the baseline population for comparisons.

(3) Contrasts -- $X_2 =$ $\begin{pmatrix} -1 \\ -1 \\ -1 \\ \vdots \\ +1 \\ +1 \\ +1 \\ +1 \end{pmatrix}$ instead of 0 & 1 } ? new variables

- $X_2 \cdot age$

- add the new variables to $\overline{X}$

& run regression

$$\Rightarrow \begin{cases} E(y) = (\beta_0 - \beta_2) + (\beta_1 - \beta_3) \cdot age & \text{: nonsmoker} \\ E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot age & \text{: smoker} \end{cases}$$

Now, can test $\beta_3 = 0$ (additive model) or

$\beta_2 = \beta_3 = 0$ (no effect due to smoking)

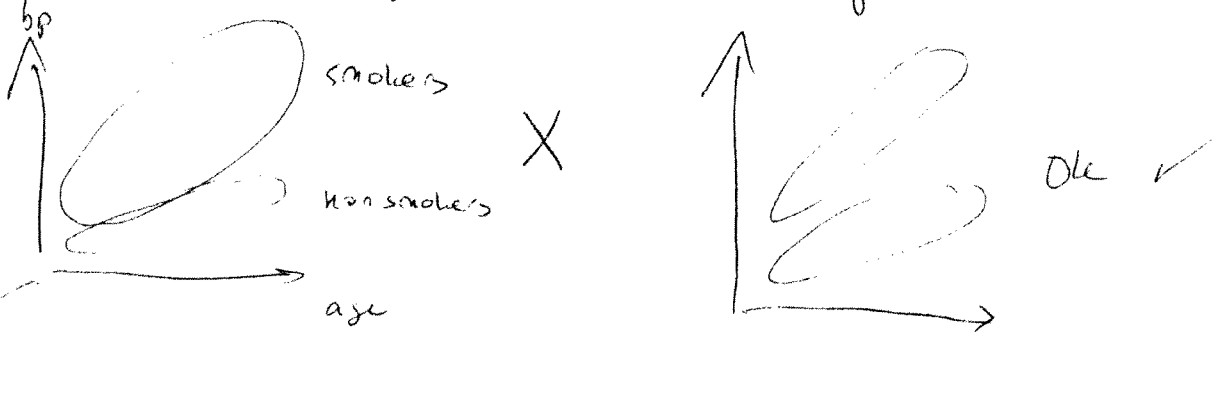
and the baseline is now the average of both populations (smokers & nonsmokers)

A few words of warning:

① It is common practice to always include the main terms if interactions are included — makes the interpretation of the interactions easier to deal with

② The basic assumptions still need to hold — e.g. constant variance.

(A) The variance is different for smokers & non-smokers, it is a violation of the basic assumptions.




What to do here?

a) weighted Least Squares

b) 2 regression models