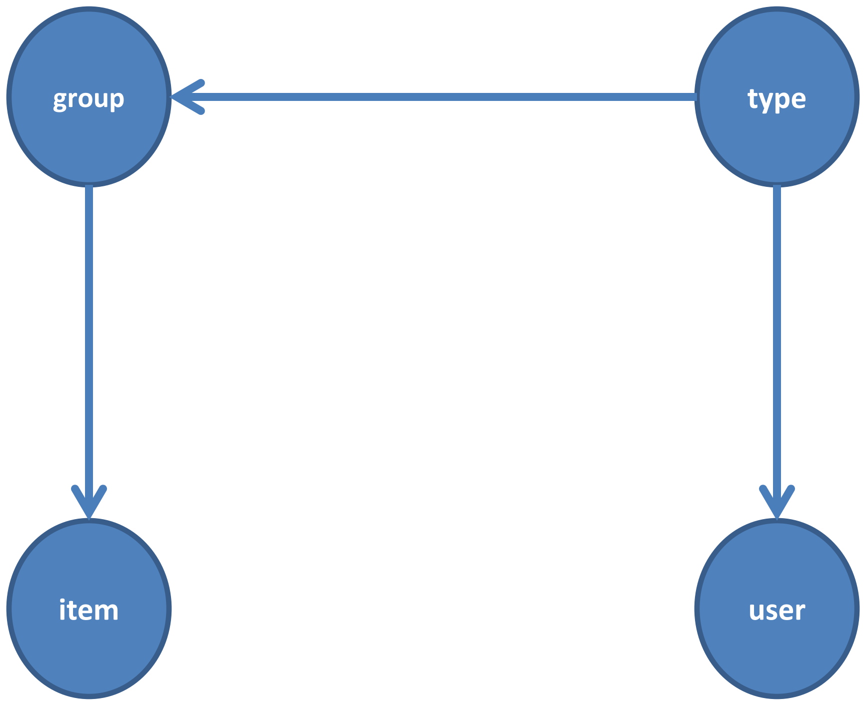
Regression analysis

Nisheeth

Recommender system example



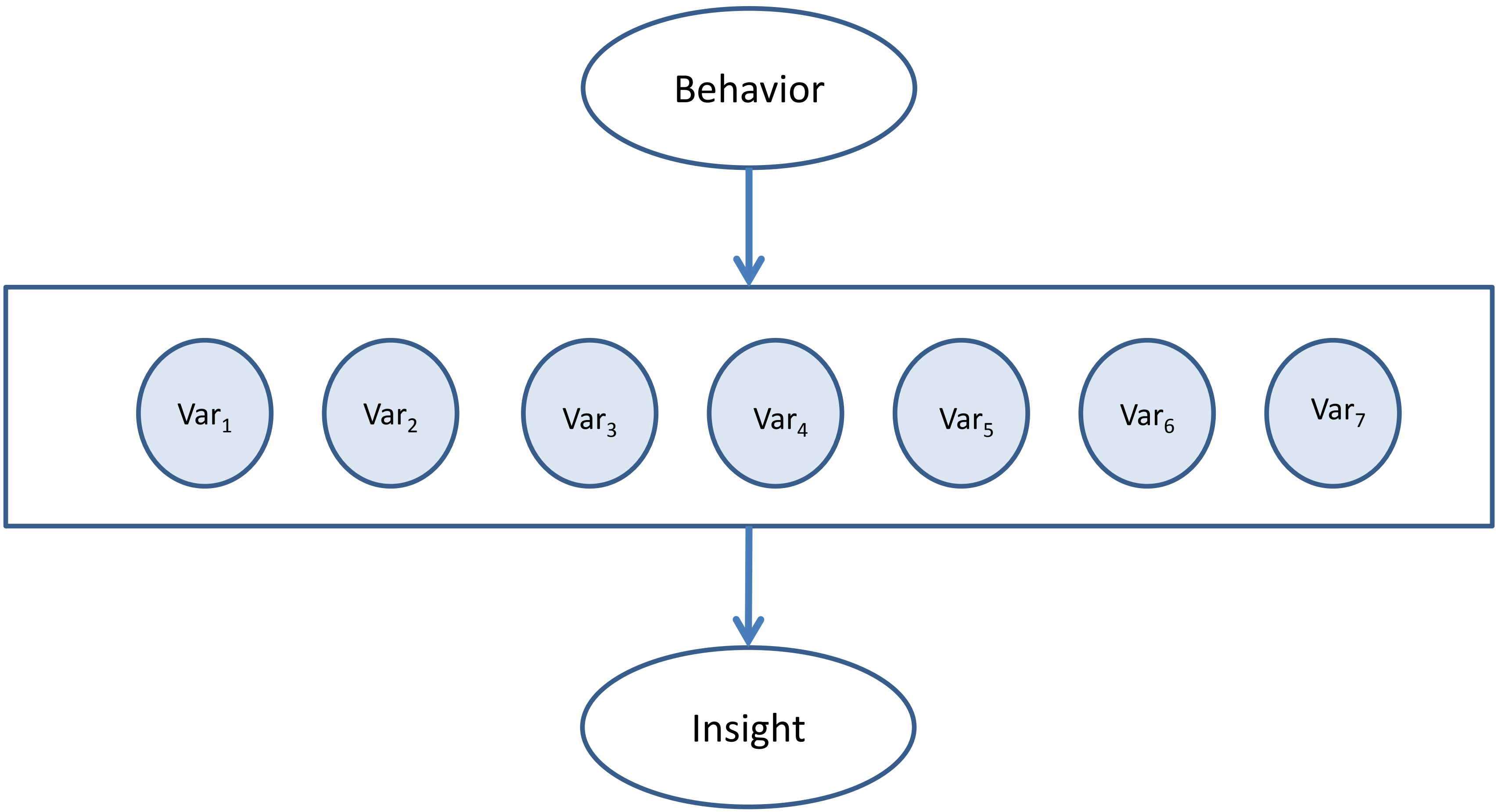
Try now

$$p(i|u) = \sum_{g,t} p(i|g)p(u|t)\overbrace{p(g|t)}^{\text{Personalization}}\underbrace{p(t)}_{\text{Context inference}}$$

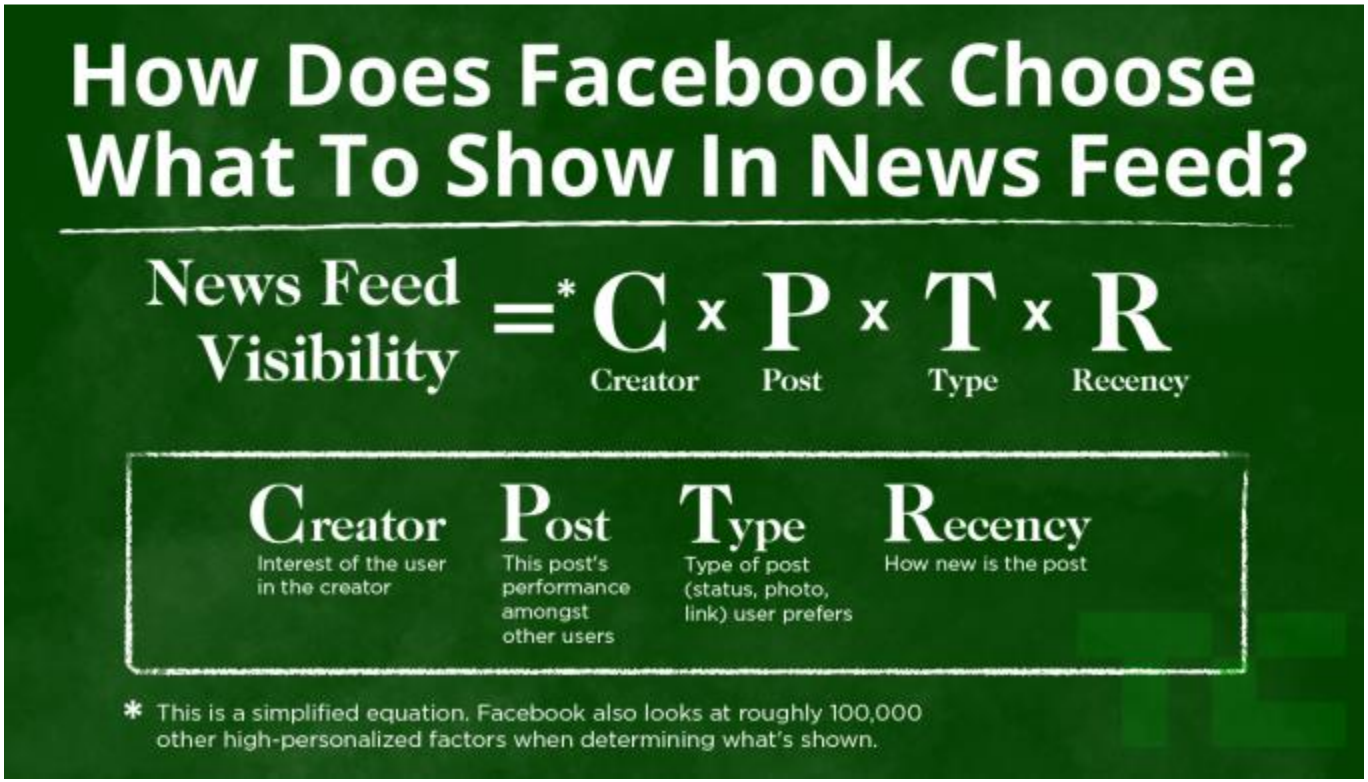
Problem


Too
Much
Information
relationships
sometimes
random
people
useless
hope
things
keep
friends
fantasy
world
knowledge
developments
Either
posts
good
Technology
truth
cover
see
continues
probably
football
thoughts
decisions
around
baseball
conversation
interesting
really
playoffs
bullshit
making
know
ahead
now
Politics
Music
Sports
hockey
season
lol
like
facts
current
First

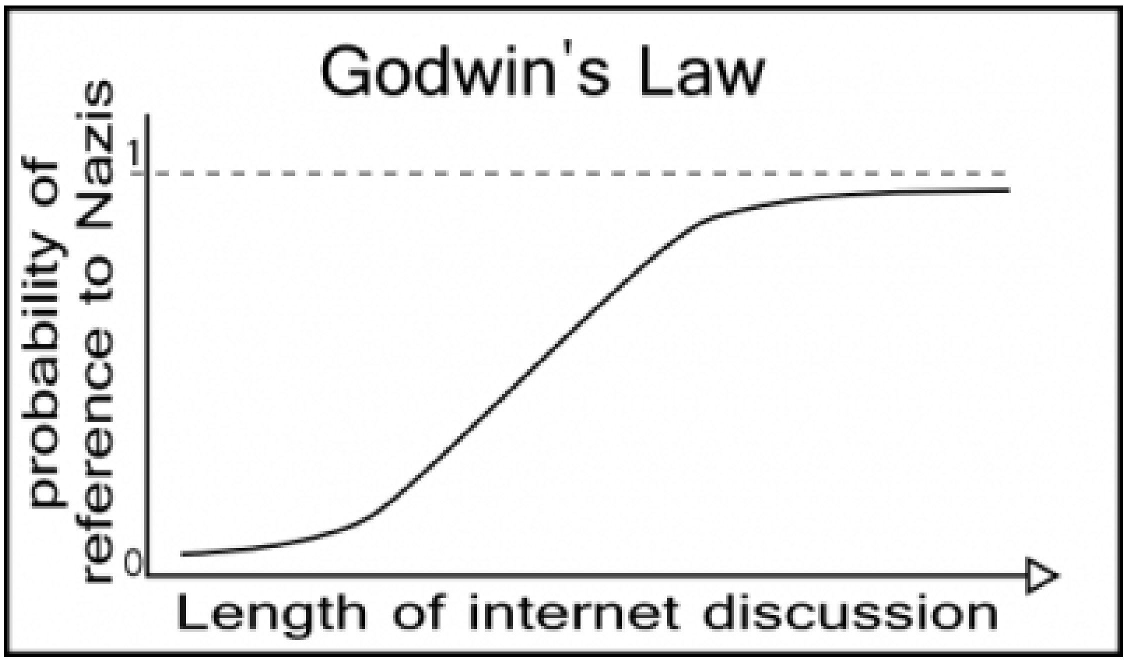
Typical human-computer interface



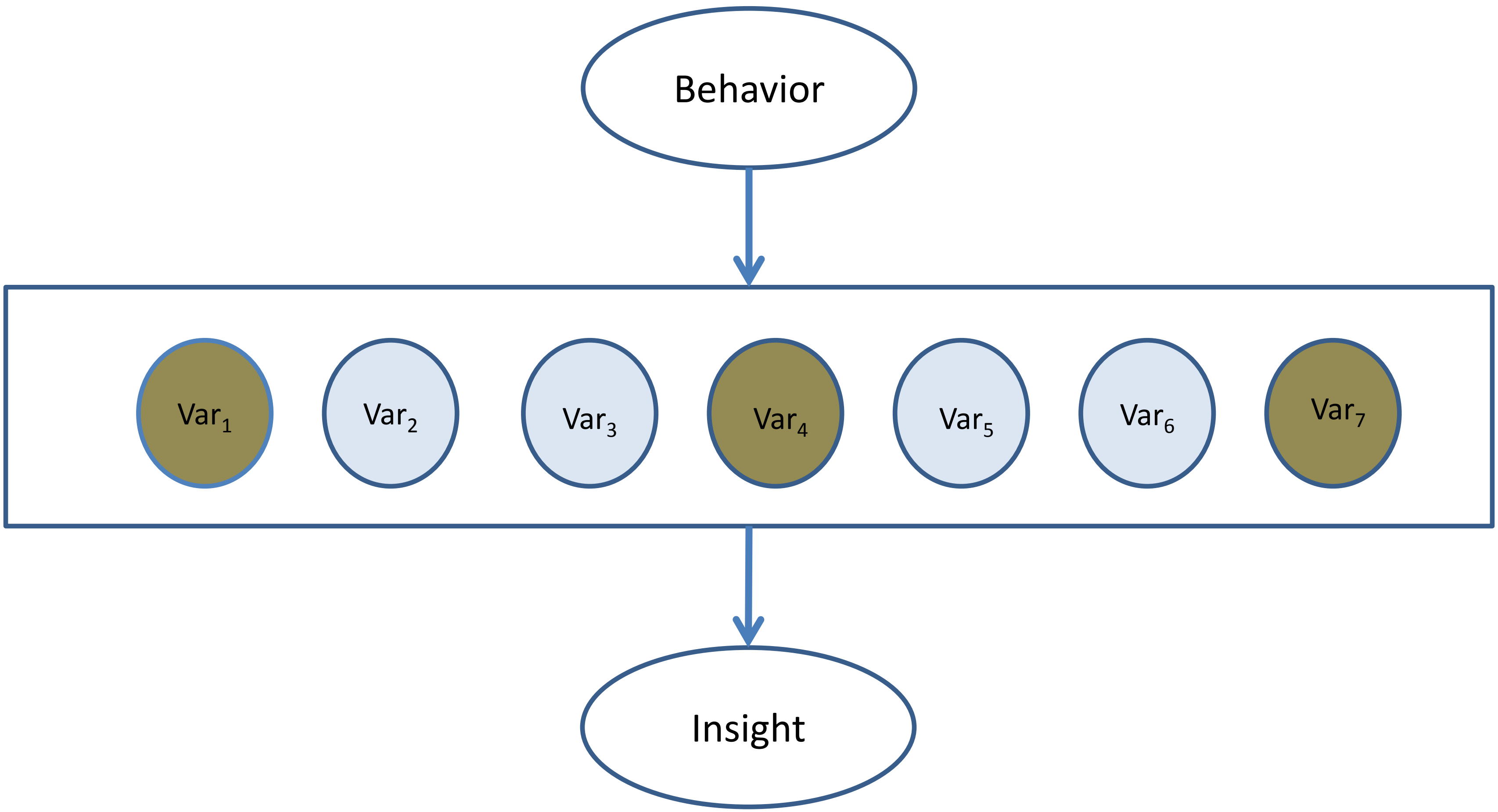
Stored variables are a function of business requirements



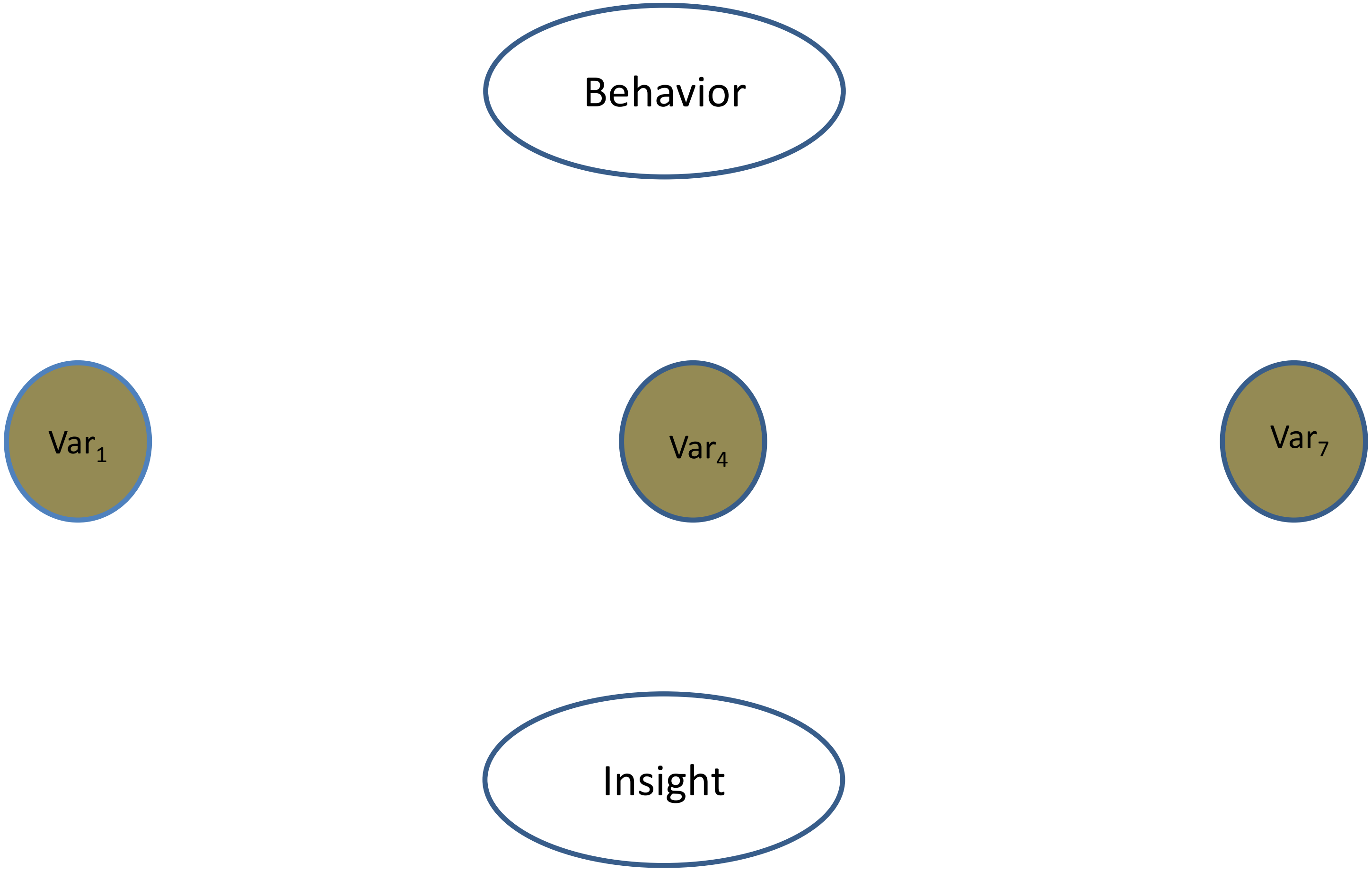
Important scientific discoveries made tangentially



What we want - I

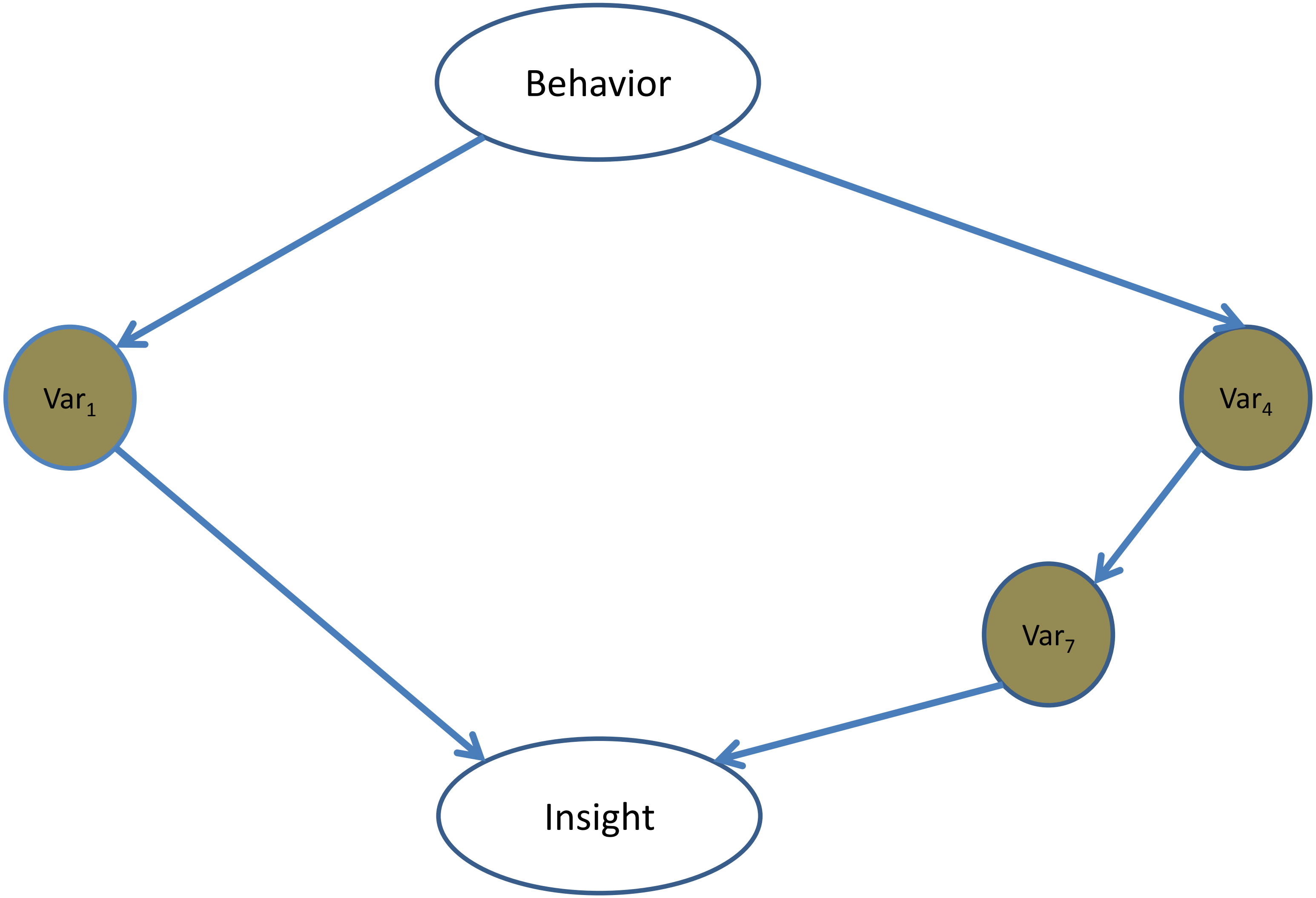


We'd get



Machine learners call this the 'feature selection' problem. We will focus on one particular way of solving it - regression analysis

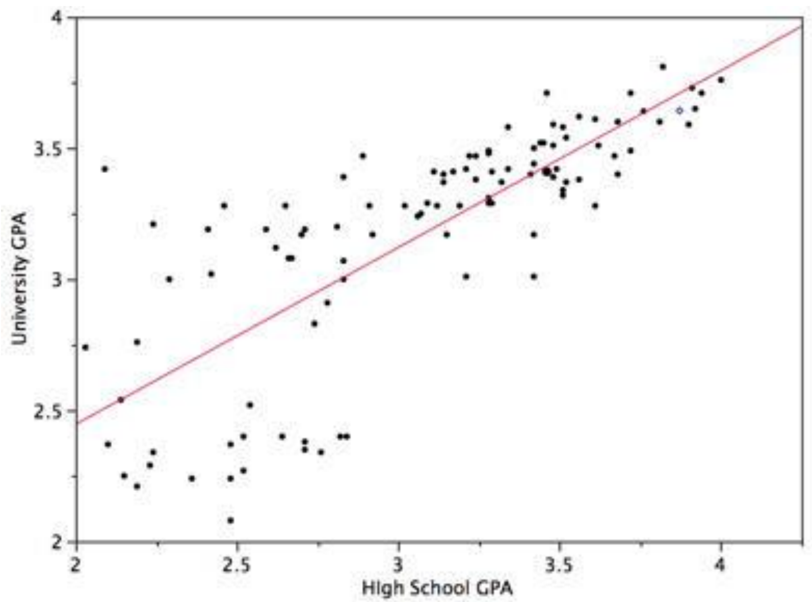
What we want - II



This is causality inference

Regression model

- Regression model estimates the nature of relationship between the independent and dependent variables.
  - Change in dependent variables that results from changes in independent variables, i.e. size of the relationship.
  - Strength of the relationship.
  - Statistical significance of the relationship.



Bivariate and multivariate models

Bivariate or simple regression model

(Education) $x \longrightarrow y$ (Income)

Multivariate or multiple regression model

(Education) x_1

(Sex) x_2

(Experience) x_3

(Age) x_4

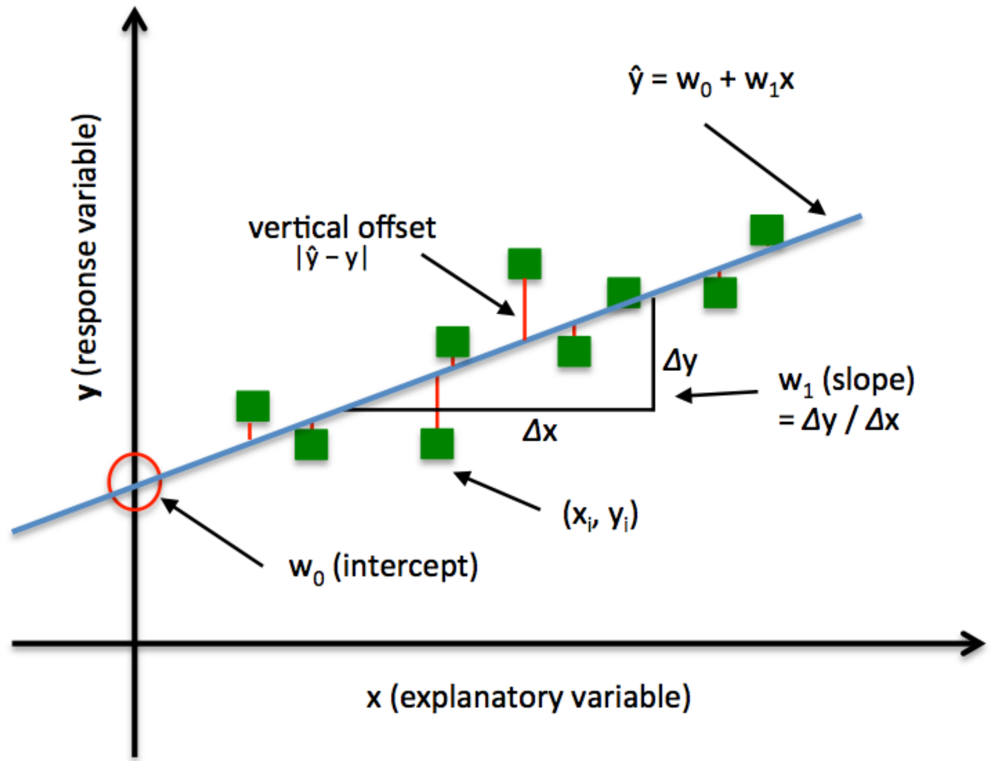
y (Income)

Bivariate or simple linear regression

- *x* is the independent variable
- *y* is the dependent variable
- The regression model is

$$y = w_0 + w_1 x + \varepsilon$$

- Two parameters to estimate – the slope of the line w_1 and the *y*-intercept w_0
- ε is the unexplained, random, or error component.

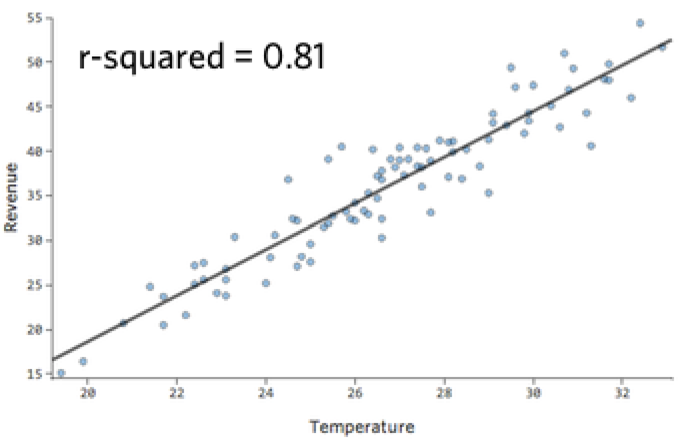


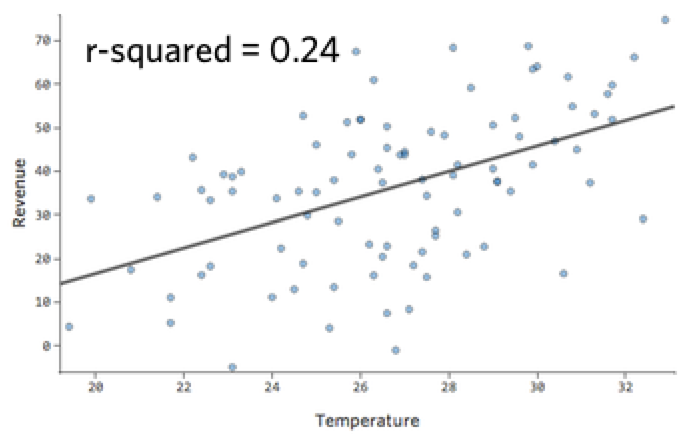
Fitting the regression model

- Any choice of **w** gives us predictions for the dependent variable f_i for each x_i
- Residual $e_i = y_i - f_i$
- Good fit = minimize $\sum_i e_i^2$
- Easy to derive estimators for coefficients using basic calculus
- $\min_{\boldsymbol{w}} \sum_{\boldsymbol{i}} (y_i - w_0 - w_1 x_i)^2$
  - $w_1 = \frac{Cov(x,y)}{Var(x)}$
  - $w_0 = \bar{y} - w_1 \bar{x}$

Assessing goodness of fit

- Sanity check E[e] = 0 → E[f] = E[y]
- $SST = \sum_i (y_i - \bar{y})^2$
- $SSR = \sum_i (y_i - f_i)^2$
- $R^2 = 1 - \frac{SSR}{SST}$

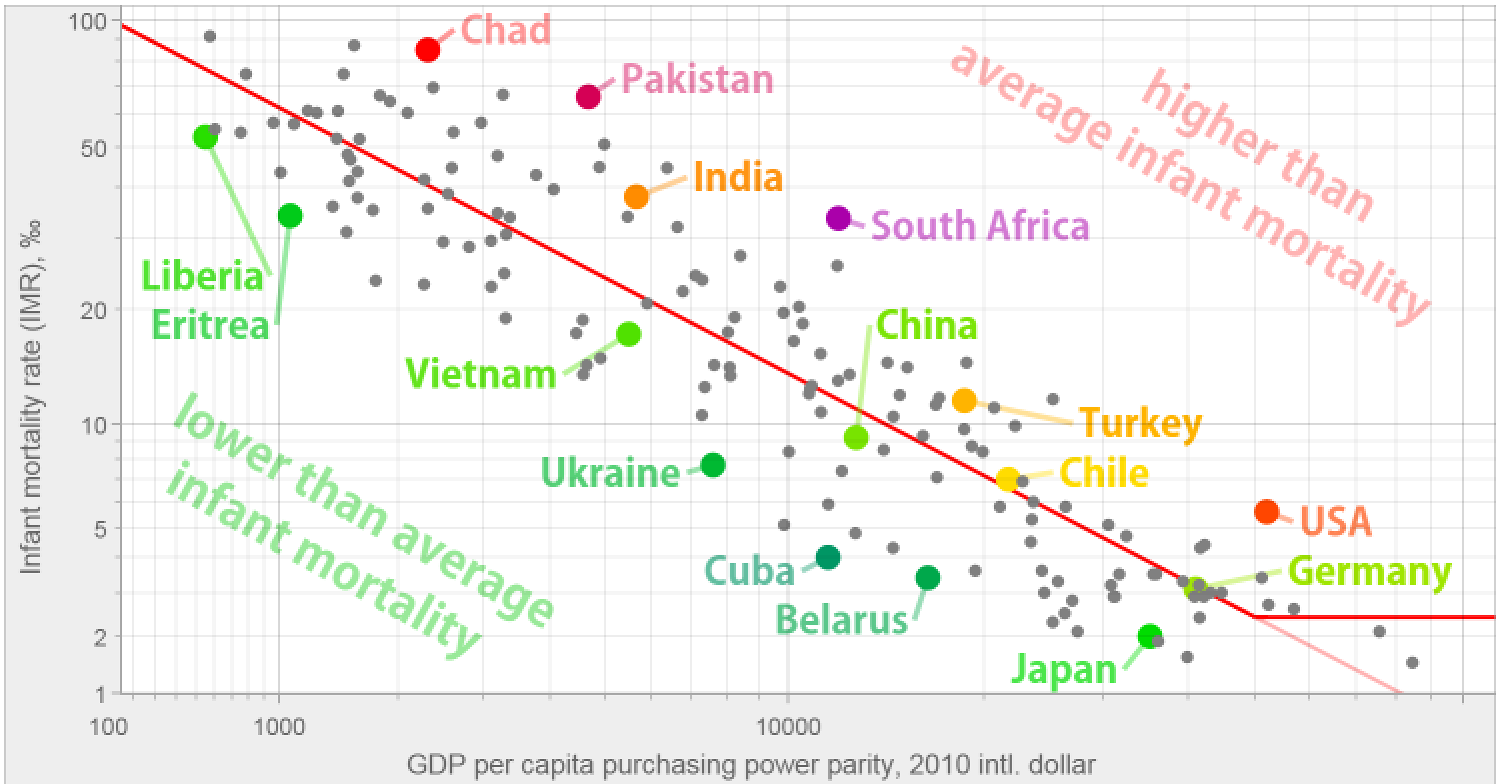




Uses of vanilla regression

- Amount of change in a dependent variable that results from changes in the independent variable(s) –
- Attempt to determine causes of phenomena.
- Prediction and forecasting
- Support or negate theoretical model.
- Modify and improve theoretical models and explanations of phenomena.

Used to tell stories that make a big difference

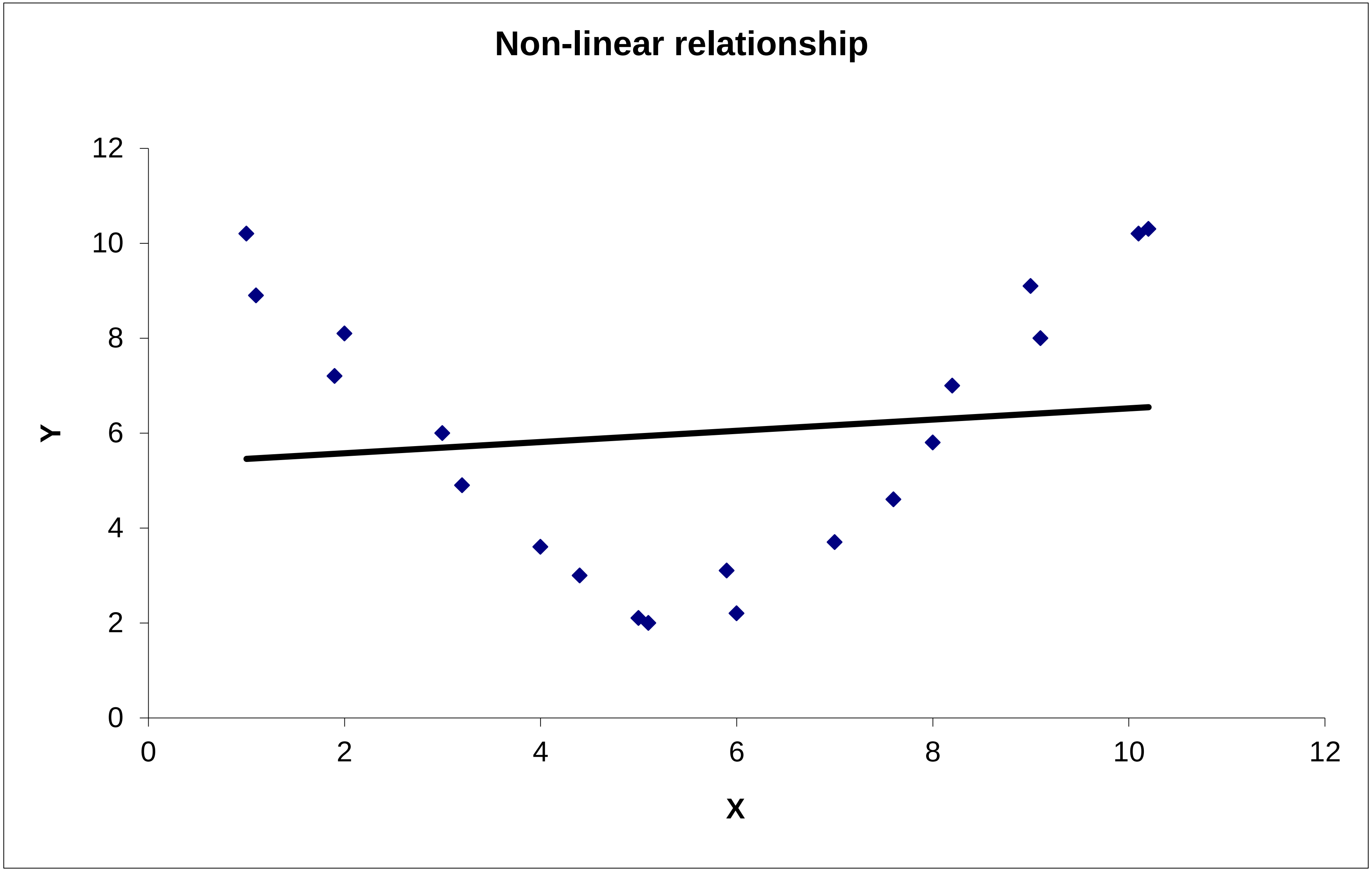


Multiple regression

- Everything works the same way as in simple regression
- $y = w_0 + w_1 x_1 + w_2 x_2 + \ \ldots + \ w_n \, x_n$
- Estimated as in the univariate case by minimizing

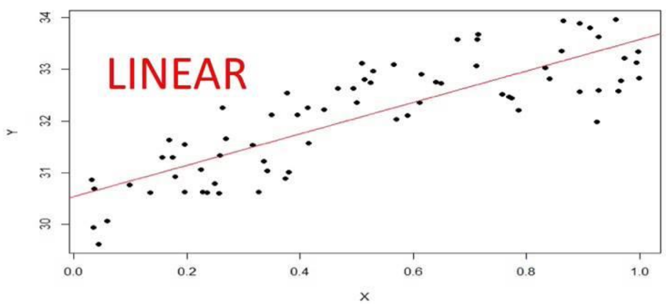
$$\sum_{i=1}^{n} \left(y_i - \ w_0 \ - \sum_{j=1}^{p} w_j x_{ij} \right)^2$$

- Each independent variable affects the dependent variable linearly in isolation
- Can also use modifications of the same variable, e.g. x^2 in place of a new variable



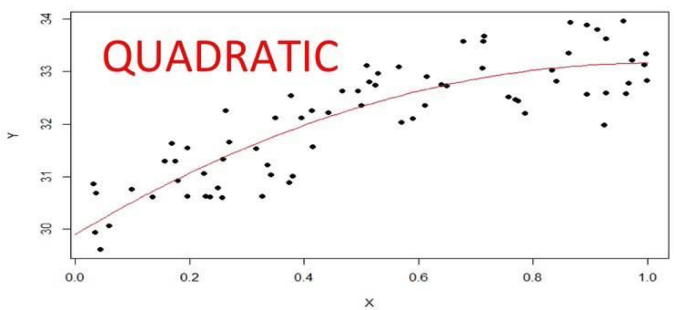
Correlation = +0.12.

All multiple linear regressions



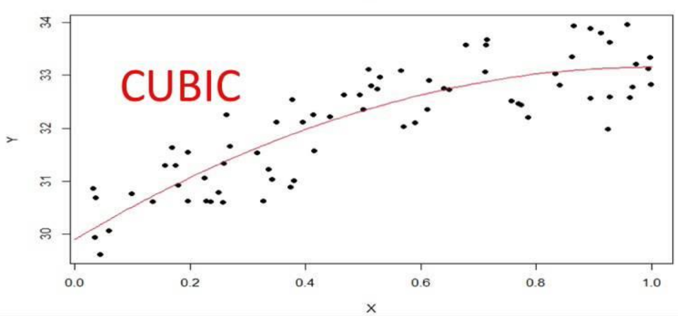
Multiple R-squared: 0.7044

$Y = 30.53 + 3.05*X$



Multiple R-squared: 0.7559

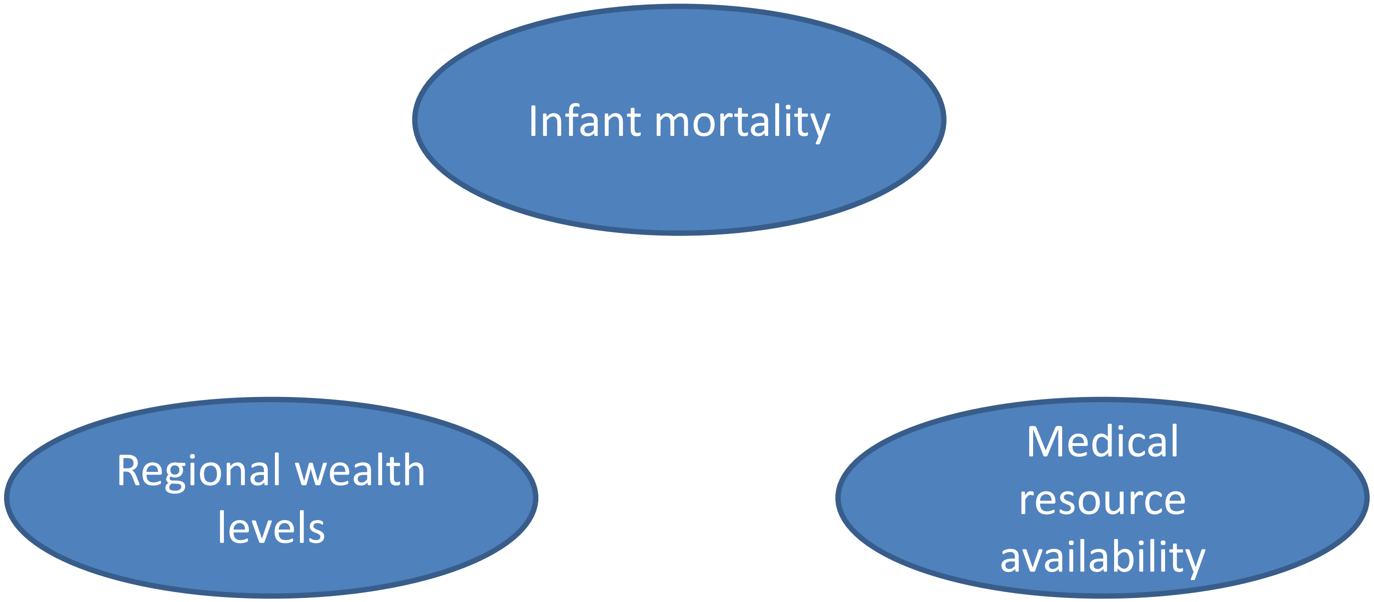
$Y = 29.90 + 6.48*X - 3.22*X^2$



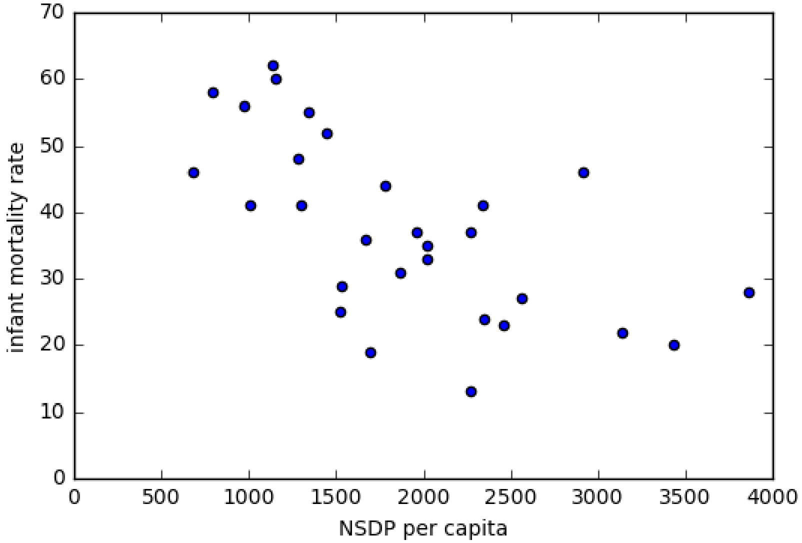
Multiple R-squared: 0.7623

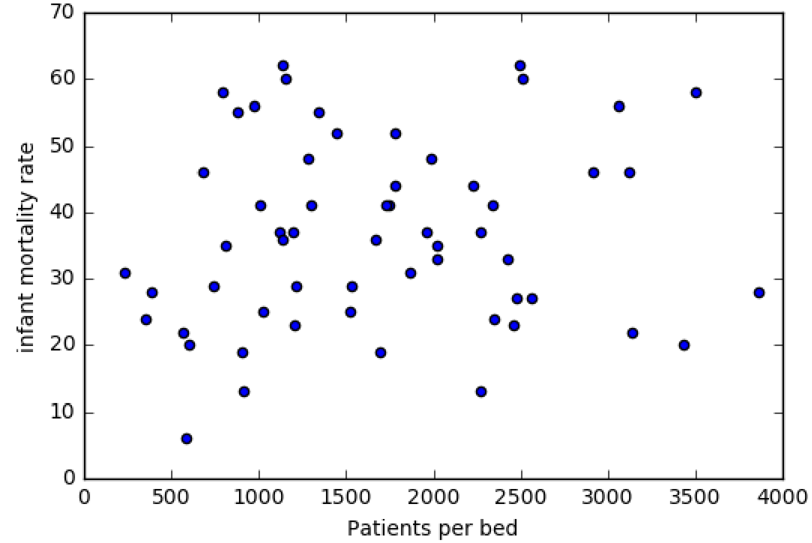
$Y = 30.17 + 3.61*X + 3.71*X^2 - 4.48*X^3$

Example


Infant mortality
Regional wealth levels
Medical resource availability

Predictable relationship





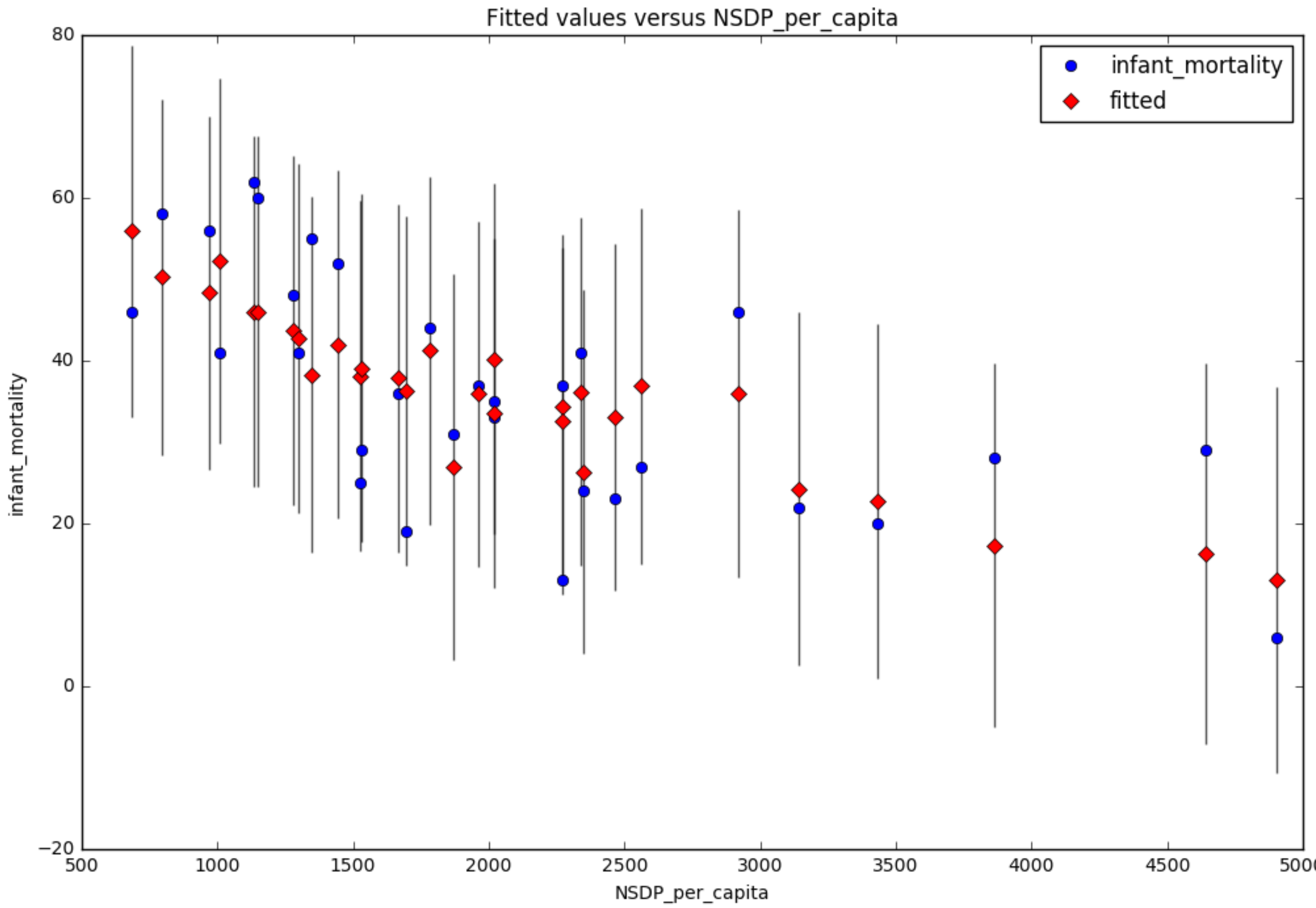
Regression result

Dep. Variable:	infant_mortality	**R-squared:**	0.524
Model:	OLS	**Adj. R-squared:**	0.490
Method:	Least Squares	**F-statistic:**	15.43
Date:	Wed, 11 Jan 2017	**Prob (F-statistic):**	3.04e-05
Time:	21:50:09	**Log-Likelihood:**	-114.39
No. Observations:	31	**AIC:**	234.8
Df Residuals:	28	**BIC:**	239.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	5.5161	23.306	0.237	0.815	-42.224 53.257
NSDP_per_capita	-0.0064	0.002	-2.936	0.007	-0.011 -0.002
np.log(patients_per_bed)	6.1056	2.825	2.162	0.039	0.320 11.892

Omnibus:	1.664	**Durbin-Watson:**	2.644
Prob(Omnibus):	0.435	**Jarque-Bera (JB):**	1.086
Skew:	-0.116	**Prob(JB):**	0.581
Kurtosis:	2.113	**Cond. No.**	2.97e+04

Looks decent too



IMR = -0.0064*(NSDP per capita) + 6.11*(log(patients per bed)) + 5.51

Interpreting the coefficients

Dep. Variable:	infant_mortality	**R-squared:**	0.524
Model:	OLS	**Adj. R-squared:**	0.490
Method:	Least Squares	**F-statistic:**	15.43
Date:	Thu, 12 Jan 2017	**Prob (F-statistic):**	3.04e-05
Time:	10:50:49	**Log-Likelihood:**	-114.39
No. Observations:	31	**AIC:**	234.8
Df Residuals:	28	**BIC:**	239.1
Df Model:	2		
Covariance Type:	nonrobust		

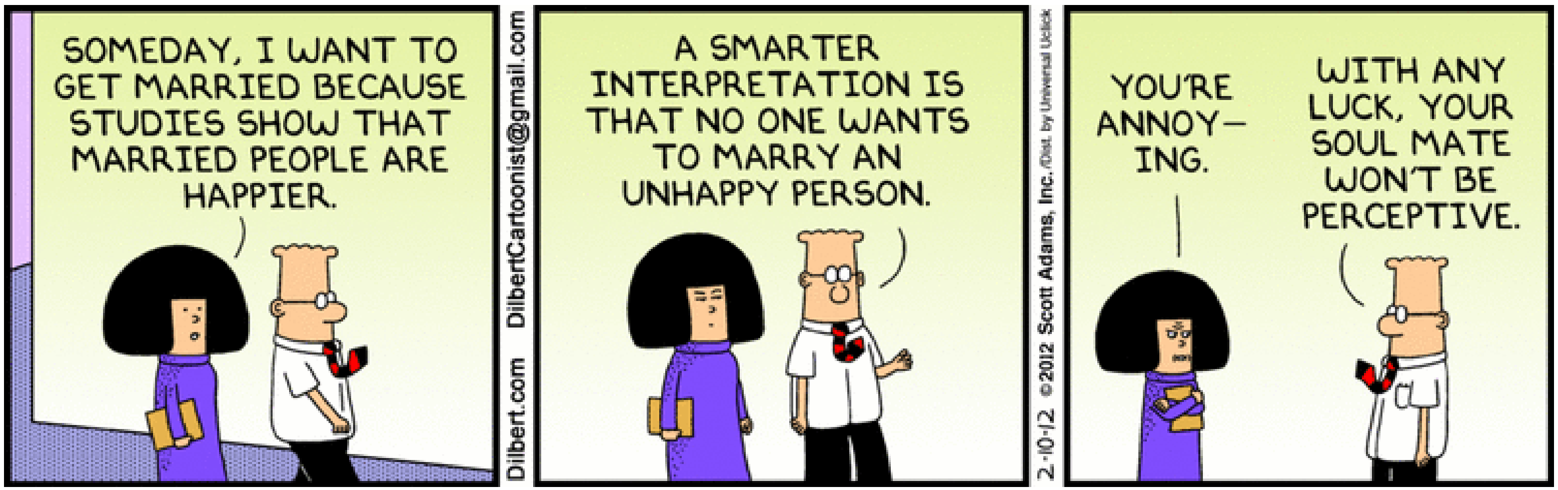
	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	36.2581	1.831	19.801	0.000	32.507 40.009
st_NSDP	-6.5784	2.241	-2.936	0.007	-11.169 -1.988
st_ppb	4.8437	2.241	2.162	0.039	0.254 9.434

Omnibus:	1.664	**Durbin-Watson:**	2.644
Prob(Omnibus):	0.435	**Jarque-Bera (JB):**	1.086
Skew:	-0.116	**Prob(JB):**	0.581
Kurtosis:	2.113	**Cond. No.**	1.93

Technical caveats

- Have to look at adjusted R^2 to assess model fit
  - R2 can never decrease by adding an extra variable
  - Have to deflate it by number of variables used for fair comparison
  - $R_{adj}^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$
- Have to watch out for problems
  - Omitted variable bias
  - Multicollinearity
  - Dummy variable trap
  - Outliers

Omitted variable bias



We omit a variable from the analysis that is

- Correlated with at least one of the independent variables and
- Determinative for the response variable mechanistically

Multicollinearity

- When two of the predictors are highly correlated
- Parameter estimation becomes unstable
- Results become suspect
- Rule of thumb: correlations higher than 0.7 between two variables → leave the less interesting one out of the analysis

Dummy variable trap

- How to handle categorical data?
- Create n dummy variables for an n category variable
- Never use all n in the analysis, leave one out
- Why? Multicollinearity

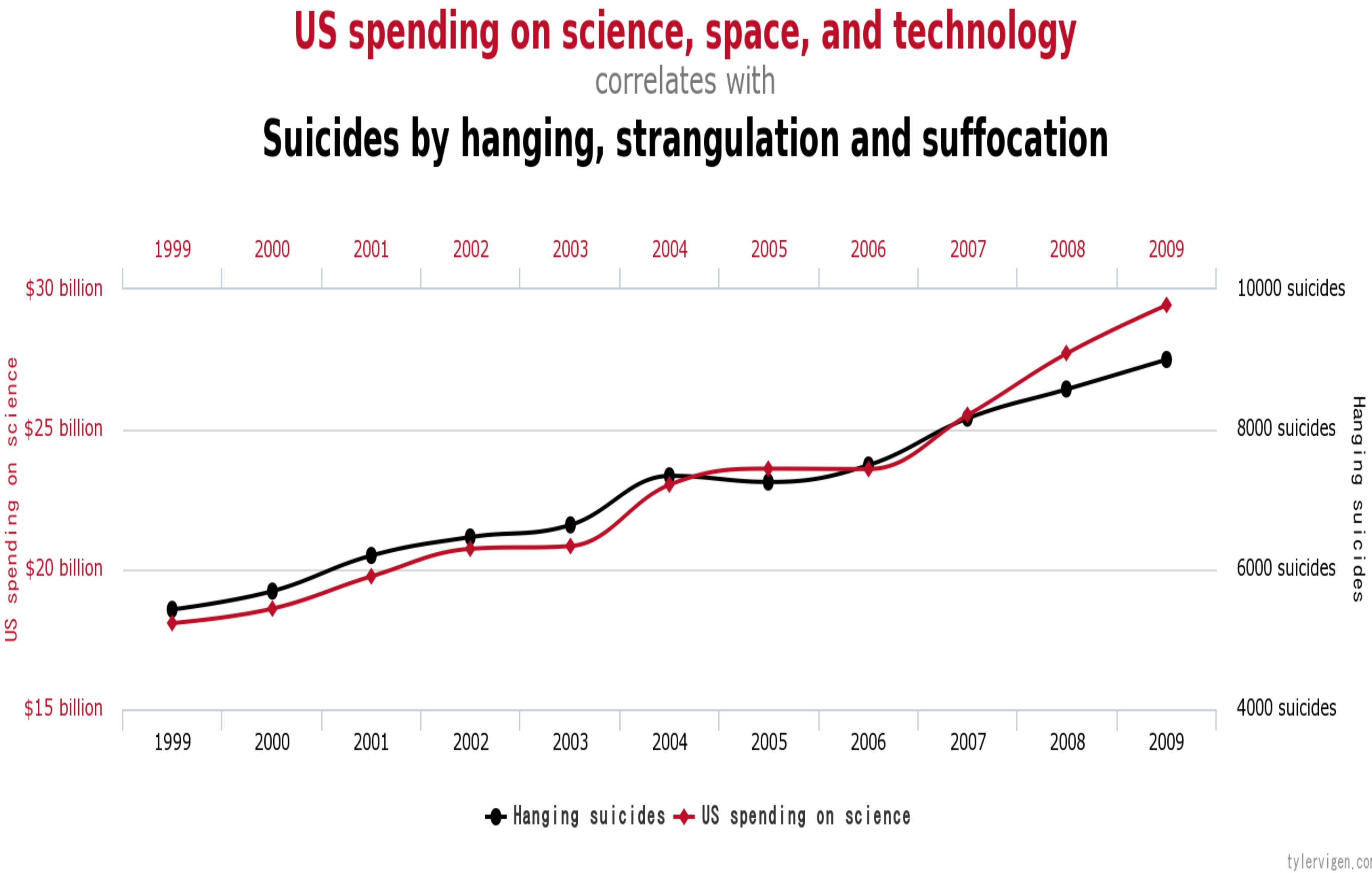
Outliers

- Rare, extreme values distort OLS fits.
  - Could be an error.
  - Could be a very important observation.
- Outlier: more than 3 standard deviations from the mean.
- Can discard, or use robust regression methods
- Caveat emptor

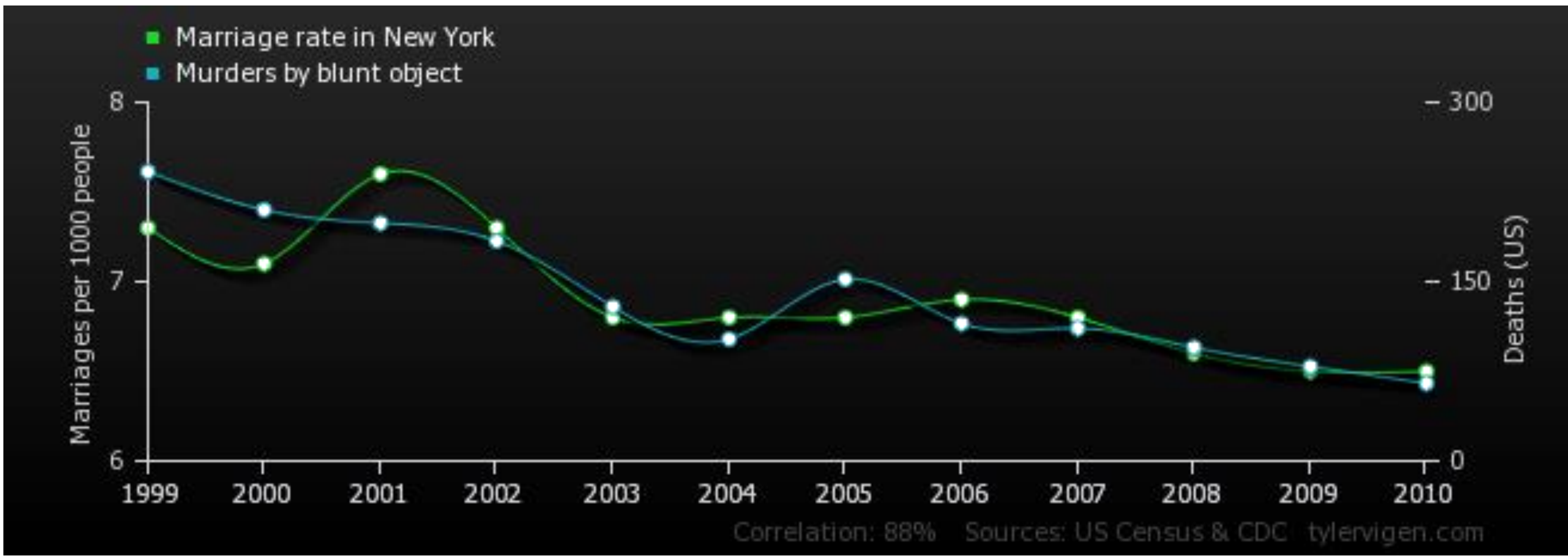
Pragmatic caveats

- Spurious correlations
- The kitchen sink problem
- Regularization

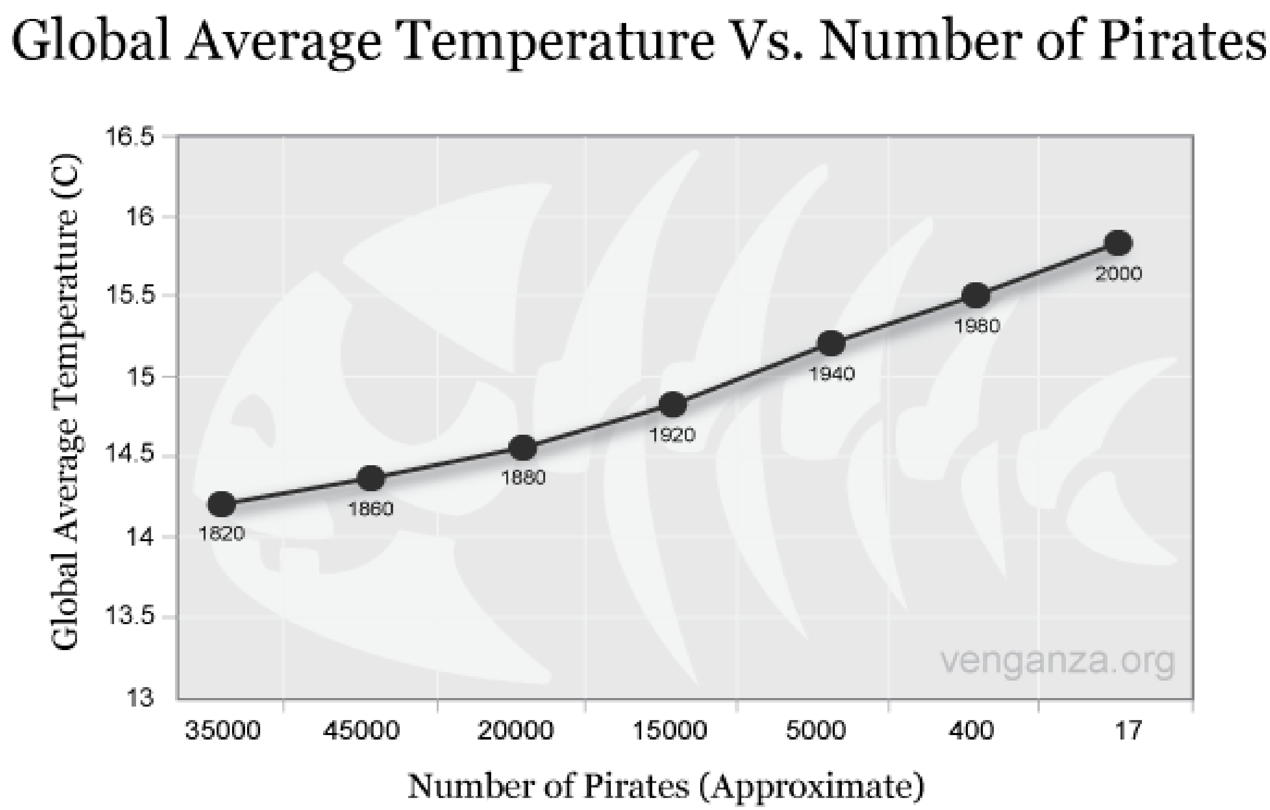
Spurious correlations



Spurious correlations



Spurious correlations

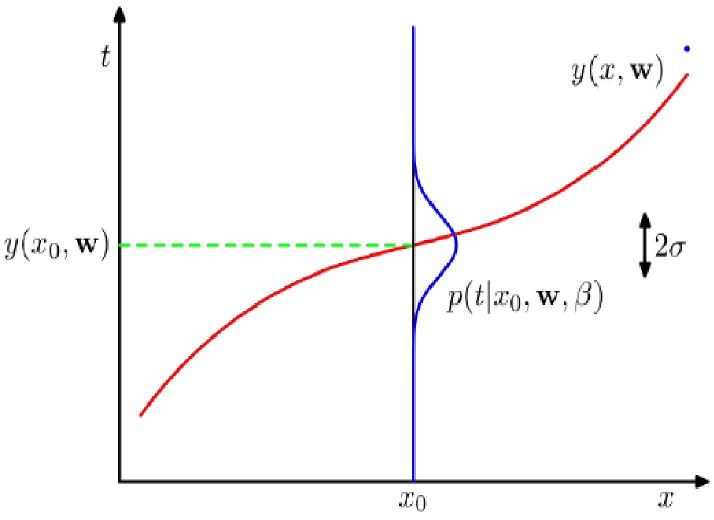


Regularization in regression models

- Regularization = trying to keep your model simple
- Do this by adding a regularization term to the regression objective function, i.e. SSE + λR
- Three basic forms in regression
  - Subset selection: R = $|\mathbf{w}|_0 = \sum_i^p I(w_i)$
  - Lasso regression: R = $|\mathbf{w}|_1 = \sum_i^p |w_i|$
  - Ridge regression: R = $|\mathbf{w}|_2 = \sum_i^p w_i^2$
- Larger λ → simpler model, with fewer non-zero coefficients

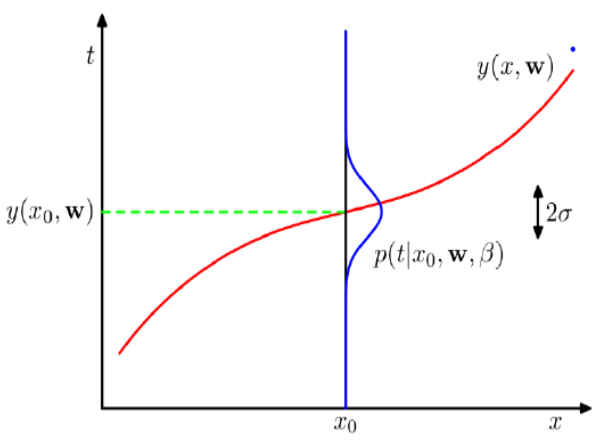
Probabilistic intuition

- Assume that $p(y|x, w, \sigma^2) = N(y|f(x, w), \sigma^2)$
- Bayes inversion would gives us $p(w|x, y, \sigma^2)$
  - If we have knowledge about the prior on w
  - Assume the prior is $N\left(w|0, \alpha = \frac{1}{\sigma_p{}^2}\right)$
- Find w by maximizing the posterior probability



Probabilistic intuition

- Equivalent to minimizing the negative log posterior
    - $\min\{-\log p(y|x, w, \sigma^2) - \log p(w|\alpha)\}$
    - $\min\left\{\frac{1}{2\sigma^2}\sum_i (y_i - w_i x_i)^2 + \frac{\alpha}{2}\sum_j w_j^2\right\}$
- You could do a full Bayesian regression instead
- How? Why?



What we get