The Language of Models

Author: Nicholas G Reich

*This material is part of the **statsTeachR** project*

Today's topics

- The language of models
- Model formulas and coefficients

Example: predicting respiratory disease severity ("lung" dataset)

Reading: Kaplan, Chapters 6 and 7.

Figure acknowledgements to Hadley Wickham.

Watch the first five minutes of Hadley's UseR! 2016 talk

" ... every model has to make assumptions, and a model by its very nature cannot question those assumptions...

models can never fundamentally surprise you because they cannot question their own assumptions."

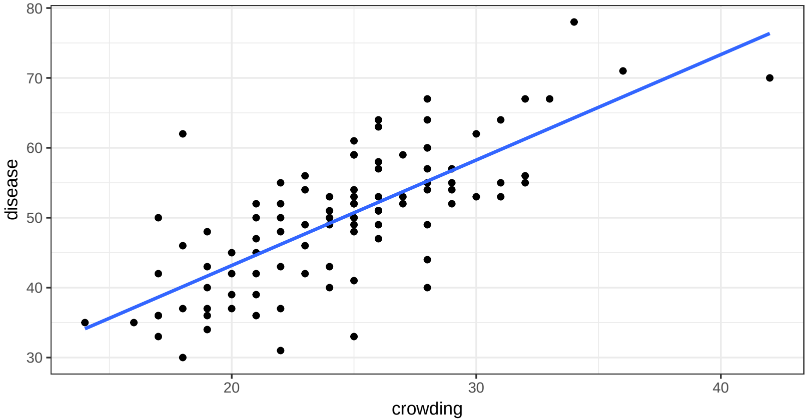
Lung Data Example

99 observations on patients who have sought treatment for the relief of respiratory disease symptoms.
The variables are:

- `disease` measure of disease severity (larger values indicates more serious condition).
- `education` highest grade completed
- `crowding` measure of crowding of living quarters (larger values indicate more crowding)
- `airqual` measure of air quality at place of residence (larger number indicates poorer quality)
- `nutrition` nutritional status (larger number indicates better nutrition)
- `smoking` smoking status (1 if smoker, 0 if non-smoker)

Lung Data Example: terms defined

```
dat <- read.table("lungc.txt", header=TRUE)
ggplot(dat, aes(crowding, disease)) + geom_point() +
    geom_smooth(method="lm", se=FALSE)
```

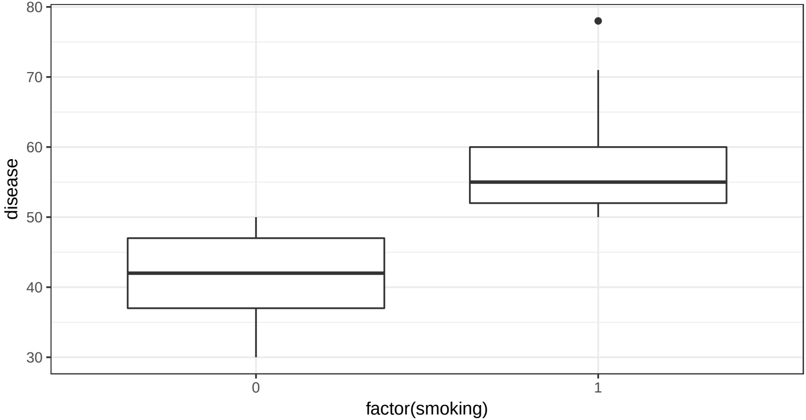


Identify: response variable, explanatory variable, model value, residual.

Lung Data Example: terms defined

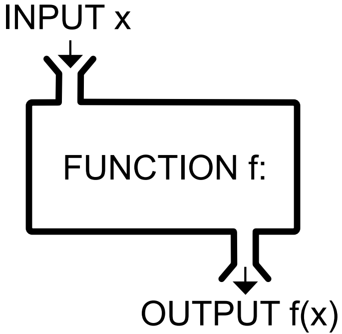
What are the "model values" for the model implied by this figure?

```r
ggplot(dat, aes(factor(smoking), disease)) + geom_boxplot()
```



Models are functions

Definition: "a **function** is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output".[1]

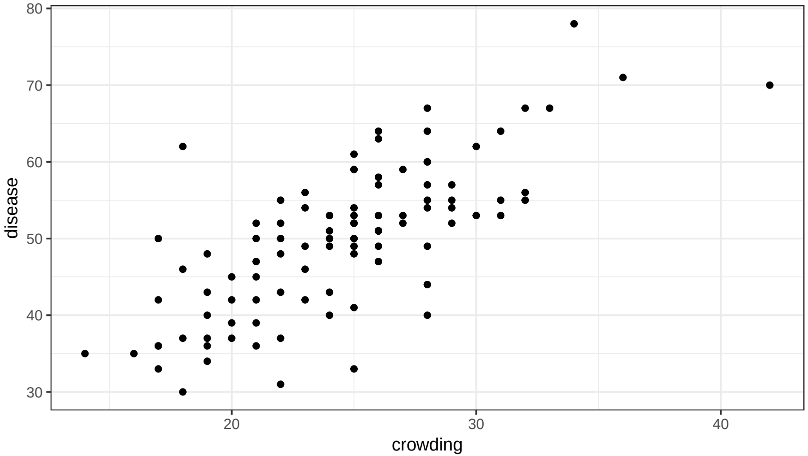


In statistical models, inputs are explanatory variables and outputs are "typical" or "expected" values of response variables.

[1] Wikipedia, https://en.wikipedia.org/wiki/Function_(mathematics)

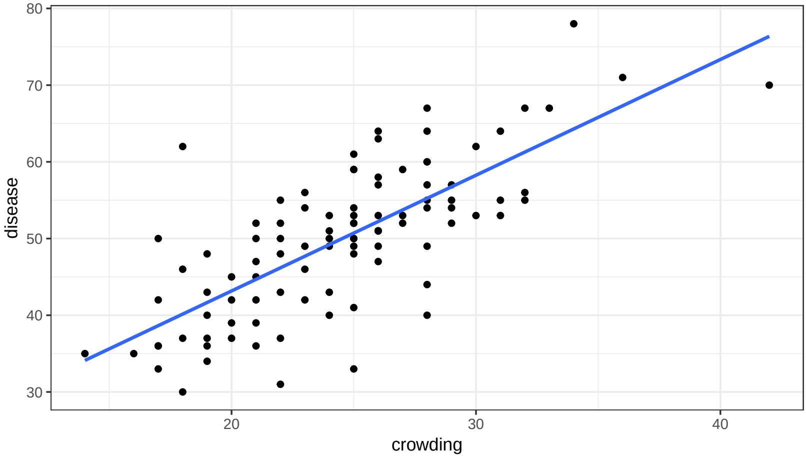
Characterize the relationship

Broadly speaking, what kind of model could describe the relationship between crowding and disease? How well would you say this model fits the data? Or predicts new observations?



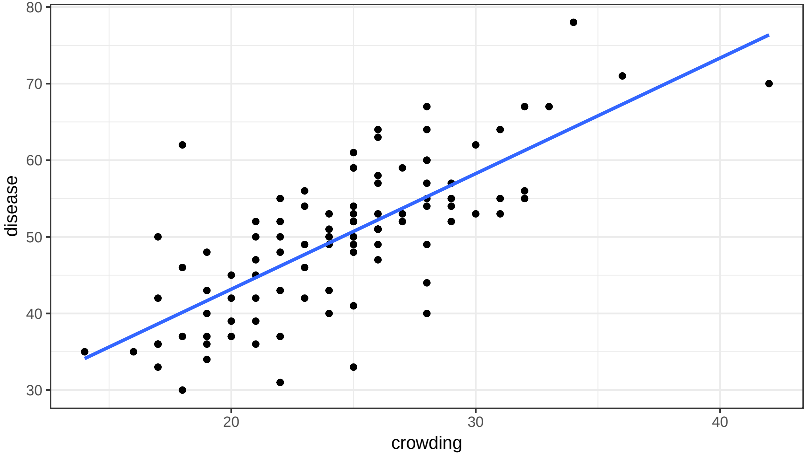
Reading model values: predicting new observations

What is the expected value of disease when crowding = 20? 30?
What range would you expect a new observation with crowding=20 to fall into?



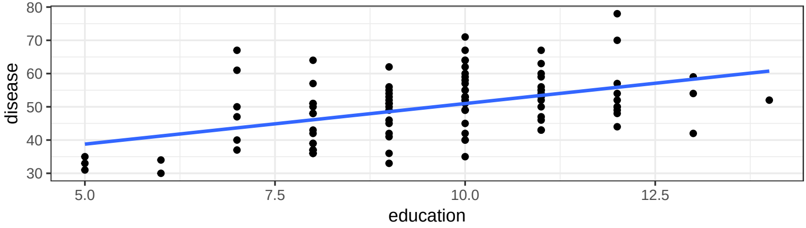
Lung Data Example: what is the model?

What do you like/dislike about this statement: “Based on this data, disease status worsens when crowding increases.”

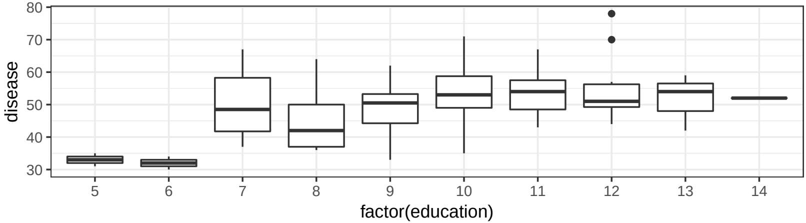


Difference between these representations of education?

```r
ggplot(dat, aes(education, disease)) + geom_point() +
    geom_smooth(method="lm", se=FALSE)
```




```r
ggplot(dat, aes(factor(education), disease)) + geom_boxplot()
```



Formulas for Statistical Models (Linear Regression)

In general, models can be expressed in this form:

$$\begin{aligned} [\text{ explanatory variable }] &\sim \text{intercept} + \text{terms} \\ [\text{ explanatory variable }] &= \text{intercept} + \text{terms} + \text{error} \end{aligned}$$

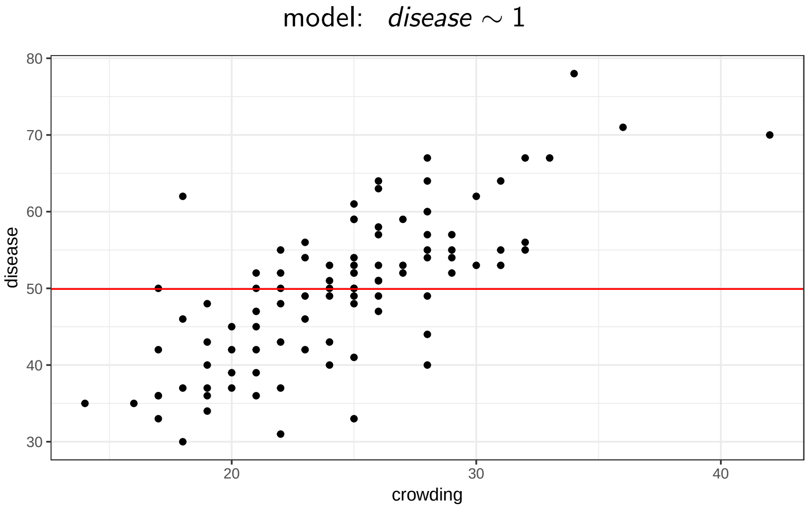
With a single predictor variable, this is simply a line:

$$\begin{aligned} Y &= a + b \cdot X + \epsilon \\ Y &= \beta_0 + \beta_1 \cdot X + \epsilon \end{aligned}$$

Different types of "terms"

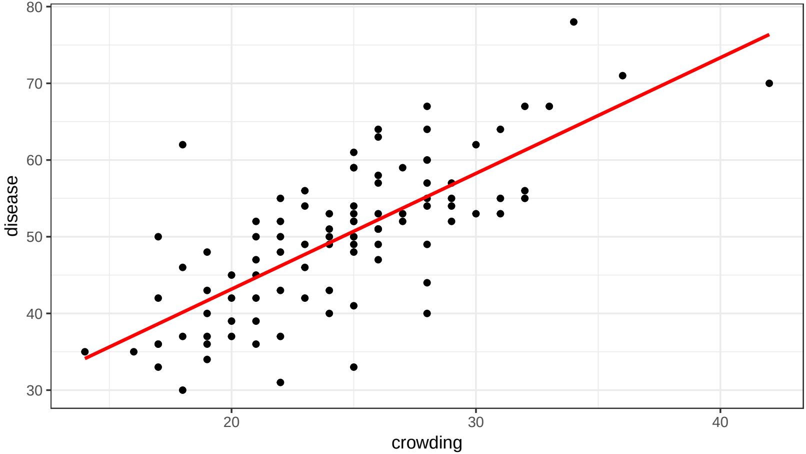
- intercept
- main effects
- interaction terms
- transformations
- smooth terms

Model terms: intercept



Model terms: main effects

model: $\textit{disease} \sim 1 + \textit{crowding}$

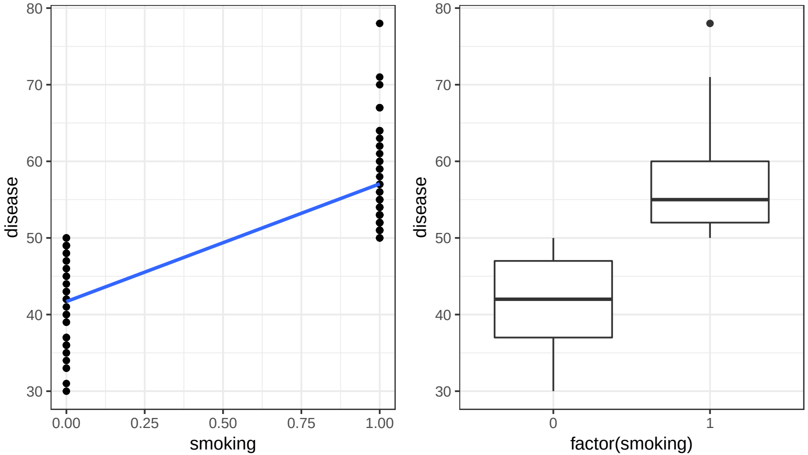


Model terms: main effects

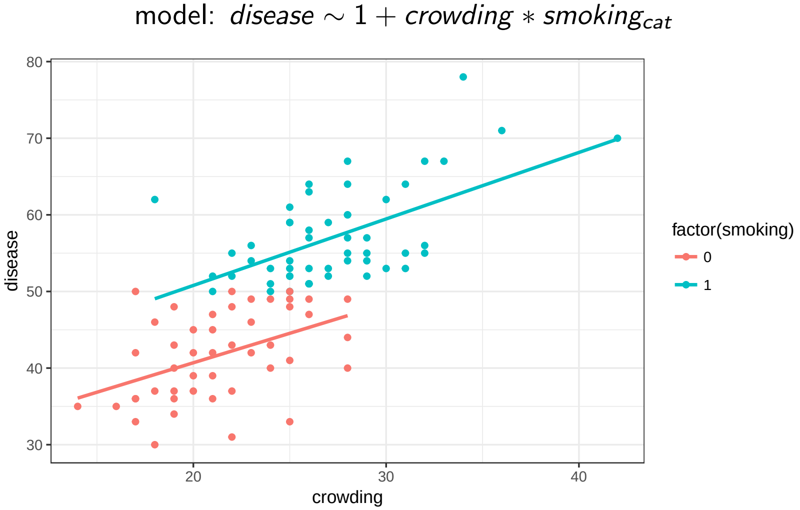
$$\text{model: } \textit{disease} \sim 1 + \textit{smoking}$$

vs.

$$\text{model: } \textit{disease} \sim 1 + \textit{smoking}_{cat}$$



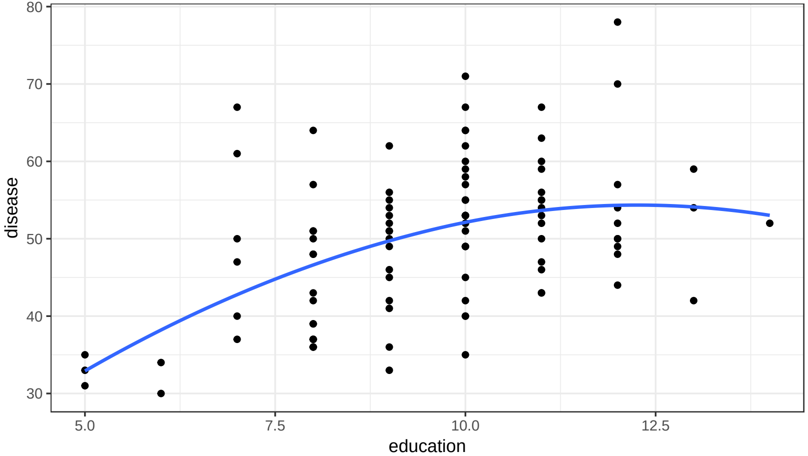
Model terms: main effects



Model terms: smooth effects

model: $disease \sim 1 + s(education)$

```r
ggplot(dat, aes(education, disease)) + geom_point() +
    geom_smooth( se=FALSE, span=2)
```



Lung Data Example

```
mlr1 <- lm(disease ~ crowding, data=dat)
kable(summary(mlr1)$coef, digits=2, format="latex")
```

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	12.99	3.48	3.74	0
crowding	1.51	0.14	10.83	0

```
mlr2 <- lm(disease ~ crowding + airqual, data=dat)
kable(summary(mlr2)$coef, digits=2, format="latex")
```

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	2.88	2.49	1.16	0.25
crowding	1.40	0.09	15.02	0.00
airqual	0.31	0.03	11.06	0.00

Why are the coefficients different?

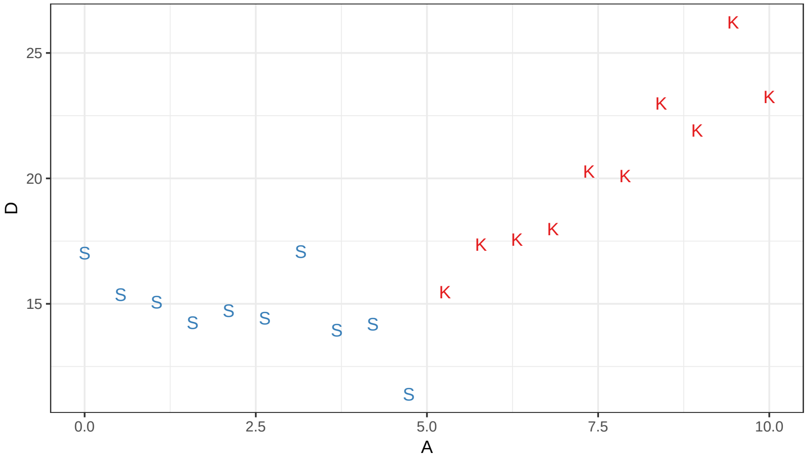
Lung Data Example

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	2.88	2.49	1.16	0.25
crowding	1.40	0.09	15.02	0.00
airqual	0.31	0.03	11.06	0.00

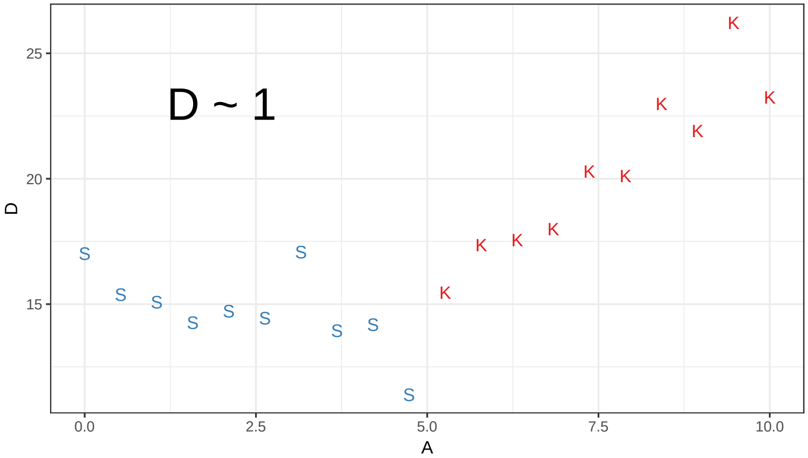
What are the interpretations of the coefficients?

Example data

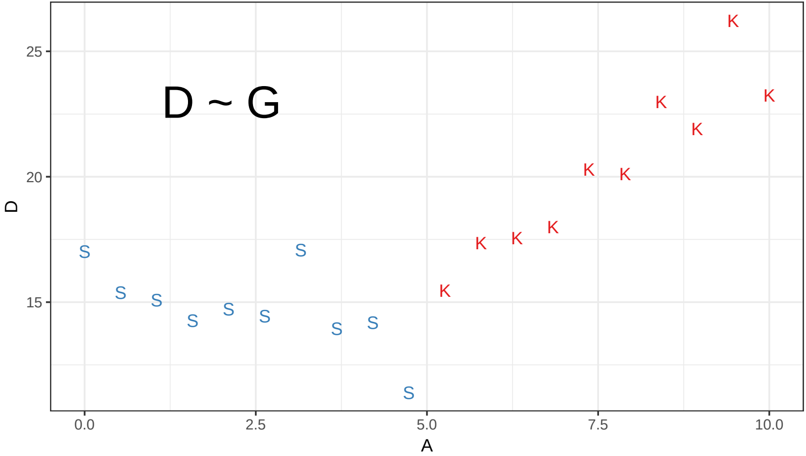
- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



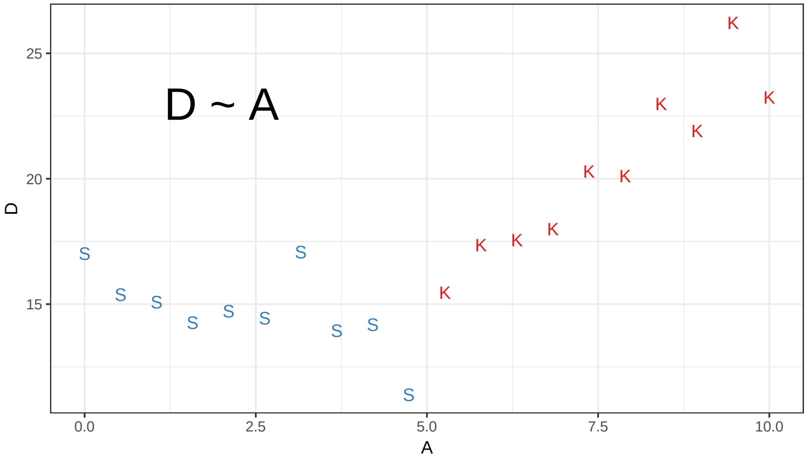
Draw the model...



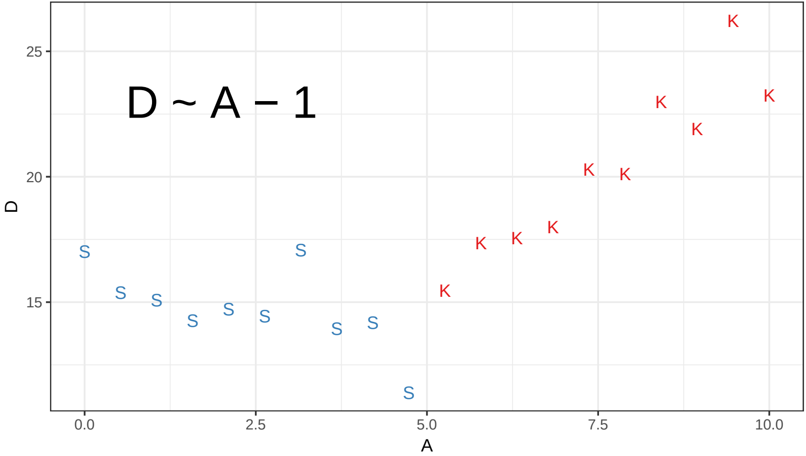
Draw the model...



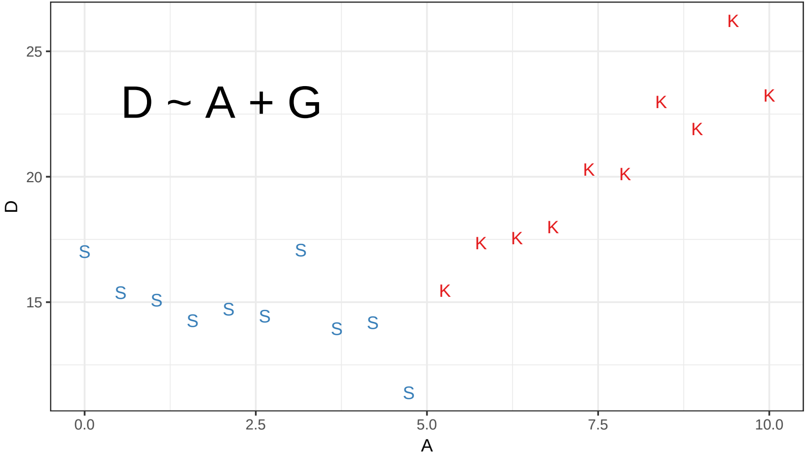
Draw the model...



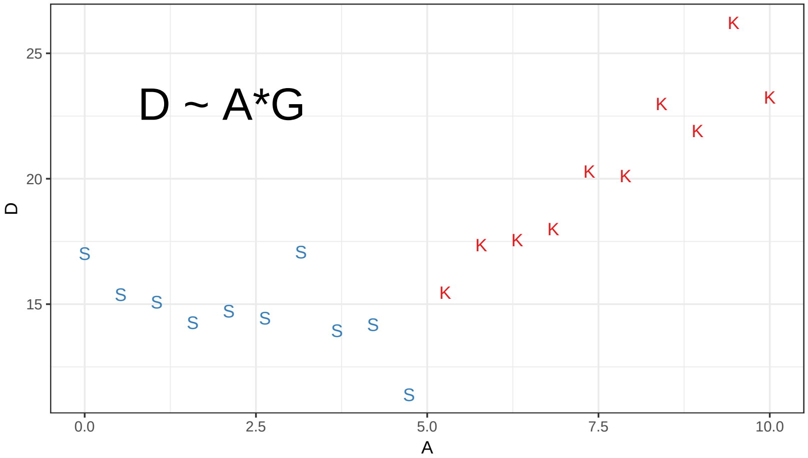
Draw the model...



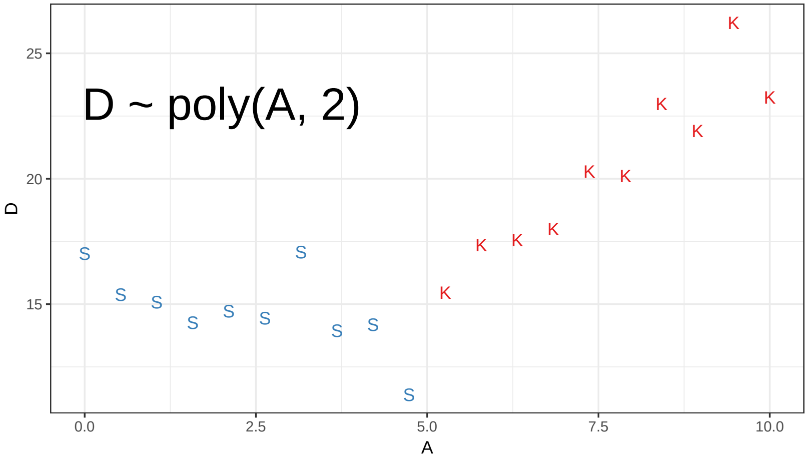
Draw the model...



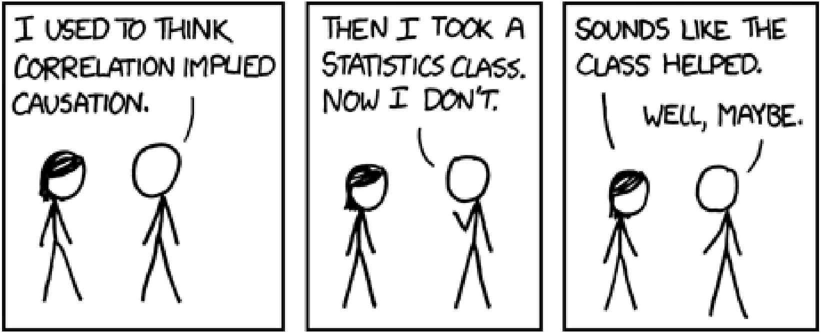
Draw the model...



Draw the model...



Parting wisdom



Up next: the mechanics and math of fitting models to data!

* Image credits: XKCD, http://xkcd.com/552/