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.

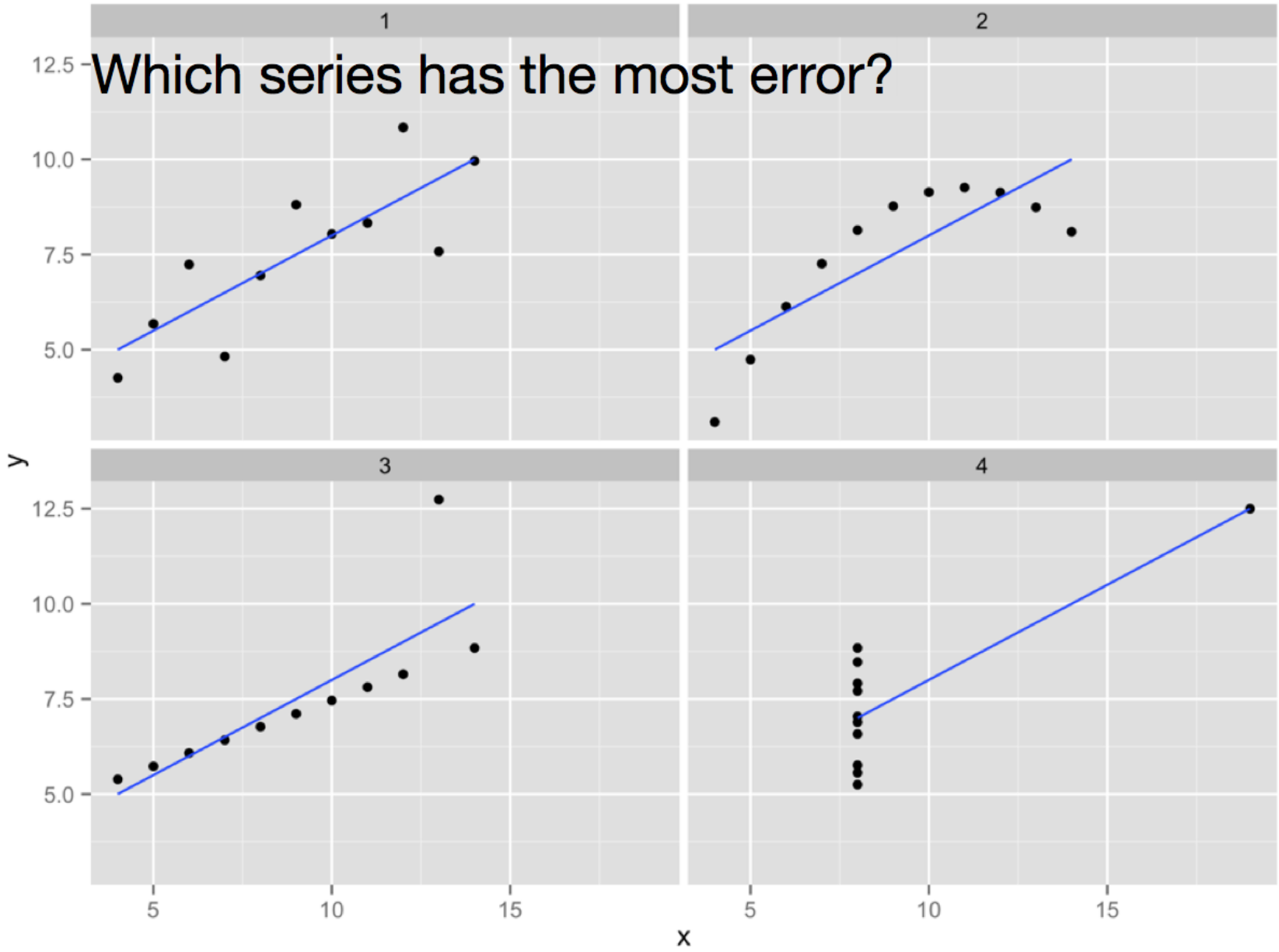


Figure acknowledgements to Hadley Wickham.

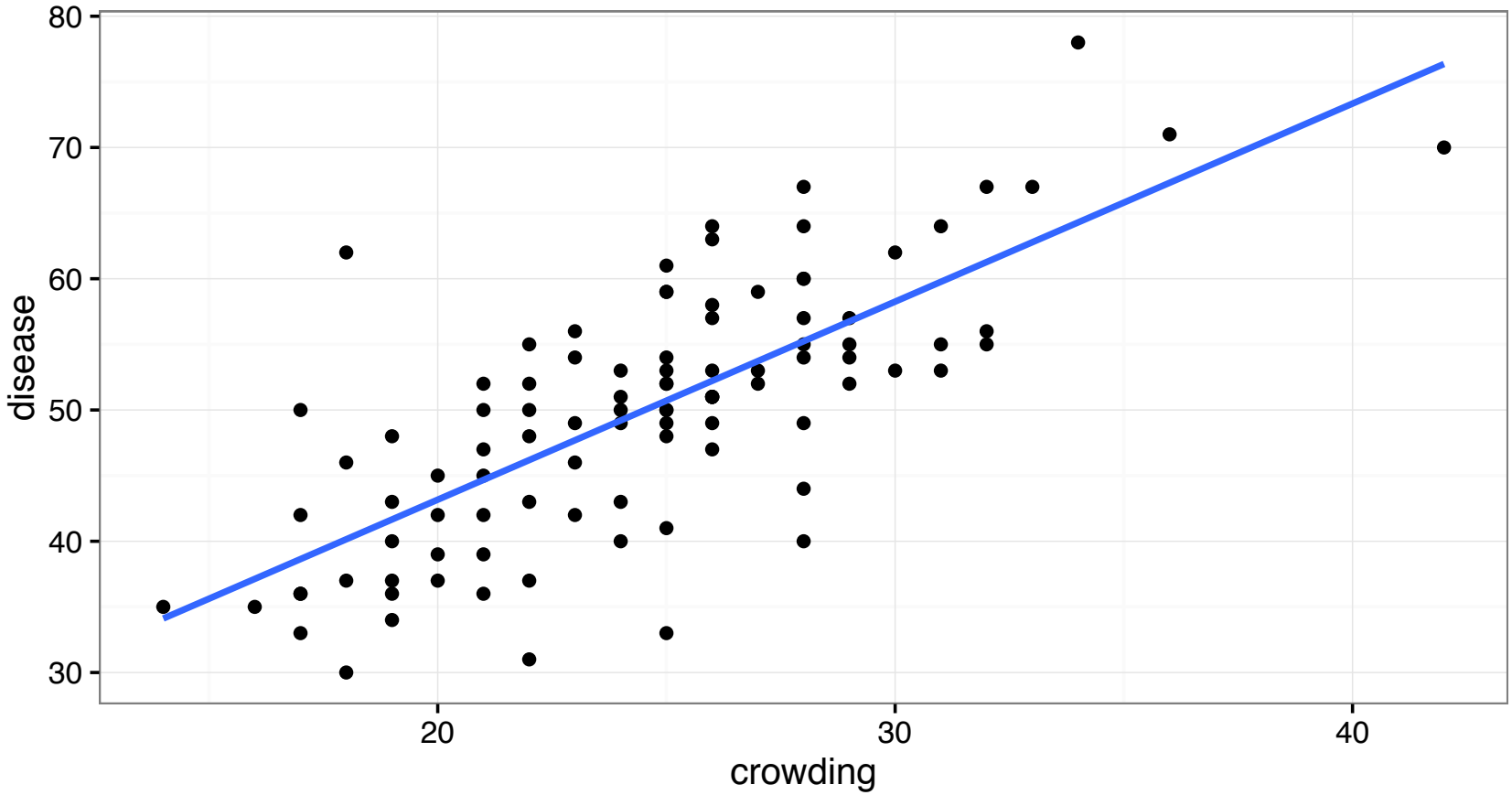
Lung Data Example

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

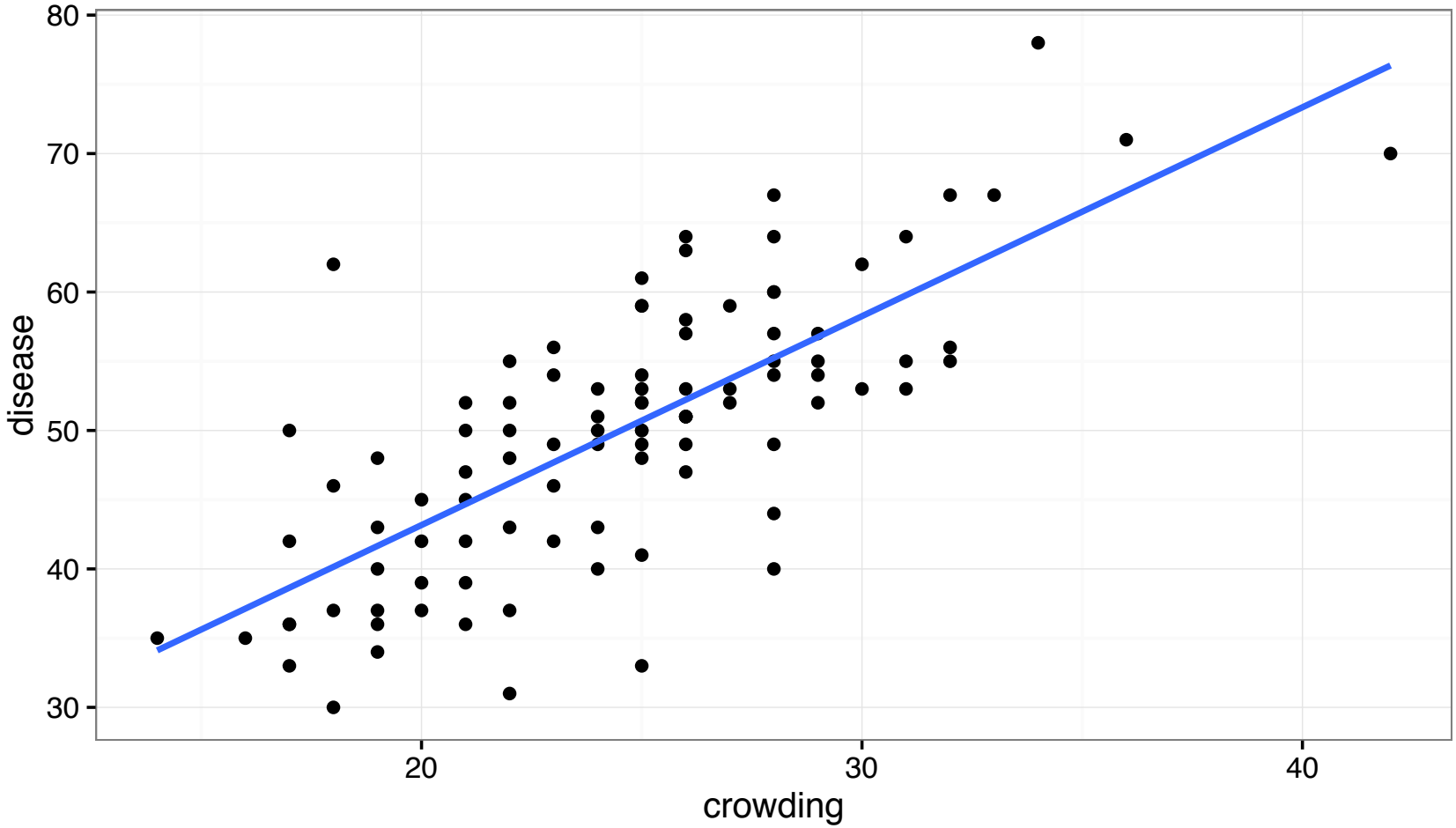
```
dat <- read.table("lungc.txt", header=TRUE)
ggplot(dat, aes(crowding, disease)) + geom_point() +
    geom_smooth(method="lm", se=FALSE)
```



Things to point out: response variable? explanatory variable? model value? residual?

Lung Data Example: what is the model?

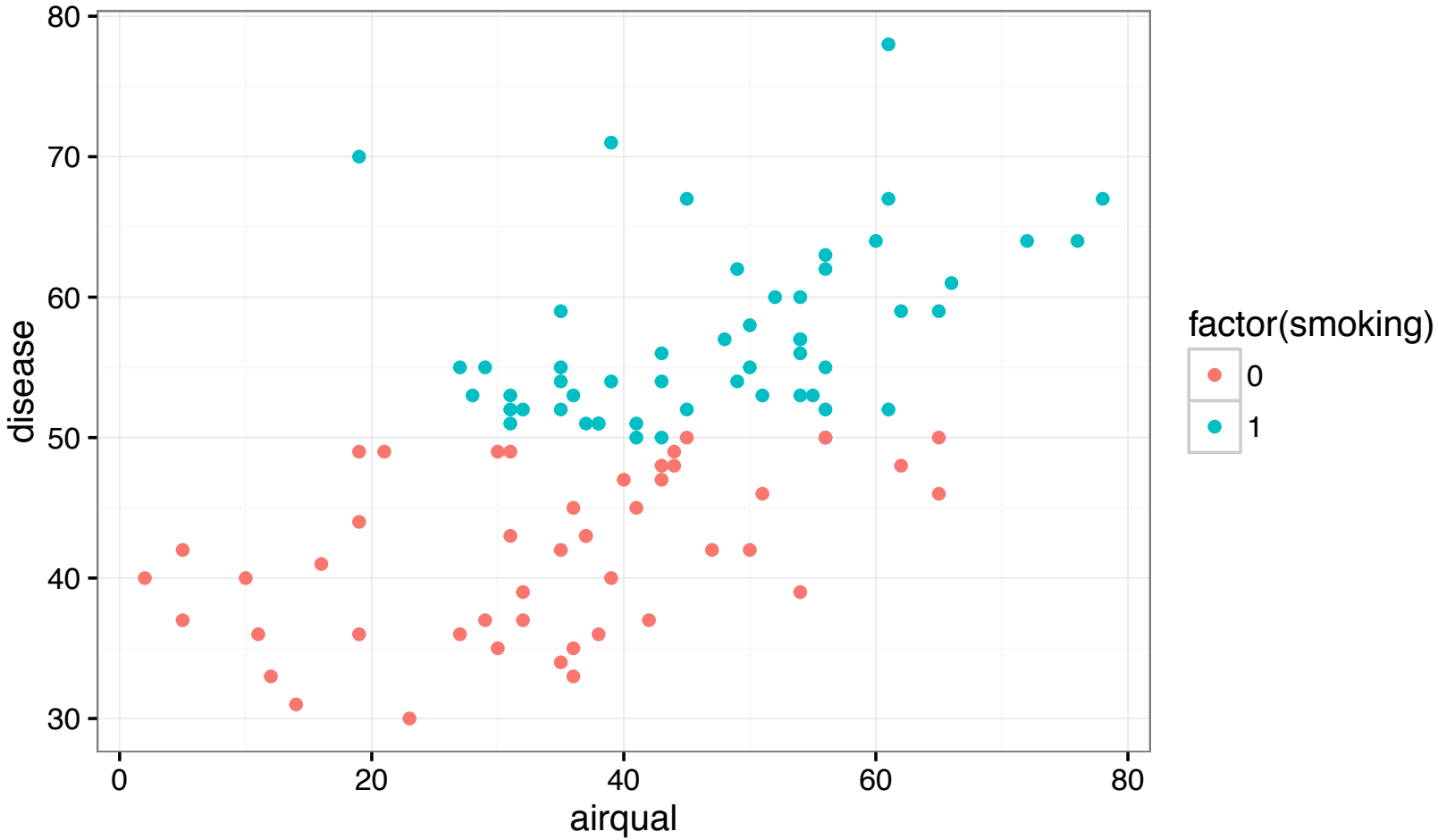
```
ggplot(dat, aes(crowding, disease)) + geom_point() +
    geom_smooth(method="lm", se=FALSE)
```



What model syntax is implied by the above figure?

Lung Data Example: what is the model?

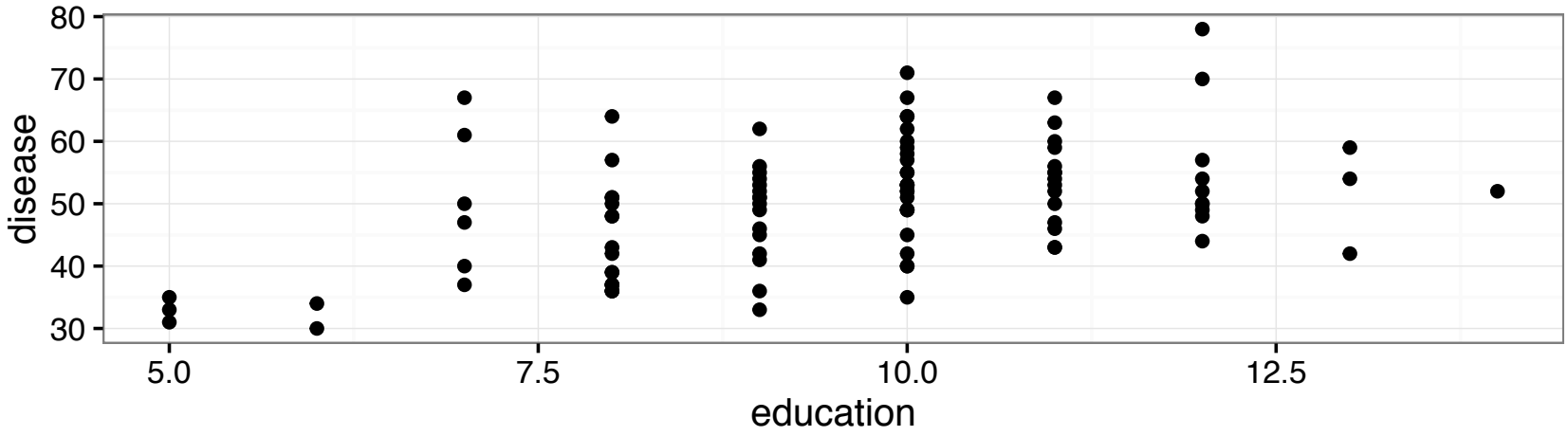
```
ggplot(dat, aes(airqual, disease, color=factor(smoking))) +
    geom_point()
```



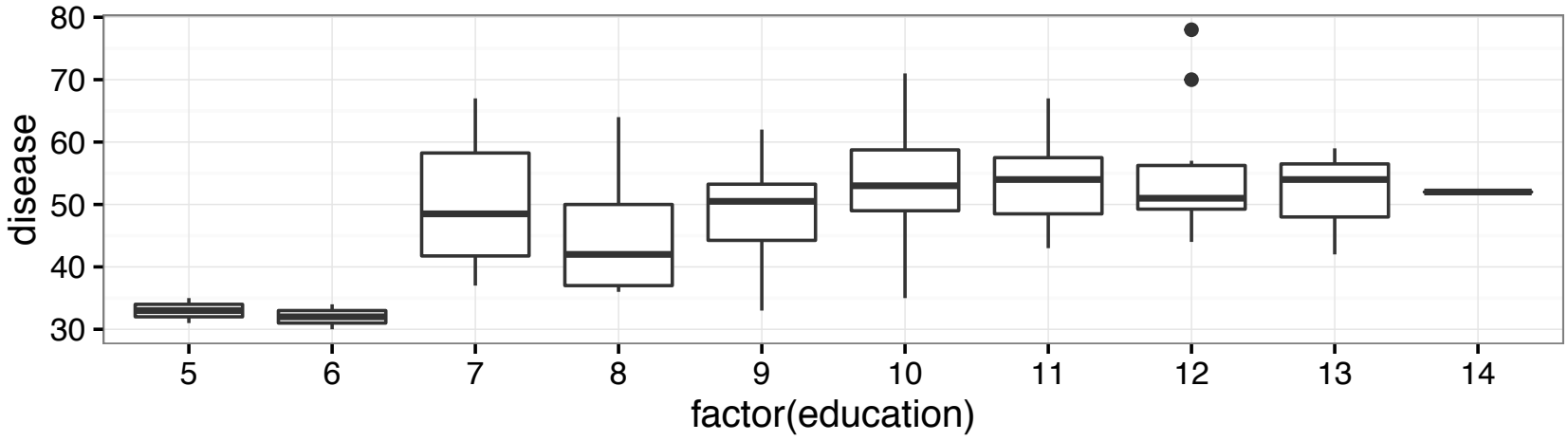
What is one possible model syntax implied by the above figure?

Which representation of education is better and why?

```
ggplot(dat, aes(education, disease)) + geom_point()
```




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



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?

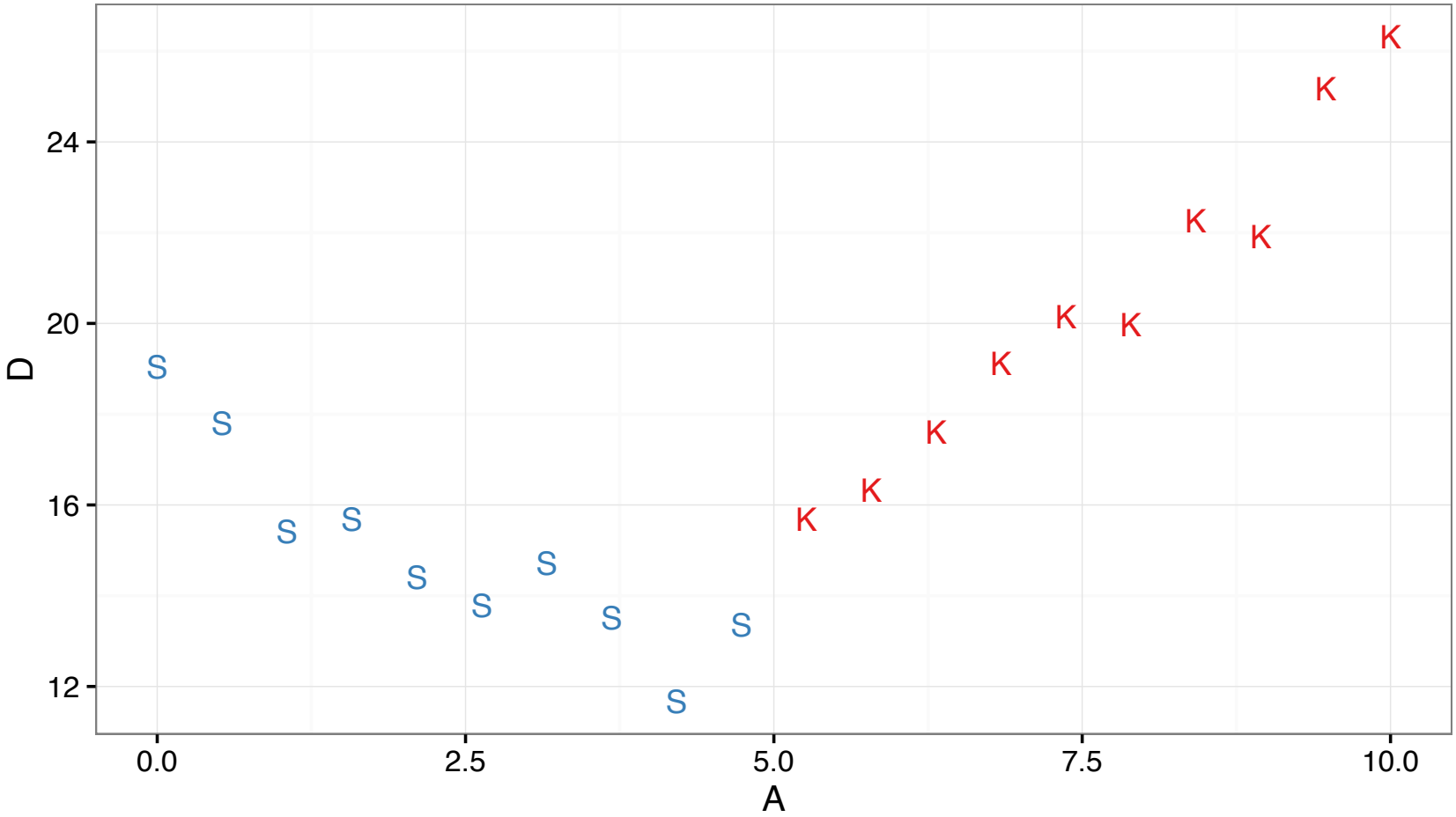
Lung Data Example

	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

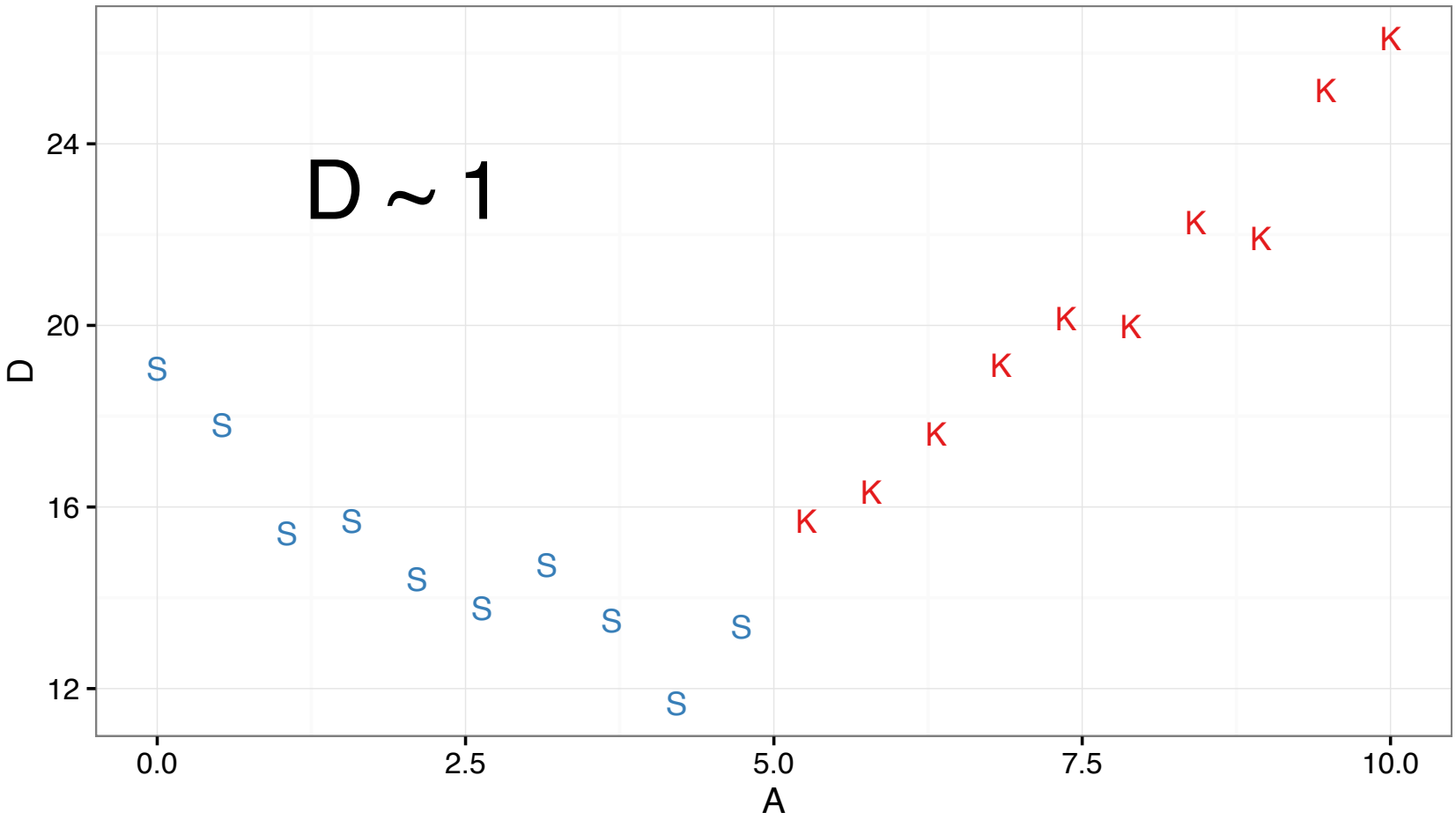
What are the interpretations of the coefficients?

Example data

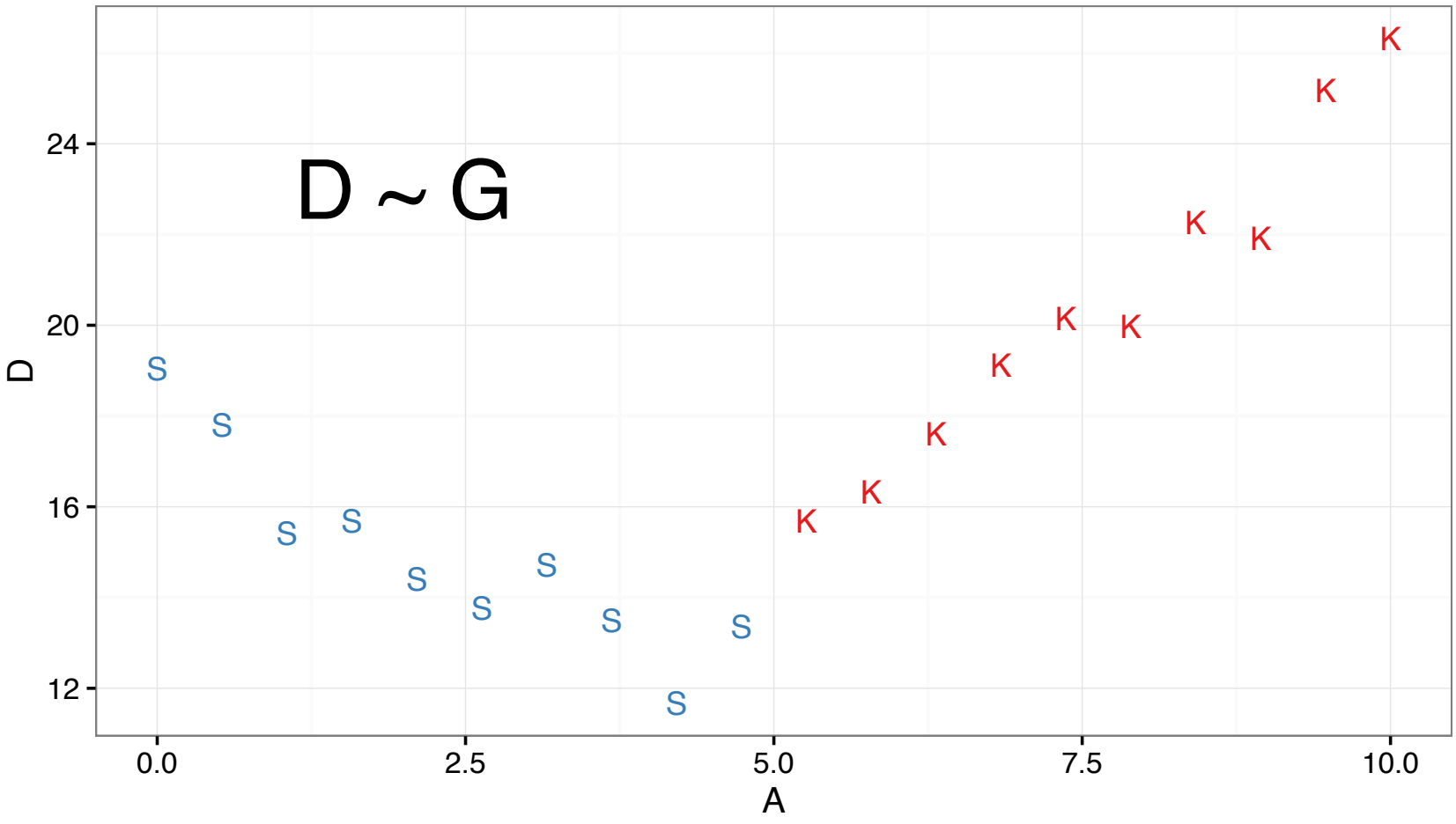
- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



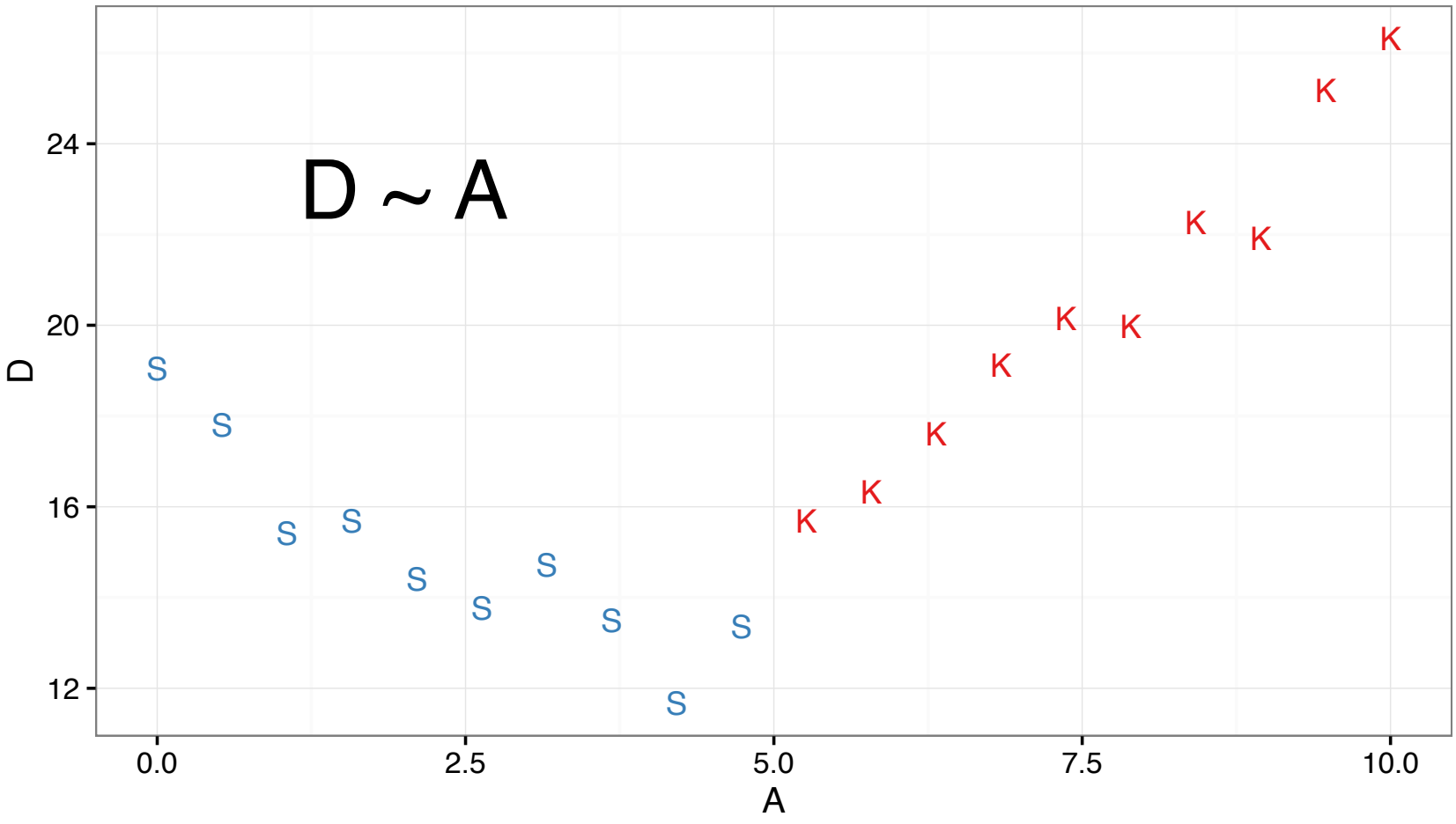
Draw the model...



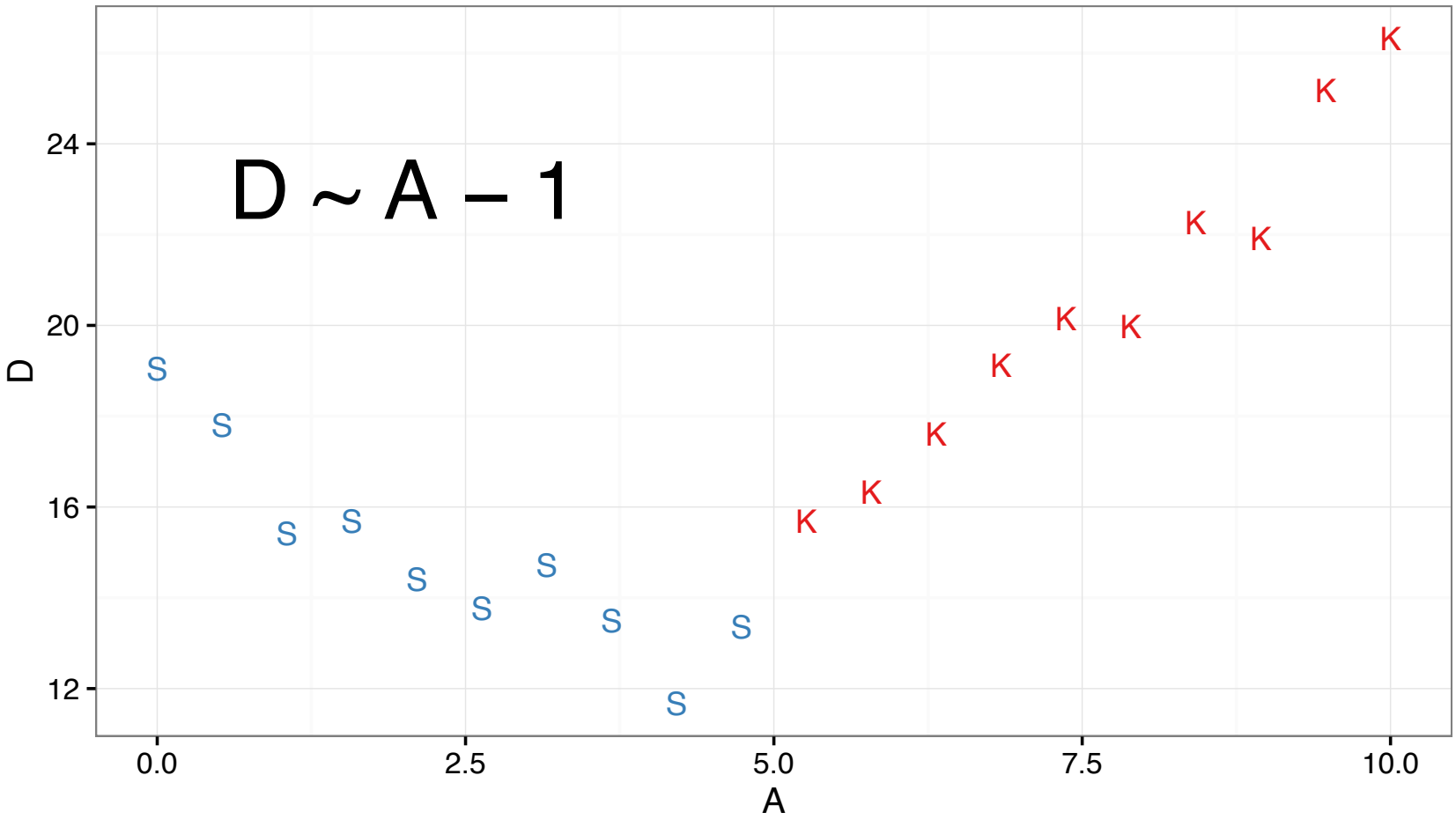
Draw the model...



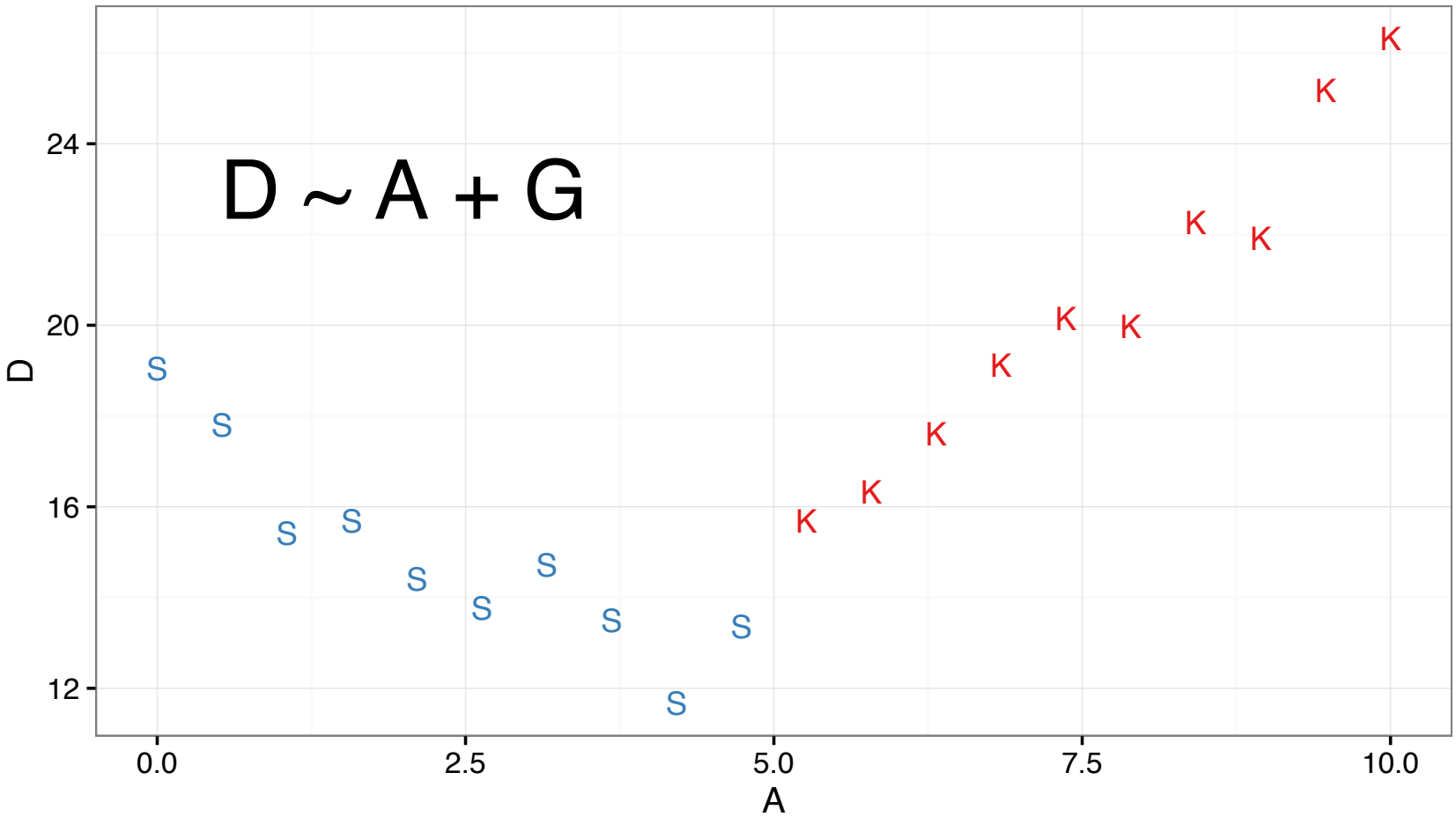
Draw the model...



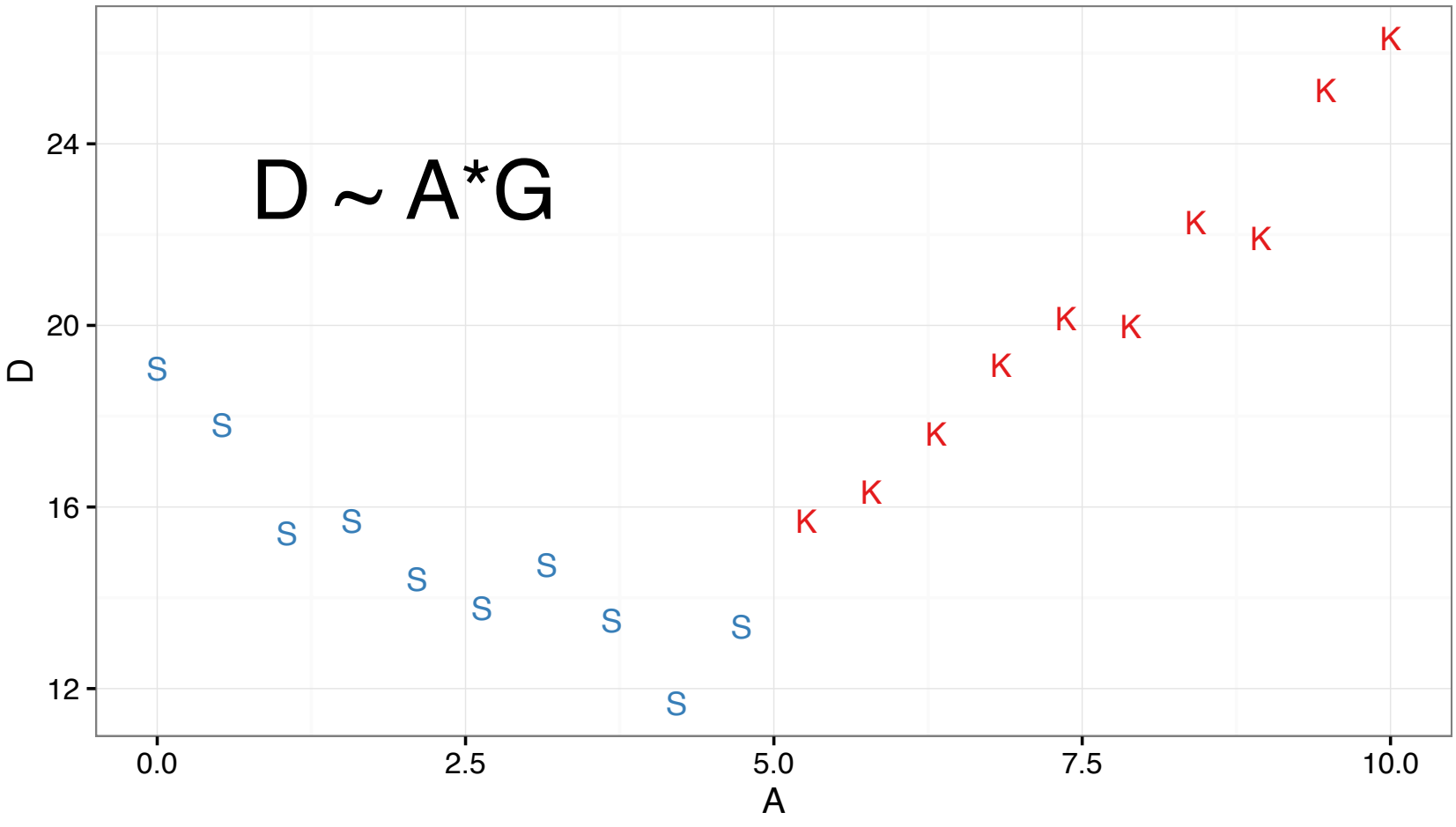
Draw the model...



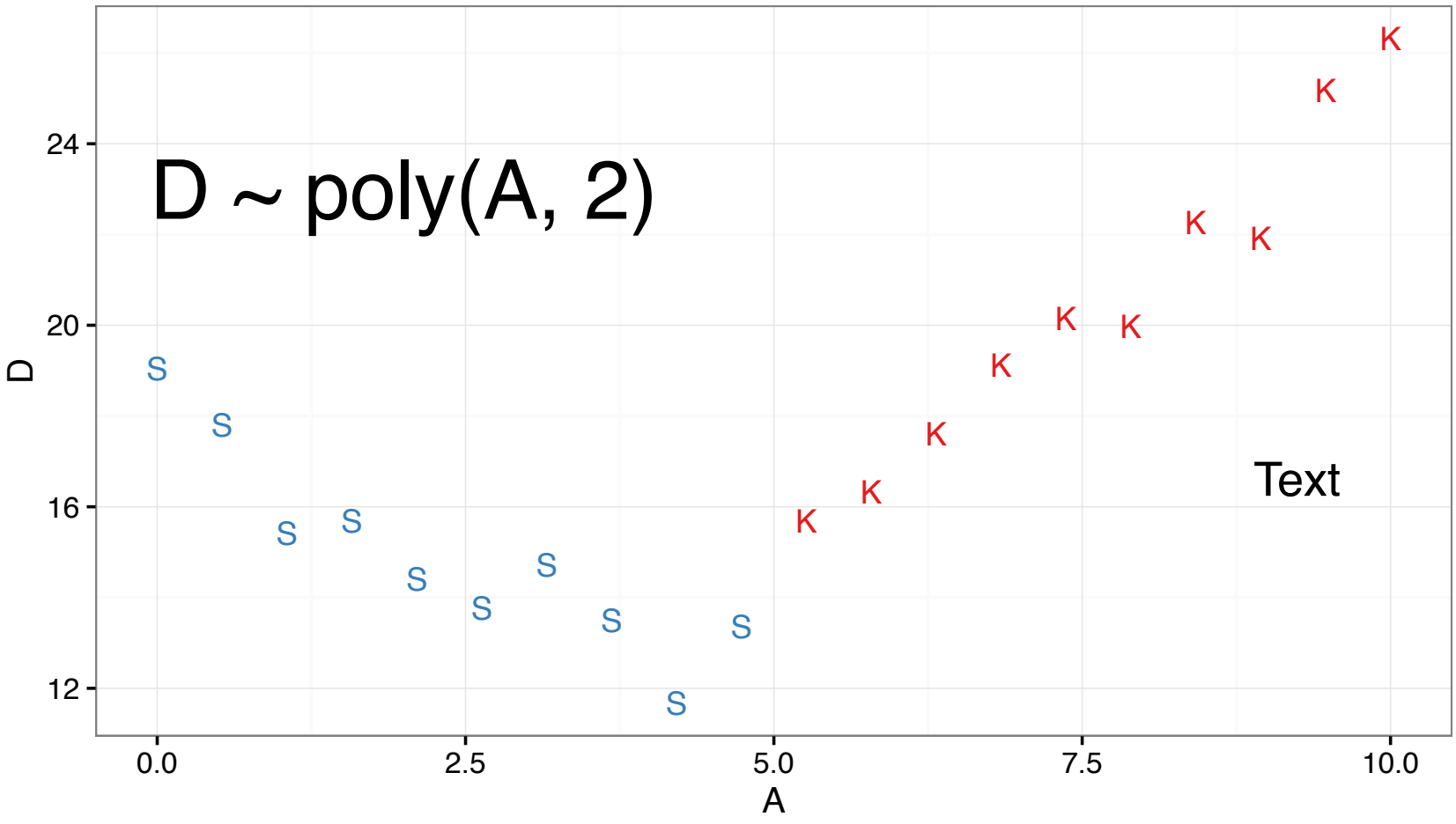
Draw the model...



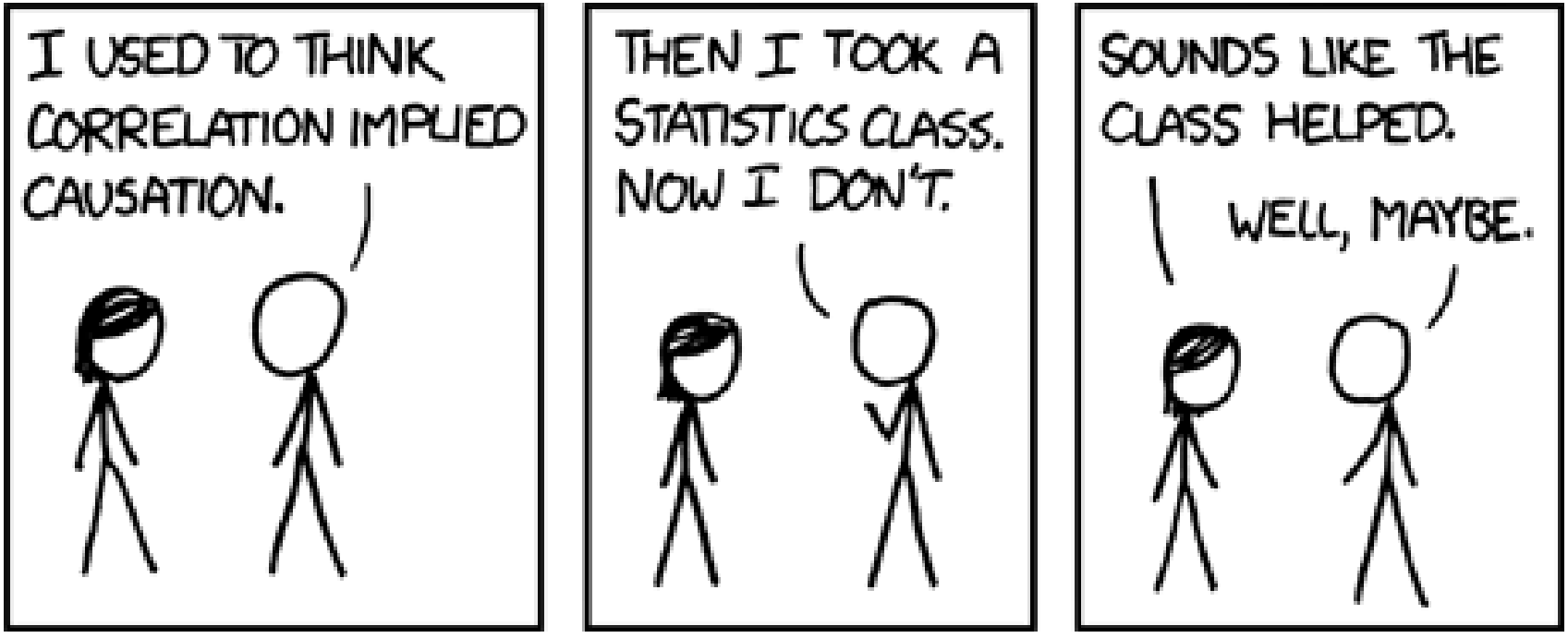
Draw the model...



Draw the model...



Parting wisdom



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

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