Slay the Word

Mark Andrews

13 July 2018

```
library(tibble)
library(dplyr)
library(ggplot2)
library(pander)
```

A univariate normal, or Gaussian, linear model is defined as follows. Assume that our data consists of n independent pairs of observations:

$$\mathcal{D} = \{(y_1, \vec{x}_1) \ldots (y_i, \vec{x}_i) \ldots (y_n, \vec{x}_n)\},$$

where each $y_i \in \mathbb{R}$ and $\vec{x}_i \in \mathbb{R}^K$. We then model this data as follows:

$$y_i \sim N(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ik}, \sigma^2), \quad \text{for } i \in \{1, 2 \ldots n\}.$$

Demonstration

Here, we generate some data.

```
N <- 50
Df <- tibble(x = rnorm(N),
             y = 1.25 + 2.25*x + rnorm(N),
             z = sample(c('yes','no'), size=N, replace=T)
)
```

```
Df %>% ggplot(mapping=aes(x=x, y=y, col=z)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  theme_classic()
```

Here, we fit the model with maximum likelihood estimation:

```
M <- lm(y ~ x + z, data=Df)
pander(summary(M))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.92	0.2073	4.438	5.457e-05
x	2.002	0.1587	12.61	1.08e-16
zyes	0.495	0.2961	1.672	0.1012

Table 2: Fitting linear model: y ~ x + z

Observations	Residual Std. Error	R^2	Adjusted R^2
50	1.019	0.7938	0.7851

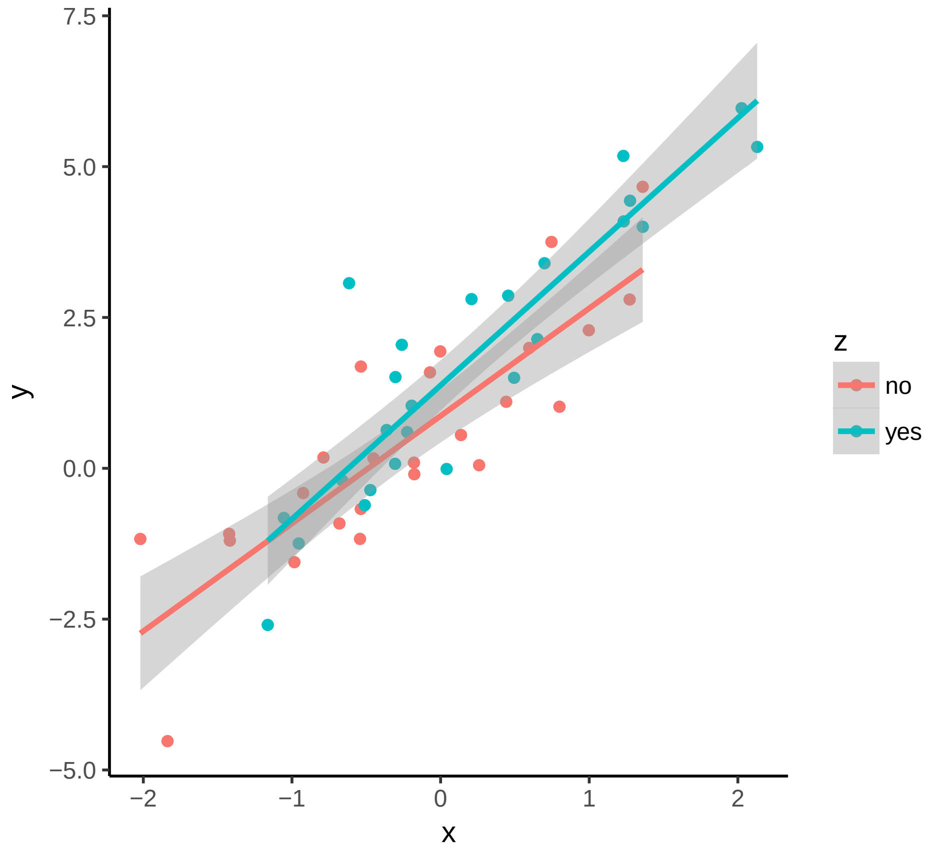

Figure 1: Scatterplot with line of best fit.