LINEAR MODELS ARE GREAT

ALEX GOLD
SOLUTIONS ENGINEER
RSTUDIO

@alexkgold

Slides at: https://github.com/akgold/dsdc_linear_models

IT ME.

WHAT DOES IT MEAN TO BE LINEAR?

$$\begin{array}{r}21.0\\21.0\\22.8\\21.4\\18.7\\18.1\\14.3\\24.4\\22.8\\19.2\\17.8\\16.4\\17.3\\15.2\\10.4\\10.4\\14.7\\32.4\\30.4\\33.9\\21.5\\15.5\\15.2\\13.3\\19.2\\27.3\\26.0\\30.4\\15.8\\19.7\\15.0\\21.4\end{array} = \beta_0 + \beta_{cyl} \begin{array}{r}6\\6\\4\\6\\8\\6\\8\\4\\4\\6\\6\\8\\8\\8\\8\\8\\8\\4\\4\\4\\4\\8\\8\\8\\8\\4\\4\\4\\8\\6\\8\\4\end{array} + \beta_{hp} \begin{array}{r}160.0\\160.0\\108.0\\258.0\\360.0\\225.0\\360.0\\146.7\\140.8\\167.6\\167.6\\275.8\\275.8\\275.8\\472.0\\460.0\\440.0\\78.7\\75.7\\71.1\\120.1\\318.0\\304.0\\350.0\\400.0\\79.0\\120.3\\95.1\\351.0\\145.0\\301.0\\121.0\end{array}$$

+ some prediction error

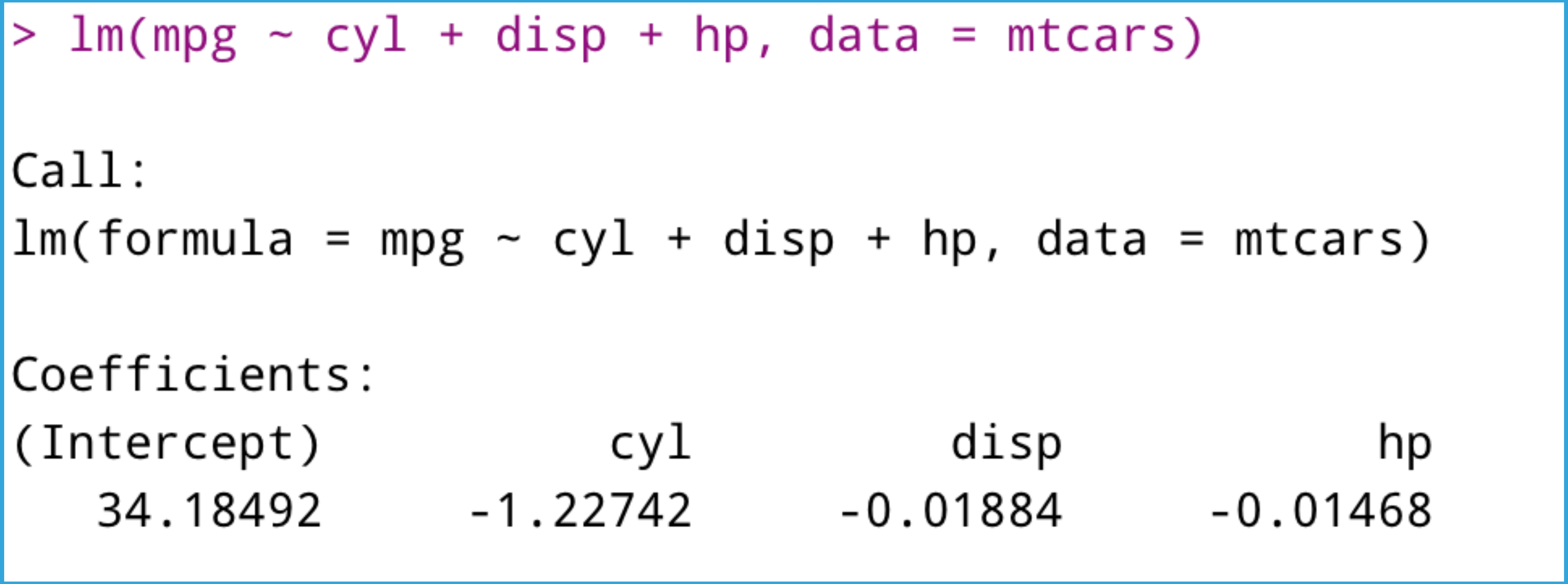
↑

Just a number

WHAT DOES IT MEAN TO BE LINEAR?

$$\boldsymbol{Y} = \beta \boldsymbol{X} + \epsilon$$

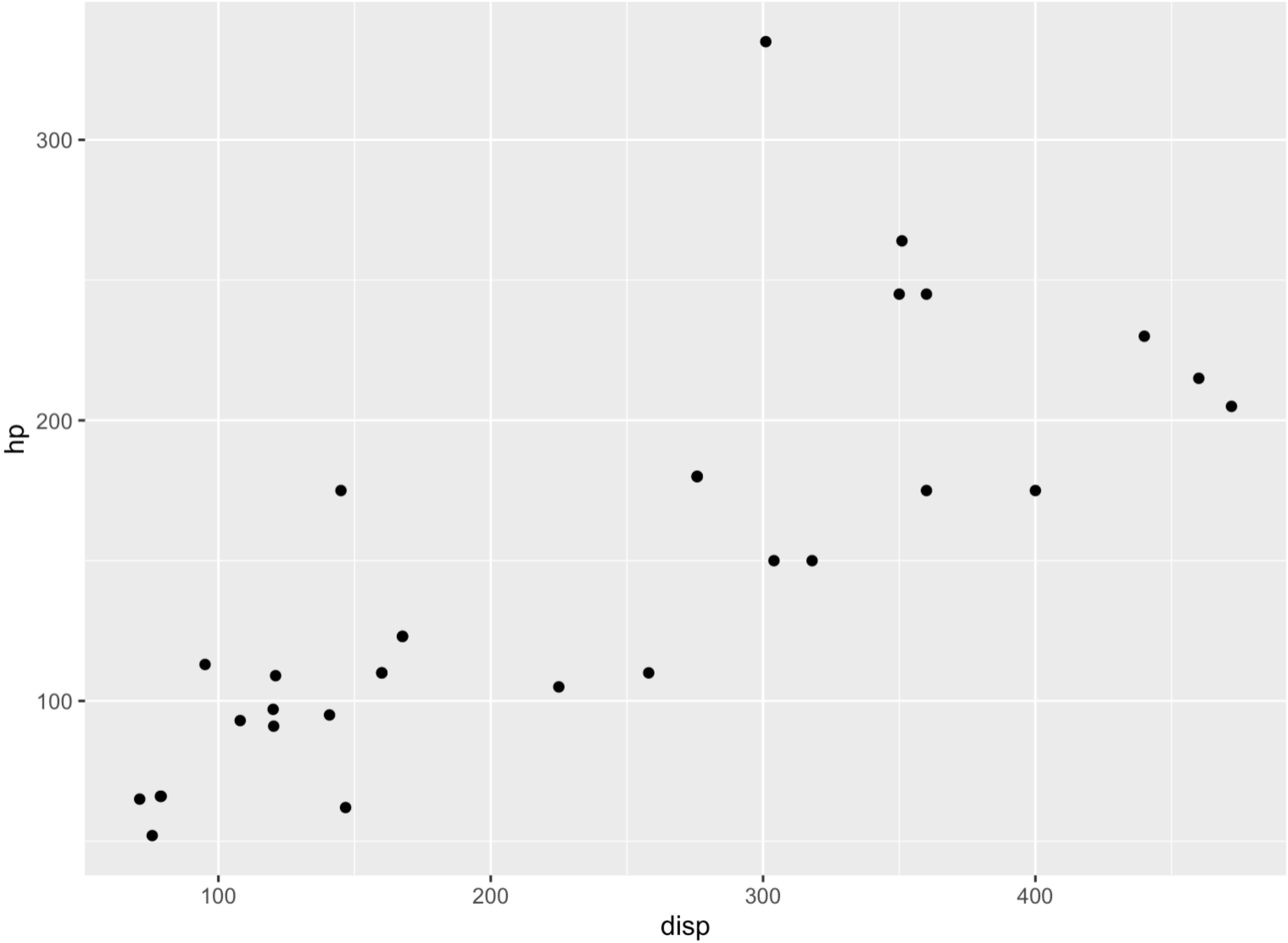
ALEX <3 LINEAR MODELS

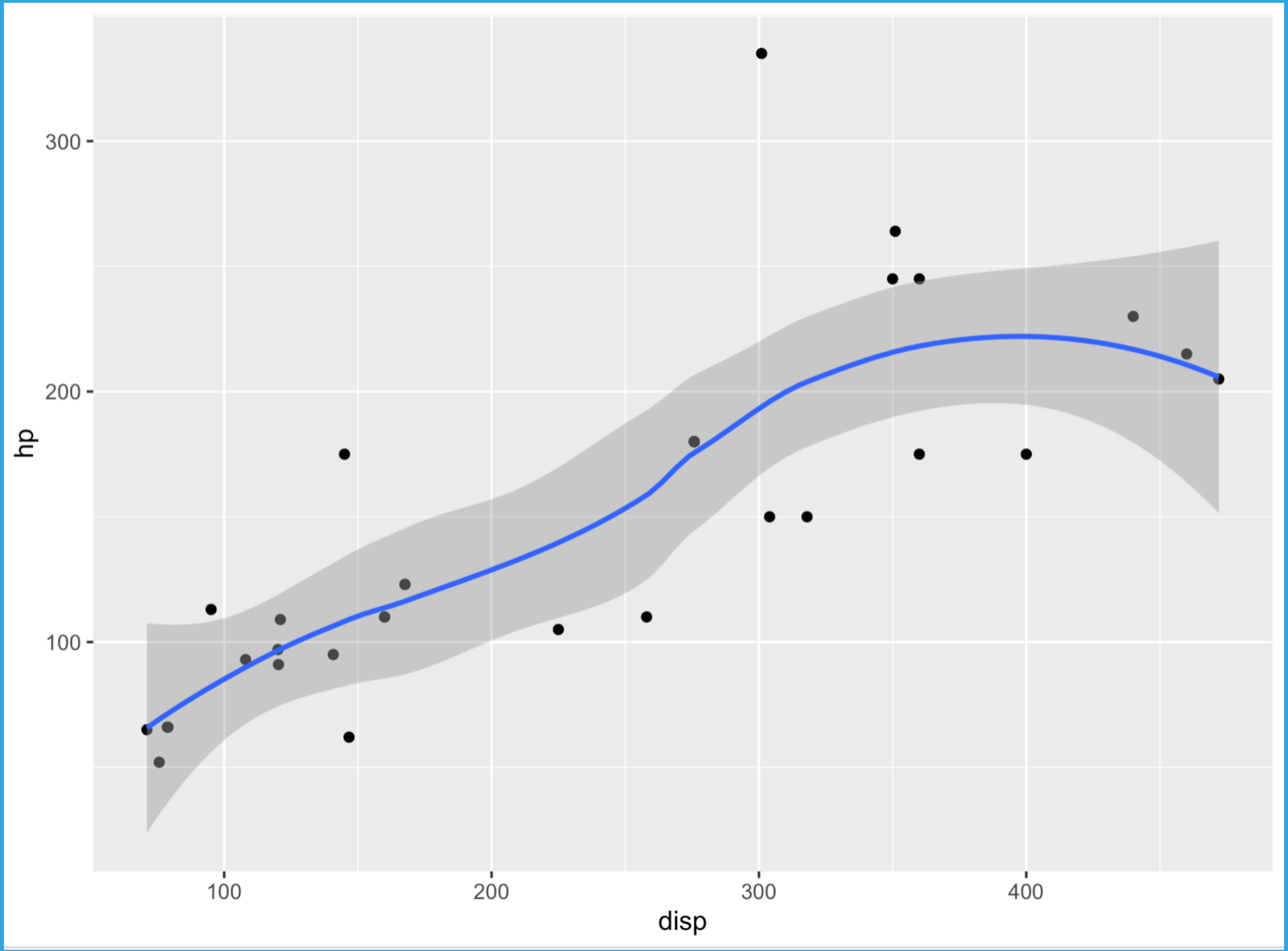


```
> lm(mpg ~ cyl + disp + hp, data = mtcars)

Call:
lm(formula = mpg ~ cyl + disp + hp, data = mtcars)

Coefficients:
(Intercept)          cyl         disp           hp
   34.18492     -1.22742     -0.01884     -0.01468
```

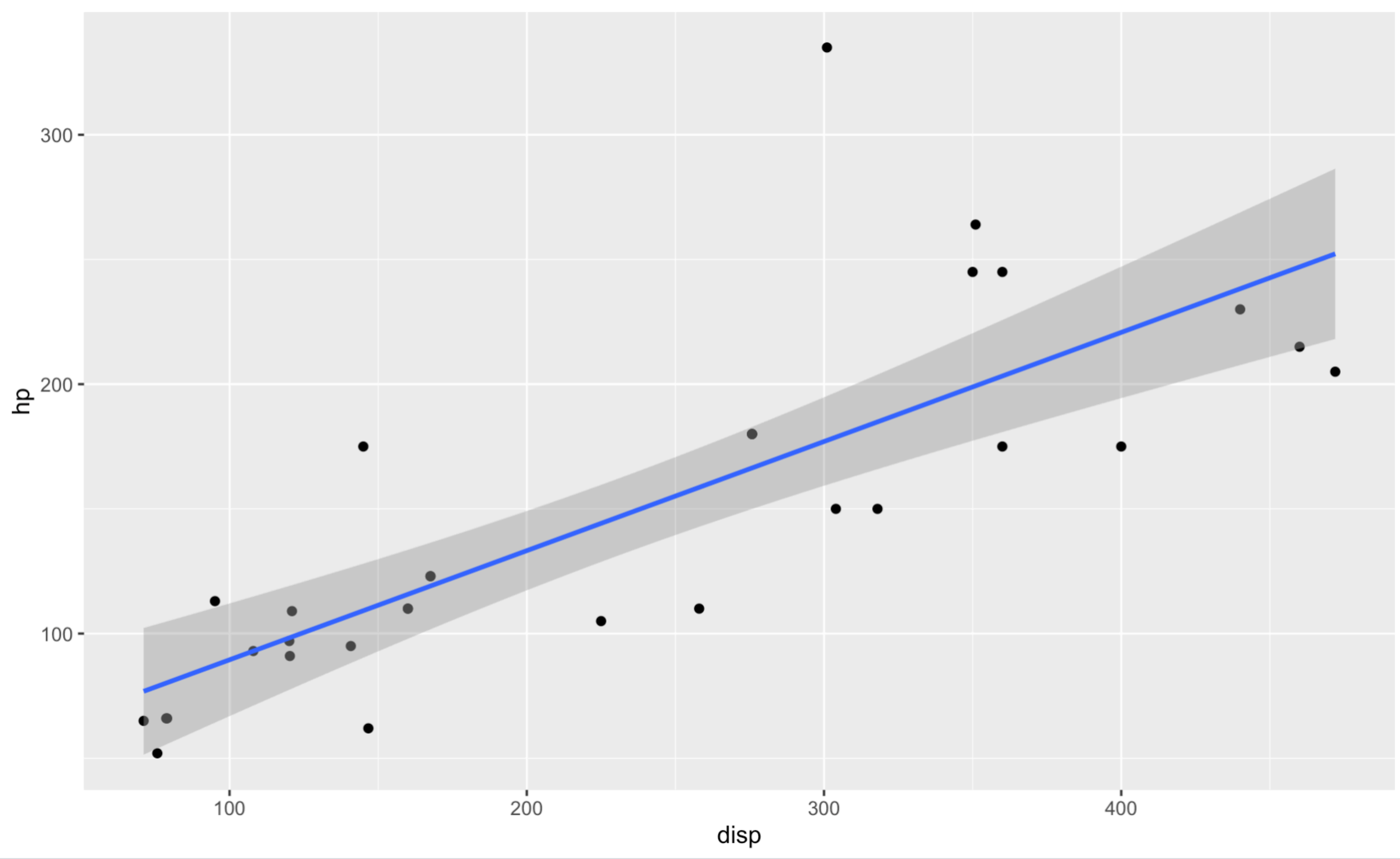
$$\boldsymbol{Y} = \underset{\uparrow}{\beta}\boldsymbol{X} + \epsilon$$

$$Y = \sum_{k} \beta_k \overset{\downarrow}{f}_k(x_i) + \epsilon$$


300
200
100
hp
100
200
300
400
disp



```
ggplot(mtcars, aes(x = disp, y = hp)) + geom_point() + stat_smooth(method = "loess")
```

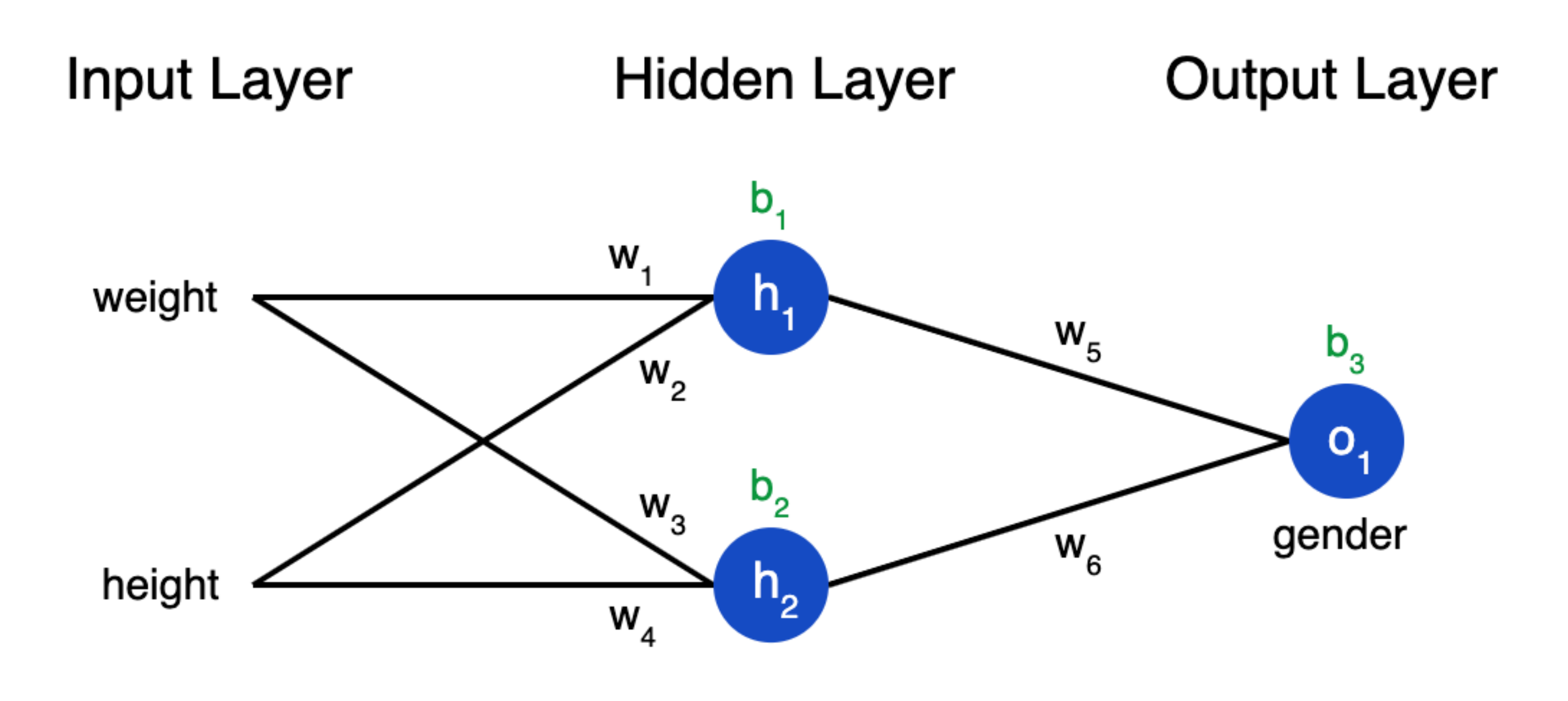


```
ggplot(mtcars, aes(x = disp, y = hp)) + geom_point() + stat_smooth(method = "lm")
```

$$gender = \beta_0 + \beta_1 weight + \beta_2 height$$

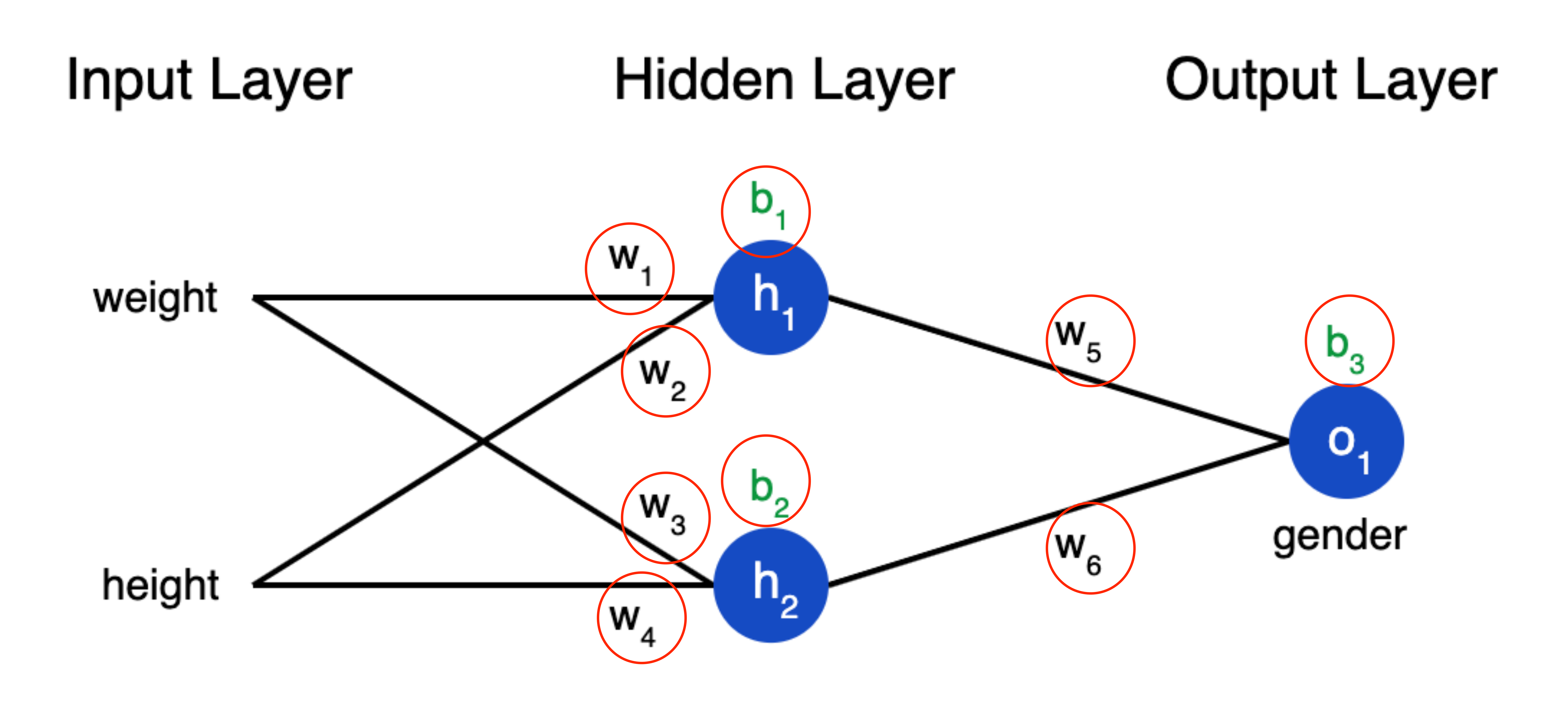
$$gender = \beta_0 + \beta_1 weight + \beta_2 height$$

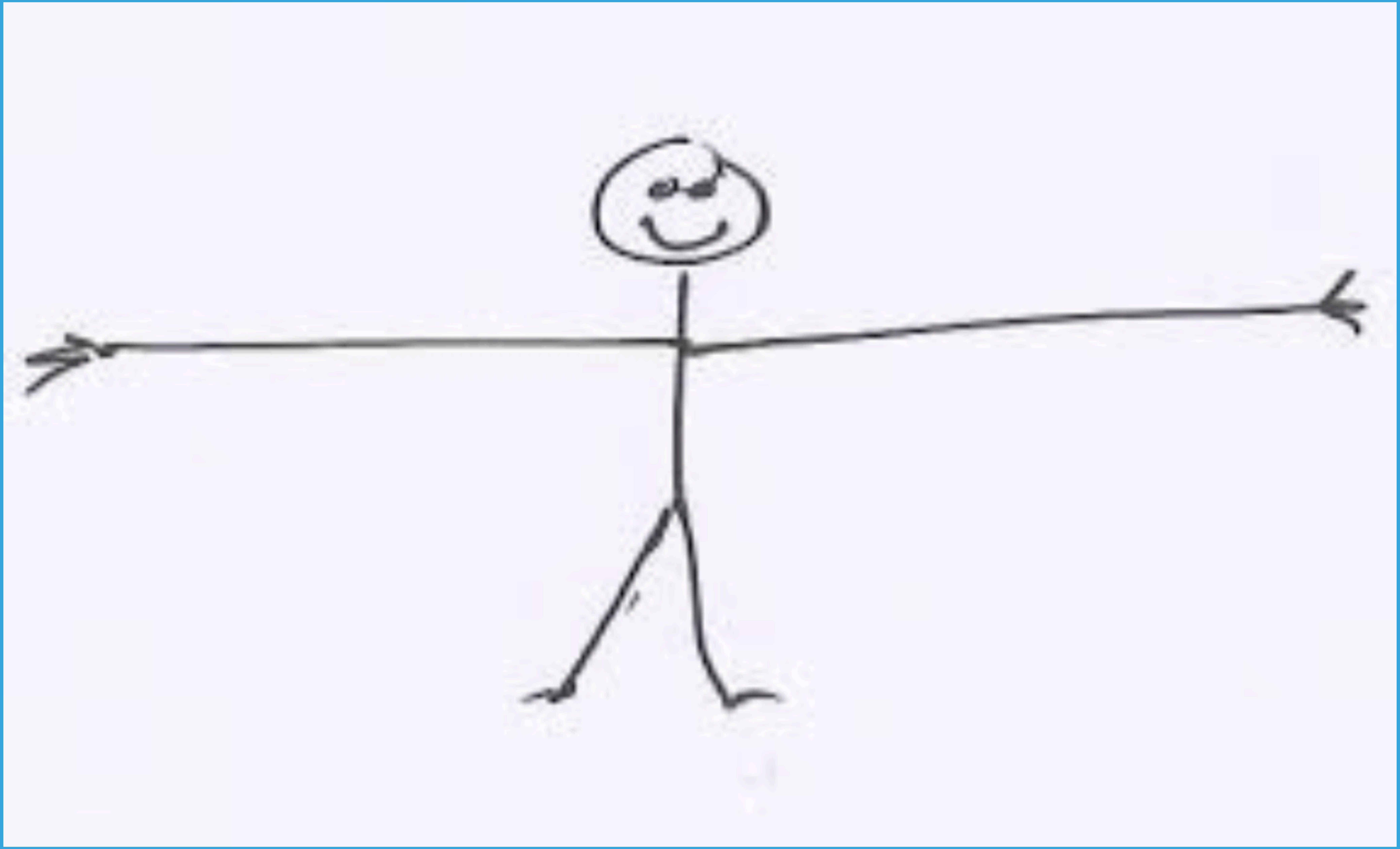
vs



$$gender = \beta_0 + \beta_1 weight + \beta_2 height$$

VS





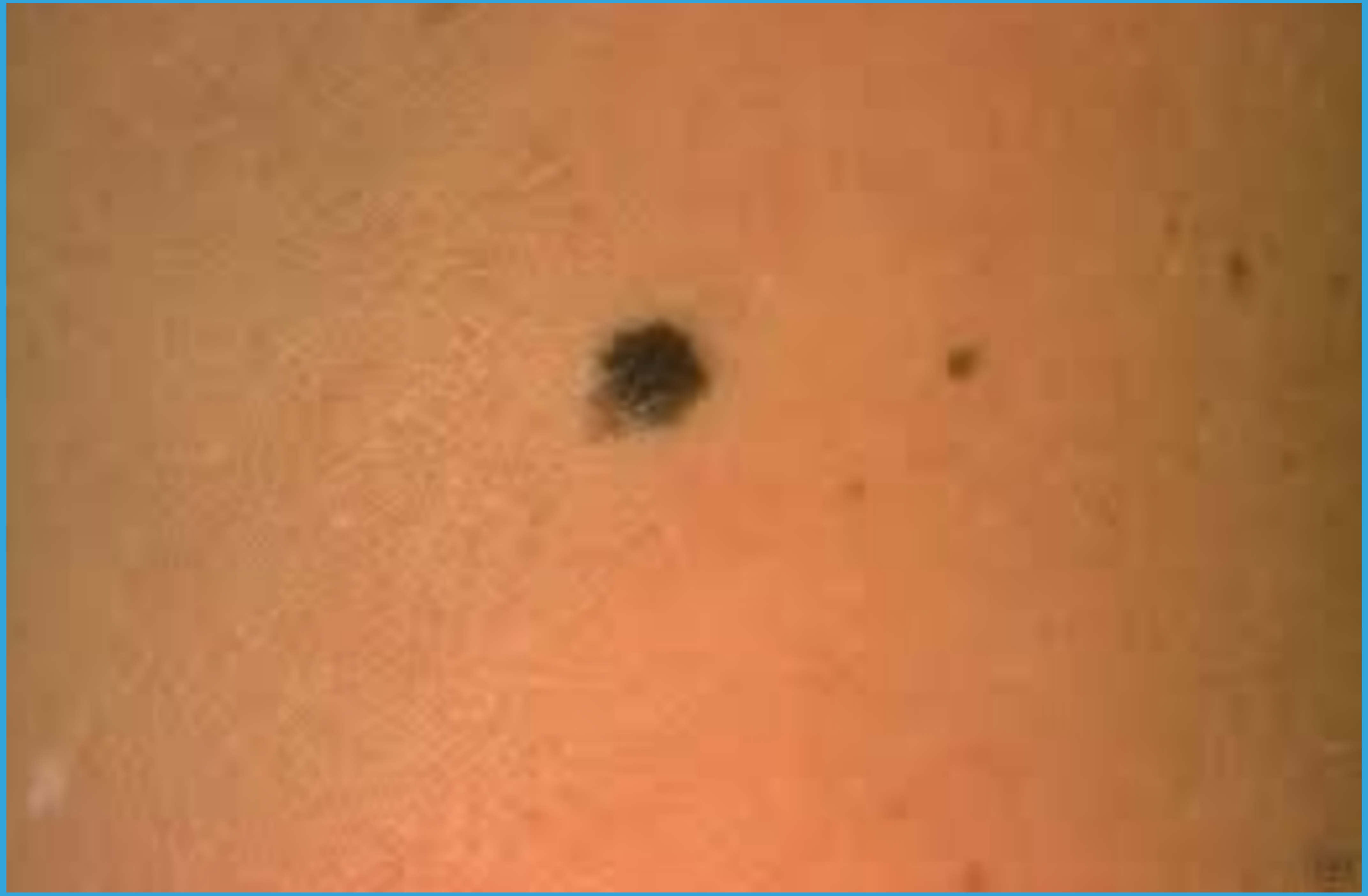
It’s all about the Data-Generating Process

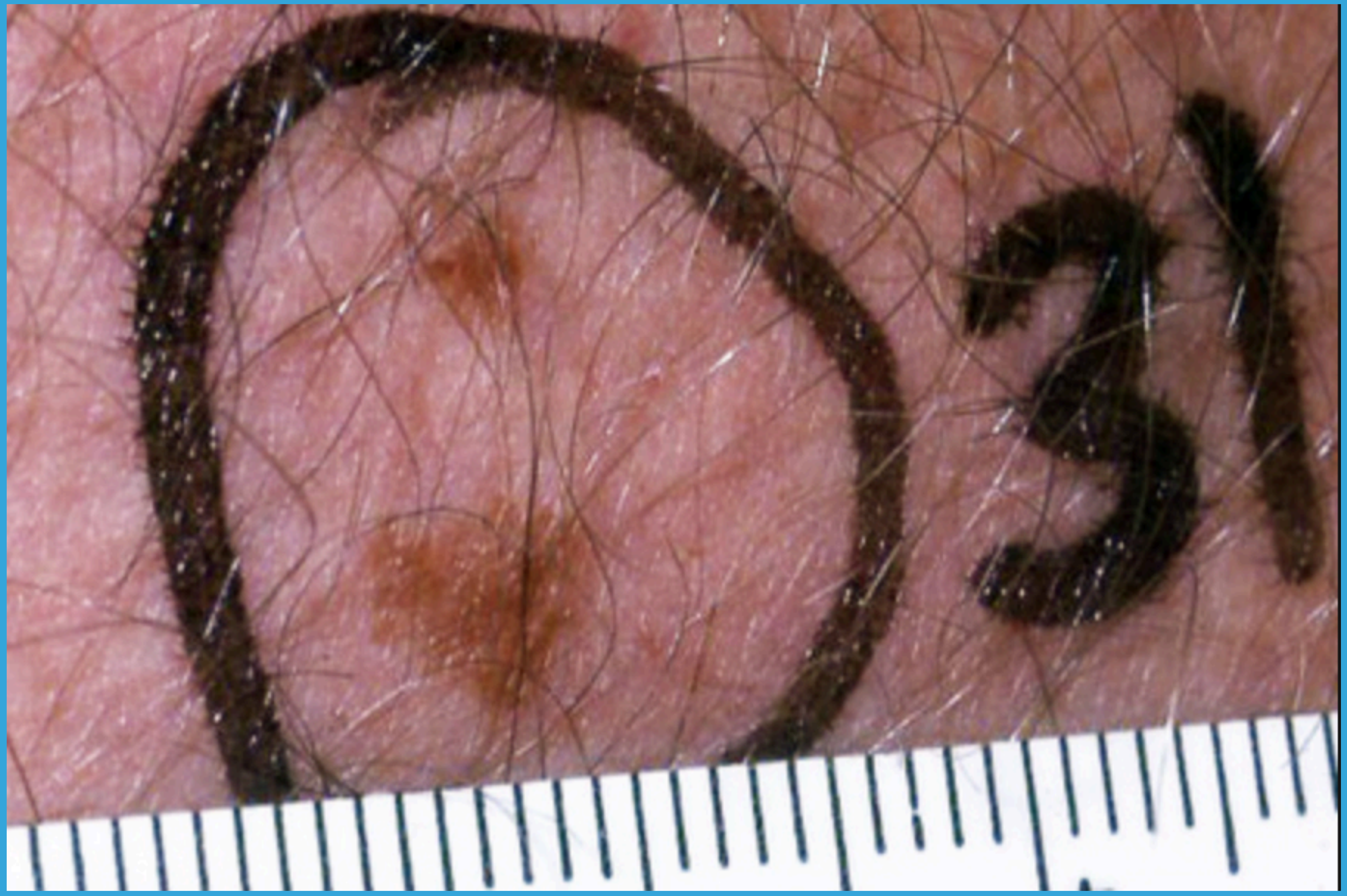
$$mpg = \beta_0 + \beta_1 cyl + disp$$

OR

$$mpg = \beta_0 + \beta_1 cyl + \beta_2 cyl^2 + \beta_3 log(disp)$$

?




31

1. INTERPRETATION MATTERS.
2. LINEARITY ISN'T RESTRICTIVE.
3. MO' SQUIGGLY = MO' OVERFITTING.
4. SMALL DATA'S OK.
5. IT'S ALL ABOUT THE DGP.

@alexkgold

https://github.com/akgold/dsdc_linear_models