What if you're doing regression $\mathcal{Y} \subseteq \mathbb{R}$ and $p = 1$ but the feature is a factor with two levels. Let $x = \mathcal{X} = \left\{\text{red, green}\right\}$. How do you model this? Try linear model.

Let red be represented as 0 and green as 1. Create a binary variable $\tilde{x} \in \left\{0, 1\right\}$. What would the hyper set look like?

$$\mathcal{H} = \left\{w_0 + w_1\tilde{x} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\right\} = \left\{w_0 + w_1\mathbb{1}_{x=\text{green}} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\right\}$$

Therefore the final model is

$$\hat{y} = b_0 + b_1\tilde{x} = b_0 + b_1\mathbb{1}_{x=\text{green}}$$

This model can be fitted with least squares.

$$\hat{y} = \begin{cases} \bar{y}_{\text{red}} & \text{if } x = \text{ red} \\ \bar{y}_{\text{green}} & \text{if } x = \text{ green} \end{cases} = \underbrace{\bar{y}_{\text{red}}}_{b_0} + \underbrace{(\overbrace{\bar{y}_{\text{green}} - \bar{y}_{\text{red}}}^{\Delta\bar{y}})}_{b_1}\mathbb{1}_{x=\text{ green}}$$

Red is called the reference level/category and thus b_1 represents the added effect of green over red.

Proof: Let $p = \frac{1}{n}\sum\mathbb{1}_{x_1=\text{ green}}$ (the proportion of greens); therefore $1-p$ is the proportion of red. Let $b_0 = \bar{y} - b_1\bar{x}$ and assume $b_1 = \bar{y}_g - \bar{y}_r$. Then

$$\bar{y} = \frac{y_1 + \cdots + y_n}{n} = \frac{\overbrace{y_{1g} + \cdots + y_{ng}}^{\text{greens}}}{n} + \frac{\overbrace{y_{1r} + \cdots + y_{nr}}^{\text{reds}}}{n} = \frac{\bar{y}_g n_g}{n} + \frac{\bar{y}_r n_r}{n} = \bar{y}_g\frac{n_g}{n} + \bar{y}_r\frac{n_r}{n}$$

Let $\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{x_{g1} + \ldots x_{gn} + x_{r1} + \cdots + x_{rn}}{n} = \frac{n_g}{n} = p$. Then

$$\bar{y} = p\bar{y}_g + (1-p)\bar{y}_r$$

For b_0:

$$b_0 = p\bar{y}_g + (1-p)\bar{y}_r - p(\bar{y}_g - \bar{y}_r) = (1-p)\bar{y}_r + p\bar{y}_r = \bar{y}_r$$

Now for b_1, note first that

$$\begin{aligned} \sum x_i y_i &= \sum y_{gi} = ng\bar{y}_g \\ n\bar{x}\bar{y} &= np\bar{y} \\ \sum x_i^2 &= n_g \\ n\bar{x}^2 &= np^2 \end{aligned}$$

Then:

$$\begin{aligned}
b_1 &= r\frac{s_y}{s_x} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \\
&= \frac{n_g\bar{y}_g - np\bar{y}}{n_g - np^2} \cdot \frac{1/n}{1/n} \\
&= \frac{p\bar{y}_g - p\bar{y}}{p - p^2} \\
&= \frac{\bar{y}_g - \bar{y}}{1 - p} \\
&= \frac{\bar{y}_g - (p\bar{y}_g + (1-p)\bar{y}_r)}{1-p} \\
&= \frac{\bar{y}_g}{1-p} - \frac{p\bar{y}_g}{1-p} - y_r \\
&= \bar{y}_g - \bar{y}_r
\end{aligned}$$

This is a line connecting the means of green and red where the difference in y is $\bar{y}_g - \bar{y}_r$. What if there were more than 2 levels in the function? For example, $x = \Big\{\text{red, green, blue}\Big\}$. Recall that x can be rewritten as $\Big\{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\Big\}$. Here one variable becomes three (dummy) variables.

$$\begin{aligned}
x_1 &= \mathbb{1}_{x=\text{ red}} \\
x_2 &= \mathbb{1}_{x=\text{ green}} \\
x_3 &= \mathbb{1}_{x=\text{ blue}}
\end{aligned}$$

We cannot use a model on $y \sim x$ here. This leads to multivariate linear regression.