Summary of Week 6

- linear regression w/ >1 predictor
- interpretation of regression coefficients
  - ↳ difference between single and multiple predictors

$$\beta_1' = \beta_1 + \beta_2\left(\mathbb{E}[Y \mid X_2 = x+1] - \mathbb{E}[Y \mid X_1 = x]\right)$$

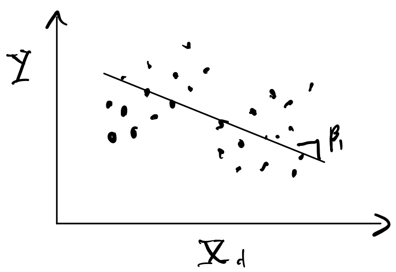
- R^2, p-values etc.
- sample distribution / colinearity → want predictors to be uncorrelated

This week

- Simpson's paradox
- More on math behind multiple predictor regression
- Categorical predictors
- Nonlinear models:
  - interactions between predictors e.g. $Y = X_1 X_2 + \varepsilon$
  - fourier modes e.g. $Y = \sin(x) + \varepsilon$

Simpson's paradox

Look at outcome Y vs predictor X_1

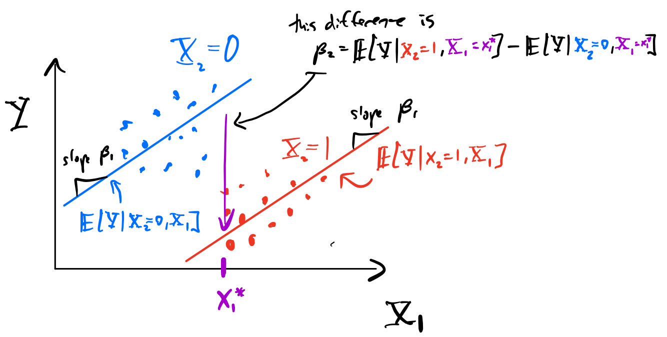


$\Longrightarrow \beta_1' < 0$

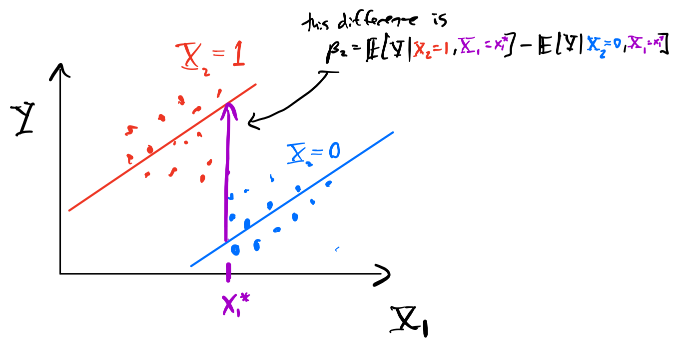
Now add predictor X_2 $\beta_1 > 0$. How can this happen?

$$0 > \beta_1' = \beta_1 + \beta_2\underbrace{\left(\mathbb{E}[X_2 \mid X_1 = x+1] - \mathbb{E}[X_2 \mid X_1 = x]\right)}_{=b} = \beta_1 + \beta_2 b$$

Case 1: $\beta_2 < 0$ AND $b > 0$



Case 2: $\beta_2 > 0$ AND $b < 0$



More Math (see page 6 in week 6 notes)

How are regression coefficients actually computed?

Use covariance like we did with one predictor.

to make things simpler, assume $\mathbb{E}[X_i] = 0$, $\beta_0 = 0$

$$Cov(X_j, Y) = Cov\left(X_j, \sum_{i=1}^{K} \beta_i X_i\right)$$

$$= \mathbb{E}\left[X_j, \sum_{i=1}^{K} \beta_i X_i\right] = \sum_{i=1}^{K} \beta_i \mathbb{E}[X_i X_j]$$

$$= \sum_{i=1}^{K} \beta_i Cov(X_i, X_j)$$

got system of K equations w/ K unknowns $\beta_1 \dots \beta_K$

Replace expectations w/ sample avg. using N data points

$$X = \underset{N \times K}{\begin{bmatrix} X_{1,1}, \dots, X_{1,K} \\ \\ X_{N,1}, \dots, X_{N,K} \end{bmatrix}}, \quad \vec{y} = \underset{N \times 1}{\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}} \quad \vec{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}$$

$$\underset{K \times N}{X^T} \underset{N \times 1}{\vec{y}} = \begin{bmatrix} \sum_i^N Y_i X_{i,1} \\ \vdots \\ \sum_i^N Y_i X_{i,K} \end{bmatrix} = \underbrace{\underset{K \times N}{X^T} \underset{N \times K}{X}}_{\underset{K \times K}{\begin{bmatrix} \sum_{i=1}^N X_{i,1} X_{i,1} & \cdots \\ \sum_{i=1}^N X_{i,1} X_{i,2} & \ddots \\ \vdots & \end{bmatrix}}} \underset{K \times 1}{\vec{\beta}} \Longrightarrow \boxed{\vec{\beta} = (X^T X)^{-1} X^T \vec{y}}$$

Optional material - not on exam

The matrix $X^T X$ is called the <u>covariance matrix</u>

Note that

$$\begin{bmatrix} \mathbb{E}[Y \mid X = x_1] \\ \vdots \\ \mathbb{E}[Y \mid X = x_n] \end{bmatrix} = X\hat{\beta}$$

$(X_{1,1} \dots X_{1,k})$

is vector of predicted averages conditioned on values of response variable

formula ⊛ also come from least squares i.e. $\vec{\beta} = \min \| y - X\beta \|_2^2$

<u>Key point</u>: ideally want $X^T X \approx I$

$$\implies \operatorname{Cov}(X_i, X_j) = 0 \quad i \neq j$$

<u>Categorical predictors</u>

Say we have a predictor like X = someone's race on a survey where you can only select one race

then $X \in \{\text{Black}, \text{white}, \text{asian}, \text{hispanic}, \text{other}\}$

$$\boxed{\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 0 & 1 & 2 & 3 & 4 \end{matrix}}$$

Can't do this because it implies intrinsic ordering. E.g. difference between black → asian less than white → other

Instead we break up into binary 4 variables: $X_{black}, X_{white}, \dots$

and consider

$$Y = \beta_0 + \beta_{black} X_{black} + \beta_{white} X_{white} + \beta_{asian} X_{asian} + \beta_{hispanic} X_{hispanic} + \beta_{other} X_{other} + \varepsilon$$

But this has new problem:

$$\beta_{black} = \mathbb{E}[Y \mid X_{black} = 1, \text{all others } 0] - \underbrace{\mathbb{E}[Y \mid X_{black} = 0, \text{all others } 0]}_{\text{this doesn't have sense because one of the } X = 1}$$

to resolve this we drop one of the predictors.

Python drops the first one in alphabetical order:

$$Y = \beta_0 + \beta_{black} X_{black} + \beta_{white} X_{white} + \boxed{\cancel{\beta_{asian} X_{asian}}} + \beta_{hispanic} X_{hispanic} + \beta_{other} X_{other} + \varepsilon$$

Selected as baseline

$$\downarrow$$

$$Y = \beta_0 + \beta_{black} X_{black} + \beta_{white} X_{white} + \beta_{hispanic} X_{hispanic} + \beta_{other} X_{other} + \varepsilon$$

in this model okay to have all predictors $= 0$ because this implies $X_{asian} = 1$, which is not in the model

Now $$\beta_{black} = \mathbb{E}[Y \mid X_{black} = 1, \text{ all others} = 0] - \mathbb{E}[Y \mid \underbrace{X_{black} = 0, \text{ all others} = 0}]$$

all $X = 0$ $\Rightarrow$ this is avg. of Y among those samples who selected asian on survey

all regression coefficients should be interpreted as avg. differences between given category (e.g. β_{black} for Black) and baseline category (e.g. asian)