Linear Regression *(find and work with a partner)*

In our linear regression setup, we are given a matrix $\boldsymbol{X}$ which consists of n data points, each with p features (also called predictors or explanatory variables), plus a "1" in the first column. Each data point $\boldsymbol{x}_i$ has an associated response variable y_i. Together these form $\boldsymbol{y}$. In multiple linear regression, $p > 1$ and our model can be described by a vector of coefficients, $\boldsymbol{b}$. In matrix-vector form, we have

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} — & \boldsymbol{x}_1 & — \\ — & \boldsymbol{x}_2 & — \\ & \vdots & \\ — & \boldsymbol{x}_n & — \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

where each example $\boldsymbol{x}_i = \big[1, x_{i1}, \ldots, x_{ip}\big]$. Our linear regression model is:

$$h_{\boldsymbol{b}}(\boldsymbol{x}) = \boldsymbol{b}^T \boldsymbol{x}$$

Beginning with the simple linear regression case ($p = 1$), we have:

$$h_{\boldsymbol{b}}(\boldsymbol{x}) = b_0 + b_1 x$$

1. *Toy example.* Let $n = 2$ and $p = 1$, with the following data (we will omit the first column of 1's in simple linear regression):

$$\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

   (a) Plot these two points – what should $\hat{b}_0$ and $\hat{b}_1$ be?

   (b) This week we derived the solution for simple linear regression:

$$\hat{b}_1 = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{Var}(\boldsymbol{x})} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

   where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Use these equations to compute $\hat{b}_0$ and $\hat{b}_1$ and verify your answer to (a).

2. In linear regression, we seek to minimize the sum of squared errors between the actual response and our prediction. We often call this RSS (residual sum of squares) or SSE (sum of squared errors). As an objective function, we often call it J and include a $\frac{1}{2}$ in front to make the derivatives work out nicely.

$$J(\boldsymbol{b}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{b}}(\boldsymbol{x_i}) - y_i\right)^2$$

(a) For the toy example on the previous page, the stochastic gradient descent updates are:

```
for i = 1,2:
```
$b_0 \leftarrow b_0 - \alpha(b_0 + b_1 x_i - y_i) \cdot 1$
$b_1 \leftarrow b_1 - \alpha(b_0 + b_1 x_i - y_i) \cdot x_i$

Assuming $\alpha = 0.1$ and our initial values are $b_0 = 0$ and $b_1 = 0$, what are b_0 and b_1 after the first step of gradient descent (i.e. first pass through all the data points)?

(b) What is the value of the objective function (cost) after this initial iteration?

(c) Use the analytic solution (normal equations) to verify your results from the first page.