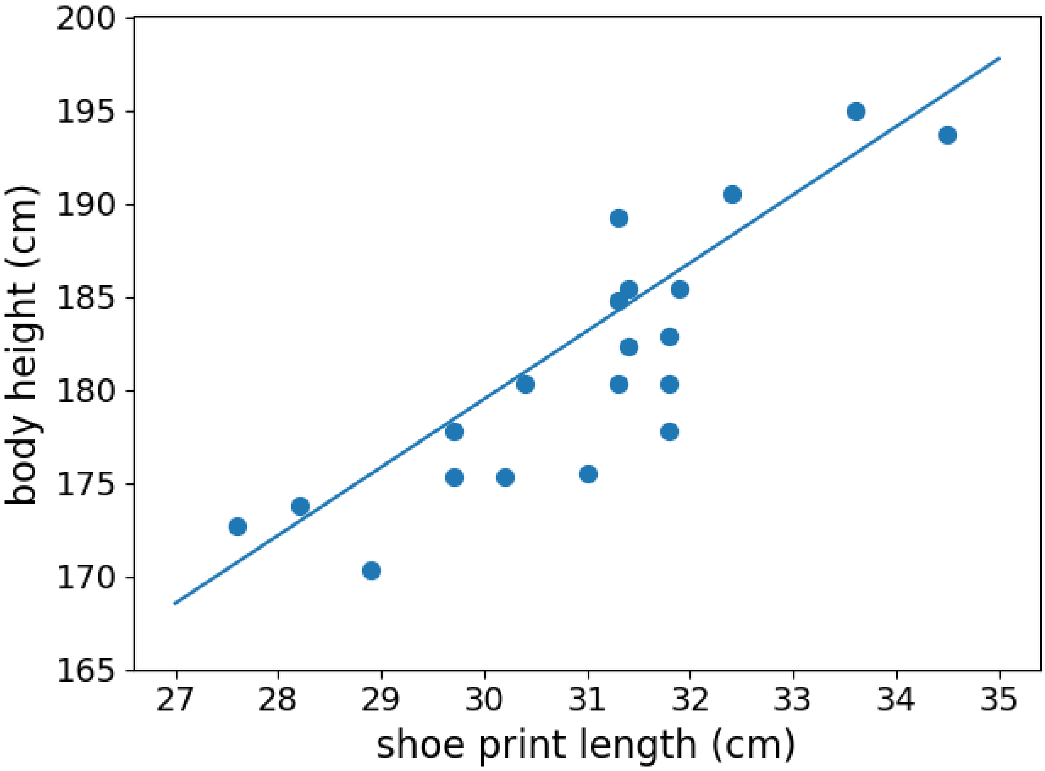
CS5228 – Tutorial 7

Classification & Regression III (Linear Models)

1. **Linear Regression.** The figure below shows the CSI example from the lecture, where we try to predict the height of a suspect given the length of a shoe print we found. The blue line represents the best fit of a Linear Regression Model.



   (a) The most commonly used loss function L for Linear Regression is the Means Squared Error (MSE):

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

   What makes MSE a suitable loss function for Linear Regression, and what part of the formula might cause "issues"?

   (b) After training a Linear Regression model over a dataset, how can we use the result to identify if there is indeed some linear relationship between the input features and the output values.

(c) Let's assume our data features strong linear relationships between the input features in output values, and we train a Linear Regression Model. Now we want to predict the values for unseen data points. In which case (i.e., for which type of data points) we might have to be very cautious when interpreting the predicted values?

(d) For our CSI example, could we train a Linear Regression model that only requires the coefficient of our single feature *length* and does not require the bias weight?

(e) In the lecture, we only skimmed over the calculation of the derivative of loss function L. Given the matrix notation of L

$$L = \frac{1}{n} \|X\theta - y\|^2$$

find $\frac{\partial L}{\partial \theta}$!

2. **Logistic Regression (optional).** We have seen that Linear Regression and Logistic Regression are very similar. Simply speaking, Logistic Regression uses the basic linear signal $\theta^T x_i$ and puts it through a function f that maps this signal to a value of range $[0, 1]$ so that the final value can be interpreted as a probability. The Sigmoid $\sigma(z)$ function is most commonly used to implement f:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

With $\theta^T x_i$ being our linear signal, we can calculate the predicted value $\hat{y}_i$ for data sample x_i as follows:

$$\hat{y}_i = \frac{1}{1 + e^{\theta^T x_i}}$$

This ultimately brought us to the *Cross-Entropy Loss*, the loss function for Logistic Regression (binary classification):

$$\begin{aligned} L &= -\frac{1}{n} \sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)] \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log \frac{1}{1 + e^{\theta^T x_i}} + (1 - y_i) \log (1 - \frac{1}{1 + e^{\theta^T x_i}}) \right] \end{aligned}$$

Well, find $\frac{\partial L}{\partial \theta}$! :)