15 Example (Optimal Bias)
Show that for a simple bias network, the bias, b, that minimizes the MSE (L^2 loss) is equal to the mean of the target values, i.e. the optimal b is equal to $\bar{y}$.

Hint: Set the derivative of the MSE loss function equal to zero.

16 Definition (Root Mean Square Error)
The square root of MSE is called the Root Mean Square Error or RMSE.

$$RMSE = \sqrt{MSE}$$

17 Definition (Variance and Standard Deviation)
The **sample variance**, s^2, of a set of n numbers, $y_1, y_2, \cdots, y_n$, is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The **sample standard deviation** is s.

18 Definition (Baseline Performance)
In deep learning it is often important to establish a **baseline performance** to compare to. One of the simplest baseline performances to compute for regression networks is the minimum RMSE of a simple bias network. If you are unable to do better than this simple baseline, you have serious issues to deal with. We show below that the standard deviation of the target values is a good estimate of the optimal performance of a simple bias network.

19 Example (Estimating Performance of Simple Bias Network)
Show that the standard deviation of the target values is a good estimate for minimum RMSE of a simple bias network.

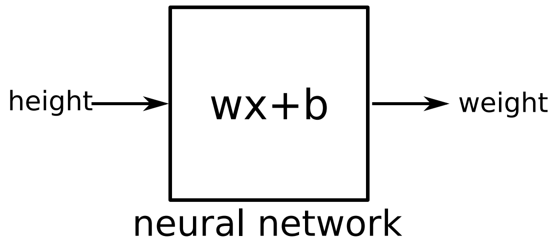
Before training a neural network, its a good idea to establish a baseline for prediction accuracy. If your neural network can't beat the simple bias network, you have a poor result.

2 Simple Linear Regression

20 Example (Simple Linear Regression)
What is the weight of a child who is 70 inches tall? Below is a table of the heights (in inches) and the weights (in pounds) of 10 randomly selected children.[2]

height	weight
67	128
67	123
72	129
69	143
69	132
70	142
67	112
67	118
66	108
68	119

Unlike the previous examples, we now have a feature, the height of a child, that we can use to predict the weight of the child. We will train the neural network shown below.



The variable x represents the value of a feature (in this case height) and w and b are the parameters of the network that must be learned from

[2] Data source: SOCR Data

the data.[3]

21 Definition (Dot Product)
Let $\mathbf{u} = \big(u_1, u_2, \ldots, u_n\big)^\top$ and $\mathbf{v} = \big(v_1, v_2, \ldots, v_n\big)^\top$ be two vectors with n components. The dot product of **u** and **v** is given by

$$\mathbf{u} \circ \mathbf{v} = \mathbf{u}^\top \mathbf{v} = u_1v_1 + v_2v_2 \ldots + u_nv_n.$$

22 Example (Dot Product)
Let $\mathbf{u} = \big(5, 3, 1\big)^\top$ and $\mathbf{v} = \big(-1, 2, 1\big)^\top$. Compute $\mathbf{u} \circ \mathbf{v} = \mathbf{u}^\top \mathbf{v}$.

23 Definition (Matrix Multiplication)
Assume A has m rows and n columns and B has n rows and r columns. Let C be the matrix product of A and B.

$$A_{m\times n}B_{n\times r} = C_{m\times r}.$$

1. The inner dimensions must match for the matrices to be compatible for multiplication.
2. The resulting matrix, C, will have the outer dimensions.
3. Each i, j element in the matrix C is a dot product of row i of A with column j of B.

24 Example (Matrix Multiplication)

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 1 & 1 \end{pmatrix}_{3\times 2} \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}_{2\times 2} = \begin{pmatrix} 3 & -1 \\ 4 & 0 \\ 5 & -1 \end{pmatrix}_{3\times 2}$$

25 Example (Matrix Form of Simple Linear Regression)
Show that the simple linear regression network can be expressed in matrix form as the function

$$\mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{w} = \hat{\mathbf{y}}$$

where

$$\mathbf{X} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \ \mathbf{w} = \begin{pmatrix} w \\ b \end{pmatrix}, \ \hat{\mathbf{y}} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{pmatrix}$$

Note: The hat is used to indicate that $\hat{\mathbf{y}}$ is predicted value for **y**.

26 Definition (Gradient Vector)
The gradient vector of a scalar-valued function $f(\mathbf{w})$ is denoted by $\frac{df}{d\mathbf{w}}$ and is given by

$$\frac{df}{d\mathbf{w}} = \left(\frac{\partial f}{\partial w_1}, \quad \frac{\partial f}{\partial w_2}, \quad \ldots, \quad \frac{\partial f}{\partial w_n}\right)^\top.$$

27 Example (Gradient Vector)
Compute the gradient vector of the function

$$f(\mathbf{w}) = f(w_1, w_2, w_3) = w_1^2 + w_1w_2 + 2w_2w_3^2.$$

28 Example (Normal Equations)
Show that the parameter values, **w**, that minimizes the L^2 loss function $MSE(\mathbf{w})$ where

$$MSE(\mathbf{w}) = \frac{1}{n}(\hat{\mathbf{y}} - \mathbf{y})^\top(\hat{\mathbf{y}} - \mathbf{y})$$

is given by the solution to the **normal equations**

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y}.$$

[3] w should not be confused for the weight of the child. The weight of the child is the output of the network, y.