一元线性回归模型

董峦 新疆农业大学

引言

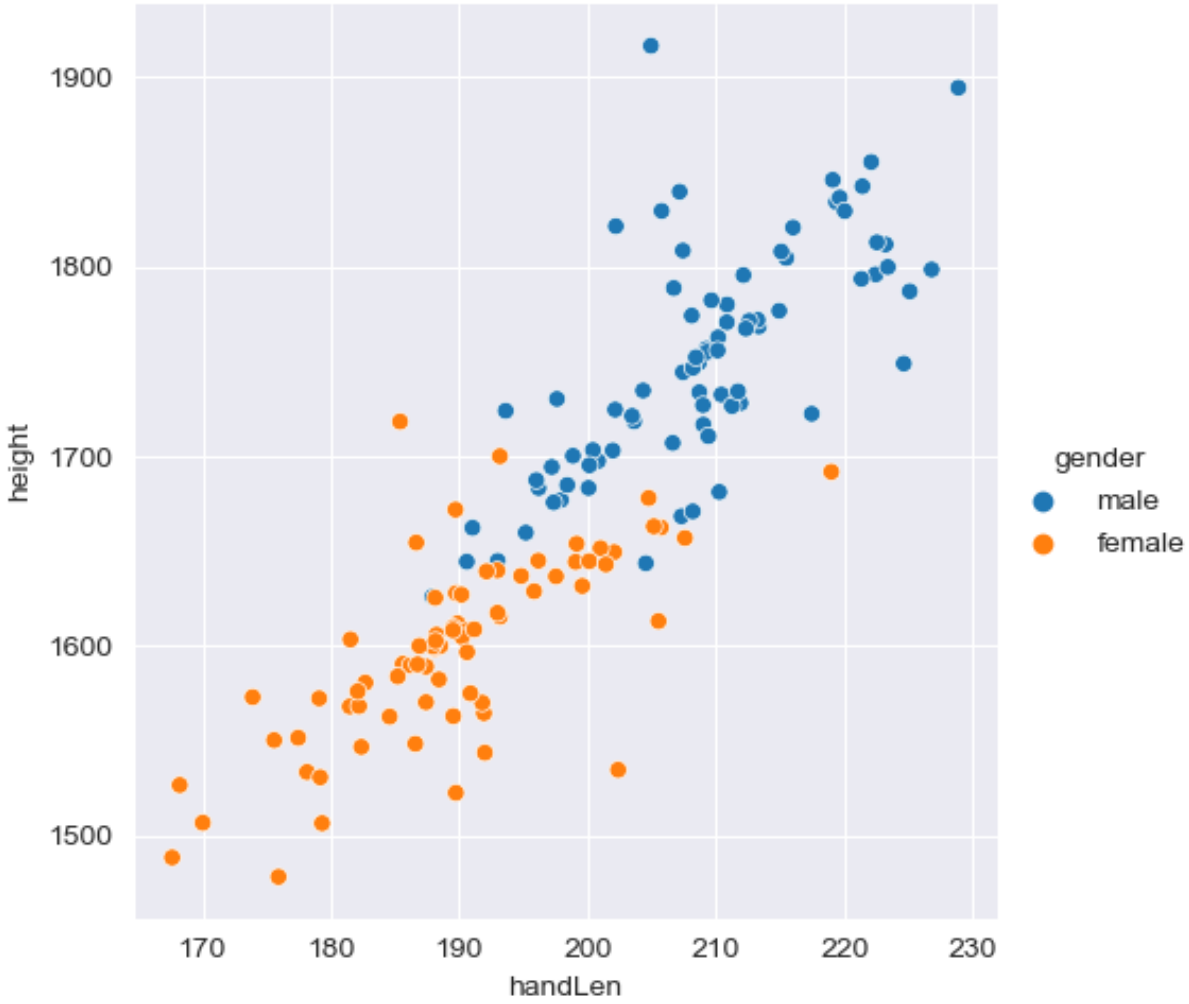
根据百度百科中“步法追踪”条目的描述，身高与脚印长度的关系是：身高（厘米）= 脚印长度（厘米）* 6.876。本人脚长约26.8厘米，按照上述公式计算出身高是184厘米，与实际情况几乎吻合。在很多影视作品和真实案件回顾里，经常有办案人员根据犯罪现场脚印推测嫌疑人身高的场景。这样看来脚长与身高似乎呈现稳定的关系，那么现实中的数据支持这个假设吗？下文分析一个特定数据集里脚长和身高的数据，讨论从脚长推测身高是否可行。

数据集

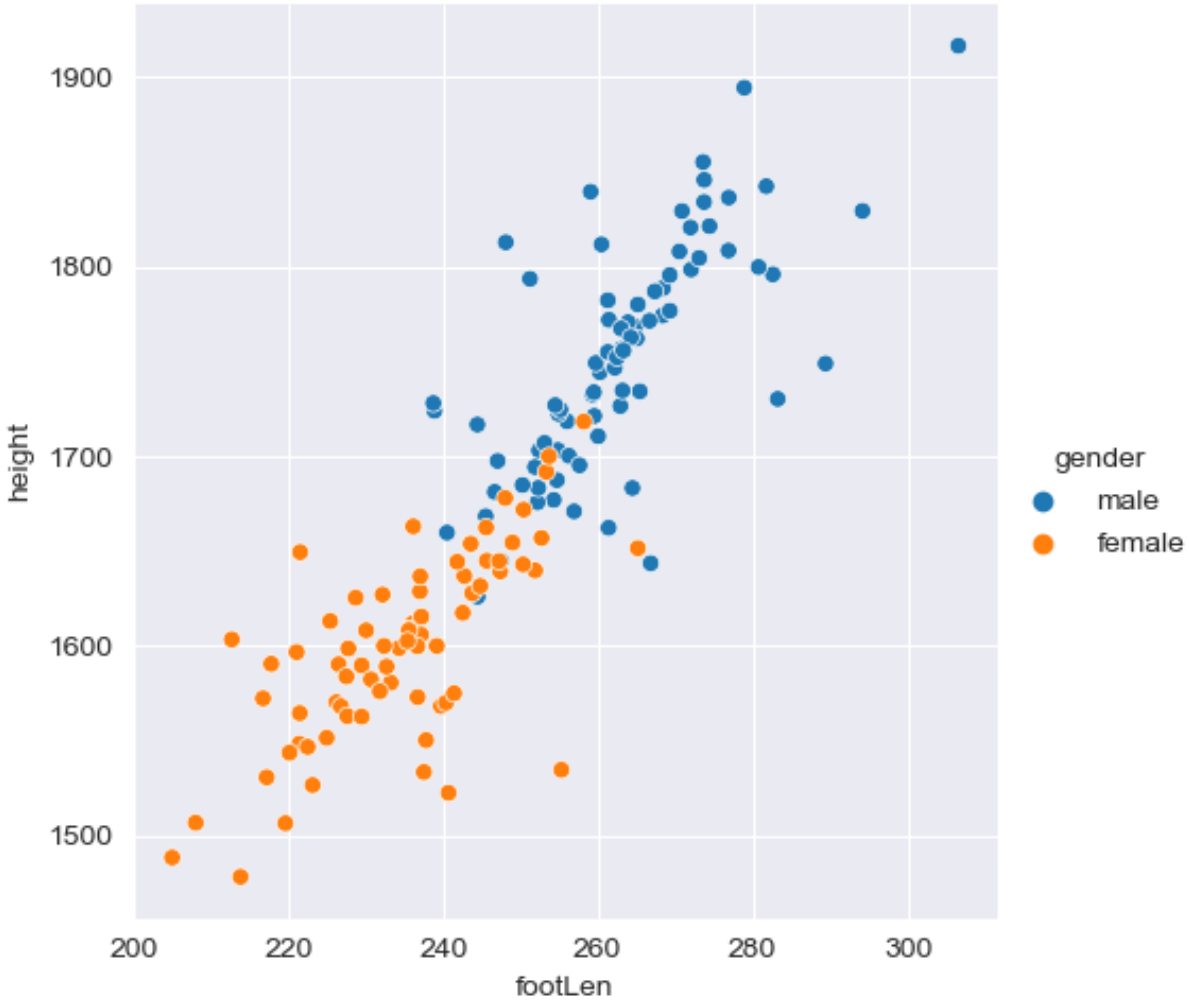
在 http://users.stat.ufl.edu/~winner/datasets.html 处有这样一个数据集：`Stature, Hand Length, and Foot Length Among Males and Females` 。根据该数据集的描述特别是研究该数据集的论文https://onlinelibrary.wiley.com/doi/abs/10.1002/ca.20146 可以知道，该数据集包含155个土耳其成年人（17~23岁）的手、脚及身高数据，其中男性和女性的数据分别有80和75个。数据是 .csv 格式，该数据集已组织成结构化形式，数据有5列，前十条数据如下图所示。每一列含义从表头可以判断出大概。第一列是序号，第二列用数值代表性别（1代表男性，2代表女性），第三列是身高数据，后两列分别是手和脚的长度数据。后三列数据单位都是毫米。

idGen	gender	height	handLen	footLen
1	1	1829.356	205.7455	294.1762
2	1	1730.262	197.5946	283.2275
3	1	1723.99	193.6019	238.9127
4	1	1748.885	224.5994	289.407
5	1	1697.573	200.8021	247.0631
6	1	1716.742	208.9908	244.4497
7	1	1795.898	222.4133	282.6041
8	1	1821.308	202.1793	274.4192
9	1	1626.03	187.8969	244.4852
10	1	1744.33	207.4033	260.2666

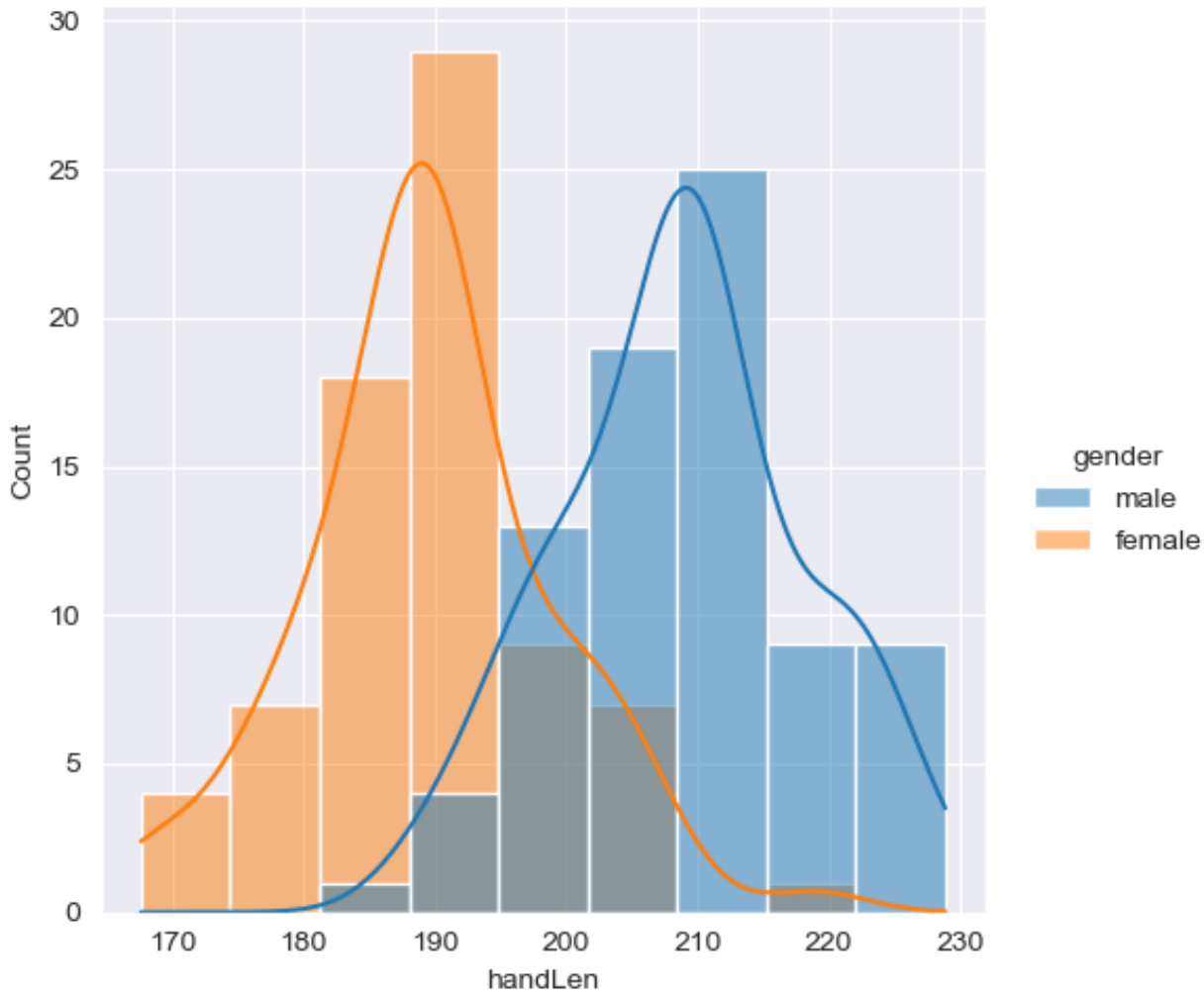
下面用可视化技术做数据的探索性分析。手与身高的关系如下图所示：

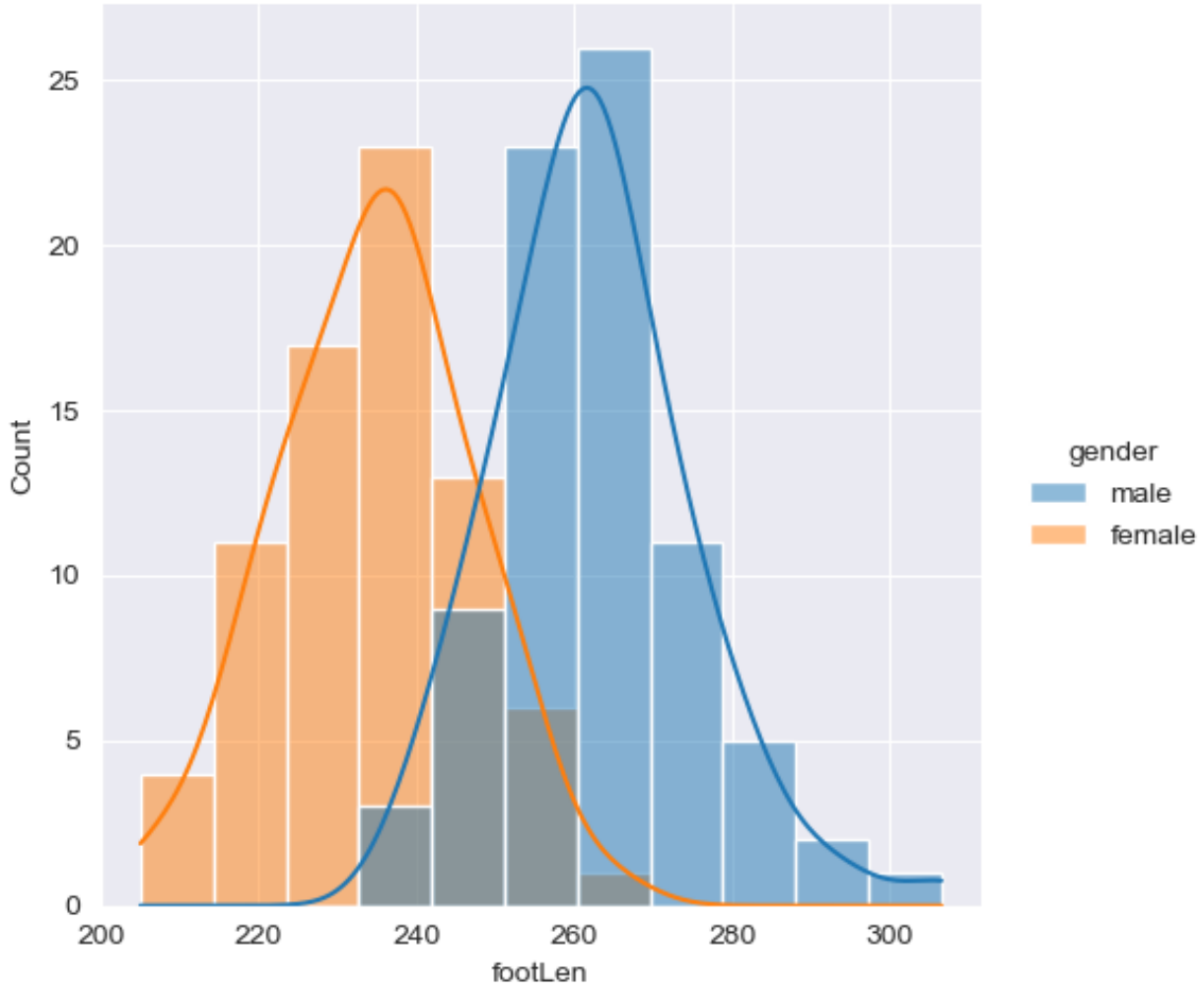

1900
1800
1700
1600
1500
height
170
180
190
200
210
220
230
handLen
gender
male
female

脚与身高的关系是：



可见男性在手、脚和身高上的数据总体大于女性，这与常识一致；手、脚的长度与身高成正比。手、脚尺寸的分布如下：


30
25
20
15
10
5
0
Count
170
180
190
200
210
220
230
handLen
gender
male
female



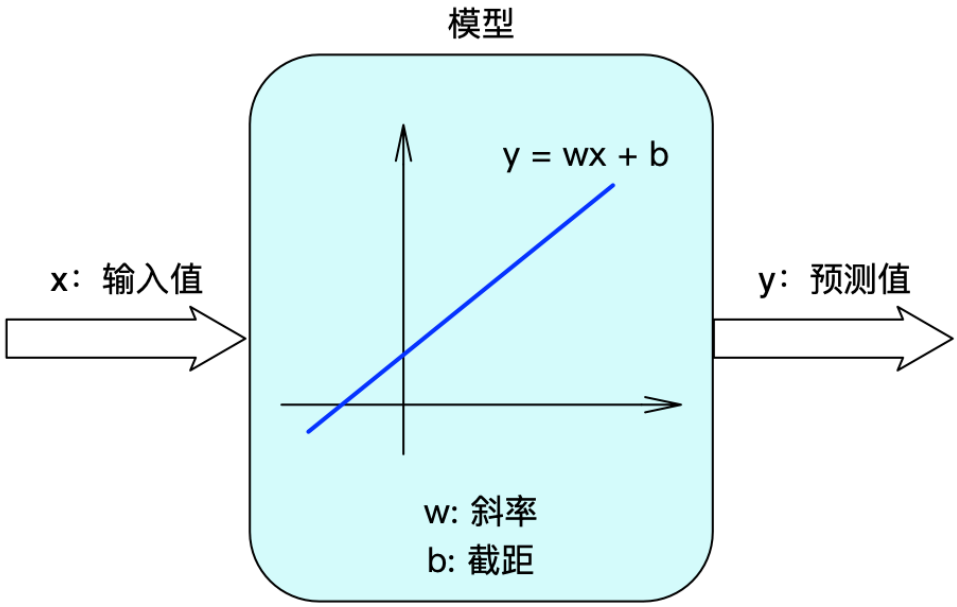
可见男性和女性在手、脚尺寸上的分布虽然都大致呈正态分布，但分布的参数（均值与标准差）是不同的。在机器学习中，从一个数据分布中得到的模型难以应用到具有不同分布的数据上，因此本文分性别讨论手、脚与身高的关系。

模型

根据上面散点图传达的信息，本文假设脚长与身高呈线性关系，即

$$y = wx + b$$

这就是本文的**模型**。其中 x 是**特征**（feature），在这里指脚长。 y 是**预测值**，在这里指身高，w、b 是参数（取 weight 和 bias 单词的首字母），所谓模型一般是指用数学语言描述的数据内在关系。确定了 w 和 b 的值，只要把脚长数据输入该模型，就能够推算身高，如下图所示：



在本例中机器学习的目标是通过数据得出 w 和 b，而得到这些参数的过程叫做“学习”。

上述模型只有一个自变量 x，自变量与应变量 y 的关系是线性的，因此该模型被称为一元线性回归模型，回归问题通俗地讲就是预测一个实数，在该模型中身高 y 是模型的输出值（或预测值），是一个实数，所以该模型是回归模型。从解析几何角度看，模型 $y = wx + b$ 是高中阶段学习过的直线方程，其中 w 是斜率，b 是截距。

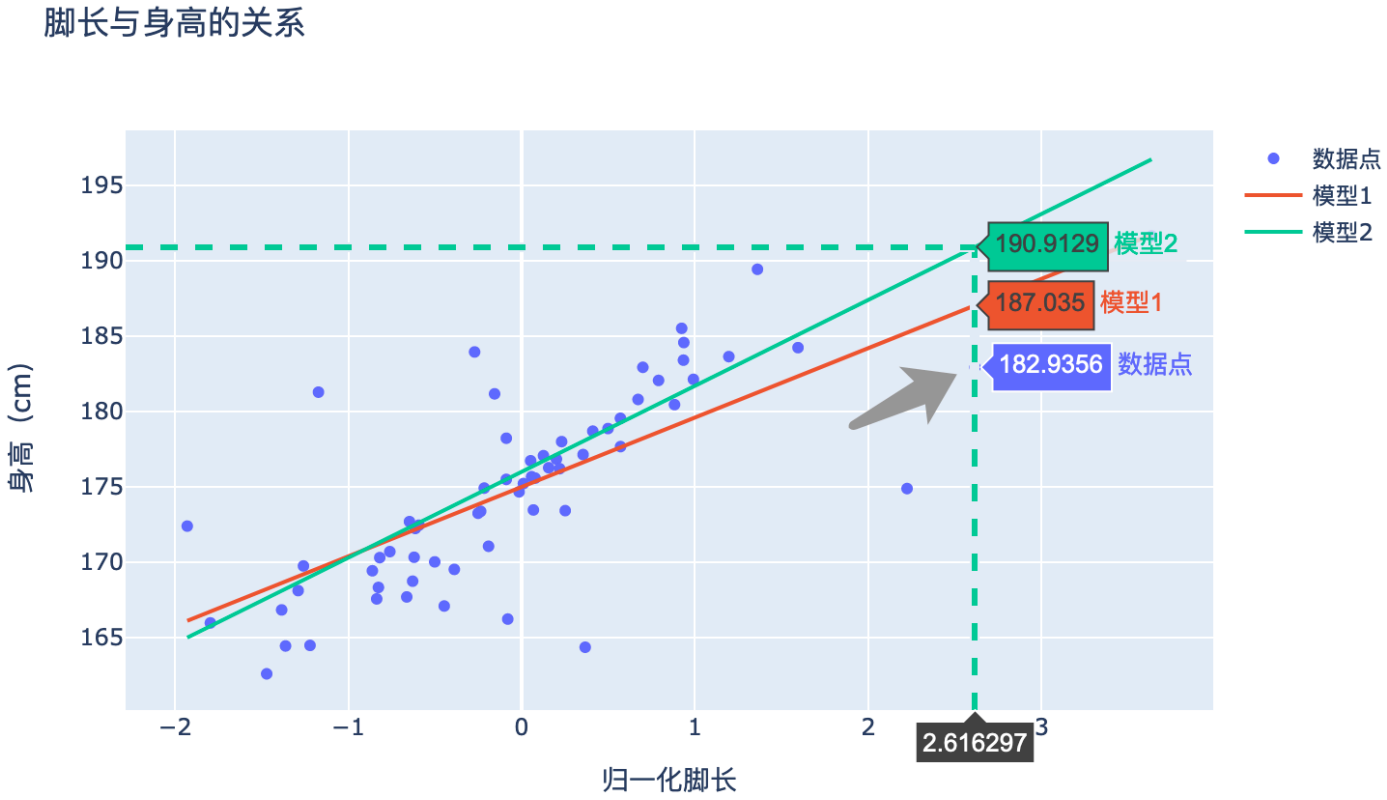
损失函数

损失函数（loss function）计算出一个标量用来评价模型的质量。本文采用均方误差（Mean Square Error，MSE）函数作为损失函数，该误差是每个数据点上预测值与实际值误差平方的均值，这么说有些拗口，用数学语言表述如下：

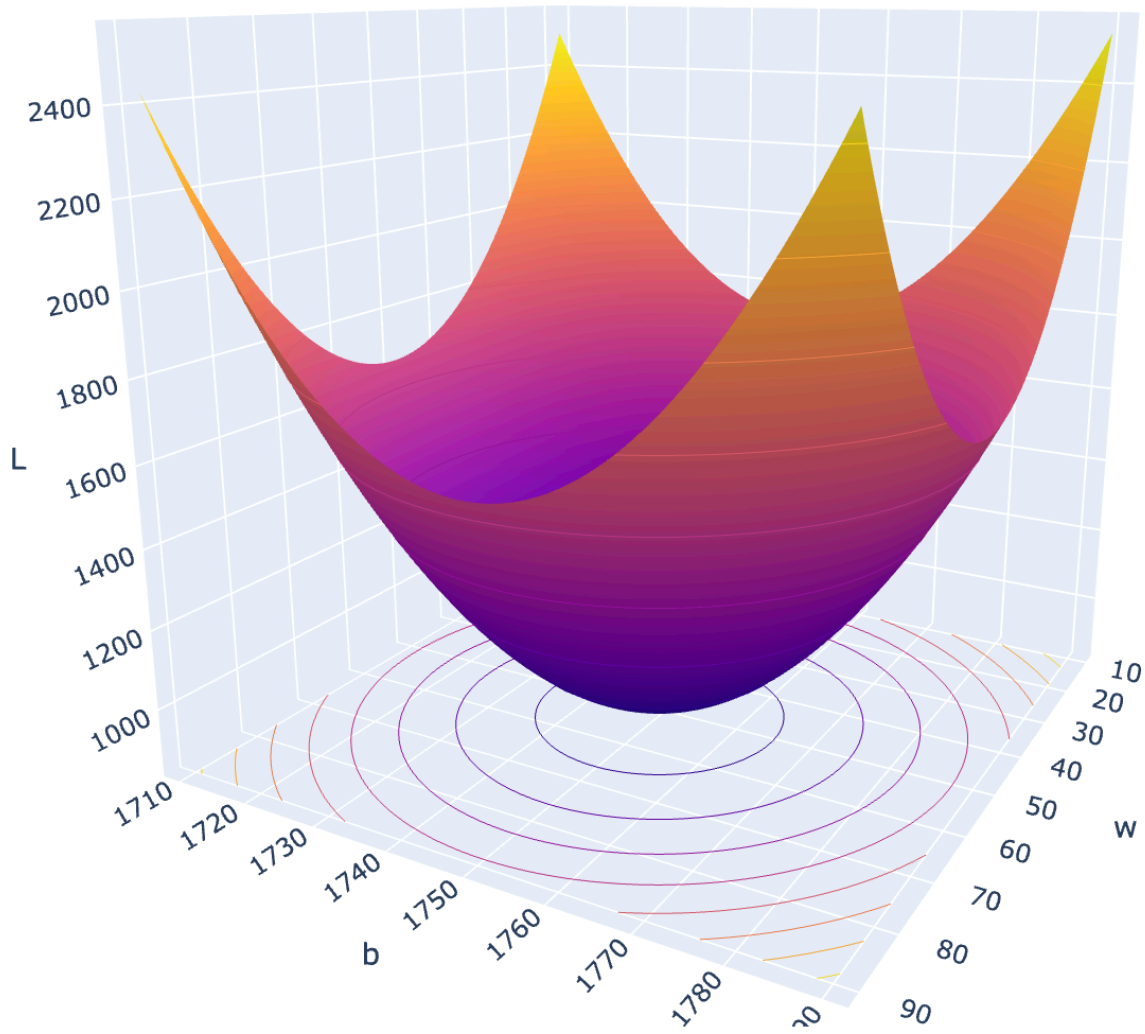
$$
\begin{aligned}
L &= \frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \\
&= \frac{1}{2m}\sum_{i=1}^{m}(wx_i + b - \hat{y}_i)^2
\end{aligned}
$$

其中，y_i 和 $\hat{y}_i$ 分别是第 i 个数据点的预测值与**目标值**（target），每个特征和目标值的组合 $(x_i, \hat{y}_i)$ 是一个**样本**（sample），m 是样本个数，L 代表损失。L 越小说明模型预测的越准确，在机器学习领域把描述模型性能的函数叫做损失函数或代价函数，这里的L 即是本文的损失函数。在损失函数前添加 $\frac{1}{2}$ 是为了在求导过程中让梯度表达式更为简洁，并不影响学习过程。

对于均方误差的含义，以下图中 $x = 2.616297, \hat{y} = 182.9356$ 这个数据点为例，模型1的预测值是187.035，模型2的预测值是190.9192，模型1在这个数据点的预测误差是$(187.035 - 182.9356)^2$，模型2在这个数据点的预测误差是$(190.9129 - 182.9356)^2$，把所有数据点的预测误差取平均值就得到了均方误差，通过该误差衡量哪个模型更好。



注意，当数据给定后损失函数 $L = f(w, b)$ 是关于模型参数 w 和 b 的函数，L 在参数空间的图像如下图所示。可见L 是凸函数，存在全局极小值点，机器学习的目标是利用某种优化方法获得该最小点对应的 w 和 b 。



然而代价函数有如此形态是有条件的，那就是特征经过了标准化（standardization），成为均值为 0 标准差为 1 的特征：

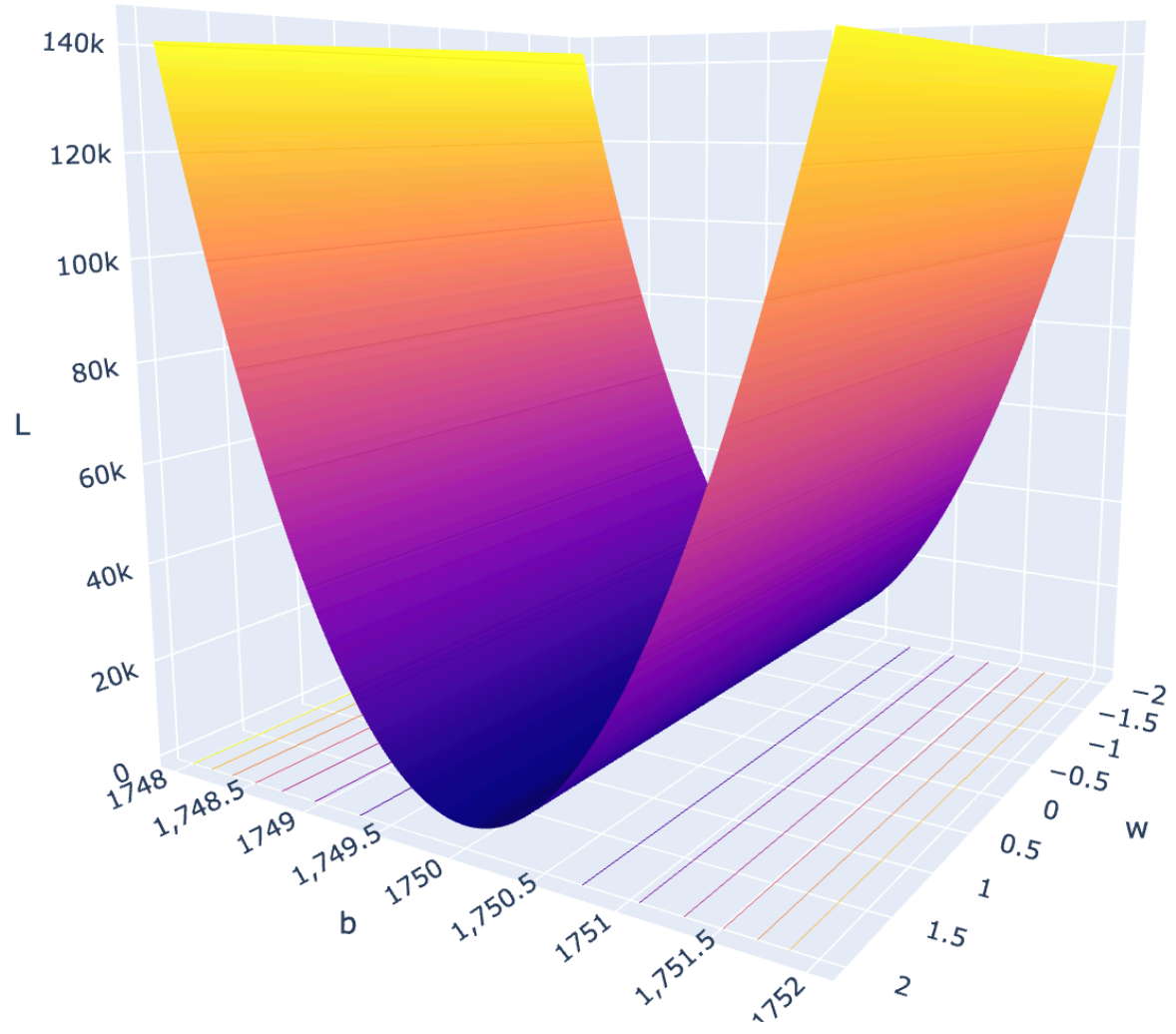
$$x_{standardization} = \frac{x - \mu_x}{\sigma_x}$$

如果特征未经过标准化，则代价函数在参数空间的形态如下图所示，像一张弯折的纸张。虽然全局极小值点仍是存在的，但从等高线的形态可以想象该曲面底部十分平坦、又扁又长，因此极小值点极难获取，优化难度很大。这是对特征做标准化的原因所在。从上面对手、脚尺寸数据的可视化可以看出，数据分布呈现正态分布特点，因此对特征做标准化是合适的。

特征的预处理除了标准化还有零均值化（zero mean）和归一化（normalization），前者是特征减去它的均值，后者是

$$x_{normalization} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

即把特征缩放到 [0, 1] 区间。实践中，标准化和零均值化用的较普遍。



优化方法

从上一节对损失函数的分析可知该函数有一个全局极小值点，根据微积分课程里的知识可知，在此处L的**梯度**为零，因此求解以下方程组可以直接求解w和b。然而这种方法的缺点是参数数量较多时，涉及资源消耗较大的矩阵求逆运算。

$$\frac{\partial L}{\partial w} = \frac{1}{m}\sum_{i=1}^{m}(wx_i + b - \hat{y}_i)x_i = 0$$

$$\frac{\partial L}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(wx_i + b - \hat{y}_i) = 0$$

本文介绍一种机器学习领域经常使用的优化方法——梯度下降（gradient descent）。梯度是一个向量，表示为

$$\nabla L = \begin{bmatrix} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \end{bmatrix}$$

其含义是L增加最快的方向，因此沿着梯度的反方向一步步移动一定能找到那个全局极小值点。**梯度下降算法**是

$$\Theta_{t+1} = \Theta_t - \alpha \nabla L_t$$

其中 Θ 指模型参数，展开来就是

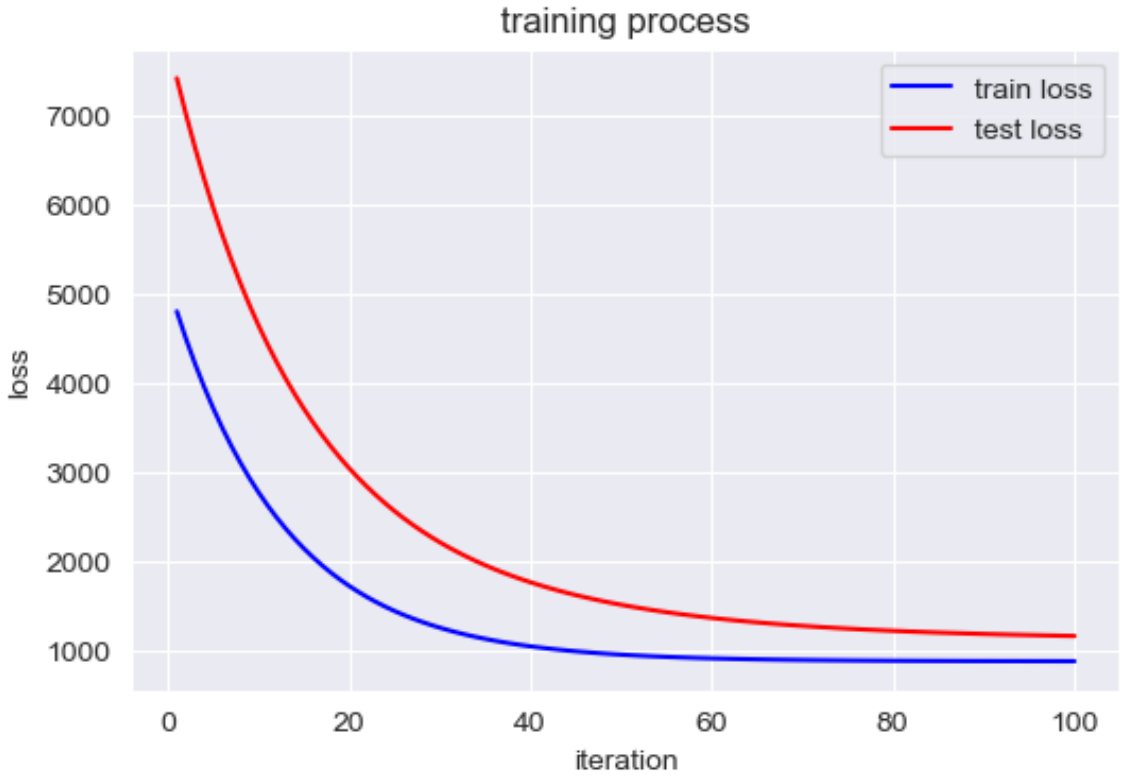
$$\begin{bmatrix} w_{t+1} \\ b_{t+1} \end{bmatrix} = \begin{bmatrix} w_t \\ b_t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_t} \\ \frac{\partial L}{\partial b_t} \end{bmatrix}$$

该迭代式的意思是：将模型参数更新为当前参数减去当前的（第 t 步）梯度乘以一个实数 α。这种操作体现了“沿着梯度的反方向一步步移动就能到达极小值点”的思想，当移动到极小值点时梯度为零，模型参数根据上述迭代式将不再更新，优化宣告结束。然而这是个理想情形，实际情况是优化过程将在最优点处“震荡”，原因在α上。

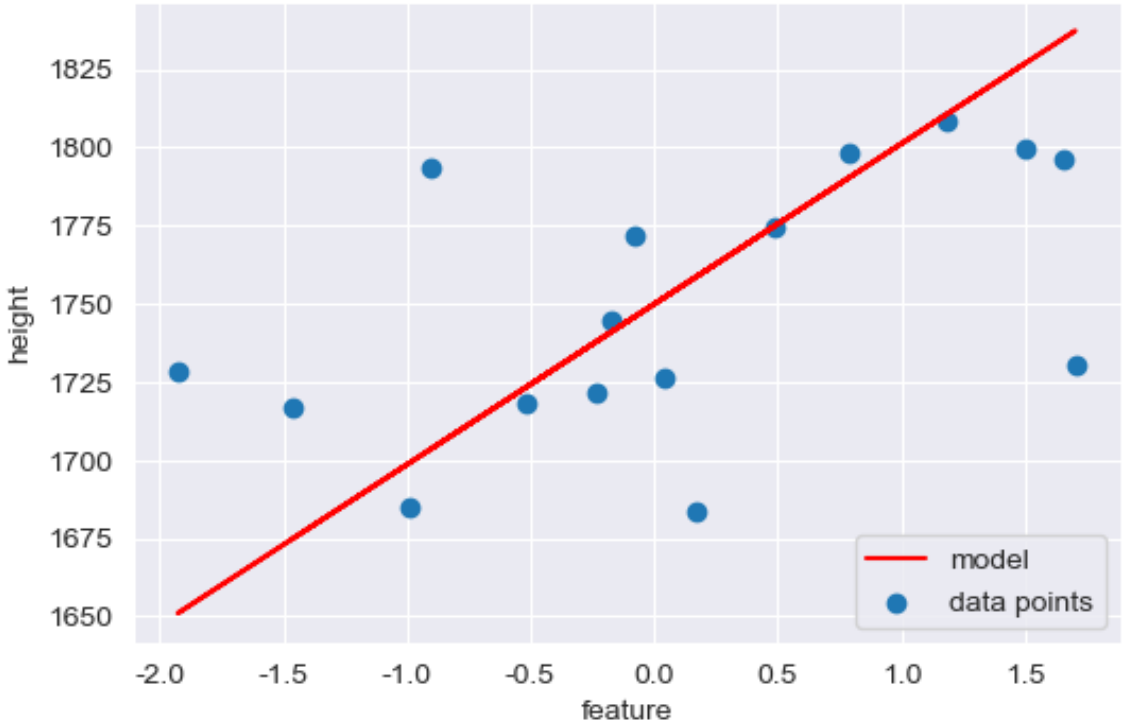
α 叫学习率（learning rate），它的值如果太小则优化过程将变得缓慢，因为向极小值移动的“步子”小了；它的值过大则优化过程将变得不稳定，因为“步子”太大可能造成优化过程不能收敛，甚至是计算出现溢出。

学习率的自适应过程是很多研究的关注点，这方面影响较大的方法是 Adam。设定学习率没有定式，如果取0.1时训练过程不稳定则取0.01试试，经过几次尝试便能确定一个当前问题下较好的数值。

将训练损失和测试损失显示出来如下图所示可以看出在约60次迭代后，模型就收敛了。测试损失的曲线与训练损失类似，但略高于后者。根据训练损失和测试损失的曲线可以推断是否发生了过拟合（overfitting）或欠拟合（underfitting），这方面内容将在后面的内容中进行讨论。下图反映的情况是一种正常情形。



经过训练模型的 $w = 51.36, b = 1749.9$，该模型与测试数据的关系如下图所示



测试模型

笔者脚长 262mm，用训练好的模型预测身高是 1748.36mm，比实际身高要低将近 10cm，这是怎么回事呢?

原因主要是该数据来自土耳其，由于人种上的差异，在该数据集上得到的模型不适用于亚洲人。类似的现象还出现在用于白人的脸部识别算法不一定在亚洲人脸上有效。由于数据分布不同造成的算法偏见（algorithmic bias）在机器学习里比较普遍，这是机器学习算法开发者需要严肃对待的问题。

对外汇报/宣传模型的性能需要报告模型在某个指标（criteria）上的数值，对于回归模型通常采用均方根误差（Root Mean Square Error, RMSE）或平均绝对误差（Mean Absolute Error, MAE），两者的定义分别是

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

小结

本文在一个具体数据集上的分析显示脚长与身高存在相关性，因此用脚长预测身高是可行的。本文用一元线性模型描述数据，以梯度下降方法得到最优的模型参数。

在笔者身高的预测上，上述模型的预测误差较大，除了数据分布的原因，另一个可能的原因是数据里存在离群点（outlier）。在该数据集里有一些数据点偏离主流数据，比如有一些脚很大但身高并不明显或者脚小但身高明显的数据点，当数据量较小时离群点对模型的训练影响较大，反之当数据量较大时离群点的影响就显得微弱了。因此数据量是影响模型质量的一个重要因素。

对于一元线性模型，由于是根据全部训练数据进行学习，而不是根据局部数据，所以彻底克服离群点的影响比较困难。反之，K近邻这类依靠局部数据进行预测的模型对离群点数据有较好的适应性。

思考与实践

- 用女性的数据得出的模型与本文基于男性数据的模型相比会有什么不同?
- 根据手长预测身高和用脚长预测身高哪种更准确?
- 数据不进行标准化时模型训练是怎样的过程？使用零均值化呢?
- 当代价函数在参数空间存在鞍点或局部极小值点时，凭借梯度下降算法能不能到达全局极小值点?