Linear Regression

Week 4

Agenda

- Simple linear regression
- Multiple linear regression
- Parameter estimation
- Goodness of fit

Example

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8

Simple Linear Regression

We assume a model:

$$Y = w_0 + w_1 X + \varepsilon$$

Given some estimates $\widehat{w}_0$ $\widehat{w}_1$, we can predict future sales using

$$\hat{y} = \widehat{w}_0 + \widehat{w}_1 x$$

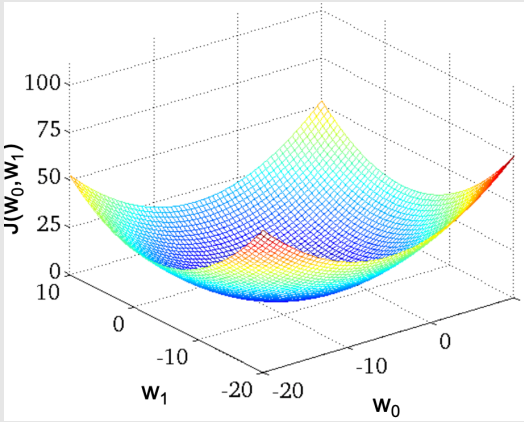
For observation (x_i, y_i), the prediction error/residual is

$$e_i = y_i - \hat{y}_i$$

Minimize the mean squared error/loss

$$MSE = J(w) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{w}_0 - \widehat{w}_1 x_i)^2$$

Estimate the parameters



Minimize the mean squared error/loss

$$MSE = J(w) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

Solving the first order derivative

$$\frac{\partial}{\partial w} J(w) = 0$$

we get

$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } \hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are sample means

Multiple Linear Regression

Multiple Linear Regression

We assume the model:

$$Y = w_0 + w_1 X_1 + \cdots + w_p X_p + \varepsilon$$

We interpret w_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\text{Sales} = w_0 + w_1 TV + w_2\ Radio + w_p Newspaper$$

Estimate the parameters

Minimize the mean squared error/loss

$$MSE = J(w) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_0 - w_1 x_{i1} \ \ldots w_p x_{ip}\right)^2$$

Estimate the parameters

Matrix representation

suppose we have training data $\{(x_i, y_i)\}_{i=1}^{n}$ where $x_i = (1, x_{i1}, \ldots, x_{ip})$

$$y = Xw$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} — & x_1 & — \\ & \vdots & \\ — & x_2 & — \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

where $w = (w_0, w_1, ..., w_p)'$. And the loss function is

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_0 - w_1 x_{i1} \ \ldots - w_p \, x_{ip}\right)^2$$

$$= \frac{1}{n}\,(y - Xw)^T (y - Xw)$$

Estimate the parameters

- Differentiate MSE w.r.t w

$$-X^T 2(y - Xw) = 0$$

- If $X^T X$ is nonsingular, then the unique solution is given by

$$\hat{w} = (X^T X)^{-1} X^T y$$

- When $X^T X$ is not invertible, the least square solution is **NOT Unique**.

Advertising example

$$y = 2.939 + 0.046 * \text{TV} + 0.189 * \text{Radio} - 0.001 * \text{Newspaper}$$

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Linear regression model is easy to interpret

Multicollinearity

- Multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others.
  - The coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

- The parameter estimation for each individual predictor may not be valid, since the variable may be redundant to others

- With the existence of perfect multicollinearity, the data matrix X has less than full rank, and therefore the moment matrix $X^T X$ cannot be inverted. The ordinary least squares estimator $(X^T X)^{-1} X^T y$ does not exist.

Estimate the parameters – Gradient Descent

Goal: to find w to minimize the cost function $J(w)$

Iterative approach:

- Begin with some initial value w_0, for example $w_0 = (0, 0, .., 0)^T$
- Repeat until converge:

Evaluate the partial derivative of $J(w)$ at current value of w and update w using

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \frac{\partial J(w)}{\partial w}\Big|_{w=w^{(t)}}$$

Estimate the parameters – Gradient Descent

Loss function: $J(w) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_{i1} \ \ldots + w_p \ x_{ip} - y_i\right)^2$

The gradient: $\frac{\partial J(w)}{\partial w_j} = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_{i1} \ \ldots + w_p \ x_{ip} - y_i\right)x_{ij}$

$$w_j^{(t+1)} \leftarrow w_j^{(t)} - \alpha \left.\frac{\partial J(w)}{\partial w}\right|_{w_j = w_j^{(t)}}$$

Or we can write it in using vectors

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \left.\frac{\partial J(w)}{\partial w}\right|_{w = w^{(t)}}$$

$$\frac{\partial J(w)}{\partial w} = \frac{2}{n}\sum_{i=1}^{n}(w^T x_i - y_i)x_i$$

Gradient Descent and Stochastic Gradient Descent

Gradient Descent:

Repeat until convergence

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i) x_i$$

Stochastic Gradient Descent:

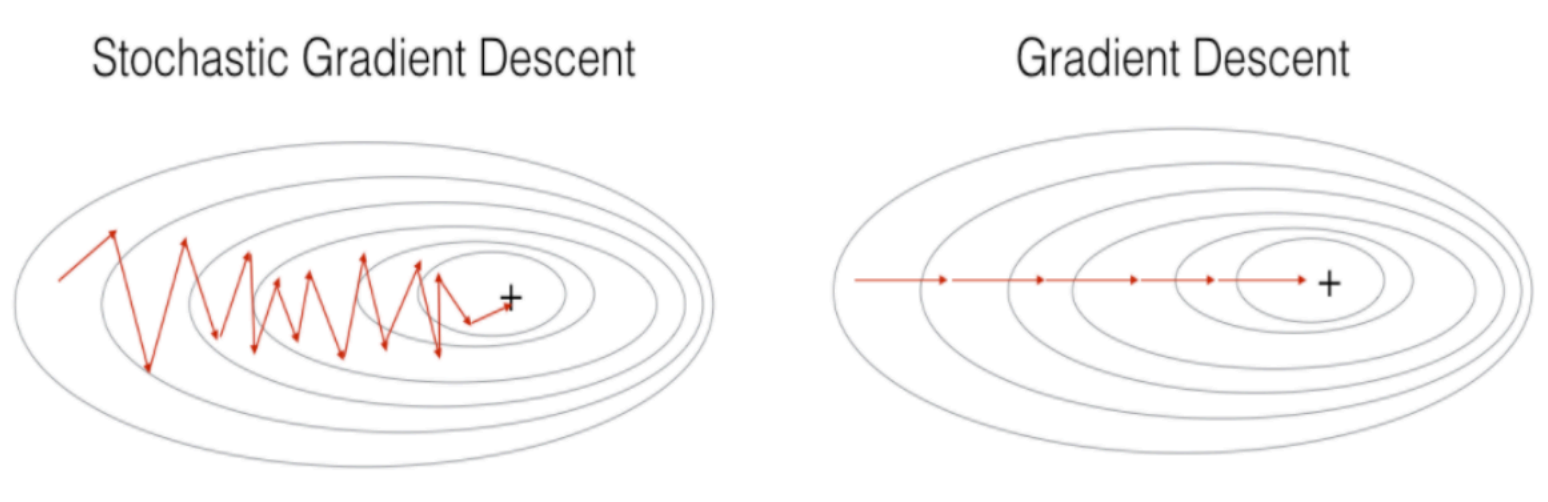
Loop{ for $i = 1$, to n {

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha (w^T x_i - y_i) x_i$$

}

}

Gradient Descent



Often stochastic gradient descent get to the minimum faster that batch gradient descent

Probabilistic interpretation

Assumptions: $Y_i = w^T X_i + \varepsilon_i$

ε_i iid $\sim N(0, \sigma^2)$ (identically and independently distributed)

So we have $p(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(\varepsilon_i)^2}{2\sigma^2}\right)$

$$p(y_i|x_i; w) = P(Y_i = y_i \,|X_i = x_i; w)$$
$$= P(\varepsilon_i = y_i - w^T x_i | X_i = x_i; w)$$
$$= \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

Probabilistic interpretation

Given ε_i's are independent, we can derive the likelihood of the model

$$L(w) = \prod_{i=1}^{N} P(y_i|x_i; w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

The log likelihood is

$$l(w) = logL(w)$$

$$= \sum_{i=1}^{N} \left[-\log\left(\sqrt{2\pi}\sigma\right) - \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right]$$

$$= -Nlog\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - w^T x_i)^2$$

Minimizing MSE is equivalent to maximizing likelihood

Model evaluation

Supervised Learning Summary

Model:	$Y = w_0 + w_1 X_1 + \cdots + w_p X_p + \varepsilon$
Objective/ Loss function:	SSE/MSE
Method:	Exact solution, Gradient Descent
Evaluation metric:	Empirical risk = Mean squared error

Generalization

Assumption: out data is generated independently and identically distributed (iid) from some unknown distribution P

$$(x_i, y_i) \sim P(X, Y)$$

The goal is minimize the expected error (true risk) under P

$$R(w) = \int P(x, y)(y - w^T x)^2 dxdy = E_{X,Y}[(y - w^T x)^2]$$

Estimate the true risk by the empirical risk on a sample data set D

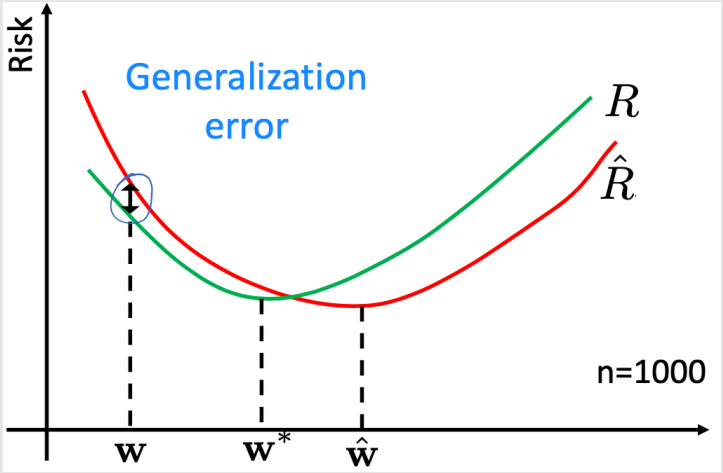
$$\hat{R}_D(w) = \frac{1}{|D|} \sum_{(x,y) \in D} (y - w^T x)^2$$

What happens if we optimize on training data?

- Suppose we are given training data D

  Parameter estimation: $\hat{w}_D = \arg\min_w \hat{R}_D(w)$

- Ideally, we want to solve: $w^* = \arg\min_w R(w)$



What if we evaluate performance on training data

With $\hat{w}_D = \arg\min_w \hat{R}_D(w), \quad w^* = \arg\min_w R(w)$.

In general it holds that $E_D\left[\hat{R}_D(\hat{w}_D)\right] \leq E_D[R(\hat{w}_D)]$

$$\frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - \hat{w}_D^T x_i)^2 \leq \frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - w^T x_i)^2 \quad \text{for } \forall\, w \text{ for } \forall\, D \text{ sampled from } P$$

$$\therefore \quad E_D\left[\frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - \hat{w}_D^T x_i)^2\right] \leq E_D \frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - w^T x_i)^2 \quad \text{for } \forall w.$$

$$E_D\left[\hat{R}_D(\hat{w}_D)\right] \leq \frac{1}{|D|}\sum_{i=1}^{|D|} E_D(y_i - w^T x_i)^2 \quad \text{for } \forall\, w$$

$$E_D\left[\hat{R}_D(\hat{w}_D)\right] \leq \frac{1}{|D|}\sum_{i=1}^{|D|} R(w) \quad \text{for } \forall\, w$$

$$E_D\left[\hat{R}_D(\hat{w}_D)\right] \leq \min_w R(w)$$

$$E_D\left[\hat{R}_D(\hat{w}_D)\right] \leq E_D\left[R(\hat{w}_D)\right]$$

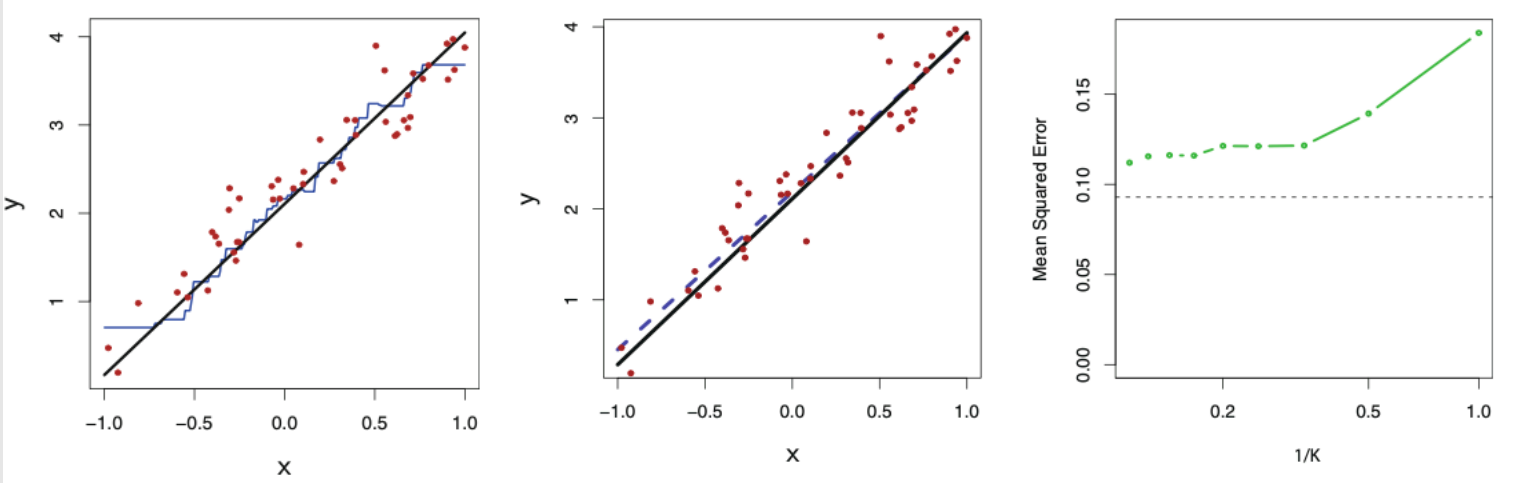
Linear regression vs. KNN

Comparison of LR and KNN for Regression

KNN: non-parametric method

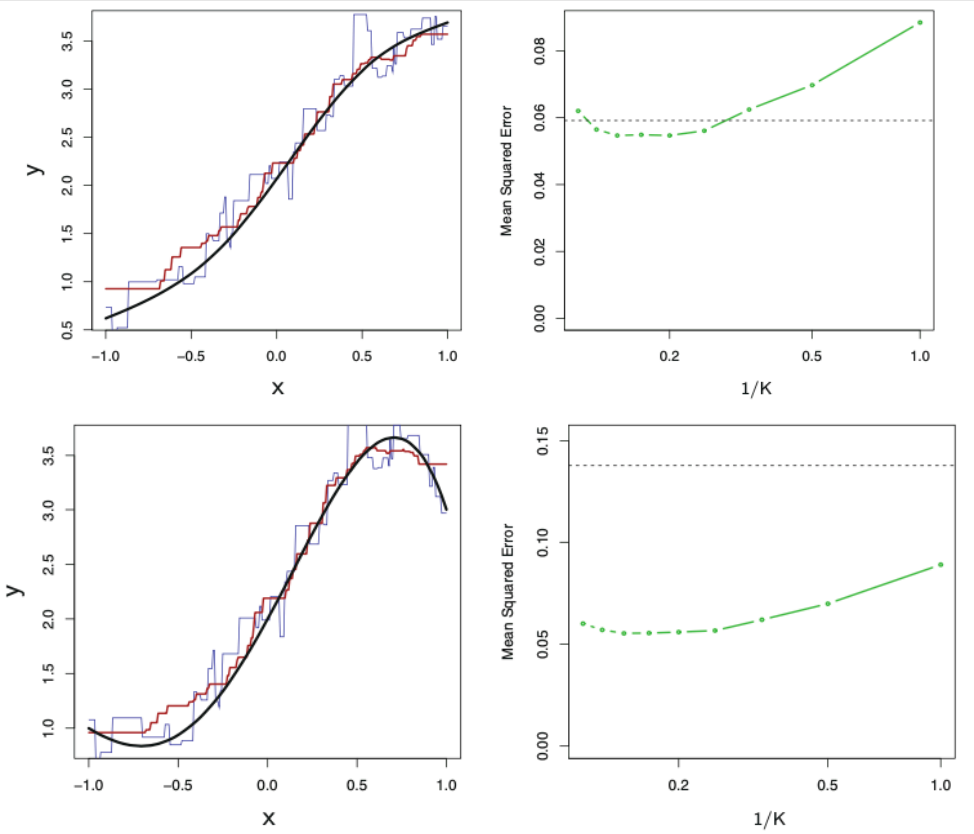
Linear Regression: assumption about $f(x)$: $f(x) = w_0 + w_1 x_1 + \cdots + w_p x_p + \varepsilon$

The parametric approach will outperform the non parametric approach if the assumption is satisfied



Comparison of LR and KNN for Regression

In practice the true relationship between X and y is rarely exactly linear



Comparison of LR and KNN for Regression

Should KNN be favored over linear regression?

Even when the true relationship is highly non-linear, KNN may still provide inferior results to linear regression

Curse of dimensionality!

Other Considerations in Regression Model

Qualitative/Categorical predictors

- Predictors with two levels, e.g. gender

$$x = \begin{cases} 1 & if\ the\ person\ is\ female \\ 0 & if\ the\ person\ is\ male \end{cases}$$

Other Considerations in Regression Model

Qualitative/Categorical predictors

- Categorical variable with more than two levels, e.g. ethnicity (consider three levels: Asian, Caucasian, African American)

$$x_1 = \begin{cases} 1 & \text{if the person is Asian} \\ 0 & \text{if the person is not Asian} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the person is Caucasian} \\ 0 & \text{if the person is not Caucasian} \end{cases}$$

$$y^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + \varepsilon^{(i)} = \begin{cases} w_0 + w_1 + \varepsilon^{(i)} & \text{if the person is Asian} \\ w_0 + w_1 + \varepsilon^{(i)} & \text{if the person is Caucasian} \\ w_0 + \varepsilon^{(i)} & \text{if the person is African American} \end{cases}$$

Potential Problem – Outliers

How does outlier influence the regression model?

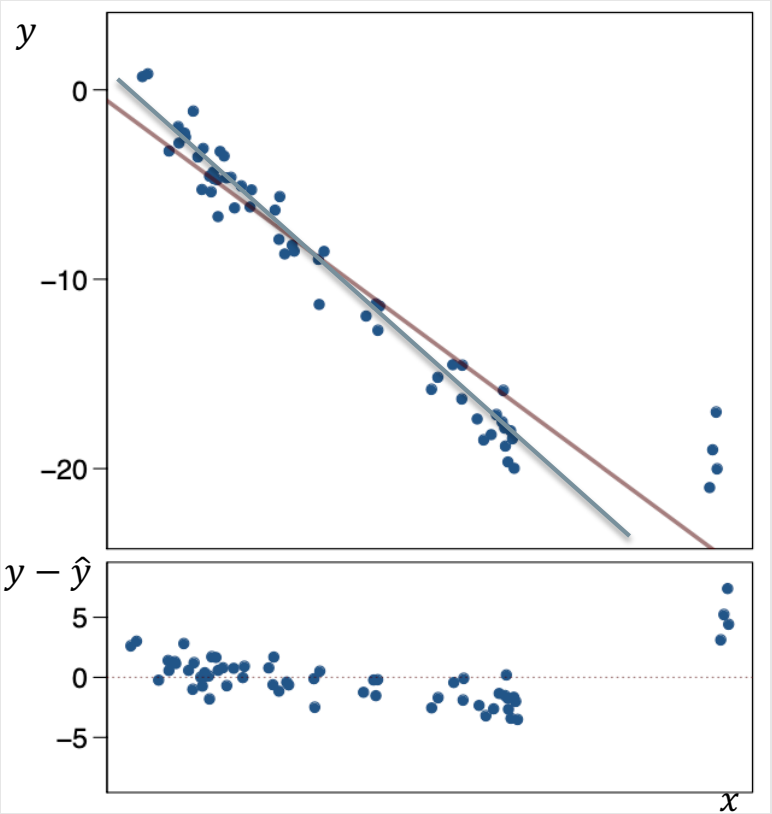
How do we identify outliers?

Residual plot, Data distribution

How do we deal with outliers?

Delete the outlier

More robust models



Potential Problem – High Leverage Points

How does a high leverage point influence the regression model?

How do we identify a high leverage point?

Has an unusual value for predictor x

How do we deal with high leverage points?

Delete the high leverage point

More robust models