Linear Model

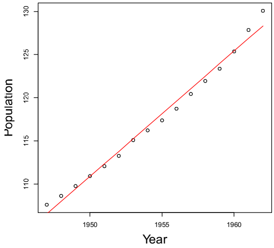
Tong Zhang

Rutgers University

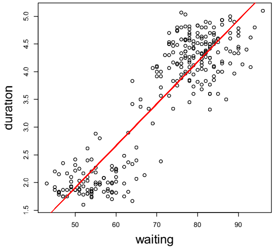
Topics

- Basic statistical model for linear regression
- Algebraic solution
- This talk: focus on three practical issues with example
- Using linear method to model nonlinearity
  - residue plot and basis expansion
- Noise variance estimation and weighted least squares
- Variable Importance
  - two different methods: cost reduction based and test based

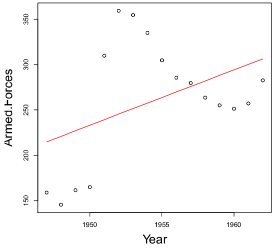
Simple Examples



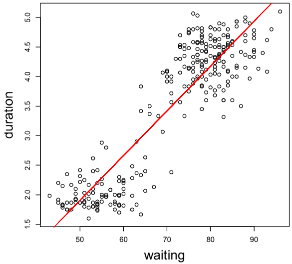
almost linear



linear + noise (?)



nonlinear



Predict Y based on X

Y: duration – response

X: waiting – covariate (feature)

: random noise

linear+noise $Y = \beta_0 + \beta_1 X + \epsilon$

Unknown model parameters: β_0 (intercept or bias) and β_1 (slope)

Questions:

How to estimate model parameters β_0 and β_1 from data?

ow good is this linear model (can we find better model)?

How important are variables β_0 and β_1?

Model Parameter Estimation

Given training data (X_i, Y_i) $(i = 1, \ldots, n)$, want to learn β_0 and β_1
eneral rule: find parameters to fit data as well as possible
efine the residues of linear regression as

$$r_i = Y_i - (\beta_0 + \beta_1 X_i).$$

want to achieve small residues
Method: minimize loss function

$$\min \sum_{i=1}^{n} L(r_i)$$

most popular loss function is squared loss $L(r_i) = r_i^2$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

this is called *linear least squares method*
other loss functions: $L(r_i) = |r_i|$, or $L(r_i) = \max(|r_i| - \epsilon, 0)^2$ etc...

Least Squares Regression

Least squares regression

$$[\hat{\beta}_0, \hat{\beta}_1] = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Statistical model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (i = 1, \ldots, n)$$

where ϵ_i is zero-mean noise for $i = 1, \ldots, n$.
Ideally noise should be iid zero-mean Gaussian

$$\epsilon_i \sim N(0, \sigma^2)$$

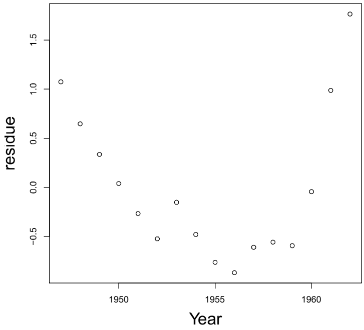
for some unknown σ^2
Remark: least squares regression is sensitive to outliers
  not robust if ϵ_i is heavier tailed than Gaussian

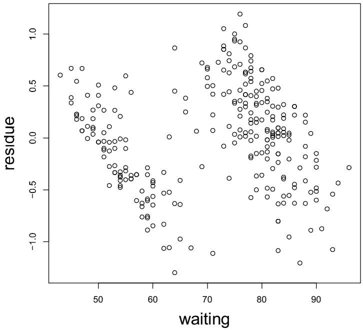
Diagnostic via Residue Plot

Residue $r_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$ should approximate ϵ_i and look random.

if not, we may add additional features to improve model.

residue does not look random – add nonlinear features





quadratic term X_i^2 piecewise linear term $\max(0, X_i - 70)$

Linear Model with Nonlinear Basis

Consider nonlinear basis functions $[f_1(X_i), \ldots, f_p(X_i)]$, we can write a general linear model as

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(X_i) + \epsilon_i \qquad (i = 1, \ldots, n)$$

Example I:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Example II:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \max(0, X_i - 68) + \beta_3 \max(0, X_i - 72) + \epsilon_i$$

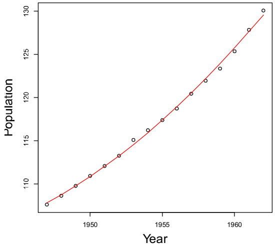
Can still use least squares method to estimate

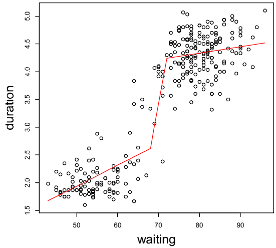
- still linear model: estimation method is linear (least squares)

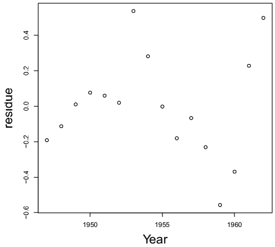
Linear method can model nonlinear functions

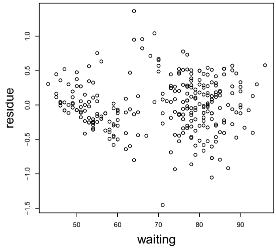
- using nonlinear basis functions

Model Nonlinearity









$\beta_0 + \beta_1 X_i + \beta_2 X_i^2$ $\quad$ $\beta_0 + \beta_1 X_i + \beta_2 \max(0, X_i - 68) + \beta_3 \max(0, X_i - 72)$

General Linear Least Squares

Consider linear regression in high dimension $X_i \in R^p$

Statistical model with regression function $f(x)$ ($x \in R^p$)

$$Y_i = \beta^T X_i + \epsilon_i : \qquad \epsilon_i \sim N(0, \sigma^2),$$

where $\beta = [\beta_1, \ldots, \beta_p] \in R^p$

can include nonlinear features and intercept (constant feature)

Minimize the empirical squared loss (residue sum-of-squares RSS):

$$\hat{\beta} = \arg\min_{\beta \in R^p} RSS(\beta) = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} (\beta^T X_i - Y_i)^2.$$

Algebraic solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

X: $n \times p$ matrix, with each row a data point of vector X_i

Y: $n \times 1$ dimensional vector $[Y_1, \ldots, Y_n] \in R^n$

Noise Distribution

Statistical model: ($X_i \in R^p$)

$$Y_i = \beta^\top X_i + \epsilon_i : \qquad \epsilon_i \sim N(0, \sigma^2)$$

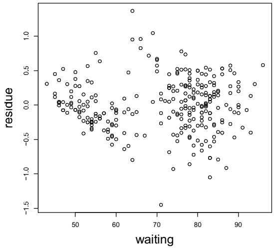
Residue: $r_i \approx \epsilon_i$

$$r_i = Y_i - \hat{\beta}^\top X_i$$

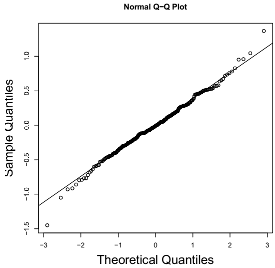
Questions:

- if noise has equal variance, how to estimate σ^2?
- does noise look like Gaussian?
- does noise have equal variance?
- if noise has unequal variance, what to do?

Noise Variance Estimation



residue plot



residue QQ-plot

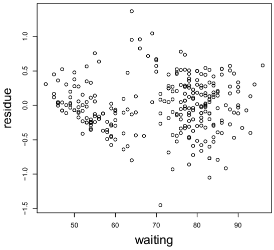
Assume noise has equal variance; then estimate noise variance with

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

divide by $n - p$ instead of n: unbiased estimate

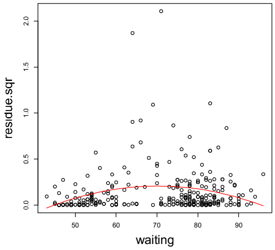
compensate "residue $<$ noise" effect due to fitting p coefficients.

More Complex Noise Variance Model



residue plot

$$Y_i \approx \beta_0 + \beta_1 X_i + \beta_2 \max(0, X_i - 68) + \beta_3 \max(0, X_i - 72))$$



squared residue plot

$$r_i^2 \approx \underbrace{v_0 + v_1(X_i - 45) + v_2(X_i - 45)^2}_{\text{model variance } \sigma_i^2}$$

What to do with unequal variance?

Weighted Least Squares Regression

Statistical model:

$$Y_i = \beta^T X_i + \epsilon_i,$$

where ϵ_i are independent noise.

Different variance for different i:

$$Var(\epsilon_i) = \sigma_i^2,$$

where assume σ_i^2 are all known.

Best unbiased linear estimator is weighted least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} w_i (Y_i - \beta^T X_i)^2,$$

where $w_i = 1/\sigma_i^2$ is inversely proportional to noise variance.

Variable Importance

Goal: to decide which variable is important

related to variable selection

Method: assign importance score to each variable and compare.

Different interpretations of variable importance:

include a particular variable can improve prediction (or stability)

he corresponding coefficient is nonzero (Hypothesis testing)

The basic ideas can also be applied to other learning algorithms including nonlinear methods

Variable Importance: method I

Model training error measured by residue sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta})^2,$$

where $\hat{\beta}$ is the solution of least squares (or other algorithm).

Test the effect of removing the j-th variable (setting $\beta_j = 0$):

- RSS_0: RSS of the original model
- $RSS_{(j)}$: RSS with j-th variable removed.
- ignificance can be measured by:

$$F_j = \frac{\text{Cost Function Increase}}{\text{Noise Variance}} = \frac{RSS_{(j)} - RSS_0}{RSS_0/(n-p)}$$

This idea can be generalized to groups of variables

- change of residue by removing a set of variables.

This idea can be applied to other learning methods

Variable Importance: method II

Equivalent alternative: $F_j = z_j^2$

Let ω_j be the j-th diagonal of $(X^T X)^{-1}$.

Estimate noise variance as:

$$\hat{\sigma}^2 = RSS_1/(n-p).$$

Define z-score for the j-th variable:

$$z_j = \frac{\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\hat{\sigma}} \cdot \frac{1}{\sqrt{\omega_j}},$$

If noise is iid Gaussian, then

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \sim t_{n-p},$$

t_{n-p}: student t distribution with degree of freedom $n-p$.

Variable Importance and Hypothesis Testing

We know that if noise is iid normal, then

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \sim t_{n-p}.$$

Hypothesis testing view of variable importance:

Null hypothesis $\beta_j = 0$. Under null hypothesis:

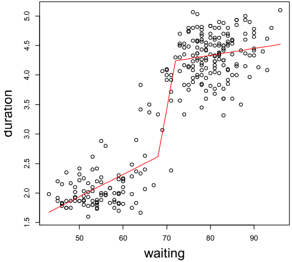
$$z_j \sim t_{n-p}.$$

Variable importance using p-value of z-score under t_{n-p} distribution

loose interpretation: the chance of β_j to be zero coefficient

Equivalent for least squares: p-value gives natural interpretation

Example



Model:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i \\
&\quad + \beta_2 \max(0, X_i - 68) \\
&\quad + \beta_3 \max(0, X_i - 72)) + \epsilon_i
\end{aligned}
$$

Variable importance:

coefficient	LS-solution	z-score	p-value
β_0	0.057978	0.171	0.865
β_1	0.037619	6.144	2.9×10^{-9}
β_2	0.368549	11.395	$< 2 \times 10^{-16}$
β_3	-0.394534	-12.737	$< 2 \times 10^{-16}$

Summary

- Basic statistical model for linear regression
  - linear regression function with Gaussian noise
  - solution via least squares method
- Model nonlinearity
  - using nonlinear basis functions
  - residue plot and simple diagnostics
- Noise variance estimation:
  - formula for constant noise variance
  - non-constant noise variance: need to use weighted least squares
- Variable Importance: two schemes
  - error reduction: removing a variable doesn't increase error a lot
  - stability: p-value and confidence interval estimation