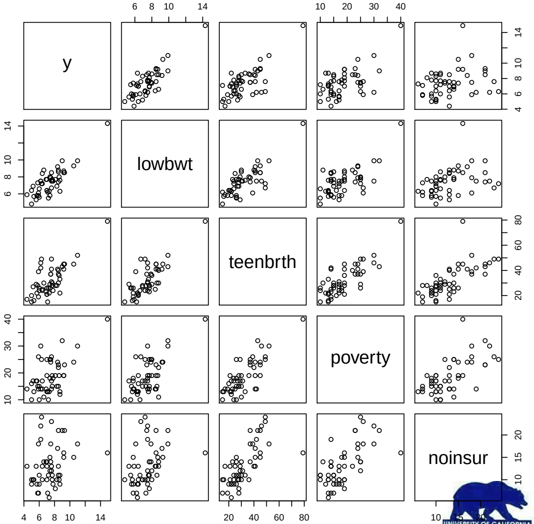
Introduction to Multivariate Statistics
Lecture 4: Linear Regression

Lexin Li

University of California, Berkeley

Outline

- Motivating example: U.S. infant mortality rate from Annie E. Casey Kids Count Data Center
  - Y: infant mortality rate
  - X_1: low birthweight rate
  - X_2: teen birth rate
  - X_3: poverty rate
  - X_4: no insurance rate
  - 50 states + D.C.
  - many other variables



Outline

- What is it about:
  - **Association/relation** between **response/output/dependent variable** (Y) and **predictor/input/feature variable** $\boldsymbol{X}$; how the value of Y changes as a function of $\boldsymbol{X}$
- Topics to cover:
  - Data visualization
  - Model, interpretation, estimation, prediction
  - Characterization of uncertainty
  - Categorical explanatory variables
  - Goodness-of-fit, model diagnosis, and remedies
  - Extensions: multivariate responses, nonlinear models, variable selection
- What to pay special attention:
  - Interpretation, interpretation, interpretation!
  - Is this a good model?

The super example

- **Body fat example**:
  - Body fat, a measure of health, is estimated through an underwater weighing technique. Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.
  - Percentage of body fat, age, weight, height, and ten body circumference measurements are recorded for 252 men.
    - Percent body fat using Brozek's equation, 457/Density - 414.2
    - Percent body fat using Siri's equation, 495/Density - 450
    - Density (gm/cm^3); Age (yrs); Weight (lbs); Height (inches); Adiposity index = Weight/Height^2 (kg/m^2); Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
    - Circumference (cm): Neck; Chest; Abdomen; Hip; Thigh; Knee; Ankle; Extended biceps; Forearm; Wrist.
  - Dichotomized body fat groups: Obese ($\geq 25\%$), Normal ($< 25\%$).

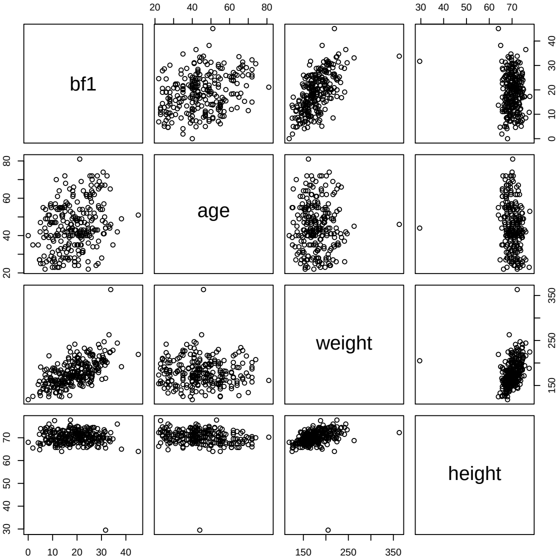
The super example

- Data description:
  - Samples: 252 men
  - Response: percentage of body fat, calculated by two different formulae
  - Dichotomized response: obese ($\geq 25\%$, 59 subjects, about 23%) *vs* normal ($< 25\%$, 193 subjects)
  - Predictors: age, weight, height
  - Predictors: 10 body circumference measurements: neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, wrist
  - Question of interest: **association** between percentage of body fat with age, weight, height, and ten body circumference measurements – body fat is an important health measure, and its measurement by an underwater weighing technique *vs* by a scale and a measuring tape
- How the data look like:

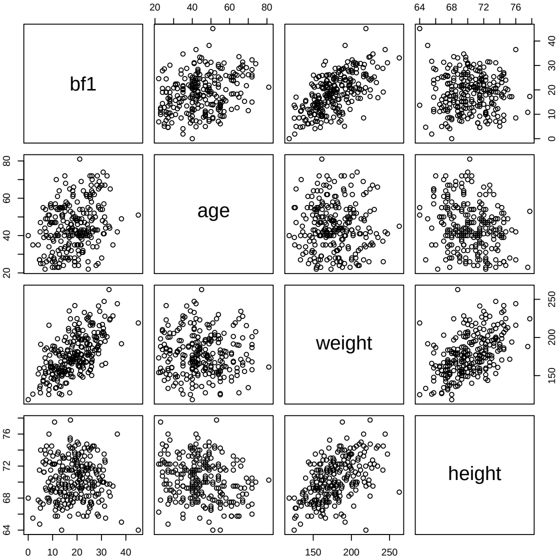
	bf1	bf2	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
[1,]	12.6	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
[2,]	6.9	6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
[3,]	24.6	25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
[4,]	10.9	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
[5,]	27.8	28.7	24	184.25	71.25	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7

...

Data visualization



Data visualization



- The two observations were **removed** from subsequent analysis, but **be cautious**!

Model

- Multiple linear regression model: population level

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

  - Response/output/dependent variable: Y; e.g., percentage of body fat
  - Predictor/input/explanatory variable/feature variable: $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\mathsf{T}$; e.g., age, weight, height, 10 body circumference measurements
  - Error ε is assumed $N(0, \sigma^2)$ and is **independent** of $\boldsymbol{X}$

  - What is meaning of Y, $\boldsymbol{X}$, and ε?
  - What is a **sample / replication**?

  - $Y|\boldsymbol{X}$ is **normally distributed**
  - $E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ – **linear mean**
  - $\mathrm{var}(Y|\boldsymbol{X}) = \sigma^2$ – **constant variance**

Model

- Given the observed data: $\{(\boldsymbol{x}_i, y_i), i = 1, \dots, n\}$, sample level

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \; i = 1, \dots, n$$

That is,

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n
\end{aligned}$$

Model

- Matrix form:

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{e}_{n\times 1}$$

$\boldsymbol{y}$: the response vector
$\boldsymbol{X}$: the design matrix
$\boldsymbol{\beta}$: the regression coefficient vector

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n\times(p+1)} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}$$

Interpretation

- **Interpretation** of β_j:
  - Represents the **partial** effect of X_j on Y, **after** the effect of all other variables have been removed
  - Regress the residual $[y_i - \sum_{k=1,k\neq j}^{p} \beta_k x_{ik}]$ on x_{ij} gives the same coefficient β_j
  - A little math:

$$
\begin{aligned}
\beta_1 &= E(Y|X_1 = x_1 + 1, X_2 = x_2, \ldots, X_p = x_p) \\
&\qquad - E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) \\
&= \{\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \ldots + \beta_p x_p\} \\
&\qquad - \{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\}
\end{aligned}
$$

- Example:
  - Response: infant birthweight (Y)
  - Predictors: mother's weight (X_1) + mother's age (X_2) + infant gender (X_3; 1=boy, 0=girl)
  - Interpretation of β_1: the average increase of birthweight for one "unit" increase of mother's weight, keeping everything else fixed

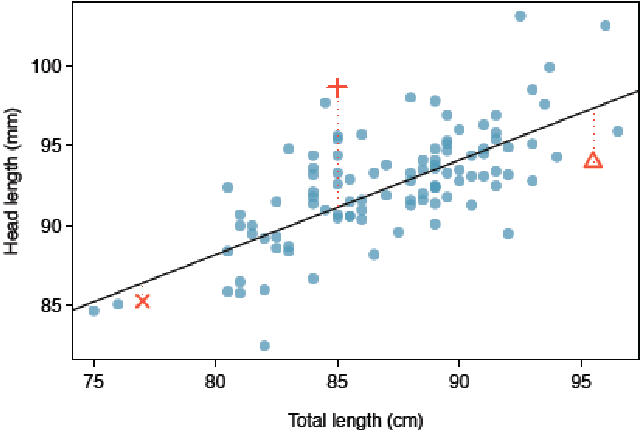
Interpretation

- Categorical predictors:
  - Interpretation of β_3: the average increase of birthweight for a boy compared to a girl, keeping everything else fixed
  - One additional predictor: mother's race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Pacific Islander, White)
  - Samples: AA, AS, WH, NH, WH, AI
  - **Indicator variables (dummy variables)**

	AS	AA	NH	WH
subject 1	0	1	0	0
subject 2	1	0	0	0
subject 3	0	0	0	1
subject 4	0	0	1	0
subject 5	0	0	0	1
subject 6	0	0	0	0

  - Interpretation: **change compared to the reference group**

Estimation



Estimation

- Ordinary least squares:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- Matrix form:

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

solution:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\ \hat{\boldsymbol{y}} &= \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}$ is the Hat matrix

Estimation

- Estimation of σ^2:
$$\hat{\sigma}^2 = \frac{SSE}{df} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}$$

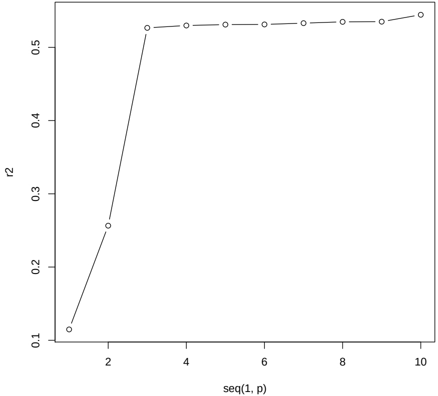
- Prediction:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_p x_{ip}, \quad i = 1, \ldots, n$$

- R^2: multiple correlation coefficient
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
Interpretation of R^2: proportion of variation in the response that has been explained by the linear model

- **Adjusted** R^2: compensating for more variables
$$R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

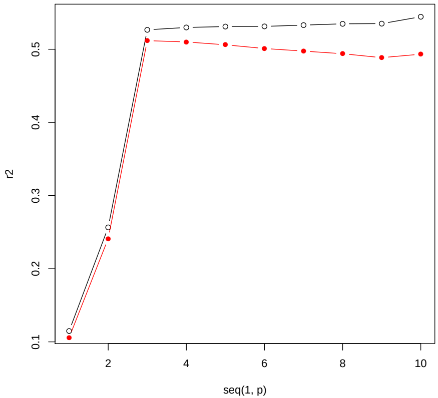
Estimation

- A simulation illustration:
  - $Y = 5 + X_1 - X_2 + 2X_3 + \varepsilon$
  - Fit a regression model of Y on X_1, then X_1, X_2, ..., then $X_1, X_2, \ldots, X_{10}$



Estimation

- A simulation illustration:
  - $Y = 5 + X_1 - X_2 + 2X_3 + \varepsilon$
  - Fit a regression model of Y on X_1, then X_1, X_2, ..., then $X_1, X_2, \ldots, X_{10}$



Inference

- Inference: quantification of uncertainty

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

- *p*-**value**:

$$\begin{aligned} \text{null hypothesis} \qquad & H_0: \ \beta_j = 0 \\ \text{test statistic} \qquad & t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \\ p\text{-value} \qquad & 2 \times [1 - pt(|t|, n-p-1)] \end{aligned}$$

where $se(\hat{\beta}_j) = \{\hat{\sigma}^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj}\}^{1/2}$

Linear regression in R

- Body fat example:
  - Response: percentage of body fat (using Brozek's formula)
  - Predictors: age, weight, height
- R:

```
# data processing
data<-read.table(file="Data_Bodyfat.txt", header=FALSE, quote="")
bf1<-data[,2]; bf2<-data[,3]
age<-data[,5]; weight<-data[,6]; height<-data[,7]
X<-as.matrix(data[,10:19])
colnames(X)<-c("neck", "chest", "abdomen", "hip", "thigh", "knee",
               "ankle", "biceps", "forearm", "wrist")
data<-cbind(bf1, bf2, age, weight, height, X)
colnames(data)<-c(c("bf1", "bf2", "age", "weight", "height"), colnames(X))

# data visualization
plot(bf1, bf2)
pairs(~bf1+age+weight+height)

# remove outliers
ids = c(seq(1,nrow(data))[data[,4]>300], seq(1,nrow(data))[data[,5]<40])
data2 = data[-ids,]
```

Linear regression in R

- R:

```
fit.lm<-lm(bf1~age+weight+height, data=data.frame(data2))
summary(fit.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.31985    9.63347   5.639 4.69e-08 ***
age          0.12575    0.02599   4.838 2.31e-06 ***
weight       0.23519    0.01373  17.124  < 2e-16 ***
height      -1.18089    0.14638  -8.067 3.17e-14 ***

Residual standard error: 4.986 on 246 degrees of freedom
Multiple R-squared:  0.5838,Adjusted R-squared:  0.5787
F-statistic:   115 on 3 and 246 DF,  p-value: < 2.2e-16
```

- Ask ourselves:
  - What does this model really tell us?
  - Interpretation of the coefficients β, R^2, adjusted R^2

Linear regression in R

- R:

```
fit.lm<-lm(bf1~age+weight+height, data=data.frame(data2))
summary(fit.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.31985    9.63347   5.639 4.69e-08 ***
age          0.12575    0.02599   4.838 2.31e-06 ***
weight       0.23519    0.01373  17.124  < 2e-16 ***
height      -1.18089    0.14638  -8.067 3.17e-14 ***

Residual standard error: 4.986 on 246 degrees of freedom
Multiple R-squared:  0.5838,Adjusted R-squared:  0.5787
F-statistic:   115 on 3 and 246 DF,  p-value: < 2.2e-16
```

- Ask ourselves:
  - What does this model really tell us?
  - Interpretation of the coefficients β, R^2, adjusted R^2
  - What are the underlying assumptions? – Are those assumptions satisfied in this data?
  - Is there anything unusual going on?

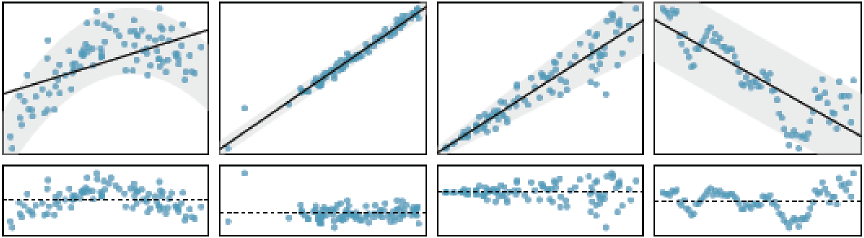
Diagnosis

- Question 1: is the linear regression model a good choice?
  - Mean function – **linear or nonlinear**
  - Variance function – **constant or nonconstant**
  - Response distribution – **normal or not**

Diagnosis

- Question 1: is the linear regression model a good choice?
  - Mean function – **linear or nonlinear**
  - Variance function – **constant or nonconstant**
  - Response distribution – **normal or not**

  - Basic idea: If the model is correct, then the **residuals** $e_i = y_i - \hat{y}_i, i = 1, \ldots, n$, should look like a sample from a normal distribution with mean zero and constant variance

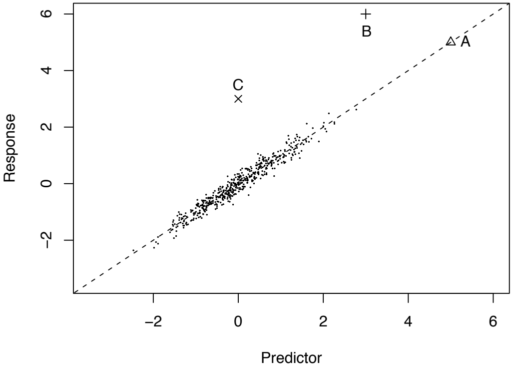


Diagnosis

- Types of residuals:
  - Ordinary residuals:
  $$e_i = y_i - \hat{y}_i$$
  measure the deviation of predicted value from observed value
  - Studentized residuals:
  $$r_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \sim t_{n-p-2}$$
  h_{ii} is the i-th diagonal element of the hat matrix $\boldsymbol{H}$;
  $\hat{\sigma}_{(i)} = \sum_{k=1,k\neq i}^{n}(y_k - \hat{y}_k)^2/(n-p-1-1)$
- **Solution: transformation**
  - Transformation of $\boldsymbol{X}$: $\log(X_j), \sqrt{X_j}, \ldots$
  - Transformation of Y: $\log(Y), \sqrt{Y}, \ldots$
  - Goal: to help achieve linearity and/or stabilize variance

Diagnosis

- Question 2: is there anything unusual?
  - **Influential observation:** data points that influence the regression line the most
  - **Outlier:** data points that stand out of the rest – possibly mistakes in data transcription, lab errors, who knows? – those points should be recognized and (hopefully) explained

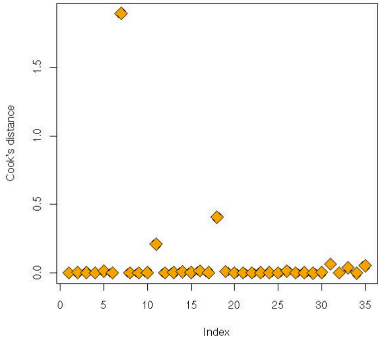


Diagnosis

- Influence measure: Cook's distance

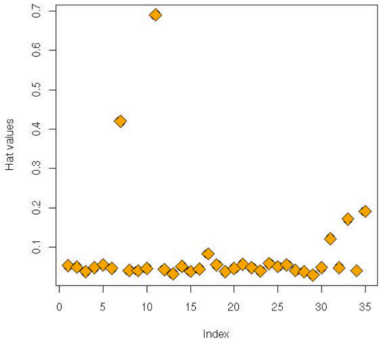
$$D_i = \frac{\sum_{j=1, j \neq i}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{(p+1)\hat{\sigma}_2}$$

where $\hat{y}_{j(-i)}$ is the j-th fitted value without the i-th observation

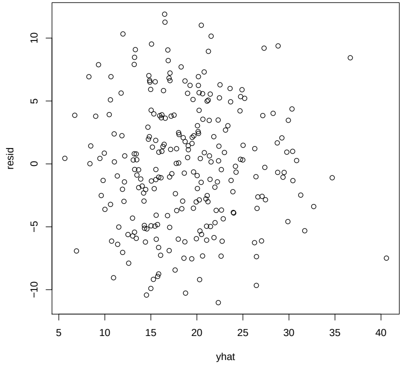


Diagnosis

- Outlier:
  - Outlier in $\boldsymbol{X}$: the $\boldsymbol{X}$ values of the observation may lie outside the "cloud" of other $\boldsymbol{X}$ values, suggesting you may be **extrapolating** the model inappropriately – h_{ii} of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}$
  - Outlier in Y: the Y value of the observation may lie very far from the fitted model – r_i

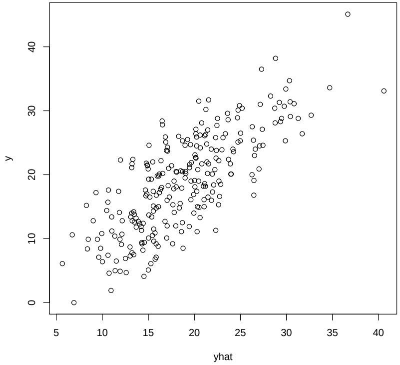


The super example



residual plot

`plot(fitted(fit.lm), resid(fit.lm))`



fitted values plot

`plot(fitted(fit.lm),y)`

Model / variable selection

- Goodness of fit test:

$$
\begin{aligned}
M_R : E(Y|\boldsymbol{X}) &= \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q \\
M_F : E(Y|\boldsymbol{X}) &= \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_p X_p
\end{aligned}
$$

Test statistic:

$$
F = \frac{[SSE(M_R) - SSE(M_F)]/(df_R - df_F)}{SSE(M_F)/df_F} \sim F_{df_R - df_F, df_F}
$$

Intuition?

- R:

```
fit.lm<-lm(bf1~age+weight+height, data=data.frame(data2))
summary(fit.lm)

Residual standard error: 4.986 on 246 degrees of freedom
Multiple R-squared:  0.5838,Adjusted R-squared:  0.5787
F-statistic:   115 on 3 and 246 DF,  p-value: < 2.2e-16
```

Model / variable selection

- Least absolute shrinkage and selection operator (**Lasso**):

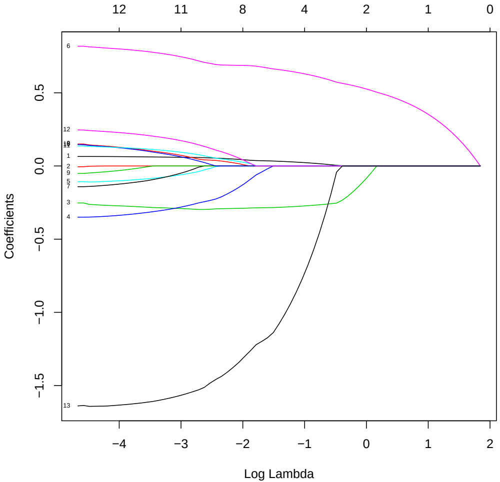
$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \underbrace{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2}_{L(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\beta})} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{P(\boldsymbol{\beta};\lambda)}$$

  - **Regularization: loss function + penalty function**
    $\leftarrow$ two **competing** terms!
  - Tuning parameter: λ

- **Caution**: use it only when necessary!

- R: `glmnet` package

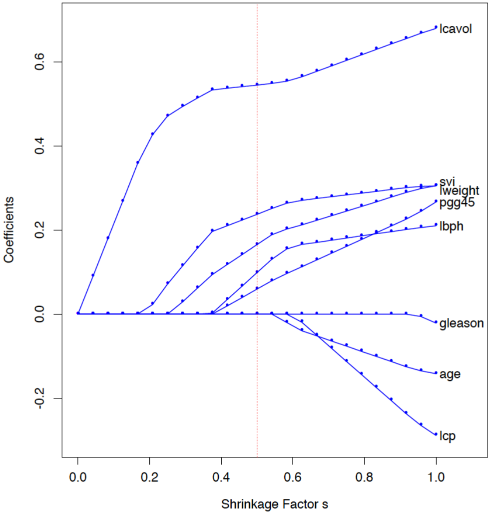
Model / variable selection



- R:

```
fit=glmnet(Xall,y)
plot(fit, xvar="lambda", label=TRUE)
```

Model / variable selection



Nonlinear models

- Basis expansion:
  - Key idea: augment / replace the input variables $\boldsymbol{X}$ with **transformations** of $\boldsymbol{X}$, and then fit a **linear model** in the new space of derived input features
  - A more flexible model:

$$Y = \beta_0 + \beta_1 h_1(X_1, \ldots, X_p) + \ldots + \beta_m h_m(X_1, \ldots, X_p) + \varepsilon$$

where $h_m(\cdot)$ are pre-specified **basis functions**

Nonlinear models

- Basis expansion:
  - Key idea: augment / replace the input variables $\boldsymbol{X}$ with **transformations** of $\boldsymbol{X}$, and then fit a **linear model** in the new space of derived input features
  - A more flexible model:
  $$Y = \beta_0 + \beta_1 h_1(X_1, \ldots, X_p) + \ldots + \beta_m h_m(X_1, \ldots, X_p) + \varepsilon$$
  where $h_m(\cdot)$ are pre-specified **basis functions**
  - Special case I: **generalized additive model**
  $$Y = \beta_0 + \beta_1 h_1(X_1) + \ldots \beta_p h_p(X_p) + \varepsilon$$

Nonlinear models

- Basis expansion:
  - Key idea: augment / replace the input variables $\boldsymbol{X}$ with **transformations** of $\boldsymbol{X}$, and then fit a **linear model** in the new space of derived input features
  - A more flexible model:

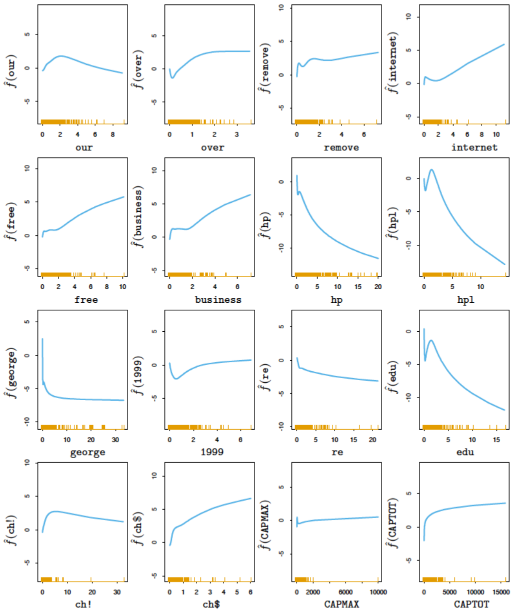
$$Y = \beta_0 + \beta_1 h_1(X_1, \ldots, X_p) + \ldots + \beta_m h_m(X_1, \ldots, X_p) + \varepsilon$$

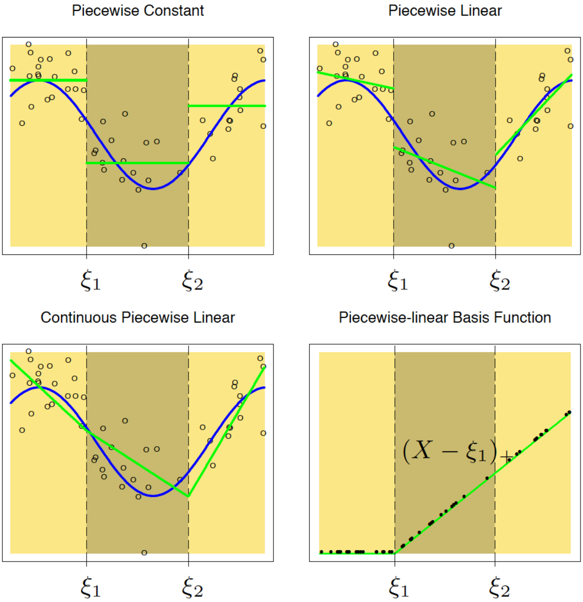
    where $h_m(\cdot)$ are pre-specified **basis functions**
  - Special case I: **generalized additive model**

$$Y = \beta_0 + \beta_1 h_1(X_1) + \ldots \beta_p h_p(X_p) + \varepsilon$$

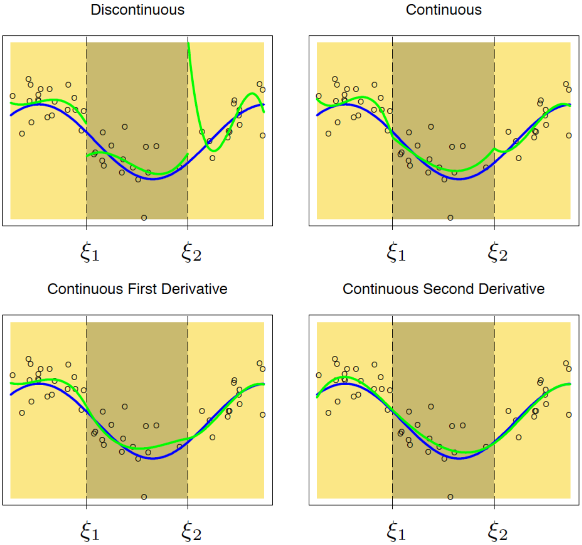
  - Special case II: **spline basis expansion** (piecewise polynomials)

$$\begin{aligned} Y &= \beta_0 + \beta_{11} h_{11}(X_1) + \ldots + \beta_{1m_1} h_{1m_1}(X_1) \\ &\quad + \ldots + \beta_{p1} h_{p1}(X_p) + \ldots + \beta_{pm_p} h_{pm_p}(X_p) + \varepsilon \end{aligned}$$


our
over
remove
internet
free
business
hp
hpl
george
1999
re
edu
ch!
ch$
CAPMAX
CAPTOT


Piecewise Constant
Piecewise Linear
Continuous Piecewise Linear
Piecewise-linear Basis Function
ξ_1
ξ_2
$(X - \xi_1)_+$

Piecewise Cubic Polynomials



Multivariate-response models

- Multivariate responses: association between $(Y_1, Y_2, \ldots, Y_q)$ and $(X_1, X_2, \ldots, X_p)$
- Example: association between infant birthweight and birth height and mother's weight, age, race, infant's gender
- Multivariate-response linear regression model: population level

$$
\begin{aligned}
Y_1 &= \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{p1}X_p + \varepsilon_1 \\
Y_2 &= \beta_{02} + \beta_{12}X_1 + \ldots + \beta_{p2}X_p + \varepsilon_2 \\
&\ldots \\
Y_q &= \beta_{0q} + \beta_{1q}X_1 + \ldots + \beta_{pq}X_p + \varepsilon_q
\end{aligned}
$$

Multivariate-response models

- Multivariate-response linear regression model: sample level

$$
\begin{aligned}
y_{11} &= \beta_{01} + \beta_{11} x_{11} + \ldots + \beta_{p1} x_{1p} + \varepsilon_{11} \\
y_{12} &= \beta_{01} + \beta_{11} x_{21} + \ldots + \beta_{p1} x_{2p} + \varepsilon_{12} \\
&\vdots \\
y_{1n} &= \beta_{01} + \beta_{11} x_{n1} + \ldots + \beta_{p1} x_{np} + \varepsilon_{1n} \\
&\vdots \\
y_{q1} &= \beta_{0q} + \beta_{1q} x_{11} + \ldots + \beta_{pq} x_{1p} + \varepsilon_{q1} \\
y_{q2} &= \beta_{0q} + \beta_{1q} x_{21} + \ldots + \beta_{pq} x_{2p} + \varepsilon_{q2} \\
&\vdots \\
y_{qn} &= \beta_{0q} + \beta_{1q} x_{n1} + \ldots + \beta_{pq} x_{np} + \varepsilon_{qn}
\end{aligned}
$$

Multivariate-response models

- Matrix form:

$$\boldsymbol{Y}_{n\times q} = \boldsymbol{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times q} + \boldsymbol{e}_{n\times q}$$

  - $\boldsymbol{Y}$: the response matrix, $n \times q$
  - $\boldsymbol{X}$: the design matrix, $n \times (p+1)$
  - $\boldsymbol{\beta}$: the regression coefficient matrix, $(p+1) \times q$

- Estimation:
  - Exactly the same principle as the univariate response linear regression
  - Improvements: reduced-rank regression and/or regularization methods

Multi-level / Hierarchical models

- Example:
  - Question of interest: Do student breakfast consumption and teaching style influence student GPA?
  - Response: GPA
  - Predictors:
    - Student level: breakfast consumption
    - Classroom level: teaching style

Multi-level / Hierarchical models

- Example:
  - Question of interest: Do student breakfast consumption and teaching style influence student GPA?
  - Response: GPA
  - Predictors:
    - Student level: breakfast consumption – Level 1 predictor
    - Classroom level: teaching style – Level 2 predictor

Multi-level / Hierarchical models

- Example:
  - Question of interest: Do student breakfast consumption and teaching style influence student GPA?
  - Response: GPA
  - Predictors:
    - Student level: breakfast consumption – Level 1 predictor
    - Classroom level: teaching style – Level 2 predictor
  - **Nested data**
- **Multi-level / hierarchical linear model**
- R package: `multilevel`

Multi-level / Hierarchical models

- **Multi-level / hierarchical linear model**
  - Level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

    - Y_{ij} = GPA measured for student i in classroom j
    - X_{ij} = breakfast consumption for student i in classroom j
    - β_{0j} = intercept for the jth classroom
    - β_{1j} = slope for the jth classroom

Multi-level / Hierarchical models

- **Multi-level / hierarchical linear model**
  - Level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

    - Y_{ij} = GPA measured for student i in classroom j
    - X_{ij} = breakfast consumption for student i in classroom j
    - β_{0j} = intercept for the jth classroom
    - β_{1j} = slope for the jth classroom
  - Level-2 model:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} G_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} G_j + U_{1j} \end{aligned}$$

    - G_j = teaching style in classroom j

Multi-level / Hierarchical models

- **Multi-level / hierarchical linear model**
  - Level-1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

    - Y_{ij} = GPA measured for student i in classroom j
    - X_{ij} = breakfast consumption for student i in classroom j
    - β_{0j} = intercept for the jth classroom
    - β_{1j} = slope for the jth classroom
  - Level-2 model:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} G_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} G_j + U_{1j} \end{aligned}$$

    - G_j = teaching style in classroom j
  - Combined model – **Mixed effects model**

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} G_j + \gamma_{11} G_j X_{ij} + U_{1j} X_{ij} + U_{0j} + r_{ij}$$

    - Fixed effects: $\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11}$
    - Random effects: U_{1j}, U_{0j}

Discussion

- What is this chapter about: **in a bigger picture**
  - Study the **association** between one **quantitative** variable (response/output/dependent variable; Y) and one or many **qualitative / quantitative** variables (predictor/input/feature variable; $\boldsymbol{X}$)

Discussion

- What is this chapter about: **in a bigger picture**
  - Study the **association** between one **quantitative** variable (response/output/dependent variable; Y) and one or many **qualitative / quantitative** variables (predictor/input/feature variable; $\boldsymbol{X}$)
  - Last chapter: study the **association** between one or a few **quantitative** variable(s) with one or a few **qualitative** variable(s)

Discussion

- What is this chapter about: **in a bigger picture**
  - Study the **association** between one **quantitative** variable (response/output/dependent variable; Y) and one or many **qualitative / quantitative** variables (predictor/input/feature variable; $\boldsymbol{X}$)
  - Last chapter: study the **association** between one or a few **quantitative** variable(s) with one or a few **qualitative** variable(s)
- Things to pay attention to:
  - What does this model tell us? – Interpretation
  - Is this a good model? – Model assumptions and model diagnosis
  - **Association $\neq$ Causation**