Regression and Time Series

Jianqing Fan

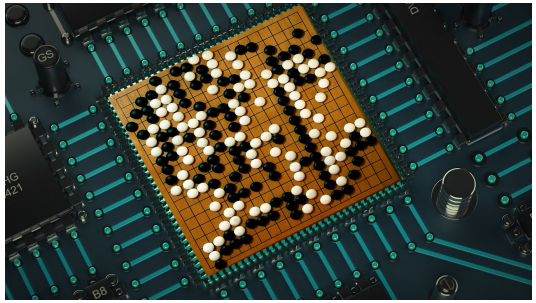
Princeton University

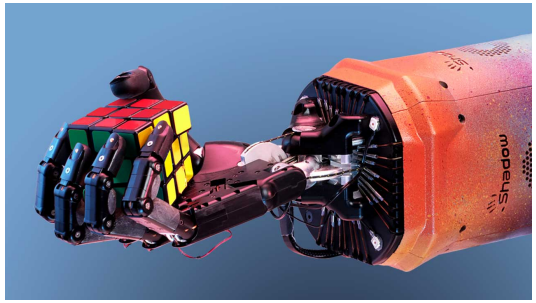
https://fan.princeton.edu



Interface of Learning and Decision

- **Games**: Atari, Go, Starcraft, Texas hold'em, ...
- **Control**: robotic hand, automatic driving, fleet management, ...
- **Prediction**: peotein folding, LLM, generative AI ...
- **Causality**: Molecular mechnism, policy eval, ...

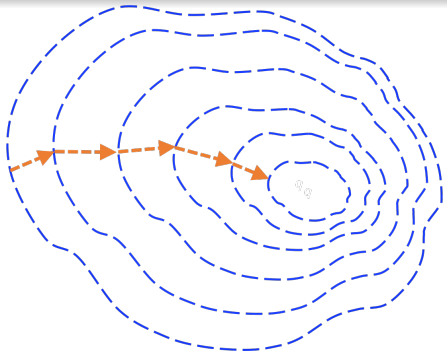




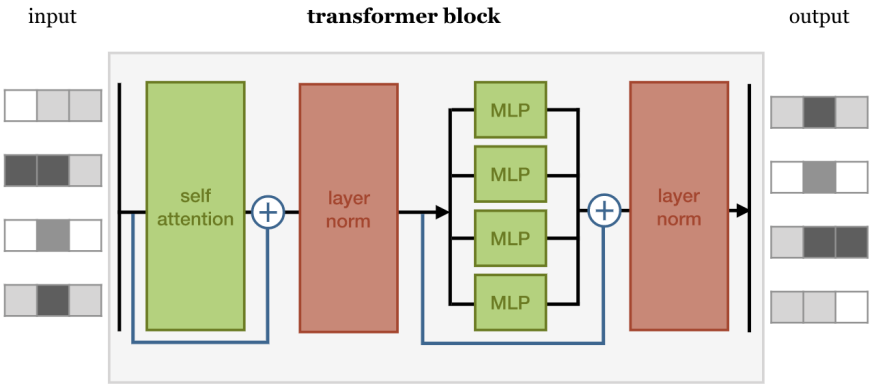
Drivers of successes

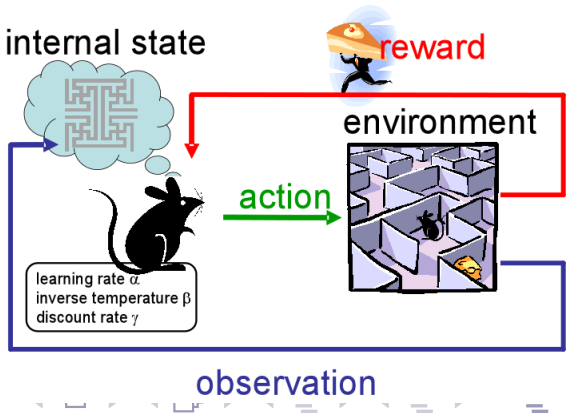
DL + **RL** + **modern computation** + **big-data**

Deep Q-Network 24/39 (Minh et al, 15)

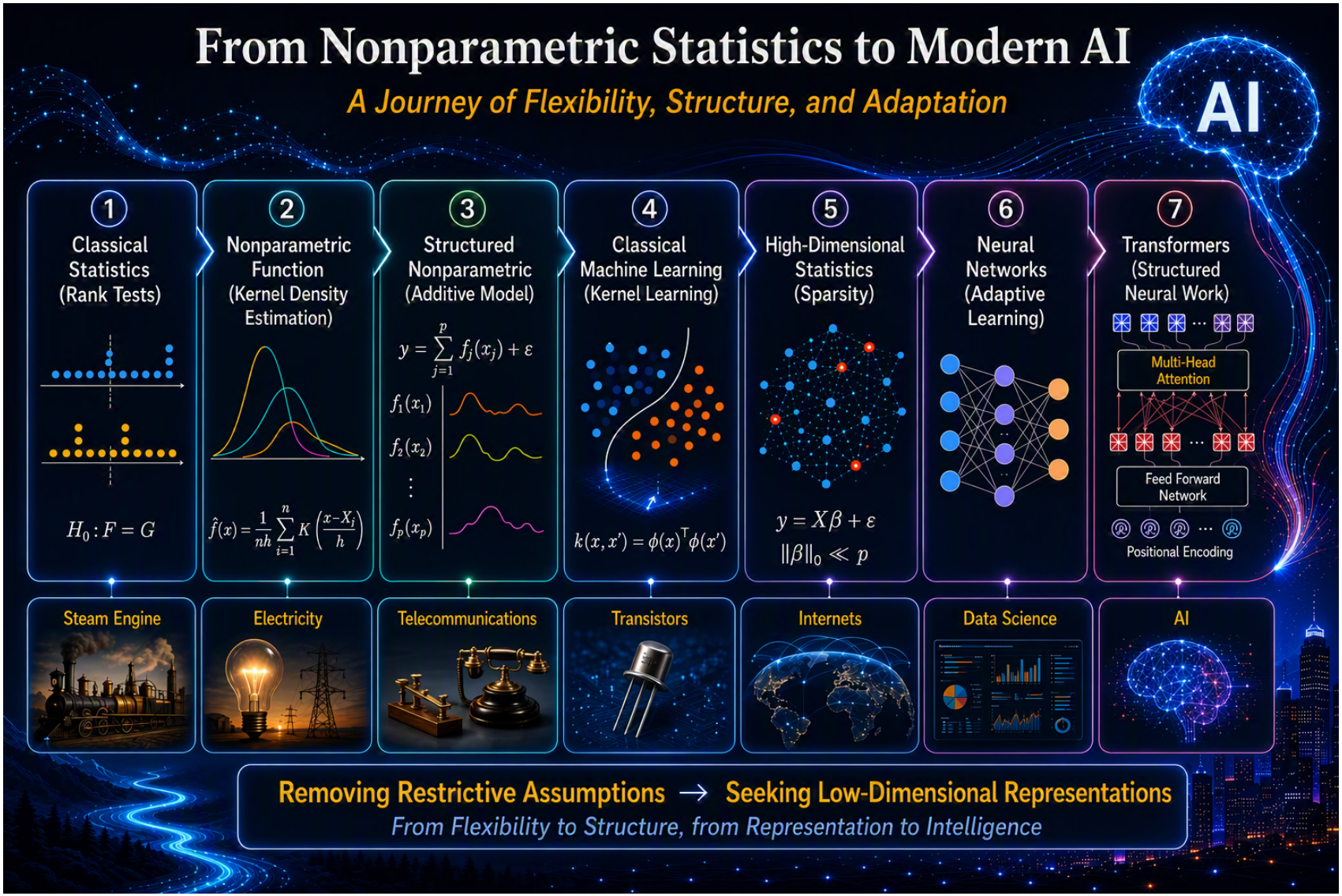


- ★ Representation power of deep neural nets + statist learning algorithms.
- ★ Policy optimization framework of Markov decision processes.
- ★ Help solve complex problems such as high-dim prediction, attribution, causality, mechnism analysis, treatment evaluation in a **data-diven** way.





Evolution of Statistics



Bias–Variance, Flexibility–Learnability

1. Multiple and Nonparametric Regression

1.1. Reg. & Optimal Prediction

1.2 Multiple Linear Regression

1.3. Assessment of Pred Error

1.4. Arts of Model Building

1.5. Ridge Reg & Regularization

1.6. Kernel Regression

1.1. Regression and Optimal Prediction

Purpose of Multiple regression

- ★ Study **assocations** between dependent & independent variables
- ★ **Predictions**
- ★ Screen irrelevant and select useful variables (**attributions**)

Example 1: Zillow is an online real estate database company founded in 2006. An important task for Zillow is to predict the house price. (Training data: 15129 cases, testing: 6484 cases)

Interest: Associations between **housing** and its **attributes**.

- **Response** Y = Housing prices
- **Covariates**
  - No. of bathrooms X_1; No. of bedrooms X_2
  - sqft-living room X_3; sqft-lot X_4
  - zipcode X_5 (70 zipcodes); view X_6 (5 categories)
  - ...

Environmental Data

Example 2: Environmental Data

Interest: Study the association between **levels of pollutants** and number of daily total **hospital admissions** for circulatory and respiratory problems.

— Dependent variable (Y) = Daily number of hospital admissions

— Collected covariates = $\{$

level of pollutant Sulphur Dioxide X_1 (in $\mu g/m^3$),

level of pollutant Nitrogen Dioxide X_2 (in $\mu g/m^3$)

level of respirable suspended particles X_3 (in $\mu g/m^3$)

Ozone level X_4

Temperature X_5 (in ^oC)

Humidity (X_6, in percent)

time X_7 (season, confounding factor),

.... $\}$

Financial Textual Data Analysis

Goal: ★learn sentiments of documents and news;

★use sentiment scores to predict returns.

Investment Strategy: Long top 50 stocks (say) and short on bottom 50 stocks based on predicted returns

—zero net position ★equally weighted or value weighted

Covariates: Textual data, denoted by **X**

(words freq. counts / doc embeddings)

Response: Financial Returns, denoted by Y

Variable Selection: What are the key phases

Best prediction and nonparametric regression

Double Expectation: $EY = E\{E(Y|\mathbf{X})\}$, for any $\mathbf{X}$

Bias-var in prediction: For any predictor $f(\cdot)$, its squared prediction error is

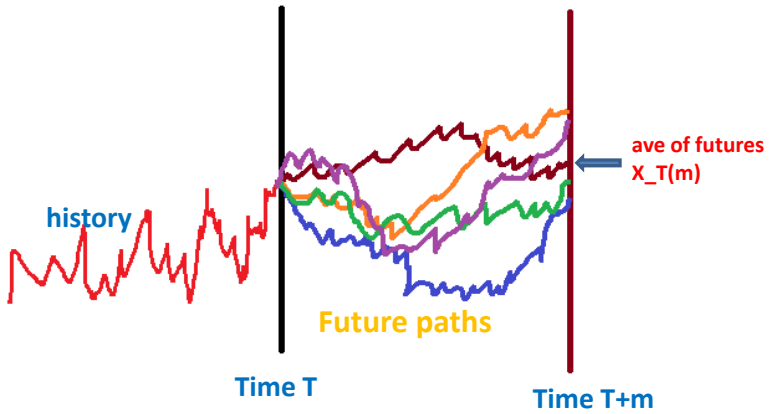
$$\mathbf{PE}(f) = E(Y - f(\mathbf{X}))^2 = \overbrace{\underbrace{E(Y - f^*(\mathbf{X}))^2}_{\mathbf{var}=\mathbf{E}\sigma^2(\mathbf{X})}}^{inherent, irreducible} + \overbrace{\underbrace{E(f^*(\mathbf{X}) - f(\mathbf{X}))^2}_{\mathbf{bias}}}^{reducible}$$

—$f^*(\mathbf{X}) = E(Y|\mathbf{X})$ is called the regression function.

Best prediction: $f^* = \text{argmin}_f E(Y - f(\mathbf{X}))^2$

Nonparametric reg.: Estimating $f^*(\cdot)$ directly

(machine learning, low bias)



Parametric reg.: assume f^* admits certain parametric forms (e.g. linear): $f^* = f(\mathbf{x}, \theta)$

Decompositions

Model: $\underbrace{Y = f^*(\mathbf{X}) + \varepsilon, \qquad E(\varepsilon|\mathbf{X}) = 0,}_{\text{always true, decomposition}}$ $\qquad \varepsilon = Y - f^*(\mathbf{X})$ is Y not explained by X

★Nonparametric: f^* flexible ★parametric: restricted

Bias-var in estimation: letting $\widehat{f}_n$ estimate f; $\bar{f}(\mathbf{x}) = E\widehat{f}_n(\mathbf{x})$, then

$$\underbrace{E(\widehat{f}_n(\mathbf{X}) - f^*(\mathbf{X}))^2}_{2^{nd} \text{ term of PE = learning error = } \textbf{MSE}} = \underbrace{E(\widehat{f}_n(\mathbf{X}) - \bar{f}(\mathbf{X}))^2}_{\textbf{var}(\widehat{f}_n(\mathbf{X}))} + E(\underbrace{\bar{f}(\mathbf{X}) - f^*(\mathbf{X})}_{\textbf{bias}})^2.$$

Role of Modeling:

★variance is small when n large, big when no. of parameters is big

★biases are small when model is complex (no. of parameters is big)

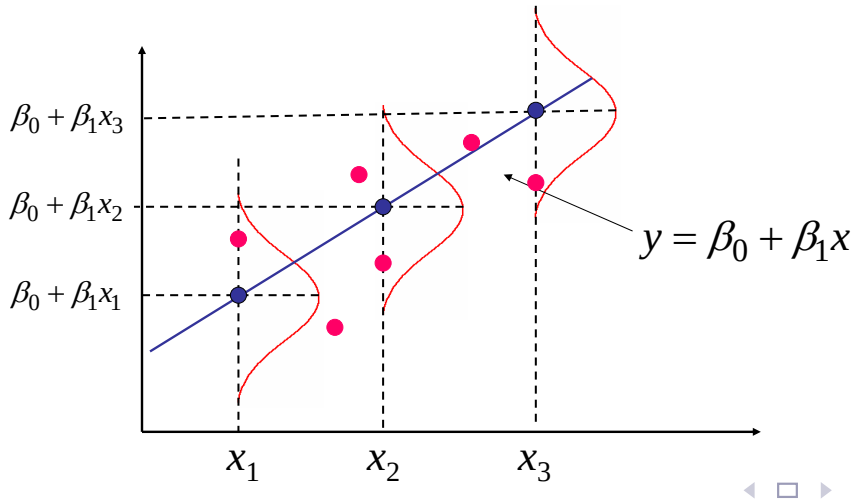
1.2 Multiple Linear Regression

- **Read materials and R-implementations here**

Multiple linear regression model

$$Y = \underbrace{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}_{f^*(\mathbf{X}),\quad \textit{assumption}} + \varepsilon$$

- Y: response / dependent variable
- X_j's: explanatory / independent variables or covariates
- ε: random error not explained / predicted by covariates
- include intercept (**bias**) by setting $X_1 = 1$

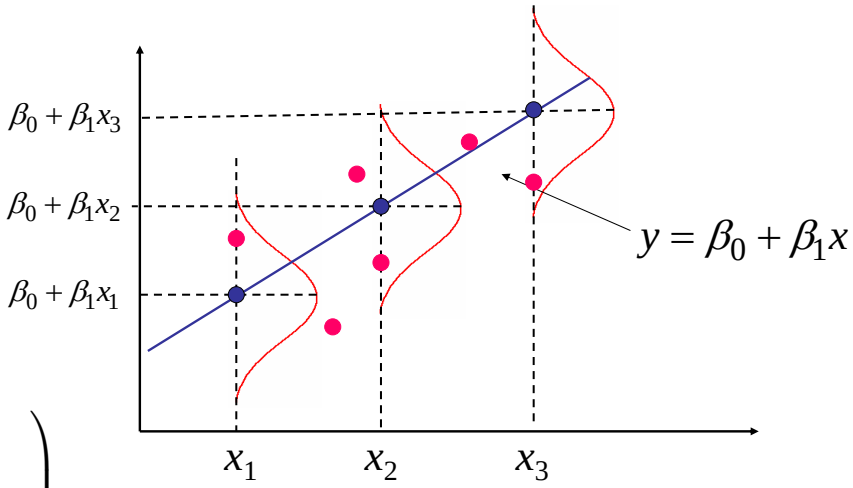


Method of Least-squares

Data: $\left\{\left(x_{i1}, x_{i2}, \cdots x_{ip}, y_i\right)\right\}_{1\le i\le n}$

Model: $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$

Matrix form: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$



$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Method of Least-Squares:

$$\min_{\beta\in\mathbb{R}^p} \quad \mathrm{RSS}\left(\beta\right) \triangleq \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2$$

★RSS stands for **residual sum-of-squares** (in-sample error)

Closed-form solution

Least-squares: Minimize wrt $\beta \in \mathbb{R}^p$

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

Setting gradients to zero yields **normal equations**:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$$

Least-squares estimator: (assume $\mathbf{X}$ has full column rank)

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Multiple R^2: $R^2 = 1 - \frac{\mathrm{RSS}(\widehat{\beta})}{\mathrm{RSS}(1)}$, proportion of variance of y explained by regression. It measures the **goodness-of-fit** (**normalized in-sample error**).

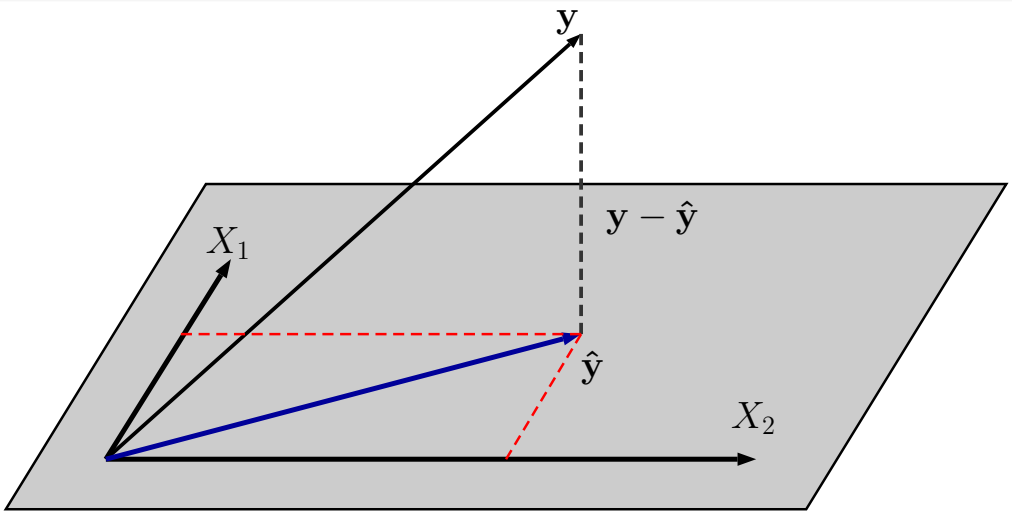
♦ $\mathrm{RSS}(1) = (n-1)\,\mathrm{var}(\mathbf{y})$

Geometric interpretation of least-squares

Fitted value: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\triangleq \mathbf{P}\in\mathbb{R}^{n\times n}}\mathbf{y}$

Theorem 2.1 [Property of projection matrix]

- ★ $\mathbf{P}\mathbf{x}_j = \mathbf{x}_j, \quad j = 1, 2, \cdots, p$
- ★ $\mathbf{P}^2 = \mathbf{P} \quad \text{or} \quad \mathbf{P}(\mathbf{I}_n - \mathbf{P}) = \mathbf{0}$
- ★ Eigenvalues of $\mathbf{P}$ are 0 or 1, with number of 1's $=$ rank($\mathbf{P}$)



project response vector $\mathbf{y}$ onto linear space spanned by $\mathbf{X}$

Statistical properties of least-squares estimator

Assumption:

- **Exogeneity**: $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$;
- **Homoscedasticity**: $\text{var}(\varepsilon|\mathbf{X}) = \sigma^2$.

Statistical Properties:

$\widehat{\beta}$ linear in $\mathbf{Y}$

★ **bias**: $\mathbb{E}(\widehat{\beta}|\mathbf{X}) = \beta$

★ **variance**: $\text{var}(\widehat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

often dropped

■ Recall $\text{cov}(\mathbf{U},\mathbf{V}) = E(\mathbf{U}-\mu_u)(\mathbf{V}-\mu_v)^T$ and $\text{var}(\mathbf{U}) = \text{cov}(\mathbf{U},\mathbf{U})$

$$\text{cov}(\mathbf{AU},\mathbf{BV}) = \mathbf{A}\,\text{cov}(\mathbf{U},\mathbf{V})\mathbf{B}^T, \qquad \text{var}(\mathbf{a}^T\mathbf{U}) = \mathbf{a}^T\,\text{var}(\mathbf{U})\mathbf{a};$$

Gauss-Markov Theorem

- How large is variance?
- Compared with other estimators?

Theorem 2.2 [Gauss-Markov Theorem]

LSE $\widehat{\beta}$ is best linear unbiased estimator (BLUE):

- $\mathbf{a}^T\widehat{\beta}$ is a linear unbiased estimator of parameter $\theta = \mathbf{a}^T\beta$
- for any linear unbiased estimator $\mathbf{b}^T\mathbf{y}$ of θ, (homework)

$$\text{var}(\mathbf{b}^T\mathbf{y}|\mathbf{X}) \geq \text{var}(\mathbf{a}^T\widehat{\beta}|\mathbf{X})$$

Estimation of σ^2: $\widehat{\sigma}^2 = \frac{\text{RSS}}{\mathbf{n}-\mathbf{p}} = \frac{\|\mathbf{y}-\mathbf{X}\widehat{\beta}\|^2}{n-p}$

$\widehat{\sigma}^2$ is is an unbiased estimator of σ^2

Statistical inference

Additional assumption: $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

Under fixed design or conditioning on $\mathbf{X}$,

$$\widehat{\beta} = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon \quad \Longrightarrow \quad \widehat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

- ★ $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, v_j\sigma^2)$ where v_j is jth diag of $(\mathbf{X}^T\mathbf{X})^{-1}$
- ★ $(n-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$ and $\widehat{\sigma}^2$ is indep. of $\widehat{\beta}$.
- ★ $1-\alpha$ **CI for** β_j: $\widehat{\beta}_j \pm t_{n-p}(1-\alpha/2)\sqrt{v_j}\widehat{\sigma}$ (homework)
- ★ $H_0 : \beta_j = 0$: test statistics $t_j = \frac{\widehat{\beta}_j}{\sqrt{v_j}\widehat{\sigma}} \sim_{H_0} t_{n-p}$.

■Holds asymptotically without normality assumption.

Proof of the classical result

1. RSS $= \mathbf{y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{y} = \varepsilon^T(\mathbf{I}_n - \mathbf{P})\varepsilon$
2. By eigendecomposition, $\mathbf{P} = \mathbf{\Gamma}\,\mathrm{diag}(\overbrace{1, \cdots, 1}^{p}, 0, \cdots, 0)\mathbf{\Gamma}^T$.
3. RSS $= \mathbf{Z}^T\,\mathrm{diag}(\overbrace{0, \cdots, 0}^{p}, 1, \cdots, 1)\mathbf{Z} \sim \sigma^2\chi^2_{n-p}$, where $\mathbf{Z} = \mathbf{\Gamma}^T\varepsilon \sim N(0, \sigma^2\mathbf{I}_n)$
4. RSS depends only on $(\mathbf{I}_n - \mathbf{P})\varepsilon$ and $\widehat{\beta}$ depends on $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon$. They are joint normal and uncorrelated $\Longrightarrow$ indep.
5. $\frac{\widehat{\beta}_j - \beta_j}{\sqrt{v_j}\widehat{\sigma}} = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{v_j}\sigma} \Big/ \sqrt{\frac{\mathrm{RSS}}{\sigma^2(n-p)}} \sim t_{n-p}$, by definition.

Zillow Data Analysis: Regression outputs

$$\text{Out-of-sample } R^2 = 1 - \frac{\text{PE of a model}}{\text{PE of naive}} = 1 - \frac{\sum_{i\in \text{TEST}}(y_i - \widehat{y}_i)^2}{\sum_{i\in \text{TEST}}(y_i - \bar{y})^2} = \underbrace{\text{Normalized test error}}_{\bar{y} = \text{in-sample ave}}$$

$$R^2 = 1 - \frac{\sum_{i\in \text{TRAIN}}(y_i - \widehat{y}_i)^2}{\sum_{i\in \text{TRAIN}}(y_i - \bar{y})^2}, \qquad \text{Adjusted } R^2 = 1 - \frac{\frac{1}{n-p}\sum_{i\in \text{TRAIN}}(y_i - \widehat{y}_i)^2}{\frac{1}{n-1}\sum_{i\in \text{TRAIN}}(y_i - \bar{y})^2)},$$

Model 1: lm(price $\sim$ bathrooms + bedrooms + sqft_living + sqft_lot)

R^2-values: In sample = 0.5101, adjusted = 0.5100, Out-sample = 0.5051

```
fit.lm1 = lm(price~bathrooms + bedrooms + sqft_living + sqft_lot, data = train_data)
#fit linear model
summary(fit.lm1)              #summarize the fit

Call:
lm(formula = price ~ bathrooms + bedrooms + sqft_living
+ sqft_lot, data = train_data)

Residuals:
Min       1Q   Median       3Q      Max
-1571803  -143678   -22595   103133  4141210

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.083e+04  8.208e+03   9.848  < 2e-16 ***
bathrooms    3.682e+03  4.178e+03   0.881    0.378
```

```
bedrooms    -5.930e+04  2.753e+03 -21.537  < 2e-16 ***
sqft_living  3.167e+02  3.750e+00  84.442  < 2e-16 ***
sqft_lot    -4.267e-01  5.504e-02  -7.753 9.52e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 257200 on 15124 degrees of freedom
Multiple R-squared:  0.5101,Adjusted R-squared:   0.51
F-statistic:  3937 on 4 and 15124 DF,  p-value: < 2.2e-16
```

- ★ The first part reminds us the models used;
- ★ the second part summarizes the residual statistics;
- ★ the third part depicts estimated coefficients $\widehat{\beta}_j$ (second column), its standard error $\sqrt{v_j}\widehat{\sigma}$ (third column), and t-statistic t_j,
- ★ the last part summarizes overall model fits: $\widehat{\sigma} = 257200$, $n - p = 15124$, $R^2 = 0.5101$, adj-R^2 = 0.51, F-statistic for testing $H_0 : \beta = 0$ is 3937 with degree of freedom $p - 1 = 4$ and $n - p = 15124$. Small P-value suggests that we reject H_0 that these four variables are not related to the house price.

Now, let us compute the out-of-sample R^2, showing roughly percentage of predicability.

```
##### out-of-sample R^2########
fit.lm1.pred.out <- predict(fit.lm1, newdata = test_data)

SS.total <- sum((test_data$price - mean(train_data$price))^2)
SS.residual <- sum( (test_data$price - fit.lm1.pred.out)^2)
1 - SS.residual / SS.total
```

[1] 0.5051489

ANOVA and F-test

ANOVA: Analysis of Variance concerns about validity of a submodel

Hypotheses

H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$
(Smaller model is adequate)

H_1: **At least one** $\beta_{q+1}, \ldots, \beta_p \neq 0$
(Larger model fits better)

Smaller (restricted) model M_0: $y = X_1\beta_1 + \epsilon$, X_1 **is** $n \times q$

Larger (full) model M_1: $y = X_2\beta_2 + \epsilon$, X_2 **is** $n \times p$ $(p > q)$

F Statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p-q)}{\mathrm{RSS}_1/(n-p)}$$

RSS_0 : residual sum of squares under M_0 (restricted)
RSS_1 : residual sum of squares under M_1 (full)

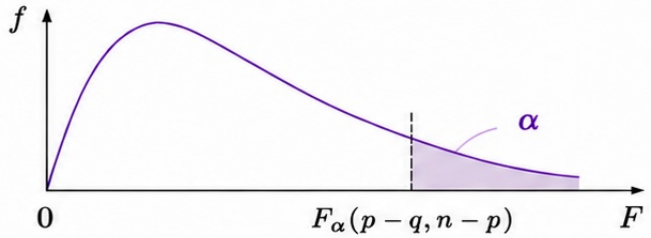
F-Distribution $F_{d1,d2}$

If $U \sim \chi^2_{d1}$ and $V \sim \chi^2_{d2}$ are independent, then

$$F = \frac{(U/d_1)}{(V/d_2)} \sim F_{d1,d2}$$

Decision Rule

Under H_0, $F \sim F_{p-q,n-p}$



Reject H_0 if

$F > F_\alpha(p-q, n-p)$

★Note that $V/d_2 = \chi^2_{d_2}/d_2 = (Z_1^2 + \cdots + Z_{d_2}^2)/d_2 \to 1$, as $d_2 \to \infty$.

Thus, F-distribution becomes a scaled $\chi^2_{d_1}$-distribution.

Correlated noise

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \text{where} \quad \text{var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{W}$$

Transform data: $\mathbf{y}^* = \mathbf{W}^{-1/2}\mathbf{y}$, $\mathbf{X}^* = \mathbf{W}^{-1/2}\mathbf{X}$, $\varepsilon^* = \mathbf{W}^{-1/2}\varepsilon$. Then

$$\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*, \quad \text{with} \quad \text{var}(\varepsilon^*|\mathbf{X}) = \sigma^2\mathbf{I}.$$

General Least-Squares: (assuming **W known**)

$$\min_{\beta\in\mathbb{R}^p} \quad ||\mathbf{y}^* - \mathbf{X}^*\beta||^2 = (\mathbf{y} - \mathbf{X}\beta)^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

Heteroscedastic errors: $\mathbf{W}_i = \sigma^2\,\text{diag}(v_1, \cdots, v_n)$

Weighted Least-squares: $\min_\beta \sum_{i=1}^n (y_i - \mathbf{X}_i^T\beta)^2/v_i$.

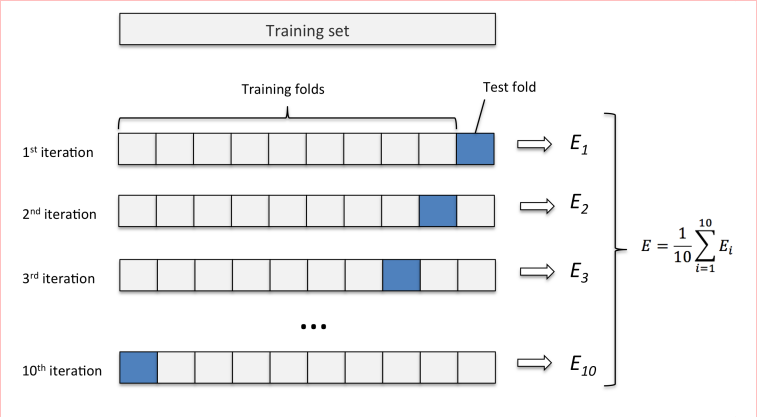
1.3 Assessment of Prediction Error

To compare multiple methods and select tuning parameters

Cross-Validation

k-fold Cross-Validation (CV)

- ★ Divide data randomly and evenly into k subsets;
- ★ Use one fold as **testing set** and remaining as **training set** to compute testing errors;
- ★ Repeat for each of k subsets and average testing errors.



PE for training size: $(1-1/k)n$

Choice of k: $k = n$ (best, but expensive; leave-one out), 10 or 5 (5-fold).

Leave-one-out: $\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{f}^{-i}(\mathbf{x}_i)]^2$,

$\widehat{f}^{-i}(\mathbf{x}_i)$ = predicted value based on the training sample $\{(\mathbf{x}_j, y_j)\}_{j\neq i}$

Bootstrap

★A simple proc. to assess variation of a method, including pred error (PE).

1. Randomly sample n_1 , without replacement, from the data as a traning set. Train a method on the subsample to get $\widehat{f}(\cdot)$
2. Evaluate the method on the remaining data: PE = $\frac{1}{n-n_1}\sum_{i\in test}(Y_i - \widehat{f}(X_i))^2$.
3. Repeat this procedure independently B time and obtain the average PE (and its SD, showing how good PE is assessed)

★L_1 and misclassification loss can be used;

★PE of size n_1 not n.

1.4. Arts of Model Building

Nonlinear and nonparametric regression

"Essentially, all models are wrong, but some are useful."

George Box (1987)

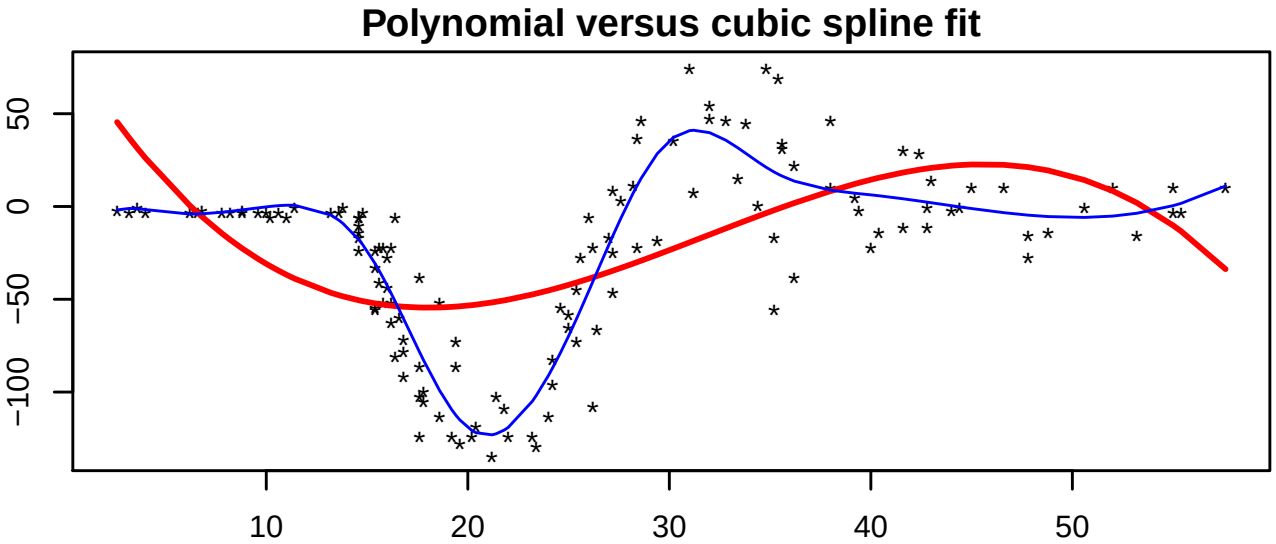
Nolinear regression

Polynomial regression: univariate

$$Y = \overbrace{\beta_0 + \beta_1 X + \cdots + \beta_d X^d}^{\approx f(X)} + \varepsilon$$

★multiple regression with $X_1 = X, \cdots, X_d = X^d$ (**basis function**)

Drawback: not suitable for functions with **varying** degrees of smoothness



motorcycle data: time vs. head acceleration (red: cubic **polynomial**; blue: cubic **splines**)

Spline regression

- ★ piecewise polynomials with degree d and continuous $(d-1)^{th}$ derivative.
- ★ **Knots**: $\{\tau_j\}_{j=1}^{K}$ where discontinuity occurs. ★data quantiles

$$Y = \text{Spline}_K(X) + \varepsilon = \sum_{j=0}^{d+K} \beta_j B_j(X) + \varepsilon$$

Basis functions: $\{1, x, \cdots, x^d, (x-\tau_j)_+^d, \quad j = 1, \cdots, K\} = \{B_j(x)\}_{j=0}^{d+K}$

- ★ cubic spline: $d = 3$, widely used;
- ★ multiple regression with $X_j = B_j(x)$. (feature)

Nonparametric: When K is large, $K_n \to \infty$

Extension to multiple covariates

- Bivariate quadratic regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 \underbrace{X_1 X_2}_{\text{interaction}} + \beta_5 X_2^2 + \varepsilon$$

- Multivariate quadratic regression: (linear in parameters)

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{\mathbf{j \le k}} \beta_{jk} X_j X_k + \varepsilon$$

- Multivariate quadratic regression with main effect and interactions

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{\mathbf{j < k}} \beta_{jk} X_j X_k + \varepsilon$$

Dummy variables

or **indicator variables**, are used to include categorical variables in a reg anal.

Example: Here are a few simple cases (Dichotomous):

$$\text{Gender}\begin{cases}\text{male}\\ \text{female}\end{cases} \qquad \text{Smoking}\begin{cases}\text{Yes}\\ \text{No}\end{cases} \qquad \text{Disease}\begin{cases}\text{present}\\ \text{not present}\end{cases}$$

For a dichotomous variable, we define $X = \begin{cases} 1, & \text{if treatment} \\ 0, & \text{if control} \end{cases}$

Example: Gender difference in income:

$$Y = \text{salary}, \quad X_1 = \text{age}, \quad X_2 = \text{year of exp.}, \quad X_3 = \begin{cases} 1, & \text{if male} \\ 0, & \text{if female} \end{cases}$$

Use of Dummy Variables

a) **No-interaction model**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error}$

$$= \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e, & \text{for female} \\ (\beta_0 + \beta_3) + \beta_1 X_1 + \beta_2 X_2 + e, & \text{for male} \end{cases}$$

■ β_3 is the gender difference after adjusting for X_1, X_2.

b) **Interaction model**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3 + \text{error}$

$$= \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error}, & \text{for female} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4) X_1 + (\beta_2 + \beta_5) X_5 + \text{error}, & \text{male} \end{cases}$$

$\beta_3, \beta_4, \beta_5$ reflect the **gender diff.** wrt salary, age and exp.

c) **Multiple Categories** (Cat, Dog, Cow): Use multiple indicators (X =annimal)

$$X_4 = \mathbf{1}_{cat}(X), \quad , X_5 = \mathbf{1}_{dog}(X), \quad X_6 = \mathbf{1}_{cow}(X)$$

Note $X_4 + X_5 + X_6 = 1$, so only two of them are used (not used = baseline).

(d) **Continuous** variable. Convert into catogories (elimentary, high, college) for nonlinearity

Model 2: (heterogeneity + interaction) Consider four variables and **their interaction** with zipcode. This amounts to fit a multiple regression with the four variables for each zip code. There are 70 different zip codes, representing different intercepts. The estimated coefficients are typically presented as the difference (contrast) to the baseline factor (zipcode: 98001).

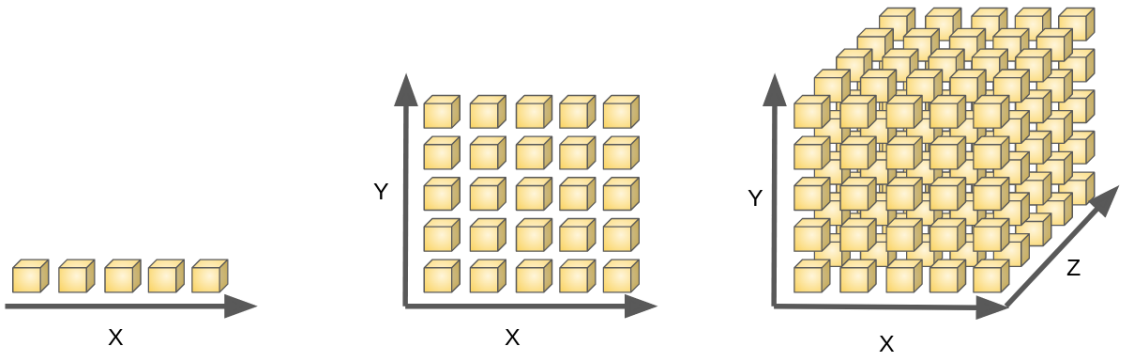
```
train_data$zipcode = as.factor(train_data$zipcode)  #treat zipcode as factor
test_data$zipcode = as.factor(test_data$zipcode)

fit.lm4 = lm(price ~ bathrooms + bedrooms + sqft_living + sqft_lot+ zipcode
+ zipcode*bathrooms + zipcode*bedrooms+zipcode*sqft_lot
+ zipcode*sqft_living, data = train_data)
summary(fit.lm4)

Residual standard error: 156900 on 14779 degrees of freedom
Multiple R-squared:  0.822, Adjusted R-squared:  0.8178
F-statistic: 195.5 on 349 and 14779 DF,  p-value: < 2.2e-16


fit.lm4.pred.out <- predict(fit.lm4, newdata = test_data)
SS.residual <- sum( (test_data$price - fit.lm4.pred.out)^2)
1 - SS.residual / SS.total
[1] 0.7950443
```

Multivariate spline regression

Idea: Tensor products of univariate basis functions

$$\{B_{i_1}(x_1)B_{i_2}(x_2)\cdots B_{i_p}(x_p)\}_{i_1=1}^{b_1}\cdots{}_{i_p=1}^{b_p}$$

Drawbacks: **curse of dimensionality**, namely, number of basis functions scales exponentially with p



d	n^d	accuracy $n^{-\frac{2}{d+4}}$
1	**100**	**100**
2	10^4	250
5	10^{10}	4000
10	10^{20}	40000
100	10^{200}	$4*10^{41}$

Structured multivariate regressions

Remedy: Add additional structure to $f(\cdot)$

Example: Additive model

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \varepsilon$$

- Number of basis functions scales **linearly** with p

Example: Bivariate interaction models:

$$Y = \sum_{1 \le i \le j \le p} f_{ij}(X_i, X_j) + \varepsilon$$

- Number of basis functions scales **quadratically** with p
- Implementation: Bivariate tensors

Q: How to write?

1.5. Ridge Regression and Regularization

Ridge Regression

Drawbacks of OLS: ★$n > p$

★large variance when collinearity: $\mathrm{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

Remedy: Ridge regression (Hoerl and Kennard, 1970)

$$\widehat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

★$\lambda > 0$ is a regularization parameter.

Tikhonov (1943) regularization

Interpretation: Penalized LS $\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$.

—Setting the gradient to zero, we get $\mathbf{X}^T(\mathbf{X}\beta - \mathbf{y}) + \lambda\beta = \mathbf{0}$.

Bayesian estimator: with prior $\beta \sim N(0, \frac{\sigma^2}{\lambda}\mathbf{I}_p)$.

Generalization: ℓ_q Penalized Least Squares

ℓ_q penalized least-squares estimate (a case of regularization):

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_q^q, \quad q \geq 0.$$

- λ tuning (regularzation) parameter, $\|\beta\|_q^q = |\beta_1|^q + \cdots + |\beta_p|^q$
- $q = 0$ is the best subset selection $\|\beta\|_0 = \#\{j : \beta_j \neq 0\}$
- Only $q = 2$ admits a closed-form solution.
- Known as Bridge estimator (Frank and Friedman, 1993);
- When $q = 1$, called Lasso estimator (Tibshirani, 1996);
- Folded concave when $0 < q < 1$ and convex when $q > 1$;

1.6 Kernel Regression

Prediction by similarity

Prediction by similarity

Theorem 2.4. Alternative expression $\widehat{\beta}_\lambda = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I})^{-1}\mathbf{y}$

Prediction at $\mathbf{x}$ is $\widehat{y} = \mathbf{x}^T\widehat{\beta}_\lambda = \mathbf{x}^T\mathbf{X}^T \underbrace{(\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I})^{-1}\mathbf{y}}_{\widehat{\alpha}} = \sum_{i=1}^n \widehat{\alpha}_i \langle \mathbf{x}, \mathbf{x}_i \rangle$.

Note $(\mathbf{X}\mathbf{X}^T)_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ and $\widehat{\alpha}$ depends on covariates only via their similarities.

- Prediction depends only **pairwise inner products**; **similarity**
- Generalize to other **similarity measures** $K(\cdot, \cdot)$, called **kernel** trick. e.g. $K\left(\text{  }, \text{  }\right) = +10 \quad \mathcal{K}\left(\text{  }, \text{  }\right) = -10$

Kernel regression

Kernel: $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$ is PSD, for any $\{\mathbf{x}_i\}_{i=1}^n$.

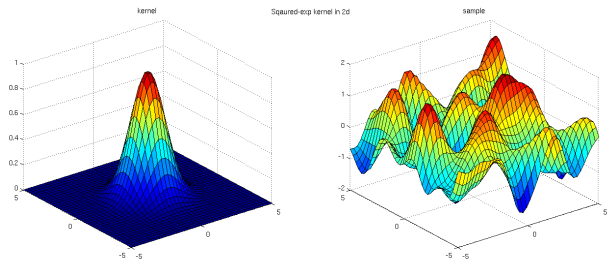
Commonly used kernels: $K(\mathbf{u}, \mathbf{v})$

- ★linear $\langle \mathbf{u}, \mathbf{v}\rangle$
- ★polynomial $(1 + \langle \mathbf{u}, \mathbf{v}\rangle)^{\mathbf{d}}$, $d = 2, 3, \cdots$;
- ★Gaussian $e^{-\gamma\|\mathbf{u}-\mathbf{v}\|^2}$
- ★Laplacian $e^{-\gamma\|\mathbf{u}-\mathbf{v}\|}$

Basis: $\{K(\cdot, \mathbf{x}_j)\}_{j=1}^n$ and express $f(\mathbf{x}) = \sum_{j=1}^n \alpha_j K(\mathbf{x}, \mathbf{x}_j)$. Fit

$$\min_{\alpha\in\mathbb{R}^n} \left\{ \|\mathbf{y} - \mathbf{K}\alpha\|^2 + \lambda\alpha^T\mathbf{K}\alpha \right\}, \quad \mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))$$

- No curse-of-dim in implementation!
- High-order interactions
- Not flexible enough



Kernel ridge regression

Kernel ridge regression

With $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j)) \in \mathbb{R}^{n \times n}$, prediction at $\mathbf{x}$ is

$$\widehat{y} = (K(\mathbf{x}, \mathbf{x}_1), \cdots, K(\mathbf{x}, \mathbf{x}_n))(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y},$$

★ $\widehat{y} = \overbrace{\widehat{f}(\mathbf{x})}^{\text{pred}} = \sum_{i=1}^{n} \overbrace{\alpha_i}^{\text{weight}} K(\mathbf{x}, \mathbf{x}_i),$ $\qquad \widehat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y};$

testing: $\mathbf{x}$ in $\widehat{f}(\mathbf{x})$; testing: $\mathbf{x}$ in $K(\mathbf{x}, \mathbf{x}_i)$; training: $\mathbf{x}_i$ in $K(\mathbf{x}, \mathbf{x}_i)$

★ tune the parameter λ to minimize prediction errors.