EECS E6690: Statistical Learning for Biological and Information Systems Lecture 2: Multiple Linear Regression

Prof. Predrag R. Jelenković
Time: Tuesday 4:10-6:40pm
303 Seeley W. Mudd Building

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.ee.columbia.edu/∼predrag

Last lecture: Intro to stat learning

Supervised vs. Unsupervised learning

Supervised:

- Let Y be the output variable, and X the input vector $X = (X_1, X_2, \ldots, X_p)$. Then
$$Y = f(X) + \epsilon$$
- Want to estimate f
- ϵ is unavoidable/irreducible noise that is independent of X, zero mean
- How to estimate f from the data? How to evaluate the estimate?
- Errors: irreducible, reducible, bias
- Overfitting and testing
John von Neumann on overfitting: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Unsupervised: No $f(\cdot)/Y$, just X

Last lecture: Estimation and Testing

Estimation:

- Select a class of function for f: Hypothesis class $\mathcal{H}$
  say, $\mathcal{H}$ are linear functions, i.e., linear regression
- Select as distance metric, i.e., **loss function**, which measures the error between $f \in \mathcal{H}$ and data
- Optimization: find $\hat{f} \in \mathcal{H}$ which minimizes the error/loss function

Testing: How good is $\hat{f}$ on unseen data? Two approaches:

- Analytical (first 3 lectures)
  - Make some analytical assumptions, e.g. Gaussian
  - Compute distributions for the parameters of interest
  - Develop statistical tests to characterize $\hat{f}$: t-test, F-test, etc
- Numerical (rest of the class)
  - Split data into training and testing
  - Use training data to find $\hat{f}$
  - Use testing data to evaluate how good is $\hat{f}$

Last lecture

- Install and get familiar with R (attend the recitation session)

Brief stat review:

- $X_1, X_2, \ldots, X_n$ – i.i.d. with mean μ and variance σ^2
- Estimators of mean and variance
  - Sample mean
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
  - Sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
  - $\bar{X}$ and S^2 are **unbiased** estimators, i.e.
$$\mathbb{E}\bar{X} = \mu, \qquad \text{and} \qquad \mathbb{E}S^2 = \sigma^2$$
  - Variability of $\bar{X}$: SE$(\bar{X})$ =Standard Error of the mean
$$\mathsf{Var}(\bar{X}) = \sigma^2/n \approx (\mathsf{SE}(\bar{X}))^2 = S^2/n$$

Last lecture: Variability

- If $X_1, \ldots, X_n$ are i.i.d. and **normal/Gaussian**, then
  - $\bar{X}$ is normal
  - S^2 has Chi - square distribution:

    $$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1},$$

    $\chi^2_{n-1} =$ sum of $(n-1)$ squares of independent standard normal variables
  - $\bar{X}$ and S^2 are independent
  - t-value and Student's t-distribution:

    $$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1},$$

    William Gosset, 1908, under pen name Student
- ... if X is not normal/Gaussian, then use the CLT

Last lecture: Hypothesis testing - t-test

t_n has a known symmetric and bell shaped density
(use $\Gamma(k+1/2) \approx \sqrt{k}\Gamma(k)$ for large k)

$$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad \text{(large } n\text{)}$$

t-**test:**

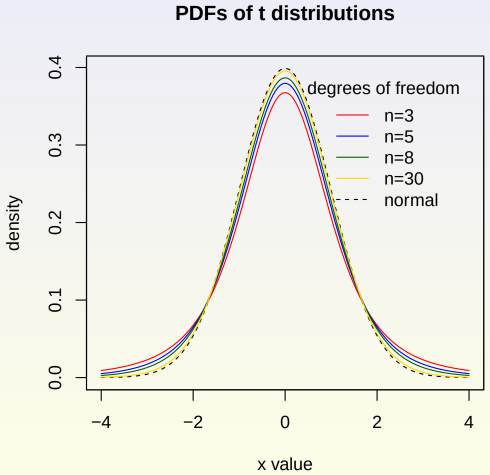
- Null hypothesis $\mathcal{H}_0 : \mu = \mu_0$
- Under $\mathcal{H}_0$, compute t-value and p-value:

$$t = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}}, \qquad p = \mathbb{P}[|t_{n-1}| \geq |t|]$$

- Since large values of t unlikely under $\mathcal{H}_0$, typically
  - pick a significance value, say $\alpha = 0.05$
  - reject $\mathcal{H}_0$ if $p < \alpha$, say $p < 0.05$
  - accept $\mathcal{H}_0$ if $p \geq \alpha$, say $p \geq 0.05$

t-distribution

- Zero mean
- Variance $(n > 2)$: $n/(n-2)$



Last lecture: Linear regression

- Simple approach to supervised learning
- Assumes linear dependence of Y on $X_1, X_2, \ldots, X_p$
  Almost never true in reality.
- Extremely useful both conceptually and practically
- Linear model in 1D ($p = 1$): $X = X_1$

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

- Estimate β_0 and β_1 by **minimizing residuals**

$$y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- Norm selection (distance measure) is important
  e.g., l_2 vs. l_1
- l_2 regression: Least squares (r_{xy} - correlation coefficient)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{S_y}{S_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Last lecture: Statistics of $\hat{\beta}_0$ and $\hat{\beta}_1$

- **Repeated sampling**
- $\hat{\beta}_0$ and $\hat{\beta}_1$ vary
- **Unbiased estimators**:

$$\mathbb{E}\hat{\beta}_0 = \beta_0 \quad \text{and} \quad \mathbb{E}\hat{\beta}_1 = \beta_1$$

- Variances: (model $Y = f(X) + \epsilon$, ϵ-Gaussian)

$$\mathsf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\mathsf{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

  where $\sigma^2 = \mathsf{Var}(\epsilon)$
- An estimate of σ^2:

$$\mathsf{RSE}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \mathsf{RSS},$$

  where RSE is the Residual Standard Error

Last lecture: Hypothesis testing and confidence intervals

- Normality assumption: $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- t-statistic:

$$\frac{\hat{\beta}_1 - \beta_1}{\mathsf{SE}(\hat{\beta}_1)} \sim t_{n-2},$$

  where

$$\mathsf{SE}(\hat{\beta}_1)^2 = \frac{1}{n-2} \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Hypothesis testing using t statistics
- $(1-\gamma)$ confidence interval (say $\gamma = 5\%, 1 - \gamma = 95\%$):

$$[\hat{\beta}_1 - \mathsf{SE}(\hat{\beta}_1) \cdot t_{\gamma/2,n-2}, \hat{\beta}_1 + \mathsf{SE}(\hat{\beta}_1) \cdot t_{\gamma/2,n-2}]$$

  where $t_{\gamma/2,n-2}$ is the $(1-\gamma/2)$-th quantile of the t_{n-2} distribution $\mathbb{P}[-t_{\gamma/2,n-2} \leq t_{n-2} \leq t_{\gamma/2,n-2}] = 1 - \gamma$, i.e,

$$\mathbb{P}[\hat{\beta}_1 - \mathsf{SE}(\hat{\beta}_1) \cdot t_{\gamma/2,n-2} \leq \beta_1 \leq \hat{\beta}_1 + \mathsf{SE}(\hat{\beta}_1) \cdot t_{\gamma/2,n-2}] = 1 - \gamma$$

Example

Recall the 1D model from Lecture 1: $Y = \beta_0 + \beta_1$ (TV advertising) $+ \epsilon$, for which we computed the estimates on $n = 200$ data points

$$\hat\beta_0 = 0.047537, \qquad \hat\beta_1 = 7.032594.$$

$\text{SE}(\hat\beta_1) = 0.457843$ and degrees of freedom, $\text{DF} = n - 2 = 198$.

Hypothesis testing: $\mathcal{H}_0 : \beta_1 = 0 \quad$ vs. $\quad \mathcal{H}_A : \beta_1 \neq 0$ Hence, t-statistics is

$$t = \frac{\hat\beta_1 - 0}{\text{SE}(\hat\beta_1)} = \frac{7.032594}{0.457843} = 15.36027$$

$\Rightarrow$ p-val $= \mathbb{P}[|t_{198}| > 15.36027] =$ 2*(1-pt(15.36027,df=198)) ≈ 0

$\Rightarrow$ Reject $\mathcal{H}_0$, (pt() is a probability distribution of t-variable in R).

Confidence interval (CI): say 95% CI, $\gamma = 5\%$

$t_{2.5\%,198} =$ |qt(.025,df=198)| $= 1.972017$, qt()=t-quantile function in R.

$$\begin{aligned} 95\%\text{ CI for true } \beta_1 :& (\hat\beta_1 \pm t_{2.5\%,198} \times \text{SE}(\hat\beta_1)) \\ =& (7.032594 \pm 1.972017 \times 0.457843) \\ =& \boxed{(6.12972, 7.935468)} \end{aligned}$$

Multidimensional linear regression

- Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Example

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \epsilon$$

- Interpretation: β_i is the average effect on Y of a one unit increase in X_i, holding all other predictors fixed

- Notes
  - Ideally the predictors are uncorrelated
  - Correlations amongst predictors cause problems
    - increased variance of coefficients
    - tricky interpretations (example: $X_1 = X_2^2$)
  - Claims of causality should be avoided for observational data

l_2 regression

- n observations: $(y_i, x_{i,1}, x_{i,2}, \dots, x_{i,p})$
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, the prediction is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p,$$

or in matrix form

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

- Minimize (over $\beta_1, \dots, \beta_p$) the residual sum of squares

$$\mathsf{RSS}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

l_2 regression: Solution

- Minimize RSS: Differentiating RSS($\boldsymbol{\beta}$), we get

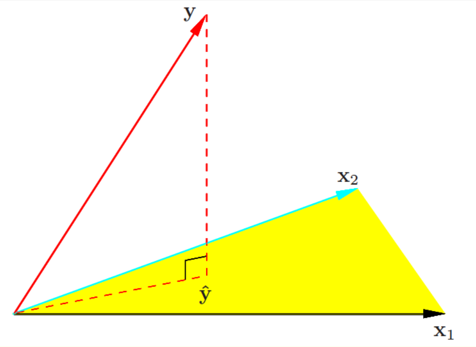
$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$$

- Solution $\boldsymbol{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$: assuming $\boldsymbol{X}^\top\boldsymbol{X}$ is full rank

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \qquad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

  If $\boldsymbol{X}^\top\boldsymbol{X}$ is singular, use the pseudo-inverse, which finds $\hat{\boldsymbol{\beta}}$ with the smallest l_2 norm, smallest $\|\hat{\boldsymbol{\beta}}\|_2^2$.

- Geometry



Example: Advertising data

```
> lm2<-lm(adv$Sales~adv$TV+adv$Radio+adv$Newspaper)
> summary(lm2)

Call:
lm(formula = adv$Sales ~ adv$TV + adv$Radio + adv$Newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.938889   0.311908   9.422   <2e-16 ***
adv$TV         0.045765   0.001395  32.809   <2e-16 ***
adv$Radio      0.188530   0.008611  21.893   <2e-16 ***
adv$Newspaper -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

> cor(adv[,2:5])
                  TV      Radio  Newspaper     Sales
TV        1.00000000 0.05480866 0.05664787 0.7822244
Radio     0.05480866 1.00000000 0.35410375 0.5762226
Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
Sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Solution: Algebraic/geometric interpretations

- Ideally $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ (RSS $= 0$), but this equation has no solution (except in trivial cases), since $\boldsymbol{y} \notin C(\boldsymbol{X})$
  $C(\boldsymbol{X}) =$ column space, i.e., hyperplane formed by columns of $\boldsymbol{X}$.
- Instead, we solve $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the l_2-projection matrix onto $C(\boldsymbol{X})$
  - $\boldsymbol{P}$ is sometimes called "hat" matrix, denoted as $\boldsymbol{H}$, since it puts a hat on $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y} \equiv \boldsymbol{H}\boldsymbol{y}$
- Equivalently, $\hat{\boldsymbol{\beta}}$ satisfies

$$\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- A unique solution exists when the columns of $\boldsymbol{X}$ are linearly independent – in that case, $\boldsymbol{X}^\top\boldsymbol{X}$ is full-rank and positive definite
- Consequences:
  - $(\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is perpendicular to $C(\boldsymbol{X})$
  - $\boldsymbol{0} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top\boldsymbol{X} = (\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y})^\top\boldsymbol{X} = \boldsymbol{y}^\top(\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X})$
  - $\sum_{i=1}^n (y_i - \hat{y}_i) = (\boldsymbol{y} - \hat{\boldsymbol{y}})\boldsymbol{1} = 0$
  - $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$

Computational Complexity

Note that $\boldsymbol{X}^\top \boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix, and thus finding $\hat{\boldsymbol{\beta}}$ requires solving linear system of $(p+1)$ equations

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y},$$

which has $O(p^3)$ computational complexity.

High dimensional data:

- Suppose $p \gg n$
  $p = \#$ of dimensions, $n = \#$ of samples
- Example: $n = 100$ samples of $p = 10,000$ gene expressions

$$\text{computational complexity} = 10^{12} (!)$$

- Can we do better than that?

Dual solution: dot products and kernels

Note that $\hat{\boldsymbol{\beta}}$ can be represented as a linear combination of data

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \left(\boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-2} \boldsymbol{X}^\top \boldsymbol{y} \right) =: \boldsymbol{X}^\top \boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i = (1, x_{i,1}, \ldots x_{i,p})$ is the ith data point, which implies

$$\hat{\boldsymbol{y}} = \boldsymbol{X} \hat{\boldsymbol{\beta}} = \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{\alpha} =: \boldsymbol{K} \boldsymbol{\alpha},$$

where $\boldsymbol{K}$ is a matrix of dot products, also known as Kernel or Gram matrix, which is symmetric and positive definite

$$K_{kj} = \langle \boldsymbol{x}_k, \boldsymbol{x}_j \rangle := \sum_{l=0}^{p} x_{k,l} x_{j,l}.$$

Hence, by minimizing the dual problem $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}\|_2^2$, one finds (assuming $\boldsymbol{K}$ being non-singular)

$$\boldsymbol{\alpha} = \boldsymbol{K}^{-1} \boldsymbol{y}$$

which has computational complexity $O(n^3)$. Direct computation of K requires $O(n^2 p)$ operations, resulting in total complexity $O(n^2(p+n)) \ll O(p^3)$ when $n \ll p$.

We will be back to dual (Kernel) solution throughout the course.

Back to the model: How good is the model fit?

- Total sum of squares: $\mathsf{TSS} = (\boldsymbol{y} - \bar{y}\mathbf{1})^\top (\boldsymbol{y} - \bar{y}\mathbf{1})$
- Explained sum of squares: $\mathsf{ESS} = (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1})^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1})$
- Then

$$\begin{aligned} \mathsf{TSS} &= (\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1}) \\ &= \mathsf{RSS} + \mathsf{ESS} + 2(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}) \\ &= \mathsf{RSS} + \mathsf{ESS} \end{aligned}$$

- A measure of quality of the model

$$R^2 = \frac{\mathsf{ESS}}{\mathsf{TSS}} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}$$

- $R^2 \uparrow$ as more explanatory variables are added to the model – need to consider the number of variables

Example: Advertising data

```
> summary(lm(adv$Sales~adv$TV+adv$Radio))

Call:
lm(formula = adv$Sales ~ adv$TV + adv$Radio)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
adv$TV       0.04575    0.00139  32.909   <2e-16 ***
adv$Radio    0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

- R^2
- Are all predictors useful? Which are?

Distribution of $\hat{\boldsymbol{\beta}}$

- Normality assumption: i.i.d. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is also normally distributed, with mean $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ and density

$$f_{\boldsymbol{y}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ is also normal with

$$\mathbb{E}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{\beta}$$

  and

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \\ &= \mathbb{E}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\left((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\right)^\top = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\end{aligned}$$

- Hence $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$

Residuals

- Residuals $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are also **normal** with zero mean

$$\mathbb{E}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] = \mathbf{0}$$

  and covariance

$$\begin{aligned}\mathbb{E}(\boldsymbol{y} - \hat{\boldsymbol{y}})(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top &= \mathbb{E}(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y})^\top \\ &= \mathbb{E}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\boldsymbol{I} - \boldsymbol{P})^\top = \sigma^2(\boldsymbol{I} - \boldsymbol{P})\end{aligned}$$

  since $\boldsymbol{P}^2 = \boldsymbol{P}$ and

$$\boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}$$

Estimating σ

- RSS $= (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{\epsilon}^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}$, and

$$\begin{aligned}
\text{rank}(I - P) &= \text{tr}(\boldsymbol{I} - \boldsymbol{P}) \\
&= \text{tr}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top) \\
&= n - \text{tr}(\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) = n - p - 1
\end{aligned}$$

- Then it can be shown (Cochran's Theorem - next class)

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p-1}$$

- Estimator (since $\chi^2_{n-p-1} = n - p - 1$)

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n-p-1}}$$

Back to testing

- $\mathcal{H}_0 : \beta_j = 0$
- Intuition: reject $\mathcal{H}_0$ if $\hat{\beta}_j$ is "large"
- How large?
- Under $\mathcal{H}_0$, $\hat{\beta}_j$ is $\mathcal{N}(0, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j})$
- Consider t-statistic

$$\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}} = \frac{\frac{\hat{\beta}_j}{\sigma\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{j,j}}}}{\sqrt{\frac{\mathsf{RSS}}{\sigma^2(n-p-1)}}} \sim t_{n-p-1}$$

F-test

- Better idea: use RSS to test instead of $\hat{\beta}$
- $\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- $\mathcal{H}_1$: exists j such that $\beta_j \neq 0$
- Under $\mathcal{H}_0$, we have a null model: $Y = \beta_0 + \epsilon$
- Let RSS_0 be the residual sum of squares under $\mathcal{H}_0$
- Under $\mathcal{H}_0$:

$$\frac{\mathsf{RSS}_0 - \mathsf{RSS}}{\sigma^2} = \frac{\mathsf{TSS} - \mathsf{RSS}}{\sigma^2} \sim \chi^2_p$$

and

$$\frac{\frac{\mathsf{TSS}-\mathsf{RSS}}{p}}{\frac{\mathsf{RSS}}{n-p-1}} \sim F_{p,n-p-1}$$

- Back to the example:

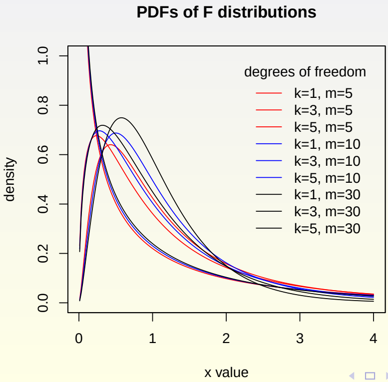
```
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

F-distribution

- $F_{k,m}$: independent $V \sim \chi^2_k$ and $W \sim \chi^2_m$

$$\frac{V/k}{W/m} \sim F_{k,m}$$

- Mean $(m > 2)$: $m/(m-2)$
- Variance $(m > 4)$: $\frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}$



F-test

- $\mathcal{H}_0 : \beta_j = 0$
- Under $\mathcal{H}_0$, we have a reduced model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots \beta_p X_p + \epsilon$$

- Refer to the reduced model by index $-j$
- Intuition: while $\text{RSS} \leq \text{RSS}_{-j}$...
  - if $\text{RSS} \ll \text{RSS}_{-j}$, then reject $\mathcal{H}_0$
  - if $\text{RSS} \approx \text{RSS}_{-j}$, then accept $\mathcal{H}_0$
- Under $\mathcal{H}_0$:

$$\frac{\text{RSS}_{-j} - \text{RSS}}{\sigma^2} \sim \chi^2_1$$

and

$$\frac{\text{RSS}_{-j} - \text{RSS}}{\frac{\text{RSS}}{n-p-1}} \sim F_{1,n-p-1}$$

F-test

- (m) denotes a sub-model obtained by a linear constraint on $\boldsymbol{\beta}$
- Examples
  - $\beta_1 = \beta_2 = \ldots = \beta_p$: $Y = \beta_0 + \beta_1(X_1 + X_2 + \cdots + X_p) + \epsilon$
  - $\beta_1 = \beta_2$: $Y = \beta_0 + \beta_1(X_1 + X_2) + \beta_3 X_3 + \ldots + \beta_p X_p + \epsilon$
- Testing: $\mathcal{H}_0$ (reduced model) vs. $\mathcal{H}_1$ (complete model)
- $q < p$ is the number of explanatory variables in the reduced model
- Under $\mathcal{H}_0$:

$$\frac{\mathsf{RSS}_{(m)} - \mathsf{RSS}}{\sigma^2} \sim \chi^2_{p-q} \quad \Rightarrow \quad \frac{\frac{\mathsf{RSS}_{(m)} - \mathsf{RSS}}{p-q}}{\frac{\mathsf{RSS}}{n-p-1}} \sim F_{p-q,n-p-1}$$

Qualitative predictors

- `Credit.csv` data set
- 400 observations:

```
"","Income","Limit","Rating","Cards","Age","Education","Gender","Student","Married","Ethnicity","Balance"
"1",14.891,3606,283,2,34,11," Male","No","Yes","Caucasian",333
"2",106.025,6645,483,3,82,15,"Female","Yes","Yes","Asian",903
"3",104.593,7075,514,4,71,11," Male","No","No","Asian",580
"4",148.924,9504,681,3,36,11,"Female","No","No","Asian",964
"5",55.882,4897,357,2,68,16," Male","No","Yes","Caucasian",331
.
.
.
"398",57.872,4171,321,5,67,12,"Female","No","Yes","Caucasian",138
"399",37.728,2525,192,1,44,13," Male","No","Yes","Caucasian",0
"400",18.701,5524,415,5,64,7,"Female","No","No","Asian",966
```

- Quantitative predictors: Income (in thousands), Limit (credit), Rating, Cards (number of), Age, Education (years of), Balance
- Qualitative predictors (factors): Gender, Student, Married, Ethnicity

Incorporating qualitative predictors

- Dependency of Balance on Gender
- Ignore all other variables
- Gender has two levels:

$$X_i = \begin{cases} 1, & \text{if } i\text{th individual is female} \\ 0, & \text{if } i\text{th individual is male} \end{cases}$$

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- Interpretation
  - β_0: average Balance among males
  - $\beta_0 + \beta_1$: average Balance among females
  - β_1: average difference in Balance between females and males
- The 1/0 encoding is arbitrary. Can use another scheme – only the interpretation changes

Example

```
> credit <- read.csv("credit.csv",header=TRUE,sep=",")
> lm3<-lm(Balance~Gender,data=credit)
> summary(lm3)

Call:
lm(formula = Balance ~ Gender, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-529.54 -455.35  -60.17  334.71 1489.20

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    509.80      33.13  15.389   <2e-16 ***
GenderFemale    19.73      46.05   0.429    0.669
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611,        Adjusted R-squared:  -0.00205
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

Incorporating qualitative predictors

- When a factor has more than two levels, a single dummy variable can not represent all possible values
- In that case, create additional dummy variables
- Example: Ethnicity

$$X_{i,1} = \begin{cases} 1, & \text{if } i\text{th individual is Asian} \\ 0, & \text{if } i\text{th individual is not Asian} \end{cases}$$

$$X_{i,2} = \begin{cases} 1, & \text{if } i\text{th individual is Caucasian} \\ 0, & \text{if } i\text{th individual is not Caucasian} \end{cases}$$

- Model:

$$Y = \begin{cases} \beta_0 + \beta_1 + \epsilon, & \text{if } i\text{th individual is Asian} \\ \beta_0 + \beta_2 + \epsilon, & \text{if } i\text{th individual is Caucasian} \\ \beta_0 + \epsilon, & \text{otherwise} \end{cases}$$

Example

```
> lm4<-lm(Balance~Ethnicity,data=credit)
> summary(lm4)

Call:
lm(formula = Balance ~ Ethnicity, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774
EthnicityCaucasian   -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188,        Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

Extensions: Interactions

- Relax additivity
- Back to `Advertising` data set
- Suppose spending money on radio advertising increases the effectiveness of TV advertising
- New model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

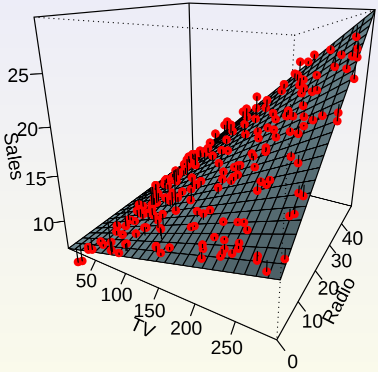
- $X_1 X_2$ – just multiply observations
- Hierarchical principle: if interaction is in the model, main effects are in the model, even if main effects are not significant

Example

```
> lm5<-lm(Sales~TV+Radio+TV*Radio,data=adv)
> summary(lm5)

Call:
lm(formula = Sales ~ TV + Radio + TV * Radio, data = adv)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_3$: the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa)

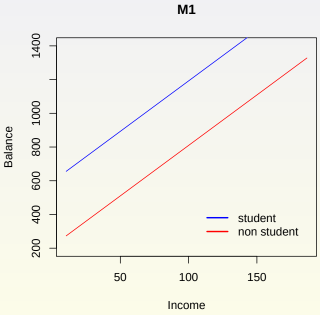
Example

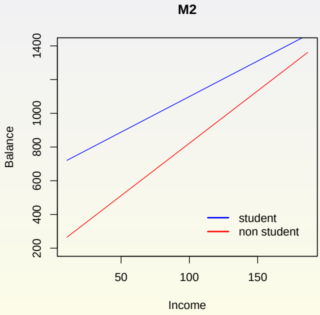


Example

- `Credit` data set
- Y = Balance, X_1 = Income, X_2 = Student $\in \{0, 1\}$
- Two models:

$$\mathsf{M}_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\mathsf{M}_2 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

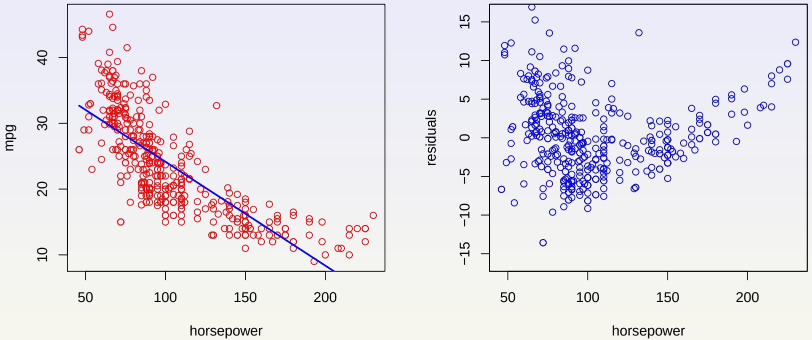




- Changes in income affect students and non-students differently

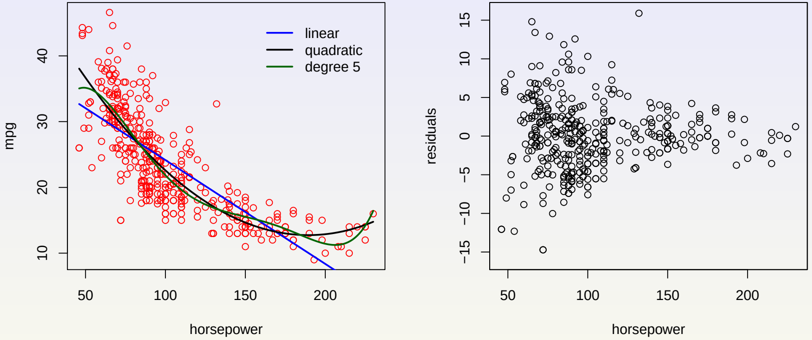
Extensions: Nonlinearities - Basis expansion

- `Auto` data set: mpg vs. horsepower



- Polynomial regression
- Model: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- The model is still linear in $\boldsymbol{\beta}$: treat X^2 as a variable
- We will have a comprehensive lecture on basis expansions later

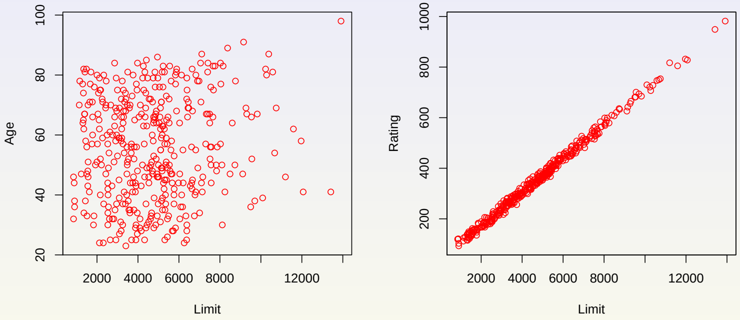
Example



- Increasing the degree can cause overfitting
- Alternative transformations

Colinearity

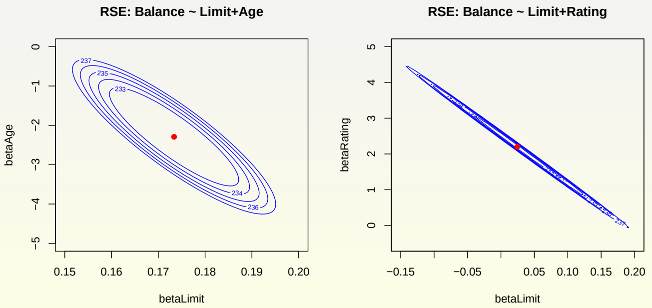
- `Credit` data set:



- Rating and Limit are co-linear
- Difficult to asses individual impact on Balance
- $\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X} \hat{\boldsymbol{\beta}}$ can be numerically unstable

Example

```
> lm8a<-lm(Balance ~ Limit + Age,data=credit)
> summary(lm8a)$coefficients
                Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) -173.410901 43.828387048 -3.956589  9.005366e-05
Limit          0.173365  0.005025662 34.495944 1.627198e-121
Age           -2.291486  0.672484540 -3.407492  7.226468e-04


> lm8b<-lm(Balance ~ Limit + Rating,data=credit)
> summary(lm8b)$coefficients
                  Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -377.53679536 45.25417619 -8.3425846 1.213565e-15
Limit          0.02451438  0.06383456  0.3840298 7.011619e-01
Rating         2.20167217  0.95229387  2.3119672 2.129053e-02
```



Prediction considerations

- Examine assumptions
- Uncertainties/Errors
  - Regression coefficients are noisy
  - Measurements are noisy even when the function in known
  - Bias: The true function might not be linear

K-NN regression: Non-parametric approach

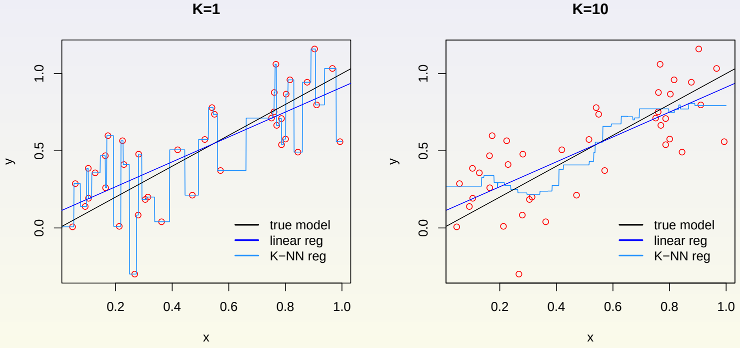
- Linear regression is not the only approach
- K-NN: K-nearest neighbors
- Intuition: "similar" argument values should lead to "similar" function values
  - distances
  - data normalization
  - high dimensionality case
- Let $\mathcal{N}_{\boldsymbol{x}}^K$ be the K-nearest neighbors set of observations for $\boldsymbol{x}$:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{K} \sum_{i \in \mathcal{N}_{\boldsymbol{x}}^K} y_i$$

- K is a parameter of the algorithm
  - Small K – flexible fit, high variance
  - Large K – smooth, high bias
  - How to select K?

Examples

- Linear model:



Reading:

ISL: Finish reading Chapter 3

ESL: Chapter 3

Homework: Homework 1 due in 2 weeks - Sep 27.