Topic VII:
Regression Models

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Fall, 2023

Introduction

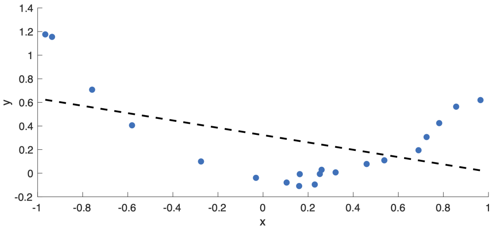
In simple linear regression, we analyze the linear relationship between the response Y and an predictor x.

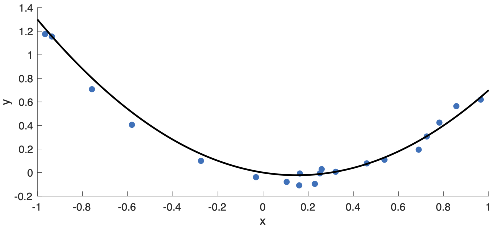
$$Y = \alpha + \beta x + \varepsilon.$$

- Estimation
- Inference

What if the relationship is not linear?

- $Y = \alpha + \beta x + \varepsilon$? No!
- $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.
- Conventional explanation: Fit a parabola to data.





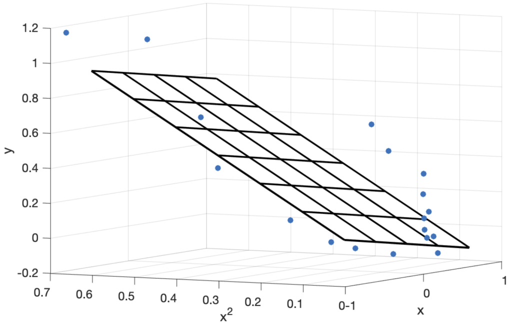
Enlarge the Feature Space

- Alternatively: Let

$$x_1 = x, \quad x_2 = x^2$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

- Bottomline: Although Y is quadratic in x, it is linear in the unknown parameters.



More generally, we use non-linear function of the inputs $\boldsymbol{\phi}(\boldsymbol{x})$ and assume

$$Y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon.$$

If $\boldsymbol{\phi}(\boldsymbol{x}) = (1, x, x^2, x^3, \ldots, x^d)$ is the vector of polynomial basis functions, it is called a **polynomial regression**.

Multiple Regression

Multiple linear regression model

For $i = 1, \ldots, n$

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

Or in vector form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} \in \mathbb{R}^{n \times 1}$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$.

- $\boldsymbol{X}$ is called the design matrix, $\boldsymbol{Y}$ is the response;
- the j-th column of $\boldsymbol{X}$, $\mathrm{x}_j = x_{\cdot j} \in \mathbb{R}^n$ is the vector of the j-th predictor;
- $\boldsymbol{\beta}$ is the unknown parameter; ε_i is the error of observation i.
- We usually set $x_{i1} = 1$ to have a non-zero intercept.

Goal: find the parameter values that fits the data the best.

Examples

$\boldsymbol{X}$ is usually called the design matrix.

$$\mathcal{N}(\mu_1,\sigma^2)\begin{cases} Y_1 \\ \vdots \\ Y_n \end{cases} \qquad \mathcal{N}(\mu_2,\sigma^2)\begin{cases} Y_{n+1} \\ \vdots \\ Y_{n+m} \end{cases} \qquad X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

- The location model: $p = 1$, $\boldsymbol{X} = (1, \cdots, 1)^\top$, $\beta_1 = \mu$.

$$Y = \mu + \varepsilon.$$

- The 2-sample model: $p = 2$, $\boldsymbol{X} = \begin{pmatrix} 1 & \cdots & 1 & 0 & \dots & 0 \\ 0 & \cdots & 0 & 1 & \dots & 1 \end{pmatrix}^\top$, $\beta = (\mu_1, \mu_2)^\top$.
- One-way (simple) analysis of variance (ANOVA), compare the mean of k groups: $p = k$, $\boldsymbol{X} =?$
- Simple linear regression: $Y = \alpha + \beta x + \varepsilon$.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = (\alpha, \beta)^\top.$$

Examples

- Higher order regression: $Y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \beta = (\alpha, \beta_1, \beta_2)^\top.$$

  This seems very powerful combined with the following fact: any continuous function can be approximated arbitrarily well by polynomials. (Weierstrass Theorem)

- Multiplicative model: $Y = \alpha x^\beta \exp(\varepsilon)$. If we take logarithm of both sides, we have $\log(Y) = \log(\alpha) + \beta \log(x) + \varepsilon$.

$$\boldsymbol{X} = \begin{pmatrix} 1 & \log(x_1) \\ \vdots & \vdots \\ 1 & \log(x_n) \end{pmatrix}, \quad \beta = (\log(\alpha), \beta)^\top.$$

Method of Least Squares

Consider the following linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- How to find a good estimator $\boldsymbol{\beta}$?
- For a given $\widehat{\boldsymbol{\beta}}$, the fitted values are $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.
- $\widehat{\boldsymbol{\varepsilon}}(\widehat{\boldsymbol{\beta}}) \triangleq \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is called the vector of the residuals.
- To estimate $\boldsymbol{\beta}$, minimize the **residual sum of squares (RSS)**:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

RSS is also called sum of squared error (SSE) or sum of squared residuals (SSR).

Least Squares Estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

- Take partial derivative of $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and set them to zero:

$$-2\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}$$

Let $f(x) = (f_1(x), \ldots, f_m(x))^\top \in \mathbb{R}^m$ for $x \in \mathbb{R}^n$, define $\frac{\partial f}{\partial x} \triangleq [\frac{\partial f_i}{\partial x_j}] \in \mathbb{R}^{m\times n}$.

Matrix differentiation

Let $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^{n\times 1}$. Let $f(x) = Ax \in \mathbb{R}^m, g(x) = x^\top Bx \in \mathbb{R}$, then

$$\frac{\partial f}{\partial x} = A \in \mathbb{R}^{m\times n}, \quad \frac{\partial g}{\partial x} = x^\top(B + B^\top) \in \mathbb{R}^{1\times n}.$$

Normal equation

$$\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}.$$

- We assumed that $n \geq p$ and $\boldsymbol{X}$ has full (column) rank, so that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.
- Given $n \geq p$ and $\boldsymbol{X}$ is full-rank, we have the **least square estimator (LSE)**

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y},$$

- The **fitted value** of $\boldsymbol{Y}$ is

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}.$$

Collinearity

Collinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this case, $\boldsymbol{X}^\top \boldsymbol{X}$ has extremely small eigenvalues, i.e., it is nearly non-invertible, and the estimation of parameter is sensitive to small changes in the model and data.

Least Squares Estimator – Computation Considerations

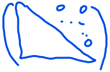
$$\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}.$$

The normal equation is usually solved via the Cholesky decomposition of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ or a QR decomposition of X.

- Cholesky decomposition

  $$\boldsymbol{X}^\top \boldsymbol{X} = LL^\top \Rightarrow LL^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y},$$

  where L is a lower triangular matrix.

- To solve the least square, first let $\boldsymbol{z} = L^\top \widehat{\boldsymbol{\beta}}$ and solve $L\boldsymbol{z} = \boldsymbol{X}^\top \boldsymbol{Y}$ for $\boldsymbol{z}$ using forward substitution, and then solve $L^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{z}$ for $\boldsymbol{\beta}$ using backward substitution.

Least Squares Estimator – Computation Considerations

$$\boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}.$$

- QR decomposition
$$\boldsymbol{X} = QR,$$
where Q is an orthogonal matrix ($QQ^\top = Q^\top Q = I$) and R is an upper triangular matrix.
- $X = QR \Rightarrow \boldsymbol{X}^\top \boldsymbol{X} = R^\top R$, QR for X is equivalent to Cholesky for $\boldsymbol{X}^\top \boldsymbol{X}$.
- But with n observations and p features, the Cholesky decomposition requires $p^3 + np^2/2$ operations, while the QR decomposition requires np^2 operations.

Depending on the relative size of n and p, the Cholesky can sometimes be faster; on the other hand, it can be less numerically stable (Lawson and Hansen, 1974).

Simple Linear Regression

Example: Simple linear regression. $Y = \alpha + \beta x + \varepsilon$.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

Least Square Estimators

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \quad \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{x}, \quad RSS = \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Geometric Interpretation

Column space of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$

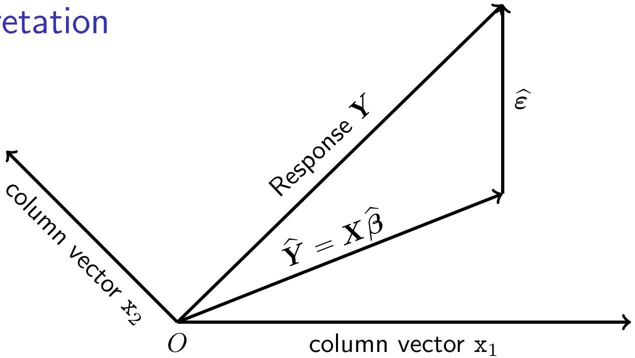
A dimension p subspace of $\mathbb{R}^n$, spanned by the column vectors of $\boldsymbol{X}$.

$$\boldsymbol{X}\boldsymbol{\beta} = \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \cdots + \beta_p \mathrm{x}_p, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Here $\mathrm{x}_j \in \mathbb{R}^n$ is the j-th column of $\boldsymbol{X}$.

- The fitted value
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$$
belongs to the column space.
- For the LSE $\widehat{\boldsymbol{\beta}}$, we know by normal equation that
$$\widehat{\boldsymbol{\varepsilon}}^\top \boldsymbol{X} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \boldsymbol{X} = 0.$$
It implies that $\widehat{\boldsymbol{\varepsilon}}$ is orthogonal to the $\boldsymbol{X}$-column space.

Geometric Interpretation



- $\widehat{\boldsymbol{Y}} = P\boldsymbol{Y}$ is the projection of $\boldsymbol{Y}$ on the X-column space, where
$$P \triangleq \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$$
is the projection matrix.
- $\widehat{\boldsymbol{\varepsilon}} = Q\boldsymbol{Y}$ where $Q = I - P$, which is orthogonal to $\widehat{\boldsymbol{Y}} = P\boldsymbol{Y}$, i.e., $\widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{\varepsilon}} = 0$.
Note that $PQ = QP = 0$.

Probabilistic Model

Now suppose we have the following assumptions

- Exogeneity: $\mathbb{E}[\varepsilon_i] = 0$.
- Homoscedasticity: $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$.
- ε_i's are independent normal random variables.

The likelihood function of y under parameter $\boldsymbol{\beta}, \sigma^2$:

$$L(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{\sum_{i=1}^n (Y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2}{2\sigma^2}\right) = \frac{1}{\sigma^n} \exp\left(-\frac{\|\boldsymbol{Y} - X\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right).$$

The MLE of $\boldsymbol{\beta}$ coincides with LSE (does not depend on σ^2).

Best Unbiased Linear Estimator (BLUE)

Assuming exogeneity and Homoscedasticity (but not normality) and under the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we have the following:

Theorem (Gauss-Markov)

1. *(Unbiasedness)* $\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\right] = \mathbb{E}\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right] = \boldsymbol{\beta}$.
2. *(Conditional standard error)* $\mathrm{Cov}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.
3. *The least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE).*

BLUE

For any given vector $\boldsymbol{a}$, $\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$ is a linear[a] unbiased estimator of the parameter $\theta = \boldsymbol{a}^\top\boldsymbol{\beta}$. Further, for any linear unbiased estimator $\boldsymbol{b}^\top\boldsymbol{Y}$ of θ, its variance is at least as large as that of $\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$.

[a] An estimator in the form of $\hat{\beta} = AY + \mu$ for some matrix A and vector μ.

Proof of BLUE. Obviously, $\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}}$ is unbiased and linear. The variance is

$$\mathrm{Var}(\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{a}.$$

Consider any linear unbiased estimator $\boldsymbol{b}^\top \boldsymbol{Y}$ of $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$. Unbiasedness requires that

$$\boldsymbol{b}^\top \boldsymbol{X} \boldsymbol{\beta} = \mathbb{E}[\boldsymbol{b}^\top \boldsymbol{Y}] = \boldsymbol{a}^\top \boldsymbol{\beta}, \quad \forall \boldsymbol{\beta}.$$

Hence $\boldsymbol{b}^\top \boldsymbol{X} = \boldsymbol{a}^\top$. Furthermore, the variance is

$$\mathrm{Var}(\boldsymbol{b}^\top \boldsymbol{Y}) = \sigma^2 \boldsymbol{b}^\top \boldsymbol{b}.$$

Plugging in $\boldsymbol{a} = \boldsymbol{X}^\top \boldsymbol{b}$, we have

$$\mathrm{Var}(\boldsymbol{b}^\top \boldsymbol{Y}) - \mathrm{Var}(\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{b}^\top Q \boldsymbol{b} \geq 0.$$

Properties of LSE

1. $E[\widehat{\boldsymbol{Y}}] = \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$.
2. $\mathrm{Cov}[\widehat{\boldsymbol{Y}}] = \mathrm{Cov}[\boldsymbol{X}\boldsymbol{\beta} + P\boldsymbol{\varepsilon}] = \sigma^2 P$.
3. $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}] = \mathrm{Cov}[Q\boldsymbol{Y}] = \mathrm{Cov}[Q\boldsymbol{\varepsilon}] = \sigma^2 Q$.
4. $\mathrm{Cov}[\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{\varepsilon}}] = 0$ as $QP = 0$.

The following is an unbiased estimator of σ^2

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-p} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p}.$$

Proof. Note that

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y}^\top Q^\top Q \boldsymbol{Y} = \boldsymbol{\varepsilon}^\top Q^\top Q \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}.$$

Use the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we have

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \mathrm{tr}(\varepsilon^\top Q \boldsymbol{\varepsilon}) = \mathrm{tr}(Q \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top)$$

Finally,

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right] = \mathbb{E}\left[\mathrm{tr}(Q \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top)\right] = \mathrm{tr}\left(Q \mathbb{E}[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top]\right) = \sigma^2 \mathrm{tr}(Q).$$

To calculate $\mathrm{tr}(Q)$, recall that trace is equal to the summation of the eigenvalues, and note that eigenvalues of Q is either 0 or 1 and the rank of Q is $n - p$ (if we assume $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible).

Generalized Least Square

What if ε is not i.i.d., but $\sim N(\mathbf{0}, \sigma^2\Sigma)$

- Σ is always positive semi-definite.
- We assume that Σ is **positive definite**, so that $\Sigma = AA^\top$, where A is invertible. (e.g. Cholesky decomposition)
- We assume that Σ is known.
- We can reduce to the standard model

$$\tilde{\boldsymbol{Y}} \triangleq A^{-1}Y = A^{-1}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = A^{-1}X\boldsymbol{\beta} + A^{-1}\boldsymbol{\varepsilon}$$

- New independent variables $\tilde{X} = A^{-1}X$, new i.i.d. error $\tilde{\boldsymbol{\varepsilon}} = A^{-1}\boldsymbol{\varepsilon}$.
- Check:

$$\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}] = \mathbb{E}[A^{-1}\boldsymbol{\varepsilon}] = A^{-1}\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$$
$$\mathrm{Cov}[\tilde{\boldsymbol{\varepsilon}}] = A^{-1}\mathrm{Cov}[\boldsymbol{\varepsilon}](A^{-1})^\top = A^{-1}\sigma^2(AA^\top)(A^{-1})^\top = \sigma^2 I$$

GLS Cont.

- Apply LSE,

$$
\begin{aligned}
&\min \|\tilde{\boldsymbol{Y}} - \tilde{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{Y} - X\boldsymbol{\beta})^\top \Sigma^{-1} (\boldsymbol{Y} - X\boldsymbol{\beta}) \\
\Rightarrow &\widehat{\boldsymbol{\beta}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \boldsymbol{Y}.
\end{aligned}
$$

- This is called the generalized least squares estimate.

$$
\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^\top \Sigma^{-1} X)^{-1}).
$$

- One can derive test statistics based on the distribution.

GLS Remarks

- A special case is when Σ is a diagonal matrix $diag(v_1, \ldots, v_n)$. In this case, we are minimizing the weighted squared error $\sum_i \varepsilon_i^2 / v_i$. This is called the **weighted least square**.
- In practice we usually don't know Σ. So we can run a standard LSE, estimate $\widehat{\Sigma}$ (need to impose some structures), and then applied GLS. This is referred to as the Cochrane-Orcutt procedure.
- If Σ is not positive definite (only semi-definite), one can use the Gram-Schmidt process (Schmidt orthonormalization) to find a set of standard normal random variables and the matrix that transform them to the original error terms.
  - Think about the geometric interpretation of covariance and use the fact that uncorrelated means independence for normal random variables.
  - Read the wikipedia page for Gram-Schmidt process.

Properties of LSE – Distributions

Now we impose the additional normal assumption, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_{n\times n})$.

- $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_{n\times n})$
- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$.
- $\widehat{\boldsymbol{Y}} = P\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + P\boldsymbol{\varepsilon} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 P), P = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^T$.
- $\widehat{\boldsymbol{\varepsilon}} = Q\boldsymbol{Y} = Q\boldsymbol{\varepsilon} \sim N(0, \sigma^2 Q), Q = I - P$. Note that $PQ = QP = 0$.
- $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ are independent: $\mathrm{Cov}(P\boldsymbol{Y}, Q\boldsymbol{Y}) = P\sigma^2 I_{n\times n} Q = 0$.
- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \widehat{\boldsymbol{Y}} \Rightarrow \widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2 = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p}$ are independent.
- Use the lemma for chi-squared distribution from Lecture 2,

$$(n-p)\widehat{\sigma}^2/\sigma^2 = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Statistical Inference for β_i

To test whether a particular variable x_i has no effect on $\boldsymbol{Y}$:

- Recall that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2 = \frac{\|\widehat{\varepsilon}\|_2^2}{n-p} = RSS/(n-p)$ are independent.
- For each β_i,

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{ii}}} = \frac{(\widehat{\beta}_i - \beta_i)/\sqrt{\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{ii}}}{\sqrt{[(n-p)\widehat{\sigma}^2/\sigma^2]/(n-p)}} \sim t_{n-p}.$$

Hypothesis Test – β_i

H_0	H_1	TS	Significance level α	p-value
$\beta_i = 0$	$\beta_i \neq 0$	$\frac{\widehat{\beta}_i}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{ii}}}$	Reject if $\lvert TS\rvert > t_{n-p}(1-\alpha/2)$	$2\mathbb{P}\{T_{n-p} > TS\}$

- The $1-\alpha$ confidence interval for β_i is

$$\left[\widehat{\beta}_i - t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{ii}}, \widehat{\beta}_i + t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{ii}}\right]$$

Simple Linear Regression

Example: Simple linear regression. $y = \alpha + \beta x + \varepsilon$.

- TS for β, $H_0 : \beta = \beta_0$, implies that $\sqrt{\frac{S_{xx}}{\widehat{\sigma}^2}}(\widehat{\beta} - \beta_0) \sim t_{n-2}$ under H_0.
- $(1-\gamma)$ CI for β:

$$\left(\widehat{\beta} - t_{n-2}(1-\gamma/2)\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}, \widehat{\beta} + t_{n-2}(1-\gamma/2)\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}\right)$$

- TS for α, $H_0 : \alpha = \alpha_0$, implies that $\sqrt{\frac{S_{xx} n}{\widehat{\sigma}^2 \sum_i x_i^2}}(\widehat{\alpha} - \alpha_0) \sim t_{n-2}$ under H_0.
- $(1-\gamma)$ CI for α:

$$\left(\widehat{\alpha} - \sqrt{\frac{\widehat{\sigma}^2 \sum_i x_i^2}{S_{xx} n}} t_{n-2}(1-\gamma/2), \widehat{\alpha} + \sqrt{\frac{\widehat{\sigma}^2 \sum_i x_i^2}{S_{xx} n}} t_{n-2}(1-\gamma/2)\right)$$

Statistical Inference for $\boldsymbol{\beta}$

To test whether the parameter is $\boldsymbol{\beta}_0$, i.e., $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{p\widehat{\sigma}^2} = \frac{\frac{\boldsymbol{\varepsilon}^\top P \boldsymbol{\varepsilon}}{\sigma^2}/p}{\frac{\boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}}{\sigma^2}/(n-p)} \sim F_{p,n-p}, \text{ under } H_0.$$

(Check independence!) This is because the numerator equals $\boldsymbol{\varepsilon}^\top P \boldsymbol{\varepsilon}$ and $tr(P) = p$ and $\widehat{\sigma}^2 = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p} = \boldsymbol{Y}^\top Q \boldsymbol{Y}/(n-p) = \boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}/(n-p)$ and $tr(Q) = n - p$.

- Reject if test statistic $\geq F_{p,n-p}(1-\alpha)$.

$(1-\alpha)$ confidence set

$$\left\{\boldsymbol{\beta} : (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq p\widehat{\sigma}^2 F_{p,n-p}(1-\alpha)\right\}$$

An ellipsoid.

Statistical Inference for the True Location of the Hyperplane

- For each experimental condition $\boldsymbol{x}_0$: Let $Y_0 = \boldsymbol{x}_0^\top \boldsymbol{\beta}$ and $\widehat{Y}_0 = \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}$

$$\frac{\widehat{Y}_0 - \mathbb{E}[Y_0]}{\widehat{\sigma}\sqrt{\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \sim t_{n-p}.$$

$\Rightarrow$ Confidence interval such that the true hyperplane cross at this $\boldsymbol{x}_0$.
Pointwise confidence band (CB)

$$\mathbb{P}(\boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} - w(\boldsymbol{x}_0) \leq \boldsymbol{x}_0^\top \boldsymbol{\beta} \leq \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} + w(\boldsymbol{x}_0)) = 1 - \alpha.$$

- Can we construct a confidence band for the entire true hyperplane ($Y = \boldsymbol{x}^\top \boldsymbol{\beta}$)?
Simultaneous CB

$$\mathbb{P}(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - w(\boldsymbol{x}) \leq \boldsymbol{x}^\top \boldsymbol{\beta} \leq \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} + w(\boldsymbol{x}), \text{ for all } \boldsymbol{x}) = 1 - \alpha.$$

Statistical Inference for the True Location of the Hyperplane

For simultaneous band, by Cauchy-Schwartz for $\langle \alpha, \beta \rangle = \alpha^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \beta$

$$
\begin{aligned}
|\boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^\top \boldsymbol{\beta}|^2 &= |\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})|^2 = \langle \boldsymbol{x}_0, (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rangle^2 \\
&\leq (\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0) \left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right)
\end{aligned}
$$

From previous slides, we know that

$$
\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{p \widehat{\sigma}^2} \sim F_{p, n-p}
$$

Confidence band

$$
|\boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^\top \boldsymbol{\beta}|^2 \leq (\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0) p \widehat{\sigma}^2 F_{p, n-p}(1 - \alpha).
$$

Simple Linear Regression

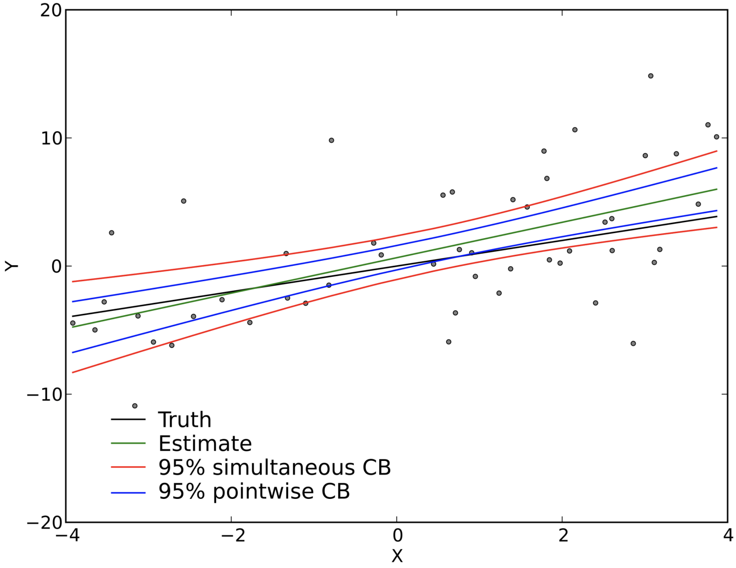
Example: Simple linear regression

Confidence interval for the new observation at x_0

$$\widehat{\alpha} + \widehat{\beta}x_0 \pm t_{n-2}(1-\gamma/2) \cdot \sqrt{\frac{RSS}{(n-2)}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Simultaneous confidence band for the regression line x_0

$$\widehat{\alpha} + \widehat{\beta}x_0 \pm \sqrt{2F_{n-2}(1-\gamma/2)} \cdot \sqrt{\frac{RSS}{(n-2)}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$


20
10
0
−10
−20
Y
−4
−2
0
2
4
X
Truth
Estimate
95% simultaneous CB
95% pointwise CB

Prediction Interval

In the previous slides, we see confidence band for mean response $\mathbb{E}[Y] = \boldsymbol{x}^\top \boldsymbol{\beta}$.

What about the response $Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$?

For each experimental condition $\boldsymbol{x}_0$:

$$\frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \sim t_{n-p}.$$

Example: Simple linear regression

(For the mean) $$\left(\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2}(1-\alpha/2) \sqrt{\frac{RSS}{(n-2)} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}\right)$$

(For the response) $$\left(\widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2}(1-\alpha/2) \sqrt{\frac{RSS}{(n-2)} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}\right)$$

Nested Model Test – F-statistic

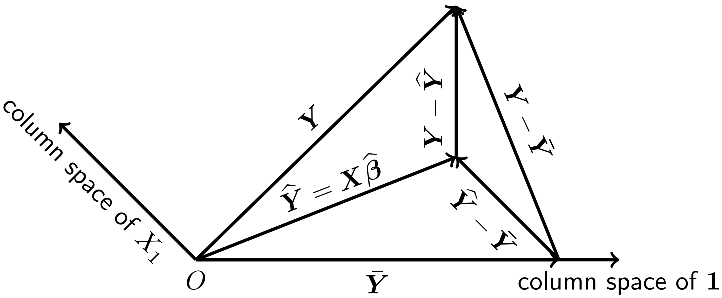
The F statistic checks whether a non-trivial linear model is not rejected.

- $X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} = (\mathbf{1}, X_1)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$.
- Consider null hypothesis $H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0$, the MLE under the null hypothesis is $\bar{\boldsymbol{Y}}$.

Nested Model Test – F-statistic

The F-statistic

$$F = \frac{\|\bar{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}\|_2^2/(p-1)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2/(n-p)} \sim F_{p-1,n-p}$$



This can be considered as testing two nested models. $H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0$ is the simpler model, and if we allow any $\boldsymbol{\beta}$, this is the full model.

Nested Model

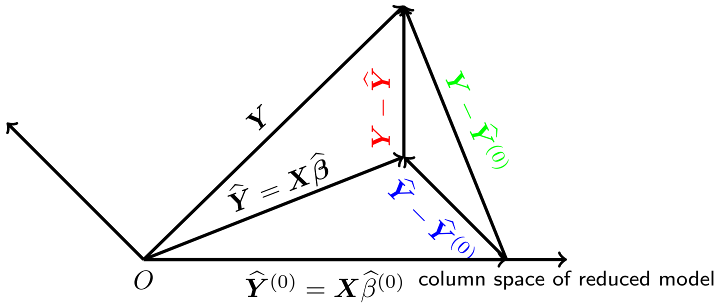
- H_1: all features play a role $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- Null hypothesis H_0: a reduced model where some parameters are redundant $B\boldsymbol{\beta} = \boldsymbol{b}$.
- For example, $\beta_2 = \beta_3 = \cdots = \beta_p = 0$. Or $\beta_1 + \beta_2 = 3$, which means we can reduce $\beta_2 = 3 - \beta_1$ and need only $p - 1$ features.
- For example, B is $(p-q) \times p$, $\boldsymbol{b} = \boldsymbol{0}$. Test if $p - q$ of the coefficients are zero.

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix}$$

- Can we represent the relationship by a simpler model?

Nested Model Test

- If the null hypothesis H_0 is true, say $\beta_2 = 0$. Then the column space of $X^{(0)}$ is a subspace of X.
- Geometric interpretation and Pythagoras theorem.



- Define $RSS = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2$ under H_1 and $RSS_0 = \|\mathbf{Y} - \widehat{\mathbf{Y}}^{(0)}\|_2^2$ under H_0.

Nested Model Test

- Recall the properties of LSE: $RSS/(n-p)$ is an unbiased estimator of σ^2, under both hypotheses. $(RSS_0 - RSS)/(p-q)$ is an unbiased estimator of σ^2 under H_0. More importantly, they are orthogonal!
- Pythagoras theorem: $\|\boldsymbol{Y} - \widehat{\mathbf{Y}}^{(0)}\|_2^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 + \|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(0)}\|_2^2$.
- This implies a statistic under H_0:

$$\frac{(RSS_0 - RSS)/(p-q)}{RSS/(n-p)} = \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(0)}\|_2^2/(p-q)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2/(n-p)} \sim F_{p-q,n-p}.$$

- Reject the simpler model if the test statitic is large.

Nested Model Test – ANOVA

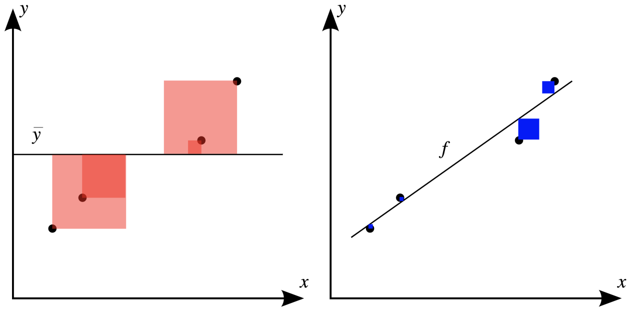
Analysis of variance (ANOVA) is a special case of the nested model test.

- $X = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix}^{\top}, \beta = (\mu_1, \mu_2, \ldots, \mu_k)^{\top}.$
- MLE is $(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot})$.
- Consider null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, the MLE under the null hypothesis is $\widehat{\boldsymbol{Y}}^{(0)} = \bar{\boldsymbol{Y}}$.
- $\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 + \|\widehat{\boldsymbol{Y}}^{(0)} - \widehat{\boldsymbol{Y}}\|_2^2 \Rightarrow SST = SSB + SSW.$

Coefficient of Determination

$$R^2 \triangleq \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \quad TSS = \|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\|_2^2.$$

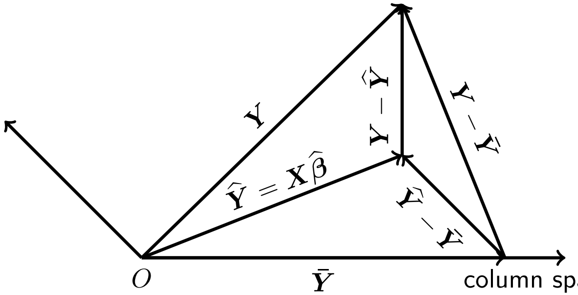
- Total Sum of Squares is the error *not captured by the sample mean*.
- $RSS = \sum(Y_i - \widehat{Y}_i)^2$ reflects the variance *not captured by the regression model*.



When $R^2 \approx 1$, it means $RSS \approx 0$, the fit is perfect!

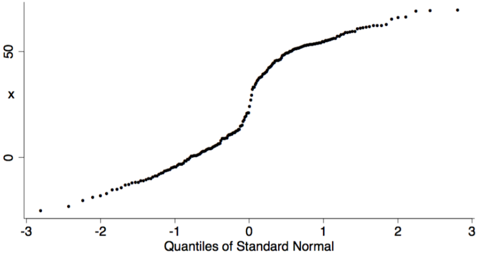
When $R^2 \approx 0$, it means the regression model is not much better than $\bar{y}$.

- One can check that
$$R^2 = \frac{\|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2/(p-1)}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2/(n-p)}.$$
- Also a F distribution? NO!



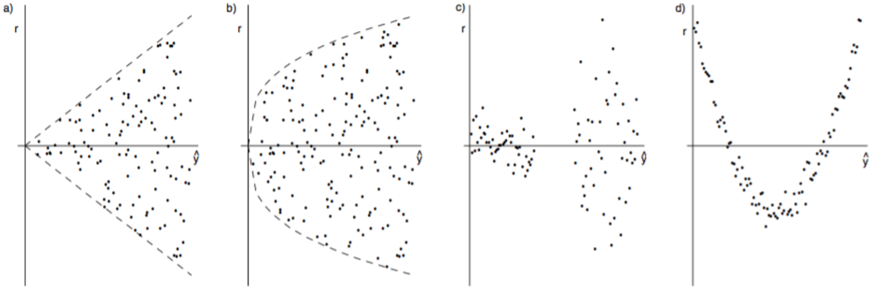
Normality

- Model assumption: ϵ_i has normal distribution.
- Normal probability plot: check whether a random sample $x_1, \ldots, x_n$ have normal distribution. Plot $\Phi^{-1}((i-0.5)/n)$ against $(x_{(i)} - \mu)/\sigma$ on the plane.
- If X is indeed normal, the plot should be roughly a straight line.
- Check the residuals: $\widehat{\epsilon_i} = Y_i - \widehat{Y}_i$ should be approximately normal when n is large.



Homoscedasticity

- Model assumption: ϵ_i are i.i.d. random variables.
- Plot $\widehat{Y}_i$ against $\widehat{\epsilon}_i$. What are the problems of the following figures?



Transformation

In practice we usually take a transformation of the response variable before fitting a regression model.

Example: $Y \propto x_1^{\beta_1} x_2^{\beta_2} \cdots x_p^{\beta_p}$, in which case we take the logarithm

$$\log(Y) = \sum_{i=1}^{p} \beta_i \log(x_i) + \varepsilon.$$

Box and Cox (1964) proposed a systematic way to find a transformation from the data.

To some extend, Box-Cox transformation can be used to address cases where normality or homoscedasticity assumptions are violated.

Box-Cox Transformation

Consider a parametric family of transformation functions. Let $Y^{(\lambda)}$ denote the transformed response, where λ is a parameter.

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(Y), & \text{if } \lambda = 0. \end{cases}$$

Box-Cox model

$$Y^{(\lambda)} = \boldsymbol{X}^{\top}\boldsymbol{\beta} + \varepsilon, \quad \text{where } \varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

The likelihood function of the Box-Cox model is

$$L(\lambda, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{Y}^{(\lambda)} - \boldsymbol{X}\boldsymbol{\beta}\|^2} \cdot J(\lambda, \boldsymbol{Y})$$

where the Jacobian is $J(\lambda, \boldsymbol{Y}) = \prod_{i=1}^n |d\boldsymbol{Y}^{(\lambda)}/d\boldsymbol{Y}| = \prod_{i=1}^n |Y_i|^{\lambda-1}$.

MLE for Box-Cox model

- Given a λ, the MLE for $\boldsymbol{\beta}, \sigma^2$ are the usual

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}^{(\lambda)}, \quad \widehat{\sigma}^2(\lambda) = \frac{1}{n}\|\boldsymbol{Y}^{(\lambda)} - \widehat{\boldsymbol{Y}}^{(\lambda)}\|^2.$$

- Plugging $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\sigma}^2(\lambda)$ into the likelihood, we have

$$\log L(\lambda) = (\lambda - 1)\sum_{i=1}^n \log(|Y_i|) - \frac{n}{2}\log \widehat{\sigma}^2(\lambda) - \frac{n}{2} \Rightarrow \widehat{\lambda}_{\text{MLE}} = \arg\max_\lambda \log L(\lambda).$$

- The MLE is $(\widehat{\lambda}_{\text{MLE}}, \widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\text{MLE}}), \widehat{\sigma}^2(\widehat{\lambda}_{\text{MLE}}))$.

Nonlinear regression

Recall our motivating example – polynomial regression

$$Y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon,$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = (1, x, x^2, x^3, \ldots, x^d)$.

Weierstrass Theorem

Any continuous $f(X)$ on $[0, 1]$ can be uniformly approximated by a polynomial function.

Spline regression

- One problem of polynomial regression: not suitable for functions with varying degrees of smoothness.
- Solution: use piece-wise polynomial.

Spline

A degree-d spline is a piecewise polynomials with degree d, which is continuous differentiable up to order $d-1$. The points where discontinuity (in the derivatives) occurs are called the <u>knots</u>.

Example: Linear spline with knots $\tau_1 < \tau_2$.

$$f(x) = \begin{cases} \beta_0 + \beta_1 x, & x \in (-\infty, \tau_1] \\ \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)^+, & x \in (\tau_1, \tau_2] \\ \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)^+ + \beta_3 (x - \tau_2)^+, & x \in (\tau_2, \infty). \end{cases}$$

The basis functions

$$B_0(x) = 1, B_1(x) = x, B_2(x) = (x - \tau_1), B_3(x) = (x - \tau_2)^+.$$

Example: Cubic spline with knots $\tau_1 < \tau_2$.

$$f(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, & x \in (-\infty, \tau_1] \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4((x-\tau_1)^+)^3, & x \in (\tau_1, \tau_2] \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4((x-\tau_1)^+)^3 + \beta_5((x-\tau_2)^+)^3, & x \in (\tau_2, \infty). \end{cases}$$

The basis functions

$$B_0(x) = 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3, B_4(x) = ((x-\tau_1)^+)^3, B_5(x) = ((x-\tau_2)^+)^3.$$

- Try to write down cubic splines with k knots.

Spline regression

$$\boldsymbol{Y} = \boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{x}) + \varepsilon.$$

- A multiple linear regression model.
- Versatile and tractable: Cubic splines are widely used.
- Later on, we will look at a fully nonparametric multiple regression called <u>kernel ridge regression</u>.

Influence of Individual Observations on the LSE

Cook's distance for the ith observation

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} \sim F_{p,n-p}.$$

- Observations with large Cook's distance alters the LSE by a lot, and thus can be considered an outlier.
- It may not be appropriate to simply discard outlier.
- For data that deviates from normal distribution (check by Q-Q plot), especially for those distributions with heavy tails, outliers appear more often.
- So we need robust methods for regression.

Robust Regression

Recall that the multiple linear regression model can be obtained by minimizing the least square error, i.e.,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2.$$

- The loss function is quadratic $\Rightarrow$ loss for outliers have much more impact on the estimation.

Consider the alternative loss function

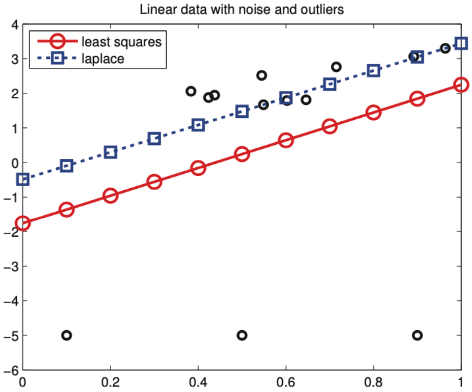
$$\widehat{\boldsymbol{\beta}}_{\mathrm{Robust}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |Y_i - \widehat{Y}_i| = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_1.$$

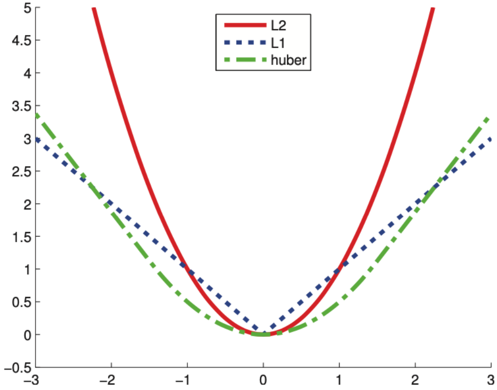
- This is called **robust regression**.

Robust Regression

Robust regression is equivalent to assuming that the response y follows a Laplace distribution.

$$p(Y|\boldsymbol{X}, \boldsymbol{\beta}, b) = \text{Laplace}(Y|\boldsymbol{X}, \boldsymbol{\beta}, b) \propto \exp(-|Y - \boldsymbol{x}^\top \boldsymbol{\beta}|/b).$$





Robust Regression

The loss function in robust regression is not differentiable at 0 and is nonlinear. Fortunately, it can be solved using linear program.

- Consider writing $\varepsilon = \varepsilon^+ - \varepsilon^-$, where $\varepsilon^+ = \max\{0, \varepsilon\}$ and $\varepsilon^- = \max\{0, -\varepsilon\}$.
- One can show that the robust regression estimation is equivalent to

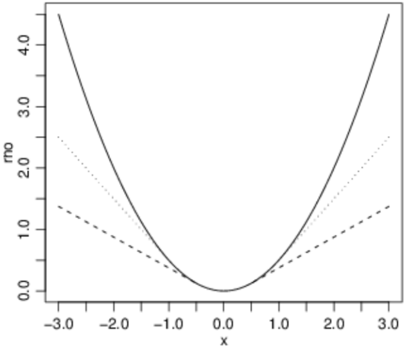
$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{Robust}} = \mathop{\arg\min}_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-} \quad & \sum_{i=1}^{n} (\varepsilon_i^+ - \varepsilon_i^-) \\ \text{suchthat} \quad & \varepsilon_i^+ \geq 0, \varepsilon_i^- \geq 0, \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i^+ + \varepsilon_i^- = Y_i. \end{aligned}$$

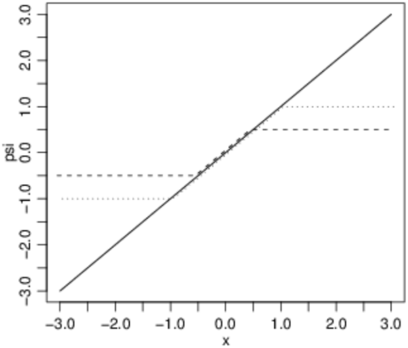
- The resulting linear program can be slow because the dimension is $p + 2n$.

Huber Regression

The loss function in robust regression is not differentiable at 0.

- A trade-off between ordinary least square and robust regression is the **Huber regression**.
- In Huber regression, the error ε is ε^2 for $\varepsilon \leq \delta$ and is $\delta|\varepsilon| - \delta^2/2$.





Robustness of the sample mean

Example: Let $X_1, \cdots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Sample mean can be obtained as the LSE for the location model $p = 1$, $X = (1, \cdots, 1)^\top$, $\beta_1 = \mu$.

$$Y = \mu + \varepsilon.$$

- We know $\bar{X}$ is a good estimator for μ under normal distribution and $\mathrm{Var}[\bar{X}_n] = \sigma^2/n$.
- If

$$X_i = \begin{cases} N(\mu, \sigma^2) & w.p.\ 1-\delta \\ f(x) & w.p.\ \delta \end{cases}$$

  and $f(x)$ has mean θ and variance τ^2.
- Then $\mathrm{Var}[\bar{X}] = (1-\delta)\sigma^2/n + \sigma\tau^2/n + \delta(1-\delta)(\theta-\mu)^2/n$.
- If $\theta \approx \mu$ and $\sigma \approx \tau$, then we are good.
- If f is Cauchy distribution, then $\mathrm{Var}[\bar{X}] = \infty$.

Median vs Mean

One can check that the median minimizes $\sum |x_i - a|$.

- Median is more robust than the mean. How about its performance?
- For $X \sim F$, we can show that $\sqrt{n}(X_{(n/2)} - \mu)$ is asymptotically normal with mean 0 and variance $1/(2f(\mu))^2$.
- The median is the obtained as the estimator of $\beta_1 = \mu$ in the robust regression for the location model $p = 1$, $X = (1, \cdots, 1)^\top$, $\beta_1 = \mu$. Robust regression is more robust than least square linear regression.
- There is a trade-off between robustness and the performance when the assumed model is correct. The median is not efficient compared to the mean.

Something in-between

Is there an estimator between the mean and the median?

- Huber estimator: minimize $\sum \rho(x_i - a)$ where

$$\rho(x) = \begin{cases} x^2/2 & |x| \le k \\ k|x| - k^2/2 & |x| \ge k \end{cases}$$

  Here ρ is differentiable.
- By tuning the parameter k, this estimator can be more "median-like" or "mean-like".
- We should expect this estimator to be asymptotically normal, because two ends (mean and median) are both so. This is indeed the case.

Reference

Jianqing Fan et al., "Statistical Foundation of Data Science,"
`https://orfe.princeton.edu/~jqfan/fan/classes/525/chapters1-3.pdf`

Kevin Murphy, "Machine Learning: A Probabilistic Prospective," MIT Press.

John Fox, "Applied Regression Analysis, Linear Models, and Related Methods," Sage Publications.

Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie, "The Elements of Statistical Learning," Springer.

Lecture notes by Prof. Kunsch and Meinshausen.