Advanced Statistics
Spring 2022

Linear Model I (Lecture 2)

Dr. Alon Kipnis

Material credit: Art Owen

Announcements

- Home Assignment 1 will be posted tonight. Due before class on March 22.
- Exploratory data analysis tutorial is available on course website
- Notes and code from the first lecture are available on course website
- Clarification concerning two-phase regression on Piazza

Recap – The Linear Model

We have data:

$$(x_i, y_i), \quad i = 1, \ldots, n$$

We propose a model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

or

$$\mathbb{E}\left[Y|X = x\right] = \beta_1 x_1 + \ldots + \beta_p x_p$$

Tasks we would like to perform:

- **Estimate** $\beta = (\beta_1, \ldots, \beta_p)$
- **Test**, e.g., whether $\beta_{105} = 0$ or not
- **Predict** y_{n+1} given x_{n+1}
- **Estimate** σ^2
- **Check** the model's assumptions
- **Make a choice** among linear models

Example: Predicting Home Prices

$$y_i = \sum_{j=1}^{p} \beta_j x_{ip} + \epsilon_i$$

$$\begin{aligned}
y_i &= \text{sale price of home } i \\
x_{i1} &= \text{constant} \\
x_{i2} &= \text{square meters of home } i \\
x_{i3} &= \text{\# of bedrooms of home } i \\
\vdots &= \vdots \\
x_{i,203} &= \text{\# of synagogues near home } i
\end{aligned}$$

Remarks:

- The model would **still be linear** even if we had that $x_{i,93} = \sqrt{\#\text{of bedrooms}}$
- **Sum** of linear models is also a **linear model**

Linear Model Notation

$$x_i \in \mathbb{R}^d, \qquad y_i \in \mathbb{R},$$
$$y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \epsilon_i,$$

where $z_{ij} = f_j(x_i)$ is a function of x_i (we call $f_j(x)$ the j-th feature of x)

Note that d (the dimension of x) does not necessarily equal p. Examples:

$$z_i = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{id} \end{pmatrix}^\top \in \mathbb{R}^{d+1}$$

or

$$z_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & x_{i1}^2 & x_{i2}^2 \end{pmatrix}^\top \in \mathbb{R}^5$$

- Names for $\{f_j(x_i)\}$: (j-th) **feature**, **predictor**, **covariate**, **independent variable**
- Names for $\{y_i\}$: **response**,**response variable**, **dependent variable**, **target**, **label**

Least Squares

Setting

- **We have data:**

$$\{(x_i, y_i)\}_{i=1}^n$$

- **We want:** to develop a model for a new response y_{n+1} given a new observation x_{n+1}
- **Our approach:**
  1. We transform each data point x_i to p features:

  $$z_{ij} = f_j(x_i), \quad z_{ij}, \qquad i = 1, \dots, n, \quad j = 1, \dots, p$$

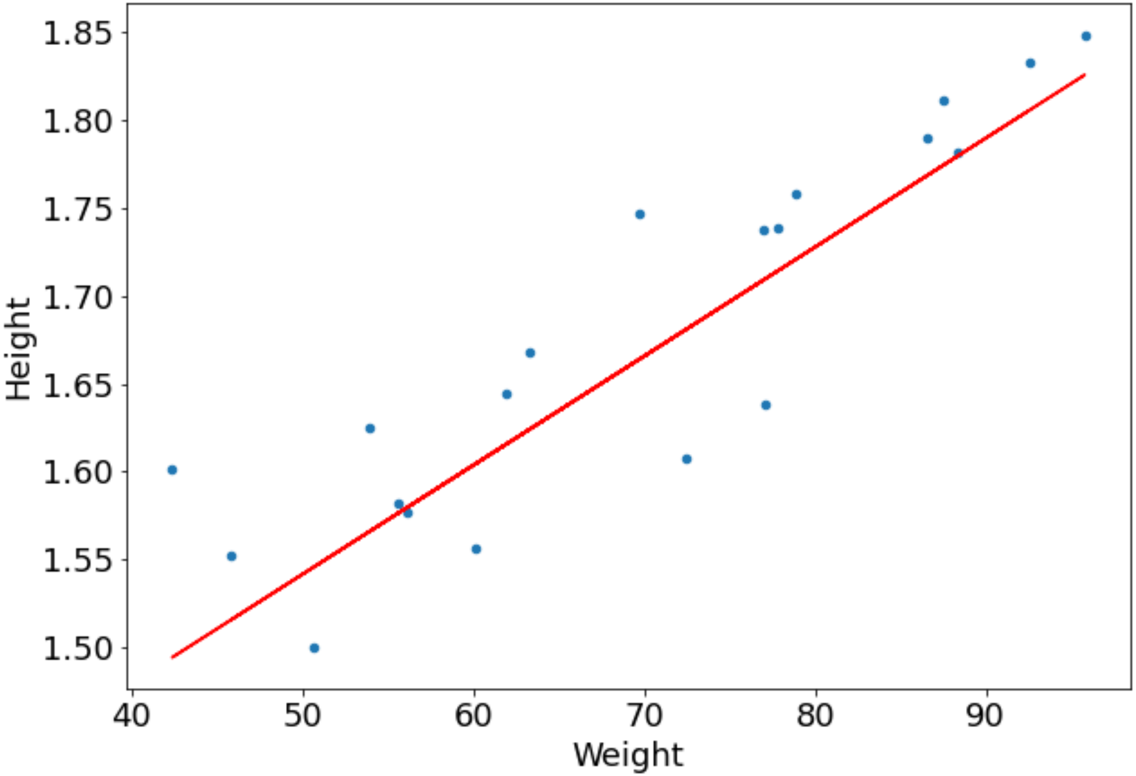
  2. We assume a linear response model:

  $$\hat{y}_{n+1} = \sum_{j=1}^{p} z_{n+1,j}\beta_j = \beta^\top z_{n+1}$$

  where $\beta = (\beta_1, \dots, \beta_p)$ is a function of $\{((z_{i1}, \dots, z_{ip}), y_i)\}_{i=1}^n$
  3. We choose the model parameters to minimize the squared error over the **given** data:

  $$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^\top z_i)^2$$

Depiction



Least Squares Notation

- Def. **Observed response variables**: $y_1, y_2, \dots, y_n$
- Def. **Features**: z_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$
- Def. **Regression coefficients**: $\beta := (\beta_1, \dots, \beta_p)$
- Def. **Squared error**:

$$S(\beta) := \sum_{i=1}^{n} \left(y_i - \beta^\top z_i \right)^2$$

- Def. **Least squares estimate**:

$$\hat{S} := \min_{\beta \in \mathbb{R}^p} S(\beta)$$

- Def. **Least squares regression coefficients**:

$$\hat{\beta} := (\hat{\beta}_1, \dots, \hat{\beta}_p) := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} S(\beta)$$

Computing least squares estimate & regression coefficients

- Using calculus:

$$\frac{\partial S}{\partial \beta_j} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n}(y_i - \beta^\top z_i)(-z_{ij}) = 0, \qquad j = 1, \dots, p$$

  (we also need to show that the solution is the minimum and not the maximum or a saddle point)

- Def. These p equations are known as the **Normal Equation** (bc. normal is a synonym to perpendicular)
- We have

$$(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^\top (z_{1j}, \dots, z_{n,j}) = 0, \qquad j = 1, \dots, p$$

it's the same var, notice j. in total you have p equations like this to solve

where

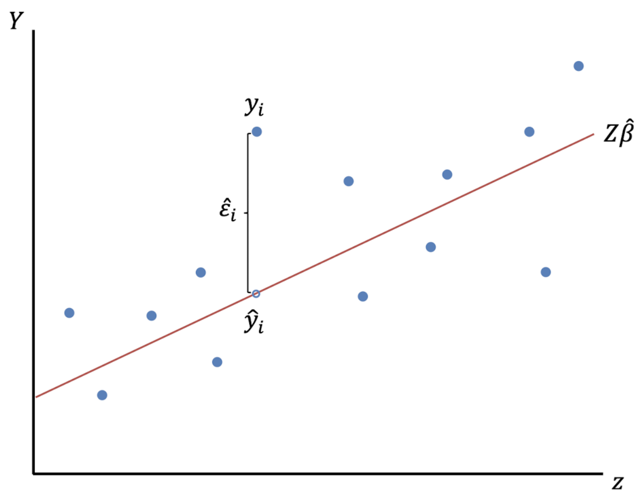
$$\hat{\epsilon}_i := y_i - \hat{\beta}^\top z_i, \qquad i = 1, \dots, n$$

are the **residuals**

Depiction of Residuals

$$\hat{y}_i = \sum_{j=1}^{p} z_{ij}\hat{\beta}_j, \qquad \hat{\epsilon}_i = y_i - \hat{y}_i, \qquad (\hat{\beta}_1, \ldots, \beta_p) = \text{argmin}\, S(\beta_1, \ldots, \beta_p)$$

With one predictor x and a constant term: $\hat{y}_i = \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x$



Matrix Notation

- Observed **response** and **features**:

$$y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \qquad Z := \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

  Z is also called the **design** or **data** matrix.
- Vector of **residuals**: $\hat{\epsilon} := y - Z\hat{\beta}$
- The Normal Equations (after dividing by -2):

$$\hat{\epsilon}^\top Z = 0 \quad \Leftrightarrow \quad Z^\top \hat{\epsilon} = 0 \quad \Leftrightarrow \quad Z^\top Z \hat{\beta} = Z^\top y$$

  - If $Z^\top Z$ is invertible, then $\hat{\beta} = (Z^\top Z)^{-1} Z^\top y$
  - The predicted value at a new point vector z_{n+1} is

$$\hat{y}_{n+1} = \hat{\beta}^\top z_{n+1} = \left((Z^\top Z)^{-1} Z^\top y \right)^\top z_{n+1} = y^\top Z (Z^\top Z)^{-1} z_{n+1}$$

    (**linear** both in the observed response vector y and the new point vector z_{n+1})

Uniqueness of Least Squares Solution

Theorem

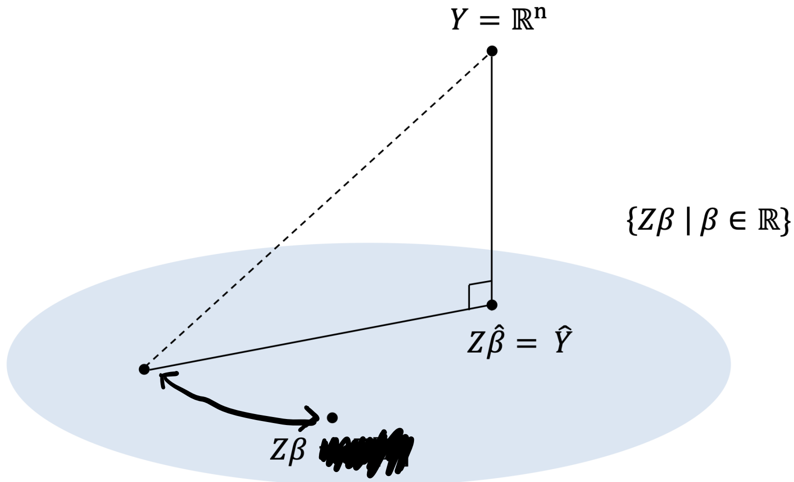
Let $Z \in \mathbb{R}^{n\times p}$ with $(Z^\top Z)^{-1}$ invertible, and let $y \in \mathbb{R}^n$. For $\beta \in \mathbb{R}^p$, define $S(\beta) = (y - Z\beta)^\top (y - Z\beta)$ and set $\hat{\beta} = (Z^\top Z)^{-1} Z^\top y$. Then $S(\beta) > S(\hat{\beta})$ for any $\beta \neq \hat{\beta}$.

Proof. We know that $Z^\top (y - \hat{\beta}^\top Z) = 0$. For arbitrary $\beta \in \mathbb{R}^p$, let $\gamma = \beta - \hat{\beta}$. Then

$$
\begin{aligned}
S(\beta) &= (y - Z\beta)^\top (y - Z\beta) \\
&= (y - Z\hat{\beta} - Z\gamma)^\top (y - Z\hat{\beta} - Z\gamma) \\
&= (y - Z\hat{\beta})^\top (y - Z\hat{\beta}) - \gamma^\top Z^\top (y - Z\hat{\beta}) - (y - Z\hat{\beta}) Z\gamma + \gamma^\top Z^\top Z\gamma \\
&= S(\hat{\beta}) + \gamma^\top Z^\top Z\gamma.
\end{aligned}
$$

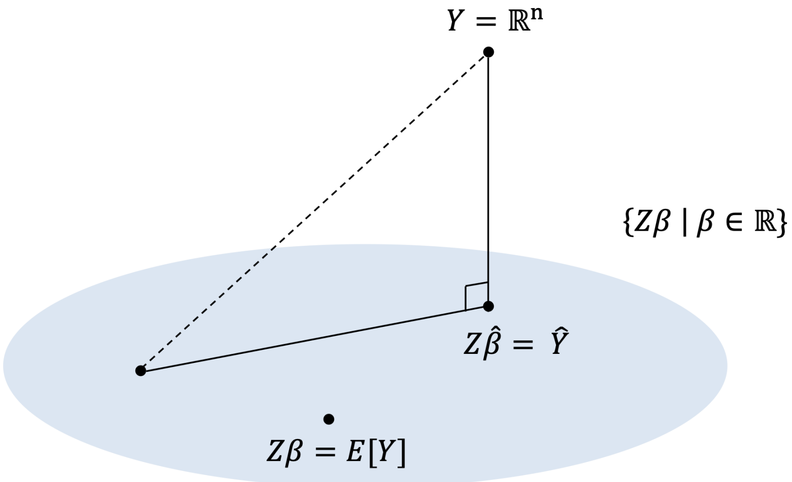
It follows that $S(\beta) = S(\hat{\beta}) + \|Z\gamma\|^2 \geq S(\hat{\beta})$, so that $\hat{\beta}$ is a minimizer of S. For uniqueness, we have $S(\hat{\beta}) = S(\beta)$ iff $Z\gamma = 0$. Since Z is invertible, this implies $\gamma = 0$ hence $\beta = \hat{\beta}$.

Geometry of Least Squares



- Consider the set $\mathcal{M} := \{Z\beta \mid \beta \in \mathbb{R}^p\} \subset \mathbb{R}^n$ (fully p dimensional because $Z^\top Z$ is invertible and so Z has rank p; convex)
- $Z\hat{\beta}$ is the closest point to Y from within $\mathcal{M}$
- From the normal equations $\hat{\epsilon}^\top Z = 0$, we get that $\hat{\epsilon} = y - Z\hat{\beta}$ is perpendicular to any line within $\mathcal{M}$

Geometry of Least Squares (cont'd)



- We can form a right angle triangle using $(y, \hat{y}, Z\beta)$ for any $\beta \in \mathbb{R}^p$, where $\hat{y} := Z\hat{\beta}$
  - For $\beta = 0$, we get: $\|y\|^2 = \|\hat{\epsilon}\|^2 + \|\hat{y}\|^2$ (take $\beta = 0$ in the proof of the theorem above, so that $S(0) = S(\hat{\beta}) + \|Z\hat{\beta}\|^2$)
  - In the next slide we will use $\beta = (\bar{y}, 0, \ldots, 0)$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

Sum-of-Squares Decomposition

- Suppose that the first feature is the all ones vector

$$Z_{i1} = (1, \dots, 1)\,, \quad i = 1, \dots, n$$

- We have

$$\underline{y} := (\bar{y}, \dots, \bar{y})^\top \in \mathcal{M}, \qquad \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i$$

From the right angle triangle $(y, \hat{y}, \underline{y})$

$$\|y - \underline{y}\|^2 = \|\hat{y} - \underline{y}\|^2 + \|y - \hat{y}\|^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- $SS_{Tot} := \sum_{i=1}^n (y_i - \bar{y})^2$ is the **Total (or centered) sum of squares**
- $SS_{Fit} := \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is the **Centered sum of squares of fitted values**
- $SS_{Res} := \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the **Residual sum of squares**

Sum-of-Squares Decomposition (cont'd)

- We write

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

as

$$SS_{Tot} = SS_{Fit} + SS_{Res}$$

- Def. **Coefficient of determination**:

$$R^2 := \frac{SS_{Fit}}{SS_{Tot}} = 1 - \frac{SS_{Res}}{SS_{Tot}}$$

  - Proportion of variation accounted for by all variables compared to the sum of squares error under the model $y_i = \beta_0 + \epsilon_i$
  - Measures how well Y is predicted or determined by $Z\hat{\beta}$:
  - $R := \sqrt{R^2}$ is called the **coefficient of multiple correlation** – it measures how well the response y correlates with the p predictors in Z taken collectively
  - When $z_i = (1, x_i) \in \mathbb{R}^2$, R is the Pearson correlation of $\{x_i\}$ and $\{y_i\}$
  - Equation (1) is an example of ANOVA decomposition

Examples

Algebra of Least Squares

Algebra of Least Squares

- The predicted value for y_i is $\hat{y}_i = Z_i \hat{\beta}$
- The vector of predicted values is

$$\hat{y} = Hy, \qquad H := Z(Z^\top Z)^{-1} Z^\top$$

  (Tukey called H the "hat" matrix)
- Properties of H:
  - Symmetric: $H = H^\top$
  - Idempotent: $H^2 = H$ (a symmetric idempotent matrix such as H is called a perpendicular projection matrix (PPM))
  - The eignevalues of a real PPM are all either 0 or 1
  - If Z is invertible, H has p non-zero eigenvalues
  - $I - H$ is PPM

Algebra of Least Squares (cont'd)

Theorem

Let A be PPM. The eigenvalues of A are all either 0 or 1.

Proof. If x is an eigenvector of H with eigenvalue λ, then $Hx = \lambda x$ and $x \neq 0$. Because H is PPM, $\lambda x = Hx = H^2 x = H(Hx) = H(\lambda x) = \lambda^2 x$, hence $\lambda^2 = \lambda$ which is satisfied iff $\lambda \in \{0, 1\}$.

Theorem

The rank of H is p

Proof. The eigenvalues of H sum to r, so
$r = \mathrm{Tr}(H) = \mathrm{Tr}(Z(Z^\top Z)^{-1} Z^\top) = \mathrm{Tr}(Z^\top Z(Z^\top Z)^{-1}) = \mathrm{Tr}(I_p) = p$

Algebra of Least Squares (cont'd)

Additional properties of $H = Z(Z^\top Z)^{-1}Z^\top$:

- $\hat{y}_i = H_i y$ (H_i is the i-th row of H)
- $H_{ij} = z_i^\top (Z^\top Z)^{-1} z_j = H_{ji}$ (the contribution of y_i to $\hat{y}_j$ equals that of y_j to $\hat{y}_i$)
- $H_{ii} = z_i^\top (Z^\top Z)^{-1} z_i \geq 0$ (Exc.)
- H projects vectors onto the **columns space** of Z
  $\mathrm{Col}(Z) := \mathcal{M} = \{Z\beta \,|\, \beta \in \mathbb{R}^p\}$
- $I - H$ projects vectors onto the **null space** of Z
  $\mathrm{Null}(Z) := \mathcal{M}^\top := \{v \in \mathbb{R}^n, \,|\, Zv = 0\}$ (the set of vectors orthogonal to vectors in $\mathcal{M}$)

The columns space and the null space are **orthogonal complements**: any $v \in \mathbb{R}^n$ can be uniquely written as $v_1 + v_2$, $v_1 \in \mathcal{M}$ and $v_2 \in \mathcal{M}^\top$. This is written as $\mathbb{R}^n = \mathcal{M} \oplus \mathcal{M}^\top$. In terms of the H matrix, $v_1 = Hv$ and $v_2 = (I - H)v$.

Distributional Results