Linear Regression Models
P8111

Lecture 12

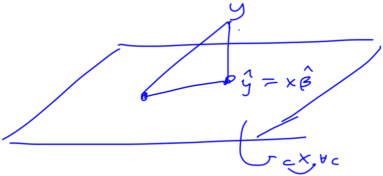
Jeff Goldsmith
March 1, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL OF PUBLIC HEALTH

Today's Lecture

- Gauss-Markov theorem
- Maximum likelihood inference
- Regression diagnostics

What's so great about LSEs?



$LSE = \underbrace{(X^T X)^{-1} X^T}_{A} y$

- Nice projection-space interpretation
- They're the "best" linear unbiased estimators

GM

- They're maximum likelihood estimators under Normally-distributed errors

$\epsilon \sim N(0, \sigma^2)$

Gauss-Markov theorem

Assume the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I$. Also assume $\boldsymbol{X}$ is a full rank design matrix.

- Among all unbiased linear estimators $\boldsymbol{C}\boldsymbol{y}$ of the regression coefficients $\boldsymbol{\beta}$, the LSE has minimum variance and is unique.

We call the LSEs "BLUE".

Gauss-Markov theorem – proof

unbiased: $Cy = \underbrace{((X^TX)^{-1}X^T + A)}\, y$

$E(Cy) = \beta$

$\Rightarrow E(((X^TX)^{-1}X^T + A)(X\beta + \epsilon))$

$\Rightarrow E(\underbrace{\cancel{(X^TX)^{-1}X^T}X\beta + AX\beta} + (X^TX)^{-1}X^T\epsilon + A\epsilon)$

$= \beta + AX\beta = \beta$

$\Rightarrow AX = 0$

Gauss-Markov theorem – proof

Cy

$\mathrm{Var}(Cy)$

$= C\underbrace{\mathrm{Var}(y)}_{\sigma^2 I}C^T$

$= \sigma^2 CC^T$

$= \sigma^2\left((X^TX)^{-1}X^T + A\right)\left((X^TX)^{-1}X^T + A\right)^T$

$= \sigma^2\Big[(X^TX)^{-1}\cancel{X^TX(X^TX)^{-1}} + \cancel{(X^TX)^{-1}\underbrace{X^TA^T}_{(AX)^T}}$

$+ \cancel{AX(X^TX)^{-1}} + AA^T\Big]$

$= \sigma^2\left[(X^TX)^{-1} + AA^T\right]$ $\quad\Big|\; x + a^2 \geq x$

$\geq \sigma^2(X^TX)^{-1}$

$=$ iff $A=0$

Gauss-Markov theorem – caveats

The Gauss-Markov theorem is great, but notice the details:

- Assumed $Var(\boldsymbol{\epsilon}) = \sigma^2 I$
- Only talking about unbiased linear estimators

Maximum likelihood estimation

Continue assuming the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I$.

$p(\text{data} \mid \text{param})$

$p(y \mid \beta)$

$\downarrow$

$L(\beta ; y)$

- Additionally, assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$
- Put differently, we're imposing the model

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 I)$$

- $\boldsymbol{y}$ is multivariate Normal with uncorrelated entries; the y_i are each independently Normally distributed

$y_i \sim N(x_i \beta, \sigma^2)$

Maximum likelihood estimation

Using independently Normal y_i's:

$$L(\beta; y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2\sigma^2}(y_i - x_i\beta)^2 \right\}$$

$$(2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum (y_i - x_i\beta)^2 \right\}$$

$$\ell(\beta; y) = \log(\quad) + -\frac{1}{2\sigma^2} \sum (y_i - x_i\beta)^2$$

$$\frac{\partial \ell}{\partial \beta} \propto \frac{\partial}{\partial \beta} \quad RSS$$

Maximum likelihood estimation

Using matrix notation:

$$L(\beta ; y) = \underbrace{(2\pi)^{-n/2}(|\sigma^2 I|)^{-1/2}} \exp\left\{-\frac{1}{2}(y - X\beta)^T(\sigma^2 I)^{-1}(y - X\beta)\right\}$$

$$-\frac{1}{2\sigma^2}\underbrace{(y - X\beta)^T(y - X\beta)}_{RSS(\beta)}$$

$$\underbrace{(X^T X)^{-1} X^T y}$$

Regression diagnostics

$$y = X\beta + \epsilon$$

$$\epsilon \sim (0, \sigma^2 I)$$

- Regression diagnostics are tools used to determine whether a given model is consistent with the data
- Usually focus on residuals
- Recall that fitted values are given by $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H}$ is the hat matrix

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}y$$

- Residuals are defined as $\boldsymbol{y} - \hat{\boldsymbol{y}} = (I - \boldsymbol{H})\boldsymbol{y}$

$$\hat{\epsilon} = (I - H)y$$

$\hat{\epsilon}$ and ϵ

- $E(\hat{\epsilon}) =$ 0

  $E((I-H)y)$

- $Var(\hat{\epsilon}) =$ $\sigma^2(I-H)$

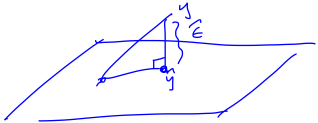
- Residuals are mean zero, but don't have constant variance nor are they uncorrelated.

$\hat{\epsilon}$ and ϵ

$\hat{y} = Hy$

- $\hat{\epsilon} = (I - H)y = \ldots = (I - \boldsymbol{H})\boldsymbol{\epsilon}$
- If ϵ is Normally distributed, so are the residuals
- Also note residuals sum to zero

Residuals and fitted values



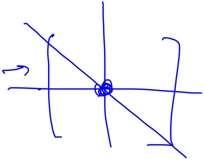
$$
\begin{aligned}
\underline{Cov(\hat{\boldsymbol{\epsilon}}^T, \hat{\boldsymbol{y}})} &= Cov((I - \boldsymbol{H})\boldsymbol{y}, \boldsymbol{H}\boldsymbol{y}) \\
&= (I - \boldsymbol{H})\sigma^2 I \boldsymbol{H}^T \\
&= \sigma^2(\boldsymbol{H} - \boldsymbol{H}) \\
&= 0
\end{aligned}
$$

- So residuals and fitted values are uncorrelated

Residuals when model is correct

- Often we plot the residuals against one of the predictors or against the fitted values
- What we look for:
  - ✓ $E(\hat{\epsilon}|x) = 0$
  - ✓ $V(\hat{\epsilon}|x) = \underline{\sigma^2(1 - h_{ii})}$

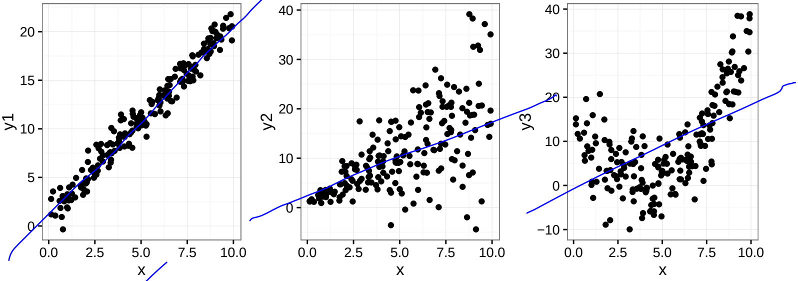
$\sigma^2(I - H) \to$



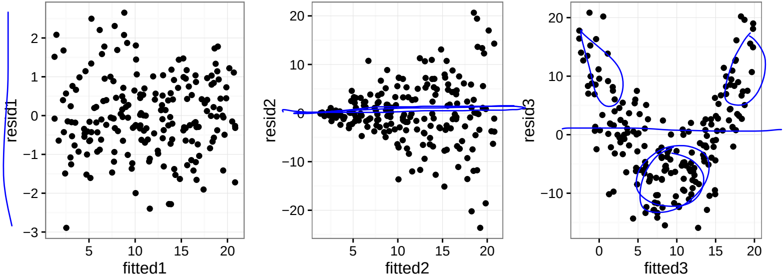
- If the model is incorrect, you may be able to spot:
  - Patterns in the residuals
  - Clear non-constant variance

Some data plots

$y_1 = X\beta_1 + \epsilon_1$ $\quad$ $y_2 = X\beta_2 + \epsilon_2$ $\quad$ $y_3 = X\beta_3 + \epsilon_3$



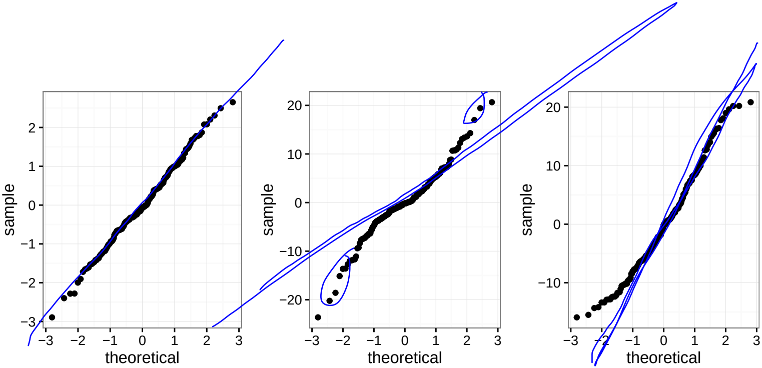
Some residual plots



Checking Normality assumption

- We often assume Normality for the errors
- Useful to check Normality of residuals
- Try a QQ plot:
  - Compute the sample quantiles of the residuals
  - Compute the quantiles of a standard Normal of size n
  - Plot these against each other
- Can also use the Shapiro-Wilk test based on correlation between observed and theoretical quantiles

Checking Normality assumption



Checking model structure

- You can plot residuals against each of the predictors, or plot outcomes against predictors
- Keep in mind the MLR uses adjusted relationships; scatterplots don't show that adjustment
- Adjusted variable plots (partial regression plots, added variable plots) can be useful

Adjusted variable plots

assoc. y + x_j

- Regress y on everything but x_j; take residuals $r_{y|-x_j}$
- Regress x_j on everything but x_j; take residuals $r_{x_j|-x_j}$
- Regress $r_{y|-x_j}$ on $r_{x_j|-x_j}$; slope of this line will match β_j in the full MLR
- Plot of $r_{y|-x_j}$ against $r_{x_j|-x_j}$ shows the "adjusted" relationship

What should you do ...

if your assumptions are violated?

- Depends on the assumption
- For problems with the errors, use LSE anyway; maybe use bootstrap for inference
- For non-linearity, try an augmented model

Today's big ideas

- Gauss-Markov, MLE, regression diagnostics

- Suggested reading: Faraway Ch 2.8, Ch 7