DSCI561: Regression I

Lecture 8: December 11, 2017

Gabriela Cohen Freue
Department of Statistics, UBC

Review from Lect 7

- Estimation and Inference in simple linear models
- Use bootstrapping to construct CI and test hypothesis
  - This can be useful when a closed form or asymptotic results of the estimator are not available

In today's lecture

- Extend the derivation of LS estimates to multiple linear regression
- Goodness of fit
- Diagnostics

LS for multiple regression

Recall from Lect 5 (simple regression):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

In multiple regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \ldots, n$$

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip}\right)^2$$

LS for multiple regression

In matrix notation:

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

We need to find values of betas that minimize the sum of squares:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial S}{\partial \beta_0} \\ \frac{\partial S}{\partial \beta_1} \\ \vdots \\ \frac{\partial S}{\partial \beta_i} \\ \vdots \\ \frac{\partial S}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

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

Then,

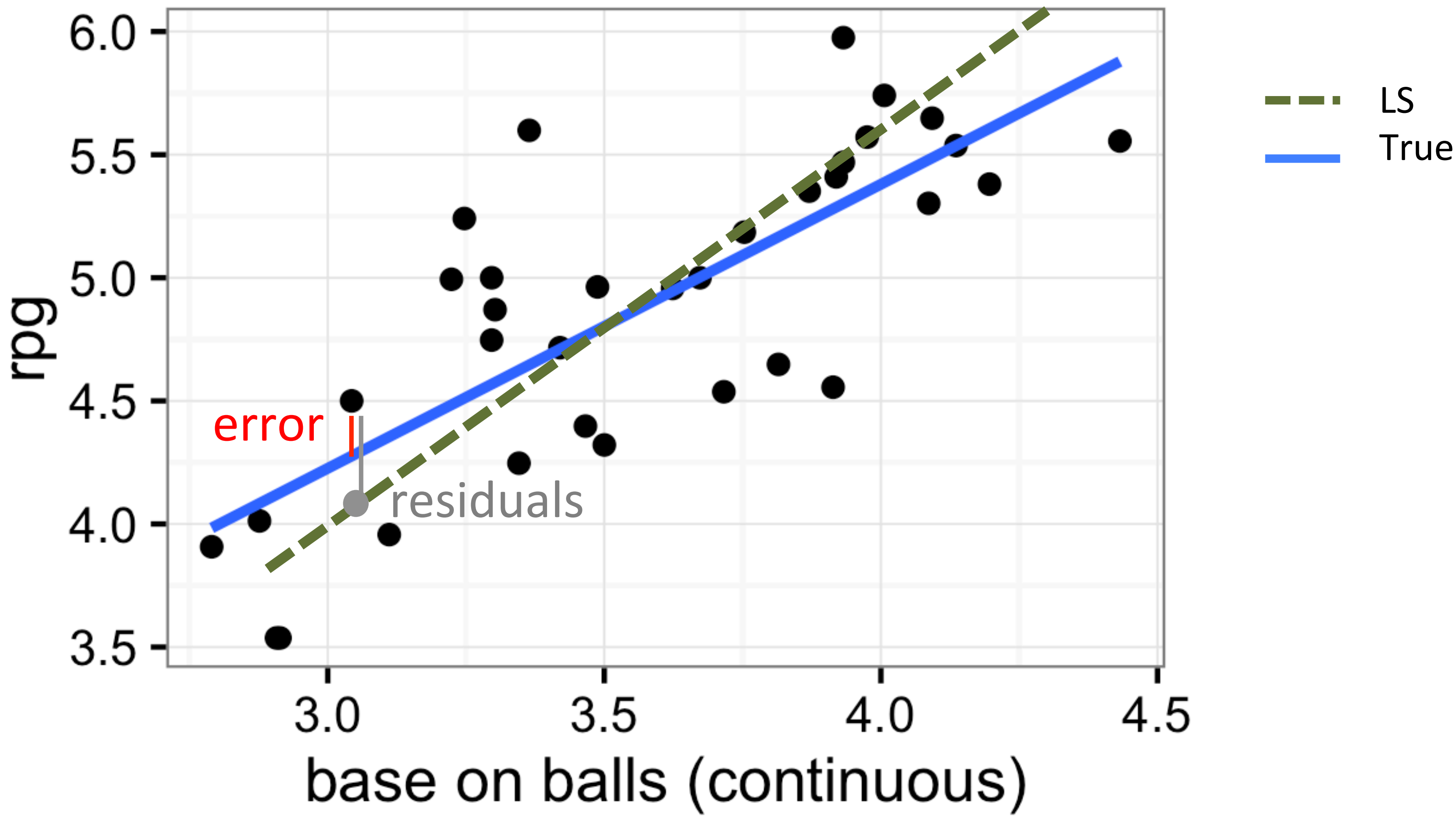
$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{y} \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

And,

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-p-1}}$$

Residuals



The residual is the vertical distance between the ***estimated*** line and the real observation

Coefficient of determination

The coefficient of determination, R^2, measures the proportion of the total variation in the y-variable explained by the regression.

$$R^2 = 1 - \frac{SS_{Resid}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The adjusted R^2, makes an adjustment to R2 so that it is not always increasing with additional explanatory variables.

$$adj R^2 = 1 - \frac{SS_{Resid}/(n-p-1)}{SS_{Total}/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_y^2}$$

Multicollinearity

- The least squares estimate satisfies: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$
- If $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular, the solution of $\hat{\boldsymbol{\beta}}$ is unique:

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

- However, $\boldsymbol{X}^T\boldsymbol{X}$ becomes nearly singular or singular when explanatory variables are collinear or multicollinear, i.e., multicollinearity problem.
- Under multicollinearity, the solution $\hat{\boldsymbol{\beta}}$ becomes very unstable (e.g., values and sign of some coefficients change as variables are added)
- Since $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, the SEs of $\hat{\boldsymbol{\beta}}$ can be large under multicollinearity.

Multicolinearity (cont.)

- Correlation between explanatory variables can be checked using pairwise plots
- Multicollinearity can be also measured through the *variance inflation factors (VIF):*

$$\mathrm{VIF}_j = \frac{1}{1 - R^2_{x_j, \boldsymbol{x}_{-j}}}, \; j = 1, \ldots, p$$

where $R^2_{x_j, \boldsymbol{x}_{-j}}$ is the coefficient of determination when x_j is regressed on the other explanatory variables in $\boldsymbol{X}$

- If $\mathrm{VIF}_j >> 1$, there is multicollinearity involving x_j in the data

Diagnostic plots

- Plot the residuals against the fitted values to check for homoscedasticity (i.e., constant variance) *vs* heteroscedasticity
- Plot the residuals against the each variable to check for possible structural deviations from the model (i.e., E[Y]=Xb)
- Normal Q-Q plot to check for normality

You should not see any pattern in all these plots, just noise

The Hat matrix

Recall from last lecture that

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

then,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}y$$

H puts a « hat » on y, thus it is called the « hat matrix »

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

h_{ij} measures the effect of the *j*-th observation on the prediction of the *i*-th response

The Hat matrix (cont.)

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2; \text{ for all } i$$

If $h_{ii} = 1 \implies h_{ij} = 0 \; \forall j \neq i$

$$\implies \hat{y}_i = y_i \text{ and } r_i = 0$$ Perfect fit!

Thus, a large h_{ii} suggests an unusually large influence of the *i*-th observation on the LS fit

Note that the hat-matrix does not depend on ***y***. Thus, outliers in the *y*-directions are not flagged by this measure

Mahalanobis Distance

Let $\boldsymbol{x}_i = (1 \; \boldsymbol{v}_i)$

$$\bar{\boldsymbol{v}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}_i$$

$$\boldsymbol{C} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{v}_i - \bar{\boldsymbol{v}})^T (\boldsymbol{v}_i - \bar{\boldsymbol{v}})$$

Define

$$MD_i^2 = (\boldsymbol{v}_i - \bar{\boldsymbol{v}}) \boldsymbol{C}^{-1} (\boldsymbol{v}_i - \bar{\boldsymbol{v}})^T = (n-1) \left[h_{ii} - \frac{1}{n} \right]$$

The Mahalanobis distance measures how far the *explanatory* part of the *i*-th observation is from the bulk of the data

Residuals

- We can complement the information from the hat matrix with that of the residuals using the *studentized residuals*:

$$t_i = \frac{r_i}{s\sqrt{1 - h_{ii}}}$$

Note: multiple outliers in the x-direction may pull the LS fit in their direction. Thus, their residuals look small compared to the residuals of the « clean » points

The Cook's squared distance

$$CD^2(i) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^T \boldsymbol{M} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))}{c}$$

$$CD^2(i) = \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}(i))^T (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}(i))}{ps^2}$$

$$CD^2(i) = \frac{1}{p} t_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

It measures the distance over which $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$ move when estimated (predicted) without the *i*-th case (single-case diaganostics)

Conclusions

- Most classical diagnostic measures (i.e., the hat-matrix, the MD_i, the studentized residuals, the Cook's distance) are useful in datasets with a *single* outlying observation
- The classical sample covariance, sample mean, and LS can be seriously affected by the presence of multiple outliers
- Multiple outliers may be difficult to flag with these measures since deleting only one point does not change the fit due to the remaining outlying points
- Some robust alternatives are available in the {robust} package, which examines residuals from a very robust fit.

anova()

Comparing the full *vs* the reduced models

$$F = \frac{\left(\mathrm{SSE}_{(\mathrm{reduced})} - \mathrm{SSE}_{(\mathrm{full})}\right) / (p-k)}{\mathrm{SSE}_{(\mathrm{full})} / (n-p-1)} \sim \mathcal{F}_{p-k, n-p-1}$$

$$R^2 = 1 - \left(1 + F\frac{p-1}{n-p}\right)^{-1}$$

- anova() can be used to compared any nested models (based on the same response). R gives an error message when models are not nested
- anova(lm()) can be used to examine the contribution of each term of the regression after controlling for previous fit variables (i.e., order matters! Type I SS). See Anova() in {car}