Outliers and Influence

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 1, 2007

Assumptions

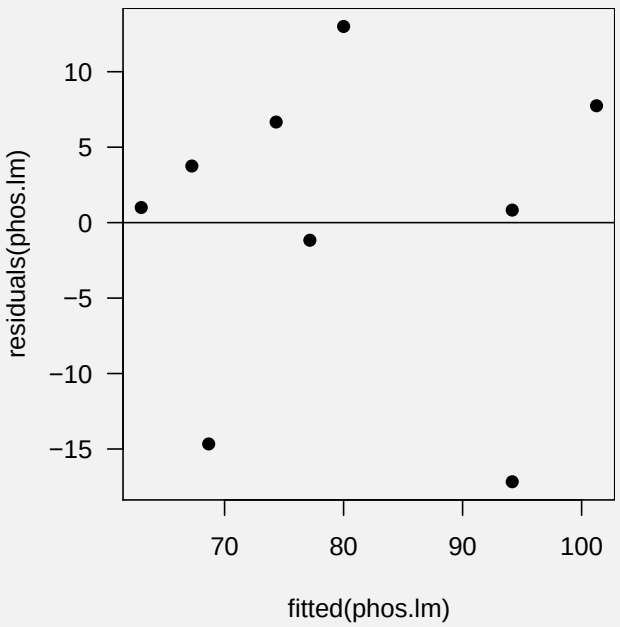
- Objectives in simple linear regression include:
  1. to describe the relationship between x and y.
  2. to estimate or predict y at a given level of x.
- Good inference depends on reasonable model assumptions.
  1. The model is correct: $E(y_i) = \beta_0 + \beta_1 x_i$.
  2. Errors e_i are independent.
  3. Errors e_i have homogeneous variance: $Var(e_i) = \sigma^2$.
  4. Errors e_i have normal distribution: $e_i \sim \mathrm{N}(0, \sigma^2)$.
- Our model diagnostic methods involve the examination of residuals.

Residuals

- A *raw residual* is defined as $r_i = y_i - \hat{y}_i; i = 1, \ldots, n$.
- A residual measures the deviation between observed value y_i and the fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- In linear regression, residuals sum to zero: $\sum_{i=1}^{n} r_i = 0$.
- Least squares regression minimizes the sum of the squared residuals: $\sum_{i=1}^{n} r_i^2 = \mathrm{SSErr}$.
- Residuals are related to the estimate of σ^2, the error variance: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$.
- The plot of residuals versus fitted values can indicate lack of model fit.

Soil P Example

```
> soilP = c(1, 4, 5, 9, 13, 11, 23, 23, 28)
> plantP = c(64, 71, 54, 81, 93, 76, 77, 95, 109)
> phos.lm = lm(plantP ~ soilP)
> par(las = 1, pch = 16)
> plot(fitted(phos.lm), residuals(phos.lm))
> abline(h = 0)
```



Observations

1. **Linearity**: *Look for high/low pattern or curvature in the residual plot.* Here, no obvious pattern.
2. **Independence**: *Check the study design, plot is not informative.*
3. **Equal variance**: *Look for differential scatter size, especially a fan-shaped pattern.* Here, no obvious pattern.
4. **Normal distribution:** *Look for linearity in normal scores plots.* Show that plot next!
5. **Possible outliers:** *Look for individual observations that stick out from the pattern.* Here, no obvious outliers.

Possible Residual Patterns

-
  - *Random scatter*: Assumptions are probably OK.
  - *Fan shape*: Unequal variances.
  - *Curvature*: Non-linearity.
  - *Extremely large residuals*: Possible outliers.
-
  - There are no golden rules or magic formulas.
  - Decision may be difficult with small sample size.
  - Regression diagnostics is *as much an art as a science*.

Demonstrate Residual Patterns with R

Do this live in class.

Types of Residuals

- *Raw Residuals:*
$$r_i = y_i - \hat{y}_i$$
- *Standardized Residuals:*
$$\frac{y_i - \hat{y}_i}{\hat{\sigma}}$$
where $\hat{\sigma} = \sqrt{MSErr}$ is an estimate of σ.
- *Studentized residuals:*
$$\frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$
where $h_i = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ is called the *leverage* for the ith observation.

Remarks About Leverage

- Leverage is related to the standard error of estimation.

$$\begin{aligned} \mathrm{SE}(\hat{\mu}_x) &= \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \times \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \hat{\sigma} \times \sqrt{h_i} \end{aligned}$$

- Leverage depends only on the explanatory variable, not the response variable.
- The sum of leverage values over the whole data set in simple linear regression is 2, *so the average is $2/n$ per data point*.
- $0 \leq h_i \leq 1$.

Influence

- *Influence* is a measure of how a data point effects the regression line.
- Specifically, a point is influential if removing it from the data set would result in substantial change to the regression line.
- Points with high leverage tend to be more influential than points with low leverage, but might not be.
- Influence *depends on both the response and explanatory variables*.
- *Cook's distance* is a measure of influence.

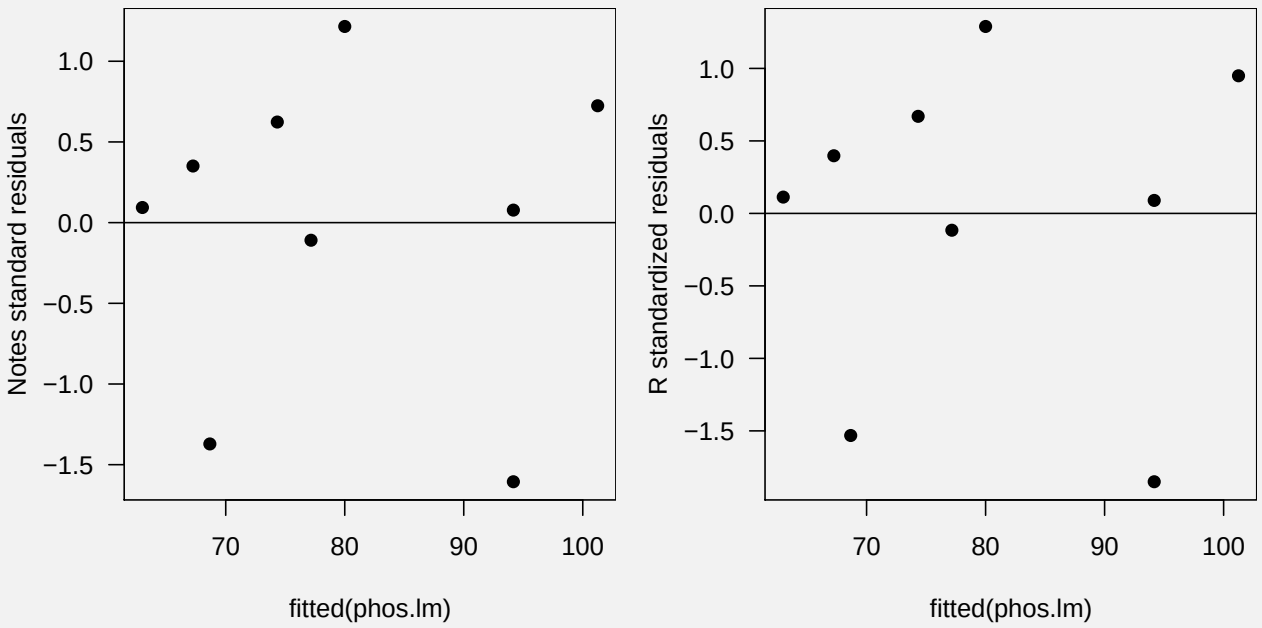
$$D_i = \frac{r_i^2}{(1+p)\hat{\sigma}^2} \times \frac{h_i}{(1-h_i)^2} = \frac{\tilde{r}_i^2}{1+p} \times \frac{h_i}{1-h_i}, \qquad p = 1 \text{ for SLR}$$

  where $\tilde{r}_i$ is the Studentized residual.
- Cook's distances more than about one might indicate data points worth extra attention.

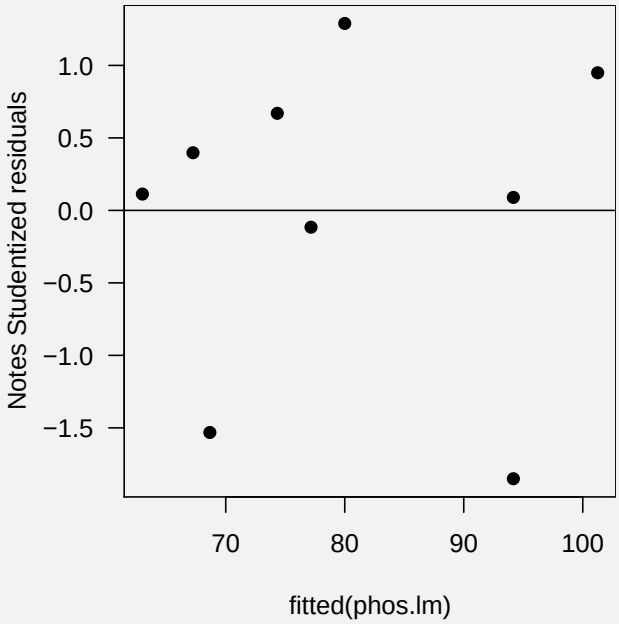
Residuals in R

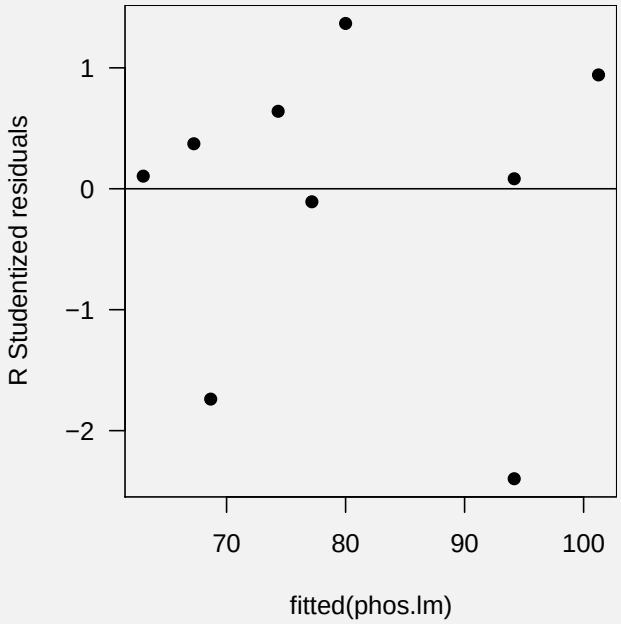
- R uses a *slightly different convention* for defining standardized and Studentized residuals.
- When computing either residual for the ith data point, R uses estimates of σ and the leverage from a regression using all points *excluding point i*.
- R functions `rstandard` and `rstudent` compute these alternative residuals.
- It turns out that `rstandard` agrees with the formula for Studentized residuals above.
- For all practical purposes, the differences are meaningless.
- As an aside, Student is capitalized because it stands for the name Student under which William Gossett, a statistician working for Guiness in the early 1900s used when publishing statistics articles. Student's t-test was his as well.

Standardized Residuals



Studentized Residuals





Influence Plots

```
> soilP = c(1, 4, 5, 9, 13, 11, 23, 23, 28)
> plantP = c(64, 71, 54, 81, 93, 76, 77, 95, 109)
> phos.lm = lm(plantP ~ soilP)
> par(mar = c(4, 4, 0, 0))
> par(las = 1, pch = 16)
> plot(phos.lm, which = 4)
```