Multiple Regression 4: Influence and Collinearity Diagnostics

Outline:

1. Influence measures
2. Collinearity diagnostics

Examples:

1. Influence Examples
2. Collinearity Example

1. Influence measures

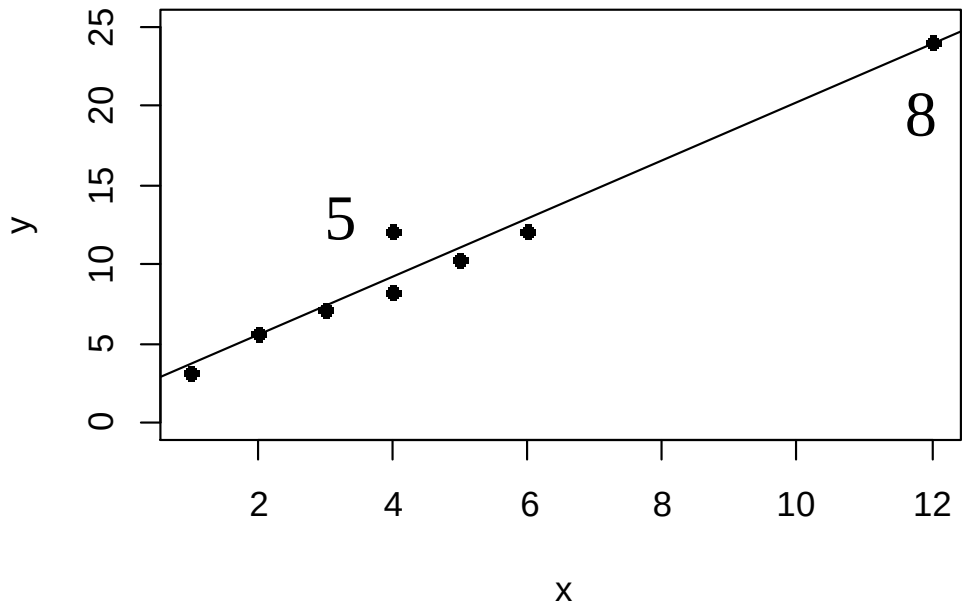
An observation is **influential** if deleting it substantially affects the conclusions of the analysis.

Influential observations exist when *both* of the following occur:
(1) an unusual Y value occurs (large residual)
(2) the X location has high leverage

No special tools are needed to spot influential observations in simple linear regression. However, we use simple linear regression to introduce diagnostic tools that are then applied to multiple regression.

We will discuss 4 measures of influence: Cook's distance, DFFITS, DFBETAS, COVRATIO. Plots of Cook's distance and leverage are given by `plot()`. All measures are computed using `influence.measures()`.

Influence Example #1

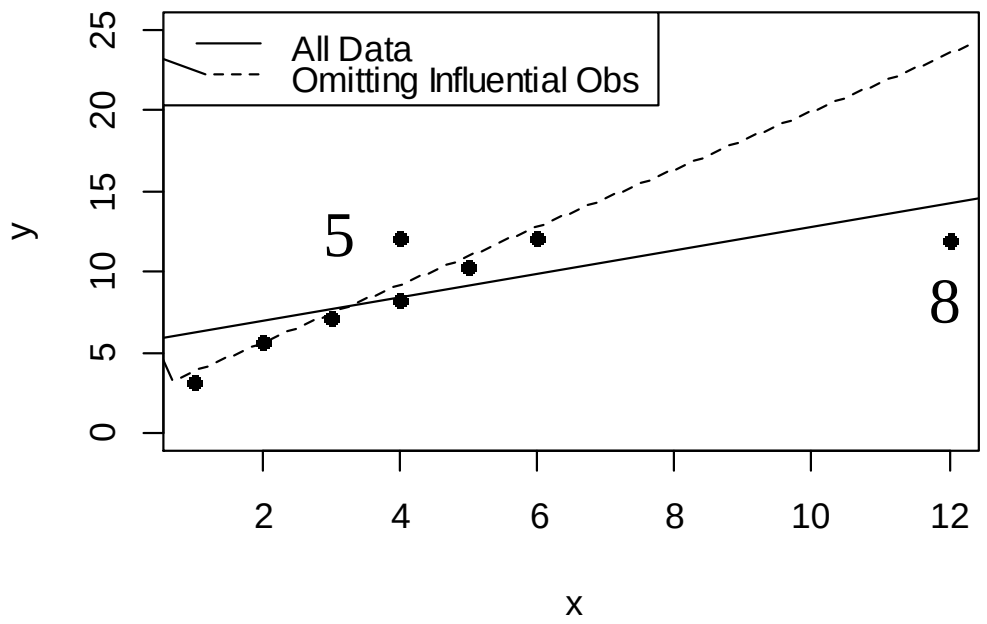


Omitting either point #5 or point #8 would not have much effect on the estimated slope. So neither point is influential.

Although point #5 has a large residual, its position in the middle of the X values minimizes its effect on the slope estimate (low leverage).

Although point #8 has a position of high leverage, it is has a small residual.

Influence Example #2



We can see that omitting point #8 has a dramatic effect on the estimated slope. So point #8 is very influential.

Point #8 has a large residual and the X value is in a position so that it can pull the line toward itself (high leverage).

Point #5 is still not influential.

Leverage ("Hat" in R) is a measure of the leverage of the i^{th} point . h_i is the i^{th} diagonal element of the n by n matrix:

$$\mathbf{X(X^TX)^{-1}X^T}$$

Important note: h_i depends entirely on the X's.

Rule of thumb: If $h_i > \frac{3(k+1)}{n}$, observation has high leverage.

The value, h_i, determines the SE of the predicted value for a particular point or residual : $SE(\hat{y}_i) = \hat{\sigma}\sqrt{h_i}$, $SE(e_i) = \hat{\sigma}\sqrt{1-h_i}$

Both formulas reflect the ability of a point to pull the line toward itself, <u>increasing</u> the variability of predicted value and <u>decreasing</u> the variability of the residual. Some properties of h_i are:

(1) $\sum_{i=1}^{n} h_i = k+1$ $\left(\text{average } h_i \text{ value is } \frac{k+1}{n}\right)$

(2) $\frac{1}{n} < h_i < \frac{1}{c}$ where c is the number of rows in X that are the same as the i^{th} row.

Standardized (or studentized) residuals:

The size of the residual is judged by the standardized (or studentized) residual:

$$s_i = \frac{e_i}{\text{SE}(e_i)} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$$

Rule of thumb: If absolute value $s_i > 2$, observation has large standardized residual.

To have actual influence a point needs to have high leverage and a large residual.

Influence Example #1

Obs#	h_i (hat)	s_i (std resid)	D_i (Cook's d)
5	0.130	2.319	0.401
8	0.806	0.191	0.076

Leverage: With n = 8, we check $h_i > \frac{3(k+1)}{n} = \frac{3(1+1)}{8} = 0.75$.

Residual: Check $|s_i| > 2$.

Point #5 has large residual, but low leverage. Not influential.
Point #8 has small residual, but high leverage. Not influential.

Influence Example #2

Obs#	h_i (hat)	s_i (std resid)	D_i (Cook's d)
5	0.130	1.617	0.195
8	0.806	-2.056	8.782

Leverage: With n = 8, we check $h_i > \frac{3(k+1)}{n} = \frac{3(1+1)}{8} = 0.75$.

Residual: Check $|s_i| > 2$.

Point #5 has "small" residual and low leverage. Not influential.
Point #8 has large residual and high leverage. Influential.

Cook's Distance is a measure of actual influence that combines leverage and the standardized residual into one number:

$$D_i = \frac{1}{k+1}(s_i)^2\left(\frac{h_i}{1-h_i}\right)$$

Rule of Thumb: If $D_i > 1$, observation considered influential.

The s_i is the standardized residual. Dividing h_i by (1- h_i) accentuates the h_i. Multiplying by the squared standardized residual by the function of h_i combines information about the residual with the information about leverage.

Influence Example #1: Point #5 has D=0.401. Point #8 has D=0.076. Neither is considered influential.

Influence Example #2: Point #5 has D=0.195 so is not influential. Point #8 has D=8.782 so is enormously influential.

An alternate interpretation of Cook's distance:

Let: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ $\hat{\beta}_{-i} = \begin{pmatrix} \hat{\beta}_{0,-i} \\ \hat{\beta}_{1,-i} \end{pmatrix}$

Think of $\hat{\beta}_{-i}$ as **fixed.**

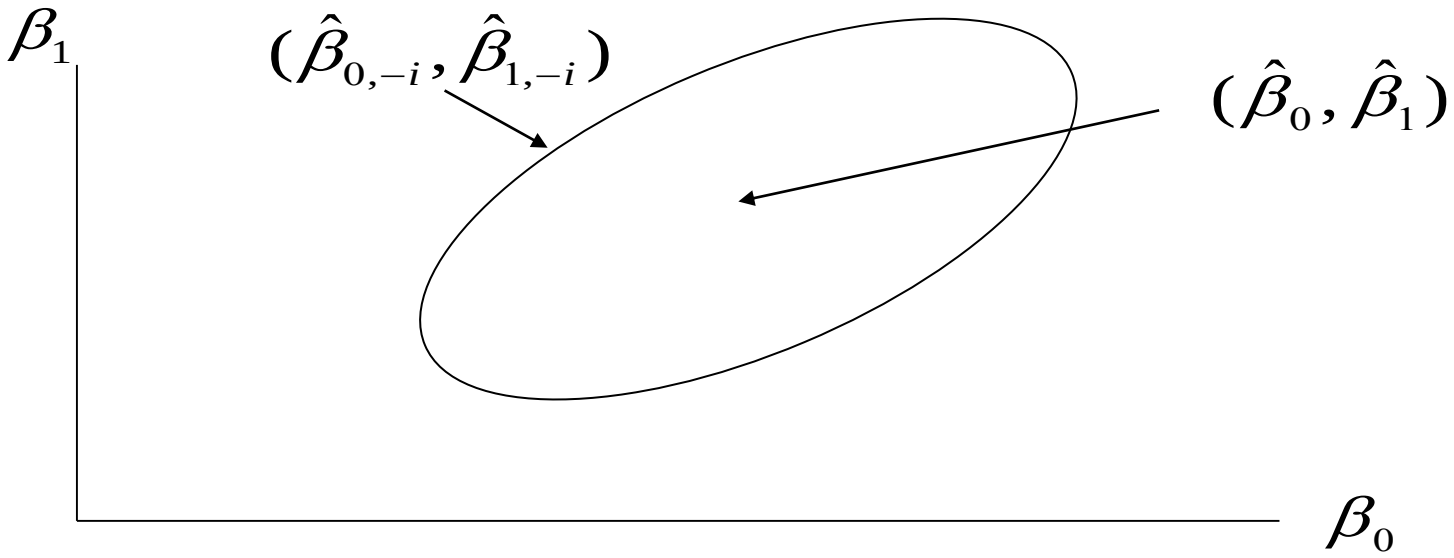
Test: $H_0 : \hat{\beta} = \hat{\beta}_{-i}$ vs $H_1 : \hat{\beta} \neq \hat{\beta}_{-i}$

$$D_i = F = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T X^T X (\hat{\beta} - \hat{\beta}_{-i})}{(k+1)\hat{\sigma}^2} = \frac{[SSE(\text{reduced}) - SSE(\text{full})]/(k+1)}{\hat{\sigma}^2}$$

$df1 = k+1$ $df2 = n-k-1$

$$pvalue = 1 - pf(D_i, df1, df2)$$

Use the fact that a 95% confidence interval is the set of values that would not be rejected in a $\alpha = 0.05$ hypothesis test. Then if the pvalue = α, , $\beta_{0,-\iota}$ and $\beta_{1,-\iota}$ are at the edge of a 95% confidence region.



Influence Example #2: For Point #8, D = F = 8.78.

df1=k+1=2, df2 = 8-2 = 6. From R we compute:

```
p = 1-pf(8.78,df1 = 2, df2 = 6)
```

p =0.0165. The same F and pvalue were obtained from the `lht()` function. Therefore removing Point #8 from the regression would move the estimate of the β's to the edge of a 98.35% confidence ellipsoid. That's a lot!

DFFITS (Difference in the Fitted Value Standardized) is another measure of influence:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(-i)i}}{\hat{\sigma}_{-i}\sqrt{h_i}}$$

This measures how many standard deviations the fitted (predicted) value moves when that point is added/deleted.

Rule of thumb: If absolute value DFFITS exceeds 2*sqrt((k+1)/n), the observation is considered influential.

Influence Example #2: 2*sqrt((1+1)/8)=1.0.

Point #8 had DFFITS=-7.042, so highly influential.

DFBETAS (Difference in the Beta's Standardized) This measures how many standard deviations the estimates of beta move when that point is added/deleted.

For the i^{th} point the DFBETA for β_j is:

$$\text{DFBETA}_i = \frac{\hat{\beta}_j - \hat{\beta}_{(-i)\,j}}{SE(\hat{\beta}_j - \hat{\beta}_{(-i)\,j})}$$

The denominator is computed using $\hat{\sigma}^2_{-i}$.

Rule of thumb: If absolute value DFBETA exceeds $2/\sqrt{n}$, the observation is considered influential.

Influence Example #2: 2/sqrt(8)=0.71

Point #8 has DFBETAS of 3.78 (intercept) and -6.47 (slope), so highly influential.

COVRATIO: The ratio of the "generalized variance" of the β's, without a point, relative to with a point. Deleting points multiplies the estimated variance the COVRATIO.

$$\text{COVRATIO}_i = \frac{Var(\hat{\beta})_{\text{without point i}}}{Var(\hat{\beta})_{\text{with point i}}}$$

Rule of Thumb: If COVRATIO deviates in either direction from 1.0 by more than 3*(k+1)/n, then observation considered influential.

For our examples: 3*(1+1)/8 = 0.75

Example #1: Deleting Point #8 , increases the variance of the β's. Deleting Point #5 decreases it .

Example #2: No points appear to be so influential on the covariance of the betas.

2. Collinearity Diagnostics

In geometry, a set of points is collinear if they lay on a single line.

In statistics, collinearity refers to an exact or approximate linear relationship between two or more predictor variables. In other words, one predictor variable can be linearly predicted from the others. See examples next slide.

When the X's are nearly collinear, the β's are hard to interpret, and very poorly estimated (have high variances).

When collinearity is detected, the easiest solution is to simply remove one or more of the violating predictors from the regression model!

The simplest kind of collinearity is when two predictor variables are correlated. This can be detected by examining pairwise scatter plots and correlation values (using corr).

Examples of EXACT Collinearity

Example #1: $X_2 = 2X_1$

X_1	X_2
1	2
2	4
3	6
4	8

Example #2: $X_1 = X_2 + X_3$

X_1	X_2	X_3
1	1	0
1	1	0
1	0	1
1	0	1

In practice, exact collinearity in multiple regression is rare. When X's are exactly collinear, the β's are not unique. If you attempt to fit a model with exact collinearity, you may get NA values for some estimated values.

In practice, we are more interested in detecting near collinearity!

Collinearity Diagnostics

Remember to start by examining pairwise correlations and scatterplots. But some types of collinearity could still be missed. So we present some additional diagnostics.

Variance Inflation Factor (VIF):

$$\text{VIF}(\beta_i) = \frac{\text{Var}(\hat{\beta}_i) \text{ estimated by the model}}{\text{Var}(\hat{\beta}_i) \text{ as if } X_i \text{ were indep. of other X's}}$$

$$= \frac{1}{1 - R^2 (\text{from the reg. of } X_i \text{ on other X's})}$$

Rule of Thumb for VIF: VIF > 4 or 10 indicates collinearity.

Use `vif()` from the car package to calculate variance inflation factors.

Other measures of collinearity are based on eigenvalues:
$\lambda_1, \lambda_2, ..., \lambda_k$.
By definition $\lambda_1 > \lambda_2 > ... > \lambda_k$ and $\sum \lambda_i = k$.
λ_i/k gives the proportion of total variance of the X's that lies in the direction of the axes of the X data cloud.

Condition Index = $\sqrt{\lambda_1}/\sqrt{\lambda_i}$
Rule of Thumb: Condition index > 30 indicates collinearity.

Variance Proportion gives the proportion of the variance of a β_i that is associated with each eigenvalue. Look at the small eigenvalues and identify variables having high "variance proportions". These variables may be nearly collinear.

Use `colldiag()` from the perturb package to calculate both of these diagnostics.