Lecture 17

Monday, July 8, 2019 2:37 PM

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Aside: Adjusted R^2 (R^2_{adj}):

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2)$$

* $R^2_{adj} < R^2$, but as $n \to \infty$, $R^2_{adj} \to R^2$

* This adjusted version <u>penalizes</u> you for adding unimportant explanatory variables in to the model. Whereas R^2 will <u>always</u> increase when additional explanatory variables are added to a model, the adjusted R^2 may increase <u>or decrease</u>.

If the reduction in SSE brought about by the additional of extra explanatory variables is not big enough to offset the penalty, then R^2_{adj} will go down. But, if the reduction in SSE is big enough, R^2_{adj} will go up.

* Note that R^2_{adj} doesn't have a useful interpretation, but it does a better job at describing goodness-of-fit than the ordinary R^2.

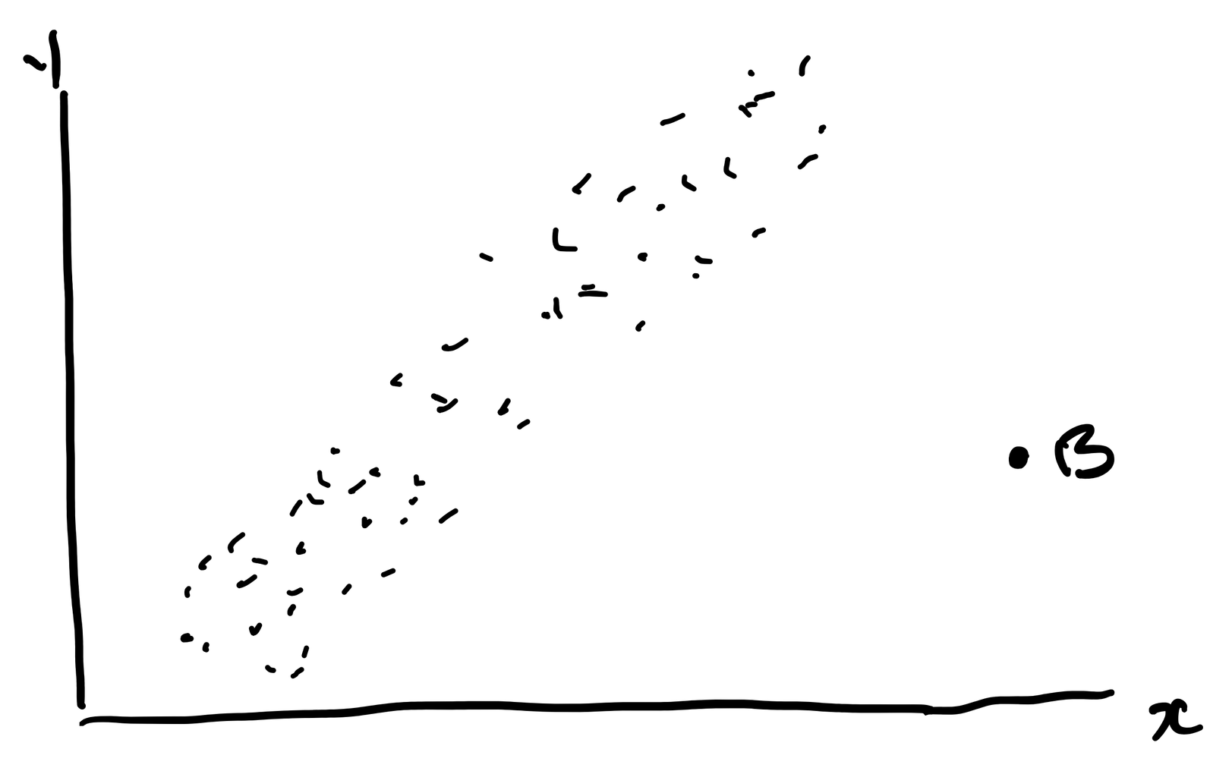
The Effect of Individual Observations

Q: Do all observations look like they come from the same model?

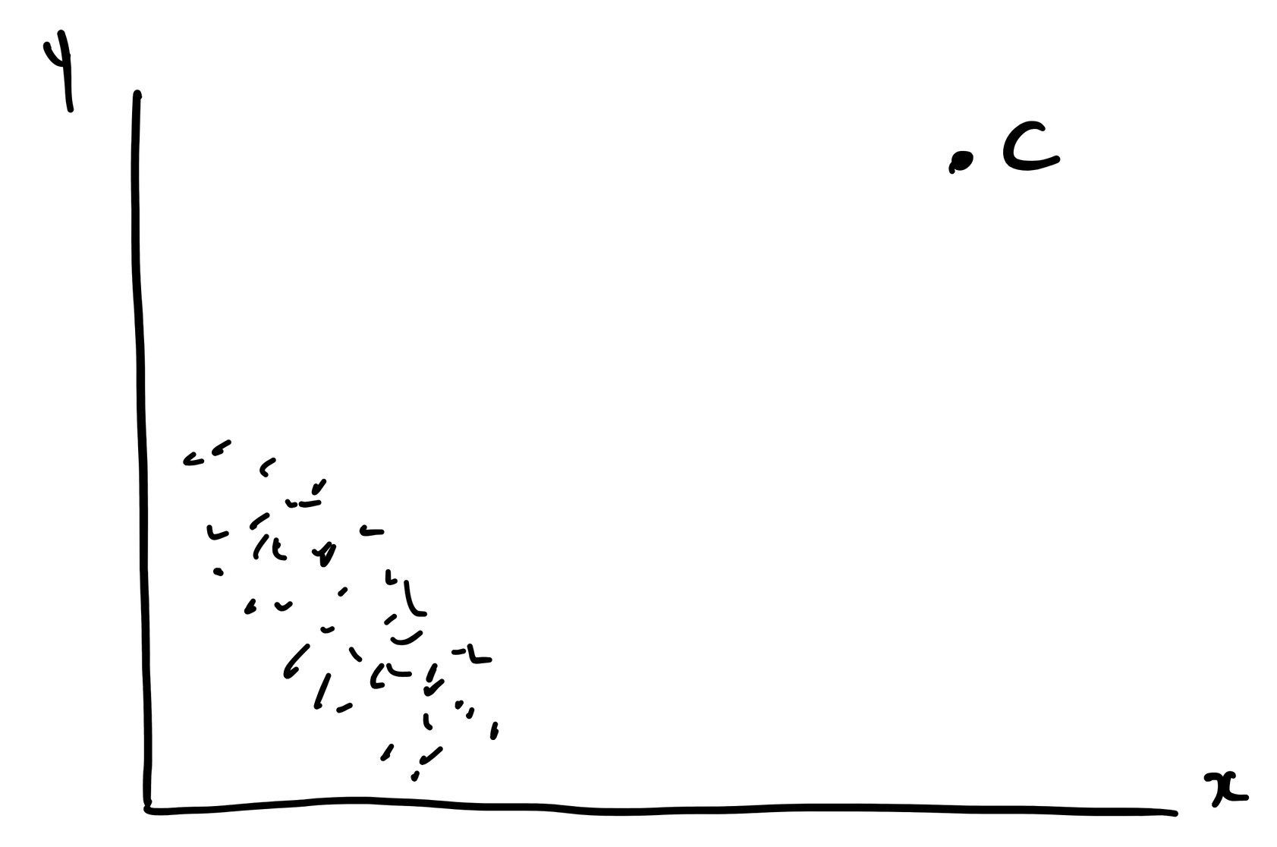
We call an observation $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ an <u>outlier</u> if it differs substantially from the other observations. Since several variables are involved we must distinguish outliers in the y (response) dimension from outliers in the x (explanatory variable) dimension.

To visualize this distinction assume we have one y and one x:

* A is an outlier in the y-dimension



* B is an outlier in the x-dimension



* C is an outlier in both the x and y dimensions

Outliers can arise for many reasons (i.e., data recording issues or unexpected natural phenomenon), but in no matter the reason, we want to be able to identify them. We may then decide to:

- delete them
- correct them
- perform analyses both with/without them
  - ← do this when an outlier cannot be explained.
- collect more data

Detecting Outliers

A simple for detecting outliers in the y dimension is to look at the <u>studentized residuals</u>:

$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

which we expect them to behave like $N(0,1)$. Thus $|d_i| > 3$ identifies an observation that has a sufficiently unusual y_i value.

This approach is simple and effective, but it won't identify outliers in the x-dimension. (i.e., point B). To detect outliers more generally, we need two new concepts:

- <u>Leverage</u>: An observation has a lot of leverage if it "pulls" the fitted values toward it (i.e., $y_i \approx \hat{\mu}_i$). We will see that the extent of the "pull" is determined by the x's.

- <u>Influence</u>: An observation has a major influence if the fitted regression model is significantly altered when the observation is omitted. (i.e. point C)

Recall that the fitted values are:

$$\hat{\vec{\mu}} = X\hat{\vec{\beta}} = X(X^TX)^{-1}X^T\vec{y} \equiv H\vec{y}$$

For a given observation i the fitted value is:

$$\hat{\mu}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

(i^{th} element of $\hat{\vec{\mu}}$) ← (i^{th} row of H) $\times \vec{y}$

The coefficient h_{ii} indicates how heavily y_i contributes to the fitted value $\hat{\mu}_i$. If h_{ii} is large (compared to the other h_{ij}'s) then y_i dominates $\hat{\mu}_i$.

It can be shown (see A5y) that $0 \le h_{ii} \le 1$ and note that if $h_{ii} \approx 1$ then $Var[e_i] = \sigma^2(1-h_{ii}) \approx 0$. Thus, if $h_{ii} \approx 1$ then $y_i \approx \hat{\mu}_i$ and we'd say that observation i has <u>high leverage</u>.

We will define h_{ii} to be the <u>leverage</u> of observation i. We will say that observation i has <u>high leverage</u> if:

- $h_{ii} \approx 1$

  and/or

- $h_{ii} > 2\bar{h}$ ← $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii}$
  - ← the leverage is much bigger than the other points' leverages.

* Since H depends only on the x's, the concept of leverage helps find outliers in the x dimension.