

Detecting Influential Data and Collinearity

1 Influential Data

Predicted values (\hat{Y}_j) are all linear combinations of B_0, \dots, B_k . Certain observations may be given unusually large weights in determining these quantities. These observations are influential data. It is advised to find out which of your observations are unusually influential; this can be tricky in a high-dimensional setting.

Note: Influential data points are not necessarily bad data points.

There are at least two very different kinds of influential data points:

1. Unusual predictor combinations $\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{j,k})$: leverage points.
2. Unusual Y_i values (observations whose Y-values do not fit the pattern in the rest of the regression): Discrepancy values or outliers.

See the following page for a pictorial representation of outliers and leverage points.

Leverage Points: We will use the so called 'hat matrix' H to find leverage points in our data. $\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{j,k})$ is a point if it has an unusually large h_{ii} value, where:

$$h_{ii} = \frac{1}{n} + \frac{X_i - \bar{X}}{\sum(x^2)} \quad (1)$$

There are different 'rules of thumb' for cutoff values. If the data set is large and a large number of predictors is being used (large k) then an $h_{ii} > \frac{2(k+1)}{n}$ is considered large. For smaller samples and $h_{ii} > \frac{3(k+1)}{n}$ is considered large. What to do with leverage points? Take a closer look (not all influential data points are bad).

Outliers: The Studentized Deleted residuals (more often called externally studentized residuals or, in PROC GLM, RSTUDENTs) are in fact functions of SSE, Y_i , $\hat{Y}_{i(i)}$ and h_{ii} :

$$t_i = \frac{Y_i - \hat{Y}_{i(i)}}{\sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}} \quad (2)$$

If the errors of the model are distributed Normal then t_i have a t-distribution with df $n-k-1$. In larger sample sizes cutoffs of ± 3 , ± 3.5 or ± 4 are used.

Influential Measures: There are two statistics that we will discuss today that measure the combination of leverage and discrepancy for an observation. The measure the degree to which an observation affects parameter estimates. The two measurements we will discuss today are DFFITS and DFBETAS. The formula for them the DFFIT for the i^{th} observation is:

$$DFFITS_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}} \quad (3)$$

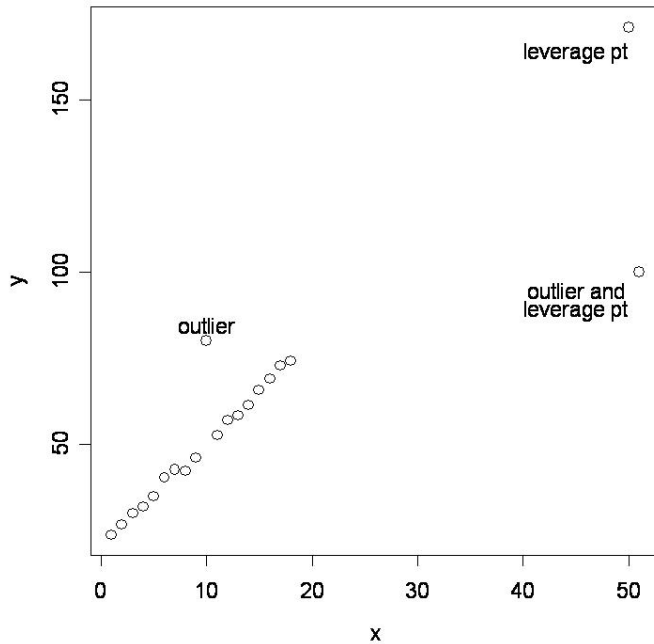


Figure 1: Graph emphasizing the differences between outliers and leverage points

And the DFBETA for the j^{th} B weight and the i^{th} observation is:

$$DFBETAS_{ij} = \frac{B_j - B_{j(i)}}{SE_{B_{j(i)}}} \quad (4)$$

Cutoff values for small to medium sized data sets are $DFFITs_i > 1$ and $DFBETAS_{ij} > \pm 1$. For large data sets they are $DFFITs_i > 2\sqrt{\frac{k+1}{n}}$ and $DFBETAS_{ij} > \pm \frac{2}{\sqrt{n}}$.

Note: All of these measures should be used as guidelines to find $\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{j,k})$ and Y_i 's that are unusual. These observations should be checked. Don't take the cutoff values too literally. Here is an example which looks at the leverage, discrepancy, and dffits:

```
proc glm data=tmp1.hanes;
model tcp = sysbp diabp;
output out=game H=leverage rstudent=esr dffits=fits;
run;

proc print data=game;
run;

proc gplot data=game;
plot (leverage esr fits)*num;
run;
quit;
```

```
proc sort data=game;
by fits;
run;
```

```
proc print;
run;
```

Here is an example which looks at the DFBETAS:

```
proc reg data=tmp1.hanes;
model tcp = sysbp diabp/influence;
ods output OutputStatistics=dvsOut;
run;
```

```
proc print data=dvsout;
run;
```

```
proc means data=dvsout;
run;
```

```
proc gplot data=dvsout;
plot (DFB_Intercept DFB_diabp DFB_sysbp)*observation;
run;
quit;
```

```
proc sort data=dvsout;
by dfb_diabp;
run;
```

2 Collinearity

One of the main uses of multiple regression is to measure the effect of one or more explanatory variables on the response while holding all of the others constant. The multiple regression model is defined as:

$$Y = B_0 + B_1X_1 + B_2X_2 + \dots + B_{p-1}X_k + \varepsilon \quad (5)$$

where Y is the response, the X_i 's are the explanatory variables, and $\varepsilon \sim N(0, \sigma^2)$ is the error term with constant variance. The estimates of the B_i 's are the predicted change in the response variable per unit increase, while the remainder of the explanatory variables are held constant. In multiple regression analysis, interpretation of the marginal relationships between the response and explanatory variables depends strongly on the assumption that the explanatory variables are not linearly related. Strong linear relationships between the explanatory variables make interpretation of the relationship between one variable and the response inaccurate because, in the presence of these relationships, it becomes impossible to change one of the variables without changing another. The presence of linear relationships is called *collinearity*, and the measurement and interpretation of the marginal effects of some explanatory variables can be lost.

The output of regression analysis for a collinear model can, at first glance, give the appearance of a highly significant model. The model could have a high R^2 , low \sqrt{MSE} , and a highly significant F statistic. The problems with collinearity are not with the fit of the model, they are with the estimates of the β_i 's. When dealing with a collinear model prediction may not be an issue if is being done within the range of X values. However; when extrapolating outside the range, consequences will be more severe than with a model, which does not have the problem of collinearity.

After the model is chosen there are other symptoms of collinearity:

- Variables which were theoretically thought, or previously shown to have a significant impact will have unusually large standard errors, and their t-tests may therefore yield insignificant p-values.
- Another symptom is that estimates of coefficients can have values inconsistent with theory and/or similar previous experiments, with "wrong signs" (a negative coefficient when the true effect is positive) or extreme magnitudes.

Variance Inflation Factor (VIF) is the statistic that we will use today to detect collinearity. The VIF is useful in part due to the ease of availability. It is readily available in almost every statistical package, and some issue a warning sign when high VIF's are observed. In order to give the formula for VIF first, define R_k^2 as the R^2 statistic of a regression model of X_k on the remaining p-2 explanatory variables. The VIF for explanatory variable k is then:

$$VIF_k = \frac{1}{1 - R_k^2} \quad (6)$$

The VIF relates to the variance of the estimated coefficient by:

$$Var(b_k) = \frac{\sigma^2}{(n - 1)s_k^2} VIF_k \quad (7)$$

VIF can be related to a statistic called Tolerance (TOL) by:

$$TOL_k = \frac{1}{VIF_K} \quad (8)$$

The difficulty of using VIF for detection of collinearity is that there are no rigorously justifiable cutoff values. Most texts dealing with regression diagnostics suggest using 5 to 10 as cutoff values ($R_k^2 = .8$ to $.9$, $TOL_K = .1$ to $.2$) for determining when corrective action should be taken.

Since GLM does not give an output of VIF we will need to use PROC REG to obtain it. GLM can be used to obtain Tolerance:

```
proc reg data=tmp1.hanes;
model tcp = sysbp diabp/vif tol;
run;
```

```
proc glm data=tmp1.hanes;
```

```
model tcp = sysbp diabp/tolerance;  
run;  
quit;
```

Principal components and Ridge Regression are two ways of dealing with collinearity, but both are difficult (see Neter, Kutner, Nachtsheim and Wasserman (1996) for details). On blackboard I've included a data set which has the problem of collinearity.