Cook's Distance

Definition

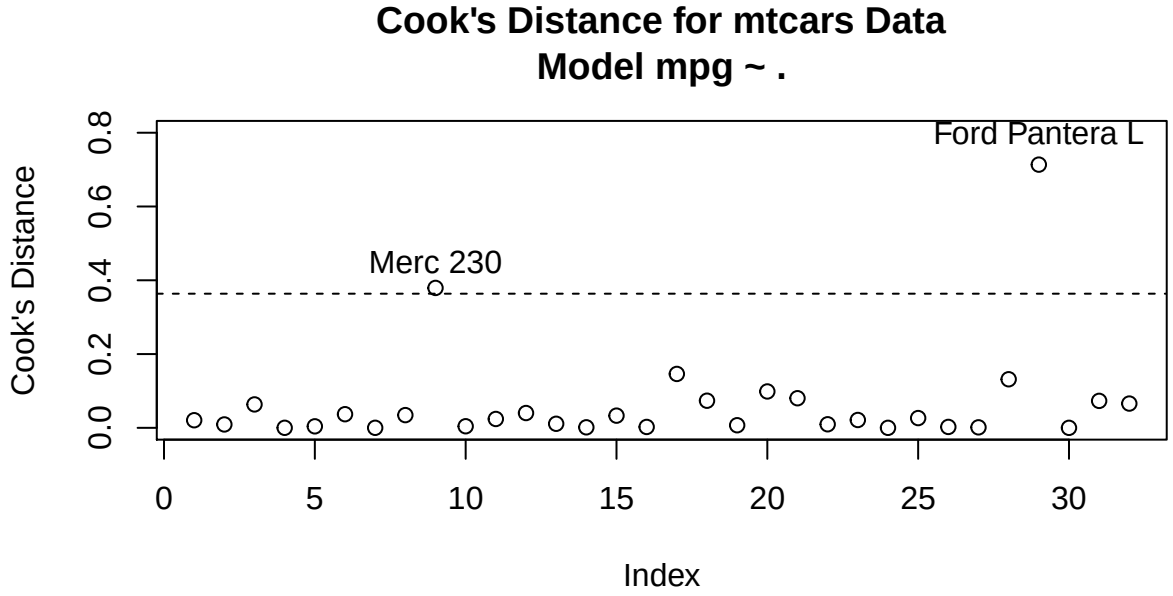
Cook's Distance calculates the leverage of an observation by finding the "distance" between $\beta_j - \beta_{j(-i)}$ where $\beta_{j(-i)}$ is the model β when the regression is fit minus that observation. The difference $\beta_j - \beta_{j(-i)}$ is tested with an F-test for the hypothesis $\beta_j - \beta_{j(-i)} = 0$ to evaluate the leverage of that observation. In the second equation, $\frac{e_i^2}{p\sqrt{MSE}}$ is a measure of *discrepancy* and $\frac{h_{ii}}{1-h_{ii}^2}$ is a measure of *leverage*.

Equations

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\sqrt{MSE}}$$

$$D_i = \frac{e_i^2}{p\sqrt{MSE}}[\frac{h_{ii}}{1-h_{ii}^2}]$$

Example Plot



Interpretation and Use

Cook's Distance is useful for finding high leverage observations in the data. A simple rule of thumb for identifying leverage points is $D_i > 4/n$. Since the numerator in the equations above is formally identical to the Wald statistic, W, then W/p has an $F_{p,n-p}$ distribution such that the F statistic for Cook's distance, $F_{p,n-p,1-\alpha}$, is an appropriate threshold approximated by $\frac{4}{n-p-1}$. Therefore, observations above this threshold should be evaluated for validity. They

may also indicate that more data in the leverage region may improve the regression model. Cook's distance helps to identify observations that may need further study or accounting in the analysis.

Solutions and Further Avenues
Since Cook's Distance identifies high leverage points, these observations should be evaluated further with other diagnostics such as studentized residuals, DFBETA, and DFFITS.

R Code

```
# library(car)
# data(mtcars)
# model <- lm(mpg ~ ., data=mtcars)
# plot(cooks.distance(model),ylim=c(0,0.8),main="Cook's Distance for mtcars
# Data\nModel mpg ~ .", ylab="Cook's Distance")
# abline(h=4/(length(mtcars)), lty=2)
# labels=row.names(mtcars)
# text(c(9,29),c(0.45,0.8),labels=labels[c(9,29)])
```