Lecture 8
Linear Regression Diagnostics, cont.

BIOST 515

January 29, 2004

Education as a predictor for mortality: an extended example

Previously, we looked at education as a predictor for mortality,

$$mort_i = \beta_0 + \beta_1 Education_i + \epsilon_i.$$

We want to extend this analysis:

- Include other factors which may cloud the relationship between education and mortality
- For example, we may expect that poverty may play a role in access to education and also higher mortality.
- No variable for poverty - will need to use other variables.

- We have median income. Why might we not want to include that?

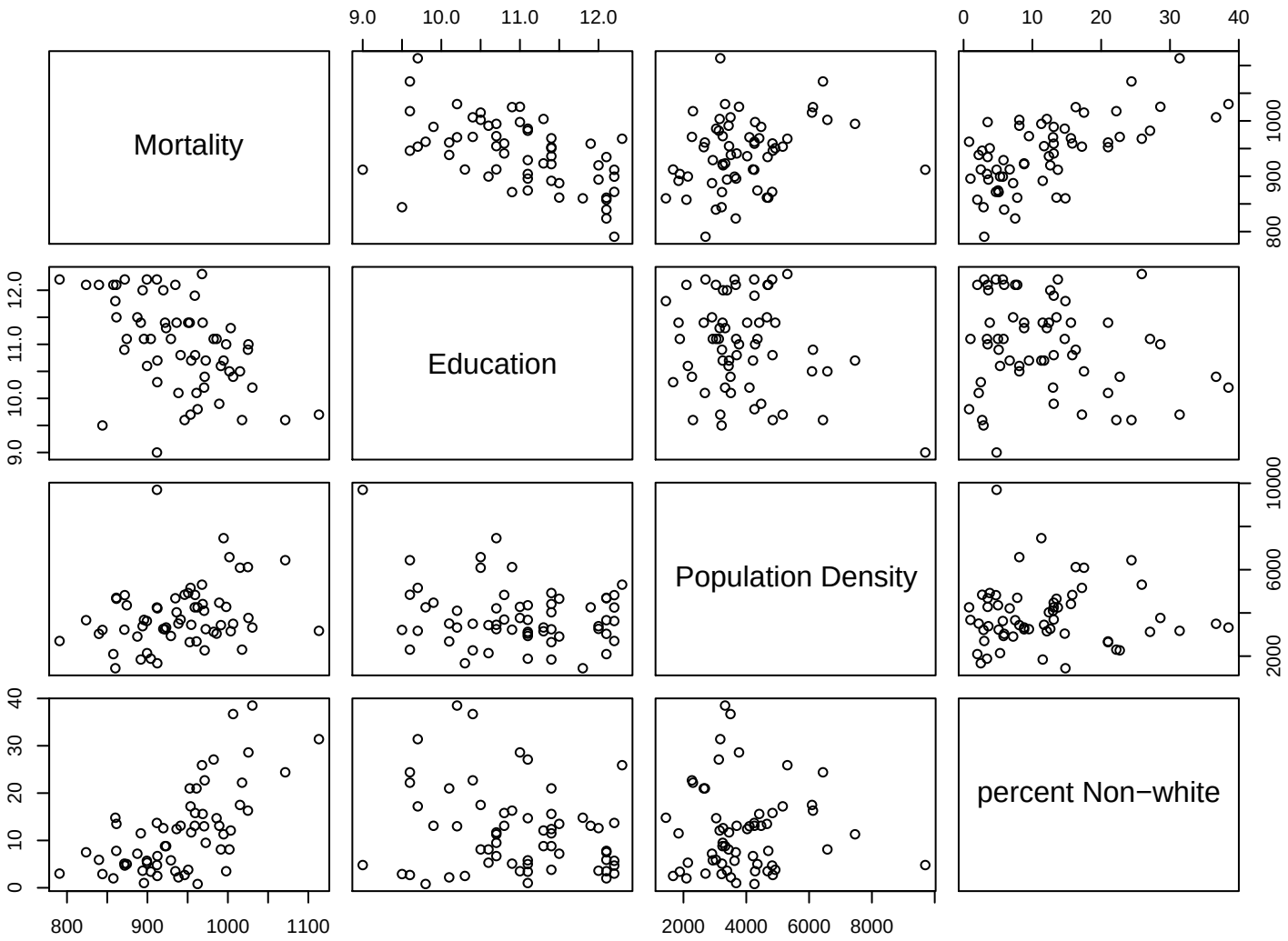
- Instead, use the percentage of the population that is non-white and the population density

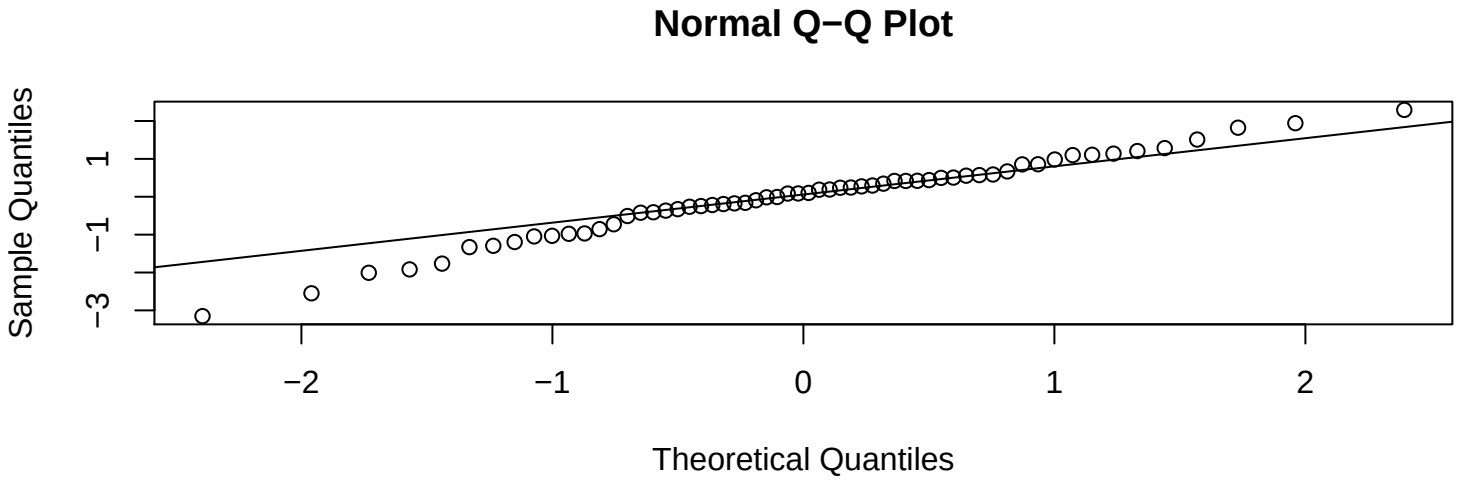
Our new model is

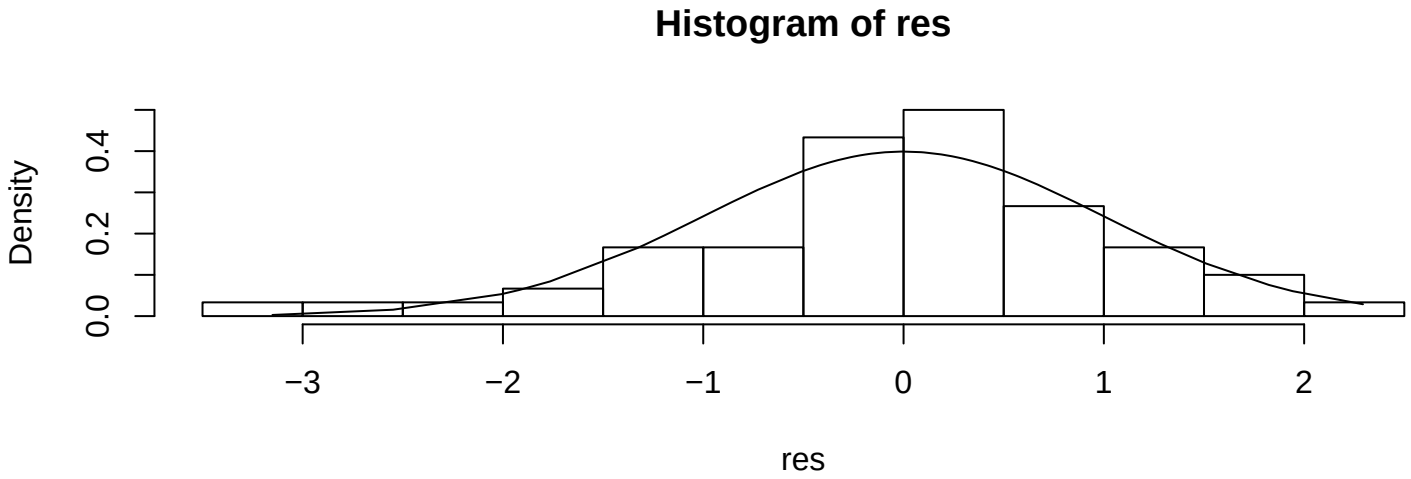
$$mort_i = \beta_0 + \beta_1 Education_i + \beta_2 PopDensity_i + \beta_3 pNonWhite_i + \epsilon_i.$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1142.4810	79.5795	14.36	0.0000
Education	−25.5200	6.6159	−3.86	0.0003
PopDensity	0.0079	0.0038	2.09	0.0408
pNonWhite	3.9917	0.6080	6.57	0.0000

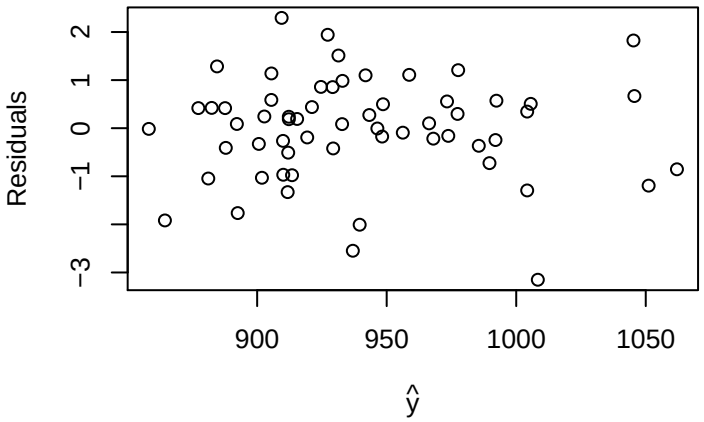
```
Residual standard error: 40.67 on 56 degrees of freedom
Multiple R-Squared: 0.5945,     Adjusted R-squared: 0.5727
F-statistic: 27.36 on 3 and 56 DF,  p-value: 4.997e-11
```

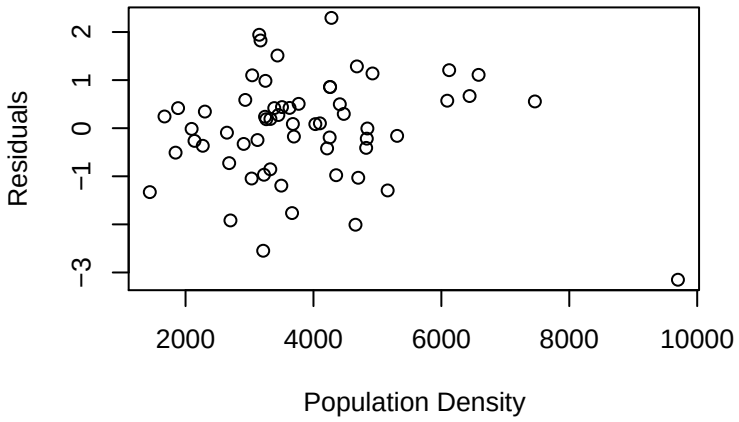

Mortality
Education
Population Density
percent Non−white
9.0
10.0
11.0
12.0
0
10
20
30
40
800
900
1000
1100
2000
4000
6000
8000
2000
6000
10000

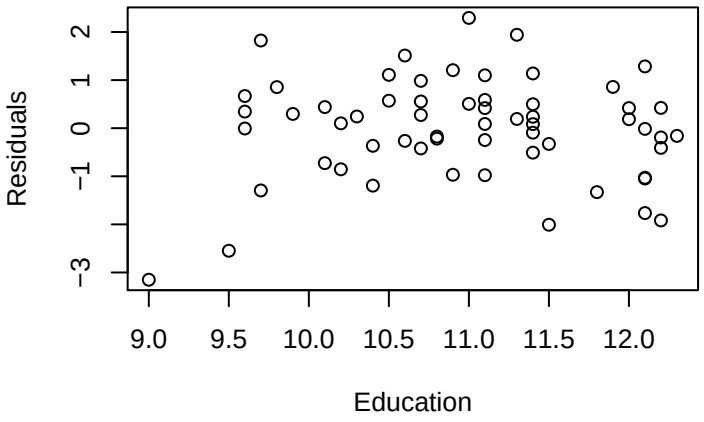

Normal Q−Q Plot
Sample Quantiles
−3
−1
1
−2
−1
0
1
2
Theoretical Quantiles

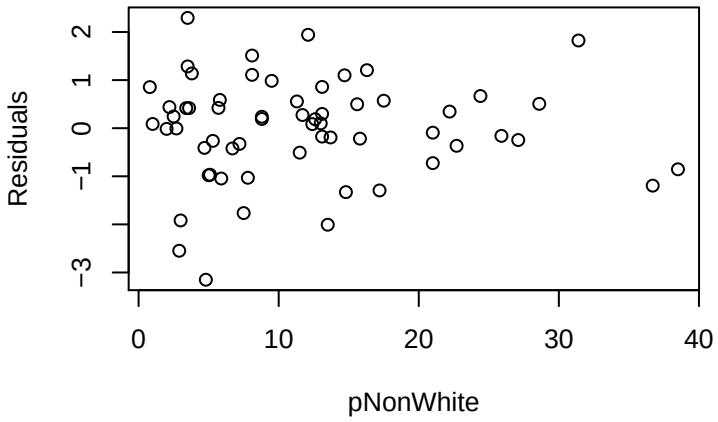

Histogram of res
Density
0.0
0.2
0.4
−3
−2
−1
0
1
2
res

Residual Plots

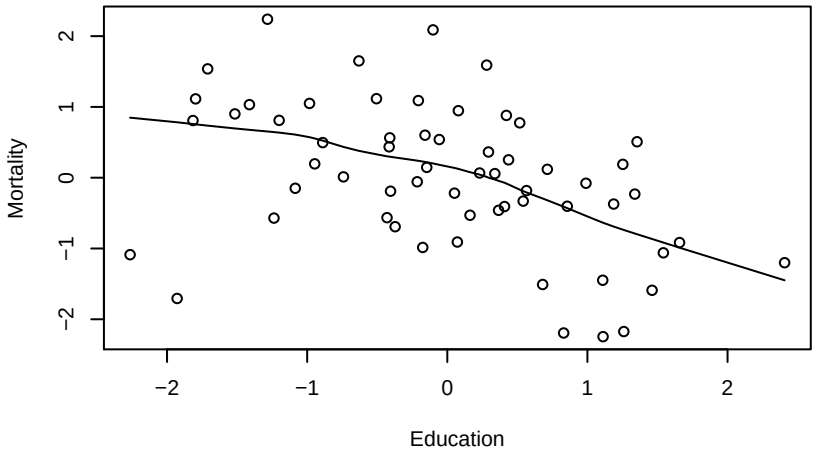


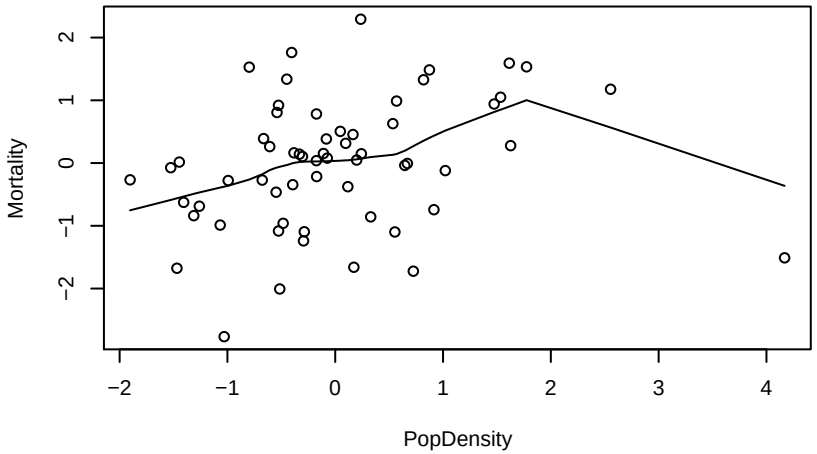


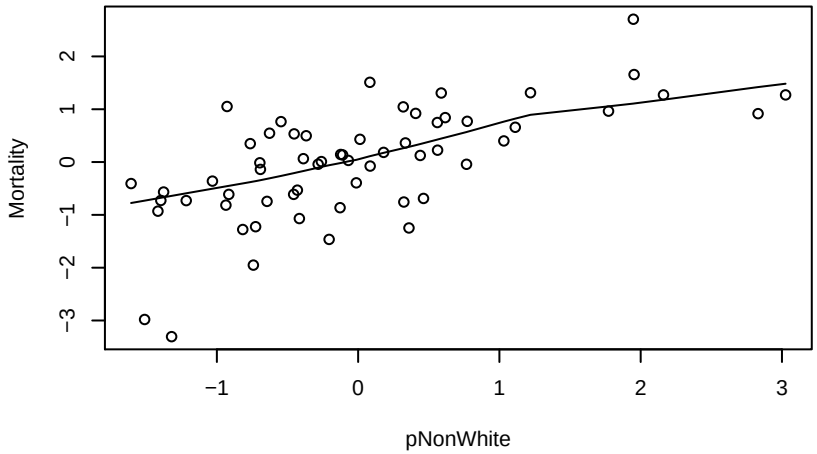




Partial regression plots







Identifying outliers and influential observations

An **outlier** is an extreme observation.

- Depending on their location in the predictor space, outliers can have severe effects on the regression model.

- We can use jackknife residuals to identify potential outliers. Any points that are greater than 3 or 4 standard deviations away from 0 may be considered potential outliers.

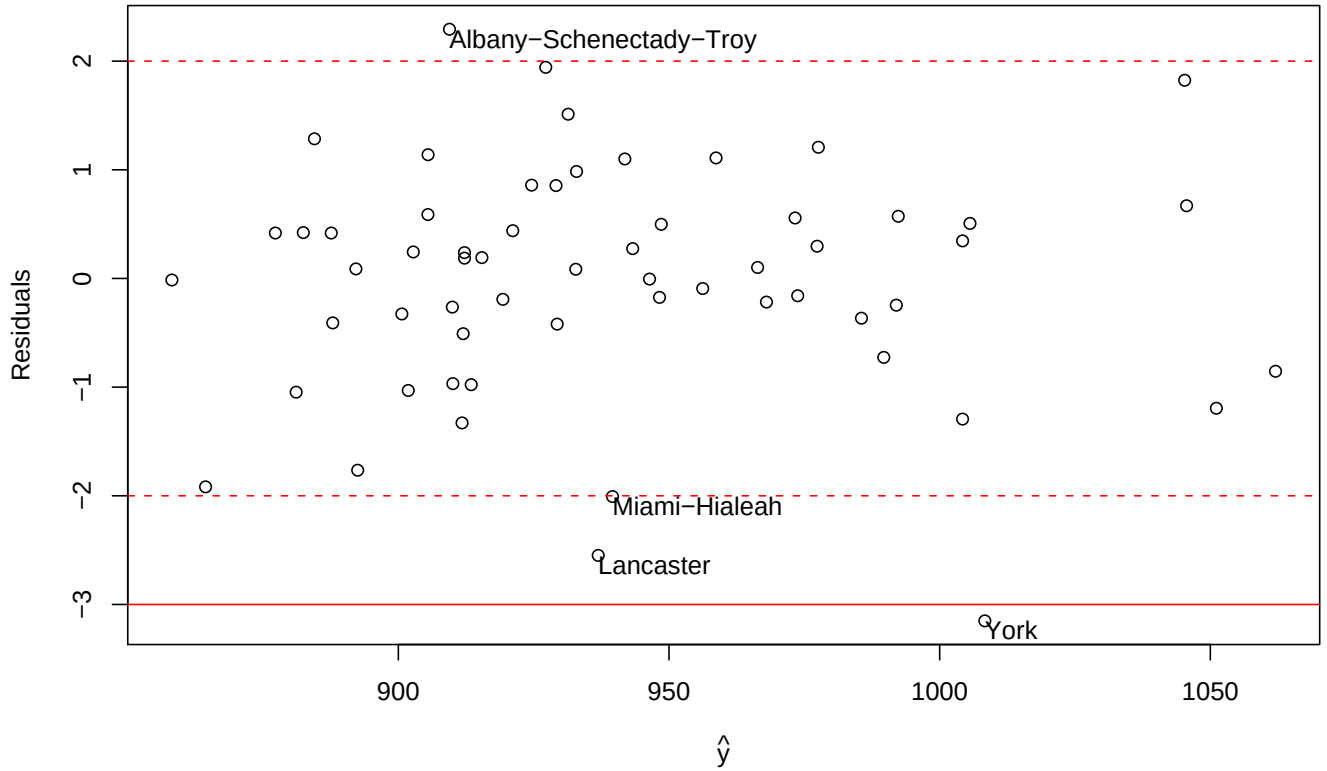
Possible explanations for an outlier

- "Bad" data that results from unusual but explainable events, eg - malfunction of measuring instrument, incorrect recording of data. In this case we should try to retrieve the correct value, but if that's not possible we may need to discard the data point.

- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disasterous to simply discard outliers.

- Poor sampling of observations in the tail of the distribution. This may be especially true if the outcome arises from a heavy-tailed distribution.

With a sample size of 60, we might expect 2 or 3 residuals to be further than 2 stand. dev. from 0 and none to be more than 3 stand. dev.



Identifying influential observations

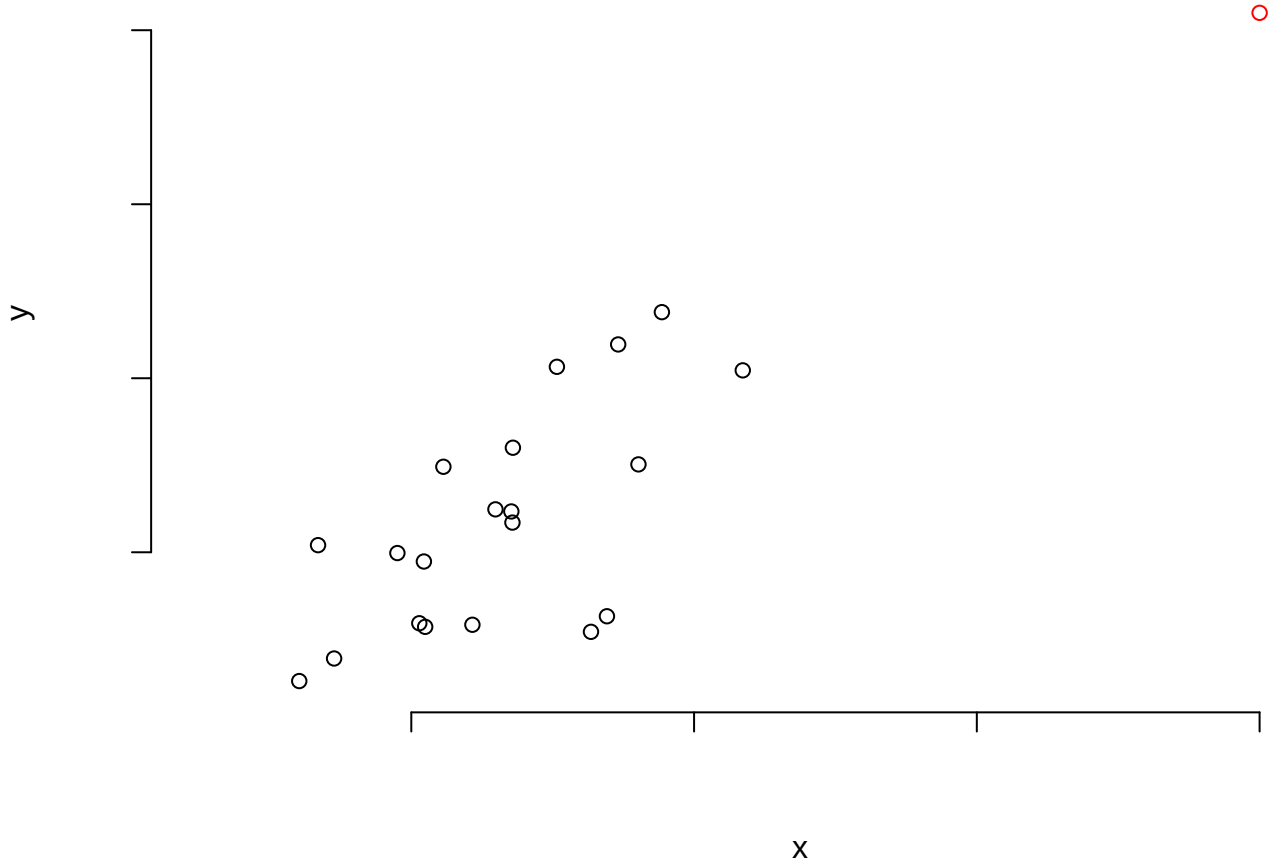
If we compute a sample mean, each observation contributes equally. This is not the case in the regression setting.

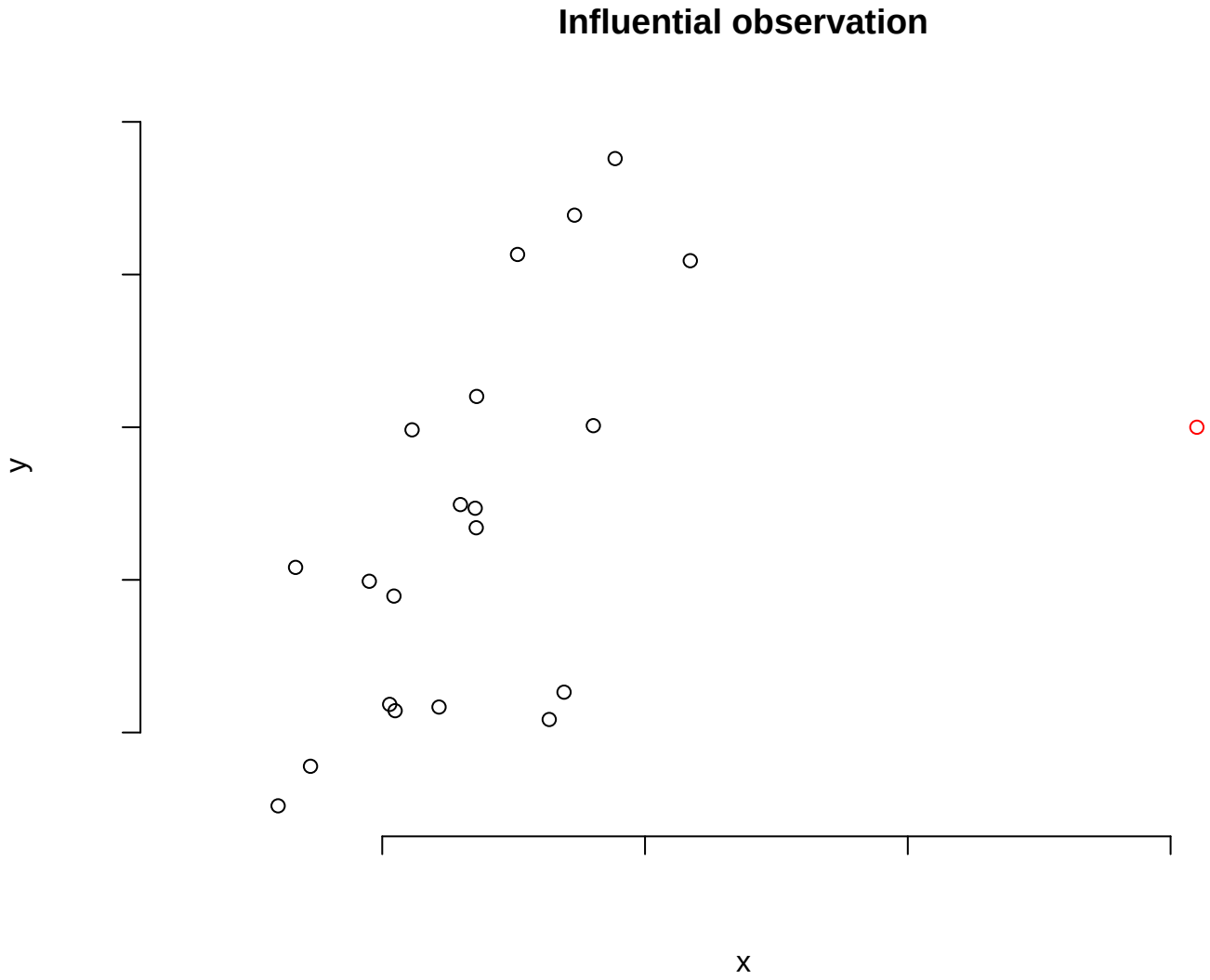
- Points that are remote in the predictor space may not influence the estimate of the regression coefficients but may influence other summary statistics, such as R^2 and the standard errors of the coefficients. These points are called **leverage** points.

- Points that have a noticeable effect on the regression coefficients are called **influence** points.

We would like to identify leverage and influence points and understand their affect on our model fits.

Leverage point




Influential observation
y
x

Leverage

The hat matrix,

$$H = X(X'X)^{-1}X'$$

plays an important role in identifying influential observations. The diagonal elements

$$h_i = x_i'(X'X)^{-1}x_i,$$

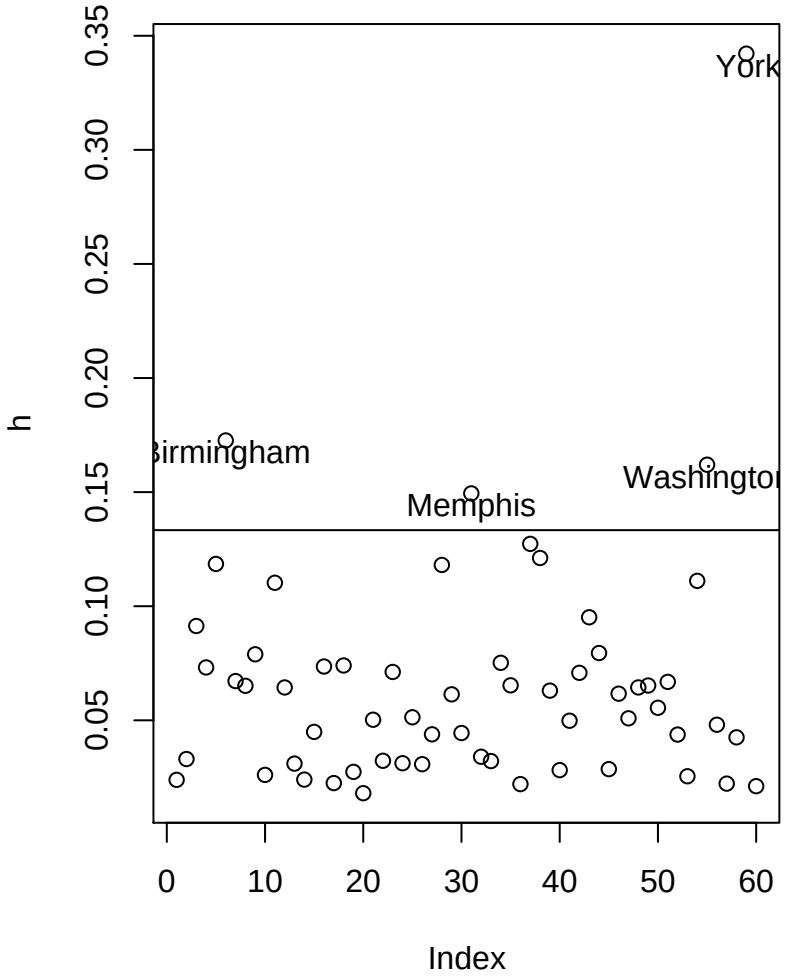
where x_i is the ith row of the X matrix, play an especially important role. h_i is a standardized measure of the distance of the X values for ith observation and the means of the X values for all n observations.

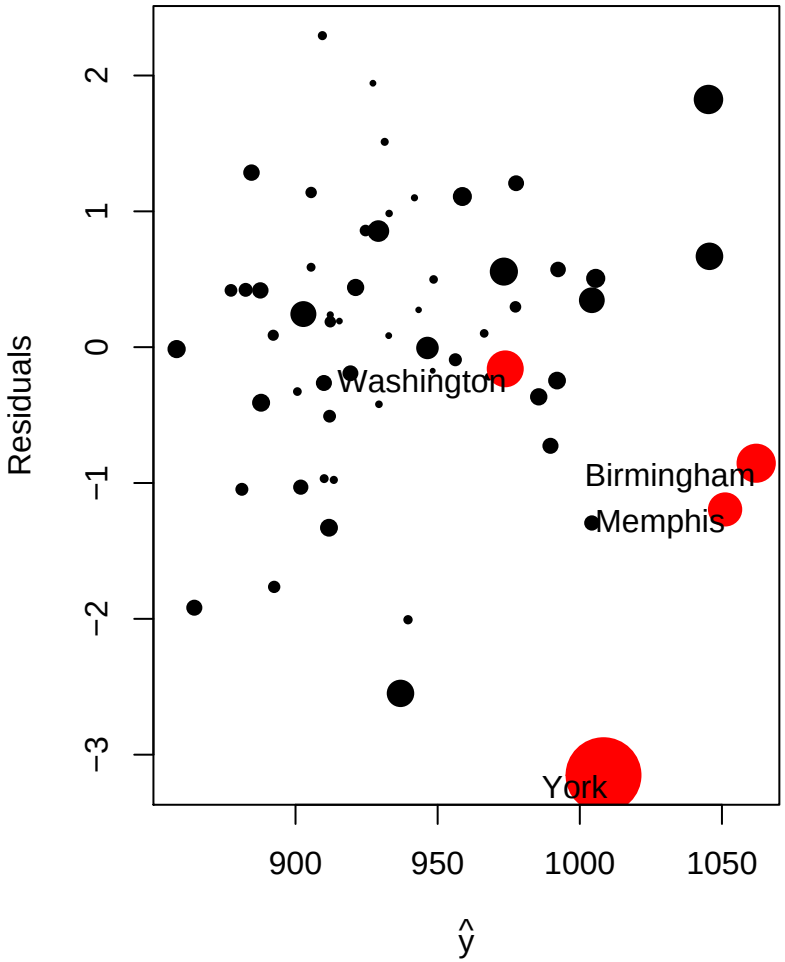
Also,

$$0 \leq h_i \leq 1 \quad \sum_{i=1}^{n} h_i = p + 1.$$

Thefore the average size of a hat diagonal is $\bar{h} = \frac{p+1}{n}$. Leverage values greater than $2\bar{h} = 2\frac{p+1}{n}$ are considered to be outlying values with regard to their X values and we would consider them leverage points.

Leverage for mortality example





Measures of Influence

Regression deletion diagnostics

- Cook's distance (Cook's D)
- DFFITS
- DFBETAS

Cook's D

Cook's D measures the influence of the ith observation on all n fitted values and is given by

$$D_i = \frac{(\hat{y} - y_{\hat{(i)}})'(\hat{y} - y_{\hat{(i)}})}{(p+1)MSE},$$

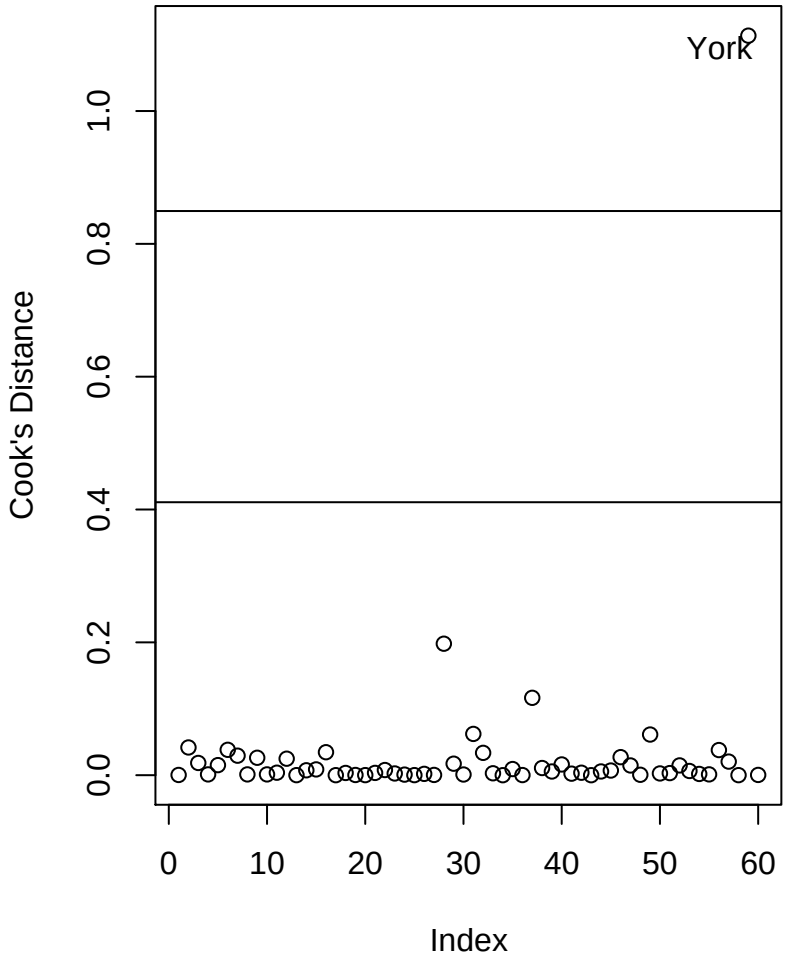
where $\hat{y}$ is the vector of fitted values when all n observations are included and $y_{\hat{(i)}}$ is the vector of fitted values when the ith observation is deleted. Cook's D can also be expressed as

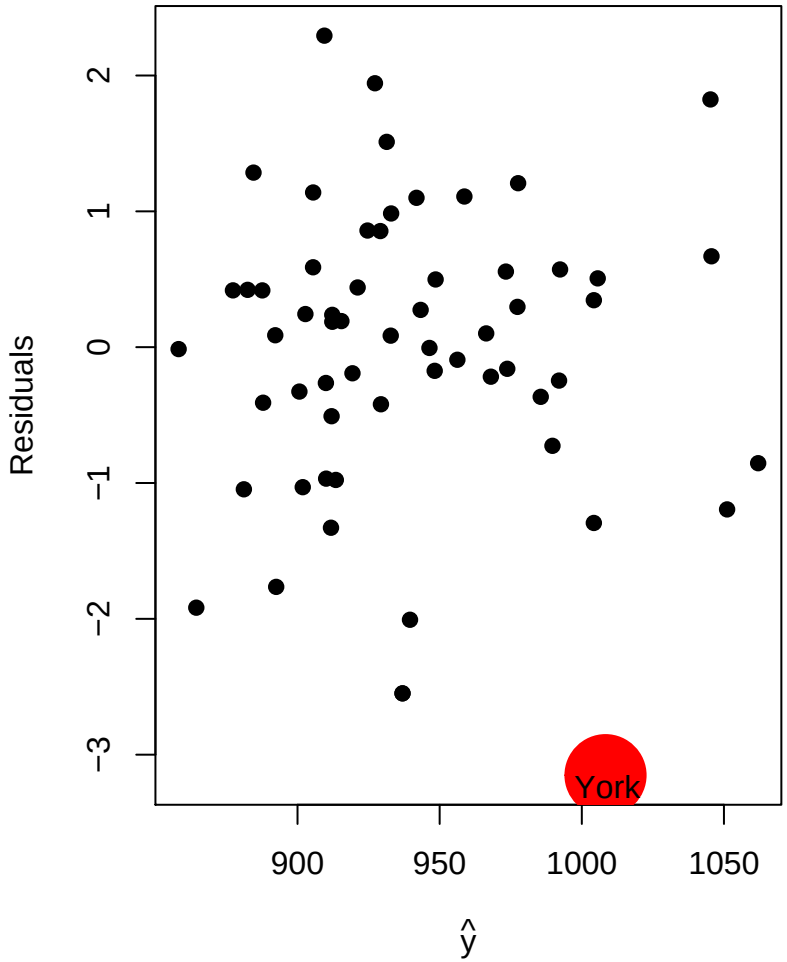
$$D_i = \frac{e_i^2}{(p+1)MSE}\left[\frac{h_i}{1-h_i}\right].$$

From this expression we see that D_i depends on both the size of the residual, e_i, and the leverage, h_i.

The magnitude of D_i is usually assessed by comparing it to $F_{p+1,n-p-1}$. If the percentile value is less than 10 or 20 % than the ith observation has little apparent influence on the fitted values. If the percentile value is greater than 50%, we conclude that the ith observation has significant effect on the fitted values.

Cook's D for mortality example





DFFITS

Cook's D measures the influence of the ith observation on all n fitted values of the outcome. In contrast, $DFFITS_i$ is a measure of the influence of the ith observation on the fitted value $\hat{y}_i$.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y_{(i)}}}{\sqrt{MSE_{(i)}h_i}},$$

where $\hat{y_{i(i)}}$ is the fitted value of y_i from the regression model fit with the ith observation deleted. The denominator is the estimated standard deviation of $\hat{y_{(i)}}$ and is based on the MSE calculated from the regression model fit when the ith observation is deleted. The resulting standardization represents the number of estimated standard deviations of $\hat{y}_i$ that the fitted value increases or decreases with the inclusion of the ith

observation in the model.

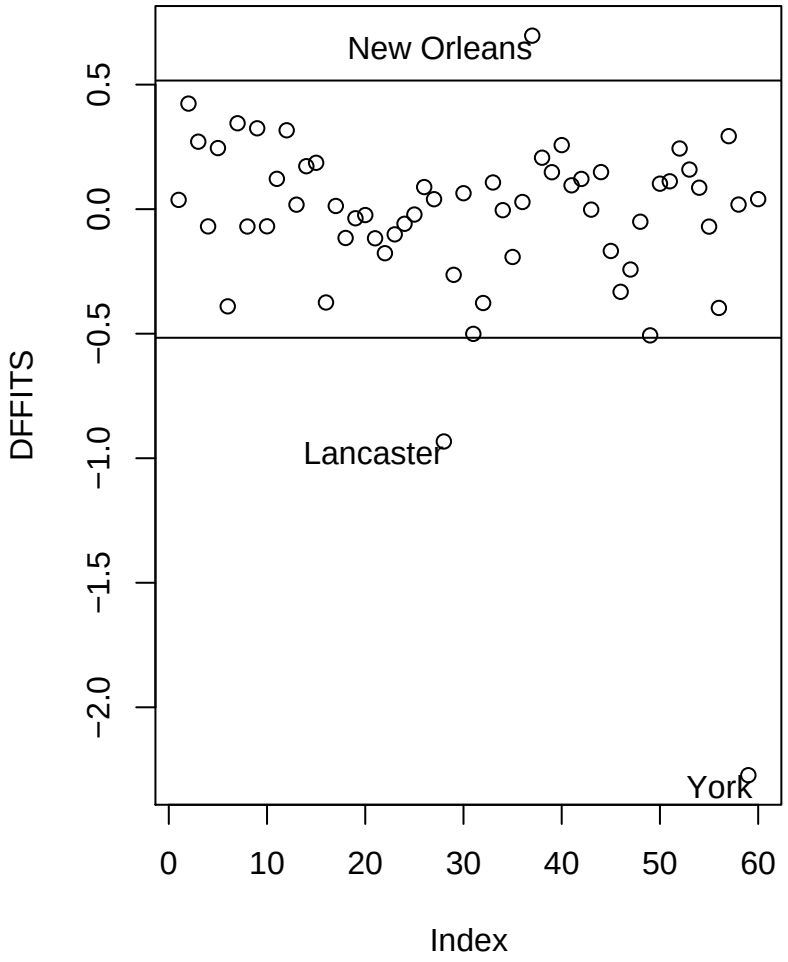
We can show that

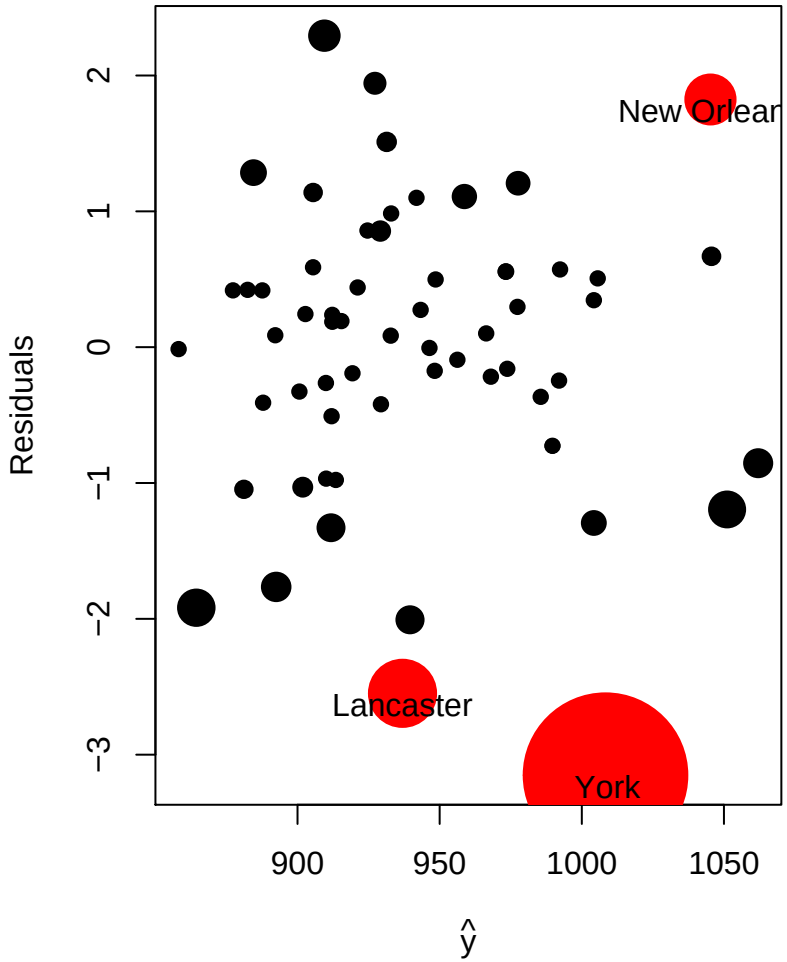
$$(DFFITS)_i = \left(\frac{h_i}{1-h_i}\right)^{1/2} r_{(-i)}$$

and can therefore be calculated without refitting the model n times.

Any observation with $|DFFITS_i| > 2\sqrt{(p+1)/n}$ warrants further investigation.

DFFITS for mortality example





DFBETAS

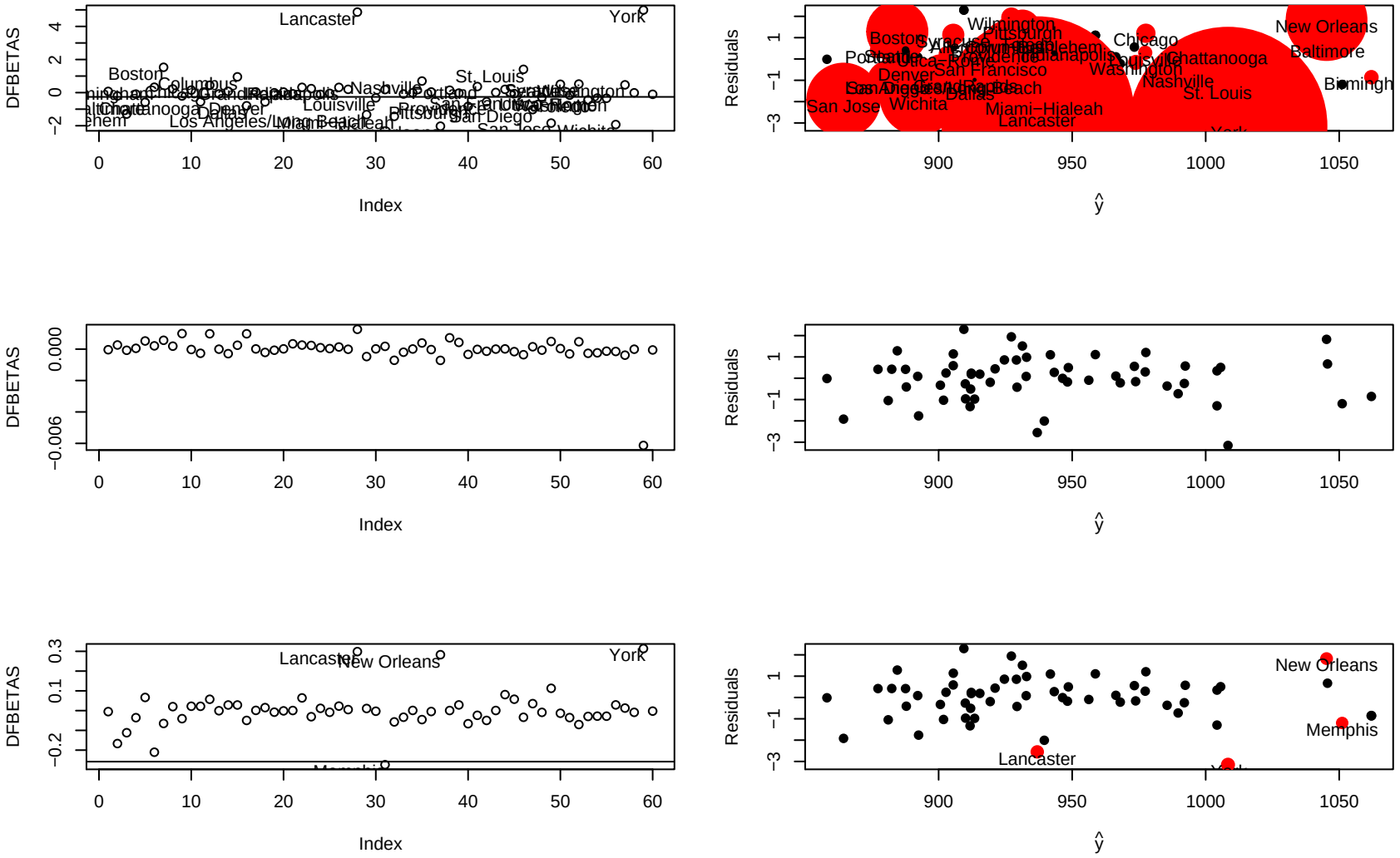
Both DFFITS and Cook's D measure an observation's influence on fitted values. Alternatively, we may be interested in an observation's influence on the coefficient estimates. We can get this by measuring the difference between the coefficient estimated ($\hat{\beta}_k$) with and without the ith observation ($\hat{\beta_{k(i)}}$) and divide this distance by an estimate of the standard error. This measures is $DFBETAS$

$$DFBETAS_{k,i} = \frac{\hat{\beta}_k - \hat{\beta_{k(i)}}}{\sqrt{C_{kk}MSE_{(i)}}}.$$

A large value of $DFBETAS_{k,i}$ is indicative of a large impact of the ith observation on the kth regression coefficient.

An observation is usually considered influential if the absolute value of $DFBETAS$ exceeds $2/\sqrt{n}$. A larger number may be considered for comparison in smaller data sets.

DFBETAS for mortality example



What if we fit the model without York?

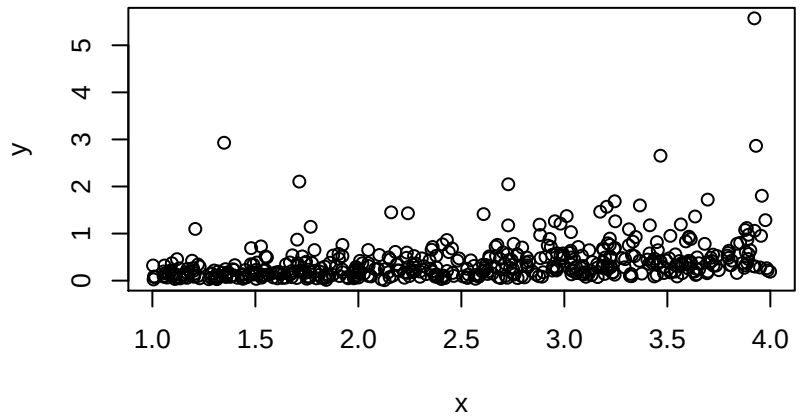
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1179.6235	74.8399	15.76	0.0000
Education	−30.5080	6.3449	−4.81	0.0000
PopDensity	0.0140	0.0040	3.50	0.0009
pNonWhite	3.6778	0.5733	6.41	0.0000

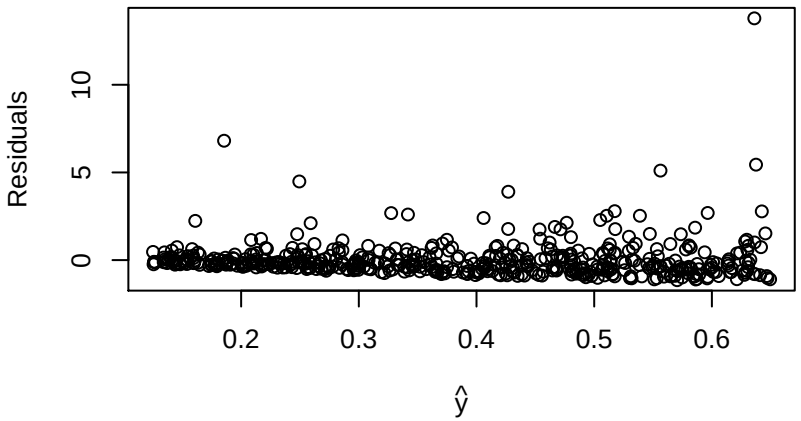
Variance stabilizing transformations

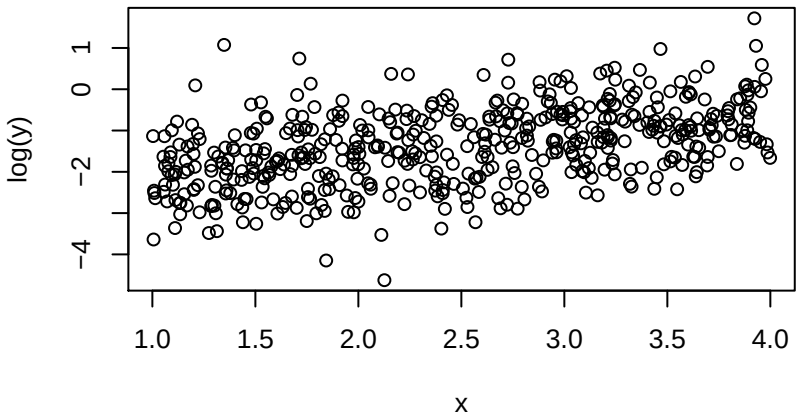
Non-constant variance can often be remedied using appropriate transformations. Ideally, we would choose the transformation based on some prior scientific knowledge, but this might not always be available.

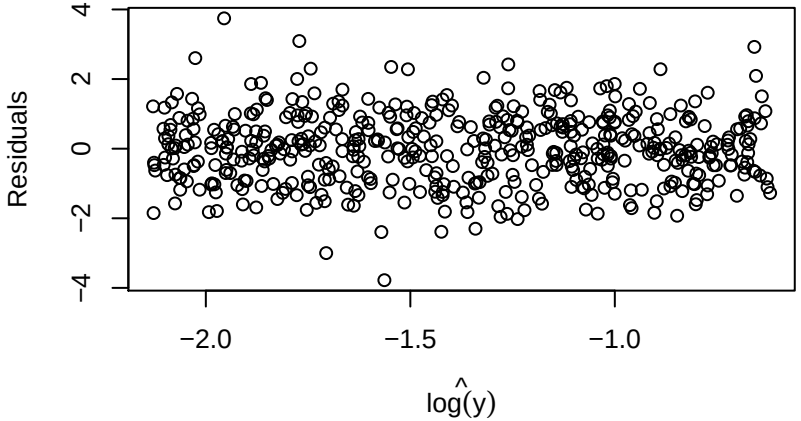
Commonly used variance stabilizing techniques

Relationship of σ^2 to $E[y]$	Transformation	comment
$\sigma^2 \propto$ constant	$y' = y$	no transformation
$\sigma^2 \propto E[y]$	$y' = \sqrt{y}$	Poisson data
$\sigma^2 \propto E[y](1 - E[y])$	$y' = sin^{-1}(\sqrt{y})$	binomial proportions,
$\sigma^2 \propto (E[y])^2$	$y' = \log(y)$	$y > 0$
$\sigma^2 \propto (E[y])^3$	$y' = y^{-1/2}$	$y > 0$
$\sigma^2 \propto (E[y])^4$	$y' = y^{-1}$	


y
0 1 2 3 4 5
1.0 1.5 2.0 2.5 3.0 3.5 4.0
x


Residuals
0 5 10
0.2 0.3 0.4 0.5 0.6
ŷ


log(y)
−4 −2 0 1
1.0 1.5 2.0 2.5 3.0 3.5 4.0
x


Residuals
−4 −2 0 2 4
−2.0 −1.5 −1.0
log(ŷ)