



Residuals, diagnostics, variation

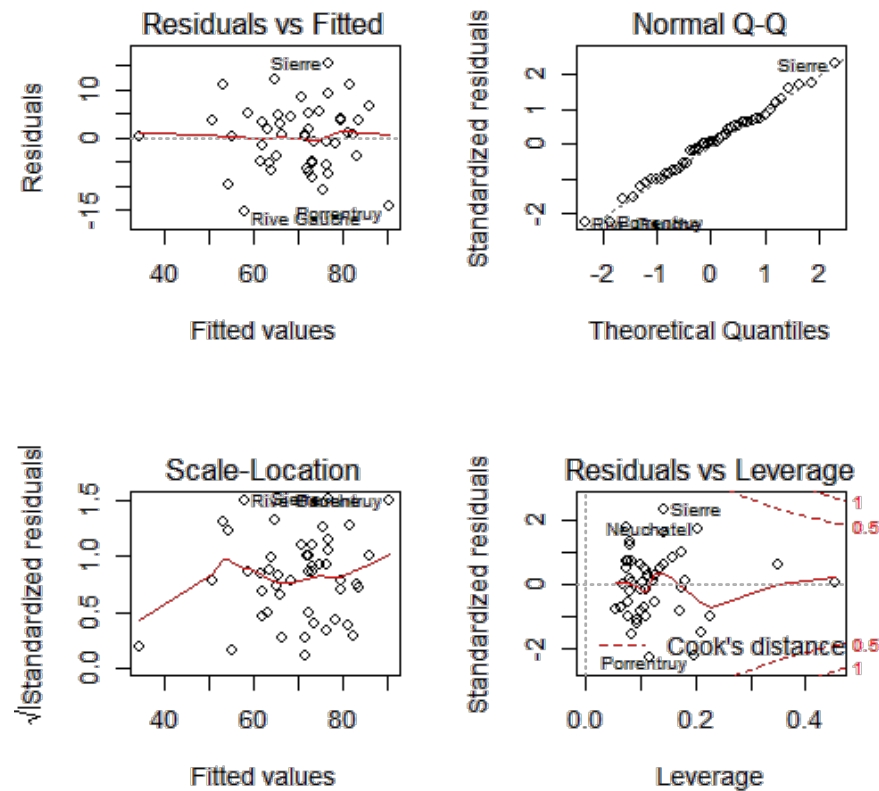
Regression

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

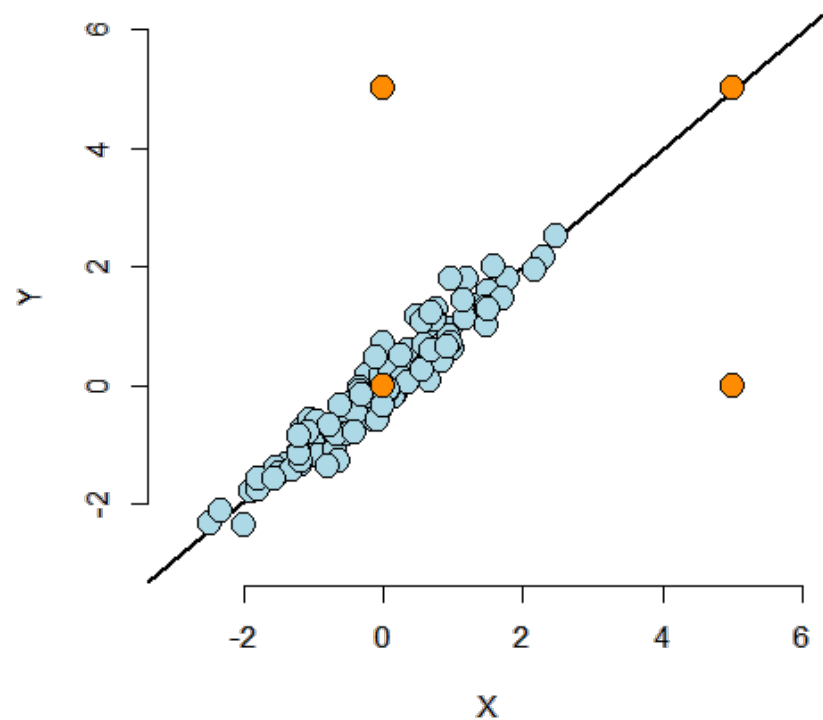
The linear model

- Specified as $Y_i = \sum_{k=1}^p X_{ik} \beta_j + \epsilon_i$
- We'll also assume here that $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$
- Define the residuals as $e_i = Y_i - \hat{Y}_i = Y_i - \sum_{k=1}^p X_{ik} \hat{\beta}_j$
- Our estimate of residual variation is $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$, the $n - p$ so that $E[\hat{\sigma}^2] = \sigma^2$

```
data(swiss); par(mfrow = c(2, 2))  
fit <- lm(Fertility ~ . , data = swiss); plot(fit)
```



Influential, high leverage and outlying points



Summary of the plot

Calling a point an outlier is vague.

- Outliers can be the result of spurious or real processes.
- Outliers can have varying degrees of influence.
- Outliers can conform to the regression relationship (i.e being marginally outlying in X or Y, but not outlying given the regression relationship).
 - Upper left hand point has low leverage, low influence, outliers in a way not conforming to the regression relationship.
 - Lower left hand point has low leverage, low influence and is not to be an outlier in any sense.
 - Upper right hand point has high leverage, but chooses not to exert it and thus would have low actual influence by conforming to the regression relationship of the other points.
 - Lower right hand point has high leverage and would exert it if it were included in the fit.

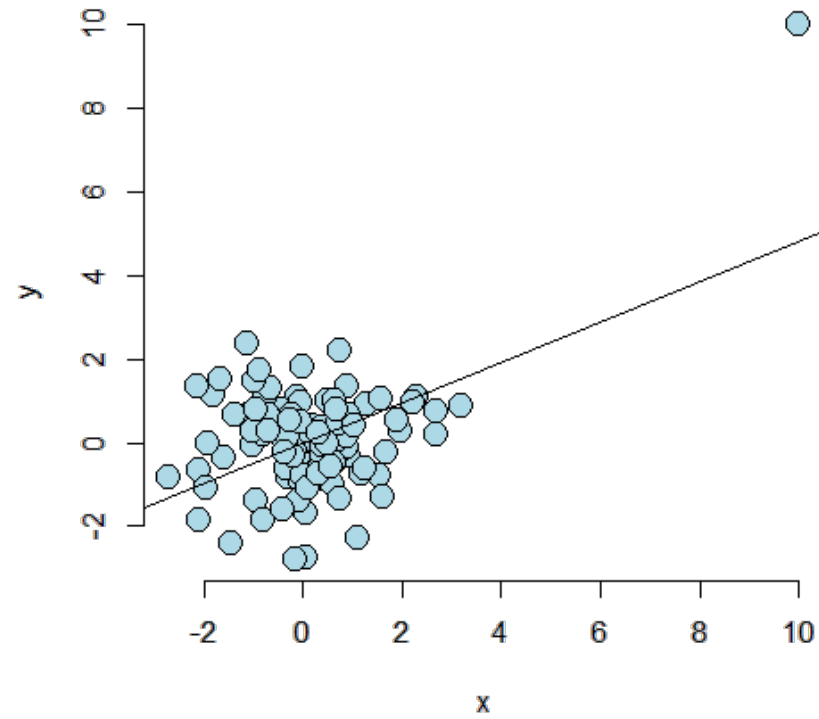
Influence measures

- Do `?influence.measures` to see the full suite of influence measures in stats. The measures include
 - `rstandard` - standardized residuals, residuals divided by their standard deviations)
 - `rstudent` - standardized residuals, residuals divided by their standard deviations, where the i^{th} data point was deleted in the calculation of the standard deviation for the residual to follow a t distribution
 - `hatvalues` - measures of leverage
 - `dffits` - change in the predicted response when the i^{th} point is deleted in fitting the model.
 - `dfbetas` - change in individual coefficients when the i^{th} point is deleted in fitting the model.
 - `cooks.distance` - overall change in the coefficients when the i^{th} point is deleted.
 - `resid` - returns the ordinary residuals
 - `resid(fit) / (1 - hatvalues(fit))` where `fit` is the linear model fit returns the PRESS residuals, i.e. the leave one out cross validation residuals - the difference in the response and the predicted response at data point i , where it was not included in the model fitting.

How do I use all of these things?

- Be wary of simplistic rules for diagnostic plots and measures. The use of these tools is context specific. It's better to understand what they are trying to accomplish and use them judiciously.
- Not all of the measures have meaningful absolute scales. You can look at them relative to the values across the data.
- They probe your data in different ways to diagnose different problems.
- Patterns in your residual plots generally indicate some poor aspect of model fit. These can include:
 - Heteroskedasticity (non constant variance).
 - Missing model terms.
 - Temporal patterns (plot residuals versus collection order).
- Residual QQ plots investigate normality of the errors.
- Leverage measures (hat values) can be useful for diagnosing data entry errors.
- Influence measures get to the bottom line, 'how does deleting or including this point impact a particular aspect of the model'.

Case 1



The code

```
n <- 100; x <- c(10, rnorm(n)); y <- c(10, c(rnorm(n)))  
plot(x, y, frame = FALSE, cex = 2, pch = 21, bg = "lightblue", col = "black")  
abline(lm(y ~ x))
```

- The point `c(10, 10)` has created a strong regression relationship where there shouldn't be one.

Showing a couple of the diagnostic values

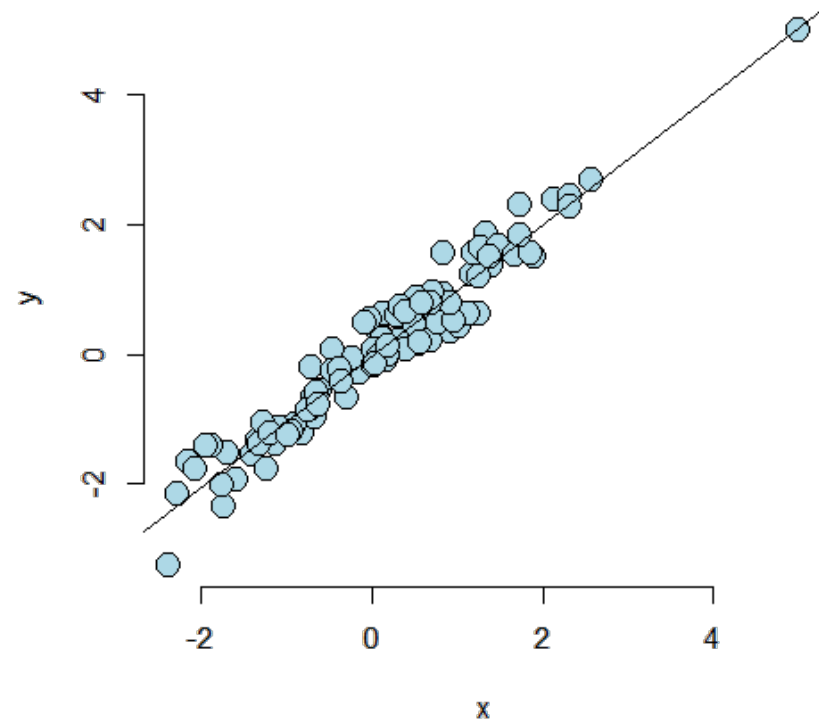
```
fit <- lm(y ~ x)
round(dfbetas(fit)[1 : 10, 2], 3)
```

1	2	3	4	5	6	7	8	9	10
6.007	-0.019	-0.007	0.014	-0.002	-0.083	-0.034	-0.045	-0.112	-0.008

```
round(hatvalues(fit)[1 : 10], 3)
```

1	2	3	4	5	6	7	8	9	10
0.445	0.010	0.011	0.011	0.030	0.017	0.012	0.033	0.021	0.010

Case 2



Looking at some of the diagnostics

```
round(dfbetas(fit2)[1 : 10, 2], 3)
```

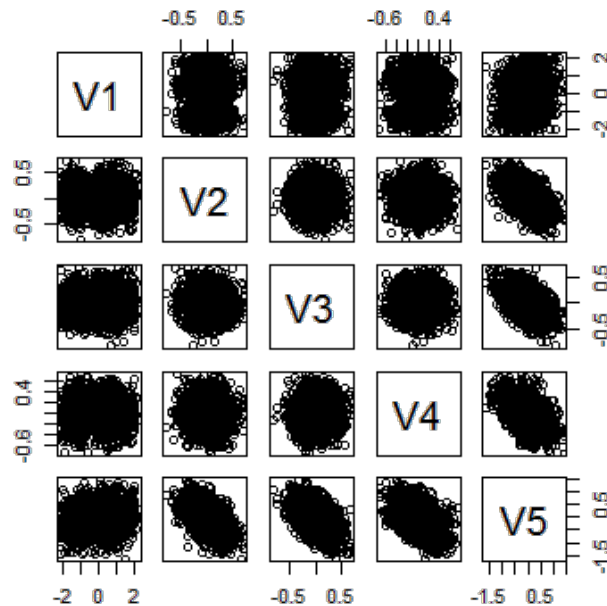
1	2	3	4	5	6	7	8	9	10
-0.072	-0.041	-0.007	0.012	0.008	-0.187	0.017	0.100	-0.059	0.035

```
round(hatvalues(fit2)[1 : 10], 3)
```

1	2	3	4	5	6	7	8	9	10
0.164	0.011	0.014	0.012	0.010	0.030	0.017	0.017	0.013	0.021

Example described by Stefanski TAS 2007 Vol 61.

```
## Don't everyone hit this server at once.  Read the paper first.  
dat <- read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/only_owl_files/only  
pairs(dat)
```



Got our P-values, should we bother to do a residual plot?

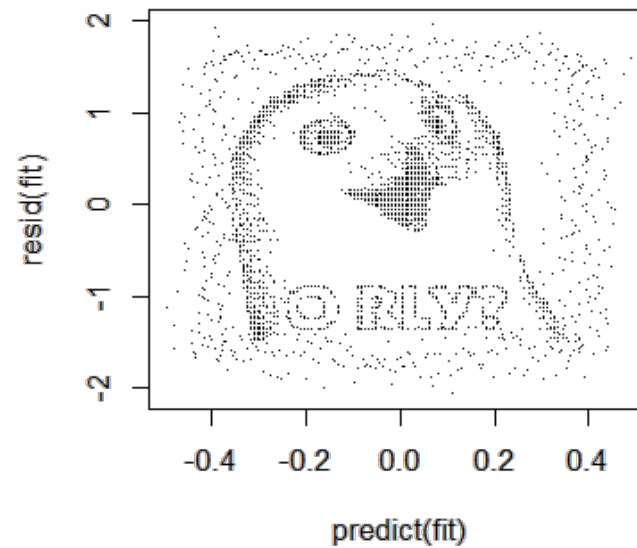
```
summary(lm(V1 ~ . -1, data = dat))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
V2	0.9856	0.12798	7.701	1.989e-14
V3	0.9715	0.12664	7.671	2.500e-14
V4	0.8606	0.11958	7.197	8.301e-13
V5	0.9267	0.08328	11.127	4.778e-28

Residual plot

P-values significant, O RLY?

```
fit <- lm(V1 ~ . - 1, data = dat); plot(predict(fit), resid(fit), pch = '.')
```



Back to the Swiss data

