# STAT22000 Autumn 2013 Lecture 6

Yibi Huang

October 14, 2013

Regression, Residuals, Outliers



## Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which "`lm`" stands for *linear model*.

```
> NEA = c(-94,-57,-29,135,143,151,245,355,392,473,486,535,571,
  580,620,690)
> fatgain = c(4.2, 3.0, 3.7, 2.7, 3.2, 3.6, 2.4, 1.3, 3.8, 1.7,
  1.6, 2.2, 1.0, 0.4, 2.3, 1.1)
> lm(fatgain ~ NEA)

Call:
lm(formula = fatgain ~ NEA)

Coefficients:
(Intercept)          NEA
   3.505123    -0.003441
```

Here you get the intercept to be 3.505 and slope to be $-0.003441$.



## Predicted Values and Residuals in R

It is better to save the model as an object.

```
> mymodel = lm(fatgain ~ NEA)
```

Then from the stored object `mymodel`, you can get the predicted values $\widehat{y}_i$ (also called the "fitted values"):

```
> mymodel$fit              # output omitted
```

and the residuals $e_i = y_i - \widehat{y}_i$:

```
> mymodel$res              # output omitted
```

Guess what we will get.

```
> fatgain - mymodel$fit - mymodel$res
```

How to add the regression line on the scatter plot?

```
> plot(NEA, fatgain)        # scatter plot
> abline(mymodel)           # add the regression line
```



Here is a more detailed output of the linear model

```
> summary(mymodel)
Call:
lm(formula = fatgain ~ NEA)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1091 -0.3904 -0.1039  0.4125  1.6439

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5051229  0.3036164  11.545 1.53e-08 ***
NEA         -0.0034415  0.0007414  -4.642 0.000381 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7399 on 14 degrees of freedom
Multiple R-squared:  0.6061,    Adjusted R-squared:  0.578
F-statistic: 21.55 on 1 and 14 DF,  p-value: 0.000381
```

We will get back to this summary in Chapter 10.



## Properties of Residuals

If predicted with a LS regression line, the residuals have the following properties

1. Residuals always **sum to zero**, $\sum_{i=1}^n e_i = 0$.
   - If the sum $> 0$, can you improve the prediction?
2. Residuals and the explanatory variable $x_i$'s have **zero correlation**.
   - If non-zero, the residuals can be predicted by $x_i$'s, not the best prediction.
   - Residuals are the part in the response that CANNOT be explained or predicted linearly by the explanatory variables.

```
> sum(mymodel$res)
[1] 6.938894e-17
> cor(NEA,l1$res)
[1] 5.786109e-17
```



## Proofs of the Two Properties of Residuals (Optional)

Recall the intercept $\widehat{a}$ and slope $\widehat{b}$ of the LS line are the $a$ and $b$ that minimize the sum of squares of errors

$$\sum_{i=1}^n (y_i - a - bx_i)^2.$$

Thus $\widehat{a}$ and $\widehat{b}$ satisfies the equations

$$\frac{d}{da}\sum_{i=1}^n (y_i - a - bx_i)^2 = -2\sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{d}{db}\sum_{i=1}^n (y_i - a - bx_i)^2 = -2\sum_{i=1}^n x_i(y_i - a - bx_i) = 0$$

i.e.,

$$\sum_{i=1}^n \underbrace{(y_i - \widehat{a} - \widehat{b}x_i)}_{=e_i} = 0 \quad \text{and} \quad \sum_{i=1}^n x_i\underbrace{(y_i - \widehat{a} - \widehat{b}x_i)}_{=e_i} = 0.$$

Thus,

$$\sum_{i=1}^n e_i = 0 \quad \text{and} \quad \sum_{i=1}^n x_i e_i = 0.$$

So far we have proved residuals sum to zero.

## Proof Cont'd

Recall the formula of the correlation coefficient

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}.$$

Thus the correlation coefficient of explanatory variable $\{x_1, x_2, \ldots, x_n\}$ and the residuals $\{e_1, e_2, \ldots, e_n\}$ is

$$r(x, e) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(e_i - \overline{e})}{s_x s_e}.$$

Thus to show $r(x, e) = 0$, we just need to show $\sum_{i=1}^{n}(x_i - \overline{x})(e_i - \overline{e}) = 0$.

$$\sum_{i=1}^{n}(x_i - \overline{x})(e_i - \overbrace{\overline{e}}^{=0}) = \sum_{i=1}^{n}(x_i - \overline{x})e_i = \underbrace{\sum_{i=1}^{n} x_i e_i}_{=0} - \overline{x}\underbrace{\sum_{i=1}^{n} e_i}_{=0} = 0$$



## Properties of Predicted Values

Observe the predicted value $\widehat{y}_i$'s are a linear transformation of the explanatory variable $x_i$'s:

$$\widehat{y}_i = \widehat{a} + \widehat{b}x_i.$$

- What is the mean of $\widehat{y}_i$'s? How is it related to the mean of of $x_i$'s? $\overline{\widehat{y}} = \widehat{a} + \widehat{b} \cdot \overline{x}$
  $= (\overline{y} - \widehat{b} \cdot \overline{x}) + \widehat{b} \cdot \overline{x}$ (since $\widehat{a} = \overline{y} - \widehat{b} \cdot \overline{x}$)
  $= \overline{y}$
  - The mean of the predicted value $\widehat{y}_i$'s is simply the mean of the observed $y_i$'s.
- How is the SD of $\widehat{y}_i$'s related to the SD of $x_i$'s?

$$s_{\widehat{y}} = |\widehat{b}| \cdot s_x = \left| r\frac{s_y}{s_x} \right| \cdot s_x = |r| \cdot s_y.$$



## Coefficient of Determination $R^2 = r^2$

So $r^2 = \dfrac{s_{\widehat{y}}^2}{s_y^2} = \dfrac{\text{Variance of } \{\widehat{y}_1, \ldots, \widehat{y}_n\}}{\text{Variance of } \{y_1, \ldots, y_n\}}$

= fraction of variation in $y_i$'s explained by $x_i$'s

- In view of this property, the square of correlation coefficient $r^2$, is also called the coefficient of determination, and is often denoted as $R^2$
- In the R output on Slide "Lecture 6 - 4," $R^2$ is shown as "`Multiple R-squared`"



| $y_i$ | = | $\widehat{y}_i$ | + | $e_i$ |
|---|---|---|---|---|
| (observed) | | (predicted) | | (residual) |

There is an important identity:

$$\mathrm{Var}(y) = \mathrm{Var}(\widehat{y}) + \mathrm{Var}(e).$$

This identity is nontrivial since in general, if $z_i = x_i + y_i$ for all $i = 1, 2, \ldots, n$, then

$$\mathrm{Var}(z) = \mathrm{Var}(x) + \mathrm{Var}(y) + r_{xy}\sqrt{\mathrm{Var}(x) \cdot \mathrm{Var}(y)}.$$

We can show that the residuals are uncorrelated with the predicted variables, $r_{\widehat{y},e} = 0$.

Since $\mathrm{Var}(\widehat{y}) = r^2\mathrm{Var}(y)$, we have $\mathrm{Var}(e) = (1 - r^2)\mathrm{Var}(y)$, i.e.,

$$\frac{\mathrm{Var}(e)}{\mathrm{Var}(y)} = \frac{\text{Variance of residuals}}{\text{Variance of responses}} = 1 - r^2$$

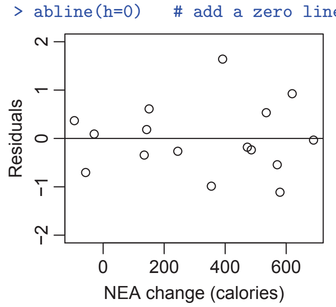


## Residual Plots — a Diagnostic Tool for Regression Model

A **residual plot** is a scatterplot of the residuals $e_i$ vs. the explanatory variable $x_i$. It is a *diagnostic tool* for the adequacy of a regression model.

E.g. here is the residual plot of the fat gain and NEA example.

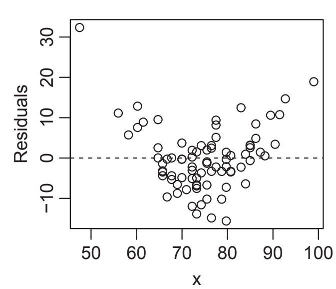
```
> plot(NEA,mymodel$res,xlab="NEA change (calories)",
   ylab="Residuals (kg)",ylim=c(-2,2))
> abline(h=0)   # add a zero line
```




A good residual plot appears "no pattern."
What does it mean by "pattern"?
Let's look at a few examples.



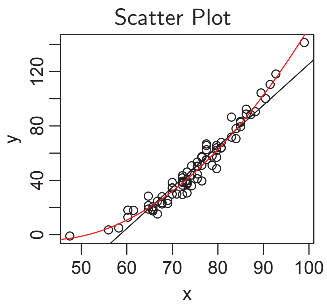
## Example 1

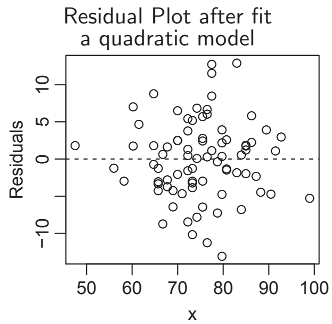



Based on the residual plot above, can you find ways to improve the prediction?

Zero correlation ≠ No association
It can be a non-linear association.

## Example 1 (Cont'd)

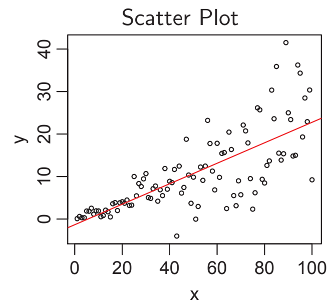


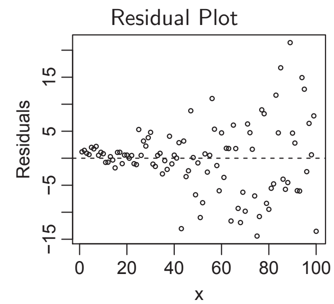




The straight line is the
LS regression line



## Example 2

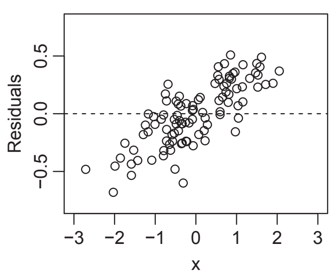



Can the linear prediction be improved in anyway?

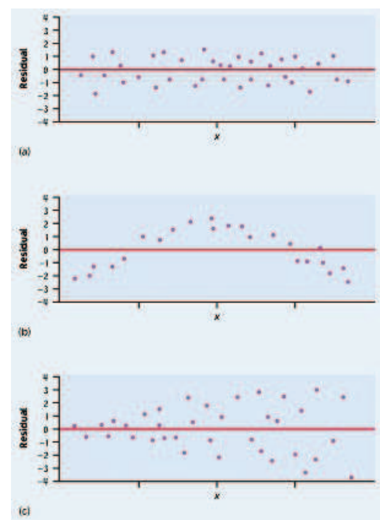


## Example 3




Can the linear prediction be improved in anyway?






(a) Residuals randomly scatter around the zero line — good!

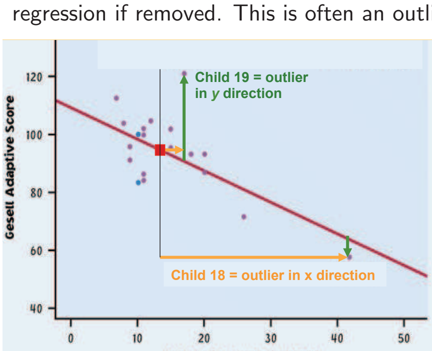
(b) Curved pattern — means the relationship you are looking at is not linear.

(c) A change in variability across a plot — predictions made in areas of larger variability will not be as good. May try weight least-square method or transforming the response.



## Outliers and Influential Points

**Outlier**: observation that lies outside the overall pattern of observations.
**Influential points**: observation that markedly changes the regression if removed. This is often an outlier on the $x$-axis.

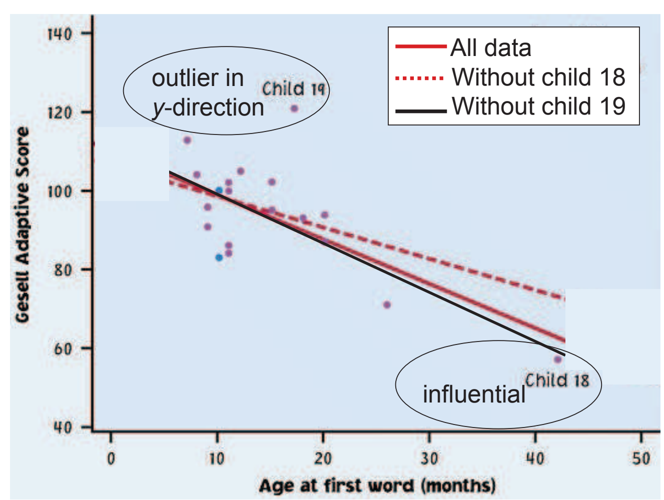



Child 19 is an outlier of the relationship.

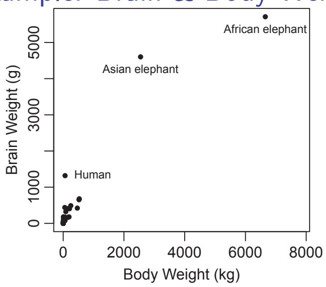
Child 18 is only an outlier in the $x$ direction and thus might be an influential point.



Are these points influential?



## Example: Brain & Body Weights for Mammals




The scatter plot shows the brain and body weights for 62 species of land mammals.

Large $r = 0.934$, but this is suspicious.

At least two influential points: African elephant and Asian elephant

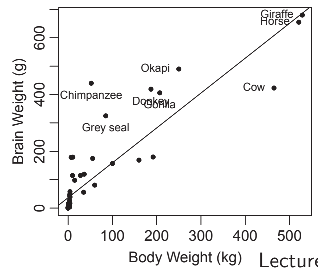
```
> mammals = read.table("mammals.txt",header=T)
> attach(mammals)
> cor(body,brain)
[1] 0.9341638
> cor(body[brain<2000], brain[brain<2000])  # exclude both elephants
[1] 0.6505592
> cor(body[brain<1000],brain[brain<1000]) # exclude 2 elephants & human
[1] 0.8884084
```

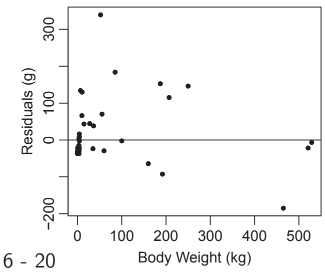


## How to Exclude Points In R?

How to exclude the 2 elephants and human in regression?

```
> myline1 = lm(brain[brain<1000] ~ body[brain<1000])
> plot(body[brain<1000],brain[brain<1000],pch=20,
  xlab="Body Weight (kg)", ylab="Brain Weight (g)")
> abline(myline1)              # add the regression line
> # Residual plot
> plot(body[brain<1000],myline1$res,pch=20,
  xlab="Body Weight (kg)", ylab="Residuals (g)")
> abline(h=0)                  # add a zero line
```






The equation for the LS regression in the previous slide is

```
> myline1
       (Intercept)   body[brain < 1000]
            36.572                1.228
```

i.e.,

$$\text{predicted brain weight} = 36.6\text{g} + 1.23 \times (\text{body weight in kg}).$$

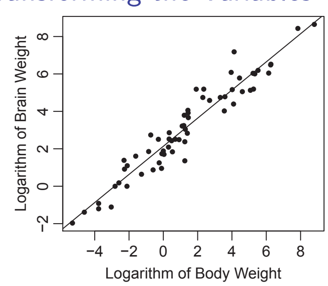
Hence the predicted brain weights are at least 36.6 g for all mammals. However, 35 out of 62 mammals in the data set have brain weights far below 36.6g:

```
> sort(brain[brain < 36])
 [1]  0.14  0.25  0.30  0.33  0.40  1.00  1.00  1.20  1.90  2.40
[11]  2.50  2.60  3.00  3.50  3.90  4.00  5.00  5.50  5.70  6.30
[21]  6.40  6.60  8.10 10.80 11.40 12.10 12.30 12.30 12.50 15.50
[31] 17.00 17.50 21.00 25.00 25.60
```

A prediction error of 10 gram is small for cows, but huge for mouses with brain weight $< 1$ gram.
For this data set, the absolute size of errors is not important.
We care more about the relative size of error: $\dfrac{\text{error}}{\text{brain weight}}$.



## Transforming the Variables




After taking log of both brain weights and body weights, the pattern is linear, with $r = 0.96$ (including elephants and human.)

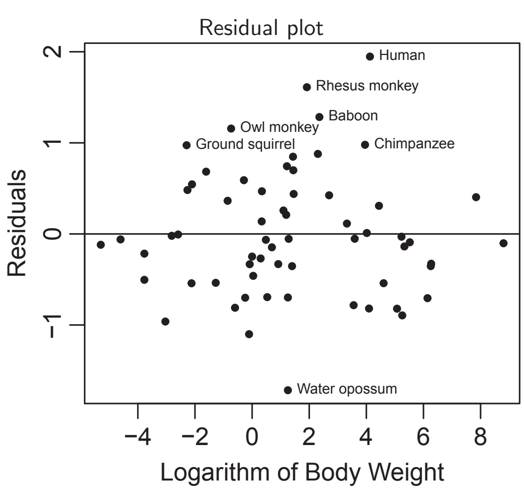
The vertical scatter is homogenous.

No influential points or outliers now.

```
> cor(log(body),log(brain))
[1] 0.9595748
> myline2 = lm(log(brain) ~ log(body))
> plot(log(body),log(brain),pch=20,
  xlab="Logarithm of Body Weight", ylab="Logarithm of Brain Weight")
> abline(myline2)
```

Sometimes transforming the variables can solve the problems of outliers or non-homogeneous scattering.








## Interpretation of the Log transformed Model

The LS regression equation in log scale is

```
> myline2
Call: lm(formula = log(brain) ~ log(body))

Coefficients:
(Intercept)    log(body)
     2.1348       0.7517
```

i.e.,

$$\text{predicted log brain weight} = 2.135 + 0.75 \times (\text{log body weight}),$$

or

$$\text{log brain weight} = 2.135 + 0.75 \times (\text{log body weight}) + \text{residual}.$$

or

$$\begin{aligned}\text{brain weight} &= e^{2.135} \times (\text{body weight})^{0.75} \times e^{\text{residual}} \\ &= 8.455 \times (\text{body weight})^{0.75} \times e^{\text{residual}}\end{aligned}$$

Observe that the error term is *multiplicative*, not *additive*.