Stat 104: Quantitative Methods for Economists

Class 34: Regression Diagnostics

Curious Theoretical Discussion

30) Suppose the assumptions of the linear regression model hold and as in class, the least squares estimates are denoted b_0b0 and b_1b1. Define the following quantities:

- quantity A = $\sum (y_i - b_0 - b_1 x_i)^2$
- quantity B = $\sum (y_i - \beta_0 - \beta_1 x_i)^2$

a) In general, quantity A is less than quantity B
b) In general, the quantities are equal
c) In general, quantity B is less than quantity

Interpret the Output

Do we need X in the model? Is β = 0 or not?

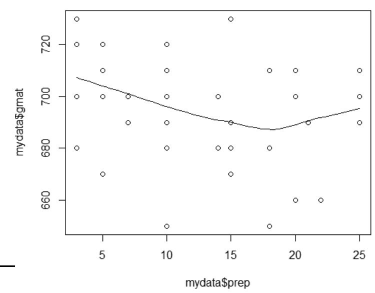
```
> fit=lm(mydata$distance~mydata$age)
> summary(fit)

Call:
lm(formula = mydata$distance ~ mydata$age)

Residuals:
   Min     1Q Median     3Q    Max
-78.23 -41.71   7.65  33.55 108.83

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  576.682     23.471   24.57   < 2e-16 ***
mydata$age    -3.007      0.424   -7.09 0.0000001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.8 on 28 degrees of freedom
Multiple R-squared:  0.642,     Adjusted R-squared:  0.629
F-statistic: 50.2 on 1 and 28 DF,  p-value: 0.000000104

> confint(fit)
                2.5 %   97.5 %
(Intercept) 528.6040 624.7599
mydata$age   -3.8761  -2.1376
```

| t | = 7 > 1.96 - reject : do need X

p-value ~ 0 < 0.05 - reject : do need X

GMAT and Number of Prep Days

- These NYU students study a lot(!). Not.



Interpret: the power of studying

```
> fit=lm(mydata$gmat~mydata$prep)
> summary(fit)

Call:
lm(formula = mydata$gmat ~ mydata$prep)

Residuals:
   Min     1Q Median     3Q    Max
-46.51 -12.60   2.47  13.68  36.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   703.31       6.58  106.87   <2e-16 ***
mydata$prep    -0.68       0.44   -1.54     0.13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20 on 36 degrees of freedom
Multiple R-squared:  0.062,     Adjusted R-squared:  0.036
F-statistic: 2.38 on 1 and 36 DF,  p-value: 0.132

> confint(fit)
                2.5 %   97.5 %
(Intercept) 689.9623 716.6549
mydata$prep  -1.5729   0.2137
```

| t | = -1.54 < 1.96 - failed to reject

p-value > 0.05 - failed to reject

0 is in the conf int - fail to reject - do not need X

Things you should know

- Be comfortable examining regression output and determining if there is a significant relationship between x and y.
- Confidence intervals for the regression parameters
- Hypothesis tests for the regression parameters

Baby Regression Diagnostics

We ***assume*** the following model holds:

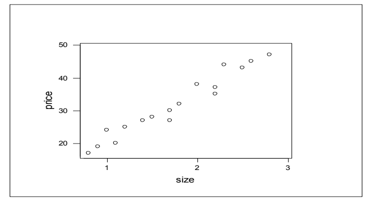
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Given X's

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad \textit{independent}$$

Given data



We (**hope, assume**) we see a linear pattern *and* a level of variation about the line.

Our model is designed to capture these two features of the data.

In practice we need to *check our assumption* that the model captures the important features of the data.

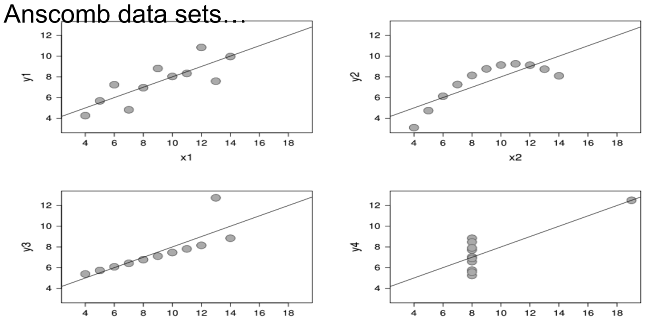
Is the model a good way to describe the data??

This is called model checking, and is done using the residuals from the regression.

Why do we have to check our model?

- All estimates, intervals, and hypothesis tests have been developed assuming that the model is correct.
- If the model is incorrect, then the formulas and methods we use are at risk of being incorrect.

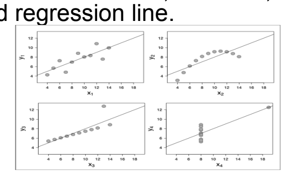
To drive this point home, let's look at the "famous" Anscomb data sets…



Anscombe's Quartet

- Francis Anscombe, "Graphs in Statistical Analysis". *American Statistician,* 1973
- **Identical** in common summary statistics: mean, variance, (Pearson) correlation, estimated regression line.

Property	Value
Mean of x in each case	9.0
Variance of x in each case	11.0
Mean of y in each case	7.5
Variance of y in each case	4.12
Correlation between x and y in each case	0.816
Linear regression line in each case	$y = 3 + 0.5x$



- Beware of not visualizing your data!
- Read more about Anscombe's Quartet Data here: http://www.automated-trading-system.com/word-of-caution-on-statistics/

i.e. if you fit a model, you'll see the same linear equation

Data Set 1

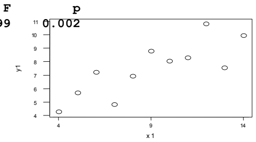
y1 = 3.00 + 0.500 x1

Predictor	Coef	SE Coef	T	P
Constant	3.000	1.125	2.67	0.026
x1	0.5001	0.1179	4.24	0.002

s = 1.237   R-sq = 66.7%   R-sq(adj) = 62.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	27.510	27.510	17.99	0.002
Error	9	13.763	1.529		
Total	10	41.273			



Data Set 2

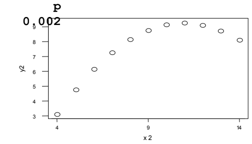
y2 = 3.00 + 0.500 x2

Predictor	Coef	SE Coef	T	P
Constant	3.001	1.125	2.67	0.026
x2	0.5000	0.1180	4.24	0.002

s = 1.237   R-sq = 66.6%   R-sq(adj) = 62.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	27.500	27.500	17.97	0.002
Error	9	13.776	1.531		
Total	10	41.276			



Data Set 3

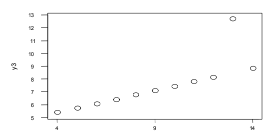
y3 = 3.00 + 0.500 x3

Predictor	Coef	SE Coef	T	P
Constant	3.002	1.124	2.67	0.026
x3	0.4997	0.1179	4.24	0.002

s = 1.236   R-sq = 66.6%   R-sq(adj) = 62.9%

Analysis of Variance

SOURCE	DF	SS	MS
Regression	1	27.470	27.470
Error	9	13.756	1.528
Total	10	41.226	



Data Set 4

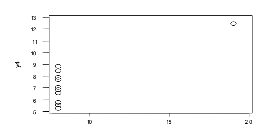
y4 = 3.00 + 0.500 x4

Predictor	Coef	SE Coef	T	P
Constant	3.002	1.124	2.67	0.026
x4	0.4999	0.1178	4.24	0.002

s = 1.236   R-sq = 66.7%   R-sq(adj) = 63.0%

Analysis of Variance

SOURCE	DF	SS	MS	
Regression	1	27.490	27.490	1
Error	9	13.742	1.527	
Total	10	41.232		



Anscomb Conclusion ?

- No data is bad; it just might not meet your assumptions. The data could simply be naughty.
- If your assumptions aren't met, the computer output might appear perfectly reasonable, but in reality be uninterpretable.

Residuals and Their Plots

- All of the assumptions of the model are really statements about the regression error terms (ε)

e by themselves are dependent on units and that's an issue.
If we divide by s, we get rid of the units

- How can we test whether the data supports these assumptions if we cannot observe the errors directly? We rely on diagnostics that use basic *least squares residuals*

$$e_i = Y_i - \hat{Y}_i$$

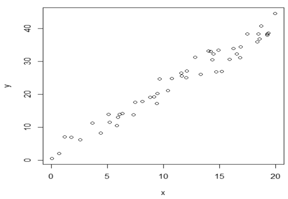
- We pretend as though the least squares residuals are the same as the true regression errors… with some limitations.
- Sometimes we use **Standardized Residuals** for convenience.
- $$\boxed{r_i = \frac{e_i}{s_e} \approx \frac{\varepsilon_i}{\sigma} \sim N(0,1)}$$

(why are these useful ?)

95% of time in -1.96 to 1.96

When things are right

Consider the data:



this plot looks like the kind of data our model is meant to describe.

Always plot Y vs X!

As a further check we examine the residuals.

Obtaining Residuals in R

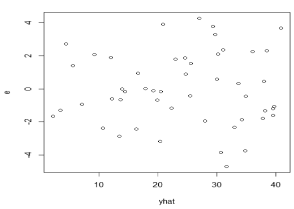
- We need the residuals, fitted values and standardized residuals

```
> fit=lm(y~x)
> e=residuals(fit)
> yhat=fitted(fit)
> sres=rstudent(fit)
```


We are looking for blobs since no relation is suppsed to be there between the two

Plot residuals versus Yhat

```
plot(yhat,e)
```



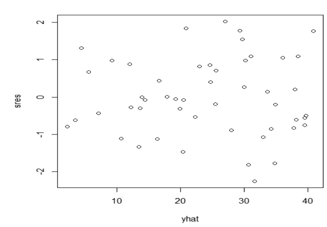
This is the way a residual plot looks when the model fits the data:

No obvious pattern!!!!!

resids unrelated to X!!!!!!

$$Y = \hat{Y} + e$$

(or) Plot standardized residuals vs Yhat



no obvious pattern!!!!!

resids unrelated to X!!!!!!

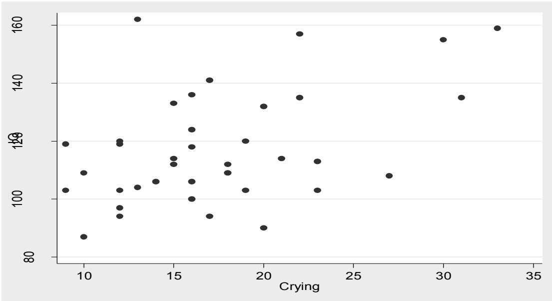
~~standardized resids between -2 and +2!!!!!!~~

plots are exactly the same, just the units differ.
both plots should be random blobs

Example: Crying Babies

- Babies who cry a lot may be more easily stimulated than other babies, and this may be an indication of higher IQ. Karelitz, et al. (1964) studied the association between IQ and crying frequency with 37 babies.
- The researchers caused the babies to cry by snapping a rubber band on the sole of their foot (bastards...).
- They recorded the frequency of cries as the number of peak cries (example: WAAAHHHHH-WAAAAHHHH is two peaks) in the most active 20 seconds of crying. Three years later, they measured the babies' IQs.

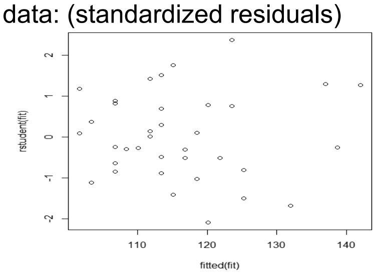
The data



Fitting a line

```
> summary(fit)

Call:
lm(formula = iq ~ crying)

Residuals:
    Min      1Q  Median      3Q     Max
-30.192  -9.791  -3.619  11.808  33.458

Coefficients:
            Estimate Std. Error t value           Pr(>|t|)
(Intercept)  86.6898     7.9650  10.884 0.000000000000883 ***
crying        1.6751     0.4313   3.884          0.000436 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 35 degrees of freedom
Multiple R-squared:  0.3012,    Adjusted R-squared:  0.2812
F-statistic: 15.09 on 1 and 35 DF,  p-value: 0.000436
```


Baby crying data: (standardized residuals)



`plot(fitted(fit),rstudent(fit))`

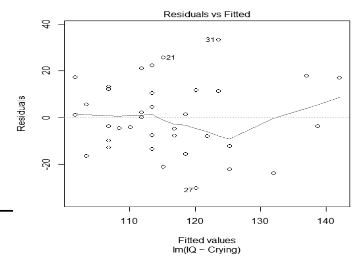
no obvious pattern!!!!!

resids unrelated to X!!!!!!

standardized resids between -2 and +2!!!!!! (uh, how do we standardize ?)

Automatic Residual Plot

- The R command `plot(fit,which=1)` will also give a basic residual plot



Finally, consider this output

The next step is diagnostics!

```
> fit=lm(y~x1+x2+x3+x4+x5+x6,data=foo)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = foo)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50089 -0.77184 -0.01539  0.81881  2.54738

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -0.006857   0.012846  -0.534                0.593
x1           4.068585   0.507623   8.015  0.00000000000000131 ***
x2           1.141218   0.256944   4.442  0.00000909229828940 ***
x3           4.031808   0.358401  11.249 < 0.0000000000000002 ***
x4           0.937294   0.127315   7.362  0.00000000000020531 ***
x5           3.984018   0.170871  23.316 < 0.0000000000000002 ***
x6           0.996700   0.022276  44.744 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9995 on 6047 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.3097,    Adjusted R-squared:  0.309
F-statistic: 452.1 on 6 and 6047 DF,  p-value: < 0.00000000000000022
```

| t | > 1.96 : We need all these x

Always best to start with a plot

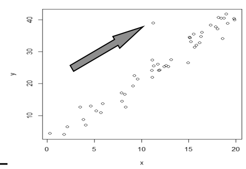
- Will do this one in class.

Outliers

Sometimes we get a point which is unusual- different from all the rest, in that the deviation away from the line seems particularly large. We call these funny points **outliers** (because it sounds better than "funny points").

Consider the data set

There seems to be one funny point !!



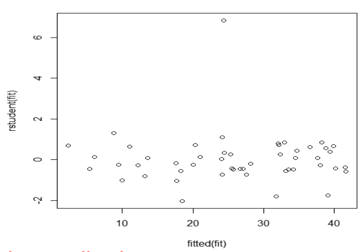
Let's see how this point shows up in the resids:

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8487 -1.4294 -0.5644  1.6264 14.6732

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  1.78015    1.05445   1.688              0.0979 .
x            2.00746    0.08019  25.034 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.031 on 48 degrees of freedom
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9274
F-statistic: 626.7 on 1 and 48 DF,  p-value: < 0.00000000000000022
```

Note the high value of R^2 still a problem with the model

The standardized residual is over 6 !

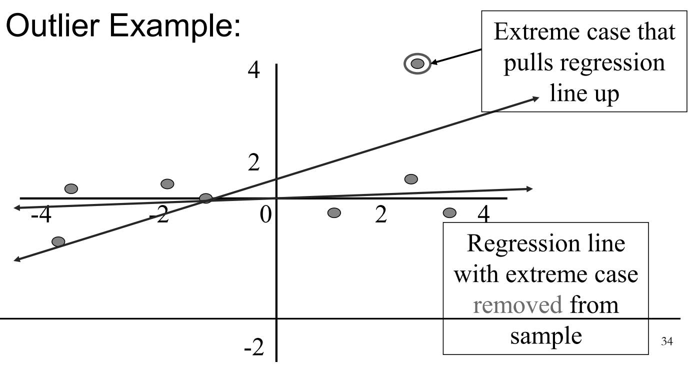
`plot(fitted(fit),rstudent(fit))`



you can see that outlier here too

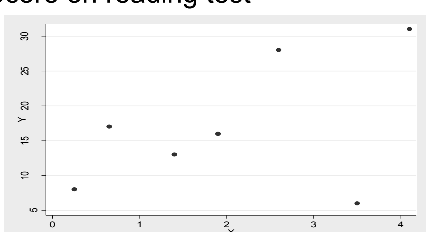
Outliers can dramatically change the line

- Outlier Example:



- Example: Study time and student achievement.
  - X variable: Average # hours spent studying per day
  - Y variable: Score on reading test

Case	X	Y
1	2.6	28
2	1.4	13
3	.65	17
4	4.1	31
5	.25	8
6	1.9	16
7	3.5	6



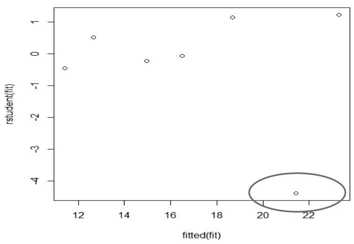
Regression Output

```
> fit=lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
      1        2        3        4        5        6        7
 9.3274  -1.9753   4.3355   7.7058  -3.4320  -0.5158 -15.4456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.662      6.402   1.665    0.157
x              3.081      2.617   1.177    0.292

Residual standard error: 9.162 on 5 degrees of freedom
Multiple R-squared:  0.217,    Adjusted R-squared:  0.0604
F-statistic: 1.386 on 1 and 5 DF,  p-value: 0.2921
```

Do you need X in the model? Doesn't look like it

Diagnostic Plot

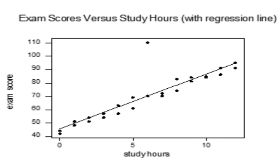
```
plot(fitted(fit),rstudent(fit))
```



Remove the outlier

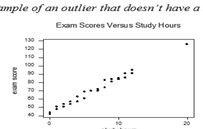
```
> fit=lm(y[-7]~x[-7])
> summary(fit)

Call:
lm(formula = y[-7] ~ x[-7])

Residuals:
      1       2       3       4       5       6 
 4.6798 -3.4467  4.8492 -0.9119 -1.8597 -3.3107 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    8.428      3.019   2.791   0.0492 *
x[-7]          5.728      1.359   4.215   0.0135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 4.259 on 4 degrees of freedom
Multiple R-squared:  0.8163,    Adjusted R-squared:  0.7703 
F-statistic: 17.77 on 1 and 4 DF,  p-value: 0.01353
```

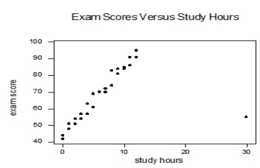
There is now a relationship! The outlier was hiding the linear relationship. Naughty outlier!



Note: Not all outliers are bad

Example of an outlier that doesn't have a large residual:

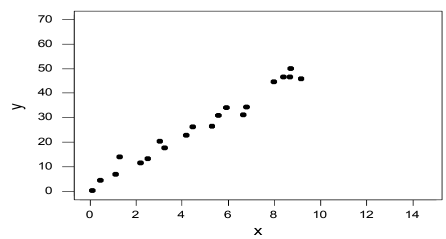




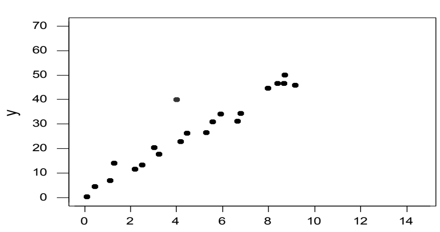
Influential observations are the worst.

Outliers in the y-space i.e. something that changes the slope, is bad

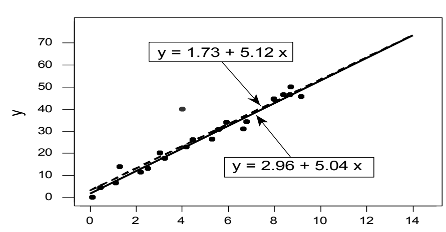
No outliers?



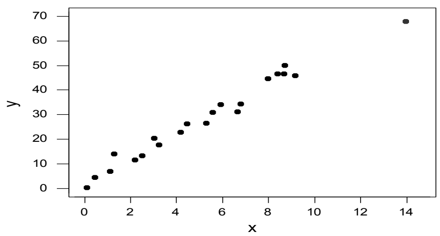
An outlier? Influential?



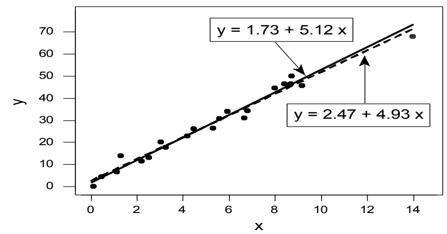
An outlier? Influential?



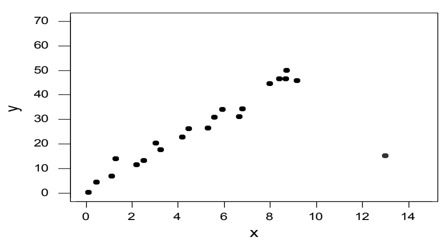
An outlier? Influential?



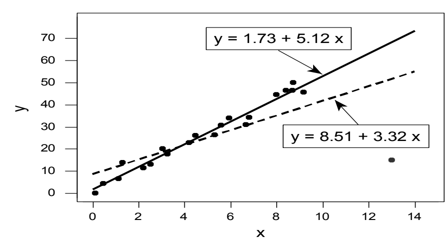
An outlier? Influential?



An outlier? Influential?



An outlier? Influential?



Unusual points in the x-space - leverage - these just move the line in a direction
Unusual points in the y-space - influential - these alter the slope of the line

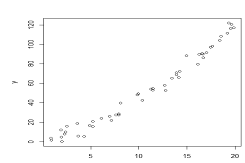
How to find Influential Points

- There is a measure called Cook's Distance which combines how extreme an observation is in the "x space" with how extreme the observation is in the "y space".
- A Cook's Distance is calculated for each row in your data set-extreme values of Cook's Distance indicate points which are probably influential (or should at least be examined).
- We will go over this after we cover multiple regression.

Nonlinearity

Another key assumption is that *Y* is a linear function of *X*.

What happens when this assumption fails ? Consider the data plotted below:



There is some nonlinearity evident in the plot !!

We run the regression and obtain the standardized residuals:

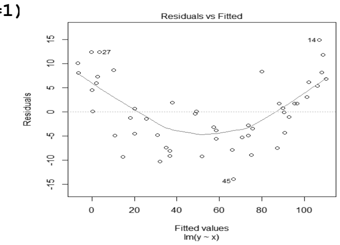
```
> fit=lm(y~x)
> sumary(fit)
Error: could not find function "sumary"
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8924  -4.9015  -0.2035   5.8075  14.8862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.8471     2.0254  -5.849 4.26e-07 ***
x             6.1471     0.1644  37.396  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.044 on 48 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9661
F-statistic:  1398 on 1 and 48 DF,  p-value: < 2.2e-16
```

Note that R^2 is pretty high.

As a *diagnostic,* we plot the residuals versus X:

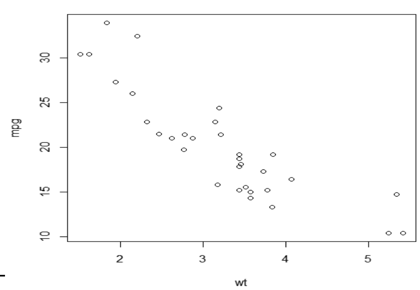
```
plot(fit,which=1)
```

there should be no relationship between the resids and X!!!!



The nonlinearity is even more evident in the residual plot !! What is wrong with fitting a linear regression to this data?

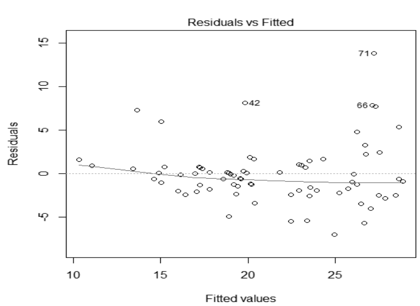
Example : Cars Data (mpg versus weight)



Regression Output

```
> fit=lm(mpg~weight)
> summary(fit)

Call:
lm(formula = mpg ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9593 -1.9325 -0.3713  0.8885 13.8174

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.4402835  1.6140031   24.44   <2e-16 ***
weight      -0.0060087  0.0005179  -11.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.439 on 72 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6515,    Adjusted R-squared:  0.6467
F-statistic: 134.6 on 1 and 72 DF,  p-value: < 2.2e-16
```

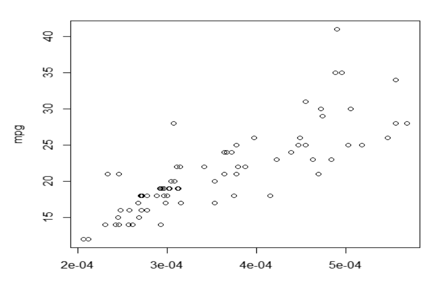

Diagnostic



What is the problem ? How do we fix it ?

We always try X first

Transformed X (using 1/X) versus Y

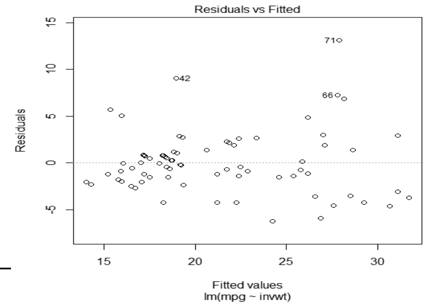


Output : Is it better ? (why ?)

```
> invwt=1/weight
> fit1=lm(mpg~invwt)
> summary(fit1)

Call:
lm(formula = mpg ~ invwt)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2259 -1.9298 -0.3319  1.3150 13.0945 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.938      1.485   2.652  0.00984 ** 
invwt       48893.647   4037.066  12.111  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.343 on 72 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6708,    Adjusted R-squared:  0.6662 
F-statistic: 146.7 on 1 and 72 DF,  p-value: < 2.2e-16
```

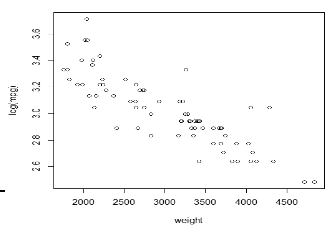

New Diagnostic Plot

- Better looking



Could have also done log(y)

- Could have logged the y variable, but interpretation becomes more difficult and can't compare models.



Example: Nations Data Set

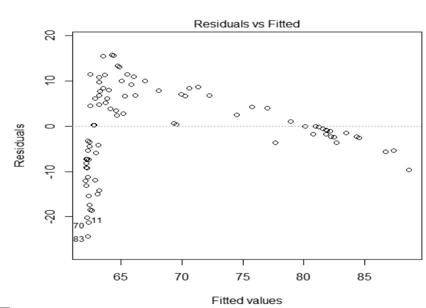
- We have a data set that has data on nations around the world. We are going to see if there is a relationship between life expectancy (females) and gross domestic product per capita (in dollars)

4	NATIONS DATA SET						
5	Country	GDP	GDPpc	Life Exp.	Persons/MD	Infant Mort.	Literacy %
6	Algeria	42.00	1570	69	1062	52	0.52
7	Angola	5.10	950	48	15136	145	0.40
8	Argentina	112.00	3400	75	326	29	0.95
9	Australia	294.00	16700	81	438	7	0.99
10	Austria	141.00	18000	80	327	7	0.99
11	Bangladesh	23.80	200	55	5264	107	0.47
12	Barbados	1.80	7000	77	1042	20	0.99
13	Belgium	178.00	17800	80	298	7	0.98
14	Belize	0.37	1635	70	2021	36	0.93
15	Brazil	369.00	2350	67	848	60	0.81
16	Burma	1.23	205	42	3177	114	0.50
17	Cambodia	2.00	280	51	27000	111	0.50
18	Cameroon	11.60	1010	59	12540	77	0.65

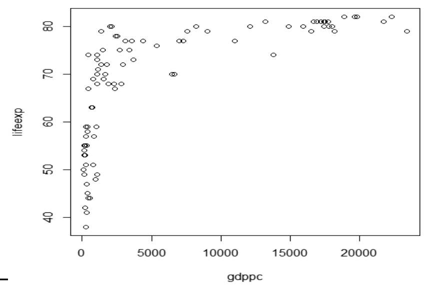
Regression Output

```
> fit=lm(lifeexp~gdppc)
> summary(fit)

Call:
lm(formula = lifeexp ~ gdppc)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.3330  -5.3513   0.1776   6.8169  15.7323 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.199e+01  1.268e+00  48.895  < 2e-16 ***
gdppc       1.138e-03  1.359e-04   8.375 7.37e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 9.268 on 89 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4407,    Adjusted R-squared:  0.4344 
F-statistic: 70.13 on 1 and 89 DF,  p-value: 7.37e-13
```


Diagnostic Plot: Uh Oh!



Maybe we should have plotted the data



We can fix this…………

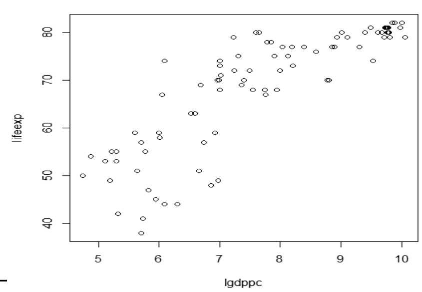
The incredible log transformation

The log transformation is used quite often in regression analysis.

There are three basic reasons for applying the log transformation:

- to accommodate non-linearity
- to reduce right skewness in the Y(or, equivalently, in the error term)
- to eliminate heteroskedasticity (non-constant variance)

Take the log of the X variable



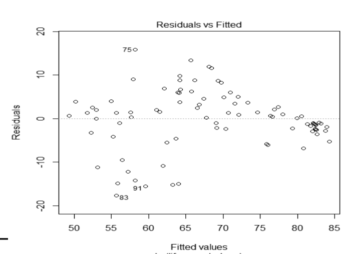
The New Regression Output

- Why is this a better model?

```
> fit=lm(lifeexp~lgdppc)
> summary(fit)

Call:
lm(formula = lifeexp ~ lgdppc)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6610  -2.4146   0.2866   3.8178  15.8226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.1846     3.4986   5.198 1.27e-06 ***
lgdppc        6.5704     0.4449  14.769  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.671 on 89 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.707
F-statistic: 218.1 on 1 and 89 DF,  p-value: < 2.2e-16
```

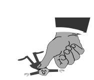

Residual Plot of New Model

- Looks better



For linear regression, if Y vs X is not linear, you want to try to make it linear

Some Rules of Thumb for Transforming X:



In general people try

$$\log(X), \sqrt{X}, 1/X, X^2$$

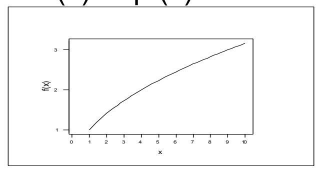
Obviously, if you have a lot of X's and you try transforming each one it will take a while.

Also, you are welcome to transform the Y variable also. But we want to make sure our model is still interpretable.

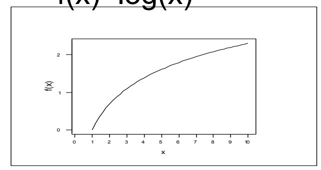
Ladder of Transformations

p		
	1.5	
	1.0	**Transformation is x^p**
	0.5	
	0.0 log	
	-0.5	
	-1.0	

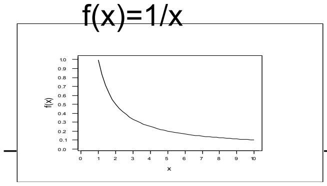
f(x)=sqrt(x)



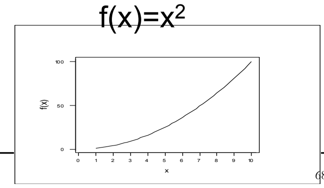
f(x)=log(x)



f(x)=1/x



$f(x)=x^2$



How to find transformations:

- Plot Y versus each X in the model-see if a non-linear relationship
- Use residual diagnostic plots to help spot ill-fitting models.
- Trial and Error- look at a lot of graphs.

Things you should know

- ❑Always plot residuals versus each X variable
- ❑If regression assumptions are violated, can’t trust confidence intervals and hypothesis tests
- ❑Possible problems are outliers, and nonlinearity