# Section 2.4 - Cautions about Regression and Correlation

Statistics 104

Autumn 2004

# Residual Plots

Underlying the regression line description is the "model"

$$Data = Fit + Error$$

where the fit is given by a straight line. It is useful to examine whether this model is a reasonable description of the data.

This is usually done by examining the residuals from the regression

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ e &= y - \hat{y} \end{aligned}$$

Facts about residuals

- When using the least squares line $\bar{e} = 0$ (average residual = 0). So the regression line doesn't tend to over or under predict.

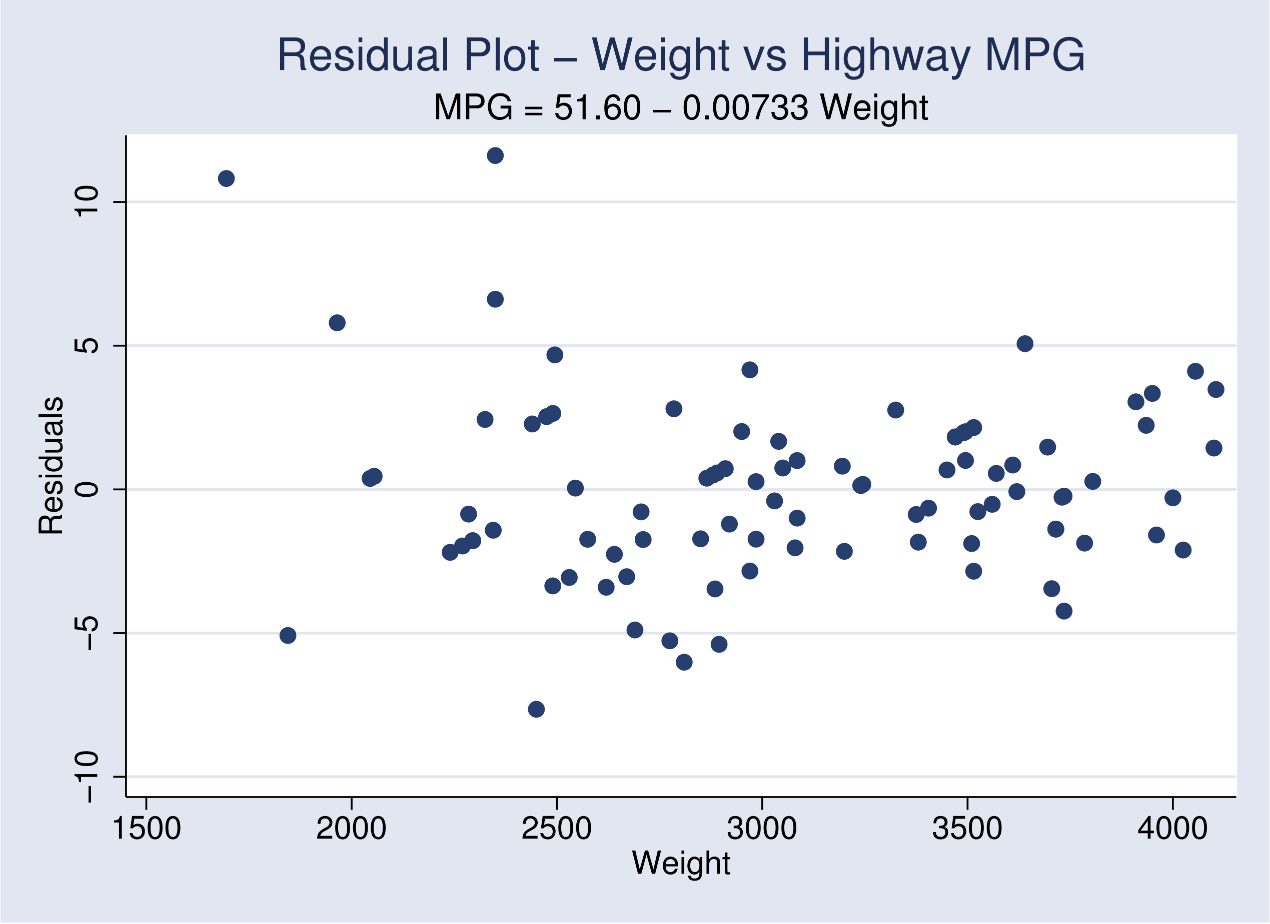
- $r_{x,e} = 0$. The correlation between the $x$'s and the residuals is 0. This implies that the regression line gets all the information about the linear pattern in the data.

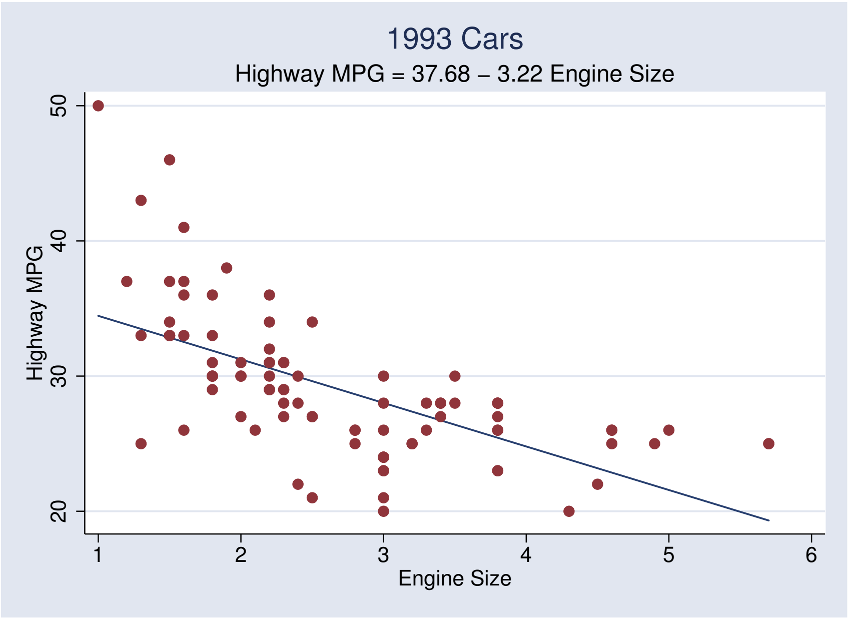
Sometimes problems can be obvious from the scatterplot of the data, but often problems can be detected more readily by examining the residuals.

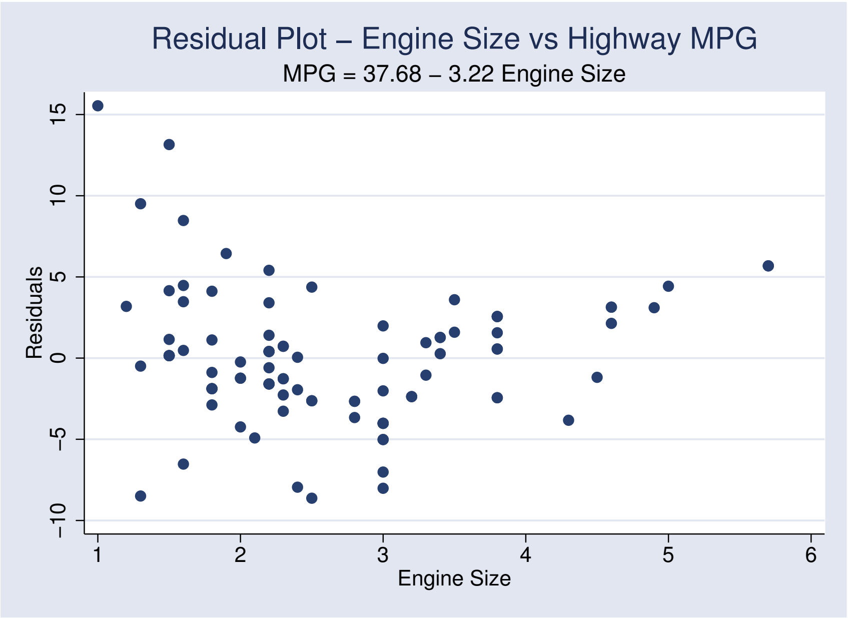
A common way of doing this examination is by a residual plot.

## Residual Plot

A scatterplot of the residuals versus the $x$'s (or the fitted values, $\hat{y}$).




1993 Cars
Highway MPG = 37.68 – 3.22 Engine Size
Highway MPG
Engine Size

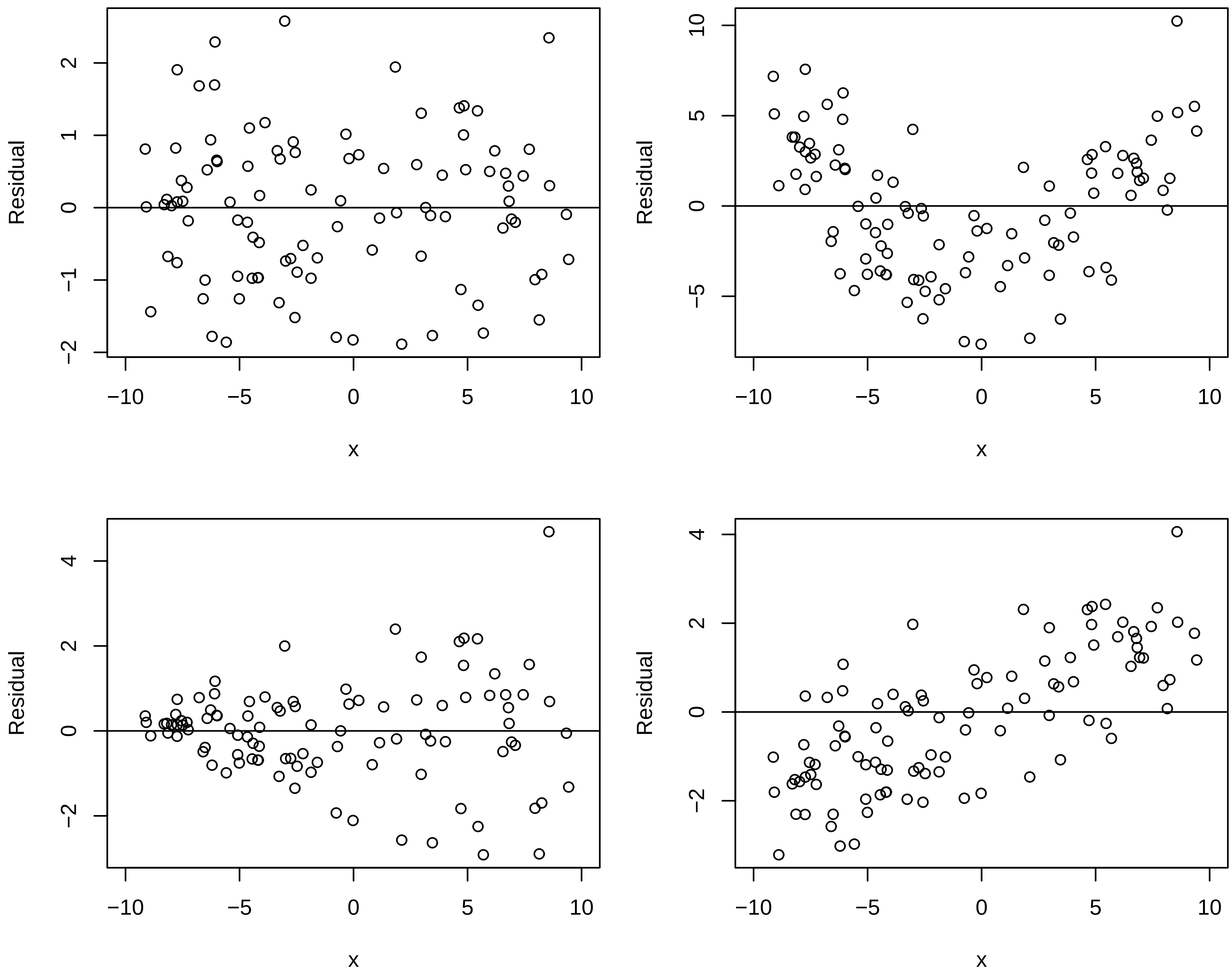

Residual Plot – Engine Size vs Highway MPG
MPG = 37.68 – 3.22 Engine Size
Residuals
Engine Size

What do you want to see in a residual plot?

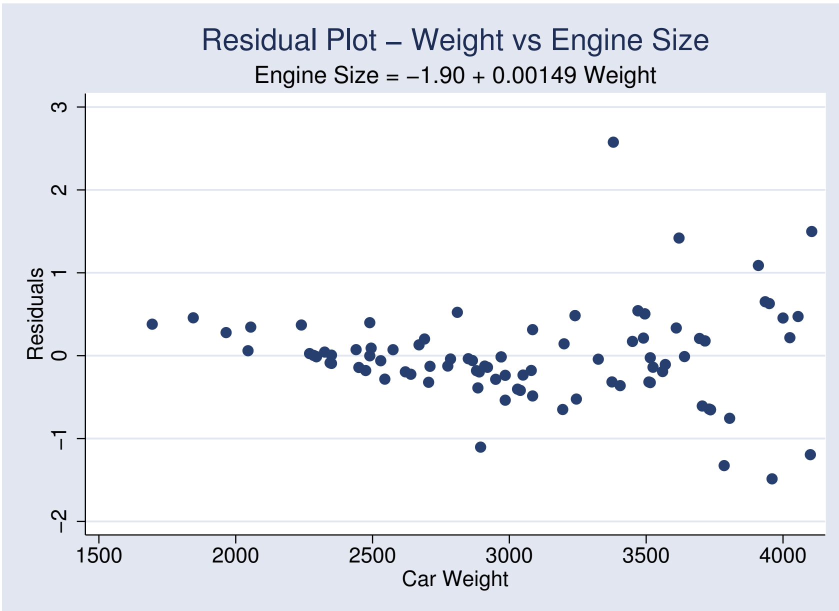
- Nothing
- No obvious pattern
- No points standing out

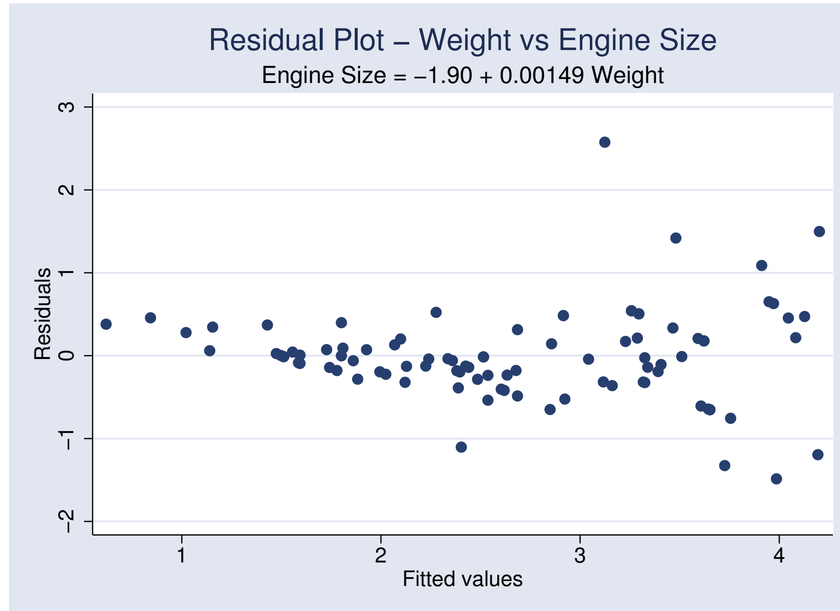
What do you **not** want to see in a residual plot?

- Curved pattern
- Fan shaped pattern
- Something standing out


Residual
x
Residual
x
Residual
x
Residual
x

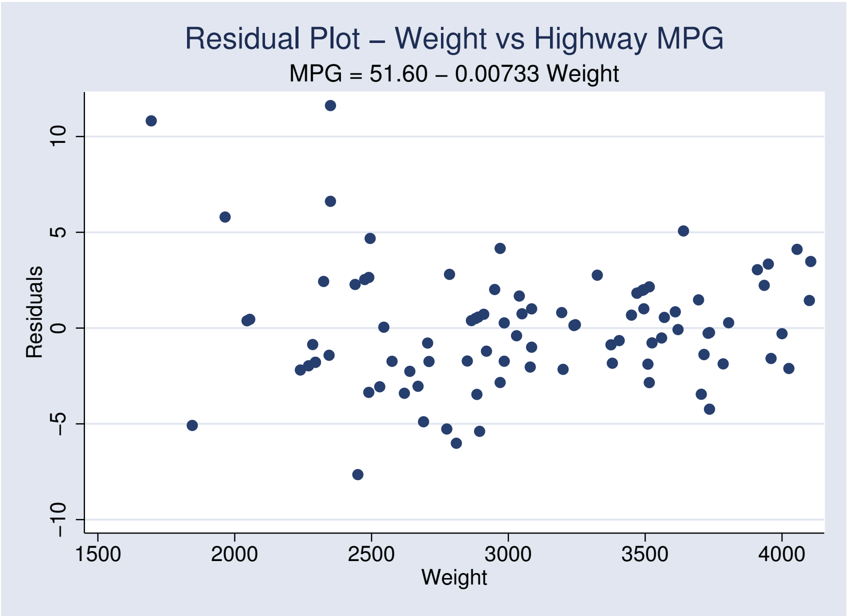
As mentioned before, you can plot the residuals versus the fits instead of the explanatory variable

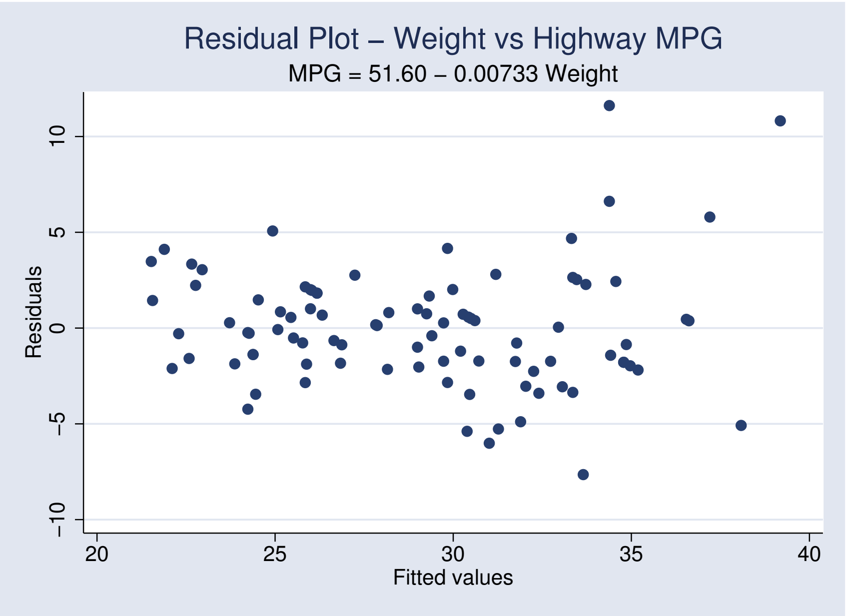



You will get effectively the same plot with the explanatory variable or the fitted values on the $x$ axis. This occurs since the fitted values can be considered a rescaling of the $x$'s as its just a linear transformation.

One slight difference you can see is that the order of the plots can flip on the $x$ axis. This will occur when $b < 0$. However this won't affect any curvature or changing variability in the plot.



## Residuals

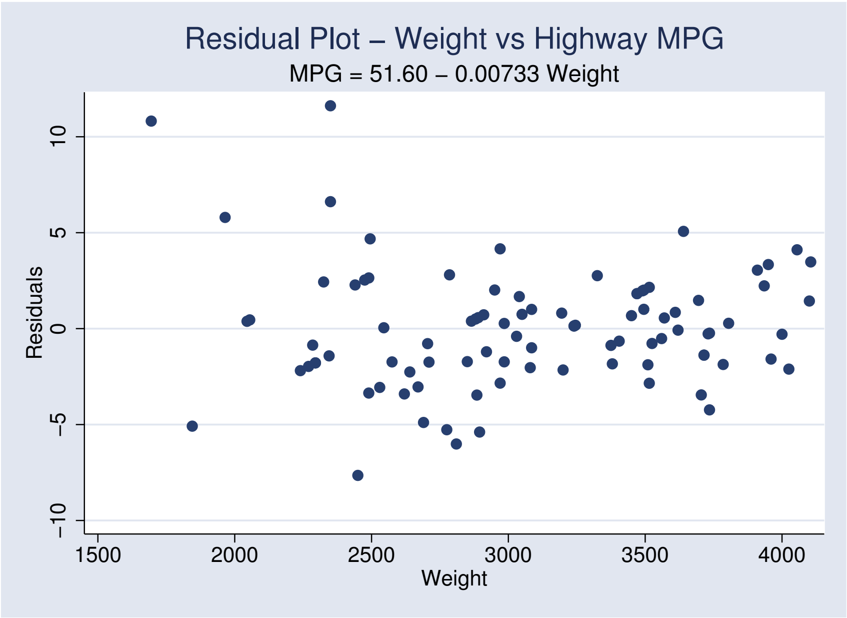
The residuals measure departures from the regression line

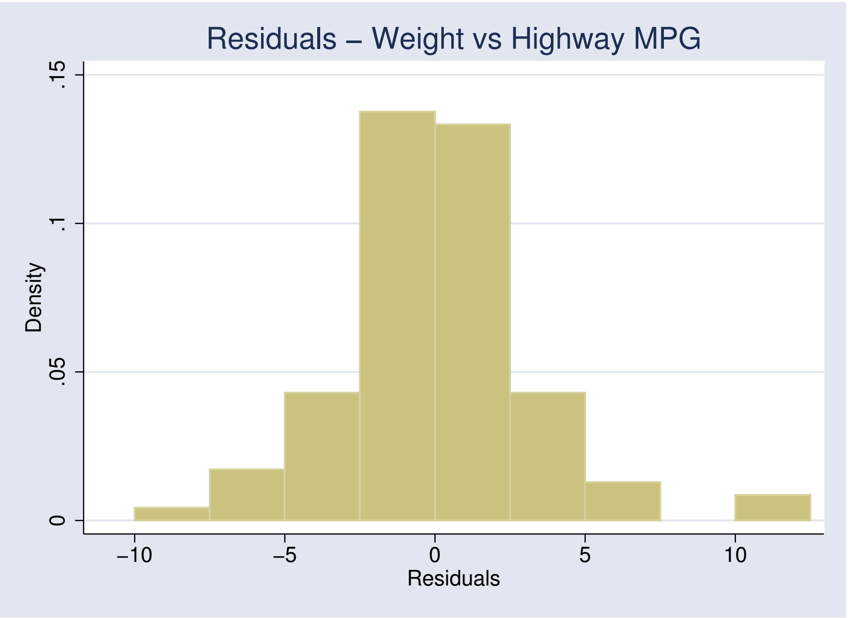
The size of a typical departure from the regression line can be measured by the standard deviation of the residuals. This is sometimes referred to as the Root Mean Square Error (Root MSE or RMSE).

```
. regress HighMPG Weight

  Source |       SS       df       MS              Number of obs =      93
---------+------------------------------           F(  1,    91) =  174.43
   Model |  1718.69528     1  1718.69528           Prob > F      =  0.0000
Residual |  896.616546    91  9.85292907           R-squared     =  0.6572
---------+------------------------------           Adj R-squared =  0.6534
   Total |  2615.31183    92  28.4273025           Root MSE      =  3.1389
```

The Mean Square Error (MSE) is the variance of the residuals. Note that it is calculated by

$$MSE = \frac{SSE}{df}$$




In this example RMSE $= 3.14$. There are a number of observations with residuals of this magnitude.

# Outliers and Influential Points

## Outliers

- Points that lie outside the overall pattern of the other observations
- When discussing outliers in regression, it usually refers to outliers in the $y$ direction, i.e. points with big residuals
- You can also can have outliers in the $x$ direction

## Influential Points

- Observations, that if removed from the analysis, would give markedly different results.
- Often outliers in the $x$ direction

Observations can be outliers, influential, both, or neither.

**Finding outliers:**

Can use the univariate approaches that we have already discussed, e.g. Boxplots (& 1.5 IQR rule), histograms, etc.

Another popular rule that is often used is to look for residuals satisfying

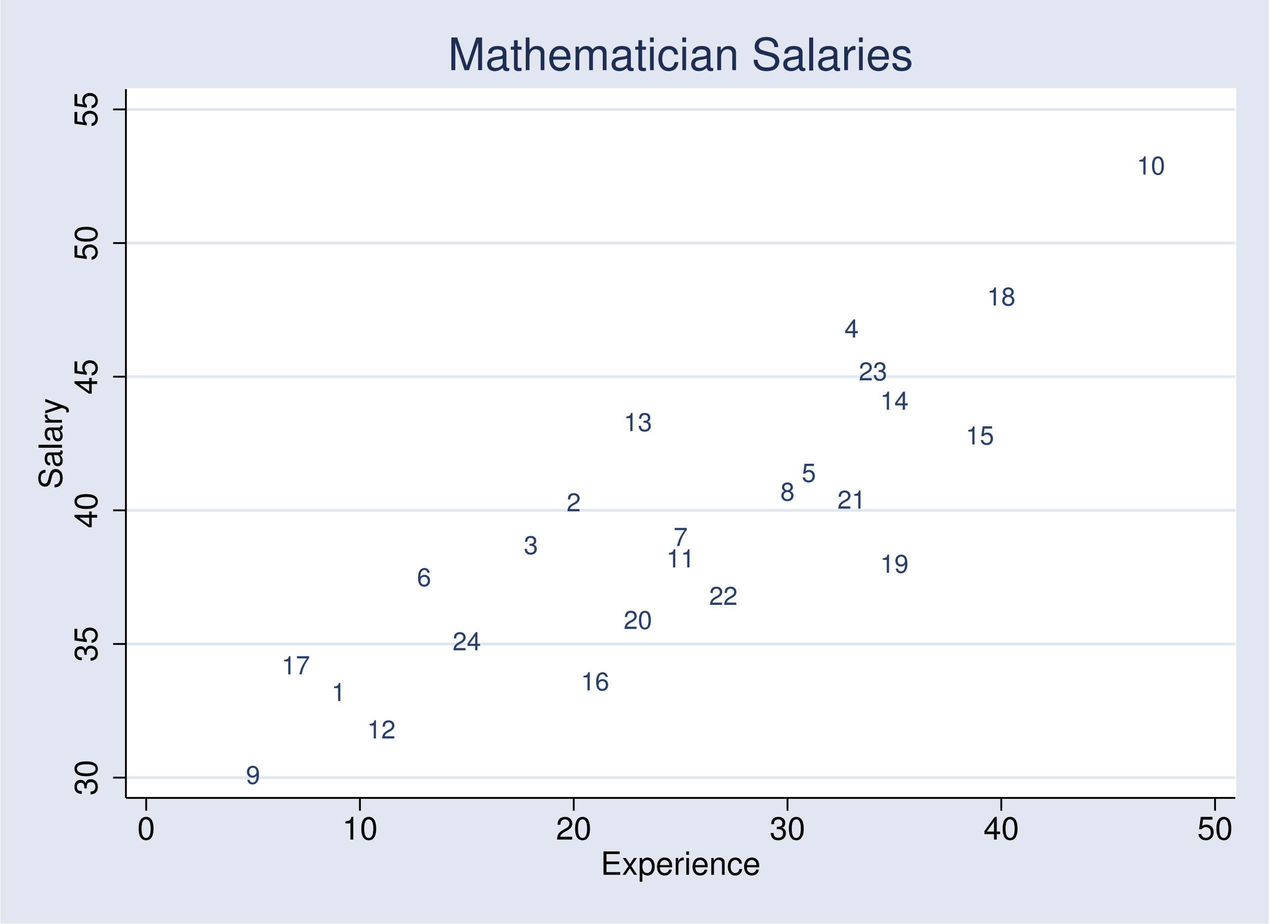
$$|e_i| > 2RMSE$$

This rule is based on the normal distribution (which we will talk about soon when we get back to Section 1.3). As we shall see, this rule has about a 5% chance of declaring a point an outlier, even if the residuals are all normally distributed.

Example: Mathematician Salaries

$x$: years experience

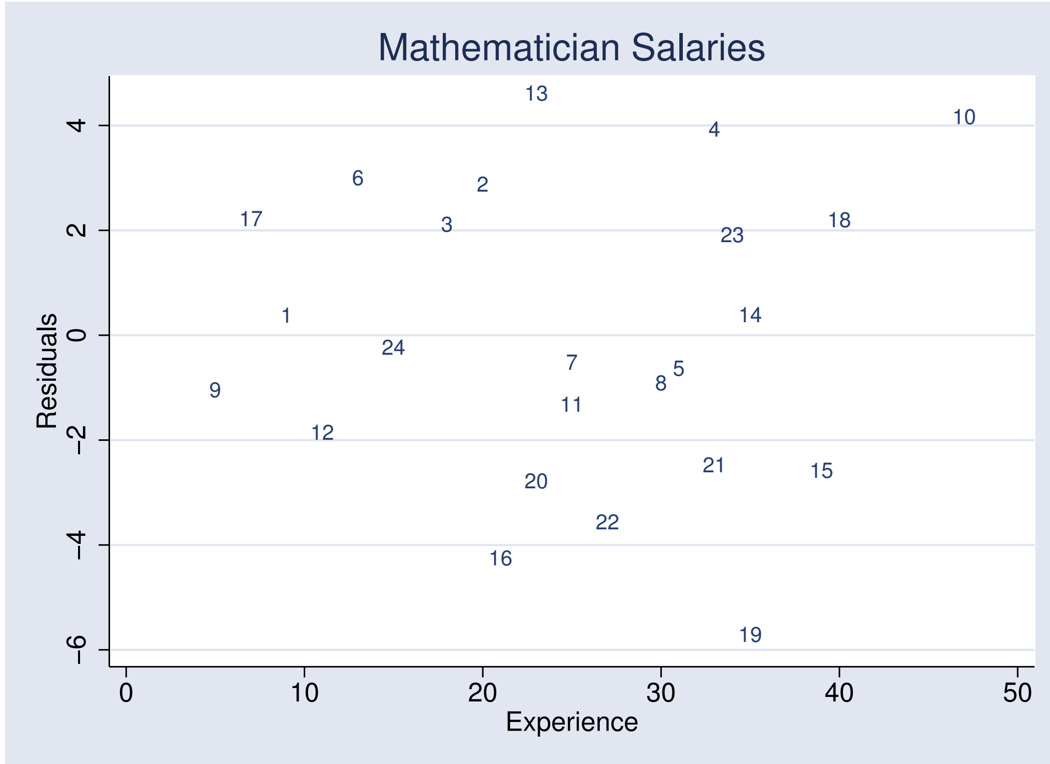
$y$: annual salary (\$1,000)

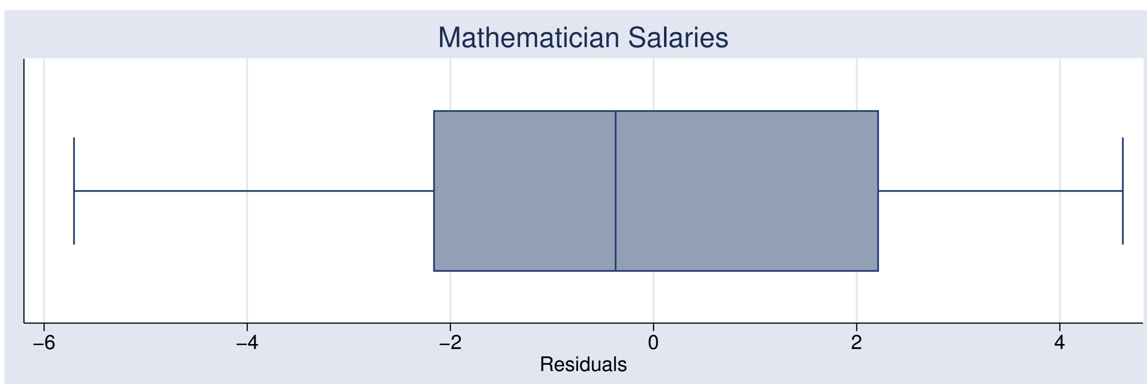


```
. regress Salary Experience

  Source |       SS       df       MS              Number of obs =      24
---------+------------------------------           F(  1,    22) =   61.69
   Model |  508.068856     1  508.068856           Prob > F      =  0.0000
Residual |  181.191175    22  8.23596249           R-squared     =  0.7371
---------+------------------------------           Adj R-squared =  0.7252
   Total |  689.260031    23  29.9678274           Root MSE      =  2.8698

------------------------------------------------------------------------------
    Salary |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
Experience |  .4187841   .0533195      7.85   0.000     .3082062     .529362
     _cons |  29.04785   1.453995     19.98   0.000     26.03245    32.06325
------------------------------------------------------------------------------
```

By the $|e_i| > 2RMSE$ rule, we are looking for $|e_i| > 5.74$.

Observation 19 just misses being picked up by this rule as $e_{19} = -5.71$.

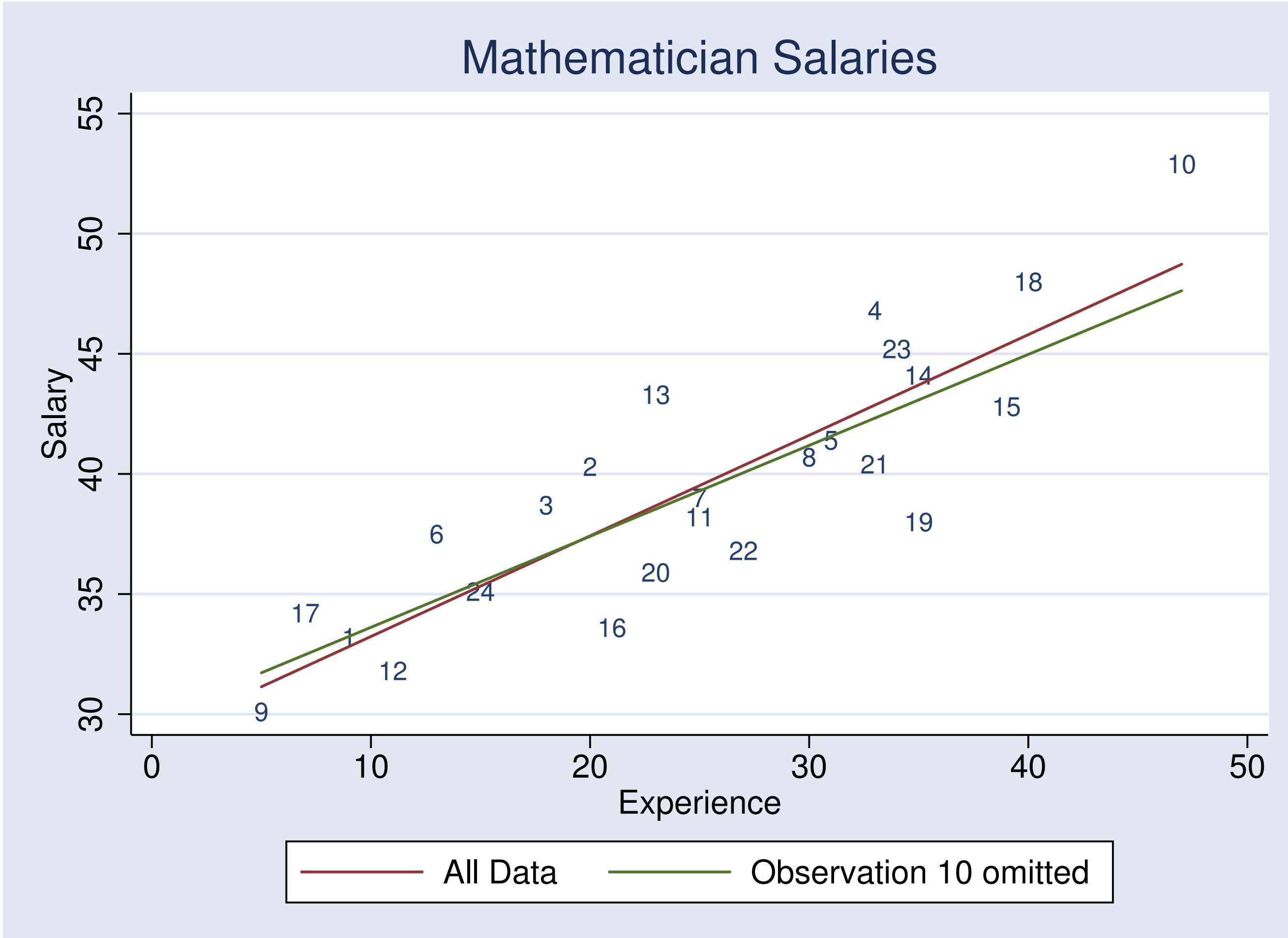

Mathematician Salaries
Residuals
Experience
13
4
10
6
2
17
3
23
18
1
14
24
7
5
8
9
11
12
21
15
20
22
16
19
4
2
0
−2
−4
−6
0
10
20
30
40
50


Mathematician Salaries
−6
−4
−2
0
2
4
Residuals

## Finding influential points:

- Drop interesting points and refit line
- Influence statistics:
  - DFits – Measure of how much the fit of each observation depends on that observation
  - Cook's D – Measure of how much the fit of all observations depends on each observation
  - DFBetas – Measure of how much $a$ and $b$ change when each observation is dropped
  - leverages – Measure of how of much and observation is an outlier in the $x$ direction.
  - etc

These are all based on the idea of dropping points and rerunning the regression. However, with smart calculations, they can determined from the original regression run.

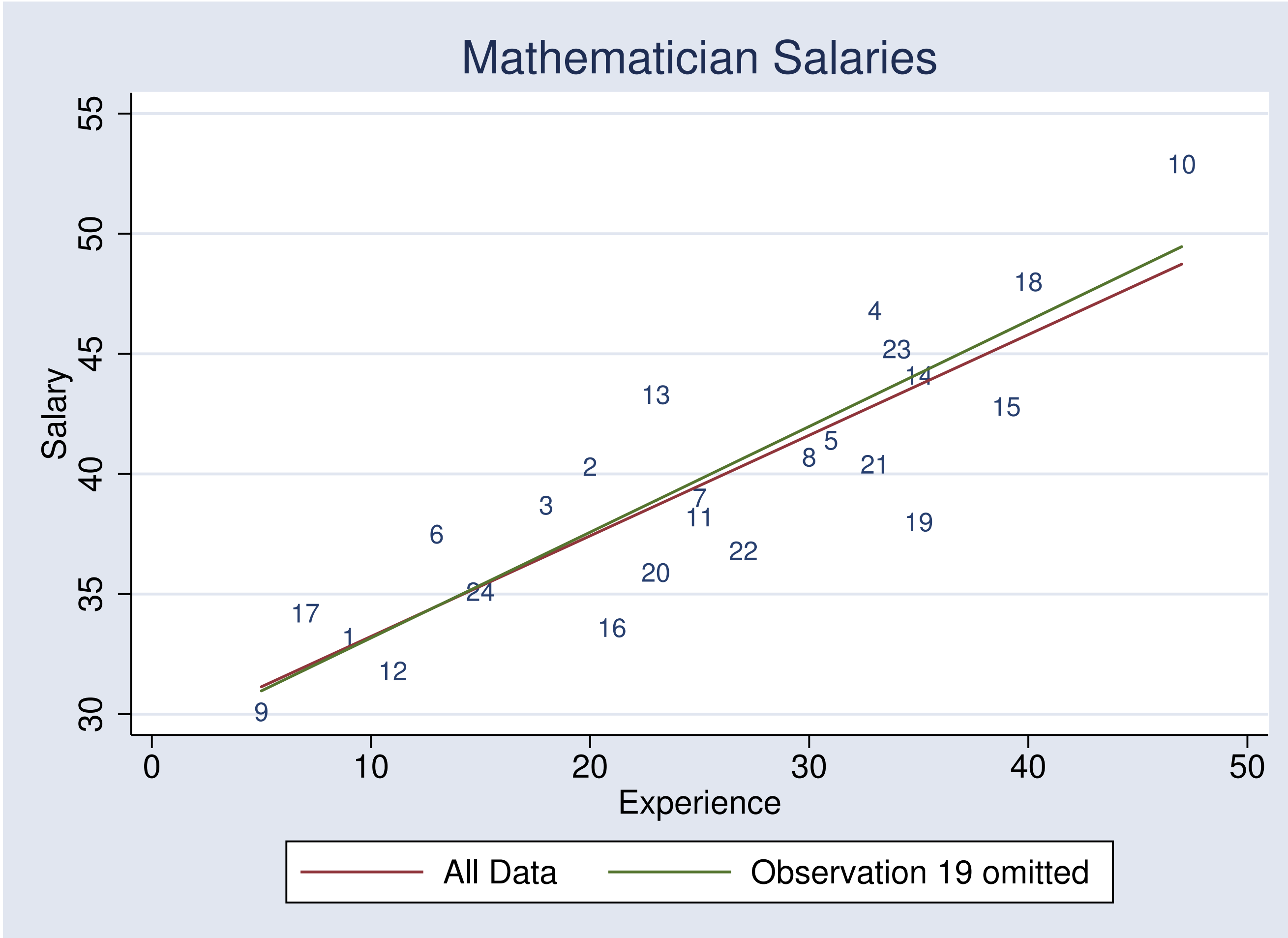



All Observations

$$\widehat{Salary} = 29.05 + 0.419 Experience$$

Observation 10 omitted

$$\widehat{Salary} = 29.83 + 0.379 Experience$$

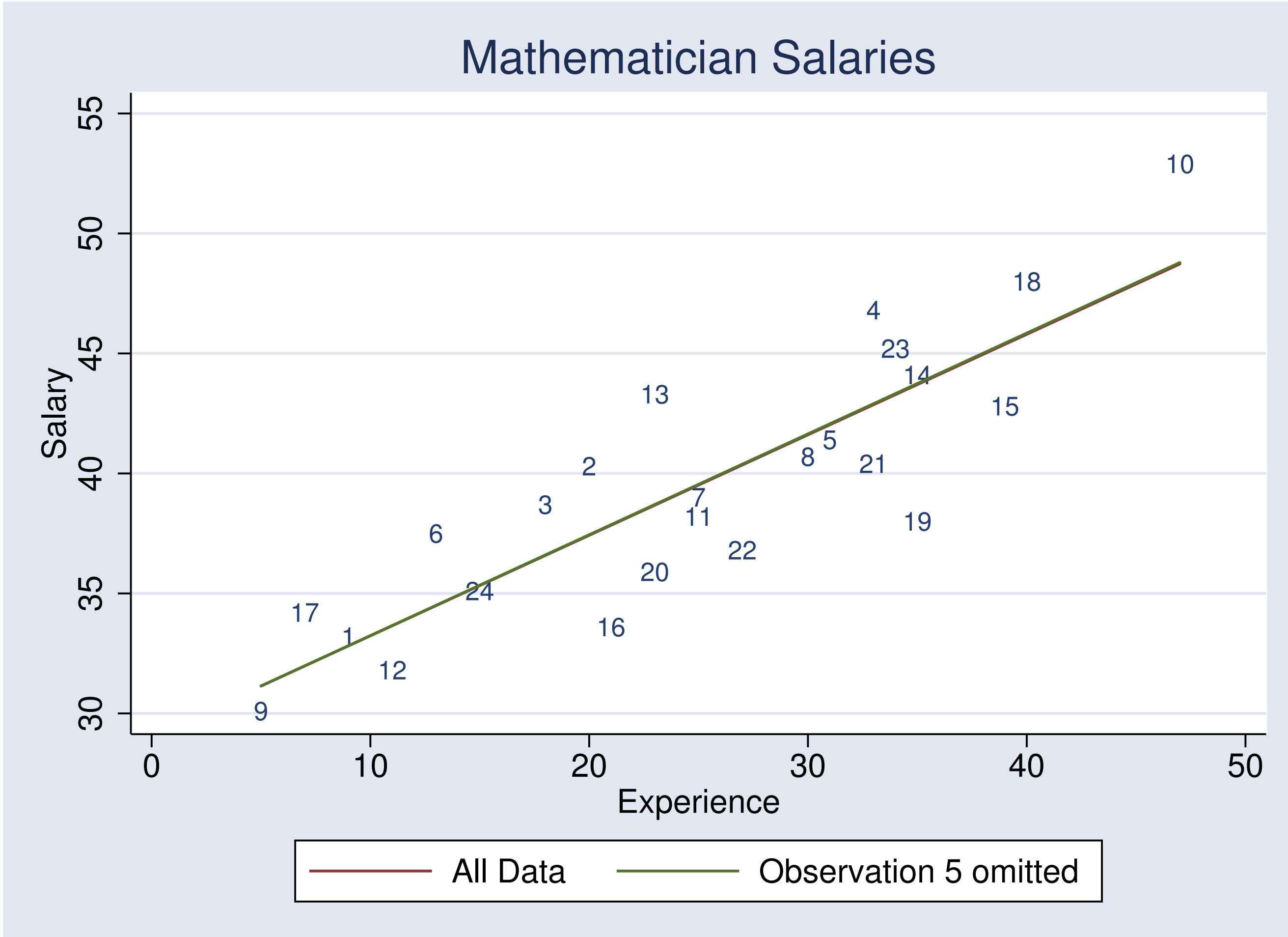



All Observations

$$\widehat{Salary} = 29.05 + 0.419 Experience$$

Observation 19 omitted

$$\widehat{Salary} = 28.77 + 0.440 Experience$$




| All Observations | Observation 5 omitted |
|---|---|
| $\widehat{Salary} = 29.05 + 0.419 Experience$ | $\widehat{Salary} = 28.04 + 0.420 Experience$ |

# Lurking Variables

A variable that is not among the explanatory or response variables in a study (or not considered for the analysis) and yet may influence the interpretation of relationships among those variables.

Example: Fisher Iris data

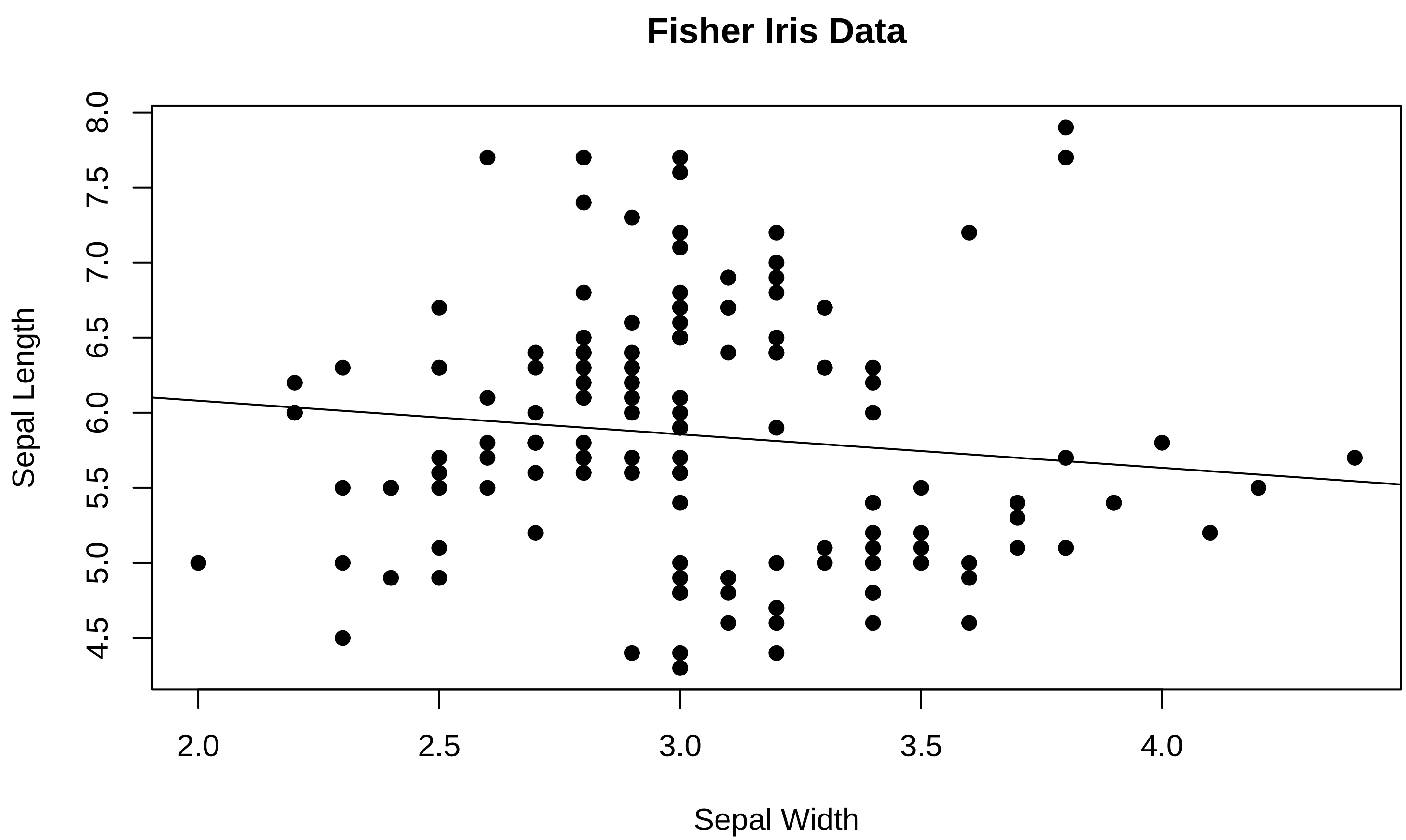
4 variables: sepal length, sepal width, petal length, petal width

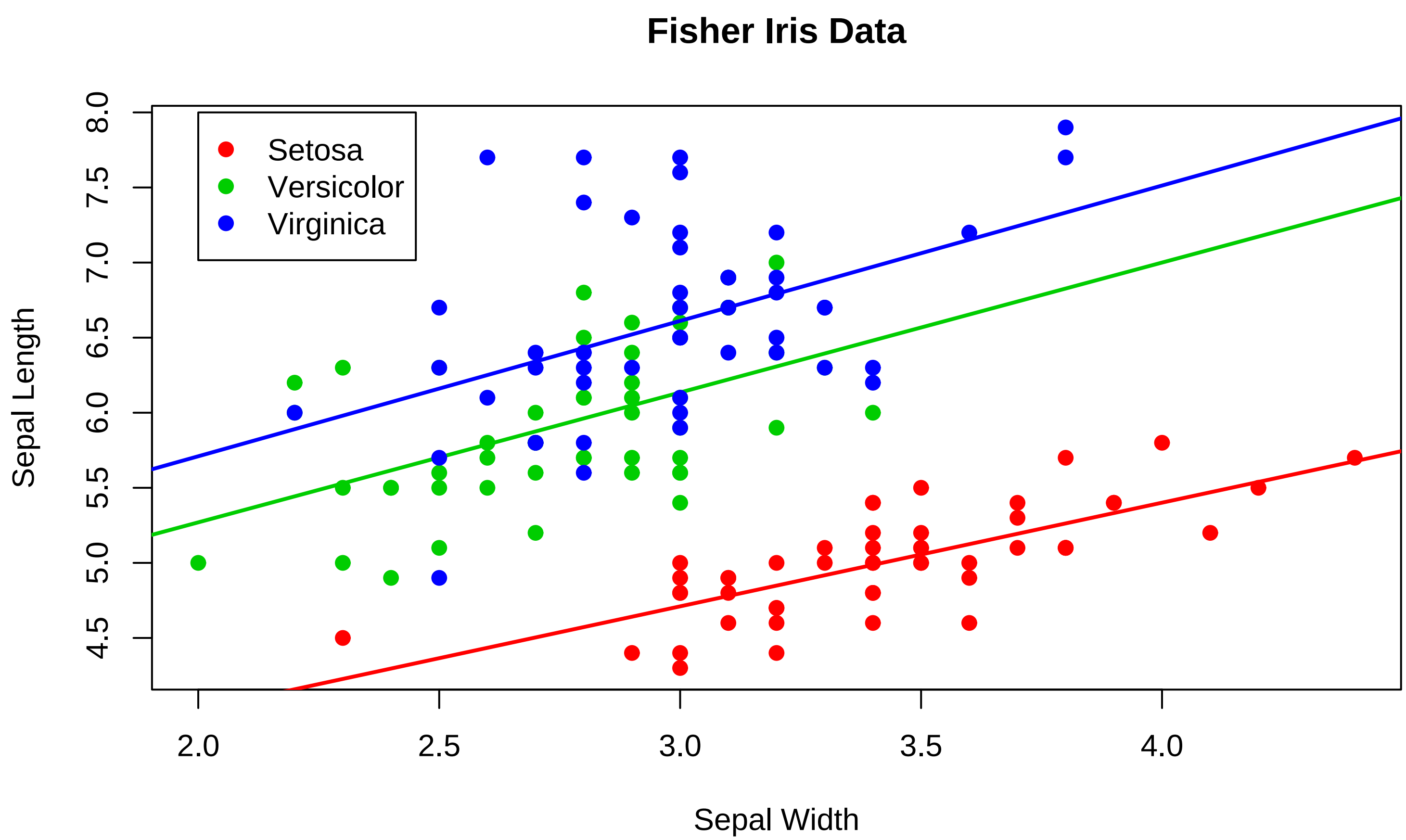
3 species:

Setosa

Versicolor

Virginica


Fisher Iris Data
Sepal Length
Sepal Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0




By ignoring species, we miss the more logical pattern of wider sepals being associated with longer sepals. There is a similar increasing trend for each species.

Example: Mathematician Salaries

There are two other possible explanatory variables in the data set, Work quality and Publication success.

Plotting the residuals against other variables can help find possible lurking variables.


In this case, both work quality and publication success are positively associated with the residuals, suggesting that both these variables should be added to the model to describe salary.

```
. regress Salary Experience WorkQuality Publication

  Source |       SS       df       MS              Number of obs =      24
---------+------------------------------           F(  3,    20) =   68.12
   Model |  627.817014     3  209.272338           Prob > F      =  0.0000
Residual |  61.4430168    20  3.07215084           R-squared     =  0.9109
---------+------------------------------           Adj R-squared =  0.8975
   Total |  689.260031    23  29.9678274           Root MSE      =  1.7528

------------------------------------------------------
     Salary |      Coef.   Std. Err.       t    P>|t|
------------+-----------------------------------------
 Experience |   .3215197   .0371087     8.66   0.000
WorkQuality |    1.10313   .3295734     3.35   0.003
Publication |   1.288941   .2984792     4.32   0.000
      _cons |   17.84693   2.001876     8.92   0.000
------------------------------------------------------
```

$$\begin{aligned}\widehat{Sal} &= 29.048 + 0.419Exp \\ \widehat{Sal} &= 17.847 + 0.322Exp + 1.103Work + 1.289Pub\end{aligned}$$

So ignoring work quality and publication success tends to lead you to overestimate the effect of experience on salary. Though the original analysis is not unreasonable if you only want to use a single predictor to describe salary.

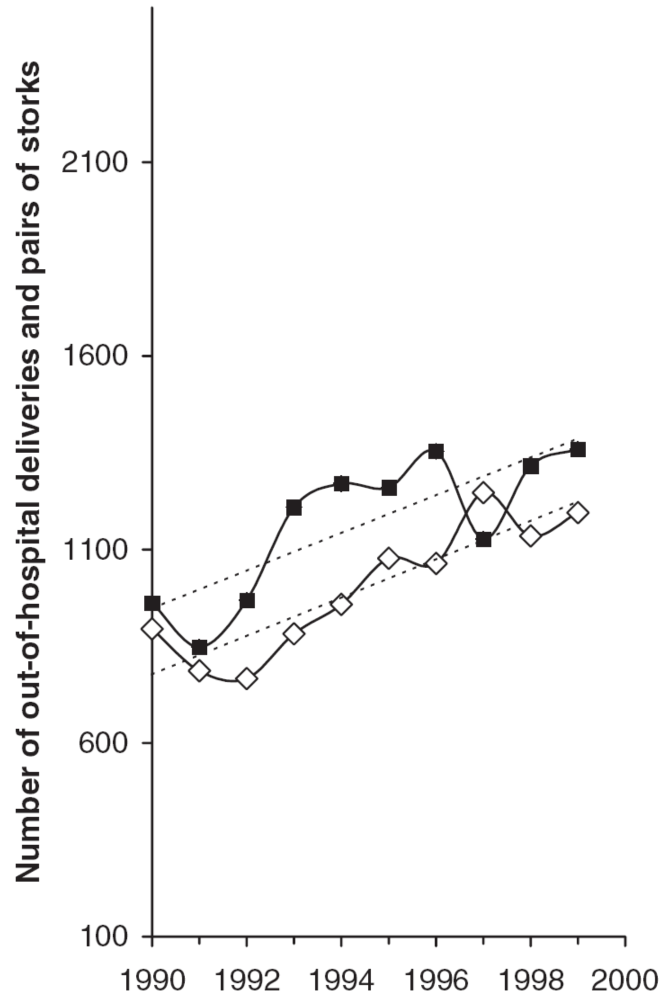
The reason that the slope is lower in the combined analaysis is that experience is positively correlated with work quality and publication success.

```
. correlate Salary Experience  WorkQuality Publication (obs=24)

             |   Salary Experi~e WorkQu~y Public~n
-------------+------------------------------------
      Salary |   1.0000
  Experience |   0.8586   1.0000
 WorkQuality |   0.6671   0.4670   1.0000
 Publication |   0.5582   0.2538   0.3228   1.0000
```

Example: Storks and Births in Berlin

Solid squares: number of pairs of storks in Brandenburg

Open diamonds: number of out of hospital deliveries in Berlin



# Association Does Not Imply Causation

An association between two variables, even if it is very strong, is not by itself good evidence that changes in one variable actually cause changes in the other.

This is an example of spurious correlation.

Conversely, a lack of correlation does not imply that a causal relationship doesn't exist.

There could a lurking variable that is masking the causal effect.