Section 2.3 - Least Squares Regression

Statistics 104

Autumn 2004

Regression

Correlation gives us a strength of a linear relationship is, but it doesn't tell us what it is.

A regression line is one approach to give that relationship.

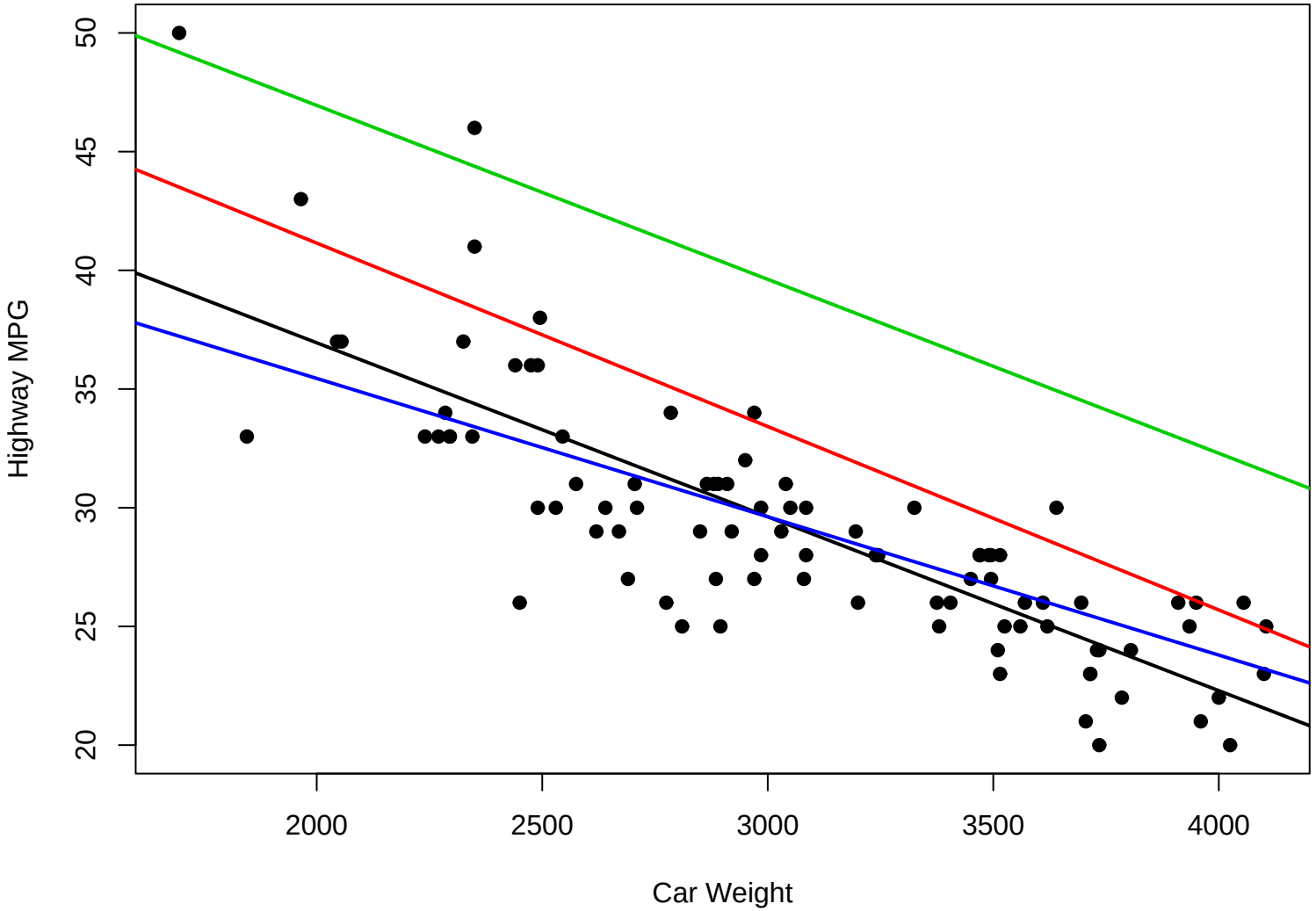
A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. The regression line is often used to predict the value of y for a given value of x.

Regression requires an explanatory variable and a response variable.

Example: Car Weight and Highway MPG

$$MPG = a + bWeight$$

How to pick a good line, as there are main possible ones.



Want a rule to do it - we will use one known as least squares.

Least Squares:

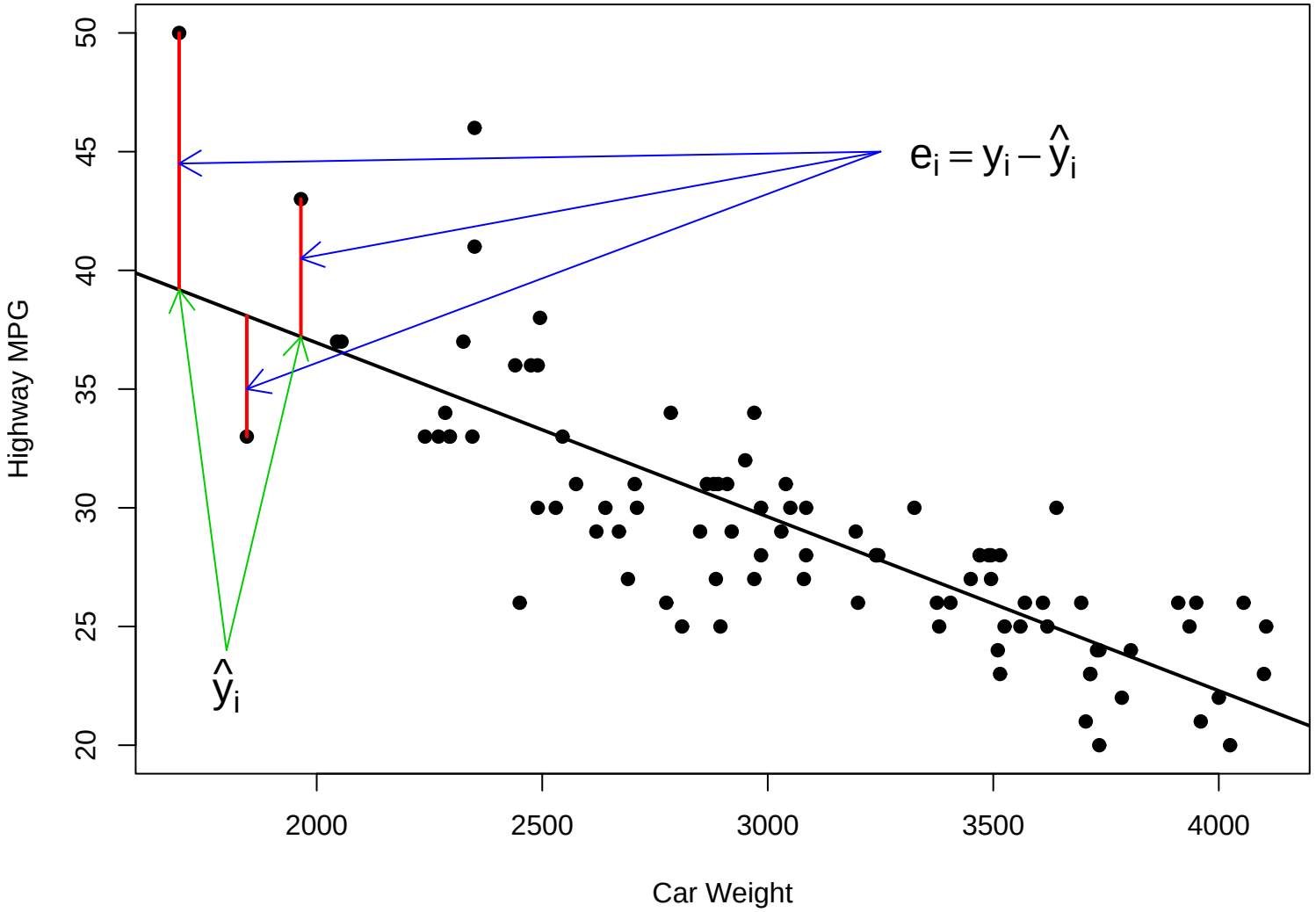
Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Pick a line $y = a + bx$ (e.g. pick intercept a and slope b)

Predicted values $\hat{y}_i = a + bx_i$

Residual $e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$

The residual measures how well a line describes each data point. We will like to have each of the e_i as small as possible.


Highway MPG
20
25
30
35
40
45
50
2000
2500
3000
3500
4000
Car Weight
$e_i = y_i - \hat{y}_i$
$\hat{y}_i$

Trying to deal with n residuals separately can get difficult quickly. We need a single measure describing the fit of each possible line.

One possible measure is the Error Sums of Squares (SSE) (or Residual Sums of Squares)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

The least squares criterion picks a and b to make this as small as possible.

The least squares solution is

$$\begin{aligned} b &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Can get these very easily from Stata

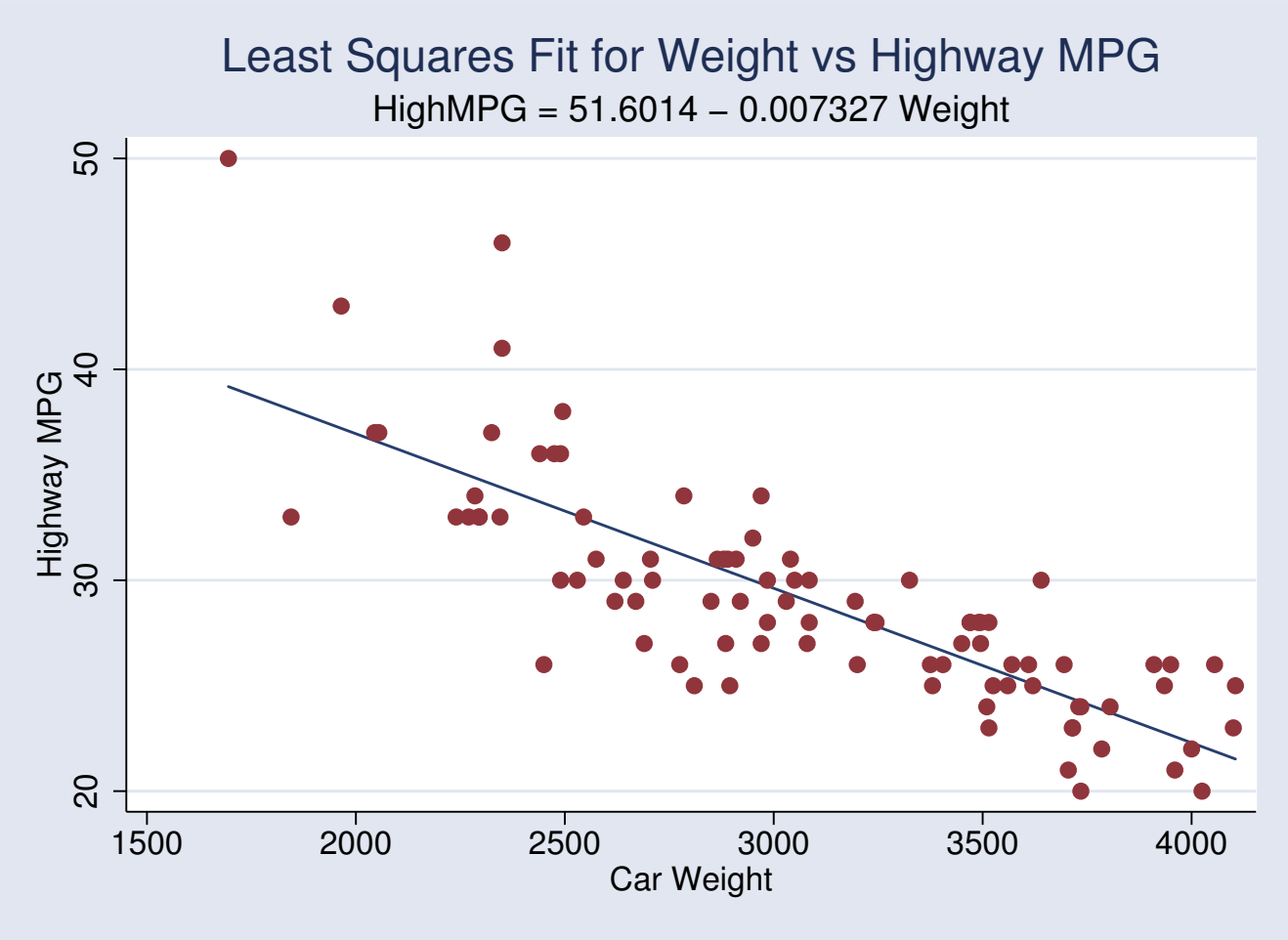
```
. regress HighMPG Weight

  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

------------------------------------------------------------------------------
 HighMPG |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |  -.0073271   .0005548   -13.21   0.000     -.008429   -.0062251
   _cons |   51.60137    1.73555    29.73   0.000     48.15391    55.04882
------------------------------------------------------------------------------
```

The line for `Weight` gives information about b and the line for `_cons` gives information about a. The column `Coef.` gives the estimates for the slope b and the intercept a.

We will talk about the other output during the term.


Least Squares Fit for Weight vs Highway MPG
HighMPG = 51.6014 – 0.007327 Weight
Highway MPG
50
40
30
20
1500
2000
2500
3000
3500
4000
Car Weight

In this example,

$$\begin{aligned} b &= -0.007327 \\ a &= 51.6014 \\ \widehat{MPG} &= 51.6014 - 0.007327 \times Weight \end{aligned}$$

Interpreting a and b

- b (slope): Expected change in response when x increases by 1.

  So for this example, for each 1 lb increase, expect the Highway MPG to go down by 0.007327. So for a 100 lbs increase, expect the MPG to do down by 0.7327.

- a (intercept): Expected (mean) value of y when $x = 0$. This doesn't always make sense. For this example, the idea implies that an average car with 0 weight will get 51.6 MPG. Anybody have a zero weight car?

Relationship between regression and correlation

The formula for the slope b can also be written as

$$b = r\frac{s_y}{s_x}$$

So another way of thinking of b is that a change of one standard deviation in x, will lead to an expected change in y of r standard deviations.

Since $-1 \leq r \leq 1$, the expected magnitude of the change in y will be less than the change in magnitude of x.

The least squares line can also be written as

$$\hat{y} = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

One thing that this formula implies is that the regression line must go through the point of means $(\bar{x}, \bar{y})$.

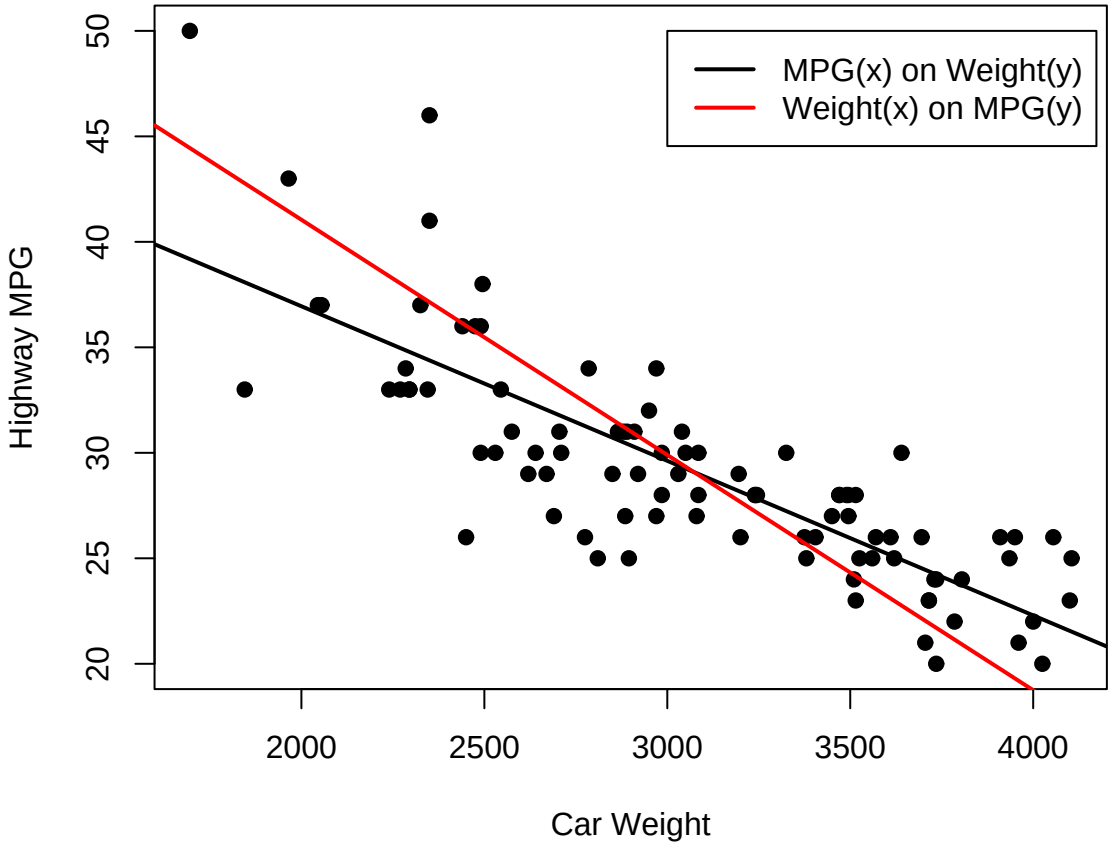
Regressing x on y and y on x

While there is no ordering of variables when it comes to correlation ($r_{xy} = r_{yx}$), it matters in regression.

The lines corresponding to the two regressions

$$\begin{aligned} y &= a + bx \\ x &= c + dy \end{aligned}$$

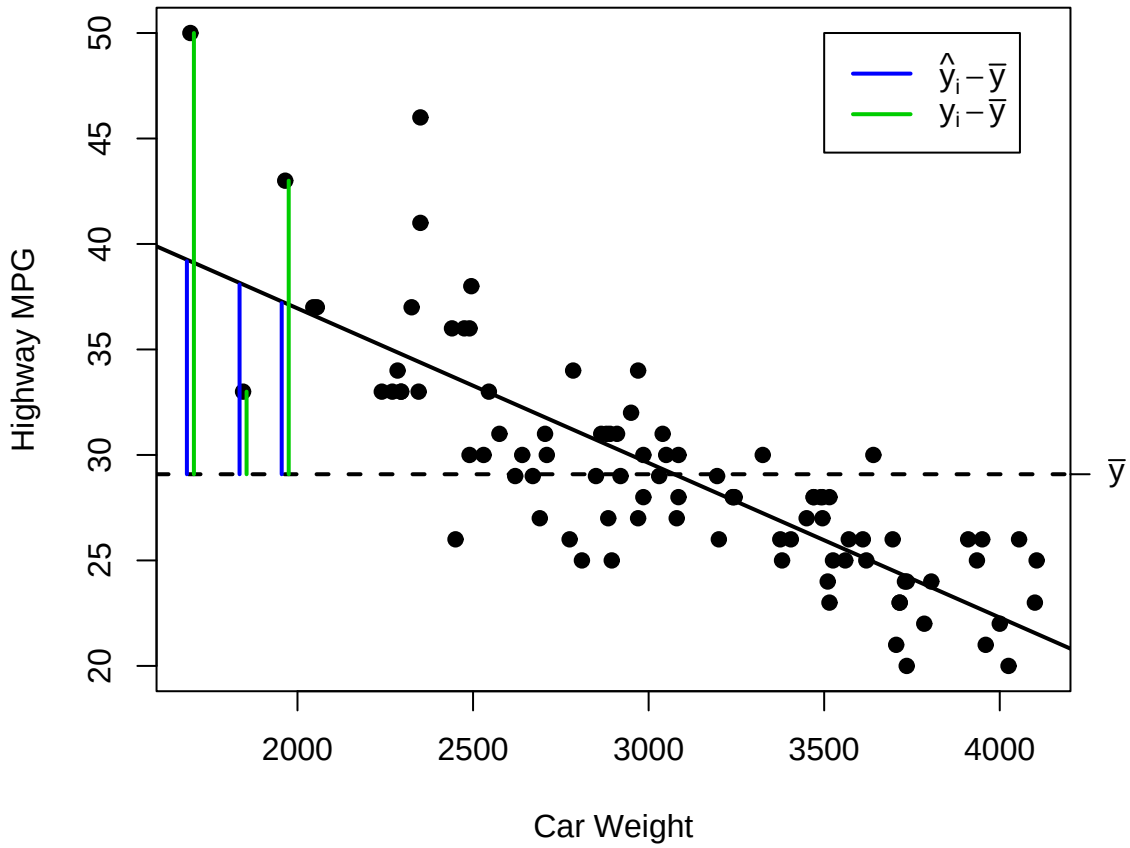
are not the same. In fact it can be shown that $bd = r^2 \leq 1$. If they were the same relationship, this product would have to be one.



r^2 in regression

There is another tie-in between regression and correlation. The square of the correlation coefficient, r^2, is the fraction of variation in the values of y that can be explained by the least squares regression of y on x.

$$
\begin{aligned}
r^2 &= \frac{\text{variance of } \hat{y}_i\text{'s}}{\text{variance of } y_i\text{'s}} \\
&= \frac{\frac{1}{n-1}\sum(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\sum(y_i - \bar{y})^2} \\
&= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\
&= \frac{SSReg}{SSTot}
\end{aligned}
$$



```
. 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

------------------------------------------------------------------------------
 HighMPG |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |  -.0073271   .0005548   -13.21   0.000     -.008429   -.0062251
   _cons |   51.60137    1.73555    29.73   0.000     48.15391    55.04882
------------------------------------------------------------------------------
```

So for this example, the weight of the cars explains 65.3% of the variability in the highway MPG observations

Why least squares?

Mathematically tractable - optimizing a quadratic is easy

It has good statistical properties (e.g. small variances for predictions under certain modelling assumptions - Gauss/Markov Theorem)

An alternative is to minimize

$$\sum |y_i - \hat{y}_i| = \sum |y_i - a - bx_i|$$

This is a difficult function to optimize since the function $|x|$ is not differentiable at $x = 0$.

While this is more difficult to do, due to increased computing power, some people fit linear models this way (L1 regression).

Tying this to univariate ideas, least squares is like finding means, and L1 regression is like finding medians.