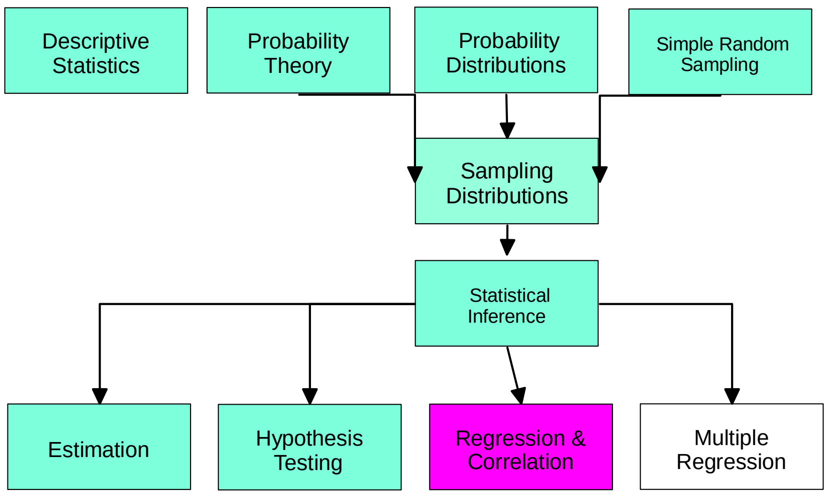
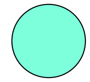
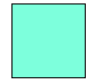
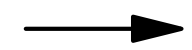
The Course So Far:

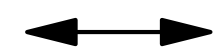


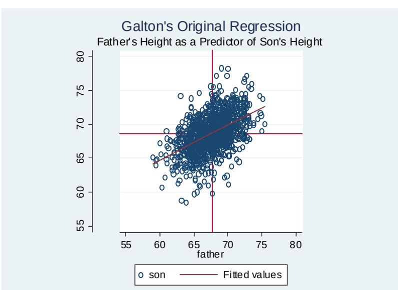
Symbols of Causal Analysis:

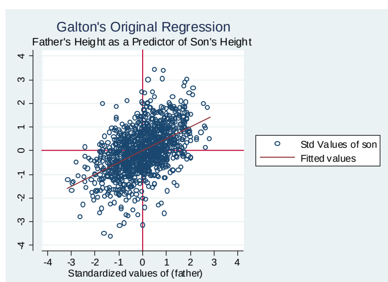

Circles representing theoretical (or latent) variables that may be causes or effects (or both) in our theory.


Squares representing empirical measures of latent variables. These are variables actually measured in the data.


Single-headed arrows representing causal effects from one variable to another.


Double-Headed Arrows representing correlation or covariance between variables but not causation between them.

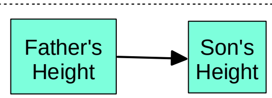


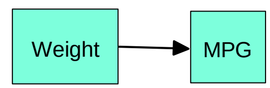



```
. summarize  father son sonstd fatherstd

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      father |      1078     67.6871    2.744869     59.008   75.43393
         son |      1078    68.68407    2.814701   58.50708   78.36479
      sonstd |      1078    5.40e-10           1  -3.615655   3.439342
   fatherstd |      1078    5.75e-11           1  -3.161936   2.822296
```

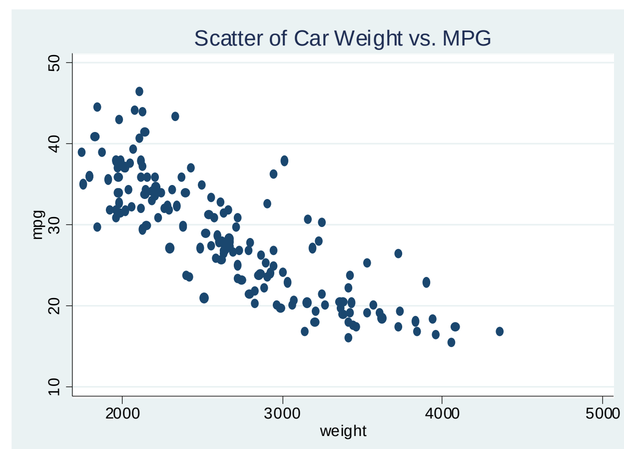
```
. regress   son father, beta

      Source |       SS       df       MS              Number of obs =    1078
-------------+------------------------------           F(  1,  1076) =  361.23
       Model |  2144.57973     1  2144.57973           Prob > F      =  0.0000
    Residual |  6388.00066  1076  5.93680359           R-squared     =  0.2513
-------------+------------------------------           Adj R-squared =  0.2506
       Total |  8532.58039  1077  7.92254447           Root MSE      =  2.4366

------------------------------------------------------------------------------
         son |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      father |    .514093   .0270487    19.01   0.000                 .5013383
       _cons |   33.88661   1.832354    18.49   0.000                        .
------------------------------------------------------------------------------
```

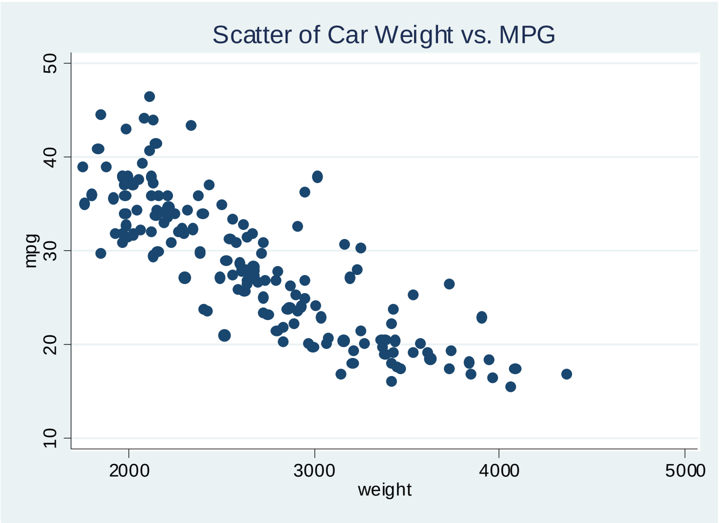
Francis Galton (1822 -1911)
The Inventor of Modern Regression Analysis





graph twoway scatter mpg weight, title("Scatter of Car Weight vs. MPG")





- Assume that we are dealing with a single relationship and that it contains only two variables:

$$Y = f(X) \rightarrow MPG = f(WGT)$$

- Choose the functional form of the relationship between Y and X:

$$Y = \alpha + \beta X \rightarrow MPG = \alpha + \beta \bullet WGT \text{ (Linear Equation)}$$

Some other possibilities are:

$$Y = \alpha e^{\beta X} \text{ which implies: } \log_e Y = \log_e \alpha + \beta X$$

$$Y = \alpha X^{\beta} \text{ which implies: } \log_e Y = \log_e \alpha + \beta \log_e X$$

These two forms, which are nonlinear can be transformed by taking natural logs of both sides. Then, the resulting logged equations are linear in the logs.

Here's another form that is linear in Y and $1/X$:

$$Y = \alpha + \beta \frac{1}{X}$$

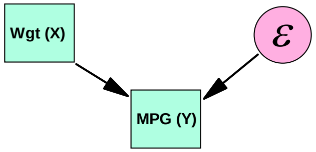
$$Y = \alpha + \beta X$$

$$Y = \alpha + \beta X + \varepsilon$$

Reasons for inserting a stochastic error term:

- The error term contains all the information that, if we knew it, would allow us to <u>completely</u> explain variation in *Y.*
- There are random errors of observation or measurement
- Over and above the total effect of all relevant factors, there is a basic and unpredictable element of randomness in human responses which can be adequately characterized only by the inclusion of a random error term.

The Complete Model

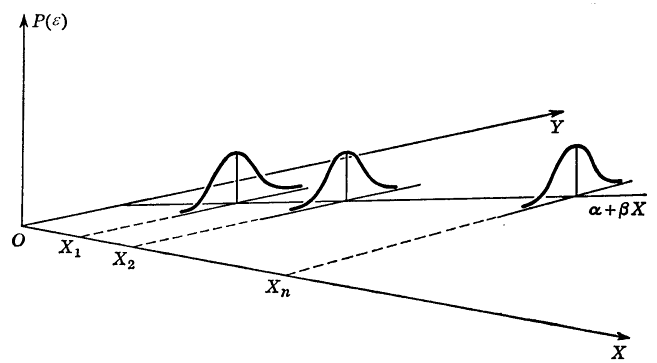


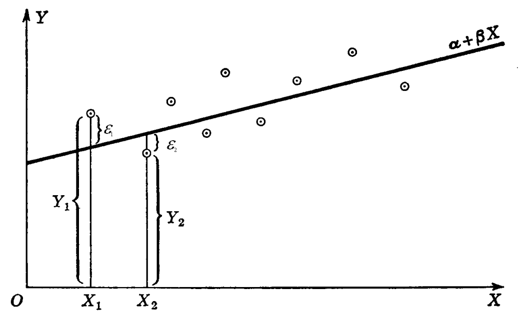
Assumptions about the stochastic model:

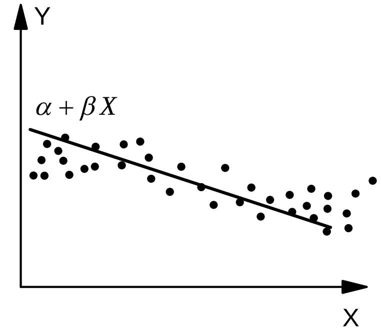
$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

$$E[\varepsilon_i \mid X] = E[\varepsilon_i] = 0 \text{ for all } i$$

$$E[\varepsilon_i \varepsilon_j] = \begin{cases} 0 & \text{for } i \neq j;\ i, j = 1, 2, \ldots, n \\ \sigma_\varepsilon^2 & \text{for } i = j;\ i, j = 1, 2, \ldots, n \end{cases}$$







$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

$$E\left[\varepsilon_i\right] = 0 \qquad \text{for all } i$$

$$E\left[\varepsilon_i \varepsilon_j\right] = \begin{cases} 0 & \text{for } i \neq j;\ i, j = 1, 2, \ldots, n \\ \sigma_\varepsilon^2 & \text{for } i = j;\ i, j = 1, 2, \ldots, n \end{cases}$$

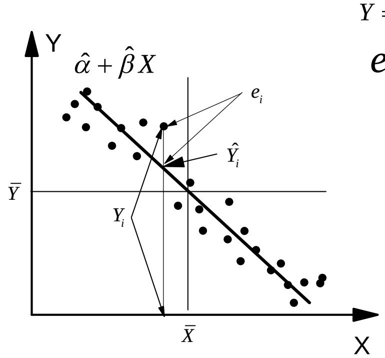
Least Squares Estimators

The Data: $X_1 \quad X_2 \ldots \quad X_n$ / $Y_1 \quad Y_2 \ldots \quad Y_n$ Arithmetic Means $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Denote the estimated line through the data as: $\hat{Y} = \hat{\alpha} + \hat{\beta} X$

$\hat{\alpha}, \hat{\beta}$ = estimates of two unknown parameters

$\hat{Y}$ = estimated value of Y for any X



e_i = the difference between the actual and estimated values of Y.

or,

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

Consequently, our goal is to minimize:

$$\sum_{i=1}^{n} e_i^2 = f\left(\hat{\alpha}, \hat{\beta}\right)$$

Derivation of Least Squares Estimators for α and β

Minimize:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad \text{or, substituting,} \quad = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2$$

To minimize the sum of squared deviations, a necessary condition is that the partial derivatives of the sum with respect to $\hat{a}$ and $\hat{\beta}$ should both be zero.
We thus have:

$$\frac{\partial}{\partial \hat{\alpha}} \sum_{i=1}^{n} e_i^2 = -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right) = 0$$

$$\frac{\partial}{\partial \hat{\beta}} \sum_{i=1}^{n} e_i^2 = -2\sum_{i=1}^{n} X_i\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right) = 0$$

Simplifying these two equations gives the standard form of the normal equations for a straight line:

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i$$

Note that dividing this equation by *n* gives $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$ which means that least-squares estimates are such that the estimated line passes through the point of means ($\bar{X}$, $\bar{Y}$).

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2$$

Now, we can subtract the mean of Y from both sides of the original equation:

$$\hat{Y} - \overline{Y} = \hat{\alpha} + \hat{\beta} X - \overline{Y} = \hat{\alpha} + \hat{\beta} X - \hat{\alpha} - \hat{\beta}\overline{X} = \hat{\beta}\left(X - \overline{X}\right)$$

Let's let lower case letters denote deviations from the means, so that

$$x_i = X_i - \overline{X} \qquad y_i = Y_i - \overline{Y} \qquad \hat{y}_i = \hat{Y}_i - \overline{Y}$$

So we can write the least squares line equation as: $\hat{y} = \hat{\beta} x$

And the residual e_i may be indicated by:

$$e_i = Y_i - \hat{Y}_i \rightarrow Y_i - \overline{Y} - \left(\hat{Y}_i - \overline{Y}\right) = y_i - \hat{y}_i = y_i - \hat{\beta} x_i$$

Now, we can rewrite our sum of squared residuals as:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta} x_i\right)^2$$

Minimizing this expression with respect to $\hat{\beta}$ gives:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

And, we can find $\hat{\alpha}$ by remembering that the regression line passes through the point of means, namely,

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

$$VAR[X] = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

$$COV[X,Y] = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

$$\hat{\beta} = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{COV[X,Y]}{VAR[X]}$$

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

$$E[\varepsilon_i] = 0 \qquad \text{for all } i$$

$$E\left[\varepsilon_i \varepsilon_j\right] = \begin{cases} 0 & \text{for } i \neq j;\ i, j = 1, 2, \ldots, n \\ \sigma_\varepsilon^2 & \text{for } i = j;\ i, j = 1, 2, \ldots, n \end{cases}$$

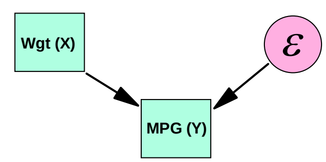
$$E\left[\hat{\beta}\right] = \beta + \sum_{i=1}^{n}\left(\frac{x_i}{\sum_{i=1}^{n} x_i^2}\right) E[e_i] = \beta$$ The least squares estimator of the slope is unbiased

$$E[\hat{\alpha}] = \alpha + \sum\left(\frac{1}{n} - \overline{X}\frac{x_i}{\sum x_i^2}\right) E[e_i] = \alpha$$ The least squares estimator of the intercept is unbiased

Turchi said that if the assumptions about the error term hold, then the least squares estimators of the slope and intercept terms are unbiased estimators. Trust him on this.

Moreover, these estimators are **best linear unbiased** estimators. That is they are *efficient* (they have the smallest variance of any linear estimator).

The Complete Model



The Model: $Y_i = \alpha + \beta x_i + \varepsilon_i \rightarrow MPG_i = \alpha + \beta \bullet wgt_i + \varepsilon_i$

That is, mileage is determined by the weight of the car and some random, but uncorrelated factors whose effect for each observation are contained in the error term, ε_i

```
. regress mpg weight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  5723.60104     1  5723.60104           Prob > F      =  0.0000
    Residual |  2603.15246   152   17.126003           R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.1384

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0101428   .0005548   -18.28   0.000     -.011239   -.0090467
       _cons |   55.89712    1.51963    36.78   0.000      52.8948    58.89944
------------------------------------------------------------------------------
```

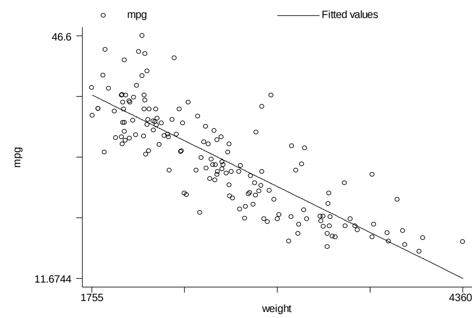
$$\widehat{MPG} = \hat{\alpha} + \hat{\beta} \bullet WGT \Rightarrow \widehat{MPG} = 55.89712 - 0.0101428 \bullet WGT$$

$$\frac{d\,\widehat{MPG}}{dWGT} = \hat{\beta} = -0.0101428$$

That is, if the weight of an automobile rises by one pound, miles per gallon will fall by 1/100. Or, if the weight rises by 1,000 pounds, MPG will fall by 10.14.

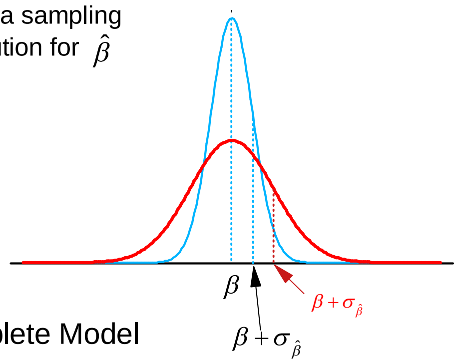
$$\widehat{MPG}_{WGT=0} = 55.89712$$

```
predict mpghat

graph  mpg mpghat weight,
connect(.s) symbol(oi)
```

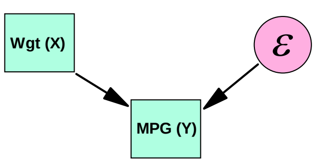


The regression coefficients $\hat{\beta}$ and $\hat{\alpha}$ are random variables, each with their own sampling distribution.

Here's a sampling distribution for $\hat{\beta}$



The Complete Model



The only source of random behavior, apart from Wgt is the error term. So, it's the distribution and variance of the error term that determines the distribution of $\hat{\beta}$ and its standard error.

$$\sigma_{\hat{\beta}} = \frac{s_\varepsilon}{\sqrt{TSS_X}} \text{ where } TSS_x = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \quad \text{and}$$

$$s_\varepsilon = \sqrt{\frac{RSS}{n-K}} \text{ where } RSS = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

and n is the sample size and K is the number of estimated parameters (2 in this case:) $\hat{\alpha}$ and $\hat{\beta}$.

Now, *if the error term, s_ε, is normally, identically and independently distributed*, we can form the *t-statistic:*

$$t = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}}$$

and test hypotheses about the true slope coefficient.

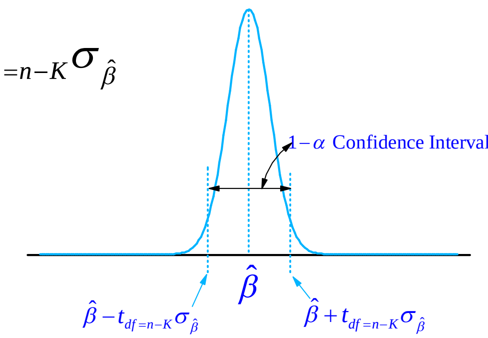
$$t = \frac{\hat{\beta} - 0}{\sigma_{\hat{\beta}}} \Rightarrow \begin{array}{l} H_o : \beta = 0 \\ H_a : \beta \neq 0 \end{array} \qquad \sigma_{\hat{\alpha}} = s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{TSS_X}}$$

$$t = \frac{\hat{\alpha} - 0}{\sigma_{\hat{\alpha}}} \Rightarrow \begin{array}{l} H_o : \alpha = 0 \\ H_a : \alpha \neq 0 \end{array}$$

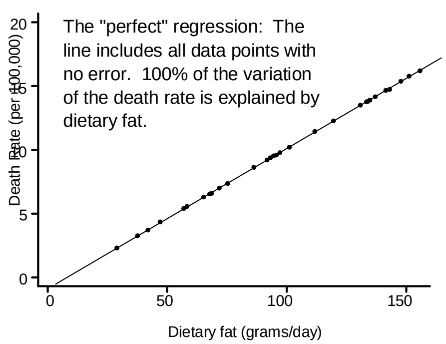
```
. regress mpg weight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  5723.60104     1  5723.60104           Prob > F      =  0.0000
    Residual |  2603.15246   152   17.126003           R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.1384

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0101428   .0005548   -18.28   0.000    -.011239   -.0090467
       _cons |   55.89712    1.51963    36.78   0.000     52.8948    58.89944
------------------------------------------------------------------------------
```

Confidence Intervals:

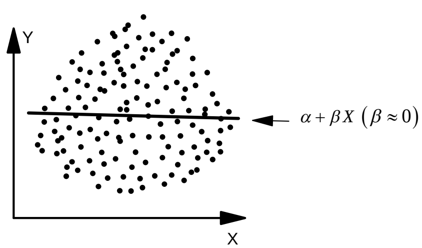
$$\hat{\beta} \pm t_{df=n-K} \sigma_{\hat{\beta}}$$




```
regress mpg wgt, level(90)
```



$$R^2 = 1.0$$



and $R^2 = 0.0$

Total Sum of Squares (TSS) is the sum of squared deviations of the dependent variable around its mean and is a measure of the total variability of the variable:

$$TSS_Y = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$$

Explained (or Model) Sum of Squares (ESS) is the sum of squared deviations of *predicted* values of *Y* around its mean:

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y} \right)^2 = \sum_{i=1}^{n} \left(Y - \hat{\alpha} - \hat{\beta} X \right)^2$$

Remember, it's RSS that least squares regression seeks to minimize.

$R^2 =$ explained variance/total variance

$$= \frac{s_{\hat{Y}}^2}{s_Y^2}$$

$$= \frac{ESS}{TSS_Y}$$

```
. regress mpg weight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  5723.60104     1  5723.60104           Prob > F      =  0.0000
    Residual |  2603.15246   152   17.126003           R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.1384

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0101428   .0005548   -18.28   0.000     -.011239   -.0090467
       _cons |   55.89712    1.51963    36.78   0.000      52.8948    58.89944
------------------------------------------------------------------------------
```

$$s_Y^2 = \frac{\left(\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

$$s_X^2 = \frac{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)}{n-1}$$

Variance formulas

$$s_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Covariance formula

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}}{\sqrt{\dfrac{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\right)^2}{n-1}} \bullet \sqrt{\dfrac{\left(\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\right)^2}{n-1}}}$$

Correlation Coefficient

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

The *correlation coefficient, r,* is a standardized measure of a *bivariate linear* relationship between two variables, *X,* and *Y.*

- **Negative:** *high values* of *Y* tend to occur with *low values* of *X*, and low *Y* with high *X.*
- **Positive:** *high values* of *Y* tend to occur with *high values* of *X,* and low *Y* with low *X.*

```
. correlate mpg weight
(obs=154)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8291   1.0000
```

r_{XY}

Claim: the bivariate correlation coefficient is simply the regression slope coefficient when one *standardized* variable is regressed on another *standardized* variable.

First, standardize *mpg & weight:*

```
egen stmpg=std(mpg)
egen stweight=std(weight)
```

Secondly, check their means and st. deviations*:*

```
. summarize stmpg stweight

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------
       stmpg |     154   -3.38e-10           1  -1.801969   2.413716
    stweight |     155   -6.80e-09           1   -1.52712   2.806283
```

Third: Regress *stmpg* on *stweight:* `regress stmpg stweight`

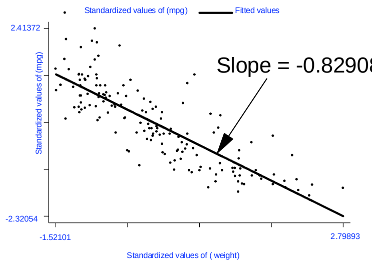
```
. regress  stmpg stweight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  105.168352     1  105.168352           Prob > F      =  0.0000
    Residual |  47.8316479   152  .314681894           R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |      153.00   153  .999999997           Root MSE      =  .56097

------------------------------------------------------------------------------
       stmpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    stweight |  -.8290807   .0453513   -18.28   0.000    -.9186811   -.7394803
       _cons |   3.96e-09   .0452039     0.00   1.000     -.089309     .089309
------------------------------------------------------------------------------
```

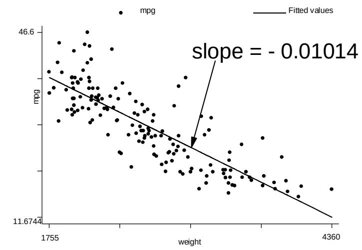
Fourth: Create a predicted version of *stmpg:*

```
. predict stmpghat
```

Fifth: Graph actual and predicted values of stmpg:

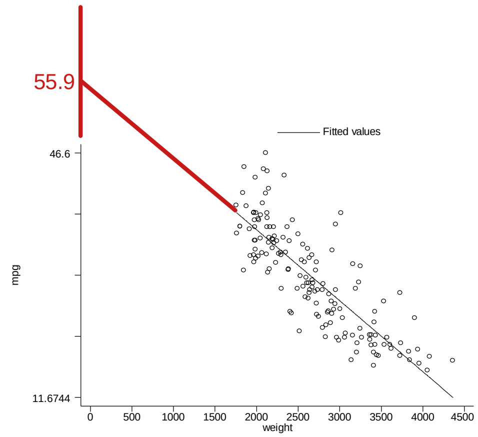
```
graph stmpg stmpghat stweight, connect(.s) symbol(oi)
```





$$\hat{\beta}^* = \hat{\beta}\frac{s_X}{s_Y} \quad \Rightarrow \quad \hat{\beta}^* = \hat{\beta}\frac{s_X}{s_Y} = r_{XY}$$

```
graph  mpg mpghat weight, connect(.s) symbol(Oi) xlabel(0 500 to 4500)
```



In bivariate regression only, the standardized regression coefficient equals the correlation coefficient.

```
regress mpg  weight,beta
```

```
. regress mpg  weight,beta

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  5723.60104     1  5723.60104           Prob > F      =  0.0000
    Residual |  2603.15246   152   17.126003           R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.1384

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      weight |  -.0101428   .0005548   -18.28   0.000                -.8290807
       _cons |   55.89712    1.51963    36.78   0.000                        .
------------------------------------------------------------------------------
```

$$\hat{\beta}^* = \hat{\beta}\frac{s_X}{s_Y} = r_{XY} = \sqrt{R^2}$$

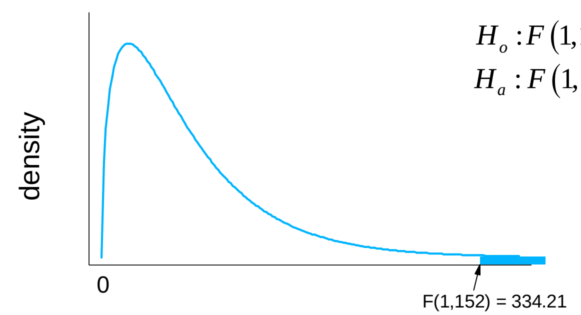
All of these equalities hold only for bivariate regression & correlation

Things will be a little more complicated in multiple regression

$$F = \frac{ESS/(K-1)}{RSS/(n-K)} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 / (K-1)}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / (n-K)}$$

Where K = number of estimated parameters (2, slope and intercept)
n = sample size

$$F = \frac{ESS/(K-1)}{RSS/(n-K)} = \frac{ESS/df_1}{RSS/df_2}$$



$$H_o: F(1,152) = 0$$
$$H_a: F(1,152) > 0$$

ESS/df_1

RSS/df_2

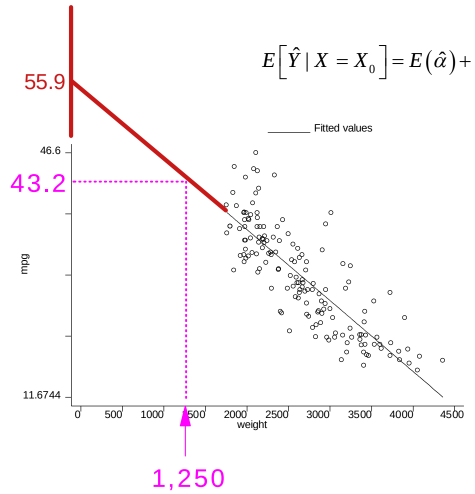
F(1,152)

```
. regress mpg weight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  334.21
       Model |  5723.60104     1  5723.60104           Prob > F      =  0.0000
    Residual |  2603.15246   152  17.126003            R-squared     =  0.6874
-------------+------------------------------           Adj R-squared =  0.6853
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.1384

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0101428   .0005548   -18.28   0.000     -.011239   -.0090467
       _cons |   55.89712    1.51963    36.78   0.000      52.8948    58.89944
------------------------------------------------------------------------------
```

Prob > F

```
graph  mpg mpghat weight, connect(.s) symbol(Oi) xlabel(0 500 to 4500)
```

$$E\left[\hat{Y} \mid X = X_0\right] = E(\hat{\alpha}) + X_0 E\left[\hat{\beta}\right] = \alpha + \beta X_0 = E\left[Y \mid X = X_0\right]$$



$$E\left[\hat{Y} \mid X = 1,250\right] = \hat{\alpha} + 1,250 \bullet \hat{\beta} = 55.89712 - 0.0101428 \bullet 1,250 = 43.2$$

$$\hat{Y}_{X_0} = \hat{\alpha} + \hat{\beta} X_0 \text{ for } X_0 = X_{3,000}$$

$$\hat{Y}_{X_0} - t_{\left(\frac{\alpha}{2}, n-2\right)} s_{\hat{Y}}, \quad \hat{Y}_{X_0} + t_{\left(\frac{\alpha}{2}, n-2\right)} s_{\hat{Y}}$$

$$s_{\hat{Y}_{X_0}} = s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{XX}}}$$

The confidence interval for the mean estimate of $\hat{Y}_{X_0}$ is:

$$\hat{Y}_{X_0} - t_{\left(\frac{\alpha}{2}, n-2\right)} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{XX}}}, \quad \hat{Y}_{X_0} + t_{\left(\frac{\alpha}{2}, n-2\right)} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{XX}}}$$

width of the confidence interval depends upon distance from the mean.

Derivation of: $Var\left(\hat{Y}_{X_0}\right)$

$$Var\left(\hat{Y}_{X_0}\right) = E\left[\hat{Y}_{X_0} - E\left(\hat{Y} \mid X_0\right)\right]^2$$
$$= E\left[\hat{\alpha} + \hat{\beta} X_0 - \alpha - \beta X_0\right]^2 = E\left[(\hat{\alpha} - \alpha) + X_0\left(\hat{\beta} - \beta\right)\right]^2$$
$$= Var(\hat{\alpha}) + X_0^2 Var\left(\hat{\beta}\right) + 2X_0 Cov\left(\hat{\alpha}, \hat{\beta}\right)$$

Noting that:

$$Var\left(\hat{\beta}\right) = s_{\hat{\beta}}^2 = E\left[\left(\hat{\beta} - \beta\right)^2\right] = \frac{s_{\varepsilon}^2}{S_{XX}}$$
$$Var(\hat{\alpha}) = s_{\hat{\alpha}}^2 = E\left[(\hat{\alpha} - \alpha)^2\right] = \frac{\sum X_i^2}{nS_{XX}} s_{\varepsilon}^2$$
$$Cov\left(\hat{\alpha}, \hat{\beta}\right) = s_{\hat{\alpha}\hat{\beta}} = E\left[(\hat{\alpha} - \alpha)\left(\hat{\beta} - \beta\right)\right] = -\frac{\bar{X}}{S_{XX}} s_{\varepsilon}^2$$
$$S_{XX} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

substituting into the variance formula gives:

$$Var\left(\hat{Y}_{X_0}\right) = s_{\varepsilon}^2 \left[\frac{\sum_{i=1}^{n} X_i^2}{nS_{XX}} + X_0^2 \frac{1}{S_{XX}} - 2\frac{X_0 \bar{X}}{S_{XX}}\right]$$

and noting that : $\sum_{i=1}^{n} X_i^2 = S_{XX} + n\bar{X}^2$

and substituting, we get:

$$Var\left(\hat{Y}_{X_0}\right) = s_{\varepsilon}^2\left[\frac{1}{n} + \frac{\bar{X}^2 + X_0^2 - 2X_0\bar{X}}{S_{XX}}\right] = s_{\varepsilon}^2\left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{XX}}\right] \Rightarrow s_{\hat{Y}_{X_0}} = s_{\varepsilon}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{XX}}}$$

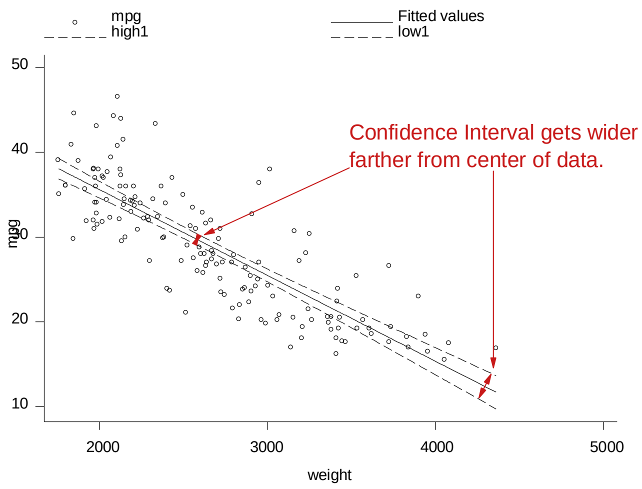
```
predict mpghat
```
(computes predicted values of Y)

```
predict SEmpghat, stdp
```
(computes $s_{\hat{Y}_{X_0}}$)

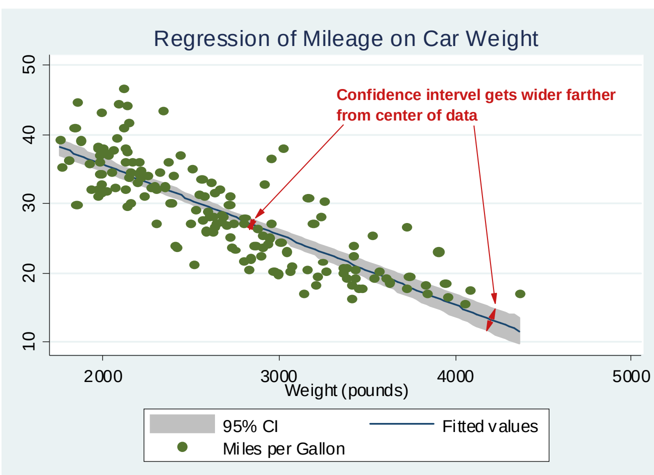
```
display invttail(df, .05/2)
```
(computes *t-value where df* = *n-K*)

```
display invttail(152,.05/2)-> 1.9756
-> 1.98

generate low1= mpghat - 1.98* SEmpghat
generate high1= mpghat + 1.98* SEmpghat
```

```
graph  mpg mpghat high1 low1 weight, symbol(oiii)
connect(.s s[-] s[-]) xlabel ylabel
```
(Stata 7)



graph twoway lfitci mpg weight || scatter mpg weight,
ti("Regression of Mileage on Car Weight")



$$E\left[Y\right] = \hat{\alpha} + \hat{\beta} X$$

Random part

$$Y_o = E\left[Y_o\right] + \varepsilon_o = \underbrace{\alpha + \beta X_o}_{} + \varepsilon_o$$

Systematic (non-random) part

$$E\left[\hat{Y} \mid X = X_0\right] = E\left(\hat{\alpha}\right) + X_0 E\left[\hat{\beta}\right] = \alpha + \beta X_0 = E\left[Y \mid X = X_0\right]$$

$\hat{\varepsilon}_0 = Y_{X_0} - \hat{Y}_{X_0}$ the forecast error, where

$$E\left(\hat{\varepsilon}_o\right) = \alpha + \beta X_o + E\left(\varepsilon_o\right) - \left[E\left(\hat{\alpha}\right) + E\left(\hat{\beta}\right) X_o\right]$$
$$= \alpha + \beta X_o + 0 - \left[\alpha + \beta X_o\right]$$
$$= 0$$

which implies that $\hat{Y}_0$ is an unbiased estimator of Y_0

$$Var\left(\hat{\varepsilon}_0\right)=Var\left(Y_{X_0}\right)+Var\left(\hat{Y}_{X_0}\right)-2Cov\left(Y_{X_0},\hat{Y}_{X_0}\right)\Rightarrow$$

$$Var\left(\hat{\varepsilon}_0\right)=Var\left(Y_{X_0}\right)+Var\left(\hat{Y}_{X_0}\right)\text{ because }Cov\left(Y_{X_o},\hat{Y}_{X_o}\right)=0$$

$$Var\left(\hat{\varepsilon}_0\right)=Var\left(Y_{X_0}\right)+Var\left(\hat{Y}_{X_0}\right)$$

$$Var\left(\hat{Y}_{X_0}\right)=s_\varepsilon^2\left[\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}\right]$$

$$Var\left(\hat{\varepsilon}_o\right)=Var\left(Y_{X_0}\right)+s_\varepsilon^2\left[\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}\right]$$

$$Var\left(Y_{X_0}\right)=Var(\alpha+\beta X_o+\varepsilon_o)=Var(\alpha+\beta X_o)+Var(\varepsilon_o)=0+s_\varepsilon^2=s_\varepsilon^2$$

$$Var\left(\hat{\varepsilon}_o\right)=s_\varepsilon^2+s_\varepsilon^2\left[\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}\right]=s_\varepsilon^2\left[1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}\right]$$

The variance of the forecast error

$$\hat{Y}_{X_0}-t_{\left(\frac{\alpha}{2},n-2\right)}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}},\quad \hat{Y}_{X_0}+t_{\left(\frac{\alpha}{2},n-2\right)}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}}$$

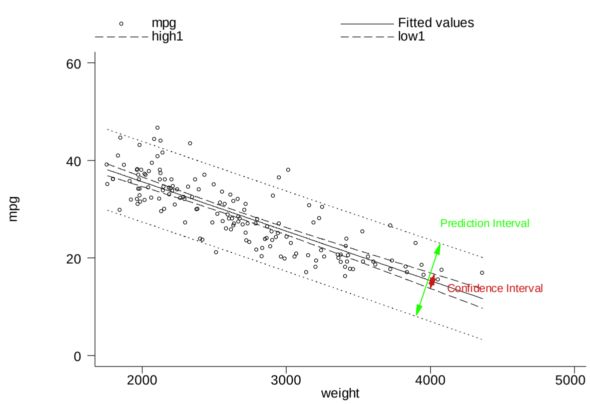
This interval is always wider because of these 1s.

$$\hat{Y}_{X_0}-t_{\left(\frac{\alpha}{2},n-2\right)}s_\varepsilon\sqrt{\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}},\quad \hat{Y}_{X_0}+t_{\left(\frac{\alpha}{2},n-2\right)}s_\varepsilon\sqrt{\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}}$$

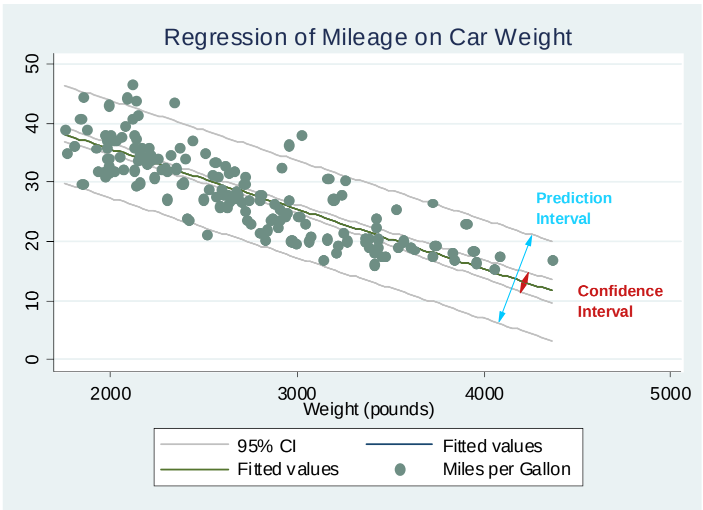
$$s_{\hat{\varepsilon}_0}=s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{S_{XX}}}$$

```
predict SEmpghat2, stdf
generate low2= mpghat - 1.98* SEmpghat2
generate high2= mpghat + 1.98* SEmpghat2
```

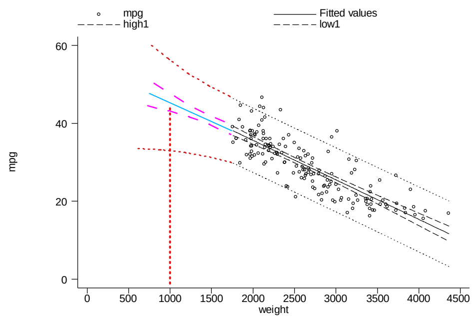
```
graph  mpg mpghat high1 low1 high2 low2 weight, symbol(oiiiii)
connect(.s s[-] s[-] s[.] s[.]) xlabel ylabel (Stata 7)
```

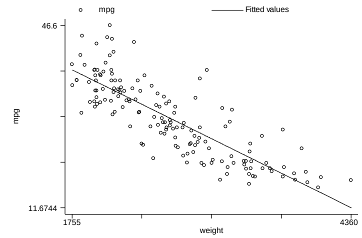


graph twoway lfitci mpg weight, ciplot(rline) || lfitci mpg weight, stdf ciplot(rline) || scatter mpg weight, ti("Regression of Mileage on Car Weight")

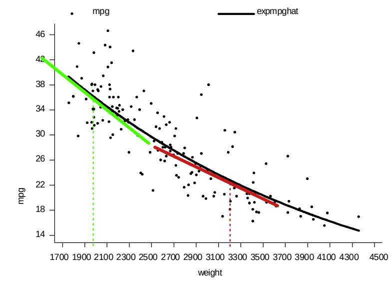


```
graph  mpg mpghat high1 low1 high2 low2 weight, symbol(oiiiii) connect(.s
s[-] s[-] s[.] s[.]) xlabel(0 500 to 4500) ylabel
```





$MPG = \alpha + \beta \bullet WGT$ (Linear Equation)

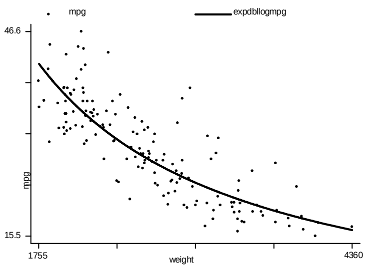


$Y = \alpha e^{\beta X}$ which implies: $\log_e Y = \log_e \alpha + \beta X$

```
generate lnmpg = ln(mpg)
```

```
. regress lnmpg weight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  399.54
       Model |  7.90792124     1  7.90792124           Prob > F      =  0.0000
    Residual |  3.00850745   152  .019792812           R-squared     =  0.7244
-------------+------------------------------           Adj R-squared =  0.7226
       Total |  10.9164287   153  .071349207           Root MSE      =  .14069

------------------------------------------------------------------------------
       lnmpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   -.000377   .0000189   -19.99   0.000    -.0004143   -.0003397
       _cons |   4.333275   .0516611    83.88   0.000     4.231208    4.435341
------------------------------------------------------------------------------
```

$Y = \alpha X^{\beta}$ which implies: $\log_e Y = \log_e \alpha + \beta \log_e X$




```
. regress lnmpg lnweight

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  394.90
       Model |   7.8824147     1   7.8824147           Prob > F      =  0.0000
    Residual |  3.03401398   152  .019960618           R-squared     =  0.7221
-------------+------------------------------           Adj R-squared =  0.7202
       Total |  10.9164287   153  .071349207           Root MSE      =  .14128

------------------------------------------------------------------------------
       lnmpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnweight |  -1.025758   .0516181   -19.87   0.000    -1.127739   -.9237759
       _cons |   11.39452   .4061919    28.05   0.000     10.59201    12.19703
------------------------------------------------------------------------------

. predict dbllogmpg
(option xb assumed; fitted values)

. generate expdbllogmpg=exp( dbllogmpg)

. graph  mpg  expdbllogmpg weight, connect(.s) symbol(oi)
```