Applied Regression Analysis

Week 2

1. Linear regression I
2. Linear regression II
3. Assumptions for linear regression
4. Hypothesis testing and confidence intervals
5. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

The Ohio State University

Suppose we have the following observations on systolic blood pressure and age for a sample of 30 individuals:

individual (i)	SBP (y)	AGE (x)	individual (i)	SBP (y)	AGE (x)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

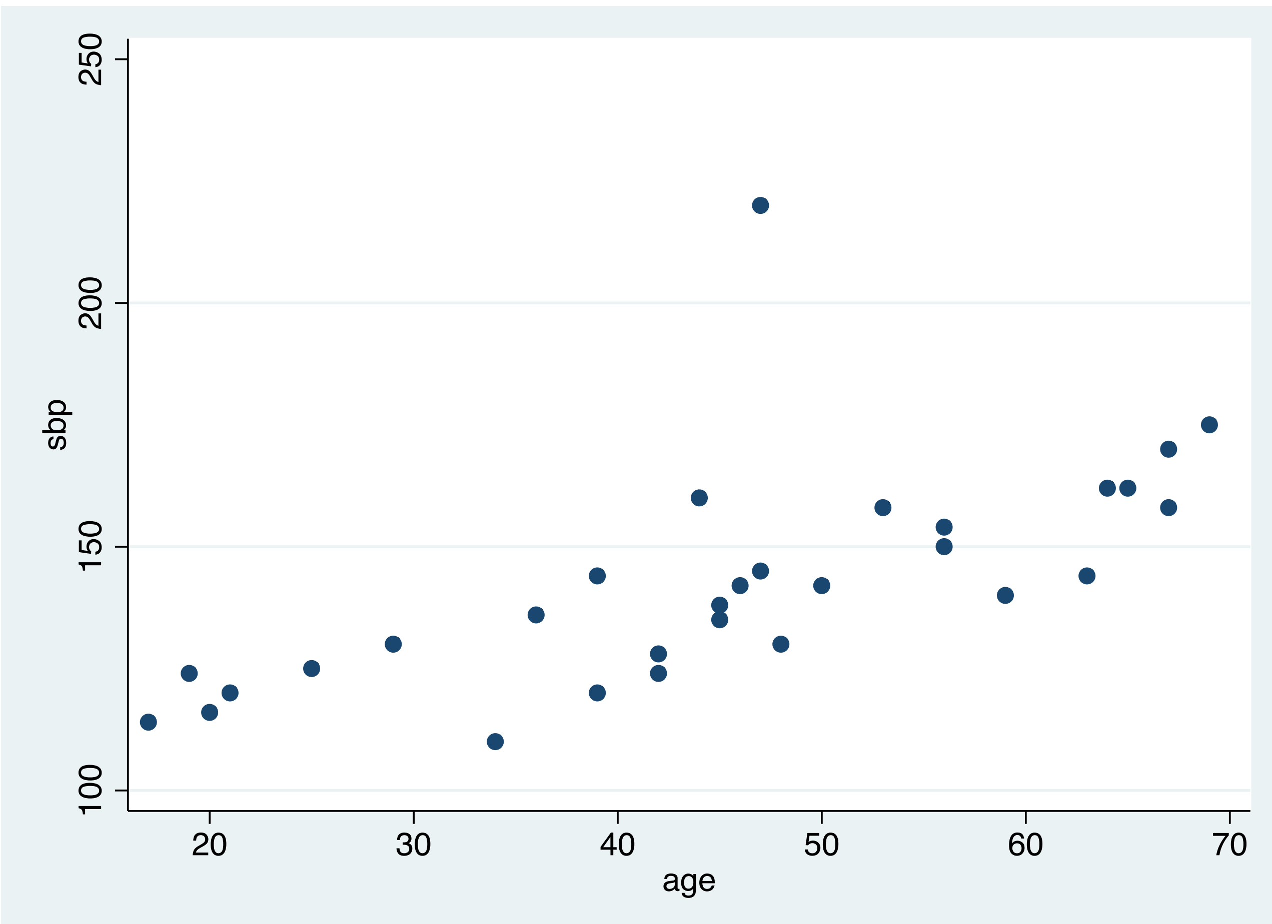
n

note: we have 30 pairs of observations that are denoted by

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30}) = (39, 144), (47, 220), \ldots, (69, 175).$$

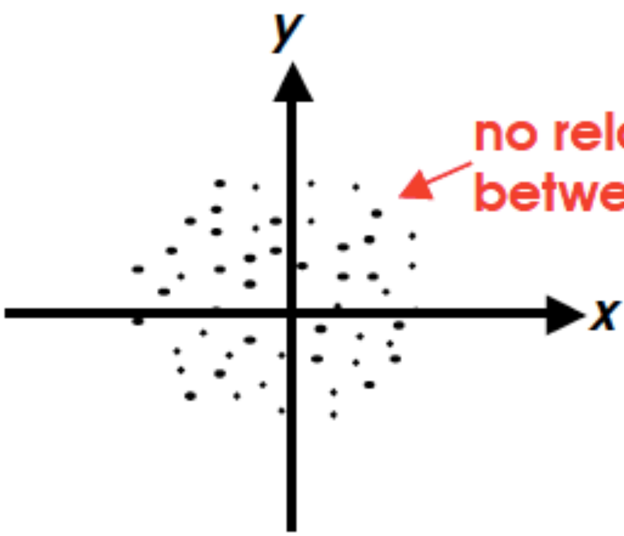
- These pairs may be considered as points in two dimensional space, so that we may plot them on a graph.
  - Such a graph is called a scatter diagram

```
. twoway (scatter sbp age)
```

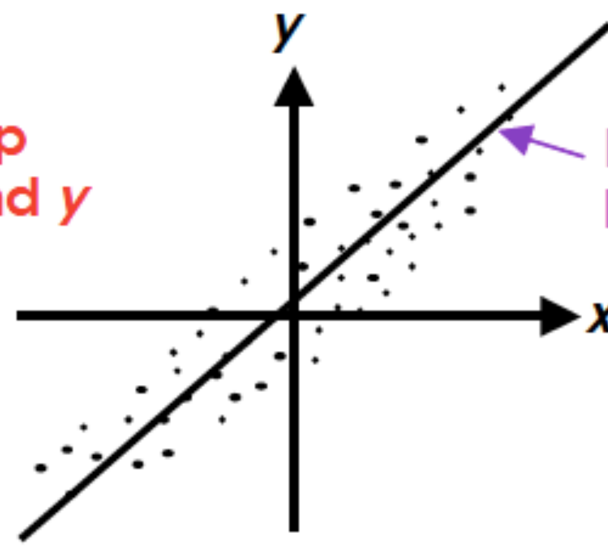


Note: AGE and SBP seem to be related. How can this relationship be measured?

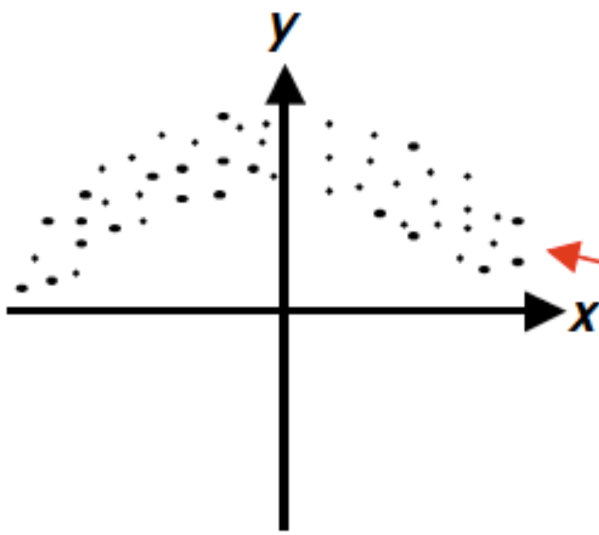
note: Scatter diagrams can take many shapes



no relationship
between *x* and *y*



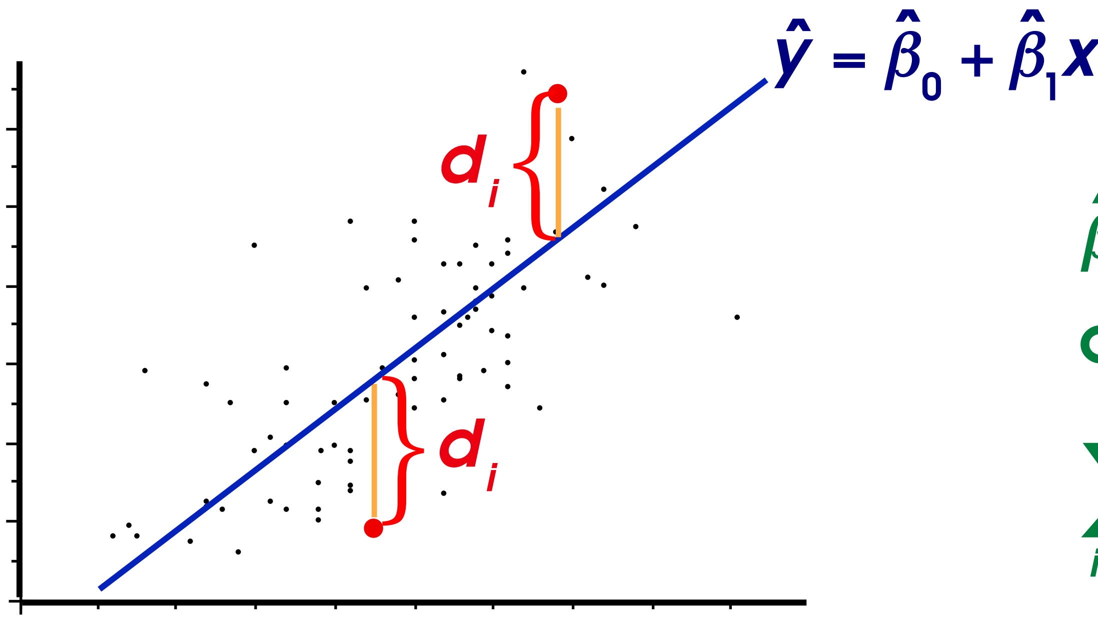
linear relationship
between *x* and *y*



non-linear relationship
between *x* and *y*

Now, given a set of data, how can we determine the line of regression?

- We are looking for that line that minimizes the <u>vertical</u> distances to the data points



$\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen such that $\sum_{i=1}^{n} d_i^2$ is minimized

i.e., we want that line such that minimizes

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

The solution to the best-fit problem is obtained by solving, simultaneously, the following equations:

$$\left.\begin{aligned}\sum y_i &= n\beta_0 + \beta_1 \sum x_i \\ \sum x_i y_i &= \beta_0 \sum x_i + \beta_1 \sum x_i^2\end{aligned}\right\}$$

come from calculus - first take derivative w.r.t. β_0 then w.r.t. β_1 and set equal to zero

Solving for β_0 and β_1 we obtain

$$\hat{\beta}_1 = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{or} \quad \hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

Example

Using the previous data on 30 individuals where we measured

$$x = \text{AGE}$$

$$Y = \text{SBP}$$

computations result in:

$$n = 30$$

$$\bar{y} = 142.53$$

$$\bar{x} = 45.13$$

$$\sum_{i=1}^{n} x_i y_i = 199{,}576$$

$$\sum_{i=1}^{n} x_i = 1{,}354$$

$$\sum_{i=1}^{n} y_i = 4{,}276$$

$$\sum_{i=1}^{n} x_i^2 = 67{,}894$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{199576 - \frac{(1354)(4276)}{30}}{67894 - \frac{(1354)^2}{30}} = 0.97$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 142.53 - (0.97)(45.13) = 98.71$$

Thus, the equation for this straight line is given by

$$\hat{y} = 98.71 + 0.97x$$

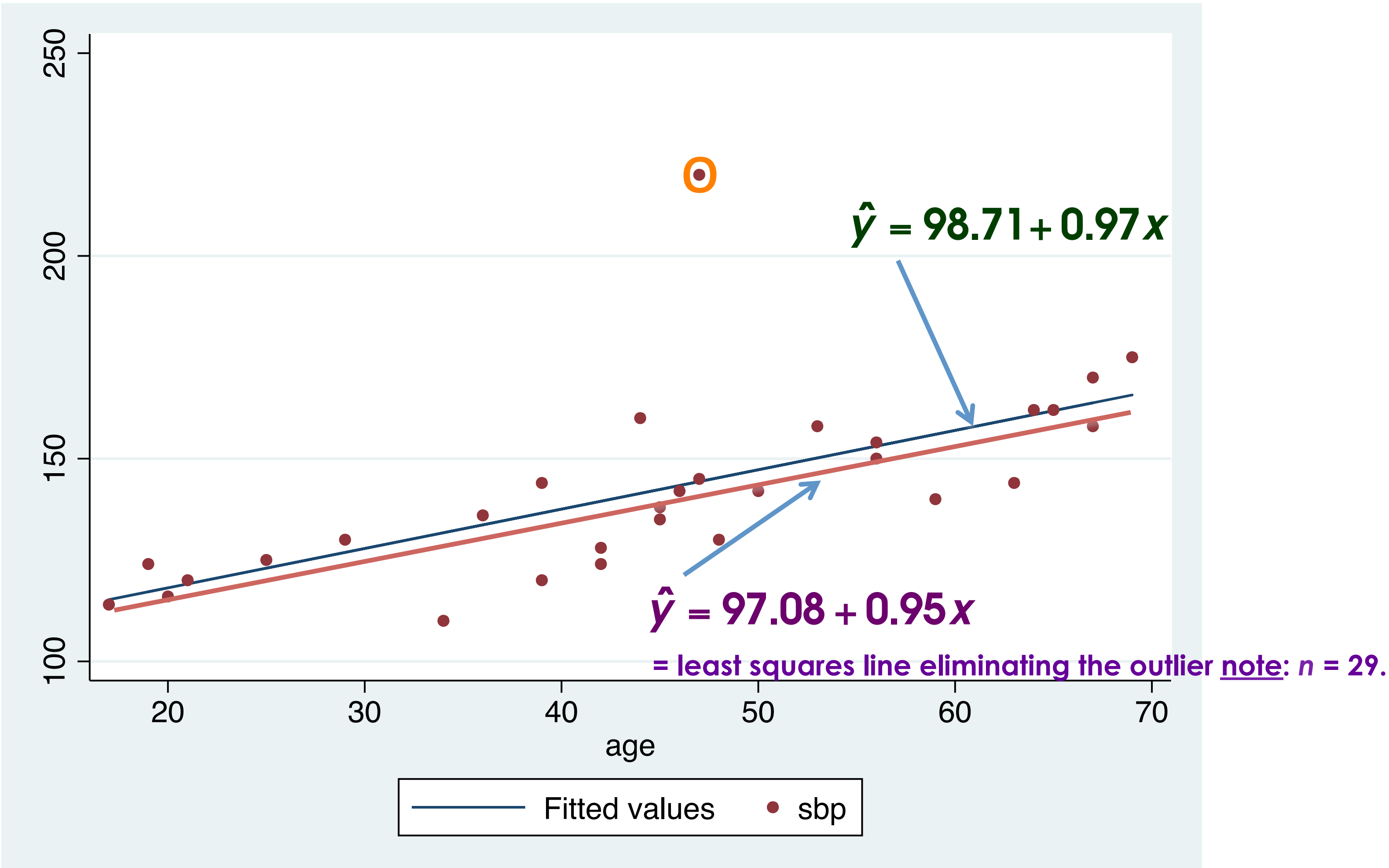
or equivalently by

$$\hat{y} = 142.53 + 0.97(x - 45.13)$$

This line should now be plotted on the scatter diagram.

e.g.,

```
. scatter yhat sbp age, c(l .) s(i o)
```



Now, recall that

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Clearly, if SSE $= 0 \Rightarrow$ perfect fit

i.e., $y_i = \hat{y}_i$, all i

as the fit gets worse, SSE gets larger

```
. sum sbp age, detail
                            sbp
-------------------------------------------------------------
      Percentiles      Smallest
 1%          110            110
 5%          114            114
10%          118            116       Obs                  30
25%          125            120       Sum of Wgt.          30

50%          141                      Mean           142.5333
                        Largest       Std. Dev.      22.58125
75%          158            162
90%          166            170       Variance       509.9126
95%          175            175       Skewness       1.291729
99%          220            220       Kurtosis       5.684303

                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           17             17
 5%           19             19
10%         20.5             20       Obs                  30
25%           36             21       Sum of Wgt.          30

50%         45.5                      Mean           45.13333
                        Largest       Std. Dev.       15.2942
75%           56             65
90%           66             67       Variance       233.9126
95%           67             67       Skewness      -.2395541
99%           69             69       Kurtosis       2.167069
```

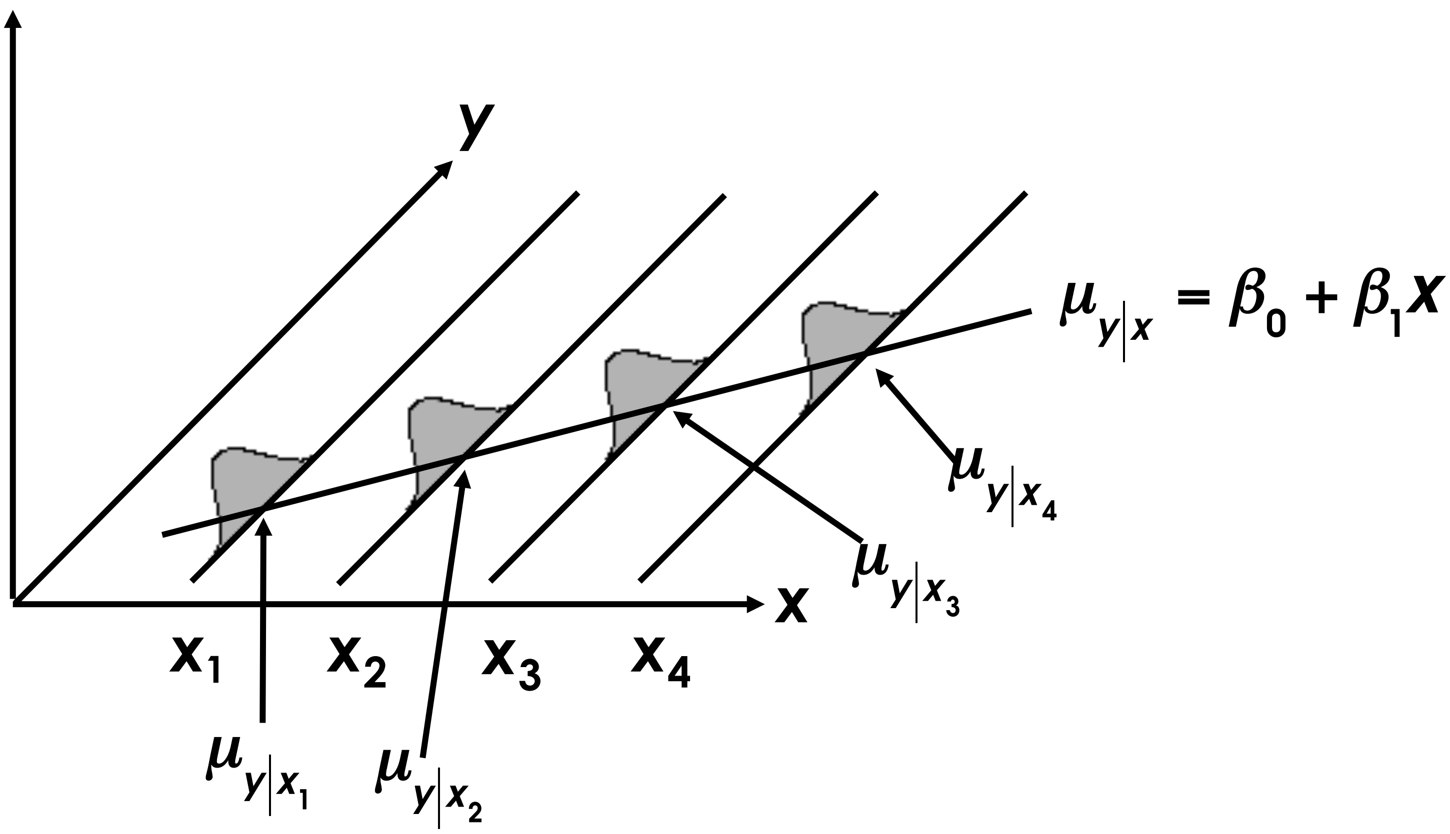
```
. regress sbp age

  Source |       SS       df       MS              Number of obs =      30
---------+------------------------------           F(  1,    28) =   21.33
   Model |  6394.02269     1  6394.02269           Prob > F      =  0.0001
Residual |  8393.44398    28  299.765856           R-squared     =  0.4324
---------+------------------------------           Adj R-squared =  0.4121
   Total |  14787.4667    29  509.912644           Root MSE      =  17.314

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9708704   .2102157      4.618   0.000       .5402629    1.401478
   _cons |   98.71472   10.00047      9.871   0.000       78.22969    119.1997
------------------------------------------------------------------------------
```

Now, one of the assumptions for regression analysis is that of homoscedasticity

(i.e., the variance of *y* is the same for any *x*)



Here, $\sigma^2_{y|x_1} = \sigma^2_{y|x_2} = \sigma^2_{y|x_3} = \sigma^2_{y|x_4}$

i.e., $\sigma^2_{y|x_i}$ is the same for all i

We will denote this common value σ^2

i.e., $\sigma^2_{y|x} \equiv \sigma^2$ for all x.

An estimate of σ^2 is given by the formula

$$s^2_{y|x} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{n-2}\left(SSE\right)$$

$$= \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta}_1^2 s_x^2\right)$$

lose 2 d.f.

one for β_0

one for β_1

sample variance of y

sample variance of x

where

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

lose 1 d.f. for estimating μ

and

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n-1}$$

in our example

$$s_y^2 = 509.91$$

$$s_x^2 = 233.91$$

$$\hat{\beta}_1 = 0.97$$

$$s_{y|x}^2 = \frac{n-1}{n-2}\left(s_y^2 - \hat{\beta}_1^2 s_x^2\right) = \frac{29}{28}\left(509.91 - .097^2\left(233.91\right)\right)$$

$$\boxed{s_{y|x}^2 = 299.77}$$

$\sqrt{s_{y|x}^2} = s_{y|x}$ is called the "standard error of estimate"

here

$$s_{y|x} = \sqrt{s_{y|x}^2} = \sqrt{299.77} = \boxed{17.31}$$

Now, if we assume that for any fixed value of x, y has a normal distribution, we can test hypotheses and construct confidence intervals for β_0 or β_1.

Under this assumption, it can be shown that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\right)$$

and

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_x^2(n-1)}\right)$$

Since we don't know σ^2 we estimate it with $s_{y|x}^2$ and use the t-distribution with $n-2$ degrees of freedom.

First consider β_1

In order to test $H_0 : \beta_1 = \beta_1^0$, where β_1^0 is some hypothesized value for β_1, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\dfrac{s_{y|x}}{s_x\sqrt{n-1}}}$$

and

$$t \sim t(n-2)$$

or, setting up confidence intervals for β_1

$$\hat{\beta}_1 - t_{1-\alpha/2}(n-2)\left[\frac{s_{y|x}}{s_x\sqrt{n-1}}\right] \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2}(n-2)\left[\frac{s_{y|x}}{s_x\sqrt{n-1}}\right]$$

e.g.

In the current example, suppose we wish to test

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_a : \beta_1 \neq 0$$

then,

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\dfrac{s_{y|x}}{s_x \sqrt{n-1}}} = \frac{0.97 - 0}{\dfrac{17.31}{(15.29)\sqrt{29}}} = 4.62$$

and we reject H_0 if $t > t_{.975}(28) = 2.0484$

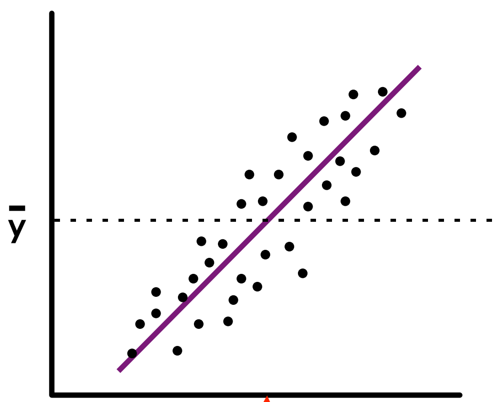
or if $t < t_{.025}(28) = -2.0484$

$\therefore$ reject H_0 at $\alpha = .05$

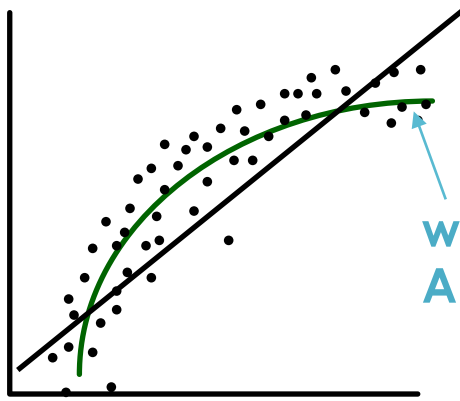
(In fact, $p < .001$)

This means that x provides significant information for the prediction of y. That is, $\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x})$ is far better than the naive model for predicting y.

A better model might exist (e.g, one with a curvilinear term), but there is a definite linear component.

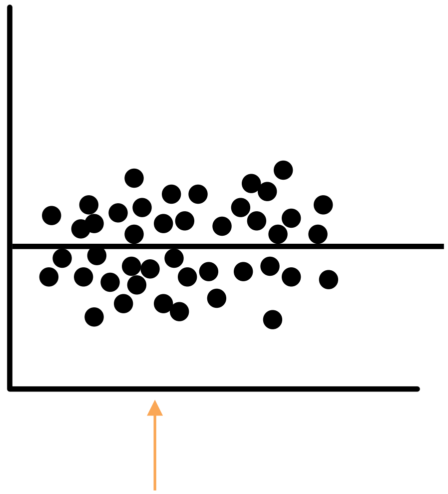


linear model certainly fits better than $\hat{y} = \bar{y}$



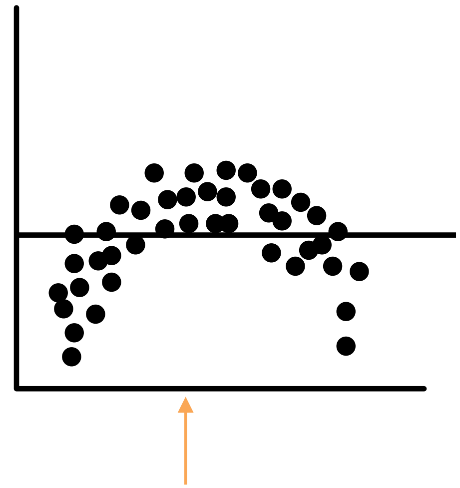
would provide A better fit

The straight line model may very well represent only a linear approximation to a truly nonlinear relationship note: if $H_0 : \beta_1 = 0$ is not rejected it means either



x provides little or no help in predicting *y*

or



The true relationship between *x* and *y* is not linear.

* Important point: whether or not $H_0 : \beta_1 = 0$ is rejected, the straight-line model may not be appropriate. Some other function may better describe the relationship between *x* and *y*.

Now Consider β_0

In order to test $H_0 : \beta_0 = \beta_0^{(0)}$, the test statistic used is

$$t = \frac{\hat{\beta}_0 - \beta_0^{(0)}}{s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}}$$

and

$$t \sim t(n-2)$$

and confidence intervals may be constructed as

$$\hat{\beta}_0 - t_{1-\alpha/2}(n-2)\left[s_{y|x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}}\right] \le \beta_0 \le \hat{\beta}_0 + t_{1-\alpha/2}(n-2)\left[s_{y|x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}}\right]$$

e.g., Continuing this example, to test

$$H_0: \beta_0 = 75$$

vs. $H_a: \beta_0 \ne 75$

$$t = \frac{\hat{\beta}_0 - \beta_0^{(0)}}{s_{y|x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}}} = \frac{98.71-75}{17.31\sqrt{\frac{1}{30}+\frac{(45.13)^2}{(29)(15.29)}}} = 2.37$$

and again reject H_0 at the α=.05 level

here $.02 < p < .05$

and

$$98.71 - 2.0484(17.31)\sqrt{\frac{1}{30} + \frac{(45.13)^2}{29(15.29)^2}} \le \beta_0 \le 98.71 + 2.0484(17.31)\sqrt{\frac{1}{30} + \frac{(45.13)^2}{29(15.29)^2}}$$

$$\boxed{78.23 \le \beta_0 \le 119.20}$$

Now, if you give me a value of x, I'll give you a confidence interval for $\mu_{y|x}$.

It can be demonstated that

$$\sigma^2_{\hat{y}_{x_0}} = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)$$

and this is estimated by

$$s^2_{\hat{y}_{x_0}} = s^2_{y|x}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)$$

and a $100(1-\alpha)\%$ confidence interval estimate for $\mu_{y|x}$ is

$$\hat{y}_{x_0} - t_{1-\alpha/2}(n-2)s_{\hat{y}_{x_0}} \le \mu_{y|x} \le \hat{y}_{x_0} + t_{1-\alpha/2}(n-2)s_{\hat{y}_{x_0}}$$

Example:

Suppose we want a 90% confidence interval for the mean SBP of 65 year old individuals

$$\hat{y}_{x_0} = \hat{y}_{65} = 142.53 + (0.97)(65-45.13)$$

$$= 161.80$$

$$s_{\hat{y}_{65}} = 17.31\left(\frac{1}{30} + \frac{(65-45.13)^2}{(29)(15.29)^2}\right)^{1/2} = 5.24$$

$$161.80 - 1.7011(5.24) \le \mu_{y|65} \le 161.80 + 1.7011(5.24)$$

$$152.89 \le \mu_{y|65} \le 170.71$$

Suppose we now wish to estimate the response *y* of a single individual based on the fitted regression function.

It can be demonstrated that the "prediction interval" (PI) is given by

$$\hat{Y}_{x_0} - t_{1-\alpha/2}(n-2)\, s_{y|x}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}} \le Y_{x_0} \le \hat{Y}_{x_0} + t_{1-\alpha/2}(n-2)\, s_{y|x}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

This is not a parameter.
Hence, we use the expression "PI" rather than "CI".

Note that this is the only difference from the previous expression.

Example

Suppose we want a 90% prediction interval for SBP for an individual whose age is 65

$$\text{again, } \hat{Y}_{65} = 161.80$$

$$S_{y|x}\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\left(n-1\right)s_x^2}} = 17.31\sqrt{1+\frac{1}{30}+\frac{\left(65-45.13\right)^2}{\left(29\right)\left(15.29\right)^2}} = 18.09$$

note how much larger this is than before (5.24)

$$161.80-\left(1.7011\right)\left(18.09\right) \le Y_{65} \le 161.80+\left(1.7011\right)\left(18.09\right)$$

$$\boxed{131.03 \le Y_{65} \le 192.57}$$

prediction interval is much wider than confidence interval was

Note that whether we are constructing confidence intervals or prediction intervals, the expressions contain the term

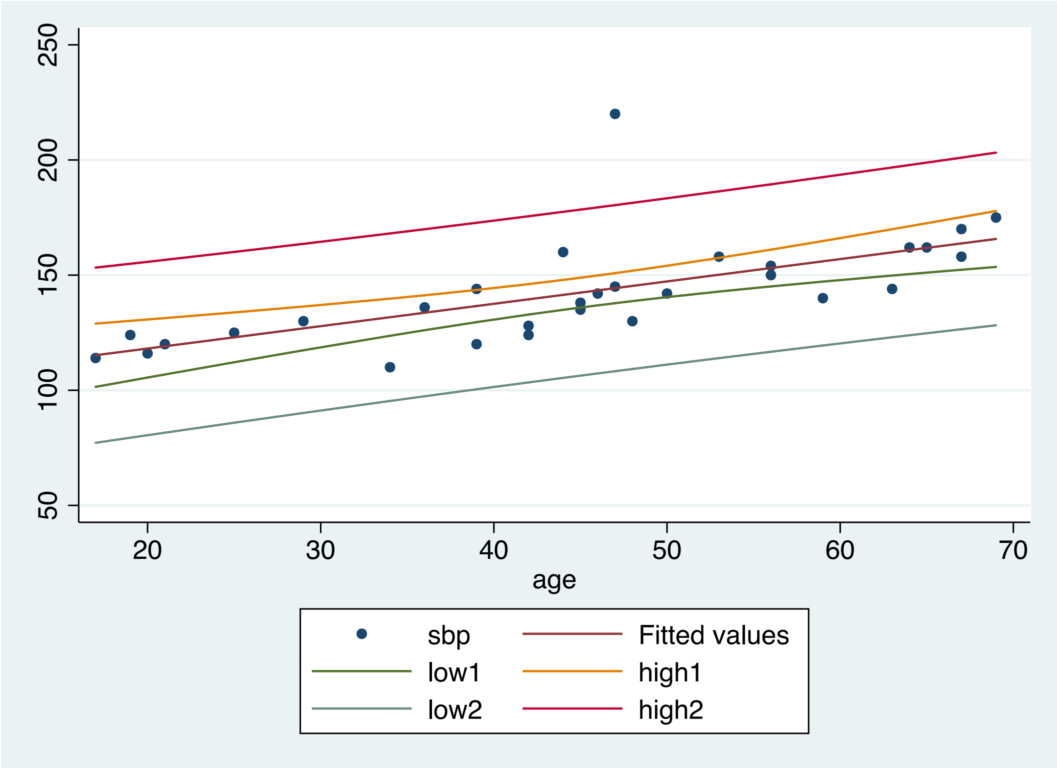
$$\frac{\left(x_0 - \bar{x}\right)^2}{\left(n-1\right)s_x^2}$$

This means that the farther x_0 is from $\bar{x}$, the larger will be the variance and the wider will be the interval.

Diagramatically:

```
. predict seyhat, stdp
. display invttail(28,0.025)
2.0484071
. generate low1= yhat-2.0484071* seyhat
. generate high1= yhat+2.0484071* seyhat
```

```
. predict sepred, stdf
. generate low2= yhat-invttail(28,0.025)* sepred
. generate high2= yhat+invttail(28,0.025)* sepred
. scatter sbp yhat low1 high1 low2 high2 age,sort connect(. l l l l l)
                                                           symbol(o i i i i i)
```



Hence, we can make more precise estimates for $\mu_{y|x}$ or Y_{x_0} when we are close to $\bar{x}$. As we move away from $\bar{x}$ our confidence intervals and prediction intervals increase in width.