Simple linear regression

Background.

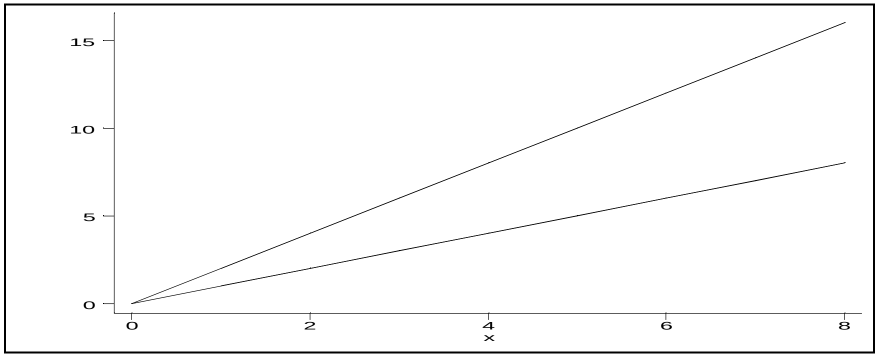
We all know what a straight line is. Along with the simple way of drawing a line (e.g., by using a ruler), there is a mathematical way to draw a line. This involves specifying the relationship between two **coordinates** x (measured on the horizontal or x axis) and y (measured on the vertical or y axis). By doing so, each point on the line is "drawn" by specification of the point's coordinates (x_i,y_i).

The equation relating the x_i to the y_i is as follows:

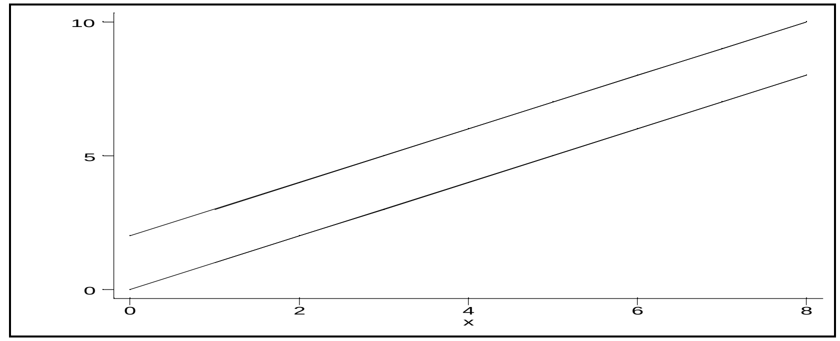
$$\boxed{y=\beta_o+\beta_1 x}$$

β_o is called the **intercept** of the line (because if x_i=0 the line "intercepts" the y axis at β_o), and β_1 is called the **slope** of the line.

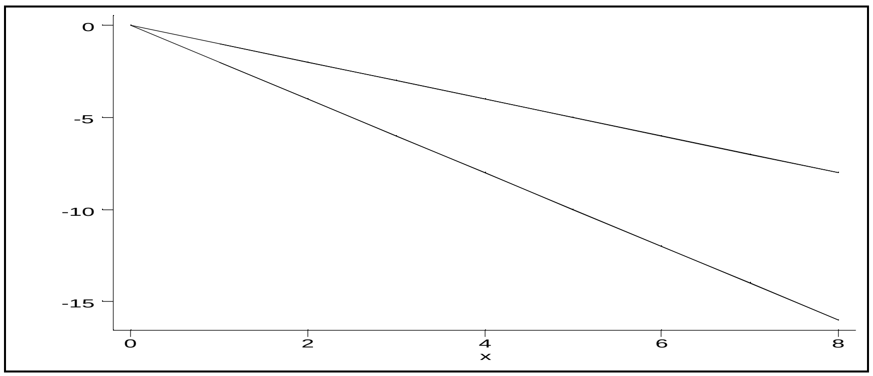
I



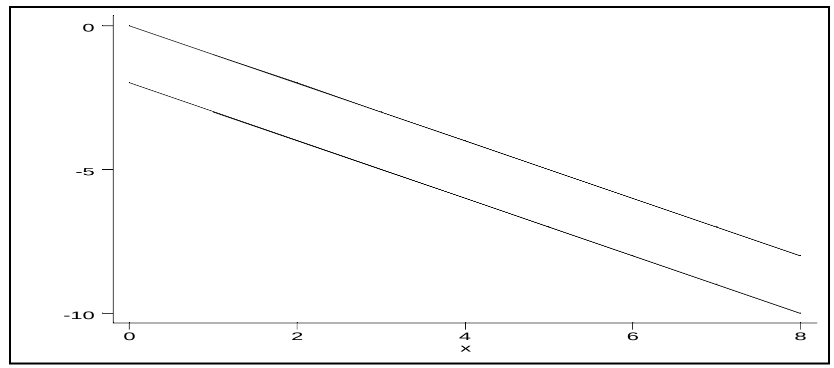
II



III

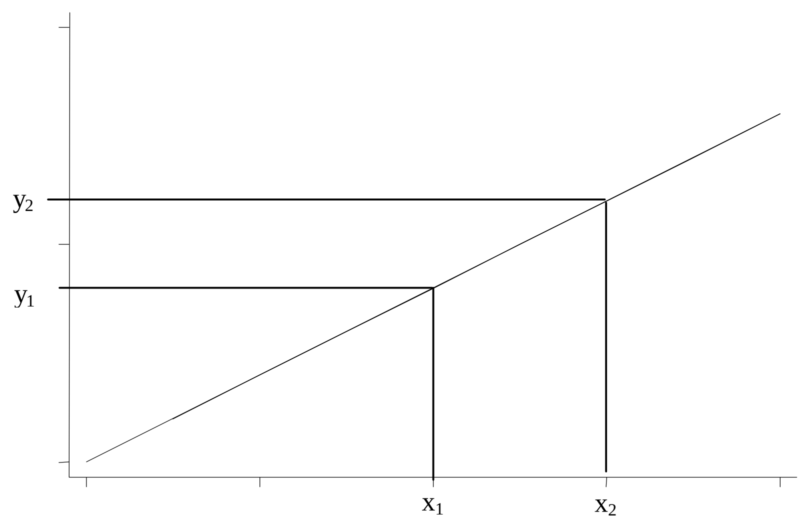


IV

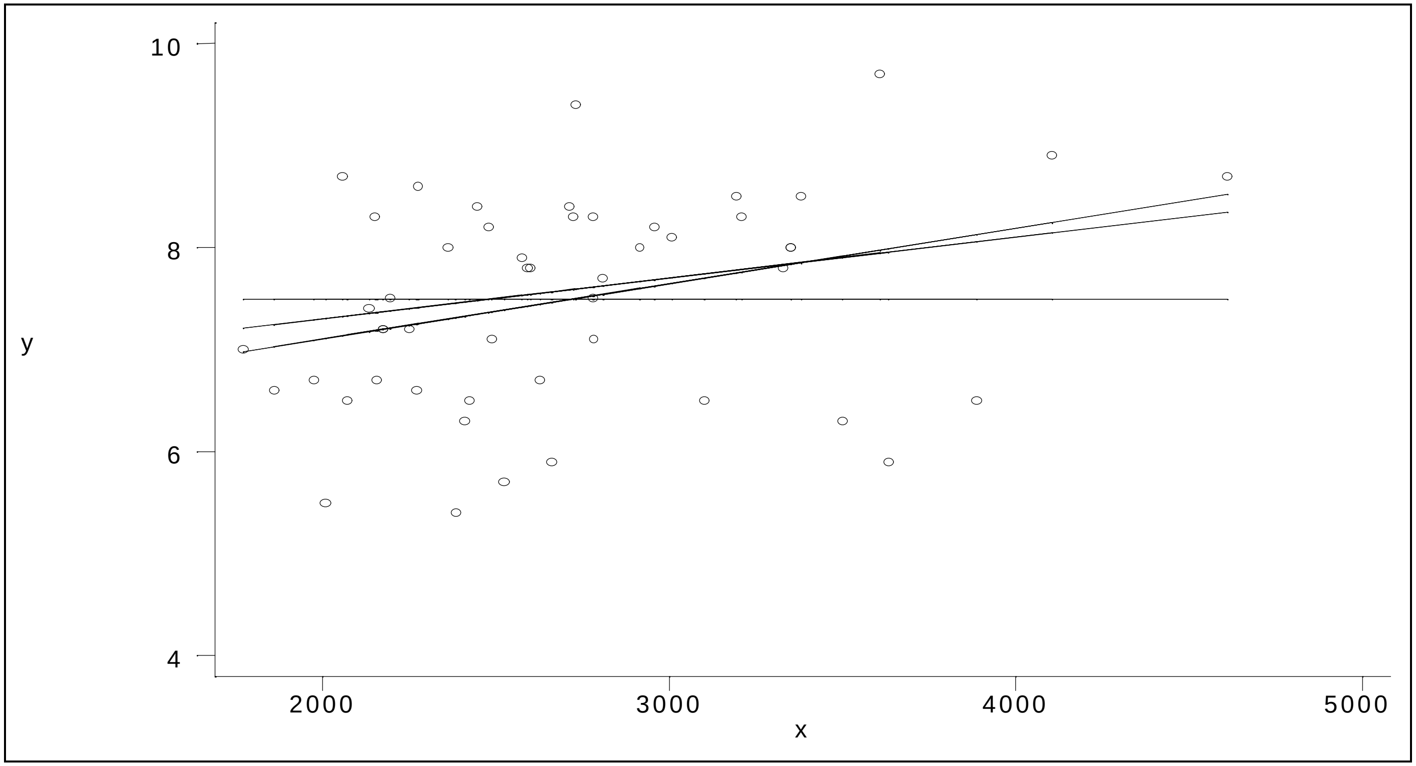


I. Both lines have the same intercept.
II. Both lines have the same slope (they are **parallel**) but different intercept.
III. Both lines have the same intercept but different **negative** slopes
IV. Both lines have the same (negative) slope but different intercepts.

The appeal of a linear relationship is the *constant slope*. This means that for a fixed increase Δx in x, there will be a fixed change Δy ($=\beta_o \Delta x$). This is going to be a fixed *increase* if the slope is positive, or a fixed *decrease* if the slope is negative, regardless of the value of x. This is in contrast to a *non-linear* relationship, such a *quadratic* or *polynomial*, where for some values of x, y will be increasing, and for some other values y will be decreasing (or vice versa).



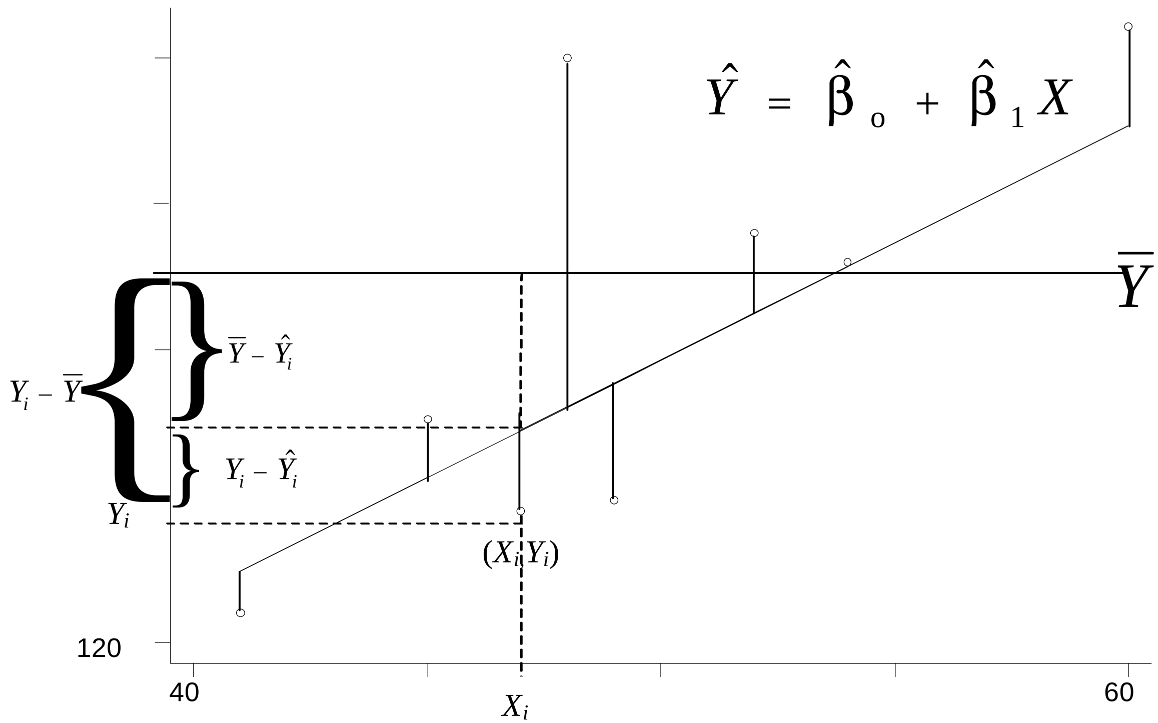
Consider the following scatter plot



Even though it seems upon inspection that *y* may be increasing for increasing *x*, the relationship is not a perfect line. If we want to draw a line through the plotted observations that we think best describes the trends in our data we may be confronted with many candidate lines.

Determining the Best-fitting straight line: The least squares method

Consider the following figure (taken from fitting a regression line to the systolic blood pressure – SBP- data of Table 5-1 in the text):



The least-squares method

The regression line (whatever it is) will not pass through all data points Y_i. Thus, in most cases, for each point X_i the line will produce an *estimated* point $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$ and most probably, $\hat{Y}_i \neq Y_i$. In fact, as we see in the previous figure, $Y_i = \hat{Y}_i + e_i$. For each choice of $\hat{\beta}_o$ and $\hat{\beta}_1$ (note that each pair $\hat{\beta}_o$ and $\hat{\beta}_1$ completely defines the line) we get a new line, and a whole new set of deviation terms e_i.

The "best-fitting line" according to the least-squares method is the one that *minimizes* the sum of square deviations $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_o + \hat{\beta}_1 X_i\right)\right)^2$.

The least-squares method (continued):

The solution is derived by use of calculus. That is, we set the last part of the above equation to zero and take partial derivatives with respect to β_o and β_1.

The resulting *least-squares estimates* of β_o and β_1 are given by the following expressions:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X}}$$

Note that since $\hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X}$, then $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X})$. This implies that if $\hat{\beta}_1$ is close to zero, our best guess for Y is the mean $\overline{Y}$.

Explaining variability

Statistical modeling is an attempt to "explain" why not all data points are equal. In other words, we are trying to account for the *variability* in the data.

The total variability in the data is given by

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left[\left(Y_i-\hat{Y}_i\right)+\left(\hat{Y}_i-\overline{Y}\right)\right]^2$$

as we can see by inspection of the previous figure. It also turns out that

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$$

This is because, the cross-product term $2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) = 0$.

Proof (Draper and Smith, p.18):

Since,

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1\left(X_i - \bar{X}\right)$$

$$\Rightarrow$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_1\left(X_i - \bar{X}\right) \qquad \text{and} \qquad Y_i - \hat{Y}_i = Y_i - \bar{Y} - \hat{\beta}_1\left(X_i - \bar{X}\right)$$

we have,

$$\begin{aligned}
2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) &= 2\sum_{i=1}^{n}\hat{\beta}_1\left(X_i - \bar{X}\right)\left\{\left(Y_i - \bar{Y}\right) - \hat{\beta}_1\left(X_i - \bar{X}\right)\right\} \\
&= 2\sum_{i=1}^{n}\hat{\beta}_1\left\{\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - \hat{\beta}_1\left(X_i - \bar{X}\right)^2\right\} \\
&= 0
\end{aligned}$$

since $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$

Explaining variability (continued)

This means that there are two parts to the total variability $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ $SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ in the data: one part that is explained or accounted for *due to the regression* $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$, and another that is left unexplained. That is, the regression cannot explain why there are still distances $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ between the estimated points and the data (this is called *error sum of squares*).

Since our goal is to reduce the part of the total variability that is unexplained, the regression line will be more useful as the variability due to regression is increasing compared to the unexplained variability. That is, the ratio $R^2 = \frac{SSR}{SSY} = \frac{SSY - SSE}{SSY}$ is as large as possible.

Degrees of freedom

We define the following quantities

$$1.\ S_Y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{(n-1)}SST \quad 2.\ S_{Y|X}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{(n-2)}SSE$$

Note that each of the sums of squares that we consider is comprised by a number of terms. For example, the total sum of squares, *SSY* is made up of *n* terms of the form $(Y_i - \overline{Y})^2$. Notice however, that once the mean $\overline{Y}$ has been estimated, only *n*-1 terms are needed to compute *SSY*. The n^{th} term is known since $SSY = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right) = 0$ for all Y_i, i=1,…,n-1. The *degrees of freedom* of *SSY* are then *n*-1. On the other hand, the sum of squares due to regression, *SSR*, is computed from a single function involving the Y_i (the estimated slope $\hat{\beta}_1$), that is, $SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \hat{\beta}_1^2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and has thus only one degree of freedom associated with it. Finally, *SSE* has *n*-2 degrees of freedom.

The Analysis of Variance Table

Source of variability	Sums of squares (SS)	Df	Mean squares (MS)	F	Prob > F
Model	$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$	1	$MSR = SSR$	$F = \dfrac{MSR}{MSE}$	$P = \mathrm{P}(F > F_{1,\, n\text{-}2;\alpha})$
Residual (error)	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	n-2	$MSE = \dfrac{SSE}{(n-2)} = S_{Y\vert X}^2$		
Corrected Total	$SSY = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$	n-1			

Assumptions of the linear regression model

1. The deviations $\varepsilon_i = Y_i - \hat{Y}$ have zero mean and variance s^2 which is unknown

2. The ε_is are *uncorrelated*, that is, for any i and j with $i \neq j$, cov(e_i,e_j)=0

Two immediate implications of these assumptions are that the mean of each data observation is $E(Y_i) = \mu_{Y|X} = \beta_o + \beta_1 X_i$, with *common* variance s^2, and that Y_i and Y_j are uncorrelated for $i \neq j$.

A final assumption that allows us to perform statistical tests is as follows:

3. The deviations ε_i are distributed according to the normal distribution, with mean 0 and variance s^2 that is, $\varepsilon_i \sim N(0,\ s^2)$.

This final assumption implies that the ε_i are not only uncorrelated but also *independent*.

Inference in simple linear regression

Tests involving the slope of the regression line

Recall that

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \Big/ \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$= \sum_{i=1}^{n} \left(X_i - \overline{X}\right) Y_i \Big/ \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} Y_1 + \cdots + \frac{\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} Y_n$$

Thus, the variance of $\hat{\beta}_1$ is $V\left(\hat{\beta}_1\right) = \dfrac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$ and the standard deviation is $S_{\hat{\beta}_1} = \dfrac{\sigma}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$.

The solution is left as an exercise. Since σ^2 is unknown $\text{s.e.}\left(\hat{\beta}_1\right) = \dfrac{s}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$ is used in the tests and confidence intervals.

Test of hypothesis for zero slope

In these models, y is our target (or *dependent* variable, the outcome of interest, or a factor that we cannot control but want to explain) and x is the explanatory (or *independent* variable).

Within each regression the primary interest is the assessment of the existence of the linear relationship between x and y. If such an association exists, then x provides information about y.

Inference on the existence of the linear association is accomplished via tests of hypotheses, and confidence intervals. Both of these center around the estimate of the slope β, since it is clear, that if the slope is zero, then changing x will have no impact on y, thus there is no association between x and y.

Hypothesis testing for zero slope (continued)

The test of hypothesis of no linear association is defined as follows:

1. H_o: No linear association between x and y: $\beta_1=0$..

2. H_a: A linear association exists between x and y:

a. $\beta_1 \neq 0$(two-sided test)

b. $\beta_1>0$

c. $\beta_1<0$ } (one-sided tests)

3. Tests are carried out at the (1-α)% level of significance

4. The test statistic is $T=\dfrac{\hat{\beta}_1}{\text{s.e.}\left(\hat{\beta}_1\right)}$ distributed as a t distribution with n-2 degrees of freedom

5. Rejection rule: Reject H_o, in favor of the three alternatives respectively, if

a. $t<t_{n-2;\alpha/2}$, or $t>t_{n-2;(1-\alpha/2)}$

b. $t>t_{n-2;(1-\alpha)}$

c. $t<t_{n-2;\alpha}$

Confidence intervals

Confidence intervals of b_1 are constructed as usual, and are based on the standard error of $\hat{\beta}_1$, the estimator, and the t statistic discussed above.

A (1-α)% confidence interval is as follows:

$$\boxed{\left(\hat{\beta}_1 - t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{\beta}_1\right), \hat{\beta}_1 + t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{\beta}_1\right)\right)}$$

<u>Inference involving the intercept</u>

In some rare occasions, tests involving the intercept are carried out. Both hypothesis tests and confidence intervals are based on the variance $S^2_{\hat{\beta}_o} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X} \right)^2} \right)$. The derivation is again left as an exercise. (*hint*. Consider the fact that $\hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X}$ and that $\overline{Y}$ and $\hat{\beta}_1$ have zero covariance as it is proven below). The statistic, $T = \frac{\hat{\beta}_o}{\text{s.e.}\left(\hat{\beta}_o\right)} \sim t_{n-2}$, where $\text{s.e.}\left(\hat{\beta}_o\right) = s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X} \right)^2}}$.

A (1-α)% confidence interval is as follows:

$$\boxed{\left(\hat{\beta}_o - t_{n-2;(1-\alpha/2)} \text{s.e.}\left(\hat{\beta}_o\right), \hat{\beta}_o + t_{n-2;(1-\alpha/2)} \text{s.e.}\left(\hat{\beta}_o\right) \right)}$$

Inference about the regression line

Recall that $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X})$. Thus, the variability of a specific point $\hat{Y}_o$ at X_o is given by

$$\begin{aligned}
S^2_{\hat{Y}_o} &= V(\overline{Y}) + (X_o - \overline{X})^2 V(\hat{\beta}_1) + (X_o - \overline{X})\operatorname{cov}(\overline{Y}, \hat{\beta}_1) \\
&= V(\overline{Y}) + (X_o - \overline{X})^2 V(\hat{\beta}_1) \qquad \text{since } \operatorname{cov}(\overline{Y}, \hat{\beta}_1) = 0. \\
&= \frac{\sigma^2}{n} + \frac{(X_o - \overline{X})^2 \sigma^2}{\Sigma (X_i - \overline{X})^2} = \sigma^2 \left\{ \frac{1}{n} + \frac{(X_o - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right\}
\end{aligned}$$

The fact that $\operatorname{cov}(\overline{Y}, \hat{\beta}_1) = 0$ is seen from the fact that

$$\begin{aligned}
\operatorname{cov}(\overline{Y}, \hat{\beta}_1) &= \operatorname{cov}\left\{ \left(\frac{Y_1}{n} + \cdots + \frac{Y_n}{n} \right), \left(\frac{(X_1 - \overline{X})Y_1}{\Sigma (X_i - \overline{X})^2} + \cdots + \frac{(X_n - \overline{X})Y_n}{\Sigma (X_i - \overline{X})^2} \right) \right\} \\
&= \left(\frac{(X_1 - \overline{X})}{n\sum (X_i - \overline{X})^2} + \cdots + \frac{(X_n - \overline{X})}{n\sum (X_i - \overline{X})^2} \right) \sigma^2 = \frac{\sum (X_i - \overline{X})}{n\sum (X_i - \overline{X})^2} \sigma^2 = 0
\end{aligned}$$

Inference about a particular value $\hat{Y}_o$ is based on the statistic

$$T = \frac{\hat{Y}_o}{\text{s.e.}\left(\hat{Y}_o\right)} \sim t_{n-2}$$

where $\text{s.e.}\left(\hat{Y}_o\right) = s\sqrt{\frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$. Notice that the estimated variability (of the regression line) *increases* as values of X are considered that are farther from the mean.

Inference about the regression line mainly involves construction of confidence intervals.

A (1-α)% confidence interval is as follows:

$$\boxed{\left(\hat{Y}_o - t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{Y}_o\right), \hat{Y}_o + t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{Y}_o\right)\right)}$$

This interval is wider away from the mean of the X's, and narrower closer to that mean.

An *F* test of overall linear association

Recall that $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$. That is, the total variation in the data is broken up into two parts. One *due to regression* and one left unexplained (*error*) respectively.

It can be shown that SSE/σ^2 or equivalently $\frac{(n-2)S_{Y|X}^2}{\sigma^2}$ follows a chi-square distribution with n-2 degrees of freedom. On the other hand, ***if*** $\boldsymbol{\beta_1=0}$, SSR/σ^2 follows a chi-square distribution with 1 degree of freedom, and is *independent* of *SSE*.

Their ratio $F=\frac{MSR}{MSE}\sim F_{1,n-2}$ is distributed according to an *F* distribution with 1 and n-2 degrees of freedom.

An *F* test of overall linear association (continued)

The *F* test of linear association, that is, the test of whether a line (other than the horizontal one going through the sample mean of the *Y*'s) is useful in explaining some of the variability of the data is based on the observation that the expected value $E(MSR) = \sigma^2 + \beta_1^2 \sum (X_i - \overline{X})^2$ while $E(SSE) = E(S^2_{Y|X}) = \sigma^2$ when the regression model is correctly specified (we will see what happens when this is not the case). If the population regression slope $\beta_1 \approx 0$, that is, if the regression does not add anything new to our understanding of the data (does not explain a substantial part of the variability), the two mean square errors *MSR* and *MSE* are estimating a common quantity (the population variance σ^2**).**

Thus the ratio should be close to 1 if the hypothesis of no linear association between *X* and *Y* is present. On the other hand, if a linear relationship exists, (β_1 is far from zero) then *SSR*>*SSE* and the ratio will deviate significantly from 1.

The F test of linear association

The F test of hypothesis of no linear association is defined as follows:

1. H_o: No linear association exists between X and Y
2. H_a: A linear association exists between X and Y
3. Tests are carried out at the (1-α)% level of significance
4. The test statistic is $F = \dfrac{MSR}{MSE}$.
5. **Rejection rule:** Reject H_o, if $F > F_{1,\, n-2;\alpha}$. This will happen if F is far from 1.0.

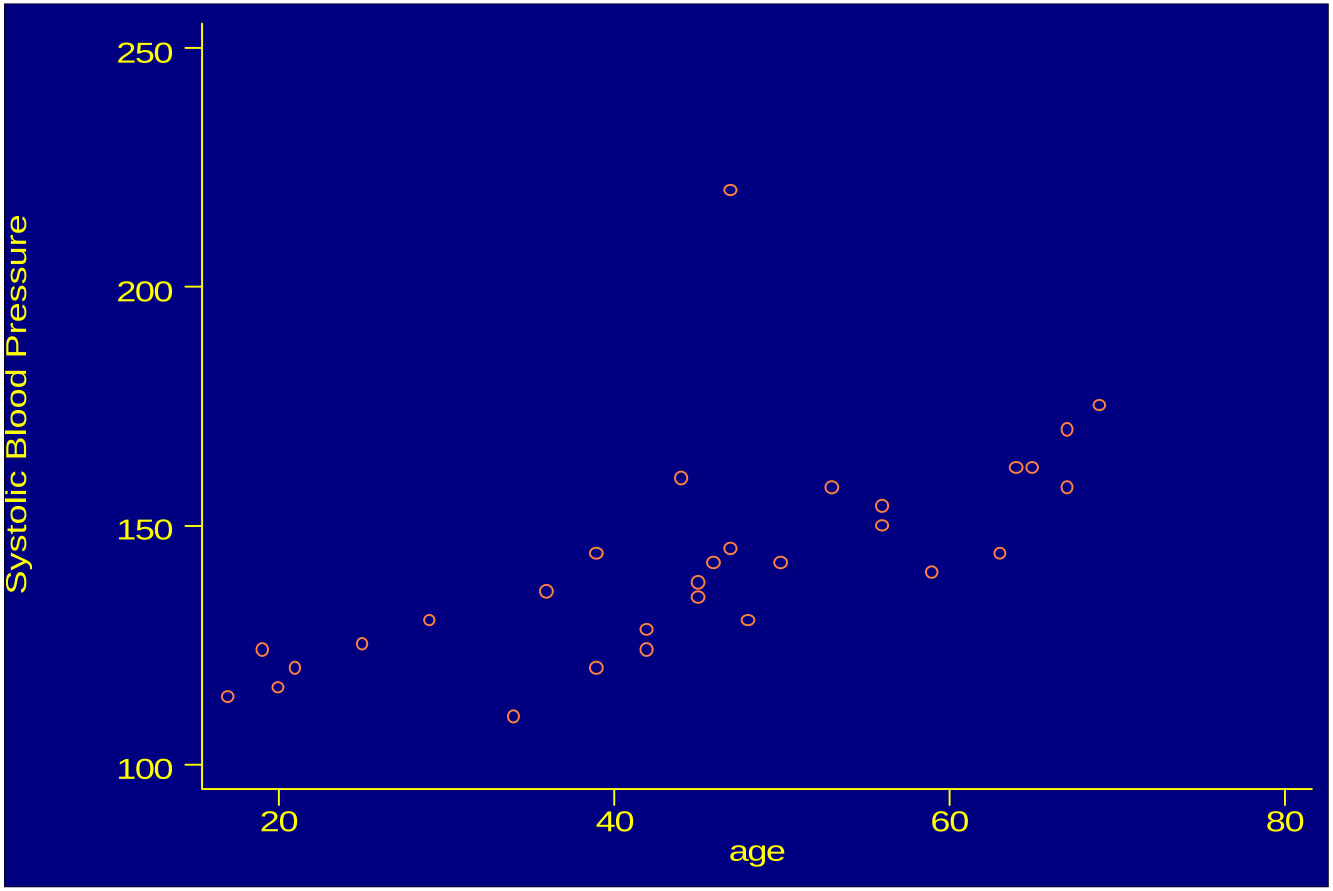
In simple linear regression, the F test is equivalent to the t test for zero slope described earlier. In fact, $T^2 = F$ where T^2 is distributed according to a t_{n-2} and F according to an $F_{1,n-2}$.

Analysis of the systolic blood pressure example

In this example, the relationship between systolic blood pressure (SBP) and age is explored. The data are listed below.

```
. list
          sbp        age              sbp        age
  1.      144         39     16.      130          4
  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
```

```
. label var sbp "Systolic Blood Pressure"
. graph sbp age, xlab ylab
```



```
. reg 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
------------------------------------------------------------------------------
```

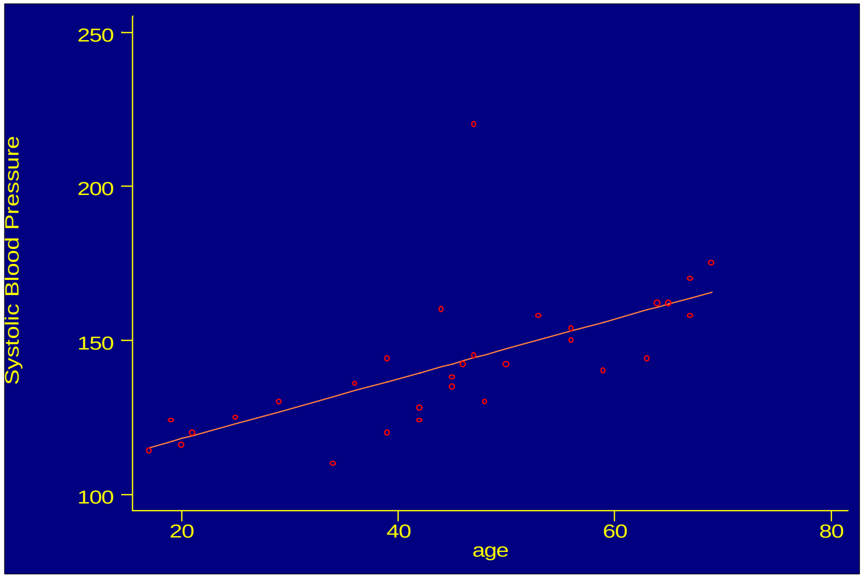
Conclusions

1. MSR=6394.02269
2. MSE=299.765856, which is the best estimate of σ^2 if the model is correct.
3. F=21.33 which is much larger than the tail of $F_{1,28;0.95}$. We thus reject the null hypothesis of no linear association between blood pressure and age.
4. The t statistic of zero slope is T=4.618. This is much larger than a $t_{28;0.975}$. Alternatively, the p value of the test is $0.000<0.05=\alpha$. Thus, we again *reject* the hypothesis of no linear association between systolic blood pressure and age. In fact, the positive estimate of the regression slope $\hat{\beta}_1 = 0.9709$ means that blood pressure *increases* with age (about one unit for every year of life).
5. The R^2 =0.4324. This means that approximately 43% of the variability (in the subjects' blood pressure) was explained by the regression model (i.e., age). Note the entry for *adjusted* R^2. This is a quantity such that adj. $R^2 = \frac{SSE/(n-2)}{SSY/(n-1)} = 1 - \left[(1-R^2) \left(\frac{(n-1)}{(n-2)} \right) \right] = 0.4121$. The adjusted R^2 is supposed to be used to compare between several models of varying complexity. It is not used often.

Predicted values

To produce the fitted values $\hat{Y}_i \quad i = 1, \cdots, 30$ we use the `predict` command as follows:

```
. predict sbphat
. graph sbphat sbp age, c(l.) s(io) xlab ylab
```



The option `c(l.)` means that the `sbphat` points should be connected by a line, while the `sbp` points should be left unconnected (a scatter plot) respectively while the option `s(io)` means that no symbol should be used for `sbphat` points while a small circle should be used for `sbp` points.

Confidence intervals about the regression line

Since STATA does not automatically produce 95% confidence intervals about the regression line these must be generated manually. To do this, we must first calculate the standard error of each estimated value, s.e.$\left[\hat{Y}_o\right] = s\sqrt{\frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$. This is computed by the option `stdp` after the `predict` command

```
. predict s, stdp
```

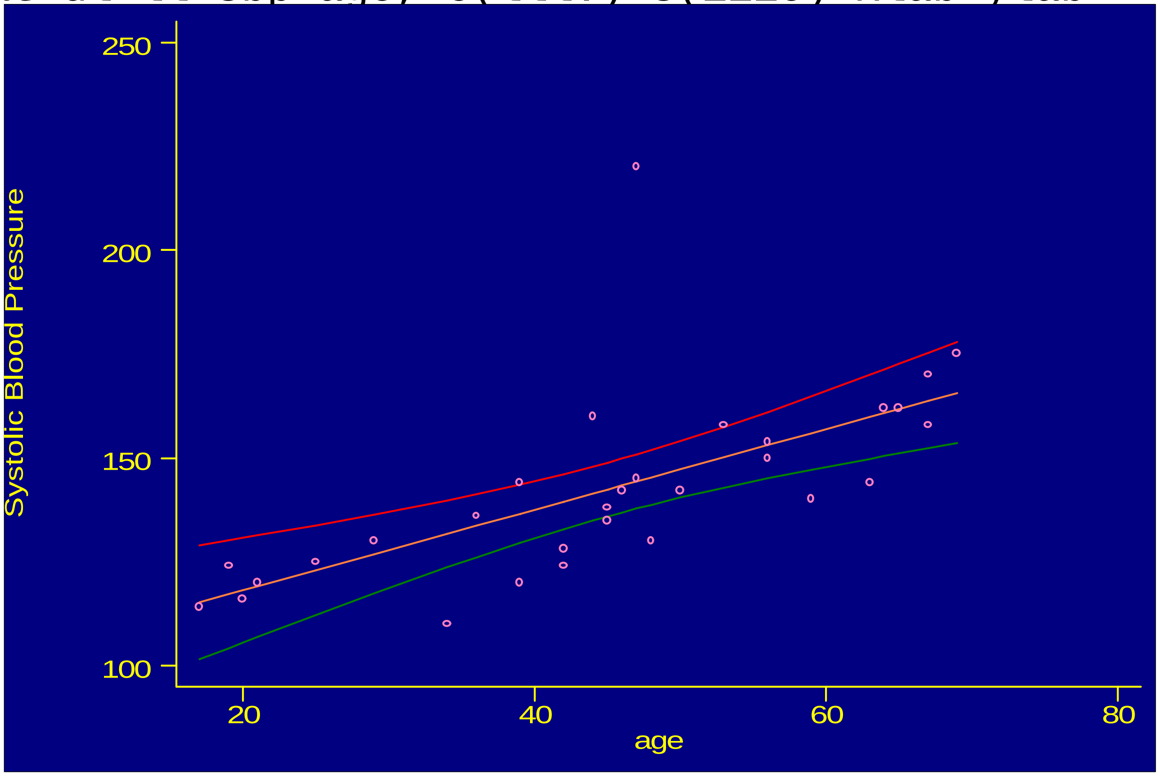
Then, the 95% upper and lower limits in each case are produced as follows:

```
. gen ul=sbphat+invt(28,0.95)*s
. gen ll=sbphat-invt(28,0.95)*s
```

In each case, `invt(28,0.95)`is the ***two-sided*** 95% tail of a t_{28} distribution (i.e., the inverse t).

Confidence intervals about the regression line (continued)

```
. sort ul
. graph sbphat ul ll sbp age, c(lll.) s(iiio) xlab ylab
```



Notice that the confidence "bands" open wider at the edges of the age interval.

Additional topic: Prediction

Statistical modeling does not only attempt to explain variability in the data, but *predict* a future observation $\hat{Y}_{X_o}$ at X_o. In doing so, it is critical to consider the sources of possible variability that enter into this prediction.

$$\underbrace{Y_{X_o} - Y}_{\substack{\text{Error in}\\ \text{predicting an}\\ \text{individual's}\\ Y \text{ at } X_o}} = \underbrace{Y_{X_o} - \mu_{Y|X_o}}_{\substack{\text{Deviation of}\\ \hat{Y}_{X_o} \text{ from true}\\ \text{mean at } X_o}} + \underbrace{\mu_{Y|X_o} - Y}_{\substack{\text{Deviation of}\\ \text{individual's}\\ Y \text{ from true}\\ \text{mean at } X_o}}$$

The variance of a new observation is $S_Y^2 + S_{\hat{Y}_{X_o}}^2 = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{\Sigma\left(X_o - \overline{X}\right)^2}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{\Sigma\left(X_o - \overline{X}\right)^2}\right)$.

Notice that the variability of the new observation is larger than that of existing observations.

Prediction intervals

When talking about future observations, we cannot construct "confidence intervals" in the strict sense (since the new observation is not a population parameter). The similar concept is called a "prediction interval". A (1-α)% such interval is based on the estimated standard deviation

$\text{s.e.}\left(\hat{Y}_{X_o}\right) = s\sqrt{1 + \frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{\Sigma\left(X_o - \overline{X}\right)^2}}$ and is constructed as follows:

$$\boxed{\left(\hat{Y}_{X_o} - t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{Y}_{X_o}\right), \hat{Y}_{X_o} + t_{n-2;(1-\alpha/2)}\text{s.e.}\left(\hat{Y}_{X_o}\right)\right)}$$

Prediction intervals of a new observations $\hat{Y}_{X_o}$

Since STATA does not automatically produce 95% prediction intervals about the regression line these must be generated manually. To do this, we must first calculate the standard error of each estimated value, $\text{s.e.}\left(\hat{Y}_{X_o}\right) = s\sqrt{1+\frac{1}{n}+\frac{\left(X_o-\overline{X}\right)^2}{\Sigma\left(X_o-\overline{X}\right)^2}}$. This is computed by the option `stdf` after the `predict` command

```
. predict sr, stdf
```

Then, the 95% upper and lower limits in each case are produced as follows:

```
. gen ulpred=sbphat+invt(28,0.95)*sr
. gen llpred=sbphat-invt(28,0.95)*sr
```

In each case, `invt(28,0.95)` is the ***two-sided*** 95% tail of a t_{28} distribution (i.e., the inverse *t*).

Prediction intervals of a new observation $\hat{Y}_{X_o}$

```
. graph sbphat ul ll ulpred llpred sbp age, c(lllll.) s(iiiiio) xlab ylab
```