Week ON2

Elmer V Villanueva

02 March 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON2")
```

Reading

Read and understand Vittinghoff et al., Chapter 3.3.

Simple linear regression

In the last lecture, we learned about covariance and correlation. We stated that the use of both covariance and correlation assumes no causal relationship between the two variables, even if, in fact, one exists. That is, it makes no sense to refer to one variable as the dependent and the other as the independent.

Let us change this assumption. Let us say that we are assuming that a particular causal relationship does exists between the two variables. Perhaps knwoledge of this causal relationship arises from previous studies, external data, or simple conjecture. In the simple case of two variables, usually labeled X and Y, we designate the former as the *independent (or predictor) variable* since it is controlled by the investigator. This means that the investigator has full control over X and chooses or preselects its values. From these values of X, the investigator obtains corresponding values of Y, which is called the *dependent (or outcome) variable.*

This process of analysis is called *regression* and provides us with more information than correlation. Regression analysis is helpful in assessing specific forms of the relationship between variables, and the ultimate objective when this method of analysis is employed usually is to *control*, *predict* or *estimate* the value of one variable corresponding to a given value of another variable. The ideas of regression were first elucidated by the English scientist Sir Francis Galton (1822-1911) in reports of his research on heredity-first in sweet peas and later in human stature. He described a tendency of adult offspring, having either short or tall parents, to revert back toward the average height of the general population. He first used the word *reversion*, and later *regression*, to refer to this phenomenon.

Formal statement of the model

We consider the simple case in which there is one dependent variable Y and one independent variable X. The model can be stated as follows:

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

where:

- y_i is the value of the outcome variable in the ith trial;
- β_0 and β_1 are parameters;
- x_i is a known constant, namely, the value of the predictor variable in the ith trial;
- ϵ_i is a random error term with mean or expected value $E(\epsilon_i) = 0$ and variance $\sigma^2(\epsilon_i) = \sigma^2$; and

- $i = 1, \ldots, n$

This regression model is said to be *simple*, *linear in the parameters*, and *linear in the predictor variable.* It is “simple” in that there is only one predictor variable, “linear in the parameters” because no parameter appears as an exponent or is multiplied or divided by another parameter, and “linear in the predictor variable” because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called a *first-order model.*

Important features of the model

The model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ has several important features.

First, the response y_i in the ith observation is the *sum* of two components: (1) the constant term $\beta_0 + \beta_1 x_i$ and (2) the random term ϵ_i. Hence, y_i is a random variable. Note from above that x_i is a constant.

Second, since $E(\epsilon_i) = 0$, it follows that the mean of y_i is

$E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i$

Thus, the response y_i, when the level of X in the ith trial is x_i, comes from a probability distribution whose mean is $E(y_i) = \beta_0 + \beta_1 y_i$. We therefore know that the regression function for our model is, more generally, $E(Y) = \beta_0 + \beta_1 X$.

Thirdly, the response y_i in the ith trial exceeds or falls short of the value of the regression function by the error term amount ϵ_i.

Fourthly, the error terms ϵ_i are assumed to have constant variance σ^2. It therefore follows thatthe responses y_i have the same constant variance $\sigma^2(Y) = \sigma^2$. Thus, the regression mondel assumes that the probability distributions of Y have the same variance σ^2, regardless of the level of the predictor variable X. The assumption of equal variance is called *homoskedasticity.*

Finally, error terms are assumed to be uncorrelated. That is, ϵ_i and ϵ_j has a covariance of zero (i.e., $Cov(\epsilon_i, \epsilon_j) = 0$ for all i, j; $i \neq j$). Since the error terms ϵ_i and ϵ_j are uncorrelated, so are the responses y_i and y_j.

In summary, the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ implies that the responses y_i come from probability distributions whose means are $E(y_i) = \beta_0 + \beta_1 x_i$ and whose variances are σ^2, the same for all levels of X. Further, any two responses y_i and y_j are uncorrelated.

Meaning of regression parameters

In the regression model $y_i = \beta_0 + \beta_1 y_i + \epsilon_i$, we stated above that β_0 and β_1 are parameters. The are also called regression coefficients. β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X. The parameter β_0 is the y-intercept of the regression line. When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = O$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a s~parate term in the regression model.

Example 1A

Digoxin is a drug often prescribed to treat heart ailments. The purpose of a study by Parker et al. [1] was to examine the interactions of digoxin with common grapefruit juice. In one experiment, subjects took digoxin with water for 2 weeks, followed by a 2-week period during which digoxin was withheld. During the next 2 weeks subjects took digoxin with grapefruit juice. For seven subjects, the average peak plasma digoxin concentration (Cmax) when taking water is given in the first column of Table 1. The second column contains the percent change in Cmax concentration when subjects were taking the digoxin with grapefruit juice [GFJ (%) change]. Use the Cmax level when taking digoxin with water to predict the percent change in Cmax concentration when taking digoxin with grapefruit juice.

Table 1. Average peak plasma digoxin concentration with water and percentage change in concentration following intake of grapefruit juice.

Cmax (ngl/ml)	GFJ % Change
2.34	29.5
2.46	40.7
1.87	5.3
3.09	23.3
5.59	-45.1
4.05	-35.3
6.21	-44.6

Let's enter the data and produce a scatterplot.

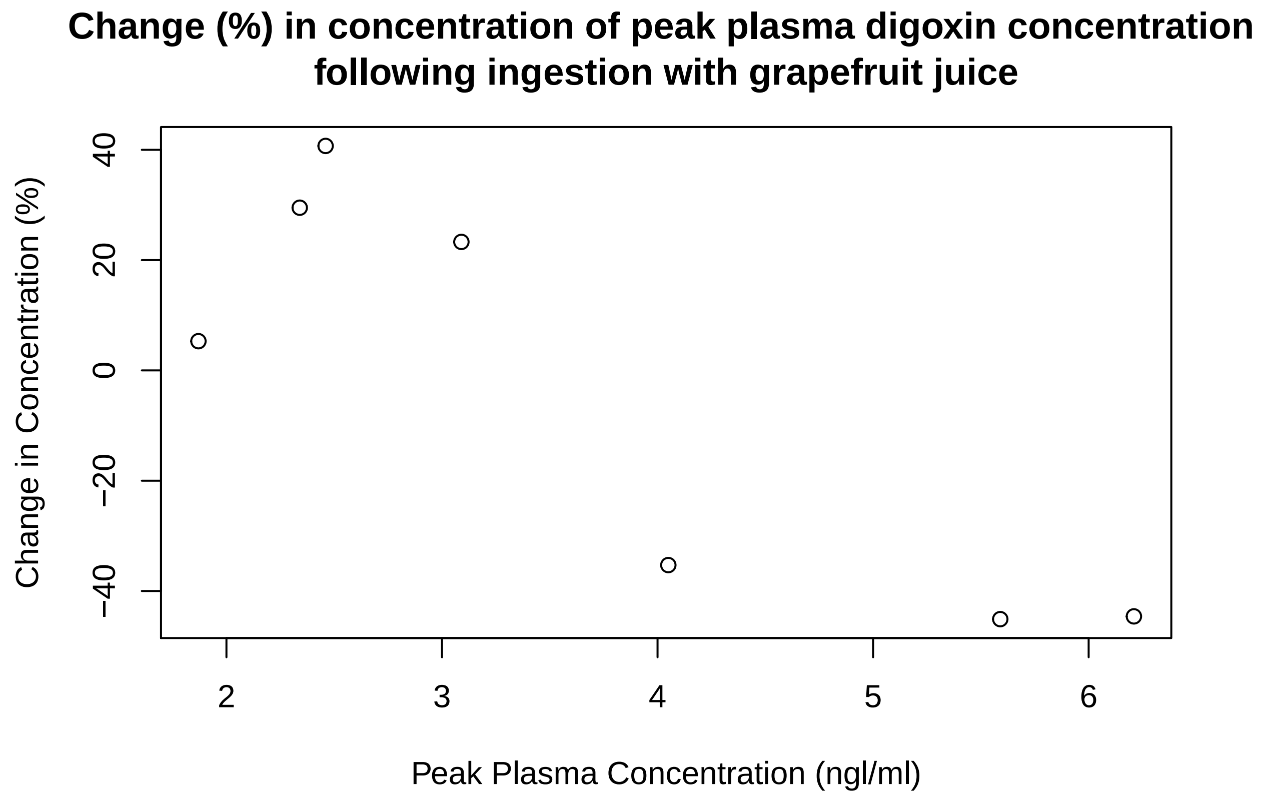
```
CMAX <- c(2.34, 2.46, 1.87, 3.09, 5.59, 4.05, 6.21)
GFJ <- c(29.5, 40.7, 5.3, 23.3, -45.1, -35.3, -44.6)
PARKER <- data.frame(CMAX, GFJ)
str(PARKER)
```

```
## 'data.frame':    7 obs. of  2 variables:
##  $ CMAX: num  2.34 2.46 1.87 3.09 5.59 4.05 6.21
##  $ GFJ : num  29.5 40.7 5.3 23.3 -45.1 -35.3 -44.6
```

```
head(PARKER)
```

```
##   CMAX   GFJ
## 1 2.34  29.5
## 2 2.46  40.7
## 3 1.87   5.3
## 4 3.09  23.3
## 5 5.59 -45.1
## 6 4.05 -35.3
```

```
plot(GFJ ~ CMAX, data = PARKER,
     main = "Change (%) in concentration of peak plasma digoxin concentration \nfollowing ingestion wit
     ylab = "Change in Concentration (%)",
     xlab = "Peak Plasma Concentration (ngl/ml)")
```



Note the way the `plot()` function is specified. This particular notation will be important later. The general form is `[DEPENDENT VARIABLE] ~ [INDEPENDENT VARIABLE(S)]`.

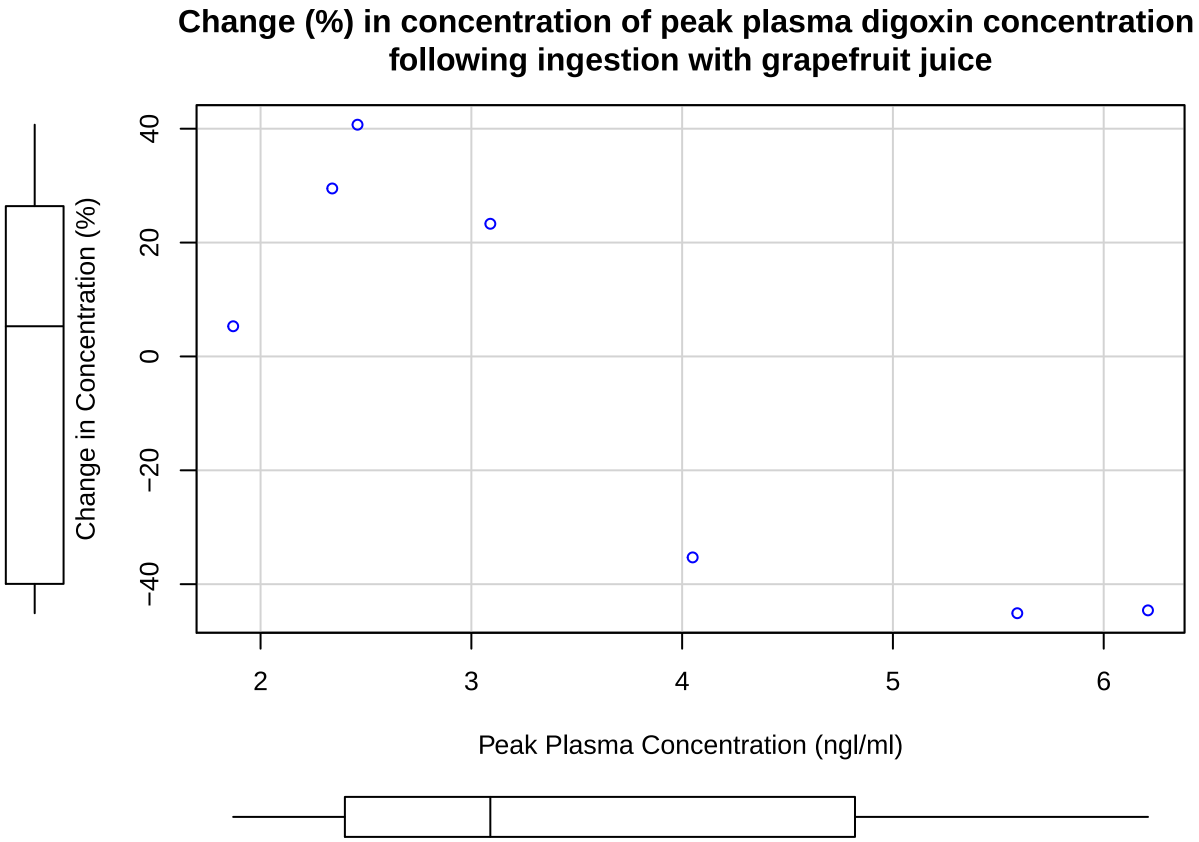
The function `scatterplot()` in the package `car` produces graph with extra features. For example, we can produce boxplots of the marginal distributions of the variables.

```
if(!require(car)){install.packages("car")}
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
library(car)
scatterplot(GFJ ~ CMAX, data = PARKER, smooth = FALSE, regLine = FALSE,
     main = "Change (%) in concentration of peak plasma digoxin concentration \nfollowing ingestion wit
     ylab = "Change in Concentration (%)",
     xlab = "Peak Plasma Concentration (ngl/ml)")
```



It looks as if it would be simple to draw, freehand, through the data points the line that describes the relationship between X and Y. It is highly unlikely, however, that the lines drawn by any two people would be exactly the same. In other words, for every person drawing such a line by eye, or freehand, we would expect a slightly different line. The question then arises as to which line best describes the relationship between the two variables. We cannot obtain an answer to this question by inspecting the lines. In fact, it is not likely that any freehand line drawn through the data will be the line that best describes the relationship between X and Y, since freehand lines will reflect any defects of vision or judgment of the person drawing the line. Similarly, when judging which of two lines best describes the relationship, subjective evaluation is liable to the same deficiencies.

What is needed for obtaining the desired line is some method that is not fraught with these difficulties.

The least-squares line

To find good estimators of the regression parameters β_0 and β_1, we employ the *method of least squares* and the resulting line is called the *least-squares line*. First we start by noting that the original regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ can be rearranged to

$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

Expressed in this form, ϵ_i can be considered the deviation of each observed value of the dependent variable y_i from its expected value $(\beta_0 + \beta_1 x_i)$.

The method of least squares requires that we consider the sum of the n squared deviations. This criterion is denoted by Q:

$Q = \sum^n \epsilon_i^2 = \sum^n (y_i - \beta_0 + \beta_1 x_i)^2$

According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, that minimise Q. We can find the estimators b_0 and b_1 in two basic ways: (1) numerical search procedures can be used that evaluate in a systematic fashion the least squares criterion Q for different estimates b_0 and b_1 until the ones that minimise Q are found; (2) analytical procedures can often be used to find the values of b_0 and b_1 that minimise Q.

Using the analytical approach, it can be shown that the values b_0 and b_1 that minimise Q for any particular set of sample data are given by the following simultaneous equations (called *normal equations*):

$\sum y_i = nb_0 + b_1 \sum x_i$ and $\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2$

When the normal equations are solved simultaneously for b_0 and b_1, we get:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \frac{1}{n}(\sum y_i - b_1 \sum x_i) = \bar{y} = b_1\bar{x}$$

An important theorem, called the *Gauss-Markov theorem*, states that under the conditions of the regression model, the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Since the necessary hand calculations are time consuming, tedious, and subject to error, the regression line equation is best obtained through the use of a computer software package.

Example 1B

Let's estimate b_0 and b_1 from the Parker data. The function we will use is `lm()`. This is a two-step process similar to that demonstrated for ANOVA in DPH101.

```
PARKER.LM <- lm(GFJ ~ CMAX, data = PARKER)
summary(PARKER.LM)
```

```
## 
## Call:
## lm(formula = GFJ ~ CMAX, data = PARKER)
## 
## Residuals:
##       1       2       3       4       5       6       7 
##   8.159  21.642 -24.982  16.227  -4.615 -24.111   7.680 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   65.855     19.646   3.352   0.0203 *
## CMAX         -19.023      4.938  -3.852   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.42 on 5 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.6976 
## F-statistic: 14.84 on 1 and 5 DF,  p-value: 0.01197
```

The first block of two lines shows the regression equation estimated by R.

The second block of three lines provides the distribution of the residuals, which is the difference between the observed value y_i and the fitted value $\hat{y}$. Residuals are highly useful and will be tackled in detail next week.

The third block of six lines contains the estimates of b_0 and b_1. In the present case, R estimated that $b_0 = 68.642$ and $b_1 = -19.529$. Thus, the linear equation for the least-squares line describing the relationship between `GFJ` and `CMAX` is

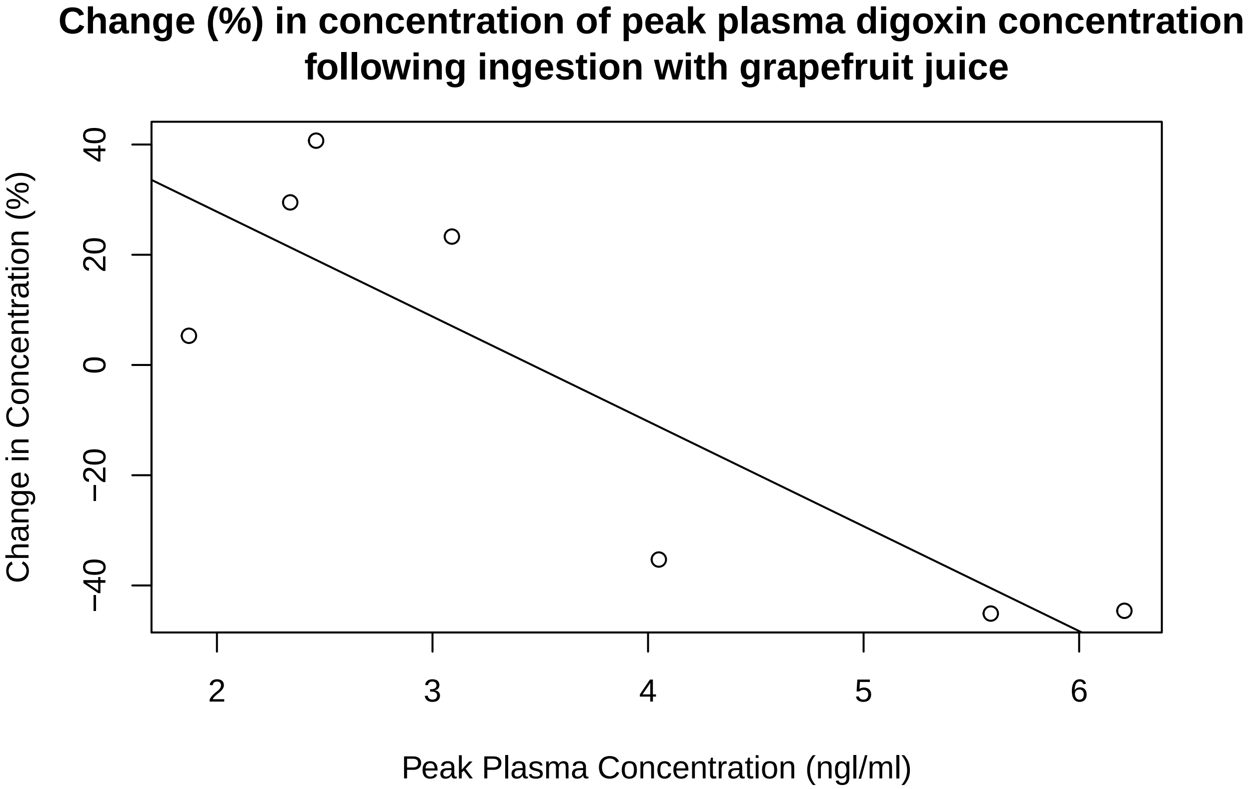
$\hat{y} = 68.642 - 19.529x$.

In this equation, we use the expression $\hat{y}$ (called "y-hat") to emphasise that these are predicted values of the response variable based on the least squares method.

Example 1C

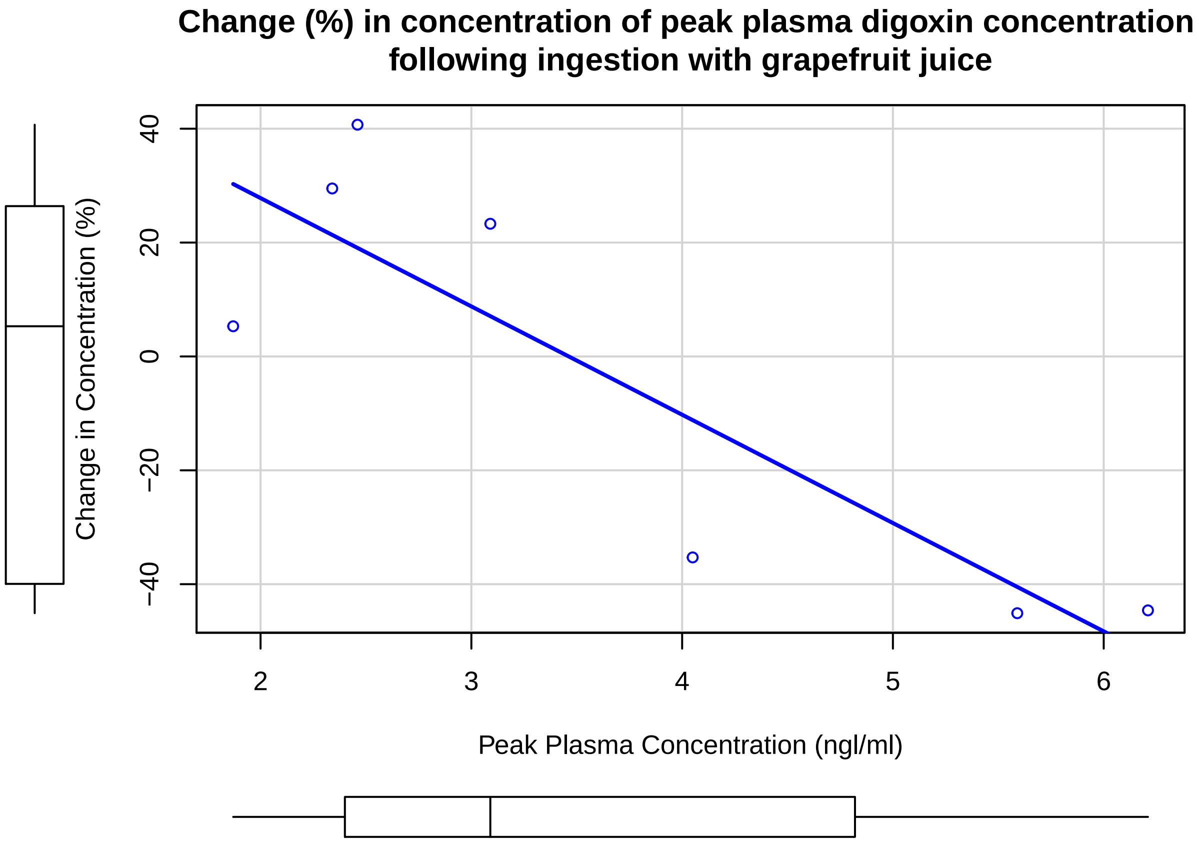
Draw the regression line onto the scatter plot.

```r
plot(GFJ ~ CMAX, data = PARKER,
     main = "Change (%) in concentration of peak plasma digoxin concentration \nfollowing ingestion wit
     ylab = "Change in Concentration (%)",
     xlab = "Peak Plasma Concentration (ngl/ml)")
abline(lm(GFJ ~ CMAX, data = PARKER))
```



Another way would be to use the `car` package.

```r
scatterplot(GFJ ~ CMAX, data = PARKER, smooth = FALSE,
     main = "Change (%) in concentration of peak plasma digoxin concentration \nfollowing ingestion wit
     ylab = "Change in Concentration (%)",
     xlab = "Peak Plasma Concentration (ngl/ml)")
```



Interpretation of least squares estimators

The least squares estimators b_0 and b_1 in physical terms. The intercept term b_0 is the percentage change in Cmax following ingestion of grapefruit juice when Cmax is zero. Since this is an extrapolation based on the data to hand, this has no particular meaning. The slope term b_1 is the estimate of the change in percentage of Cmax following ingestion of grapefruit juice per unit baseline Cmax. That is to say, for every 1 ngl/ml increase in Cmax, a person who ingests grapefruit juice is expected to experience a percentage change of -19.529%.

The relationship among regression, covariance and correlation

It can be shown that an alternative way to express the slope term b_1 is

$$b_1 = \frac{Cov(Y,X)}{\sigma_X^2} = Cor(Y,X)\frac{\sigma_Y}{\sigma_X}$$

Example 2A

The head of a lab wanted to know how long it took for 25 lab technicians to process blood samples for the presence of a particular genetic marker. Her data appear below.

Table 2. Lot size and hours to completion of DNA analysis

Lot Size	Hours
80	399
30	121
50	221

Lot Size	Hours
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342
70	323

Using the data given above, regress hours to completion on lot size. Interpret the corresponding regression coefficients.

Let's enter the data and produce a scatterplot.

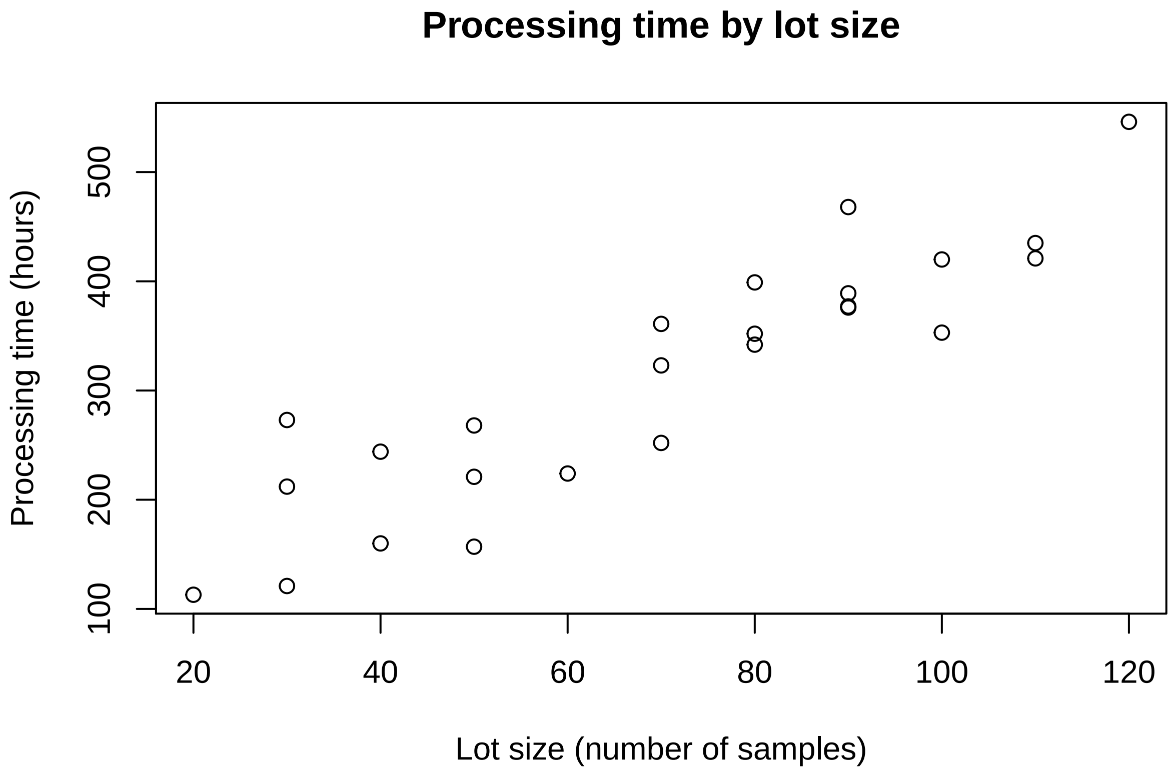
```
LOT <- c(80, 30, 50, 90, 70, 60, 120, 80, 100, 50, 40, 70, 90,
         20, 110, 100, 30, 50, 90, 110, 30, 90, 40, 80, 70)
HOURS <- c(399, 121, 221, 376, 361, 224, 546, 352, 353, 157, 160, 252, 389,
           113, 435, 420, 212, 268, 377, 421, 273, 468, 244, 342, 323)
LAB <- data.frame(LOT, HOURS)
str(LAB)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ LOT  : num  80 30 50 90 70 60 120 80 100 50 ...
##  $ HOURS: num  399 121 221 376 361 224 546 352 353 157 ...
```

```
head(LAB)
```

```
##   LOT HOURS
## 1  80   399
## 2  30   121
## 3  50   221
## 4  90   376
## 5  70   361
## 6  60   224
```

```
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
```



The least squares estimate is derived as follows.

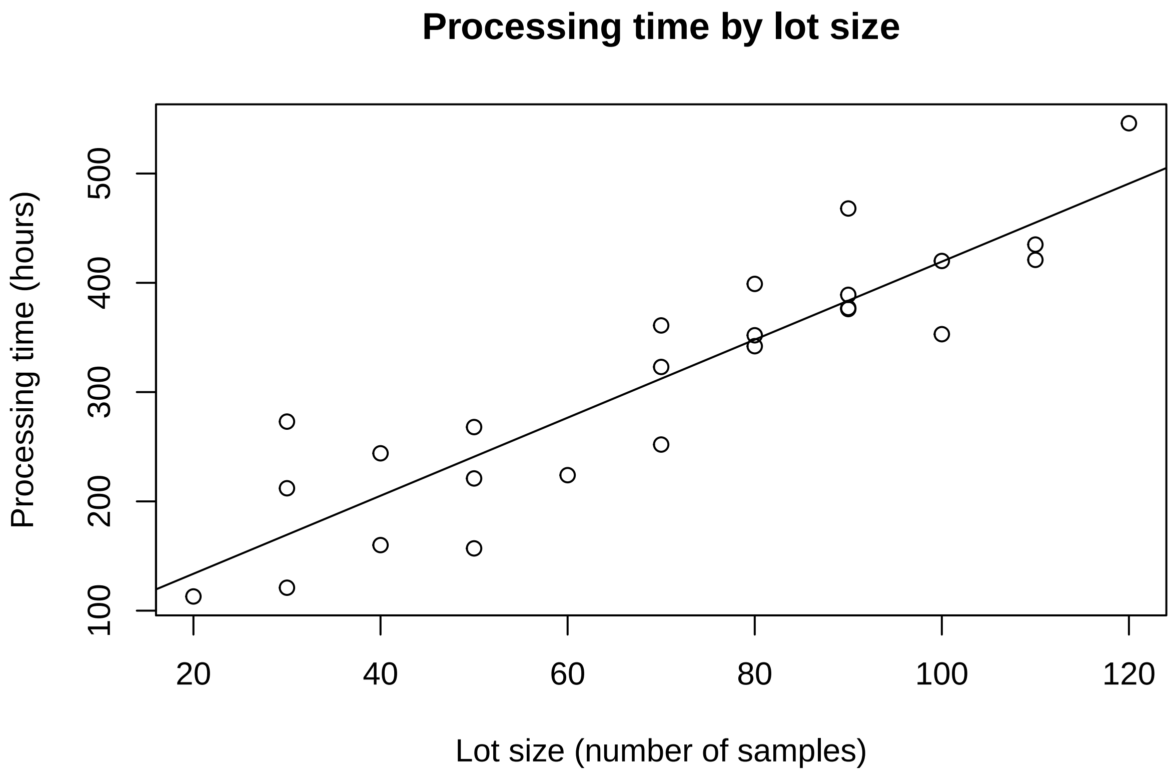
```
LAB.LM <- lm(HOURS ~ LOT, data = LAB)
summary(LAB.LM)
```

```
##
## Call:
## lm(formula = HOURS ~ LOT, data = LAB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## LOT            3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

The regression equation is $Processing\ Time = 62.4 + 3.6 \times \mathit{Lot\ Size}$. It will take 62.4 hours to process a lot size of zero. For every extra sample processed, the time to completion will increase by 3.6 hours.

Finally, the regression equation is overlaid onto the original scatterplot below.

```
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
abline(lm(HOURS ~ LOT, data = LAB))
```



Using the regression equation

If the results of the evaluation of the sample regression equation indicate that there is a relationship between the two variables of interest, we can put the regression equation to practical use by predicting what value Y is likely to assume given a particular value of X.

There are two forms of predictions. The first, called *interpolation*, attempts to provide a point estimate for in-scope values. The second form is called *extrapolation* and is used for values that are out of scope. Extrapolated estimates must be used with caution.

Example 2B

The lab hires a new lab technician and, after a period of training, has been assigned to perform DNA analysis. The lab head is interested in the length of time it will take this new person to process 57 samples and 600 samples.

```
predict(LAB.LM, data.frame(LOT = c(57, 600)))
```

```
##         1         2
##  265.8674 2204.4871
```

Using the model estimated earlier, we predict that it will take the new technician 265.9 hours to process 57 samples. The processing of 600 samples will take 2,204.5 hours.

Inferences concerning β_0 and β_1

Under the assumptions of the regression model, we can tests hypotheses and construct confidence intervals about β_0 and β_1.

Hypothesis testing

The sampling distributions of b_0 and b_1 are normally dsitrbuted with the following means and variances:

For b_0, $\mu_{b_0} = \beta_0$ and $\sigma^2_{b_0} = \dfrac{\sigma^2_{y|x} \sum X_i^2}{n \sum (x_i - \bar{x}^2)}$.

For b_1, $\mu_{b_1} = \beta_1$ and $\sigma^2_{b_1} = \dfrac{\sigma^2_{y|x}}{\sum (x_i - \bar{x}^2)}$.

In both these equations, $\sigma^2_{y|x}$ is the unexplained variance of the subpopulation of y values for a given x. Since $\sigma^2_{y|x}$ is often unknown, we estimate $\sigma^2_{y|x}$ by its sample counterpart $s^2_{y|x}$.

Remembering that the square root of the variance of a sampling distribution is called the *standard error*, the standard errors for β_0 and β_1 are

$$SE(b_0) = s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}^2)}}$$

$$SE(b_1) = \frac{s_{y|x}}{\sqrt{\sum (x_i - \bar{x}^2)}}$$

Knowing the estimates of the mean and standard error allows us to test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_A : \beta \neq 0$ using the t-test: $t = b/SE(b)$ with t being distributed as a Student's t with $n - 2$ degrees of freedom.

Example 2C

Using the data on lab sample processing times and assuming $\alpha = 0.05$, test the following null hypotheses: $H_{OA} : b_0 = 0$ and $H_{OB} : b_1 = 0$

The results of the previous simple linear regression can be retrieved.

```
summary(LAB.LM)
```

```
##
## Call:
## lm(formula = HOURS ~ LOT, data = LAB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## LOT            3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Note that t-values and p-values are already provided in the output. In the test for $H_{OA} : b_0 = 0$ the caluclated value for $p = 0.0259$. This means that on repeated sampling, we would observe at least as extreme a result as this for the intercept estimate about once per 39 trials under the null hypothesis. Similarly, the hypothesis test for the slope parameter shows $p = 4.45 \times 10^{-10}$. This means that we would observe at least as large a slope estimate as 3.6 with a probability of abour one in billion trials, assuming that the null hypothesis was true.

It must be emphasised that failure to reject the null hypothesis that $b_1 = 0$ does not mean that X and Y are not related. Not only is it possible that a type II error may have been committed but it may be true that X and Y are related in some nonlinear manner. On the other hand, when we reject the null hypothesis that $b_1 = 0$,we cannot conclude that the *true* relationship between X and Y is linear. Again, it may be that although the data fit the linear regression model fairly well (as evidenced by the fact that the null hypothesis that $b_1 = 0$ is rejected), some nonlinear model would provide an even better fit. Consequently, when we reject $H_0 : b_1 = 0$, the best we can say is that more useful results may be obtained by taking into account the regression of Y on X than in ignoring it.

Confidence intervals

Using the formulae for the standard errors shown above also allows us to construct $(1 - \alpha) \times 100\%$ confidence intervals following the general formula estimator $\pm$ (reliability factor)(standard error).

The $(1 - \alpha) \times 100\%$ confidence interval for β_0 is $b_0 \pm t_{(\alpha/2;df=n-2)} \times SE(b_0)$.

The $(1 - \alpha) \times 100\%$ confidence interval for β_1 is $b_1 \pm t_{(\alpha/2;df=n-2)} \times SE(b_1)$.

Exaxmple 2D

Produce 95% and 99% confidence intervals for the intercept and slope estimates derived for the lab data.

To produce confidence intervals, we use the `confint()` function.

```
confint(LAB.LM, level = 0.95)
```

```
##                2.5 %     97.5 %
## (Intercept) 8.213711 116.518006
## LOT         2.852435   4.287969
```

```
confint(LAB.LM, level = 0.99)
```

```
##                  0.5 %     99.5 %
## (Intercept) -11.122986 135.854703
## LOT           2.596135   4.544269
```

The interpretation of the confidence intervals are as follows. We are 95% confident that the intercept estimate lies between the interval 8.2 and 116.5 hours because on repeated sampling, 95% of intervals constructed in this manner will contain the true intercept. In addition, we are 95% confidence that the slope estimate lies between 2.9 and 4.2 because on repeated sampling, 95% of intervals constructed in this manner will contain the true slope.

Example 3

The purpose of a study by Brown and Persley [2] was to characterise acute hepatitis A in patients more than 40 years old. They performed a retrospective chart review of 20 subjects who were diagnosed with acute hepatitis A, but were not hospitalised. Of interest was the use of age (years) to predict bilirubin levels (mg/dl). The following data were collected.

Table 3. Age and bilirubin levels in 20 non-hospitalised acute hepatitis A patients

Age (years)	Bilirubin (mg/dl)	Age (years)	Bilirubin (mg/dl)
78	7.5	44	7.0
72	12.9	42	1.8
81	14.3	45	0.8
59	8.0	78	3.8
64	14.1	47	3.5
48	10.9	50	5.1
46	12.3	57	16.5
42	1.0	52	3.5
58	5.2	58	5.6
52	5.1	45	1.9

Conduct the required analysis and test the hypotheses at the $\alpha = 0.10$ level of significance. Compute the p-value for each test and interpret it. Then, calculate and interpret the 95% confidence intervals. Finally, predict the bilirubin levels of patients aged 40 and 70.

First, we enter the data.

```
AGE <- c(78, 44, 72, 42, 81, 45, 59, 78, 64, 47,
         48, 50, 46, 57, 42, 52, 58, 58, 52,45)
BILIRUBIN <- c(7.5, 7.0, 12.9, 1.8, 14.3, 0.8, 8.0, 3.8, 14.1, 3.5,
               10.9, 5.1, 12.3, 16.5, 1.0, 3.5, 5.2, 5.6, 5.1, 1.9)
HEPA <- data.frame(AGE, BILIRUBIN)
str(HEPA)
```

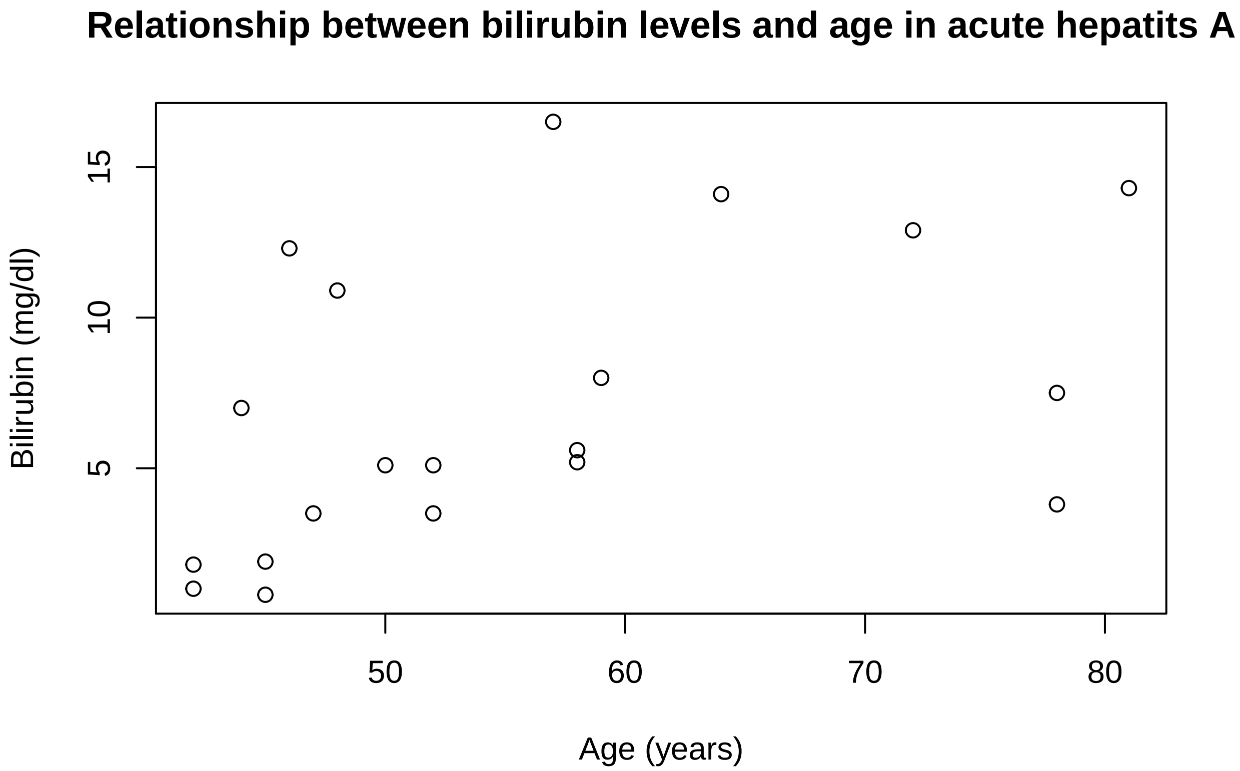
```
## 'data.frame':    20 obs. of  2 variables:
##  $ AGE      : num  78 44 72 42 81 45 59 78 64 47 ...
##  $ BILIRUBIN: num  7.5 7 12.9 1.8 14.3 0.8 8 3.8 14.1 3.5 ...
```

```
head(HEPA)
```

```
##   AGE BILIRUBIN
## 1  78       7.5
## 2  44       7.0
## 3  72      12.9
## 4  42       1.8
## 5  81      14.3
## 6  45       0.8
```

Then, we produce a scatterplot.

```
plot(BILIRUBIN ~ AGE, data = HEPA,
     main = "Relationship between bilirubin levels and age in acute hepatits A",
     ylab = "Bilirubin (mg/dl)",
     xlab = "Age (years)")
```



The points show a general spread with a slight positive relationship.

Next, we regress bilirubin on age using a simple linear regression.

```
HEPA.LM <- lm(BILIRUBIN ~ AGE, data = HEPA)
summary(HEPA.LM)
```

```
##
## Call:
## lm(formula = BILIRUBIN ~ AGE, data = HEPA)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.203 -2.927 -1.529  2.812  9.263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.9842     4.5962  -0.649   0.5244
## AGE           0.1793     0.0803   2.233   0.0385 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.414 on 18 degrees of freedom
## Multiple R-squared:  0.2169, Adjusted R-squared:  0.1734
## F-statistic: 4.987 on 1 and 18 DF,  p-value: 0.03848
```

```
confint(HEPA.LM, level = 0.90)
```

```
##                       5 %      95 %
## (Intercept) -10.95429006 4.9859136
## AGE           0.04007251 0.3185747
```

The results show that the level of bilirubin increases by 0.18 (90% CI 0.04, 0.32) ml/dl per year of life ($p = 0.0385$). The result is statistically significant at $\alpha = 0.10$. We are 90% confident that the level of bilirubin increase is between 0.04 and 0.32 mg/dl. The intercept, being out of scope and impossible, has no practical meaning.

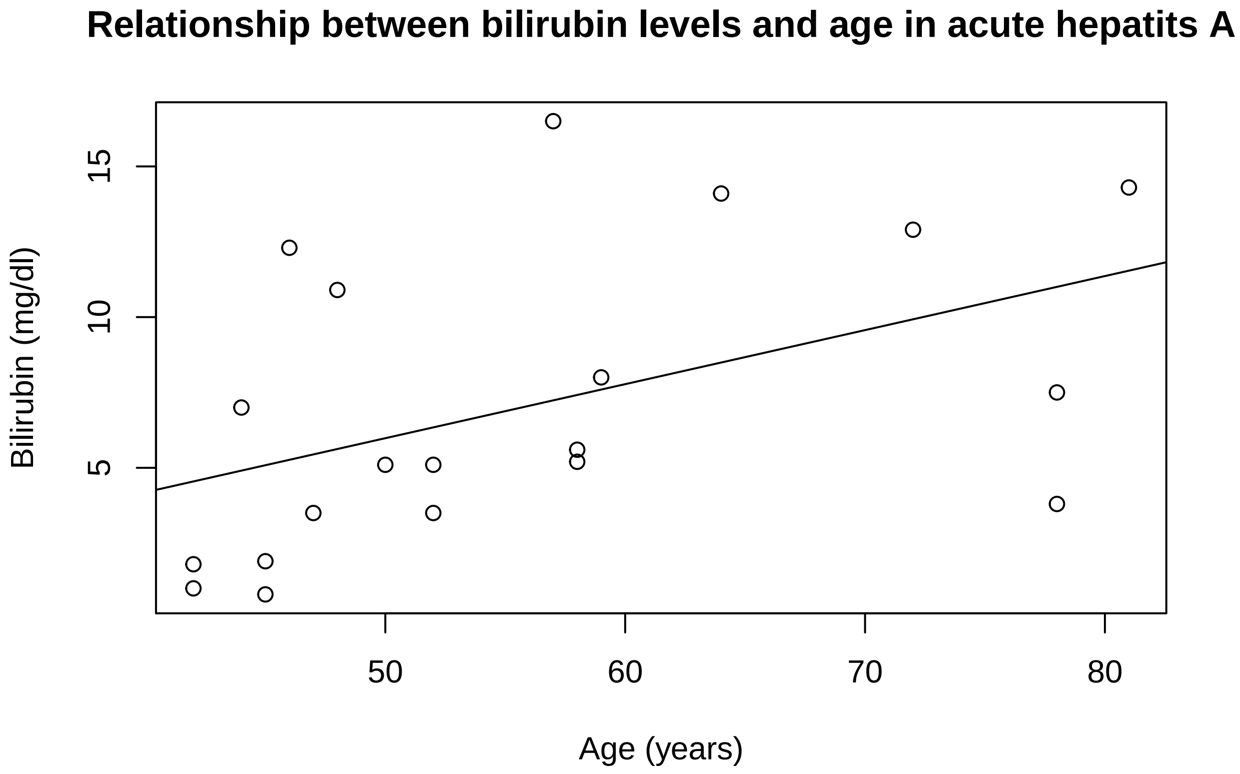
```
predict(HEPA.LM, data.frame(AGE = c(40, 70)))
```

```
##        1        2
## 4.188755 9.568463
```

The predicted bilirubin levels for a 40-year-old patient and a 70-year-old patient are 4.19 and 9.57 ml/dl, respectively.

Finally, a scsatterplot of the relationship is shown.

```
plot(BILIRUBIN ~ AGE, data = HEPA,
     main = "Relationship between bilirubin levels and age in acute hepatits A",
     ylab = "Bilirubin (mg/dl)",
     xlab = "Age (years)")
abline(HEPA.LM)
```



References

1. Parker RB, Yates R, Soberman JE, Laizure C. Effects of grapefruit juice on intestial P-glycoprotein: Evaluation using digoxin in humans. *Pharmacotherapy* 2003;23:979-987.

2. Brown GR, Persley K. Hepatitis A epidemic in the elderly. *Southern Medical Journal* 2002;95:826-833.

THE END