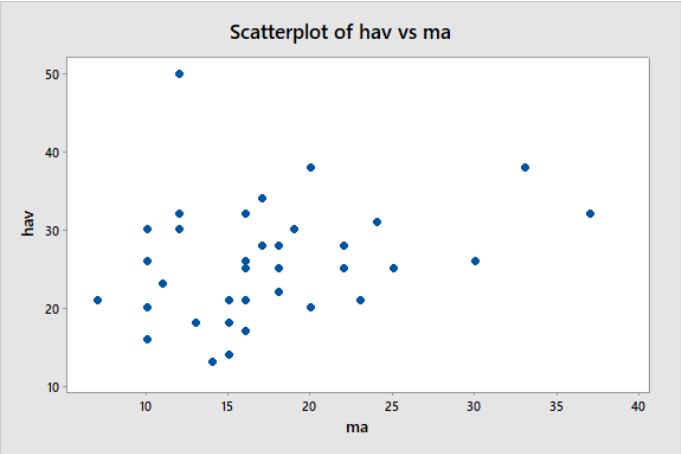
Supplementary exercises 2.11, 2.48, 10.38 and 10.39 of IPS7e

Data: Measurements on 38 patients of MAV and HA angles (two deformations of the foot, the former of which is less severe and is hypothesized to be useful as a predictor of the latter). Two response variables, and samples can be assumed i.i.d. from a population for which the 38 patients are considered representative.

Exercise 2.11

(a) For the scatterplot, we put the MA value on the x-axis because there is interest in predicting HAV from MA. This is, however, of primary interest in the context of regression, and for the purpose of studying the strength and direction of association between the two variables the axes might as well be reversed. Both variables are measured as responses, so none of them would (in our usage) be labeled as explanatory.



(b) The association is clearly positive, and maybe approximately linear; the points are so scattered that it is hard to tell whether it is linear or not. There is one observation (no. 31; MA=12 and HAV=50) which seems outside the pattern of the other observations; we may consider this observation a suspected outlier. Even without this observation the association is rather weak.

(c) There may be a positive association between the two measurements but it seems too weak to allow any useful prediction of HAV from an MA-value. To quantify our impression of the association, we could compute the correlation (Session 12 of VHM 801): $r = 0.302$; as expected, it is positive but quite low.

Exercise 2.48

(a) Although least-squares estimation of the regression line is based on formal assumptions, we defer listing of those to a later question. The fitted line menu in Minitab provides the desired plot.

```
MTB > Fitline 'hav' 'ma';
SUBC>   Confidence 95.0.
```

Regression Analysis: hav versus ma

The regression equation is
hav = 19.72 + 0.3388 ma

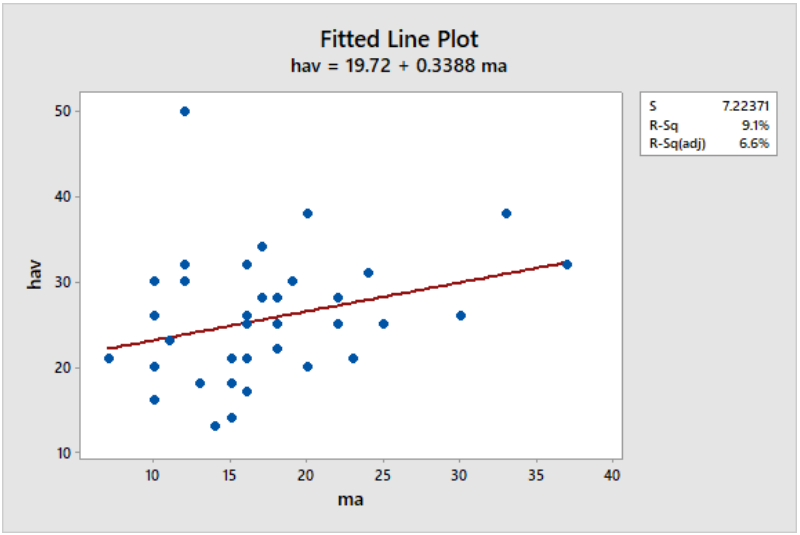
Model Summary

S	R-sq	R-sq(adj)
7.22371	9.13%	6.60%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	188.71	188.714	3.62	0.065
Error	36	1878.55	52.182		
Total	37	2067.26			

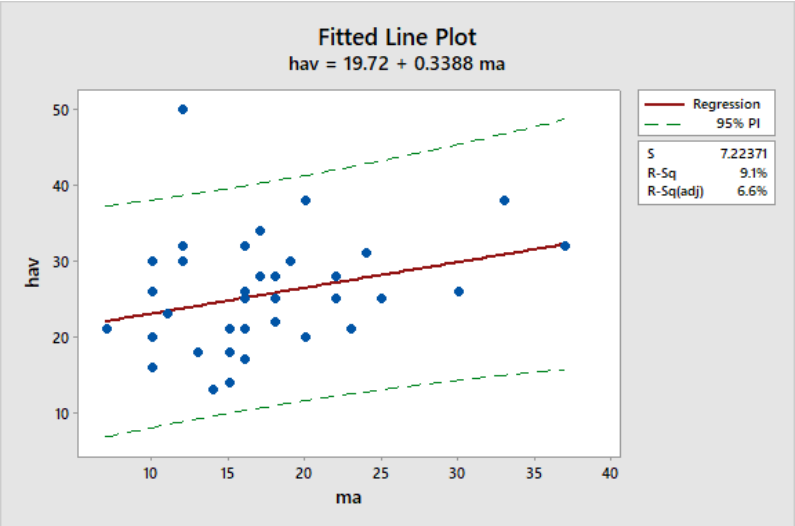
Fitted Line: hav versus ma



(b) The listing above gives the prediction equation as: HAV $= 19.72 + 0.3388 \cdot$ MA. Therefore, from an MA-value of 25 we predict the HAV to be:

$$\widehat{\text{HAV}} = 19.72 + 0.3388 \cdot 25 = 28.19.$$

(c) The intention of the question for a numerical measure of the "accuracy" of the prediction is probably $R^2 = 0.091$. (Note that if we distinguish between accuracy and precision, this would be referring to the precision.) However, there are two better ways of quantifying the precision in the equation. One is by the residual standard deviation, s=7.224. It measures the spread of the points about the line. Roughly speaking, the points should be within a band of $\pm 2s$ of the line; this band is very wide and includes almost the entire range of HAV-values. Even better, we can compute 95% prediction intervals for new observations; these can be overlaid the fitted line plot by choice of the relevant option.



The display shows that the prediction intervals are so wide that they will be pretty much useless for any practical application.

Exercise 10.38
Note that parts (a) and (b) were already answered above.

(c) The linear regression model is:

$$\text{HAV}_i = \beta_0 + \beta_1 \text{MA}_i + \varepsilon_i,$$

where β_0 is the intercept, β_1 is the slope and the errors $\varepsilon_1, \ldots, \varepsilon_{38}$ are assumed i.i.d. from $\text{N}(0, \sigma)$.

(d) The "question of interest" is perhaps not so obvious from the description. The intention is clearly to carry out a hypothesis test for association between MA and HAV, but from the description it appears that the real interest is whether the relation can be used for prediction. That depends not only on whether an association exists, but also on its strength. Here we will assess the statistical significance of the association.

H_0 : $\beta_1 = 0$ no association between MA and HAV),
H_a : $\beta_1 \neq 0$ some association (but no particular direction) between MA and HAV.

(e) We have two essentially equivalent ways of testing the hypothesis, a t-test and the F-test in the ANOVA table. The ANOVA table was already included in the listing for the fitted line plot, with the results:

$$F = 3.62, \quad \text{df} = (1, 36), \quad P = 0.065.$$

We need to carry out the analysis through the Regression menu to get more information about the parameters, including the t-test.

```
MTB > Regress;
SUBC>   Response 'hav';
SUBC>   Nodefault;
SUBC>   Continuous 'ma';
SUBC>   Terms ma;
SUBC>   Constant;
SUBC>   Unstandardized;
SUBC>   Tmethod;
SUBC>   Tanova;
SUBC>   Tsummary;
SUBC>   Tcoefficients;
SUBC>   Tequation;
SUBC>   TDiagnostics 0.
```

Regression Analysis: hav versus ma

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	188.7	188.71	3.62	0.065
ma	1	188.7	188.71	3.62	0.065
Error	36	1878.5	52.18		
Lack-of-Fit	17	1105.9	65.05	1.60	0.161
Pure Error	19	772.6	40.66		
Total	37	2067.3			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.22371	9.13%	6.60%	0.60%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	19.72	3.22	6.13	0.000	
ma	0.339	0.178	1.90	0.065	1.00

Regression Equation

hav = 19.72 + 0.339 ma

Fits and Diagnostics for Unusual Observations

Obs	hav	Fit	Resid	Std Resid	
5	38.00	30.90	7.10	1.09	X
31	50.00	23.79	26.21	3.70	R
37	32.00	32.26	-0.26	-0.04	X

R Large residual
X Unusual X

These listings include the t-test for H_0:

$$t = 1.90, \quad \text{df} = 36, \quad P = 0.065.$$

We conclude (from either test) that there is no evidence against H_0 at the 5% level. However, the P-value is close so we may say that there is some indication of a weak association. Note that the only advantage of the t-test over the F-test is that it can be used with a one-sided alternative hypothesis.

Exercise 10.39
From the Minitab listing above, the 95% CI for β_1 is computed as

$$\hat{\beta}_1 \pm t^* \cdot \text{SE}(\hat{\beta}_1) = 0.3388 \pm 2.028 \cdot 0.1782 = 0.339 \pm 0.361 = (-0.023, 0.700),$$

where $t^* = t_{.975}(36) = 2.028$ is obtained from Minitab (from the t-distribution table we would use $t_{.975}(30) = 2.042$).

By the fact that zero is contained in the confidence interval, we would be able to conclude that the test for H_0 previously computed is non-significant at the 5% level. Confidence intervals for the regression parameters can be displayed by Minitab's Regression menu by choosing the Expanded tables under Results.

Addendum: Analyses without suspected outlier
The observation in question was noted previously (HAV = 50 and MA = 12). The HAV value is the largest in the dataset, and it occurs for a patient with a relatively low MA angle. The fitted line plots show this observation as clearly above the other data points. We carry out this analysis in Minitab by copying the HAV-values to another column, replacing the value 50 by missing, and rerunning the regression analyses.

```
MTB > Copy 'hav' c3;
SUBC>   Varnames.
MTB > let c3(31)='*'
MTB > Fitline 'hav_1' 'ma';
SUBC>   Confidence 95.0.
```

Regression Analysis: hav_1 versus ma

The regression equation is
hav_1 = 17.66 + 0.4189 ma

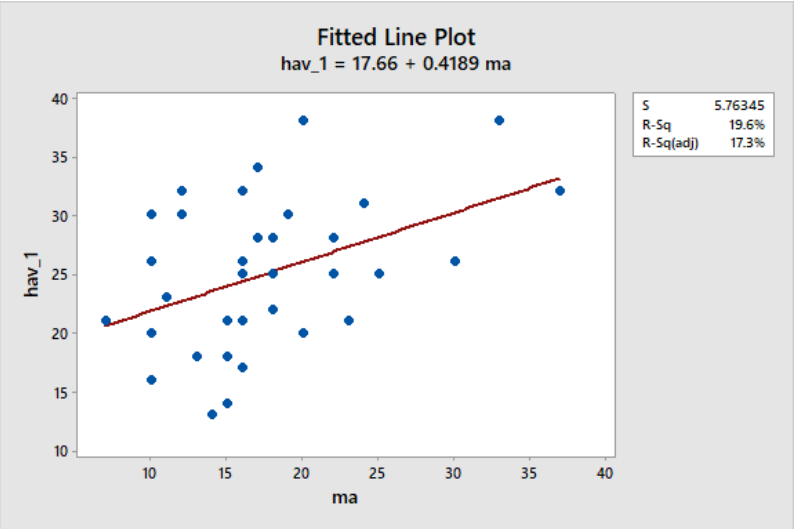
Model Summary

S	R-sq	R-sq(adj)
5.76345	19.64%	17.35%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	284.20	284.205	8.56	0.006
Error	35	1162.61	33.217		
Total	36	1446.81			

Fitted Line: hav_1 versus ma



Comments:
Without the suspected outlier, the association between MA and HAV is clearly significant ($P = 0.006$). Even if the standard deviation about the line dropped to 5.76 it is still quite large. The slope of the line increased to 0.42 and the intercept dropped to 17.7. Thus, the fitted line starts a little lower and is somewhat steeper, but the difference is not huge when comparing the plots.

As to the question whether the suspected outlier should be deleted from the data, it is possible to show (using methods beyond this course) that the observation is very unlikely to have occurred by chance alone. A test for whether this observation can be described by the same model as the others, gives a P-value of 0.002. This still does not mean that it *must* be dropped, but it provides justification to do so. The practical difference between the results of the two analyses is only minor, because even the model without the suspected outlier is not useful for prediction.