1. Prove the following:
    a. (1 point) If $\hat{\theta}_1$ is an unbiased estimator for θ, and X is a random variable with mean μ=0, then $\hat{\theta}_2 = \hat{\theta}_1 + X$ is also an unbiased estimator for θ.
    b. (1 point) If $\hat{\theta}_1$ is an unbiased estimator for θ such that $\mathrm{E}\left[\hat{\theta}_1\right] = a\theta + b$, where $a = /0$, then $\hat{\theta}_2 = \frac{\hat{\theta}_1 - b}{a}$ is also an unbiased estimator for θ.

Solution:

a. We have

$\mathrm{E}\left[\hat{\theta}_2\right] = \mathrm{E}\left[\hat{\theta}_1\right] + E[X]$ by linearity of expectation

$\mathrm{E}\left[\hat{\theta}_2\right] = \theta + 0 = \theta$ since $\hat{\theta}_1$ is unbiased and μ=0

Therefore $\hat{\theta}_2$ is unbiased estimator for θ

b. We have

$$\mathrm{E}\left[\hat{\theta}_2\right] = \frac{\mathrm{E}\left[\hat{\theta}_1\right] - b}{a} = \frac{a\theta + b - b}{a} = \theta$$

Thus $\hat{\theta}_2$ is unbiased estimator for θ

2. To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$\overline{x_1} = 3{,}51$	$\overline{x_2} = 3{,}24$
$S_1 = 0{,}51$	$S_2 = 0{,}52$

    a. (1,5 points) Build a 99% confident interval for the difference in average satisfaction levels of customers of the two companies as measured on this five-point scale, and explain its meaning.
    b. (1,5 points) Perform a 1% hypothesis test to decide whether customers of Company 1 are more satisfied than those of Company 2. Explain the result of the test.
    c. (1 point)Calculate the p-value of the previous test

Solution:

a. We have to large independent samples, therefore the confident Interval for the difference of means is:

$$\mu_1 - \mu_2 \in \left[\bar{x}_1 - \bar{x}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = 0{,}27 \pm 2{,}576\sqrt{\frac{0{,}51^2}{174} + \frac{0{,}52^2}{355}} = 0{,}27 \pm 0{,}12$$

So the Interval is [0,15; 0,39]. The Interval does not contain value 0, therefore both mean values cannot be equal with a 99% leve lof confidence, being the average level of customer satisfaction for Company 1 between 0,15 and 0,39 points higher than that for Company 2.

b. We test the following hypothesis with α=0,01:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

The test statistic is:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0{,}27}{\sqrt{\frac{0{,}51^2}{174} + \frac{0{,}52^2}{355}}} = 5{,}684$$

This is a right-tailed test with a single critical value $z_\alpha = z_{0.01} = 2{,}326$, therefore the rejection area is [2,326,∞]. The test statistic falls in the rejection area so we reject H_0 which means that customers from company 1 are more satisfied than customers from company 2, as we had already found out in question a.

c. P-Value = $P(Z > z_0) = P(Z > 5{,}684) = 1 - P(Z < -5{,}684) \sim 0$

3. We want to explain a certain varible Y by means of variables X1, X2, X3 and X4. First we obtain their dispersion matrix which is shown in figure 1, and then we try to build multiple linear regression models starting with the four variables and then removing one at a time until we try only with variables X1 and X2. The summary of each model and their corresponding residual grahps are shown in figures 2 to 4.

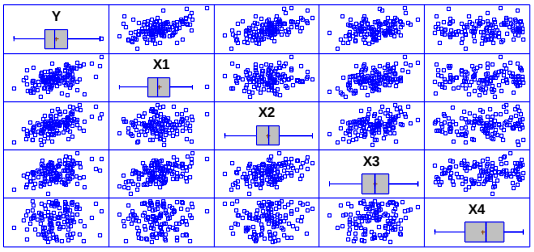

Figure 1.

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	47,6932	20,2909	2,35047	0,0203
X1	25,9569	4,5788	5,66893	0,0000
X2	29,7292	3,95144	7,52365	0,0000
X3	-0,106052	0,115617	-0,917275	0,3608
X4	2,14434	13,1196	0,163446	0,8704

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	179183,	4	44795,8	27,26	0,0000
Residual	202098,	123	1643,07		
Total (Corr.)	381281,	127			

R-squared = 46,995 percent
R-squared (adjusted for d.f.) = 45,2713 percent

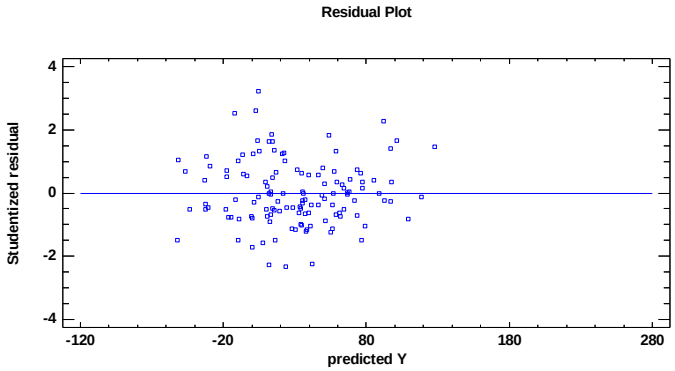

Figure 2.

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	48,6092	19,425	2,5024	0,0136
X1	25,9502	4,56061	5,69006	0,0000
X2	29,7668	3,92923	7,57575	0,0000
X3	-0,104584	0,114814	-0,910902	0,3641

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	179139,	3	59713,1	36,63	0,0000
Residual	202142,	124	1630,18		
Total (Corr.)	381281,	127			

R-squared = 46,9835 percent
R-squared (adjusted for d.f.) = 45,7009 percent

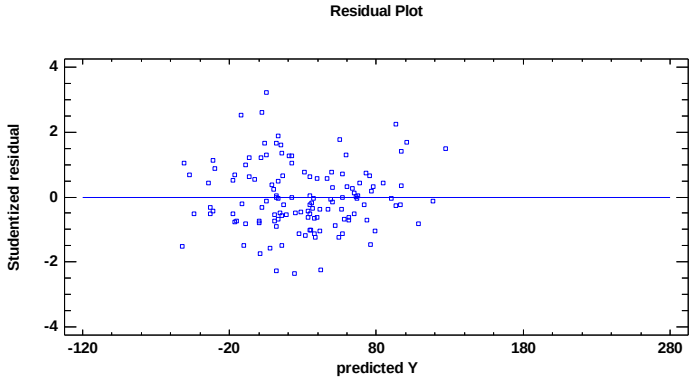

Figure 3.

			Standard	*T*	
Parameter		*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT		31,2164	3,5684	8,74801	0,0000
X1		23,9501	3,99456	5,99569	0,0000
X2		28,2857	3,57457	7,91304	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	177787,	2	88893,3	54,60	0,0000
Residual	203494,	125	1627,96		
Total (Corr.)	381281,	127			

R-squared = 46,6288 percent
R-squared (adjusted for d.f.) = 45,7748 percent

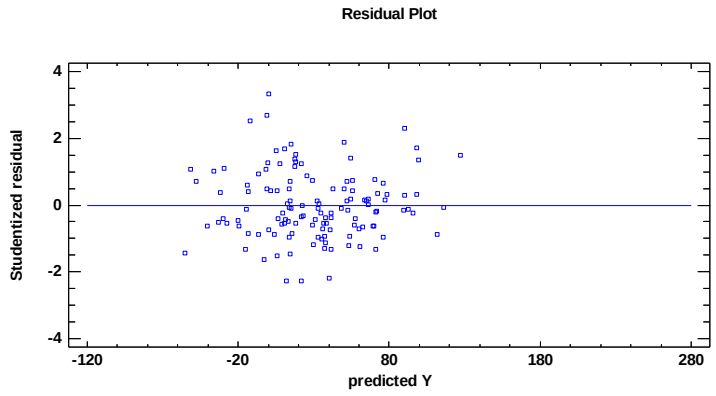

Figure 4.

a. (1 point) What conclusions can be drawn from the dispersion matrix?
b. (2 points) Write the expression for a valid model and interpret the coefficients of the regressos as well as the coefficient R^2.
c. (0,5 points) Explain what variables are removed from one model to the next one and why
d. (0,5 points) Explain what happens to the R^2 coefficients

Solution:

a. The dispersion matrix shows that variables X1 and X2 have a linear relationship with variable Y. Variable X3 could have a linear relationship with variable Y but very weak, and variable X4 does not have any linear relationship with variable Y at all.
b. $Y = 31{,}2164 + 23{,}9501 * X1 + 28{,}2857 * X2$
(P-value=0) (P-value=0)

R^2_{adj}= 45,77%

Is the only valid model because all its variables are significant (P-value<0,05) and the residual plot does not show any pattern. The value for the coefficient for X1, 23,9501, indicates that if variable X1 increases by one unit, keeping constant variable X2, then variable Y will increase in 23,9501 units. The value for the coefficient for X2, 28,2857, indicates that if variable X2 increases by one unit, keeping constant variable X1, then

variable Y will increase in 28,2857 units. R^2_{adj}= 45,77% indicates that this model explain 45,77% of variable Y.

c. The first model obtained had two variables that were not significant (P-value>0,05), so the one with the highest P-value (X4) was removed to calculate the second model where still there was a variable non significant (X3), so we also removed it, and this time the model obtained had two significant variables (X1 and X2).
d. Coeffient R^2_{adj} increases from one model to the next becuse the model improves each time we remove a non significant variable. Coefficient R^2 decreases from one model to the next because the number of variables present in the model decreases.