Solutions of the exercises on Linear Regression and Analysis of Variance (part 3 Applied Statistics)

Exercise 1

a. $y = \beta_0 + \beta_1 x + \varepsilon$, where $y =$ "yearly amount spend on food" and $x =$ "the yearly income" and the independent disturbances ε all have a $N(0, \sigma^2)$-distribution (the parameters β_0, β_1 and σ^2 are unknown parameters

b. Using the coefficients table of the output: $\hat{\beta}_0 = -0.916$ and $\hat{\beta}_1 = 0.481$.
Estimate of σ^2: 'residual mean square' = 3.170

c. First we compute $r^2 = 1 - \frac{SS(Errors)}{SS(Total)}$ We will use the ANOVA table (but one could also determine the sample variance of the y-values and use $SS(Total) = (n-1)S_y^2$)
$\Rightarrow r^2 = 1 - \frac{44.384}{116.000} = 0.6174$, Since $\hat{\beta}_1$ is positive, then so is r: $r = \sqrt{0.6174} = 0.786$

d. 95%-CI$(\beta_1) = \left(\widehat{\beta}_1 - c\frac{S}{\sqrt{\sum_i (x_i - \bar{x})^2}}, \widehat{\beta}_1 + c\frac{S}{\sqrt{\sum_i (x_i - \bar{x})^2}} \right)$, where $P(T_{n-2} \geq c) = \frac{1}{2}\alpha$ (Property 3.1.5)
From the output we find $\hat{\beta}_1 = 0.481$ and the standard error $\frac{S}{\sqrt{\sum_i (x_i - \bar{x})^2}} = 0.101$ and from the t_{16-2}-table:
$c = 2.145$, so 95%-CI$(\beta_1) = (0.481 - 2.145 \times 0.101, 0.481 + 2.145 \times 0.101) = (0.264, 0.698)$

e. 1. $y = \beta_0 + \beta_1 x + \varepsilon$, where $y =$ "yearly amount spend on food" and $x =$ "the yearly income" and the independent disturbances ε all have a $N(0, \sigma^2)$-distribution.
2. Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ with α = 1%
3. Test statistic: $T = \frac{\widehat{\beta}_1}{\sqrt{S^2 / \sum_i (x_i - \bar{x})^2}}$
4. Under $H_0 : T \sim t_{n-2} = t_{14}$
5. Observed value of T: $\frac{0.481}{0.101} = 4.76$ (output difference by rounding)
6. We will reject H_0 if $T \leq -c$ or $T \geq c$ at significance level 1%, so $c = 2.977$, such that $P(T_{14} \geq c) = 0.005$
7. 4.76 lies in the Rejection Region $\Rightarrow$ reject H_0
8. We consider it proven, at a 1% level of significance, that the yearly income is related to the yearly amount spend on food.

Exercise 2

a. $y = \beta_0 + \beta_1 x + \varepsilon$, where $y =$ "Turnover in Spring + Summer" and $x =$ "Turnover in Autumn + Winter" and the independent disturbances ε all have a $N(0, \sigma^2)$-distribution (the parameters β_0, β_1 and σ^2 are unknown parameters)

b. Just using the output: the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are respectively -11.489 and 0.98737, giving the equation of the estimated regression line: $y = -11.489 + 0.98737x$.

c. The proportion explained variation is: $R^2 = 1 - \frac{SS(Errors)}{SS(Total)} = 1 - \frac{7670.40}{91805.34} = 91.6\%$.
This proportion is close to 1, so that the regression model is helpful to predict the turnover in Summer+ Spring based on the turnover in Autumn + Winter.

d. Prediction: $\hat{y} = -11.489 + 0.98737 \times 140 = 126.7$

e. 95%-$PI(Y|x^* = 140)$ has interval boundaries $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm c\sqrt{S^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]}$,
where $c = 2.160$ is such that $P(T_{15-2} \geq c) = \frac{1}{2}\alpha = 2.5\%$, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 126.7$ (see part d.)
$S^2 = 590.03$ ($MS(Errors)$ in ANOVA-table), $n = 15$, $\bar{x} = 156.88$ (given) and
$\sum_i (x_i - \bar{x})^2 = (n-1)s_x^2 = 14 \cdot (78.51)^2 = 86393.5$:
the interval is $(126.7 - 2.160 \times 25.13, 126.7 + 2.160 \times 25.13) = (72.4, 181.0)$
(*Compare to the confidence interval for the expected turnover in Spring/Summer:*
The 95%-$CI(E(Y|x^* = 140))$ *has interval boundaries* $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm c\sqrt{S^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]}$, *where:* $x^* = 140$

and all other "ingredients" are the same as in part e. (the standard error is smaller now: 6.425):
95%-CI$(\beta_0 + \beta_1 x^*) = (126.7 - 2{,}160 \cdot 6.425,\ 126.7 + 2{,}160 \cdot 6.425) \approx (112.8,\ 140.6)$)

f. The value 140.62 is pretty much in the middle of the prediction interval, established in e. Meaning that we cannot be sure that 140.62 is sufficient: if you would order 181 hl we can be 95% confident it is sufficient. Note that 140.6 hl is the upper bound of the 95% confidence interval, but this is an interval for the mean turnover (mean over many years) and does not give information about the turnover in a specific year.

Exercise 3

a. Estimates: $\hat{\beta}_0 = 162.281$, $\hat{\beta}_1 = -81.304$.

b. The standard errors: $se(\hat{\beta}_0) = 8.963$ and $se(\hat{\beta}_1) = 7.305$
These are the estimates of the standard deviations of the underlying distributions of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
We have $var(\hat{\beta}_1) = \sigma^2/S_{xx}$ with $S_{xx} = \sum_i (x_i - \bar{x})^2$
The standard deviation of $\hat{\beta}_1$ is thus $SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$, estimated by the standard error $se(\hat{\beta}_1) = \frac{S}{\sqrt{S_{xx}}} = 7.305$.

c. We are estimating σ^2 by means of $S^2 = \frac{\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$: we have to divide the residual sum of squares (numerator) by the corresponding degrees of freedom $(n-2)$: we get 9.125

d. 95% confidence interval for β_1 : boundaries are $\hat{\beta}_1 \pm c \times se(\hat{\beta}_1)$
Elaboration: $\hat{\beta}_1 = -81.304$ and $se(\hat{\beta}_1) = 7.305$ (from the output)
$c = 2.306$ (t-distribution, $n-2 = 8$ degrees of freedom, 95% probability between $-c$ and c)
95% confidence interval becomes: $(-81.304 - 2.306 \times 7.305, -81.304 + 2.306 \times 7.305) = (-98.1\ , -64.5)$

e. Let Y denote the hardness and let denote x the corresponding annealing temperature. The eight steps of the test:

1. $Y = \beta_0 + \beta_1 x + \varepsilon$, with independent disturbances ε being $N(0, \sigma^2)$ -distributed.
2. We test H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$
3. Test statistic: $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$
4. Under $H_0 : T \sim t_{n-2} = t_8$
5. Outcome of T: $T = \frac{-81.304}{7.305} = -11.130$ (from output)
6. We reject H_0 if $T \leq -c$ or . $\alpha = 5\%$, table $\Rightarrow c = 2.306$
7. Rejection region contains outcome -11.130 so reject H_0 .
8. Using level of significance 5% we have proven that the hardness depends on the annealing temperature.

f. Looking at the scatter plot we should see only 'chaos', no pattern whatsoever. We have to learn to judge residual plots. At first glance there is no pattern, perhaps caused by the fact that there are only a few points. Having a closer look you can distinguish a banana-shaped or V-shaped cloud of points vaguely. This has its origin in the plot of y versus x: a slight curvature of the cloud of points can be distinguished as soon as you draw a straight line in the middle of the points. So a close examination of the plots may give rise to some doubt about the fit of the model. We return to this data set in another exercise. Then we investigate whether quadratic regression fits the data better.

Exercise 4

a. The model equation becomes: $Y = \beta_1 x + \varepsilon$ ('regression through the origin')
We estimate the parameter β_1 by means of least squares:
Minimize $\sum_i (y_i - \beta_1 x_i)^2$ as function of β_1.
Differentiate: $-2\sum_i (y_i - \beta_1 x_i) x_i$
Equating the derivative to zero gives an equation for the estimate $\hat{\beta}_1$: $-2\sum_i (y_i - \hat{\beta}_1 x_i) x_i = 0$
Solving: $\sum_i (y_i - \hat{\beta}_1 x_i) x_i = \sum_i x_i y_i - \hat{\beta}_1 \sum_i x_i^2 = 0$ and hence $\hat{\beta}_1 = \sum_i x_i y_i / \sum_i x_i^2$.
You can check that indeed a minimum has been attained: the second derivative is $+\sum_i x_i^2 > 0$

b. Now we have to study the new estimator $\hat{\beta}_1 = \sum_i x_i Y_i / \sum_i x_i^2$ with independent random variables Y_i distributed according to a normal distribution with expecation $\beta_1 x_i$ and (common) variance σ^2.
We compute its variance: $var(\hat{\beta}_1) = var(\sum_i x_i Y_i / \sum_i x_i^2) = \frac{var(\sum_i x_i Y_i)}{(\sum_i x_i^2)^2} = \sum_i x_i^2 var(Y_i) / (\sum_i x_i^2)^2$

$$= \sigma^2 \sum_i x_i^2 / (\sum_i x_i^2)^2 = \sigma^2 / \sum_i x_i^2$$

In general $\sum_i x_i^2 > \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$ so the "new" variance of $\hat{\beta}_1$ tends to be smaller than the "old" variance $\sigma^2 / \sum_i (x_i - \bar{x})^2$.

Exercise 5

a. Let Y denote the mortality rate per 100 000 males. Let us first investigate whether the variable *North* really affects the mortality rate Y , in addition to the variable *Calcium.* The eight steps of the test:

(1) $Y = \beta_0 + \beta_1 Calcium + \beta_2 North + \varepsilon$, with independent disturbances ε which are $N(0, \sigma^2)$ -distributed.
(2) We test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$
(3) Test statistic: $T = \hat{\beta}_2 / se(\hat{\beta}_2)$
(4) Under $H_0 : T \sim t_{61-3} = t_{58}$
(5) Outcome of T : $\frac{176.711}{36.891} = 4.79$
(6) We reject H_0 if $T \leq -c$ or . $T \geq c$. $\alpha = 5\%$, $t_{58} \Rightarrow c = 2.00$ (interpolation)
(7) The rejection region contains the outcome 4.79 so reject H_0 .
(8) Using level of significance 5% we have proven that the variable *North* really affects the mortality rate *Y*, in addition to the variable *Calcium.*

The test for investigating whether C*alcium* really affects mortality rate *Y*, in addition to *North,* is as follows:
(1) $Y = \beta_0 + \beta_1 Calcium + \beta_2 North + \varepsilon$, with independent disturbances ε which are $N(0, \sigma^2)$ -distributed.
(2) We test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$
(3) Test statistics: $T = \hat{\beta}_1 / se(\hat{\beta}_1)$
(4) Under $H_0 : T \sim t_{61-3} = t_{58}$
(5) Outcome of T : $\frac{-2.034}{0.483} = -4.21$
(6) We reject H_0 if $T \leq -c$ or $T \geq c$. $\alpha = 5\%$, $t_{58} \Rightarrow c = 2.00$ (interpolation)
(7) The Rejection region contains the outcome -4.21 so reject H_0 .
(8) Using level of significance 5% we have proven that the variable *Calcium* really affects the mortality rate *Y*, in addition to the variable *North*

b. We shall test whether it is useful to use the predictors *Calcium* and *North* for prediction of *Y*. (According to some textbooks this should be the first step in a statistical analysis.) The eight steps of the test:
(1) $Y = \beta_0 + \beta_1 Calcium + \beta_2 North + \varepsilon$, with independent disturbances ε which are $N(0, \sigma^2)$ -distributed.
(2) We test $H_0 : \beta_1 = \beta_2 = 0$ against $H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0$
(3) Test statistic: $F = \frac{SS(regression)/k}{SS(error)/(n-k-1)} = \frac{SS(regression)/2}{SS(error)/58}$
(4) Under $H_0 : F \sim F_{58}^2$
(5) Outcome of F : $\frac{1248317.8/2}{864855.86/58} = \frac{624158.905}{14911.308} = 41.86$
(6) We reject H_0 if $F \geq c$. $\alpha = 5\%$, $F_{58}^2 \Rightarrow c = \frac{2}{20} \times 3.23 + \frac{18}{20} \times 3.15 = 3.16$ (interpolation between F_{40}^2 and F_{60}^2)
(7) Rejection Region contains outcome 41.86 so reject H_0 .

(8) Using level of significance 5% we have proven that at least one of the predictor variables is useful for the prediction of the mortality rate Y.

c. We are searching for a pattern in the plot. There is no pattern, only chaos, this OK. No doubt about the fit. Note we are looking in the vertical direction, judging the residuals. The points are not evenly spread in the horizontal direction. This makes the judgement about the residuals more difficult.
Furthermore we can search for outliers. Nearly all standardized residuals are contained by the interval $(-2,2)$. On the average only 5% of the standardized residuals should be lying outside $(-1.96,1.96) \approx (-2.2)$. To our opinion the largest residual is not extreme enough to worry about.

d. $R^2_{adj} = 1 - \frac{n-1}{n-k-1} \times (SS_E/SS_T) = 1 - \frac{60}{58} \times \frac{864855.86}{2113173.7} = 57.7\%$
This means that 57.7% of the spread of the dependent variable is explained by the two predictor variables. This is not a good value, weak relationship, if we want predict the dependent variable because it is far away from the optimal value 100%. But this time it is natural: It is a bad situation if you could predict the mortality rate perfectly by means of only the calcium concentration and geographic variable *North.*

Exercise 6

a. The eight steps of the test:

(1) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, with independent disturbances ε which are $N(0,\sigma^2)$-distributed.

(2) We test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$

(3) Test statistic: $T = \hat{\beta}_2 / se(\hat{\beta}_2)$

(4) Under $H_0 : T \sim t_{10-3} = t_7$

(5) Outcome of T: $\frac{131.898}{40.648} = 3.25$

(6) We reject H_0 if $T \leq -c$ or $T \geq c$, where $\alpha = 5\% \Rightarrow c = 2.365$

(7) The rejection region contains outcome 3.25 so reject H_0.

(8) Using level of significance 5% we have proven that the quadratic term has to be added to the model.
We thus have proven that quadratic regression is better than simple linear regression.

b. We have to judge whether the residual plot shows chaos. That seems to be the case. It is a little bit troublesome that there are more points in the left part of the plot, suggesting unequal spread of the residuals. I guess we need more data for assessing such a pattern.

c. Modified scheme of eight steps:

(1) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, with independent disturbances ε which are $N(0,\sigma^2)$-distributed.

(2) We test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$

(3) Test statistic: $T = \hat{\beta}_2 / se(\hat{\beta}_2)$

(4) Under $H_0 : T \sim t_{10-3} = t_7$

(5) Outcome of T: $\frac{131.898}{40.648} = 3.25$

(6) The (two-sided) p-value is 0.014

(7) Since p-value $\leq \alpha = 5\%$ we have to reject the null hypothesis.

(8) Using level of significance 5% we have proven that the quadratic term has to be added to the model.
We thus have proven that quadratic regression is better than simple linear regression.

If we would choose $\alpha = 1\%$ then p-value$= 0.014 > \alpha$, and hence we don't reject H_0.

Exercise 7

a. For answering the question we have to rearrange the model equation:
$Y = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) \times x_1 + \varepsilon$
If we fix x_2 then Y depends on x_1 in a linear way, $\beta_0 + \beta_2 x_2$ is the (new) constant (intercept) and $\beta_1 + \beta_3 x_2$ is the (new) slope. Note that this slope is affected by x_2, this disappears when we skip the interaction term.

b. We copy the output and replace all signs '?'.

Variables in the equation			
Variables	coefficient	Std. error	T
Intercept	14.9600		
x_1	1.5321	0.5910	2.59
x_2	−0.4323	1.7964	−0.24
$x_1 \times x_2$	−0.0553	0.1554	−0.36

Analysis of Variance				
	df	SS	MS	F
Regression	3	92.110	30.70	67.71
Residual	8	3.627	0.4534	
Total	11	95.737		

c. The estimated equation: $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$

We get from the output: $y = 14.9600 + 1.5321 x_1 - 0.4323 x_2 - 0.0553 x_1 x_2$

We calculate a (point)prediction by using the estimated equation with $x_1 = 12$ and $x_2 = 3$. This renders the following prediction: $\hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 3 - 0.0553 \times 12 \times 3 = 30.06$

d. Prediction for $x_1 = 12$ en $x_2 = 4$: $\hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 4 - 0.0553 \times 12 \times 4 = 28.96$

This is strange because the last prediction is smaller than the first prediction.

We should think about the interpretation of the individual parameters β_i:

The parameter β_i in general reflects the mean change in the dependent variable Y when we increase x_i with 1 unit and fix the other predictor variables.

Fixing the other predictor variables while changing one specific predictor variable does not reflect reality often, because many times predictor variables are changing simultaneously:

Here raising advertisement costs and increasing the number of sales representatives may be simultaneous actions. Then distinguishing the respective causes of the predictor variables may be difficult because of the dependence the predictors.

In statistics it is a famous phenomenon that estimates $\hat{\beta}_i$ turn out to have the wrong sign (you expected a positive value but you get a negative value or vice versa). Many times this phenomenon can be explained by strong relationships between predictor variables, to some extent a number of predictor variables share the same information. Sometimes the 'strangeness' can be solved by simplifying the model.

e. Boundaries for 95% confidence interval for β_2 : $\hat{\beta}_2 \pm c \times se(\hat{\beta}_2)$,

with -0.4323 and 1.7964 for $\hat{\beta}_2$ and $se(\hat{\beta}_2)$ respectively, and $c = 2.31$ (t_8-distribution).

We get: $(-0.4323 - 2.31 \times 1.7964, -0.4323 + 2.31 \times 1.7964) = (-4.582, 3.717)$

f. The eight steps of the test:

1. Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + \varepsilon$, with independent disturbances ε which are $N(\mu, \sigma^2)$-dist.
2. We test $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$
3. Test statistic: $T = \hat{\beta}_3 / se(\hat{\beta}_3)$
4. Under H_0 we have $T \sim t_8$
5. Observed value $t = \frac{-0.0553}{0.1554} \approx -0.36$
6. We reject H_0 if $T \leq -c$ or $T \geq c$. Level of significance 5%, t-table $\Rightarrow c = 2.31$
7. The rejection region does not contain -0.36, hence we don't reject H_0.
8. Using level of significance 5% we did not prove that the interaction term should be part of the model.

g. When you apply again the t-test for testing $H_0: \beta_2 = 0$ against $H_0: \beta_2 \neq 0$ then again you need not reject the null hypothesis. This does not mean that you have to skip both terms $\beta_2 x_2$ and $\beta_3 x_1 \times x_2$. Skip first the interaction term (a model with the interaction term and without the term $\beta_2 x_2$ is weird). Continue with the model without interaction and test $H_0: \beta_2 = 0$ against $H_0: \beta_2 \neq 0$ in order to investigate whether we can skip x_2 completely. For this procedure we thus need more output.

h. Scatter plots of the data is always helpful for appreciating the model applied. Furthermore a model check by means of a scatter plot of residuals is missing.

Exercise 10 (Exercise 8 and 9 cancelled)

a. **1. Model:** $\boldsymbol{Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon}$, $i = 1,2,3$ and $j = 1,2,3$, where:

the α_i`s are the levels of the factor Material and the $\beta_j$`s are the levels of the factor Temperature

$\sum_{i=1}^{3} \alpha_i = 0$, $\sum_{j=1}^{3} \beta_j = 0$, $\sum_{i=1}^{3} \gamma_{ij} = 0$ for each j and $\sum_{j=1}^{3} \gamma_{ij} = 0$ for each i

and for the observations the errors (disturbances) ε are assumed independent and all $N(0, \sigma^2)$.

The results of analysis of variance by SPSS (the 3 columns of the SPSS-file are given in the first table under b.)

Tests of Between-Subjects Effects

Dependent Variable: Life (in hours)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	59416,222[a]	8	7427,028	11,000	,000
Intercept	400900,028	1	400900,028	593,739	,000
Material	10683,722	2	5341,861	7,911	,002
Temperature	39118,722	2	19559,361	28,968	,000
Material * Temperature	9613,778	4	2403,444	3,560	,019
Error	18230,750	27	675,213		
Total	478547,000	36			
Corrected Total	77646,972	35			

a. R Squared = ,765 (Adjusted R Squared = ,696)

1.

2. The p-value of the test on $H_0: \alpha_i = \beta_j = \gamma_{ij} = 0$ for $i, j = 1,2,3$ against H_1: "at least one coefficient $\neq 0$" has a p-value $PF \geq 11.000|H_0) \approx 0.000 < 0.0005 < \alpha = 5\%$. Hence we can conclude (at 5% level of significance) that at least one of the effects of material, temperature or interaction is significant.

3. The p-value of the test on $H_0: \gamma_{ij} = 0$ for $i, j = 1,2,3$ against $H_1: \gamma_{ij} \neq 0$ for at least one pair (i, j), in the model with the factors Material and Temperature, has a p-value $1.9\% < \alpha = 5\%$. Hence we can conclude (at 5% level of significance) that adding the interaction effects to the model is beneficiary in explaining the life of batteries.

4. The p-value of the test on $H_0: \alpha_i = 0$ for $i = 1,2,3$ against H_1: $\alpha_i \neq 0$ for at least one value of i, in the model with the other factors Temperature and Interaction, has a p-value $0.2\% < \alpha = 5\%$. Hence we can conclude (at 5% level of significance) that adding the Material factor to the model increases the explanation of the life of batteries significantly.

5. The p-value of the test on $H_0: \beta_j = 0$ for $j = 1,2,3$ against $H_1: \beta_j \neq 0$ for at least one value of j, in the model with the other factors Material and interaction, has a p-value $1.9\% < \alpha = 5\%$. Hence we can conclude (at 5% level of significance) that adding the temperature factor to the model enhances the explanation of the life of batteries significantly.

b. Below you see first the data matrix (in SPSS), then first the results of the linear regression with all (8) variables and (extra!) then without the (4) interaction terms.

Life	**Material**	**Temp**	x_1	x_2	x_3	x_4	$x_5 = x_1 \times x_3$	$x_6 = x_1 \times x_4$	$x_7 = x_2 \times x_3$	$x_8 = x_2 \times x_4$
130	0	0	0	0	0	0	0	0	0	0
155	0	0	0	0	0	0	0	0	0	0
74	0	0	0	0	0	0	0	0	0	0
180	0	0	0	0	0	0	0	0	0	0
34	0	1	0	0	1	0	0	0	0	0
40	0	1	0	0	1	0	0	0	0	0
80	0	1	0	0	1	0	0	0	0	0
75	0	1	0	0	1	0	0	0	0	0
20	0	2	0	0	0	1	0	0	0	0
70	0	2	0	0	0	1	0	0	0	0
82	0	2	0	0	0	1	0	0	0	0
58	0	2	0	0	0	1	0	0	0	0
150	1	0	1	0	0	0	0	0	0	0
188	1	0	1	0	0	0	0	0	0	0
159	1	0	1	0	0	0	0	0	0	0
126	1	0	1	0	0	0	0	0	0	0
136	1	1	1	0	1	0	1	0	0	0

122	1	1	1	0	1	0	1	0	0	0
106	1	1	1	0	1	0	1	0	0	0
115	1	1	1	0	1	0	1	0	0	0
25	1	2	1	0	0	1	0	1	0	0
70	1	2	1	0	0	1	0	1	0	0
58	1	2	1	0	0	1	0	1	0	0
45	1	2	1	0	0	1	0	1	0	0
138	2	0	0	1	0	0	0	0	0	0
110	2	0	0	1	0	0	0	0	0	0
168	2	0	0	1	0	0	0	0	0	0
160	2	0	0	1	0	0	0	0	0	0
174	2	1	0	1	1	0	0	0	1	0
120	2	1	0	1	1	0	0	0	1	0
150	2	1	0	1	1	0	0	0	1	0
139	2	1	0	1	1	0	0	0	1	0
96	2	2	0	1	0	1	0	0	0	1
104	2	2	0	1	0	1	0	0	0	1
82	2	2	0	1	0	1	0	0	0	1
60	2	2	0	1	0	1	0	0	0	1

Running Linear regression with all 8 variables (including interaction):

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,875[a]	,765	,696	25,985

a. Predictors: (Constant), x2*x4, x2*x3, x1*x4, x1*x3, High temp, material 2, material 3, Medium temp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134,750	12,992		10,371	,000
	material 2	21,000	18,374	,213	1,143	,263
	material 3	9,250	18,374	,094	,503	,619
	Medium temp	-77,500	18,374	-,787	-4,218	,000
	High temp	-77,250	18,374	-,784	-4,204	,000
	x1*x3	41,500	25,985	,281	1,597	,122
	x1*x4	-29,000	25,985	-,196	-1,116	,274
	x2*x3	79,250	25,985	,536	3,050	,005
	x2*x4	18,750	25,985	,127	,722	,477

a. Dependent Variable: Life (in hours)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	59416,222	8	7427,028	11,000	,000[a]
	Residual	18230,750	27	675,213		
	Total	77646,972	35			

a. Predictors: (Constant), x2*x4, x2*x3, x1*x4, x1*x3, High temp, material 2, material 3, Medium temp

b. Dependent Variable: Life (in hours)

Runn

Run

0-model (without interaction)

Results, running linear regression in SPSS with 4 variables (without interaction):

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	49802,444	4	12450,611	13,862	,000[a]
	Residual	27844,528	31	898,211		
	Total	77646,972	35			

a. Predictors: (Constant), High temp, material 3, material 2, Medium temp

b. Dependent Variable: Life (in hours)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,801[a]	,641	,595	29,970

a. Predictors: (Constant), High temp, material 3, material 2, Medium temp

1. The F-test in the ANOVA table (all 8 variables) reports the value 11.000 for $F = \frac{MS(Regression)}{MS(Errors)}$, with a p-value $P(F \geq 11.000|H_0) < 0.0005$, these values are the same as reported in a.1 (two-factor analysis).
2. Both the coefficients for the factor "Material" seem to be not significant (in the presence of the other in the model). We cannot conclude that Material a factor can be cancelled: there can be a strong relation between these variables ("collinearity"), such that one variable is not significant in the presence of the other in the model. This confirmed by the p-value of the test on the factor "Material" in a.3: the factor has significant effects. Possibly one of the variables can be cancelled: e.g. cancelling x_2 (material 3 has the largest p-value) would mean that we only distinguish material 2 and not-material 2. One can take the value of R^2_{adj} after removing x_2 in consideration.
   Both the coefficients of the factor Temperature are significant (p-value = 0.000 < α).
   3 out of 4 interaction coefficients can be potentially be cancelled from the model since the p-value > $\alpha = 5\%$.

Extra:

3. Note that adding the interaction (from 4 to 8 variables) increased the R^2_{adj} (from 59.5% to 69.6%), justifying this increase in number of variables: but canceling x_2 or one or more interaction variables, might increase the value. Check that if we (e.g.) cancel x_8 the value R^2_{adj} increases to 70.1%.
4. Based on (only) the information of the "full" model with interaction, we cannot test on the factor "Interaction", as described in a.3 of analysis of variance. But running Linear regression for the null-model without interaction as well we can (see property 3.2.8 and example 3.3.10):

$$F = \frac{\frac{[SS_0(Errors) - SS(Errors)]}{k-m}}{\frac{SS(Errors)}{n-k-1}} = \frac{(27844.528 - 18230.750)/4}{18230.750/27} \approx 3.56$$

(almost) the same value of F we found in a.3: the RR is $F \geq 2.73$ ($\alpha = 5\%$, F^4_{27}-table), so reject the null hypothesis of "no effect of interaction", as before.