University of Amsterdam

SSO Assignment 2

Assignment 2

October 15, 2021

Students:
Alex Dekker
10761012

Laura Hilhorst
11048999

Youri Moll
10714235

Course:
Statistics, Simulation and
Optimization

1

The *lm()* function and the *summary()* function of R are used to answer subquestions a, b and c.

a

In order to execute the Step up strategy, first all the individual explanatory variables are inspected with respect to *total* as response variable. *Takers* had the highest R^2 value (0.787), and was significant. Afterwards, *expend* was added, resulting in a model with an R^2 of 0.8195. No third variable is used, as both *ratio* and *salary* provided non-significant results. The resulting model is as follows:

```
# Resulting model:
summary(lm(total~takers+expend))
# Total = 993.8317 - 2.8509*takers + 12.2865*expend + error
# with R^2 = 0.8195 and hat(sigma) = 32.46
```

Afterwards, the Step Down method is used. *Expend, salary* and *ratio* are all insignificant. *Expend* is removed first, as it has the highest p-value (0.674). The resulting model contains only significant variables. The resulting model is as follows:

```
# Resulting model:
summary(lm(total~ratio+salary+takers)
# Total = 1057.8982 - 4.6394*ratio + 2.5525*salary - 2.9134*takers + error
# With R^2 = 0.8239 and hat(sigma) = 32.41
```

While the second one has a slightly higher value for R^2, and about the same $\hat{\sigma}$, the first one is preferred, as it only uses two variables to explain the data instead of three.

b

First, the data is expended with *takers*^{2}:

```
    sat$takers2=(sat$takers)^2
```

Then, the same steps are executed as in the previous section. Both strategies resulted in the same model:

```
# Resulting model:
summary(lm(total~expend+takers+takers2))
# Total = 1052 + 7.914*expend - 6.381*takers + 0.04741*takers2 + error
# With R^2 = 0.8859 and hat(sigma) = 26.08
```

c

The second solution is preferred (b), as the R^2 is much higher (0.8859 vs 0.8195), even though one more variable is used. *Takers* is most significant and negatively correlated. *Takers*^{2} is also very significant and has a positive correlation, this is possibly to compensate for *takers*, because the data is quite non-linear.

d

The chosen model is put into R's *fitted()* function, and a dataframe containing the new data row is added with the given parameters. Finally, the *predict()* function is used to predict the *total* score:

```
best_model =lm(total~expend+takers+takers2, data=sat)
fitted(best_model)
newxdata = data.frame(expend=5, takers=25, takers2=625)
predict(best_model,newxdata,interval="confidence",level=0.95)
```

This provided the following prediction:
Total = 961.5703, with 949.0796 as lower bound, and 974.061 as upper.

2

In the following sections, the ANOVA implementation provided by r was used (aov).

a

ANOVA in r assumes the null-hypothesis to be no difference between the two groups.

Using ANOVA through r we have found a P-value of 0.174. This is not enough to reject the null-hypothesis under a 95% confidence interval.

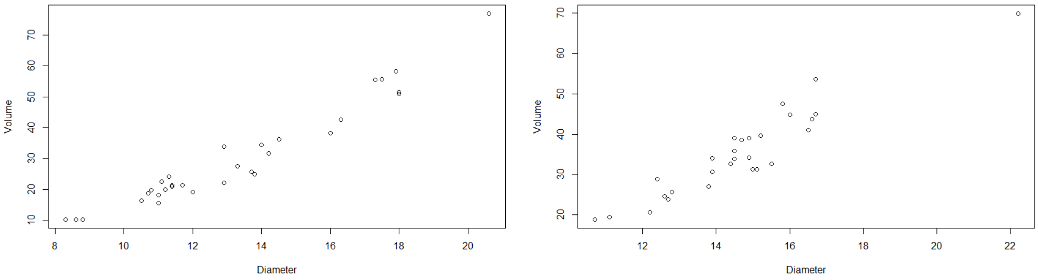
When only the tree type is used to make predictions regarding the tree volume the best a model can do is output the mean. This is 30.171 for beech with the 95% confidence interval being bounded between 25.086 and 35.256. For oak the mean is 35.250 and the 95% confidence interval is between 29.899 and 40.601.

b

Like before, we examine the P-value provided by the ANOVA model to determine of tree type is significant in determining volume by this time we also include the height and width. With these parameters included we find a p value of $P = 1.57 * 10^{-07}$. Which is definitely significant enough to reject H_0 of no difference between the two groups.

When predicting the volume for a tree with average diameter and height which are (13.25, 76) and $(14.64, 75.68)$ for beech and oak respectively, the predicted value is the mean volume or 30.17 for beech and 35.27 for oak.

We expect there to be a linear relationship between the diameter of a tree and its volume i.e. a larger diameter implies a greater volume. This assumption is proven to be correct when the relationship between these two attributes is visualised. This visualisation can be seen in Figure 1.



(a) The volume of beeches plotted against the diameter.

(b) The volume of beeches plotted against the diameter.

Figure 1: The relationship between volume and diameter for both tree types.

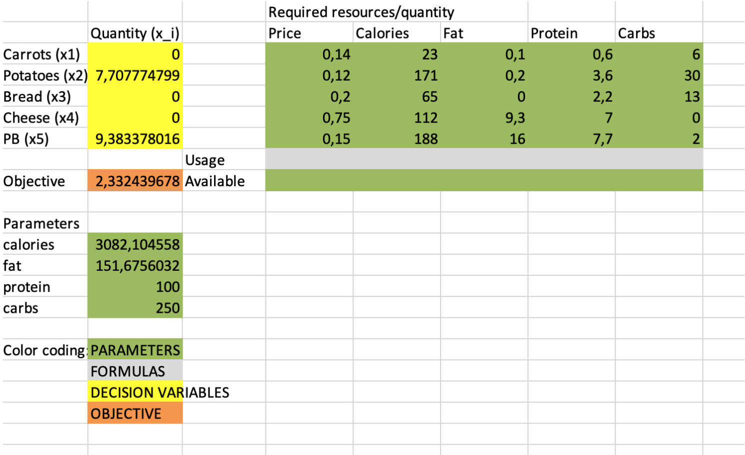
c

An assumption that has been built into the model so far is that the height and diameter of a tree are independent of each other. This is of course incorrect since there is a limit to how high a tree with a certain diameter can grow without falling. Because of this, we propose an improvement to the model that takes this into account. In this new model, we make height and diameter not independent which should result in better predictions.

We tested this new model by splitting the dataset into a random train and test set and trained both the old and new model on the training set. Both of these models were then used to make predictions regarding the test set, which gives us two sets of predictions. These predictions are then tested for their accuracy and the r^2 score is calculated for both models. This score was 0.909 for the old model and 0.933 for the new model. And as can be seen, incorporating this dependency relationship between height and diameter into the model yields slightly better results.

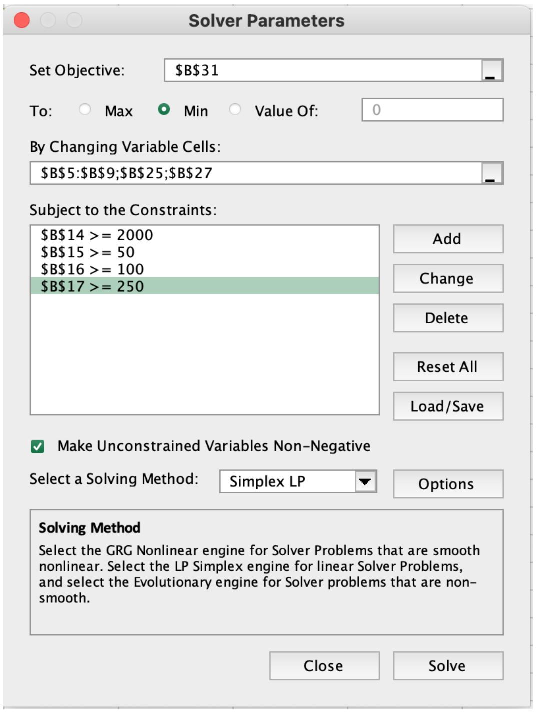
3

a



			Required resources/quantity				
	Quantity (x_i)		Price	Calories	Fat	Protein	Carbs
Carrots (x1)	0		0,14	23	0,1	0,6	6
Potatoes (x2)	7,707774799		0,12	171	0,2	3,6	30
Bread (x3)	0		0,2	65	0	2,2	13
Cheese (x4)	0		0,75	112	9,3	7	0
PB (x5)	9,383378016		0,15	188	16	7,7	2
		Usage					
Objective	2,332439678	Available					
Parameters							
calories	3082,104558						
fat	151,6756032						
protein	100						
carbs	250						
Color coding:	PARAMETERS						
	FORMULAS						
	DECISION VARIABLES						
	OBJECTIVE						

Figure 2: Excel sheet for 3a



Solver Parameters

Set Objective: $B$31

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$B$5:$B$9;$B$25;$B$27

Subject to the Constraints:
$B$14 >= 2000
$B$15 >= 50
$B$16 >= 100
$B$17 >= 250

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 3: Excel solver for 3a

Problem: Optimal product mix to satisfy the diet at the lowest price
Decision variables:
x_1 = carrots
x_2 = potatoes
x_3 = bread
x_4 = cheese

x_5 = peanut butter

Objective: min $0.14x_1 + 0.12x_2 + 0.2x_3 + 0.75x_4 + 0.15x_5$

Constraints:
Calories >= 2000, $23x_1 + 171x_2 + 65x_3 + 112x_4 + 188x_5$
Fat >= 50, $0.1x_2 + 0.2x_2 + 9.3x_4 + 16x_5$
Protein >= 100, $0.6x_1 + 3.7x_2 + 2.2x_3 + 7x_4 + 7.7x_5$
Carbohydrates $> =$ 250, $6x_1 + 30x_2 + 13x_3 + 2x_5$
x_i >= 0, i = 1...5

As can be seen from Figure 2, the optimal product mix consists of 7.7 servings of potatoes and 9.4 servings of peanut butter at at cost of 2.33. The solver parameters are stated in Figure 3.

b

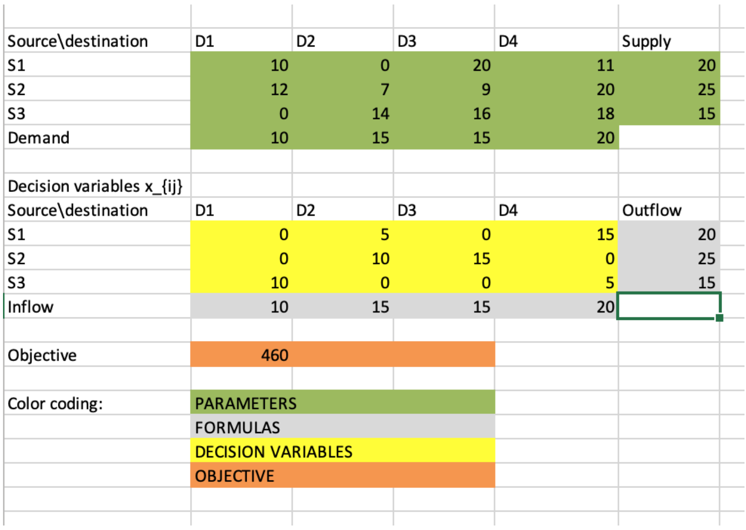
We added decision variable x_6 consisting of peanut butter at price 0.25, and all values the same as the 'normal' peanut butter. We added constraint $x_5 <= 5$. The full solver parameters can be found in figures 8 and 9 in the appendix.
Given this new situation, the solver determined that the optimal product mix is a diet consisting of 17.08 portions of potatoes and 5 portions of peanut butter at price 2.8.

c

We added integer constraints to all decision variables. The full solver parameters can be found in figures 10 and 11 in the appendix. The cheapest diet according to the solver was 20 portions of potatoes and 4 portions of peanut butter at price 3. This is more expensive than the diet in 3a. It overshoots the other requirements by quite a bit, but it is the cheapest option to satisfy the protein requirement without using fractions.

4

a



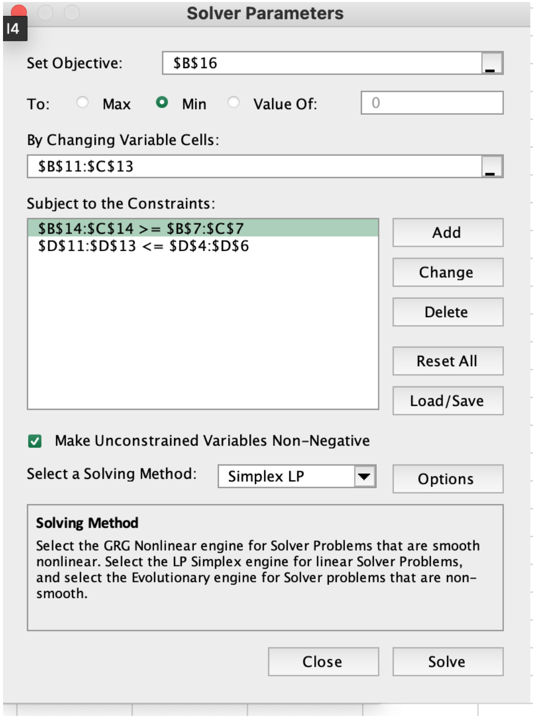
Source\destination	D1	D2	D3	D4	Supply
S1	10	0	20	11	20
S2	12	7	9	20	25
S3	0	14	16	18	15
Demand	10	15	15	20	

Decision variables x_{ij}

Source\destination	D1	D2	D3	D4	Outflow
S1	0	5	0	15	20
S2	0	10	15	0	25
S3	10	0	0	5	15
Inflow	10	15	15	20	

Objective	460
Color coding:	PARAMETERS
	FORMULAS
	DECISION VARIABLES
	OBJECTIVE

Figure 4: Excel sheet for 4a



Solver Parameters

Set Objective: $B$16

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $B$11:$C$13

Subject to the Constraints:

$B$14:$C$14 >= $B$7:$C$7
$D$11:$D$13 <= $D$4:$D$6

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

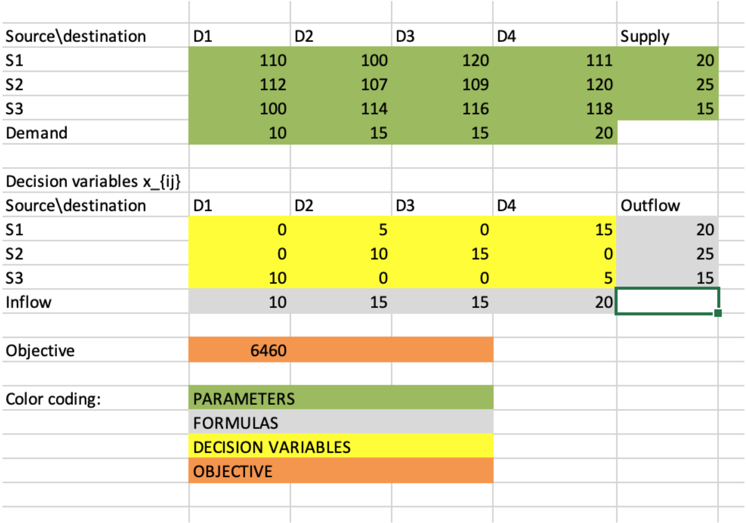
Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 5: Excel solver for 4a

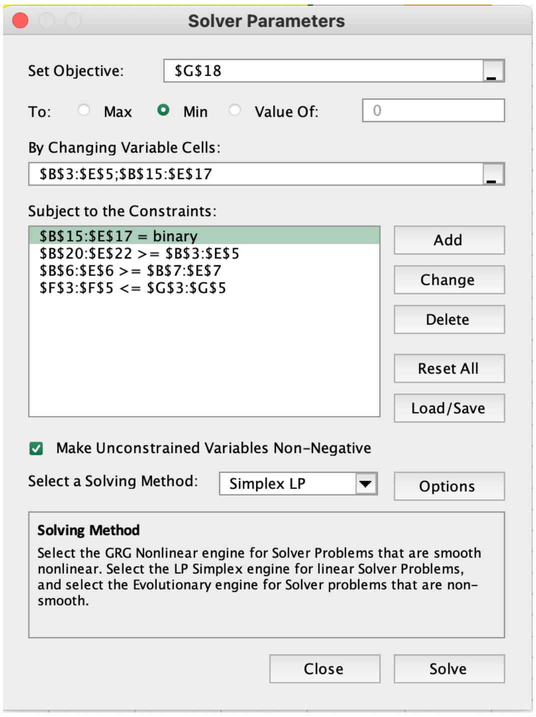
Problem: Find the lowest cost to fulfil demand while respecting supply.
Decision variables: x_{ij}, quantity transported from i to j, $\forall$ i,j
Objective: min $\sum_{i,j} c_{ij}x_{ij}$ (c = cost, x = number of units)
Constraints:
Maximum supply: $\sum_j x_{ij} <= a_i$, i = 1...n
Maximum demand: $\sum_i x_{ij} >= b_j$, i = 1...m
$x_{ij} >= 0$, $\forall$ i,j
The cheapest transport solution, taking into account supply and demand, is depicted in Figure 4 at a price of 460. The solver parameters are stated in Figure 5.

b



Source\destination	D1	D2	D3	D4	Supply
S1	110	100	120	111	20
S2	112	107	109	120	25
S3	100	114	116	118	15
Demand	10	15	15	20	
Decision variables x_{ij}					
Source\destination	D1	D2	D3	D4	Outflow
S1	0	5	0	15	20
S2	0	10	15	0	25
S3	10	0	0	5	15
Inflow	10	15	15	20	
Objective	6460				
Color coding:	PARAMETERS				
	FORMULAS				
	DECISION VARIABLES				
	OBJECTIVE				

Figure 6: Excel sheet for 4b



Solver Parameters

Set Objective: $G$18

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$B$3:$E$5;$B$15:$E$17

Subject to the Constraints:
$B$15:$E$17 = binary
$B$20:$E$22 >= $B$3:$E$5
$B$6:$E$6 >= $B$7:$E$7
$F$3:$F$5 <= $G$3:$G$5

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 7: Excel sheet for 4b

We added additional decision variables, namely a binary variable for each edge x_{ij}, whose value is 0 if the edge is unused and 1 if the edge is used. For every one of these variables that has value 1, 100 is added to the total price.

In this case, we see that adding the additional cost to each used edge caused the solver to use fewer edges: 5 instead of 6 in 4a. However, it does come at a vastly higher price of

995, including the 500 extra cost for used edges. Details of the transport solution can be found in figure xx.

5

a

This call center staffing problem can be described as a covering problem. We have Universe U with the half-hours u; 24 entries of half-hour duration time slots, with 25 shifts i; 8 8.5 hour shifts (with one empty slot after 4 hours), and 17 4 hour shifts. The objective is to minimize the number of costs. The costs are 160 euros for 8.5 hour shifts (8 times 20) and 96 euros for 4 hour shifts (4 times 24). The constraints are provided by the given table. The required staffing b_u needs to be at least the given number of workers at that time.

Problem: Find the lowest cost to fulfil the given scheduling problem.
Decision variables: x_i, number of shifts i taken.
Objective: min $\sum_i c_i x_i$
Constraints: $b_u = [10, 11, 13...]^T$ (values are second column of Table 1 as described in the assignment).

The solver used is found in Figure 12.
This resulted in a total cost of 3296 euros, with the shift coverage shown in Figure 13.

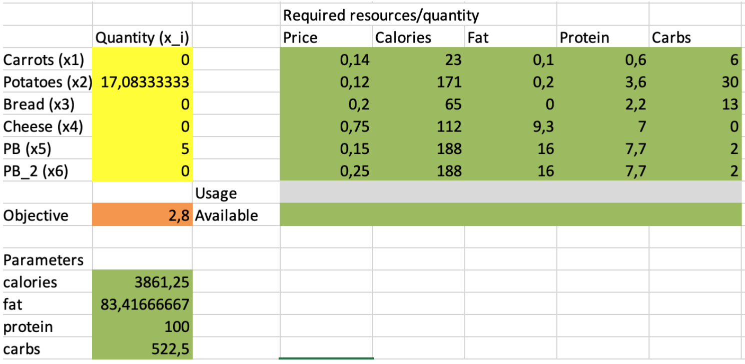
b

With only 8 shifts i, and the required staffing no longer being a constraint, the problem significantly changes. Because the sum of absolute values needs to be minimized, the problem is no longer linear. In order to make it linear, two new rows are added to the Excel sheet: $e_{pos(u)}$ and $e_{min(u)}$. Both take the difference between the scheduled number of workers and the demanded number of workers, but the first only takes the positive differences (scheduled > demanded) and the latter takes the negative errors (scheduled < demanded). Now two new constraints are added: the left hand side constraint, which is the difference between demanded and scheduled, and the right hand side constraint, which is the difference between $e_{pos(u)}$ and $e_{min(u)}$.

Problem: Find the schedule that is as close as possible to the demanded schedule.
Decision variables: x_i, number of shifts i taken. **AND** $e_{pos(u)}$ and $e_{min(u)}$
Objective: min $\sum_u e_{pos(u)} + e_{min(u)}$
Constraints: $b_u - S_u = e_{pos(u)} - e_{min(u)}$

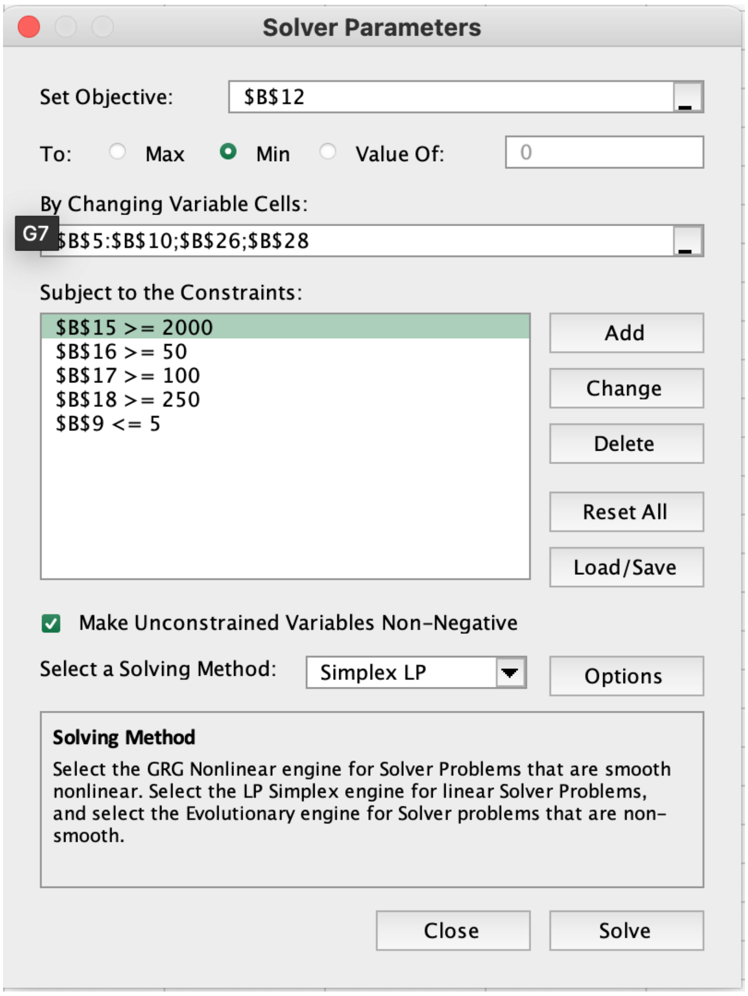
The solver used is found in Figure 14.
This resulted in a minimized the sum of absolute differences between the demanded and scheduled number of workers per time interval of **60**, with the shift coverage shown in Figure 15.

Appendix:



			Required resources/quantity				
	Quantity (x_i)		Price	Calories	Fat	Protein	Carbs
Carrots (x1)	0		0,14	23	0,1	0,6	6
Potatoes (x2)	17,08333333		0,12	171	0,2	3,6	30
Bread (x3)	0		0,2	65	0	2,2	13
Cheese (x4)	0		0,75	112	9,3	7	0
PB (x5)	5		0,15	188	16	7,7	2
PB_2 (x6)	0		0,25	188	16	7,7	2
		Usage					
Objective	2,8	Available					
Parameters							
calories	3861,25						
fat	83,41666667						
protein	100						
carbs	522,5						

Figure 8: Excel sheet for 3b



Solver Parameters

Set Objective: $B$12

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:

G7 $B$5:$B$10;$B$26;$B$28

Subject to the Constraints:

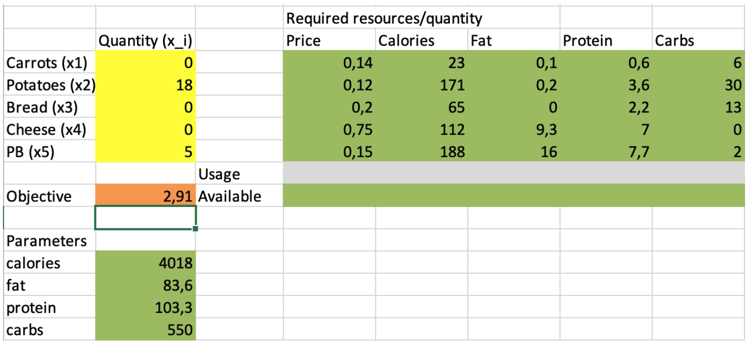
$B$15 >= 2000
$B$16 >= 50
$B$17 >= 100
$B$18 >= 250
$B$9 <= 5

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

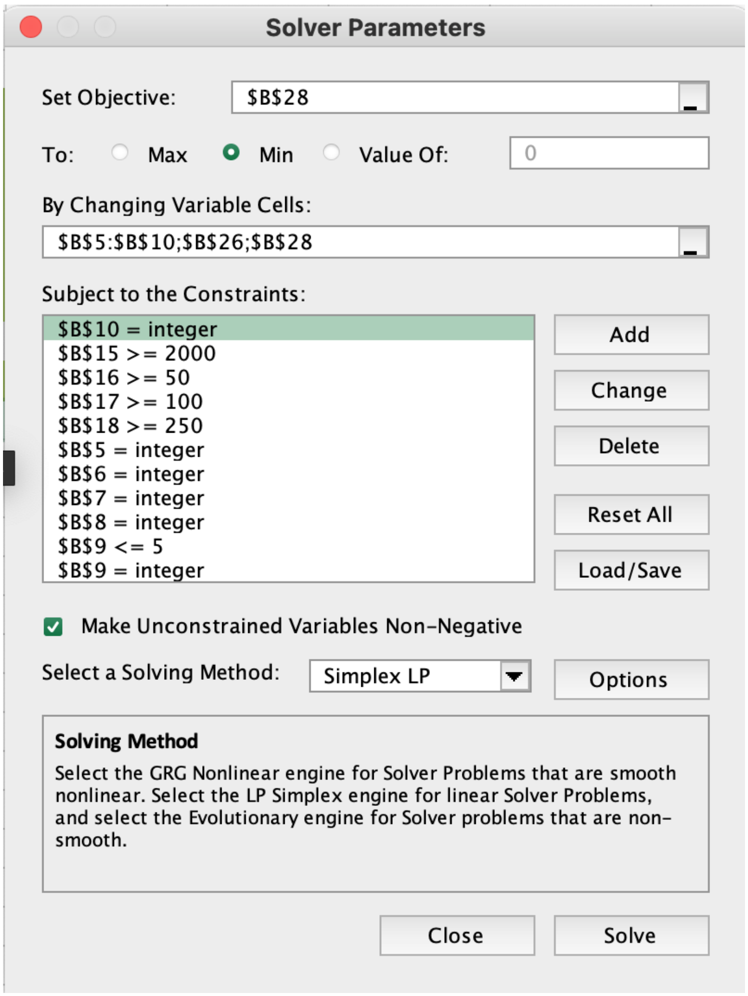
Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 9: Excel solver for 3b



			Required resources/quantity				
	Quantity (x_i)		Price	Calories	Fat	Protein	Carbs
Carrots (x1)	0		0,14	23	0,1	0,6	6
Potatoes (x2)	18		0,12	171	0,2	3,6	30
Bread (x3)	0		0,2	65	0	2,2	13
Cheese (x4)	0		0,75	112	9,3	7	0
PB (x5)	5		0,15	188	16	7,7	2
		Usage					
Objective	2,91	Available					
Parameters							
calories	4018						
fat	83,6						
protein	103,3						
carbs	550						

Figure 10: Excel sheet for 3c



Solver Parameters

Set Objective: $B$28

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$B$5:$B$10;$B$26;$B$28

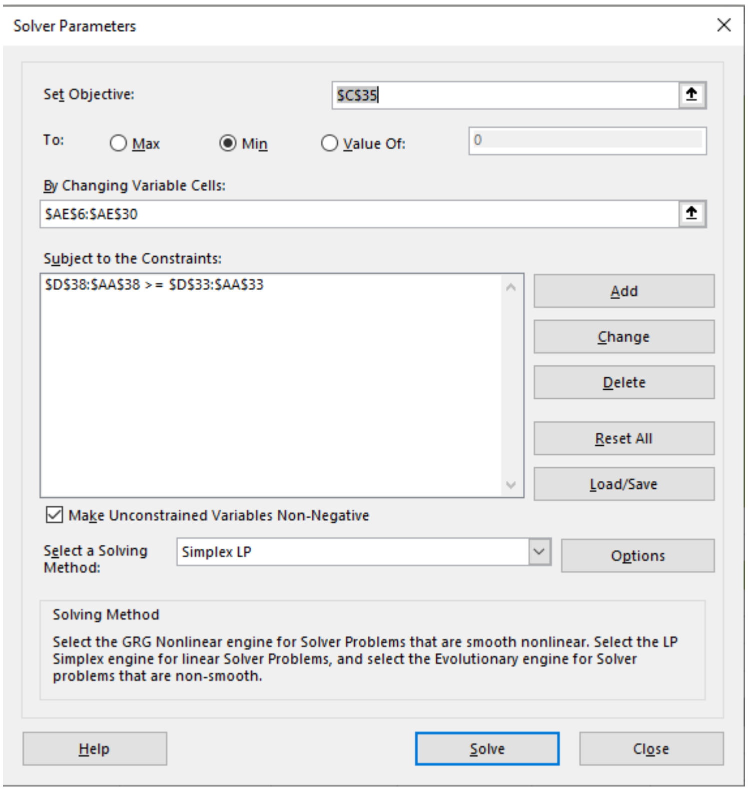
Subject to the Constraints:
$B$10 = integer
$B$15 >= 2000
$B$16 >= 50
$B$17 >= 100
$B$18 >= 250
$B$5 = integer
$B$6 = integer
$B$7 = integer
$B$8 = integer
$B$9 <= 5
$B$9 = integer

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 11: Solver for 3c



Solver Parameters

Set Objective: $C$35

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$AE$6:$AE$30

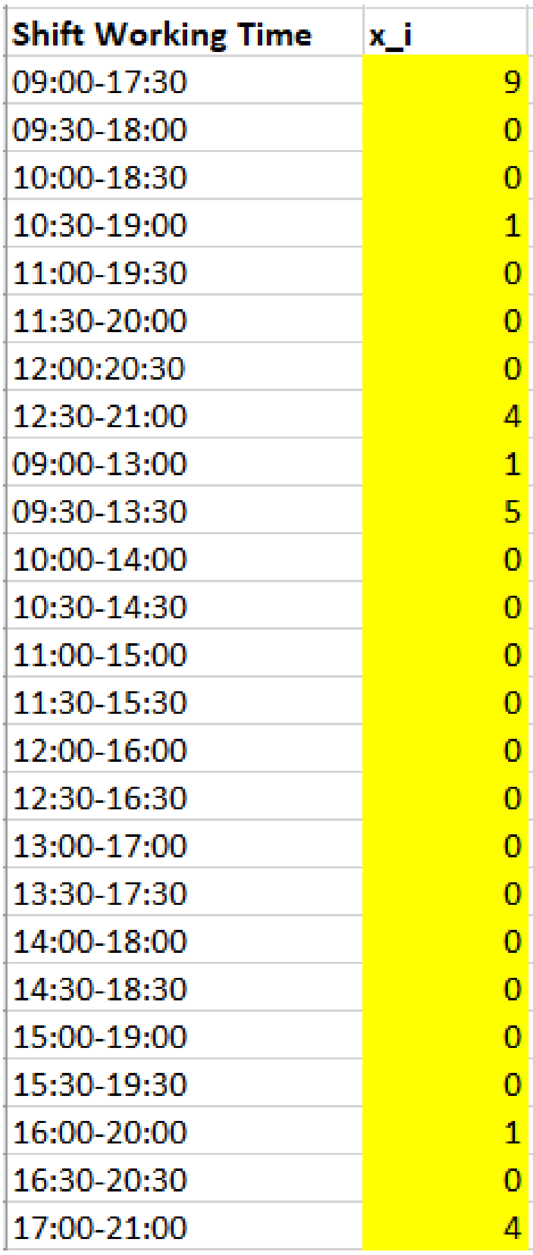
Subject to the Constraints:
$D$38:$AA$38 >= $D$33:$AA$33

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

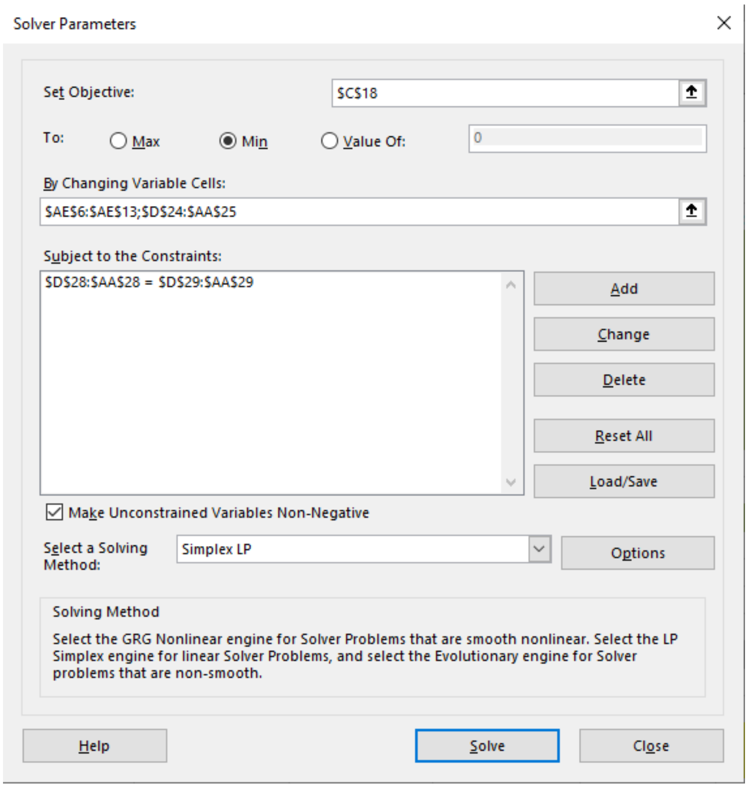
Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 12: Solver for 5a



Shift Working Time	x_i
09:00-17:30	9
09:30-18:00	0
10:00-18:30	0
10:30-19:00	1
11:00-19:30	0
11:30-20:00	0
12:00:20:30	0
12:30-21:00	4
09:00-13:00	1
09:30-13:30	5
10:00-14:00	0
10:30-14:30	0
11:00-15:00	0
11:30-15:30	0
12:00-16:00	0
12:30-16:30	0
13:00-17:00	0
13:30-17:30	0
14:00-18:00	0
14:30-18:30	0
15:00-19:00	0
15:30-19:30	0
16:00-20:00	1
16:30-20:30	0
17:00-21:00	4

Figure 13: End result 5a



Solver Parameters

Set Objective: $C$18

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$AE$6:$AE$13;$D$24:$AA$25

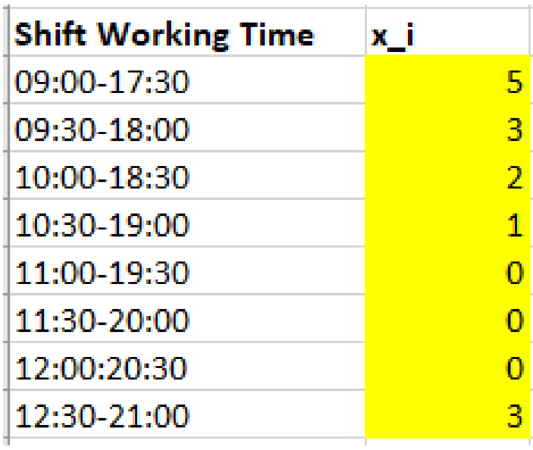
Subject to the Constraints:
$D$28:$AA$28 = $D$29:$AA$29

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 14: Solver for 5b



Shift Working Time	x_i
09:00-17:30	5
09:30-18:00	3
10:00-18:30	2
10:30-19:00	1
11:00-19:30	0
11:30-20:00	0
12:00:20:30	0
12:30-21:00	3

Figure 15: End result 5b