Assignment 2 Group 26

Vanshita Sharma Kumar Lucian Truşcă Xander Poortvliet

2025-10-10

Excercise 2.1

2.1a

We first will start with the full multi-regression model

```
model_full = lm(total ~ expend + ratio + salary + takers, data=sat)
```

```
## The AIC score for the full model = 497.3694
```

Step Up method: With the forward selection we will first start with no predictors and add variables one by one based on the lowest AIC

```
model_StepUp = lm(total ~ expend + takers, data=sat)
```

```
## The AIC score for the step-up method  = 494.7994
```

Step-down Method: we start from the full model and iteratively remove variables that worsen AIC the least.

```
model_StepDown <- lm(total ~ expend + takers, data=sat)
```

```
## The AIC score for the step-down method  = 494.7994
```

Model interpretation: SAT performance is best explained by school spending and participation rate. Other variables (ratio, salary) don't significantly improve model fit.

2.1b

```
sat$takers2 = sat$takers^2
## 2) Stepwise model selection (AIC)
# forward (start from intercept)
m0    = lm(total ~ 1, data = sat)
scope = ~ expend + ratio + salary + takers + takers2
m_fwd = step(m0, scope = list(lower = ~1, upper = scope),
              direction = "forward", trace = 0)
```

Where the result for the AIC is 473.9 (rounded up from 473.85).

```
## [1] 473.8576

## The AICS without takers2 is:  494.7994

## Analysis of Variance Table
##
## Model 1: total ~ expend + takers
## Model 2: total ~ takers + takers2 + expend
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     47 49520
## 2     46 31298  1     18222 26.783 4.872e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In a nested-model ANOVA comparing M_1 : total $\sim$ expend+takers to M_2 : total $\sim$ expend+takers+takers^2, adding the quadratic term reduces the residual sum of squares from 49,520 to 31,298, a drop of 18,222 with one additional parameter (df $= 1$), yielding $F(1, 46) = 26.783$ and $p = 4.872 \times 10^{-6}$. This highly significant improvement leads us to conclude that **takers**^{2} is a useful predictor: it captures curvature in the relationship between SAT scores and participation that the linear-only specification misses.

2.1c

Comparing the reduced model M_1 to the expanded model M_2, the ANOVA shows a large and statistically significant drop in residual sum of squares as seen previously, where this drop implies the rejection of $H_0 : \beta_{\text{takers}^2} = 0$ and confirming that the quadratic term is informative; this statistical improvement is mirrored by information criteria, with AIC falling from ≈ 492.8 for M_1 to ≈ 471.9 for M_2, indicating that the model including takers^2 provides a substantially better fit despite its extra parameter.

2.1d

```r
model_best = lm(total ~ takers + takers2 + expend, data = sat)

state = data.frame(
  expend=5,
  ratio=20,
  salary=36.000,
  takers=25,
  takers2=25^2
)

predict(model_best, newdata = state, interval = "prediction", level = 0.95)
```

```
##        fit      lwr     upr
## 1 961.5703 907.6003 1015.54
```

```r
predict(model_best, newdata = state, interval = "confidence", level = 0.95)
```

```
##        fit      lwr     upr
## 1 961.5703 949.0796 974.061
```

The 95% prediction and confidence intervals are as follows:

- Prediction: [907.6003, 1015.54]
- Confidence: [949.0796, 974.061]

Excercise 2.2

2.2a

```r
model_a = lm(volume ~ type, data = data)
predict(model_a, newdata = data.frame(type = c("beech", "oak")))
```

```
##        1        2
## 30.17097 35.25000
```

```r
summary(model_a)
```

```
##
## Call:
## lm(formula = volume ~ type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.971  -9.960  -2.771   5.940  46.829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   30.171      2.539  11.881   <2e-16 ***
## typeoak        5.079      3.686   1.378    0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 57 degrees of freedom
## Multiple R-squared:  0.03223,    Adjusted R-squared:  0.01525
## F-statistic: 1.898 on 1 and 57 DF,  p-value: 0.1736
```

```r
anova(model_a)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df  Sum Sq Mean Sq F value Pr(>F)
## type       1   379.5  379.52  1.8984 0.1736
## Residuals 57 11394.8  199.91
```

An ANOVA comparing mean wood volume between "beech" and "oak" trees shows no statistically significant difference ($F = 1.90$, $p = 0.17$). This indicates that, based only on tree type, there is no evidence that oak trees are more voluminous than beech tres. The estimated mean volumes were approximately {beech=30.17, oak = 35.25}. While it appears oaks may be slightly larger on average, this difference is not significant at the 5% level.

2.2b

```r
# fitting an ANCOVA model, now including diameter and height
model_b <- lm(volume ~ diameter + height + type, data = data)

str(data)
```

```
## 'data.frame':    59 obs. of  4 variables:
##  $ diameter: num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ height  : int  70 65 63 72 81 83 66 75 80 75 ...
##  $ volume  : num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
##  $ type    : Factor w/ 2 levels "beech","oak": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
summary(data)
```

```
##     diameter         height          volume         type   
##  Min.   : 8.30   Min.   :63.00   Min.   :10.20   beech:31  
##  1st Qu.:11.55   1st Qu.:71.00   1st Qu.:21.35   oak  :28  
##  Median :13.90   Median :76.00   Median :31.30             
##  Mean   :13.91   Mean   :75.85   Mean   :32.58             
##  3rd Qu.:15.65   3rd Qu.:80.50   3rd Qu.:39.30             
##  Max.   :22.20   Max.   :87.00   Max.   :77.00             
```

```r
summary(model_b)
```

```
## 
## Call:
## lm(formula = volume ~ diameter + height + type, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.1859 -2.1396 -0.0871  1.7208  7.7010 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -63.78138    5.51293 -11.569 2.33e-16 ***
## diameter      4.69806    0.16450  28.559  < 2e-16 ***
## height        0.41725    0.07515   5.552 8.42e-07 ***
## typeoak      -1.30460    0.87791  -1.486    0.143    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.243 on 55 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9482 
## F-statistic: 354.9 on 3 and 55 DF,  p-value: < 2.2e-16
```

```r
anova(model_b)
```

```
## Analysis of Variance Table
## 
```

```
## Response: volume
##           Df  Sum Sq Mean Sq   F value    Pr(>F)
## diameter   1 10826.5 10826.5 1029.5139 < 2.2e-16 ***
## height     1   346.2   346.2   32.9192 4.254e-07 ***
## type       1    23.2    23.2    2.2083     0.143
## Residuals 55   578.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#mean values from data (for estimating volume later)
mean_diam = mean(data$diameter)
mean_height = mean(data$height)
```

An ANCOVA including diameter and height as covariates showed that both diameter (p « 0.001) and height (p « 0.001) significantly affect tree volume. However, tree type was not significant (p = 0.143), indicating that, once diameter and height are accounted for, Beech and Oak trees have similar expected volumes.

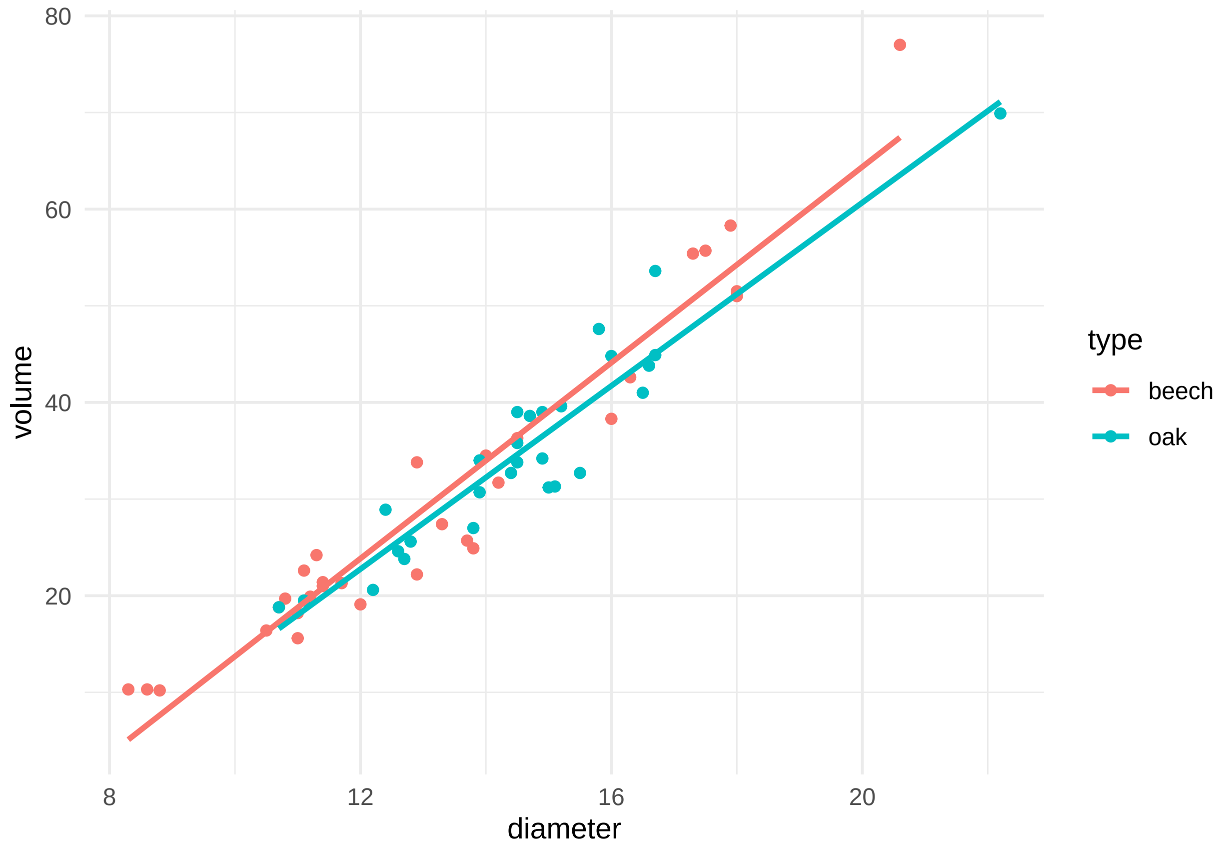
Now we estimate/predict volumes for these avg values

```r
newdata = data.frame(
  type = c("beech", "oak"),
  diameter = mean_diam,
  height = mean_height
)

predicted_vol = predict(model_b, newdata = newdata, interval = "confidence")
predicted_vol
```

```
##        fit      lwr      upr
## 1 33.20049 32.01179 34.38918
## 2 31.89589 30.64274 33.14904
```

Using our fitted ANCOVA model, predicted mean volumes for "beech" and "oak" at average diameter and height are nearly identical, with overlapping confidence intervals. This confirms that tree type does not have a statistically significant influence on volume when tree size is controlled for.

```r
library(ggplot2)

ggplot(data, aes(x = diameter, y = volume, color = type)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  theme_minimal()
```



The scatterplot shows that volume increases linearly with diameter for both tree types, and the regression lines are nearly parallel.

```
model_interaction <- lm(volume ~ type * diameter + height, data = data)
anova(model_b, model_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: volume ~ diameter + height + type
## Model 2: volume ~ type * diameter + height
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     55 578.39
## 2     54 572.87  1    5.5165 0.52  0.474
```

When we added an interaction between tree type and diameter, the model didn't improve (F = 0.52, p = 0.474). This means the way volume changes with diameter is about the same for both beech and oak, so the ANCOVA assumption of parallel slopes is reasonable.

2.2c

The volume of a cylinder is given by $V = \pi r^2 H$, where r is the radius of the circular base, H is the height, and π is the mathematical constant (approximately 3.14159).

```
trees$type = factor(trees$type)
trees$calc_vol = pi*((trees$diameter/2)^2)*trees$height
```

```r
model_c = lm(volume ~ calc_vol + type, data = trees)
summary(model_c)
```

```
##
## Call:
## lm(formula = volume ~ calc_vol + type, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6321 -1.4601 -0.3746  1.5045  5.3354
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.056e-01  7.843e-01  -0.645    0.522
## calc_vol     2.723e-03  5.926e-05  45.958   <2e-16 ***
## typeoak      4.529e-01  6.061e-01   0.747    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.292 on 56 degrees of freedom
## Multiple R-squared:  0.975,  Adjusted R-squared:  0.9741
## F-statistic:  1092 on 2 and 56 DF,  p-value: < 2.2e-16
```

```r
anova(model_c)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df  Sum Sq Mean Sq   F value Pr(>F)
## calc_vol   1 11477.1 11477.1 2183.8014 <2e-16 ***
## type       1     2.9     2.9    0.5583 0.4581
## Residuals 56   294.3     5.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As $p <= 0.05$, this yields a better result.

Excercise 2.3

To solve the problem of the optimal product mix with excel. We choose the number of servings of each food to minimize total cost while meeting minimum nutrient requirements. It is important to note that for all questions (including 2.4 and 2.5), the options menu has the same configuration as in the image below.

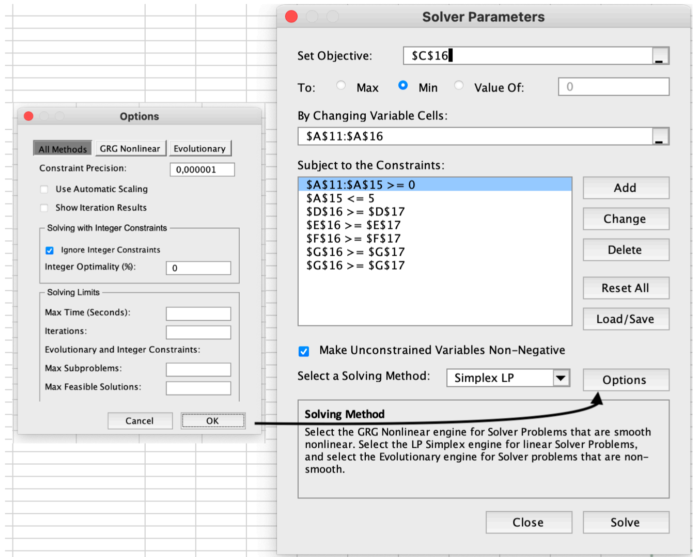


Figure 1: Options Menu

Notation

- Foods $F = \{\text{carrots}, \text{potatoes}, \text{bread}, \text{cheddar}, \text{pb}\}$.
- Parameters per serving $f \in F$:
  - price c_f,
  - calories a_f^{cal},
  - fat a_f^{fat},
  - protein a_f^{prot},
  - carbs a_f^{carb}.
- Minimum requirements: $(b_{\text{cal}}, b_{\text{fat}}, b_{\text{prot}}, b_{\text{carb}}) = (2000, 50, 100, 250)$.
- Decision variables: $x_f \geq 0 =$ servings of food f.

2.3a

The solution uses continuous servings, so constraints can be met exactly. We can visualise our excel solver.

$$\begin{aligned}
\min_{x \geq 0} \quad & \sum_{f \in F} c_f \, x_f \\
\text{s.t.} \quad & \sum_{f \in F} a_f^{\text{cal}} x_f \;\geq\; b_{\text{cal}}, \\
& \sum_{f \in F} a_f^{\text{fat}} x_f \;\geq\; b_{\text{fat}}, \\
& \sum_{f \in F} a_f^{\text{prot}} x_f \;\geq\; b_{\text{prot}}, \\
& \sum_{f \in F} a_f^{\text{carb}} x_f \;\geq\; b_{\text{carb}}.
\end{aligned}$$

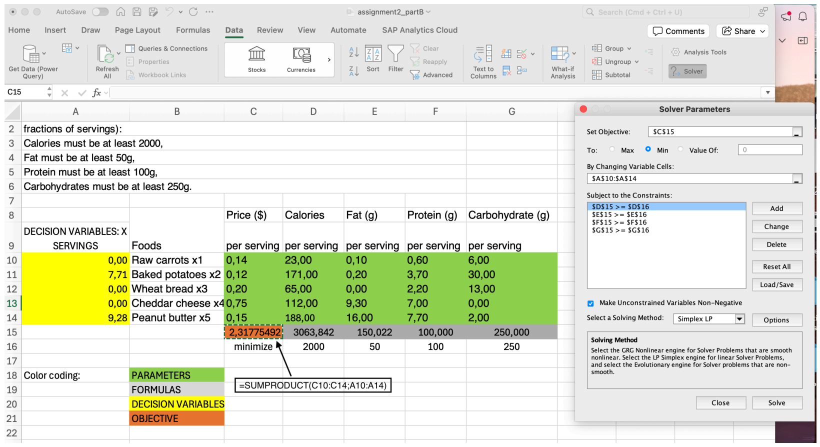


Figure 2: 2.3a Excel Solver

According to the optimal solution for the linear optimization problem, the cheapest feasible diet is a two-food combination of 7.71 servings of baked potatoes and 9.28 servings of peanut butter, costing approximately \$2.32 per day. In this optimal solution, the protein and carbohydrate requirements are exactly met, with an excess (slack) in calories and fat, meaning these nutrients are well above their minimum thresholds.

2.3b

Let peanut butter be split into two variables: $x_{\text{pb}}^{(1)}$ = the **first** (cheap) PB servings, and $x_{\text{pb}}^{(2)}$ = any **additional** PB servings. Prices: $c_{\text{pb}}^{(1)} = 0.15$, $c_{\text{pb}}^{(2)} = 0.25$. Cap: $0 \leq x_{\text{pb}}^{(1)} \leq 5$.

All nutrients per serving are identical for both PB tiers.

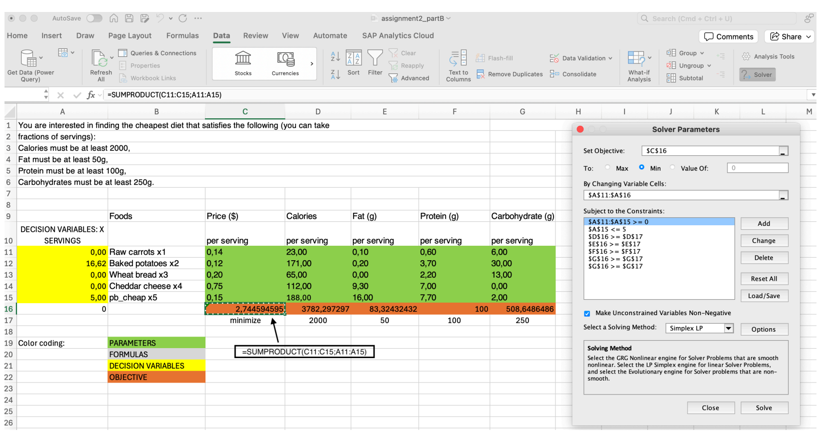


Figure 3: 2.3b Excel Solver

Interpretation. This stays linear by replacing PB with two variables: buy up to 5 cheap units, then any extra at the higher price. In the optimum, the model purchases exactly the 5 cheap PB units and substitutes the rest with the next-best cheap source (potatoes), increasing total cost to 16.62 vs. (a).

2.3c

Same as (a), but restrict servings to integers:

$$x_f \in \mathbb{Z}_{\geq 0} \quad \text{for all } f \in F.$$

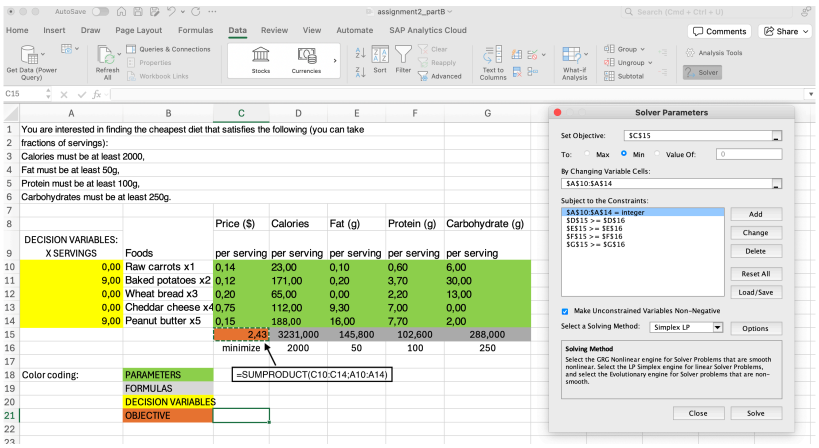


Figure 4: 2.3c Excel Solver

Since the solution from part a was already optimal using fractional servings, limiting the model to whole numbers makes it less efficient. In this case, the cheapest diet is still a mix of baked potatoes and peanut butter, but because the solver can't use fractional amounts, it has to round up to full servings. This makes the diet a bit more expensive, with the total cost rising slightly from about $2.31 to $2.43 per day.

Excercise 2.4

2.4a

This model minimizes total shipping cost from three sources (S1–S3) to four destinations (D1–D4). Each source has a *supply limit* and each destination has a *demand requirement*. The Solver chooses shipments x_{ij} (from source i to destination j) so that **total cost is minimal** while all supply and demand constraints are met.

Model.

$$\begin{aligned}
\min_{x} \quad & \sum_{i \in S} \sum_{j \in D} c_{ij}\, x_{ij} \\
\text{s.t.} \quad & \sum_{j \in D} x_{ij} \;\leq\; a_i \qquad \forall i \in S \quad \text{(supply)} \\
& \sum_{i \in S} x_{ij} \;\geq\; b_j \qquad \forall j \in D \quad \text{(demand)} \\
& x_{ij} \;\geq\; 0 \qquad\qquad \forall (i,j) \in S \times D\,.
\end{aligned}$$

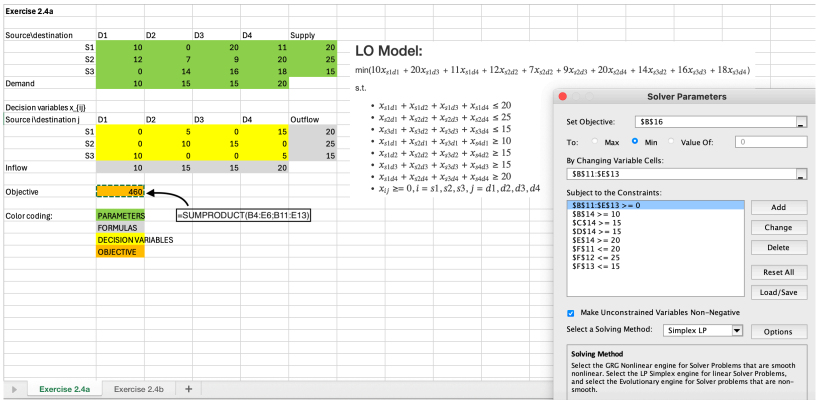


Figure 5: 2.4a Excel Solver

The optimal flows is the following, S1 ships 5 to D2 and 15 to D4; S2 ships 10 to D3 and 15 to D2; S3 ships 10 to D1 and 5 to D4. With a minimum total transportation cost of **$460**. *Interpretation.* The solution uses the cheapest lanes as much as possible (e.g., S2→D2, S3→D1) and avoids expensive ones (e.g., S1→D3). Total cost **$460** is the most economical plan that exactly meets demand without exceeding supply.

2.4b

Question 2.4b extends the previous question by adding a **fixed cost of 100** each time a route (i, j) is used. Binary variables $y_{ij} \in \{0, 1\}$ indicate whether a route is opened. The objective now minimizes **transport cost + fixed activation cost**.

The LO model is as follows:

$min(10x_{s1d1} + 100y_{s1d1} + 20x_{s1d3} + 100y_{s1d3} + 11x_{s1d4} + +100y_{s1d4} + 12x_{s2d1} + 100y_{s2d1} + 7x_{s2d2} + 100y_{s2d2} + 9x_{s2d3} + 100y_{s2d3} + 20x_{s2d4} + 100y_{s2d4} + 14x_{s3d2} + 100y_{s3d2} + 16x_{s3d3} + 100y_{s3d3} + 18x_{s3d4} + 100y_{s3d4})$

s.t.

- $x_{s1d1} + x_{s1d2} + x_{s1d3} + x_{s1d4} \leq 20$
- $x_{s2d1} + x_{s2d2} + x_{s2d3} + x_{s2d4} \leq 25$
- $x_{s3d1} + x_{s3d2} + x_{s3d3} + x_{s3d4} \leq 15$
- $x_{s1d1} + x_{s2d1} + x_{s3d1} + x_{s4d1} \geq 10$
- $x_{s1d2} + x_{s2d2} + x_{s3d2} + x_{s4d2} \geq 15$
- $x_{s1d3} + x_{s2d3} + x_{s3d3} + x_{s4d3} \geq 15$
- $x_{s1d4} + x_{s2d4} + x_{s3d4} + x_{s4d4} \geq 20$
- $x_{ij} \geq= 0, i = s1, s2, s3, j = d1, d2, d3, d4$
- $y_{ij} = \{1, 0\}$

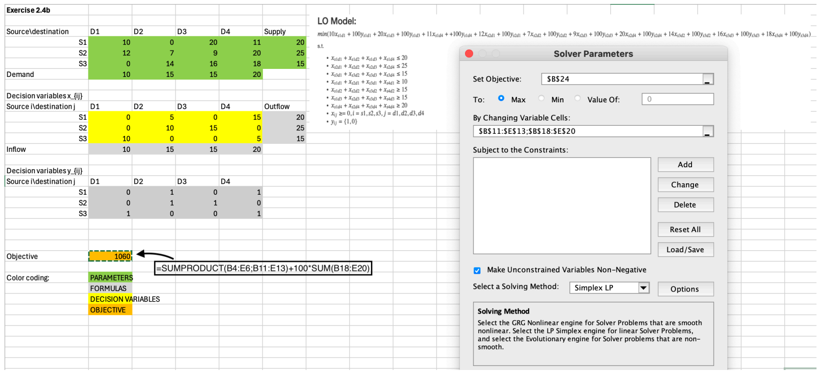


Figure 6: 2.4b Excel Solver

The optimal plan opens only cost-effective routes (those with $y_{ij} = 1$) and sends the required flows on them. The minimum total cost (including fixed charges) will be **$1060**. *Interpretation.* With activation costs, the model prefers **fewer routes** carrying larger volumes to avoid paying many fixed fees. This raises total cost from **$460** to **$1060**, but yields a more consolidated network.

Excercise 2.5

2.5a

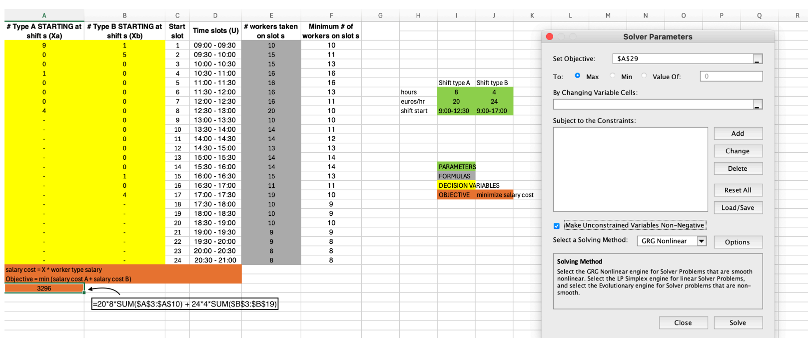


Figure 7: 2.5a Shift scheduling

The schedule covers every time slot, with several slots having the exact number of workers required. The abrupt rise in workers from 11 to 19 at 17:00 is due to the overlap of new workers finishing up the rest of 4 hours of the day (part time 4-hour shift workers). The dip in workers from 19 workers to 10 after 17:30 is caused by the early 8-hour shifts finishing while the new 4-hour shifts are starting.

Most of the coverage comes from type A (8-hour) shifts, with a few shorter type B shifts added to handle busier times. The optimal (minimal) total daily cost comes to about €3,296, after meeting all constraints and requirements.

2.5b

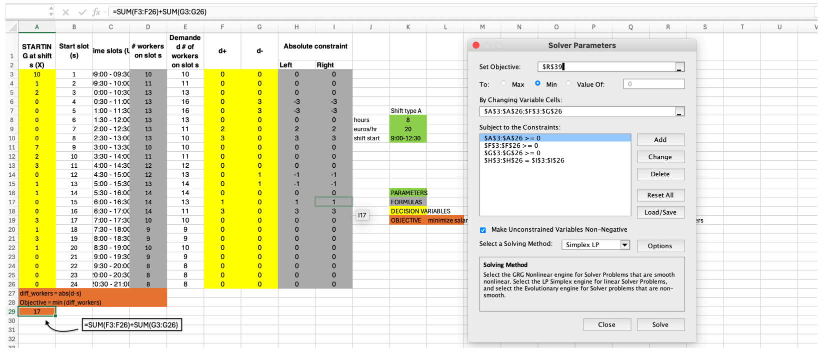


Figure 8: 2.5b Shift scheduling

The schedule is optimized to minimize the absolute sum of differences between scheduled hours en demanded hours. This is done by adding $d+$ and $d-$ to make the non-linear relation linear again. In the end, the biggest difference occurs when there are either 3 more or 3 less people scheduled than demanded. The sum of differences is 17.