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 reject $H_0 : \beta_{\text{takers}^2} = 0$ and 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

QUESTION D

Excercise 2.2

2.2a

```r
data$type <- factor(data$type)
model_a = lm(volume ~ type, data = data)
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
```

A one-way ANOVA comparing mean wood volume between Beech and Oak trees (n = 59) found no statistically significant difference in average volume between the species ($F = 1.90, p = 0.17$); thus, based on these data, we cannot conclude that Oaks are more voluminous than Beeches at the 5% significance level. The estimated mean volumes were approximately $\bar{V}_{\text{Oak}} \approx$ [insert mean] and $\bar{V}_{\text{Beech}} \approx$ [insert mean]; although Oaks appear slightly larger on average, this observed difference is not statistically significant.

2.2b

```r
data$type = factor(data$type)
model_b <- lm(volume ~ diameter + height + type, data = data)
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
```

NOT DONE YET

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).

```r
trees$type = factor(trees$type)
trees$calc_vol = pi*((trees$diameter/2)^2)*trees$height
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, the options menu has the same configuration as in the image below.

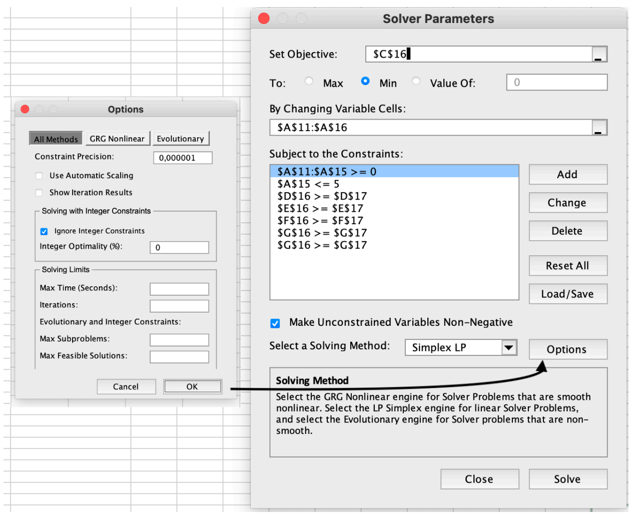

Figure 1: Options Menu

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.

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

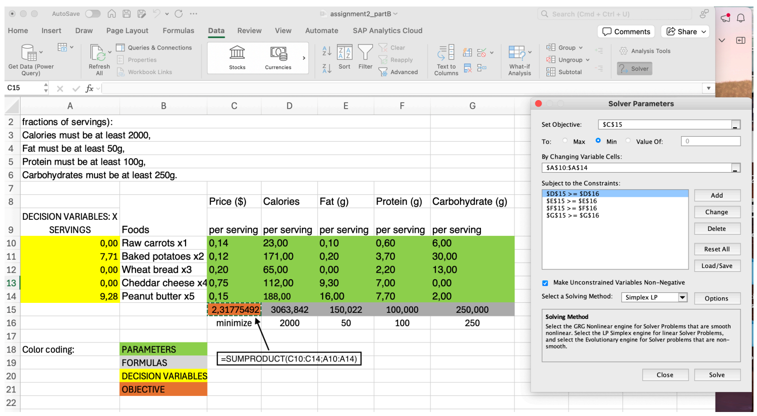

Figure 2: 2.3a Excel Solver

The optimal continuous solution selects **7.7147** servings of potatoes and **9.2800** servings of peanut butter, with **0** servings of carrots, bread, and cheddar; the **total cost is $2.3178**. The **binding constraints are protein and carbohydrates**, while **calories** ($\approx$ 3063.8) and **fat** ($\approx$ 150.0 g) have slack.

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: $0 \le x_{\text{pb}}^{(1)} \le 5$.

All nutrients per serving are identical for both PB tiers.

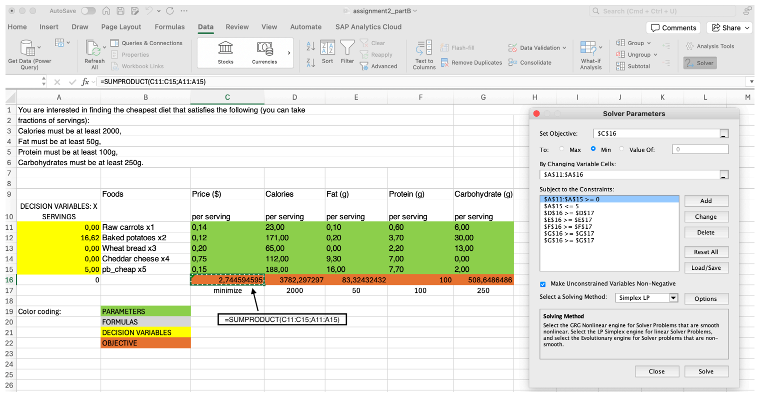

Figure 3: 2.3b Excel Solver

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:

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

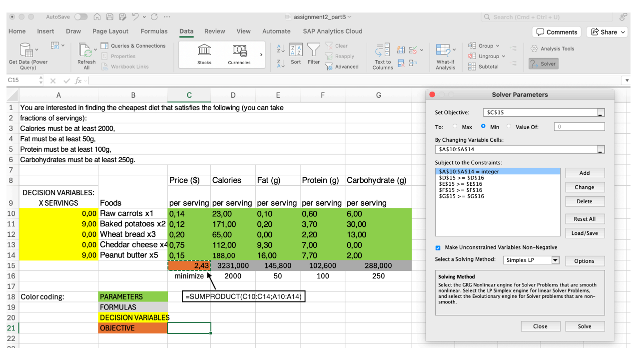

Figure 4: 2.3c Excel Solver

Interpretation. Integrality removes the ability to hit thresholds exactly, so the solution "rounds up" and typically costs more than the fractional LP.

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.

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

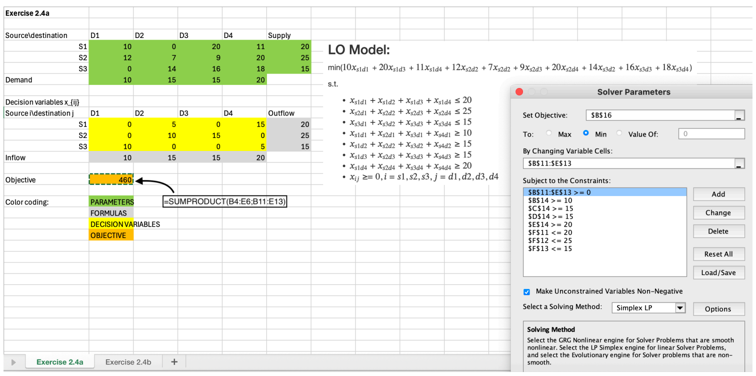

Figure 5: 2.4a Excel Solver

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**.

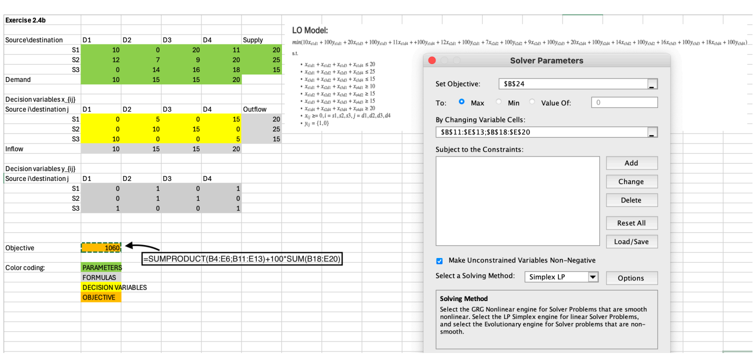

Figure 6: 2.4b Excel Solver

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