Take Home Assignment 1

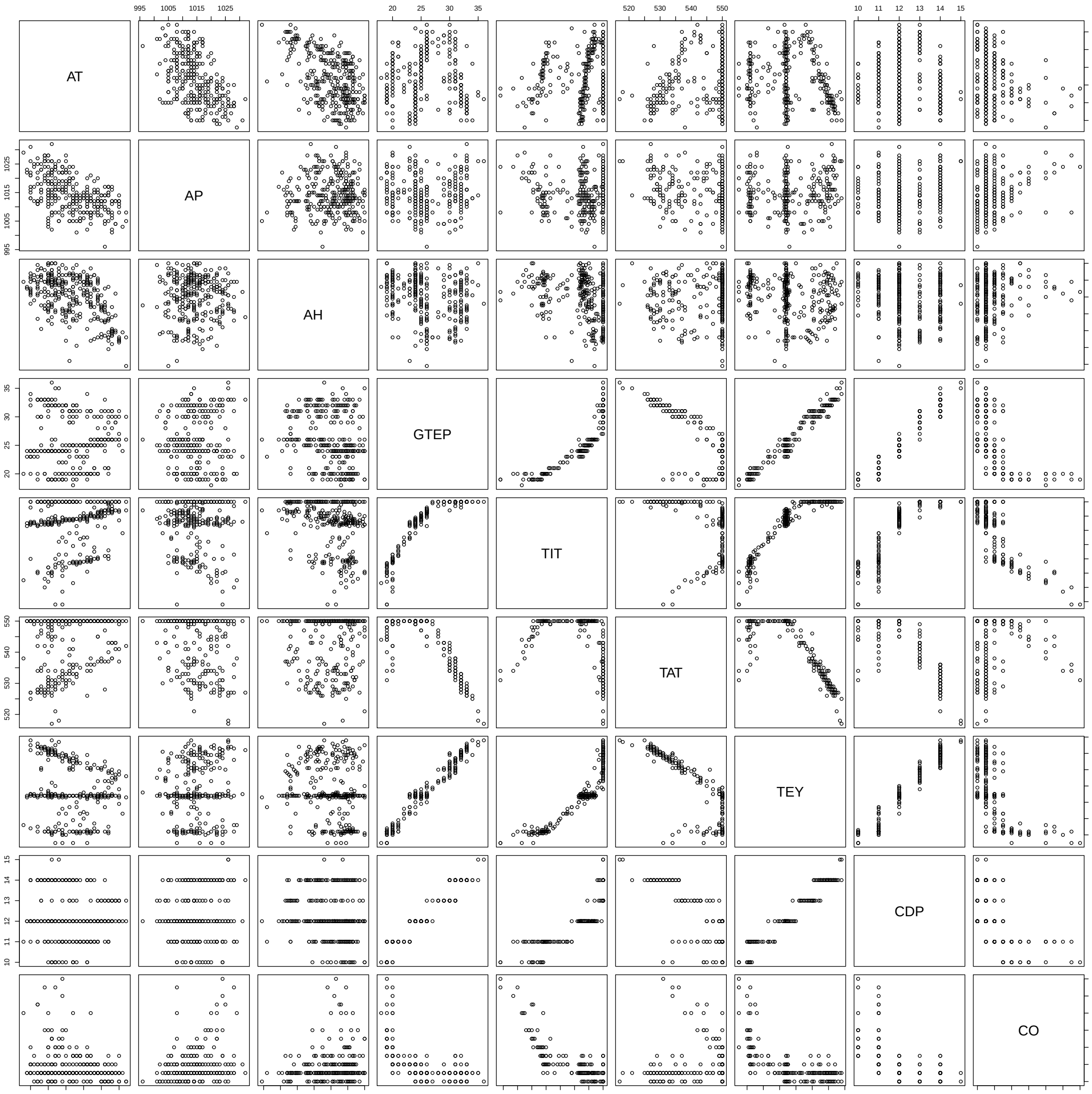
Jan Ryszkiewicz

Task 1

subtask 1

We perform basic inspection:

```r
load("Data ST523 813 E2025 Exam.rdata")
df <- Data
pairs(df)
```




```r
nrow(df)
```

```
[1] 300
```

```r
ncol(df)
```

```
[1] 9
```

```r
summary(df)
```

```
       AT              AP              AH              GTEP            TIT      
 Min.   : 3.00   Min.   : 996   Min.   : 39.00   Min.   :18.00   Min.   :1028  
 1st Qu.:11.00   1st Qu.:1010   1st Qu.: 72.00   1st Qu.:23.00   1st Qu.:1083  
 Median :16.00   Median :1014   Median : 82.00   Median :25.00   Median :1089  
 Mean   :16.74   Mean   :1015   Mean   : 79.82   Mean   :25.85   Mean   :1085  
 3rd Qu.:22.25   3rd Qu.:1019   3rd Qu.: 90.00   3rd Qu.:30.00   3rd Qu.:1100  
 Max.   :32.00   Max.   :1032   Max.   :100.00   Max.   :36.00   Max.   :1100  
      TAT             TEY              CDP              CO      
 Min.   :517.0   Min.   :105.0   Min.   :10.00   Min.   : 0.0  
 1st Qu.:536.0   1st Qu.:130.0   1st Qu.:12.00   1st Qu.: 1.0  
 Median :550.0   Median :134.0   Median :12.00   Median : 1.0  
 Mean   :543.8   Mean   :136.5   Mean   :12.29   Mean   : 1.7  
 3rd Qu.:550.0   3rd Qu.:151.0   3rd Qu.:13.00   3rd Qu.: 2.0  
 Max.   :550.0   Max.   :168.0   Max.   :15.00   Max.   :12.0  
```

We fit a default model:

```r
model = lm(CO ~ AT + AP + AH + GTEP + TIT + TAT + CDP + TEY, data = df)
```

subtask 2

Display all the estimated parameters:

```r
coef(model)
```

```
  (Intercept)            AT            AP            AH          GTEP 
124.201440581  -0.021611280   0.009736961  -0.009621796  -0.370650220 
          TIT           TAT           CDP           TEY 
  0.029651998  -0.255240883  -0.463761682  -0.068439127 
```

And the one for ambient temperature in particular:

```r
coef(model)["AT"]
```

```
         AT 
-0.02161128 
```

From this we know that the estimated change in CO for a 1°C increase in ambient temperature, other predictors constant, is -0.0216 units.

subtask 3

Perform F-Test:

```r
summary(model)
```

```
Call:
lm(formula = CO ~ AT + AP + AH + GTEP + TIT + TAT + CDP + TEY,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8724 -0.5804 -0.0571  0.4270  4.3507 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 124.201441  17.933449   6.926 2.78e-11 ***
AT           -0.021611   0.028025  -0.771 0.441242    
AP            0.009737   0.012831   0.759 0.448558    
AH           -0.009622   0.005824  -1.652 0.099605 .  
GTEP         -0.370650   0.158512  -2.338 0.020048 *  
TIT           0.029652   0.058304   0.509 0.611437    
TAT          -0.255241   0.072992  -3.497 0.000544 ***
CDP          -0.463762   0.221422  -2.094 0.037083 *  
TEY          -0.068439   0.071399  -0.959 0.338581    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9744 on 291 degrees of freedom
Multiple R-squared:  0.7386,    Adjusted R-squared:  0.7314 
F-statistic: 102.8 on 8 and 291 DF,  p-value: < 2.2e-16
```

As we can see:

`F-statistic: 102.8 on 8 and 291 DF,  p-value: < 2.2e-16`

P-value is very low which indicates the rejection of H_0 - global null hypotesis at significance $\alpha = 0.05$

We have 291 residual degrees of freedom (n - p) and 8 model degrees of freedom (p - 1)

Value of the F-test statistic is 102.8

subtask 4

Let's start by fitting 2 submodels:

```
M_A = lm(CO ~ AT + AP + AH, data = df)
M_B = lm(CO ~ GTEP + TIT + TAT + TEY, data = df)
```

Now try comparing ambient only model to the default:

```
anova(M_A, model)
```

```
Analysis of Variance Table

Model 1: CO ~ AT + AP + AH
Model 2: CO ~ AT + AP + AH + GTEP + TIT + TAT + CDP + TEY
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    296 971.35
2    291 276.31  5    695.04 146.4 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see by the corresponding F-statistic (146.4) and p-value ($<$ 2.2e-16) The default model explains the data *much better* i.e.

$$\exists_{i \neq AT,AP,AH} \text{ such that } \beta_i \neq 0$$

So we reject the H_0 that all other predictiors aside from AT,AP,AH are $= 0$.

What about the process only model?:

```
anova(M_B, model)
```

```
Analysis of Variance Table

Model 1: CO ~ GTEP + TIT + TAT + TEY
Model 2: CO ~ AT + AP + AH + GTEP + TIT + TAT + CDP + TEY
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    295 286.13
2    291 276.31  4    9.8158 2.5844 0.0373 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The rejection of the null hypothesis—that the reduced (process-only) model explains the data equally well—is supported by the F statistic (~2.58) and the corresponding p-value (0.037). This suggests that, at a significance level of $\alpha = 0.05$, adding other predictors leads to a statistically significant improvement in model fit. However, for $\alpha = 0.01$, the evidence would not be strong enough to reject the null hypothesis.

Therefore we conclude that we *can reduce* the default model to the Process Only, however only when accepting our significance level to be < 0.04

subtask 5

From the last subtask we can conclude that M_B is better than M_A therefore M_B is chosen as our final model for this subtask.

By inspecting adjusted R^2 between models we can see the diffrence of total adjusted explained variation between models.

```
R_Ambient = summary(M_A)$adj.r.squared
R_Process = summary(M_B)$adj.r.squared
R_default = summary(model)$adj.r.squared

R_Ambient
```

```
[1] 0.07171319
```

```
R_Process
```

```
[1] 0.7256326
```

```
R_default
```

```
[1] 0.731403
```

The absolute reduction of unexplained variation for Process Only model M_B:

```
explained_variation_Model_B = summary(M_B)$r.squared
CO = df$CO
TSS = sum((CO - mean(CO))^2)
RSS = sum(residuals(M_B)^2)
absolute_reduction_Model_B = TSS - RSS

explained_variation_Model_B
```

```
[1] 0.7293031
```

```r
absolute_reduction_Model_B
```

```
[1] 770.8733
```

This basically concludes *Task 1* but I wanted to explore a certain idea with the presented models.

Let us introduce another alternative model M_Alt that includes CDP as one of its predictors:

```r
M_Alt = lm(CO ~ GTEP + TIT + TAT + CDP + TEY, data = df)

anova(M_Alt,M_B)
```

```
Analysis of Variance Table

Model 1: CO ~ GTEP + TIT + TAT + CDP + TEY
Model 2: CO ~ GTEP + TIT + TAT + TEY
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    294 281.35
2    295 286.13 -1   -4.7816 4.9967 0.02615 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyzing this anova table we can conclude that the alternative model is better as the perforemd partial F test returned low p'value

However we could further reduce the M_Alt:

```r
anova(M_Alt)
```

```
Analysis of Variance Table

Response: CO
           Df Sum Sq Mean Sq  F value    Pr(>F)
GTEP        1 341.46  341.46 356.8233 < 2.2e-16 ***
TIT         1 401.59  401.59 419.6583 < 2.2e-16 ***
TAT         1  24.64   24.64  25.7507 6.899e-07 ***
CDP         1   5.96    5.96   6.2230   0.01316 *
TEY         1   2.00    2.00   2.0886   0.14946
```

```
Residuals 294 281.35    0.96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see that the F test corresponding to the inclusion of TEY predictor shows a relatively high p'value that might indicate corresonding $\beta_{TEY} \approx 0$

It is worth performing R^2 inspection for the alternate model M_Alt that includes CDP predictor and removes TEY as it seemed redundant:

```r
M_Alt = lm(CO ~ GTEP + TIT + TAT + CDP, data = df)

explained_variation_ModelAlt = summary(M_Alt)$r.squared
RSS = sum(residuals(M_Alt)^2)
absolute_reduction_ModelAlt = TSS - RSS

explained_variation_ModelAlt
```

```
[1] 0.7319358
```

```r
absolute_reduction_ModelAlt
```

```
[1] 773.6562
```

The absolute reduction in the residual sum of squares is only slightly higher for model M_Alt, at the cost of including one additional predictor *(CDP)*, and removing one other *(TEY)*. Depending on our circumstances, we can choose to either stay by *M_Alt* for marginally better performance or opt for the different *M_B*, which performs almost equivalently.

Task 2

subtask 1

The model presented in task can also be written as:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{45} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & \mathbb{1}_{2,1} & \mathbb{1}_{3,1} & \mathbb{1}_{4,1} & X_1 \\ 1 & \mathbb{1}_{2,2} & \mathbb{1}_{3,2} & \mathbb{1}_{4,2} & X_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbb{1}_{2,45} & \mathbb{1}_{3,45} & \mathbb{1}_{4,45} & X_45 \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ b \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{45} \end{bmatrix}.$$

$$\mathbb{1}_{i,j} = \begin{cases} 1 & \text{if } \alpha_j \text{ present in sample } i \\ 0 & \text{otherwise} \end{cases}$$

Where column 1 of X corresponds to always present μ and each column $2, 3, 4$ is filled with $0, 1$ depending on corresponding $\alpha_{j(i)}$ (indicator function)

Also α_i defined as:

$$\begin{aligned}
\alpha_2 &= \text{Temporary – Research/Academic (relative to Permanent)} \\
\alpha_3 &= \text{Temporary – Private Consultant (relative to Permanent)} \\
\alpha_4 &= \text{Freelance (relative to Permanent)}
\end{aligned}$$

We know that:

$$\begin{aligned}
\text{Cov}(\hat{\alpha}_2, \hat{\alpha}_3) &= 22{,}000{,}000 \\
\text{Cov}(\hat{\alpha}_2, \hat{\alpha}_4) &= 20{,}000{,}000 \\
\text{Cov}(\hat{\alpha}_3, \hat{\alpha}_4) &= 21{,}000{,}000
\end{aligned}$$

And:

$$\begin{aligned}
\hat{\text{SE}}(\hat{\mu}) &= 20{,}000 \\
\hat{\text{SE}}(\hat{\alpha}_2) &= 24{,}000 \\
\hat{\text{SE}}(\hat{\alpha}_3) &= 23{,}000 \\
\hat{\text{SE}}(\hat{\alpha}_4) &= 22{,}000
\end{aligned}$$

First we want to calculate the CL for $\alpha_2 - \alpha_3$ with confidence 90%

Which is:

$$\begin{aligned}
c &= [0, 1, -1, 0, 0], \\
\hat{\psi} &= c^T \hat{\beta} = \hat{\alpha}_2 - \hat{\alpha}_3, \\
\hat{\text{SE}}(\hat{\psi}) &= \sqrt{Var(\hat{\alpha}_2 - \hat{\alpha}_3)} \\
&= \sqrt{cVar(\hat{\beta})c^T}, \\
&= \sqrt{Var(\hat{\alpha}_2) + Var(\hat{\alpha}_3) - 2Cov(\hat{\alpha}_2, \hat{\alpha}_3)}, \\
\text{CI}_{90\%} &= \hat{\psi} \pm t_{45-5,\, 1-0.10/2} \cdot \hat{\text{SE}}(\hat{\psi})
\end{aligned}$$

That then by substitution becomes:

$$
\begin{aligned}
\hat{\psi} &= -40{,}000 - (-10{,}000) \\
&= -30{,}000 \\
\hat{\mathrm{SE}}(\hat{\psi}) &= \sqrt{(24{,}000)^2 + (23{,}000)^2 - 2 \cdot 22{,}000{,}000} \\
&\approx 32{,}573 \\
t_{40,\,(1-0.10)/2} &\approx 1.68385 \\
\mathrm{CI}_{90\%} &= \hat{\psi} \pm t_{45-5,\,1-0.10/2} \cdot \hat{\mathrm{SE}}(\hat{\psi}) \\
\mathrm{CI}_{90\%} &= -30{,}000 \pm 1.68385 \cdot 32{,}573 \\
&= -30{,}000 \pm 54848
\end{aligned}
$$

Therefore the Confidence interval is:

$$\mathrm{CI}_{90\%} = [-84{,}848 \ , \ 24{,}848]$$

That concludes the first subtask of Task 2.

subtask 2

Next we want to look for statistical evidence for: $\alpha_2 \leq \alpha_3$

Therefore we perform One Sided hypotesis Test on:

$$
\begin{aligned}
H_0 &= \alpha_2 - \alpha_3 > 0 \\
H_a &= \alpha_2 - \alpha_3 \leq 0
\end{aligned}
$$

H_0 - temporary researchers are earning **more** than temporary private consultants

H_a - the opposite

We are using the same $\hat{\psi}$ from previous subtask:

$$
\begin{aligned}
\hat{\psi} &= \hat{\alpha}_2 - \hat{\alpha}_3 \\
T &= \frac{(\hat{\alpha}_2 - \hat{\alpha}_3) - 0}{\hat{\mathrm{SE}}(\hat{\psi})} \\
&= \frac{-30{,}000}{32{,}573} \\
&\approx -0.921
\end{aligned}
$$

Then:

```r
qt(0.05, df = 40)
```

```
[1] -1.683851
```

$$
\begin{gathered}
-0.921 > -1.683851 \\
T > t_{40,\,0.05}
\end{gathered}
$$

We cannot reject H_0

Therefore there is not enough statistical evidence that temporary researchers are earning less than temporary private consultants.

Task 3

We have $x_i \in [-2, 2]$ for $i = 1, \dots, n$. The X and β matrces are as follows:

$$
X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.
$$

Therefore, the variance of $\hat{\beta}$ is

$$
\operatorname{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1} = \sigma^2 \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1}.
$$

Following the inverse this becomes

$$
\operatorname{Var}(\hat{\beta}) = \frac{\sigma^2}{n \sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}.
$$

And then, variance of the slope estimate $(\hat{\beta}_1)$ is:

$$
\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2 n}{n \sum_i x_i^2 - (\sum_i x_i)^2}
$$

Which further reduces to:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2} \\ &= \frac{\sigma^2}{\sum_i x_i^2 - \frac{2}{n}(\sum_i x_i)^2 + \frac{1}{n}(\sum_i x_i)^2} \\ &= \frac{\sigma^2}{\sum_i x_i^2 - 2\sum_i x_i \cdot \bar{x} + n\bar{x}^2} \\ &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\end{aligned}$$

We want to minimize the $\mathrm{Var}(\hat{\beta}_1)$. Which can be done by maximizing $\sum_i (x_i - \bar{x})^2$ as σ^2 is constant. This can be achieved by spreading $x_1, \dots, x_n$ as much as possible that will ideally set $\bar{x} = \frac{1}{n}\sum_i (x_i) = 0$

We know that:

$$x_i \in [-2, 2], \quad \forall i = 1, \dots, n$$

Then we can choose:

- $x_1, \dots, x_{n/2} = -2$
- $x_{n/2+1}, \dots, x_n = 2$

By that we minimize $\mathrm{Var}(\hat{\beta}_1)$ now equal to:

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i (\pm 2 - 0)^2} = \frac{\sigma^2}{4n}$$

And we achive maximum possible precission

Q.E.D.

Task 4

Only for Master's