Assignment

Nel Skowronek

Exercise 1

1. Exploring the Dataset

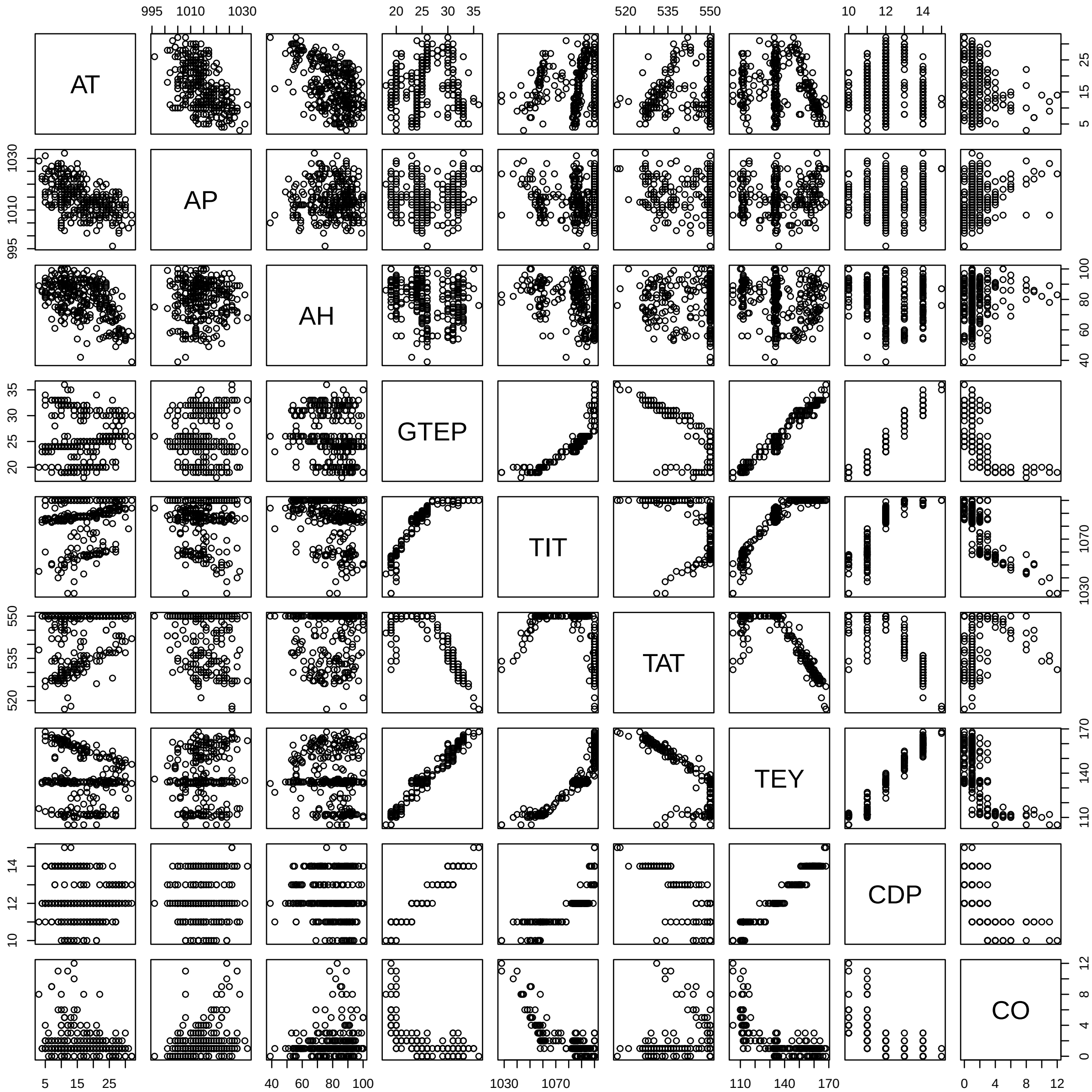
```r
load("Data ST523 813 E2025 Exam.rdata")
dim(Data)
```

```
[1] 300   9
```

```r
summary(Data)
```

```
       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  
```

```r
pairs(Data)
```



So we have 9 non-categorical variables and a total of 300 observations.

2. Linear Model

```
fit = lm(CO ~ AT + AP + AH + GTEP + TIT + TAT + TEY + CDP, data = Data)
summary(fit)
```

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

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 ***
TEY          -0.068439   0.071399  -0.959 0.338581
CDP          -0.463762   0.221422  -2.094 0.037083 *
---
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
```

$$n = 300 \quad p = 9$$

We can see the estimated intercept and other 8 parameters in the "Estimate" column above.
We have $300 - 9 = 291$ degrees of freedom.
Judging by the coefficient of Ambient Temperature predictor we can expect a 0.021611 decrease in CO-level for every $1^{\circ}C$ increase in Ambient Temperature.

3. F - test

- Hypothesis:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_p = 0 \quad H_1 : \exists_{j \in 2...p} \beta_j \neq 0$$

```r
fit.0 = lm(CO ~ 1, data = Data)
anova(fit.0, fit)
```

```
Analysis of Variance Table

Model 1: CO ~ 1
Model 2: CO ~ AT + AP + AH + GTEP + TIT + TAT + TEY + CDP
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    299 1057.00
2    291  276.31  8    780.69 102.77 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see the RSS drastically dropped in the wider model, compared to the null, which already gives us an indication that it might explain the data much better.

- The observed test statistic can be read from the table: $F = 102.77$
- The probability of getting such a high value is very small: $p_{val} < 2.2 \cdot 10^{-16}$
- Under the Null Hypothesis F should follow the $F_{8,291}$ distribution.

(Because the difference in parameters / degrees of freedom is 8, and the wider model has 291 degrees of freedom, which we've shown before)

The p-value is much smaller than the significance level therefore we can definitely reject the null hypothesis - we know that our test statistic must be much higher than the 0.95th quantile of the null distribution.

4. Reduced Models

Model A

Similarly as before, we state the hypothesis:

$$H_0 : \beta_5 = \beta_6 = ... = \beta_p = 0 \quad H_1 : \exists_{j \in 5...p} \beta_j \neq 0$$

```r
fit.A = lm(CO ~ AT + AP + AH, data = Data)
anova(fit.A, fit)
```

```
Analysis of Variance Table

Model 1: CO ~ AT + AP + AH
Model 2: CO ~ AT + AP + AH + GTEP + TIT + TAT + TEY + CDP
  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
```

And similarly as before we can see that the test statistic is very high (146.4), while the p-value is very low (close to 0) - which strongly speaks against the null hypothesis.
The conclusion is - we cannot reduce the full model to Model A.

Model B

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_9 = 0 \quad H_1 : \exists_{j \in 2,3,4,9} \beta_j \neq 0$$

```
fit.B = lm(CO ~ GTEP + TIT + TAT + TEY, data = Data)
anova(fit.B, fit)
```

```
Analysis of Variance Table

Model 1: CO ~ GTEP + TIT + TAT + TEY
Model 2: CO ~ AT + AP + AH + GTEP + TIT + TAT + TEY + CDP
  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
```

Here the situation is slightly different. The test statistic is much lower (2.5844), and the p-value is much higher than with the previous 2 models (0.0373) - but it's still quite low. Depending on our significance level we might reject the null hypothesis or not. For signif. level $\alpha = 0.05$ we would reject the null and say that the model can't be reduced. For signif. level $\alpha = 0.01$ however we would be able to say that the full model can be reduced to Model B.

Exercise 2

Model:

$$Y_i = \mu + \alpha_{j(i)} + \beta' \cdot X_i + \epsilon_i$$
$$i \in [1...n]$$
$$n = 45$$

Model Matrix and Coefficients

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & X_1 \\ \vdots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \mathbf{1} & \mathbf{0} & \vdots \\ 1 & \mathbf{0} & \mathbf{0} & \mathbf{1} & X_n \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \\ \alpha_{TempResearch} \\ \alpha_{TempPrivate} \\ \alpha_{Freelance} \\ \beta' \end{bmatrix}$$
$$p = 5$$

1. Confidence Interval

$$c^T = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix}$$
$$c^T\beta = \alpha_{TempResearch} - \alpha_{TempPrivate}$$
$$CI = c^T\hat{\beta} \pm \widehat{SE}_{c^T\hat{\beta}} \cdot t_{n-p;1-\frac{\alpha}{2}}$$

The only thing in the above formula that we don't have is the $\widehat{SE}_{c^T\hat{\beta}}$. We usually used RSS to get it, but since we don't have access neither to the predictors nor responses we can use the output of the program to calculate it differently:

$$\widehat{SE}_{c^T\hat{\beta}} = \sqrt{Var(c^T\hat{\beta})}$$
$$Var(c^T\hat{\beta}) = Var(\hat{\alpha}_{TempResearch} - \hat{\alpha}_{TempPrivate}) =$$
$$= Var(\hat{\alpha}_{TempResearch}) + Var(\hat{\alpha}_{TempPrivate}) - 2Cov(\hat{\alpha}_{TempResearch},\ \hat{\alpha}_{TempPrivate}) =$$
$$= \widehat{SE}^2_{\hat{\alpha}_{TempResearch}} + \widehat{SE}^2_{\hat{\alpha}_{TempPrivate}} - 2Cov(\hat{\alpha}_{TempResearch},\ \hat{\alpha}_{TempPrivate})$$

Now we have everything we need to calculate the confidence interval.

```
n = 45
p = 5

a.re = -40000
a.pr = -10000
```

```r
psi = a.re - a.pr

SE.re = 24000
SE.pr = 23000
COV.re.pr = 22000000

SE = sqrt(SE.re ^ 2 + SE.pr ^ 2 - 2 * COV.re.pr)

CI.lower = psi - SE * qt(0.95, n - p)
CI.upper = psi + SE * qt(0.95, n - p)

c(CI.lower, CI.upper)
```

```
[1] -84848.07  24848.07
```

2. Hypothesis Testing

$$H_0 : \begin{array}{c}\alpha_{TempResearch} \geq \alpha_{TempPrivate} \\ \implies -c^T\beta \leq 0\end{array} \qquad H_1 : -c^T\beta > 0$$

We've shown on the lecture that if H_0 holds then:

$$T = \frac{-c^T\widehat{\beta} - 0}{\widehat{SE}_{c^T\widehat{\beta}}}$$
$$P(T > t_{n-p;1-\alpha}) \leq \alpha$$

Which means we will have statistical evidence to reject H_0 if the test statistic is bigger than the 0.95 quantile of the given above t_{n-p} - distribution.

```r
T = -psi / SE
t.95 = qt(0.95, n - p)

c(T, t.95)
```

```
[1] 0.9210083 1.6838510
```

As we can see, the test statistic - even though it's positive - is not bigger than the quantile, therefore we don't have sufficient statistical evidence to reject the null hypothesis and say that temporary researches are earning less than temporary private consultants.

Exercise 3

Model Matrix and Coefficients

For a simple linear relationship we have the model matrix and the parameter vector as so:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_{intercept} \\ \beta_{slope} \end{bmatrix}$$

Since we want to estimate β_{slope} as precisely as possible, we would like to minimize $Var(\hat{\beta}_{slope}) = Var(\hat{\beta})_{22}$.

We know that for the Least Square Estimator $\hat{\beta}$ $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

$$(X^TX)^{-1} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} = \begin{bmatrix} n & \sum x_i \\ \sum x_1 & \sum x_i^2 \end{bmatrix}^{-1}$$

We can find it's reverse using any method.

$$\left[\begin{array}{cc|cc} n & \sum x_i & 1 & 0 \\ \sum x_i & \sum x_i^2 & 0 & 1 \end{array} \right]$$

$$\left[\begin{array}{cc|cc} 1 & \bar{x} & \frac{1}{n} & 0 \\ \sum x_i & \sum x_i^2 & 0 & 1 \end{array} \right]$$

$$\left[\begin{array}{cc|cc} 1 & \bar{x} & \frac{1}{n} & 0 \\ 0 & \sum x_i^2 - \bar{x}\sum x_i & -\bar{x} & 1 \end{array} \right]$$

$$\left[\begin{array}{cc|cc} 1 & \bar{x} & \frac{1}{n} & 0 \\ 0 & \sum x_i(x_i - \bar{x}) & -\bar{x} & 1 \end{array} \right]$$

$$\left[\begin{array}{cc|cc} 1 & \bar{x} & \frac{1}{n} & 0 \\ 0 & 1 & \frac{-\bar{x}}{\sum x_i(x_i - \bar{x})} & \frac{1}{\sum x_i(x_i - \bar{x})} \end{array} \right]$$

At this point we don't even need to calculate further, because we only want $Var(\hat{\beta})_{22}$ which we can see is equal to $\frac{\sigma^2}{\sum x_i(x_i - \bar{x})}$