# MATH38141 Project 1

Wei Chung-Yu

Nov 2023

## 1 Q1a

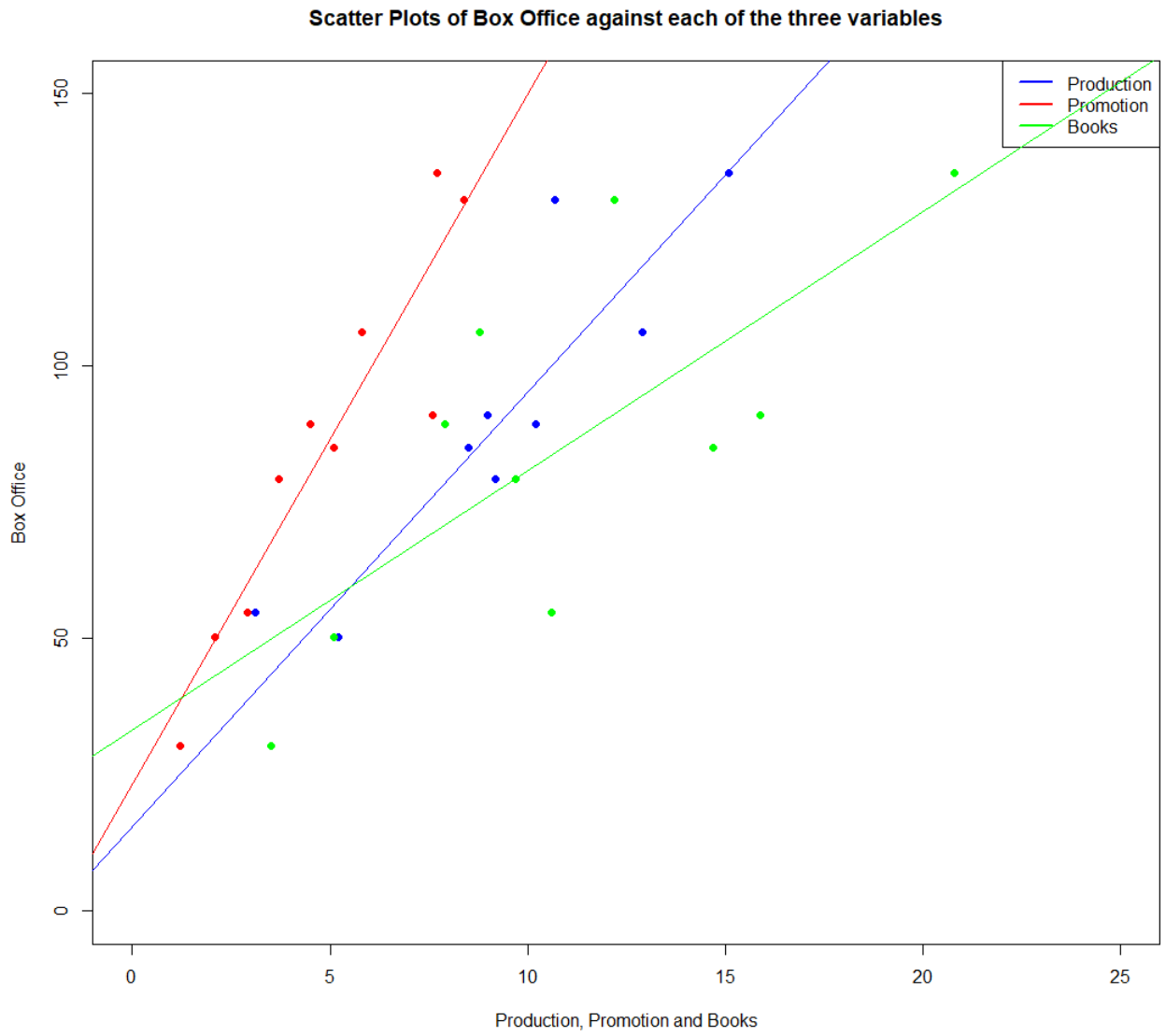


Figure 1

From this scatter plot, it is easy to tell the data values show a linear pattern between the Box office and the other three variables in general. Even though all three variables show linear pattern, the data

points for promotion and production show strong linear pattern as the data points mainly lies on the regression lines, and the data points for variable books show weak linear pattern, since the data points does not concentrated on the regression lines. In summary, the increase in production, promotion and books leads to the increase in Box office.

## 2 Q1b

$$\text{BoxOffice} = \beta_0 + \beta_1 \text{Production} + \beta_2 \text{Promotion} + \beta_3 \text{Books} + \epsilon$$

## 3 Q1c

First make a matrix of X, Y and $X^T$

$$X = \begin{bmatrix} 1 & 8.5 & 5.1 & 14.7 \\ 1 & 12.9 & 5.8 & 8.8 \\ 1 & 5.2 & 2.1 & 5.1 \\ 1 & 10.7 & 8.4 & 12.2 \\ 1 & 3.1 & 2.9 & 10.6 \\ 1 & 3.5 & 1.2 & 3.5 \\ 1 & 9.2 & 3.7 & 9.7 \\ 1 & 9.0 & 7.6 & 15.9 \\ 1 & 15.1 & 7.7 & 20.8 \\ 1 & 10.2 & 4.5 & 7.9 \end{bmatrix} \qquad Y = \begin{bmatrix} 85.1 \\ 106.3 \\ 50.2 \\ 130.6 \\ 54.8 \\ 30.3 \\ 79.4 \\ 91.0 \\ 135.4 \\ 89.3 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 8.5 & 12.9 & 5.2 & 10.7 & 3.1 & 3.5 & 9.2 & 9.0 & 15.1 & 10.2 \\ 5.1 & 5.8 & 2.1 & 8.4 & 2.9 & 1.2 & 3.7 & 7.6 & 7.7 & 4.5 \\ 14.7 & 8.8 & 5.1 & 12.2 & 10.6 & 3.5 & 9.7 & 15.9 & 20.8 & 7.9 \end{bmatrix}$$

$$XTX = \begin{bmatrix} 10.0 & 87.40 & 49.00 & 109.20 \\ 87.4 & 899.74 & 496.77 & 1067.64 \\ 49.0 & 496.77 & 295.46 & 626.58 \\ 109.2 & 1067.64 & 626.58 & 1434.94 \end{bmatrix}$$

$$XTX^{-1} = \begin{bmatrix} 0.76114269 & -5.239098 \times 10^{-2} & 0.02743088 & -3.092096 \times 10^{-2} \\ -5.239098 \times 10^{-2} & 1.957698 \times 10^{-2} & -0.02422813 & 5.483330 \times 10^{-7} \\ 0.02743088 & -2.422813 \times 10^{-2} & 0.07798774 & -1.811511 \times 10^{-2} \\ -3.092096 \times 10^{-2} & 5.483330 \times 10^{-7} & -0.01811511 & 1.095973 \times 10^{-2} \end{bmatrix}$$

The least square error can be calculated by using this formula.

$$\hat{B} = (X^T X)^{-1} X^T Y$$

$$\hat{B} = \begin{bmatrix} 12.02862033 \\ 4.22824115 \\ 7.54180158 \\ -0.06394467 \end{bmatrix}$$

By using R, we can get the result

$B_0 = 12.02862033$, $B_1 = 4.22824115$, $B_2 = 7.54180158$, $B_3 = -0.06394467$

For the confidence interval, we can use this formula to calculate for each regression coefficients.

$$\hat{B}_i \pm t_{n-p-1}^{1-\frac{\alpha}{2}} \sqrt{s^2 g_{ii}}$$

where $g^{ii}$ is the ith diagonal element of the matrix $(X^T X)^{-1}$. In this case $a$ is 5%, $p$ is the length of $\hat{B}$ which is 4, and

$$S^2 = \frac{SSE}{n-p}$$

Calculate $\hat{Y}$ using this formula:

$$\hat{Y} = X \cdot \hat{\beta} = \begin{bmatrix} 85.49187 \\ 109.75267 \\ 49.52714 \\ 119.84181 \\ 46.32958 \\ 35.65382 \\ 78.21284 \\ 106.38376 \\ 132.61688 \\ 88.58962 \end{bmatrix}$$

Calculate $\epsilon$ using this formula:

$$\epsilon = \hat{Y} - Y = \begin{bmatrix} 0.3918716 \\ 3.4526673 \\ -0.6728602 \\ -10.7581910 \\ -8.4704210 \\ 5.3538199 \\ -1.1871585 \\ 15.3837625 \\ -2.7831151 \\ -0.7103757 \end{bmatrix}$$

The Residual sum of squares(SSE) can be calculated by this formula:

$$SSE = \sum \epsilon^2 = 474.9972$$

and we can get $S^2$ equals to 79.16619.

$t_5^{0.975} = 2.446912$ by using R. With these data, we can calculate the 95% confidence intervals for all regression coefficients:

$$\hat{B}_0 \pm 2.446912\sqrt{79.16619 g_{11}} = (-6.96559, 31.02283)$$

$$\hat{B}_1 \pm 2.446912\sqrt{79.16619 g_{22}} = (1.182023, 7.274460)$$

$$\hat{B}_2 \pm 2.446912\sqrt{79.16619 g_{33}} = (1.461832, 13.621771)$$

$$\hat{B}_3 \pm 2.446912\sqrt{79.16619 g_{44}} = (-2.343174, 2.215285)$$

# 4 Q1d

According to the signs of estimates for $\hat{B}$, it seems that the BoxOffice increases as Production and Promotion increase because $B_1$ and $B_2$ have the positive sign. It is reduced by Books because $B_3$ has the negative sign, however, the impact of books can be positive since the confidence interval (-2.343174, 2.215285) lies on both sides of 0.

# 5 Q1e

To calculate for the $R^2$ statistic, we need to first calculate the total corrected sum of squares(SYY):

$$SYY = (Y^T Y) - n(\bar{Y})^2 = 10273.66$$

$$SSR = SYY - SSE = 9798.667$$

$$R^2 = SSR/SYY = 0.9537656 = 95.37656\%$$

The $R^2$ reflects that 95.37656% of the total variation in Y is explained by the regression model, so the model fits the data well.

# 6 Q1f

0 is included in the confidence interval for $B_0$ at 95% level. Therefore we can not reject the null hypothesis.

# 7 Q1g

The reduced model is Model 1.

| **Source** | **S.S.** | **d.f.** | **m.s.** | **F-ratio** |
|---|---|---|---|---|
| Regression fitting reduced model | SSR$\omega$=Syy-SSE$\omega$ = 9798.26 | k = 2 | - | - |
| Extra | SSEXT = 0.3728 | p-k = 1 | MSEXT= SSEXT / (p-k) = 0.4 | F = MSEXT / MSE$\Omega$=0.004712694 |
| Residual fitting full model | SSE$\Omega$ = 474.9972 | n-p-1 = 6 | MSE$\Omega$ = SSE$\Omega$ / (n-p-1) = 79.17 | - |
| Total | Syy = 10273.66 | n-1 = 9 | - | - |

Table 1: ANOVA Table

# 8 Q1h

By forming the ANOVA table for Model 1 and Model 2. We can get the result:

$$P(F_{1,6} > 0.004712694) = 0.9474996$$

which shows that the interaction effects are insignificant. The high p-value suggests that the inclusion of Books in the regression model does not significantly improve the fit of the model compared to the reduced model without Books. So Books should not be included into the model.

## 9 Q1i

Regressing BoxOffice on Books alone. We can get the result:

$$P(F_{1,8} > 9.166514) = 0.01636844$$

The p-value = 0.01636844 which is smaller than 0.05. Since the p-value is less than 0.05, we can reject the null hypothesis that Books has no effect on BoxOffice in the context of the simple linear regression model.

This result contradicts with part(h) because the books does have impact on Boxoffice, even if Books is not significant when Production and Promotion is inside the model, when we do not consider the other two variables, the SLM is still significant.

## 10 Q2a

Multiple linear regression model $\Omega : \text{ANS} = \beta_0 + \beta_1\text{NSF} + \beta_2\text{INV} + \beta_3\text{ASA} + \beta_4\text{SSD} + \beta_5\text{NCS} + \epsilon$

Simple linear regression model $\omega : \text{ANS} = \beta_0 + \beta_2\text{INV} + \epsilon$

The assumptions that have been made are: the errors are independent random variables with zero mean and homogeneous variance.

## 11 Q2b

$$X_\omega = \begin{bmatrix} 1 & 3.0 & 294 & 8.2 & 8.2 & 11 \\ 1 & 2.2 & 232 & 6.9 & 4.1 & 12 \\ 1 & 0.5 & 149 & 3.0 & 4.3 & 15 \\ 1 & 5.5 & 600 & 12.0 & 16.1 & 1 \\ 1 & 4.4 & 567 & 10.6 & 14.1 & 5 \\ 1 & 4.8 & 571 & 11.8 & 12.7 & 4 \\ 1 & 3.1 & 512 & 8.1 & 10.1 & 10 \\ 1 & 2.5 & 347 & 7.7 & 8.4 & 12 \\ 1 & 1.2 & 212 & 3.3 & 2.1 & 15 \\ 1 & 0.6 & 102 & 4.9 & 4.7 & 8 \\ 1 & 5.4 & 788 & 17.4 & 12.3 & 1 \\ 1 & 4.2 & 577 & 10.5 & 14.0 & 7 \\ 1 & 4.7 & 535 & 11.3 & 15.0 & 3 \\ 1 & 0.6 & 163 & 2.5 & 2.5 & 14 \\ 1 & 1.2 & 168 & 4.7 & 3.3 & 11 \\ 1 & 1.6 & 151 & 4.6 & 2.7 & 10 \\ 1 & 4.3 & 342 & 5.5 & 16.0 & 4 \\ 1 & 2.6 & 196 & 7.2 & 6.3 & 13 \\ 1 & 3.8 & 453 & 10.4 & 13.9 & 7 \\ 1 & 5.3 & 518 & 11.5 & 16.3 & 1 \\ 1 & 5.6 & 615 & 12.3 & 16.0 & 0 \\ 1 & 0.8 & 278 & 2.8 & 6.5 & 14 \\ 1 & 1.1 & 142 & 3.1 & 1.6 & 12 \\ 1 & 3.6 & 461 & 9.6 & 11.3 & 6 \\ 1 & 3.5 & 382 & 9.8 & 11.5 & 5 \\ 1 & 5.1 & 590 & 12.0 & 15.7 & 0 \\ 1 & 8.6 & 517 & 7.0 & 12.0 & 8 \end{bmatrix} \qquad X_\Omega = \begin{bmatrix} 1 & 294 \\ 1 & 232 \\ 1 & 149 \\ 1 & 600 \\ 1 & 567 \\ 1 & 571 \\ 1 & 512 \\ 1 & 347 \\ 1 & 212 \\ 1 & 102 \\ 1 & 788 \\ 1 & 577 \\ 1 & 535 \\ 1 & 163 \\ 1 & 168 \\ 1 & 151 \\ 1 & 342 \\ 1 & 196 \\ 1 & 453 \\ 1 & 518 \\ 1 & 615 \\ 1 & 278 \\ 1 & 142 \\ 1 & 461 \\ 1 & 382 \\ 1 & 590 \\ 1 & 517 \end{bmatrix}$$

$$\hat{B}_\Omega = (X_\Omega^T X_\Omega)^{-1} X_\Omega^T Y = \begin{bmatrix} -18.5685907 \\ 16.2051112 \\ 0.1747971 \\ 11.5133677 \\ 13.5652689 \\ -5.3234386 \end{bmatrix}$$

$$\hat{B}_\omega = (X_\omega^T X_\omega)^{-1} X_\omega^T Y = \begin{bmatrix} -81.43580 \\ 0.949796 \end{bmatrix}$$

$\text{SSE}_\Omega = \sum \epsilon^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}))^2 = 6524.62$
$\text{SSE}_\omega = \sum \epsilon^2 = \sum_{i=1}^{n} (y_i - (\alpha_0 + \alpha_1 x_i))^2 = 101631.8$

## 12 Q2c

$\text{SSE}_\Omega \leq \text{SSE}_\omega$ because $\Omega \supset \omega$. The size of the difference

$$(extras.s.) SSEXT = SSE_\omega - SSE_\Omega$$

## 13 Q2d

Confidence interval for ASA:

$$\hat{\beta}_{ASA} \pm \left(\frac{S}{\sqrt{S_{xx}}}\right) t_{n-2}^{(\alpha/2)} = 11.5133677 \pm \left(\frac{S}{\sqrt{S_{xx}}}\right) t_{21}^{0.05} = (7.16186458, 15.864871)$$

Confidence interval for SSD:

$$\hat{\beta}_{SSD} \pm \left(\frac{S}{\sqrt{S_{xx}}}\right) t_{n-2}^{(\alpha/2)} = 13.5652689 \pm \left(\frac{S}{\sqrt{S_{xx}}}\right) t_{21}^{0.05} = (10.52268068, 16.607857)$$

Where $S_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$ and $s = \sqrt{\frac{SSE_\Omega}{n-p}} = 17.62658$.

Since $7.16186458 < 15 < 15.864871$, 15 lies inside the confidence interval at 10% significance level. We accept $H_0$, that the regression coefficient of ASA is 15.

Since 10 does not lies inside the confidence interval at 10% significance level. We have enough evidence to reject $H_0$, and the coefficient of SSD is not 10.

## 14 Q2e

| shop | NSF | INV | ASA | SSD | NCS |
|---|---|---|---|---|---|
| 1 | 3.0 | 500 | 5.0 | 5.0 | 5 |
| 2 | 5.0 | 500 | 10.0 | 10.0 | 10 |

$$\hat{\beta} = (X_\Omega^T X_\Omega)^{-1} X_\Omega^T Y = \begin{bmatrix} -18.5685907 \\ 16.2051112 \\ 0.1747971 \\ 11.5133677 \\ 13.5652689 \\ -5.3234386 \end{bmatrix}$$

The CI for difference of shop 1 and shop 2 at 5% significance level is defined by

$$\hat{l} \pm s.e.(\hat{l} - l)t_{n-p-1}^{(\alpha/2)} = \hat{l} \pm s.e.(\hat{l} - l)t_{21}^{(0.0025)}$$

$$s.e.(\hat{l} - l) = s\left\{a^T X^*(X^T X)^{-1} X^{*T} a + a^T a\right\}^{1/2}$$

$$\hat{l} = a^T X^* \hat{\beta}$$

$$a = [1, -1]^T, s = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2}{n-p}} = 17.62658$$

$$X^* = \begin{bmatrix} 1 & 3 & 500 & 5 & 5 & 5 \\ 1 & 5 & 500 & 10 & 10 & 10 \end{bmatrix}$$

Hence,

$$\hat{l} = -131.1862$$

$$s.e.(\hat{l} - l) = 35.03966$$

$$t_{21}^{(0.0025)} = 2.079614$$

$$\text{CI} = (-204.0552, -58.31725)$$

$$\Omega : \text{ANS} = \beta_0 + \beta_1 \text{NSF} + \beta_2 \text{INV} + \beta_3 \text{ASA} + \beta_4 \text{SSD} + \beta_5 \text{NCS} + \epsilon$$

We get predicted difference $= ANS_1 - ANS_2 = 216.2213 - 347.4075 = -131.1862$
The CI for the difference of annual net sales between shop 2 and shop 1 is (-204.0552, -58.31725) at 5% significance level. Since the confidence interval represents the range of values that are plausible for the true mean difference in net sales between the two shops, the actual difference (-131.1862) aligns with the model's predictions.

Therefore, we can conclude that two shops does not perform significantly differently at a 5% significance level.

# 15 Q2f

## 15.1 (i)

Models:
$\Omega_1 : \text{ANS} = \beta_0 + \beta_2 \text{INV} + \beta_5 \text{NCS} + \epsilon$
$\omega : \text{ANS} = \beta_0 + \beta_2 \text{INV} + \epsilon$

## 15.2 (ii)

$\text{SSE}_\omega = \sum \epsilon^2 = \sum_{i=1}^n (y_i - (\alpha_0 + \alpha_1 x_i))^2 = 101631.8$
$\text{SSE}_{\Omega 1} = \sum \epsilon^2 = 40575.46$
SSEXT= 101631.8-40575.46 = 61056.34

$$F - ratio = \frac{\frac{SSEXT}{p-k}}{MSE_{\Omega 1}} = \frac{\frac{61056.34}{3-1}}{1690.644} = 36.114 \quad (1)$$

The F-ratio is associated with p-value $= \text{P}(F_{2,23} > 36.114) = 0.000003329$, so we reject $H_0$. Include NCS in the model will make the ANS significantly different.

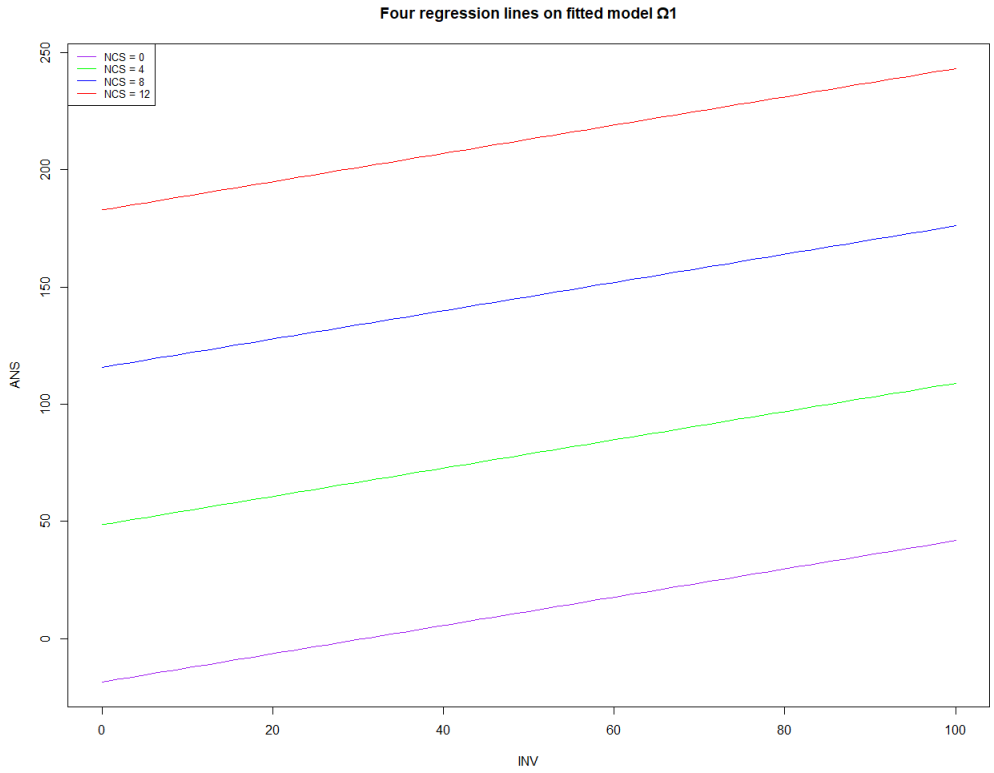


Figure 2: ANS vs INV for the four values of NCS

### 15.3 (iii)

From the graph we can see that all four regression lines increases linearly, so ANS increases as INV increases. However, the higher the NCS level, the lower the starting point (intercept) of the line on the ANS axis.