Assignment 2. Solution

Question 1 [20 marks].

```
Budget=c(162,5,154,56,37,190,191,136,75,70,58,47,54,49,42,264,91,51,39,38,39,34,
26,24,19,24,19,19,13,24,20,21,78,75,126,151,20,20,8,26,164,33,57)
Opening=c(63.99,16.45,18.45,8.85,19.45,32.83,60.68,73.42,59.26,19.25,7.53,141.52,
16.77,14.90,30.86,50.59,15.38,33.12,55.29,27.85,22.79,44.36,31.61,6.12,23.12,
12.31,27.47,49.95,20.38,20.39,24.03,11.55,33.97,38.13,53.27,85.44,8.87,22.73,12.72,
47.11,43.33,12.32,21.48)
Theatres=c(3799,2848,3182,2807,2457,3616,3377,3838,3493,3566,2119,4016,3361,2733,
2757,3614,3724,3207,3661,2883,2954,3265,3143,2568,1876,2441,2951,3079,3235,3244,
2807,3105,3265,3165,3599,4101,2714,3024,1197,3126,4053,2744,3253)
Ratings=c(9.1,6.9,4.9,7.1,7.5,5.4,4.7,7.6,8.4,6.4,7.1,4.3,5.7,7.0,6.1,8.4,5.5,7.7,
6.1,6.8,7.1,7.6,5.1,4.5,6.7,6.2,5.6,7.0,5.6,4.0,5.6,5.3,6.3,8.3,7.1,6.8,5.6,7.2,7.9,
6.5,8.1,6.1,7.0)
USRevenue=c(294.4,56.5,134.8,28.9,47.9,83.3,159.3,255.0,255.8,52.9,35.9,294.6,64.1,
34.9,60.4,203.5,48.3,102.0,176.0,89.3,48.7,276.4,146.0,13.5,79.0,31.8,75.5,116.6,
59.1,39.4,45.4,24.8,101.7,120.6,104.6,180.5,19.2,72.6,32.9,70.3,216.8,47.7,90.0)
Movies=data.frame(USRevenue, Budget, Opening, Theatres, Ratings)

> model <- lm(USRevenue ~ Budget + Opening + Theatres + Ratings, data=Movies)
> summary(model)

Call:
lm(formula = USRevenue ~ Budget + Opening + Theatres + Ratings,
data = Movies)

Residuals:
    Min      1Q  Median      3Q     Max
-69.820 -22.803  -3.520   8.785 131.422

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -159.5254    57.6257  -2.768  0.00866 **
Budget         0.1314     0.1285   1.023  0.31268
Opening        2.0450     0.3100   6.597 8.67e-08 ***
Theatres       0.0235     0.0157   1.497  0.14254
Ratings       17.4440     5.3377   3.268  0.00230 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.57 on 38 degrees of freedom
Multiple R-squared:  0.7795,Adjusted R-squared:  0.7563
```

```
F-statistic: 33.58 on 4 and 38 DF,  p-value: 5.298e-12

> anova(model)
Analysis of Variance Table

Response: USRevenue
          Df Sum Sq Mean Sq F value    Pr(>F)
Budget     1  69874   69874  44.622 6.687e-08 ***
Opening    1 122680  122680  78.344 9.051e-11 ***
Theatres   1   1065    1065   0.680  0.414721
Ratings    1  16724   16724  10.680  0.002302 **
Residuals 38  59505    1566
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

(a) [4 marks] $F = 33.58$ and p-value$= 5.298 \times 10^{-12}$.

Using $\alpha = 0.05$, we conclude that there is very strong evidence that a model with all the predictors (Budget, Opening, Theatres, Ratings) is better than a model with just an intercept.

(b) [4 marks]

```
> anova(lm(USRevenue ~ Budget + Opening, data=Movies))

Analysis of Variance Table

Response: USRevenue
          Df Sum Sq Mean Sq F value    Pr(>F)
Budget     1  69874   69874  36.160 4.527e-07 ***
Opening    1 122680  122680  63.488 8.738e-10 ***
Residuals 40  77294    1932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

or

```
> anova(lm(USRevenue ~ Budget, data=Movies),
lm(USRevenue ~ Budget + Opening, data=Movies))

Analysis of Variance Table
```

```
Model 1: USRevenue ~ Budget
Model 2: USRevenue ~ Budget + Opening
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     41 199974
2     40  77294  1    122680 63.488 8.738e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test the hypothesis that in the model with Budget and Opening, $\beta_{Opening} = 0$ vs $\beta_{Opening} \neq 0$.

From the anova output $f = 63.488$ and the p-value$= 8.738 \times 10^{-10} << 0.05$, which indicates that there is very strong evidence to reject H_0.

We conclude that there is very strong evidence that a model with Budget and Opening is better than a model with just Budget.

(c) [4 marks]

```
> anova(lm(USRevenue ~ Budget, data=Movies),
lm(USRevenue ~ Budget + Opening + Theatres + Ratings, data=Movies))

Analysis of Variance Table

Model 1: USRevenue ~ Budget
Model 2: USRevenue ~ Budget + Opening + Theatres + Ratings
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     41 199974
2     38  59505  3    140469 29.901 4.227e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

In the full model, test the hypothesis that

$$\beta_{Opening} = \beta_{Theatres} = \beta_{Ratings} = 0$$

vs

$$\beta_{Opening} \neq 0 \text{ or } \beta_{Theatres} \neq 0 \text{ or } \beta_{Ratings} \neq 0$$

From the anova output, the value of the F-statistic is given by $f = 29.901$ and the p-value$= 4.227 \times -10$.

There is very strong evidence that a model with all the predictors is preferred over a model with Budget as the predictor, (p-value $<< 0.05$).

(d) [4 marks] This is the partial t-test, with $t = 1.497$ and $p = 0.14254 > 0.05$, which indicates that there is no evidence to reject the null hypothesis.

We conclude that there is no evidence that Theatres is related to USRevenue in the presence of Budget, Opening and Ratings.

(e) [2 marks] Obtain a 99% prediction interval for the `USRevenue` based on the model with all four predictors.

```
> newdata <- data.frame(Budget=30,Theatres=3065,Opening=19.83,
Ratings=7.6)
> predict(model,newdata,interval="predict",level=0.99)
       fit       lwr      upr
1 89.58642 -21.80826 200.9811
```

The 99% prediction interval for USRevenue is given by $(-21.80826, 200.9811)$.

(f) [2 marks] Upload R `summary` and `anova` outputs in one file.

Question 2 [20 marks]. Consider the general linear model

$$y = X\beta + \varepsilon, \quad \text{where } \mathbb{E}(\varepsilon) = 0 \text{ and } Var(\varepsilon) = \sigma^2 I_{n\times n},$$

and the transformation of ε given by $q = \beta + L\varepsilon$.

Part A. [7 marks]

A1. [3 marks]

$$\mathbb{E}(q) = \mathbb{E}(\beta + L\varepsilon) = \beta + L\mathbb{E}(\varepsilon) = \beta$$

and

$$\begin{aligned} Var(q) &= \mathbb{E}[(q - \mathbb{E}(q))(q - \mathbb{E}(q))^\top] = \mathbb{E}[L\varepsilon(L\varepsilon)^\top] \\ &= \mathbb{E}[L\varepsilon\varepsilon^\top L^\top] = L\mathbb{E}(\varepsilon\varepsilon^\top)L^\top = LVar(\varepsilon)L^\top \\ &= L(\sigma^2 I)L^\top = \sigma^2 LL^\top. \end{aligned}$$

A2. [2 marks]

$$\begin{aligned} \mathbb{E}(q^\top q) &= \mathbb{E}(q^\top I_{p\times p}\ q) \\ &= \mathbb{E}(q)^\top I_{p\times p}\ \mathbb{E}(q) + tr(I_{p\times p}Var(q)) \\ &= \beta^\top\beta + tr(I_{p\times p}\sigma^2 LL^\top) \\ &= \beta^\top\beta + \sigma^2 tr(LL^\top). \end{aligned}$$

A3. [2 marks] We assume the error term is multivariate normal $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then $q = \beta + L\varepsilon$, as a linear (more precisely, an affine) transformation of ε, is also multivariate normal,

$$q \sim \mathcal{N}(\beta, \sigma^2 LL^\top).$$

Part B. [13 marks]

B1. [1 mark] The least squares estimator in full-rank general linear model is given by $b = (X^\top X)^{-1}X^\top y$.
Substituting the linear model for y, we obtain

$$\begin{aligned} b &= (X^\top X)^{-1}X^\top(X\beta + \varepsilon) \\ &= (X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top\varepsilon \\ &= \beta + L\varepsilon, \end{aligned}$$

where $L = (X^\top X)^{-1}X^\top$.

B2. [4 marks] Recall the definition of the residual sum of squares

$$SS_{res} = (y - \widehat{y})^\top (y - \widehat{y}) = (y - Xb)^\top (y - Xb).$$

Simplifying the vector of residuals, we obtain

$$\begin{aligned} y - Xb &= X\beta + \varepsilon - Xb \\ &= X\beta + \varepsilon - X(\beta + L\varepsilon) \\ &= X\beta + \varepsilon - X\beta + XL\varepsilon \\ &= \varepsilon + XL\varepsilon \\ &= (I - XL)\varepsilon \end{aligned}$$

and

$$(y - Xb)^\top = \varepsilon^\top (I - XL)^\top.$$

We can show that the square $n \times n$ matrix $I - XL$ is symmetric and idempotent. Indeed,

$$\begin{aligned} (I - XL)^\top &= I^\top - L^\top X^\top \\ &= I - (X(X^\top X)^{-1})X^\top \\ &= I - XL \end{aligned}$$

and

$$\begin{aligned} (I - XL)(I - XL) &= I - XL - XL + XLXL \\ &= I - 2XL + X(X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top \\ &= I - 2XL + XL \\ &= I - XL. \end{aligned}$$

Using all the above, we get

$$\begin{aligned} SS_{res} &= (y - Xb)^\top (y - Xb) \\ &= \varepsilon^\top (I - XL)^\top (I - XL)\varepsilon \\ &= \varepsilon^\top (I - XL)\varepsilon. \end{aligned}$$

B3. [2 marks] We can apply the rule for computing the expectation of quadratic forms $\mathbb{E}(\varepsilon A \varepsilon^\top) = \mu^\top A \mu + tr(A\Sigma)$ for the case $A = I - XL$. Here $\mu = \mathbb{E}(\varepsilon) = 0$ and $\Sigma = Var(\varepsilon) = \sigma^2 I$.

$$\begin{aligned}\mathbb{E}[\varepsilon(I - XL)\varepsilon^\top] &= tr[(I - XL)\sigma^2 I] \\ &= \sigma^2 tr(I - XL) \\ &= \sigma^2[tr(I) - tr(XL)].\end{aligned}$$

In the above expression, I is $n \times n$ identity matrix. We know that the trace of a matrix is the sum of its diagonal elements. Therefore, $tr(I) = n$. On the other hand, $tr(XL) = tr(LX) = tr((X^\top X)^{-1}X^\top X) = tr(I_{p\times p}) = p$. Finally, we get

$$\mathbb{E}[\varepsilon(I - XL)\varepsilon^\top] = \sigma^2(n - p).$$

B4. [2 marks] Recall that the usual estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{(y - Xb)^\top(y - Xb)}{n - p} = \frac{SS_{res}}{n - p}.$$

By results of parts B2 and B3, we observe that

$$\widehat{\sigma}^2 = \frac{\varepsilon(I - XL)\varepsilon^\top}{n - p}$$

and

$$\mathbb{E}(\widehat{\sigma}^2) = \frac{\mathbb{E}[\varepsilon(I - XL)\varepsilon^\top]}{n - p} = \frac{\sigma^2(n - p)}{n - p} = \sigma^2.$$

We conclude that the estimator $\widehat{\sigma}^2$ is unbiased for σ^2.

B5. [4 marks] Assuming that the error term is multivariate normal random vector, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, it follows that the least squares estimator $b = \beta + L\varepsilon$ is also multivariate normal, $b \sim \mathcal{N}(\beta, \sigma^2 LL^\top)$, and hence $b_j = \beta_j + l_j^\top \varepsilon \sim \mathcal{N}(\beta_j, l_j^\top \sigma^2 l_j)$.
We can standardise b_j and obtain

$$\frac{b_j - \beta_j}{\sigma\sqrt{l_j^\top l_j}} \sim \mathcal{N}(0, 1).$$

Confidence interval for β_j coefficient can be derived based on the ratio

$$\frac{\frac{b_j - \beta_j}{\sigma\sqrt{l_j^\top l_j}}}{\sqrt{\frac{(n-p)\widehat{\sigma}^2}{\sigma^2}/(n - p)}} = \frac{b_j - \beta_j}{\widehat{\sigma}\sqrt{l_j^\top l_j}} \sim t_{n-p},$$

which follows the t distribution with $(n-p)$ degrees of freedom because it has the form $\frac{Z}{\sqrt{Y/\nu}}$, where $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_\nu$ and Z is independent of Y.
Therefore,

$$\mathbb{P}(-t_{\alpha/2,n-p} \leqslant \frac{b_j - \beta_j}{\widehat{\sigma}\sqrt{l_j^\top l_j}} \leqslant t_{\alpha/2,n-p}) = 1 - \alpha.$$

Consequently, a $100(1-\alpha)\%$ confidence interval for β_j is

$$b_j \pm t_{\alpha/2,n-p}\widehat{\sigma}\sqrt{l_j^\top l_j}.$$