1. [15 marks] A data set on gas consumption of cars includes the following variables for $n = 30$ cars:

y – gas consumption (in miles/gallon),
x_1 – engine size (cylinder displacement in cubic inches),
x_2 – engine horsepower,
x_3 – engine torqne,
x_4 – engine compression ratio,
x_5 – rear axle ratio.

The following two multiple linear regression models are fitted to the data using R:

Model 1:	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
Model 2:	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$

Use the provided R code and output to answer the following questions. For hypothesis testing questions, state the null and alternative hypotheses, the observed test statistics, the p-values and the conclusions. Use significance level $\alpha = 0.05$ for all tests.

(1a) [1] For Model 1, give the fitted/estimated model.

$\hat{Y} = 32.621 - 0.078 x_1 + 0.007 x_2 + 0.040 x_3$

(1b) [2] For Model 1, test for significance of regression.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_1: At least one of $\beta_1, \beta_2, \beta_3$ is not zero.

F-statistic: 28.83 on 3 and 26 df ; P-value $= 1.995 \times 10^{-8} < 0.05$

Reject H_0, regression is significant.

(1c) [2] For Model 1, use t test to assess the contribution of the engine size x_1.

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$t = -2.116$, P-value $= 0.044 < 0.05$

Reject H_0; Contribution from x_1 is significant.

(1d) [2] For Model 1, construct a 95% confidence interval for β_1.

$\hat{\beta}_1 \pm t_{0.025,26}\, S.e.(\hat{\beta}_1) \Rightarrow -0.078 \pm 2.056 \times 0.037$

or $(-0.154, -0.002)$

(1e) [1] For Model 1, what percentage of variation in y has been explained by the regression?

$$R^2 = 0.7688 = 76.88\%$$

(1f) [1] Using fitted Model 1, compute by hand the predicted gas consumption for a car with engine size $x_1 = 350$, horsepower $x_2 = 170$ and torqne $x_3 = 275$.

$$\hat{Y} = 32.621 - 0.078 \times 350 + 0.007 \times 170 + 0.040 \times 275 = 17.511$$

(1g) [4] For Model 2, use a partial F test to assess the (combined) contribution of engine compression ratio and rear axle ratio given all of the other regressors are included.

$H_0: \beta_4 = \beta_5 = 0$

H_1: At least one of β_4, β_5 is not zero.

$$F_{obs} = \frac{(263.31 - 245.49)/2}{245.49/24} = 0.871$$

$$\text{P-value} = P(F_{2,24} > 0.871) > P(F_{2,24} > 3.403) = 0.05$$

Do not reject H_0; Contribution from x_4 and x_5 not significant.

(1h) [2] For Model 2, suppose you wish to test $H_0 : \beta_2 + 2\beta_3 = \beta_4$. Give the reduced model under H_0 *without* using β_4. Write it using a similar format to that of Model 1 and Model 2 on page 2.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Under H_0: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (\beta_2 + 2\beta_3) x_4 + \beta_5 x_5 + \varepsilon$

$$= \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_4) + \beta_3 (x_3 + 2x_4) + \beta_5 x_5 + \varepsilon$$

2. [8 marks] In matrix form, the multiple linear regression model for the vector of responses $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is the random error, $\mathbf{X}$ is the fixed matrix of regressor variable values. The least squares estimate of regression parameter β is $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. The hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

(2a) [1] Express the fitted value $\hat{\mathbf{y}}$ in terms of $\mathbf{H}$ and $\mathbf{y}$.

$$\hat{\underline{Y}} = X\hat{\beta} = X(X^TX)^{-1}X^T\underline{Y} = H\underline{Y}$$

(2b) [3] Let $\mathbf{e}$ be the residual of the least squares fit. Show that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\epsilon$.

$$\begin{aligned}\underline{e} = \underline{Y} - \hat{\underline{Y}} &= \underline{Y} - H\underline{Y} \\ &= (I - H)\underline{Y} \\ &= (I - H)(X\underline{\beta} + \underline{\varepsilon}) \\ &= (X - HX)\underline{\beta} + (I - H)\underline{\varepsilon} \\ &= (X - X)\underline{\beta} + (I - H)\underline{\varepsilon} \\ &= (I - H)\underline{\varepsilon}\end{aligned}$$

(2c) [2] Find the mean of $\hat{\beta}$ and express the covariance matrix of $\hat{\beta}$ in terms of σ^2 and $\mathbf{X}$.

$$E(\hat{\underline{\beta}}) = (X^TX)^{-1}X^TE(\underline{Y}) = (X^TX)^{-1}X^T(X\underline{\beta}) = \underline{\beta}$$

$$\begin{aligned}V(\hat{\underline{\beta}}) &= (X^TX)^{-1}X^TV(\underline{Y})\left[(X^TX)^{-1}X^T\right]^T \\ &= (X^TX)^{-1}X^TV(\underline{\varepsilon})X(X^TX)^{-1} \qquad V(\underline{\varepsilon}) = \sigma^2 I \\ &= (X^TX)^{-1}X^T\sigma^2 I X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

(2d) [1] What is the distribution of $\hat{\beta}$.

$$N_p(\underline{\beta}, \sigma^2(X^TX)^{-1})$$

(2e) [1] What is the function $S(\beta)$ that $\hat{\beta}$ minimizes? Express $S(\beta)$ using $\mathbf{X}$, $\mathbf{y}$ and β.

$$S(\underline{\beta}) = \underline{\varepsilon}^T\underline{\varepsilon} = (\underline{Y} - X\underline{\beta})^T(\underline{Y} - X\underline{\beta})$$

3. [7 marks] Suppose $(X_1, X_2, X_3)^\top \sim N(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu}^\top = (1, 0, 2)$ and

$$\Sigma = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

(3a) [2] Let $Y_1 = 3X_1 - 2X_2 + X_3$. Find $E(Y_1)$ and $Var(Y_1)$.

$E(Y_1) = 3E(x_1) - 2E(x_2) + E(x_3) = 3\times1 - 2\times0 + 2 = 5$

$V(Y_1) = 9V(x_1) + 4V(x_2) + V(x_3) - 12\,\mathrm{Cov}(x_1, x_2) + 6\,\mathrm{Cov}(x_1, x_3) - 4\,\mathrm{Cov}(x_2, x_3)$

$= 9\times2 + 4\times1 + 5 - 12\times1 = 15$

(3b) [2] Find $Cov(X_1 - X_2, X_1 - X_3)$.

$= \mathrm{Cov}(x_1, x_1) + \mathrm{Cov}(x_1, -x_3) + \mathrm{Cov}(-x_2, x_1) + \mathrm{Cov}(-x_2, -x_3)$

$= V(x_1) - \mathrm{Cov}(x_1, x_3) - \mathrm{Cov}(x_1, x_2) + \mathrm{Cov}(x_2, x_3)$

$= 2 - 0 - 1 + 0$

$= 1$

(3c) [1] Give the distribution of Y_1 in (3a), including values of its parameters.

$N(5, \sigma^2 = 15)$

(3d) [1] What is the distribution of $Y_2 = X_2^2 + \frac{1}{5}(X_3 - 2)^2$?

$x_2^2 = z_1^2, \quad \frac{1}{5}(x_3 - 2)^2 = z_2^2$

$\mathrm{cov}(x_2, x_3) = 0 \Rightarrow x_2 \perp x_3 \Rightarrow z_1^2 \perp z_2^2$

$\therefore Y_2 = z_1^2 + z_2^2 = \chi_2^2$

(3e) [1] What is the distribution of $Y_3 = (X_1 + X_2 - 1)^2/(X_3 - 2)^2$?

$E(x_1 + x_2 - 1) = 0, \quad V(x_1 + x_2 - 1) = V(x_1) + V(x_2) + 2\,\mathrm{Cov}(x_1, x_2) = 5$

$\therefore x_1 + x_2 - 1 \sim N(0, 5)$

$x_3 \sim N(2, 5)$ and $x_3 \perp (x_1 - x_2 - 1)$

$\therefore Y_3 = \dfrac{(x_1 + x_2 - 1)^2/5}{(x_3 - 2)^2/5} = \dfrac{\chi_1^2/1}{\chi_1^2/1} \sim F_{1,1}$

THE END

Problem 1 R output

```
> #### [1] various t and F critical values #####
>
> qt(0.975,  26)
[1] 2.055529
> qf(0.950,2,24)
[1] 3.402826


> #### [2] model 1 related R commands and output #####
>
> mdl1=lm(y~x1+x2+x3)
> summary(mdl1)

Call: lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3724 -2.0962  0.1354  1.6751  6.6286

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.620594   2.050397  15.909 6.45e-15 ***
x1          -0.077808   0.036767  -2.116   0.0441 *
x2           0.007284   0.053111   0.137   0.8920
x3           0.039820   0.065881   0.604   0.5508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.182 on 26 degrees of freedom
Multiple R-squared: 0.7688,     Adjusted R-squared: 0.7422
F-statistic: 28.83 on 3 and 26 DF,  p-value: 1.995e-08

> anova.lm(mdl1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 866.50  866.50 85.5590 1.048e-09 ***
x2         1   5.60    5.60  0.5525    0.4639
x3         1   3.70    3.70  0.3653    0.5508
Residuals 26 263.31   10.13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> vif(mdl1)
      x1       x2       x3
52.30027 16.10358 85.76325

> ##### [3] model 2 related R commands and outout #####
>
> mdl2=lm(y~x1+x2+x3+x4+x5)
> anova.lm(mdl2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 866.50  866.50 84.7132 2.421e-09 ***
x2         1   5.60    5.60  0.5471    0.4667
x3         1   3.70    3.70  0.3617    0.5532
x4         1  15.59   15.59  1.5243    0.2289
x5         1   2.24    2.24  0.2186    0.6443
Residuals 24 245.49   10.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```