1. An experiment is conducted to estimate the annual demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained:

<table>
<thead>
<tr>
<th>Cars sold ($\times 10^3$)</th>
<th>Cost ($$k$)</th>
<th>Unemployment rate (%)</th>
<th>Interest rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>7.2</td>
<td>8.7</td>
<td>5.5</td>
</tr>
<tr>
<td>5.9</td>
<td>10.0</td>
<td>9.4</td>
<td>4.4</td>
</tr>
<tr>
<td>6.5</td>
<td>9.0</td>
<td>10.0</td>
<td>4.0</td>
</tr>
<tr>
<td>5.9</td>
<td>5.5</td>
<td>9.0</td>
<td>7.0</td>
</tr>
<tr>
<td>8.0</td>
<td>9.0</td>
<td>12.0</td>
<td>5.0</td>
</tr>
<tr>
<td>9.0</td>
<td>9.8</td>
<td>11.0</td>
<td>6.2</td>
</tr>
<tr>
<td>10.0</td>
<td>14.5</td>
<td>12.0</td>
<td>5.8</td>
</tr>
<tr>
<td>10.8</td>
<td>8.0</td>
<td>13.7</td>
<td>3.9</td>
</tr>
</tbody>
</table>

(a) Write down a linear model for cars sold in terms of the other variables. Calculate the least squares estimator, $b$.

**Solution [2 marks]:** Our model is $y = X\beta + \varepsilon$ where

\[
y = \begin{pmatrix}(5.5, 5.9, \ldots, 10.8)\end{pmatrix}^T
\]

\[
X = \begin{bmatrix} 1 & 7.2 & 8.7 & 5.5 \\ 1 & 10.0 & 9.4 & 4.4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 8.0 & 13.7 & 3.9 \end{bmatrix}
\]

\[
\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T
\]

```
> n <- 8
> p <- 4
> X <- matrix(c(rep(1,n),
+ 7.2,10,9,5.5,9,9.8,14.5,8,
+ 8.7,9.4,10,9,12,11,12,13.7,
+ 5.5,4.4,4,7,5.6,2,5.8,3.9),n,p)
> y <- as.vector(c(5.5,5.9,6.5,5.9,8,9,10,10.8))
> (b <- solve(t(X) %*% X) %*% t(X) %*% y)

[,1]
[1,] -7.4044796
[2,] -0.0549251
[3,] 1.1174846
[4,] 0.3861206
```

(b) Calculate the variance of the least squares estimator. You may express this in terms of the variance of the errors, $\sigma^2$.

**Solution [2 marks]:** The variance is $\sigma^2$ multiplied by

```
> round(solve(t(X) %*% X), 3)

[1,] 13.497 -0.055 -0.699 -1.030
```
(c) A confidence interval is found for the average number of $10,000 cars sold in a year which has unemployment rate 8.5% and interest rate 7%, and is calculated to be (4062, 7947). Find the confidence level used.

**Solution [4 marks]:** Let $\alpha$ be the level used. Then

$$\left( x^* \right)^T b - t_{\alpha/2} s^2 \sqrt{\left( x^* \right)^T (X^T X)^{-1} x^*} = 4.062$$

$$t_{\alpha/2} = \frac{(x^*)^T b - 4.062}{s^2 \sqrt{\left( x^* \right)^T (X^T X)^{-1} x^*}}$$

```r
> s2 <- sum((y - X %*% b)^2)/(n-p)
> xst <- as.vector(c(1,10,8.5,7))
> talph <- (t(xst) %*% b - 4.062) / sqrt(s2 * t(xst) %*% solve(t(X) %*% X) %*% xst)
> 1-2*pt(talph, n-p, lower.tail=FALSE)
[,1]
[1,] 0.9800019
```

The confidence level is 98%.

(d) The joint 95% confidence region for all the parameters is an ellipsoid. Find the centre of this ellipsoid.

**Solution [2 marks]:** The centre is the least squares estimator for the parameters, $b$, which was found in (a).

(e) Test for model relevance ($H_0 : \beta = 0$).

**Solution [2 marks]:**

```r
> SSReg <- t(y) %*% X %*% solve(t(X) %*% X) %*% t(X) %*% y
> SSRes <- s2*(n-p)
> Fstat <- (SSReg/p)/(SSRes/(n-p))
> Fstat
[,1]
[1,] 317.4027
> pf(Fstat, p, n-p, lower.tail = FALSE)
[,1]
[1,] 2.952959e-05
```

The model is clearly relevant.

(f) Perform one step of backward elimination on the full model. (You do not need to check the intercept term.)

**Solution [4 marks]:**

```r
> X1 <- X[, -2]
> R1 <- t(y) %*% X %*% solve(t(X1) %*% X1) %*% t(X1) %*% y
> X2 <- X[, -3]
> R2 <- t(y) %*% X %*% solve(t(X2) %*% X2) %*% t(X2) %*% y
> X3 <- X[, -4]
> R3 <- t(y) %*% X %*% solve(t(X3) %*% X3) %*% t(X3) %*% y
> pf((SSReg-R1)/(SSRes/(n-p)), 1, n-p, lower.tail=FALSE)
[,1]
[1,] 0.2871663
> pf((SSReg-R2)/(SSRes/(n-p)), 1, n-p, lower.tail=FALSE)
[,1]
[1,] 0.002055812
> pf((SSReg-R3)/(SSRes/(n-p)), 1, n-p, lower.tail=FALSE)
```
We should remove the cost parameter.

(g) Calculate the difference in Akaike’s information criterion between the full model and the model that results from question 1f.

Solution [2 marks]:

\[
AIC_{full} = n \times \log(\frac{SS_{Res}}{n}) + 2p \\
X_{new} = X[,1,3,4]] \\
b_{new} = solve(t(X_{new}) \times X_{new}) \times t(X_{new}) \times y \\
SS_{Res_{new}} = \sum((y - X_{new} \times b_{new})^2) \\
AIC_{new} = n \times \log(\frac{SS_{Res_{new}}}{n}) + 2p \\
AIC_{full} - AIC_{new}
\]

[1] -0.5549902

(h) Use \texttt{lm}, \texttt{predict} and \texttt{anova} to check your answers above. Use R’s diagnostic plots to check that your model assumptions are justified.

Solution [4 marks]:

\[
cars <- data.frame(sales=y, cost=X[,2], unemployment=X[,3], interest=X[,4]) \\
model <- lm(sales ~ cost + unemployment + interest, data=cars) \\
\]

# (a), (d)

\[
\text{model$coefficient} \\
(\text{Intercept}) \quad \text{cost} \quad \text{unemployment} \quad \text{interest} \\
-7.4044796 \quad 0.1207646 \quad 1.1174846 \quad 0.3861206
\]

# (c)

\[
predict(model, newdata=list(cost=10, unemployment=8.5, interest=7), \\
+ interval="confidence", level=0.98) \\
\]

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.00463</td>
<td>4.062057 7.947202</td>
</tr>
</tbody>
</table>

# (a)

\[
emptymodel <- lm(sales ~ 0, data=cars) \\
anova(emptymodel, model)
\]

Analysis of Variance Table

Model 1: sales ~ 0
Model 2: sales ~ cost + unemployment + interest

<table>
<thead>
<tr>
<th>Df</th>
<th>RSS</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>503.76</td>
<td>503.76</td>
<td>1</td>
<td>8.404796</td>
</tr>
<tr>
<td>4</td>
<td>1.58</td>
<td>502.18</td>
<td>4</td>
<td>150.4846</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 '.' 0.1 ' ' 1

# (f), (g)

\[
drop1(model, scope=-., test="F")
\]

Single term deletions

Model:

\[
sales ~ cost + unemployment + interest
\]

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>1.5821</td>
<td>-4.9653</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cost</td>
<td>1</td>
<td>0.5953</td>
<td>2.1775</td>
<td>-4.4103</td>
<td>0.15051</td>
</tr>
<tr>
<td>unemployment</td>
<td>1</td>
<td>20.0562</td>
<td>21.6383</td>
<td>13.5602</td>
<td>50.7062</td>
</tr>
<tr>
<td>interest</td>
<td>1</td>
<td>1.1014</td>
<td>2.6836</td>
<td>-2.7384</td>
<td>2.7846</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 '.' 0.1 ' ' 1
> -4.9653 + 4.4103
[1] -0.555
> # (h)
> newmodel <- lm(sales ~ unemployment + interest, data=cars)
> par(mfrow=c(2,2))
> plot(newmodel, which=1)
> plot(newmodel, which=2)
> plot(newmodel, which=3)
> plot(newmodel, which=5)

There are no evident outliers, and no notable patterns in the residual plots. Given we only have eight observations, the Q-Q plot is also OK.

2. For this question we use a dataset distributed with R. Create the dataframe `statedata` and read about it using the following commands.

```r
data(state)
helper(state)
statedata <- data.frame(state.x77, row.names=state.abb, check.names=TRUE)
str(statedata)
```

Find a parsimonious linear model for life expectancy using the other variables. First plot the data and consider the need for variable transformations. Next perform model selection using forward selection, backwards elimination, and stepwise selection with the AIC. For the model obtained by stepwise selection, perform diagnostics to justify using a linear model.

In your answer include the R code and output you used.
Solution [10 marks]: We first plot the data. Looking at the life expectancy against the other variables, there is evidence of a linear relationship with income, illiteracy, murder, and high school grad. There is no obvious relationship with population, frost and area. Population and area both have distributions skewed to right. log(population) and log(area) would be less skewed and might fit better with the other variables. There is a hint of heteroskedasticity in income, and a whiff of non-linearity in high school grad, but neither enough for immediate concern.

> pairs(statedata)
> statedata$logPopulation <- log(statedata$Population)
> statedata$logArea <- log(statedata$Area)

Forward selection: Using forward selection with a 95% significance criterion, we add variables in the order: murder, high school grad, log(population), frost. Here is the first step:

> model0 <- lm(Life.Exp ~ 1, data=statedata)
> add1(model0, scope= ~ . + Population + Income + Illiteracy + Murder + HS.Grad + + Frost + Area + logPopulation + logArea, test="F")

Single term additions

Model:
Life.Exp ~ 1

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>88.299</td>
<td>30.435</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>1</td>
<td>0.409</td>
<td>87.890</td>
<td>32.203</td>
<td>0.2233</td>
</tr>
<tr>
<td>Income</td>
<td>1</td>
<td>10.223</td>
<td>78.076</td>
<td>6.2847</td>
<td>0.01562 *</td>
</tr>
<tr>
<td>Illiteracy</td>
<td>1</td>
<td>30.578</td>
<td>57.721</td>
<td>11.179</td>
<td>25.4289</td>
</tr>
<tr>
<td>Murder</td>
<td>1</td>
<td>53.838</td>
<td>34.461</td>
<td>14.609</td>
<td>74.9887</td>
</tr>
</tbody>
</table>
If \( \log(\text{population}) \) and \( \log(\text{area}) \) are not considered, then we add the following variables: murder, high school grad, frost. Population is the most significant of the remaining variables, but just misses being added to the model using 95% significance, so taking the log of the population did help.

**Backward elimination:** Using backward elimination with a 95% significance criterion, we remove variables in the order: population, income, area, illiteracy, \( \log(\text{area}) \). The final model has variables murder, high school grad, frost and \( \log(\text{population}) \). That is, it is the same as the model found using forward selection. Here is the first step

```R
> model0 <- lm(Life.Exp ~ ., data=statedata)
> drop1(model0, scope = ~ ., test="F")
```

**Stepwise selection:** Starting with a full model or just an intercept, stepwise selection using the AIC gives the same model as before (murder, high school grad and \( \log(\text{population}) \) and frost). The relevant commands are

```R
fullmodel <- lm(Life.Exp ~ ., data=statedata)
step(fullmodel, scope="~.")
model0 <- lm(Life.Exp ~ 1, data=statedata)
step(model0, scope="~. + Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area + \log(\text{Population}) + \log(\text{Area})
```

If \( \log(\text{population}) \) and \( \log(\text{area}) \) are not considered, then we remove the following variables: area, illiteracy, income, population. The resulting model has variables murder, high school grad and frost, which is what we obtained using forward selection with the same set of variables.
If we do not consider log(population) and log(area) then the final model has variables population, murder, high school grad and frost (whether you start with the full model or just an intercept). Unlike the forward and backward schemes, this model includes population.

Diagnostics: We check that our model meets our assumptions using diagnostic plots. Everything looks good here.

Hawaii and Maine give the worst fits, though the range of residuals is fine given the number of observations. Curiously, Hawaii and Maine are also the most westerly and the most easterly states, so outliers geographically if not statistically.

```r
> model <- lm(Life.Exp ~ Murder + HS.Grad + logPopulation + Frost, data = statedata)
> opar <- par(mfrow=c(2,2))
> plot(model, which=1)
> plot(model, which=2)
> plot(model, which=3)
> plot(model, which=5)
> par <- opar
```

Including log(population) in the model instead of population degrades the fit, with the residuals looking a bit less normal. If we include neither log(population) nor population in the model then the residuals are less normal again, though they still show no sign of trend or heteroskedasticity, and in this case plots are still acceptable.

3. The **QR decomposition** of an \( m \times n \) matrix \( X \), \( m \geq n \), is an expression \( X = QR \) where \( Q \) is an \( m \times m \) orthogonal matrix and \( R \) is an \( m \times n \) upper-triangular matrix. (So the last \( m - n \) rows of \( R \) are all 0.) The **QR decomposition** can be used to find the eigenvalues of \( X \), and is used to solve
the normal equations $X^T X \beta = X^T y$. Put $z = R \beta$ then

\[
X^T X \beta = R^T Q R \beta = R^T z
\]

Thus to solve the normal equations we can first solve $R^T z = X^T y$, which is easy because $R^T$ is lower triangular, then solve $R \beta = z$, which is easy because $R$ is upper triangular.

(a) Show that if $R = \begin{bmatrix} R_1 & r_1 \\ 0 & r_n \end{bmatrix}$ is an $n \times n$ upper triangular matrix of full rank, and

\[
\begin{bmatrix} R_1 & r_1 \\ 0 & r_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_n \end{bmatrix} = \begin{bmatrix} z_1 \\ z_n \end{bmatrix}
\]

then $x_n = z_n/r_n$ and $R_1 x_1 = z_1 - x_n r_1$.

Solution [1 mark]: Multiplying out the LHS we get

\[
\begin{bmatrix} R_1 x_1 + r_1 x_n \\ r_n x_n \end{bmatrix} = \begin{bmatrix} z_1 \\ z_n \end{bmatrix}
\]

We now just equate the bottom and top parts on each side.

(b) Write an R function to solve $Rx = z$, assuming that $R$ is an $n \times n$ upper triangular matrix of full rank. You may not use the function solve.

One way to do this is using a recursive algorithm.

Demonstrate that your function works, by showing that it gives the same output as the following:

```r
A <- matrix(c(1,0,0, 1,2,0, 3,3,1), 3, 3)
b <- c(3, 4, 2)
solve(A, b)
```

Solution [9 marks]: Here is a recursive solution; one could also use a loop.

```r
> upper_tri <- function(A, b) {
+   # solve A %*% x == b assuming A is an n*n upper triangular matrix of full rank
+   n <- sqrt(length(A))
+   if (n == 1) {
+     return(b/A)
+   } else {
+     x_n <- b[n]/A[n,n]
+     x_1 <- upper_tri(A[-n,-n], b[-n] - A[-n,n]*x_n)
+     return(c(x_1, x_n))
+   }
+ }
> A <- matrix(c(1,0,0, 1,2,0, 3,3,1), 3, 3)
> b <- c(3,4,2)
> solve(A, b)
[1] -2 -1  2
> upper_tri(A, b)
[1] -2 -1  2
```