Tutorial Problems

Question 1

A researcher is interested in whether a teenager's income (measured in UK pound per week) can be used to predict the amount they will gamble (gambling expenditure measured in UK pound per year), at least for the teenagers who do regularly gamble. The dataset contains the information of 36 teenagers who spend at least 1 pound on gambling per week. Below are some summary statistics for the two variables in the dataset.

	Mean	Standard deviation	Correlation
Gamble (Y)	25.151	33.985	0.64
Income (X)	4.968	3.807	

Based only on given information, complete the summary table of the simple linear regression model between gamble and income. Noting that for X data, the sample standard deviation (as listed in the table above) is equal to $1/(n-1)s_x$, with $s_x = \sqrt{S_{xx}} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Similar calculation holds for Y data.

```
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  ?          7.303      ?       0.65
income       ?          1.173      ?    2.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.42 on ? degrees of freedom
Multiple R-squared:  ?,Adjusted R-squared:  0.3957
F-statistic: ? on ? and ? DF,  p-value: ?
```

Question 2

A vector of random variables $X = (X_1, X_2, X_3)^\top$ has covariance matrix

$$\Sigma = \begin{pmatrix} 9 & -4 & 1 \\ -4 & 25 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

(a) Find the correlation coefficient between X_1 and X_2. (Assumed knowledge)

(b) Find the variance of $Y = -2X_1 + 3X_2$. Write Y as $a^\top X$ and show that the variance is $a^\top \Sigma a$.

(c) What is the standard deviation of X_3? (Assumed knowledge)

Computer Problems

For the following question, we will continue to use the `olympic.txt` dataset that consists of the winning heights or distances (in inches) for the High Jump, Discus and Long Jump events at the Olympics up to 1996 from the last week's tutorial.

Question 1

(a) (From Tutorial 2) Read and store the `olympic.txt` dataset in R as the data frame `olympic`. Create a new data frame `olympicMetric` that has measurements in metres, by using the conversion 1 m = 39.3701 inches, and the full year (e.g. 1900 rather than 0). Show the first 6 rows of the `olympicMetric`.

(b) (From Tutorial 2) Fit the simple linear regression model with HighJump being the response and LongJump to be the covariate, and obtain the `summary` output.

```
olympicLm <- lm(HighJump ~ LongJump, data = olympicMetric)
```

For parts (c)-(f), we will conduct the F-test for the null hypothesis $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ using the general full-reduced model approach.

(c) What are the models under $H_0 : \beta_0 = 0$ and the models under H_1?

(d) Fit the model under H_0 in R using

```
olympic_H0 <- lm(HighJump ~ -1 + LongJump, data = olympicMetric)
```

and obtain the corresponding SSE of this model.

(e) From the two fitted models, obtain the F-statistic manually using the formula

$$F = \frac{(\text{SSE}(H_0) - \text{SSE}(H_1))/(df(H_0) - df(H_1))}{\text{SSE}(H_1)/df(H_1)}$$

Compute the corresponding p-value of the test. Verify your calculations with

```
anova(olympic_H0, olympicLm)
```

(f) Verify the relationship between the computed F-statistic and the t-statistic associated with the intercept in the summary table of `olympicLm`.

(g) Find point estimates for missing values of HighJump. Is it more appropriate to construct prediction intervals or confidence intervals for missing values of HighJump? Explain your answer.

(h) Check the model assumptions using the graphical diagnostic plots from the lectures.