Problem 1

(A) The rank is 4.

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(B) $X^T X\beta = X^T y$

$$\begin{bmatrix} 3 & 2 & 2 & 2 & 1 & 1 \\ 2 & 3 & 2 & 1 & 2 & 1 \\ 2 & 2 & 3 & 1 & 1 & 2 \\ 2 & 1 & 1 & 3 & 2 & 2 \\ 1 & 2 & 1 & 2 & 3 & 2 \\ 1 & 1 & 2 & 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix}$$

This has infinitely many solutions because there are four distinct equations, but there are six unknowns. There are too many unknown variables, and not enough distinct equations to find distinct solutions for each unknown variable.

(C) $\alpha_1 - \alpha_2 = [0, 1, -1, 0, 0, 0]\beta$

$$c^T = [0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0]$$

$$c = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \epsilon\, C(X^T)$$

Yes, $\alpha_1 - \alpha_2$ is estimable.

(D) $\beta_1 - 2\beta_2 + \beta_3 = [0 \quad 0 \quad 0 \quad 1 \quad -2 \quad 1]\beta$

$$c^T = [0 \quad 0 \quad 0 \quad 1 \quad -2 \quad 1]$$

$$c = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} - 2\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \epsilon\, C(X^T)$$

Yes, $\beta_1 - 2\beta_2 + \beta_3$ is estimable.

(E)

```
> library(estimability)
>
> #Define model matrix
> X <- cbind(rep(1,6),
+            c(rep(1, 3),rep(0,3)),
+            c(rep(0, 3),rep(1,3)),
+            c(1,0,0,1,0,0),
+            c(0,1,0,0,1,0),
+            c(0,0,1,0,0,1))
>
> #coefficient vectors
> cvec1 <- c(0, 1, -1, 0, 0, 0)
> cvec2 <- c(0, 0, 0, 1, -2, 1)
> nb <- nonest.basis(X)
>
> #checking estimability
> is.estble(cvec1,nb)
[1] TRUE
> is.estble(cvec2,nb)
[1] TRUE
```

Problem 2

(A)

```
> #Fit data from teengamb with gamble as response variable
> out <- lm(gamble ~ ., data = teengamb)
> out$coefficients
 (Intercept)          sex       status       income       verbal
 22.55565063 -22.11833009   0.05223384   4.96197922  -2.95949350
```

```
> summary(out)

Call:
lm(formula = gamble ~ ., data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-51.082 -11.320  -1.451   9.452  94.252

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.55565   17.19680   1.312   0.1968
sex         -22.11833    8.21111  -2.694   0.0101 *
status        0.05223    0.28111   0.186   0.8535
income        4.96198    1.02539   4.839 1.79e-05 ***
verbal       -2.95949    2.17215  -1.362   0.1803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267,	Adjusted R-squared:  0.4816
F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

Approximately 52.67% of the variation in gamble is accounted for by the other predictors.

(B) Largest residual was 94.2522174, which is case number 24.

```
> max(out$residuals)
[1] 94.25222
> out$residuals
          1           2           3           4           5           6           7           8
 10.6507430   9.3711318   5.4630298 -17.4957487  29.5194692  -2.9846919  -7.0242994 -12.3060734
          9          10          11          12          13          14          15          16
  6.8496267 -10.3329505   1.5934936  -3.0958161   0.1172839   9.5331344   2.8488167  17.2107726
         17          18          19          20          21          22          23          24
-25.2627227 -27.7998544  13.1446553 -15.9510624 -16.0041386  -9.5801478 -27.2711657  94.2522174
         25          26          27          28          29          30          31          32
  0.6993361  -9.1670510 -25.8747696  -8.7455549  -6.8803097 -19.8090866  10.8793766  15.0599340
         33          34          35          36          37          38          39          40
 11.7462296  -3.5932770 -14.4016736  45.6051264  20.5472529  11.2429290 -51.0824078   8.8669438
         41          42          43          44          45          46          47
 -1.4513921  -3.8361619  -4.3831786 -14.8940753   5.4506347   1.4092321   7.1662399
```

(C) Mean is approximately 0. Median residual is -1.451392.

```
> mean(out$residuals)
[1] -1.556914e-16
> median(out$residuals)
[1] -1.451392
```

(D) Approximately 0.

```
> cor(out$residuals,out$fitted.values)
[1] -6.215823e-17
```

(E) Approximately 0.

```
> cor(out$residuals,teengamb$income)
[1] 3.247058e-17
```

(F) The difference between the predicted expenditure on gambling for a male compared to a female is approximately 25.90921. The mean value for males was approximately 29.775, whereas the mean value for females was 3.866.

```
> female_data <- teengamb[teengamb$sex == 1, ]
> male_data <- teengamb[teengamb$sex == 0, ]
>
> male_out <- lm(gamble ~ ., data = male_data)
> female_out <- lm(gamble ~ ., data = female_data)
>
> mean_males <- mean(male_out$fitted.values)
> mean_females <- mean(female_out$fitted.values)
>
> diff <- mean_males - mean_females
> diff
[1] 25.90921
```

Problem 3

```
> out_wages_1 <- lm(wage ~ educ + exper, data = uswages)
> out_wages_1$coefficients
(Intercept)        educ       exper
-242.799412   51.175268    9.774767
```

For the model with wages as the response variable and years of education and experience as the explanatory variables, we found the model below, where x_1 is the years of education and x_2 is the number of years of experience.

$$\hat{y} = -242.80 + 51.18x_1 + 9.77x_2$$

The regression coefficient for years of education is approximately 51.18, which means that the wage would increase on average by 51.18 for every additional year of education.

```
> out_wages_2 <- lm(log(wage) ~ educ + exper, data = uswages)
> out_wages_2$coefficients
(Intercept)        educ       exper
 4.65031905  0.09050628  0.01807855
```

For the model with log(wages) as the response variable and years of education and experience as the explanatory variables, we found the model below, where x_1 is the years of education and x_2 is the number of years of experience.

$$\widehat{\ln(y)} = 4.6503 + 0.0905x_1 + 0.0181x_2$$

When solved for y becomes:

$$\hat{y} = 104.62(1.0947)^{x_1}(1.0182)^{x_2}$$

The regression coefficient for years of education is approximately 0.0905, which becomes a multiplicative factor of 1.0947. This means that the wage will increase on average by 9.47% for each additional year of education.

9.47% growth per year sounds more natural, because you are growing at a percentage based on your previous year of education instead of simply adding a flat rate per each additional year.

Appendix (R Code)

```
######################################
# Author: Evan Whitfield
# Date Last Edit: 1-24-25
# Purpose: To answer problems for ST503 HW2
######################################
#
# Problem 1
#
##################################
library(estimability)

#Define model matrix
X <- cbind(rep(1,6),
        c(rep(1, 3),rep(0,3)),
        c(rep(0, 3),rep(1,3)),
        c(1,0,0,1,0,0),
        c(0,1,0,0,1,0),
        c(0,0,1,0,0,1))

#coefficient vectors
cvec1 <- c(0, 1, -1, 0, 0, 0)
cvec2 <- c(0, 0, 0, 1, -2, 1)
nb <- nonest.basis(X)

#checking estimability
is.estble(cvec1,nb)
is.estble(cvec2,nb)


##################################
#
# Problem 2
#
##################################

library(faraway)

#Fit data from teengamb with gamble as response variable
out <- lm(gamble ~ ., data = teengamb)
out$coefficients
summary(out)

#Determining statistics for the residuals
mean(out$residuals)
```

```
median(out$residuals)
max(out$residuals)
out$residuals

#Finding correlation between the residuals and the fitted values
cor(out$residuals,out$fitted.values)

#Finding correlation between residuals and income variable
cor(out$residuals,teengamb$income)

#Sub-setting the data based on gender
female_data <- teengamb[teengamb$sex == 1, ]
male_data <- teengamb[teengamb$sex == 0, ]

#Determing linear model for each gender
male_out <- lm(gamble ~ ., data = male_data)
female_out <- lm(gamble ~ ., data = female_data)

#Calculating mean(expected) fitted value
mean_males <- mean(male_out$fitted.values)
mean_females <- mean(female_out$fitted.values)

#calculating the difference beween the expected values of each gender
diff <- mean_males - mean_females
diff

##############################################
#
# Problem 3
#
##############################################

out_wages_1 <- lm(wage ~ educ + exper, data = uswages)
out_wages_1$coefficients

out_wages_2 <- lm(log(wage) ~ educ + exper, data = uswages)
out_wages_2$coefficients
```