MATH 8050 Homework 4

Due April 5 at 11:59pm

Instructions: For the problems that follow, please show all of your work. For problems that you solve using software, please provide the code you used either with the problem or in an appendix at the end of your homework document. To submit your answers for homework, please upload a **single** PDF to Canvas.

Problem 1

We will study the `ais` data set, which contains several variables related to body and blood composition of athletes in Australia. To load this data into R, run the following (you might need to install the `DAAG` package):

```
library(DAAG)
data("ais")
```

(a) Fit a linear model with lean body mass (`lbm`) as the response variable and `sport` as a predictor. Print a summary of the model

(b) Plot the boxplots for `lbm` according to `sport`.

(c) How do you interpret the results of the F-test provided in the model summary for this model?

(d) How do you interpret the coefficient estimate for Field (track and field)? What about the coefficient for Tennis?

(e) Use model contrasts to obtain a point estimate and a 90% confidence interval for the difference in lean body mass between `Tennis` and `W_Polo` (water polo) athletes.

(f) Add the `ht` (height) variable to the model. How do you interpret the coefficient estimate for `ht`? How do you interpret the coefficient estimate for `Field` now?

Problem 2

Consider the `UN11` data set in the `alr4` package in R (you will probably need to install the `alr4` package). This data set contains $n = 237$ observations with 6 total variables: `region`, `group`, `fertility`, `ppgdp`, `lifeExpF`, and `pctUrban`. To load this data:

```
library(alr4)
data("UN11")
```

(a) Fit a model with `fertility` as the response variable and `ppgdp`, `lifeExpF`, and `pctUrban` as predictor variables.

(b) Make a plot of the fitted values versus the residuals for the model you just fit. Are there any trends? Do you think this linear model is appropriate for the data?

(c) Plot the scatterplot pairs for each of `fertility`, `ppgdp`, `lifeExpF`, and `pctUrban`. What do you notice about the relationships between the variables?

(d) Fit the model with `log(ppgdp)` instead, and interpret the coefficient estimate for `log(ppgdp)`. Plot the new scatterplot pairs with `fertility`, `log(ppgdp)`, `lifeExpF`, and `pctUrban`. Also provide an updated model fit and residual plot. Do you think the model fit is more appropriate now?

(e) Now fit the model with both `log(fertility)` and `log(ppgdp)`, and interpret the coefficient estimate for `log(ppgdp)`. Plot the new scatterplot pairs with `log(fertility)`, `log(ppgdp)`, `lifeExpF`, and `pctUrban`. Also provide an updated residual plot. Do you think the model fit is more appropriate now?

Problem 3

Consider the usual multiple linear regression model with $p-1$ predictors,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \text{ for all i.}$$

We have seen that this model can be written in vector/matrix form as:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of all observations of the response and $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients $\beta_0, \ldots, \beta_p - 1$.

(a) Using vector and matrix notation, derive the expectation of the vector of residuals $\underline{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, where $\hat{\boldsymbol{Y}}$ is the $n \times 1$ vector of all fitted values from the regression model.

Use the following facts to answer the next two questions:

$$Var(A\boldsymbol{Y}) = AVar(\boldsymbol{Y})A, \text{ where } A \text{ is some matrix of constants.}$$

$$Cov(A\boldsymbol{Y}, B\boldsymbol{Y}) = AVar(\boldsymbol{Y})B^T, \text{ where } A \text{ and } B \text{ are matrices of constants.}$$

$Var(\boldsymbol{Y}) = Cov(\boldsymbol{Y}, \boldsymbol{Y})$ (meaning that covariance between a vector and itself is just the variance of that vector)

(b) Using vector and matrix notation, derive the variance of the vector of residuals $\underline{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

(c) Find the distribution of $B\boldsymbol{Y} + \underline{c}$, where $\underline{c} = c \times \mathbf{1}_n$ and c is some constant.

Problem 4

Consider the `swiss` data set contained in R, which has $n = 47$ observations of socio-economic factors in Swiss provinces. The data has 6 variables: `Fertility`, `Agriculture`, `Examination`, `Education`, `Catholic`, and `Infant.Mortality`. To learn more about these variables and their units, run `help("swiss")` in your R console.

(a) Perform variable selection on this data, with `Fertility` as the response variable and all others as predictor variables. Perform selection using forward, backward, and stepwise approaches with AIC as the selection criterion.

(b) Answer the following questions based on your findings in part (a): (i) Do the selected models from each of these approaches agree? (ii) Which variables do you think best explain/model `Fertility` in this data? Explain your reasoning.

MATH 8050 - HW4

Parampreet Singh

2024-04-05

DATA ANALYSIS HW4

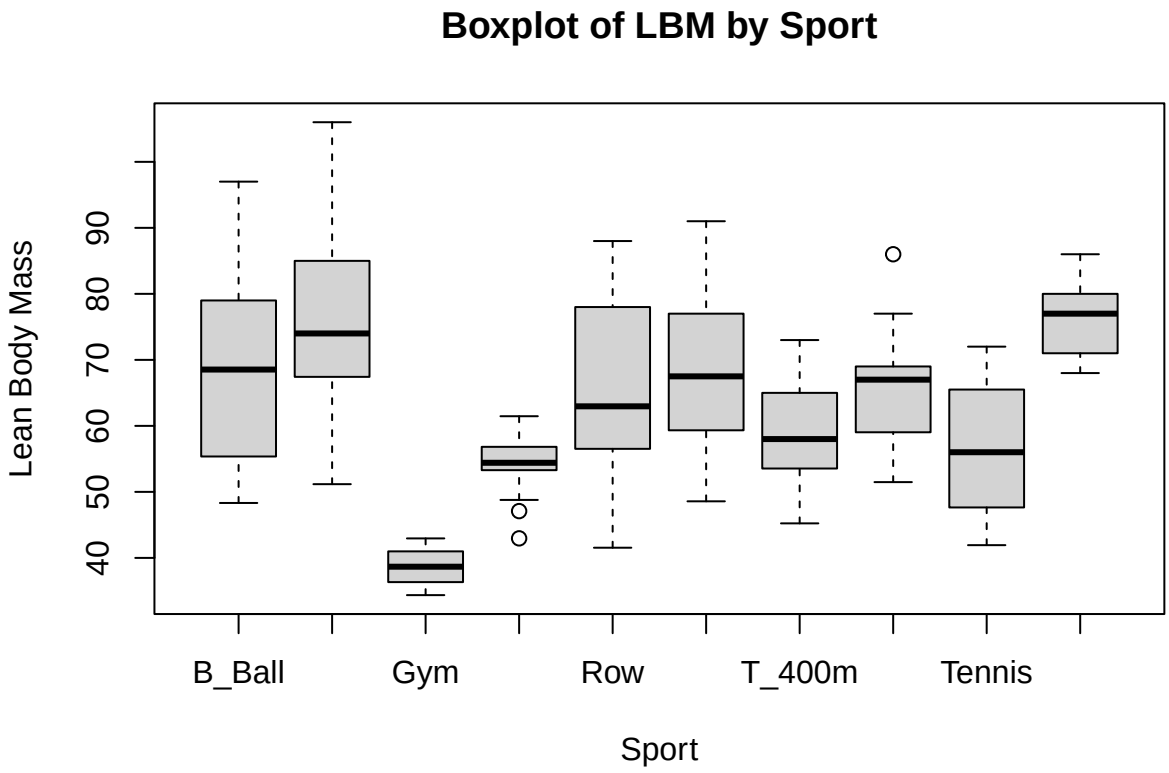
PROBLEM 1

```
library(DAAG)
data("ais")

# A: Linear model with lbm as the response and sport as the predictor
model_lbm_sport <- lm(lbm ~ sport, data = ais)
summary(model_lbm_sport)
```

```
##
## Call:
## lm(formula = lbm ~ sport, data = ais)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.294  -7.857  -0.118   6.720  29.536
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68.2628     2.1336  31.995  < 2e-16 ***
## sportField     8.2014     3.2468   2.526  0.01234 *
## sportGym     -29.6028     5.7448  -5.153 6.34e-07 ***
## sportNetball -13.9998     3.0822  -4.542 9.82e-06 ***
## sportRow      -1.6350     2.7618  -0.592  0.55456
## sportSwim     -0.8773     3.1185  -0.281  0.77875
## sportT_400m   -9.6714     2.9114  -3.322  0.00107 **
## sportT_Sprnt  -2.5395     3.4841  -0.729  0.46696
## sportTennis  -12.0173     3.8597  -3.114  0.00213 **
## sportW_Polo    7.6784     3.3535   2.290  0.02313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.67 on 192 degrees of freedom
## Multiple R-squared:  0.3637, Adjusted R-squared:  0.3338
## F-statistic: 12.19 on 9 and 192 DF,  p-value: 3.562e-15
```

```
# B: Boxplots for lbm according to sport
boxplot(lbm ~ sport, data = ais,
        main = "Boxplot of LBM by Sport",
```

```
        xlab = "Sport",
        ylab = "Lean Body Mass")
```



C: Interpretation of the F-test - The F-test in the summary output is used to determine whether there is a significant effect of sport on lean body mass across the different sports. The null hypothesis for the F-test is that all group means are equal, or in other words, sport has no effect on lean body mass.

The provided output shows an F-statistic value of 12.19 with a p-value of 3.562e-15. The very small p-value (much less than any common significance level like 0.05 or 0.01) leads us to reject the null hypothesis. This means there is strong evidence that there are differences in lean body mass among at least some of the sports categories.

D: Interpretation of the coefficient estimate for Field and Tennis - The coefficient estimate for each sport represents the difference in lean body mass between athletes of that sport and the reference category (which is not shown and is likely the one not listed, often the one alphabetically first if the reference was not manually set).

For sportField, the estimate is 8.2014 with a p-value of 0.01234. Since the p-value is less than 0.05, it suggests that athletes in track and field sports have, on average, 8.2014 units higher lean body mass compared to the reference sport category, holding all else equal, and this difference is statistically significant.

For Tennis, the estimate is -12.0173 with a p-value of 0.00213. This negative coefficient indicates that tennis athletes have, on average, 12.0173 units lower lean body mass compared to the reference sport category, holding all else equal, and this difference is statistically significant.

To sum up, the lean body mass for track and field athletes is significantly higher than the reference group, while for tennis athletes, it's significantly lower.

```r
# E: Estimate the contrast between Tennis and W_Polo
library(emmeans)

# Estimate the emmeans for 'sport'
emm <- emmeans(model_lbm_sport, specs = "sport")

# Calculate the pairwise contrasts
contrast_estimate <- pairs(emm)

# We can use the summary function to extract the contrasts
summary_contrast <- summary(contrast_estimate)

# Find the row in the summary that corresponds to the Tennis vs. W_Polo contrast
tennis_vs_wpolo_row <- which(summary_contrast$contrast == "Tennis - W_Polo")

# Obtain the 90% CI for the Tennis vs. W_Polo contrast
tennis_vs_wpolo_ci <- confint(contrast_estimate, level =
                                0.90)[tennis_vs_wpolo_row, ]

# Output the results
print(tennis_vs_wpolo_ci)
```

```
##  contrast        estimate   SE  df lower.CL upper.CL
##  Tennis - W_Polo    -19.7 4.13 192    -31.9    -7.53
##
## Confidence level used: 0.9
## Conf-level adjustment: tukey method for comparing a family of 10 estimates
```

```r
# F: Adding the ht (height) variable to the model
# Fit the model including height
model_lbm_sport_ht <- lm(lbm ~ sport + ht, data = ais)

# Get a summary of the new model to interpret the coefficients
summary(model_lbm_sport_ht)
```

```
##
## Call:
## lm(formula = lbm ~ sport + ht, data = ais)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.014  -4.226   0.447   4.093  20.605
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -143.17381   11.08885 -12.912  < 2e-16 ***
## sportField     17.23448    1.95977   8.794 8.42e-16 ***
## sportGym        9.88606    3.94523   2.506  0.01305 *
## sportNetball    0.09121    1.94953   0.047  0.96273
## sportRow        5.40806    1.65935   3.259  0.00132 **
## sportSwim       8.21686    1.88767   4.353 2.19e-05 ***
## sportT_400m     5.13147    1.87219   2.741  0.00671 **
```

```
## sportT_Sprnt   11.21561    2.16362   5.184 5.51e-07 ***
## sportTennis     4.22914    2.41482   1.751  0.08150 .
## sportW_Polo     8.16754    1.96512   4.156 4.88e-05 ***
## ht              1.12073    0.05840  19.190  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.251 on 191 degrees of freedom
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7713
## F-statistic: 68.79 on 10 and 191 DF,  p-value: < 2.2e-16
```

Interpretation of the coefficient for ht: The coefficient estimate for ht is 1.12073, with a highly significant p-value (p < 2.2e-16). This suggests a strong relationship between height and lean body mass. The interpretation is that for each one-centimeter increase in height, the lean body mass is expected to increase by an average of 1.12073 kilograms, assuming all other factors remain constant.

Interpretation of the coefficient for Field now: The coefficient for Field (assuming it represents Track and Field athletes) is 17.23481 with a p-value < 2.2e-16, which is also highly significant. This suggests that, holding height constant, Track and Field athletes have a lean body mass that is, on average, 17.23481 kilograms more than the reference category.

PROBLEM 2

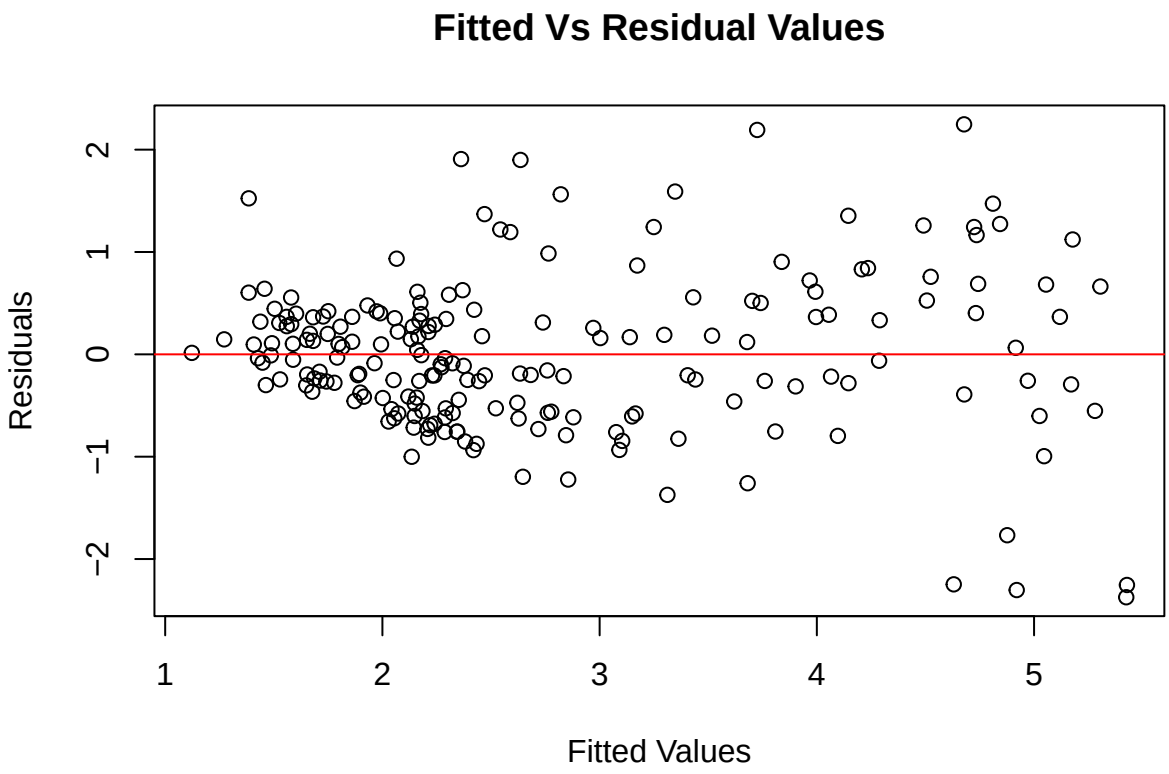
```
# A: Fit the model with fertility as the response variable
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:DAAG':
##
##     vif

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

##
## Attaching package: 'alr4'

## The following object is masked _by_ '.GlobalEnv':
##
##     ais

## The following object is masked from 'package:DAAG':
##
##     ais
```

```r
data("UN11")

model_fertility <- lm(fertility ~ ppgdp + lifeExpF + pctUrban, data = UN11)
summary(model_fertility)
```

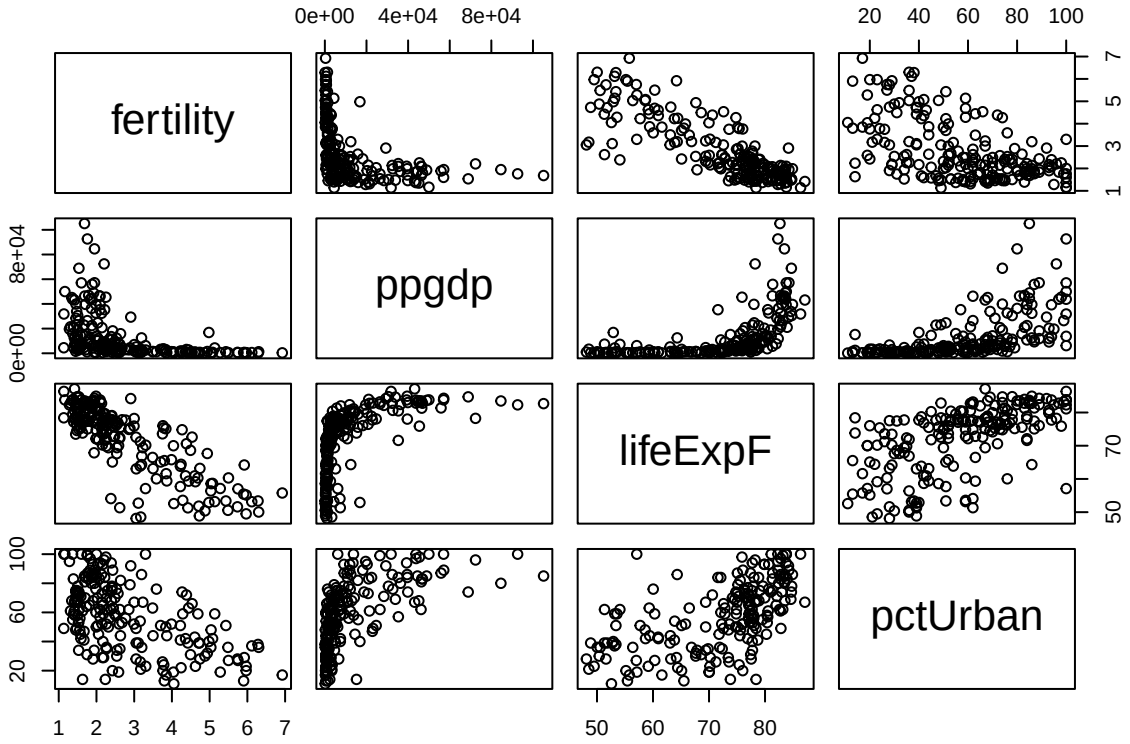
```
##
## Call:
## lm(formula = fertility ~ ppgdp + lifeExpF + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37338 -0.46706 -0.03743  0.39032  2.24726
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.064e+01  4.430e-01  24.011   <2e-16 ***
## ppgdp        3.996e-06  3.828e-06   1.044    0.298
## lifeExpF    -1.052e-01  6.982e-03 -15.064   <2e-16 ***
## pctUrban    -5.594e-03  3.147e-03  -1.778    0.077 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7592 on 195 degrees of freedom
## Multiple R-squared:  0.6836, Adjusted R-squared:  0.6788
## F-statistic: 140.5 on 3 and 195 DF,  p-value: < 2.2e-16
```

```r
# B: Plot of fitted values vs. residuals
plot(fitted(model_fertility), residuals(model_fertility),
     main="Fitted Vs Residual Values",
     xlab="Fitted Values",
     ylab="Residuals")
abline(h = 0, col = "red")
```



Interpretation: The residual plot of the linear model shows no clear patterns, which implies the linear model might be a suitable fit, as it doesn't seem to violate the linearity assumption significantly.

```
# C: Scatter plot pairs
pairs(~fertility + ppgdp + lifeExpF + pctUrban, data = UN11)
```



In the scatterplot examining fertility versus per capita GDP, a negative trend emerges, indicating that fertility rates tend to fall as per capita GDP climbs. This trend appears to curve, suggesting a nonlinear relationship.

When observing fertility against female life expectancy, a negative link is also present; nations with longer female life expectancies often have lower fertility rates. The pattern here seems fairly straight, hinting at a linear association.

The connection between fertility and the percentage of the urban population isn't completely linear. Initially, there's considerable variability, yet a general negative direction is noticeable, with fertility rates tending to drop as the urban population proportion increases.

Looking at per capita GDP versus female life expectancy, a positive association is discernible. Higher per capita GDP often aligns with longer female life expectancy, and the pattern of this relationship seems to curve, implying nonlinearity.

The relationship between per capita GDP and the percentage of the urban population is also positive; higher GDP per capita is linked with a larger urban population share, with the relationship again being nonlinear.

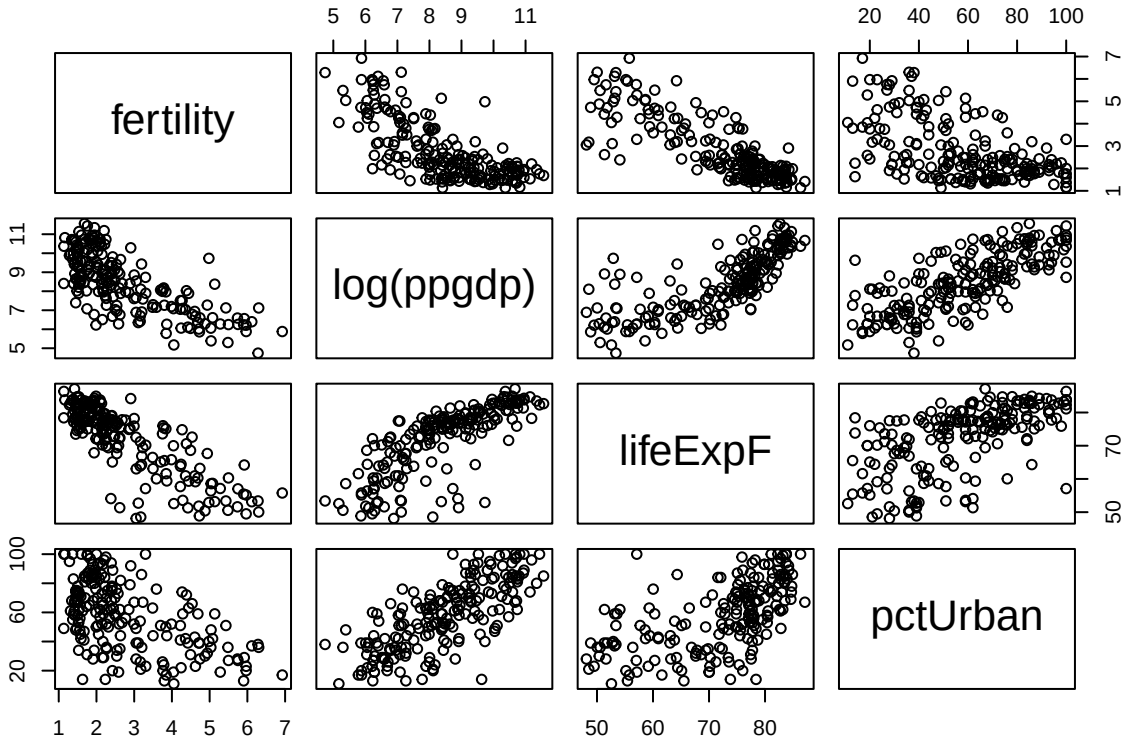
Lastly, female life expectancy correlates positively with the percentage of urban population; countries with more urbanized populations generally see higher female life expectancies.

The observations of nonlinear trends suggest that simple linear models might not adequately capture the nuances of these relationships without transformations of the variables, such as applying a logarithmic scale to per capita GDP. Additionally, the presence of positive correlations among per capita GDP, female life expectancy, and the urban population percentage raises the issue of potential multicollinearity, which could complicate the interpretation of these variables in a multivariate model.

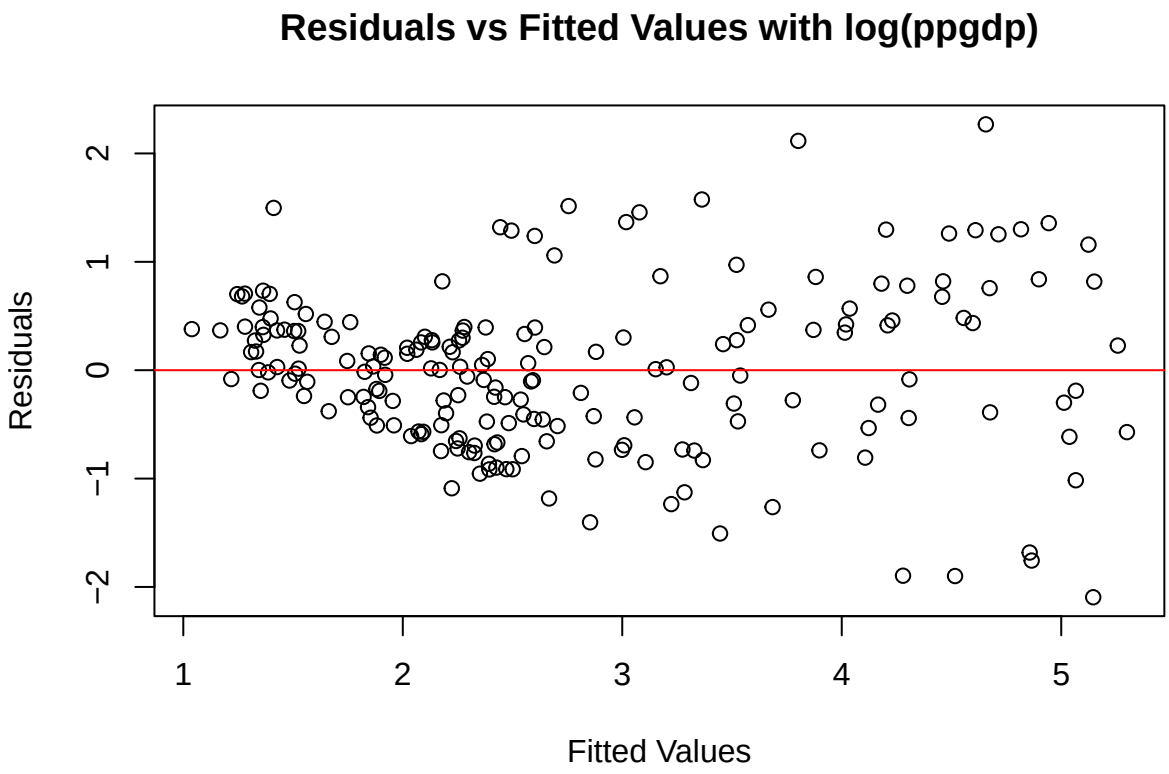
```r
# D: Model with log(ppgdp) and related plots
model_fertility_log_ppgdp <- lm(fertility ~ log(ppgdp) + lifeExpF + pctUrban,
                                data = UN11)
summary(model_fertility_log_ppgdp)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + lifeExpF + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09478 -0.49773  0.00959  0.39810  2.26841
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.681994   0.406247  26.294  < 2e-16 ***
## log(ppgdp)  -0.197913   0.064416  -3.072  0.00243 **
## lifeExpF    -0.087656   0.008229 -10.652  < 2e-16 ***
## pctUrban     0.001578   0.003405   0.463  0.64371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7436 on 195 degrees of freedom
## Multiple R-squared:  0.6966, Adjusted R-squared:  0.6919
## F-statistic: 149.2 on 3 and 195 DF,  p-value: < 2.2e-16
```

```r
pairs(~fertility + log(ppgdp) + lifeExpF + pctUrban, data = UN11)
```




```
plot(fitted(model_fertility_log_ppgdp), residuals(model_fertility_log_ppgdp),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted Values with log(ppgdp)")
abline(h = 0, col = "red")
```



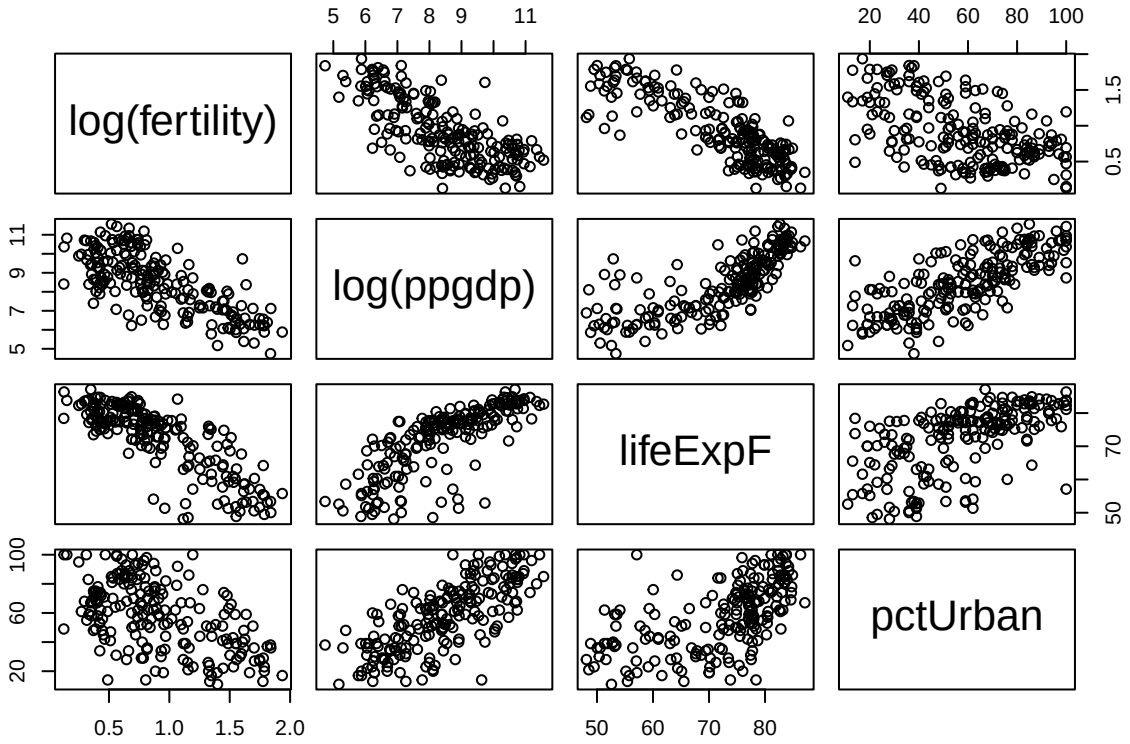
The regression analysis shows a significant inverse correlation between a country's GDP per capita and its fertility rate. A one-unit increase in the logged GDP per capita correlates with a 0.1979 unit decrease in the fertility rate, assuming female life expectancy and urbanization rates remain constant. The improved linearity in scatter plots after logging GDP suggests that the transformed model captures the relationship between the variables more effectively. While the spread of residuals still exhibits some variability, the previous cone-shaped pattern is less pronounced, marking an advancement over the initial model. However, residual analysis indicates room for further refinement.

```
# E:  Model with both log(fertility) and log(ppgdp)
model_log_fertility_log_ppgdp <- lm(log(fertility) ~ log(ppgdp) + lifeExpF +
                                    pctUrban, data = UN11)
summary(model_log_fertility_log_ppgdp)
```

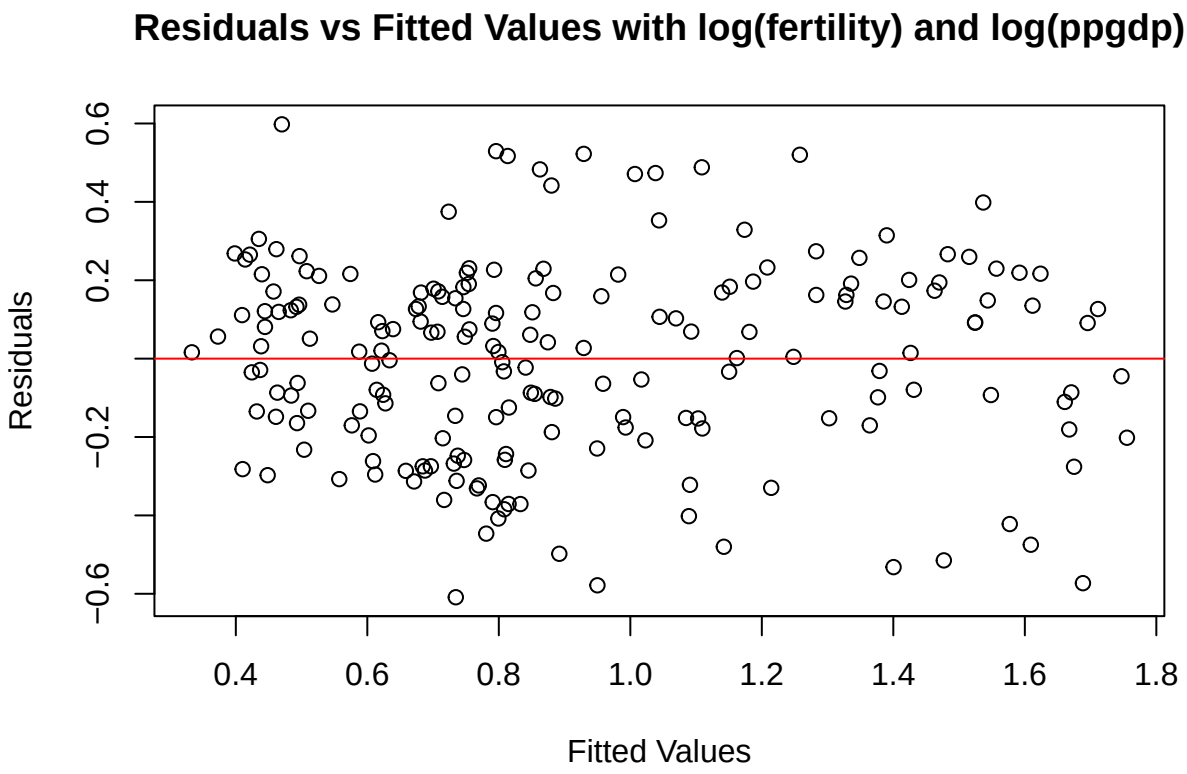
```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp) + lifeExpF + pctUrban,
##     data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60882 -0.17027  0.02719  0.17002  0.59780
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5478477  0.1355562  26.173  < 2e-16 ***
## log(ppgdp)  -0.0758004  0.0214944  -3.527 0.000525 ***
```

```
## lifeExpF    -0.0283677  0.0027458 -10.331  < 2e-16 ***
## pctUrban     0.0009794  0.0011363   0.862 0.389798
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2481 on 195 degrees of freedom
## Multiple R-squared:  0.6937, Adjusted R-squared:  0.689
## F-statistic: 147.2 on 3 and 195 DF,  p-value: < 2.2e-16
```

```r
pairs(~log(fertility) + log(ppgdp) + lifeExpF + pctUrban, data = UN11)
```




```r
plot(fitted(model_log_fertility_log_ppgdp),
     residuals(model_log_fertility_log_ppgdp),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted Values with log(fertility) and log(ppgdp)")
abline(h = 0, col = "red")
```



The regression analysis indicates a significant negative association between a country's GDP per capita and its fertility rate; a one-unit increase in the log-transformed GDP per capita is associated with a 0.0758 unit decrease in fertility rate, holding female life expectancy and urbanization constant. The residual plot shows a marked improvement with a more uniform variance, suggesting an appropriate model fit without discernible trends.

PROBLEM 4

```
# Part A:
# Load the Swiss data
data(swiss)

# Full model with all predictors
full_model <- lm(Fertility ~ ., data = swiss)

# Forward selection
forward_model <- step(object = lm(Fertility ~ 1, data = swiss),
                      scope = list(lower = formula(lm(Fertility ~ 1,
                                                      data = swiss)),
                                   upper = formula(full_model)),
                      direction = "forward",
                      trace = 0)

# Backward selection
backward_model <- step(full_model, direction = "backward", trace = 0)

# Step wise selection (both forward and backward steps)
```

```r
stepwise_model <- step(object = lm(Fertility ~ 1, data = swiss),
                       scope = list(lower = formula(lm(Fertility ~ 1,
                                                       data = swiss)),
                                    upper = formula(full_model)),
                       direction = "both",
                       trace = 0)

forward_model
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality +
##     Agriculture, data = swiss)
##
## Coefficients:
##      (Intercept)         Education          Catholic  Infant.Mortality
##          62.1013           -0.9803            0.1247            1.0784
##      Agriculture
##          -0.1546
```

```r
backward_model
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Coefficients:
##      (Intercept)       Agriculture         Education          Catholic
##          62.1013           -0.1546           -0.9803            0.1247
## Infant.Mortality
##           1.0784
```

```r
stepwise_model
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality +
##     Agriculture, data = swiss)
##
## Coefficients:
##      (Intercept)         Education          Catholic  Infant.Mortality
##          62.1013           -0.9803            0.1247            1.0784
##      Agriculture
##          -0.1546
```

Part B(i): The models converge on Education, Catholic, and Infant.Mortality as key predictors for Fertility, though Agriculture's inclusion in only two models hints at its less definitive predictive role.

Part B(ii): Education, Catholicism, and Infant Mortality emerge as the most consistent predictors of Fertility across all models, highlighting their strong correlation with Fertility in Swiss provinces. Education inversely correlates with Fertility, suggesting that higher educational attainment may lead to lower Fertility rates.

A higher percentage of Catholics correlates with increased Fertility, while greater Infant Mortality rates also correspond with higher Fertility, potentially indicating compensatory reproductive behavior. These persistent associations across different models reinforce the significance of these factors in understanding Fertility trends.

Answer 3.

(a) $e = Y - \hat{Y}$

$e = Y - X(X^T X)^{-1} X^T Y$

$e = (I - H) Y \quad [X(X^T X)^{-1} X^T = H]$

$e = (I - H)(X\beta + \varepsilon)$

$e = (X\beta - HX\beta) + \varepsilon(I - H)$

$e = X\beta - X\underbrace{(X^T X)^{-1} X^T X}_{A^{-1}A = I}\beta + \varepsilon(I - H)$

$e = X\beta - X\beta + \varepsilon(I - H)$

$e = \varepsilon(I - H)$

$E(e) = E[\varepsilon(I - H)]$

$= \underbrace{(I - H)}_{\text{constant}} E(\varepsilon)$

As we know, $E(\varepsilon) = 0$

So, $E(e) = 0$

(b) Variance of vector e $(n \times 1)$:

$Var(e) = Cov(e, e) = Cov((I - H)Y, (I - H)Y)$

$= (I - H)\, Cov(Y)(I - H)$

$= (I - H)\, \sigma^2 I_n (I - H)$

$= \sigma^2 [(I - H)(I - H)]$

$= \sigma^2 [I^2 - IH - HI + HH]$

As H is an independent matrix, $HH = H$

So, $\text{Var}(e) = \sigma^2 [I_n - H - H + H]$

$\text{Var}(e) = \sigma^2 [I_n - H]$

(c) To find: distribution of $BY + C$ 1_n

[$1_n \rightarrow n \times 1$ vector of 1s]

$Y \sim N(X\beta, \sigma^2 \Sigma)$

Normal distribution is not affected by linear transformations.

So,

$BY + C \sim N(E, V)$, where:

$E \rightarrow$ Expectation

$V \rightarrow$ Variance

$$E = E(BY + C) = B E(Y) + E(C)$$

[We know, $E(Y) = X\beta$]

$$\Rightarrow E = BX\beta + C$$

$$V = \text{Var}(BY + C) = B \,\text{Var}(Y) B' + \text{Var}(C)$$

$$\text{Var}(C) = \begin{pmatrix} 11 \\ 0 \end{pmatrix}$$

$$= B(\sigma^2 I) B' + 0$$

$$= \sigma^2 B B'$$

So, $BY + C \sim N(BX\beta + C, \sigma^2 BB')$ *