GU4205/5205-Linear Regression Models-Lab1c

Authors: Dr. Banu Baydil, Dr. Ronald Neath

Fall 2022

Section 1: Fitting a Multiple Linear Regression Model

In this section we will work with the 2011 UN data.

```
library(alr4)
attach(UN11)
dim(UN11)
```

```
## [1] 199   6
```

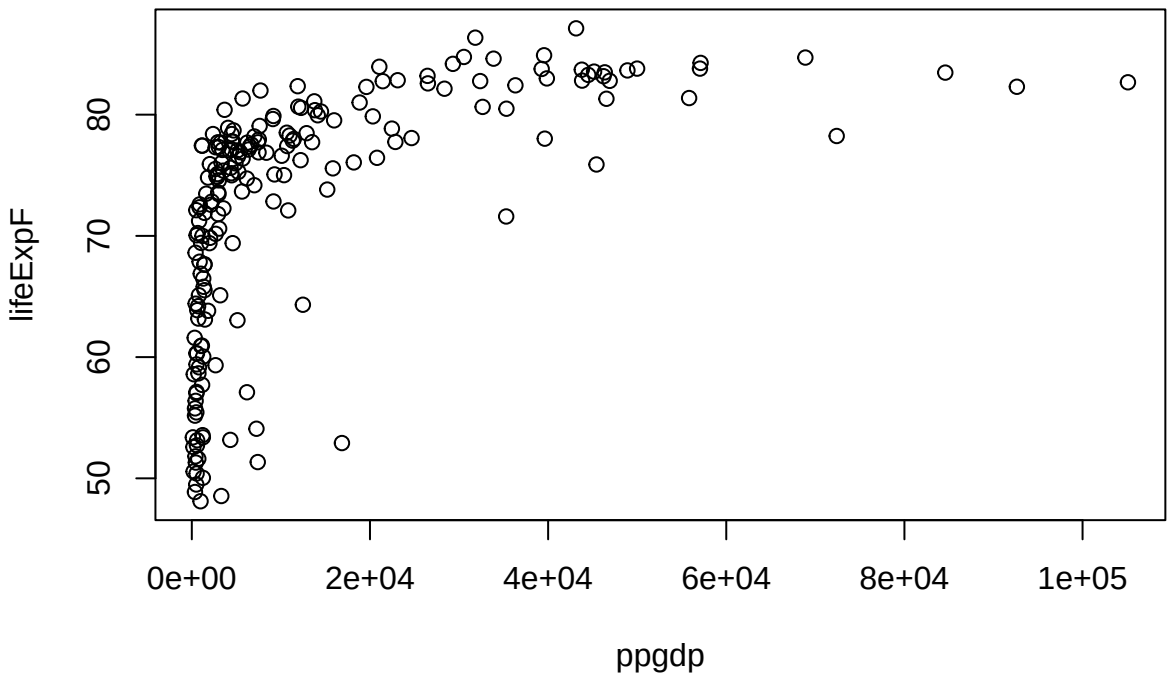
```
names(UN11)
```

```
## [1] "region"    "group"     "fertility" "ppgdp"     "lifeExpF"  "pctUrban"
```

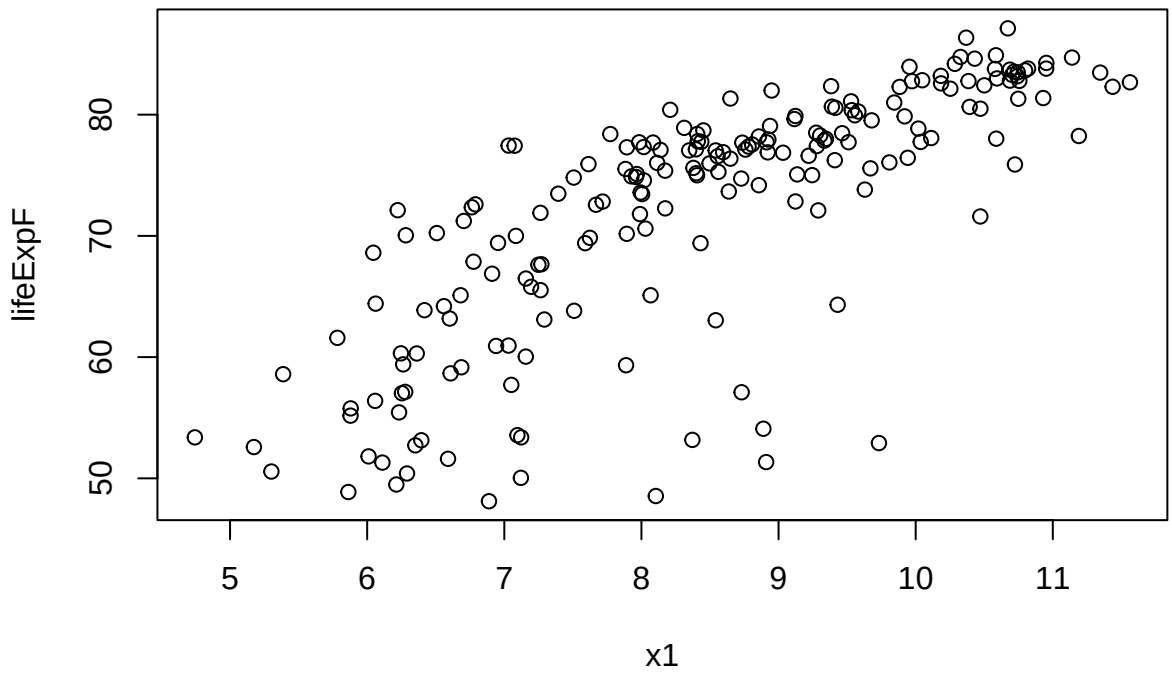
```
help(UN11)
```

Let us consider a regression of y=lifeExpF on x1=ppgdp:

```
plot(lifeExpF ~ ppgdp, data=UN11)
```




```
# Let's try log transformation!
x1=log(ppgdp)
plot(lifeExpF ~ x1, data=UN11)
```



Better. Still some curvature in the mean function?

Let's fit the SLR model:

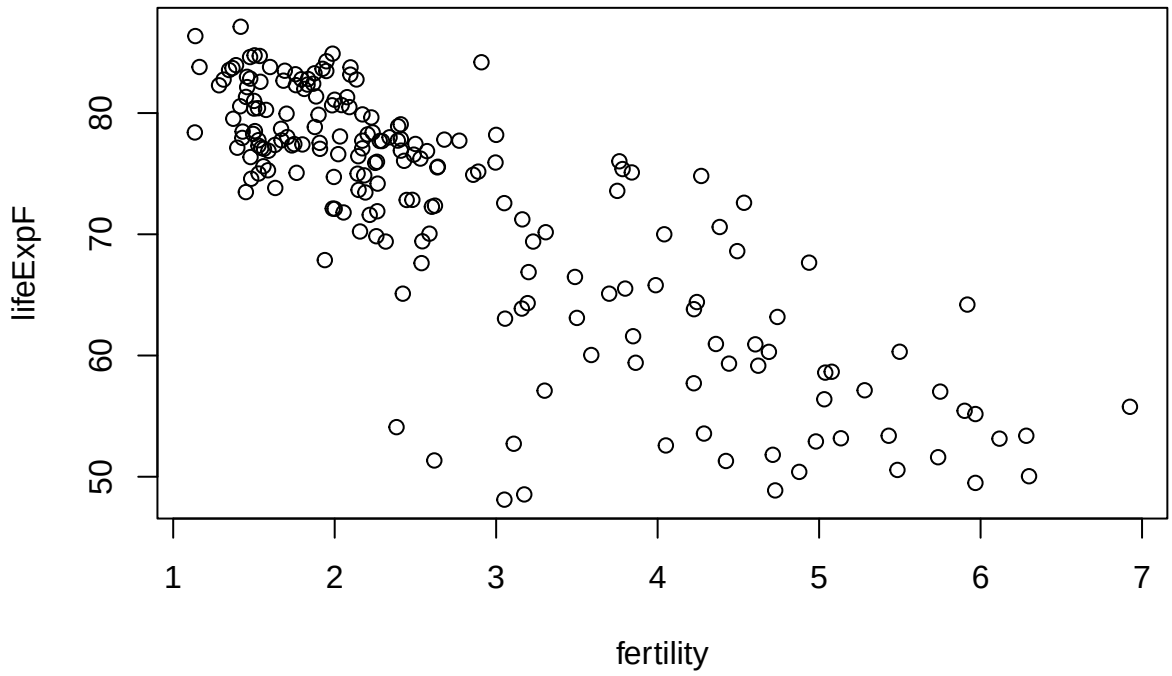
```
m1 <- lm(lifeExpF ~ log(ppgdp), data=UN11)
summary(m1)
```

```
##
## Call:
## lm(formula = lifeExpF ~ log(ppgdp), data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.749  -2.879   1.280   3.987  12.345
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.8148     2.5314   11.78   <2e-16 ***
## log(ppgdp)    5.0188     0.2942   17.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.448 on 197 degrees of freedom
## Multiple R-squared:  0.5964, Adjusted R-squared:  0.5943
## F-statistic: 291.1 on 1 and 197 DF,  p-value: < 2.2e-16
```

Note that beta1.hat=5.02, beta0.hat=29.8, sigma.hat=6.45, and R.squared=.596. About 60% of variability in lifeExpF is explained by log(ppgdp).

Next, let us consider y=lifeExpF and x2=fertility

```
plot(lifeExpF ~ fertility, data=UN11)
```



and fit the SLR model:

```
m2 <- lm(lifeExpF ~ fertility, data=UN11)
summary(m2)
```

```
##
## Call:
## lm(formula = lifeExpF ~ fertility, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3806  -2.6855   0.5826   3.6434  12.8156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.4805     0.9369   95.51   <2e-16 ***
## fertility    -6.2242     0.3054  -20.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.757 on 197 degrees of freedom
## Multiple R-squared:  0.6783, Adjusted R-squared:  0.6767
## F-statistic: 415.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

Note that beta1.hat=-6.22, beta0.hat=89.5, sigma.hat=5.76, R.squared=.678. About 68% of variability in lifeExpF is explained by fertility.

Small p-value indicates that there is evidence that country fertility level is negatively associated with female life expectancy.

Given two countries, one with higher fertility rate by 1 birth per woman, the model predicts that the country with the higher fertility has shorter female life expectancy by 6.2 years.

Let us next fit a regression model with both predictors!

```
m12 <- lm(lifeExpF ~ log(ppgdp) + fertility, data=UN11)
summary(m12)
```

```
##
## Call:
## lm(formula = lifeExpF ~ log(ppgdp) + fertility, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6362  -1.6854   0.4221   2.7301  11.7978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.4484     3.7446  16.944  < 2e-16 ***
## log(ppgdp)    2.4150     0.3386   7.132 1.86e-11 ***
## fertility    -4.1991     0.3938 -10.664  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.142 on 196 degrees of freedom
## Multiple R-squared:  0.7446, Adjusted R-squared:  0.742
## F-statistic: 285.7 on 2 and 196 DF,  p-value: < 2.2e-16
```

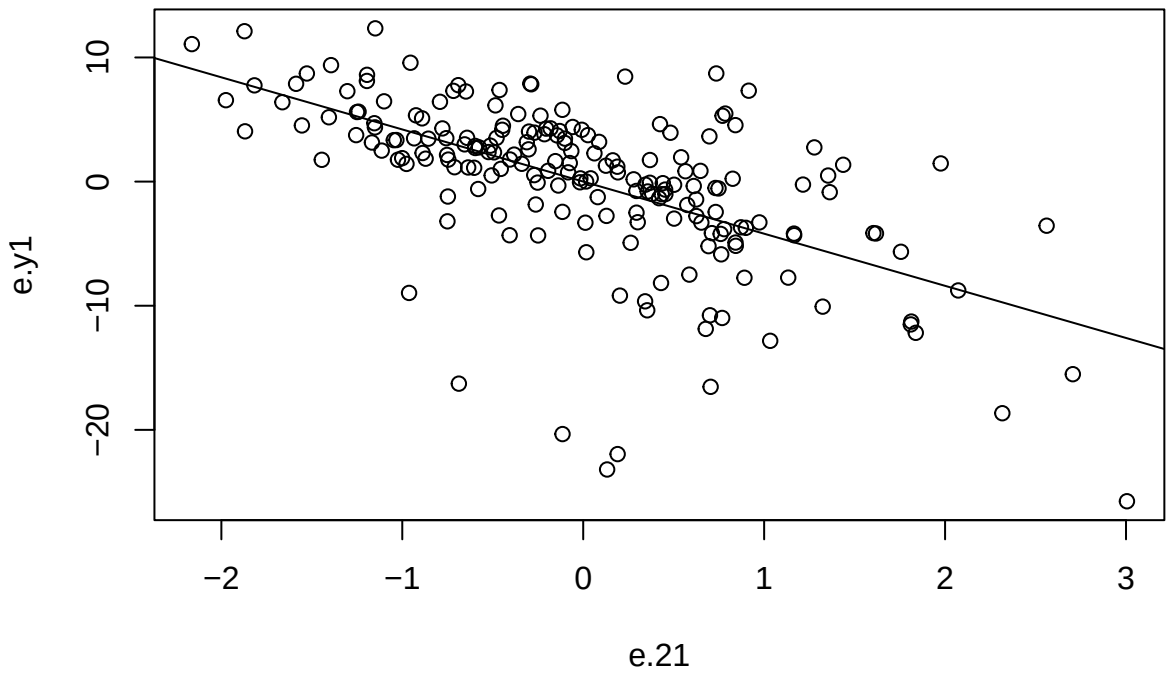
Note that beta1.hat=2.4, beta2.hat=-4.2, sigma.hat=5.14, R.squared=.74.

Given two countries, one with higher fertility rate by 1 birth per woman and both having the same ppgdp, the model predicts that the country with higher fertility has shorter life expectancy by 4.2 years.

Section 2: Added-variable Plot for Fertility, Accounting for log(ppgdp)

We can do this as follows:

```
e.y1 <- resid(lm(lifeExpF ~ log(ppgdp), data=UN11))
e.21 <- resid(lm(fertility ~ log(ppgdp), data=UN11))
plot(e.y1 ~ e.21); abline(lm(e.y1 ~ e.21));
```

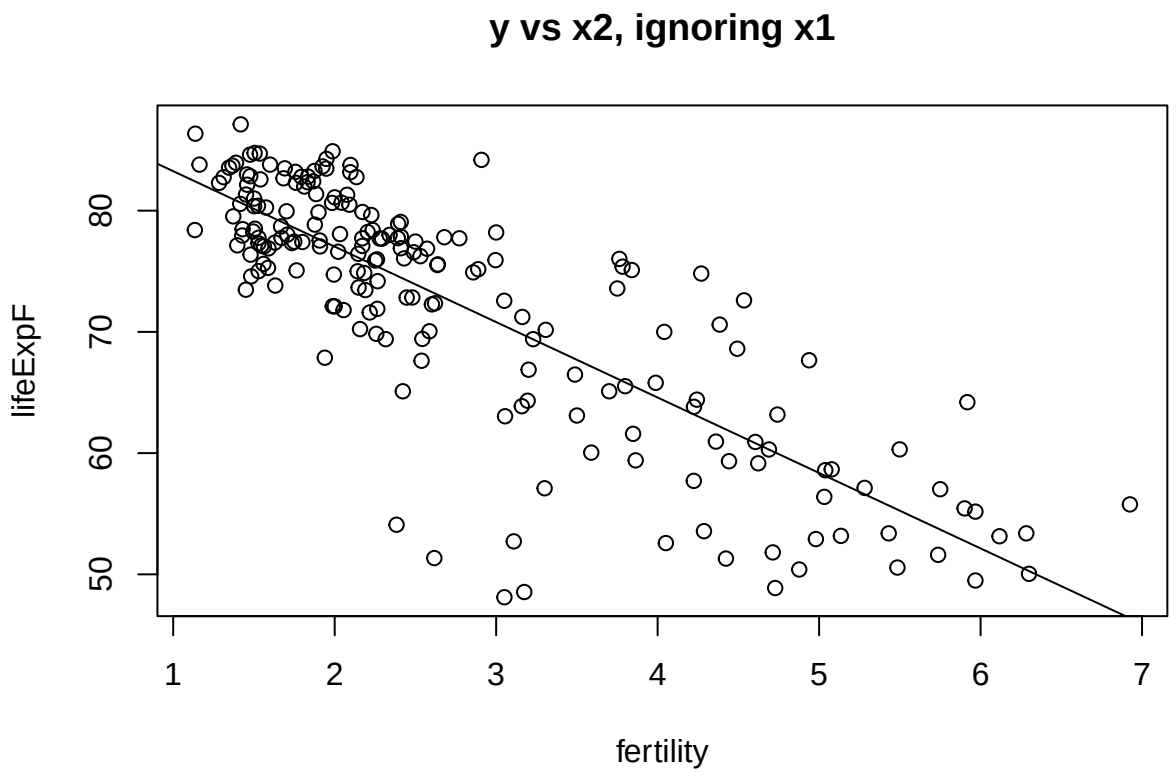



```
coef(lm(e.y1 ~ e.21)) # should get intercept=0 and slope=-4.2
```

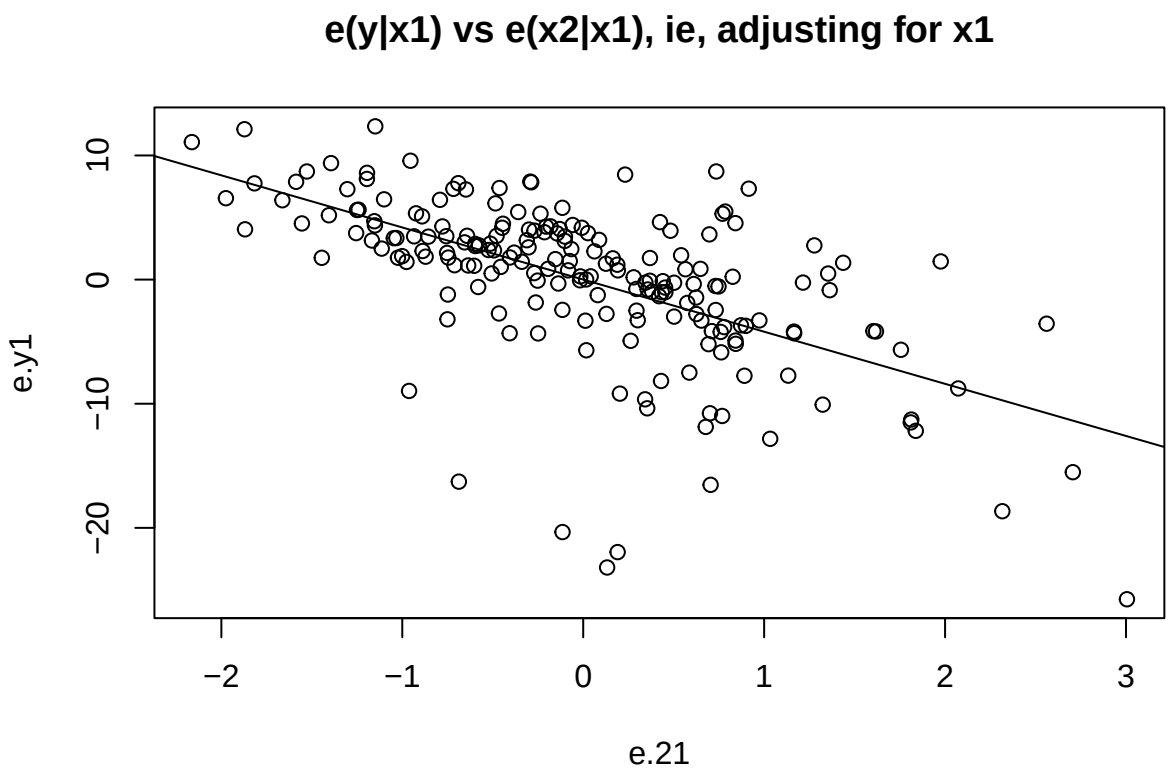
```
##   (Intercept)          e.21
##  7.680727e-16 -4.199124e+00
```

Relationship Between lifeExpF and fertility, before and after adjusting for log(ppgdp)

```
plot(lifeExpF ~ fertility, data=UN11, main="y vs x2, ignoring x1")
abline(m2)
```




```
plot(e.y1 ~ e.21, main="e(y|x1) vs e(x2|x1), ie, adjusting for x1")
abline(lm(e.y1 ~ e.21))
```



Let us demonstrate that proportion of variability explained by x1 and x2 equals the proportion explained by x1, plus the proportion of that unexplained by x1 that is explained by x2, after adjusting for x1:

```
#to show: Rsq.12 = Rsq.1 + (1 - Rsq.1) * Rsq.2_1

Rsq.1 <- summary(m1)$r.squared
Rsq.1
```

```
## [1] 0.5963835
```

```
Rsq.2_1 <- summary(lm(e.y1 ~ e.21))$r.squared
Rsq.2_1
```

```
## [1] 0.3671866
```

```
Rsq.12 <- summary(m12)$r.squared
Rsq.12
```

```
## [1] 0.7445861
```

```
Rsq.12==Rsq.1 + (1 - Rsq.1) * Rsq.2_1
```

```
## [1] TRUE
```

Section 3: Creating a New Dataframe

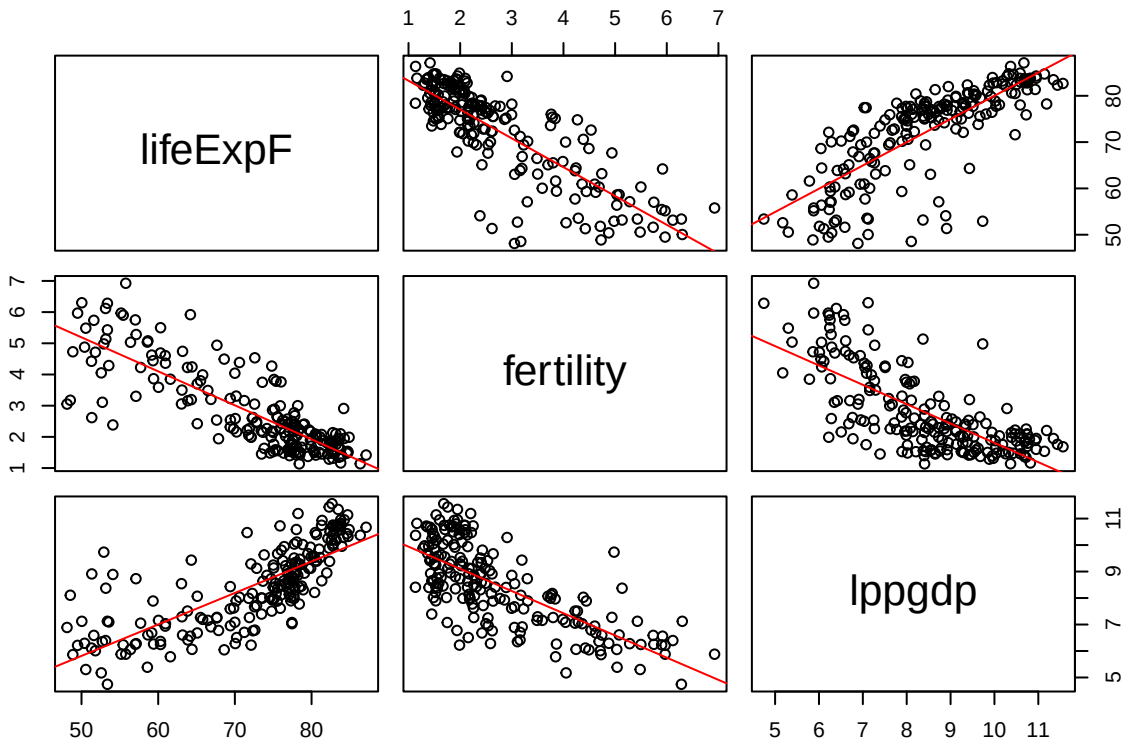
Let us create a new data set from UN11 Data set:

```
Data <- data.frame(lifeExpF=UN11$lifeExpF)

Data$fertility <- UN11$fertility

Data$lppgdp <- log(UN11$ppgdp)

rownames(Data) <- rownames(UN11)
```

Section 4: Scatter Plot Matrix

Scatter plot matrix is given by:

```
panel.ls <- function(x,y)
{
  points(x,y); abline(lm(y~x), col="red");
}
pairs(Data, panel=panel.ls)
```



Section 5: Sample Correlations:

Sample correlation matrix is given by:

```
round(cor(Data), 4)
```

```
##           lifeExpF fertility  lppgdp
## lifeExpF    1.0000   -0.8236  0.7723
```

```
## fertility  -0.8236    1.0000 -0.7211
## lppgdp      0.7723   -0.7211  1.0000
```

Section 6: Confidence and Prediction Intervals

Let us first fit a linear model and save the model fit in m13.

```
m13 <- lm(lifeExpF ~ lppgdp + fertility, data=Data)
summary(m13)
```

```
##
## Call:
## lm(formula = lifeExpF ~ lppgdp + fertility, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6362  -1.6854   0.4221   2.7301  11.7978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.4484     3.7446  16.944  < 2e-16 ***
## lppgdp        2.4150     0.3386   7.132 1.86e-11 ***
## fertility    -4.1991     0.3938 -10.664  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.142 on 196 degrees of freedom
## Multiple R-squared:  0.7446, Adjusted R-squared:  0.742
## F-statistic: 285.7 on 2 and 196 DF,  p-value: < 2.2e-16
```

Next we can compute a 95% confidence interval for mean female life expectancy of countries with lppgdp=2, fertility=3; and a 95% prediction interval for female life expectancy of a country with lppgdp=2, fertility=3.

```
new.case <- data.frame(lppgdp=2, fertility=3)
predict(m13, newdata=new.case, interval="confidence", level=.95) #CI
```

```
##        fit      lwr      upr
## 1 55.68105 51.43501 59.92709
```

```
predict(m13, newdata=new.case, interval="prediction", level=.95) #PI
```

```
##        fit      lwr      upr
## 1 55.68105 44.68644 66.67567
```