Applied Econometrics HW 3

Tianyu Gao

2022-10-04

1

a)

For 1744 samples, the sample size is large enough, so we can approach the t-distribution using normal distribution.

$$t = \frac{\beta_{age^4} - 0}{SE(\beta_{age^4})} = \frac{.0005}{.0009} \approx .556$$
$$p - value \approx .711$$

Therefore we fail the reject the null hypothesis that the age^4 term is 0. To determine the appropriate degree of the polynomial, we first need to draw the scatter plot of the dependent variable and independent variable to find out whether a linear term is enough to describe the relationship, if not, we can turn to investigating the terms of higher powers (let's say independent variable to the power x) using residual plot assuming the term independent variable to the power n-1 is including, i.e, drawing the scatter plot of residuals and the variable we are interested in (the linear term) to find out whether powers bigger than x have an unexplained influence on the residuals.

```
5/9
```

```
## [1] 0.5555556
```

```
pnorm(.556)
```

```
## [1] 0.7108946
```

b)

I prefer the latter whose highest-power is quadratic, and this is because while the two models have similar R^2 and SER, the latter one is simpler.

For quadratic regression,

$$R^2_{adj} = 1 - \frac{\frac{SSR}{n-p-1}}{\frac{TSS}{n-1}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$
$$= 1 - (1 - .222)\frac{1743}{1744 - 4} \approx .221$$

```
1 - (1 - .222)*1743 / 1740
```

```
## [1] 0.2206586
```

For cubic regression,

$$R^2_{adj} = 1 - \frac{\frac{SSR}{n-p-1}}{\frac{TSS}{n-1}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$
$$= 1 - (1 - .225)\frac{1743}{1739} \approx .223$$

```
1 - (1 - .225) * 1743 / 1739
```

```
## [1] 0.2232174
```

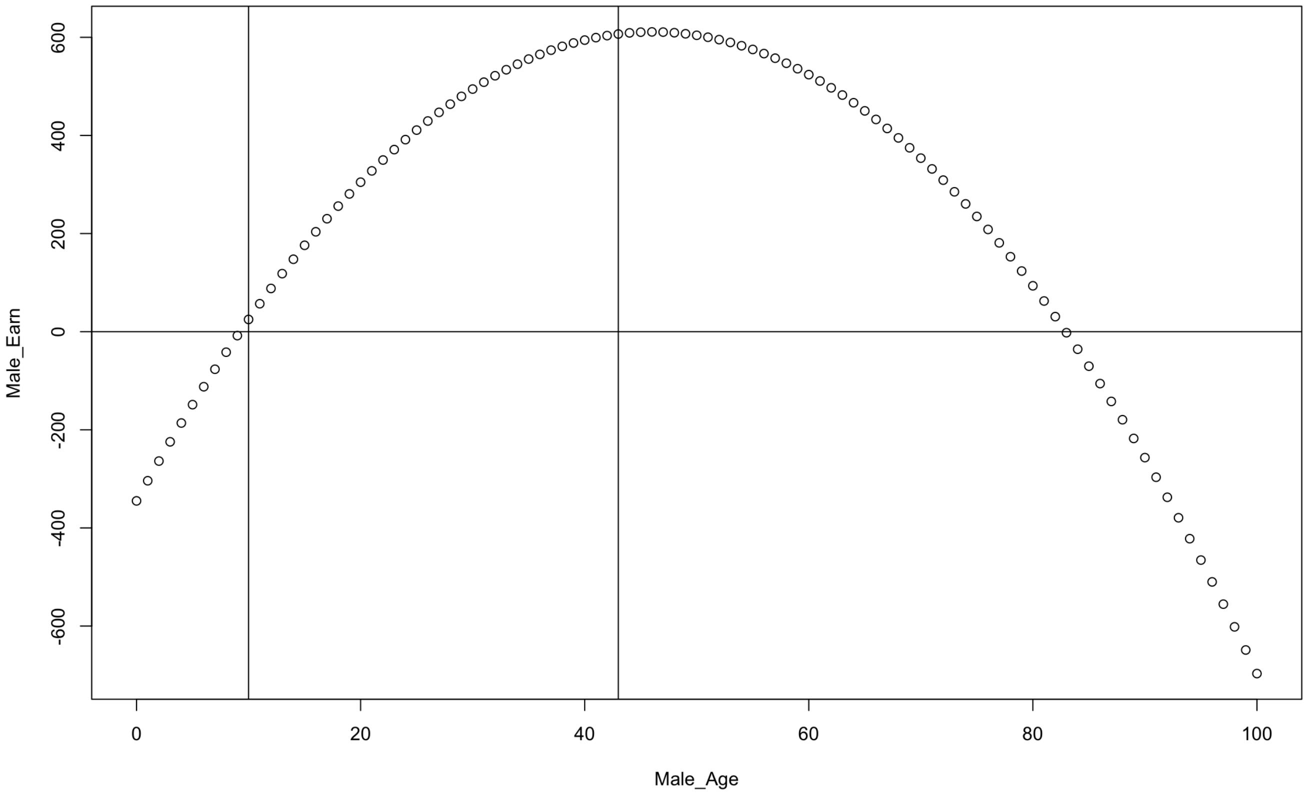
Actually, for R^2_{adj}, cubic regression is slightly bigger than quadratic regression. But considering the interpret-ability and the relation ship between the terms of different powers, I would use quadratic algorithm.

c)

For male observations, we have

$$\widehat{Earn} = -344.88 + 41.48Age - .45Age^2$$

```
Male_Age = seq(from = 0, to = 100, by = 1)
Male_Earn = -344.88 + 41.48 * Male_Age -.45 * Male_Age^2
plot(Male_Age, Male_Earn)
abline(a = 0, b = 0)
abline(v = 10)
abline(v = 43)
```

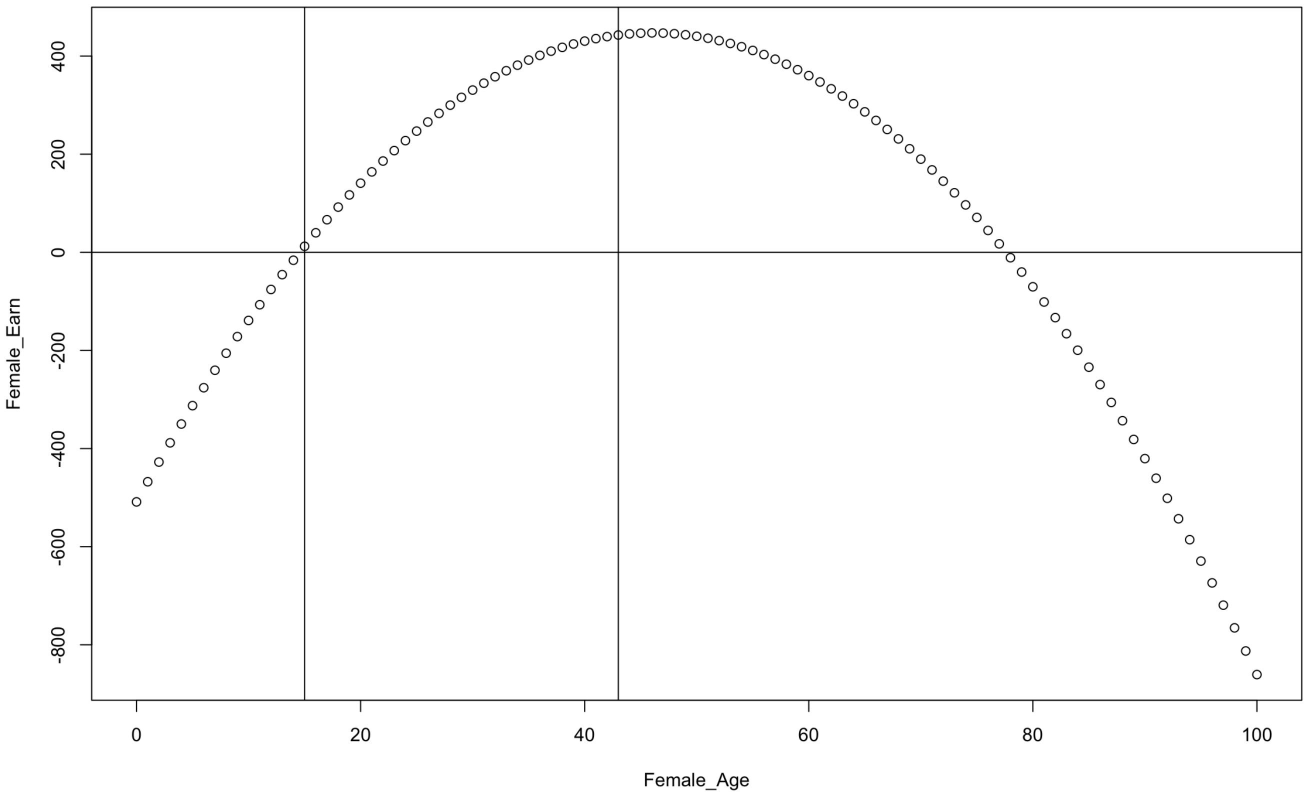


The simulation curve makes perfect sense that it has similar trend to the raw data. Because the earning starts at around 18 in raw data (obviously it is illegal to hire underage people), and the salary grows from 18 to 40, to make the curve is smooth (as the polynomial function is n-order differentiable) the simulated data before 10 is negative. As the experience and social networking grow, earnings of people grow in general. The peak continues until 60 and suddenly decreases because people earn much who have enough money while their energy doesn't allow them to work as hard as young guys retire and people do not have enough money and work on easy jobs continue working after their 60s which implies a decrease in the salary. Also, because polynomial curve is always smooth and differentiable at any order, the curve drop here.

For women:

$$\widehat{Earns} = -344.88 + 41.48Age - .45Age^2 - 163.81 = -508.69 + 41.48Age - .45Age^2$$

```
Female_Age = seq(from = 0, to = 100, by = 1)
Female_Earn = -508.69 + 41.48 * Female_Age -.45 * Female_Age^2
plot(Female_Age, Female_Earn)
abline(a = 0, b = 0)
abline(v = 15)
abline(v = 43)
```



The trend is similar to female individuals. Also, the earning grows and drops. It is worth noticing that the earnings of the female individuals are significantly lower than those of the males, which implies a structural discrimination in patriarchy society towards female individuals, including assuming male workers are much more compatible the female ones and female salary should be deducted in advance in case they get pregnant and not being able to go to work. Also, employers tend to think female workers tend to focus more on their family so they are inclined to promoting male workers.

d)

Step 1:

In the dataset label the male observations $Female = 0$ and the female observations $Female = 1$

Step 2:

Use residual plots and scatter plots to find the appropriate power of Age, in this step, we can also run different regression on different-powered terms and use model selection methods such as step-wise selection to figure out the appropriate model. Suppose we find the best model with linear term and quadratic term.

Step 3:

Suppose the highest power is k

Estimate the following regression model:

$$\widehat{Earn} = \beta_0 + \sum_{i=1}^{k} \beta_i Age^i + \beta_3 Female$$

In later discussion, we suppose the quadratic model is appropriate.

Step 4:

Make sure the $p - value$ for Age and Age^2 is small enough so we cannot reject the null hypothesis.

Step 5:

The β_1 and β_2 represent the effect on earnings holding the constant the gender variable. For people at $Age\ \frac{k}{2}$, one year change in Age leads to $\beta_1 + k\beta_2$ change in estimated earning.

For age k,

$$
\begin{aligned}
SE(effect) &= SE(\beta_1 + k\beta_2)\\
&= \sqrt{\frac{Var(\beta_1 + k\beta_2)}{n}}\\
&= \sqrt{\frac{Var(\beta_1) + k^2Var(\beta_2) + 2kCOV(\beta_1, \beta_2)}{n}}\\
&= SE(\beta_1) + kSE(\beta_2) + \sqrt{\frac{2kCOV(\beta_1, \beta_2)}{n}}
\end{aligned}
$$

2

Suppose the regression model is:

$$Log\_Earning = \beta_0 + \beta_1 Gender + \beta_2 Marital + \beta_3 Education + X\vec{\beta} + U$$

where X is the matrix of other explanatory variables and $\vec{\beta}$ is the column vector containing other coefficients.

In this model, I do not specify the subscripts, which means $Earning, Gender, Marital, Education$ are column vectors containing all the observations.

a)

Holding other variables constant, the variable of $Education$ increases 1 unit, the $Earning$ increase $\beta_3 \times 100\%$ by proportion.

b)

$$Gender_i = \begin{cases} 1, when\ observation\ i\ is\ female \\ 0, otherwise \end{cases}$$

$$Marital_i = \begin{cases} 1, when\ observation\ is\ married \\ 0, otherwise \end{cases}$$

For single male, the coefficient measures the effect of gender and marital isβ_0

For single female, the coefficient measures the effect of gender and marital is$\beta_0 + \beta_1$

For married male, the coefficient measures the effect of gender and marital is$\beta_0 + \beta_2$

For married female, , the coefficient measures the effect of gender and marital is$\beta_0 + \beta_1 + \beta_2$

c)

$$Log\_Earning_i = \beta_0 + \beta_1 Gender_i + \beta_2 Marital_i + \beta_3 Education_i + \beta_4 Gender_i \times Marital_i + X_i\vec{\beta} + U$$

In the equation above, $Log_Earning_i, Gender_i, Marital_i, Education_i$ are scalars from observation i. X_i is a vector containing other information from observation i. And the parameters need re-estimating

Assume we use the same definition from session b,

For single male, the coefficient measures the effect of gender and marital isβ_0

For single female, the coefficient measures the effect of gender and marital is$\beta_0 + \beta_1$

For married male, the coefficient measures the effect of gender and marital is$\beta_0 + \beta_2$

For married female, , the coefficient measures the effect of gender and marital is$\beta_0 + \beta_1 + \beta_2 + \beta_4$

The latter one is more general is because it separates the effect of gender and marital status, we can discuss whether the effect of marital status toward different gender is different.

3

a)

When $\beta_0 = 0$ and $\beta_1 = 1$, we have

$$ur_{i,t} = ur_{i,t-k} + z_{i,t}$$

The state i is state i-k + an error term. In other words, $ur_{i,t}$ is quite similar to $ur_{i,t-k}$

b)

$$\frac{2.25}{.61} \approx 3.689 > 1.96$$
$$\frac{.60}{.13} \approx 4.615 > 1.96$$

```
2.25/.61
```

```
## [1] 3.688525
```

```
.6/.13
```

```
## [1] 4.615385
```

Therefore both coefficients are significant.

The change in unemployment rate in 1970 explains 40% of change in unemployment rate in 1995. On average, when the unemployment in 1970 increases 1 unit, the unemployment rate of corresponding states in 1995 increases 0.6. The unemployment rate in 1995 where the unemployment rate in 1970 is 0 is 2.25.

c)

The data points seem to distribute along two lines. One is the regression line, another one is the 45^o curve. In this case, we can use a dummy variable and its interaction term with unemployment rate in 1970. Also, the WA point is far from the center of the curve $(\overline{ur}_{1970}, \overline{ur}_{1995})$, and the MD point is relatively closer, so the influence of WA point is larger which makes the curve incline to it. It is also a problem of influential outliers. For Maryland, its unemployment rate in 1970 is approximately 3.0 and 5.1 in 1995, while for Washington, the unemployment in 1970 is approximately 9.5 and 5.3 in 1995.

There seems not be a persistence of persistence. However, if we run a dummy variable regression above, we might be able to separate different trends in the data. One is persistence, another is non-persistence. To test the hypothesis, we can first fit this model:

$$ur_{1995} = \beta_0 + \beta_1 D_1 + \beta_2 ur_{1970} + \beta_2 D_2 ur_{1970}$$

when

If we simply want to test whether all the data have persistence, we can fit this model:

$$ur_{1995} = \beta_{intercept} + \beta_{slope} ur_{1970} + ur_{1970}$$

If both of null hypothesis for $\beta_{intercept} = 0$ and $\beta_{slope} = 0$ fail, we can say that there is a persistence in the data.

d)

$$ur_{i+k} = 2.25 + .6ur_i \Rightarrow ur_{i+2k} = 2.25 + .6 * 2.25 + .6 * .6ur_i$$
$$\Rightarrow ur_{i+nk} = \sum_{j=0}^{n-1} 2.25^j + .6^n ur_i$$

When $n \to \infty, ur_{i+nk} = \infty$,and when n is relatively large (but not approaches infinity),

$$ur_{i+nk} = \sum_{j=0}^{n-1} 2.25^j$$

Because we omit the ur_i term in the comparison. Therefore, the model is not appropriate in the long run because all the states would have identical unemployment rate.

e)

$$\begin{cases} 1 - \frac{\sigma_V^2}{\sigma_V^2+\sigma_X^2} \to 0 \iff \sigma_V >> \sigma_X \\ 1 - \frac{\sigma_V^2}{\sigma_V^2+\sigma_X^2} \to 1 \iff \sigma_V << \sigma_X \end{cases}$$

Therefore when the variance of natural rates is much larger than the variance of the variance of transitory shock, in other words, US is in love and peace (no state-scale conflict), we have $\beta_1 \to 1$ otherwise $\beta_1 \to 0$.

f)

$$\frac{.27}{.07} = 3.857$$

```
.27/.07
```

```
## [1] 3.857143
```

$$\frac{.27 - 1}{.07} = 10.429$$

```
(.27 - 1)/.07
```

```
## [1] -10.42857
```

Therefore β_1 is closer to 0, which means US in 1995 is relatively peaceful.

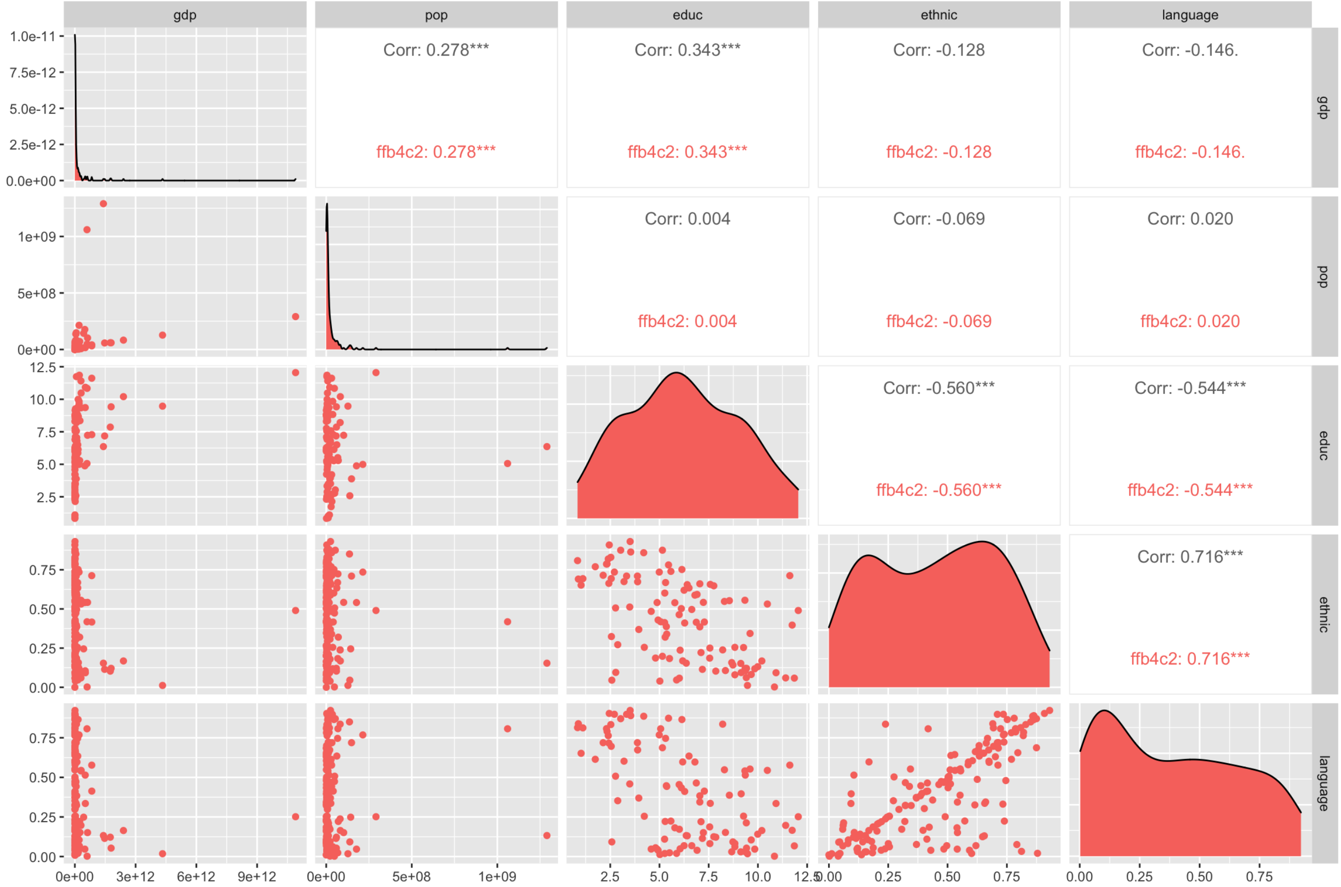
4

```
library(ggplot2)
library(haven)
library(GGally)
library(lmtest)
library(sandwich)
setwd("/Users/kevintsukuyo/Documents/Course Files/2022F/Applied Econometrics/HW3")
terrorism = read_stata('terrorism.dta')
```

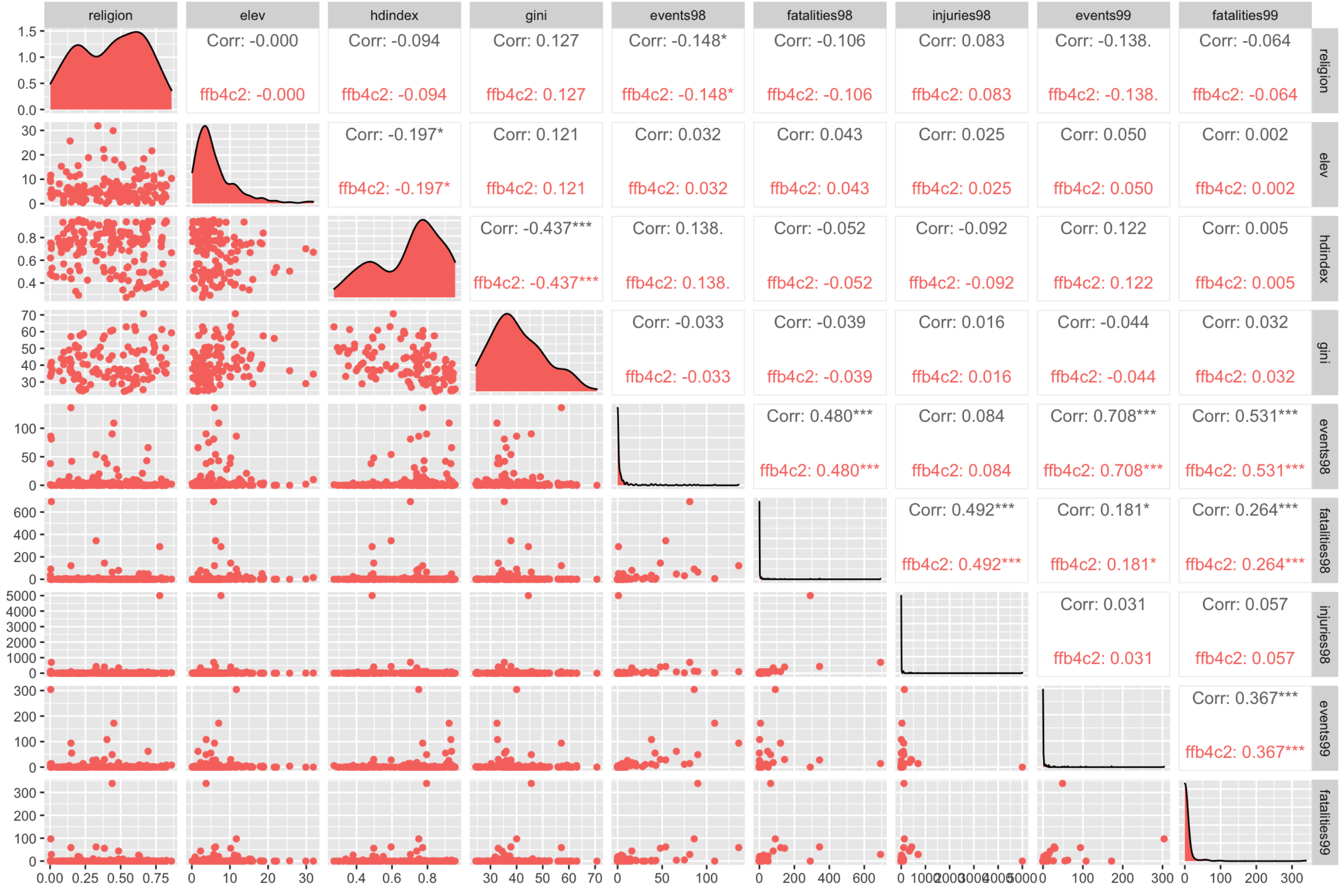
```
ncol(terrorism)
```

```
## [1] 49
```

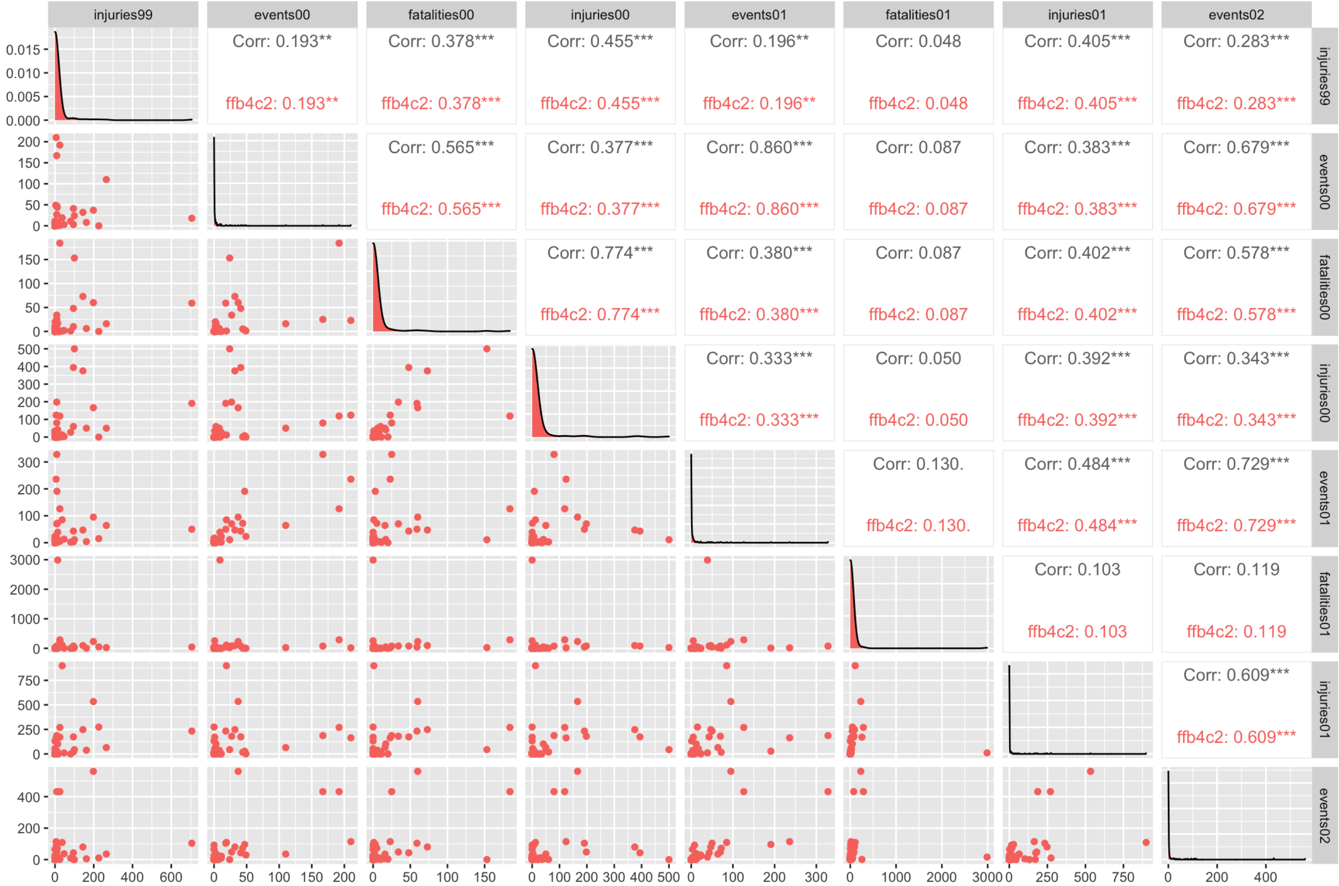
```
ggpairs(terrorism, columns = 3:7, ggplot2::aes(colour='ffb4c2'))
```



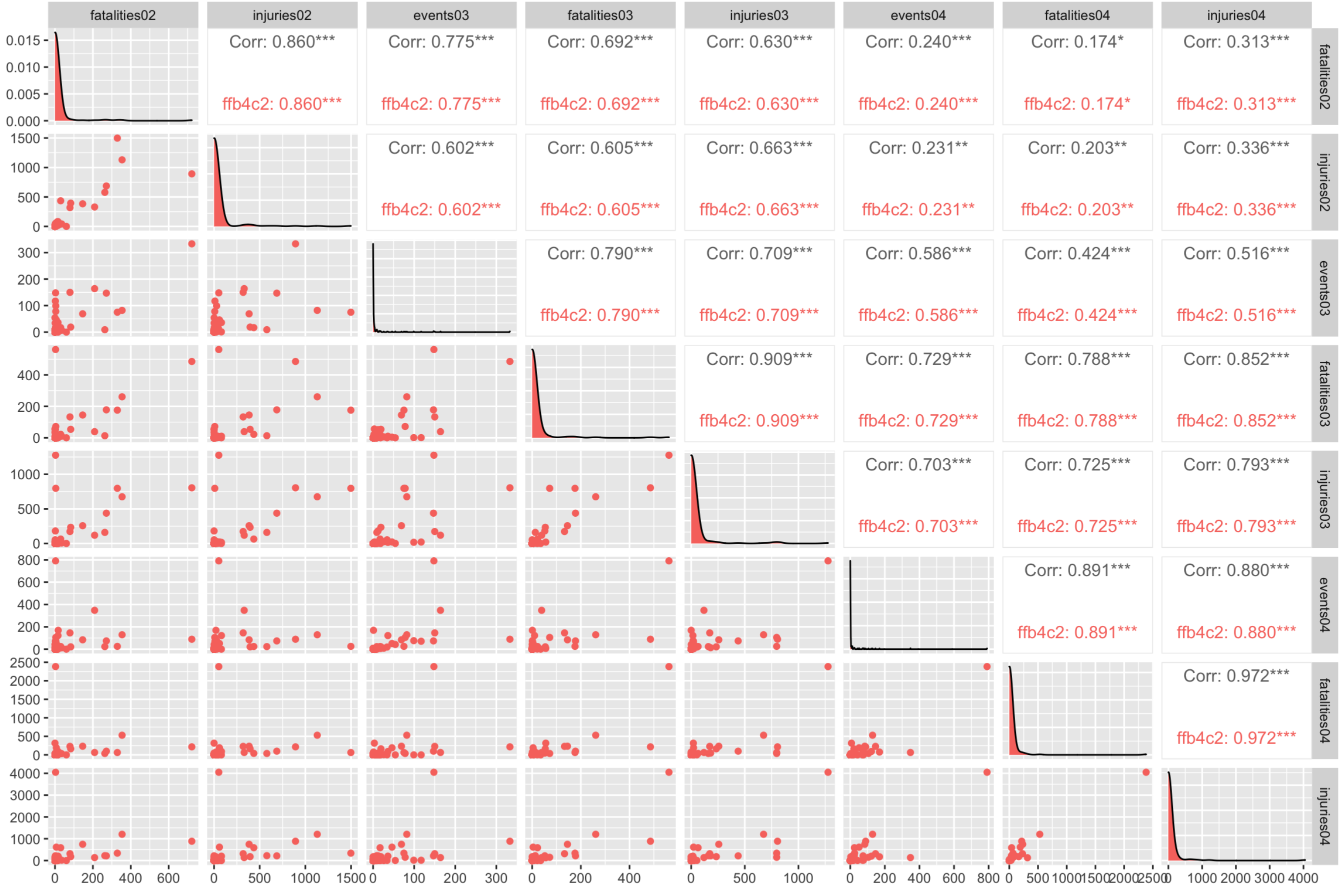
```
ggpairs(terrorism, columns = 8:16, ggplot2::aes(colour='ffb4c2'))
```



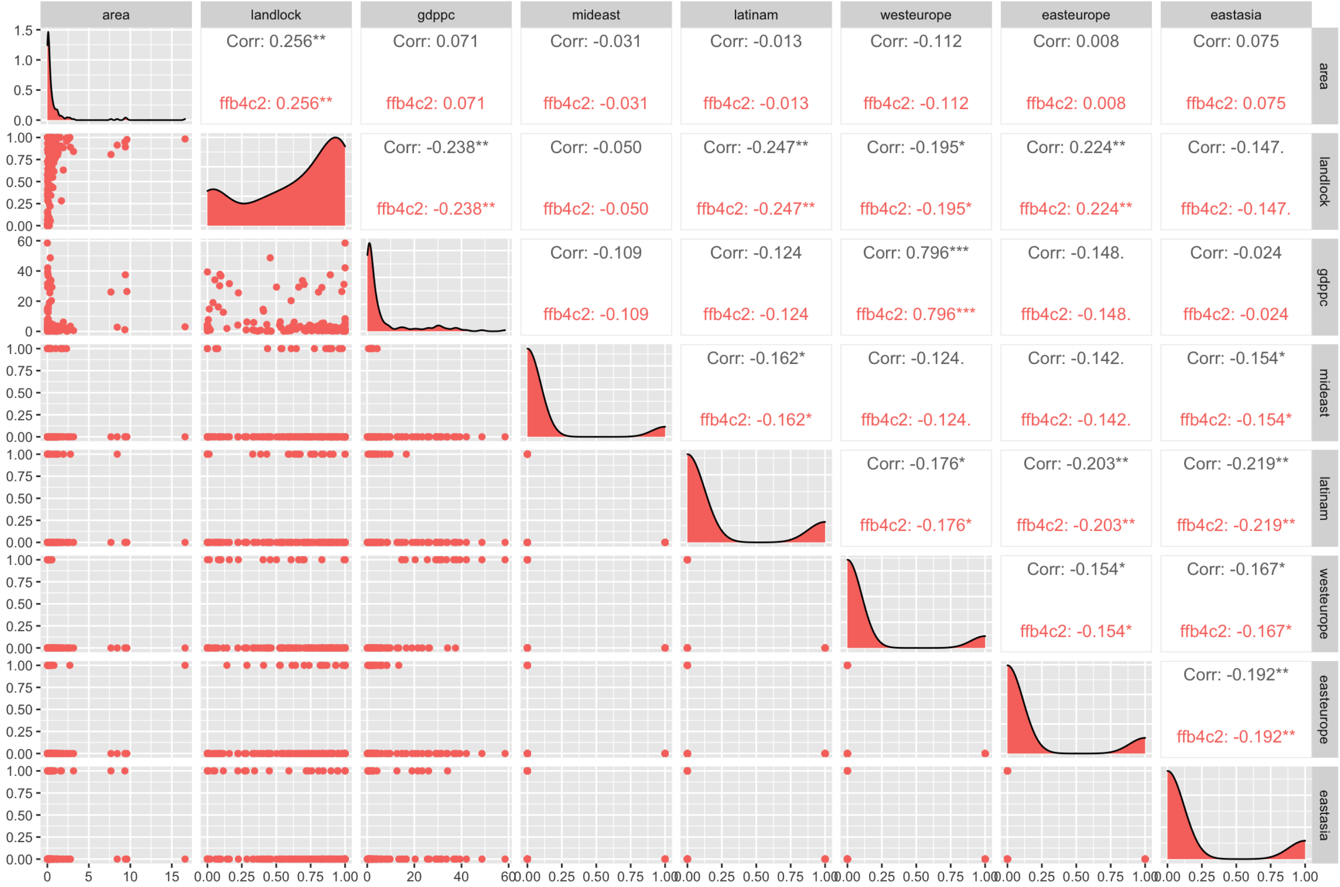
```
ggpairs(terrorism, columns = 17:24, ggplot2::aes(colour='ffb4c2'))
```



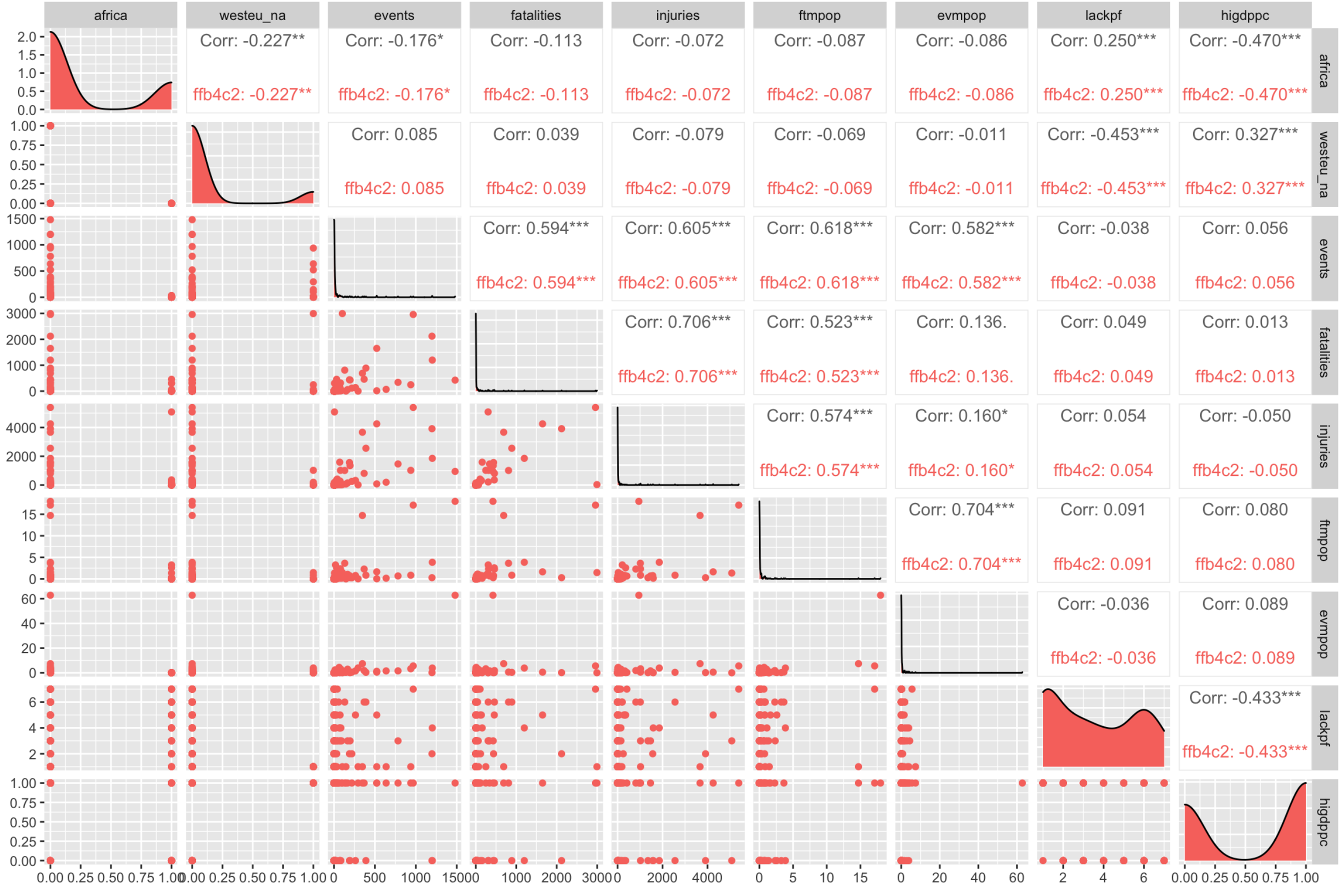
```
ggpairs(terrorism, columns = 25:32, ggplot2::aes(colour='ffb4c2'))
```



```
ggpairs(terrorism, columns = 33:40, ggplot2::aes(colour='ffb4c2'))
```

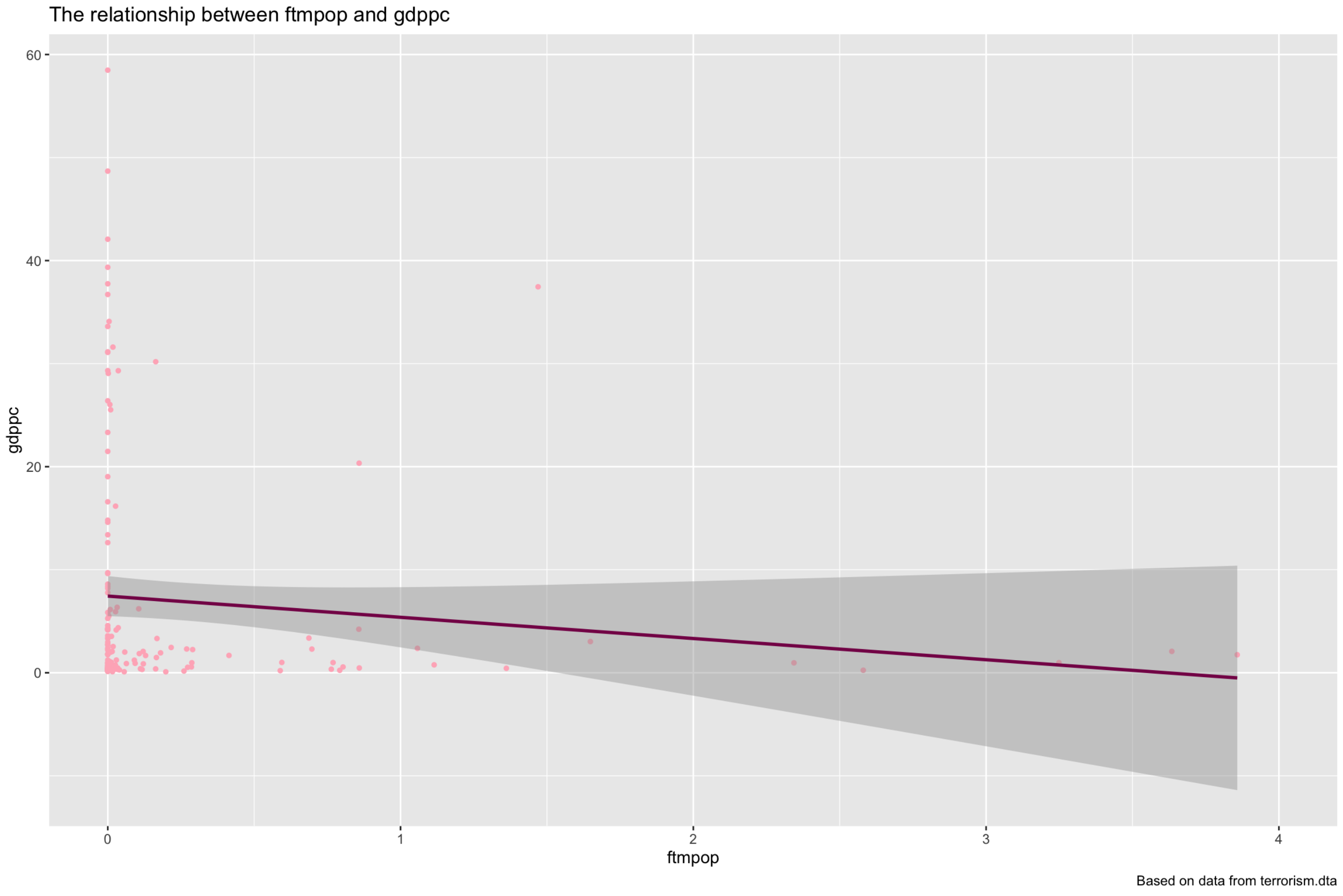


```
ggpairs(terrorism, columns = 41:49, ggplot2::aes(colour='ffb4c2'))
```



a)

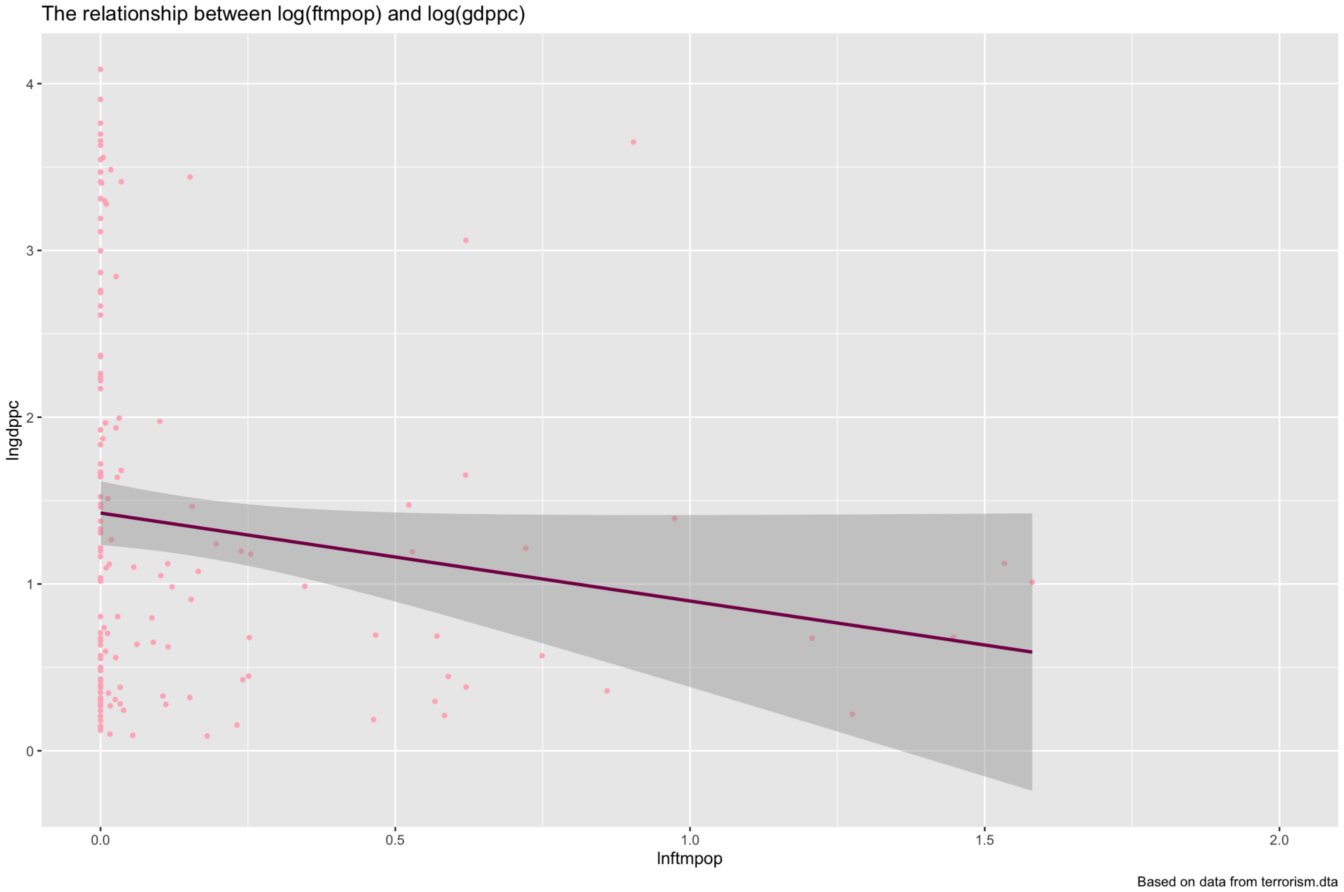
```
ggplot(data = terrorism, mapping = aes(x = ftmpop, y = gdppc))+
  geom_point(size = 1, color = '#ffb4c2')+
  labs(title = "The relationship between ftmpop and gdppc", caption = "Based on data from terrorism.dta")+
  geom_smooth(method = lm, color = '#860056')+
  xlim(0, 4)
```



The linear relationship is not significant, may be we need to do some transformation on data.

b)

```
# Here when I take logarithm, I added 1 on both terms, and this would not change the trend of the # # data, but ca
n prevent the data from turning into infinity
terrorism$lnftmpop = log(terrorism$ftmpop+1)
terrorism$lngdppc = log(terrorism$gdppc+1)
ggplot(data = terrorism, mapping = aes(x = lnftmpop, y = lngdppc))+
  geom_point(size = 1, color = '#ffb4c2')+
  labs(title = "The relationship between log(ftmpop) and log(gdppc)", caption = "Based on data from terrorism.dt
a")+
  geom_smooth(method = lm, color = '#860056')+
  xlim(0, 2)
```



Still, there is no linear relationship, but after log scaling, the points are now more separated.

c)

I would suggest nftmpop and lngdppc for OLS because the transformed data separate more equally, unlike the raw data which distribute in a small interval. Using the processed data can reduce the sum of squared residual to increase the significance.

d)

i)

```
m1 = lm(lnftmpop~lngdppc + lackpf + I(lackpf^2) + ethnic + religion, data = terrorism)
summary(m1)
```

```
##
## Call:
## lm(formula = lnftmpop ~ lngdppc + lackpf + I(lackpf^2) + ethnic +
##     religion, data = terrorism)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29274 -0.16332 -0.07853  0.00451  1.31515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.20588    0.16372  -1.257   0.2106
## lngdppc      0.03500    0.03544   0.987   0.3251
## lackpf       0.17885    0.07254   2.465   0.0149 *
## I(lackpf^2) -0.01924    0.00880  -2.186   0.0304 *
## ethnic       0.06539    0.11722   0.558   0.5778
## religion    -0.07451    0.10880  -0.685   0.4945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3018 on 144 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.07294,    Adjusted R-squared:  0.04075
## F-statistic: 2.266 on 5 and 144 DF,  p-value: 0.05101
```

Under homoskedacity assumption, the model is

$$ln(ftmpop_i) = -.206 + .035lngdppc_i + .179lackpf_i - .019lackpf_i^2 + .065ethnic_i - .074religion_i$$

```
coeftest(m1, vcov = vcovHC(m1, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2058754  0.1320299 -1.5593 0.121118
## lngdppc      0.0349962  0.0305562  1.1453 0.253983
## lackpf       0.1788503  0.0639612  2.7962 0.005877 **
## I(lackpf^2) -0.0192360  0.0076637 -2.5100 0.013178 *
## ethnic       0.0653907  0.1086710  0.6017 0.548300
## religion    -0.0745140  0.1166165 -0.6390 0.523860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model under hetroskedasticity assumption is the same (though rounded on different digits), which means in the model is homoskedasic.

ii)

```
terrorism$lnftmpop[is.na(terrorism$lnftmpop) | terrorism$lnftmpop == 'INF'] = NA
m2 = lm(lnftmpop~lackpf + higdppc + higdppc*lackpf, data = terrorism, rm.na = TRUE)
summary(m2)
```

```
##
## Call:
## lm(formula = lnftmpop ~ lackpf + higdppc + higdppc * lackpf,
##     data = terrorism, rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40741 -0.17704 -0.07782 -0.05331  2.67693
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.1749253  0.1342528   1.303    0.194
## lackpf          0.0007036  0.0277169   0.025    0.980
## higdppc        -0.1520354  0.1505493  -1.010    0.314
## lackpf:higdppc  0.0542285  0.0341527   1.588    0.114
##
## Residual standard error: 0.4143 on 170 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.04277,    Adjusted R-squared:  0.02588
## F-statistic: 2.532 on 3 and 170 DF,  p-value: 0.05879
```

Under the homoskedatisity assumption, the model is:

$$lnftmpop_i = .175 + .001 lackpf_i - .152 higdppc_i + .054 lackpf_i * higdppc_i$$

```
coeftest(m2, vcov = vcov(m2, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value Pr(>|t|)
## (Intercept)     0.17492528  0.13425284  1.3030   0.1944
## lackpf          0.00070362  0.02771694  0.0254   0.9798
## higdppc        -0.15203541  0.15054928 -1.0099   0.3140
## lackpf:higdppc  0.05422851  0.03415269  1.5878   0.1142
```

The model under hetroskedasticity assumption is the same (though rounded on different digits), which means in the model is homoskedasic.

iii)

```
m3 = lm(lngdppc~lackpf + I(lackpf^2) + higdppc + ethnic + religion + higdppc * ethnic + higdppc * religion, data
 = terrorism)
summary(m3)
```

```
##
## Call:
## lm(formula = lngdppc ~ lackpf + I(lackpf^2) + higdppc + ethnic + 
##     religion + higdppc * ethnic + higdppc * religion, data = terrorism)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.23143 -0.32895 -0.05046  0.30526  1.96038 
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.99652    0.27809   7.179 3.59e-11 ***
## lackpf           -0.63759    0.12417  -5.135 9.14e-07 ***
## I(lackpf^2)       0.06685    0.01542   4.337 2.73e-05 ***
## higdppc           1.37983    0.25534   5.404 2.68e-07 ***
## ethnic           -0.21433    0.26847  -0.798  0.42599    
## religion         -0.15116    0.28738  -0.526  0.59972    
## higdppc:ethnic   -1.13607    0.41979  -2.706  0.00764 ** 
## higdppc:religion  0.54334    0.40418   1.344  0.18099    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5617 on 142 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.7552, Adjusted R-squared:  0.7431 
## F-statistic: 62.57 on 7 and 142 DF,  p-value: < 2.2e-16
```

The homoskedastic model:

$lngdppc_i = 1.996 - .638lackpf_i + .067lackpf_i^2 + 1.380higdppc_i - .214ethnic_i - .151religion_i - 1.13higdppc_i * ethnic_i + .543higdppc_i * relig$

```
coeftest(m3, vcov = vcovHC(m3, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)    
## (Intercept)       1.996517   0.288706  6.9154 1.464e-10 ***
## lackpf           -0.637591   0.143984 -4.4282 1.883e-05 ***
## I(lackpf^2)       0.066850   0.015872  4.2117 4.478e-05 ***
## higdppc           1.379832   0.243440  5.6681 7.773e-08 ***
## ethnic           -0.214335   0.119742 -1.7900   0.07559 .  
## religion         -0.151157   0.173945 -0.8690   0.38632    
## higdppc:ethnic   -1.136074   0.438384 -2.5915   0.01055 *  
## higdppc:religion  0.543343   0.408479  1.3302   0.18560    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model under hetroskedasticity assumption is the same (though rounded on different digits), which means in the model is homoskedasic.

e)

When lackpf decreases by 2 units, the effect toward lnftmpop is:

$$.349 * 2 - 2 * 2 * .036 * 100\% = .5.54\%$$

```
.349 * 2 - 2 * 2 * .036
```

```
## [1] 0.554
```

Therefore the lnftmpop will increase 5.54%

f)

```
summary(m2)
```

```
##
## Call:
## lm(formula = lnftmpop ~ lackpf + higdppc + higdppc * lackpf,
##     data = terrorism, rm.na = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40741 -0.17704 -0.07782 -0.05331  2.67693
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1749253  0.1342528   1.303    0.194
## lackpf         0.0007036  0.0277169   0.025    0.980
## higdppc       -0.1520354  0.1505493  -1.010    0.314
## lackpf:higdppc  0.0542285  0.0341527   1.588    0.114
##
## Residual standard error: 0.4143 on 170 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.04277,    Adjusted R-squared:  0.02588
## F-statistic: 2.532 on 3 and 170 DF,  p-value: 0.05879
```

$$lnftmpop_i = .175 + .001 lackpf_i - .152 higdppc_i + .054 lackpf_i * higdppc_i$$

$$\frac{.054}{.034} \approx 1.588$$

```
.054 / .034
```

```
## [1] 1.588235
```

$$p - value = 2 * \Phi(-1.588) \approx .11 > .05$$

```
pnorm(-1.588235) * 2
```

```
## [1] 0.1122332
```

Therefore we fail to reject the null hypothesis that the two slopes are the same at the 5% significance level.

g)

For countries having higher GDP per capita, when index of lack of political freedom increases one unit, the number of fatalities from terrorist incidents in the country will be extra 5.5% larger than the countries with lower GDP holding other parameters constant.

```
exp(.054) -1
```

```
## [1] 0.0554846
```

h)

```
summary(m3)
```

```
##
## Call:
## lm(formula = lngdppc ~ lackpf + I(lackpf^2) + higdppc + ethnic +
##     religion + higdppc * ethnic + higdppc * religion, data = terrorism)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23143 -0.32895 -0.05046  0.30526  1.96038
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.99652    0.27809   7.179 3.59e-11 ***
## lackpf           -0.63759    0.12417  -5.135 9.14e-07 ***
## I(lackpf^2)       0.06685    0.01542   4.337 2.73e-05 ***
## higdppc           1.37983    0.25534   5.404 2.68e-07 ***
## ethnic           -0.21433    0.26847  -0.798  0.42599
## religion         -0.15116    0.28738  -0.526  0.59972
## higdppc:ethnic   -1.13607    0.41979  -2.706  0.00764 **
## higdppc:religion  0.54334    0.40418   1.344  0.18099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5617 on 142 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.7552, Adjusted R-squared:  0.7431
## F-statistic: 62.57 on 7 and 142 DF,  p-value: < 2.2e-16
```

$lngdppc_i = 1.996 - .638 lackpf_i + .067 lackpf_i^2 + 1.380 higdppc_i - .214 ethnic_i - .151 religion_i - 1.13 higdppc_i * ethnic_i + .543 higdppc_i * relig$

This theory does not hold. The coefficients for ethnic and religion are negative, which means the higher diversity of ethnic and religion a country has, no matter which economic condition is, the terrorist attack will be fewer.