Homework 3

Rohan Sadale

```r
library(alr4)
```

```
## Warning: package 'alr4' was built under R version 3.2.3
```

```
## Warning: package 'car' was built under R version 3.2.3
```

```
## Warning: package 'effects' was built under R version 3.2.3
```

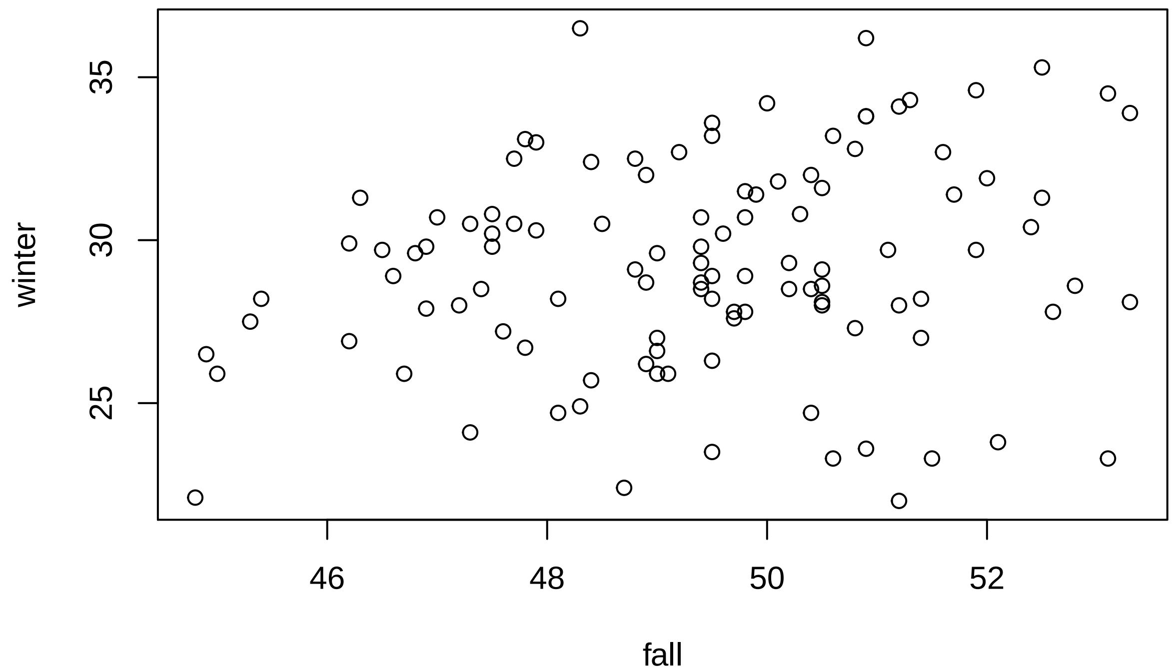
2.6

- 2.6.1

```r
summary(ftcollinstemp)
```

```
##       year           fall           winter     
##  Min.   :1900   Min.   :44.8   Min.   :22.00  
##  1st Qu.:1928   1st Qu.:47.9   1st Qu.:27.40  
##  Median :1955   Median :49.5   Median :29.10  
##  Mean   :1955   Mean   :49.4   Mean   :29.25  
##  3rd Qu.:1982   3rd Qu.:50.8   3rd Qu.:31.45  
##  Max.   :2010   Max.   :53.3   Max.   :36.50
```

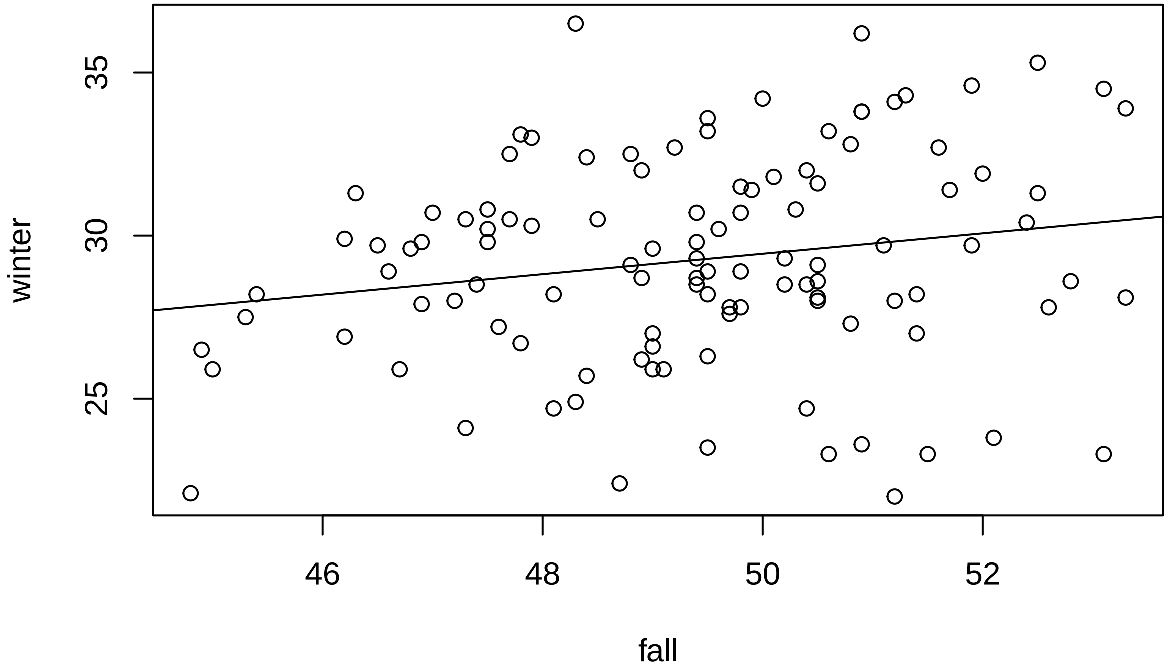
```r
plot(winter~fall, data = ftcollinstemp)
```



There is a very small linear trend in the plot. Most of the values are scattered especially when temperature in Fall becomes greater than 47 F. However most of time the temp in winter lies around 30 F.

- 2.6.2

```
plot(winter~fall, data = ftcollinstemp)
model1 <- lm(winter~fall, data = ftcollinstemp)
abline(model1)
```




```r
#Tesing null hypothesis
summary(model1)
```

```
##
## Call:
## lm(formula = winter ~ fall, data = ftcollinstemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

```r
t <- (0.3132 - 0) / 0.1528
p_values <- 2*pt(-t, 109)
print( p_values)
```

```
## [1] 0.04279025
```

As we can see p-value is less than 0.05. Thus we can reject NULL hypothesis.

- 2.6.3

```r
summary(model1)
```

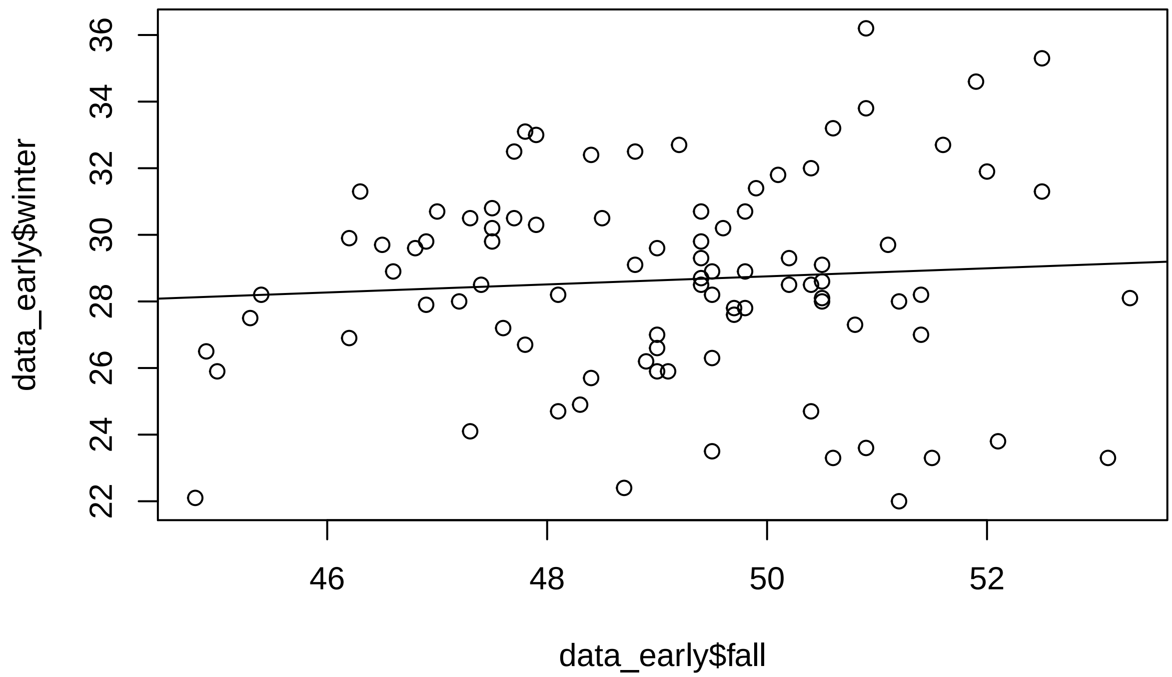
```
##
## Call:
## lm(formula = winter ~ fall, data = ftcollinstemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

From the summary, we can see that R-squared value is 0.0371. This shows that about 3% of variability in the observed values of Winter temp can be explained by the fall temp.

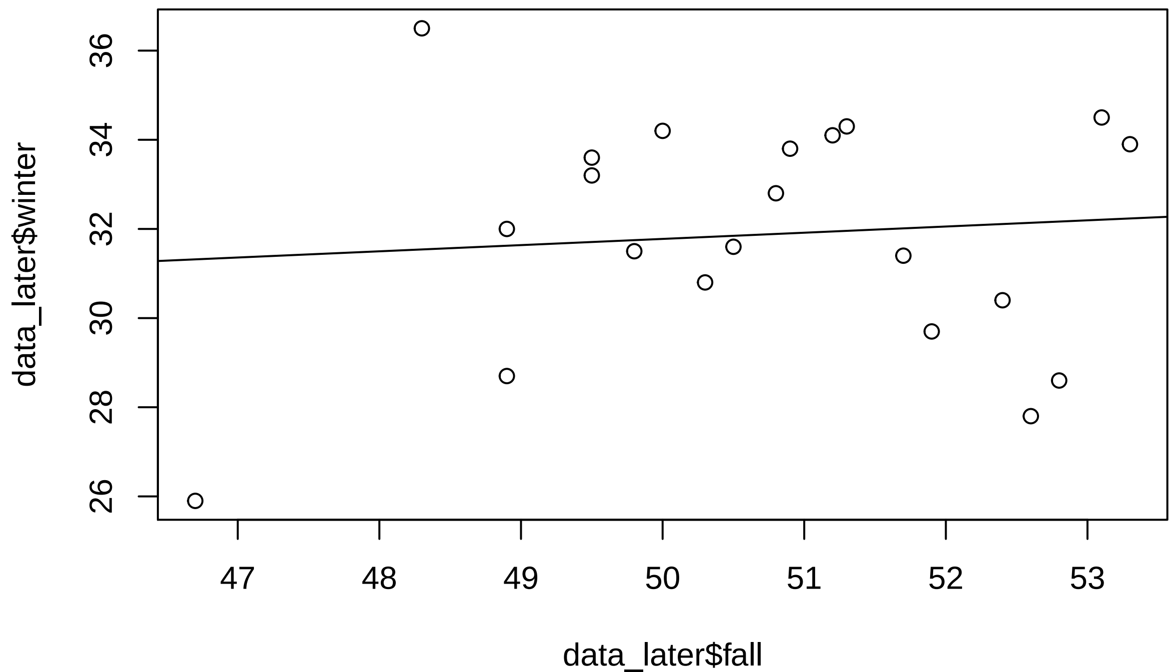
- 2.6.4

```r
data_early = ftcollinstemp[ftcollinstemp$year<1990,]
data_later = ftcollinstemp[ftcollinstemp$year>=1990,]

plot(data_early$winter ~ data_early$fall)
model2 <- lm(data_early$winter ~ data_early$fall)
abline(model2)
```




```r
plot(data_later$winter ~ data_later$fall)
model3 <- lm(data_later$winter ~ data_later$fall)
abline(model3)
```




```r
summary(model2)
```

```
##
## Call:
## lm(formula = data_early$winter ~ data_early$fall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8976 -1.6349  0.0118  2.0079  7.3387
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.7079     8.2600   2.749  0.00725 **
## data_early$fall   0.1209     0.1681   0.719  0.47397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.057 on 88 degrees of freedom
## Multiple R-squared:  0.005842,   Adjusted R-squared:  -0.005455
## F-statistic: 0.5171 on 1 and 88 DF,  p-value: 0.474
```

```r
summary(model3)
```

```
##
```

```
## Call:
## lm(formula = data_later$winter ~ data_later$fall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4174 -1.7097  0.3768  1.8988  4.9602
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.8260    17.7973   1.395    0.179
## data_later$fall   0.1390     0.3509   0.396    0.696
##
## Residual standard error: 2.699 on 19 degrees of freedom
## Multiple R-squared:  0.00819,    Adjusted R-squared:  -0.04401
## F-statistic: 0.1569 on 1 and 19 DF,  p-value: 0.6965
```

In case of model2(year < 1990), we have df $= 88$. Whereas in case of model3(Year > 1990) df $= 19$. There are very less data points in model3 as compared to model2. The R-squared value of model2 is better than model3 which can suggest that in model2 winter temp can be better explained by fall temp as compared to model3. However, in both models p-values are greater than 0.05, thus we fail to reject the null hypothesis i.e. accept null hypothesis.

2.16

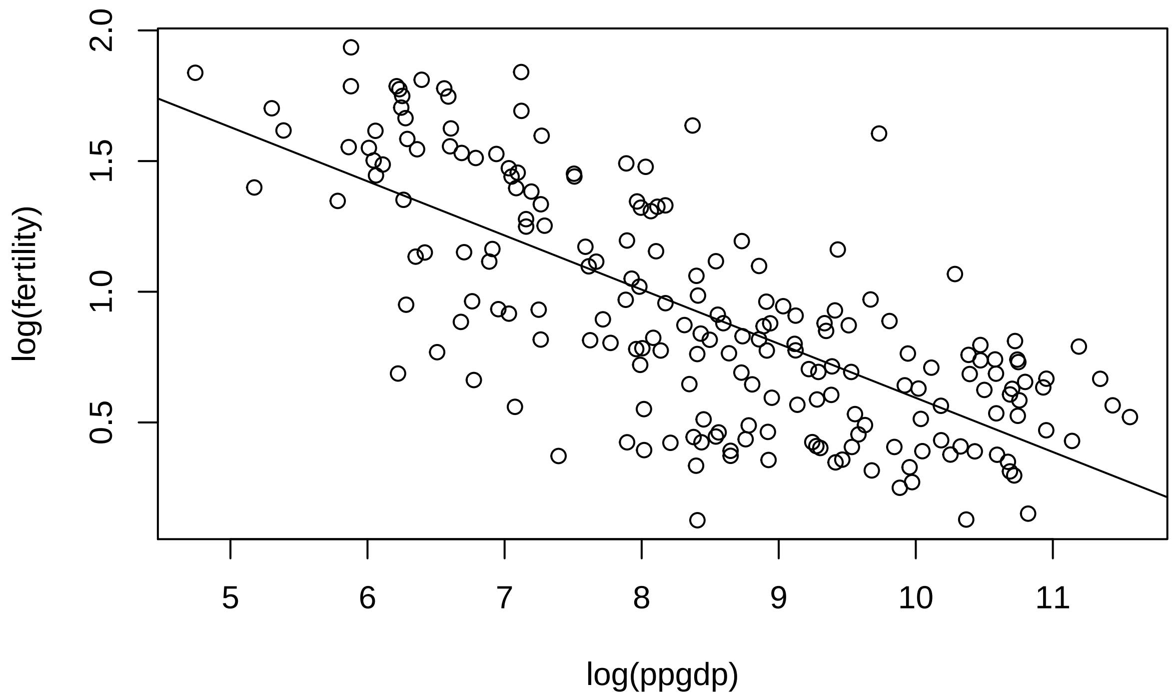
- 2.16.1

```r
model4 <- lm(log(fertility) ~ log(ppgdp), UN11)
print(model4)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)
##
## Coefficients:
## (Intercept)   log(ppgdp)
##      2.6655      -0.2071
```

- 2.16.2

```r
plot(log(fertility) ~ log(ppgdp), UN11)
abline(model4)
```



+ 2.16.3

```
summary(model4)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79828 -0.21639  0.02669  0.23424  0.95596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66551    0.12057   22.11   <2e-16 ***
## log(ppgdp)  -0.20715    0.01401  -14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 197 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.5236
## F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
t <- (-0.20715-0)/0.01401
p_values <- pt(-abs(t), 197)
```

As p-value is less than 0.05, we reject the null hypothesis

- 2.16.4

```r
summary(model4)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79828 -0.21639  0.02669  0.23424  0.95596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66551    0.12057   22.11   <2e-16 ***
## log(ppgdp)  -0.20715    0.01401  -14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 197 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.5236
## F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

Coefficient of Determination is 0.526 This states that with 52.6% of variability in observed values of log(fertility) can be explained by log(ppgdp)

- 2.16.5

```r
predict <- predict(model4, data.frame(ppgdp=1000),interval="prediction",level=0.95)
print(c(exp(predict[2]), exp(predict[3])))
```

```
## [1] 1.869889 6.317070
```

- 2.16.6

```r
print(UN11[UN11$fertility == max(UN11$fertility),][1])
```

```
##       region
## Niger Africa
```

```r
print(UN11[UN11$fertility == min(UN11$fertility),][1])
```

```
##                        region
## Bosnia and Herzegovina Europe
```

```r
# Two Largest Negative
print(sort(residuals(model4))[1])
```

```
## Bosnia and Herzegovina
##             -0.7982759
```

```r
print(sort(residuals(model4))[2])
```

```
##   Moldova
## -0.762329
```

```r
# Two Largest Positive
print(sort(residuals(model4))[198])
```

```
##    Angola
## 0.7047167
```

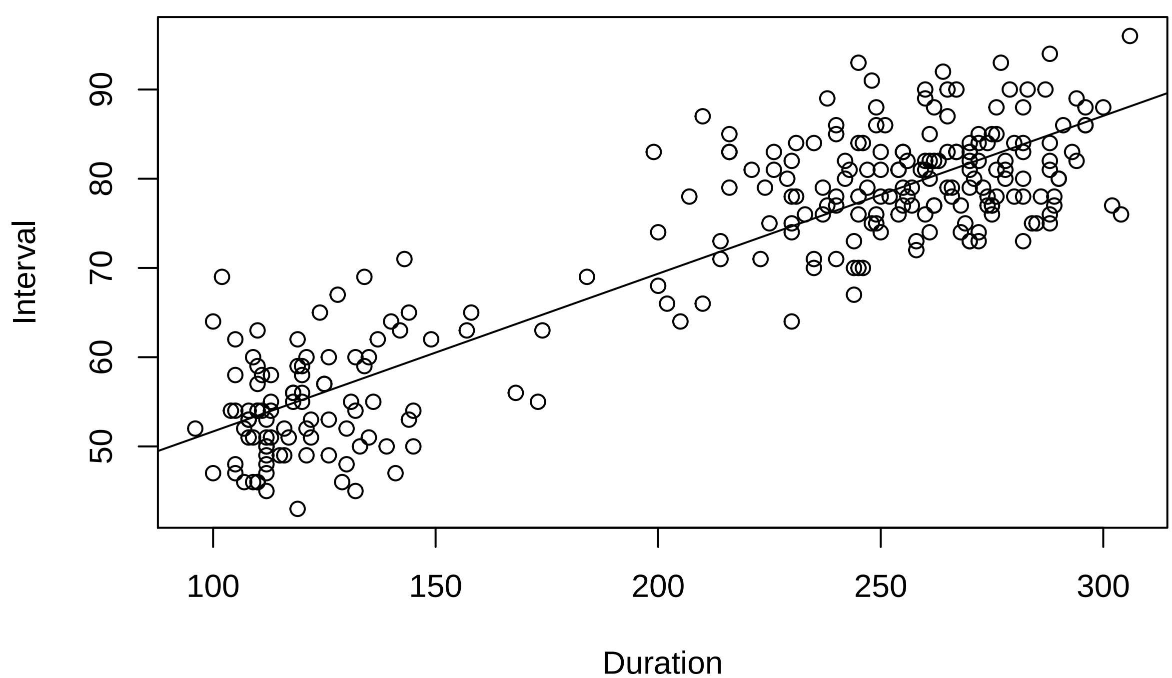
```r
print(sort(residuals(model4))[199])
```

```
## Equatorial Guinea
##         0.9559557
```

2.20

- 2.20.1

```r
plot(Interval ~ Duration, oldfaith)
model5 <- lm(Interval ~ Duration, oldfaith)
abline(model5)
```



```r
print(coefficients(model5))
```

```
## (Intercept)    Duration
##  33.9878076   0.1768629
```

The plot shows that the Interval(time to next eruption) increases linearly with the Duration of the current eruption. The slope states that the for 1 sec increase in duration the time to next eruption increases by 0.17 sec

- 2.20.2

```r
predict <- predict(model5, data.frame(Duration=250),interval="prediction",level=0.95)
print(predict)
```

```
##        fit      lwr      upr
## 1 78.20354 66.35401 90.05307
```

- 2.20.3

```r
predict <- predict(model5, data.frame(Duration=250),interval="prediction",level=0.90)
print(predict)[3]
```

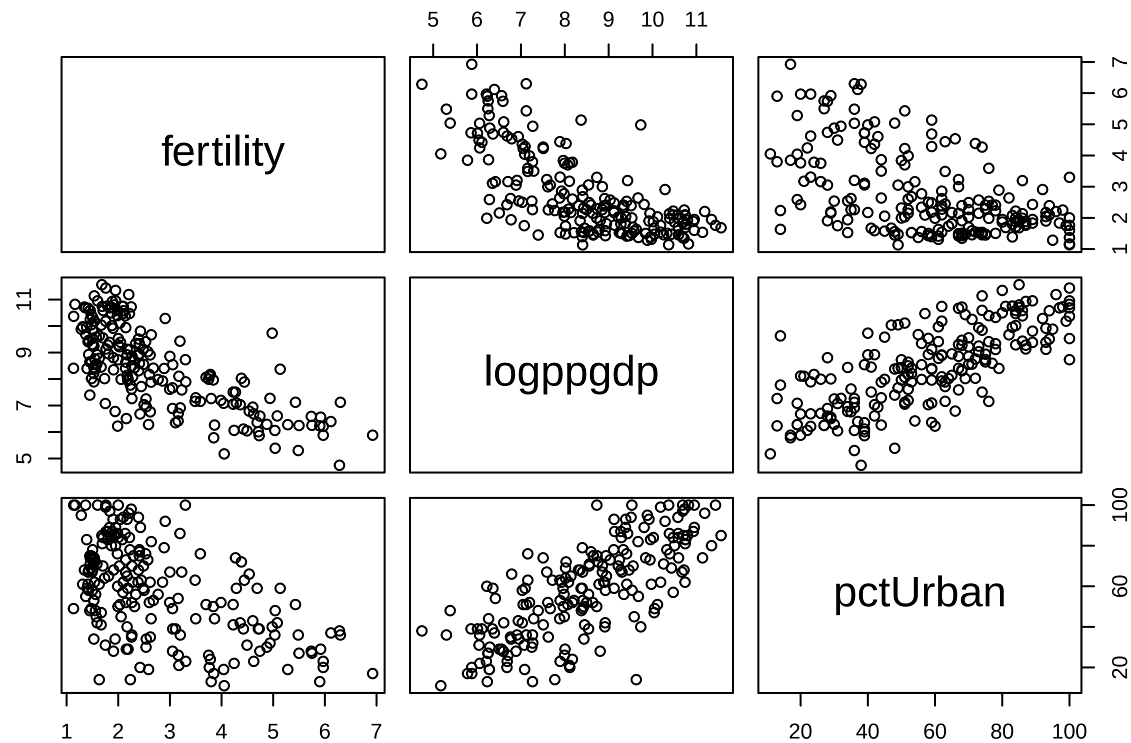
```
##        fit      lwr      upr
## 1 78.20354 68.26967 88.13741
```

```
## [1] 88.13741
```

3.2

- 3.2.1

```r
m <- data.frame(fertility = UN11$fertility, logppgdp = log(UN11$ppgdp), pctUrban = UN11$pctUrban)
plot(m)
```



From the scatter plot we can see that log(ppgdp) and pctUrban are strongly correlated with each other, and there is a plausibility for a good fit of a simple linear regression model. We can also see that fertility is more correlated to log(ppgdp) than to pctUrban.

- 3.2.2

```
model6 <- lm(UN11$fertility ~ log(UN11$ppgdp))
print(model6)
```

```
##
## Call:
## lm(formula = UN11$fertility ~ log(UN11$ppgdp))
##
## Coefficients:
##     (Intercept)  log(UN11$ppgdp)
##          8.0097          -0.6201
```

```
model7 <- lm(UN11$fertility ~ (UN11$pctUrban))
print(model7)
```

```
##
## Call:
## lm(formula = UN11$fertility ~ (UN11$pctUrban))
##
## Coefficients:
```

```
##   (Intercept)  UN11$pctUrban
##       4.55982       -0.03105
```

```
summary(model7)
```

```
##
## Call:
## lm(formula = UN11$fertility ~ (UN11$pctUrban))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4932 -0.7795 -0.1475  0.6517  2.9029
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.559823   0.213681  21.339   <2e-16 ***
## UN11$pctUrban -0.031045   0.003421  -9.076   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.128 on 197 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2913
## F-statistic: 82.37 on 1 and 197 DF,  p-value: < 2.2e-16
```

We can see that the slope coefficients are significantly different from 0. This is because p-value is too low and we reject NULL hypotesis.

- 3.2.3

```
model8 <- lm(fertility~pctUrban, data=UN11)
model9 <- lm(log(ppgdp)~pctUrban, data = UN11)
model10 <- lm(residuals(model8) ~ residuals(model9))
#summary(model8)
#summary(model9)
summary(model10)
```

```
##
## Call:
## lm(formula = residuals(model8) ~ residuals(model9))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.986e-16  6.596e-02   0.000        1
## residuals(model9) -6.151e-01  6.399e-02  -9.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3158
## F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

Looking at the R-squared value, we can say that log(ppgdp) explains 31.93% of remaining variability in fertility after adjusting for pctUrban. Thus log(ppgdp) is useful after adjusting for pctUrban.

```
model11 <- lm(fertility~log(ppgdp), data=UN11)
model12 <- lm(pctUrban~log(ppgdp), data = UN11)
model13 <- lm(residuals(model11) ~ residuals(model12))
#summary(model11)
#summary(model12)
summary(model13)
```

```
##
## Call:
## lm(formula = residuals(model11) ~ residuals(model12))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.313e-17  6.596e-02   0.000    1.000
## residuals(model12) -4.393e-04  4.255e-03  -0.103    0.918
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:  5.411e-05,  Adjusted R-squared:  -0.005022
## F-statistic: 0.01066 on 1 and 197 DF,  p-value: 0.9179
```

As R-squared value is very small, pctUrban is not useful for explaining remaining variability in fertility after adjusting for log(ppgdp).

```
model14 = lm(fertility~log(ppgdp)+pctUrban, data=UN11)
summary(model14)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
## log(ppgdp)  -0.6151425  0.0641565  -9.588   <2e-16 ***
## pctUrban    -0.0004393  0.0042656  -0.103    0.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9328 on 196 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5151
## F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

From the summary(model14) we can see that slope for log(ppgdp) and pctUrban is same as what we obtained from models - model10(log(ppgdp) after adjusting for pctUrban) and model13 (pctUrban after adjusting for log(ppgdp)) respectively. We can also see the coefficient of determination of model14 is similar to coefficient of model developed by fertility vs log(ppgdp). This suggests that addition of pctUrban variable to the regression isn't useful.

- 3.2.4

```r
coefficients(model14)
```

```
##   (Intercept)    log(ppgdp)      pctUrban
##  7.9932698831 -0.6151424675 -0.0004392792
```

```r
coefficients(model10)
```

```
##       (Intercept) residuals(model9)
##     -1.985664e-16     -6.151425e-01
```

From above we can say that estimated coefficient for log(ppgdp) is the same as the estimated slope in the added-variable plot for log(ppgdp) after pctUrban.

- 3.2.5

```r
isTRUE(all.equal(residuals(model14), residuals(model10)))
```

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

```r
isTRUE(all.equal(residuals(model14), residuals(model13)))
```

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

- 3.2.6

```r
summary(model14)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
## log(ppgdp)  -0.6151425  0.0641565  -9.588   <2e-16 ***
## pctUrban    -0.0004393  0.0042656  -0.103    0.918
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9328 on 196 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5151
## F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

```r
summary(model10)
```

```
##
## Call:
## lm(formula = residuals(model8) ~ residuals(model9))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.986e-16  6.596e-02   0.000        1
## residuals(model9) -6.151e-01  6.399e-02  -9.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3158
## F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

From the summary we can see that t-value for the coefficient for log(ppgdp) is not quite the same from the added-variable plot and from the regression with both regressors. The reason is because difference in degrees of freedom. In case of model with two regressors, we are taking 3 degrees of freedom out(two for prediction and one for response)(df = 199-3 = 196). On the other hand, in case of model with one regressor, we are taking 2 degrees of freedom out (199-2 = 197).