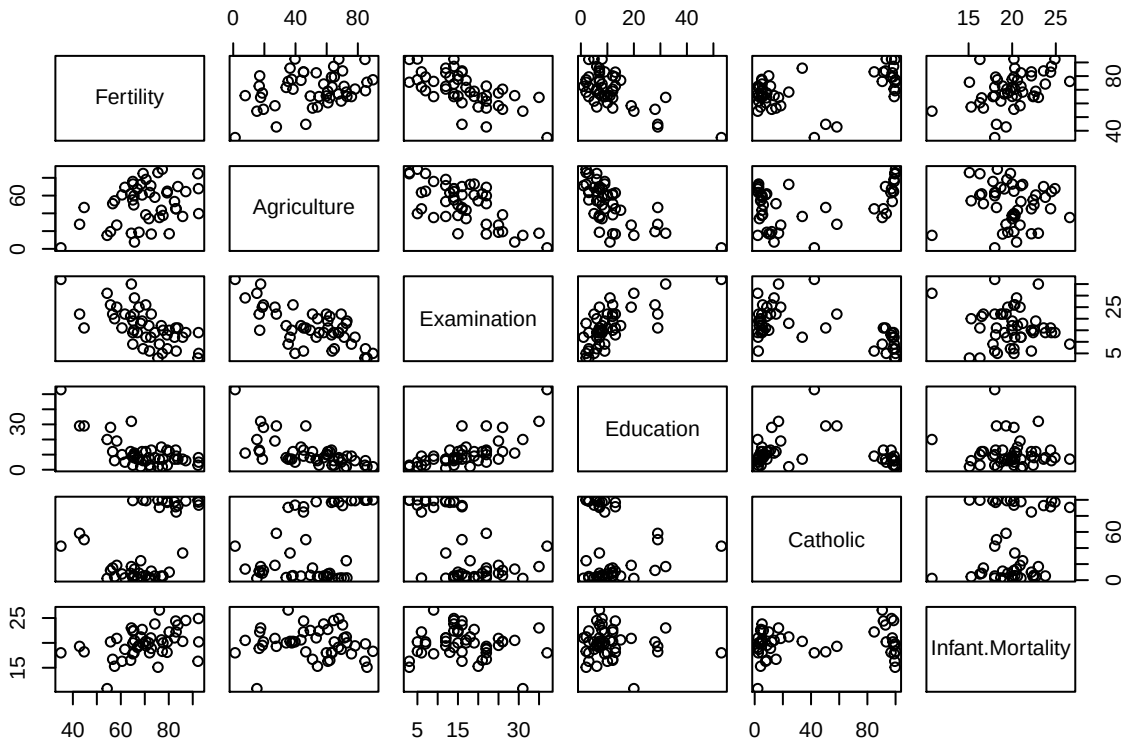
Lab2

Anya Conti

4/19/2017

```
data("swiss")
pairs(swiss)
```




```
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

```
?swiss
```

1. The sample size is 47.

```r
dim(swiss)
```

```
## [1] 47  6
```

```r
model.summaries <- data.frame(vars=c("Agriculture","Examination",
                                     "Education","Catholic","Infant.Mortality"),
                              MSE=NA,
                              coef.estimate=NA)

model.summaries
```

```
##               vars MSE coef.estimate
## 1      Agriculture  NA            NA
## 2      Examination  NA            NA
## 3        Education  NA            NA
## 4         Catholic  NA            NA
## 5 Infant.Mortality  NA            NA
```

```r
# fit all 5 models and fill in the table
for (i in 1:5)
{
  # make the formula
  model.formula <- paste("Fertility~",model.summaries$vars[i])
  fit <- lm(model.formula,data=swiss)

  # look at the anova
  print(anova(fit))

  # look at estimates
  print(summary(fit))

  # fill in the table
  model.summaries$MSE[i] <- anova(fit)[2,3]
  model.summaries$coef.estimate[i] <- fit$coef[2]
}
```

```
## Analysis of Variance Table
##
## Response: Fertility
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Agriculture  1  894.8  894.84  6.4089 0.01492 *
## Residuals   45 6283.1  139.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = model.formula, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5374  -7.8685  -0.6362   9.0464  24.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.30438    4.25126  14.185   <2e-16 ***
## Agriculture  0.19420    0.07671   2.532   0.0149 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 45 degrees of freedom
## Multiple R-squared:  0.1247, Adjusted R-squared:  0.1052
## F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
##
## Analysis of Variance Table
##
## Response: Fertility
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Examination  1 2994.4 2994.39  32.209 9.45e-07 ***
## Residuals   45 4183.6   92.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = model.formula, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9375  -6.0044  -0.3393   7.9239  19.7399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.8185     3.2576  26.651  < 2e-16 ***
## Examination  -1.0113     0.1782  -5.675 9.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.642 on 45 degrees of freedom
## Multiple R-squared:  0.4172, Adjusted R-squared:  0.4042
## F-statistic: 32.21 on 1 and 45 DF,  p-value: 9.45e-07
##
## Analysis of Variance Table
##
## Response: Fertility
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Education  1 3162.7  3162.7  35.446 3.659e-07 ***
## Residuals 45 4015.2    89.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = model.formula, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.036  -6.711  -1.011   9.526  19.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## Education    -0.8624     0.1448  -5.954 3.66e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
##
## Analysis of Variance Table
##
## Response: Fertility
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Catholic   1 1543.3 1543.29  12.325 0.001029 **
## Residuals 45 5634.7  125.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = model.formula, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.309  -4.060   0.511   6.851  16.682
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.42826    2.30510  27.950  < 2e-16 ***
## Catholic     0.13889    0.03956   3.511  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 45 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.1976
## F-statistic: 12.33 on 1 and 45 DF,  p-value: 0.001029
##
## Analysis of Variance Table
##
## Response: Fertility
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Infant.Mortality  1 1245.5 1245.51  9.4477 0.003585 **
## Residuals        45 5932.4  131.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = model.formula, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.672  -5.687  -0.381   7.239  28.565
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       34.5155    11.7113   2.947  0.00507 **
## Infant.Mortality   1.7865     0.5812   3.074  0.00359 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.48 on 45 degrees of freedom
## Multiple R-squared:  0.1735, Adjusted R-squared:  0.1552
## F-statistic: 9.448 on 1 and 45 DF,  p-value: 0.003585
```

2. Education has the smallest mse (mean squared error) closely followed by examination. MSE is the SSE (sum of squared errors) divided by n. Since n is the same for each model, the ratios of their mean squared errors compared with each other would be the same as the ratios of the sum of squared errors compared to each other. The SSE measures the unexplained variance of the dependent variable (fertility in this case) based on the model. Thus comparing the mse's helps to show which model leaves the most or least unexplained variance in fertility.

3.

```
model.summaries
```

```
##               vars       MSE coef.estimate
## 1      Agriculture 139.62480     0.1942017
## 2      Examination  92.96816    -1.0113173
## 3        Education  89.22746    -0.8623503
## 4         Catholic 125.21488     0.1388857
## 5 Infant.Mortality 131.83209     1.7864860
```

For every 1% increase of males in the province involved in agriculture as an occupation, there is an estimated 0.1942 unit (not specified) increase in fertility.
For every 1% increase of draftees in the province receiving the highest mark on army examination, there is an estimated 1.0113 unit (not specified) decrease in fertility.
For every 1% increase of education beyond primary school for draftees, there is an estimated 0.8624 unit (not specified) decrease in fertility.
For every 1% increase in population that is Catholic (vs Protestant), there is an estimated 0.1389 unit (not specified) increase in fertility.
For every 1% increase live births who live less than 1 year, there is an estimated 1.7865 unit (not specified) increase in fertility.

4. Here is the summary of both simple linear models. Both have p-values less than 0.0001 and are statistically significantly different from 0 at any confidence level typically used. Also, from question 2, they are the variables which leave the least unexplained variance in fertility.

```
summary(lm(Fertility ~ Examination , data = swiss))
```

```
##
## Call:
## lm(formula = Fertility ~ Examination, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9375  -6.0044  -0.3393   7.9239  19.7399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.8185     3.2576  26.651  < 2e-16 ***
## Examination  -1.0113     0.1782  -5.675 9.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.642 on 45 degrees of freedom
```

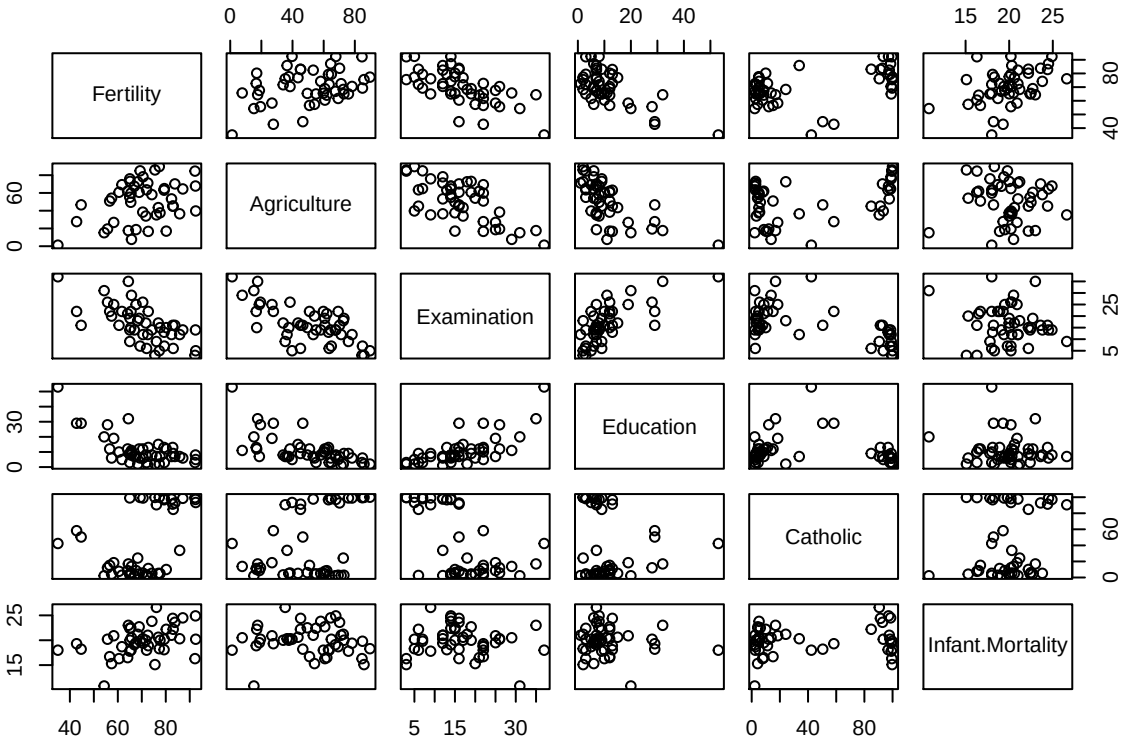
```
## Multiple R-squared:  0.4172, Adjusted R-squared:  0.4042
## F-statistic: 32.21 on 1 and 45 DF,  p-value: 9.45e-07
```

```
summary(lm(Fertility ~ Education , data = swiss))
```

```
##
## Call:
## lm(formula = Fertility ~ Education, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.036  -6.711  -1.011   9.526  19.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## Education    -0.8624     0.1448  -5.954 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

Based on the pairs plot, they appear to have a medium positive linear correlation. This brings up a question of multicollinearity.

```
pairs(swiss)
```



Their correlation coefficient is 0.6984, which again seems to point to a possible issue of multicollinearity.

```
cor(swiss)
```

```
##                  Fertility Agriculture Examination   Education   Catholic
```

```
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility              0.41655603
## Agriculture           -0.06085861
## Examination           -0.11402160
## Education             -0.09932185
## Catholic               0.17549591
## Infant.Mortality       1.00000000
```

Here is the model with both variables in it. Note that while both variables are still statistically significantly different from 0 at an alpha level of 0.05, the p-values have gone down by a lot (for examination 0.021 compared to 9.45e-7, and for education 0.0075 compared to 3.66e-7). In addition, while both coefficients are still negative, they are both much closer to 0 compared to before which indicates a smaller effect (for examination -0.5572 compared to -1.0113 before and for education -0.5395 compared to -0.8624 before).

```
Mod.EdEx <- lm(Fertility ~ Examination + Education, data = swiss)
summary(Mod.EdEx)
```

```
##
## Call:
## lm(formula = Fertility ~ Examination + Education, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9935  -6.8894  -0.3621   7.1640  19.2634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  85.2533     3.0855  27.630   <2e-16 ***
## Examination  -0.5572     0.2319  -2.402   0.0206 *
## Education    -0.5395     0.1924  -2.803   0.0075 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.982 on 44 degrees of freedom
## Multiple R-squared:  0.5055, Adjusted R-squared:  0.483
## F-statistic: 22.49 on 2 and 44 DF,  p-value: 1.87e-07
```

In addition, based on the anova table, the mean squared error of the model is only slightly smaller than the mean squared errors of the individual models. (80.67 compared to 89.227 and 92.968).

```
anova(Mod.EdEx)
```

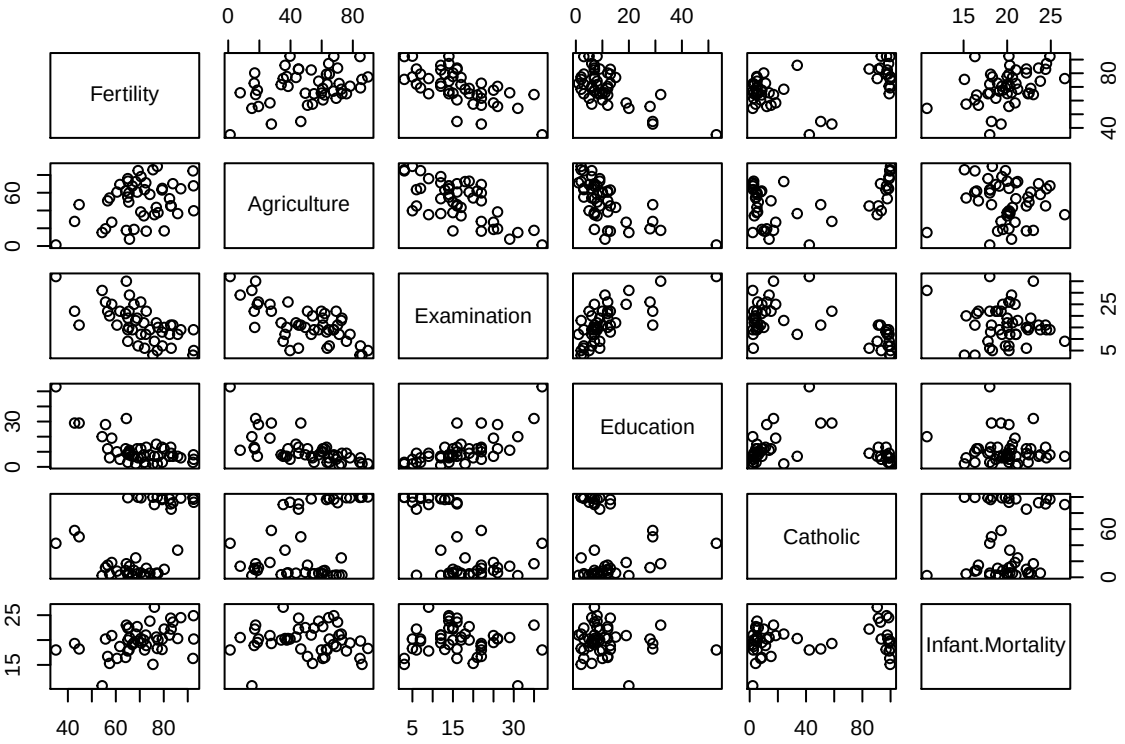
```
## Analysis of Variance Table
##
## Response: Fertility
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Examination  1 2994.4 2994.39 37.1176 2.468e-07 ***
## Education    1  634.0  633.96  7.8584  0.007497 **
## Residuals   44 3549.6   80.67
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is a problem of multicollinearity. The correlation between the variables means that some of the variance in fertility explained by one variable is also explained by the other. Therefore, including both variables in the model is almost redundant.

5. Based on the pairs plot, they appear to have a medium negative linear correlation.

```
pairs(swiss)
```



Their correlation coefficient is -0.63, so they are also fairly correlated.

```
cor(swiss)
```

```
##                  Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility              0.41655603
## Agriculture           -0.06085861
## Examination           -0.11402160
## Education             -0.09932185
## Catholic               0.17549591
## Infant.Mortality       1.00000000
```

In the simple linear, the coefficient is 0.1942 which is positive. Note: the p-value is 0.0149.

```
summary(lm(Fertility ~ Agriculture , data = swiss))
```

```
##
## Call:
```

```
## lm(formula = Fertility ~ Agriculture, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5374  -7.8685  -0.6362   9.0464  24.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.30438    4.25126  14.185   <2e-16 ***
## Agriculture  0.19420    0.07671   2.532   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 45 degrees of freedom
## Multiple R-squared:  0.1247, Adjusted R-squared:  0.1052
## F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
```

When education is added in, the coefficient for agriculture changes to -0.06648 (though the p-value this time is only 0.411 so not statistically significantly different from 0). Again, the correlation between the two variables creates this effect. Essentially, in the muliple regression model, the coefficient for agriculture is the effect on fertility of a 1% increase in males who are involved in agriculture HOLDING EDUCATION CONSTANT. So it compares what effect increased agriculture has on fertility for regions of similar education level.

```
Mod.AgEd <- lm(Fertility ~ Agriculture + Education, data = swiss)
summary(Mod.AgEd)
```

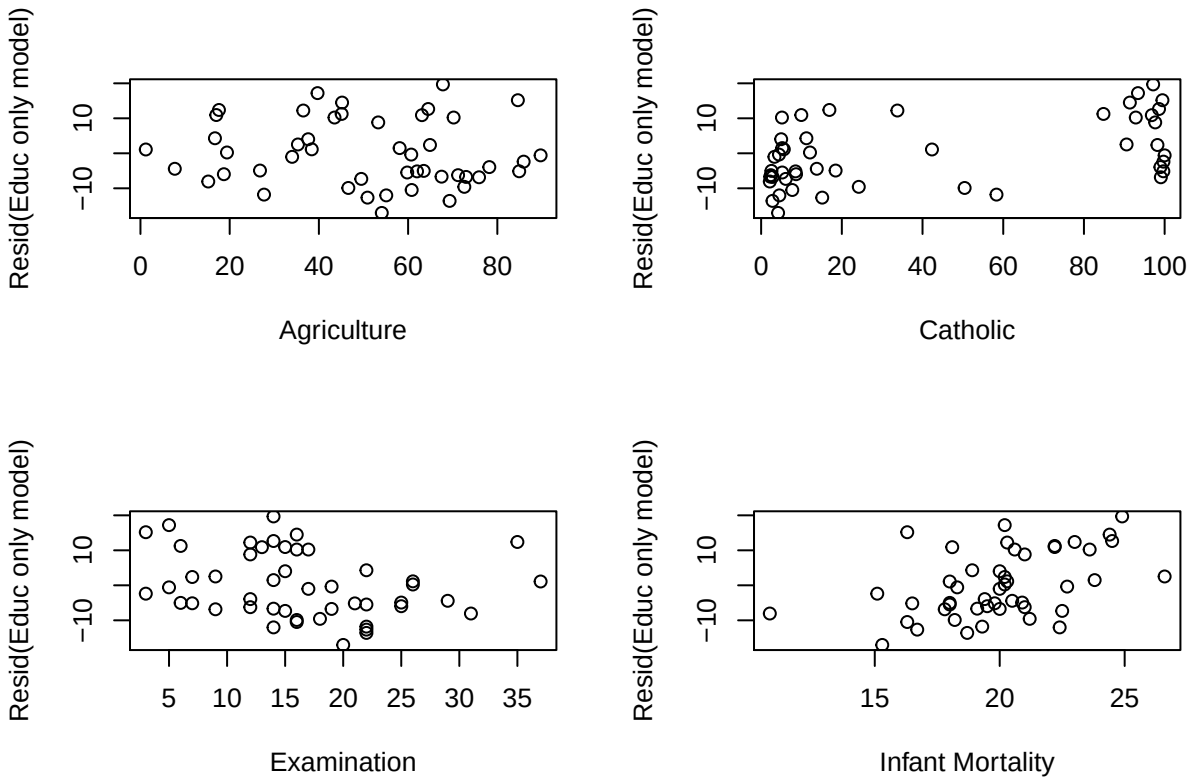
```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3072  -6.6157  -0.9443   8.7028  20.5291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.08005    5.78180  14.542  < 2e-16 ***
## Agriculture -0.06648    0.08005  -0.830    0.411
## Education   -0.96276    0.18906  -5.092  7.1e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.479 on 44 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.4242
## F-statistic: 17.95 on 2 and 44 DF,  p-value: 2e-06
```

Essentially, Agriculture might have a negative effect on fertility because more agriculture tends to also have less education, but the change in fertility might actually be due to education, so this model accounts for just looking at the effect of agriculture holding education constant.

6. The residual of the education only model seems be somewhat correlated with Catholic and Infant Mortality which means those variables can predict the unexplained variance of fertility left after education. Neither Examination or Agriculture seem to have an apparent correlation with the residuals.

```
par(mfrow=c(2,2))
fit.Educ <- lm(Fertility~Education,data=swiss)
plot(swiss$Agri,resid(fit.Educ),xlab="Agriculture",ylab="Resid(Educ only model)")
```

```r
plot(swiss$Cath,resid(fit.Educ),xlab="Catholic",ylab="Resid(Educ only model)")
plot(swiss$Exam,resid(fit.Educ),xlab="Examination",ylab="Resid(Educ only model)")
plot(swiss$Infa,resid(fit.Educ),xlab="Infant Mortality",ylab="Resid(Educ only model)")
```



7. This model helps to explain the influence of each variable on fertility rate, holding the other variables constant.

```r
Mod.NoEx <- lm(Fertility ~ Education + Agriculture + Catholic + Infant.Mortality , data=swiss)
summary(Mod.NoEx)
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Agriculture + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
```

```
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

```
anova(Mod.NoEx)
```

```
## Analysis of Variance Table
##
## Response: Fertility
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## Education          1 3162.7  3162.7 61.5523 9.206e-10 ***
## Agriculture        1   62.0    62.0  1.2060   0.27839
## Catholic           1 1385.4  1385.4 26.9622 5.686e-06 ***
## Infant.Mortality   1  409.8   409.8  7.9757   0.00722 **
## Residuals         42 2158.1    51.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. Since the slope of the line is closer to a slope of 0 than a slope of 1, it shows less extreme residuals (closer to 0) in general compared to the residual of the model with just education.

```
plot(resid(fit.Educ),resid(lm(Fertility~Agriculture+Education+Catholic+Infant.Mortality,data=swiss)),
     xlab="resid(Educ Only)",ylab="resid(All but Exam)",
     xlim=c(-25,25),ylim=c(-25,25)) # Note that the last line makes the x and y axes the same
abline(a=0,b=1) # This adds a line with intercept = 0 and slope = 1.
```