Thursday Exam 2 - STAT 324 (Swiss Dataset)

## **Overview of the swiss Dataset**

To answer the first five questions on the test, you can rely entirely on the regression modeling and selection procedures implemented in this packet using the swiss dataset. Begin by reviewing the overview and variable selection sections to understand the context and role of each variable—particularly that Fertility is the response variable and all others are quantitative predictors. This structure supports your ability to interpret regression coefficients and evaluate model utility based on the variables’ relationships.

The swiss dataset is a built-in R dataset containing standardized fertility measure and socio-economic indicators for 47 French-speaking provinces of Switzerland around 1888. All variables are numeric, and there are no missing values, making it convenient for regression analysis.

Our primary goal is to predict the Fertility measure based on other socio-economic indicators.

**Variable Selection**

We will use the following variables from the swiss dataset:

* Fertility (Common standardized fertility measure) - **Response Variable**
* Agriculture (% of males involved in agriculture as occupation) - Predictor
* Examination (% draftees receiving highest mark on army examination) - Predictor
* Education (% education beyond primary school for draftees) - Predictor
* Catholic (% ‘catholic’ (as opposed to ‘protestant’)) - Predictor
* Infant.Mortality (live births who live less than 1 year) - Predictor

## Preliminary Analysis

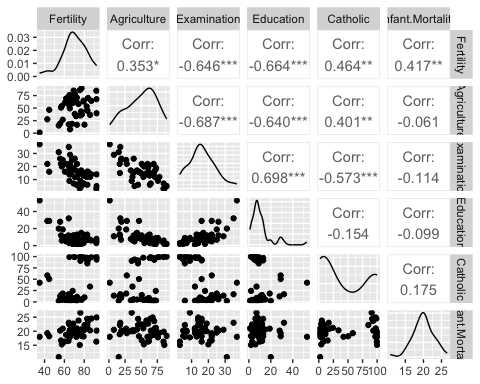
### **1. Data Summaries**

swiss\_df <- swiss   
summary(swiss\_df)

Fertility Agriculture Examination Education   
 Min. :35.00 Min. : 1.20 Min. : 3.00 Min. : 1.00   
 1st Qu.:64.70 1st Qu.:35.90 1st Qu.:12.00 1st Qu.: 6.00   
 Median :70.40 Median :54.10 Median :16.00 Median : 8.00   
 Mean :70.14 Mean :50.66 Mean :16.49 Mean :10.98   
 3rd Qu.:78.45 3rd Qu.:67.65 3rd Qu.:22.00 3rd Qu.:12.00   
 Max. :92.50 Max. :89.70 Max. :37.00 Max. :53.00   
 Catholic Infant.Mortality  
 Min. : 2.150 Min. :10.80   
 1st Qu.: 5.195 1st Qu.:18.15   
 Median : 15.140 Median :20.00   
 Mean : 41.144 Mean :19.94   
 3rd Qu.: 93.125 3rd Qu.:21.70   
 Max. :100.000 Max. :26.60

### **2. Visualizations**

ggpairs(swiss\_df)



### **3. Regular Model — MLR (Multiple Linear Regression)**

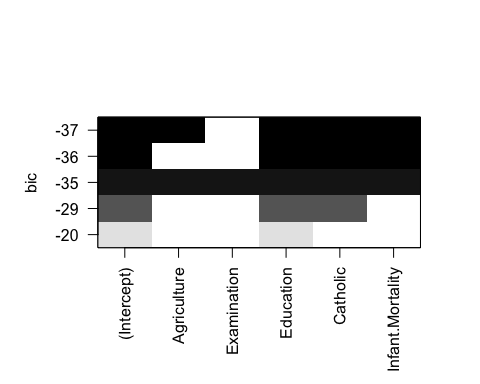
Fit a multiple linear regression model with Fertility as the response and all other selected predictors.

model\_mlr\_swiss <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data = swiss\_df)  
summary(model\_mlr\_swiss)  
Call:  
lm(formula = Fertility ~ Agriculture + Examination + Education +   
 Catholic + Infant.Mortality, data = swiss\_df)  
Residuals:  
 Min 1Q Median 3Q Max   
-15.2743 -5.2617 0.5032 4.1198 15.3213   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 66.91518 10.70604 6.250 1.91e-07 \*\*\*  
Agriculture -0.17211 0.07030 -2.448 0.01873 \*   
Examination -0.25801 0.25388 -1.016 0.31546   
Education -0.87094 0.18303 -4.758 2.43e-05 \*\*\*  
Catholic 0.10412 0.03526 2.953 0.00519 \*\*   
Infant.Mortality 1.07705 0.38172 2.822 0.00734 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 7.165 on 41 degrees of freedom  
Multiple R-squared: 0.7067, Adjusted R-squared: 0.671   
F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10

### **4. BIC Best Subset**

We use regsubsets to find the best subset of predictors based on the Bayesian Information Criterion (BIC).

regfit\_full\_swiss <- regsubsets(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data = swiss\_df, nvmax = 5) # Max number of predictors  
  
plot(regfit\_full\_swiss)



### **5. AIC Stepwise Regression**

Perform stepwise model selection using AIC, starting from the full model.

initial\_model\_swiss <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data = swiss\_df)  
  
step\_mod\_swiss <- MASS::stepAIC(initial\_model\_swiss, direction = "both", trace = TRUE)

Start: AIC=190.69  
Fertility ~ Agriculture + Examination + Education + Catholic +   
 Infant.Mortality  
  
 Df Sum of Sq RSS AIC  
- Examination 1 53.03 2158.1 189.86  
<none> 2105.0 190.69  
- Agriculture 1 307.72 2412.8 195.10  
- Infant.Mortality 1 408.75 2513.8 197.03  
- Catholic 1 447.71 2552.8 197.75  
- Education 1 1162.56 3267.6 209.36  
  
Step: AIC=189.86  
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality  
  
 Df Sum of Sq RSS AIC  
<none> 2158.1 189.86  
+ Examination 1 53.03 2105.0 190.69  
- Agriculture 1 264.18 2422.2 193.29  
- Infant.Mortality 1 409.81 2567.9 196.03  
- Catholic 1 956.57 3114.6 205.10  
- Education 1 2249.97 4408.0 221.43

## **Analysis of Best Model**

### **Evaluating of Best Model**

best\_model\_swiss <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality, data = swiss\_df)  
summary(best\_model\_swiss)

Call:  
lm(formula = Fertility ~ Agriculture + Education + Catholic +   
 Infant.Mortality, data = swiss\_df)  
  
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 \*\*\*  
Agriculture -0.15462 0.06819 -2.267 0.02857 \*   
Education -0.98026 0.14814 -6.617 5.14e-08 \*\*\*  
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

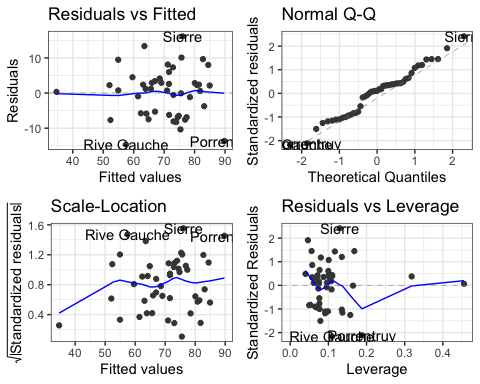
### **Evaluating Variance Inflation Factor**

vif(best\_model\_swiss)

Agriculture Education Catholic Infant.Mortality   
 2.147153 1.816361 1.299916 1.107528

### **Evaluating Residual Plots**

autoplot(best\_model\_swiss) +   
 theme\_bw()



### **Evaluating Outliers, Leverage and Influential Points**

# Add standardized residuals  
augmented\_df <- augment(best\_model\_swiss) ## from broom package  
rownames(augmented\_df) <- NULL  
augmented\_df %>%   
 dplyr::select(-`.rownames`, -Fertility,-Agriculture,-Education,-Catholic,-Infant.Mortality) %>%  
 arrange(desc(.cooksd)) %>%   
 head(10) %>%   
 round(3)

# A tibble: 10 × 6  
 .fitted .resid .hat .sigma .cooksd .std.resid  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 89.8 -13.7 0.188 6.86 0.207 -2.12  
 2 76.1 16.1 0.13 6.73 0.174 2.41  
 3 57.5 -14.7 0.109 6.84 0.115 -2.17  
 4 54.9 9.48 0.168 7.07 0.085 1.45  
 5 74.0 -8.26 0.157 7.12 0.059 -1.25  
 6 82.7 9.66 0.117 7.08 0.054 1.43  
 7 84.5 8.01 0.135 7.13 0.045 1.20  
 8 52.4 -7.68 0.137 7.14 0.042 -1.15  
 9 75.4 -10.4 0.079 7.06 0.039 -1.51  
10 63.5 13.4 0.046 6.93 0.035 1.91