Stat 6021: Homework Set 6

Tom Lever

10/13/22

1. For this first question, you will use the data set **swiss** which is part of the **datasets** package. Load the data. For more information about the data set, type **?swiss**. This data set encapsulates a standardized Fertility measure and socioeconomic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

```
data_set <- swiss
nrow(data_set)
```

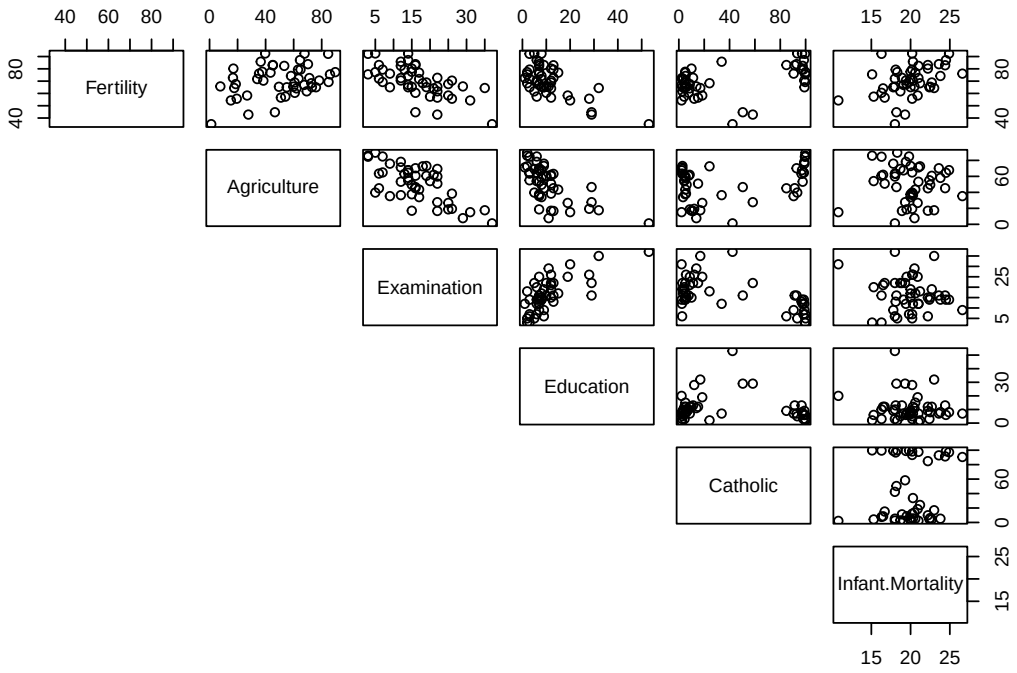
```
## [1] 47
```

```
head(data_set, n = 3)
```

```
##              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
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
```

   (a) Create a scatterplot matrix and find the correlation between all pairs of variables for this data set. Answer the following questions based on the output:

```
pairs(data_set, lower.panel = NULL)
```




```
correlation_matrix <- round(cor(data_set), 3)
correlation_matrix
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility            1.000       0.353      -0.646    -0.664    0.464
## Agriculture          0.353       1.000      -0.687    -0.640    0.401
## Examination         -0.646      -0.687       1.000     0.698   -0.573
## Education           -0.664      -0.640       0.698     1.000   -0.154
## Catholic             0.464       0.401      -0.573    -0.154    1.000
## Infant.Mortality     0.417      -0.061      -0.114    -0.099    0.175
##                  Infant.Mortality
## Fertility                   0.417
## Agriculture                -0.061
## Examination                -0.114
## Education                  -0.099
## Catholic                    0.175
## Infant.Mortality            1.000
```

i. Which predictors appear to be linear related to the Fertility measure?

According to Keith G. Calkins (https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm), "[Linear] correlation coefficients whose magnitude[s] are between 0.9 and 1.0 indicate variables which can be considered very highly [linearly] correlated. Correlation coefficients whose magnitude[s] are between 0.7 and 0.9 indicate variables which can be considered highly correlated. Correlation coefficients whose magnitude[s] are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude[s] are between 0.3 and 0.5 indicate variables which have... low correlation[s]. Correlation coefficients whose magnitude[s] are less than 0.3 have little if any (linear) correlation."

```
ifelse(abs(correlation_matrix) > 0.3, correlation_matrix, "")
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility        "1"       "0.353"     "-0.646"    "-0.664"  "0.464"
## Agriculture      "0.353"   "1"         "-0.687"    "-0.64"   "0.401"
## Examination      "-0.646"  "-0.687"    "1"         "0.698"   "-0.573"
## Education        "-0.664"  "-0.64"     "0.698"     "1"       ""
## Catholic         "0.464"   "0.401"     "-0.573"    ""        "1"
## Infant.Mortality "0.417"   ""          ""          ""        ""
##                  Infant.Mortality
## Fertility        "0.417"
## Agriculture      ""
## Examination      ""
## Education        ""
## Catholic         ""
## Infant.Mortality "1"
```

Fertility has a low correlation with Agriculture, Catholic, and Infant.Mortality, a moderate correlation with Examination and Education, and a perfect correlation with itself.

ii. Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

```
ifelse(abs(correlation_matrix) > 0.7, correlation_matrix, "")
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility        "1"       ""          ""          ""        ""
## Agriculture      ""        "1"         ""          ""        ""
## Examination      ""        ""          "1"         ""        ""
## Education        ""        ""          ""          "1"       ""
## Catholic         ""        ""          ""          ""        "1"
## Infant.Mortality ""        ""          ""          ""        ""
##                  Infant.Mortality
## Fertility        ""
## Agriculture      ""
## Examination      ""
## Education        ""
## Catholic         ""
## Infant.Mortality "1"
```

All predictors are perfectly correlated with themselves. No other predictors and highly or very highly correlated.

```
ifelse(abs(correlation_matrix) > 0.5, correlation_matrix, "")
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility        "1"       ""          "-0.646"    "-0.664"  ""
## Agriculture      ""        "1"         "-0.687"    "-0.64"   ""
## Examination      "-0.646"  "-0.687"    "1"         "0.698"   "-0.573"
## Education        "-0.664"  "-0.64"     "0.698"     "1"       ""
## Catholic         ""        ""          "-0.573"    ""        "1"
## Infant.Mortality ""        ""          ""          ""        ""
##                  Infant.Mortality
## Fertility        ""
## Agriculture      ""
## Examination      ""
## Education        ""
```

```
## Catholic         ""
## Infant.Mortality "1"
```

Agriculture and Examination are moderately correlated. Agriculture and Education are moderately correlated. Examination and Education are moderately correlated. Examination and Catholic are moderately correlated.

(b) Fit a multiple linear regression model with the Fertility measure as the response variable and all other variables as predictors. Use the `summary` function to obtain the estimated coefficients and results from the various hypothesis tests for this model.

```
library(TomLeversRPackage)
linear_model <- lm(Fertility ~ ., data = data_set)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = data_set)
##
## 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
##
## E(y | x) =
##     B_0 +
##     B_Agriculture * Agriculture +
##     B_Examination * Examination +
##     B_Education * Education +
##     B_Catholic * Catholic +
##     B_Infant.Mortality * Infant.Mortality
## E(y | x) =
##     66.9151816789687 +
##     -0.172113970941456 * Agriculture +
##     -0.258008239834725 * Examination +
##     -0.870940062939424 * Education +
##     0.104115330743767 * Catholic +
##     1.07704814069099 * Infant.Mortality
## Number of observations: 47
## Estimated variance of errors: 51.3425104986362
## Multiple R:  0.840675324719792   Adjusted R:  0.819128181298091
## Critical value t(alpha/2 = 0.05/2, DFRes = 41): 2.01954097044138
```

i. What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

```
analyze_variance(linear_model)
```

```
## Analysis of Variance Table
##
## Response: Fertility
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## Agriculture        1  894.84  894.84 17.4288 0.0001515 ***
## Examination        1 2210.38 2210.38 43.0516 6.885e-08 ***
## Education          1  891.81  891.81 17.3699 0.0001549 ***
## Catholic           1  667.13  667.13 12.9937 0.0008387 ***
## Infant.Mortality   1  408.75  408.75  7.9612 0.0073357 **
## Residuals         41 2105.04   51.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## DFR: 5, SSR: 5072.91196317299, MSR: 1014.5823926346
## F0: 19.761059262218, F(alpha = 0.05, DFR = 5, DFRes = 41): 2.44342860009504
## p: 5.59379854113429e-10
## DFT: 46, SST: 7177.95489361708
## R2: 0.706735001592728, Adjusted R2: 0.670970977396719
## Number of observations: 47
```

We assume that errors are random, are independent, and follow a normal distribution with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$. The multiple linear regression model from part 1b is useful in predicting Fertility if at least one of the predictors Agriculture, Examination, Education, Catholic, Infant.Mortality contributes significantly to the model. We conduct a test of the null hypothesis $H_0 : \beta_{Agriculture} = \beta_{Examination} = \beta_{Education} = \beta_{Catholic} = \beta_{Infant.Mortality} = 0$ (i.e., all coefficients of the predictors are 0). The alternate hypothesis is $H_1 : \beta_{predictor} \neq 0$ for at least one predictor (i.e., at least one of the coefficients of the predictors is not 0). If we reject the null hypothesis, at least one predictor contributes significantly to the model.

```
significance_level <- 0.05
test_null_hypothesis_involving_MLR_coefficients(
    linear_model,
    significance_level
)
```

```
## Since probability 5.59379854113429e-10 is less than significance level 0.05,
## we reject the null hypothesis.
## We have sufficient evidence to support the alternate hypothesis.
```

Alternatively, we can test whether a test statistic F_0 is greater than a critical value $F_{\alpha=0.05,\ df_R=5,\ df_{Res}=41}$. This test statistic is the ratio of the regression mean square to the residual mean square for our linear model. If the result of this test is true, then it is likely that at least one coefficient of a predictor is not equal to 0.

Since the test statistic $F_0 = 19.761$ is greater than the critical value $F_{\alpha=0.05,\ df_R=5,\ df_{Res}=41} = 2.443$, we reject the null hypothesis and support the alternate hypothesis. Since we reject the null hypothesis and support the alternate hypothesis, at least one of the predictor variables contributes significantly to the model. Since at least one of the predictors contributes significantly to the model, the model is useful in predicting Fertility.

ii. Look at the numerical values of the estimated slopes as well as the probabilities. Do they seem to agree with or contradict what you had written in your answer to part 1a? Briefly explain what is going on here.

Examination has the most moderate correlations with other predictors, has a correlation with Fertility close to the correlation between Education and Fertility, and has a probability greater than a significance level 0.05 / may be removed from our linear model; these patterns may be in agreement. Agriculture has the lowest magnitude of correlation coefficient with Fertility and after Examination has the highest probability / lowest significance; these patterns may be in agreement. Education has the highest magnitude of correlation coefficient with Fertility and has the lowest probability / most significance to a multiple linear model / in the presence of all predictors; these patterns may be in agreement. All predictors are correlated to Fertility; all predictors are significant to the multiple linear model, except for Examination, which has the most moderate correlations with other predictors, and may be eliminated from the multiple linear model; these patterns may be in agreement.

The estimated coefficient for Agriculture of -0.172 ostensibly disagrees with the low positive correlation between Fertility and Agriculture, though this coefficient indicates the degree to which Fertility changes in the context of this multiple linear regression model and in the presence of all predictors, the low positive correlation indicates a poor positive linear relationship between Fertility and Agriculture, and Agriculture has the lowest significance to the linear model. The estimated coefficient for Examination of -0.258 may agree with the moderate negative correlation between Fertility and Examination, given that Examination has the most moderate correlations with other predictors. The estimated coefficient for Education of -0.870 may agree with the moderate negative correlation between Fertility and Education, given that Education has the highest magnitude of correlation coefficient with Fertility and lowest probability / most significance. The estimated coefficient for Catholic of 0.104 may agree with low positive correlation between Fertility and Catholic, given that Catholic has the highest magnitude of correlation coefficient with Fertility after Education and Examination and highest probability / most significance after Education and Examination. The estimated coefficient for Infant.Mortality of 1.077 may with low positive correlation between Fertility and Infant.Mortality. All predictors appear to have nonzero coefficients and to be linearly related to Fertility.

2. Data from $n = 113$ hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital. The response variable is $InfctRsk$, the percentage of patients who get an infection while hospitalized. The predictors are $Stay$, the average length of stay, Age, the average patient age, $Xrays$, a measure of how many X-rays are done in the hospital, and $Services$, a measure of how many different services the hospital offers. We consider the following multiple regression equation: $E(InfctRsk) = \beta_0 + \beta_1 Stay + \beta_2 Age + \beta_3 Xrays + \beta_4 Services$. Some R output is shown in the prompt for this homework. You may assume that the regression assumptions are met.

   (a) What is the value of the estimated coefficient of the variable $Stay$? Write a sentence that interprets this value.

   The value of the estimated coefficient of the variable $Stay$ $\hat{\beta}_1 = 0.237209$. According to the estimated multiple linear regression model corresponding to the R output, for an increase in predictor $Stay$ / the average length of stay of one unit, the predictor $InfctRsk$ / the percentage of patients who get an infection while hospitalized increases by 0.237209.

   (b) Derive the test statistic, p-value, and critical value for the variable Age. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable Age?

   The hypotheses for testing the significance of regression coefficient β_{Age} are $H_0 : \beta_{Age} = 0$ (i.e., the regression coefficient corresponding to predictor Age is 0), and $H_1 : \beta_{Age} \neq 0$ (i.e., the regression coefficient corresponding to predictor Age is not 0). If the null hypothesis is rejected, the predictor Age is significant to the multiple linear regression model and cannot be deleted from the model. If the null hypothesis is not rejected, the predictor Age is insignificant to the multiple linear regression

model and can be deleted from the model. The test statistic for testing the null hypothesis is

$$t_0 = \frac{\hat{\beta}_{Age}}{SE\left(\hat{\beta}_{Age}\right)} = \frac{-0.014071}{0.022708} = -0.619$$

```
test_statistic <- -0.014071 / 0.022708
test_statistic
```

```
## [1] -0.6196495
```

We test at significance level $\alpha = 0.05$. The null hypothesis is rejected if the magnitude $|t_0|$ of the test statistic is greater than a critical value

$$t_{\alpha/2=0.05/2,\ df_{Res}=n-p=113-5=108} = 1.982$$

```
significance_level <- 0.05
number_of_confidence_intervals <- 1
number_of_observations <- 113
number_of_variables <- 5
residual_degrees_of_freedom <- number_of_observations - number_of_variables
critical_value <- calculate_critical_value_t(
    significance_level,
    number_of_confidence_intervals,
    residual_degrees_of_freedom
)
critical_value
```

```
## [1] 1.982173
```

Since the magnitude of the test statistic is less than the critical value, we fail to reject the null hypothesis. The predictor *Age* is insignificant to the multiple linear regression model and can be deleted from the model.

Alternately, the null hypothesis is rejected if the probability that a random test statistic from a Student's t distribution with degrees of freedom df_{Res} is less than the negative magnitude of our test statistic or greater than the positive magnitude of our test statistic (i.e., the p-value) is less than our significance level.

```
calculate_p_value_from_t_statistic_and_residual_degrees_of_freedom(
    test_statistic,
    residual_degrees_of_freedom
)
```

```
## [1] 0.5367937
```

Since the p-value is greater than our significance level, we fail to reject the null hypothesis. The predictor *Age* is insigificant to the multiple linear regression model and can be deleted from the model.

(c) A classmate states: "The variable *Age* is not linearly related to the predicted infection risk". Do you agree with your classmate's statement? Briefly explain.

I disagree. With examining a scatterplot of infection risk versus *Age*, we do not know whether or not the predictor *Age* is linearly related to the response $InfctRsk$. The above null-hypothesis test result allows us to conclude only that the predictor *Age* is insignificant to the multiple linear regression and can be deleted from the model in the presence of all predictors.

(d) Using the Bonferroni method, construct 95-percent joint confidence intervals for β_1, β_2, and β_3.

Let significance level $\alpha = 0.05$.

Let series / family confidence level $CL_f = 1 - \alpha = 0.95$ be the proportion of families of confidence-level estimates that have the characteristic that all confidence intervals in the family contain the population parameter when repeated samples are selected and families of confidence intervals are calculated for each sample.

Analysis of data frequently consists of constructing a series / family of joint confidence-interval estimates with the characteristic that CL_f of families of confidence intervals have the characteristic that all confidence intervals in the family contain the population parameter when repeated samples are selected and families of confidence intervals are calculated for each sample.

If we want a family confidence level to be at least $CL_f = 1 - \alpha$ for a family of g confidence intervals, each confidence interval for the coefficient β_i of the linear model for a population is constructed at $CL_1 = 1 - \alpha/f$ as

$$CI_i = \left[\hat{\beta}_i - t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_i\right),\ \hat{\beta}_i + t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_i\right)\right]$$

We construct 95-percent joint confidence intervals for coefficients β_1, β_2, and β_3 of the linear model for a population of hospital observations. $g = 3$.

$$CI_1 = \left[\hat{\beta}_1 - t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_1\right),\ \hat{\beta}_1 + t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_1\right)\right]$$

$$\hat{\beta}_1 = 0.237209$$

$$t_{\alpha/(2g)=0.05/(2*3),\ df_{Res}=108} = 2.432$$

$$SE\left(\hat{\beta}_1\right) = 0.060957$$

$$CI_1 = [(0.237209) - (2.432)(0.060957),\ (0.237209) + (2.432)(0.060957)]$$

$$CI_1 = [0.0890,\ 0.385]$$

```
number_of_confidence_intervals <- 3
critical_value <- calculate_critical_value_t(
    significance_level,
    number_of_confidence_intervals,
    residual_degrees_of_freedom
)
critical_value
```

```
## [1] 2.431841
```

$$CI_2 = \left[\hat{\beta}_2 - t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_2\right),\ \hat{\beta}_2 + t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_2\right)\right]$$

$$\hat{\beta}_2 = -0.014071$$

$$SE\left(\hat{\beta}_2\right) = 0.022708$$

$$CI_2 = [(-0.014071) - (2.432)(0.022708),\ (-0.014071) + (2.432)(0.022708)]$$

$$CI_2 = [-0.0693,\ 0.0412]$$

$$CI_3 = \left[\hat{\beta}_3 - t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_3\right),\ \hat{\beta}_3 + t_{\alpha/(2g),\ df_{Res}} SE\left(\hat{\beta}_3\right)\right]$$

$$\hat{\beta}_3 = 0.020383$$

$$SE\left(\hat{\beta}_3\right) = 0.005524$$

$$CI_3 = [(0.020383) - (2.432)(0.005524),\ (0.020383) + (2.432)(0.005524)]$$

$$CI_3 = [-0.00695,\ 0.0338]$$

(e) Fill in the values for the ANOVA table for this multiple linear regression model.

Source of Variation	df	SS	MS
Regression	4	84.624	21.156
Error	108	116.8128	1.0816
Total	112	201.437	1.799

The regression degrees of freedom is the number of predictors. $df_R = k = 4$. The residual degrees of freedom is difference between the number of observations and the number of variables. $df_{Res} = n - p = 113 - 5 = 108$. The total degrees of freedom is one less than the number of observations. $df_T = 112$.

The residual standard error $\hat{\sigma} = 1.04$. An unbiased estimator for the variance $Var(\epsilon) = \sigma^2$ of the error terms of a multiple linear model for a population of hospital observations $\hat{\sigma}^2 = 1.04^2 = 1.0816$. The expected value of the residual mean square MS_{Res} is the variance $Var(\epsilon)$. The residual mean square $MS_{Res} = \hat{\sigma}^2 = 1.0816$. The residual mean square $MS_{Res} = SS_{Res}/df_{Res}$. The residual sum of squares $SS_{Res} = MS_{Res}df_{Res} = 1.0816 * 108 = 116.8128$.

The F statistic $F_0 = 19.56$. $F_0 = MS_R/MS_{Res}$. The regression mean square $MS_R = F_0 MS_{Res} = 19.56 * 1.0816 = 21.156$. $MS_R = SS_R/k$. $SS_R = MS_R k = 21.156 * 4 = 84.624$.

The total sum of squares $SS_T = SS_R + SS_{Res} = 84.624 + 116.8128 = 201.437$.

The total mean square may be thought of as $MS_T = SS_T/df_T = 201.437/112 = 1.799$.

(f) What is the coefficient of determination R^2 for this model? Write a sentence that interprets this value in context.

The coefficient of determination

$$R^2 = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{116.8128}{201.437} = 0.580$$

is the proportion of variation in infection risk that is explained by the linear relationship between infection risk and the predictors $Stay$, Age, $Xrays$, and $Services$.

(g) What is the adjusted coefficient of determination R^2_{adj} for this model?

The adjusted coefficient of determination

$$R^2_{adj} = 1 - \frac{MS_{Res}}{MS_T} = 1 - \frac{1.0816}{1.799} = 0.399$$

Per https://www.ibm.com/docs/fi/cognos-analytics/11.1.0?topic=terms-adjusted-r-squared, "Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models.

"It identifies the percentage of variation in the target field that is explained by the input or inputs. R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects... in the model [increases]. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific effect does not improve the model.

"Adjusted R^2 is calculated by dividing the residual mean square error... (which is the sample variance of the target field)... by the total mean square error... The result is then subtracted from 1.

"Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0.09 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between 0 and 1."