Practice Problems

Solutions to the practice questions ***will not be given****, but the solutions will be discussed during the review session. The session will be recorded.*

1. In the following regression models, identify which models are linear and which are non-linear. Y is the response, X_i are the predictors and β_i are the coefficients:

   (a) $Y = \beta_0 + \beta_1 X_1^3 + \beta_2 X_2^5$

   (b) $\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \log_2 X_2$

   (c) $Y = \beta_0 + \beta_1 X^2 + \beta_2 X^3 + \beta_3 X^3$

   (d) $Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2$

   (e) $Y = \beta_0 + e^{\beta_1 X_1} + \beta_2 X_2$

   (f) $Y = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$

2. Explain the difference between the regression model error terms and the model residuals.

3. For each of the following scenarios, a simple linear regression model was fitted:

   (a) $R^2 < 0.03$ and p-value < 0.05

   (b) $R^2 < 0.03$ and p-value > 0.05

   (c) $R^2 > 0.75$ and p-value < 0.05

   (d) $R^2 > 0.75$ and p-value > 0.05

   With a 5% significance level, what are your conclusions?

4. Consider the following regression model

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

   After fitting the above model, we find that X_1 is statistically significant. However, when we fit the simple linear regression model

   $$Y = \beta_0 + \beta_1 X_1$$

   we find that X_1 is not statistically significant. What is your conclusion for the independent variables X_1, X_2, X_3 and X_4?

5. What are the effects of multicollinearity in a linear regression model when its goal is the explanation of the variance of the response.

6. A linear regression model that is good at explaining the variation of the response, is it also good for prediction? Discuss.

7. What is the meaning of the term ”heteroscedasticity”?

8. What would be then consequences for the OLS estimator if heteroscedasticity is present in a regression model but ignored?

9. Suggest a remedial measure for a model that exhibits heteroscedasticity?

10. If OLS is used in the presence of autocorrelation, which of the following will be likely consequences?

    (a) Coefficient estimates may be misleading

    (b) Hypothesis tests could reach the wrong conclusions

    (c) Forecasts made from the model could be biased

    (d) Standard errors may inappropriate.

11. What will be the properties of the OLS estimator in the presence of multicollinearity?

12. Which one of the following is NOT an example of mis-specification of functional form?

    (a) Using a linear specification when y scales as a function of the squares of x

    (b) Using a linear specification when a double-logarithmic model would be more appropriate

    (c) Modeling y as a function of x when in fact it scales as a function of 1/x

    (d) Excluding a relevant variable from a linear regression model.

13. What are the consequences in the estimated coefficients, if the residuals from a fitted regression using a small sample are not normally distributed?

14. What are the consequences on a regression model, If a relevant variable is omitted from the regression equation.

1. **United Nations Data**

This data set contains the variables:

- PPgdp: the *2001 gross national product per person in US dollars*
- Fertility: the *birth rate per 1000 females* in the population in the year 2000.

The data are for 193 localities, mostly UN member countries.

(a) Based on the scatterplot of Fertility vs PPgdp (Figure 2a), does a straight line mean function seem to be a satisfactory summary of this graph?

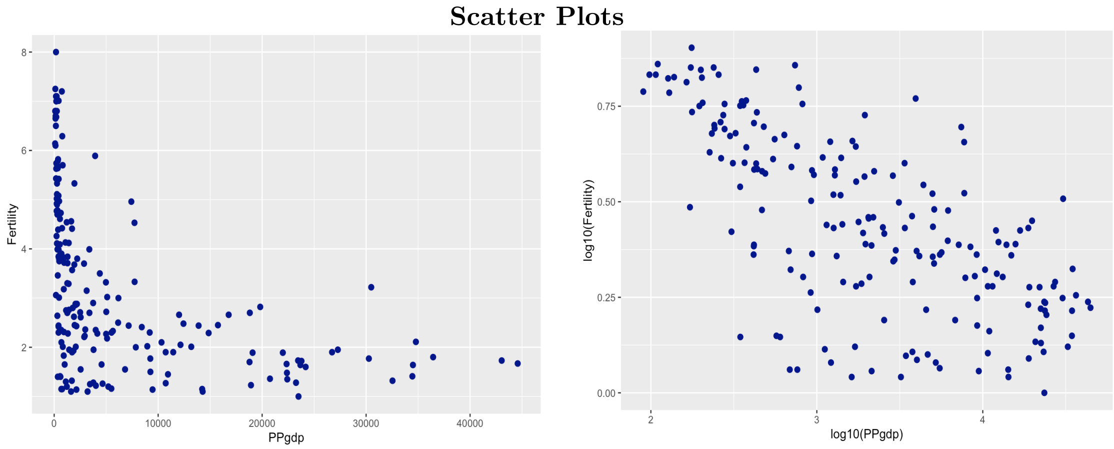


Fig. 2a: Fertility vs PPgdp **Fig. 2b:** $\log_{10}$(Fertility) vs $\log_{10}$(PPgdp)

(b) Based on the scatterplot of $\log_{10}$(Fertility) vs $\log_{10}$(PPgdp) (Figure 2b), does a straight line mean function seem to be a better summary of this graph?

Consider the *transformed variables*: $\log_{10}$(Fertility) and $\log_{10}$(PPgdp). We fit a simple linear regression model in R and obtain the following output:

```
un.log.reg = lm(log10(Fertility) ~ log10(PPgdp))
summary(un.log.reg)
```

```
##
## Call:
## lm(formula = log10(Fertility) ~ log10(PPgdp))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48587 -0.08148  0.03058  0.11327  0.39130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.17399    0.05879   19.97   <2e-16 ***
## log10(PPgdp) -0.22116    0.01737  -12.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1721 on 191 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4563
## F-statistic: 162.1 on 1 and 191 DF,  p-value: < 2.2e-16
```

	Mean	S
$Y = \log_{10}(Fertility)$	0.4421188	$S_{YY} = 10.45504$
$X = \log_{10}(PPgdp)$	3.309251	$S_{XX} = 98.14416$

(c) *State* the estimated regression line.

(d) Conduct a test to determine whether or not there is a linear association between $\log_{10}$`(Fertility)` and $\log_{10}$`(PPgdp)`. Use $\alpha = 0.05$. State the alternatives, decision rule (using the appropriate p-value in the results provided) and conclusion.

(e) Obtain the following:

i. a point estimate of the difference in the mean $\log_{10}$`(Fertility)` when $\log_{10}$`(PPgdp)` differ by **one**.

ii. the value of the coefficient of determination and explain its meaning.

iii. a point estimate for σ^2.

iv. the value of the correlation coefficient r with the appropriate *sign*.

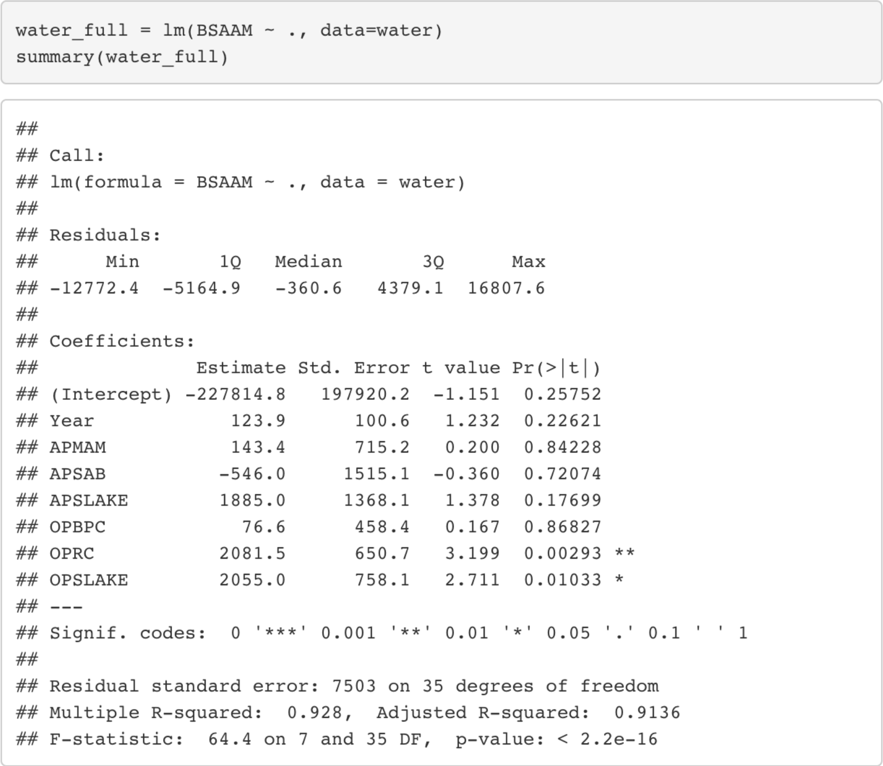
(f) Increasing $\log_{10}$`(PPgdp)` by one unit is the same as multiplying `PPgdp` by 10. If two localities differ in `PPgdp` by a factor of 10, give a 95% *confidence interval* on the difference in $\log_{10}$`(Fertility)` for these two localities.

(g) For a locality not in the data we know that `PPgdp=1000`. Obtain a 95% **prediction** interval for prediction $\log_{10}$`(Fertility)`. Give also a 95% prediction interval for prediction `Fertility`.

2. **California Water Supply**

One factor affecting water availability is *stream runoff*. If runoff could be predicted, engineers, planners and policy makers could do their jobs more efficiently. This data set contains 43 years worth of precipitation measurements taken at six sites in the Owens Valley (labeled `APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, `OPRC`, and `OPSLAKE`)[1], and *stream runoff volume* at a site near Bishop, California.

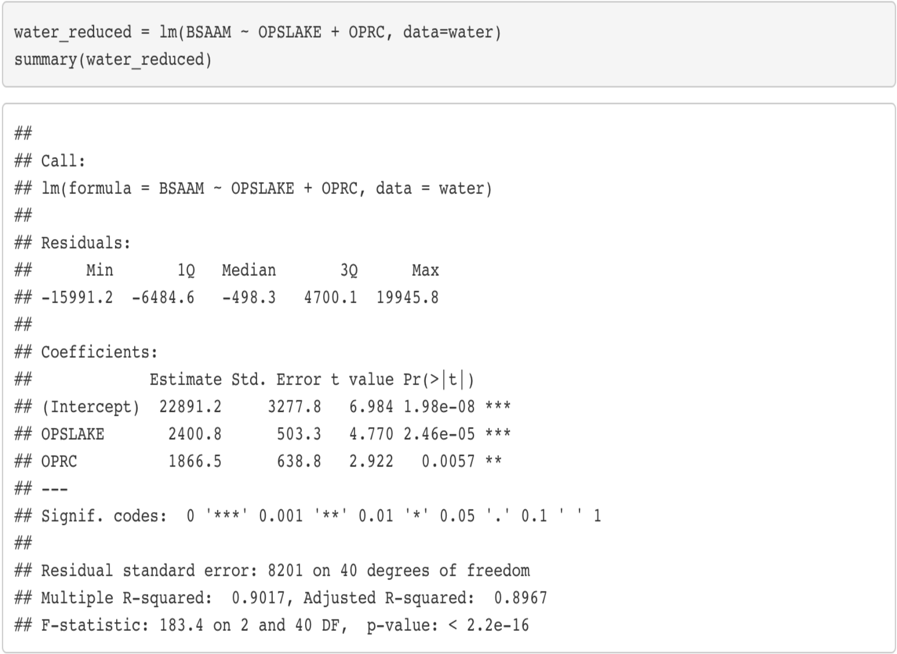
We fit a Multiple Linear Regression model with the *stream runoff volume* (the `BSAAM` variable) as the response.The year of data collection (variable `Year`) the 6 sites above are the predictors. The results are summarized below:



```
water_full = lm(BSAAM ~ ., data=water)
summary(water_full)
```

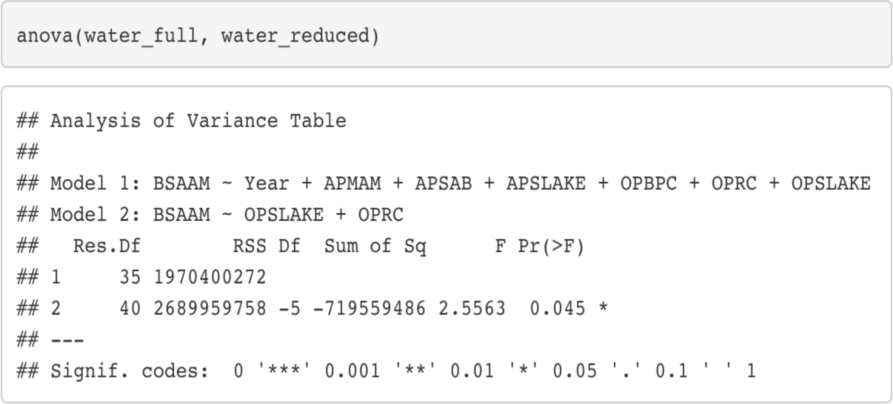
```
##
## Call:
## lm(formula = BSAAM ~ ., data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12772.4  -5164.9   -360.6   4379.1  16807.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -227814.8   197920.2  -1.151  0.25752
## Year            123.9      100.6   1.232  0.22621
## APMAM           143.4      715.2   0.200  0.84228
## APSAB          -546.0     1515.1  -0.360  0.72074
## APSLAKE        1885.0     1368.1   1.378  0.17699
## OPBPC            76.6      458.4   0.167  0.86827
## OPRC           2081.5      650.7   3.199  0.00293 **
## OPSLAKE        2055.0      758.1   2.711  0.01033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7503 on 35 degrees of freedom
## Multiple R-squared:  0.928,  Adjusted R-squared:  0.9136
## F-statistic:  64.4 on 7 and 35 DF,  p-value: < 2.2e-16
```

[1] These are just different names of sites in California where measurements were taken.



```
water_reduced = lm(BSAAM ~ OPSLAKE + OPRC, data=water)
summary(water_reduced)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPSLAKE + OPRC, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15991.2  -6484.6   -498.3   4700.1  19945.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22891.2     3277.8   6.984 1.98e-08 ***
## OPSLAKE       2400.8      503.3   4.770 2.46e-05 ***
## OPRC          1866.5      638.8   2.922   0.0057 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8201 on 40 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8967
## F-statistic: 183.4 on 2 and 40 DF,  p-value: < 2.2e-16
```



```
anova(water_full, water_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: BSAAM ~ Year + APMAM + APSAB + APSLAKE + OPBPC + OPRC + OPSLAKE
## Model 2: BSAAM ~ OPSLAKE + OPRC
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1     35 1970400272
## 2     40 2689959758 -5 -719559486 2.5563  0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

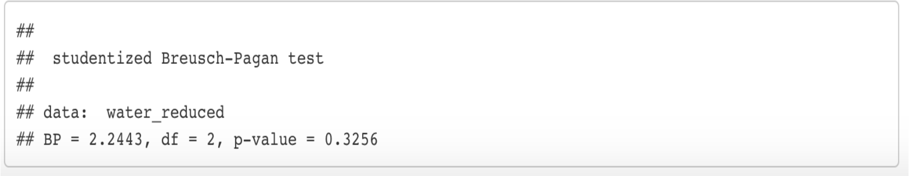
(a) Using the appropriate R output, test whether the variables `Year`, `APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, can be dropped from the regression model given that the other variables are retained. Use $\alpha = 5\%$. State the alternatives, decision rule and conclusion.

(b) You want to test whether $\beta_{\texttt{APMAM}} = 140$ and $\beta_{\texttt{OPBPC}} = 80$. State the null/alternative hypotheses, the appropriate test statistic and its distribution.[2] You do not need to perform the hypothesis test.

(c) You want to test whether $\beta_{\texttt{Year}} = \beta_{\texttt{OPBPC}} = \beta_{\texttt{OPRC}} = 0$. State the null/alternative hypotheses, the appropriate test statistic and its distribution. You do not need to perform the hypothesis test.

(d) Consider the Breusch-Pagan test results for the reduced model:

[2] By distribution we mean whether it is a T distribution or an F distribution along with the appropriate degrees of freedom.



```
##
##  studentized Breusch-Pagan test
##
## data:  water_reduced
## BP = 2.2443, df = 2, p-value = 0.3256
```

State the alternatives, decision rule and conclusion from this test. Use $\alpha = 0.05$.