Instruction

- Answer **ALL 15** questions on the answer sheet provided. Start by writing your NAME and UPI in the fields given.
- All questions have a **single** correct answer and carry the same mark value.
- If you give more than one answer to any question you will receive zero marks for that question.
- Incorrect answers are not penalised.

We will look at two datasets for this assignment, the dataset **swiss** and the dataset **CO2**.

**swiss** contains six variables for each of the 47 French-speaking provinces in Switzerland measured in 1888. The variables are:

- **Fertility**: common standardised fertility measure (our response);
- **Agriculture**: % of males involved in agriculture as occupation;
- **Examination**: % draftees receiving highest mark on army examination;
- **Education**: % education beyond primary school for draftees;
- **Catholic**: % of catholic population;
- **Infant.Morality**: % live births who live less than 1 year.

**CO2** contains five variables with 84 observations. The variables are:

- **conc**: A numeric vector of ambient carbon dioxide concentrations in $mL/L$;
- **Plant**: A plant identifier. There are 12 plants in the dataset, each measured at 5 different concentrations;
- **Type**: A factor with two levels, **Quebec** and **Mississippi**, 6 plants per type;
- **Treatment**: A factor with two levels, **chilled** and **nonchilled**;
- **uptake**: A numeric vector of carbon dioxide uptake rates in $\mu mol/m^2s$ (our response).
1. Assume we built a linear model with Fertility as response. Which visualisation tool is the most suitable to suggest suitable transformations for the covariates?

   (zz) GAM plots
   (1) Box-Cox plots
   (1) Residual v. Fitted values plots
   (1) Autocorrelation function plots
   (1) Coplots

![Figure 1: Trellis plot for Question 2](image)

2. Fig. 1 shows a trellis plot of CO₂ uptake against concentration conditional on type (panel) and treatment (colour). Which of the following statements is wrong?

   (zz) Plants in Mississippi have a higher uptake than plants in Quebec.
   (1) The data seem planar.
   (1) A parallel lines model appears suitable.
   (1) There seems to be a positive relationship between conc and uptake.
   (1) There are no unusually big residuals.
3. Fig. 2 shows the pairs plot for the **Swiss** data. Which of the following statements is **wrong**?

(zz) The slope for **Fertility** in the full model is visible in the corresponding panel with **Infant.Mortality**.

(1) **Education** has a potential outlier.

(1) There is indication for weak collinearity in the data.

(1) **Agriculture** and **Examination** are negatively correlated.

(1) There seem to be only few provinces with equal number of protestants and catholics.
4. Below is the summary output for the full Fertility model

```r
Call: 
lm(formula = Fertility ~ ., data = swiss) 
--- 
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 ** 
--- 
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 
```

Which of the following statements is **wrong**?

(zz) The covariates are unrelated to Fertility.

(1) Holding everything else constant, Fertility decreases with increasing Agriculture.

(1) Holding everything else constant, Fertility increases with increasing Infant.Mortality.

(1) The estimated coefficient for Catholic is 0.10412.

(1) The standard error for the coefficient for Examination is about as big as the absolute of the estimate for its coefficient.

5. Further, regarding the multiple $R^2$ of 0.7067 which of the following statements is **wrong**?

(zz) Removing Examination from the model can increase the multiple $R^2$.

(1) A low $R^2$ such as this leads to rather large prediction intervals.

(1) Deleting a low-leverage outlier can increase the $R^2$.

(1) The adjusted $R^2$ is never larger than the multiple $R^2$.

(1) If the covariates perfectly explain the variation in the response, then $R^2$ is equal to 1.
6. We use the above model to build a confidence interval for the average Fertility rate in an average village with the values below:

```r
> new.data <- data.frame(Agriculture=70,Examination=20,Education=6,
Catholic=95,Infant.Mortality=25)
> predict(swiss.lm,newdata=new.data,se.fit=F,interval="c")

  fit lwr  upr
1 81.29856  73.69039  88.90673
```

Which of the following statements is **wrong**:

(zz) The prediction interval would be smaller.

(1) If the values for the fictitious village lie outside the range of the observed covariates, then the confidence interval is not reliable.

(1) The average fictitious village has a Fertility between 73.7 and 88.9.

(1) To get the prediction interval, we replace "c" by "p".

(1) The relatively low \( R^2 \) does not affect the validity of the confidence interval.

7. The pairs plot showed correlations between the covariates. We look at the variance inflation factor to assess whether there are any issues with collinearity.

```r
> diag(solve(cor(swiss[-1])))

  Agriculture Examination Education Catholic Infant.Mortality
1.284129 3.675420 2.774943 1.937160 1.107542
```

Which of the following statements is **wrong**:

(zz) Some of the variance inflation factors indicate collinearity.

(1) The variance inflation factor for Examination is the highest.

(1) The variance of the coefficient for Agriculture depends on the variance of Agriculture.

(1) The variance of the coefficient for Education depends on the error variance.

(1) The variance of the coefficient for Catholic depends on the sample size.
8. Fig. 3 shows a collection of diagnostic plots for the Fertility data. Which of the following statements is wrong:

(zz) There are some large studentised residuals.
(1) There is at least one high-leverage point.
(1) The residuals are approximately normal.
(1) The model appears planar.
(1) We can assume constant scatter.
Figure 4: Transformation plots for Question 9.

9. Fig. 4 shows a few curves to test whether a transformation is useful for one or two of the variables. Which of the following is wrong?

(zz) The Box-Cox-Plot suggests a log-transform for the response.
(1) Fitting a polynomial to Examination could improve the fit.
(1) Fitting a polynomial to Education would only improve the fit for a few outliers.
(1) Catholic does not need a transformation.
(1) The distribution of observations suggests that no transformation is needed for Education.
Figure 5: Influence plots for Question 10.

10. Fig. 5 visualises the various influence measures for the Fertility data. Which of the following statements is wrong?

  (zz) Observation 21 strongly influences the estimate of the coefficient for Catholic.
  (1) Observation 6 strongly influences the fit.
  (1) Observation 45 is a high-leverage point.
  (1) Observation 19 strongly influences the size of the standard errors.
  (1) No observation has strong influence on the estimate of the coefficient for Examination.
11. We use allpossregs to find a collection of suitable models.

<table>
<thead>
<tr>
<th></th>
<th>rssp</th>
<th>sigma2</th>
<th>adjRsq</th>
<th>Cp</th>
<th>AIC</th>
<th>BIC</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4015.236</td>
<td>89.227</td>
<td>0.428</td>
<td>35.205</td>
<td>82.205</td>
<td>85.905</td>
<td>379.031</td>
</tr>
<tr>
<td>2</td>
<td>3054.169</td>
<td>69.413</td>
<td>0.555</td>
<td>18.486</td>
<td>65.486</td>
<td>71.037</td>
<td>307.676</td>
</tr>
<tr>
<td>3</td>
<td>2422.245</td>
<td>56.331</td>
<td>0.639</td>
<td>8.178</td>
<td>55.178</td>
<td>62.579</td>
<td>254.531</td>
</tr>
<tr>
<td>4</td>
<td>2158.069</td>
<td>51.383</td>
<td>0.671</td>
<td>5.033</td>
<td>52.033</td>
<td>61.284</td>
<td>236.903</td>
</tr>
<tr>
<td>5</td>
<td>2105.043</td>
<td>51.343</td>
<td>0.671</td>
<td>6.000</td>
<td>53.000</td>
<td>64.101</td>
<td>247.849</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agriculture</th>
<th>Examination</th>
<th>Education</th>
<th>Catholic</th>
<th>Infant.Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Which of the following statements is **wrong**?

(zz) BIC favours a model with two variables.

(1) **Examination** only occurs in the full model.

(1) Cp and AIC favour the same model.

(1) Cross validation favours a model with four variables.

(1) The most suitable model with two variables contains **Education** and **Catholic**.
12. We now move back to the $CO_2$ data. Before fitting a model we will check whether our factors interact. Fig. 6 shows the interaction plot. Which statement is wrong?

(a) Two factors interact if their lines are parallel.
(b) The plot suggests interaction between the factors.
(c) Mississippi has a lower mean uptake than Quebec for chilled plants.
(d) Mississippi has a lower mean uptake than Quebec for non-chilled plants.
(e) In general, non-chilled plants appear to have a higher uptake.
13. We fit a parallel-lines model to the $CO_2$ data:

Call:
`lm(formula = uptake ~ conc + Type + Treatment, data = CO2[-1])`
---

Coefficients:

| Estimate   | Std. Error | t value | Pr(>|t|) |
|------------|------------|---------|----------|
| (Intercept)| 39.613812  | 1.282083| <2e-16   ***|
| conc       | 0.005669   | 0.001781| 3.183    0.00238 **|
| TypeMississippi | -15.080000 | 0.942983 | -15.992 <2e-16 ***|
| Treatmentchilled | -8.060000  | 0.942983 | -8.547 9.8e-12 ***|

Residual standard error: 3.652 on 56 degrees of freedom
Multiple R-squared: 0.8582, Adjusted R-squared: 0.8506
F-statistic: 113 on 3 and 56 DF, p-value: < 2.2e-16

Which of the following statements is **wrong**?

(zz) The fitted line for chilled Quebec plants has slope $-8.060000$.
(1) The fitted line for unchilled Mississippi plants has slope $0.005669$.
(1) The intercept for unchilled Quebec plants is $39.613812$.
(1) The intercept for unchilled Mississippi plants is $24.53381$.
(1) The intercept for chilled Quebec plants is $31.55381$. 
14. We add interactions between the factors to our model

```r
> model2 <- lm(uptake~conc+Type*Treatment,data=CO2[-1])
> summary(model2)
```

Call:
`lm(formula = uptake ~ conc + Type * Treatment, data = CO2[-1])`

---

Coefficients:                           Estimate Std. Error t value Pr(>|t|)
(Intercept)             37.260479    1.027002  36.281  < 2e-16 ***
conc                    0.005669    0.001339   4.234 8.79e-05 ***
TypeMississippi        -10.373333    1.002605  -10.346 1.62e-14 ***
Treatmentchilled        -3.353333    1.002605   -3.345  0.00149 **
TypeMississippi:Treatmentchilled -9.413333    1.417898   -6.639  1.47e-08 ***
---
Residual standard error: 2.746 on 55 degrees of freedom
Multiple R-squared: 0.9213, Adjusted R-squared: 0.9156
F-statistic: 160.9 on 4 and 55 DF, p-value: < 2.2e-16

> anova(model1,model2)

Analysis of Variance Table

```r
Model 1: uptake ~ conc + Type + Treatment
Model 2: uptake ~ conc + Type * Treatment

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>4</td>
<td>746.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>1</td>
<td>414.65</td>
<td>332.29</td>
<td>1.47e-08 ***</td>
</tr>
</tbody>
</table>
```

Which of the following statements is **wrong**?

(zz) `anova` suggests that `model1` is adequate.

1. Under `model1` the expected uptake of a chilled plant with concentration \(200\text{mL/L}\) in Mississippi is \(17.60757\mu\text{mol/m}^2s\).

1. Under `model2` the expected uptake of a chilled plant with concentration \(200\text{mL/L}\) in Mississippi is \(15.25424\mu\text{mol/m}^2s\).

1. The summary output for `model2` shows significant offsets for the intercepts.

1. In both models unchilled plants from Quebec have the largest intercept.
15. We use model2 to compute prediction and confidence intervals for plants with concentration 500 for all four factor level combinations.

```r
> test.set <- data.frame(conc=rep(500,4),
  Type=c("Quebec","Quebec","Mississippi","Mississippi"),
  Treatment=c("chilled","nonchilled","chilled","nonchilled"))
> predict(model2,newdata=test.set,interval="p")
          fit     lwr     upr
1 36.74155 31.05657 42.42653
2 40.09488 34.40991 45.77986
3 16.95488 11.26991 22.63986
4 29.72155 24.03657 35.40653
> predict(model2,newdata=test.set,interval="c")
          fit     lwr     upr
1 36.74155 35.31314 38.16996
2 40.09488 38.66647 41.52329
3 16.95488 15.52647 18.38329
4 29.72155 28.29314 31.14996
```

Which of the following statements is **wrong**?

(zz) A chilled plant from Quebec under a $CO_2$ concentration of 500$ml/L$ has an uptake between 35.3 and 38.2$\mu mol/m^2s$.

(1) The average chilled plant from Mississippi under a $CO_2$ concentration of 500$ml/L$ has an uptake between 15.5 and 18.4$\mu mol/m^2s$.

(1) A nonchilled plant from Quebec under a $CO_2$ concentration of 500$ml/L$ has an uptake between 34.4 and 45.8$\mu mol/m^2s$.

(1) The expected uptake of a nonchilled plant from Mississippi under a $CO_2$ concentration of 500$ml/L$ is approximately 29.7$\mu mol/m^2s$.

(1) The average nonchilled plant from Quebec under a $CO_2$ concentration of 500$ml/L$ has an uptake between 38.6 and 41.5$\mu mol/m^2s$. 