Understanding Data and Statistical Design (60117)

Lab 8: Multiple Linear Regression II

This lab is marked from 18.

Please submit via Canvas.

Due by the conclusion of the lab class

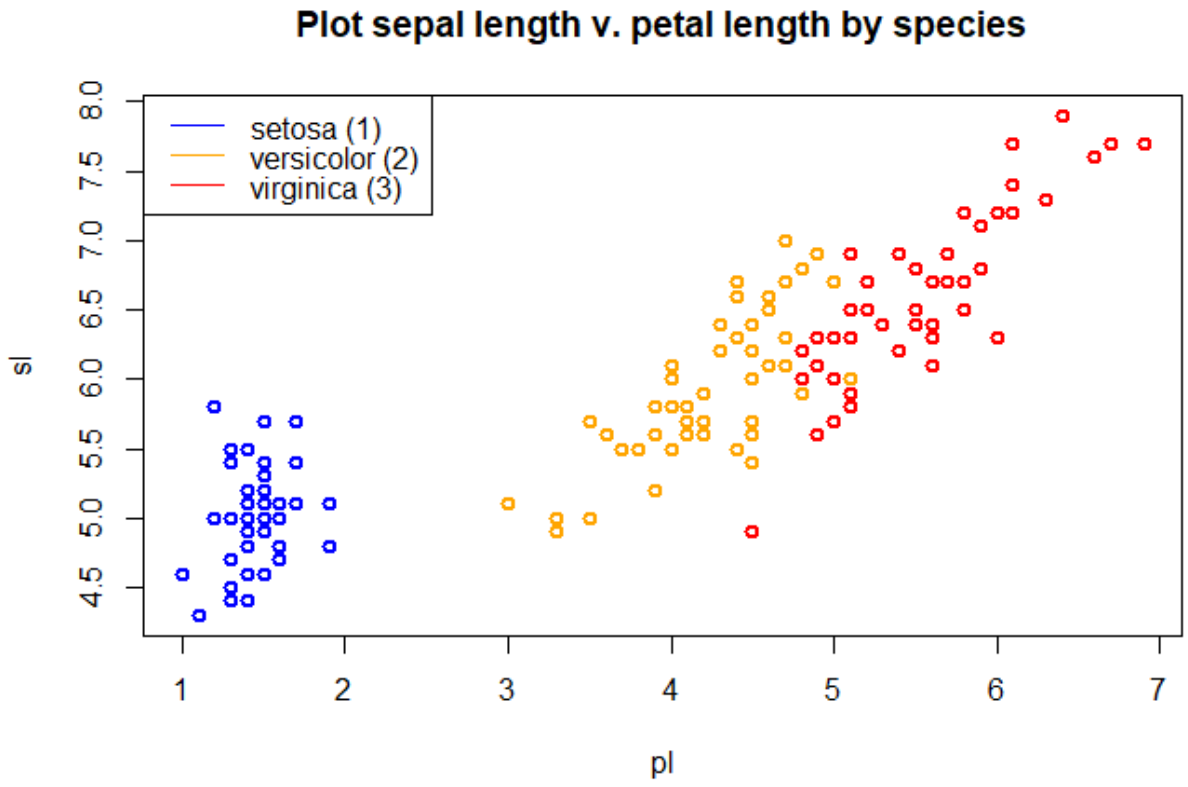
QUESTION 1. Iris flower petal characteristics

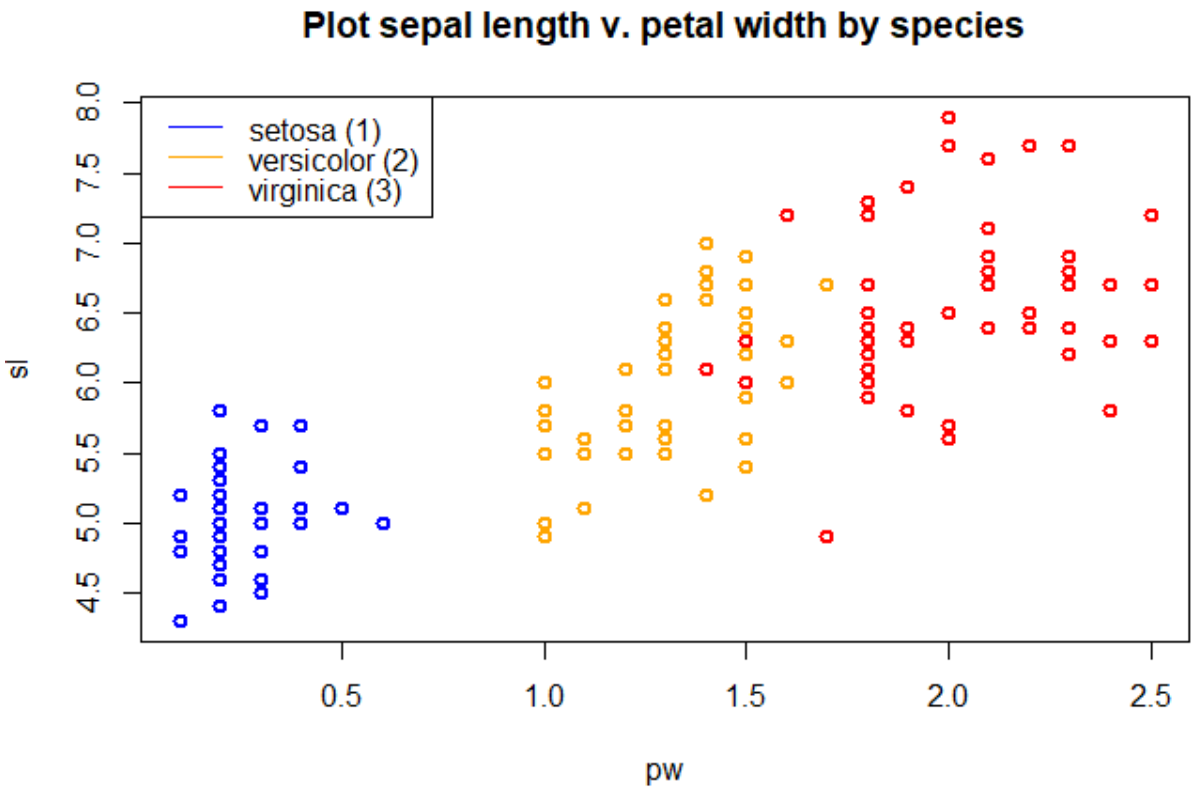
In this question we continue the analysis of iris flowers, but this time with the inclusion of a categorical predictor.

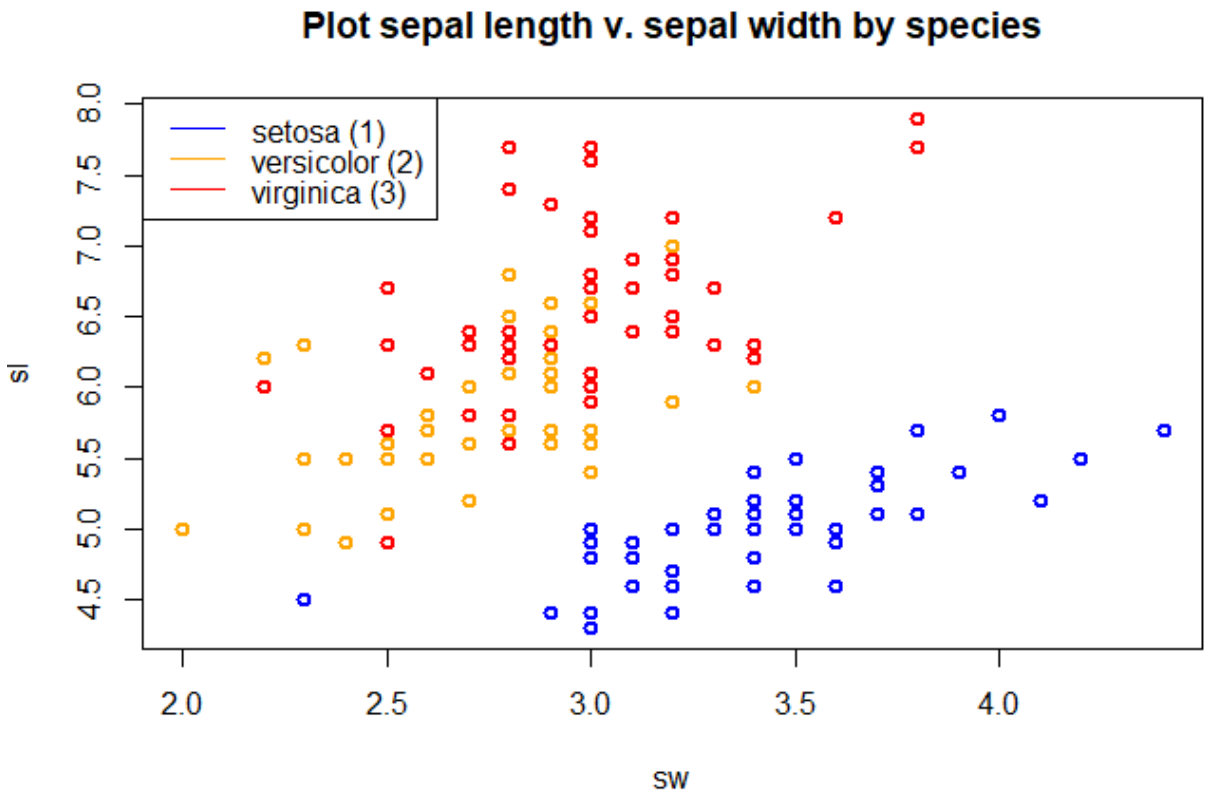
Name	Type	Description
sl	response	length of flower sepal
pl	continuous predictor	length of flower petal
pw	continuous predictor	width of flower petal
sw	predictor	width of flower sepal
spec	continuous predictor	iris species: setosa (1), versicolor (2), virginica (3)

The data is from the "iris" data set built into R (see accompanying R code file).

We begin by looking at the relationship between the variables with some scatter plots. R produced the output below.







(a) Citing evidence in the scatter plots, comment on whether the categorical variable *spec* could be a useful predictor in a model also containing *pl*, *pw* and *sl* **[3 marks]**.

Yes, because the distribution of the response variable and the predictors shows a difference between the species, it is a very useful predictor.

(b) Citing evidence in the scatter plots, comment on the need for interaction terms involving *spec* and the three continuous predictors **[3 marks]**.

Interaction terms would be useful to capture how the relationship between iris flower measurements (sepal and petal length and width) and sepal length varies among iris species. The graphs suggest that species(spec) interacts with petal length(pl) and potentially with sepal width(sw), as trends and clustering are different for each species.
The interaction terms between spec and pl, and possibly spec and sw, should be considered in the regression model to allow the slopes of these relationships to change with each iris species, thus improving the model's accuracy.

To avoid over complicating things, we will drop the predictor *pw* and consider the population model

$$SL = \beta_0 + \gamma_2 * spec2 + \gamma_3 * spec3 + \beta_{pl} * pl$$
$$+(\beta_{sw} + \delta_2 * spec2 + \delta_3 * spec3) * sw + \epsilon.$$

where $spec2$ and $spec3$ are the binary dummy variables (reference category $spec = 1$).

Model for $spec = 1$.

$$SL = \beta_0 + \beta_{pl} * pl + \beta_{sw} * sw + \epsilon$$

Model for $spec = 2$.

$$SL = \beta_0 + \gamma_2 + \beta_{pl} * pl + (\beta_{sw} + \delta_2) * sw + \epsilon$$

Model for $spec = 3$.

$$SL = \beta_0 + \gamma_3 + \beta_{pl} * pl + (\beta_{sw} + \delta_3) * sw + \epsilon$$

Now fit this regression model to the sample data.

(c) Write down the regression equation for each of the three iris varieties **[3 marks]**.

For `setosa` (spec1), which is the reference category:

$$\hat{SL} = 1.66659 + 0.82205 \cdot pl + 0.62357 \cdot sw$$

For `versicolor` (spec2):

$$\hat{SL} = (1.66659 + 0.28256) + 0.82205 \cdot pl + (0.62357 - 0.44850) \cdot sw$$
$$\hat{SL} = 1.94915 + 0.82205 \cdot pl + 0.17507 \cdot sw$$

For `virginica` (spec3):

$$\hat{SL} = (1.66659 - 0.64587) + 0.82205 \cdot pl + (0.62357 - 0.28621) \cdot sw$$
$$\hat{SL} = 1.02072 + 0.82205 \cdot pl + 0.33736 \cdot sw$$

(d) Provide interpretations of the estimates for γ_3, β_{sw} and δ_2 **[3 marks]**.

γ_3: suggests holding other variables constant, Virginia's sepal length (SL) is 0.64587 units shorter on average than setosa. This effect is not statistically significant (p-value: 0.2744).

β_{sw}: The coefficient 0.62357 indicates that for every one-unit increase in sepal width (sw), sepal length (SL) increases by an average of 0.62357 units. This effect is statistically significant (p-value: 2.72e-07).

δ_2: The interaction term, -0.44850 represents the difference in the relationship between sepal width (sw) and sepal length (SL) for versicolor compared to the reference category, setosa. For versicolor, the effect of sepal width on sepal length

is reduced by 0.44850 units. This interaction effect is statistically significant (p-value: 0.0182).

(e) Determine if there is any statistical evidence of multicollinearity **[2 marks]**. Why should one expect multicollinearity in this model **[1 mark]**?

```
                  GVIF Df GVIF^(1/(2*Df))
pl          21.721813  1        4.660667
spec      7635.522574  2        9.347806
sw           4.029191  1        2.007285
spec:sw   6342.655138  2        8.924169
```

By calculating the VIF, spec and spec:sw have a coefficient greater than 5, 9.34 and 8.9 respectively, demonstrating a high degree of multicollinearity.
This may be due to the interaction terms since, by using dummy variables, the variables can be linearly dependent on each other.

(f) Using significance level 0.05, test if there is significant interaction between iris species and sepal width. Write down the null and alternative hypotheses, the test statistic and p–value, the result of the test and a conclusion using a minimum of mathematical language **[3 marks]**.

```
> anova(model1)
Analysis of Variance Table

Response: sl
           Df Sum Sq Mean Sq  F value    Pr(>F)
pl          1 77.643  77.643 828.4914 < 2.2e-16 ***
spec        2  7.843   3.922  41.8463 4.929e-15 ***
sw          1  2.716   2.716  28.9826 2.923e-07 ***
spec:sw     2  0.564   0.282   3.0094   0.05246 .
Residuals 143 13.401   0.094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ho: There is no interaction between iris species and sepal width.

Ha: There is an interaction between iris species and sepal width.

Statistic: 3.0094

Pvalue: 0.05246

Result: We don’t have enough evidence to reject the Ho, since the p_value > α

Conclusion: There is no interaction between spec and sw.

(g) Using diagnostic plots of the residuals, assess whether the modelling assumption have been satisfied **[not assessed]**.

(h) Consider the new data points:

- $pl = 1.35, sw = 3.75, spec = 1$
- $pl = 3.85, sw = 2.65, spec = 2.$

Use R to calculate the fitted values, 95% mean prediction intervals and 95% individual prediction intervals for the new values **[not assessed]**.

(i) Identify influential points, refit the model after excluding these points and analyse model fit **[not assessed]**.