Statistics and Experimental Design - 2026-27

3 Linear Regression - Inferential Approach

PRACTICAL EXERCISES

1. Consider the dataset `iris`, available in `R`, concerning morphometric measurements of iris flowers. Consider only the observations of the four morphometric variables: petal and sepal width and length (all in cm) for $n = 150$ iris flowers. Assume that these observations constitute a random sample from a larger population.

   (a) Consider the linear relationship between petal width (`Petal.Width`) and the remaining three morphometric variables.

      i. Construct scatter plots for every possible pair of variables. Comment.

      ii. Fit a multiple linear regression model for petal width using the remaining three variables as predictors. Comment on the coefficient of determination obtained.

      iii. Interpret the estimated values of the parameters associated with each of the predictor variables.

      iv. Consider the sign of the coefficient b_j associated with the predictor `Sepal.Length`, in the multiple linear regression model fitted above . Given the scatter plot relating the response variable response variable `Petal.Width` with the predictor `Sepal.Length`, obtained in item 1(a)i), what would be the sign of the slope of that regression line? Comment.

      v. Obtain 95% confidence intervals for β_1, β_2 and β_3. Comment.

      vi. Test whether it is reasonable to assume that an increase in length of the sepals, keeping the remaining predictors fixed, is associated with a decrease in mean petal width.

   (b) Now consider the relationship between petal width (`Petal.Width`, response variable y) and petal length (`Petal.Length`, predictor variable x), both in cm. Answer the following questions.

      i. Fit the regression line of petal width (y) on petal length (x).

      ii. Obtain estimates of the variances and standard errors of the estimators of the parameters of the population regression line, β_0 and β_1.

      iii. Use a hypothesis test on the slope of the population regression line β_1 to assess the following statement: "there is no significant linear relationship between petal lengths and widths in iris flowers."

      iv. Use a hypothesis test to assess the following statement: "for each additional centimetre in petal length, petal width increases, on average, by $0.5cm$".

      v. Use a hypothesis test to assess the following statement: "for each additional centimetre in petal length, petal width increases, on average, by less than $0.5cm$".

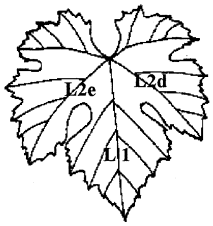
      vi. Estimate the expected petal width for iris flowers whose petal length is 4.5cm. Construct a confidence interval for this expected value.

      vii. Construct a 95% prediction interval for the width of a petal whose length is 4.5cm. Compare it with the confidence interval obtained in the previous item and comment.

      viii. Study the residual plots to detect any potential problems with the model assumptions. Comment on your conclusions.

2. Accurate measurement of leaf areas requires destructive techniques. The aim is to obtain a model that allows leaf areas to be estimated for grapevine cultivars, using predictor variables that can be measured without removing the leaves from the vine. For this purpose, at the ISA Horticulture Section, 200 leaves were randomly selected from each of three cultivars: Fernão Pires, Vital and Água Santa. For each leaf,

leaf area (variable **Area**, in cm^2), main vein length (variable **NP**, in cm), left lateral vein length (variable **NLesq**, in cm), and right lateral vein length (variable **NLdir**, in cm) were measured. The data file is available on the course *web* page in **txt** format under the name **videiras.txt**. Answer the following questions for the full set of 600 pairs of observations, without distinguishing between cultivars.



(a) Construct scatter plots for every pair of the 4 observed variables. Comment.

(b) Calculate the correlation matrix for all pairs of observed variables. Comment.

(c) Describe and fit the multiple linear regression model obtained by modelling `Area` using the three available predictors.

(d) Assuming the validity of the model, test, at a significance level of $\alpha = 0.01$, the hypothesis that, for each additional centimetre in main vein length (with the lateral vein lengths unchanged), there is a mean increase in leaf area of 7 cm^2. Repeat the test using a significance level of $\alpha = 0.05$. Comment and calculate the p-value of this test.

(e) Is it reasonable to assume that the coefficients of the two lateral veins are equal? Justify formally.

(f) The veins of three new leaves on the vine were measured. The results obtained were:

No. leaf	NP	NLesq	NLdir
1	12.1	11.6	11.9
2	10.6	10.1	9.9
3	15.1	14.9	14.0

For each new leaf, calculate:

i. the estimated leaf area;

ii. a 95% confidence interval for the expected leaf area associated with these predictor values;

iii. a 95% prediction interval for the leaf area of each individual leaf.

(g) For the fitted multiple regression, comment on the value of the coefficient of determination. In particular, test the overall fit of the model.

(h) Now consider the problem of predicting leaf areas (variable **Area**) using only measurements of main vein length (variable **NP**).

i. Construct the scatter plot of leaf areas and main vein lengths.

ii. Consider the logarithmic transformations of both variables and construct the scatter plot using the log-transformed data.

iii. Fit the regression line of log-areas on log-main-vein lengths and plot it on the scatter plot obtained in the previous item. Comment.

iv. What nonlinear relationship between `Area` and `NP` corresponds to the line fitted in the previous item? Using the `curve` command, plot it on the scatter plot constructed in item 2(h)i. Comment.

v. Construct a 95% confidence interval for the slope of the population regression line relating the log-transformed variables. Interpret the biological meaning of the result obtained.

vi. A researcher argues that, since leaf area is a two-dimensional measure and vein length is one-dimensional, it would be plausible for area to be proportional to the square of length. Perform a hypothesis test to assess this claim.

vii. Assess the assumptions of the fitted linear regression model by analysing the behaviour of the residuals and the remaining diagnostic tools. Comment. Now consider the linear regression between the original (non-log-transformed) variables and repeat the residual analysis. Comment.

3. In CAED Report 17, Iowa State University, 1963, the following meteorological and corn production data for the state of Iowa (USA), for the years 1930–1962, are reported. The data file is available on the course *web* page in **txt** format under the name **milho.txt**.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
Year		Prec. 'pre-season' (in.)	Temp. May (^{o}F)	Prec. June (in.)	Temp. June (^{o}F)	Prec. July (in.)	Temp. July (^{o}F)	Prec. August (in.)	Temp. August (^{o}F)	Prod. corn (bu/acre)
1930	1	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	2	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
1932	3	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	4	16.76	60.5	1.64	77.8	3.45	76.1	3.01	70.5	40.0
1934	5	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	6	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	7	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	8	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	9	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
1939	10	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	11	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	12	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	13	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	14	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	15	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	16	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	17	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	18	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	20	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	21	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	23	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	24	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	25	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	26	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	27	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	28	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	29	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	30	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	31	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	32	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	33	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

(a) Fit a Linear Model to predict corn production (in *bu/acre*), using all the remaining variables as predictors. Perform the F-test for overall model fit. Comment on the result.

(b) Determine the adjusted R^2. Comment.

(c) Repeat the fit from the first item, now excluding the time variable x_1 from the set of predictors. Compare the fitting results in the two cases. Comment.

(d) Use a t-test for the coefficient β_1 in the model with all predictors to determine whether it is possible to conclude that the models with and without predictor x_1 have significantly different fits.

(e) Use a partial F-test to answer the question in the previous item. Compare the *p-values* obtained from these two tests and discuss their relationship.

(f) Based only on the fit of the full model, obtained in item 3a), state, with justification, which predictor variable could be removed with the least effect on model quality.

(g) Test whether the model with all predictor variables and the model including only predictor variables that can be known by the end of June differ significantly. Comment.

(h) Identify a more parsimonious model using the method of sequential variable elimination based on tests of $\beta_j = 0$ ($\alpha = 0.10$). Repeat using the Akaike Information Criterion (AIC) as the selection criterion.

(i) Study the residual plots and other diagnostics.

4. For commercial purposes, plum weight is commonly estimated from their diameters. In order to obtain a relationship between diameter and weight, valid for a particular variety, $n = 41$ fruits were measured, diameter (`diametro` in *mm*) and weighed (`peso` in *g*), with the values given in the data file available on the course *web* page in **txt** format under the name **ameixas.txt**.

   (a) Construct the scatter plot of `peso` *vs* `diametro`. Comment on the underlying relationship obtained. Fit a simple linear regression of `peso` on `diametro` and plot the fitted regression line on the scatter plot.

   (b) Fit a second-degree polynomial to the relationship between the two variables: $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Give the estimates of the parameters of this model. Plot the fitted parabola over the scatter plot obtained in the previous item.

   (c) Formally test whether the parabolic model in the previous item fits significantly better than the initial linear model. Comment.

   (d) Inspect the residuals of the fitted parabolic model and comment.

   (e) Investigate whether it is worthwhile considering a third-degree polynomial for the relationship between fruit diameter and weight.

5. In a study of a tree species, the aim is to establish relationships among tree trunk height, diameter at breast height and trunk volume. Measurements of these variables were made on $n = 31$ trees, with the results designated by the names *Altura* (measured in feet), *Diametro* (measured in inches) and *Volume* (measured in cubic feet). The following are some basic descriptive statistics, as well as the correlation coefficients between the variables:

```
> apply(arvores,2,summary)
        Diametro Altura Volume
Min.        8.30     63  10.20
1st Qu.    11.05     72  19.40
Median     12.90     76  24.20
Mean       13.25     76  30.17
3rd Qu.    15.25     80  37.30
Max.       20.60     87  77.00
```

```
> apply(arvores,2,var)
  Diametro     Altura     Volume
  9.847914  40.600000 270.202796

> cor(arvores)
           Diametro    Altura    Volume
Diametro 1.0000000 0.5192801 0.9671194
Altura   0.5192801 1.0000000 0.5982497
Volume   0.9671194 0.5982497 1.0000000
```

   (a) A multiple linear regression model was initially fitted to predict trunk volumes from their heights and diameters, with the following results.

```
Call: lm(formula = Volume ~ Diametro + Altura, data=arvores)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07
Diametro      4.7082     0.2643  17.816  < 2e-16
Altura        0.3393     0.1302   2.607   0.0145
---
```

```
Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. Perform the overall model fit test. Discuss the result.

ii. State whether it is possible to simplify this model by obtaining a simple linear regression that is not significantly worse than this model. Use significance levels $\alpha = 0.05$ and $\alpha = 0.01$. Comment.

iii. Regardless of your answer to the previous item, give, for each of the simple linear regression submodels considered, the Coefficients of Determination and the value of the F statistic in the overall fit test.

(b) Based on previous experience, it was suggested that the fit could be further improved by applying a logarithmic transformation to all variables. The resulting fit is given below.

```
Call: lm(formula = log(Volume) ~ log(Diametro) + log(Altura) , data=arvores)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(Diametro)  1.98265    0.07501  26.432  < 2e-16 ***
log(Altura)    1.11712    0.20444   5.464 7.81e-06 ***
---
Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-Squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. What underlying relationship does this model represent, in terms of the original (non-log-transformed) variables?

ii. Discuss the following statement: "the fit to the log-transformed data is better, given the larger Coefficient of Determination and the larger value of the F statistic".

(c) It was finally decided to try a model (without transforming the variables) in which the variables *Height* and *Volume* swap roles relative to the initial model, that is, to determine whether trunk height can be adequately described by a linear relationship with Diameter and Volume. The following results were obtained with this model:

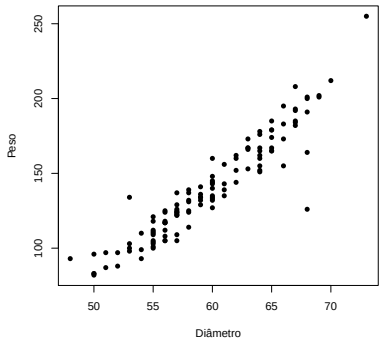
```
Call: lm(formula = Altura ~ Diametro + Volume, data=arvores)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.2958     9.0866   9.167 6.33e-10
Diametro     -1.8615     1.1567  -1.609   0.1188
Volume        0.5756     0.2208   2.607   0.0145

Residual standard error: 5.056 on 28 degrees of freedom
Multiple R-Squared: 0.4123,     Adjusted R-squared: 0.3703
F-statistic:  9.82 on 2 and 28 DF,  p-value: 0.0005868
```

Discuss the result of this test, taking into account the relatively low value of the Coefficient of Determination associated with the fit. How can the fact that this new relationship between the same three variables used in the initial model produce a much poorer fit?

6. In an experiment, the aim is to relate pear weight (in g) to its diameter (in mm), based on measurements from $n = 120$ pears. The following are some summary statistics for the results obtained, together with the corresponding scatter plot:

```
     Weight            Diameter
Min. :   82.0     Min. :  48.00
Mean :  139.8     Mean :  59.71
Max. :  255.0     Max. :  73.00
Var. : 1131.675   Var. :  27.4688375
```

The correlation coefficient between the observed variables is $r = 0.9397929$.



(a) A simple linear regression model of weight on diameter was fitted.

i. What is the coefficient of determination of the fitted regression line, and how should it be interpreted?

ii. Test whether the model should be considered useless ($\alpha = 0.05$).

iii. Justify which observation has the greatest leverage For that observation, calculate the leverage. How does the value compare with the mean leverage?

(b) A second-order polynomial model was subsequently fitted to predict weight from diameter, and the following fit was obtained:

```
Call: lm(formula = Weight ~ Diameter + I(Diameter^2), data = Pera)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   134.80456  118.14750   1.141   0.2562
Diameter       -5.89331    3.95255  -1.491   0.1386
I(Diameter^2)   0.09935    0.03289   3.021   0.0031
---
Residual standard error: 11.17 on 117 degrees of freedom
Multiple R-squared: 0.8917, Adjusted R-squared: 0.8898
F-statistic: 481.5 on 2 and 117 DF,  p-value: < 2.2e-16
```

Briefly justify whether this quadratic model is significantly better than the linear model considered previously.

7. The aim is to study the soil characteristics that affect the production of above-ground biomass of *Spartina alterniflora* , a perennial herbaceous plant of above-ground biomass of *Spartina alterniflora*, a perennial herbaceous plant 45 randomly selected soils were sampled, and the following variables were observed: biomass (`BIO`, in $g\ m^{-2}$), salinity (`SAL`), acidity (`pH`) and the concentrations of potassium (`K`), sodium (`Na`) and zinc (`Zn`), the latter in *ppm*.

The following are summary indicators of the observed values (minimum, mean, median, quartiles and standard deviation), together with the correlation matrix of the observed variables.

```
     BIO           SAL             pH            K              Na            Zn
Min.   : 236  Min.   :24.00  Min.   :3.200  Min.   : 350.7  Min.   : 7886  Min.   : 0.2105
1st Qu.: 416  1st Qu.:27.00  1st Qu.:3.450  1st Qu.: 528.6  1st Qu.:11345  1st Qu.:13.9852
Median : 824  Median :30.00  Median :4.450  Median : 773.3  Median :14752  Median :19.5880
Mean   :1001  Mean   :30.27  Mean   :4.602  Mean   : 797.6  Mean   :16597  Mean   :17.8752
3rd Qu.:1560  3rd Qu.:33.00  3rd Qu.:5.200  3rd Qu.: 954.1  3rd Qu.:20436  3rd Qu.:22.6758
Max.   :2436  Max.   :38.00  Max.   :7.450  Max.   :1441.7  Max.   :35186  Max.   :31.2865
sd     : 660  sd     : 3.72  sd     :1.247  sd     : 297.6  sd     : 6882  sd     : 8.2798
```

```
> cor(biomassa)
        BIO     SAL      pH       K      Na      Zn
BIO  1.0000 -0.1032  0.7742 -0.2046 -0.2721 -0.6244
SAL -0.1032  1.0000 -0.0513 -0.0206  0.1623 -0.4208
pH   0.7742 -0.0513  1.0000  0.0192 -0.0377 -0.7222
K   -0.2046 -0.0206  0.0192  1.0000  0.7921  0.0736
Na  -0.2721  0.1623 -0.0377  0.7921  1.0000  0.1170
Zn  -0.6244 -0.4208 -0.7222  0.0736  0.1170  1.0000
```

(a) Initially, the aim is to model biomass using a simple linear regression. Answer the following questions, giving justification.

i. Which is the best predictor variable for modelling biomass using a regression line? What is the coefficient of determination associated with it, and how should this value be interpreted?

ii. For the regression line selected in the previous item,

A. Calculate the equation of the fitted line.

B. Show that the estimate of the common variance of the random errors is 178 565.9.

C. Construct a 95% prediction interval for an individual observation of the above-ground biomass of *Spartina alterniflora* at a site where the soil has pH 4.5.

(b) A multiple linear regression model was then fitted for biomass using the other five observed variables, with the following results:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.252e+03  1.235e+03   1.014  0.31674
SAL         -3.029e+01  2.403e+01  -1.260  0.21508
pH           3.055e+02  8.788e+01   3.477  0.00126
K           -2.851e-01  3.484e-01  -0.818  0.41817
Na          -8.673e-03  1.593e-02  -0.544  0.58926
Zn          -2.068e+01  1.505e+01  -1.373  0.17746
---
Residual standard error: 398.3 on 39 degrees of freedom
Multiple R-squared: 0.6773,     Adjusted R-squared: 0.6359
F-statistic: 16.37 on 5 and 39 DF,  p-value: 1.082e-08
```

i. Without performing calculations, comment on the following statement: *even without performing additional tests, the analysis of the last column of the results table, containing the p-values (p-values), implies that a simple linear regression on the variable* `pH` *would produce a fit that does not differ significantly from the model with the five predictors.* Justify.

ii. Regardless of your answer to the previous item, formally test whether the fit of the multiple linear regression model with the five predictors differs significantly ($\alpha = 0.10$) from the fit of the simple linear regression model considered in question 7a. Comment.

iii. Can it be assumed that, all else being equal, for each additional ppm in soil zinc concentration, the above-ground biomass of this plant decreases, on average, by 40 $g\ m^{-2}$? Answer this question using a 95% confidence interval.

iv. Starting from the model above, a sequential elimination algorithm based on t-tests of the model coefficients β_j, selected a submodel with four predictors. State, with justification:

A. Which are the four predictors;

B. What is the coefficient of determination of this submodel.

8. In a study of lettuces, it was considered useful to model soil electrical conductivity from its chemical composition. In 27 soil samples, in addition to electrical conductivity, (variable `CE`, in microsiemens per centimetre, $\mu S/cm$), the concentration of sodium cations (variable `Na`), potassium (variable `K`), calcium

(variable `Ca`) and magnesium (variable `Mg`), all in $cmol/kg$. The correlations between each pair of variables are given below.

```
         CE       Na        K      Ca      Mg
CE  1.00000 -0.08618 -0.05741 0.17778 0.55346
Na -0.08618  1.00000  0.33811 0.20362 0.30437
K  -0.05741  0.33811  1.00000 0.08943 0.38161
Ca  0.17778  0.20362  0.08943 1.00000 0.64198
Mg  0.55346  0.30437  0.38161 0.64198 1.00000
```

A multiple linear regression model was fitted using all cation concentrations as predictors. The following results were obtained:

```
Call: lm(formula = CE ~ Na + K + Ca + Mg, data = Alfaces)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -815.76     446.45  -1.827  0.08126 .
Na           -102.21      84.52  -1.209  0.23935
K            -405.78     207.13  -1.959  0.06291 .
Ca            -57.68      29.61  -1.948  0.06432 .
Mg            668.17     145.34   4.597  0.00014 ***
---
Residual standard error: 68.07 on 22 degrees of freedom
Multiple R-squared: 0.5147, Adjusted R-squared: 0.4264
F-statistic: 5.832 on 4 and 22 DF,  p-value: 0.002344
```

(a) Test the overall fit of the model. Discuss the quality of this fit. In your discussion, also take into account the values of the ordinary and adjusted coefficients of determination.

(b) Interpret, in the context of the problem under study, the meaning of the value -57.68 in the first column of the table, giving the units of measurement of this value.

(c) A researcher claims that when the concentration of sodium cations increases, mean electrical conductivity decreases. Assuming the validity of the model, and placing the burden of proof on the researcher's hypothesis, what conclusion can be reached based on the data ($\alpha = 0.05$)?

(d) It was stated that, based on the results of the model fitted above, the only one of the four predictor variables that is important for modelling electrical conductivity is the variable `Mg`.

   i. Without doing any calculations, state whether you consider this statement to be justified.

   ii. Regardless of your answer to the previous item, formally compare formalmente o model presented above with the simple linear regression in which the response variable `CE` has as its only predictor the variable `Mg`. Comment.