

HW 2

Exercise 1 (Sweetness of orange juice.).

To study the sweetness of orange juices, researchers collect some data on the sweetness index (y) and the amount of pectin (x) in the orange juice (in g/L). The dataset is `ORANGEJUICE.csv`.

- (a) Fit the model and find a 90% confidence interval for the true slope of the line. Interpret the result.
- (b) Fit the model and determine whether there is a positive or negative linear relationship between the amount of pectin x and the sweetness y . That is, determine if there is sufficient evidence (at $\alpha = 0.05$) to indicate that β_1 , the slope of the straight-line model, is significantly different from zero.

Exercise 2 (Car program).

A company bought a car or two each year. The data is shown in the file `COMPANYCAR.csv`.

- (a) Fit the simple linear regression model, $E(y) = \beta_0 + \beta_1$, to the data.
- (b) List assumptions required for the regression analysis.
- (c) Find the value of SSE.
- (d) Find the estimated standard error of the regression model, s .
- (e) Give a practical interpretation of s .
- (f) Find a 95% confidence interval for the true slope of the line.
- (g) Interpret the confidence interval in (f).
- (h) Find the p -value for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Use this result to test the simple linear regression model is statistically useful for predicting the annual cost using the year of initial operation. (Test using $\alpha = 0.05$)
- (i) Find and interpret the coefficient of determination, R^2 .
- (j) A researcher wants to estimate of the average annual cost of company cars with the year in 2020. Which interval is desired by the researcher, a 95% prediction interval for y or a 95% confidence interval for $E(y)$? Use R to calculate the desired interval.
- (k) Give a practical interpretation of the interval in part (j).

Exercise 3 (Fill in the blanks in the table and answer questions).

Look at the output of the linear regression model. Fill in the blanks in the table and answer questions:

```

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1067.78 -284.90 -26.95  247.33 1002.14 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -188.87003  223.02827 -0.847   0.401    
x             0.52725   0.04288  1. ---- 2. ---- 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3. ----- on 50 degrees of freedom
Multiple R-squared:  0.7514,   Adjusted R-squared:  0.7465 
F-statistic: 151.2 on 1 and 50 DF,  p-value: < 2.2e-16

```

Analysis of Variance Table

	Df	SS	MS	F	P
Source	1	29582640	4. -----	151.16	9.8981e-17
Error	50	9785481	5. -----		
Total	51	39368121			

- (a) Please fill in the blanks in the table.
 - (1)
 - (2)
 - (3)
 - (4)
 - (5)
- (b) Find and interpret the coefficient of determination, r^2 .
- (c) Calculate the coefficient of correlation, r .

Exercise 4 (Position effects in memories).

Here is an experiment. There are 9 words on screen and participants are requested to recall these words as many as possible. The results are stored in the file WORDMEMORY.csv: each row represents a word and the recall rate of the word at this position is recorded.

- (a) Find a 99% confidence interval for the mean recall proportion for words in the fifth position. Interpret the result.
- (b) Find a 99% prediction interval for the recall proportion of a particular word in the fifth position. Interpret the result.
- (c) Compare the two intervals, part (a) and part (b). Which interval is wider?
Will this always be the case? Explain.