Analysis of Scientific Data – Summer Semester, 2022

Week 6 Tutorial Solutions

PART A - Linear Regression

Question 1

Lung function data for 25 cystic fibrosis patients in a study by O'Neill *et al.* (1983) was presented in Douglas Altman's *Practical Statistics for Medical Research* (1991). The data contains a number of variables recorded for each patient and here we are interested in the relationship between two of these, **weight** (weight, kg) and **pemax** (maximum expiratory pressure, cm of H_2O).

a) The R output for the linear model is presented below. Compute the values of A and C, and then give a value for B.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.5456          A   5.003 4.63e-05 ***
weight        1.1867     0.3009   3.944 B

Residual standard error: 26.38 on C degrees of freedom
```

$C = n - 2 = 25 - 2 = 23.$

$A = \text{se}(b_0).$

Since $t = (b_0 - 0)/\text{se}(b_0),$

$5.003 = (63.5456 - 0)/A,$

so $A = \frac{63.5456}{5.003} = 12.70.$

$B = 2 * P(t_{23} \geq 3.944).$

In R using `1-pt(3.944, df=23)`

```
[1] 0.000323207
```

$B = 2 \times 0.000323 = 0.000646.$

b) Based on the linear model, what is the estimated mean maximum expiratory pressure for patients with a weight of 50 kg?

Regression equation is $\hat{y} = 63.5456 + 1.1867x$. Substituting $x = 50$ gives

$\hat{y} = 63.5456 + 1.1867 \times 50 = 122.88$ cm H_2O

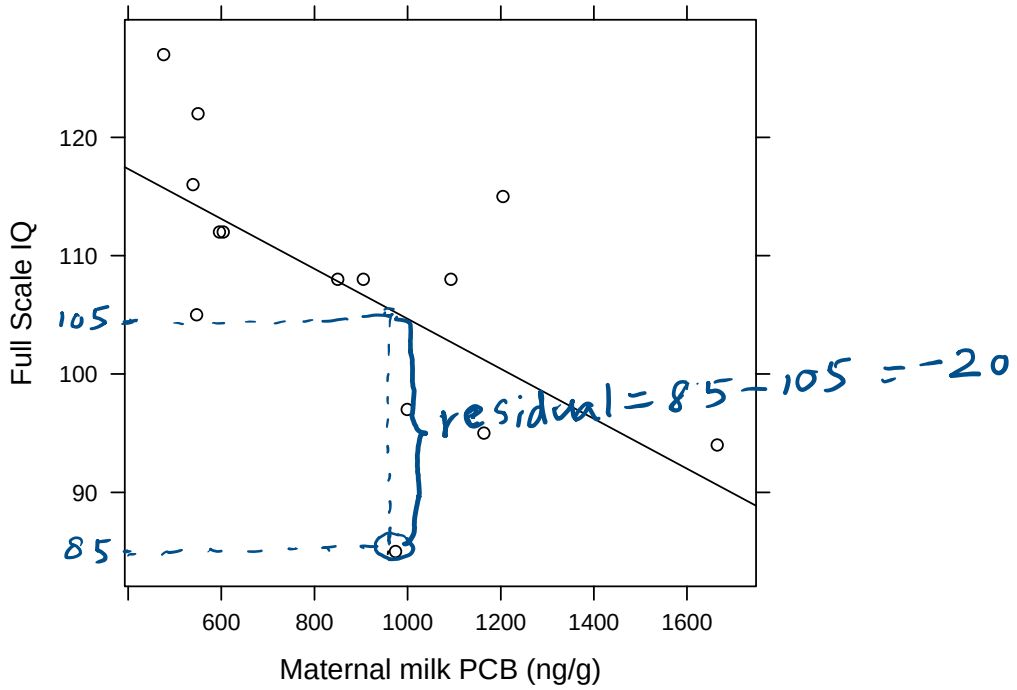
c) Conduct a hypothesis test to determine whether there is any evidence of a linear association between **pemax** and **weight**. What do you conclude?

Test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$. From the table in (b), the p-value is 0.000646.

Thus, we have strong evidence to suggest that there is a linear relationship between **pemax** and **weight**.

Question 2

Polychlorinated biphenyls (PCBs) were once used in industry but were banned in the 1970s because of concerns about their toxicity. Despite the ban, PCBs can still be detected in most people because they are persistent in the environment. A team of researchers recorded the amount of PCBs detected in maternal milk from mothers who had eaten fish from a particular lake considered to be contaminated with PCBs. They subsequently administered an IQ test to the children when they were 11 years old. The results are shown in the following scatter plot along with the least-squares line fitting a linear relationship between the two variables:



A regression analysis in R produced the following (edited) summary:

```
Coefficients:
              Estimate Std. Error
(Intercept) 125.773972   7.008028
PCB          -0.021109   0.007538

Residual standard error: 9.314 on 12 degrees of freedom
Multiple R-squared: 0.3952, Adjusted R-squared: 0.3448
F-statistic: 7.842 on 1 and 12 DF
```

a) Based on the degrees of freedom, how many pairs of observations were used in the analysis?

df $= n - 2 = 12$, so 14 observations. (Can check by counting the dots.)

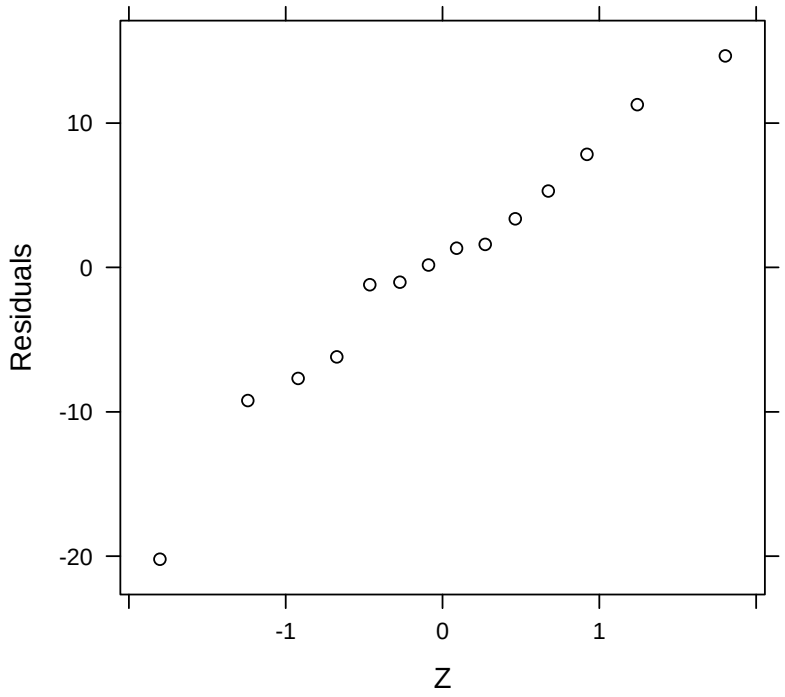
b) Based on the coefficients for the least-squares line provided by R, estimate the mean IQ of children if their mothers had a maternal milk PCB measurement of 1400 ng/g.

Regression equation is $\hat{y} = 125.773972 - 0.021109x$.

Substituting $x = 1400$ gives

$\hat{y} = 125.773972 - 0.021109 \times 1400 = 96.2217$

c) Which assumption of linear regression does the following plot check? Comment on the validity of the assumption here.



Normal probability plot shows points are close to a straight line. Normality of errors assumption is not violated.

d) Circle the point with a residual of -20 in the original scatter plot.

See the scatter plot

e) Is there evidence of a negative association between maternal milk PCB levels and IQ outcome?

If β_1 is the underlying slope of the relationship between IQ and PCB, we want to test H_0: $\beta_1 = 0$ vs H_1: $\beta_1 < 0$. The t statistic is

$$t = \frac{b_1 - 0}{\text{se}(b_1)} = \frac{-0.0021109 - 0}{0.007538} = -2.800345$$

The p-value is $P(T_{12} \leq -2.800345)$.
Using R `pt(-2.800345, df=12)`

p-value is 0.008, giving strong evidence to suggest that there is a negative association between IQ outcome and maternal milk PCB.

PART B – Linear Regression and ANOVA using RStudio

The effect of alcohol on the human body depends on the blood alcohol concentration. The blood alcohol concentration (BAC measured in g/dL) depends on many factors such as the number of standard drinks (Drinks), sex (Sex), and body mass (Mass measured in Kg). A study assigned a target alcohol dosage (Low, Medium, and High) for 20 individuals. The actual standard drinks consumed was recorded along with the blood alcohol concentration from a urine sample after a fixed waiting period. The data is given in the "BloodAlcohol.csv" file in the Week 6 folder.

(See Week6PartB.R file for R codes)

a) What is the estimated least square line to examine the relationship between blood alcohol concentration and body mass?

$\widehat{BAC} = 0.1585 - 0.0012Mass$

b) Based on the model what is the 95% confidence interval for the mean blood alcohol concentration for a person with body mass 70kg?

(0.0658, 0.0841) g/dL

c) The association between blood alcohol concentration and body mass may be obscured by the differences in the number of standard drinks consumed and the sex. Assuming there is no interaction fit a multiple linear regression model for blood alcohol concentration with body mass, number of drinks and sex as predictors (explanatory variables). What is the coefficient estimate for the variable Sex. Interpret it.

The population regression equation is

$BAC = \beta_0 + \beta_1 Mass + \beta_2 Drinks + \beta_3 Sex + U$

Coefficient estimate for the variable Sex = 0.0045

For a given body mass and a number of standard drinks consumed, the mean blood alcohol concentration for males is 0.0045 g/dL is higher than that of for females.

d) Based on the model in part c), what is the estimated blood alcohol concentration for a male person with body mass 60kg who has had 1.8 standard drinks?

0.079 g/dL

e) Based on the model in part c), what is the p-value to test whether there is a statistically significant positive linear relationship between blood alcohol concentration and the variable Drinks?

We need to test the following hypotheses.

$H_0: \beta_2 = 0$ Vs $H_0: \beta_2 > 0$

Note: Consider the row corresponding to the variable Drinks in the R output. The p-value given in there is to test the hypotheses $H_0: \beta_2 = 0$ Vs $H_0: \beta_2 \neq 0$. Thus, the p value to test for a positive linear relationship should be half of the corresponding p-value from the R output.

p-value = 0.1621/2 = 0.08105

f) The researcher uses One-Way ANOVA model to test if there is a difference in population mean blood alcohol concentration levels between the different dosage groups. What is the R^2 value of this model? Interpret it.

$$R^2 = \frac{0.003227}{0.003227 + 0.006250}$$

$$R^2 = 0.3405$$

34.05% of variability in blood alcohol concentration is explained by the variability in alcohol dosage level.

g) The researcher further performs multiple comparisons using Tukey's Honestly Significant Difference and 5% level of significance. Comment on the results.

At 5% level of significance, the population mean blood alcohol concentration levels are different between low and high alcohol dosage groups only.

PART C – One-Way ANOVA

Early developmental experiences, such as incubation conditions, can have important consequences for post-hatching fitness in birds. A paper reported on a study where wood duck eggs were collected from nest boxes and experimentally incubated at various fixed temperatures, each falling within the range of temperatures of naturally incubated wood duck nests. The response variables recorded included egg mass (g) at the time of hatching. A one-way analysis of variance of egg mass by temperature in R gave the following results:

	Df	Sum Sq
Temperature	3	117.1
Residual	23	312.5

a) How many fixed temperatures were used in this design?

Since DF Temperature = 3 = $k - 1$, there must have been 4 fixed temperatures.

b) How many eggs were included in this analysis?

DF Total = $3 + 23 = 26 = n - 1$, so there were 27 eggs.

c) The paper reported that the mean egg mass for the 6 eggs at 35 °C was "44.7 ± 1.88 g" where they indicated that "1.88" was the standard error of the mean (SEM). Based on this statement, what was the sample standard deviation of the egg mass for those 6 eggs?

SEM = $\frac{s}{\sqrt{n}}$, so $1.88 = \frac{s}{\sqrt{6}}$, giving $s = \sqrt{6} \times 1.88 = 4.61$ g

d) What is the R^2 value for this model?

$$R^2 = \frac{\text{SS Temp}}{\text{SS Total}} = \frac{117.1}{117.1 + 312.5} = 0.27$$

Thus, incubation temperature explains 27% of the variability in egg mass.

e) What is the F statistic and p-value used to test for the difference in mean egg mass between the different temperatures?

$$F = \frac{117.1/3}{312.5/23} = 2.872852$$

Using R:
```
1-pf(2.872852, 3, 23)
[1] 0.05828635
```

$P(F_{3,23} \geq 2.872852) = 0.0583$, only weak evidence of a difference in mean egg mass between the four incubation temperatures.