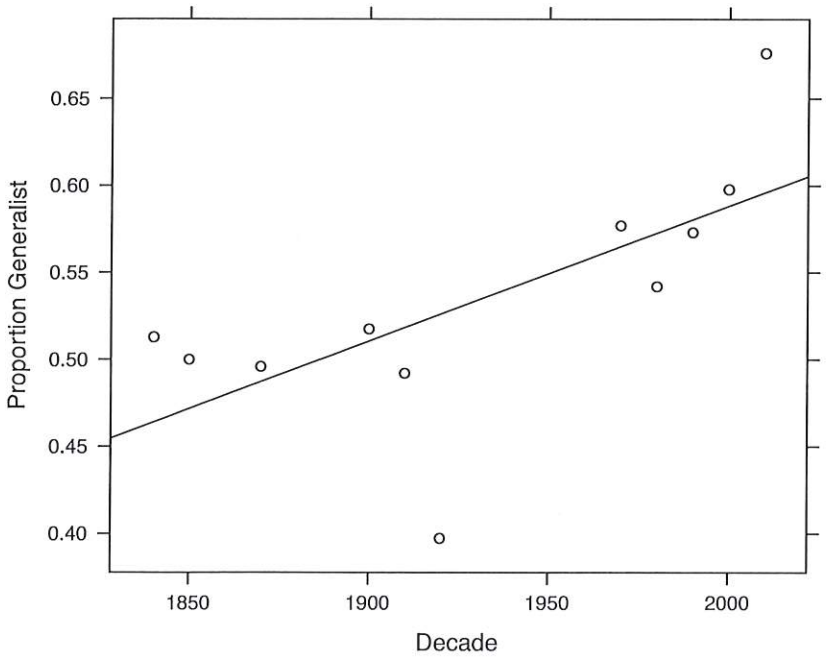
Question 2

11 marks

Environmental changes strongly impact the distribution of species but little is known about how communities of species in different habitats transform over long time frames. To address this, a recent study analyzed changes in the species composition of a southeastern German butterfly community over nearly two centuries (1840–2013). They classified all species observed over this period, according to various ecological and behavioural traits, as either a habitat specialist (preferring a particular type of habitat) or a habitat generalist.

For each of 11 decades of available data, the researchers calculated the proportion of the species that were generalists. The figure below displays the proportion of generalists observed for each decade along with the least-squares line:



The output on the next page shows the results of a linear regression in R for the relationship between the proportion of generalist species and decade:

```
lm(formula = PropGeneral ~ Decade)

Residuals:
      Min        1Q    Median        3Q       Max
-0.128950 -0.016963  0.008562  0.020019  0.079940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9651151  0.5515491
Decade       0.0007767  0.0002855
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05601 on 9 degrees of freedom
Multiple R-squared:  0.4513,  Adjusted R-squared:  0.3903
F-statistic: 7.401 on 1 and 9 DF,  p-value: 0.02359
```

(a) Briefly interpret the value -0.9651151 in the regression output. [1 mark]

Estimated "proportion" of generalist species in year 0. (unrealistic negative proportion)

(b) Briefly interpret the value 0.0007767 in the regression output. [1 mark]

Estimated proportion of generalist species increases by 0·0007767 per year

(c) Does the regression analysis provide evidence of a change in the proportion of butterfly species that are generalist over time? State the null and alternative hypotheses, and report the appropriate test statistic and P-value. What do you conclude? [3 marks]

let β_1 = true slope in the regression model.

test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$$\text{test statistic} = \frac{\text{estimate} - \text{hypothesis}}{\text{s.e.(estimate)}} = \frac{0.0007767}{0.0002855} = 2.7205$$

$$\text{p-value} = 2 \times \min\{\mathbb{P}(T_9 \geq 2.7205), \mathbb{P}(T_9 \leq 2.7205)\}$$

degrees of freedom $= n - 2 = 11 - 2 = 9$ $= 2 \times 0.0118 = 0.0236$

moderate evidence against the null, suggestion the proportion changes over time.

(d) Give a 95% confidence interval for the underlying slope of the linear relationship between the proportion of generalist species and time. [2 marks]

estimate $\pm$ (critical value) $\times$ s.e.(estimate)

95% CI $0.95 = 1 - \alpha \Rightarrow \alpha = 0.05$

critical value $t_{1-\alpha/2;\, n-2} = t_{0.975;\, 9} = 2.2622$

$0.0007767 \pm 2.2622 \times 0.0002855$

95% confident the true slope is between 0.0001308 and 0.0014.

(e) In 1920 the observed proportion of habitat generalists was 0.397. What is the residual associated with 1920 in the regression model? [1 mark]

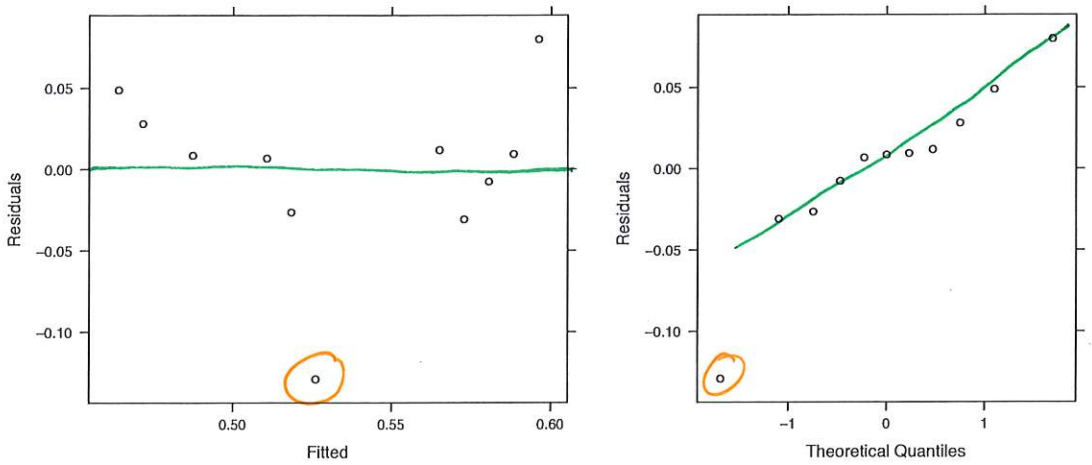
fitted value $= -0.9651151 + 0.0007767 \times 1920$

$= 0.5261$

residual = observed − fitted value

$= 0.397 - 0.5261 = -0.1291$

(f) The following figures were generated by R to help check the assumptions underlying the linear regression:



Comment on the validity of the assumptions underlying linear regression for this data with reference to these figures. [3 marks]

Assumption

1. mean response is linear with explanatory variable.
2. response has normal distribution
3. response has constant variance

(4. Independence).

Figure Left : - no obvious trend consistent with linearity assumption.

- even spread of residuals consist with constant variance assumption.

Figure Right - with exception of residual (-0.1291), points line close to a straight line consistent with normality.

Question 5 [18 marks]

Children with malignant disease are at increased risk of bone disorders and cardiovascular disease. A study aimed to explore if vitamin D status may influence this risk, by measuring vitamin D levels in children with malignant disease and comparing this to a control group of children (with no malignant disease). The results are shown in the following table:

Vitamin D	Deficient	Not Deficient	
Control children	6	54	60
Children with malignant disease	13	48	61

(a) Calculate the sample proportion of children with malignant disease who have a vitamin D deficiency. *[2 marks]*

$\frac{13}{61} \simeq 0.2131$

(b) The researchers believed that the incidence of vitamin D deficiency would be higher for children with malignant disease. Briefly explain why a chi-squared test would not be appropriate for testing this belief. *[2 marks]*

(c) Is there evidence that the incidence of vitamin D deficiency is higher for children with malignant disease compared to the control children? *[6 marks]*

p_1 = proportion of control who are deficient

p_2 = " disease "

Test $H_0: p_1 = p_2$ vs $H_1: p_1 < p_2$

observe $n_1 = 60$, $\hat{p}_1 = \frac{6}{60} = 0.1$ $n_2 = 61$, $\hat{p}_2 = \frac{13}{61} \simeq 0.2131$

$$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad \hat{p}_{pool} = \frac{6+13}{60+61} = 0.157$$

$$= 0.0662$$

$$\text{test stat} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s.e.(\hat{p}_1 - \hat{p}_2)} \quad \leftarrow \text{under } H_0 = \frac{(0.1 - 0.2131) - 0}{0.0662} = -1.7099$$

$$p\text{-value} = P(Z \leq -1.7099) = 0.0436$$

There is moderate evidence against H_0, suggesting higher incidence of vitamin D deficiency in children with disease.

Inspired by these initial results, the researchers carried out a further study with new groups of children in which they classified vitamin D levels into *three* groups (deficient, insufficient and sufficient) to gain a better understanding of its role. The results from this second study are given below.

Vitamin D	Deficient	Insufficient	Sufficient
Control children	11	28	50
Children with malignant disease	20	33	40

(d) Using this data, is there evidence of an association between vitamin D levels and malignant disease in children? *[8 marks]*

H_0: no association between 'disease status' and 'vitamin D level'

H_1: some association between ...

Expected counts (row total x col total / total) | Using MATLAB/R

	Def	Ins	Suff
Control	15.16	29.83	44.01
disease	15.84	31.17	45.99

$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i} = 4.0479$$

degrees of freedom = (#rows − 1) × (#cols − 1)
= (2−1) × (3−1) = 2

p-value = $\mathbb{P}(\chi^2_2 \geq 4.0479) = 0.1321$

no evidence against the null hypothesis, suggesting no association between disease and vitamin D levels.

END OF EXAMINATION