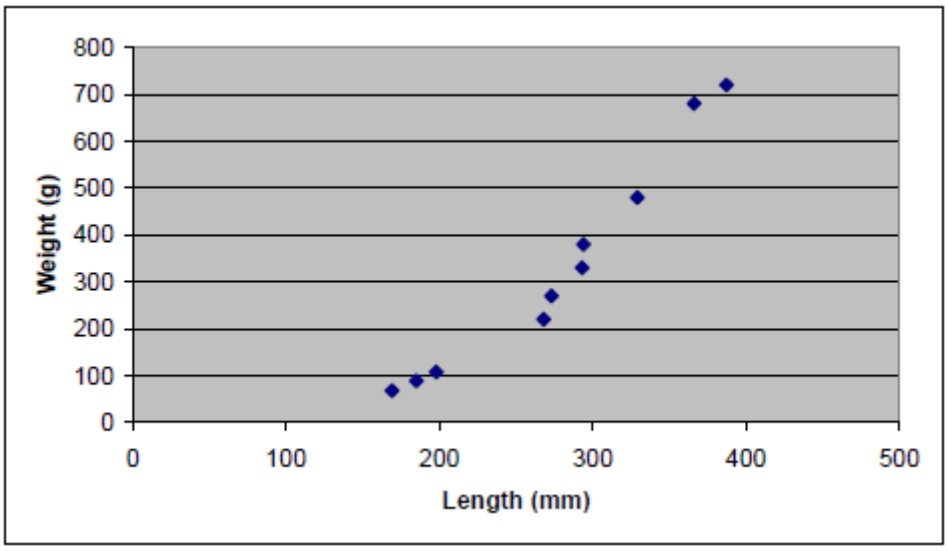
RSS Higher Certificate in Statistics, Specimen B

Module 4: Linear Models

Solutions

Question 1

(i) グラフは以下のようである，このグラフより，最初に行う近似としては，少なくとも直線のあてはめが妥当であることが言える．ただし，重さと長さの関係に曲線関係があるかもしれない．



(ii) (a) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 1075861 - \frac{2762 \times 3345}{10} = 1519720,$$

$S_{xy} = 812594 - 2762^2/10 = 497296$ より $\hat{\beta}_1 = 3.056$ である．

また，$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 334.5 - 3.056 \times 276.2 = -509.56$ である．

(b) 始めの 3 点は全く当てはまっていない．切片は -509.56 であるが，仮に 3 点がぴったり当てはまっているならば切片はずっと 0 に近くなるように見えることに注意せよ．残りの点はわりとよく当てはまっている．しかしながら 2 次の関係も調べるのがよいだろう．

(iii) 決定係数は $R^2 = S_{xy}^{\ 2} / S_{xx} S_{yy}$ で与えられ，$S_{yy} = 1610009 - 3345^2/10 = 491106.5$ より

$$R^2 = \frac{1519720^2}{497296 \times 491106.5} = 0.9457$$ である．

このようにスズキの重さの変動全体の 94.6%が，長さとの線形関係によって説明される．

Question 2

(i) モデルは

$$y_{ij} = \mu + t_i + \varepsilon_{ij} \quad i = 1,2,...,k,\ \ j = 1,2,...,r_i \quad \{\varepsilon_{ij}\} \sim \text{ind N}(0,\sigma^2)$$

と表わされる．k 個の処理があり，添字 $i = 1, 2, \ldots, k$ で表わされる．実験あるいは調査で，i 番目のグループ，つまり i 番目の処理を受けている群には r_i 個の個体がある．y_{ij} は i 群の j 番目の個体に対する観測値（応答）である．また μ は母集団全体の平均であり，t_i は処理 i における（μ からの偏差としての）効果の母平均で $\sum t_i = 0$ を満たす．
残差（誤差）項 ε_{ij} は独立（無相関）で，分散が σ^2 の正規分布にしたがう．

これは加法モデルである．（すべての項の和を考慮し，すべての項で応答変数の変動を説明する．）

(ii) ここでの「処理」は「High」と「Low」と「Work」である．$r_1 = r_2 = r_3 = 12$ であり，和は，High : 5528，Low : 3754，Work : 3511　である．全合計は 12793，$\sum\sum {y_{ij}}^2 = 5719139$，

修正項は $\dfrac{12793^2}{36} = 4546134.694$ であるから，

全平方和は 5719139 – 4546134.694 = 1173004.306,

処理平方和は $\dfrac{5528^2}{12} + \dfrac{3754^2}{12} + \dfrac{3511^2}{12} - 4546134.694 = 202067.056$

であり，残差平方和はこれらの引き算で得られる．
よって，分散分析表は以下のようになる．（SS と MS はわずかに丸め誤差がある．）

SOURCE	DF	SS	MS	F value
Treatments	2	202067	101034	3.43　Compare $F_{2,33}$
Residual	33	970937	29422	$= \hat{\sigma}^2$
TOTAL	35	1173004		

$F_{2,33}$ の上側 5%点は約 3.3 なので，処理効果は有意である．したがって，すべての処理の効果が同じであるという帰無仮説を棄却するための根拠が認められる．

処理の違いを調べるために，まず処理の平均を計算すると（わかりやすく小さい順に並べると）

Work: 292.58　Low: 312.83　High: 460.67

となる．これらの任意の組における最小有意差は

$$t_{33}\sqrt{\frac{2\times 29422}{12}} = 70.026 t_{33} \quad where\ \ t_{33} = \begin{cases} 2.035 & at\ 5\% \\ 2.736 & at\ 1\% \\ 3.617 & at\ 0.1\% \end{cases}$$

であるので，5%最小有意差は 142.50，1%最小有意差は 191.59，0.1%最小有意差は 253.28 となる．したがって「High」の平均の値が大きく，5%水準においては「High」との間に有意差が認められる．「Low」と「Work」にはさほど差はない．

レポート

群内変動に対してグループの平均を比較する分析を行った結果，「High」は他の 2 グループよりも根気があり，他の 2 グループはかなり似ているということに対する根拠が認められた．群内変動は非常に大きい．

Question 3

(a)

(i) 複数の被験者が同じ得点を取った場合は平均順位を用いる．これは (ii) で数人の被験者に見られる．

(ii) 順位および順位差は以下のようになる．

Subject	1	2	3	4	5	6	7	8	9	10	11	12
IT rank	5	12	1	10	11	6	8	3	4	2	9	7
LT rank	2	12	6½	8½	10½	1	4½	4½	6½	3	10½	8½
Difference d_i	3	0	−5½	1½	½	5	3½	−1½	−2½	−1	−1½	−1½

$\sum d_i^2 = 9+0+\cdots+2.25=93$ より，スピアマンの順位相関係数は

$$r_s = 1-\frac{6\sum d_i^2}{n(n^2-1)} = 1-\frac{558}{1716} = 0.675$$

である．数表より，これは（両側）5%水準において 0 とは有意な差がある．

（注：タイがある場合に r_s にこの計算式を使うとわずかな誤差が生じるが，ここでは大差ないようである．）

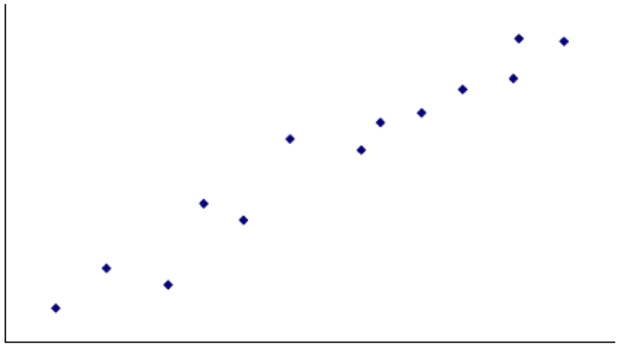
(iii) 相関がないという帰無仮説を棄却する根拠が認められるので，知能と水平思考能力との間の関連性はありそうである．ただ，その関係はそれほど強いものではないようである．

(b) 積率相関係数の式は

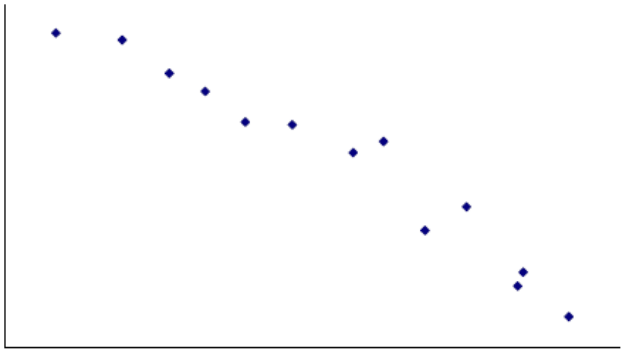
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

のように表される．これは x_i と y_i の間の線形関係の強さを示していて，$r = \pm 1$ のとき線形関係があり $r = 0$ のとき線形関係がないことを表している．基本的に x と y は共に確率変数である．

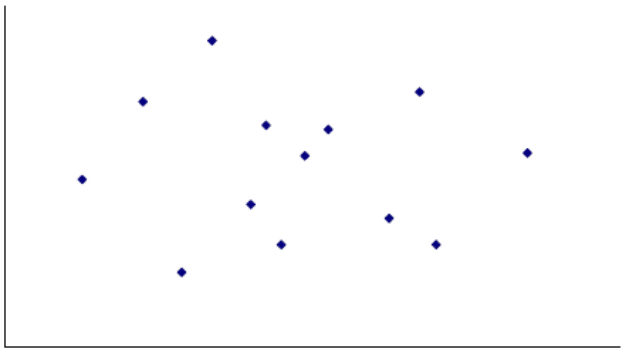
(i) r が +1 に近い場合．（増加する）直線関係の近くに散布している．



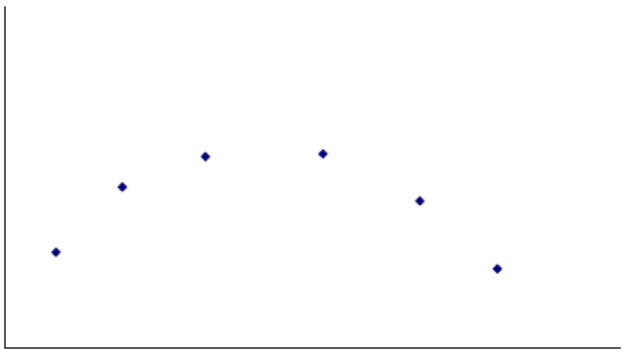
(ii) r が -1 に近い場合．(a)とは違い x が増加すると y は減少する．



(iii) 独立したデータ $(r \approx 0)$



(iv) 線形でない関係がある場合．たとえば $y = x^2$.



Question 4

(i) 実験のマテリアルの間（プロット間，ユニット間）に確率的な自然変動が存在することを認識し，それを考慮しなければならない．処置における平均応答の妥当な推定値を得たり，処置がなされたプロット間の自然誤差を適切に推定するために，処置に対する応答を評価する際いくつかの単位プロットからの結果が用いられなければならない．これが反復が重要である理由である．

個々のユニットへの処置割付けを無作為化することにより，特定の処置を都合のよいユニットやあるいは都合の悪いユニットに意識的に割付けるようなことがなくなるので，バイアスを排除する手助けとなる．各々のユニットに割付けられる確率は，どの処置についても等しい．通常の分散分析モデル（詳しくは Question 2 を参照のこと）では，観測値が同じ分散を持ち，相関がないことが要求される．無作為化はこれを実現するための一助となる．（注：ブロック化のような他の工夫も必要とされるかもしれない．これらは Graduate Diploma の一部である Higher Certificate の Module 6 でさらに詳しく扱われる．）

(ii) 全合計は 147.70，29 個すべての観測値の平方和は 831.8900 である．（問題中に与えられている．）

「修正項」は $\frac{147.70^2}{29} = 752.2514$ であるから，全平方和は 831.8900 − 752.2514 = 79.6386,

種における平方和は $\frac{38.80^2}{9} + \frac{47.50^2}{12} + \frac{61.40^2}{8} - 752.2514 = 74.2855$，残差平方和はこれらの引き算で得られる．したがって分散分析表は以下のようになる．

SOURCE	DF	SS	MS	F value
Species	2	74.2855	37.1428	180.4 Compare $F_{2,26}$
Residual	26	5.3531	0.2059	$= \hat{\sigma}^2$
TOTAL	28	79.6386		

$F_{2,26}$ の上側 0.1%点は 9.12 なので，種の効果は非常に高有意である．すべての種が類似したふるまいをみせるという無帰仮説を棄却するのにかなり強い根拠がある．

以下の (a) (b) では，t 検定により有意性を調べる．

(a) *pinus* 全体の平均は $\frac{38.8+47.5}{21} = 4.110$ で，*eucalyptus* の平均は 7.675 である．

差の分散は $\frac{\hat{\sigma}^2}{21} + \frac{\hat{\sigma}^2}{8} = 0.035542$ で推定される．したがって t 統計量は $\frac{7.675-4.110}{\sqrt{0.035542}} = 18.91$ となり，これは t_{26} からの観測値としてこれは明らかに極端な値で

ある．したがって *pinus* と *eucalyptus* の間には成育した木の高さの平均に違いがあるということの圧倒的な根拠がある．

(b)　2 つの *pinus* の平均はそれぞれ $\frac{38.8}{9}=4.311$ と $\frac{47.5}{12}=3.958$ であり，

差の分散は $\frac{\hat{\sigma}^2}{9}+\frac{\hat{\sigma}^2}{12}=0.040036$ で推定される．したがって t 統計量は $\frac{4.311-3.958}{\sqrt{0.040036}}=1.76$ となり，t_{26} からの観測値としてこれは有意でない（両側 5%点は 2.056）．したがって 2 つの *pinus* 間に生育した木の高さの平均に差があるという根拠は認められない．

残差（観測値－あてはめ値）はここでは（観測値－処理平均値）であり，下表のようである．

PC	–0.11	–0.01	–0.81	–0.41	0.69	0.49	0.29	0.19	–0.31			
PK	–0.01	–0.11	0.29	0.74	0.19	–0.66	–0.31	–0.26	–0.01	0.04	–0.26	0.34
ED	0.27	0.42	0.62	–1.08	–0.18	0.02	–0.43	0.32				

(PC:Pinus caribea　PK:Pinus kesiya　ED:Eucalyptus deglupta)

以下のドットプロットを見ると，PK（Pinus kesiya）が最も「正規的」に見える．ED（Eucalyptus deglupta）はばらついており，外れ値が存在している可能性がある．