RSS Higher Certificate in Statistics, 2010

Module 4: Linear Models

Solutions

Question 1

(a) (i) 比較を目的とした実験において，各処理（例：肥料）を実験材料（例：与えられた種の小麦）に*無作為に*割り当てる必要がある，というのが無作為化の考えである．この実験のこの例を用いると，同じ種類の小麦同士の確率的なばらつきが平均化されるため，収穫量の相違が処理（肥料）の違いを表す傾向が強くなる．実験が行われている畑の土壌の豊かさが持つ一定の“肥沃勾配”のような，バイアスのあらゆる原因（ことによると思いもよらない原因）の排除が，無作為な割り当てによって促進される．データの解析は，処理間の収穫量のばらつきと処理内のランダムなばらつきを比較することに焦点を当てている．後者のばらつきに対して前者のばらつきが大きいほど，収穫量の違いの原因は偶然生じたものというよりは処理の影響にあると言えよう．

解析の精度を上げるために，各処理をいくつかの*反復*（ランダムなばらつきをなくすために同一と考えられ繰り返し用いられる条件．たとえば，標準的な土地区画や個々の植物）に適用する必要がある．これは各処理に割り当てられている実験材料内の無作為なばらつきを平均化するためである．

試験で受験者が挙げる例では，実験材料と処理（上記の例では小麦と肥料）をはっきりと識別してほしかった．系統的なばらつきと，無作為な処理の割り当てを実験材料に反映しているランダムなばらつきとを比較する，という主要な目的に言及してほしかった．

(ii) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ ， $i = 1,...,k$ ， $j = 1,...,r$

$i = 1,...,k$: 処理の添字

$j = 1,...,r$: 反復の添字

μ : 真の全体の平均収穫量（母平均）

α_i : 全体の平均収穫量に対する i 番目の真の平均効果（母平均効果）

ε_{ij} : i 番目の処理の j 回目の反復における誤差， $N(0,\sigma^2)$ に従う独立正規確率変数

(b) (i) 表の 20 個の観測値の合計は 50+75+85+100=310，平均は 310/20=15.5，平方和は 544+1181+1505+2044=5274.

“修正項”は $\frac{310^2}{20} = 4805$

よって全平方和は 5274 - 4805=469（自由度 19）.

処理（繊維内の綿の割合，%）平方和は $\frac{50^2}{5}+\frac{75^2}{5}+\frac{85^2}{5}+\frac{100^2}{5}-4805=265$ （自由度 3）．
残差平方和は 469 – 265=204，自由度は 19 – 3=16

以上から，分散分析表は以下のようになる．

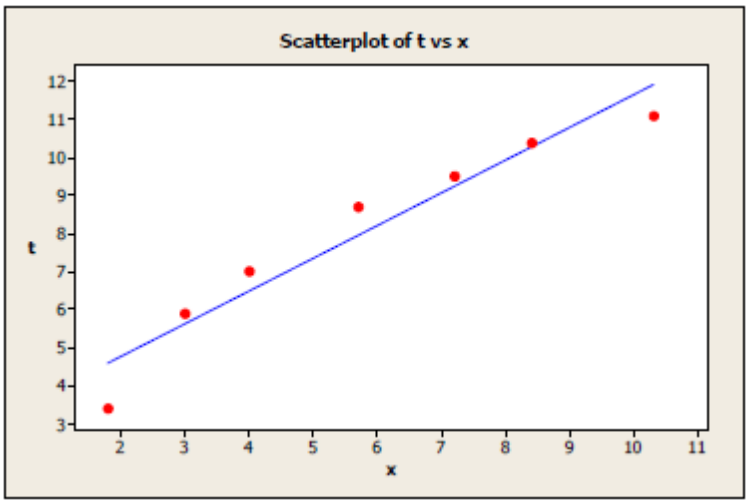
SOURCE	DF	SS	MS	F value
% cotton	3	265	88.33	6.93 Compare $F_{3,16}$
Residual	16	204	12.75	$=\hat{\sigma}^2$
TOTAL	19	469		

検定の正式な有意水準は問題で指定されていない．しかし，$F_{3,16}$ の上側 0.5%点は 6.30 で，分散分析から得られる F 値はこれを上回っているので，綿の割合の影響は大変高有意である．繊維内の綿の割合の違いが張力に影響を与えるという非常に強い証拠がある．

(ii) 綿の割合が間隔尺度に基づく比率であるから，綿の割合が増加するときに張力(TS)の傾向があると予想するかもしれない．データの表から明らかに，綿の割合に伴って TS の平均が増加していることが読み取れ，この増加が線形的であるかを調べるのが自然だろう．その手順としてまず，綿の割合の値 1 つあたりに TS の観測値が 5 つずつあることに注意しながら，TS（従属変数）の 20 個の観測値を綿の割合（独立変数）に回帰させる．この回帰の平方和(SS)は，綿の割合の影響の線形性に起因するだろう．(b)(i)での分散分析全体でこの SS は綿の割合ごとの平方和を少し下回るが，その値 265 は，綿の割合への非線形（2 次，3 次など）依存性に起因する SS を表している．よって，単純な線形回帰モデル（または比例モデル）の妥当性を検定することができる．この詳細は解答に不要である．

Qustion 2

(i)



[注：原点は入れていない．]

t の値は x に伴い強い増加傾向を示しているが，直線というよりは，勾配が減少していく曲線に見える．直線モデルでは，x の範囲の中央付近で t の値が小さく評価され，x の値が小さいときや大きいときは t の値が大きく評価されると言えよう．

(ii) 関係式 $\exp(t) = Ax^B$ の（e を底とした）対数をとると $t = \log A + B\log x$．
これを $t = a + b\log x$ と比べれば $a = \log A,\ b = B$ となる．

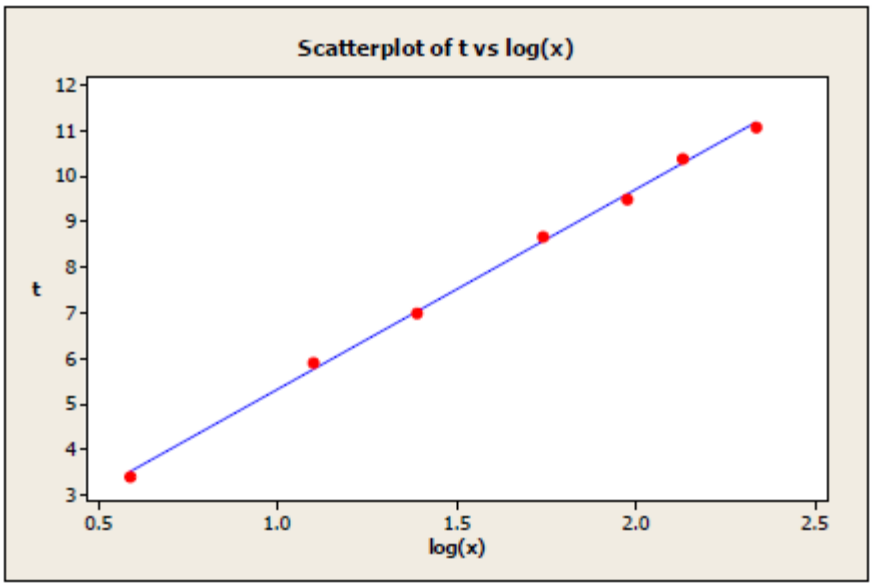
(iii) 問題文のデータを標準的な単純線形回帰の公式に代入して，

$$\hat{b} = \frac{n\sum t\log x - \sum t\sum \log x}{n\sum(\log x)^2 - \left(\sum \log x\right)^2} = \frac{(7\times 100.101) - (56\times 11.2476)}{7\times 20.3687 - 11.2476^2} = \frac{70.8414}{16.0724} = 4.4076$$

$$\hat{a} = \bar{t} - \hat{b}\overline{\log x} = 8 - \left(4.4076\times\frac{11.2476}{7}\right) = 0.917(87)$$

[0.9178 はあとで使う．これは，$\hat{b}$ を四捨五入せずに有効数字 4 桁で計算したときの $\hat{a}$ の値である．]

従って，$\log x$ を独立変数として用いた回帰直線は t =0.918+4.4076 $\log x$ である．



[注：原点は入れていない．]

(iv) 上述したように，元の散布図は勾配が減少する非線形増加傾向に従っている．回帰直線はプロットの両端のデータを上回っているが中央のデータを下回っている．しかし，$\log x$ を独立変数として使う場合，データの点は非常によい直線傾向を示している．7 点全てが回帰直線のかなり近く，しかもばらばらに上にも下にもあるので，ばらつきは小さくランダムで，異なる $\log x$ の値にわたり適度に一定でもあるようだ．以上から，2 つめのモデルの方が好ましい．

x =6 のとき，1 つめのモデルでは $t = 3.027 + 0.8617 \times 6 = 8.20$ (有効数字 3 桁)と予測される．2 つめのモデルでは $t = 0.918 + 4.4076 \log 6 = 8.815$ と予測される．

中央付近の x の値におけるこの予測では，1 つめのモデルが 2 つめと比べ約 0.6 小さく評価している．この差は，2 つめのモデルで観測されるばらつきをはるかに上回り，2 つめのモデルが優位であることをより強く支持している．

Question 3

(a) サンプルの積率相関係数は

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

で定義される．ここで，$\bar{x},\bar{y}$ はそれぞれ，$x_1,x_2,\ldots,x_n$ の値の標本平均と $y_1,y_2,\ldots,y_n$ の値の標本平均を表し，総和はデータをわたる．

ただし，通常（手）計算に使用される同等な公式は

$$r=\frac{\sum x_iy_i-\frac{\sum x_i\sum y_i}{n}}{\sqrt{\left(\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}\right)}}=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{\left(n\sum x_i^2-\left(\sum x_i\right)^2\right)\left(n\sum y_i^2-\left(\sum y_i\right)^2\right)}}$$

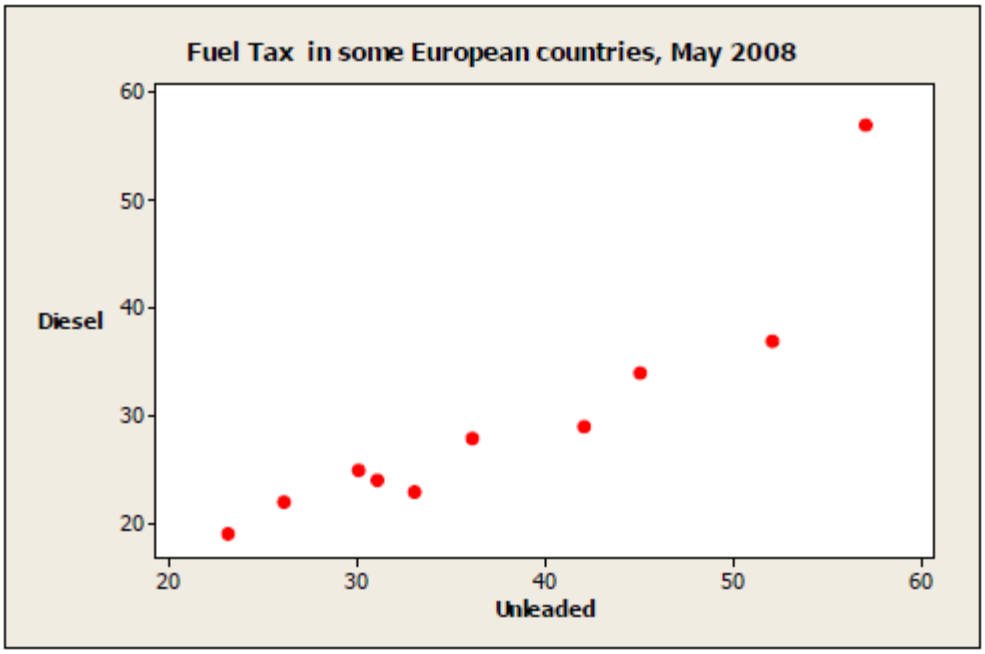
である．

$i=1,2,\ldots,n$ に対して，$x_1,x_2,\ldots,x_n$ を小さい順に並べたときの x_i のランク（順位）を $R(x_i)$ とし（つまり x_i は $R(x_i)$ 番目に小さい），同様に $y_1,y_2,\ldots,y_n$ を小さい順に並べたときの y_i のランクを $R(y_i)$ とする．その後，上の r の式の x_i を $R(x_i)$ で，y_i を $R(y_i)$ で置き換えたものをスピアマンの（標本）順位相関係数 r_s とする．平均値 $\bar{x},\bar{y}$ はともに，ランク $1,2,\ldots,n$ の平均 $(n+1)/2$ になることに注意せよ．

r は標本において，x,y といった 2 つの数量や変数間の相互な線形的な関係性を測っている．2 変数が基本的な線形関係を持っていると考えられる場合には便利である．

r_s は標本において，x,y といった 2 つの数量や変数間の関係性の強さを測っている．2 変数が基本的な単調な（線形かもしれないし線形でないかもしれない）関係を持っていると考えられる場合に用いられる．だが，単調であるが非線形の関係の場合には，r は変数間の関係性の真の強さよりも小さく計算される傾向がある．

(b) (i)



[注：原点は入れていない．]

表に記載された国については非常に明確な単調増加の関係がある．ただし，右端の点（イギリス）が残りの点の傾向（ほぼ線形性）からやや外れていると思われる．データには，外れ値を認めた上で線形傾向があるか，緩い曲線の関係があるか，いずれかだとわかる．

(ii) 便宜上，上記の r の 3 番目の式を使って，

$$r=\frac{10\times12191-375\times298}{\sqrt{\left(10\times15193-375^2\right)\left(10\times9974-298^2\right)}}=\frac{10160}{\sqrt{11305\times10936}}=\frac{10160}{11119}=0.914$$

*試験用の王立統計学会の統計表*から，標本サイズ 10，1%水準での r の臨界値は片側検定で 0.7155 である．0.914>0.7155 だから，基本的な母集団の相関は 0 であるという仮説が棄却され，ディーゼルの燃料税と無鉛ガソリンの燃料税の間に正の相関の証拠があると結論づけた．

(iii)

Unleaded (x)	*R(x)*	*Diesel (y)*	*R(y)*	$d = R(x) - R(y)$
36	6	28	6	0
42	7	29	7	0
23	1	19	1	0
52	9	37	9	0
26	2	22	2	0
30	3	25	5	−2
45	8	34	8	0
33	5	23	3	2
31	4	24	4	0
57	10	57	10	0

式 $r_s = 1 - \dfrac{6\sum d^2}{n\left(n^2-1\right)}$ から r_s を求めると $r_s = 1 - 48/990 = 0.9515$ となる. 表より, 標本サイズ 10, 1%水準での臨界値は片側検定で 0.7455 である． $0.9515 > 0.7455$ だから，基本的な母集団において x と y に関連はないという帰無仮説が棄却され，ディーゼルの燃料税と無鉛ガソリンの燃料税の間に正の関係性（単調な関係）の証拠があると結論づけた.

(iv) (b)(i)の散布図を見ると，x と y の間の関係が単調であるのは明らかだが，線形性についてはいくぶん疑問である． $r_s > r$ に注目したい．2 つの検定の結果が一致しているが，順位検定の方がこれらのデータにより適している．しかし，関連の強さは，より不適切な検定でも有意な結果が得られるほど強い.

主要な注意点は，データがまとめられている国は，定義された母集団からの無作為抽出であると暗黙的に前提としていることだ．ここでの“母集団”はヨーロッパ諸国が自然だろうが，非復元抽出されなければならない．しかし，このような仮定は，しばしば経済データの統計的解析で作られている.

Question 4

(i) 通常の仮定は，残差（誤差）の各項は独立していなければいけないこと，それらの分布は平均が 0，分散が一定であること，通常の推定や検定を行う場合には正規分布に従うとすること．

(ii) (a) 3 つの散布図全てがほぼ直線的に見える．S の各値（B:12.25, A:6, C:6）に沿ってコメントするなら，モデル B の散布図はモデル A や C の散布図よりもばらつきが大きく，この点でモデル A と C の散布図は類似していることがわかる．

(b) モデル B では，x_1 の係数が 0 であるという帰無仮説を検定する検定推定量の値は，帰無分布の t_6 より，$(5.0000-0)/0.9129=5.48$ である．
表より，5%水準両側検定の臨界値は 2.447 である．検定統計量の値はこれを上回っているので，係数が 0 でないという対立仮説がより望ましく，5%水準では帰無仮説は棄却される．それは 1%水準（臨界値は 3.707）でも棄却され，検定統計量の値は実際に，5.959 の 0.1%臨界値よりあまり小さくはない. だからここでは, 帰無仮説が決定的に棄却され，x_1 の係数が 0 でないという強力な証拠がある．

同様にモデル C においては，x_2 の係数が 0 であるという帰無仮説を検定する検定統計量の値は，再び帰無分布の t_6 より $(2.4000-0)/0.2000=12.00$ である．従ってここでは，帰無仮説はかなり決定的に棄却され，x_2 の係数が 0 でないという非常に強い証拠がある．

(c) モデル A では，x_2 の存在下で x_1 の有意性を調べる部分的な t 検定の検定統計量の値は，帰無分布の t_5 より $(1.000-0)/1.000=1.00$ である．5%水準両側検定の臨界値は 2.571 である．そのため，x_2 の存在下では x_1 は省略できると結論づけてよいだろう．
x_1 の存在下で x_2 の有意性を調べる検定も同様である．この検定統計量の値は，再び帰無分布の t_5 より $(2.000-0)/0.4472=4.47$ である．これは 2.571 を優に超えているし，実際には 1%水準での臨界点(4.032)を超えてもいるので，x_1 の存在下でも x_2 はなお必要であると，かなり強く導けるだろう．

x_1 と x_2 への回帰の全体的な有意性の検定，つまり x_1 の係数，x_2 の係数ともに 0 になるという帰無仮説（対立仮説は少なくとも一方が 0 でない）の検定では，F = (回帰の平均平方)/(残差の平均平方)を考えればよい．ここでの F 値は $2610.0/36.0=72.5$，帰無分布は $F_{2,5}$ である．72.5 は通常用いる臨界点（たとえば上側 5%点は 5.79）全てを大幅に上回っているから，帰無仮説は決定的に棄却される．

“$R^2=96.7\%$”とは，y のばらつき全体の 96.7%が y の x_1 と x_2 上への線形重回帰で説明されることを意味する[1 つの解釈として，x_1 と x_2 を組み合わせた最適な線形予測と y との相関の 2 乗が，0.967 になる].

$$R^2 = \frac{(\text{残差平方和})}{(\text{全平方和})} = \frac{5220}{5400}$$

(d) (b)では，モデル B と C それぞれについて，1 つだけの変数（それぞれ x_1, x_2）が有意である強い証拠を述べてきた．(c)では，モデル A が x_1 と x_2 を含んでいるものの，x_2 の存在下では x_1 を取り除けばよいことを示してきた．だがモデル A からは，x_1 の存在下では x_2 も必要だという強い証拠も導けるので，モデル B（x_1 のみ）では不十分だとわかる．モデル C では，x_2 の係数は 0 と有意差があり（更に定数項も同様），より複雑なモデル A と比べても，ほぼ同じくらいよい“説明”（R^2 の値）や等しい誤差平均平方を実現している．したがって，モデル C を選択するべきだ．