RSS Higher Certificate in Statistics, Specimen A

Module 4: Linear Models

Solutions

Question 1

(i) モデルは $y_i = a + bx_i + e_i$ ($i = 1, 2, \ldots, n$) で表される.
ただし $\{e_i\}$ は期待値 0, 一定の分散 σ^2 の無相関な確率変数とする.

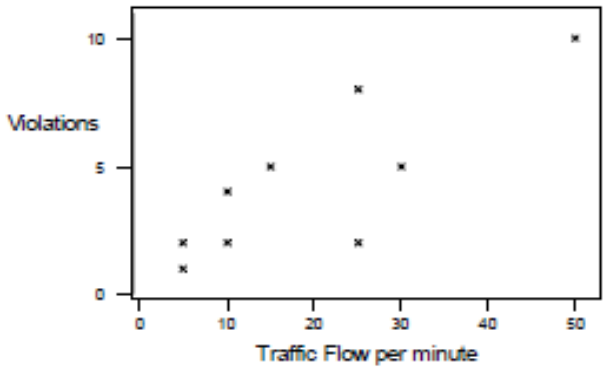
（さらに回帰係数の推測を行うには $\{e_i\}$ の正規性が必要である. これは Question 4 の場合とみなせる.）

(ii) (a) $y_i = \beta x_i + e_i$ に対し, $S = \sum e_i^2 = \sum (y_i - \beta x_i)^2$ が最小となるようにする.

$\frac{dS}{d\beta} = -2\sum x_i (y_i - \beta x_i)$ で，これを 0 とおくと $\sum x_i y_i = \beta \sum x_i^2$ となるので，最小 2 乗推定量は $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ となる.

（最小となることは, $\frac{d^2 S}{d\beta^2} = 2\sum x_i^2 > 0$ によって確認される.）

(b) 以下の散布図を見よ. ほぼ直線的な増加傾向が見られるが, x が増加するほどばらつきが大きいようである. モデルを信頼するのにデータのプロットの数が十分ではない.

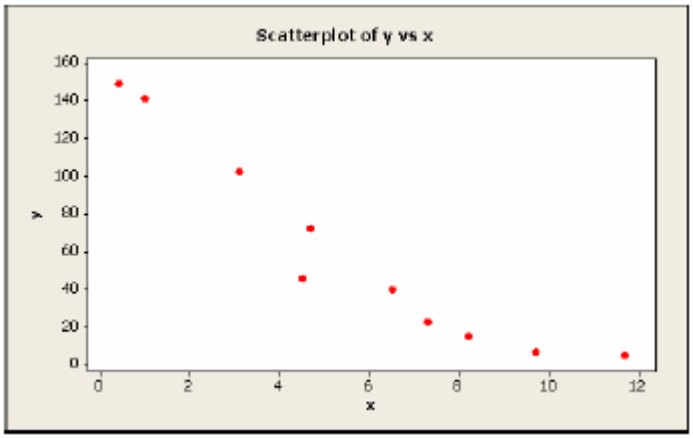


要約統計量は $n = 10, \sum x_i = 180, \sum y_i = 40, \sum x_i^2 = 5150, \sum y_i^2 = 244, \sum x_i y_i = 1055$ であるから $\hat{\beta} = 1055/5150 = 0.205$. よって回帰直線は $y = 0.205\,x$ である.
したがって $x = 20$ に対する違反車数の期待値は $0.205 \times 20 = 4.1$ である.

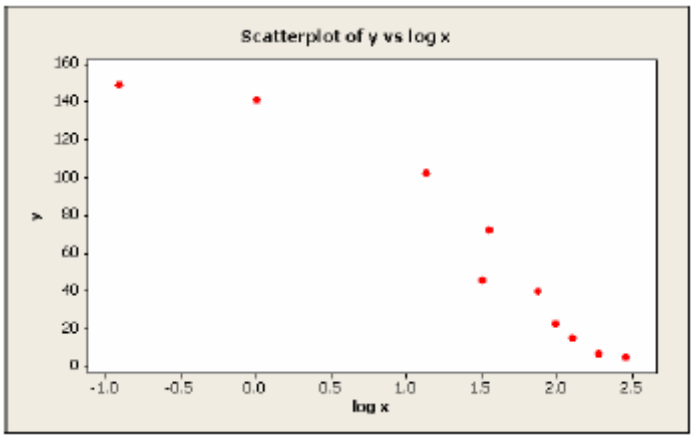
論理的に，交通量がないならば違反車もないので，x=0 のとき y=0 である，つまり 0 切片モデルは理にかなっていると思われる. 散布図はこのことに矛盾しない.

Question 2

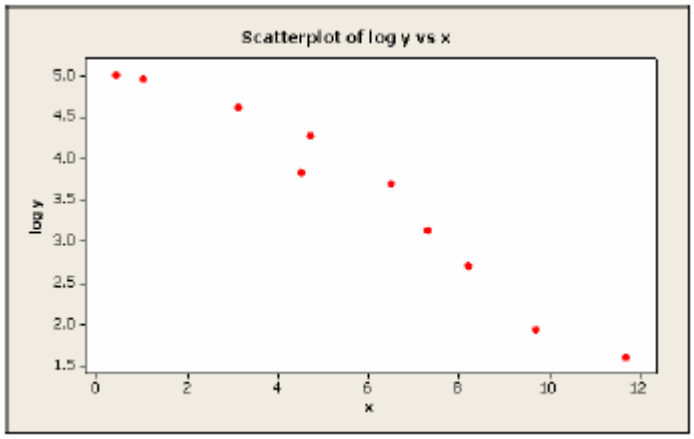
(i) (a)



(b)



(c)



これら全てのプロットは曲線部分を含んでいる.

(a)よりも(c)の方がわずかに“まっすぐ”で，(b)を用いるのは最もよくない.

よって直線で最もよく表されている関係は (c) (x, log y) である.

(ii) (a)から(c)の決定係数 R^2 は全て大きいが，その中でも (c) が最も大きい.

係数に関する t 値もまた (a)から(c) で全て大きいが，その中でも (c) が最も大きい．

(c) は目的変数の平均に関する残差分散が最も小さい．

(c) のみ“大きな”残差がない．

(c) は（わずかの差で）最適なプロットであるとみられる．

$\log x$ に対し $\log y$ を回帰させるのは賢明でないようだ．プロットがより歪むように見える．

(iii) (c)を用いると $\log y = 5.41 - 0.322x$ なので

$$y = \exp\left(5.41 - 0.322x\right) = e^{5.41}e^{-0.322x} = 223.63e^{-0.322x}$$

(iv) (iii)の式より，x=5 を代入すると $y = 223.63e^{-1.61} = 44.7\left(100\text{人}/km^2\right)$ となる．よって推定値は 4470 人／km^2.

(v) データの上端部では曲線傾向がみられるが，それ以外の範囲での予測は適切であるとしてよい．同様の理由により x の変域外の補外法は信用できず，線形モデルだと人口密度が少なく推定されやすい．

隣の市や町の中心部はどこにあるだろうか．かなり遠く離れていない限り，大いに相互作用はあり得るだろう．中心からの方角に応じて人口密度が変化するといった，方向の影響もあるかもしれない．

Question 3

要約統計量：

Dial type	1	2	3	
Total	276	327	294	
Number of tests n_i	8	6	7	Total 21
Mean number of errors $\bar{x}_i$	34.5	54.5	42.0	

(i) 分散分析

Source of variation	df	SS	Mean Square	*F* ratio
Dial type	2	1377	688.50	6.67
Residual (Error)	18	1858	103.22	
Total	20	3234		

F 値 6.67 を $F_{2,18}$ と比較すると，非常に高有意である（上側 1%点は 6.01）．したがって，3 つのタイプの文字盤での誤りの平均個数はみな同じであるという帰無仮説は棄却される．少なくとも 1 つの平均が他の 2 つの平均とかけ離れていると推定される．

全てのデータセットの母集団分布が, 同じ分散σ^2の正規分布であることを仮定している．

(ii) $\bar{x}_2 - \bar{x}_1 = 20.0$ であり，この推定値の標準誤差は $\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}} = \sqrt{103.22\left(\frac{1}{8}+\frac{1}{6}\right)} = 5.487$ である（このように Type 1 と Type 2 の文字盤の誤りの平均の差は非常に高度に有意である．実際，t_{18} と比較すべき検定統計量の値は 20.0/5.487=3.645 である．） t_{18} の両側 5%点は 2.101 なので，真の人口平均差 $\mu_2 - \mu_1$ の 95%信頼区間は

$$20.0 \pm (2.101 \times 5.487) = 20.0 \pm 11.53 = (8.47, 31.53)$$

で与えられる．

繰り返し抽出を行うという観点からの解釈では，「95%信頼区間」とは，実験データの組を用いて上のような方法で計算された各区間のうち 95%が$\mu_2 - \mu_1$の真の値を含む，ということを意味している．

(iii) 21 個の観測値から残差が計算され，正規確率プロットとして，もしくは当てはめ値対残差のプロットをすることによって傾向を調べられる．

3 つの文字盤内の等分散性をチェックできるが，ここにあるような少量のデータに対しては検定の感度は低く，有効でない．

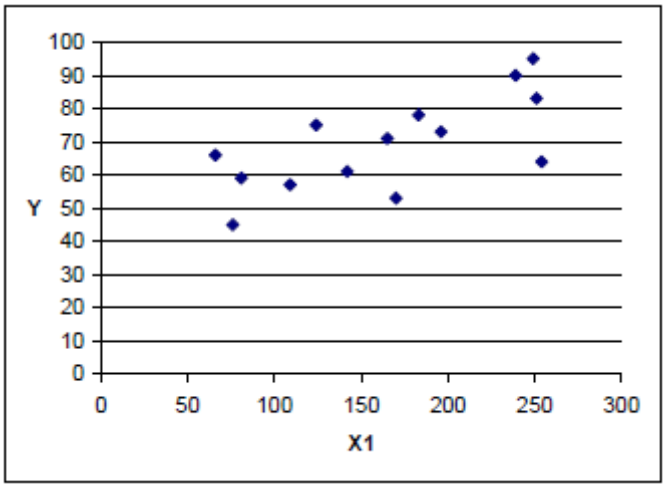
外れ値は背景条件の変化や誤差を記録することでチェックできる．再解析を行う前にあらゆる外れ値はデータから取り除かれる．

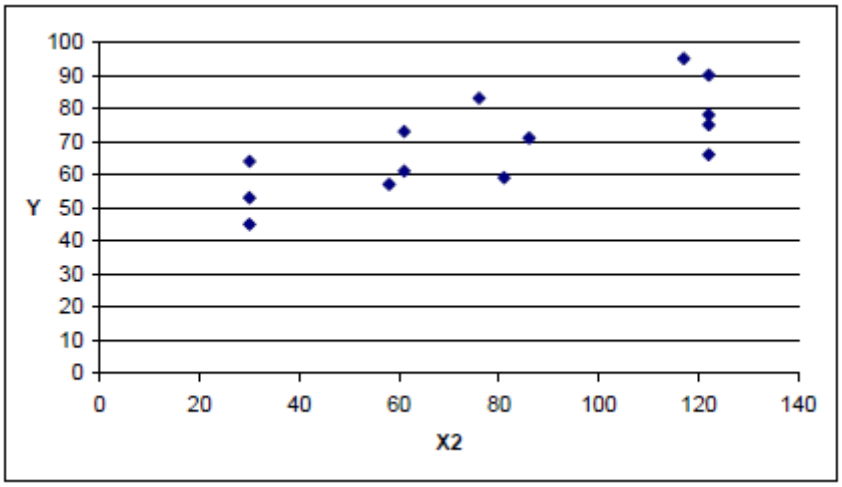
変換（たとえば log ）によってデータは（より等分散性を満たす）正規分布のようにふるかもしれない．

Question 4

(i) モデルは $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, ε_i ~N (0, σ^2) と表される．$\beta_0, \beta_1, \beta_2$ はパラメータ（定数）でデータから推定される．$\{\varepsilon_i\}$ は期待値 0，等分散 σ^2 の独立した正規誤差である．β_0は応答 Y_i の全体の平均を，β_1, β_2 は，もう一方の変数 x を一定に固定したときのそれぞれ x_1，x_2 の単位増加量に対する Y の増加を示している．

(ii) (a)





グラフから，x_1 か x_2 が増えるとおおよそ線形的に Y が増える傾向が見えてくるが，かなりばらつきがある．x_1 か x_2 のどちらか一つの単回帰だとあまりよくなさそうだが，x_1 と x_2 の重回帰を試す価値はある．

(b) コンピュータの出力の「Predictor」の係数を $\beta_0, \beta_1, \beta_2$ とすると，当てはめられた方程式は

$$Y = 24.82 + 0.1380x_1 + 0.2723x_2$$

となる．係数の推定値の標準偏差，また仮説 "$\beta_0 = 0$" などを調べるための t 検定の結果な

ども，このコンピュータの出力で与えられる．ここでの t 値はみな非常に大きい．残差の自由度が 11 なので，t 値は正式には t_{11} と比較して検定される．また，p-値は少なくとも少数第 3 位までは 0 である．これは，帰無仮説 "$\beta_j = 0$" $(j = 0,1,2)$ それぞれにおいて，得られた検定量の値が非常に大きく，またそのような値をとる確率が小さいことを示している．したがって帰無仮説は棄却され，このモデルでは $\beta_0, \beta_1, \beta_2$ 全てを残す必要がある．

分散分析で与えられる残差平均平方は 9.8 であり，これは実験誤差 σ^2 の推定値である．9.8 の平方根は 3.132 でこれも，s の値として出力される．

R^2（95.9%）は，全変動のうちモデルにより説明される部分の割合で，(回帰平方和)/(全平方和)である．（R はしばしば“決定係数”と呼ばれる．）ここでは，この割合（パーセンテージ）は非常に高く，x_1 と x_2 が含まれるモデルが大変良くデータを説明していることを意味している．

分散分析表の F 値 128.2 は $F_{2,11}$ と比較される．これは非常に大きい値で，かなり高有意である（p-値は少なくとも少数第 3 位までは 0）．x_1 と x_2 を共に Y の“予測”に使わなかったならば，このように高有意にはならなかったであろう．こうして，両方を含むモデルがよいモデルだというもうひとつの説明ができる．

定数項の値 24.823 は，おそらく一般の総費用に対する「基本」費用であろう．

(c) x_1 のみが使われるとき，Y の変動全体のたった 1224.0/2622.9=47.43%しか説明されない．信頼性の高い予測のためには x_2 を加えることが重要である．

（注 (1) x_2 のみを使う場合も同様の結果が得られる．つまり x_1 も加える必要がある．）

（注 (2) 積 x_1x_2（すなわち窓ガラスの面積）を予測因子の一つとして用いることができる．（これはあまり良くないということが示される．））