2018年6月 過去問

＜問2＞

コメント

定義に沿って計算を行うだけの問題と思いきや、単純に定義に当てはめるだけだと苦労する問題だった。分散に関する公式をうまく引き出せるか否かがポイントとなる。

(1) 期待値の定義に当てはめれば良い → 解答参照

(2) こちらも期待値の定義に当てはめれば良い。$X_i X_j$ の全ての組み合わせについて考えるのがミソ。 → 解答参照

(2) 分散の定義通りに従って解くと非常に計算が重くなる。↓

$$V[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$$

$$= E[\bar{X}^2] - E[\bar{X}]^2$$

ここで $E[\bar{X}^2]$ を真面目に計算すると、$\bar{X}$ が取りうる全通りについて和を取ることとなり、コストが高くなる。

方針としては、総和の分散は、分散の和の形に展開すると良い。

$$V[\bar{X}] = V\left[\frac{1}{5}\sum X_i\right]$$

$$= \frac{1}{25} V\left[\sum X_i\right]$$

$$= \frac{1}{25}\left(\sum_i V[X_i] + 2\sum_{i<j} Cov[X_i, X_j]\right) \quad (\because *)$$

$V[X_i]$、$Cov[X_i, X_j]$ については解答参照.

(＊)について

$$V[X_1+X_2+\cdots+X_5] = E\{(X_1+\cdots+X_5 - E[X_1+\cdots+X_5])^2\}$$

$$= E\left[\{(X_1-E[X_1])+\cdots+(X_5-E[X_5])\}^2\right]$$

$$= E\left[\sum_{i=1}^{5}(X_i-E[X_i])^2 + 2\sum_{i<j}(X_i-E[X_i])(X_j-E[X_j])\right]$$

$$= \sum_{i=1}^{5} E\left[(X_i-E[X_i])^2\right] + 2\sum_{i<j} E\left[(X_i-E[X_i])(X_j-E[X_j])\right]$$

$$= \sum_{i=1}^{5} V[X_i] + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

<問4>

コメント

二項検定の検定統計量の形を導き出せるかを問う問題。これはぜひ解けておきたい。

(1) 問いの表から、20代女性、20代男性の利用率を出して、該当する図を選べば良い。

(2) 二項検定の検定量を問う問題。どの変数が正規分布に従うかを意識する。

男性の利用数 n_m が $Bin(111, p_m)$ に従い、

女性の利用数 n_F が $Bin(106, p_F)$ に従うとする。

$$\hat{p}_m = \frac{n_m}{111} \sim N\left(p_m, \frac{p_m(1-p_m)}{111}\right)$$

$$\hat{p}_F = \frac{n_F}{106} \sim N\left(p_F, \frac{p_F(1-p_F)}{106}\right)$$

これより、

$$\hat{p}_m - \hat{p}_F \sim N\left(p_m - p_F, \frac{p_m(1-p_m)}{111} + \frac{p_F(1-p_F)}{106}\right)$$

帰無仮説を $p_m = p_F = p$ とすると、

$$\hat{p}_m - \hat{p}_F \sim N\left(0, p(1-p)\left(\frac{1}{111} + \frac{1}{106}\right)\right)$$

これより、

$$z = \frac{\hat{p}_m - \hat{p}_F}{\sqrt{p(1-p)\left(\frac{1}{111} + \frac{1}{106}\right)}} \sim N(0, 1)$$

あとは、$\hat{p}_m = \frac{38}{111}$, $\hat{p}_F = \frac{60}{106}$, $p = \frac{98}{217}$

を代入する。

＜問 6 ＞

コメント

偏差値の定義と、混合正規分布の基礎的な性質を知っていれば解ける。偏差値の定義を覚えておくことがこの問題のポイント

(1) $N(65, 5^2)$ の分布における 64点の偏差値は、

$$50 + 10 \times \underbrace{\frac{64-65}{5}}_{\text{正規化した値}}$$ となる。

Bさんにおいても同様に計算する。

(2) 平均、分散、混合比率を加味しながら選択する。

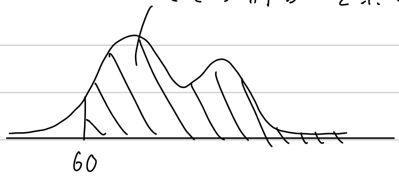
	文		理
平均	60	<	80
分散	5^2	>	3^2
比率	$\frac{200}{300}$	>	$\frac{100}{300}$

→ 2 〃

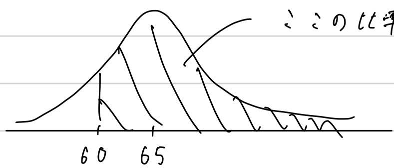
注) 文系の比率が2倍だからといって、ピークの値が2倍にはならない。
なぜなら分散が大きい分、ピークが小さくなるため。

(3)

ここの割合いを求める。



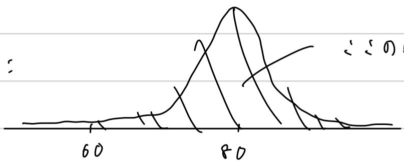
N_1:



ここの比率 × $\frac{200}{300}$

$+$

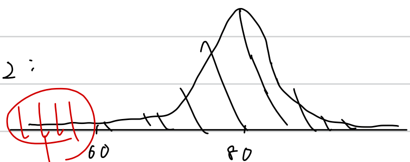
N_2:



ここの比率 × $\frac{100}{300}$

もう少し楽しようとすると、

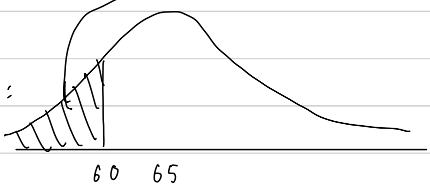
N_2:



この部分はほぼ 0 なのを使って、

$1 - \langle この部分 \rangle \cdot \frac{200}{300}$ でも OK.

N_1:



<問 8>

コメント

まだ未学習分野である時系列解析の問題である。本問題を解くためには、以下の２点を押さえる必要がある。

- 偏自己相関のプロットと、自己相関のプロットの違い
- 不偏性の評価の仕方

(1) $\alpha = 0.5$ のとき、u_{t+1} は u_t に正の相関の影響を受ける。
自己相関係数のプロットは、ラグ h に対して、
u_t と u_{t-h} の相関係数のプロットとなる。
u_t と u_{t-1} の相関は高く、u_t と u_{t-2} の相関もある。
よって ⑤ のようなプロットになる。
一方で、偏自己相関係数のプロットは ② である。
ラグ h の偏自己相関係数は、u_t から $u_{t-1}, \cdots, u_{t-h+1}$ の影響を除いたものと、u_{t-1} から $u_{t-1}, \cdots, u_{t-h+1}$ の影響を除いたものの相関係数である。(ref. ワークブック P247)

例：$h=2$ の偏自己相関係数は、u_t と u_{t-2} の相関から u_{t-1} の影響を除くことになるので、AR(1) モデルで 0 になる。

(2) → 解答を参照

(3) 選択肢の正否を判断するには、次のことを示せば良い。

1. $\overline{y_T}$ は不偏推定量である。
2. $\overline{y_T}$ の分散を T で表示する。

1. については、$u_1, \cdots, u_T$ が定常であることを考慮すると、

$$E(u_t) = \alpha E(u_t) + E(\varepsilon_t)$$

$$\Leftrightarrow (1-\alpha) E(u_t) = 0$$

$$\Leftrightarrow E(u_t) = 0 \quad (\because 0 < \alpha < 1)$$

従って、

$$E(\overline{y_T}) = \frac{1}{T}\sum_{t=1}^{T} E(y_t)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\{E(\mu) + E(u_t)\}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mu$$

$$= \mu$$

よって $\overline{y_T}$ は不偏推定量である。

2. について、

$$V(\overline{y_T}) = V\left(\frac{1}{T}\sum_{t=1}^{T} y_t\right)$$

$$= \frac{1}{T^2}\left\{\sum_{t=1}^{T} V(y_t) + 2\sum_{1\le i<j\le T} Cov(y_i, y_j)\right\} \quad (\because 問2)$$

$$= \frac{1}{T^2}\left\{T \cdot V(u_t) + 2\sum_{1\le i<j\le T} E\left[(y_i - E(y_i))(y_j - E(y_j))\right]\right\}$$

$$= \frac{1}{T^2}\left\{T \cdot V(u_t) + 2\sum_{1\le i<j\le T} E\left[u_i \cdot u_j\right]\right\}$$

$$= \frac{1}{T^2}\left\{ T \cdot V(U_t) + 2\sum_{i=1}^{T}\sum_{j=i+1}^{T} E[U_i \cdot U_j] \right\}$$

ここで、$E(U_t, U_{t+1}) = E[U_t(\alpha U_t + \varepsilon_{t+1})]$

$$= \alpha E(U_t^2) \cdot E(U_t) \cdot E(\varepsilon_{t+1})$$

$$= \alpha V(U_t)$$

$$= \alpha \sigma_u^2$$

また、$E(U_t, U_{t+2}) = E[U_t(\alpha U_{t+1} + \varepsilon_{t+2})]$

$$= \cdots$$

$$= \alpha E[U_t \cdot U_{t+1}]$$

$$= \alpha^2 V(U_t)$$

$$= \alpha^2 \sigma_u^2$$

帰納的に、$E(U_i, U_j) = \alpha^{j-i} \sigma_u^2$

$$V(\bar{y}_T) = \frac{1}{T^2}\left\{ T \cdot V(U_t) + 2\sum_{i=1}^{T}\sum_{j=i+1}^{T} E[U_i \cdot U_j] \right\}$$

$$= \frac{1}{T^2}\left\{ T\sigma_u^2 + 2 \cdot \sum_{i=1}^{T}\sum_{j=i+1}^{T} \sigma_u^2 \alpha^{(j-i)} \right\}$$

（ ここから キビシイ … ）

＜問10＞

コメント

さまざまな正則化の性質を理解しているかを問う問題である。L1正則化は、L2正則化よりも0になるパラメータが多くなる（スパースになる）ことを理解していれば解ける。

(1) パラメータ推定方法ごとの性質は次の通り

方法	性質
OLS	パラメータがめったに0にならない. 絶対値が大きい
OLS+変数減少法	パラメータの多くが0になる. 絶対値が大きい
L_1 正則化	パラメータの多くが0になる. 絶対値が小さい
L_2 正則化	パラメータの多くが0にならない. 絶対値が小さい

＊変数減少法は、「多変量解析」P.71

(2) → 解答参照

<問12>

コメント

未学習の分野である成分分解の内容である（やってないよね？）。とはいえ、コレログラムのプロットの意味さえ分かっていれば、他は図をよく見ることで問題は解くことができる。

(1) コレログラムは、ラグ h に対する自己相関係数のプロットである。
$h=6, 12, 18, \cdots$ に対して正の相関があるはずなので、④か⑤にしぼられる。
$h=6$ よりも $h=12$ の方が相関が大きいので、⑤とわかる。

(2) ⟶ 解答参照

＜論述問２＞

コメント

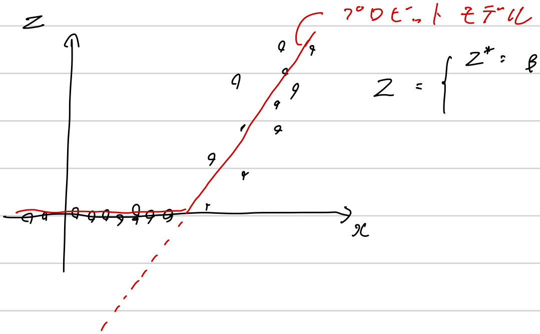
前半はプロビットモデルの推論時の計算ができるかを問う問題で、後半はトービットモデル（まだ学んでない）の尤度計算とAICの定義を問う問題である。

〔1〕(1) 与えられたパラメータを $P(Y=1) = \Phi(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$ に代入して計算すれば良い

(2) これも値を代入して計算するだけである。$\Phi(\cdot)$ の微分をすること以外は大したことをしていない。

→ 解答参照

〔2〕 トービットモデルについては、ワークブック P157 , 158 を参照.

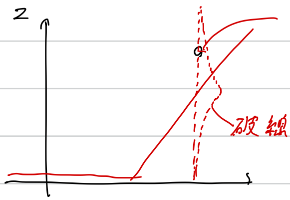


$$Z = \begin{cases} Z^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon & (Z^* \geq 0) \\ 0 & (Z^* < 0) \end{cases}$$

尤度を計算するとき、$Z > 0$ のデータと $Z = 0$ のデータで別の式を使うのがポイント

[$Z > 0$ のデータ]

通常のガウス誤差の尤度を使う。



確率密度が尤度になる。

破線は、ある x における $p(z \mid x)$ のガウス分布

<尤度> $= \mathcal{N}(z_i \mid \beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$

$$= \mathcal{N}(z_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \mid 0, \sigma^2)$$

$$= \mathcal{N}(z_i^* \mid 0, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{z_i^{*2}}{2\sigma^2} \right\}$$

$$= \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_i^*/\sigma)^2}{2} \right\}$$

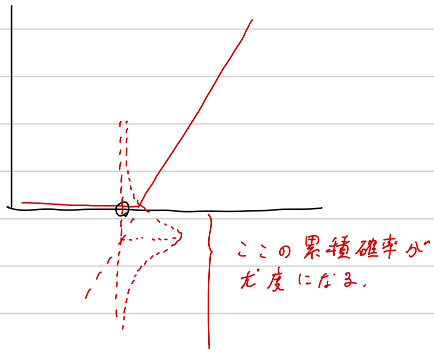
$$= \frac{1}{\sigma} \mathcal{N}(z_i^* \mid 0, 1)$$

$$= \frac{1}{\sigma} \phi\{ z_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \}$$

< $Z > 0$ の全データでの尤度 >

$$= \prod_{i:\, z_i > 0} \frac{1}{\sigma} \phi\{ z_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \}$$

[Z = 0 のデータ]



$$\langle 尤度 \rangle = P\left(z_i < 0 \mid z_i \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)\right)$$

$$= P\left(\frac{z_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{\sigma} < -\frac{\beta_0 + \beta_1 X_1 + \beta_2 X_2}{\sigma} \;\middle|\; \frac{z_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{\sigma} \sim N(0,1)\right)$$

$$= \Phi\left(-\frac{\beta_0 + \beta_1 X_1 + \beta_2 X_2}{\sigma}\right)$$

$\langle Z = 0$ の全データでの尤度 $\rangle$

$$= \prod_{i : z_i = 0} \Phi\left(-\frac{\beta_0 + \beta_1 X_1 + \beta_2 X_2}{\sigma}\right)$$

Z > 0, Z = 0 の全データをかけ合わせた全体の尤度は.

$$L = \prod_{i : z_i > 0} \frac{1}{\sigma} \phi\left\{z_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2)\right\}$$

$$\cdot \prod_{i : z_i = 0} \Phi\left(-\frac{\beta_0 + \beta_1 X_1 + \beta_2 X_2}{\sigma}\right)$$

[2] AIC の定義は次式である．

$$AIC = -2 \langle 対数尤度 \rangle + 2 \langle 自由パラメータ数 \rangle$$

(PRML では，
$AIC = \ln p(D \mid w_{ML}) - M$ であったので注意)

「日照時間」+「平均気温」モデルでのパラメータ数は，
β_0 ，β_1 ，β_2 ，σ^2 の4つである点に注意．