課題 1.1

問題

乗算法を用いて (0,1] の一様乱数列を生成するプログラムを作成せよ．このプログラムから生成される数列が一様乱数であることを (0,1] 区間を 100 等分してそれぞれの区間に値が入る回数がほぼ同じになることにより示せ．

方法

乗算法 $I_{j+1} = aI_j (\mathrm{mod}\ m)$ により乱数列 $\{I_j\}$ を生成し, $m-1$ で割ることで $(0,1]$ の一様乱数列とする．また, 各区間に含まれる乱数の個数を数え, 結果をヒストグラムで図示し, 考察する．なお, 乗算法におけるパラメータは $a = 69621, m = 2^{32}$ とする．

結果

prob1_1.c を用いて,初期値を $a_0 = 2312$ として乱数を 1000 万個生成したときの各区間に含まれる乱数の個数についてのヒストグラムは図 1 となる．

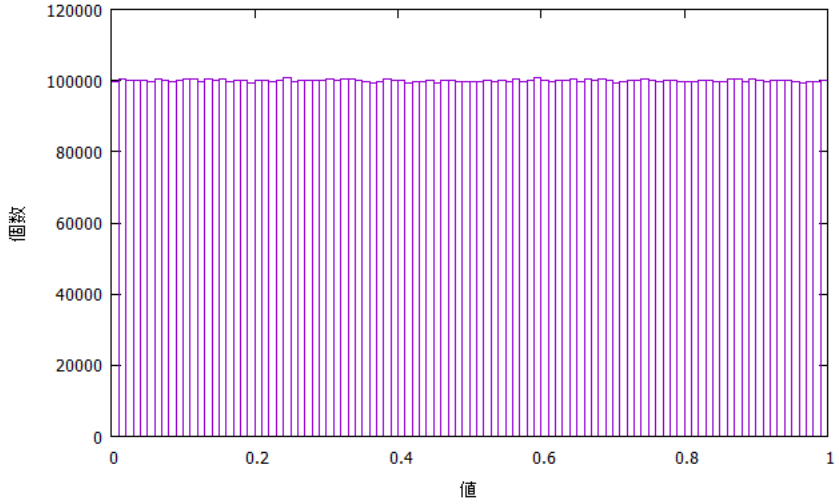

図 1　各区間に含まれる乱数の個数のヒストグラム

考察

図 1 より各区間に値が入る回数はほぼ同じであり, プログラムにより生成される数列が一様乱数であることがわかる．

課題 1.2

問題

Box-Muller 法を用いて平均 μ, 分散σ^2 に従う正規乱数を生成するプログラムを作成せよ．更に平均

1($\mu = 1$), 分散 4($\sigma = 2$) の場合について乱数を n 個生成し, 標本平均値と標本分散を n の関数として図示せよ.

方法

乗算法により 2 つの独立な一様乱数列 $\{u_{1j}\}, \{u_{2j}\}$ を生成し,Box-Muller 法

$$n_{1j} = \sqrt{-2\ln(u_{1j})}\cos(2\pi u_{2j}) \tag{1}$$

$$n_{2j} = \sqrt{-2\ln(u_{1j})}\cos(2\pi u_{2j}) \tag{2}$$

により, 標準正規乱数列 $\{n_{1j}\}, \{n_{2j}\}$ を生成する. $N_{1j} = \sigma n_{1j} + \mu$と変数変換を行うと平均$\mu$, 分散$\sigma^2$の正規乱数列 $\{N_{1j}\}$ が生成される. なお, 一様乱数の生成方法, パラメータは課題 1.1 と同様であり.奇数番目を $\{u_{1j}\}$, 偶数番目を $\{u_{2j}\}$ とする. 乗算法の初期値は 2312 である.

プログラムにより平均 1, 分散 4 の n 個の正規乱数を生成し, 標本平均値, 標準分散値を求め, 乱数の個数 n と標本平均値または標本分散値との関係を図示する.

結果

prob1_2.c を用いて, 標本平均値, 標本分散値を求める. 乱数の個数 n と標本平均値または標本分散値との関係は図 2 となる.

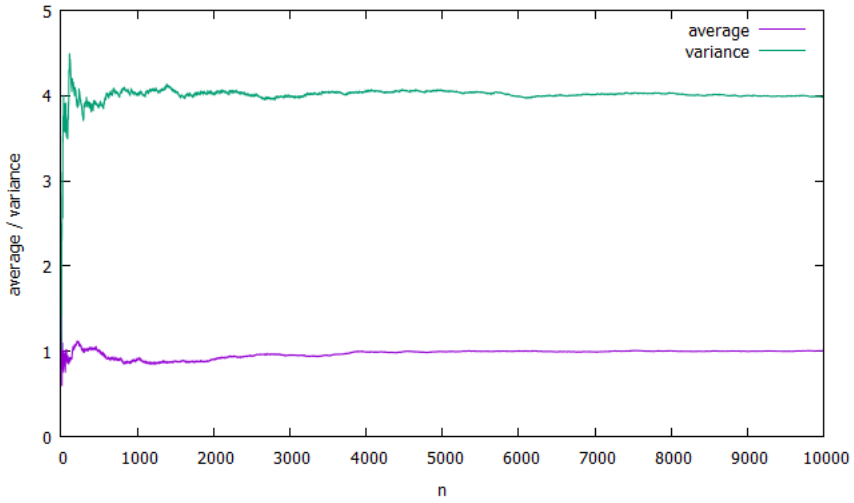

図 2　乱数の個数 n と標本平均値(average), 標本分散値 (variance) の関係

考察

図 2 より乱数の個数 n が増加すると標本平均値は$\mu = 1$, 標本分散値は$\sigma^2 = 4$ に収束することがわかる.

課題 1.3

問題

$a > 1, b > 1$ について不完全ベータ関数を数値積分により計算するプログラムを利用して自由度 $\nu = 5, 10, 25, 60, 120$ について t 分布の累積分布関数を図示せよ. プログラムが正しくできている事を t 分布表を用いて確認せよ.

方法

正則化済み不完全ベータ関数$\beta(x; a, b)$ の分母, 分子をそれぞれシンプソン則を用いて数値積分し, 最後に分数に戻すプログラムを組む. このプログラムにより, 自由度νの t 分布の累積分布関数 $\Pr[T \leq t] = \beta(\frac{t+\sqrt{t^2+\nu}}{2\sqrt{t^2+\nu}}; \frac{\nu}{2}, \frac{\nu}{2})$ を計算する. 各自由度νにおいて, t を -4 から 4 まで 0.1 ずつ値を増やしながら累積分布関数$\Pr[T \leq t]$ を計算し, グラフとする.
また, プログラムにより各自由度νにおいて, t を 0.001 刻みで増加させたときに p 値 $(= 1-\Pr[T \leq t])$ が $p = 0.1, 0.01, 0.001$ を初めて下回る t の値を調べる. これにより t 分布表を作成し, 実際の t 分布表と比較することで, プログラムが正しいことを確認する.

結果

各自由度νにおける t 分布の累積分布関数は図 3 となる.

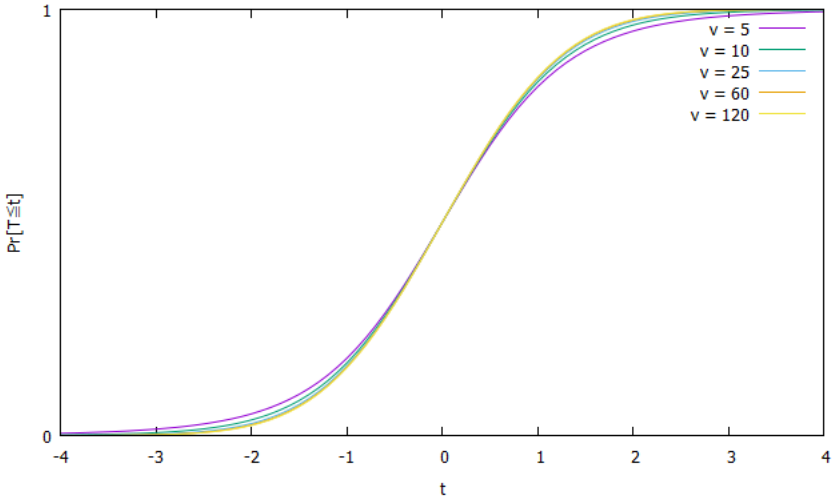

図 3　各自由度νにおける t 分布の累積分布関数

また,prob1_3.c を用いて作成した t 分布表は表 1, 実際の t 分布表は表 2 である.

表 1　プログラムにより作成した t 分布表

$v \setminus p$	0.1	0.01	0.001
5	1.476	3.366	5.894
10	1.373	2.764	4.144
25	1.317	2.486	3.451
60	1.296	2.391	3.232
120	1.289	2.358	3.160

表 2　実際の t 分布表

$v \setminus p$	0.1	0.01	0.001
5	1.476	3.365	5.893
10	1.372	2.764	4.144
25	1.316	2.485	3.450
60	1.296	2.390	3.232
120	1.289	2.358	3.160

考察

プログラムにより作られた t 分布表と実際の t 分布表がほぼ一致していることから, プログラムが正しいことが確認できる. 小数第 3 位の違いは, 丸め誤差の影響や, 実際の t 分布表で省略されている小数第 4 位以降の影響であると思われる.

課題 1.4

問題

まずμ_0とσを好きな値に設定して,Box-Muller法を用いて平均μ_0と分散σ^2に従う 100 個の乱数を作りなさい. そして, これらの乱数列が, 最初に決めたμ_0の平均値を持つことを t 検定によって検定しなさい. この時の, t 値と p 値はいくらであるかを示せ.

方法

$\mu_0 = 4, \sigma = 3$ として,Box-Muller法を用いて平均 4, 分散 9 の正規乱数を $100(= n)$ 個生成する. なお, 正規乱数の生成方法やパラメータは課題 1.2 と同様であり, 乗算法の初期値は 2312 である. これらの乱数列の標本平均値μがμ_0と大きなズレがあるかを検定する. 帰無仮説 H_0, 対立仮説 H_1は

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

として, 両側検定を行う. なお, 有意水準は$\alpha = 0.05$ とする.

乱数列より標本平均値$\bar{x}$, 標本分散値$\hat{\sigma}^2$を求め, t 値, p 値を

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \tag{3}$$
$$p = \Pr[T \leq -|t|] + \Pr[T \geq |t|] \tag{4}$$
$$= 2\left(1 - \beta\left(\frac{|t| + \sqrt{t^2 + n - 1}}{2\sqrt{t^2 + n - 1}}; \frac{n-1}{2}, \frac{n-1}{2}\right)\right) \tag{5}$$

により, 計算する. 両側検定なので p 値は t 分布の両すそ部分を含める. p 値が有意水準以下ならば帰無仮説は棄却され, そうでなければ棄却されない.

結果

prob1_4.c を用いて 標本平均値$\bar{x}$, 標本分散値$\hat{\sigma}^2$, t 値, p 値を求め, t 検定を行う. プログラムにより求め

た標本平均値$\bar{x}$, 標本分散値$\hat{\sigma}^2$, t 値, p 値は表 3 となる.

表 3　各統計量の値

$\bar{x}$	$\hat{\sigma}^2$	t	p
13.787338	8.748336	-0.718997	0.473865

$0.05 < p$ なので帰無仮説 H_0は棄却されない.

考察

帰無仮説 H_0が棄却されなかったので, 帰無仮説が正しくないとは断定できない. つまり, 今回の検定からは.標本平均値と真の平均値に大きなズレがあるかどうかは何も言えない.

課題 2.1

問題

(6) 式をパラメータαとβによって偏微分し$\frac{\partial E}{\partial \beta} = 0, \frac{\partial E}{\partial \alpha} = 0$ と置くことにより, (7) 式と (8) 式を求めよ.

$$E = \sum_{j=1}^{n} (y_j - (\alpha + \beta x_j))^2 \tag{6}$$

$$\hat{\beta} = \frac{\sum_{j=1}^{n} (y_j - \bar{y})(x_j - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \tag{7}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{8}$$

解答

$$\frac{\partial E}{\partial \beta} = -2 \sum_{j=1}^{n} x_j (y_j - (\alpha + \beta x_j)) \tag{9}$$

$$\frac{\partial E}{\partial \alpha} = -2 \sum_{j=1}^{n} (y_j - (\alpha + \beta x_j)) \tag{10}$$

$\frac{\partial E}{\partial \beta} = 0, \frac{\partial E}{\partial \alpha} = 0$ となるα, βの値を$\hat{\alpha}, \hat{\beta}$とおくと

$$(\sum_{j=1}^{n} x_j y_j) - \hat{\alpha}(\sum_{j=1}^{n} x_j) - \hat{\beta}(\sum_{j=1}^{n} x_j^2) = 0 \tag{11}$$

$$(\sum_{j=1}^{n} y_j) - n\hat{\alpha} - \hat{\beta}(\sum_{j=1}^{n} x_j) = 0 \tag{12}$$

$\hat{\beta}$について解くと

$$\hat{\beta} = \frac{n(\sum_{j=1}^{n} x_j y_j) - (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j)}{n(\sum_{j=1}^{n} x_j^2) - (\sum_{j=1}^{n} x_j)^2} \tag{13}$$

$\bar{x}, \bar{y}$を標本平均値として, 分子を変形すると

$$
\begin{aligned}
& n\ (\sum_{j=1}^{n} x_j y_j) - (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j) \\
&= n(\sum_{j=1}^{n} x_j y_j) - (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j) - (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j) + (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j) && (14) \\
&= n(\sum_{j=1}^{n} x_j y_j) - n\bar{x}(\sum_{j=1}^{n} y_j) - n\bar{y}(\sum_{j=1}^{n} x_j) + n^2 \bar{x}\bar{y} && (15) \\
&= n\sum_{j=1}^{n}(x_j y_j - \bar{x} y_j - x_j \bar{y} + \bar{x}\bar{y}) && (16) \\
&= n\sum_{j=1}^{n}(y_j - \bar{y})(x_j - \bar{x}) && (17)
\end{aligned}
$$

また, 分母を変形すると

$$
\begin{aligned}
n(\sum_{j=1}^{n} x_j^2) - (\sum_{j=1}^{n} x_j)^2 &= n(\sum_{j=1}^{n} x_j^2) - 2(\sum_{j=1}^{n} x_j)^2 + (\sum_{j=1}^{n} x_j)^2 && (18) \\
&= n(\sum_{j=1}^{n} x_j^2) - 2n\bar{x}(\sum_{j=1}^{n} x_j) + n^2 \bar{x}^2 && (19) \\
&= n(\sum_{j=1}^{n} x_j^2 - 2\bar{x} x_j + \bar{x}^2) && (20) \\
&= n\sum_{j=1}^{n}(x_j - \bar{x})^2 && (21)
\end{aligned}
$$

したがって,

$$
\hat{\beta} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{22}
$$

$\hat{\alpha}$は式 (12) より,

$$
\begin{aligned}
\hat{\alpha} &= \frac{1}{n}(\sum_{j=1}^{n} y_j) - \hat{\beta}\frac{1}{n}(\sum_{j=1}^{n} x_j) && (23) \\
&= \bar{y} - \hat{\beta}\bar{x} && (24)
\end{aligned}
$$

課題 2.2

問題

データ $(X_i, Y_i)(i = 1, \ldots, n)$ が与えられたとき, 単回帰 $Y_i = \beta X_i + \alpha$の回帰係数αとβを求めるプログラムを作成し, 回帰係数βについて t 検定を行うための p 値を求めよ. このプログラムが正しい結果を導いてることを, 乱数を用いたデータ生成モデルを用いて確認せよ. なぜ作成したプログラムが正しい答えを与えていると言えるか根拠を示せ.

方法

乗算法により独立な一様乱数列 $\{x_i\}, \{y_i\}$ をそれぞれ 100 個生成する. 一様乱数の生成方法, パラメータは課題 1.2 と同様であり, 乗算法の初期値は 231 である データ (x_i, y_i) について, 課題 2.1 の式 (7), 式 (8) により, 回帰係数$\hat{\alpha}, \hat{\beta}$を求める.

$\hat{\beta}$について t 検定を行うが, $\hat{\beta}$の符号により帰無仮説を変更する. $\hat{\beta}$が正の場合, 帰無仮説 H_0, 対立仮説 H_1は

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$

として, 右側検定を行う. なお, 有意水準は$\alpha = 0.05$ とする.

推定誤差の数列 $\{u_j\}$ を $u_j = y_j - (\hat{\alpha} + \hat{\beta} x_j)$ により定めると, 推定誤差の不偏分散値$\hat{\sigma}^2$は

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^{n} (y_j - (\hat{\alpha} + \hat{\beta} x_j))^2 \tag{25}$$

により求める. $\hat{\beta}$は平均$\beta(= 0)$, 分散$\hat{\sigma}^2 / \sum_{j=1}^{n} (x_j - \bar{x})^2$の正規分布に従うので, t 値, p 値を

$$t = \frac{\hat{\beta}}{\hat{\sigma} / \sqrt{\sum_{j=1}^{n} (x_j - \bar{x})^2}} \tag{26}$$
$$p = \Pr[T > t] \tag{27}$$
$$= 1 - \beta \left(\frac{t + \sqrt{t^2 + n - 2}}{2\sqrt{t^2 + n - 2}}; \frac{n-2}{2}, \frac{n-2}{2} \right) \tag{28}$$

により, 計算する. p 値が有意水準以下ならば帰無仮説は棄却され, そうでなければ棄却されない.

$\hat{\beta}$が負の場合, 帰無仮説 H_0, 対立仮説 H_1は

$$H_0 : \beta \geq 0$$
$$H_1 : \beta < 0$$

として, 左側検定を行う.なお, 有意水準は同様の値である. 推定誤差の不変分散量$\hat{\sigma}^2$は式 (25), t 値は式 (26) と同様に計算するが, p 値は左側検定であるので,

$$p = \Pr[T < t] \tag{29}$$
$$= \beta \left(\frac{t + \sqrt{t^2 + n - 2}}{2\sqrt{t^2 + n - 2}}; \frac{n-2}{2}, \frac{n-2}{2} \right) \tag{30}$$

により, 計算する. p 値が有意水準以下ならば帰無仮説は棄却され, そうでなければ棄却されない.

更に, 新たな乱数列 $\{z_i\}$ を $z_i = x_i + y_i$により定め, データ (x_i, z_i) について同様に回帰係数$\hat{\alpha}, \hat{\beta}$を求め, $\hat{\beta}$について同様の検定を行う.有意水準は同様の値とする.

データ (x_i, y_i) はそれぞれ独立な一様乱数であることから, $\beta = 0$ であり, t 検定では帰無仮説 H_0が棄却されないと予想される. また, データ (x_i, z_i) は z_iの定義より, $\beta = 1$ であり, t 検定では帰無仮説 H_0が棄却されると予想される.

結果

データ (x_i, y_i) において,prob2_2.cを用いて計算した回帰係数の値は$\hat{\alpha} = 0.488405, \hat{\beta} = 0.027066$ であり,回帰直線は $y = 0.027066x + 0.488405$ となる. データ (x_i, y_i) と回帰直線をプロットしたものが図 4 である.

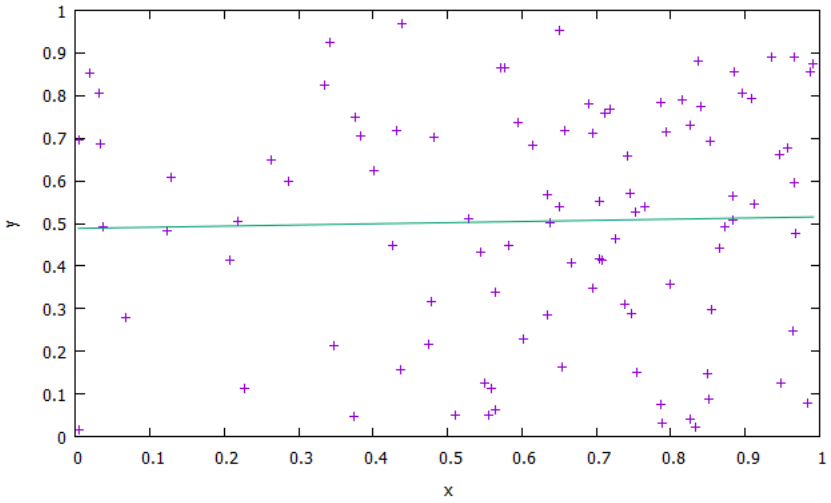

図 4　データ (x_i, y_i) と回帰直線

$\hat{\beta} > 0$ であるので,帰無仮説 H_0,対立仮説 H_1は

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$

として,右側検定を行う. prob2_2.cにより求めた不変分散値$\hat{\sigma}^2$, t 値, p 値は表 4 となる.

表 4　各統計量の値

$\hat{\sigma}^2$	t	p
0.075254	0.262641	0.396705

$0.05 < p$ なので帰無仮説 H_0は棄却されない.

また,データ (x_i, z_i) において,prob2_2.cにより計算された回帰係数の値は$\hat{\alpha} = 0.488405, \hat{\beta} = 1.027066$ であり,回帰直線は $z = 1.027066x + 0.488405$ となる. データ (x_i, z_i) と回帰直線をプロットしたものが図 5 である.

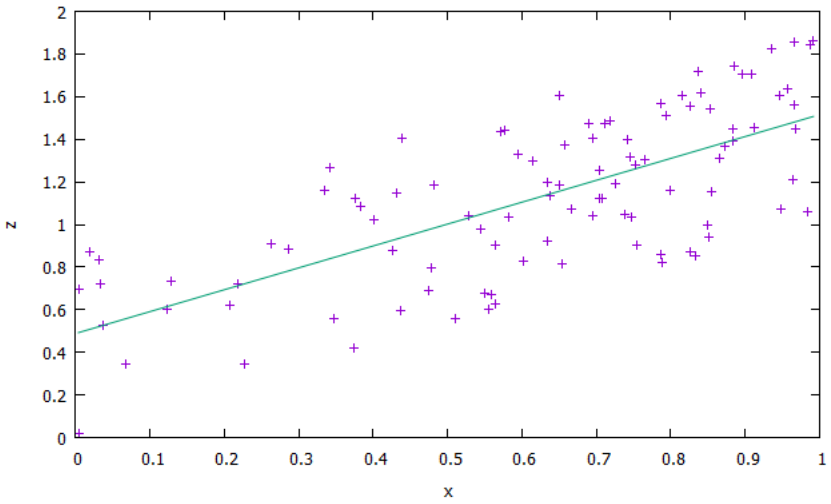

図 5　データ (x_i, z_i) と回帰直線

$\hat{\beta} > 0$ であるので, 帰無仮説 H_0, 対立仮説 H_1 は

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$

として, 右側検定を行う. prob2_2.cにより求めた不変分散値$\hat{\sigma}^2$, t 値, p 値は表 5 となる.

表 5　各統計量の値

$\hat{\sigma}^2$	t	p
0.075254	9.966305	0.000000

$p \leq 0.05$ なので帰無仮説 H_0 は棄却される.

考察

「方法」で述べた予想と同じ結果が得られたので, 作成したプログラムは正しい答えを与えていると言える.

課題 2.3

問題

世界銀行データバンク (http://databank.worldbank.org/data/home.aspx) にてダウンロードできるデータを用いて , A 年 (A は任意の年) における国ごとの, 国民一人当たりの GDP(current USD per capita)X_iと国民ひとり当りのエネルギー消費量(kg od oil equicalent per capita)Y_i との間にべき乗関係 $\log_{10} Y_i = \beta \log_{10} X_i + \alpha$が認められるか調べなさい (単回帰プログラムを用いてこのべき指数βを決定し, このべき指数βに対する t 検定が, 帰無仮説を 5% 有意水準で棄却できるか確認する).

方法

2002 年のデータを用いて X_iと Y_iの間のべき乗関係の回帰係数を調べる. Energy_GDP.csvよりデータ (X_i, Y_i) を読み取り, 2 つの配列 $\{x_i\}, \{y_i\}$ を, $x_i = \log_{10} X_i, y_i = \log_{10} Y_i$により定める. このデータ (x_i, y_i) について, 課題 2.1 の式 (7), 式 (8) により回帰係数$\hat{\alpha}, \hat{\beta}$を求め, 課題 2.2 と同様に$\hat{\beta}$について t 検定を行う.なお, 有意水準は 0.05 とする.

結果

データ (x_i, y_i) において,prob2_3.cにより計算された回帰係数の値は$\hat{\alpha} = 1.109393, \hat{\beta} = 0.588847$ であり, 回帰直線は $y = 0.588847x + 1.109393$ となる. データ (x_i, y_i), すなわち $(\log_{10} X_i, \log_{10} Y_i)$ と回帰直線をプロットしたものが図 6 である.

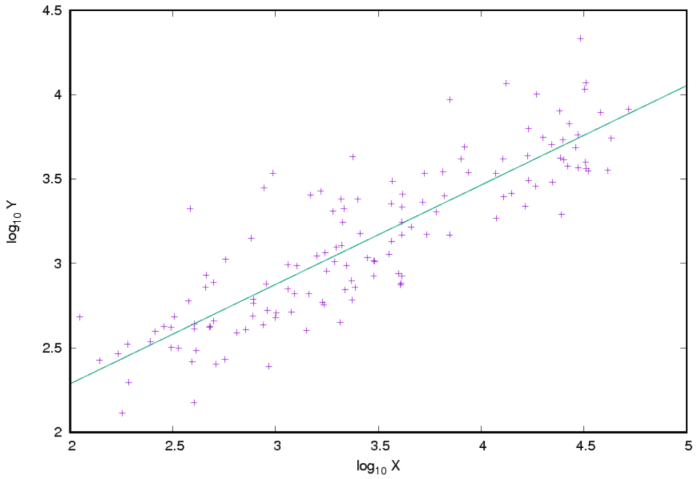

図 6　データ $(\log_{10} X_i, \log_{10} Y_i)$ と回帰直線

$\hat{\beta} > 0$ であるので, 帰無仮説 H_0, 対立仮説 H_1は

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$

として, 右側検定を行う. prob2_3.cにより求めた不変分散値$\hat{\sigma}^2, t$ 値, p 値は表 6 となる.

表 6　各統計量の値

$\hat{\sigma}^2$	t	p
0.057734	19.583787	0.000000

$p \leq 0.05$ なので帰無仮説 H_0は棄却される.

考察

帰無仮説 H_0が棄却されたので, $\hat{\beta} > 0$ が正しいと主張できる. つまり, 国民一人当たりの GDP と国民ひとり当りのエネルギー消費量の間に正のべき乗関係が認められる.