統計基礎

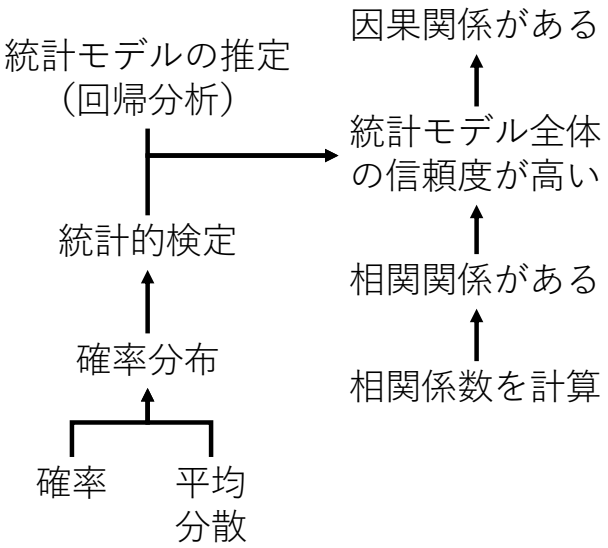
古谷知之

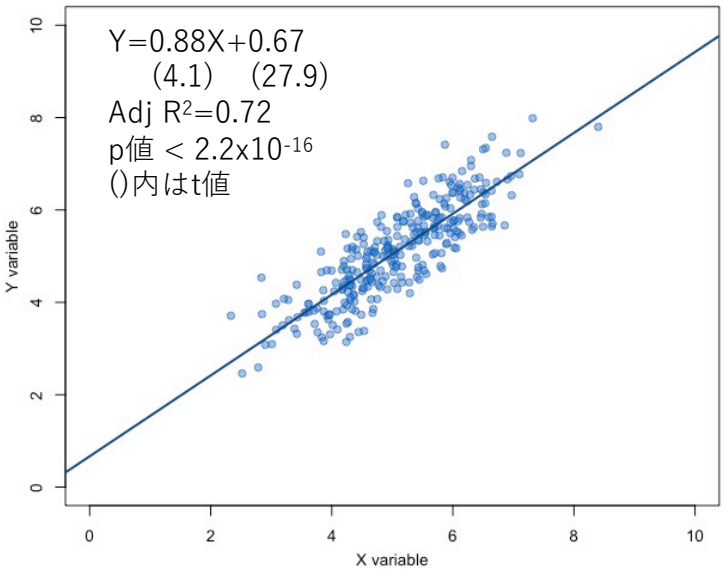
授業概要

＊履修者の状況に応じて変更される場合がありますが、概ね以下のような流れで授業を進めます。

第1回　ガイダンス・確率
第2回　確率変数と確率分布(1)
第3回　確率変数と確率分布(2)
第4回　母集団と平均
第5回　単回帰分析(1)
第6回　単回帰分析(2)
第7回　仮説検定(1)
第8回　仮説検定(2)
第9回　重回帰分析
第10回　R演習(1)
第11回　R演習(2)
第12回　R演習(3)
第13回　R演習(4)
第14回　最終試験

授業の全体像





授業内容

- 単回帰分析の実装
- 残差の性質
- 残差解析と外れ値の検出

- 重回帰分析の実装
- データの標準化と標準化偏回帰係数
- 多重共線性

単回帰分析で検討すべきこと

- 回帰係数の値：偏回帰係数
- 回帰係数の統計的有意性：t検定
- 回帰係数の信頼度：信頼区間
- 予測への適用可能性：決定係数
- 外れ値の検出：残差解析

回帰分析の主な手順

- 因果関係を証明したいテーマを設定する
- 因果関係を示すのに適した従属変数と独立変数を用意する
- 回帰係数、回帰係数のt値とp値、決定係数、自由度修正済み決定係数などの推定量・統計量を計算する
- 回帰係数のt検定を行い、採用する従属変数を選定する
- 推定結果が事前に想定した仮説とあわない（回帰係数が統計的に有意でない、決定係数の値が小さい、など）場合には、他の変数を用意するなどして分析をやり直す
- 決定係数が大きく、回帰係数がいずれも統計的に有意な結果が得られるた場合には、最終的に推定結果を解釈する

分析テーマの例

- 経済理論：消費は所得に依存する
- 「所得が増えれば消費が増える」という単回帰モデルを構築
- 独立変数（説明変数）：所得x_i
- 従属変数（被説明変数）：消費y_i
- 分析単位：年収階級i
- モデル式

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 期待される分析結果
  - $\beta_1 > 0$かつβ_1が5%水準で統計的に有意（t値が1.96以上）
  - 自由度修正済みR^2が1に近い

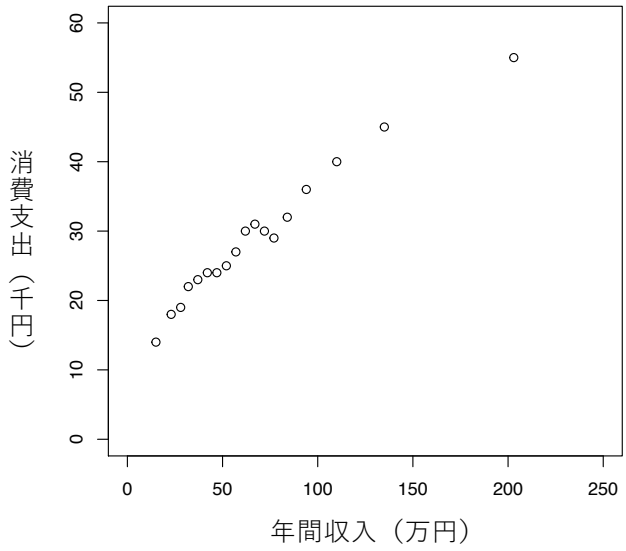
分析に用いるデータ

- 『家計調査』2017年9月、第２－６表「年間収入階級別１世帯当たり１か月間の収入と支出」
- データは政府統計ポータルサイトe-statから入手可能
- 「二人以上の世帯」
- 所得x_i＝「年間収入(10万円)」、支出y_i＝「消費支出(千円)」を用いる
- 「年間年収階級」の18階級を分析単位とする

散布図を描いてみる

- 散布図を見る限り、相関関係はありそう
- 年間収入が増加するに従い、消費支出も増加するように見える

- 理論に基づかなくても因果関係がある場合があるので、データを見ることが大事



単回帰分析の分析結果

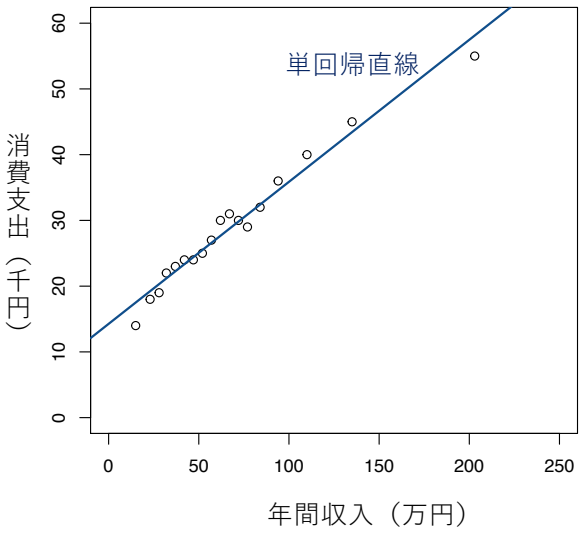
- 単回帰分析の結果は以下のようになった ()内はt値

$$\hat{y} = 14.3 + 0.216x$$

(18.56)　(23.02)

自由度修正済み$R^2 = 0.97$

- $\hat{\beta}_1 = 0.22 > 0$かつ$\hat{\beta}_1$のt値=23.02と5%水準で統計的に有意
- 自由度修正済み$R^2 = 0.97$と1に近い



回帰係数の信頼区間

- 回帰分析の独立変数の数がk個（単回帰分析のとき$k=1$）、標本数がn個のとき独立変数xの偏回帰係数$\hat{\beta}$について、$\frac{\beta-\hat{\beta}}{SE}$は自由度$\mathrm{n}-k-1$の$t$分布に従う
- ここでSEは標準誤差(standard error)であり、不偏分散s_x^2（標本分散σ_x^2）を用いて計算される

$$\mathrm{SE}=\sqrt{s_x^2/n}$$
$$s_x^2=\frac{n-1}{n}\sigma_x^2$$

回帰係数の信頼区間

- 偏回帰係数$\hat{\beta}$の95%信頼区間は以下のようになる

$$\left[\hat{\beta}-t_{n-k-1,0.025}\sqrt{s_x^2/n}\,,\hat{\beta}+t_{n-k-1,0.025}\sqrt{s_x^2/n}\right]$$

または

$$\left[\hat{\beta}-t_{n-k-1,0.025}\,SE,\hat{\beta}+t_{n-k-1,0.025}\,SE\right]$$

- 単回帰分析のとき、自由度$n-2$のt分布に従う
- 標本数が多く有意水準5%のとき、$\hat{\beta}\pm 1.96SE$が目安となる

回帰係数の信頼区間

- 前述の推定結果では、年間収入に対する偏回帰係数の標準誤差$SE = 0.0094$が得られたことから、信頼区間は

$$[0.216 - 1.96\times0.0094, 0.216 + 1.96\times0.0094]$$

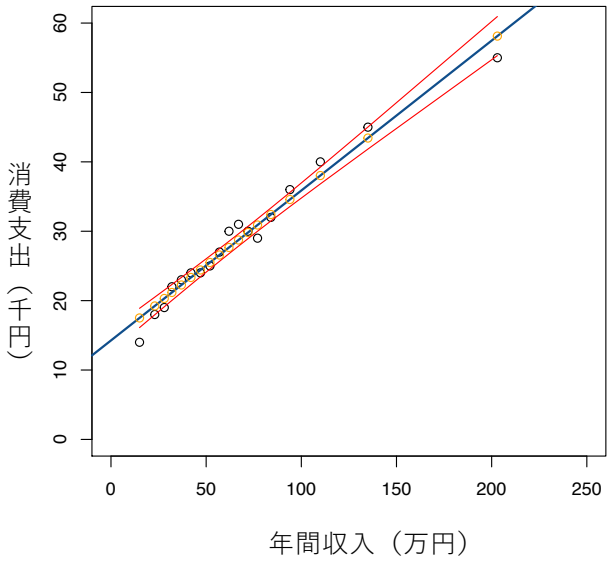
すなわち

$$[0.198, 0.234]$$

となる

予測値と残差

- 単回帰式（$\hat{y} = 14.3 + 0.216x$）に年間収入を代入すると消費支出の予測値$\hat{y}$が得られる
- 消費支出の予測値$\hat{y}$と実績値yとの差が残差（予測誤差）ε
- 予測値$\hat{y}$をプロットすると左図のようになる（オレンジ色）
- 95%信頼区間は赤色の曲線で挟まれた区間



単回帰分析の結果のまとめ方（例）

- 単回帰分析の結果は、以下のように整理すると分かりやすい
- 回帰係数、t値、自由度修正済みR^2、標本数の記述は必須
- 標準誤差と95%信頼区間を記述するとより親切

説明変数	回帰係数	t値	標準誤差	95%信頼区間
切片	14.26	18.56	0.769	[12.75, 15.77]
年間収入(万円)	0.216	23.02	0.00938	[0.198, 0.234]
自由度修正済みR^2 0.97				
標本数18				

残差の性質

- 推定された回帰モデルの残差$e_i = y_i - \hat{y}_i$（実績値と予測値との差分）は、回帰分析で説明がつかない部分を意味する
- 残差e_iは互いに独立に（各i毎に）平均0、分散σ^2の正規分布に従う。$e_i \sim N(0, \sigma^2)$
- 残差平方和S_eは次のように求められる

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

- これの自由度は$n-2$なので、回帰の残差分散s_e^2は

$$s_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}$$

となる

残差の性質

- 更に残差$e_i = y_i - \hat{y}_i$は、以下の 2 つの性質を持つ

- 残差の和は0になる

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

- 残差と従属変数とは直交する

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0$$

残差の性質

- 誤差εが正規分布に従う仮定$\varepsilon \sim N(0, \sigma^2 I)$の下で、残差$e$も正規分布に従う

$$e \sim N\left(0, \sigma^2 (I - X(X^T X)^{-1} X^T)\right)$$

$$X^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$$

- ここで、 $H = X(X^T X)^{-1} X^T$はハット行列と呼ばれる

レバレッジ（梃子比）

梃子＝てこ

- ハット行列Hの第i対角要素をレバレッジ（梃子比） h_{ii}という
- 梃子比は以下の性質を持つ

$$\frac{1}{n} < h_{ii} < 1$$

- 従属変数の個数をk個とすると、以下の性質が成り立つ

$$\sum_{i=1}^{n} h_{ii} = k + 1$$

レバレッジとスチューデント化残差

- レバレッジ（梃子比）を用いて、残差の分散とスチューデント化残差を以下のように得ることができる
- 残差e_iの分散

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

- スチューデント化残差

$$\frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

クックの距離

- １つのデータを除去して推計される予測値と、全データを用いて推計される予測値との差の平方和を、誤差分散の推定値で割ったもの

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot s_e^2}$$

- ここで、$\hat{y}_j$は全データを用いて得られる予測値、$\hat{y}_{j(i)}$はデータiを除去して得られる予測値、pは説明変数の数（単回帰分析のとき$p = 2$）、s_e^2は残差分散の推定値である。

残差解析と外れ値の検出

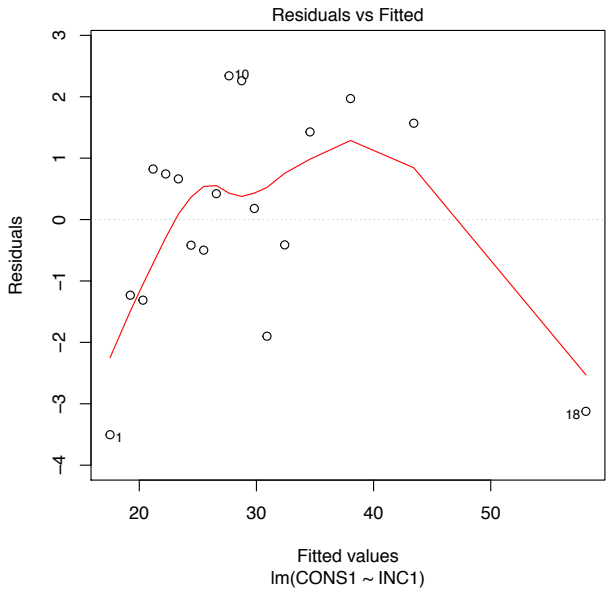
- 残差についての性質を調べることで、回帰分析用いられたデータが外れ値なのかどうかを判断する材料を提供できる。主に以下の手法が用いられる

- 残差プロット
- 残差の正規Q-Qプロット
- S-Lプロット
- 梃子（てこ）比とクックの距離

残差プロット

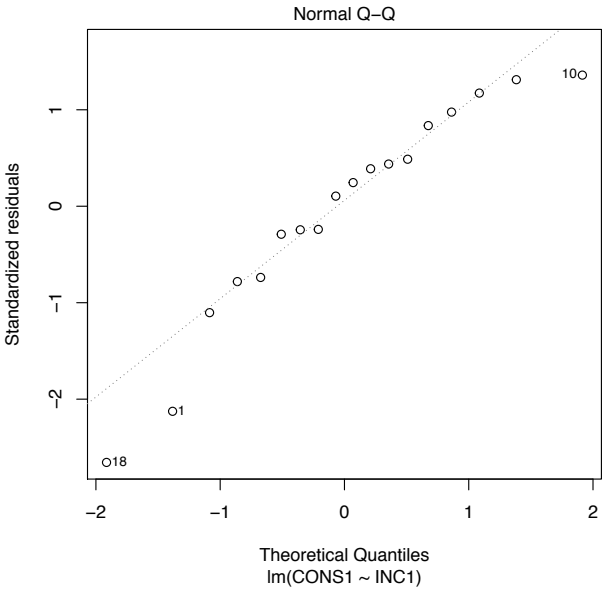
- 予測値$\hat{y}_i$を横軸、残差e_iを縦軸に描いた散布図
- 縦軸の絶対値が$2\hat{\sigma}$を超える残差が多く見られるようなら、データを採用するという仮定を疑うべき→そのデータは外れ値である可能性がある

- 右の結果の場合、$i = 1, 10, 18$が外れ値の可能性がある



残差の正規QQプロット

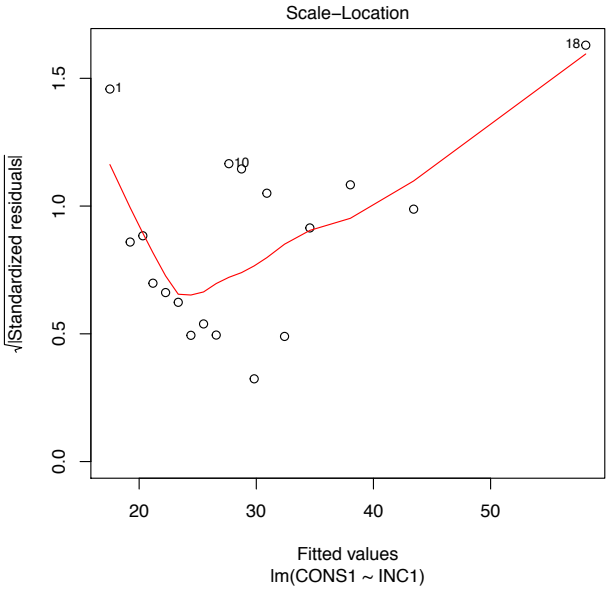
- スチューデント化残差に対する正規Q-Qプロット
- Q-Qプロットではデータ正規分布に従うとき45°線上にデータが乗ってくる性質がある。残差が正規分布に近ければ45°線上にプロットされる
- 縦軸で絶対値が 2 を超える残差が多い場合は、外れ値の可能性がある
- 右の結果の場合、$i = 1, 18$が外れ値の可能性がある



S-Lプロット

- スチューデント化残差の絶対値の平方根を予測値に対して描いた散布図
- 縦軸が$\sqrt{2}$を超えるようなら、データが外れ値であることへの注意が必要

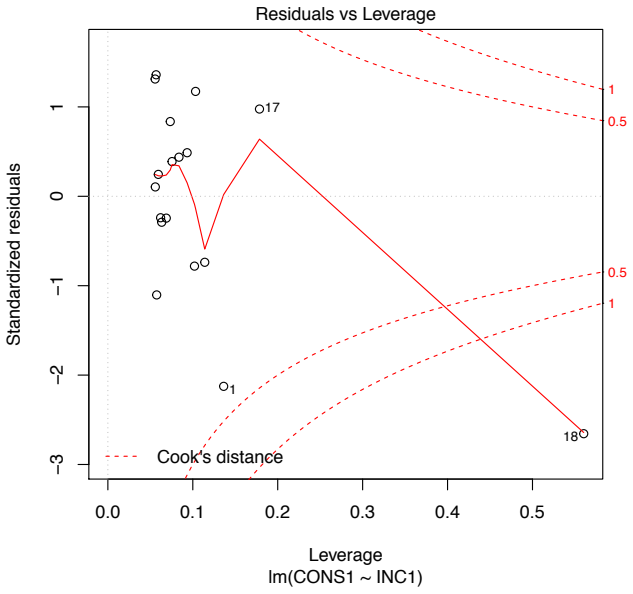
- 右の結果の場合、 $i = 1, 18$が外れ値の可能性がある



梃子比とクックの距離

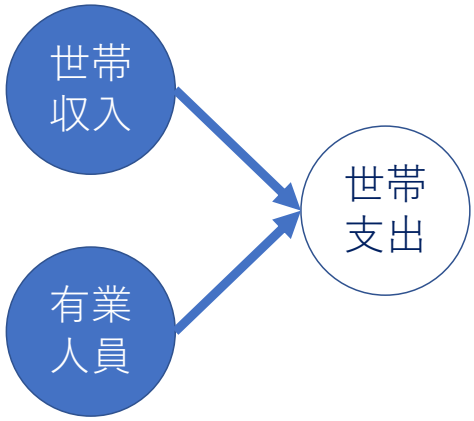
- 梃子比h_{ii}が$2(k+1)/n$より小さければ注意が必要
- 実践的にはクックの距離D_iが $0.2 < D_i \leq 0.5$なら「要注意」、 $0.5 < D_i$なら当該データを「解析から除去」するのが望ましい

- 右の結果の場合、 $i = 18$を除去したほうがよく、 $i = 1, 17$は要注意データだと判断できる



変数を増やしてみよう

- 世帯の支出が増える原因は、所得が増えるだけではない
- 他の原因を追加できないか？
- 例えば、同じ世帯で働いている人の人数が増える場合も、支出が増える可能性がある
  - 共働き、同居する子供が働く等



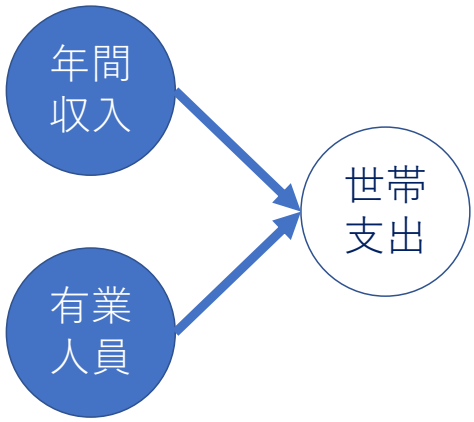
重回帰分析

- 複数の独立変数（説明変数）を用いて回帰分析を行うことを、重回帰分析という

- 独立変数をそれぞれ年間収入x_1、有業人員x_2と表す
- このとき、重回帰モデル式は

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

と表すことができる

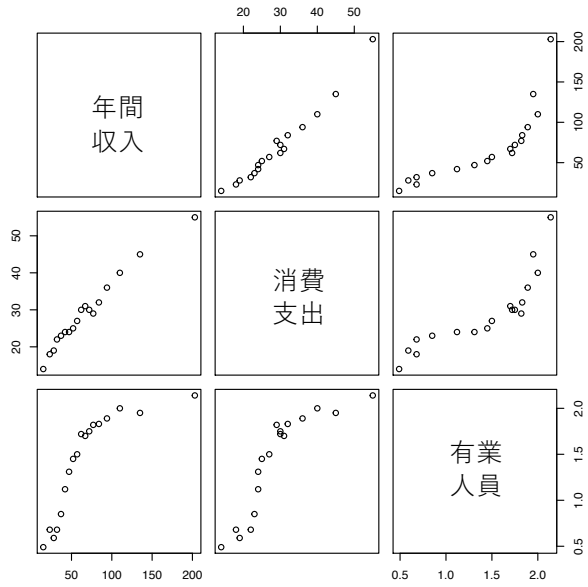


分析に用いるデータ

- 『家計調査』2017年9月、第２－６表「年間収入階級別１世帯当たり１か月間の収入と支出」
- データは政府統計ポータルサイトe-statから入手可能
- 「二人以上の世帯」
- 所得x_{1i}＝「年間収入(10万円)」、有業人員数x_{1i}＝「有業人員(人)」、支出y_i＝「消費支出(千円)」を用いる
- 「年間年収階級」の18階級を分析単位とする

散布図を描いてみる

- 有業人員と消費支出との間にも正の相関関係がありそう？
  - 直線的な関係ではない？
- 有業人員と年間収入との間にも正の相関関係がありそう？
  - これについては後で検討する



重回帰分析の分析結果

- 単回帰分析の結果は以下のようになった ()内はt値

$$\hat{y} = \underset{(10.82)}{12.04} + \underset{(13.18)}{0.187}x_1 + \underset{(2.50)}{2.98}x_2$$

自由度修正済み$R^2 = 0.977$

- $\hat{\beta}_1 = 0.19 > 0$かつ$\hat{\beta}_1$のt値=13.18と5%水準で統計的に有意
- $\hat{\beta}_2 = 2.98 > 0$かつ$\hat{\beta}_2$のt値=2.50と5%水準で統計的に有意
- 自由度修正済み$R^2 = 0.977$と1に近い

- 偏回帰係数$\hat{\beta}_1$と$\hat{\beta}_2$を比較したとき、消費支出に与える影響はどちらが大きいのか？

重回帰分析の分析結果

- 単回帰分析の結果は以下のようになった

$$\hat{y} = \underset{(10.82)}{12.04} + \underset{(13.18)}{0.187}x_1 + \underset{(2.50)}{2.98}x_2$$

()内はt値

自由度修正済み$R^2 = 0.977$

- 偏回帰係数$\hat{\beta}_1$と$\hat{\beta}_2$を比較したとき、消費支出に与える影響はどちらが大きいのか？
- 年間収入x_1が１単位(10万円)増えると消費支出が187円(=0.187X1000円）増える
- 有業人員x_2が１単位(１人)増えると消費支出が2980円(=2.98X1000円）増える
- 単位が異なると偏回帰係数の影響を比較しづらい？

データの標準化と標準化偏回帰係数

- 単位が異なる複数の変数を用いる場合や、単位に意味がない変数（例：５段階評価等）を用いる場合
- 偏回帰係数を比較するために、独立変数と従属変数を平均0・標準偏差1となるデータに標準化する
- 変数xに対する標準化データz_xは以下のように得られる

$$z_x = \frac{x - \bar{x}}{sd(x)}$$

- 標準化した変数を用いて回帰分析をした結果、得られた偏回帰係数を標準化偏回帰係数という

データの標準化

- 例えば消費支出y_i（千円）のデータ

14	18	19	22	23	24	24	25	27	30	31	30	29	32	36	40	45	55

- 平均29.11、標準偏差10.05より

$$\frac{y_i - 29.11}{10.05}$$

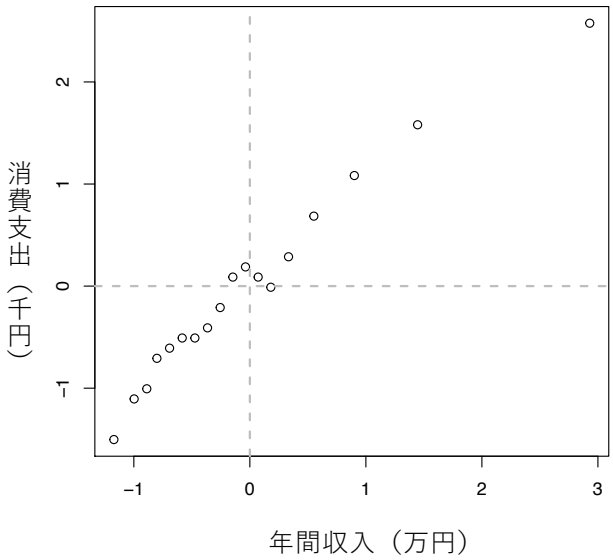
を計算すると、標準化後のデータz_yが得られる

-1.50	-1.11	-1.01	-0.71	-0.61	-0.51	-0.51	-0.41	-0.21	0.09	0.19	0.09	-0.01	0.29	0.69	1.08	1.58	2.58

- 標準化後のデータz_yは、必ず平均0・標準偏差1となる

データの標準化

- 標準化後の年間収入と消費支出の散布図をプロットすると、右図のようになる
- いずれのデータも、平均0±1の辺りに分布していることがわかる



標準化後のデータを用いた回帰分析

- 標準化後の年間収入z_{x1}、有業人員z_{x2}および消費支出z_yを用いて再度重回帰分析をすると、以下のような結果が得られる

$$\widehat{z_y} = \underset{(0.00)}{0.00} + \underset{(13.18)}{0.853} z_{x1} + \underset{(2.50)}{0.162} z_{x2}$$

()内はt値

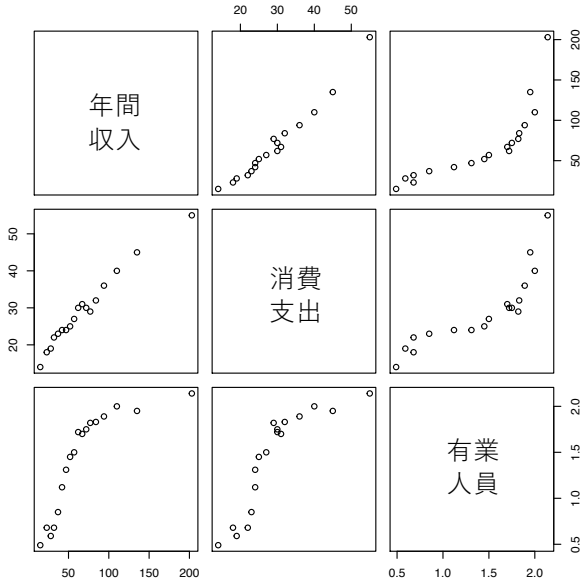
自由度修正済み$R^2 = 0.977$

- 標準化後の独立変数に対する標準化偏回帰係数はいずれも正で5%水準で統計的に有意である
- 年間収入の方が回帰係数が大きいからといって、消費支出に与える影響は有業人員より年間収入のほうが大きい訳ではない
- 1標準偏差変化した場合の変化量を示しているに過ぎず、元のデータの分散を把握しておく必要がある。

変数間の相関

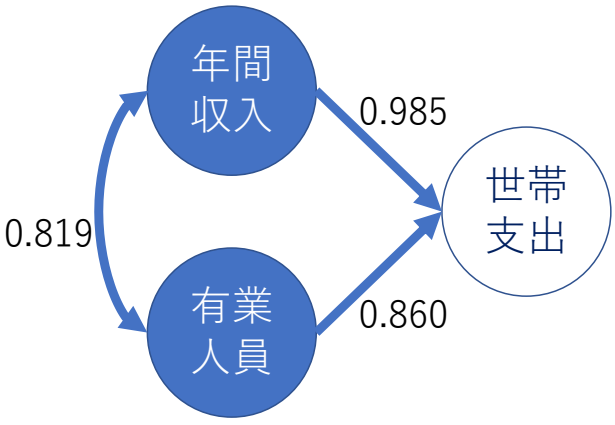
- 独立変数である年間収入と有業人員は、互いに正の相関関係にあるかもしれない
- 実際に、３変数間の相関係数を求めると以下のようになる

	消費支出	年間収入	有業人員
消費支出	1.000		
年間収入	0.985	1.000	
有業人員	0.860	0.819	1.000



変数間の相関

- （当然だが）世帯内で働いている人の数（有業人員）が増えると、世帯の年間収入は増加するだろう
- 独立変数同士が相関する場合、両方の変数を同時に独立変数として採用してよいのか？
- どちらかを採用するとしても、どちらを採用するべきか？



多重共線性

- 独立変数間に関連がある場合、以下のような状況が生じることがある

- t値が過小評価される（実際に有意でも有意でなくなる等）
- 偏回帰係数の標準誤差（分散）が大きくなる（回帰が歪む）
- 決定係数の値が大きくなる
- 偏回帰係数の符号が本来なるべき符号とは逆の符号になる

多重共線性の測定

- 分散拡大係数 (variance inflation factor: VIF) を用いて多重共線性の深刻度を測定する
- 独立変数x_1とx_2の標準偏差σ_{x1}とσ_{x2}および共分散σ_{x1x2}から相関係数rが次式より計算できる

$$r = \frac{\sigma_{x1x2}}{\sigma_{x1}\sigma_{x2}}$$

- 相関係数rを二乗した重相関係数r^2を用いて、VIFは以下のように計算される

$$VIF = \frac{1}{1 - r^2}$$

多重共線性の測定

- VIFが10以下のとき多重共線性がないと判断される（理想的には2以下）。VIFが10以上のとき、どちらかの変数を外して再度回帰分析を行う。

- 年間収入x_1と有業人員x_2の$VIF = \frac{1}{1-0.819^2} \approx 3.04$となる
- 従って、この 2 変数の間に多重共線性があるとはいえない