データ解析 2016：レポート試験　（問合せ先：瀧川 takigawa@ist.hokudai.ac.jp）

提出先:　8/4の授業日 or 6-16瀧川居室ドアポスト or takigawa@ist.hokudai.ac.jpへメール提出
〆切：　8/5(金) 17:00まで (難しい場合は事前に相談してください！)
注意：レポートには必ず名前と学生番号を記載すること！！

【問題】(意図：やっぱりパソコン使っても良いので一回は手計算してみないと定着しづらいよね…)
(1) (回答必須) 表1のデータに対して回帰分析を行い、空欄(小数点以下2桁)を埋めて結果の一覧を完成させよ。電卓、そろばん、Excel、R、Python、Wolfram Alphaなど好きな計算環境を用いて良い。
(2) (回答必須) 上記の各々の数値を得る計算の方法を簡単に説明せよ。(ヒント：次ページ)
(3) (余裕ある人のみ) 各々の係数についてp値および95%信頼区間を算出し、手順を簡単に説明せよ。

【講義Webサイト】
いままでの講義スライドや下の表1のデータファイルを置いています。
注: httpじゃなくhttpsでないと資料にアクセスできません。(たぶん)
https://art.ist.hokudai.ac.jp/~takigawa/course/da2016
(閲覧ID da2016 パスワード pokemongo)

表1

	目的変数	説明変数			
	Y	X_1	X_2	X_3	X_4
番号	新生児の体重	母親の体重(kg)	母親の年齢	懐妊期間(日)	喫煙習慣の有無 (有ならば1,無ならば0)
1	3087	48	28	304	0
2	3229	52	24	286	1
3	3204	61	33	273	1
4	3346	58	30	295	0
5	3579	56	21	290	0
6	2325	46	26	262	0
7	3159	55	30	318	1
8	3589	63	37	298	0
9	2969	52	25	299	1
10	2819	40	22	313	0
11	3191	59	34	285	1
12	3346	57	28	306	0
13	2444	45	30	291	1
14	3662	64	21	274	0
15	3241	53	29	283	0

回帰分析の結果一覧

回帰統計	
重相関 R	
重決定 R^2	
自由度調整済 R^2	
標準誤差 $\sqrt{\hat{\sigma}^2}$	
観測数 n	
説明変数の数 p	

	係数 β_i	標準誤差	t 比
切片(定数項)			
説明変数 X_1			
説明変数 X_2			
説明変数 X_3			
説明変数 X_4			

・Excelの利用について：研究科の学生は以下でMicrosoft Office 2016が個人PCでも使えます

【申請方法】
♦教育情報システム（ELMS）にログインのうえ、ポータル内の注目コンテンツ『大学提供ソフトウェア利用申請』リンクから「包括契約ソフトユーザー申請システム」に入り、『Microsoftライセンス』を選択し、利用規約同意後のシステム内「利用者メニュー」にあります「Microsoftユーザー申請・インストール手順」に従って、申請及びインストール願います。
♦利用可能ソフトウェアは別紙のとおりとなっております。
※※※※※※※※※※※※※※※※※※
インストールされる「Microsoft Office365 ProPlus」に含まれるソフトウェア
（Windows及びMac OSで利用可能です）
*MicrosoftWord
*MicrosoftExcel
*MicrosoftPowerPoint
*MicrosoftOneNote ＊
*MicrosoftOutlook
*MicrosoftLync（利用できません）
*MicrosoftPublisher ＊
*MicrosoftAccess ＊
＊Windowsのみの機能

・ヒント：回帰分析(多変量版)の手順は以下の感じ （各々の詳細は講義スライド等で復習）

やりたいこと： 予測式 $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ を作りたい

1) まずデータからβ_0, β_1, β_2, β_3, β_4の推定値を求める (行列公式 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$)
2) それぞれのYに対して予測式を用いて予測値 $\hat{Y}$ を求める
3) 残差変動 $\Sigma_i (Y_i - \hat{Y}_i)^2$ を自由度n-p-1で割って(n=15,p=4)、誤差分散の不偏推定量 $\hat{\sigma}^2$ を得る。そのルートをとった値が回帰全体の「標準誤差」
4) Yの平均値 $\bar{Y}$ を計算し、それを用いて全変動 $\Sigma_i (Y_i - \bar{Y})^2$ と回帰変動 $\Sigma_i (\hat{Y}_i - \bar{Y})^2$ を求める。「回帰変動/全変動」の値が「決定係数 R^2」、「1-(残差変動/(n-p-1))/(全変動/(n-1))」の値が「自由度調整済み決定係数」、決定係数のルートが「重相関 R」($\hat{Y}_i$とY_iの相関係数に等しい)
5) 係数の推定量 $\hat{\boldsymbol{\beta}}$ の分散共分散行列 $\hat{\Sigma}_\beta$ を求める (行列公式 $\hat{\Sigma}_\beta = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$)
その対角成分の値のルートがそれぞれ対応する切片および説明変数の「標準誤差」
係数/標準誤差が各々の「t比」
6) 「t比」の値は自由度n-p-1のt分布に従うことから、この分布で得られたt比以上の値が出る確率の2倍が「p値」。2倍するのはt分布が対象で両端の裾の確率を見るから。(次ページ参照)
7) 自由度n-p-1のt分布でP[ltl > a]=0.05となるようなa (境界値:critical value)を求める。このとき「係数 ＋ a × 標準誤差」が0.95%(1-0.05%)信頼区間の上限、「係数 – a × 標準誤差」が下限

・ヒント2：自由度5の分布でt比-3のときの「p値(両側の上側確率)」(約0.03)を求める

Excel 2010以降

セルに「=T.DIST.2T(ABS(-3),5)」と入力し0.030099248を得る。

R

「2 * pt(abs(-3),df=5,lower.tail=FALSE)」と入力し0.03009925を得る。

Python (要numpyとscipy←Anacondaで入れれば勝手に入っています)

```
from scipy import stats
import numpy as np
2 * stats.t.sf(np.abs(-3), 5)
```

を実行し0.030099247897462569を得る。

Wolfram Alpha (www.wolframalpha.com/)

「P[|X|>3] for X ~ student t with 5 dof」と入力して0.03009924789746…を得る

・ヒント3：自由度17のt分布で、両側の上側確率が0.05(5%)となるようなtの値(約2.11)を求める

Excel 2010以降

セルに「=T.INV.2T(0.05,17)」と入力し2.109815578を得る。

R

「qt(0.05/2,df=17)」と入力し-2.109816を得る。

Python (要numpyとscipy←Anacondaで入れれば勝手に入っています)

```
from scipy import stats
import numpy as np
stats.t.isf(0.05/2,17)
```

を実行し2.1098155778331811を得る。

Wolfram Alpha (www.wolframalpha.com/)

「Quantile[StudentTDistribution[17], 0.975]」と入力して2.10982を得る

t分布表 (統計学の本の付録についている)を使う。dfは「自由度(degree of freedom)」

	有意確率								
両側	0.10	0.05	0.01	0.001		0.10	0.05	0.01	0.001
片側	0.05	0.025	0.005	0.0005		0.05	0.025	0.005	0.0005
df					df				
1	6.3138	12.706	63.657	636.62	18	1.7341	2.1009	2.8784	3.922
2	2.9200	4.3027	9.9248	31.598	19	1.7291	2.0930	2.8609	3.883
3	2.3534	3.1825	5.8409	12.941	20	1.7247	2.0860	2.8453	3.850
4	2.1318	2.7764	4.6041	8.610	21	1.7207	2.0796	2.8314	3.819
5	2.0150	2.5706	4.0321	6.859	22	1.7171	2.0739	2.8188	3.792
6	1.9432	2.4469	3.7074	5.959	23	1.7139	2.0687	2.8073	3.767
7	1.8946	2.3646	3.4995	5.405	24	1.7109	2.0639	2.7969	3.745
8	1.8595	2.3060	3.3554	5.041	25	1.7081	2.0595	2.7874	3.725
9	1.8331	2.2622	3.2498	4.781	26	1.7056	2.0555	2.7787	3.707
10	1.8125	2.2281	3.1693	4.587	27	1.7033	2.0518	2.7707	3.690
11	1.7959	2.2010	3.1058	4.437	28	1.7011	2.0484	2.7633	3.674
12	1.7823	2.1788	3.0545	4.318	29	1.6991	2.0452	2.7564	3.659
13	1.7709	2.1604	3.0123	4.221	30	1.6973	2.0423	2.7500	3.646
14	1.7613	2.1448	2.9768	4.140	40	1.6839	2.0211	2.7045	3.551
15	1.7530	2.1315	2.9467	4.073	60	1.6707	2.0003	2.6603	3.460
16	1.7459	2.1199	2.9208	4.015	120	1.6577	1.9799	2.6174	3.373
17	1.7396	2.1098	2.8982	3.965	∞	1.6449	1.9600	2.5758	3.291

・補足：授業ではあまり触れなかったが、多変量の場合、各係数を「偏回帰係数」などとも呼ぶ。

回帰結果にはこの偏回帰係数の他、各変量を最初に平均0,分散1に標準化してから回帰係数を計算した場合に得られる「標準化偏回帰係数」もしばしば含まれる。例えば以下の出力はIBM SPSSという統計ソフトウェアの回帰分析の出力例である。(従属変数とは目的変数の別名)

係数[a]

モデル		標準化されていない係数		標準化係数			共線性の統計量	
		B	標準誤差	ベータ	t値	有意確率	許容度	VIF
1	(定数)	-3.237	.203		-15.920	.000		
	世帯の収入(千単位)	.014	.001	.256	10.334	.000	.535	1.869
	年齢	.030	.006	.113	5.038	.000	.652	1.533
	現職の雇用期間(年)	.169	.009	.539	18.963	.000	.407	2.458
	不履行予測, モデル 1	6.423	.180	.776	35.611	.000	.692	1.446

a. 従属変数 クレジットカードの負債(千単位)

この標準化回帰係数は説明変数を標準化して回帰を再計算しなくても以下の要領で、標準化されていない通常の偏回帰係数から計算できる。

※標準化偏回帰係数

偏回帰係数の大きさは各々の変量のダイナミックレンジに依存する！たとえば、x1は1～100までの値をとり、x2は-1～1までの場合、β1のほうがβ2より小さくなりやすいがこれは必ずしもβ1のほうがβ2より目的変数を説明しないことを意味しない。

なので、目的変数、および、各説明変数を標準化してから計算した偏回帰係数を**標準化偏回帰係数**と言う。

標準化偏回帰係数 $\tilde{\beta}_i = \hat{\beta}_i \frac{\mathrm{sd}(x_i)}{\mathrm{sd}(y)}$ として非標準化係数から計算できる