1. 相関と線形回帰

この章の目標は**相関**と**線形回帰**を導入することです。これらは連続的な予測変数と連続的な結果変数の間の関係を分析する時に使われる，標準的なツールなのです。

1.1 相関

このセクションでは，データの変数 *間*の関係をどうやって記述するかについて論じようと思います。そのために，変数間の**相関**について主に話すことになります。でもまず，あるデータについて触れておかなければなりません。

1.1.1 データ

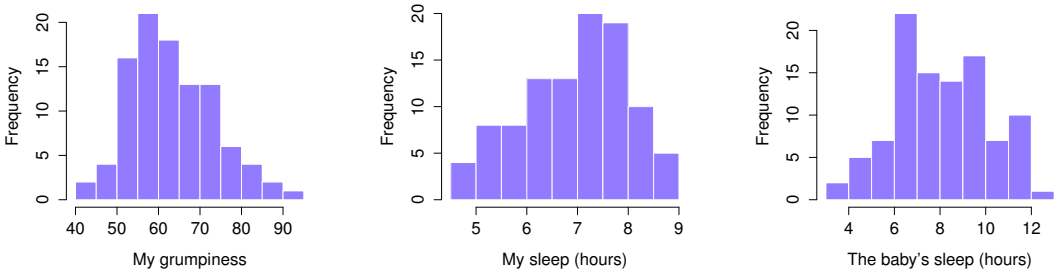
Table1.1　親子関係データの記述統計量

変数名	最小値	最大値	平均値	中央値	標準偏差	四分位範囲 (IQR)
ダンの不機嫌さ	41	91	63.71	62	10.05	14
ダンの睡眠時間	4.84	9.00	6.97	7.03	1.02	1.45
ダンの息子の睡眠時間	3.25	12.07	8.05	7.95	2.07	3.21

どの親でも身近な話題である，睡眠についてのトピックを見てみましょう。ここで使うデータセットは架空のものですが，事実に基づいたものです。Suppose I'm curious to find out how much my infant son's sleeping habits affect my mood. 私の不機嫌さを，0(全く不機嫌でない) から 100(非常に非常に不機嫌である) までのスケールで正確に評定できるとしましょう。さらに私の不機嫌さを

と，睡眠パターン，そして息子の睡眠パターンについて，測定していたとしましょう。そうですね，100 日分ぐらい。そして，マニアックなことに，私はそのデータを parenthood.csv というファイル名で保存したのです。そのデータを JASP に読み込むと，4 つの変数がああることがわかります。dan.sleep, baby.sleep, dan.grump そして day です。初めて JASP セットに読み込むときは，それぞれのデータ変数型が正しく読み込まれないかもしれませんが，そのときは次のように修正してください。dan.sleep, baby.sleep, dan.grump そして day 変数は連続変数であると指定し，ID は名義的な I(整数の) 変数にするのです。

次に，基本的な記述統計量を見てみましょう。この興味深い三つの変数それぞれがどのようになっているか，可視化した記述を与えてくれるのが，図 **??** に示したヒストグラムのプロットです。もう一つ注意点を。JASP はいくつもの複数の異なる統計量を計算することができるからといって，全てをレポートする必要はありません。これをレポートにするときには，おそらくこれらの統計量の中から自分 (と私の読者) にとって最も興味があるものをプッ苦アップして，それを素敵で単純な表のような形にして示すでしょう。表 **??**のようにね[*1]。私がそれを表にする時には，全て “人間が読める” 変数名を付与します。これはいいやり方です。私は十分に寝れていないことにも注意してください。これは良いやり方ではないですね。でも他の親御さんたちからはよくあることだと言われます。



(a) (b) (c)

Figure1.1 parenthood データセットに含まれる 3 つの変数のヒストグラム

1.1.2 The strength and direction of a relationship

2 つの変数がどれぐらい密接に関係しているかを表す一般的な方法として，散布図を描くことができます。理想を言えば，もうちょっと言葉を足したいところですが。例えば，dan.sleep と dan.grump

[*1]実は，表にしてもまだ悩ましいものです。実際にほとんどの人は，中心化傾向の測度 *ひとつだけ*，変数の *ある*側面だけをピックアップします。

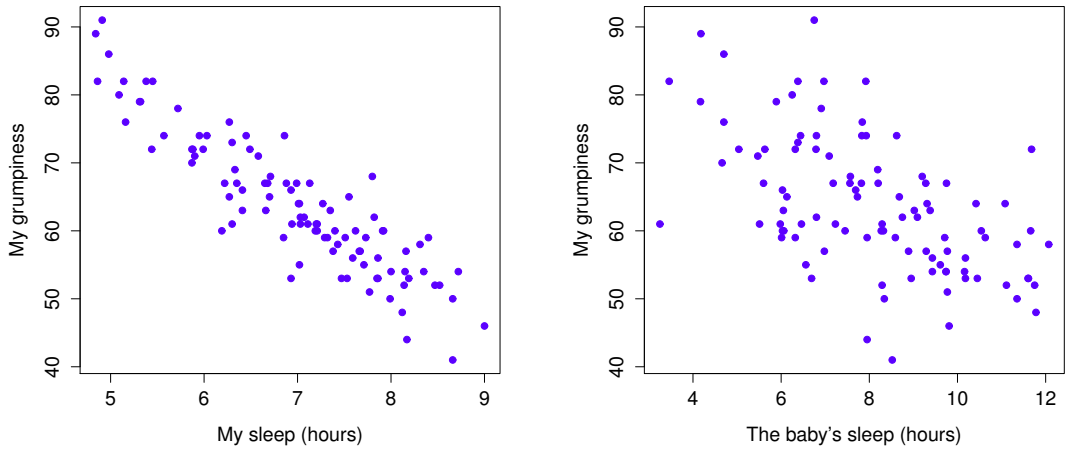


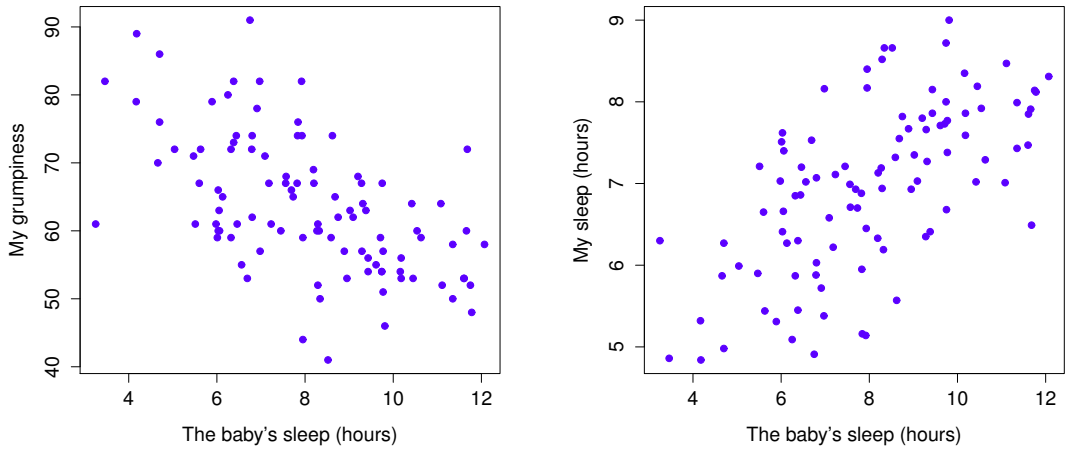
Figure1.2　dan.sleep と dan.grump の関係を示す散布図 (左) と，baby.sleep と dan.grump の関係を表す散布図 (右)

(図 **??**, 左) の関係と，baby.sleep と dan.grump (図 **??**, 右) の関係を比較してみましょう。両者を並べてみてみたら，明らかに両者が *質的に*同じであることがわかります。よく寝るとイラつきが減るのです! しかし，dan.sleep と dan.grump の関係が baby.sleep と dan.grump の関係よりも *より強い*こともまた，明らかです。左側のプロットは右側のものよりも "弱い" のです。もし私の機嫌を予測したいと思ったのなら，私の息子が何時間ぐらい寝たかを知ることで多少の助けにはなりますが，それよりも私が何時間寝たかを知った方がより役に立つわけです。

比較として，図 **??**をみてみましょう。"baby.sleep と dan.grump" の散布図 (左) を，"baby.sleep v dan.sleep" の散布図 (右) と比べてみると，全体的な関係の強さは同じですが，方向が逆になっています。つまり，もし私の息子がよく寝てくれたら，私は *より多く*寝ることができますが (正の関係，右側です)，彼がよく眠ると私のムカつきは *より少なく*なるのです (負の関係，左側です)。

1.1.3　相関係数

この考え方を，もう少し正確にして **相関係数**の考え方を導入しましょう (もっと正確に言えば，ピアソンの相関係数です)。これは伝統的に，r で表されます。変数 X と Y の間の相関係数 (r_{XY} と表されます) は，次のセクションでもう少し正確に定義しますが，その測度は -1 から 1 の間の値になります。もし $r = -1$ であれば，完全に負の関係にあることになり，$r = 1$ であれば，完全に正の関係にあると言えます。もし $r = 0$ であれば，全くの無関係です。図 **??**に異なる相関関係がどういう



(a) (b)

Figure1.3 散布図が示すのは `baby.sleep` と `dan.grump` の関係 (左)，それと `baby.sleep` と `dan.sleep` の関係 (右) です

. . . . . .

見え方をするか，いくつか示してあります。

ピアソンの相関係数の式は，いくつかの表記方法があります。その式を最も簡単に書き下す方法は，2 つのステップに分けることだと思います。まず，**共分散**の考え方を導入します。変数 X と Y の共分散は分散の一般化として導入でき，数学的には 2 つの変数の関係を記述するものです。それほど有益なものではありません。

$$\mathrm{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)$$

X に関する量と Y に関するの量掛け算をして ("積")，それを平均しています[a]。共分散の式は X と Y の "積和平均" だと考えることができます。

共分散は良い特徴を持っていて，もし X と Y が完全に関係ない場合，共分散はゼロになります。もしその関係が正であれば (図 **??**に示したように)，共分散は正の値になりますし，その関係が負

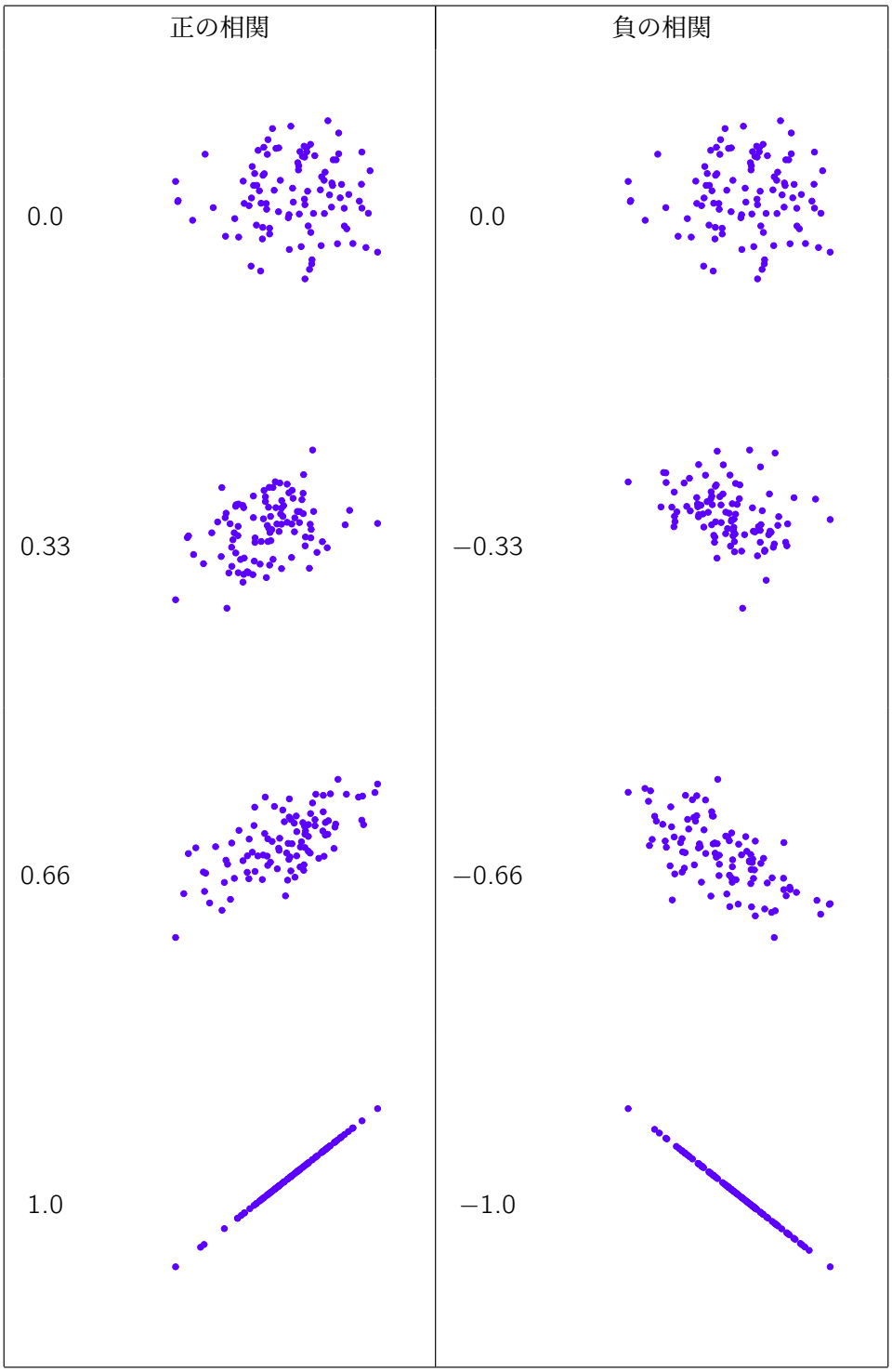


Figure1.4　相関の強さと方向を変えた影響を図示したもの。左の列は，相関が 0，.33，.66，そして 1 になっています。右側の列は，相関が 0, -.33, -.66, そして-1 のものです。

であれば共分散の値もまた負になるのです。言い換えれば，共分散は相関の基本的な考え方をうまく捉えています。

残念ながら，共分散の大きさは簡単に解釈できません。というのもそれは X と Y の持っている単位に依存しますし，さらに悪いことに，共分散の実際の単位は少し奇妙なのです。例えば，もし X が `dan.sleep` 変数 (単位：時間) で，Y が `dan.grump` 変数 (単位；不機嫌さ) のとき，その共分散は “時間 × 不機嫌さ” になります。これがどういう意味なのか，私にはさっぱりです。

ピアソンの相関係数 r は共分散を標準化するために解釈の問題を解決してくれます。まさに，元のデータを標準偏差で割った標準化した z スコアと同じように。しかし，2 つの変数が共分散に関わっていますので，標準化するのはそれぞれの標準偏差で割っていることになります[b]。言い換えるなら，X と Y の相関係数は次のように書くことができるのです。

$$r_{XY} = \frac{\mathrm{Cov}(X, Y)}{\hat{\sigma}_X \, \hat{\sigma}_Y}$$

[a]分散と標準偏差でもみてきましたが，実践的には N ではなく $N - 1$ で割るのです
[b]これはちょっと単純化しすぎですが，ここでの目的には合います。

共分散を標準化することによって，先ほど指摘した共分散の良い特徴を失うことなく，r の値が強度を表すようになります。すなわち，$r = 1$ は完全な正の相関を表しますし，$r = -1$ は完全に負の相関を表すことになります。この点については後ほど，セクション **??**でもう少し広げます。でもその前に，JASP でどうやって相関を計算するのかをみておきましょう。

1.1.4 JASP で相関を計算する

JASP で相関を計算するのは，‘回帰’-‘相関行列’ ボタンをクリックするだけです。4 つの変数全てを右のボックスにうつし，図**??**の出力を得ます。各相関係数 (‘ピアソンの r’ とあるもの) に，p-値がついていることに注意してください。明らかに，なんらかの検定が行われていますが，今それは無視することにしましょう。それについては，近々お話しすることになるでしょう!

1.1.5 相関係数を解釈する

当然，実際の生活では相関係数が 1 になるのを目にすることはありません。では，例えば $r = .4$ であればどう解釈したら良いのでしょう? 丁寧に答えるなら，あなたがそのデータをどう使うかによりますし，あなたのフィールドでどれぐらい強いといえるのかによります。工学系の友人と相関係数と話した時，.95 以下であれば全く無意味だということでした (工学系の中でも，ちょっと話を盛ってる気がしますけど)。一方，心理学でもそうですが，それぐらいの強さを期待できる相関係数

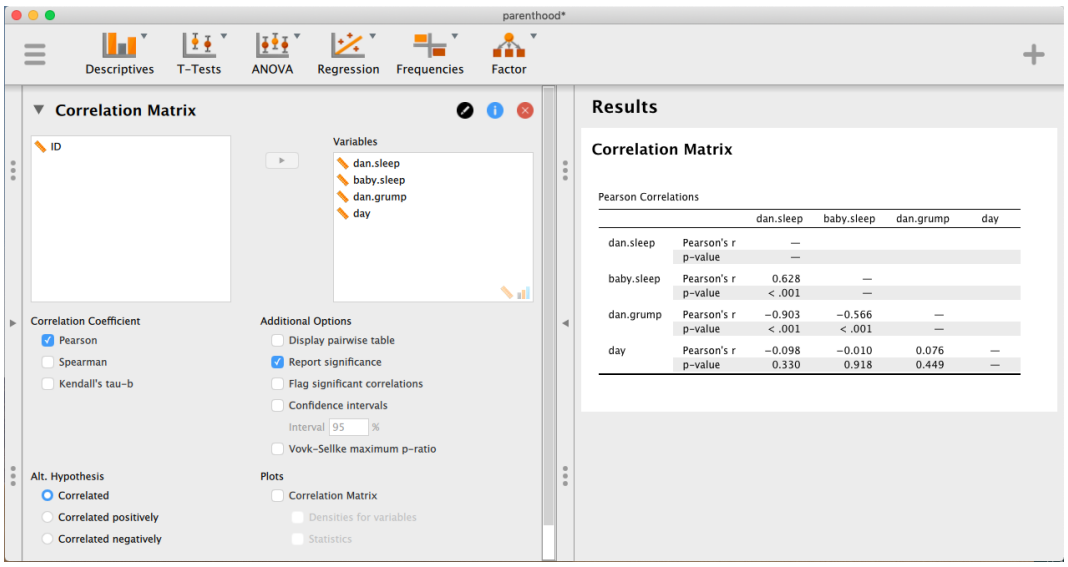


Figure1.5　`parenthood.csv` ファイルにある変数間の相関係数を示した JASP スクリーンショット

が実際にみられることもあります。たとえば，人はどのように類似性を判断するかの理論を検証するのに使うベンチマークデータセットは非常にクリーンで，少なくとも .9 の相関を達成しないようであれば成功したと言えないそうです。とはいえ，知能の基本的な相関 (たとえば，洞察の時間，反応時間) を探しているときは，相関係数が .3 を超えることがあればうまく行った方です。要するに，相関係数の解釈は文脈に依存するということです。とはいえ，表 **??**が典型的な例です。

しかし，データにいかなる解釈をする場合でもその前に，*常に*散布図を確認しろというのは強調しすぎるということはありません。相関はあなたが思うような意味を持ってないかもしれないのです。“Anscome の四人組” (**Anscombe1973**) という 4 つのデータセットについての古典的な例が，このことを示しています。それぞれのデータは 2 つの変数 X と Y をもっています。4 つのデータセットは全て，X の平均が 9 で Y の平均が 7.5 になっています。すべての X の標準偏差もほとんど同じで，Y 変数についてもそうです。そしてどのデータセットも，X と Y の相関係数は $r = 0.816$ になっています。ファイル `anscombe.csv` に用意したので，みなさん自身で確認してみてください。

この 4 つのデータセットがほとんど同じ性質であるとおもったでしょうか。ちがいますよ。4 つのデータセットすべてについて，X と Y の散布図を描いたのが図 **??**ですが，この 4 つはお互い まったく違うものであることがわかるでしょう。ここでの教えは，実生活において我々がよく忘れてしま

Table1.2　相関を解釈する大雑把な目安。*大雑把*なものであると言ってることに注意。関係の強弱に明確なルールはありません。文脈に依存するものです。

相関	強さ	方向
-1.0 to -0.9	とても強い	負
-0.9 to -0.7	強い	負
-0.7 to -0.4	中程度	負
-0.4 to -0.2	弱い	負
-0.2 to 0	無視できるレベル	負
0 to 0.2	無視できるレベル	正
0.2 to 0.4	弱い	正
0.4 to 0.7	中程度	正
0.7 to 0.9	強い	正
0.9 to 1.0	とても強い	正

うことですが，“データはいつも図にする”(第 **??**章) ということです。

1.1.6　スピアマンの順序相関

ピアソンの相関係数は便利な特徴が多くありますが，欠点もあります。ある問題点は特に目立ちます：それが表しているのは，二変数間の *直線的な*関係だということです。言い換えると，わかることはデータがひとつの完全なる直線に収まる程度の指標なのです。時にこれは私たちが “関係がある” という言葉で意味することの良い近似になりますし，ピアソンの相関はそれを計算するのにもってこいです。でも時には，そうならないのです。

ピアソンの相関係数が適切でない場合について，非常にありがちなケースは変数 X が他の変数 Y の増加を反映しているものの，本当の関係は線形ではないような場合です。例として，試験勉強の努力量と結果の関係を考えてみましょう。もし努力量 (X) がゼロなら，成績 (Y) も 0% になるでしょう。でも少し努力すれば，それは *かなり*改善されます。授業に合わせて勉強していくのは良いことですし，授業内容だけに絞って行ってもそれほどの努力なして 35% ぐらいまでは成績が上がるでしょう。でも成績尺度の逆の端では，同じ効果を見込むことはできません。誰でも知っているように，55% の成績を取るよりも 90% の成績を取る方が，*ぐっと*多くの努力を必要とします。何が言いたいかというと，試験勉強と成績のようなデータをとったとすると，そこでピアソンの相関係数を見ることがミスリーディングであるという例になるのです。

これを示すために，図 **??**を用意しました。ここでは 10 人の学生がある授業を履修していた時の，勉強時間と成績の関係をプロットしてあります。この (架空の) データセットで興味深い点は，努力

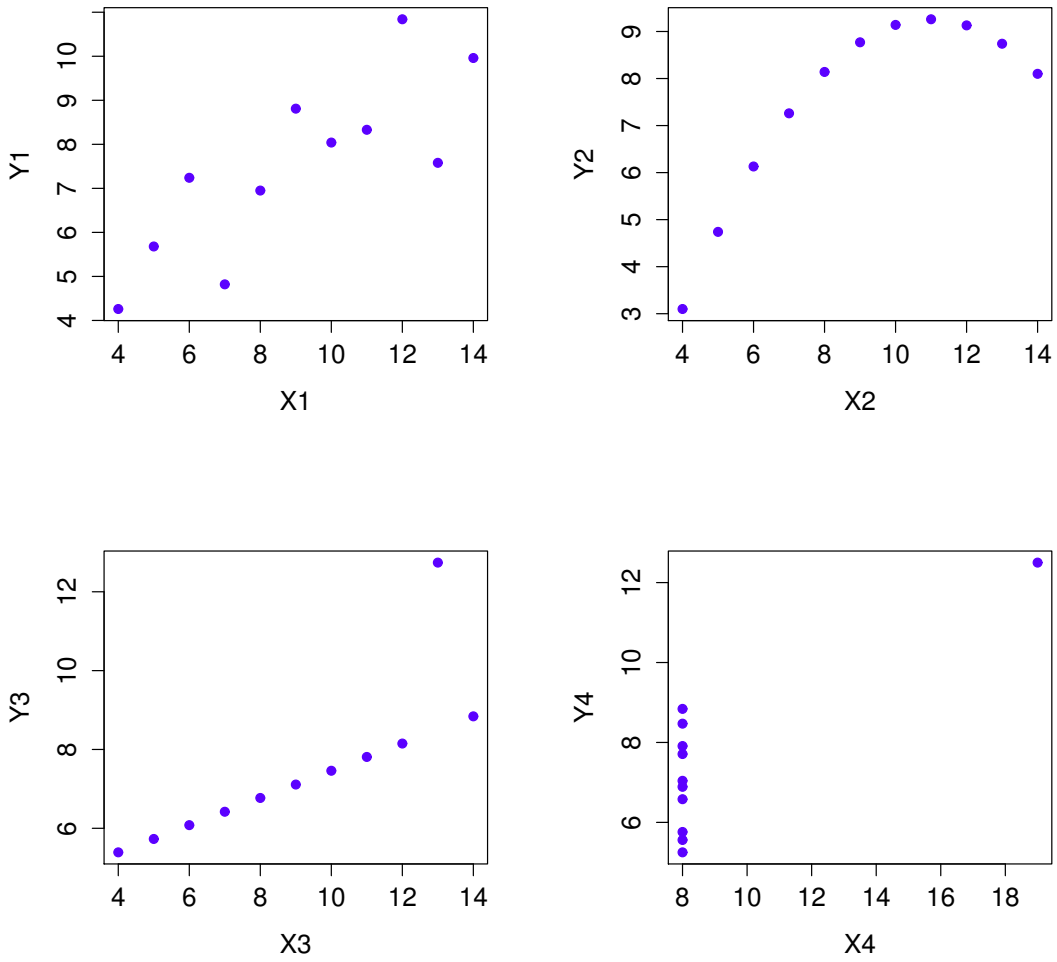


Figure1.6　Anscombe の四人組。この四つのデータセットはいずれも相関係数 $r = .816$ だが，それぞれ質の違うものである。

すると *かならず*成績が上がるというところです。多かろうが少なかろうが，努力量を増やしさえすれば成績が下がることは決してありません。ピアソンの相関係数を計算してみると，勉強時間と成績との関係を表すことができ，その係数は `0.91` であることがわかります。しかし勉強時間を増やせば いつでも成績が上がるという関係を，実際に反映しているとは言えません。ここで本当に私たちが言いたいことは，相関は *完全*であるということなのですが，そのためには “関係” がどうなっているのかについて別の表現が必要だということです。私たちが探しているのはこの事実をうまく反映する何かであって，そこには完璧な**順序的な関係**があるということです。すなわち，学生 1 が学生 2

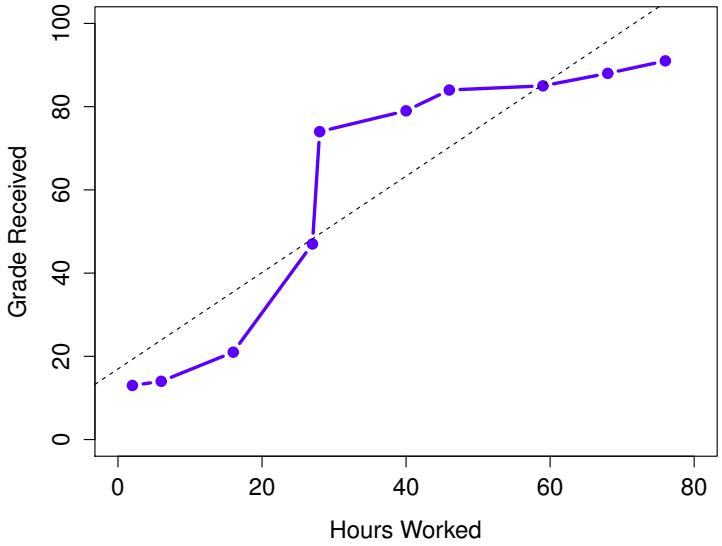


Figure1.7　たった 10 人の学生からなる仮想データセットで，労働時間と成績の関係をしめしたもの (各円が 1 人の学生を表しています)。真ん中にある点線は二変数の線形関係を表しています。ピアソンの相関は強くて $r = .91$ です。でもここで注目すべきは，二変数が完全に単調なかんけいにあるということ。この小さなサンプルでは，労働時間の上昇が常に成績の上昇を意味していて，これが点線で描かれています。これはスピアマンの相関で言うと $\rho = 1$ であることを意味します。でも，このような小さなデータセットでは，どちらのバージョンがより良く実際の関係を表しているといえるのかは，未解決の問題です。

よりもたくさん勉強したら，学生 1 の成績の方が良くなるということを保証したいわけです。相関が $r = .91$ である，という言い方ではなくね。

どうしたらいいでしょう? 実は簡単なことなんです。順序的な関係を見る時，データセットが順序尺度水準であるとして扱えばいいのです! つまり，“勉強時間” という言葉ではなく，10 人の学英の勉強時間順についての順位だといえばいいのです。つまり，学生 1 は誰よりも勉強時間が少ない (2 時間) おですから，最低ランクであるといえます (rank=1)。学生 4 はその次で，半期全体でも 6 時間しか勉強していないのですから，その次に低いランク (rank =2) とします。ここで “rank = 1” が “低い順位” を意味していることに注意してください。日常用語で私たちが “1 位だ” というと，“最下位だ” ではなく “最上位だ” ということを意味しますよね。だから注意してほしいのですが，“一番小さな値から一番大きな値にむけて” 順序づける (小さいことは rank 1 とする) こともできますし，“一番大きな値から一番小さな値に向けて” 順序づける (大きいことを rank 1 とする) こともできます。今回のケースでは小さい方から大きい方に順序づけていきましたが，どういう設定にしたか忘れ

がちなので，ちょっと注意して覚えておいてください!

さて，では努力量と結果について最下位から一位まで順序づけた学生データを見てみましょう。

	順位 (勉強時間)	順位 (試験の成績)
学生 1	1	1
学生 2	10	10
学生 3	6	6
学生 4	2	2
学生 5	3	3
学生 6	5	5
学生 7	4	4
学生 8	8	8
学生 9	7	7
学生 10	9	9

ふむ。両者が *一致*しましたね。もっとも勉強時間が多かった学生は最も良い成績をもらっていますし，勉強時間が最も少なかった学生は成績も最も悪い，などとなっています。上の表に示したように，2 つの順位は一致しているので，その相関を今度みてみると完全な関係，相関係数 `1.0` となります。

ここで再導入したものは**スピアマンの順位相関係数**と呼ばれ，ピアソンの相関係数 r と区別するために ρ で表されます。スピアマンの ρ を JASP で計算するには，単に ‘相関行列’ の画面で ‘スピアマン’ のチェックボックスをクリックするだけです。

1.2 散布図

散布図はシンプルですが効果的なツールで，相関のセクション (セクション **??**) でみたように，*2* つの変数間関係を可視化してくれます。“散布図” という言葉を使う時，私たちが普通思いつくのは応用例です。この種のプロットでは，各データ点がある点に対応しています。ドットのプロットの水平の位置は一方の変数における値で，垂直方向の位置はもうひとつの変数における値です。多くの状況では，その *因果的な*関係 (すなわち，A が B の原因，あるいは B が A の原因である，あるいは他の変数 C が A と B を統制しているなど) がどうなっているかについて，はっきりした意見を言うことは難しいでしょうそうした状況では，どの変数を x 軸において，どの変数を y 軸に置くべきかと言うことは本質的な問題ではありません。しかしある変数が原因になっているに違いないとか，少なくとも方向性についてある見通しがある，と言う幾分強いアイデアがあることも多いでしょう。

もしそうなら，原因変数を x 軸に，影響される変数を y 軸におくといいでしょう。これを覚えておいて，JASP でどうやって散布図を書くかみていきます。使うデータは同じく `parenthood` データセット (ファイルは `parenthood.csv`) で，相関の導入のときに使ったものです。

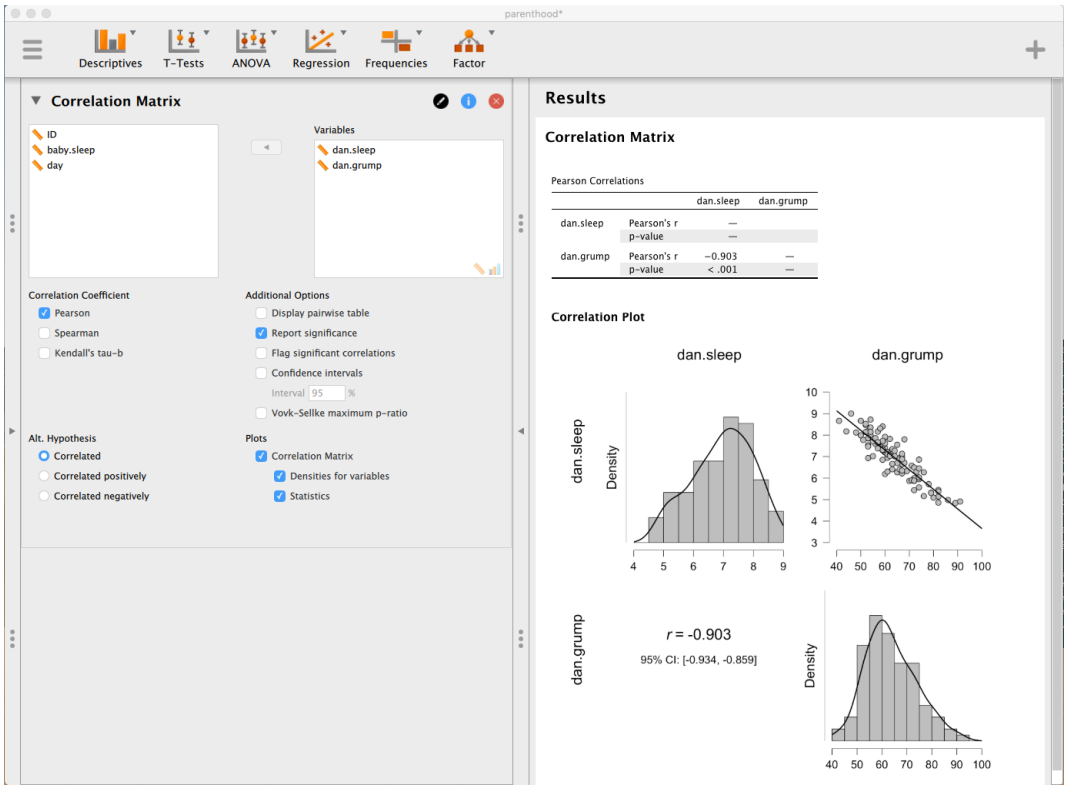


Figure1.8　JASP で '相関行列' から散布図をかく

ここでのゴールは，私の睡眠時間 (`dan.sleep`) と，次の日どれぐらい機嫌が悪いか (`dan.grump`) の全体的な関係を散布図で描くことです。JASP を使ってプロットを書くときは，'回帰分析'-'相関行列' の下にある 'プロット' オプションのボタンを押せばよく，図 fig:scatterplot1 が得られます。JASP は点の中に線を引きますが，これは後ほど，セクション (**??**) で説明します。こうした散布図を描くとき，'変数の記述統計量' も追加できて，変数がどのように分布しているかを示すヒストグラムや密度曲線を追加することができます。'統計量' オプションを押すこともできて，そうすると相関係数の推定値の 95% 信頼区間も得られます。

1.3

線形回帰モデルとは?

今から見る回帰モデルはとてもパワフルなツールなのですが，本質以外のものを削ぎ落としてみれば，線形回帰モデルは基本的にピアソンの相関 (セクション **??**) の少し凝ったバージョンに過ぎません。

会期の基本的なアイデアは相関と密接に結びついているので，`parenthood.csv` ファイルにもどって相関がどういうものだったかを見直してみましょう。思いおこせば，このデータセットでなぜダンがいつも不機嫌なのかを見つけようとしていて，仮説としては睡眠が足りてないのではないか，というものでした。散布図を描いて睡眠時間と翌日の不機嫌さの関係を検証しようとし，図 **??**で，この関係が相関係数 $r = -.90$ として表されることを見ましたが，私たちがこっそりイメージしていたのは図 **??**a のような何かだったのではないでしょうか。つまり，我々はデータの真ん中を通る直線を心の中に描いていたのです。統計学では，この線のことを**回帰直線**と呼んでいます。われわれは馬鹿ではないので，この回帰線はデータの真ん中を通っていることに注意してくださいね。図 **??**b にあるような，馬鹿っぽいプロットを想像することはないのです。

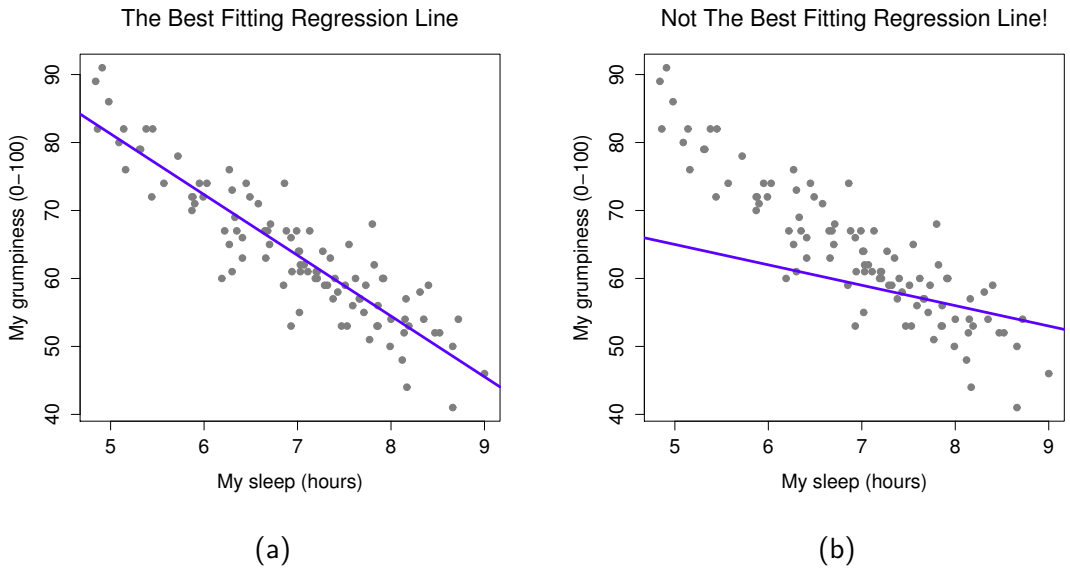


Figure1.9　パネル a が示すのは，図 **??** からきた睡眠と不機嫌さの散布図に，回帰直線を書き足したものです。驚くな彼，この線はデータの真ん中を通っているのです。それに比べて，パネル b が示すのは同じデータですが，回帰線を間違った書き方をしています。

これはそれほど驚くことではありません。図 **??**b に描いた直線はデータに“フィット”してないので，データを要約するという目的には役に立たないのです。これは非常にシンプルに見えますが，これに関する数学を少し見てみればとても強力なものであることがわかります。さてそのために，高校

数学を少し思い出してみましょう。直線は一般に次のように表現されるのでした。

$$y = a + bx$$

少なくとも，私が何年も前に高校生だった頃はこうだったはずです。2 つの *変数*は x と y で，2 つの *係数* a と b があります[*2]。係数 a は直線の y *切片*であり，係数 b は直線の *傾き*を表しています。高校時代の古ぼけた思い出を掘り返してみれば (すみません，私たちにとって随分前のことなのです)，切片は "$x = 0$ のときの y の値" として解釈できるのでしたね。同様に，傾き b はもし x の値が 1 単位増加したら，y の値が b 単位増加することを意味していますし，負の傾きであれば y の値が増加ではなく減少することを意味します。おうイエス，これで全て思い出しましたね。こうやって覚えていれば，回帰線も全く同じ式を使うのですから驚くことはありませんね。もし Y が結果変数 (従属変数) で,，X が予測変数 (独立変数) であれば，この回帰式は次のように描くことができます。

$$\hat{Y}_i = b_0 + b_1 X_i$$

ふむぅ。同じ数式に見えますが，ごくわずか，今回のバージョンは違うところがあります。これを理解してみましょう。まず，私は X_i と Y_i と書いています。ただの X,Y ではないですね。これはなぜかというと，私たちは実際のデータを扱っているからですね。この式では，X_i は予測変数の第 i 番目の観測を表しています (つまり私のこの研究では，第 i 日目に寝た時間の数字です)。そして私はそこまで明示的に伝えていませんでしたが，私はこの式がデータセット全ての点 (全ての i) について成立していると仮定しているのです。次に，Y_i ではなく $\hat{Y}_i$ になっているところに気づいたでしょう。これがどうしてかというと，*実際のデータである* Y_i と，*予測値の* $\hat{Y}_i$(つまり回帰線によって成される予測) とを区別したいからです。3 つ目に，係数を a,b から b_0,b_1 に書き換えました。これは回帰モデルにおける係数を表すのに統計学者が好むやり方だ，というに過ぎません。なぜ b が選ばれたのかわかりませんが，とにかくそうするようです。どんな場合でも b_0 は常に切片を表していて，b_1 は傾きを表しています。

大変結構。次に言わずにいられないのは，良い回帰線であろうと悪い回帰線であろうと，データが完全に直線にのることはないということです。つまり言い方を変えれば，データ Y_i は回帰モデルの予測である $\hat{Y}_i$ に一致しないということです。統計家は文字を追加し，あらゆるものに文字と数字を付与するのが好きなので，モデルの予測と実際のデータ点との差分を *残差*として，ε_i であらわします[*3]。数学を使うと，残差の定義は次のようになります。

$$\varepsilon_i = Y_i - \hat{Y}_i$$

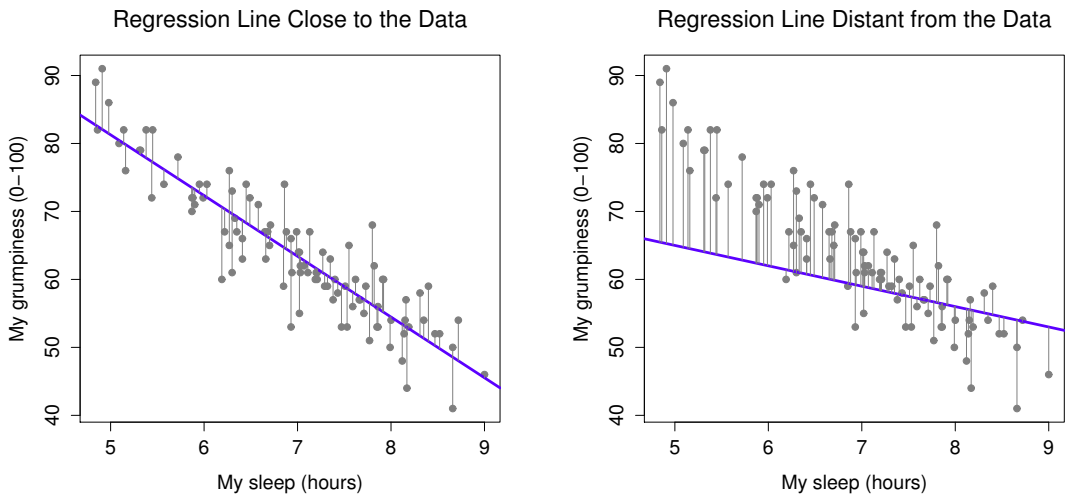
そうすると，完全な線形モデルは次のように書くことになります。

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

[*2] これは $y = mx + b$ とかかれることもあります。そのとき m が傾きの係数で b が切片 (定数) の係数です

[*3] ε はギリシア文字でイプシロンといいます。残差をあらわすのには，ε_i か e_i を使うのがならわしです。

1.4 回帰モデルの推定



(a) (b)

Figure1.10　最適なフィッティングをした回帰直線に伴う残差を描いたもの (パネル a) と，当てはまりの悪い回帰線に伴う残渣を描いたもの (パネル b)。残差は良い回帰線の方がかなり小さくなります。もう一度，驚くことではないですが，良い直線というのはデータの中心を通るものだということを確認しましょう。

オウケイ，図を見直してみましょう。でも今度は全ての観測度数に対して，残渣を表す線を追加してみます。もし回帰線が良いものであれば，残差 (黒い点線の長さ) は全て小さくなっていることが，図 **??**a に示されています。しかし回帰線が悪いものであれば，残差はいくぶん伸びてしまうことが図 **??**b に見て取れます。ふむぅ。私たちが “ほしい” 回帰モデルは，残差が *小*さいほうですよね。そう，そうでないとおかしいのです。実際，“ベストフィット” する回帰線は，残差を最小にするものだといえるでしょう。あるいは，統計学者はあらゆるものの二乗を取るのが好きなので，次のように言えばいいかもしれません。

> 推定された回帰係数，$\hat{b}_0$ と $\hat{b}_1$ は，残差の平方を最小にするもので，$\sum_i (Y_i - \hat{Y}_i)^2$ とか $\sum_i \epsilon_i{}^2$ と書くことができる。

そうそう，これで良くなりました。そしてこのように位置をずらして書いたので，おそらくこれが正解なんでしょう。そしてこれツェ異界なので，回帰係数は *推定値*だという事実を書いておくのは大

事なことでしょう (私たちは母集団のパラメータを推測しようとしていたのでした!)。だから，小さなハットをつけて，b_0 と b_1 ではなく $\hat{b}_0$ と $\hat{b}_1$ をつけましょう。最後にもう一つ言っておきたいことがあって，回帰モデルの推定する方法は 1 つだけではありませんから，この推定プロセスにより専門的な名前をつけておきます。**最小二乗法 (OLS)** です。

ここで，回帰係数 $\hat{b}_0$ and $\hat{b}_1$ の "ベストな" チョイスとは何かについて，はっきりと定義できるようになりました。当然次の質問は，もし最適な回帰係数というのが残差の平方を最小化するものだとしたら，どうやってその不思議な数字を *見つけたら*よいのかということです。この問いに応えるのは実はちょっと複雑です，回帰のロジックを理解する助けになりません[*4]。今回は見逃してあげましょう。長くて面倒な方法を最初に見せてから，JASP がやってくれる素晴らしいショートカットを "ご開帳する" ので h なく，近道をしていきなり JASP でこの面倒な仕事をやっつけちゃいましょう。

1.4.1 JASP による線形回帰

線形回帰を走らせるには，`parenthood.csv` データファイルを使って，JASP の '回帰'–'線形回帰' 分析と進めましょう。'従属変数' のところに `dan.grump` をして石，'共変量' ボックスのところに `dan.sleep` を持っていきます。図 **??**に示した結果より，切片 $\hat{b}_0 = 125.956$ と傾き $\hat{b}_1 = -8.937$ を得ます。言い換えると，図 **??** にプロットしたような，ベストフィットした回帰線の式は次のようなものだったわけです。

$$\hat{Y}_i = 125.956 + (-8.937 \ X_i)$$

1.4.2 推定されたモデルを解釈する

最も重要なことは，この機器係数を解釈する方法を理解することです。傾きの $\hat{b}_1$ から始めましょう。傾きの定義を覚えていれば，回帰係数 $\hat{b}_1 = -8.94$ が意味することは，もし X_i が 1 点上昇すれば，Y_i は 8.94 点さがるということですね。つまり，もう一時間余分に寝ることができれば，私の気分は改善し，不機嫌さが 8.94 ポイントさがるということです。切片はどうでしょうか? $\hat{b}_0$ は "X_i が 0 のときに期待される Y_i の値" ということなので，直感的にもわかりますね。もし私が一睡もしていなかったら ($X_i = 0$)，私の不機嫌さは尺度を飛び出して，$Y_i = 125.96$ という非常識な値になるということです。それは避けた方がいいと思いますねえ。

[*4] あるいは，少なくともほとんどの人にとって役に立たないと思っています。しかし万一，これを読んでいる人の中に，線形代数のカンフー師範がいれば (そして正直に言って，統計の導入クラスの中に数人はそういう人がいるのですが)，推定問題を解決するために次のように書き換えれば *その人の助けに*なるでしょう。$\hat{\boldsymbol{b}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{y}$, とするのです。ここで，$\hat{\boldsymbol{b}}$ は推定された回帰係数の入ったベクトル，X は "デザイン行列" で予測変数 (と，すべて 1 が入っている列を追加したもの。X は厳密に回帰するものの変数で，その違いについてはまだ触れてませんが) のベクトルで，$\boldsymbol{y}$ は結果変数を含んだベクトルです。他の人にとっては，これは役にも立たないし恐ろしいものに見えるかもしれません。しかし，線形回帰の一部は線形代数の用語で書かれているので，この章のこれのように脚注の中で見てもらえればと思います。もしこの数学がわかれば，たいしたものです。そうでなかったら，無視してくれたらいいです。

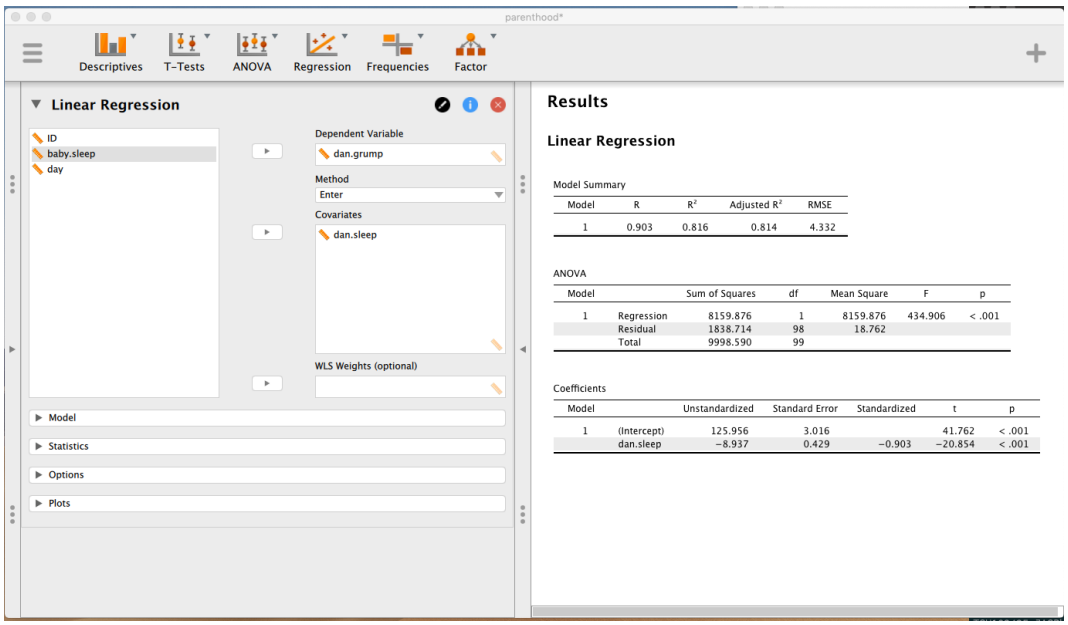


Figure1.11　単純な線形回帰分析をする JASP のスクリーンショット

1.5 重回帰分析

既に述べてきた単純な回帰分析は，興味関心ある単一の予測変数，例では `dan.sleep` があることを前提としていました。実際，これまで話してきた統計ツールは いずれも，分析にはひとつの予測変数とひとつの結果変数を使うものでした。しかし多くの (たぶんほとんどの) 研究では，複数の予測変数について検証したいと思うのでしょう。もしそうなら，線形回帰のフレームワークを複数の予測変数をもつものに拡張することができればいいと思いませんか。**重回帰**モデルがその目的に合うのでは，と思いません?

重回帰の考え方は非常に単純です。やるべきことは，回帰方程式に項を追加するだけです。興味がある変数が 2 つあるとしましょう。`dan.sleep` と `baby.sleep` の両方で，`dan.grump` 変数を説明したいとします。以前，Y_i で i-日目の私の不機嫌さを表現したのでした。今度は 2 つの X 変数を持っています。最初の変数は私の睡眠時間に，第二の変数は息子の睡眠時間に対応しています。さて X_{i1} を私の i-日目の睡眠時間，X_{i2} をその日の我が子の睡眠時間を表すものとしましょう。そうすると，重回帰モデルは次のように書くことができます。

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \epsilon_i$$

既に述べたように，ε_i は i-番目の観測における残差を表しており，$\varepsilon_i = Y_i - \hat{Y}_i$ です。このモデルでは，3 つの係数を推定しなければなりません。切片 b_0，私の睡眠時間に対応した係数 b_1，私の子の睡眠時間に対応した係数 b_2 です。しかし，推定しなければならない係数の個数が変わったとて，どう推定するかの基本的な考え方に変化はありません。推定すべき係数 $\hat{b}_0$, $\hat{b}_1$ そして $\hat{b}_2$ は，残差の平方を最小化するように求めれば良いのです。

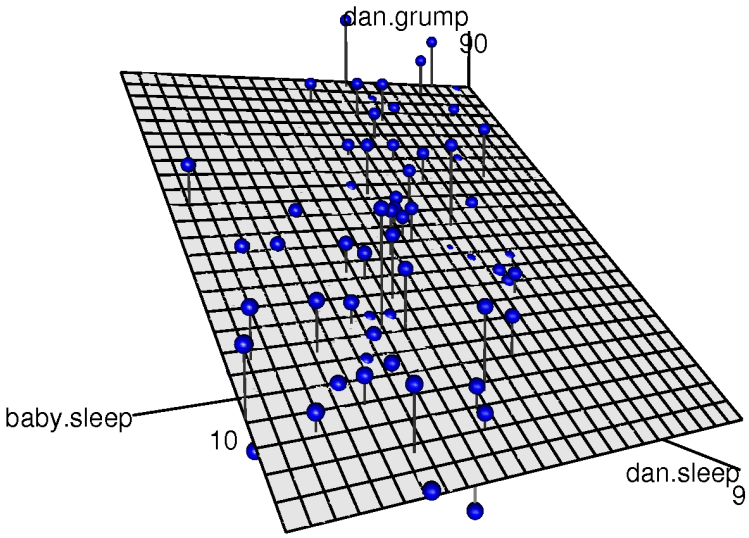


Figure1.12　重回帰モデルの 3 次元の可視化。モデルには 2 つの説明変数，`dan.sleep` と `baby.sleep` があり，結果変数は `dan.grump` です。3 つの変数を一緒にするので，3 次元空間になります。各データ (点) は空間の位置を示します。単回帰モデルを 2 次元空間に線で示したのと同じように，重回帰モデルでは 3 次元空間に平面で表すことになります。回帰係数を推定する時は，全ての青い点ができるだけ近くに来るような平面を探すことになります。

1.5.1 JASP でやってみよう

重回帰を JASP でやる方法は，単回帰と違いがありません。JASP の ‘共変量’ ボックスに，変数を追加するだけです。たとえば，dan.sleep と baby.sleep の両方を予測変数として，不機嫌さの説明につい書いたい時は，baby.sleep を dan.sleep の隣にある ‘共変量’ ボックスに持っていきます。JASP はデフォルトで切片の推定を含んでいます。今回得られる係数は次のようになります。

```
%（Intercept）    dan.sleep   baby.sleep
（切片）       ダンの睡眠時間  子の睡眠時間
  125.966         -8.950       0.011
```

dan.sleep にかかる係数はかなり大きいようで，睡眠時間が減っていくたびに不機嫌さが増えていくようです。一方，baby.sleep の係数はすごく小さいので，私の息子がどれぐらい寝るかというのは大した問題ではないようです。私の期限がどうなるかは，*私*がどれほど寝られるかにかかっているようです。重回帰モデルがどうなっているかの感覚を掴むために，図 **??**を描いてみました。3 次元プロットは 3 つの変数と重回帰モデルそのものを示しています。

1.5.2 式を一般化する

上で示した式は，2 つの予測変数を含んでいるときの重回帰モデルです。おどろくなかれ，もしやりたいのならもっと多くの X と，係数 B を追加してもいいのです。言い換えると，もし K 個の予測変数をモデルの中に入れたいのなら，回帰方程式は次のようになるでしょう。

$$Y_i = b_0 + \left(\sum_{k=1}^{K} b_k X_{ik} \right) + \varepsilon_i$$

1.6 回帰モデルの適合を評価する

線形回帰モデルの係数をどうやって推定するのかをみてきました。問題は，この回帰 g どれぐらい良いものなのかわからないということです。たとえば，セクション **??**で回帰モデルを作り，睡眠時間が長ければ私の気分が良くなると *主張*しましたが，これはまったく馬鹿げたことかもしれません。思い出してほしいのですが，回帰モデルは私の気分がどんなものかについての予測 $\hat{Y}_i$ を作りますが，それは実際の気分 Y_i ではないのです。もしこの 2 つがとても近ければ，回帰モデルはとても

良い仕事をしてくれるでしょう。もしそれがあ違うものであれば，役に立たないでしょう。

1.6.1 R^2 値

もう一度，これの数学的表現を見てみましょう。まず，残差の二乗和を次のようにしたのでした。

$$\text{SS}_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

これが小さければいいんですけどね。特に，結果変数の分散全体にたいして小さいことが望まれます。

$$\text{SS}_{tot} = \sum_i (Y_i - \bar{Y})^2$$

ではここで，て計算ではなく,' この値それぞれを計算してみましょう。JASP ファイル `parenthood_rsquared.jasp` を用意しましたので，これを本のデータフォルダから開いてください。データファイルには 5 つの変数があることに気づくでしょう。そのうち 2 つは元の dan.sleep と `dan.grump` 変数で，既に使ったものです。残りの 3 つは**計算された**変数です。

1. `Y.pred` は回帰方程式から予測された不機嫌さの値です。数式 '`125.97 + (-8.94 * dan.sleep)`' を使って計算されます。2. `resid` は残差，$\varepsilon_i = Y_i - \hat{Y}_i$ で，回帰式からの**予測された** 不機嫌さの値と，**実際の**不機嫌さの値の間の違いを表しています。数式 '`dan.grump - Y.pred`' で計算されます。3. `sq.resid` は残差の平方で，数式 '`resid2`' で計算されます。

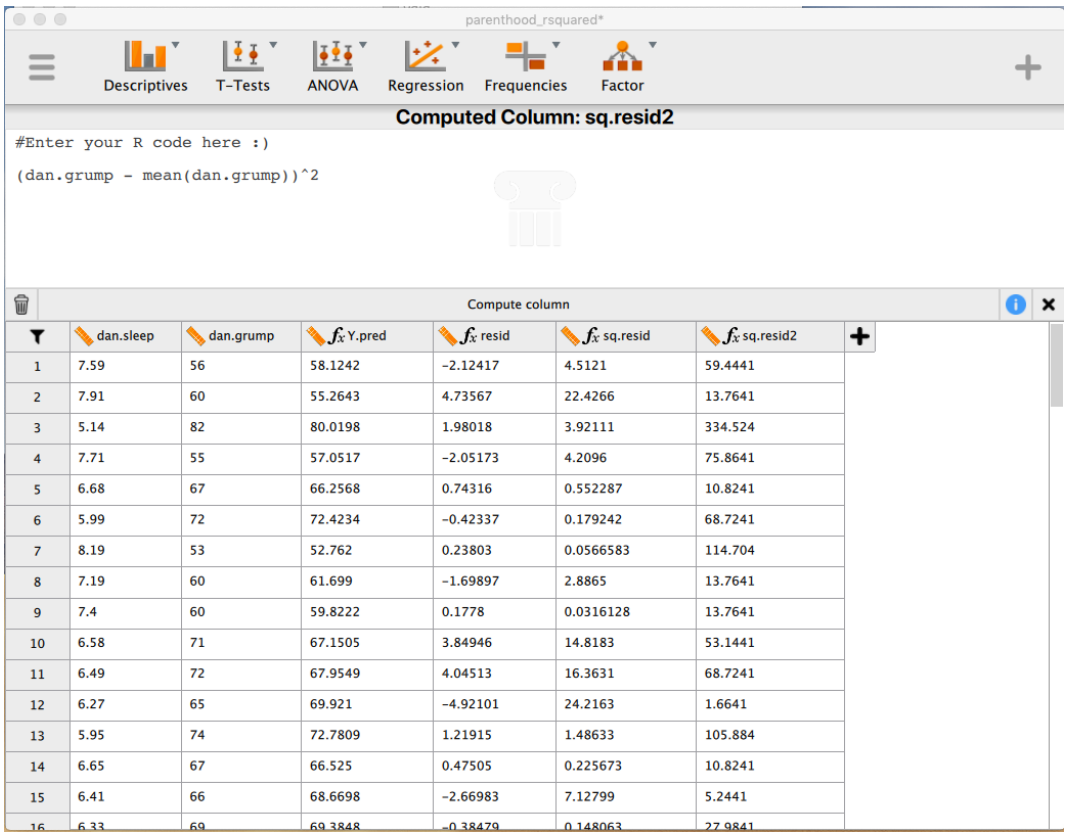
SS_{res} は残差の平方なので，JASP を使って `sq.resid` 列を探します。'記述統計'-'記述統計' をクリックし，`sq.resid` を '変数' ボックスに移します。次に '統計量' オプションから ' 合計' を選択します。すると '`1838.714`' という値を得ることができます。

すばらしぃ。大きな数字ですが，そこにそれほど意味はありません。どうせなら，二乗和を計算しちゃいましょう。これもかなりシンプルです。SS_{tot} を同様に計算してみます。今度は，新しい計算列を作る必要があります。"+" の記号を選ぶところから始めましょう。'変数名' のところに '`sq.resid2`' とタイプします。(なぜかはすぐにわかります)。"R" ボタンを選択して，"列を作る" をクリックします。R コードとして，次のように入力します (図**??**をみてください)。

```
(dan.grump - mean(dan.grump))^2
```

それから "列を計算する" をクリックします。これは残差の値を生成するのですが，その残差 (誤差) は**本当に良くない**予測モデルからのものです。すなわち，ここでのモデルは単に全ての不機嫌さ値の**平均**を使うものだからです。SS_{tot} にするには，上でやったのと同じように `sq.resid2` の平方和を計算する必要があります。

ここで '`9998.590`' という値が得られるでしょう。ふむぅ。さっきのよりさらに大きな数字になり



	dan.sleep	dan.grump	Y.pred	resid	sq.resid	sq.resid2
1	7.59	56	58.1242	−2.12417	4.5121	59.4441
2	7.91	60	55.2643	4.73567	22.4266	13.7641
3	5.14	82	80.0198	1.98018	3.92111	334.524
4	7.71	55	57.0517	−2.05173	4.2096	75.8641
5	6.68	67	66.2568	0.74316	0.552287	10.8241
6	5.99	72	72.4234	−0.42337	0.179242	68.7241
7	8.19	53	52.762	0.23803	0.0566583	114.704
8	7.19	60	61.699	−1.69897	2.8865	13.7641
9	7.4	60	59.8222	0.1778	0.0316128	13.7641
10	6.58	71	67.1505	3.84946	14.8183	53.1441
11	6.49	72	67.9549	4.04513	16.3631	68.7241
12	6.27	65	69.921	−4.92101	24.2163	1.6641
13	5.95	74	72.7809	1.21915	1.48633	105.884
14	6.65	67	66.525	0.47505	0.225673	10.8241
15	6.41	66	68.6698	−2.66983	7.12799	5.2441
16	6.33	69	69.3848	−0.38479	0.148063	27.9841

Figure1.13　R で残差を計算する JASP のスクリーンショット

ましたが，これは私たちの回帰モデルが良い予測をした (すなわち平均値だけを予測変数としたモデルよりも残差をぐっと減らした) ことを意味します。でもわかりにくいですね。

たぶんもう少し改良できます。ここでやりたいことは，2 つの意味をなさない数字から，1 つの数字に置き換えることです。素敵で意味のわかる 1 つの数字は，特に理由はないですが R^2 と呼ぶことにします。R^2 値が好ましい理由は，回帰モデルがデータを予測するときに誤差を全く生じないばあい，その値が 1 に等しくなるからです。言い換えると，そのとき残差は 0 になっているわけです。つまり $\mathrm{SS}_{res} = 0$ なら $R^2 = 1$ であることが期待できます。同様に，もしモデルが全く使い物にならないのであれば，R^2 は 0 に等しくなります。“使い物にならない” というのはどういうことでしょう? 回帰モデルが家からでて，髪を切って，仕事をして，ということを提案してくれるものならいいのですが，多分もう少し実用的な定義をしないといけないでしょう。今回の場合，残差の平方和が全平方和より小さくならない，つまり $\mathrm{SS}_{res} = \mathrm{SS}_{tot}$ ということです。まって，なんでそんなことが言えるの，って? それはこの式から，R^2 値を次のように書き直すことができるからです。

$$R^2 = 1 - \frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}}$$

and equally simple to calculate by hand: そして，簡単に次のように手計算できます。

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ &= 1 - \frac{1838.714}{9998.590} \\ &= 1 - 0.1839 \\ &= 0.8161. \end{aligned}$$

R^2 値は，ときどき**決定係数**と呼ばれることもあります[*5]。これは簡単に解釈できるのです。すなわち，彫られた予測変数で結果変数の分散を説明する *割合*を表しているのです。今回は，$R^2 = .8161$ が得られたわけですから，予測変数 (`dan.sleep`) は結果変数 (`dan.grump`) の分散の 81.61% を説明するのです。

当然ですが，回帰モデルから R^2 値がほしいときに，この計算をいつもやらなければならないということはないのです。JASP がデフォルトで計算してくれてますよ! 図 **??** をもう一度見てください。'モデルの要約' と書かれた表に R^2 が既にあるじゃないですか!

1.6.2 回帰と相関の関係

ここにきて先ほどの回帰の結果に戻ってみましょう。この非常にシンプルな式は，基本的には相関と同じものなのです。以前，r という記号を使ってピアソンの相関係数を表しました。相関係数 r と線形回帰からきた R^2 値の間にはなんらかの関係があるのではないでしょうか? もちろんです。相関係数の二乗である r^2 は，予測変数が 1 つだけのとき R^2 と一致します。言い換えると，ピアソンの相関係数は線形回帰モデルを 1 つの説明変数で使ったときとほぼ同じことなのです。

1.6.3 自由度調整済み R^2 値

最後にもう 1 つ指摘しておきましょう。モデルの評価指標としてレポートに書くときに多いのは，少し違う "自由度調整済み R^2" として知られているものです。この調整済み R^2 を使うモチベーションは，モデルに説明変数を追加すると *必ず* R^2 値が増加してしまう (あるいは少なくとも減ることはない) からです。

調整済み R^2 値は以下にあるように，少し計算式が違います。N 件のデータからなる K 個の予測

[*5] そして "ときどき" というのは，"ほとんどない" ということでもあります。実際には，みんなこれを "R 二乗値" といいます。

変数を持つ回帰モデルの場合，調整済み R^2 は次のようになります。

$$\text{adj. } R^2 = 1 - \left(\frac{\text{SS}_{res}}{\text{SS}_{tot}} \times \frac{N-1}{N-K-1} \right)$$

この調整は，自由度を計算に入れています。この調整済み R^2 をつかう大きな利点は，予測変数をモデルに追加することで偶然以上のパフォーマンスの向上が期待できるときにのみ，この調整済み R^2 が増加するようになっているところです。調整済み R^2 の大きな弱点は，R^2 の時のようなエレガントな解釈が できないということです。R^2 は回帰モデルによって，説明される結果変数の分散のうちモデルが説明する割合と解釈できるのでした。私の知る限り，調整済み R^2 にそれと同じように解釈できる筋道はありません。

R^2 と調整済み R^2 のどちらをレポートするべきか，という当然の質問が湧き上がってきますね。これはおそらく，個人的な好みだと思います。解釈できることを考えるなら，R^2 の方が良いでしょう。バイアスを補正したいのなら，調整済み R^2 がよいでしょう。私自身のことについていうなら，私は R^2 のほうが好みです。私の感覚的には，モデルの性能を解釈できる速度である方がより重要だと思うからです。それに，セクション **??**でこの後見るように，R^2 値をの改善が予測変数を追加したことだけによるもので，モデルを良いものにしたわけではないのではないかと心配するのであれば，仮説検定を使うことができます。

1.7 回帰モデルの仮説検定

回帰モデルがどんなものか，回帰係数をどのように推定するか，モデルのパフォーマンスをどう評価するかについて，説明してきました (最後の一つは，効果量の測定と同じです。)。次に話す必要があるのは，仮説検定です。ここでは 2 つの異なる (しかし関係のある) 仮説検定について話さなければなりません。すなわち，回帰モデルが全体的にヌルモデルと比べて意味のあるパフォーマンスを見せているかどうか，そしてある回帰係数が 0 から有意に異なっているかどうか，です。

1.7.1 モデル全体を検定する

ではすでに，回帰モデルの推定が終わっているとしましょう。最初の検定の仮説は，予測変数と結果変数の間に *なんの関係*もないという帰無仮説で，対立仮説は *データは回帰モデルが予測するように分布している*というものです。

フォーマルには，明らかに“回帰する”というモデルに対する“ヌルモデル”では，予測変数を一つも含んでおらず，切片 b_0 だけ含まれるというもので，以下のように表されます。

$$H_0 : Y_i = b_0 + \varepsilon_i$$

検証したい回帰モデルは K 個の予測変数を持ち，“対立モデル”は次のような重回帰モデルの式で表現されます。

$$H_1 : Y_i = b_0 + \left(\sum_{k=1}^{K} b_k X_{ik} \right) + \varepsilon_i$$

この２つの仮説を互いに戦わせるにはどうしたらいいでしょう？ そのひみつは，全分散 SS_{tot} を残差分散 SS_{res} と回帰モデルによる分散 SS_{mod} の和に分割できることにあります。技術的なところは後ほど，第 **??**章の ANOVA で見るので省略します。次のことだけ見ておいてください。

$$\mathrm{SS}_{mod} = \mathrm{SS}_{tot} - \mathrm{SS}_{res}$$

そして平方和を自由度で割ることで，平均平方に置き換えます。

$$\mathrm{MS}_{mod} = \frac{\mathrm{SS}_{mod}}{df_{mod}}$$

$$\mathrm{MS}_{res} = \frac{\mathrm{SS}_{res}}{df_{res}}$$

さて，自由度はいくらになるでしょう？ あなたが想像するように，モデルに伴う自由度 df は，予測変数の数に密接に関係します。事実，自由度は $df_{mod} = K$ であることがわかります。残差の全自由度は $df_{res} = N - K - 1$ です。

さて，平均平方を得たら次のように F 統計量を計算します。

$$F = \frac{\mathrm{MS}_{mod}}{\mathrm{MS}_{res}}$$

そしてこれに関する自由度は K と $N - K - 1$ になります。¥

F 統計量については第 **??**章で詳しく見ることになりますが，今はたんに F が大きくなれば帰無仮説が対立仮説に比べて弱くなることを示している，と思っていてくれれば結構です。JASP でこの検定をするためにはどうすれば良いか，お見せするのは簡単ですが，まずそれぞれの回帰係数の検定の方を見ておきましょう。

1.7.2 回帰係数の検定

ここで導入した F 検定は，モデル全体が偶然以上のパフォーマンスを示すかどうかをチェックす

るものでした。もしあなたの回帰モデルが F 検定で有意な結果にならなければ，とても良い回帰モデルとは言えない (あるいは，もしかするとですが，良いデータではなかったのかも) でしょう。しかし検定で有意にならなかったことがモデルに問題があることを強く主張していたとしても，*有意になったことが* (つまり帰無仮説を退けたことが) モデルが良かったことを示すわけではありません! なぜそうなるのか，不思議に思いません? 答えは，重回帰モデルの回帰係数を見ればわかります。私たちは既に，上のセクション**??** で，次のような回帰係数を得たのでした。

```
% (Intercept)    dan.sleep   baby.sleep
  切片          ダンの睡眠時間    赤ちゃんの睡眠時間
  125.966         -8.950       0.011
```

推定された回帰係数が，**赤ちゃんの睡眠時間** についてはわずか (0.011) で，**ダンの睡眠時間**についてはもう少し大きい (-8.950) ことがわかります。この 2 つの変数が絶対的に同じ尺度で測定されているなら (これらはいずれも “寝た時間” で測定されていますが)，これの意味がわかるわけです。現に，私は自分の機嫌の悪さを予測するためには，*私*がどれぐらい睡眠できたかということにのみ影響を受けていると推察しています。

ここで先ほど論じた，t 検定をもう一度使うことができます。この検定は，帰無仮説として真の回帰係数がゼロである $(b = 0)$ とし，それに対する対立仮説はゼロではない $(b \neq 0)$ とすることです。つまり，次のようになります。

$$H_0: \quad b = 0$$
$$H_1: \quad b \neq 0$$

どうやってこの検定をしましょうか? ふむぅ。中心極限定理は推定された回帰係数 $\hat{b}$ の標本分布が，平均 b を中心にした正規分布になっているかも，と教えてくれます。これが意味するところは，もし帰無仮説が真であれば，$\hat{b}$ の標本分布は平均がゼロで，標準偏差はわからないことになります。回帰係数の標準誤差についての良い推定値として，$\mathrm{SE}(\hat{b})$ があるとすれば，ラッキーですよね。これは*まさに*，第 **??**章でやった一標本の t-検定の状況です。ですから t 統計量を次のように定義できます。

$$t = \frac{\hat{b}}{\mathrm{SE}(\hat{b})}$$

理由は後回しにしますが，ここでの自由度は $df = N - K - 1$ になります。悔しいことに，回帰係数の標準誤差 $\mathrm{SE}(\hat{b})$ を推定するのは，第 **??**章でやった t-検定の平均値の標準誤差の計算と同じように簡単にいくわけではありません。実際この計算式は変な形をしていますし，見たところでなんの助けにもなりません[*6] 我々の目的としては，推定された回帰係数に伴う標準誤差は，予測変数と結果変数のどちらにも依存していることと，分散の均質性の仮定 (簡単に説明しましたね) が破られることに幾分敏感であることを指摘すれば十分です。ともかく，t 統計量は第 **??**章で説明した t 統計量と

[*6] より進んだ読者のために。残差ベクトルは $\boldsymbol{\epsilon} = \boldsymbol{y} - \mathsf{X}\hat{\boldsymbol{b}}$ になります。K 個の予測変数と切片があるので，推定された残差分散は $\hat{\sigma}^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}/(N - K - 1)$ になります。推定された係数の分散共分散行列は $\hat{\sigma}^2(\mathsf{X}'\mathsf{X})^{-1}$ で，これの対角項は $\mathrm{SE}(\hat{\boldsymbol{b}})$ であり，これが標準誤差になります。

同じように解釈できます。両側検定を仮定していると思いますから (つまり，$b > 0$ でも $b < 0$ でも気にしないとおもいますから)，t の値が極端であれば (すなわち，0 よりずっと小さいか，0 よりずっと大きいとき)，帰無仮説を棄却するべきだということになります。

1.7.3 仮設検定を JASP でやってみる

ここまで話してきた統計量全てを計算するときにあなたがやるべきことは，JASP で関係するオプションをチェックして回帰分析を実行するだけです。幸にして，これらのオプションは普通デフォルトで選ばれています。図 **??**にあるように，便利なアウトプットがいっぱい出てきます。

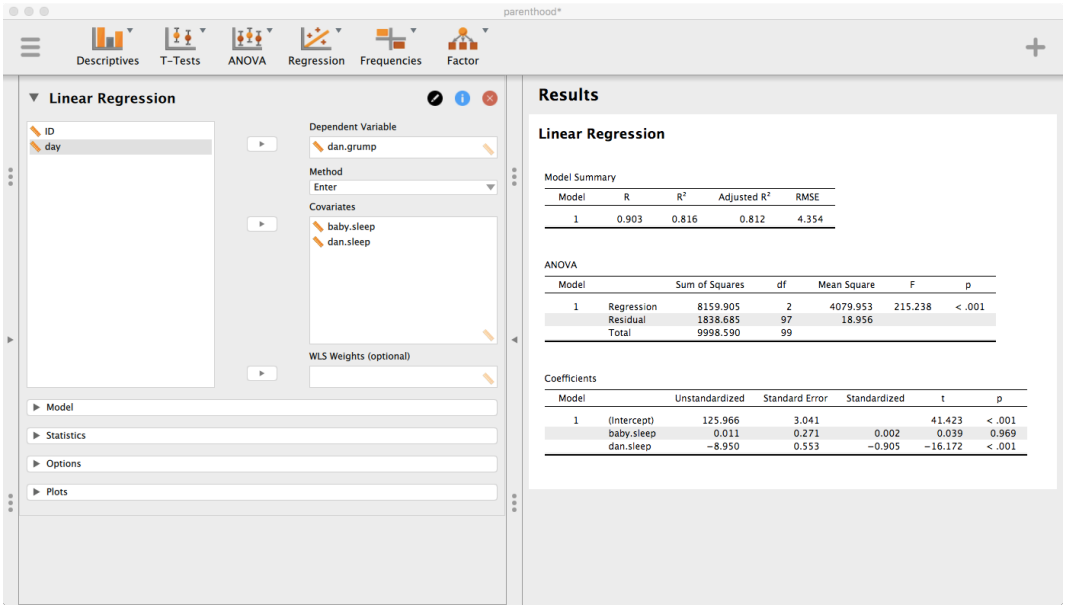


Figure1.14　重回帰分析と関係する仮設検定の JASP スクリーンショット

JASP の出力結果の下にある '係数' が図 fig:reg2 に示されていますが，これが回帰係数を示しています。この表の各行が回帰モデルの係数それぞれに対応しています。最初の行は切片で，そのあとに各予測変数の係数が続きます。列に関係する情報が示されています。最初の列 ('非標準化' というラベルがついています) は実際の係数 b です (切片は 125.966，**ダンの睡眠時間**という予測変数の係数は-8.950 です)。二列目は標準誤差 $\hat{\sigma}_b$ です。三列目は '標準化回帰係数' で，これについてはセクション**??**で説明します。四列目は t 統計量を提供していており，表中の数字はいつも $t = \hat{b}/\mathrm{SE}(\hat{b})$ となっ

ていることに注意しましょう。最後に，最終列がそれぞれの検定に対応する p 値を示しています。[*7]

係数表は t 検定でつかう自由度をリスト化してくれないのですが，それは常に $N - K - 1$ で，'ANOVA' とラベル化された出力のなかに示されています。この表からモデルが偶然以上に有意なパフォーマンスを示したことがわかりますし $(F(2, 97) = 215.238, p < .001)$，それは驚くどのことではありません。$R^2 = .816$ という値は回帰モデルが結果変数の分散の 81.6% を説明することを示していますから。しかし，各係数についての t 検定結果を見てみると，**赤ちゃんの睡眠時間** 変数が有意な影響を持っていなかったことがわかります。このモデルのなかでうまく機能していたのは，**ダンの睡眠時間** 変数でした。これを併せて考えると，この回帰モデルはデータに対して実際にはいいモデルではなかった，といえるでしょう。説明変数全体の中から，**赤ちゃんの睡眠時間** 変数を取り除いた方がよさそうです。言い換えると，単純な回帰モデルがよりよさそうだということです。

1.8 回帰係数について

線形回帰の前提となる仮定とそれに合致しているかどうかをチェックする方法について説明する前に，短い 2 つの議論をしておこうと思います。いずれも回帰係数に関係することです。最初に言わなければいけないのは，係数の信頼区間の計算です。そのあとで，どの予測変数が最も重要なものなのかをどう判断するかという，やや不明確な問題について議論します。

1.8.1 係数の信頼区間

母数のように，回帰係数 b も標本データから完全な正確さでも止められるものではありません。これは，なぜ検定が必要なのかという問いに対する答えの一部です。このとき，b の真の値についての不確実さを表現する信頼区間を報告できると，とても便利ですね。これは特に，研究の関心が X が *どれぐらいつよく* Y に影響しているかを見つけることにある場合は大事なことです。そうした時は，回帰の重み b に一番関心があるのですから。

幸い，回帰係数についての信頼区間は便利なやり方で計算できます。

$$\text{CI}(b) = \hat{b} \pm \left(t_{crit} \times \text{SE}(\hat{b}) \right)$$

[*7] JASP は多重検定をしていますが，いかなる多重比較の補正もやっていません。これは標準的な一標本 t 検定で両側検定になっています。多重検定の補正をしたいのなら，自分でやらなければなりません。

ここで $\mathrm{SE}(\hat{b})$ は回帰係数の標準誤差で，t_{crit} は t 分布に関する臨界値です。たとえば，95% 瀋陽区間が欲しいとしたら，臨界値は自由度 $N - K - 1$ の t 分布における 97.5 パーセンタイル点になります。つまり以前考えた信頼区間の計算アプローチと同じだということです。

JASP で信頼区間を表示させるには，回帰モデル画面から '統計' メニューから '信頼区間' を選択します。デフォルトでは 95% CI になっていますが，簡単に好きな値，例えば 99% に変えることができます。

1.8.2 標準化した回帰係数を計算する

あなたに必要になるかもしれないもうひとつのこととして，"標準化された" 回帰係数を計算したいことがあるかもしれません。これはよく β で表わされます。標準化係数の根拠は次のようなものになります。多くの場合，分析に用いる変数が異なる単位に基づいています。たとえば，もし私のモデルが教育期間 (何年教育を受けたか) と年収から IQ スコアを予測しようとしていたとしましょう。あきらかに教育期間と年収は同じ単位ではありません。学校に行ってる期間というのは 10 数年ですし，年収は 10,000 ドルぐらいの単位で変化します。測定の単位は回帰係数に強く影響します。係数 b は単位に応じて解釈できます。2 つの予測変数と，結果変数の両方について，です。しかし異なる予測変数間の係数を比較するのは難しいですね。違う係数間の比較がしたいという状況もあるのです。結果変数に最も強い影響を与える予測変数を知るための，標準化された測定が必要ですね。これこそ**標準化係数**がしようとしていることです。

基本的なアイデアは極めてシンプルです。標準化した係数は，回帰分析をする前に全ての変数を z-値にしたものを使うことで得られます[*8]。ここでのアイデアは，全ての予測変数を z 値に変換し，同じ尺度で回帰変数にすることで，異なる単位を持つ変数がもたらす問題を取り除くというものです。元の変数がどんなものであれ，β 値 1 つぶんは，予測変数が 1 標準偏差増加すると，それに応じて結果変数が 1 標準偏差分増加する問うことを意味します。つまり，もし変数 A の β の絶対値が B のそれより大きければ，A のほうが結果変数により強く関係しているということを意味します。少なくとも，そういうことです。ちょっと注意を促しておきますと，これは "1 標準差分の変化" は基本的に全ての変数において同じである，という仮定に深く関わっているものです。これが本当かどうかは，それほど明白なことではありません。

解釈の問題は横に置いて，どのように計算するかを見てみましょう。あなたができることは，全ての変数をあなた自身が標準化することで，それから回帰分析をすることになりますが，もっと簡単

[*8] 厳密には全ての *回帰変数*を，です。。つまり，モデルに含まれる回帰係数に関係するあらゆる "もの" が対象です。回帰モデルは，既に述べたように，各予測変数に対応する 1 つの回帰変数がありますし，その逆もまた然りです。しかしこれは一般的に正しいわけではなく，その例については第 **??**章で触れます。しかしこの区別について，今はそれほど気にしなくていいでしょう。

なやり方があります。結論から言うと，予測変数 X と結果変数 Y に関する β 係数は非常に単純な数式で，次のように表せます。

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

ここで σ_X は予測変数の標準偏差，σ_Y は結果変数 Y の標準偏差です。これは問題をグッと単純にしてくれますね。

同じぐらい単純なやり方ですが，JASP は β 係数をデフォルトで算出してくれますので，図 **??**の‘係数’についての表で第 3 の列を見てください。これがはっきり示しているように，**ダンの睡眠時間**変数は**赤ちゃんの睡眠時間**変数よりも強い影響を持っています。とはいえ，これは標準化係数 β よりも元の回帰係数 b を使った方がわかりやすい例でもあります。結局，私の睡眠時間と赤ちゃんの睡眠時間は *すでに*同じ尺度で測定されていたからです。眠った時間の長さですよね。これを z 値にすることで問題を複雑にする必要なんかあります?

1.9 回帰分析の仮定

線形回帰モデルはいくつかの仮定の上に成り立っています。セクション **??**では，この仮定に合致しているかどうかをチェックする方法について説明しますが，まずはそれらをみておきましょう。

- *正規性の仮定*。統計における多くのモデルと同じように，単回帰であれ重回帰であれ，正規性の仮定お上に成立しています。特にその *残差*が正規分布していることを仮定しています。予測変数 X と結果変数 Y が正規分布に従っていなくてもいいのですが，残差 ε は正規分布に従っていなければなりません。これはセクション **??**をみてください。
- *線形性*。線形回帰モデルの基本的な仮定ですが，変数 X と Y の間には線形関係がなければなりません! 単回帰であれ重回帰であれ，その関係は線形でなければならないのです。
- *分散の均一性*。厳密に言うなら，回帰モデルは各残差 ε_i が平均 0 の正規分布，そして (ここでの目的では特に重要なことですが) ひとつひとつの残差について同じ標準偏差 σ の正規分布であることが亜重要です。実際には，各残差が同じ分布からきているかどうかを検証するのは不可能です。その代わり，残渣の標準偏差が全ての $\hat{Y}$ について同じであるかをみることはできますし，(強いて言えば) モデルにおける全ての予測変数 X についてもそうか，と考えることはできます。
- *予測変数が相関しないこと*。重回帰分析においては，説明変数同士があまりにも強く相関しないようにすることが重要です。これは重回帰分析の “理論的な” 仮定ではありませんが，実際には求められることです。予測変数が互いに強く相関すると (これは “共線性” と呼ばれます)，モデルの評価の時に影響が出ます。

- *残差が互いに独立していること*。これは“包括的な”仮定であり，“残差に何も面白いことは起こっていない”という効果のことです。もし何かおかしなことがあれば (例えば，全ての残差が観測されてない他の変数に強く依存しているなど)，物事を歪めてしまうのです。
- *“悪さをする”外れ値がないこと*。改めて，これもモデルの理論的な仮定ではありませんが (あるいはむしろ，他の全てに共通する種類の話かもしれませんが)，回帰モデルが 1 つ 2 つの異常なデータ点に強く影響されているというようなことがないようにしなければなりません。なぜなら，モデルの適切さに関わりますし，いくつかのケースがもつデータの信憑性に関わるからです。これについてはセクション **??**を参照してください。

1.10 モデルのチェック

このセクションの主な狙いは，**回帰分析の診断**で，回帰モデルの仮定が満たされている顔チェックしたり，もしその過程が満たされていないのであればどのようにモデルを修正すれば良いか，なにも“おもしろいこと”が起きていないことをチェックするにはどうすれば良いか，という職人芸を考えることです。ここでモデルチェックの“職人芸”といったのには理由があります。これは簡単なことではなくて，診断に利用できる標準的なツールはたくさんあるし，あなたのモデルの問題点 (もし何かあれば，と言うことですが) を修正してくれたりもするんですが，これらを実行する時にかなりの判定をすることになるので，あなたが本当にやらなければならないことはその練習なのです。微に入り細に入って詳細を検証すると迷子になっちゃいますし，細々としたこと全てを覚えておこうとすると疲れてしまいます。ツールの *全てを学ぼうとすると*，非常にフラストレーションが貯まるという副作用がありますので，*あらゆるモデルチェックを対象にするのは諦めよう*，となるのです。これはちょっとまずいですねえ!

このセクションでは，あなたの回帰モデルが想定していたことをチェックするための，いくつかの異なる方法を提示します。あなたがすべきこと全てをカバーするものではありませんが，多くの人が実践することよりも多少くわしく解説します。本来導入のクラスでこの全てをやることはないんですけどね。ですが，どんなツールが使えるのかを知ってもらうことは重要だと思いますので，そのいくつかを紹介したいと思います。最後に，**Fox2011** にあるこのセクションはもっとヘヴィで，その本では R をつかって回帰分析を実行するためにつかう `car` パッケージを用いて説明されていたことを指摘しておきましょう。`car` パッケージは，回帰分析の診断に優れたいくつかのツールを提供するのですが，その本を見れば見事いわかりやすく解説してくれています。あまり大それたことは言いたくないのですが，JASP にはなくて R にたけあるより優れた診断技術を見るためだけでも，**Fox2011** を読む価値があると思いますよ。

1.10.1 3 種類の残差

回帰分析の診断の主なものは，残差を見ることで行われます。ここまでの話で，あなたは頭の中で悲観的な統計学理論を作り上げ，残差をしっかり見ないといけないとうこと ならば考えないといけない残差にはいくつかの種類があるのでは，という推測ができるようになっているでしょう。このセクションでは以下の 3 種類の残差について考えます。“普通の残差”，“標準化された残差”，そして “Student 化された残差” です。図にあるように，第 4 の “ピアソン残差” というのもあります。しかしこの章で考えるモデルについては，ピアソン残差は普通の残差とおなじものになります。

第一の，最も単純な残差が**普通の残差**です。これは実際に生の残差で，この章でこれまで話してきたものと同じです。普通の残差とは，単に予測値 $\hat{Y}_i$ と観測値 Y_i の差分のことです。第 i-番目の普通の残差を示すために，これまで ε_i という表記を使ってきましたが，これからもそうしていきたいと思います。これを覚えておいていただいて，ごく単純な方程式から，

$$\varepsilon_i = Y_i - \hat{Y}_i$$

とすることができます。もちろんこれは以前見たことがあると思いますし，特に残差の種類に言及しない限り，残差といえばこれのことです。ここに目新しいものはありません。ただ繰り返しただけです。普通の残差の欠点は，結果変数が何であるか，回帰モデルがどれほど良いものであったかに依存した，毎回異なる単位で現れることです。つあり，切片なしで回帰モデルを実行すると，普通の残差の平均は 0 ですが，分散は毎回の回帰モデルごとに変わってしまうわけです。たいていの文脈では，特に残差の *パターン*にのみ興味があってその実際の値には興味がないという場合は，**標準化された残差**を考えてょ卯が便利です。それは標準偏差が 1 になるように標準化されているのですから。

> これを計算するには，普通の残差をこれらの残差から推定された (母集団の) 標準偏差で割ることになります。技術的な理由から，かくがくしかじかあって，次のような形になります。
>
> $$\varepsilon'_i = \frac{\varepsilon_i}{\hat{\sigma}\sqrt{1-h_i}}$$
>
> ここで $\hat{\sigma}$ は推定された母集団における普通の残差の標準偏差で，h_i は i 番目の観測における “推定値” です。推定値とは何であるかについては説明していませんが[a]，それほど大きな意味はありません。今のところは，普通の残差を z スコアに変換するための標準化された残差だ，と思っておいていただければ十分です。事実，それ以上でもそれ以下でもなくて，ちょっとファンシーになったぐらいのものですよね。
>
> [a] あるいは望まれないかもしれませんが，その機会があるかも

3 番目の残差は **Student 化された残差** (別名 “ジャックナイフ残差” です) であり，標準化された残差よりもファンシーな感じです。ここでのアイデアは，あらためて，普通の残差をある量で割ったものになります。ねらいは残差のある標準化された推定値にすることです。

この計算をするときの式は，ちょっと異なっていて次のようになります。

$$\varepsilon_i^* = \frac{\varepsilon_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$$

ここでの推定された標準偏差が $\hat{\sigma}_{(-i)}$ になっていることに注意してください。これが意味することは，第 i 番目の観測がデータから削除された時に，あなたが *手に入れるであろう*残差の標準偏差を推定するということです。悪夢のような計算をしているように思えるかもしれません。だって，N 回の新しい回帰分析をやれと言われてるようにみえるからです (最近の計算機でも，大きなデータセットであれば少し文句を言うかもしれません)。ところが幸い，とにかく頭のいい人たちが，この標準偏差は実際次のような計算で得られることを示してくれました。

$$\hat{\sigma}_{(-i)} = \hat{\sigma}\sqrt{\frac{N-K-1-\varepsilon_i'^2}{N-K-2}}$$

これなら軽いもんでしょ?

動く前に，この残差が必要になることはそんなにないこと，ほとんど全ての会期診断の中心にあったとしてもね，ということは伝えておきたいと思います。ほとんどの場合，さまざまな診断や仮定のチェックが提供され，これらの計算の面倒を見てくれるでしょう。とはいえ，標準的でない手法が必要になるかもしれないことを考えて，実際にどう計算するかを知っていることはいつだっていいことではあります。

1.10.2 3 種類の異常値

線形回帰分析をする時に危険な問題の 1 つは，ごく少ない “普通でない”，あるいは “異常な” 観測値のせいでかなり敏感になってしまっていることです。これについては以前，セクション **??** の外れ値の文脈で議論しました。そこでは ‘探索’-‘記述統計’ のところにあるボックスプロットのオプションを使って，自動的に特定されるものでしたが，今回もう少し正確に話を進めたいと思います。線形回帰の文脈では，“異常値” と呼ばれるものは，概念上 3 つに区別されます。いずれも興味深いものですが，分析においては異なる意味合いをもたらすものです。

最初の異常な観測値は**外れ値**です。外れ値の定義は (この文脈では)，回帰モデルの予測するものから大きく外れる観測値ということです。その例を図 **??**に示しました。実際には，外れ値はかなり大きな標準化された残差，ε_i^* を持っているものとするとして，この概念を操作化できます。外れ値は興味深いものです。大きな外れ値はデータのゴミ *かもしれません*。つまりデータセットで正しく記録されなかったのかもしれないし，他の欠陥が検出される可能性があります。外れ値だからと言う理由で，この観測データを取り除くべきではありません。これが外れ値であるという事実は，そのケースをじっくり見て，なぜ違いが生じたのかを見つける手がかりでもあるのです。

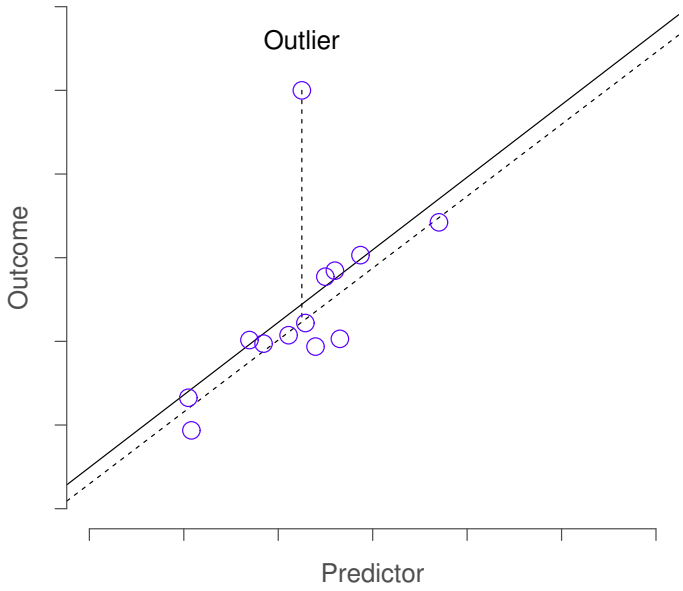


Figure1.15　外れ値の図。点線が外れ値を含まない時の回帰線とそれに対応する残差 (標準化された残差) です。実践は異常値も含んだ回帰直線。外れ値は結果変数 (y 軸上) における普通でない値であって，予測変数 (x 軸上) のものではなく，回帰線から大きく離れているものです。

普通じゃない観測値をみつける 2 つめの方法は，それが高い**レバレッジ**をもっているかどうかです。これは他のすべての観測値からぐっと異なる観測が生じた時に生じます。これは必ずしも大きな残差を伴って生じるものではありません。他のすべての観測値からぐっと異なる，ということを正確に言いかえるなら，それは回帰直線にはとても近いところにあるのに，といえます。この例を図 **??** に示しました。観測値のレバレッジとは，その *ハット値*の言葉で操作的に定義され，h_i と洗わされます。ハット値の式は，少し複雑ではあります[*9]が，その解釈はそれほど複雑ではありません。h_i は i-番目の観測値が回帰線がぐっと伸びた時の行き着く先を “制御” しているものなのです。

一般に，ある変数が，予測変数からなる項の他の変数から遠く離れているなら，大きなハット値をもっていることになります (非常に大雑把に説明するなら，高いレバレッジは平均の 2-3 倍以上と

[*9]あらためて，線形代数マニア向けのせつめいです。“予測値行列” が行列 H で表すとすると，観測値ベクトル $\boldsymbol{y}$ を予測値ベクトル $\hat{\boldsymbol{y}}$ に変換する,$\hat{\boldsymbol{y}} = \mathrm{H}\boldsymbol{y}$ として定義されます。この名前は “$bmy にハットをかぶせる” 行列であることから来ています。第 i-番目のハット *値*は，行列の i-番目の対角要素です (技術的なことですが，これは h_i より h_{ii} で表すべきことです)。オウ，これをどうやって計算するか瓦解になるところですかね。こうです。:$\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$. あらかわいい。ですよね？

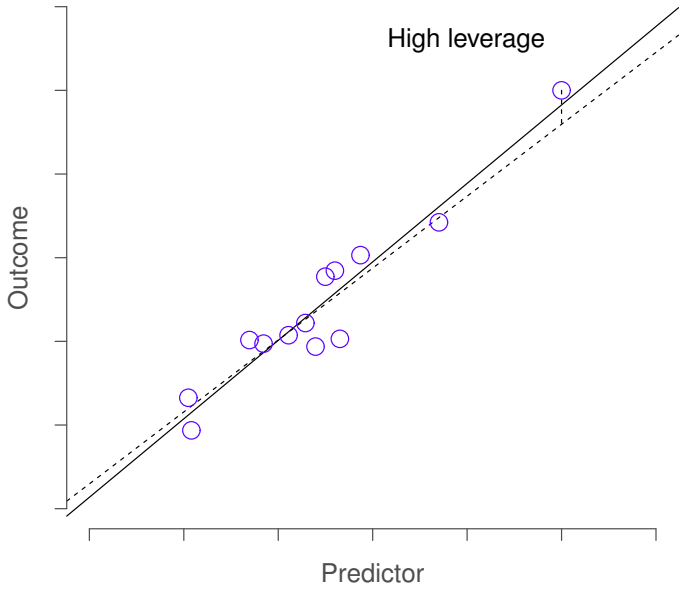


Figure1.16　高いレバレッジ点の図。この時の異常値は，予測変数 (x 軸) と結果変数 (y 軸) のどちらにおいても普通じゃない値ですが，この異常さはその他の観測値間に見られる相関パターンとかなり一貫したものです。観測値は回帰線の近くにありますし，それを邪魔するものではありません。

いうことです。そしてハット値の総和は $K + 1$ に制限されています)。高いレバレッジ点についてはもっと細かく見る価値がありますが，それが異常値でない限り，懸念材料になる可能性は低いでしょう。

ここで第 3 の普通じゃなさが出てきます。観測の**影響度**です。これは高いレバレッジ点を持っている外れ値のことです。すなわち，ある面では他のすべての点から大きく外れており，また回帰線からも遠く離れているのです。この状況を図 **??**に示しました。以前の 2 つの図と比較してみてください。外れ値は回帰線からそれほど多く離れていませんし，それほど高いレバレッジも持っていません。しかし外れ値でもあり高いレバレッジも持っているようであれば，回帰直線に大きな影響を与えます。なぜこの点が高い影響力を持っているか，そしてなぜそれが大きな問題になるかということの理由がこれです。影響度の測度として知られているのは，**Cook の距離**です。

$$D_i = \frac{\varepsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_i}$$

これは外れ値の測度 (左の項) とレバレッジの測度 (右の項) をかけたものであることに注意してください。

Cook の距離が大きくなるのは，実際かなり外れ値の状態で，かつ大きなレバレッジを持っている時，ということになります。大まかに言うと，Cook の距離が 1 以上であれば大きいと言えるでしょう (私がよく使う手っ取り早いルールです)。

JASP では，Cook の距離についての情報は '統計量' メニューの下にある 'ケースごとの診断' をクリックすることで計算できます。このデータを可視化するには 2 つの方法があります。まず，それぞれのケースについて (つまり，データの行ごとに)Cook の距離を見るために，'全て' を選択します。すると図 **??**のようになります。あるいは，Cook の距離がある閾値を超えた時**だけ**表示させることもできます。JASP のデフォルトの閾値は 1 です。どちらの場合も，この閾値を超えるデータのケー

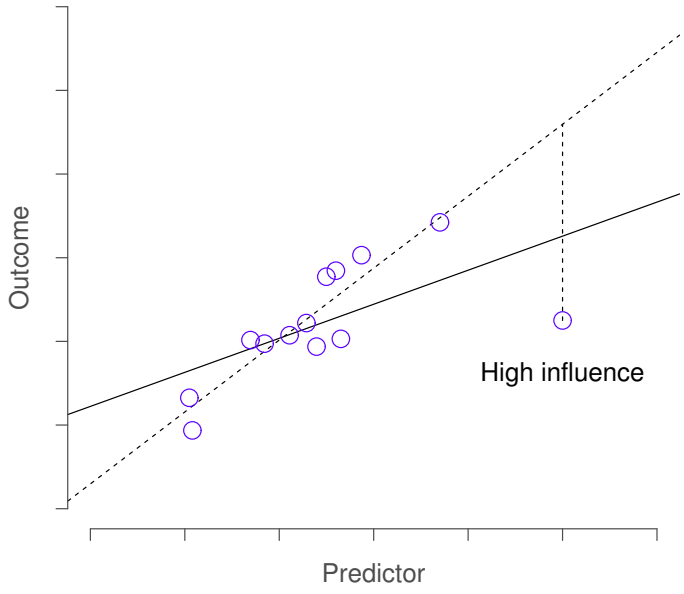


Figure1.17　高い影響度を持つ点の図。この場合は，異常値は予測変数 (x 軸) においてかなり異常亜値で，回帰線から遠く離れてしまっています。結果的に，回帰線はかなりそれに影響されてしまいます。結果変数 (y 軸) でみてみれば，全体的に典型的な観測点なのですが。

スがないことがわかりますね。

Casewise Diagnostics

Case Number	Std. Residual	dan.grump	Predicted Value	Residual	Cook's Distance
1	−0.497	56.000	58.140	−2.140	0.002
2	1.104	60.000	55.292	4.708	0.017
3	0.464	82.000	80.045	1.955	0.005
4	−0.477	55.000	57.060	−2.060	0.001
5	0.168	67.000	66.281	0.719	0.000
6	−0.095	72.000	72.407	−0.407	0.000
7	0.053	53.000	52.773	0.227	0.000
8	−0.393	60.000	61.700	−1.700	0.001
9	0.047	60.000	59.797	0.203	0.000
10	0.890	71.000	67.148	3.852	0.003
11	0.959	72.000	68.001	3.999	0.027
12	−1.139	65.000	69.912	−4.912	0.008

Figure1.18　JASP の出力で各ケース/データの行ごとに Cook の距離を表示する

次に出てくる質問はもし大きな Cook の距離が得られたらどうするべきか，ということですね。普通，これに対してバッチリ素早く適用できるルールはありません。おそらく最初にするべきことは，Cook の距離が最大のケース[*10]を除外して回帰分析を実行し，モデルパフォーマンスと回帰係数がどうなるかを見ることです。もし大きな違いがあるのなら，データセットやあなたが研究中に書き散らしたであろうノートを掘り返して考え直すべき時です。そして *なぜ*こんなに違いが出たのかを明らかにしましょう。もしあるデータ点が結果を歪めていると言うことが明らかになったら，それを除外することを考えてもいいですが，なぜこのケースが特に質的に他と異なるのか，別に取り分けて分析しなければならないかをしっかりと説明しなければ，良いやり方とはいえないでしょう。

1.10.3　残差の正規性をチェックする

本書で議論してきた多くの統計ツールと同じように，回帰モデルも正規性の仮定に依存しています。今回の場合は，残差が正規分布に従っていると仮定したことになります。最初にすべきことは，JASP で QQ プロットを書くことで，‘プロット’-‘Q-Q プロット’ から標準化された残差のオプションを選んでください。出力は図 **??**に示した通りで，標準化された残差が回帰分析の理論的な値に対応した関数としてプロットされています。

チェックすべきもう 1 つのことは，予測値と残差の関係をみることです。JASP でこれをするには，さまざまな ‘残差プロット’ をえらぶことで，それぞれ予測変数，結果変数，予測値と残差の関係を散布図にしたものを提供してくれます。図**??**を参照してください。これらのプロットでは，‘点’ の一様分布が得られており，‘点’ の塊やパターンがはっきり見えないことがわかります。これらのプ

[*10] これを簡単に JASP で実行する方法はいまのところないので，R のもつ `car` パッケージのような，よりパワフルな回帰プログラムから，より発展的な回帰分析を実行するといいでしょう。

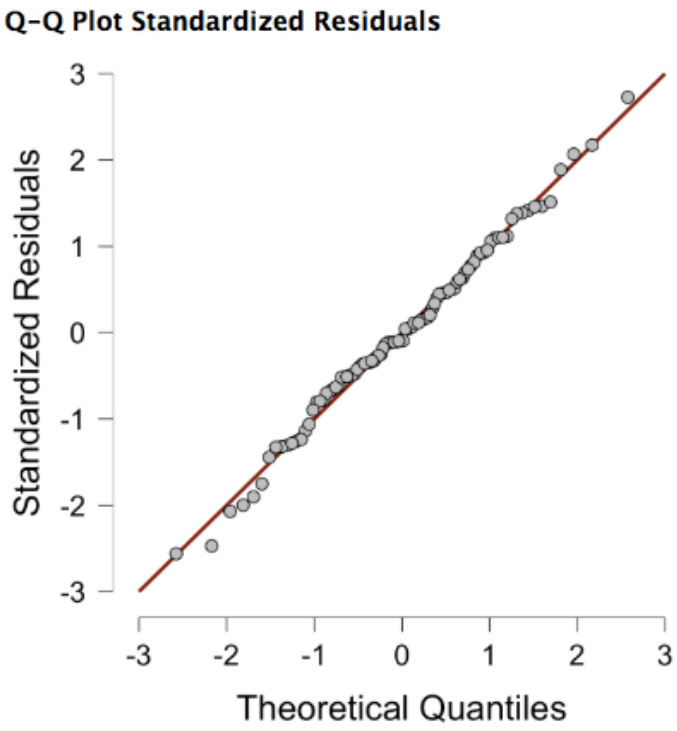


Figure1.19　モデルの理論的な値に対して標準化された残差の量を JASP でプロットしたもの。

ロットを見ると，プロット全体に点が均一に広がっていることから，特に心配すべきことはないように思います。プロット (b) ではやや一様分布でない傾向が見られますが，強い逸脱でもないですしそれほど心配することではないでしょう。

もし心配なら，この問題 (およびその他の多くの問題) に対する解決策のひとつは，一つ以上の変数を変換することです。変換はこのテキストの範囲を超えていますが。

1.11 モデル選択

残っている大きな問題の 1 つが，“モデル選択” です。すなわち，データセットの中に複数の変数がある場合，どれを説明変数にしてどれを含めないのがいいでしょうか。言い換えるなら，**変数選択**の問題です。一般的に，選択は複雑な問題ですが，モデルに含むべき変数のサブセットを選択すると言う問題に限れば，すこしシンプルになります。とはいえ，この限定的なトピックでもそこまで詳細に語るつもりはありません。そのかわり，これ居ついて考えるための 2 つの広範囲にわたる原則，そ

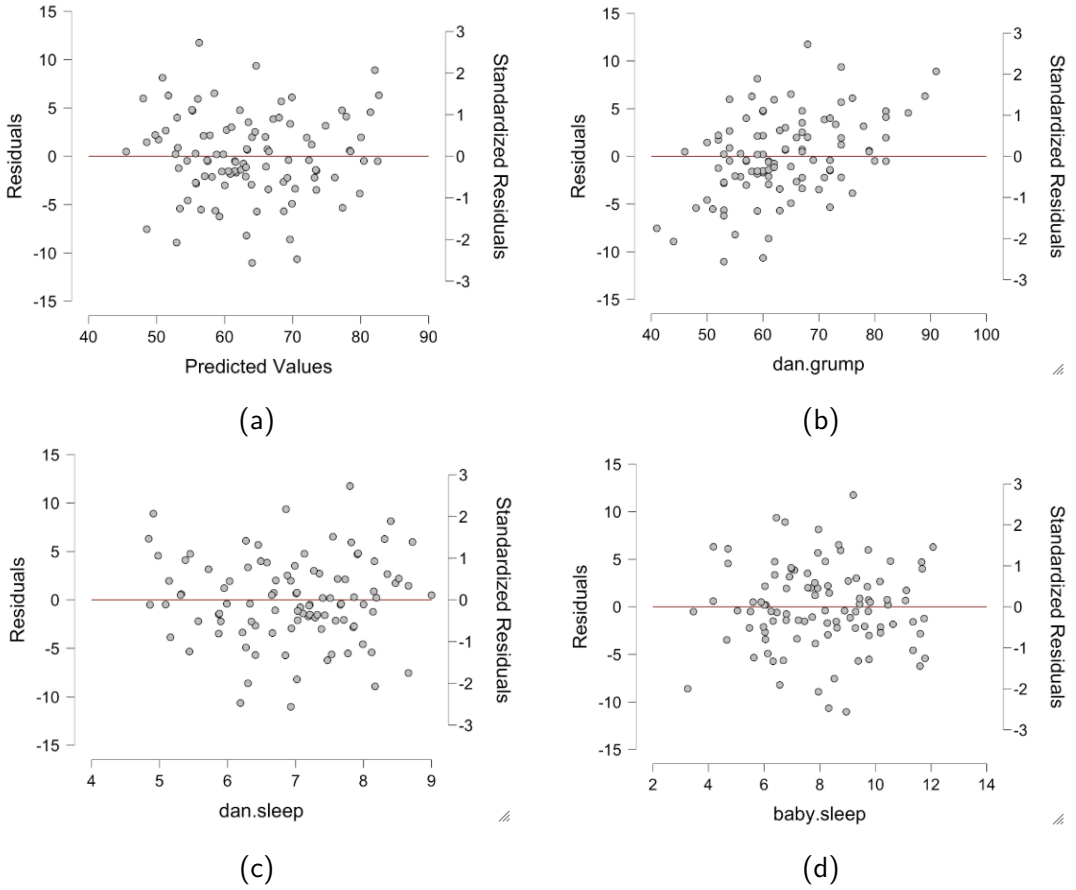


Figure1.20　Residuals plots produced in JASP

してあなたのモデルの中に含めるべき変数のサブセットを選択する助けになるような，あるしっかりしたツールについての議論を展開したいと思います。まず，2 つの原則についてみてみましょう。

- 自分の選択に実質的な根拠がるのはいいことです。つまり，多くの場合，研究者たるあなたはあり得る回帰モデルの中で理論的に興味のある少数のモデルを取り上げる理由があるはずです。そのモデルはあなたの研究領域において，意味のある解釈ができるものでしょう。このことの重要性を低く見積もってはいけません。統計家は化学的な問題の手助けをするものであって，その逆ではないのです。
- あなたの選択が統計的推測に関するものである場合，単純化と適合度との間にトレードオフがあります。多くの予測変数をモデルに追加すると，モデルは複雑になります。各予測変数は新しい自由パラメータ (つまり新しい回帰係数) を追加することになり，新しいパラメータはそれぞれモデルが確立変数を “吸い取る” 容量を増やすことになります。結果的に，あなたがどんな変数を追加しようとも，適合度 (たとえば R^2) は時には明らかに，時には偶然に増加し

ていきます。もしあなたのモデルを新しいデータに対して一般化したいのであれば，あまりにも変数が多くなるのは避けるべきなのです。

この後ろの原則は，**オッカムの剃刀**と呼ばれることがあり，次のようなことわざで言い表されます。*必要以上に中身を増やすな*。これの意味は，R^2 をぶち上げるためだけにあんまり関係のない予測変数を入れるのはやめましょう，ということです。ふむぅ。んー，もとの方が良さそうですね。

ともあれ，我々が必要としているのは数学的な基準で，回帰分析における変数選択という文脈におけるオッカムの剃刀を実現するための量的な原則です。いくつかのやり方があることがわかっています。そのうちの 1 つは，**赤池情報量規準** (**Akaike1974**) というものです。これは現在，JASP の標準的な出力にはなっていませんが，JASP がモデルのデータフィットを計算する手続きの中でごく簡単に計算されます。

In the context of a linear regression model, the AIC for a model that has n observations and K predictor variables (not including the intercept) can be computed[a] as

$$AIC = n \ln(SS_{res}) + 2K$$

[a]Strictly speaking, this formula is not completely correct. Akaike's original definition was in terms of something called a *maximum likelihood estimate* for the model, and as such, there are some other terms that appear in the computation. However, many of them don't depend on the model, and given that the purpose of the AIC is to *compare models*, these terms will be present in all models and will mathematically "wash out". Thus, I am presenting a 'bare bones' version of the formula that is sufficient for our purposes.

Here's the basic principle behind using AIC for model comparison: the smaller the AIC value, the better the model performance. If we ignore the low level details it's fairly obvious what the AIC does. On the left we have a term that decreases as the model predictions get better; on the right we have a term that increases as the model complexity increases. The best model is the one that both fits the data well (small SS_{res}, left hand side) and uses as few predictors as possible (small K, right hand side). In short, this is a simple mathematical implementation of Ockham's razor.

Let's demonstrate how AIC can be used to compare the two regression models we have computed in this chapter. Consider first the regression model with only one predictor: `dan.sleep` (see Figure **??**). In this model, we have $n = 100$ observations, $K = 1$ predictor, and $SS_{res} = 1838.714$. Thus,

$$\begin{aligned} AIC_1 &= n \ln(SS_{res}) + 2K \\ &= 100 \ln(1838.714) + 2(1) \\ &= 753.68 \end{aligned}$$

Now consider the second model using two predictors: `dan.sleep` and `baby.sleep` (see Figure **??**). In this model, we have $n = 100$ observations, $K = 2$ predictors, and $SS_{res} = 1838.685$. This gives us

$$\begin{aligned} AIC_2 &= n \ln(SS_{res}) + 2K \\ &= 100 \ln(1838.685) + 2(2) \\ &= 755.68. \end{aligned}$$

Since $AIC_1 < AIC_2$, this tells us that Model 1 is the better fit, which confirms our intuitions. Adding `baby.sleep` doesn't add much to the model fit, but it increases model complexity. AIC balances these two requirements; the penalty for adding an additional parameter is not outweighed by the meager improvement in model fit.

1.12 Summary

- Want to know how strong the relationship is between two variables? Calculate a correlation (Section **??**).
- Drawing scatterplots (Section **??**).
- Basic ideas in linear regression and how regression models are estimated (Sections **??** and **??**).
- Multiple linear regression (Section **??**).
- Measuring the overall performance of a regression model using R^2 (Section **??**).
- Hypothesis tests for regression models (Section **??**)
- Calculating confidence intervals for regression coefficients and standardised coefficients (Section **??**).
- The assumptions of regression (Section **??**) and how to check them (Section **??**).
- Selecting a regression model (Section **??**).