OpenIntro

第 8 章：線形回帰入門

OpenIntro Statistics 第 4 版（日本語版）

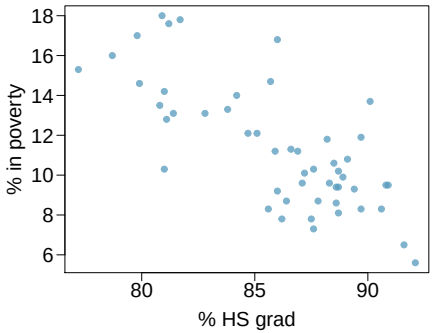
原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

数値変数のモデル化

本節では、2 つの数値変数間の関係を定量化する方法と、数値または質的な説明変数を用いて数値の目的変数をモデル化する方法を学ぶ。

貧困率と高校卒業率

下の散布図は、米国全 50 州とワシントン DC の高校卒業率と、貧困ライン（2012 年時点で 4 人家族の年収 23,050 ドル未満）以下で生活する住民の割合との関係を示している。



目的変数は？

説明変数は？

関係は？

米国の高校卒業率から貧困率を予測する線形モデルは次のとおりである：

$$\hat{poverty} = 64.78 - 0.62 \times HS_{grad}$$

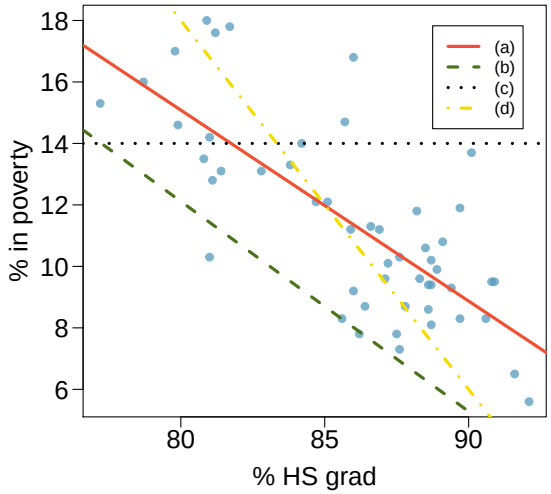
「ハット（ˆ）」は推定値であることを示す。

ジョージア州の高校卒業率は 85.1%である。このモデルは同州の貧困率をどのように予測するか？

$$64.78 - 0.62 \times 85.1 = 12.018$$

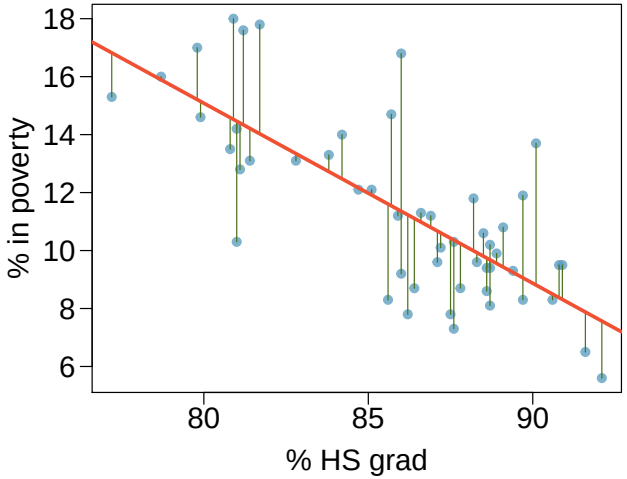
直線の目視確認

貧困率と高校卒業率の線形関係に最もよく当てはまる直線はどれか。1 つ選べ。



残差

残差とはモデルの当てはめから残ったものである：データ = 当てはめ値 + 残差

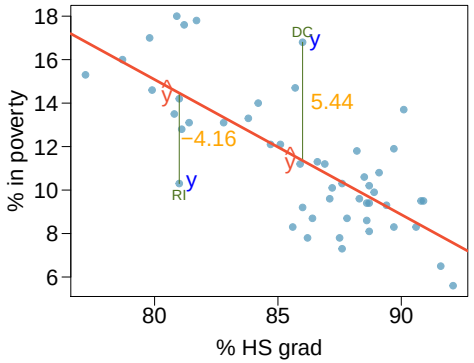


残差（続き）

残差

残差は観測値（y_i）と予測値 $\hat{y}_i$ との差である。

$$e_i = y_i - \hat{y}_i$$



- ワシントン DC の貧困率は予測値より 5.44%高い。
- ロードアイランド州の貧困率は予測値より 4.16%低い。

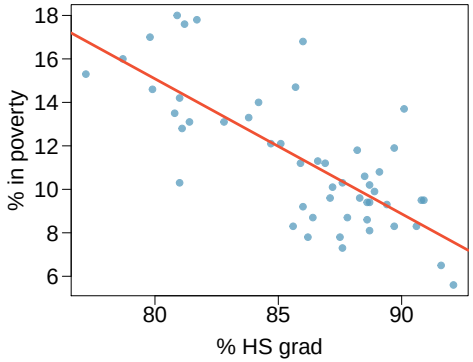
関係の定量化

- 相関は 2 つの変数間の線形な関連の強さを表す。
- -1（完全な負の相関）から $+1$（完全な正の相関）の値をとる。
- 値が 0 の場合、線形な関連がないことを示す。

相関の推測

貧困率と高校卒業率の相関として最もよい推測はどれか？

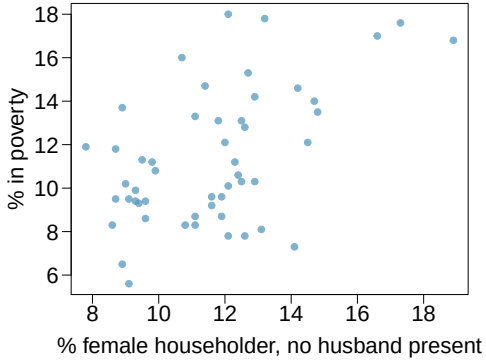
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



相関の推測

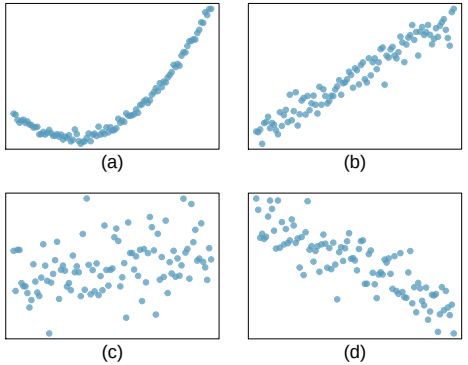
貧困率と「夫なし女性世帯主」の割合の相関として最もよい推測はどれか？

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



相関の評価

以下のうち、最も強い相関（相関係数が$+1$ または-1 に最も近い）を持つのはどれか？



最良の直線を見つけるための客観的な基準

- 残差が小さい直線を求める：
  1. 方法 1：残差の絶対値の和を最小化する

$$|e_1| + |e_2| + \cdots + |e_n|$$

  2. 方法 2：残差の二乗和を最小化する——最小二乗法

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- なぜ最小二乗法か？
  1. 最も広く使われている
  2. 手計算およびソフトウェアによる計算が容易
  3. 多くの場面で、2 倍の大きさの残差は通常 2 倍以上に悪いとみなされる

最小二乗直線

$$\hat{y} = \beta_0 + \beta_1 x$$

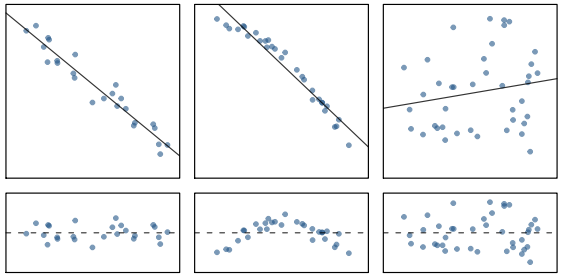
- $\hat{y}$：目的変数 y の予測値
- β_0：切片（母数）
  - b_0：切片（点推定量）
- β_1：傾き（母数）
  - b_1：傾き（点推定量）
- x：説明変数

最小二乗直線の条件

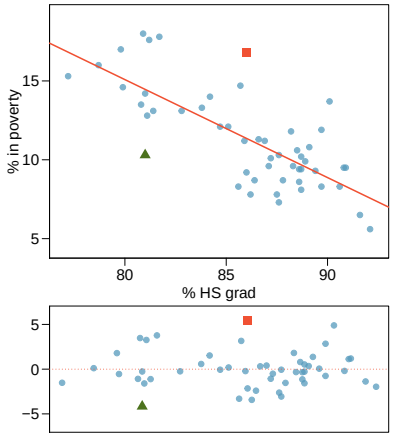
1. 線形性
2. 残差がほぼ正規分布に従う
3. 等分散性（一定のばらつき）

条件（1）：線形性

- 説明変数と目的変数の関係は線形でなければならない。
- 非線形な関係へのモデル当てはめ手法も存在するが、本講義の範囲を超える。この話題に興味があれば、openintro.org のオンライン補足資料で新しい手法を学べる。
- 散布図または残差プロットを用いて確認する。



残差プロットの見方



▲ *RI*（ロードアイランド州）：

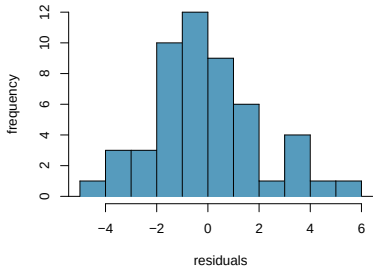
$$\%\ HS\ grad = 81, \quad \%\ in\ poverty = 10.3$$
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 \times 81 = 14.46$$
$$e = 10.3 - 14.46 = -4.16$$

■ *DC*（ワシントン *DC*）：

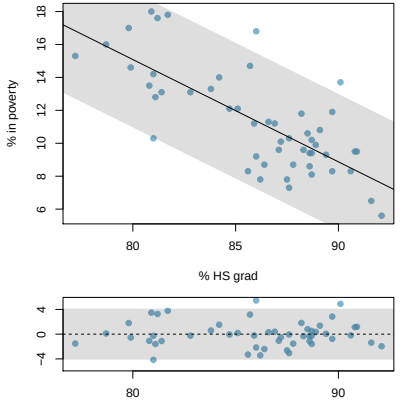
$$\%\ HS\ grad = 86, \quad \%\ in\ poverty = 16.8$$
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 \times 86 = 11.36$$
$$e = 16.8 - 11.36 = 5.44$$

条件（2）：残差がほぼ正規分布に従う

- 残差はほぼ正規分布に従う必要がある。
- データの傾向から外れた異常な観測値がある場合、この条件が満たされないことがある。
- ヒストグラムを用いて確認する。



条件（3）：等分散性（一定のばらつき）

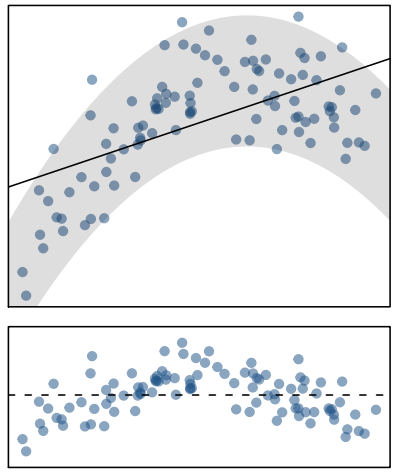


- 最小二乗直線の周りの点のばらつきはおおむね一定でなければならない。
- これは、0 の線の周りの残差のばらつきもおおむね一定であることを意味する。
- 等分散性（*homoscedasticity*）とも呼ばれる。
- 残差プロットを用いて確認する。

条件の確認

この線形モデルが明らかに違反している条件はどれか？

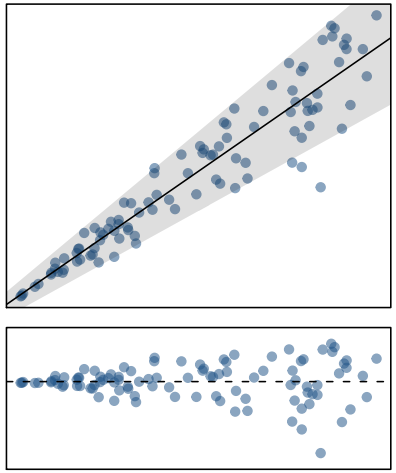
(a) 等分散性
(b) 線形関係
(c) 正規残差
(d) 極端な外れ値がない



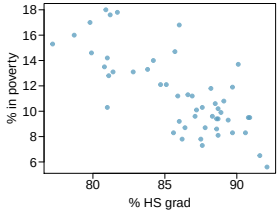
条件の確認

この線形モデルが明らかに違反している条件はどれか？

(a) 等分散性
(b) 線形関係
(c) 正規残差
(d) 極端な外れ値がない



与えられたデータ



	高校卒業率 (%) (x)	貧困率 (%) (y)
平均	$\bar{x} = 86.01$	$\bar{y} = 11.35$
標準偏差	$s_x = 3.73$	$s_y = 3.1$
	相関	$R = -0.75$

傾き

傾き

回帰の傾きは次の式で計算できる：

$$b_1 = \frac{s_y}{s_x} R$$

具体的には…

$$b_1 = \frac{3.1}{3.73} \times (-0.75) = -0.62$$

解釈

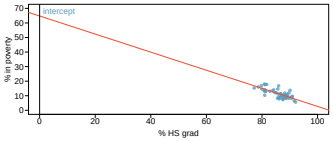
高校卒業率が 1%ポイント上昇するごとに、貧困率は平均的に 0.62%ポイント低下すると予測される。

切片

切片

切片は回帰直線が y 軸と交わる点である。回帰直線は常に $(\bar{x}, \bar{y})$ を通るという性質を用いて計算する：

$$b_0 = \bar{y} - b_1 \bar{x}$$



$$\begin{aligned} b_0 &= 11.35 - (-0.62) \times 86.01 \\ &= 64.68 \end{aligned}$$

切片の正しい解釈はどれか？

(a) 高校卒業率が 1%ポイント上昇するごとに、貧困率は平均的に 64.68%増加すると予測される。

(b) 高校卒業率が 1%ポイント低下するごとに、貧困率は平均的に 64.68%増加すると予測される。

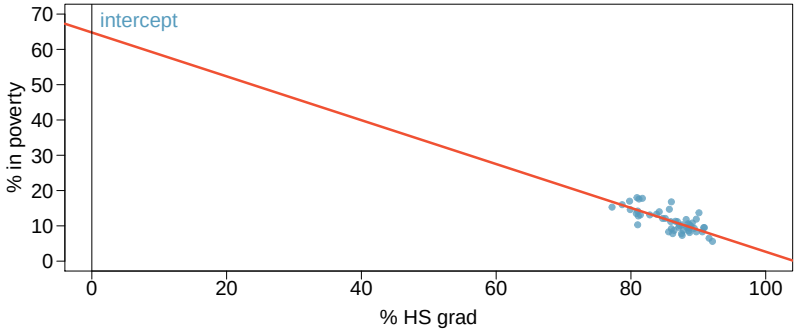
(c) 高校卒業者がいない場合、住民の 64.68%が貧困ライン以下で生活する。

(d) 高校卒業者がいない州では、住民の平均 64.68%が貧困ライン以下で生活すると予測される。

(e) 高校卒業者がいない州では、貧困率は平均的に 64.68%増加すると予測される。

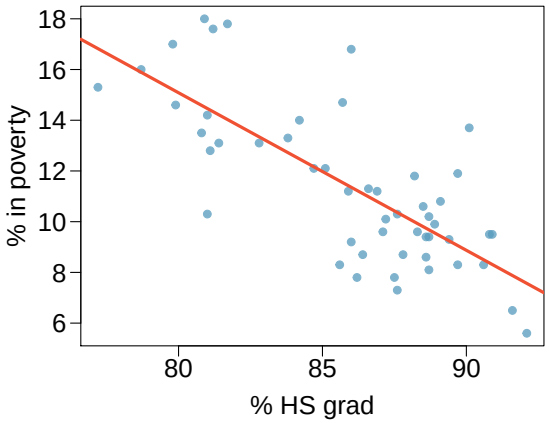
切片についての補足

データセットには高校卒業者が一人もいない州は存在しないため、切片はあまり意味を持たず、有用でもなく、切片の予測値がデータの大部分から大きく離れているため信頼性も低い。



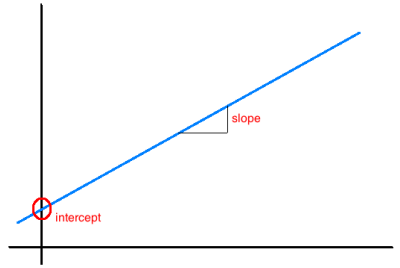
回帰直線

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



傾きと切片の解釈

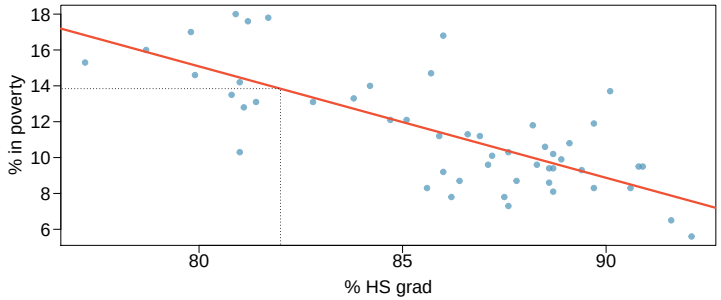
- 切片：$x = 0$ のとき、y は切片に等しいと予測される。
- 傾き：x が 1 単位増加するごとに、y は平均的に傾きの分だけ増加／減少すると予測される。



注：これらの記述は、研究が無作為化比較実験でない限り、因果関係を意味しない。

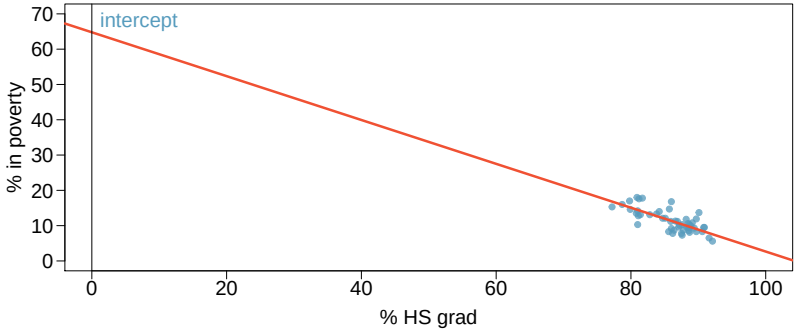
予測

- 線形モデルを用いて、説明変数の特定の値に対する目的変数の値を予測することを予測と呼ぶ。線形モデルの式に x の値を代入するだけでよい。
- 予測値には不確かさが伴う。

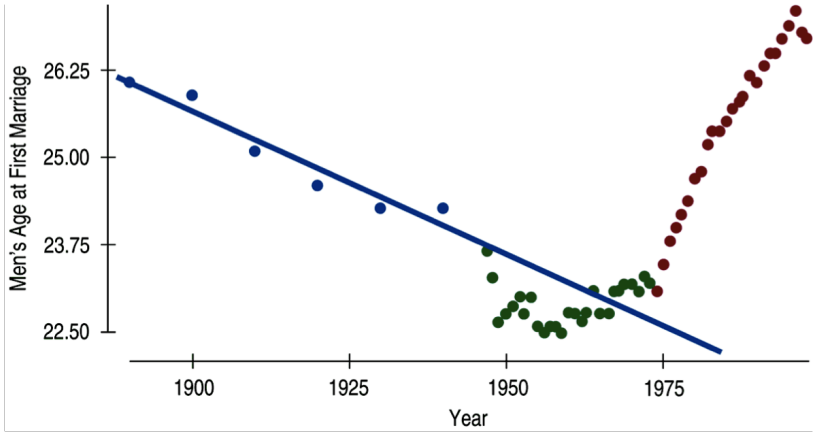


外挿

- モデルの推定を元のデータの範囲外の値に適用することを外挿と呼ぶ。
- 切片が外挿になる場合もある。



外挿の例



外挿の例


BBC NEWS
Watch One-Minute World News
News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news
Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK
E-mail this to a friend
Printable version
Women 'may outsprint men by 2156'
Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.
An Oxford University study found that women are running faster than they have ever done over 100m.
Women are set to become the dominant sprinters
At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.
The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.
However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.
"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

外挿の例

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

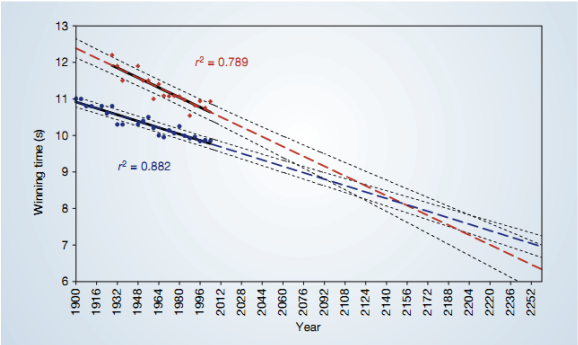


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

決定係数 R^2

- 線形モデルの当てはまりの強さは、通常 R^2（決定係数）を用いて評価する。
- R^2 は相関係数の二乗として計算される。
- 目的変数のばらつきのうち、モデルで説明できる割合を示す。
- 残りのばらつきは、モデルに含まれていない変数またはデータの固有のランダム性によって説明される。
- ここで扱っているモデルでは、$R^2 = (-0.75)^2 = 0.56$ である。

R^2 の解釈

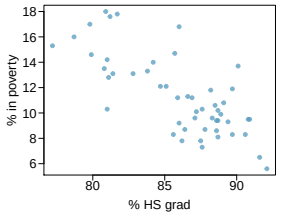
$R = -0.75$、$R^2 = 0.56$ の正しい解釈はどれか？

(a) 51 州における高校卒業率のばらつきの 56%はモデルで説明できる。

(b) 51 州における貧困率のばらつきの 56%はモデルで説明できる。

(c) 高校卒業率が貧困率を正しく予測できる割合は 56%である。

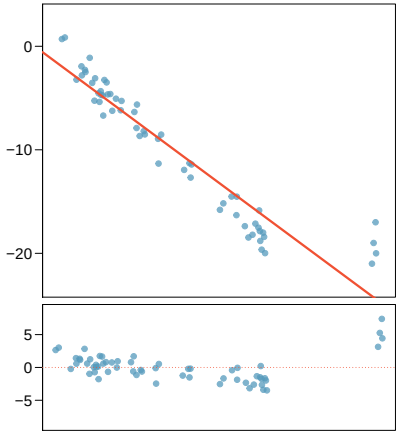
(d) 51 州における貧困率のばらつきの 75%はモデルで説明できる。



外れ値の種類

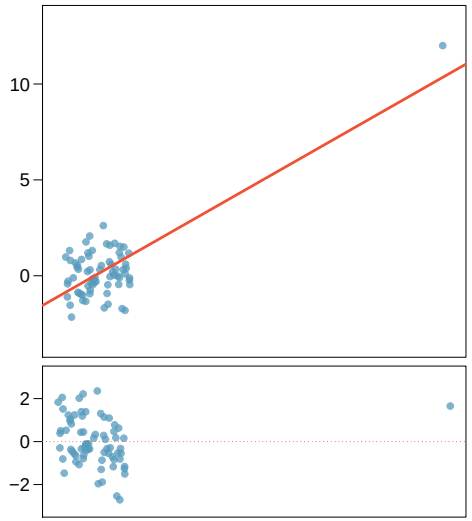
この散布図において、外れ値は最小二乗直線にどのような影響を与えているか？

この問いに答えるために、外れ値がある場合とない場合の回帰直線がどうなるかを考える。外れ値がない場合、回帰直線はより急な傾きになり、大きな観測値の集まりに近くなる。外れ値がある場合、直線は上方に引っ張られ、大きな観測値の集まりの一部から遠ざかる。



外れ値の種類

この散布図において、外れ値は最小二乗直線にどのような影響を与えているか？

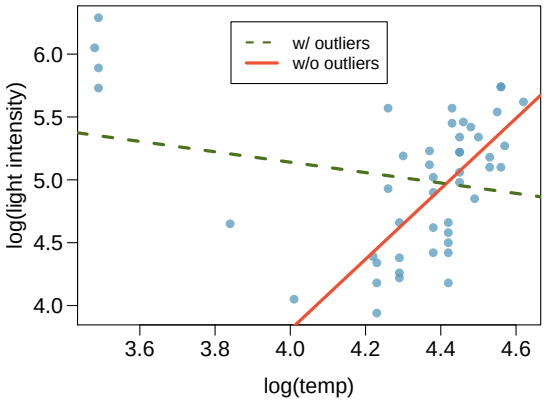


用語

- 外れ値とは、点の集まりから離れた位置にある点のことである。
- 点の集まりの中心から水平方向に離れた外れ値を高レバレッジ点（high leverage points）と呼ぶ。
- 回帰直線の傾きに実際に影響を与える高レバレッジ点を影響点（influential points）と呼ぶ。
- ある点が影響点かどうかを判断するには、その点がある場合とない場合の回帰直線を視覚的に比較する。直線の傾きが大きく変わる場合、その点は影響点である。大きく変わらない場合は影響点ではない。

影響点

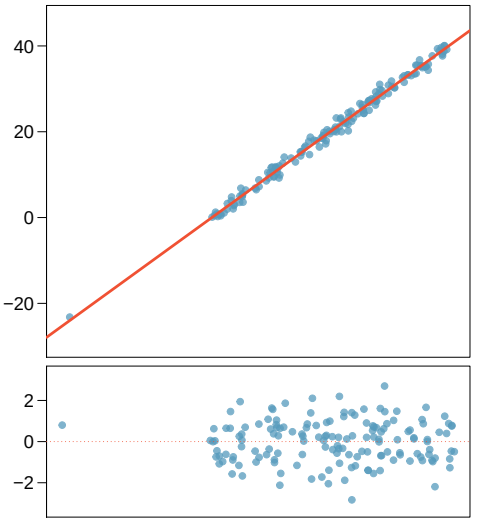
星団 CYG OB1 に属する 47 個の星について、表面温度の対数と光度の対数のデータが得られている。



外れ値の種類

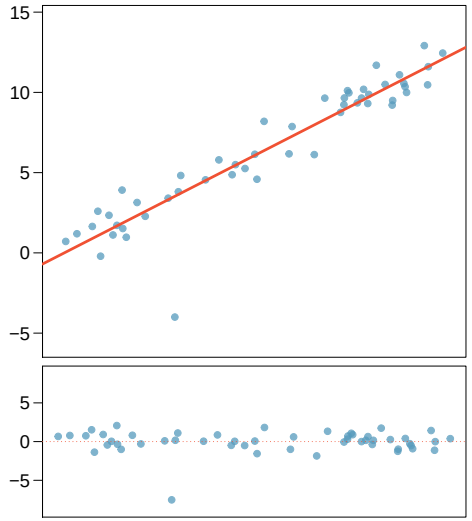
この外れ値を最もよく表す記述はどれか？

(a) 影響点
(b) 高レバレッジ点
(c) 上記のいずれでもない
(d) 外れ値はない



外れ値の種類

この外れ値は回帰直線の傾きに影響を与えているか？



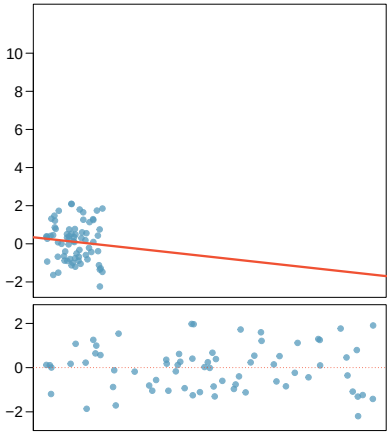
まとめ

次のうち正しい記述はどれか？

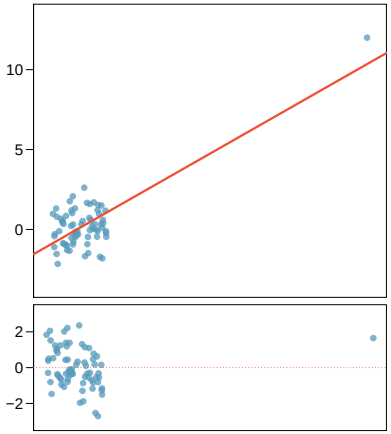
(a) 影響点は常に回帰直線の切片を変化させる。
(b) 影響点は常に R^2 を低下させる。
(c) 低レバレッジ点の方が高レバレッジ点よりも影響点になりやすい。
(d) データセットに影響点が含まれる場合、説明変数と目的変数の関係は常に非線形である。
(e) 上記のいずれも正しくない。

まとめ（続き）

$R = 0.08, R^2 = 0.0064$

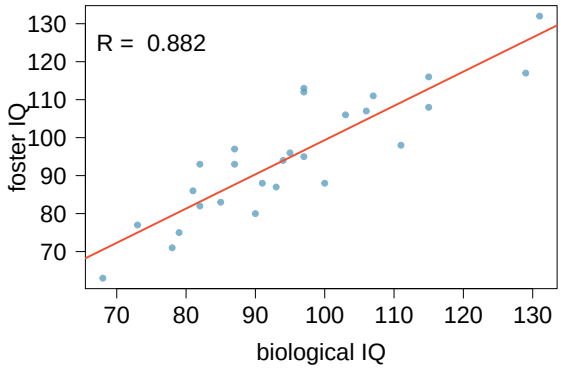


$R = 0.79, R^2 = 0.6241$



遺伝か環境か？

1966 年、Cyril Burt は「知能の差異の遺伝的決定：一緒に育てられた双子と別々に育てられた双子の研究」という論文を発表した。データは、一卵性双生児 27 組（代表的な無作為標本と仮定）の IQ スコアで構成されており、一方は養親に、もう一方は実の親に育てられた。



次のうち誤っている記述はどれか？

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) 実の親に育てられた双子の IQ が 10 ポイント高い場合、養親に育てられた双子の IQ は平均的に 9 ポイント高いと関連づけられる。

(b) 養親に育てられた双子の IQ の約 78%はモデルで正確に予測できる。

(c) 線形モデルは $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$ である。

(d) 平均より高い IQ を持つ養育双子は、平均より高い IQ を持つ実親双子と対応している傾向がある。

傾きの検定

出生時に引き離されたすべての双子の代表的な標本としてこの 27 組を想定した場合、実の親に育てられた双子の IQ が養親に育てられた双子の IQ の有意な予測因子であるかどうかを検定するための適切な仮説はどれか？

(a) $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

傾きの検定（続き）

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- 回帰の推測では常に t 検定を用いる。

  復習：検定統計量 $T = \frac{\text{点推定量} - \text{帰無値}}{SE}$

- 点推定量 $= b_1$ は観測された傾きである。
- SE_{b_1} は傾きに関連する標準誤差である。
- 傾きに関連する自由度は $df = n - 2$ であり、n は標本サイズである。

  復習：推定するパラメータ *1* つにつき自由度を *1* つ失う。単回帰では β_0 と β_1 の *2* つのパラメータを推定するため、$df = n - 2$ となる。

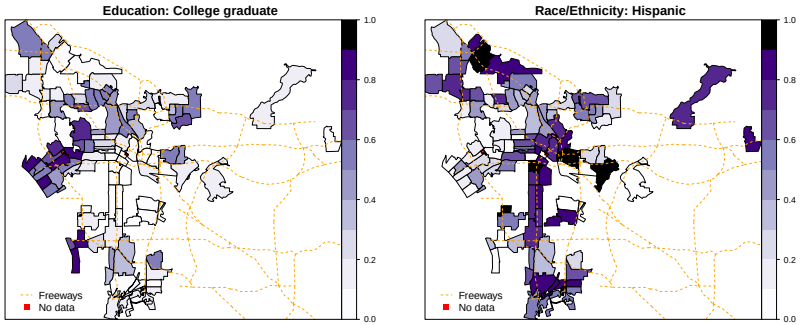
傾きの検定（続き）

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25 \\
p\text{値} &= P(|T| > 9.36) < 0.01
\end{aligned}
$$

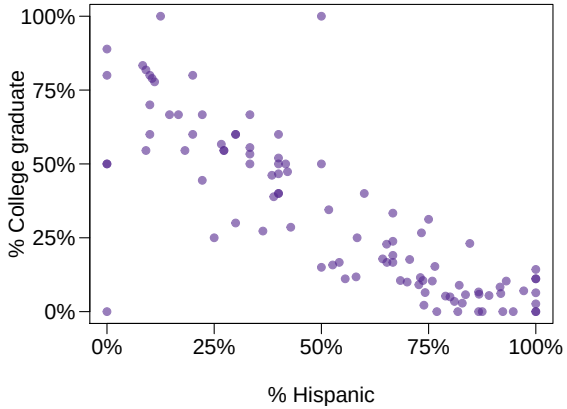
LA 地区における大学卒業率とヒスパニック系住民割合

LA の 100 の郵便番号地域の標本において、大学卒業率とヒスパニック系住民割合の関係について何が言えるか？



LA 地区における大学卒業率とヒスパニック系住民割合（別の視点）

LA の 100 の郵便番号地域の標本において、大学卒業率とヒスパニック系住民割合の関係について何が言えるか？



LA 地区における大学卒業率とヒスパニック系住民割合：線形モデル

傾きの最もよい解釈はどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
%Hispanic	-0.7527	0.0501	-15.01	0.0000

(a) LA 地区の郵便番号地域でヒスパニック系住民が 1%増加すると、大学卒業率は 75%低下する。

(b) LA 地区の郵便番号地域でヒスパニック系住民が 1%増加すると、大学卒業率は 0.75%低下する。

(c) ヒスパニック系住民が 1%追加されるごとに、LA 地区の郵便番号地域の大学卒業率は 0.75%低下する。

(d) ヒスパニック系住民がいない郵便番号地域では、大学卒業率は 75%と予測される。

LA 地区における大学卒業率とヒスパニック系住民割合：線形モデル

このデータは LA 地区の郵便番号地域においてヒスパニック系住民割合と大学卒業率の間に統計的に有意な関係があることの説得力ある証拠を提供しているか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

これらの郵便番号地域が無作為に選択されていない場合、この p 値はどの程度信頼できるか？

傾きの信頼区間

信頼区間は 点推定量 $\pm$ ME として計算され、単回帰における傾きの自由度は $n-2$ であることを思い出してほしい。27 組の双子の観測に基づくモデルにおいて、傾きパラメータの 95%信頼区間として正しいものはどれか？

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

まとめ

- 単回帰モデルの傾きに対する推測：
  - 仮説検定：

$$T = \frac{b_1 - \text{帰無値}}{SE_{b_1}} \qquad df = n - 2$$

  - 信頼区間：

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- 帰無値は、説明変数と目的変数の間に何らかの関係があるかを確認するため、通常 0 に設定される。
- 回帰出力には b_1、SE_{b_1}、および帰無値が 0 の場合の傾きの t 検定の両側 p 値が含まれる。
- 切片の推測を行うことはほとんどないため、傾きの推定と推測に焦点を当てる。

注意事項

- 扱っているデータの種類（無作為標本、非無作為標本、または母集団）を常に把握しておくこと。
- すでに母集団データがある場合、統計的推測とそれに基づく p 値は意味を持たない。
- 偏った非無作為標本を使用している場合、推測の結果は信頼できない。
- 最終的な目標は、独立した観測値を持つことである。