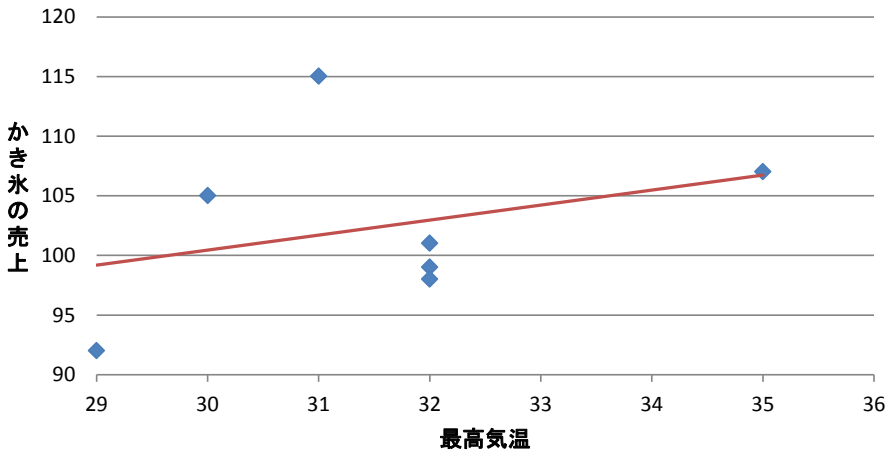
回帰モデル編のイントロ

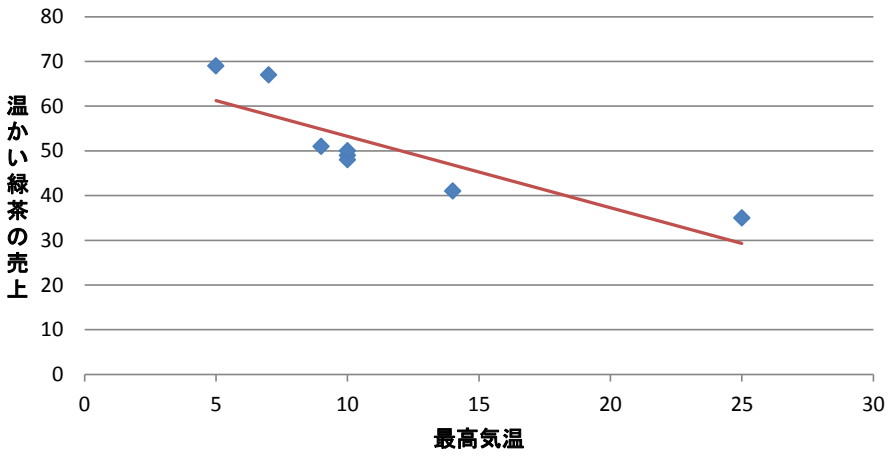
- 回帰モデル＝胃カメラでたとえると
  - 胃-胃カメラ-医者
    - ✓ 医者が胃カメラを使って胃を検査し、医学的診断に基づき、病気を治療する
  - データ-回帰モデル-分析者
    - ✓ 分析者が回帰モデルを使ってデータを分析し、データの特徴や回帰モデルを説明し、実務に役立てる

- 習得すること
  - 基本編：回帰モデルの基本的な使い方を知ろう
    - ✓ 胃カメラの一般的な使い方
  - 応用編：データの特徴に合わせて、回帰モデルを改善しよう
    - ✓ 胃の個別具体的な症状に合わせた胃カメラの使い方

回帰モデルとは？

- 目的変数yと説明変数xの関係を説明するモデル。ここでは線形関係 $y = \alpha + \beta x$ を扱う。
- 理論的にシンプルなこともあり、統計的手法として、まず最初に勉強することが多い
- 解釈しやすく、最も実用性の高い手法の一つ
  - ➢ビジネスにおける回帰モデルの利用シーンとして、需要や売上の予測が挙げられる
- ここでは、まず説明変数が1つの単回帰を学び、次に複数の重回帰を学ぶ





回帰モデルを走らせてみよう

- Rの組み込みデータcarsについて
  - 車の速度(マイル/時間)と停車するまでの距離(フィート)

- 演習 データcarsの基礎分析
  - データをロードしよう
    - ✓ data(cars)
  - データをざっと眺めてみよう
    - ✓ View(cars)
  - データの変数の数と件数を把握しよう
    - ✓ str(cars)
  - 要約統計量を求めよう
    - ✓ summary(cars)
  - 散布図を描いてみよう
    - ✓ plot(cars$speed, cars$dist)
  - 相関係数を求めてみよう
    - ✓ cor(cars$speed, cars$dist)

- 問
  - 基礎分析から、speedとdistにどんな関係が見られるだろうか？
  - speedからdistを予測できそうだろうか？

単回帰モデルを走らせてみよう

- speedからdistを予測するために、1変数の回帰モデル（単回帰、Y = α + βX）を走らせてみよう

```
lm01 <- lm(dist ~ speed, #linear modelの略
           data = cars)

lm(目的変数 ~ 説明変数,
   data = データ名)
```

- 最小二乗法による推定結果の出力
  - 残差の二乗和を最小化するようにα と βを求める

```
summary(lm01)
```

単回帰モデルの推定結果

```
Call:
  lm(formula = dist ~ speed, data = cars)

Residuals:
   Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
  speed       3.9324     0.4155   9.464 1.49e-12 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

推定モデル

残差の分布

係数

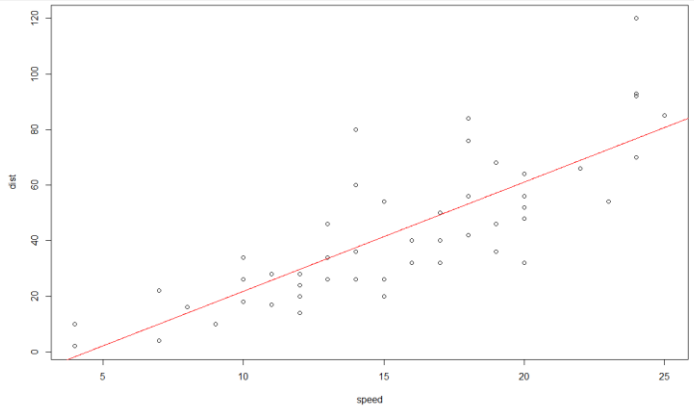
	推定値	標準偏差	t値	p値
(切片)	-17.5791			
speed	3.9324			

有意水準のコード

残差の標準偏差 自由度48 (= 50-2)

決定係数 0.6511

自由度(1, 48)の***F値***、p値



単回帰モデルの推定結果の読み取り　一覧

- 残差
  - ➢実績値Yと推定値（モデルの予測値）Yの差
    - ✓　モデルで説明しきれなかった分

- 係数
  - ➢αとβの推定結果

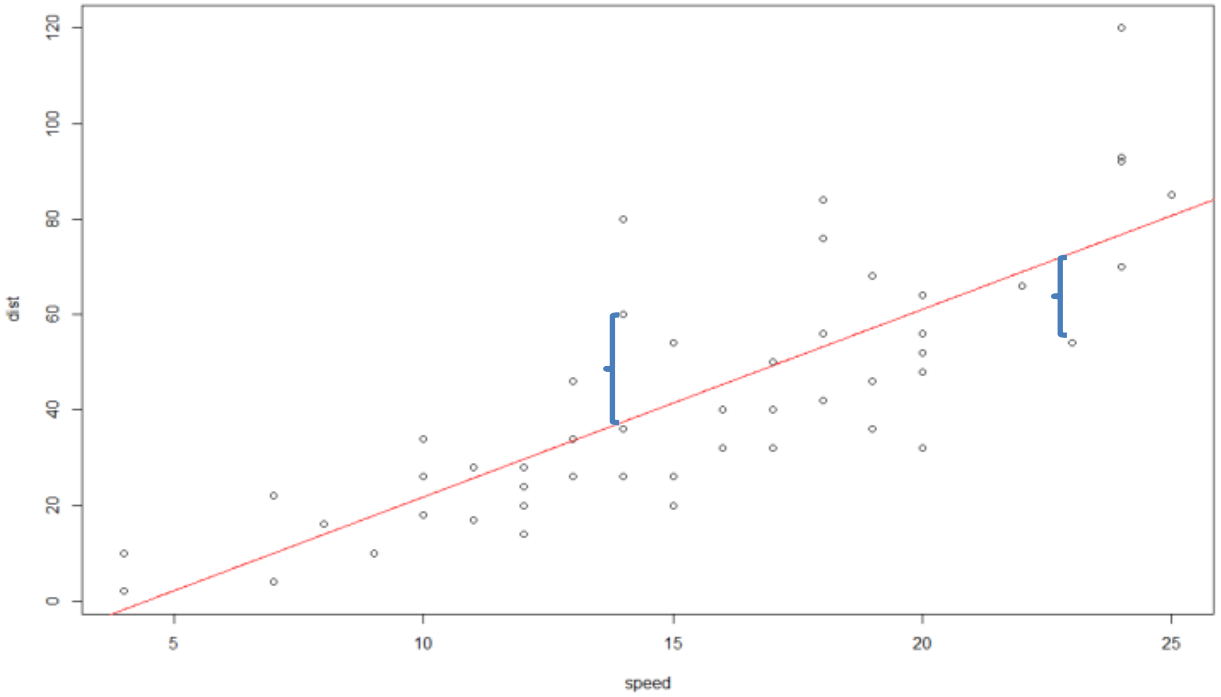
- 残差の標準偏差（単独ではほとんど使わない）
  - ➢残差の平方和を(件数–2)で除した値の平方根で、小さいほど誤差がばらついていない
  - ➢予測の信頼区間などに使用する

- 決定係数
  - ➢モデルで説明されたバラつき部分／Yの全てのバラつき
    - ✓　0～1の間を取り、1の場合データは完全に回帰式上にある。0の場合、XとYは無関係。（β＝0で、y = α + 0Xの水平線）

- F値（あまり使わない）
  - ➢F値は、yのバラつきのうちXで説明できる分／ yのバラつきのうちXで説明できない分
  - ➢F値が大きいほど、Xで説明できる分が大きい
  - ➢F値で帰無仮説β=0（スロープパラメータの0制約）を検定する

単回帰モデルの推定結果の読み取り　残差

- 残差とは
  - u = Y – (α + βX) = 実績値Y - 推定値Y
  - グラフでいうと、推定式からの縦のずれで、プラスの値もマイナスの値もとりうる



単回帰モデルの推定結果の読み取り　残差

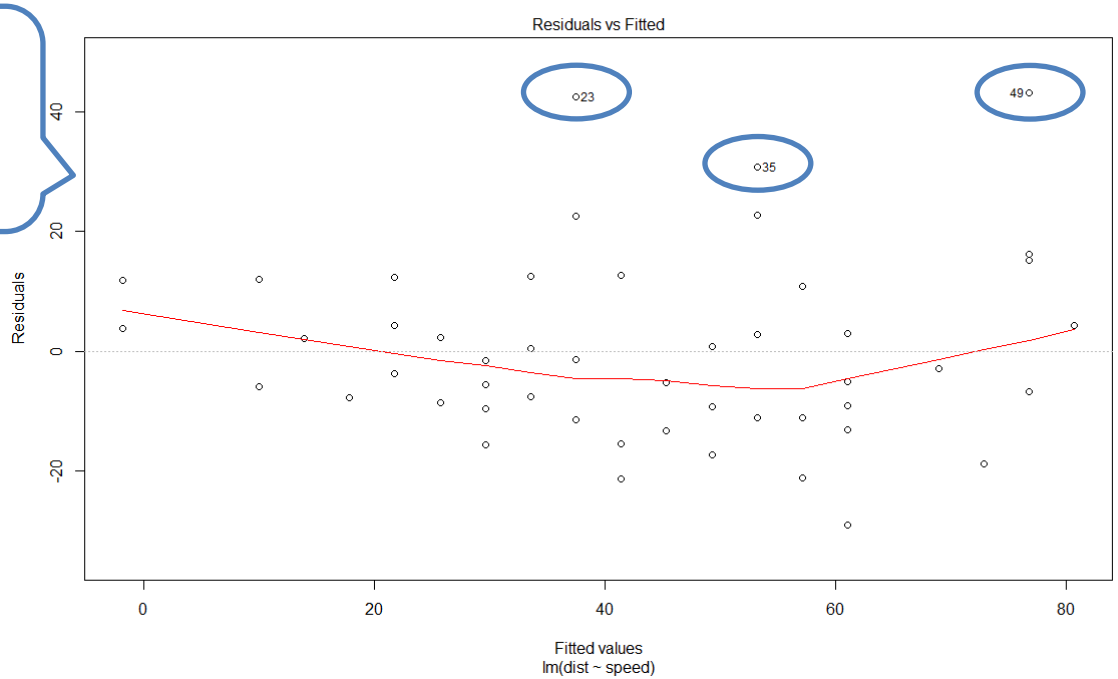
- なぜ残差が発生するのか？
  - ➢現実に対して、線形性を仮定している
    - ✓ 現実はもっと複雑な関係
  - ➢データの計測誤差
    - ✓ 丸め誤差
    - ✓ いい加減な回答
  - ➢その他
    - ✓ 突発的なショックやノイズ

単回帰モデルの推定結果の読み取り　残差

- 残差のチェックポイント
  - ➢残差から、推定値から大きく乖離した値（外れ値）に目星をつける
    - ✓ てこ比等、統計指標による外れ値の検出もある

```
plot(lm(dist ~ speed, data = cars),
     which = 1)
```

縦軸：残差と横軸：予測値のプロット。残差の全体像を把握するのに使う
23, 35, 49が外れ値として指摘されている



単回帰モデルの推定結果の読み取り　残差

- 23, 35, 49の抽出

	speed	dist	pred	resid
23	14	80	37.47463	42.52537
35	18	84	53.20426	30.79574
49	24	120	76.79872	43.20128

- speed = 14のレコードの比較
  - distが26~80と値がばらついている

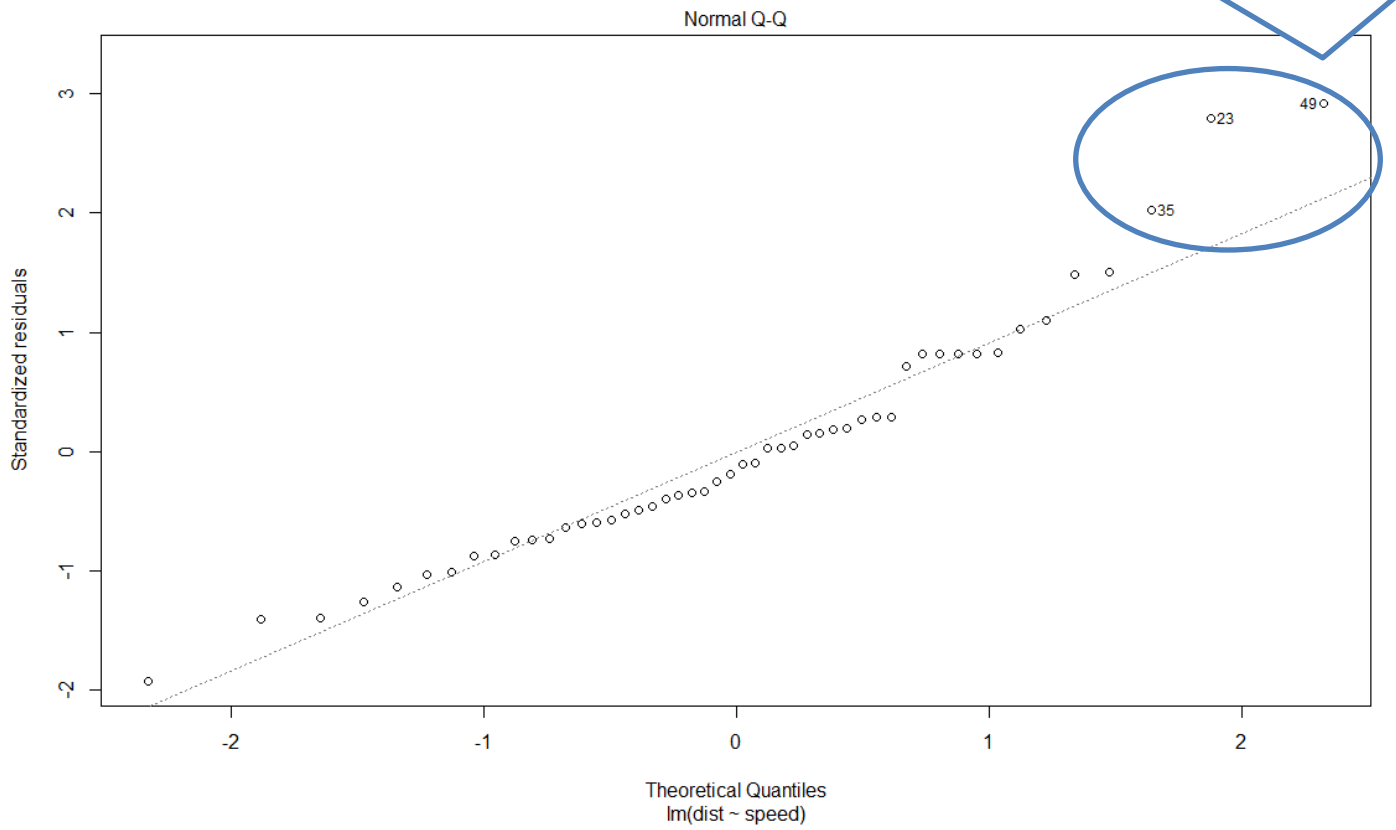
	speed	dist	pred	resid
20	14	26	37.47463	-11.474628
21	14	36	37.47463	-1.474628
22	14	60	37.47463	22.525372
23	14	80	37.47463	42.525372

単回帰モデルの推定結果の読み取り　残差

- 残差のチェックポイント

```
plot(lm(dist ~ speed, data = cars),
     which = 2)
```

残差が正規分布に従うと仮定するので、正規Q-Qプロットでその検証
縦軸が標準化された残差、横軸が理論値
残差が正規分布に従うと、点が直線状にプロットする



単回帰モデルの推定結果の読み取り　係数

- y = α + βxのαとβが推定すべきパラメータ
  - α：切片、定数項
  - β：回帰係数、係数、傾き、スロープパラメータ

- 仮説検定の復習
  - 帰無仮説をあらかじめ決められた有意水準で棄却するかしないか決定すること
  - 通常、帰無仮説は棄却されて意味があるようになっている。帰無仮説が正しいとしたら、X%の確率でしか起こらないような稀なことが起こっており、それは帰無仮説が疑わしいことを意味する
  - 線形回帰におけるパラメータの仮説検定は、自由度(n-2)のt分布に従う

- 有意水準
  - t分布の端のX%のこと。この外側を棄却域という。この場合、帰無仮説は棄却される。弊社では、5%が最もよく用いられる（次に1%）。

- 標準偏差
  - 推定値のバラつきのこと。大きい（小さい）ほど、不確かな（確かな）推定値と解釈され、帰無仮説が棄却しにくく（しやすく）なる。

- パラメータのt検定
  - 誤差項にいくつかの仮定があり、理論的な導出を経て、自由度(n-2)のt分布に従う
  - 帰無仮説がβ=0、対立仮説β≠0　これをt値の大小で判断する
  - 乱暴に言って、サンプル数が数十以上であれば、tが絶対値で2より大きいと帰無仮説が棄却される。

単回帰モデルの推定結果の読み取り　係数

- p値
  - 帰無仮説が正しい時、極端なt値が観測される確率
  - 乱暴に言って、t値が絶対値で2より大きいと0.05以下になる。

- βの解釈
  - 有意であり、符号（プラスマイナス）の解釈が自然である前提で
  - Xが1単位増加すると、β分Yが増加する
  - 速度が1マイル毎時増加すると、停車までの距離が3.9324フィート長くなる

単回帰モデルの推定結果の読み取り　決定係数

- 決定係数(R-square)とは
  - R-square = 1 – (残差の分散) / (実績値の分散)
  - 解釈
    - ✓ データの分散が、モデルによりどれだけ減ったかを表す指標
    - ✓ 0～1の間を取る
    - ✓ 1であれば、データが完全に回帰直線上にある
    - ✓ 0であれば、XとYは無関係。水平線（β＝0）
  - 実績値Yと推定値Yの相関係数の二乗
    - ✓ モデルの精度評価は、RMSE(Root Mean Square Error、平均二乗誤差)などもよく用いられる
  - Adjusted R-squareは重回帰で扱う

- 前回の復習
  - ➢1変数の線形回帰モデルの基礎編
    - ✓ carsのデータで、speedでdistを予測するlm関数
    - ✓ 統計指標の学習

      係数、係数のt検定、決定係数等
    - ✓ モデルの含意

      速度が1マイル毎時増加すると、停車までの距離が3.9324フィート長くなる

      速い車は急に止まれない
    - ✓ グラフ化はとても重要

- 前回の積み残し
  - ➢予測とその信頼区間
  - ➢Anscombeのデータ

本日の内容

- 1変数の線形回帰モデルの基本編
  - 予測とその信頼区間
- 1変数の線形回帰モデルの応用編
  - Anscombeのデータ

- 2変数以上の重回帰モデルの基本編
  - Duncanの社会調査のデータ
  - 多重共線性

前回の復習用の演習

- 仮想データteaについて
  - 温かい緑茶の売上(tea)とその時の気温(temperature)の関係

- 演習1 基礎分析
  - データをロードしよう
  - データをざっと眺めてみよう
  - データの変数の数と件数を把握しよう
  - 要約統計量を求めよう
  - 散布図を描いてみよう
  - 相関係数を求めてみよう

- 問
  - 基礎分析から、teaとtemperatureにどんな関係が見られるだろうか？
  - temperatureからteaを予測できそうだろうか？

- 演習2　回帰モデル
  - 目的変数tea、説明変数temperatureとして、線形回帰モデルを推定してみよう
  - データの散布図と推定結果を図示してみよう
    - ✓ 推定結果に影響を与えている値はあるだろうか？
  - 決定係数はいくらか？　回帰係数βの有意性と符号はどのように推定されたか？
  - 気温が1度下がると（上がると）、売上はいくら増える（減る）だろうか？

予測値の信頼区間

- 点予測
  - ➢temperatureが10の時、teaの予測値は？
  - ➢モデルが算出した予測値は、残差が0となるときの値

- 予測値の信頼区間とは
  - ✓ 構築データのサンプリングをやり直せば値は変わり得る
  - ✓ 突発的なノイズにより、モデルのパラメータが影響を受ける
  - ➢データのサンプリングやノイズによる「予測値のブレ」に対する予測値の信頼区間
  - ➢95%信頼区間の場合、この推定を100回繰り返した場合、95回の推定はこの範囲を通ると解釈される

➢予測値の信頼区間のプログラム

```
# モデル構築
lmtea <- lm(tea ~ temperature, data = tea)

# 信頼区間付きでモデル適用
pred <- predict(lmtea,  interval = “conf”, level = 0.95)  #モデル, 信頼区間, 水準
head(pred)

# データをプロット
plot(tea$temperature, tea$tea , xlim = c(0,30), ylim = c(0, 100)) #xlimとylimはこの後も共通して指定

# 予測値 fit
par(new = T)
plot(tea$temperature, pred[,1], type = "l", col = "red", xlim = c(0,30), ylim = c(0, 100), xlab = "", ylab = "")
#x軸とy軸のlabelは表示しない

# 下限 lwr
par(new = T)
plot(tea$temperature, pred[,2], type = "p", col = "blue", pch = 3, xlim = c(0,30), ylim = c(0, 100), xlab = "",
ylab = "")
# 上限 upr
par(new = T)
plot(tea$temperature,pred[,3],  type = "p", col = "blue", pch = 3, xlim = c(0,30), ylim = c(0, 100), xlab = "",
ylab = "")
```

予測値の信頼区間

➢temperatureが10の時、teaの予測値の95%信頼区間は？

```
# モデル構築
lmtea <- lm(tea ~ temperature, data = tea)

# tea = 10 のデータフレーム作成
test <- data.frame(temperature = 10)

# 信頼区間付きで適用
predict(lmtea,
        newdata = test,
        interval = "conf",
        level = 0.95)
```

➢cf. 予測区間

- ✓ モデルは正しいとして、予測値＋残差の信頼区間
- ✓ interval=“prediction”とすると予測値の信頼区間と同様に算出できる

単回帰(基本編) 総合演習

- 演習
  - ➢回帰モデル
    - ✓ carsのデータで、speedからdistを予測する回帰モデルを走らせてみよう
  - ➢予測の信頼区間
    - ✓ speedが2, 15.4, 30, 50の時のdistの予測とその信頼区間を求めよう

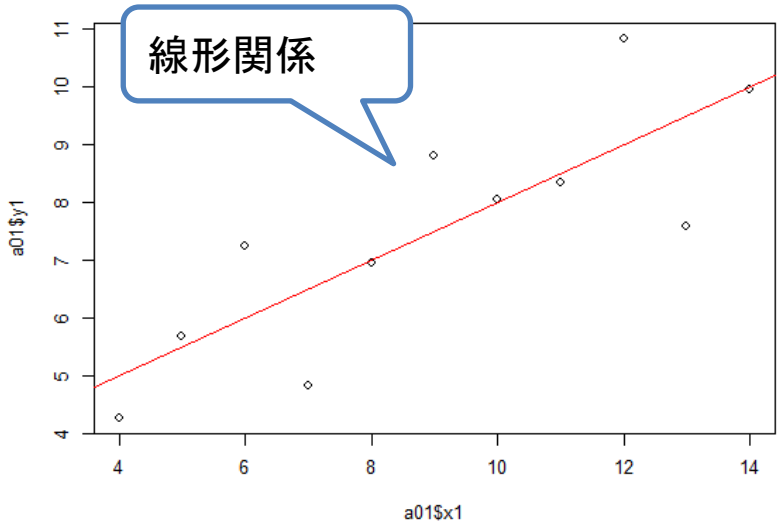
- 問
  - ➢信頼区間の下限と上限の幅は、上記4つの値でどのように異なるだろうか？

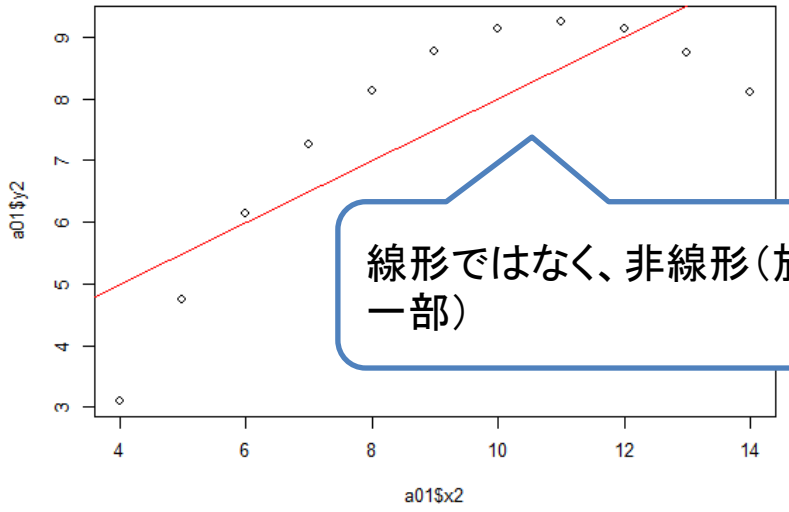
単回帰（応用編） Anscombeのデータ

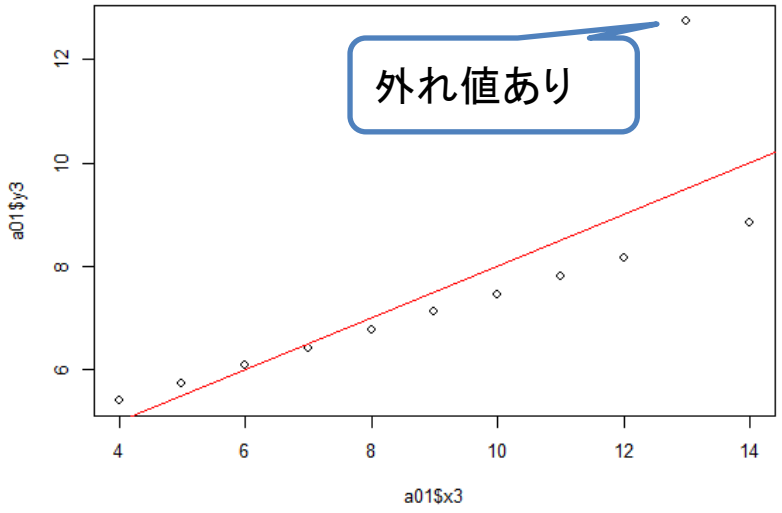
- Anscombe, 1973, “Graphs in Statistical Analysis”, American Statistican のデータ

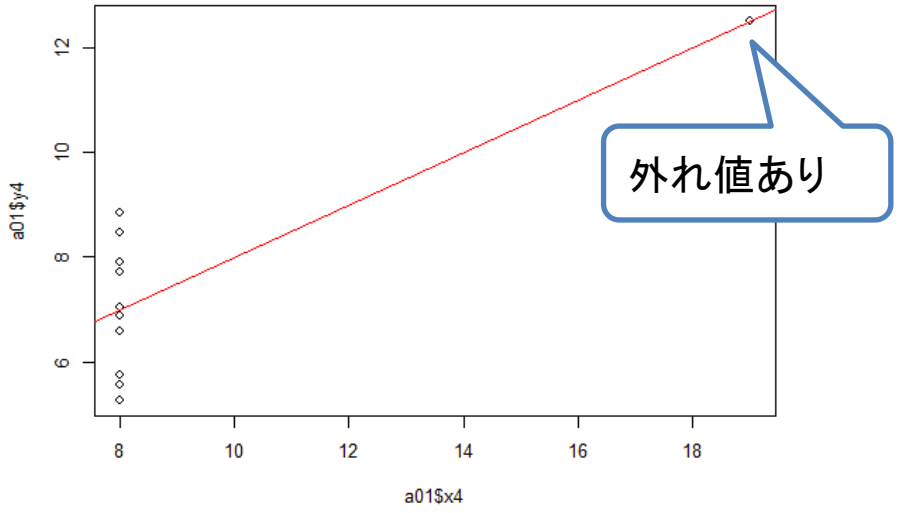
- 演習
  - あえて最初はグラフ化しない
  - 平均値を求めてみよう mean(data$x)
  - 4つのペアの相関係数と回帰モデルをそれぞれ求めてみよう
    - ✓ Y1とX1, Y2とX2, Y3とX3, Y4とX4
  - 最後に、散布図と回帰直線をそれぞれグラフ化してみよう

単回帰(応用編) Anscombeのデータ









単回帰（応用編） Anscombeのデータ

- 演習　線形回帰がうまくいかない場合の対策例
  - Y3とX3、Y4とX4 外れ値を除外して、再推定しよう
  - このようなデータクリーニングをしない場合の推定結果と比較してみよう

※単回帰の枠組みではないが、Y2とX2はY2 = X2 + X2^2で推定すべきであろう

単回帰(応用編) 演習の含意

- 平均値も相関係数も推定結果も同じでも、グラフ化すると分布が全然違う！

- たった1つの外れ値でも推定結果が影響を受けるので、線形回帰はグラフ化が大事！

単回帰(応用編) 外れ値の対応方法

- モデル上の外れ値
  - 目検でわかるような残差の絶対値が明らかに大きいもの、hat関数など
    - ✓ データの除外、データの加工、ダミー変数などで対応

- データそのものの外れ値
  - X%点の外側、極端な外れ値
    - ✓ 除外、丸め処理など

- 重回帰モデルとは
  - ➢単回帰の説明変数が複数になった版

- Rの組み込みデータDuncanについて
  - ➢1950年における職業調査データで、全4変数、53件。
  - ➢incomeを説明するモデルを作りたい

分類	変数名	変数のタイプ	説明
目的変数	income	数値	1950年における男性の3500ドル以上の勤労収入の割合（%）
説明変数	type	文字	prof:専門職、wc:ホワイトカラー、bc:ブルーカラー
	education	数値	1950年における男性の高卒の割合（%）
	prestage	数値	NORC調査による尊敬される職業格付け（%）

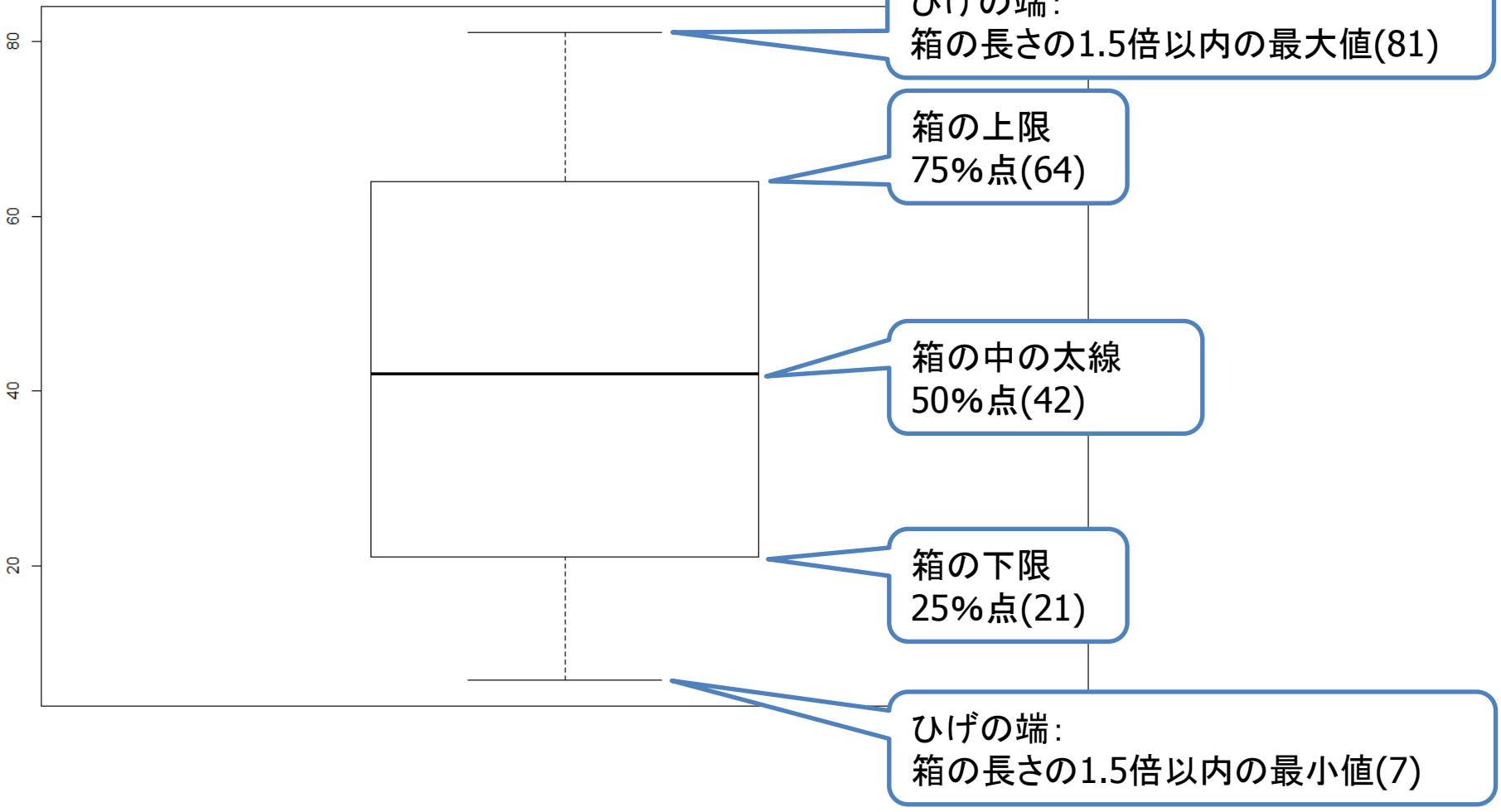
- 演習
  - ➢これまでと同じ枠組みでやってみよう
    - ✓ データをロードしよう
    - ✓ データをざっと眺めてみよう
    - ✓ データの変数の数と件数を把握しよう
    - ✓ 要約統計量を求めよう
    - ✓ 散布図を描いてみよう(2変数の分析)
    - ✓ 相関係数を求めてみよう(2変数の分析、typeは文字変数なので除く)

- Duncanの基礎分析
  - ➢文字×数値(2変数以上の分析)
    - ✓ type別の分布

      箱ひげ図

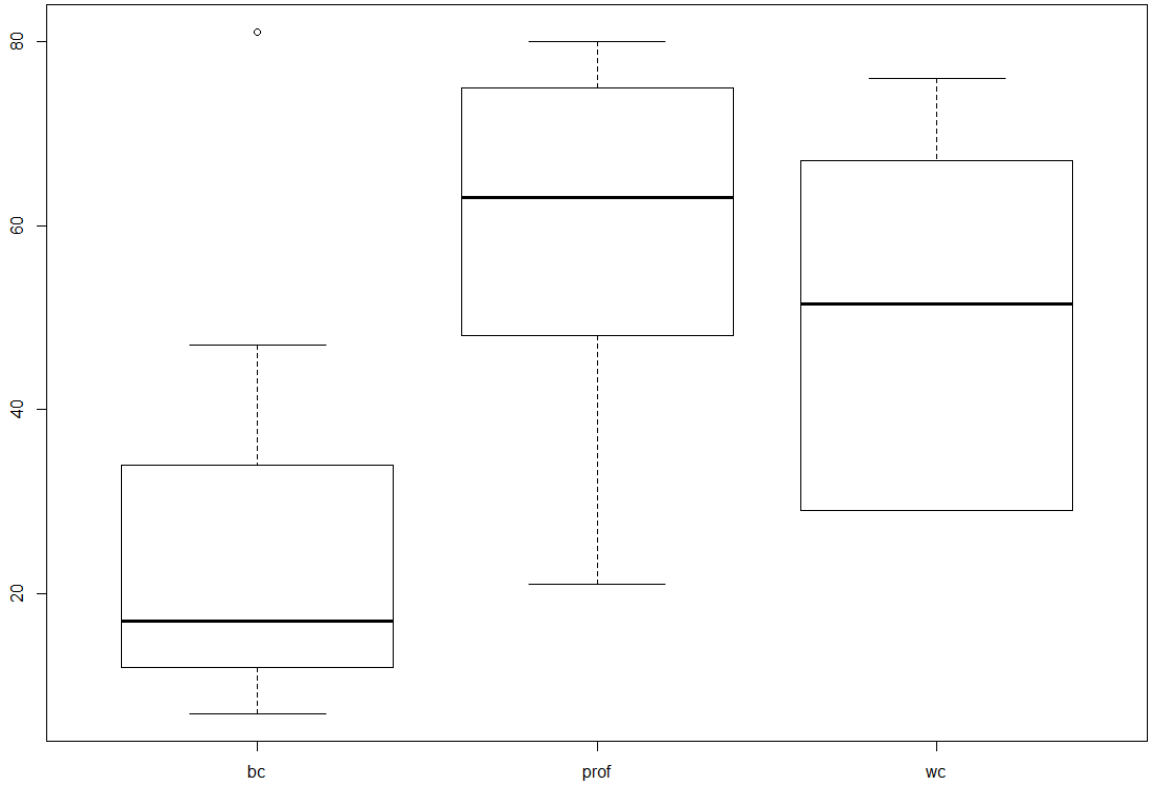
箱ひげ図の説明

- 分布を箱ひげ図で示し、直感的に理解する



箱ひげ図の説明

- bcとそれ以外でincomeの分布が異なることを箱ひげ図で示す



ダミー変数

- ダミー変数の考え方
  - 一時的、突発的なイベント
    - ✓ 学校行事、バーゲン、天災、リーマンショック
  - 性別など量的変数でない変数
    - ✓ 0:女性、1:男性など
- ダミー変数の作り方
  - 0 :イベント非発生/非該当者、1:イベント発生/該当者のように0/1の値で表す

重回帰モデルを走らせてみよう

- incomeを説明する重回帰モデルを作ってみよう
  - まずは単回帰、次に重回帰

- 新しい統計指標
  - 偏回帰係数
    - ✓ 他の変数の影響を除いた前提で、目的変数とその説明変数との単回帰係数
  - 自由度調整済み決定係数
    - ✓ 決定係数は、説明変数を増やしただけ必ず値が増加するという欠点がある。説明変数を増やした分をペナルティとする
    - ✓ 重回帰では、自由度調整済み決定係数を使用する

重回帰モデルを走らせてみよう

- 多重共線性
  - 説明変数同士の相関係数が絶対値で1に近い場合、推定値が有意にならなかったり、不自然な符号になったりする現象のこと
  - 類似の説明変数を一度に扱うことが多い重回帰では、生じやすい
- VIF(Variance Inflation Factor、分散拡大係数)によるとらえ方
  - VIF = 1 / (1 – r12^2)
    - ✓ r12は説明変数1と説明変数2の相関係数
  - r12が大きいほどVIFも大きくなり、多重共線性が生じやすくなる

多重共線性とその解決案

- 財務データなど、リッチなデータほど類似の指標が多数含まれ、多重共線性に直面しやすい

- 解決案
  - ➢類似の変数を同時に投入しない
    - ✓ 相関係数、変数のクラスタリング、主成分分析などで、似た性質の変数は使用せず、代表的な変数のみ使用する
  - ➢類似の変数をそのまま投入できる手法
    - ✓ Ridge回帰（FEGでも研究中）

参考文献

- 回帰分析入門 Rで学ぶ最新データ解析　東京図書　豊田秀樹ら著
  - 理論と演習のバランスが良く、回帰モデルのセッションの復習に向いている。今後学習予定のロジスティックモデルなどもカバーされている。

- R Cookbook, O'reilly, Paul Teetor
  - Rのプログラムの辞書みたいな本。回帰モデルについても、プログラムと出力の読み取り方を一つ一つ丁寧に解説している。

- R in Action, Manning, Robert I. Kabacoff
  - R Cookbookに比べ、手法や応用例が豊富で、より実践的。一方で、理論面は手薄。

重回帰モデルを走らせてみよう

- 演習
  - ➢SASによる回帰分析 p.88 20人の1～3年の成績（100点満点の試験）
  - ➢相関係数を求めてみよう
  - ➢説明変数X1とX2をそれぞれ単回帰で推定してみよう
    - ✓ 符号は自然だろうか？
    - ✓ 3年生の成績を予測するのに有効なのは、どちらの変数だろうか？
  - ➢説明変数X1とX2を重回帰で推定してみよう
    - ✓ 単回帰の推定結果と比較して、符号が変わっていないだろうか？
  - ➢X1とX2の重回帰は多重共線性のため、不自然な解釈になったり、有意にならなかったりすることがある
    - ✓ Variance Inflation Factor(VIF)
  - ➢2年間の平均成績と成績の増加分の説明変数を作り、重回帰してみよう
    - ✓ 生の変数を使うのではなく、解釈などの観点から加工変数を作ると、うまくいくことがある

重回帰モデルを走らせてみよう

- 重回帰モデルとは
  - ➢単回帰の説明変数が複数になった版
- Rの組み込みデータmtcarsについて
  - ➢1974年のMotor Tred US Magazine

分類	変数名	説明
目的変数	mpg	Miles / Gallon、燃費
説明変数	cyl	シリンダーの数
	disp	排気量
	hp	馬力
	drat	Rear axle ratio
	wt	重量
	qsec	1/4マイル時間
	vs	V/S
	am	変速機(0:オートマ、1:マニュアル)
	gear	前進ギア段数
	carb	キャブレターの数

重回帰モデルを走らせてみよう

- 重回帰モデルとは
  - 単回帰の説明変数が複数になった版

- Rの組み込みデータmtcarsについて
  - 1974年のMotor Tred US Magazine
  - 10変数の説明は、抜粋版でいいかも

- 演習 データmtcarsの基礎分析
  - データをロードしよう
  - データをざっと眺めてみよう
  - データの変数の数と件数を把握しよう
  - 要約統計量を求めよう
  - 散布図を描いてみよう
  - 相関係数を求めてみよう

- 問
  - 基礎分析から、mpgと相関の強い変数の候補を挙げてみよう