データ解析 第14回
回帰分析

北九州市立大学経済学部

齋藤 朗宏

今日の内容

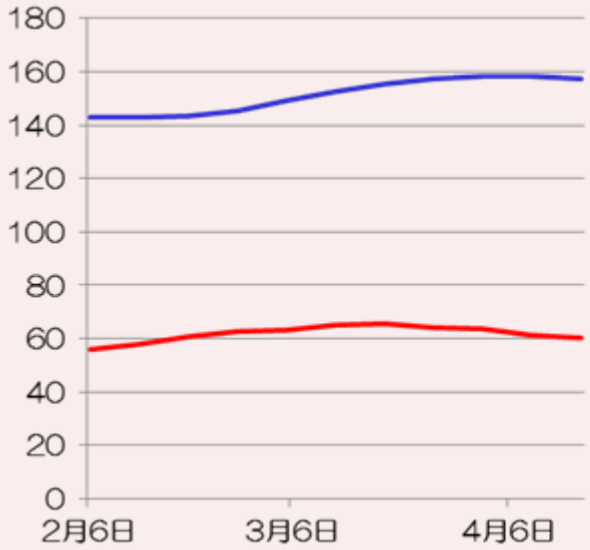
- 単回帰分析
- 重回帰分析

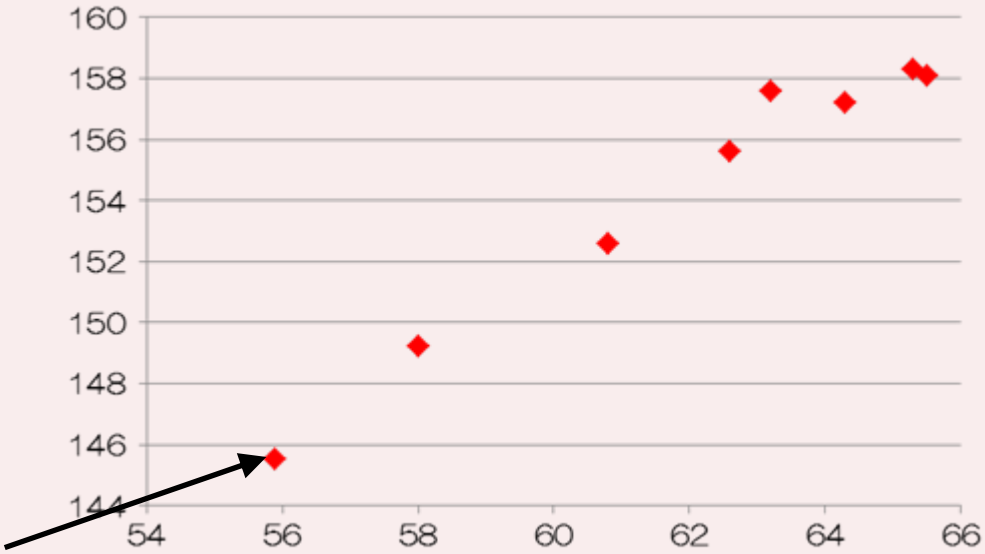
原油価格と石油製品価格

- 原油問題関係府省連絡会議Webサイト
  - http://www.cas.go.jp/jp/seisaku/genyu/

- 原油価格・石油製品価格の現状について
  - http://www.cas.go.jp/jp/seisaku/genyu/dai6/siryou2.pdf
- 石油製品価格，原油価格推移
  - 原油価格と石油製品価格は，概ね2～3週間程度ピークがずれている．
  - 原油価格を知っていれば，3週間後のガソリンの価格が予測できる！

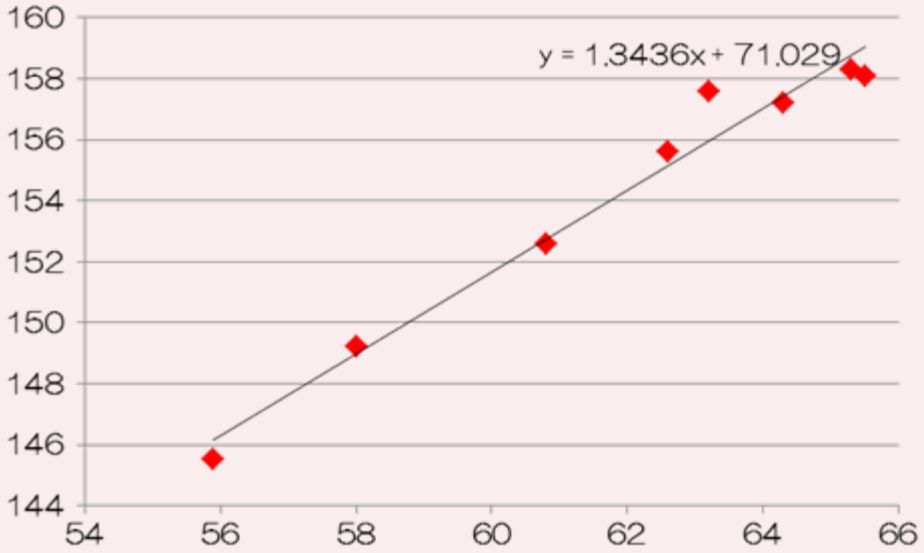
原油価格と石油製品価格





ある週の原油価格が55.9円，その3週間後のガソリンの価格が145.5円

価格の予想



ガソリンの価格＝3週間前の原油の価格
×1.34＋71.03

単回帰分析

回帰分析

- ある変数（予測変数・説明変数）の値から他の変数（目的変数・基準変数）の値を予測・説明する分析を回帰分析（regression analysis）と呼ぶ．特に，一次関数を用いて予測・説明を行う場合，線形回帰と呼ぶ．また，予測変数が１つの場合を単回帰分析，複数ある場合を重回帰分析と呼ぶ．
  - 「原油価格」から「ガソリン価格」を予測・説明．
    →単回帰分析
  - 「原油価格」と「原油在庫」から，「ガソリン価格」を予測・説明．
    →重回帰分析

単回帰分析

- 単回帰分析では，以下のようなモデル式をまず考える．

$$y_i = \alpha x_i + \beta + e_i$$

- この式は，前回説明した一元配置の分散分析と記号は少々異なるが，ほぼ同じ形である．分散分析と回帰分析は，線形モデルとして，数学的性質を同じくする．
- ここで，y_iは目的変数（ガソリン価格など），x_iは予測変数（原油価格など）であり，α, β はそれぞれ一次関数における傾き，切片であり，e_iは誤差である．

単回帰分析

$$y_i = \alpha x_i + \beta + e_i$$

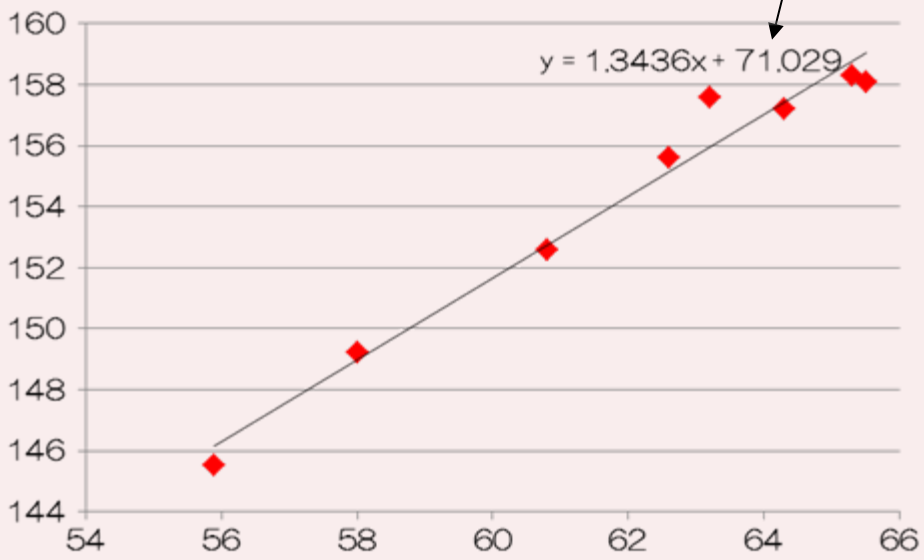
予測部分　　誤差

- この式は，x_iを利用してy_iを，$\hat{y}_i = \alpha x_i + \beta$という形で予測，説明する事を意味している．
- しかし，予測には誤差が必ず含まれる（同じドバイの原油価格であったとしても，3週間後に同じガソリン価格になるとは限らない）その誤差がe_iで表現される．
- この α, β の値をデータから推定することで，予測式を完成させる．

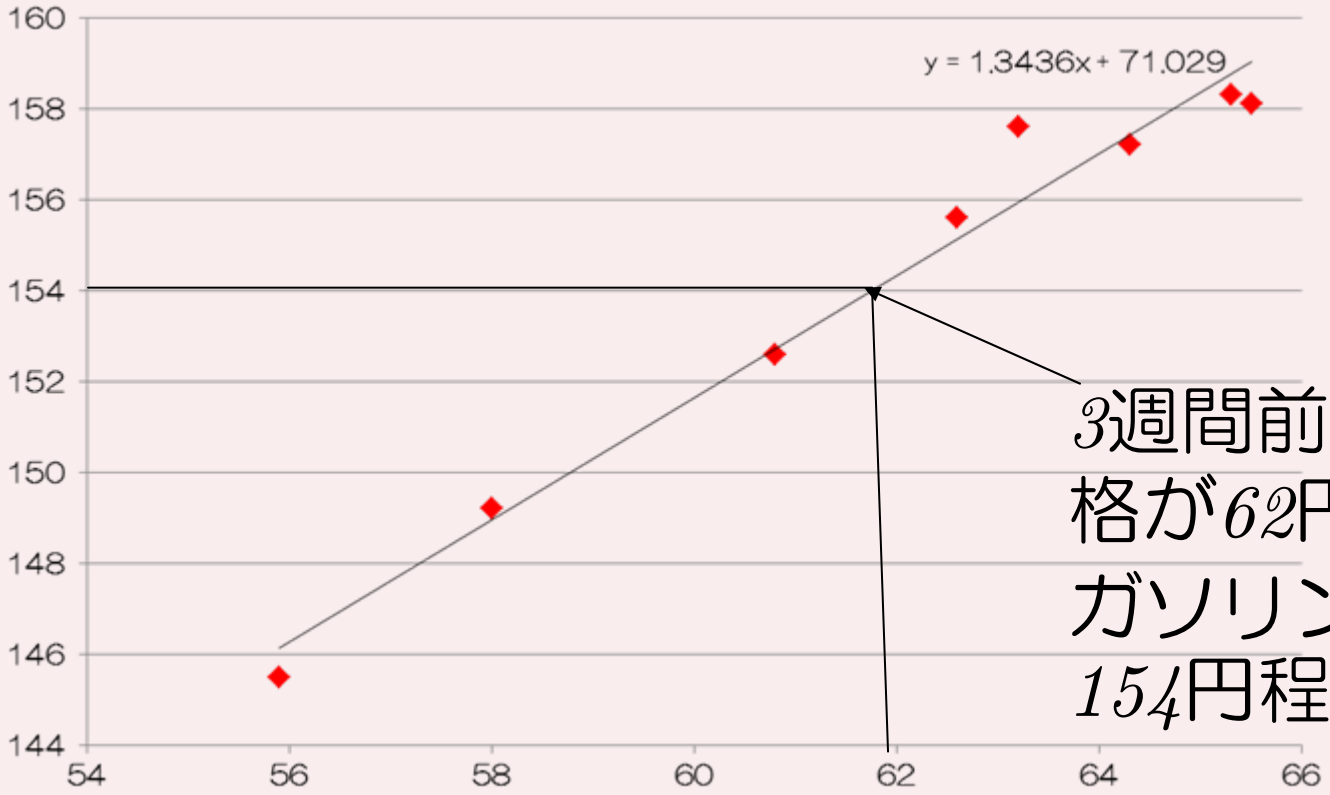
単回帰分析

回帰直線(予測式),

$$\hat{y} = \alpha x + \beta$$



単回帰分析

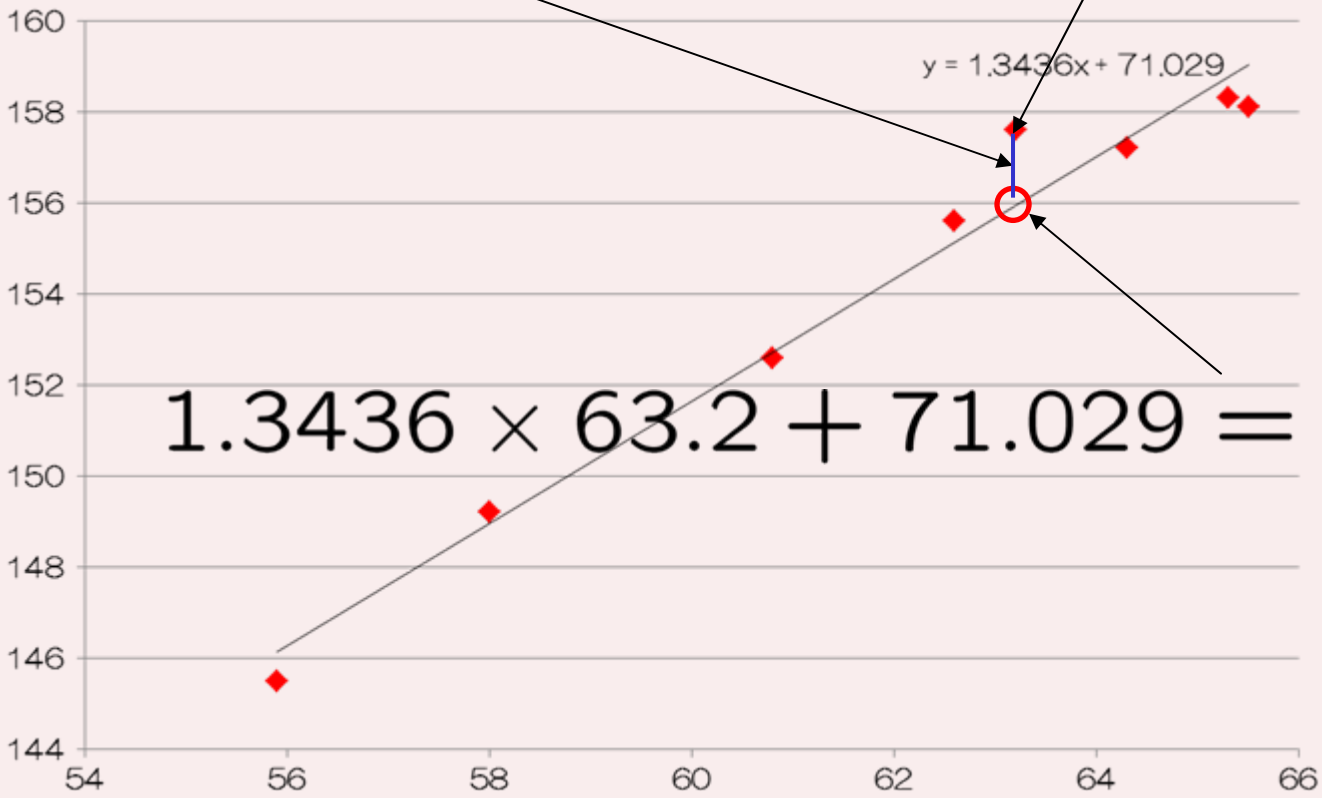


3週間前の原油価格が62円の場合，ガソリン価格は154円程度．

単回帰分析

$e_5 = 1.655$

$155.945 + 1.655 = 157.6$



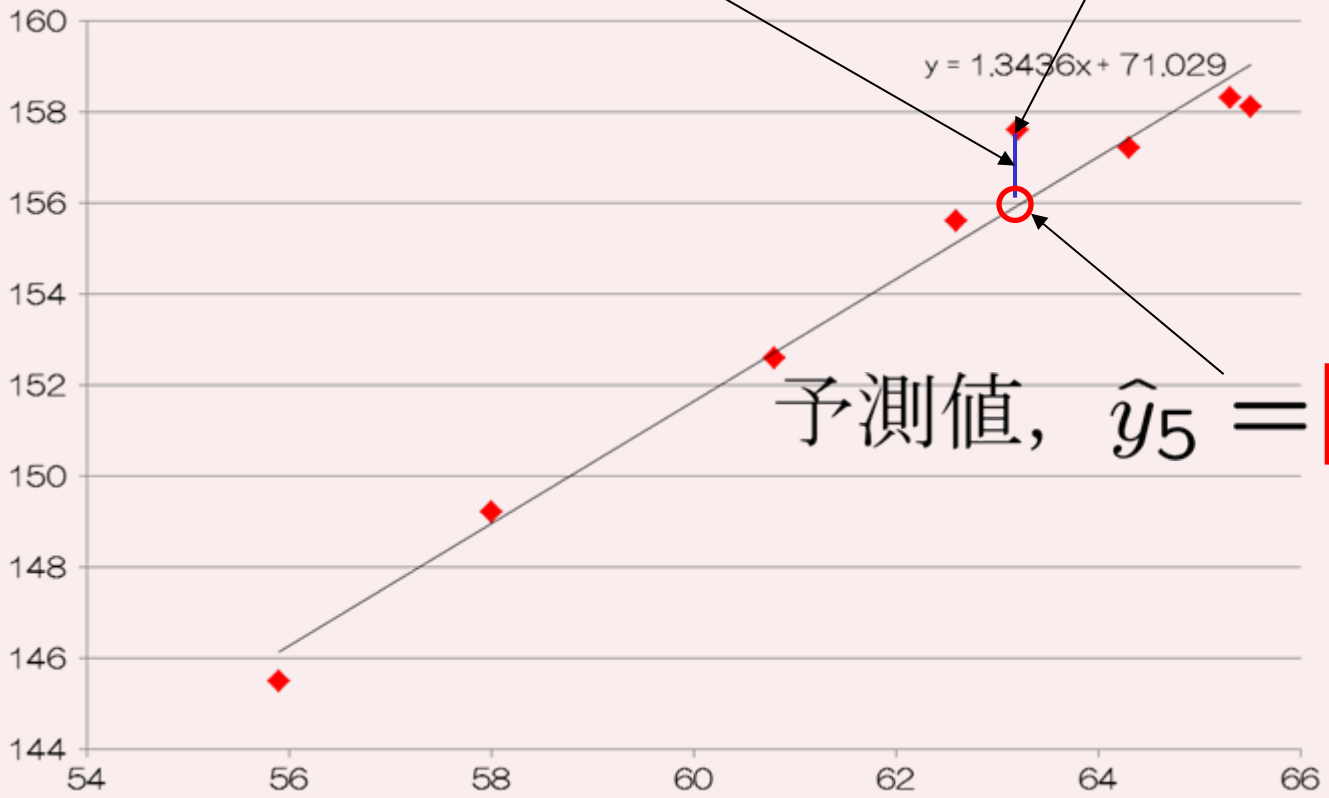
$1.3436 \times 63.2 + 71.029 = 155.945$

単回帰分析

予測値と実測値
との誤差(残差), e_5

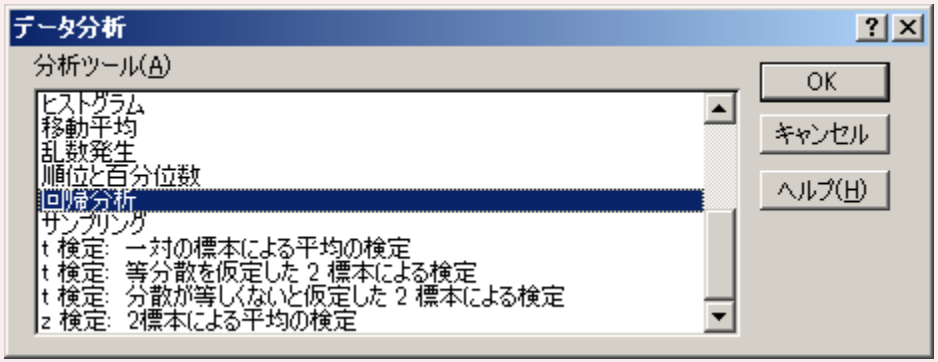
実測値,

$$y_5 = \alpha x_5 + \beta + e_5$$



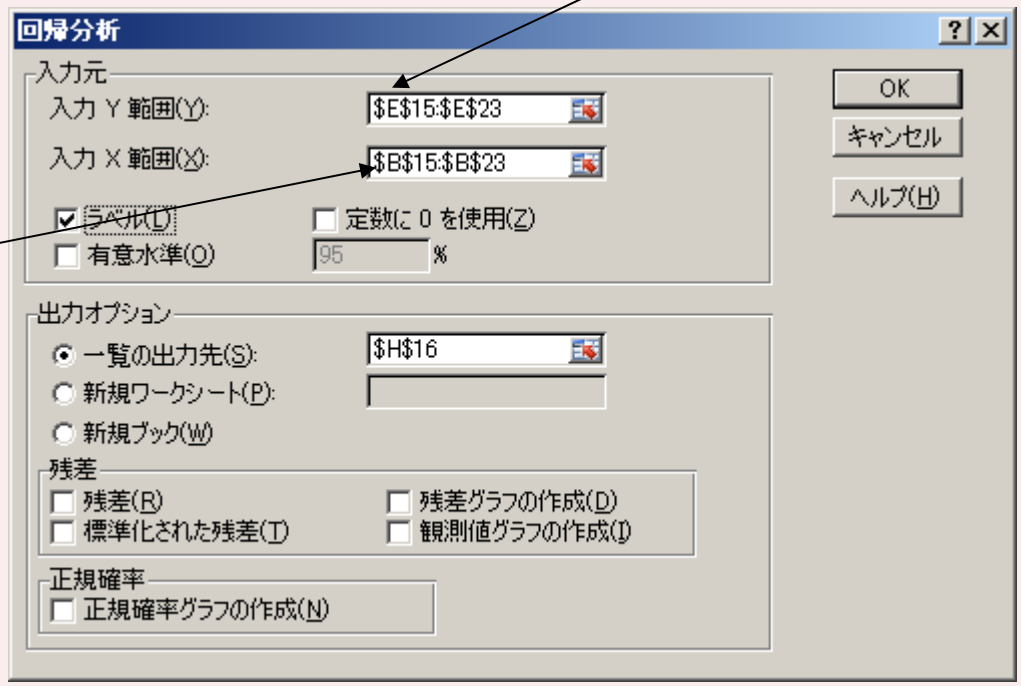
予測値, $\hat{y}_5 = \alpha x_5 + \beta$

Excelによる単回帰分析



目的変数

予測変数



単回帰分析の出力

概要	
回帰統計	
重相関 R	0.985403
重決定 R2	0.971018
補正 R2	0.966188
標準誤差	0.872934
観測数	8

予測値と実測値の間の相関係数

重相関Rの２乗．２乗しているので，正の値．データの変動の97%を予測式で説明できていると解釈する．決定係数と呼ばれる．
高い方が望ましい．

予測値と実測値との間の標準誤差
関数“$STEYX$”のヘルプなど参照．

単回帰分析の出力

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	153.1867	153.1867	201.0285423	7.69E-06
残差	6	4.572087	0.762014		
合計	7	157.7588			

分散分析と見方はほぼ一緒．有意Fが５％以下ならば，この回帰式は５％で有意と考える．
つまり，変数「ドバイの原油価格」の高低が，変数「ガソリン価格」の値と関係していると考える．

単回帰分析の出力

		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
β	切片	71.02859	5.878557	12.08266	1.95E-05	56.64428	85.41291	56.64428	85.41291
α	ドバイ原油	1.343566	0.094761	14.17845	7.69E-06	1.111694	1.575438	1.111694	1.575438

係数の推定値

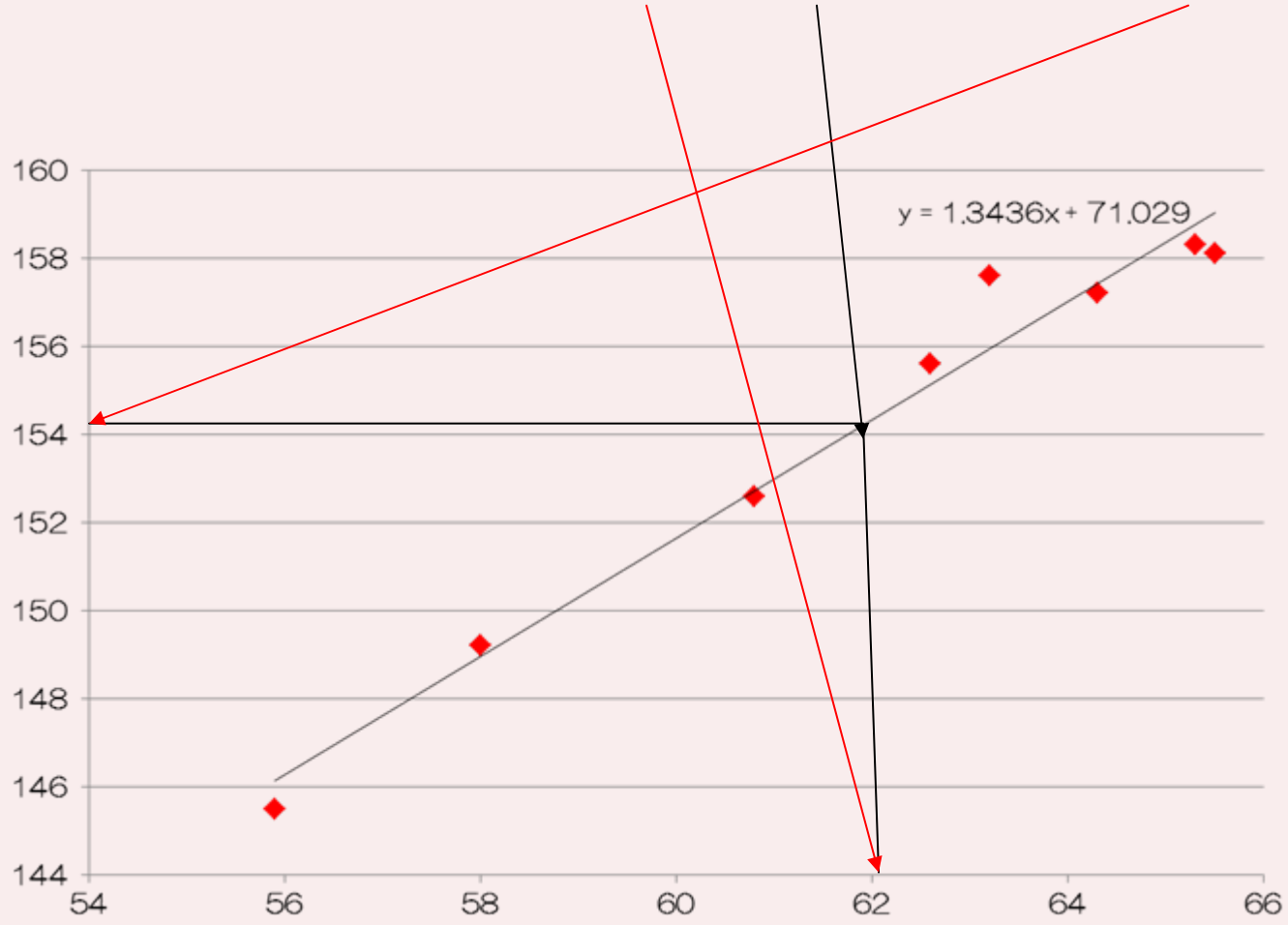
係数の信頼区間

$$\text{ガソリン価格予測値}(\hat{y}) = 1.34(\alpha) \times \text{原油価格}(x) + 71.03(\beta)$$

切片，傾きが0という帰無仮説に関する検定．
傾きに関する帰無仮説が棄却されれば，傾きが0ではない，即ち変数「ドバイの原油価格」は「ガソリン価格」の予測に意味があると考える．

単回帰分析

$$\text{ガソリン価格予測値}(\hat{y}) = 1.34 \times 62 + 71.03 = 154.11$$

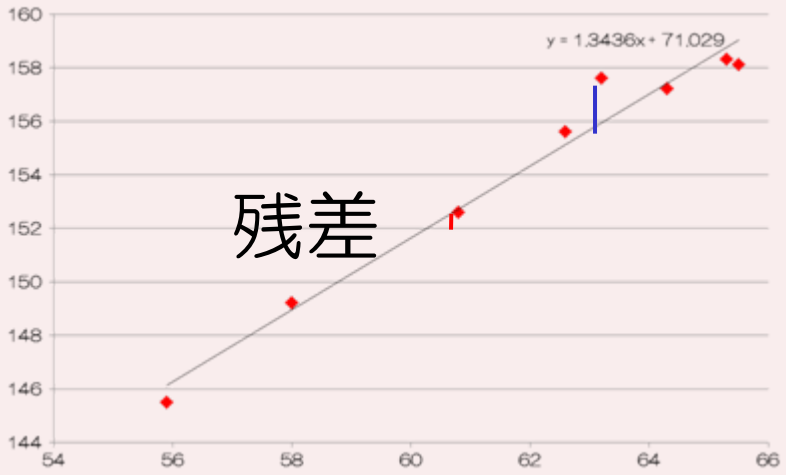


原油価格が62円の日がなくても，原油価格が62円であった場合のガソリン価格の予測値が計算できる．

単回帰分析の出力

残差出力

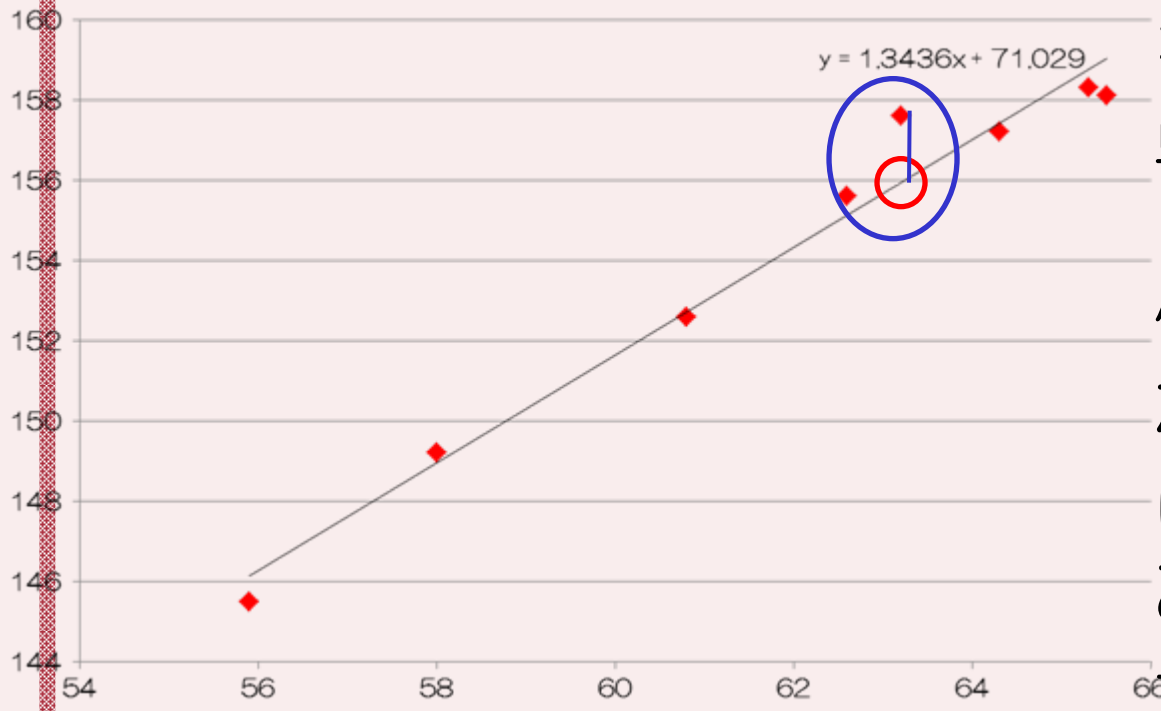
観測値	予測値: レギュラーガソリン	残差
1	146.1339265	−0.63393
2	148.9554148	0.244585
3	152.7173992	−0.1174
4	155.1358178	0.464182
5	155.9419573	1.658043
6	158.7634457	−0.46345
7	159.0321588	−0.93216
8	157.4198798	−0.21988



予測値と実測値との差が残差．概ね1円以下の範囲に収まっている．3番目，2月20日の残差が一番小さく5番目，3月5日の残差が一番大きい．

重回帰分析

重回帰分析



ドバイの原油価格を予測変数とした単回帰分析でかなりの部分ガソリン価格は説明できたが，3月5日の価格には若干の誤差があった．そこで，予測変数「アメリカの原油在庫の変化」を加えることを考える．

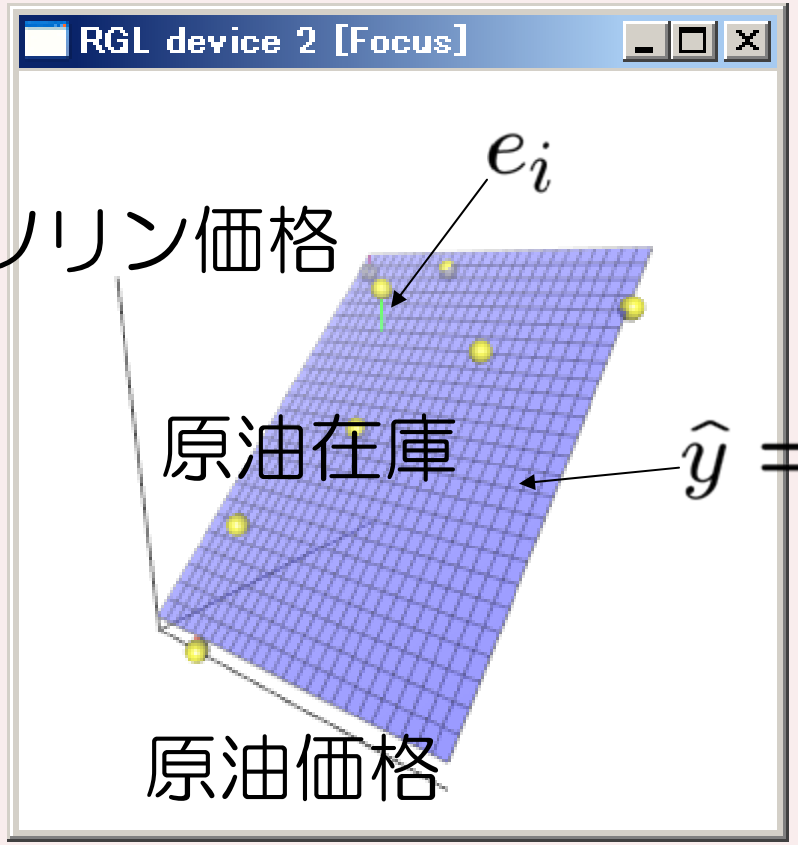
重回帰分析

- 重回帰分析では，以下のようなモデル式をまず考える．

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \beta + e_i$$

- 目的変数，切片の部分に違いはない．予測部分と誤差との関係も単回帰分析と同じ．
- 予測変数が２つになったのに伴い，α, x の部分が変化している．x_{1i} は，１つ目の予測変数の i 番目の値（「2月6日」の「原油価格」など），α_1 はそれにかかる傾き，x_{2i} は２つ目の予測変数の i 番目の値，α_2 はそれにかかる傾きである．
- 予測変数が３つ以上でも，同じようにモデル化．

重回帰分析

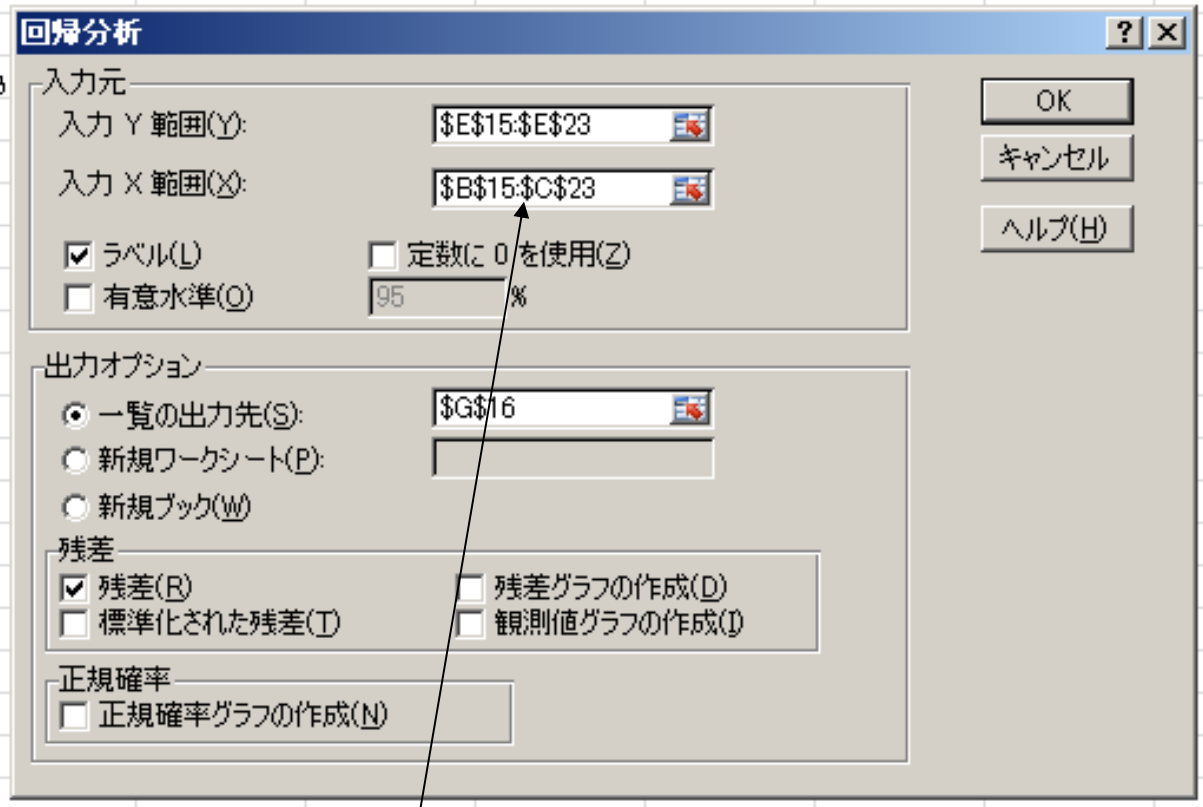


$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \beta$$

予測変数が２つになったことで，回帰直線は回帰平面となる．

Ｅｘｃｅｌによる重回帰分析

	ドバイ原油	米原油在庫	為
2月6日	55.9	30.4	
2月13日	58	−17.1	
2月20日	60.8	163.3	
2月27日	62.6	416	
3月5日	63.2	83.2	
3月12日	65.3	175	
3月19日	65.5	−116.2	
3月26日	64.3	710.2	



予測変数が２つに

Ｅｘｃｅｌによる重回帰分析

回帰統計	
重相関 R	0.985403
重決定 R2	0.971018
補正 R2	0.966188
標準誤差	0.872934
観測数	8

単回帰分析

回帰統計	
重相関 R	0.985584
重決定 R2	0.971375
補正 R2	0.959925
標準誤差	0.950353
観測数	8

重回帰分析

１行目を見ると，予測値と実測値の相関は向上している．それに伴い，決定係数の値も向上している．しかし，決定係数は，予測変数が多ければ多いほどいい値が出る傾向がある．それを修正したのが補正Ｒ２の部分である．補正$R2$は若干悪化しており，ここから，予測変数を増やしてもあまり意味がないと解釈できる．

Ｅｘｃｅｌによる重回帰分析

		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
β	切片	71.45729	6.626532	10.78351	0.000119	54.42325	88.49134	54.42325	88.49134
$\alpha 1$	ドバイ原油	1.335618	0.107971	12.37015	6.12E-05	1.05807	1.613166	1.05807	1.613166
$\alpha 2$	米原油在庫	0.000353	0.001413	0.249514	0.812889	-0.00328	0.003985	-0.00328	0.003985

係数の推定値

係数の信頼区間

$$\text{ガソリン価格予測値}(\hat{y}) = 1.34(\alpha_1) \times \text{原油価格}(x_1) + 0.00035(\alpha_2) \times \text{米原油在庫}(x_2) + 71.46(\beta)$$

Ｐ値から，「米原油在庫」の係数が０であるという帰無仮説は棄却できない．つまり，「米原油在庫」を予測変数として使うことが合理的であるという証拠はないことになる．

多重共線性問題

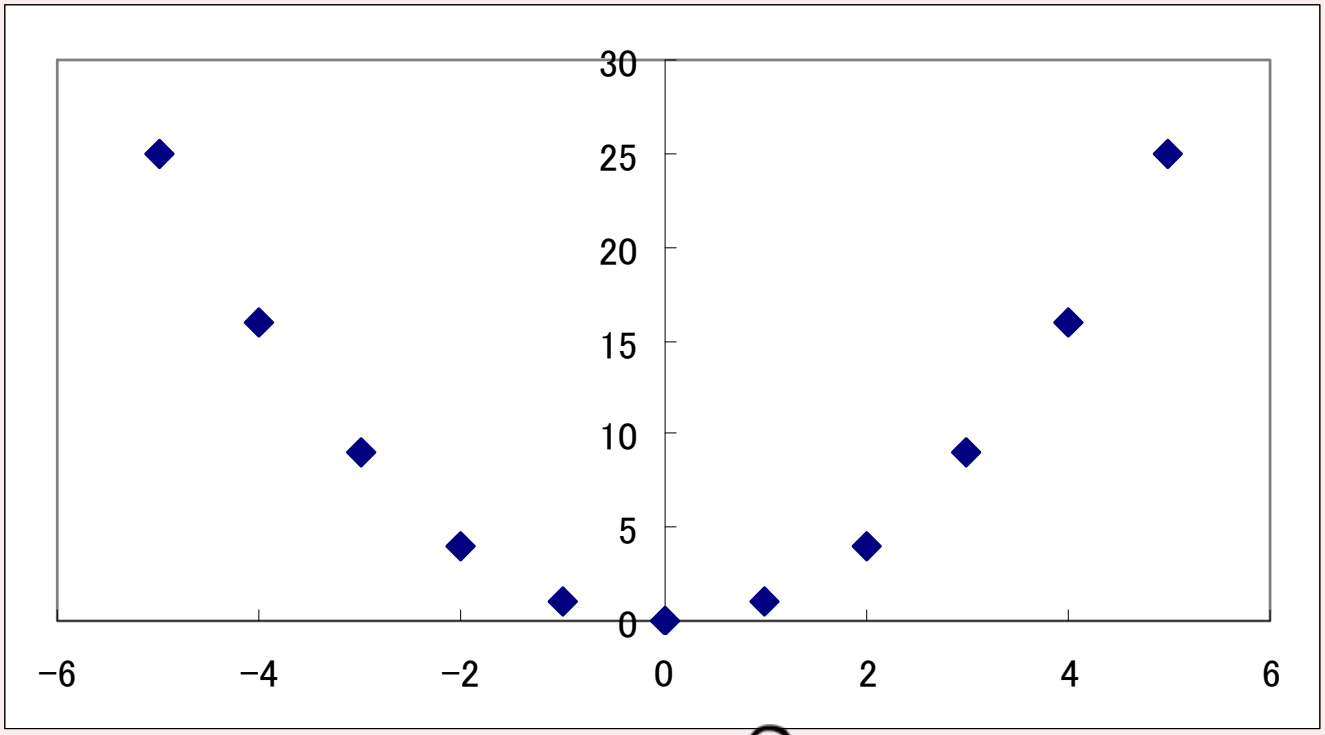
	ドバイ原油	米原油在庫	為替相場(米ドル)	レギュラーガソリン
ドバイ原油	1			
米原油在庫	0.295021	1		
為替相場(米ドル)	0.985403	0.32456279	1	
レギュラーガソリン	0.985403	0.30875365	0.964514974	1

- 変数「米原油在庫」を用いたが，あまり意味はなかった．ガソリン価格との相関を見ると，被為替相場の方が相関は高かったので，そちらを使った方が良かったのでは？
- それはダメ．何故なら，ドバイ原油と為替相場，予測変数間の相関が高すぎるから．

多重共線性問題

- 重回帰分析において，予測変数間の相関があまりに高すぎる場合，多重共線性問題という問題が起こる可能性があることが知られている．これは，α, β 等の推定値が不安定になる（僅かなデータの変化で大きく変わってしまう）問題である．これを避けるためには予測変数間で相関の高い変数はあまり使わないこと．
- そもそも，「ドバイ原油」の情報と「為替相場」の情報はかなり重複しているので，「為替相場」の情報を加えたところで，推定精度はあまり向上しない．

回帰分析の問題点



- このデータには，$y = x^2$ という明確な関連性があるが，相関係数を算出すると $\rho = 0$ となる．こういったデータには，これまでの回帰分析は役に立たない．

回帰分析の問題点

回帰統計	
重相関 R	0
重決定 R2	0
補正 R2	-0.11111111
標準誤差	9.763879011
観測数	11

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	0	0	0	1
残差	9	858	95.33333		
合計	10	858			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	10	2.94392	3.396831	0.00791296	3.34039	16.65961	3.34039	16.65961
X 値 1	0	0.930949	0	1	-2.10595	2.105954	-2.10595	2.105954

- 全く予測ができていない．これは，これまで述べた回帰分析が，相関係数同様直線的な関係しか表現できないためである．

実習

- 配布したアデレード大学のデータを用い，手の幅から身長を予測するモデルを作成せよ．また，Excelの出力から予測式を作成し，手の幅が17cmであるときの身長の予測値を求めよ．
- 手の幅と心拍数から身長を予測するモデルを作成せよ．また，Excelの出力から予測式を作成し，手の幅が17cm，心拍数が90であるときの身長の予測値を求めよ．
- このデータについて，単回帰分析と重回帰分析ではどちらが適切だったか検討せよ．