

回帰分析

予測と発展的なモデル

村田 昇

講義概要

- 第1回: 回帰モデルの考え方と推定
- 第2回: モデルの評価
- 第3回: モデルによる予測と発展的なモデル

回帰分析の復習

線形回帰モデル

- 目的変数を説明変数で説明する関係式を構成
 - 説明変数: x_1, \dots, x_p (p次元)
 - 目的変数: y (1次元)
- 回帰係数 $\beta_0, \beta_1, \dots, \beta_p$ を用いた一次式

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- 誤差項を含む確率モデルで観測データを表現

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i \quad (i = 1, \dots, n)$$

問題設定

- 確率モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{確率分布}$$

- 式の評価: 残差平方和の最小化による推定

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

解とその一意性

- 解の条件: 正規方程式

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$$

- 解の一意性: Gram行列 $\mathbf{X}^\top \mathbf{X}$ が正則

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

解析の事例

気温に影響を与える要因の分析

- データの概要

日付	気温	降雨	日射	降雪	風向	風速	気圧	湿度	雲量
2023-09-01	29.2	0.0	24.01	0	SSE	4.3	1012.1	71	2.0
2023-09-02	29.6	0.0	22.07	0	SSE	3.1	1010.3	72	8.0
2023-09-03	29.1	3.5	18.64	0	ENE	2.8	1010.6	74	9.3
2023-09-04	26.1	34.0	7.48	0	N	2.6	1007.5	96	10.0
2023-09-05	29.3	0.0	22.58	0	S	3.5	1005.2	77	3.5
2023-09-06	27.5	0.5	13.17	0	SSW	2.6	1003.6	79	10.0
2023-09-07	27.0	0.5	11.01	0	ENE	2.5	1007.9	72	10.0
2023-09-08	21.9	107.5	2.10	0	NW	3.4	1007.8	98	10.0
2023-09-09	24.8	1.0	8.81	0	S	2.2	1006.8	93	7.5
2023-09-10	27.8	0.0	17.57	0	S	3.1	1009.1	83	6.3
2023-09-11	28.1	0.0	17.19	0	SSE	3.1	1010.1	79	9.0
2023-09-12	27.7	0.0	20.02	0	SSE	2.8	1010.0	76	4.8
2023-09-13	28.0	0.0	22.00	0	SE	2.4	1010.9	74	4.5
2023-09-14	28.2	0.0	14.54	0	SSE	2.8	1009.9	80	7.0
2023-09-15	27.4	10.5	9.21	0	NE	2.0	1010.9	88	8.5
2023-09-16	27.9	0.0	11.78	0	SSE	2.0	1011.5	86	10.0
2023-09-17	28.7	0.0	14.84	0	S	3.2	1011.5	80	4.0
2023-09-18	28.9	0.0	19.59	0	S	4.2	1011.6	74	1.8
2023-09-19	29.0	0.0	19.93	0	S	3.3	1010.1	72	2.3
2023-09-20	27.2	6.0	10.65	0	N	1.9	1009.3	82	8.3
2023-09-21	26.7	2.0	6.65	0	S	4.1	1006.7	87	9.5
2023-09-22	24.8	59.5	6.83	0	ENE	2.5	1008.1	93	10.0
2023-09-23	22.1	4.0	4.48	0	NE	2.6	1012.5	89	10.0
2023-09-24	22.2	0.0	15.81	0	N	3.0	1017.2	67	7.0
2023-09-25	22.4	0.0	15.49	0	N	2.5	1017.1	69	6.5
2023-09-26	24.6	0.0	16.08	0	NNW	2.0	1012.7	71	6.0
2023-09-27	25.3	0.0	11.59	0	SSE	1.9	1008.1	81	9.0
2023-09-28	27.4	0.0	14.03	0	ESE	1.9	1004.7	79	5.8
2023-09-29	26.3	0.0	10.11	0	SSE	3.0	1009.0	75	8.5
2023-09-30	25.6	0.0	7.98	0	S	2.5	1007.5	77	7.0

- 気温を説明する 5 種類の線形回帰モデルを検討
 - モデル 1: 気温 = F(気圧)
 - モデル 2: 気温 = F(日射)
 - モデル 3: 気温 = F(気圧, 日射)
 - モデル 4: 気温 = F(気圧, 日射, 湿度)
 - モデル 5: 気温 = F(気圧, 日射, 雲量)

分析の視覚化

- 関連するデータの散布図

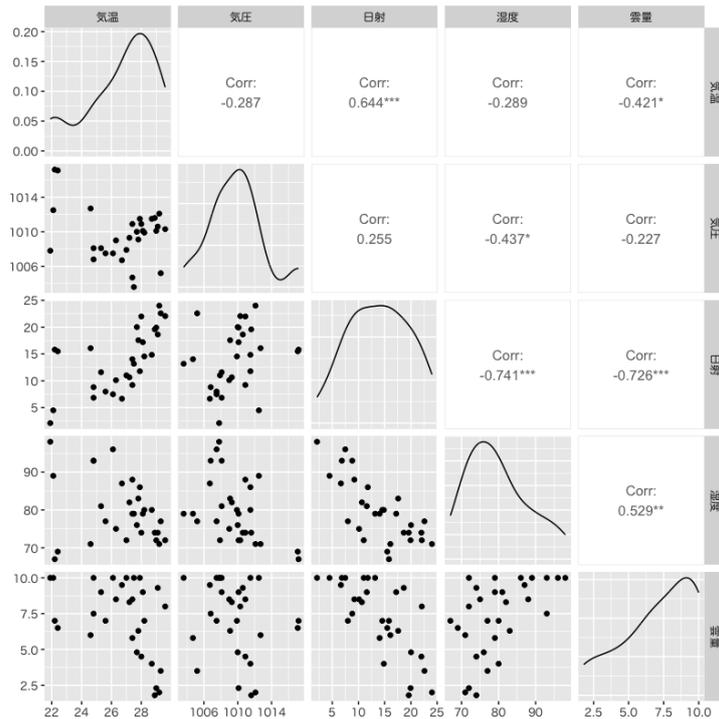


Figure 1: 散布図

- 観測値とあてはめ値の比較

寄与率

- 決定係数 (R-squared)

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- 自由度調整済み決定係数 (adjusted R-squared)

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- 不偏分散で補正

モデルの評価

- 決定係数 (R^2 ・Adjusted R^2) によるモデルの比較

Characteristic	モデル 1		モデル 2		モデル 3		モデル 4		モデル 5	
	Beta	95% CI ^l								
気圧	-0.21	-0.49, 0.06			-0.36	-0.55, -0.18	-0.32	-0.53, -0.12	-0.36	-0.55, -0.17
日射			0.25	0.14, 0.37	0.30	0.20, 0.40	0.35	0.21, 0.49	0.32	0.18, 0.46
湿度							0.05	-0.06, 0.16		
雲量									0.05	-0.26, 0.36
R ²	0.082		0.414		0.632		0.644		0.633	

Adjusted R² 0.049 0.393 0.604 0.603 0.591

^lCI = Confidence Interval

F 統計量による検定

- 説明変数のうち 1 つでも役に立つか否かを検定する
 - 帰無仮説 $H_0: \beta_1 = \dots = \beta_p = 0$
 - 対立仮説 $H_1: \exists j \beta_j \neq 0$ (少なくとも 1 つは役に立つ)
- F 統計量: 決定係数 (または残差) を用いて計算

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

- p 値: 自由度 $p, n-p-1$ の F 分布で計算

モデルの評価

- F 統計量によるモデルの比較

Characteristic	モデル 1		モデル 2		モデル 3		モデル 4		モデル 5	
	Beta	95% CI ^l	Beta	95% CI ^l	Beta	95% CI ^l	Beta	95% CI ^l	Beta	95% CI ^l
気圧	-0.21	-0.49, 0.06			-0.36	-0.55, -0.18	-0.32	-0.53, -0.12	-0.36	-0.55, -0.17
日射			0.25	0.14, 0.37	0.30	0.20, 0.40	0.35	0.21, 0.49	0.32	0.18, 0.46
湿度							0.05	-0.06, 0.16		
雲量									0.05	-0.26, 0.36
R ²	0.082		0.414		0.632		0.644		0.633	
Statistic	2.51		19.8		23.1		15.7		14.9	
p-value	0.12		<0.001		<0.001		<0.001		<0.001	

^lCI = Confidence Interval

t 統計量による検定

- 回帰係数 β_j が回帰式に寄与するか否かを検定する
 - 帰無仮説 $H_0: \beta_j = 0$
 - 対立仮説 $H_1: \beta_j \neq 0$ (β_j は役に立つ)
- t 統計量: 各係数ごと, ζ は $(X^T X)^{-1}$ の対角成分

$$t = \frac{\hat{\beta}_j}{\hat{\sigma} \zeta_j}$$

- p 値: 自由度 $n-p-1$ の t 分布を用いて計算

モデルの評価

- t 統計量によるモデルの比較

Characteristic	モデル 1				モデル 2				モデル 3				Beta
	Beta	SE ^l	Statistic	p-value	Beta	SE ^l	Statistic	p-value	Beta	SE ^l	Statistic	p-value	
(Intercept)	243	137	1.78	0.086	23	0.855	27.1	<0.001	386	91.0	4.25	<0.001	346

気圧	-0.21	0.135	-1.58	0.12						-0.36	0.090	-3.99	<0.001	-0.32
日射					0.25	0.057	4.45	<0.001	0.30	0.048	6.35	<0.001	<0.001	0.35
湿度														0.05
雲量														

¹SE = Standard Error

診断プロットによる評価

- モデル 2
- モデル 3
- モデル 4

回帰モデルによる予測

予測

- 新しいデータ (説明変数) x に対する **予測値**

$$\hat{y} = (1, x^T)\hat{\beta}, \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = w(x)^T y, \quad w(x)^T = (1, x^T)(X^T X)^{-1} X^T$$

- 重みは元データと新規データの説明変数で決定

予測値の性質

- 推定量は以下の性質をもつ多変量正規分布

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \beta \\ \text{Cov}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

- この性質を利用して以下の3つの値の違いを評価

$$\begin{aligned} \hat{y} &= (1, x^T)\hat{\beta} && \text{(回帰式による予測値)} \\ \tilde{y} &= (1, x^T)\beta && \text{(最適な予測値)} \\ y &= (1, x^T)\beta + \epsilon && \text{(観測値)} \end{aligned}$$

- \hat{y} と y は独立な正規分布に従うことに注意

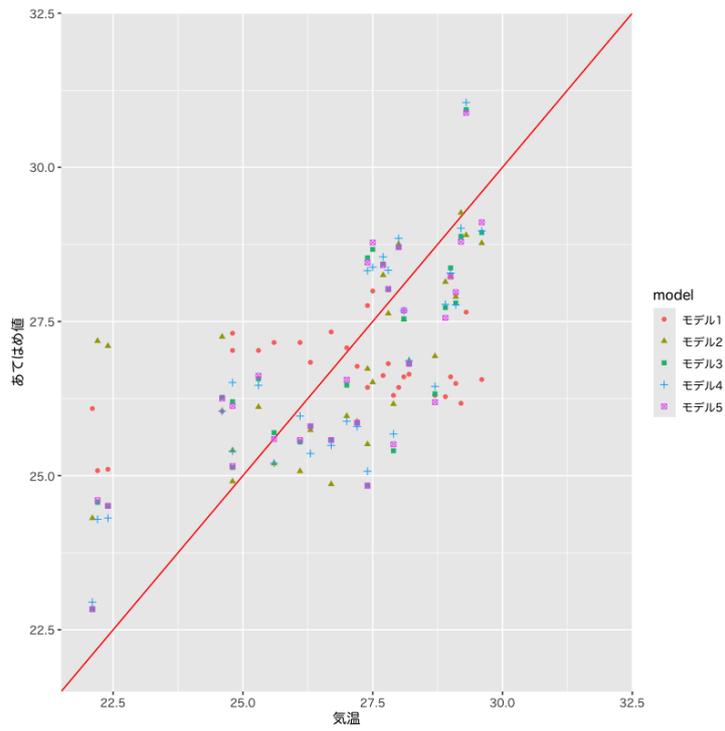


Figure 2: モデルの比較

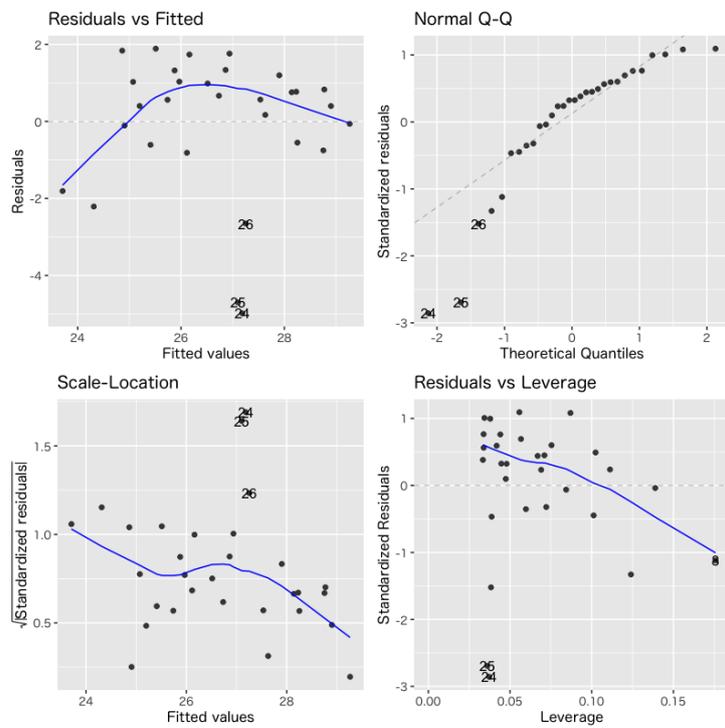


Figure 3: モデル 3 の診断

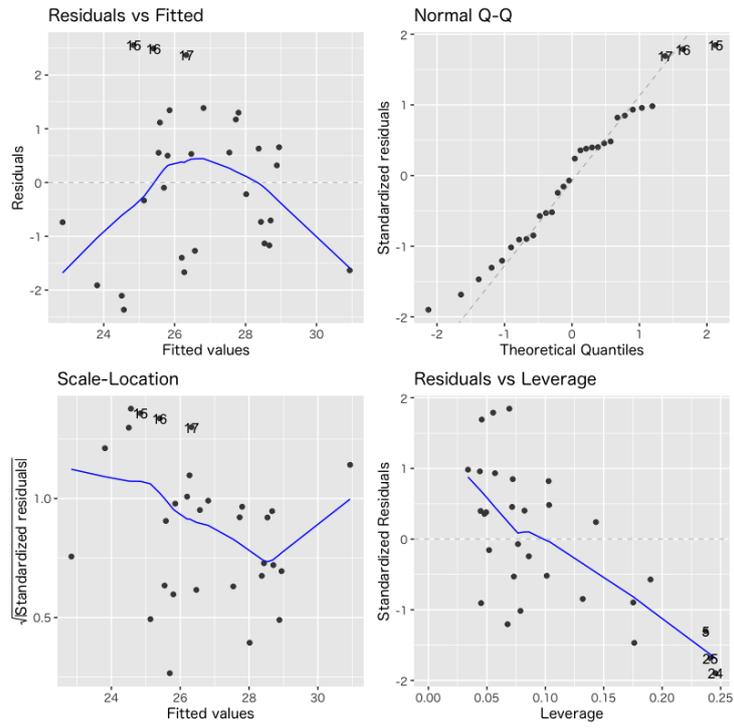


Figure 4: モデル 3 の診断

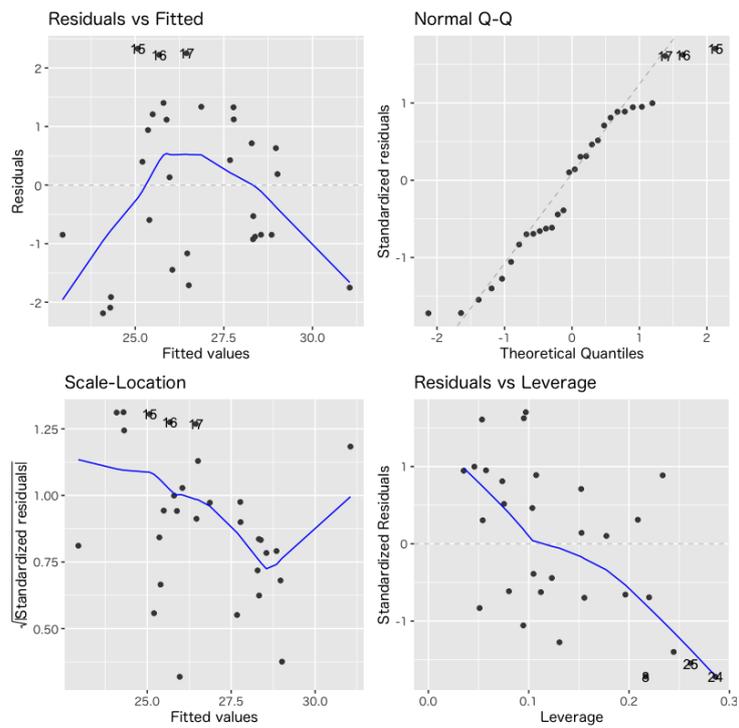


Figure 5: モデル 4 の診断

信頼区間

最適な予測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}\mathbb{E}[\tilde{y} - \hat{y}] &= (1, \mathbf{x}^\top)\boldsymbol{\beta} - (1, \mathbf{x}^\top)\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\ \text{Var}(\tilde{y} - \hat{y}) &= \underbrace{\sigma^2(1, \mathbf{x}^\top)(X^\top X)^{-1}(1, \mathbf{x}^\top)^\top}_{\hat{\boldsymbol{\beta}} \text{ の推定誤差による分散}} = \sigma^2\gamma_c(\mathbf{x})^2\end{aligned}$$

- 標準化による表現

$$\frac{\tilde{y} - \hat{y}}{\sigma\gamma_c(\mathbf{x})} \sim \mathcal{N}(0, 1)$$

信頼区間

- 未知の分散を不偏分散で推定

$$Z = \frac{\tilde{y} - \hat{y}}{\hat{\sigma}\gamma_c(\mathbf{x})} \sim \mathcal{T}(n-p-1) \quad (t \text{ 分布})$$

- 確率 α の信頼区間

$$I_\alpha^c = (\hat{y} - C_\alpha \hat{\sigma}\gamma_c(\mathbf{x}), \hat{y} + C_\alpha \hat{\sigma}\gamma_c(\mathbf{x}))$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

– 最適な予測値 \hat{y} が入ることが期待される区間

予測区間

観測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}\mathbb{E}[y - \hat{y}] &= (1, \mathbf{x}^\top)\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] - (1, \mathbf{x}^\top)\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\ \text{Var}(y - \hat{y}) &= \underbrace{\sigma^2(1, \mathbf{x}^\top)(X^\top X)^{-1}(1, \mathbf{x}^\top)^\top}_{\hat{\boldsymbol{\beta}} \text{ の推定誤差による分散}} + \underbrace{\sigma^2}_{\text{誤差の分散}} = \sigma^2\gamma_p(\mathbf{x})^2\end{aligned}$$

- 標準化による表現

$$\frac{y - \hat{y}}{\sigma\gamma_p(\mathbf{x})} \sim \mathcal{N}(0, 1)$$

予測区間

- 未知の分散を不偏分散で推定

$$Z = \frac{y - \hat{y}}{\hat{\sigma}\gamma_p(\mathbf{x})} \sim \mathcal{T}(n-p-1) \quad (t \text{ 分布})$$

- 確率 α の予測区間

$$I_\alpha^p = (\hat{y} - C_\alpha \hat{\sigma}\gamma_p(\mathbf{x}), \hat{y} + C_\alpha \hat{\sigma}\gamma_p(\mathbf{x}))$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

- 観測値 y が入ることが期待される区間
- $\gamma_p > \gamma_c$ なので信頼区間より広くなる

実習

R : 予測値と区間推定

- 関数 `stats::predict()` を用いた予測

```
#!/ モデルの作成
toy_train <- tibble(x1 = ..., x2 = ..., y = ...)
toy_lm <- lm(y ~ x1 + x2, data = toy_train)
toy_train_fitted <- predict(toy_lm) # あてはめ値の計算
#!/ 新しいデータの予測
toy_test <- tibble(x1 = ..., x2 = ...) # 予測したいデータの説明変数
toy_test_fitted <- predict(toy_lm, # 予測値の計算
                          newdata = toy_test)
toy_test_conf <- predict(toy_lm, newdata = toy_test, # 信頼区間
                        interval = "confidence", level = 0.95)
toy_test_pred <- predict(toy_lm, newdata = toy_test, # 予測区間
                        interval = "prediction", level = 0.95)
#!/ 信頼区間, 予測区間の水準の既定値は 0.95
```

- 関数 `broom::augment()` を用いた予測

```
#!/ モデルの作成
toy_train <- tibble(x1 = ..., x2 = ..., y = ...)
toy_lm <- lm(y ~ x1 + x2, data = toy_train)
toy_train_fitted <- augment(toy_lm) # あてはめ値の計算
#!/ 新しいデータの予測
toy_test <- tibble(x1 = ..., x2 = ...) # 予測したいデータの説明変数
toy_test_fitted <- augment(toy_lm, # 予測値の計算
                          newdata = toy_test)
toy_test_conf <- augment(toy_lm, newdata = toy_test, # 信頼区間
                        interval = "confidence", conf.level = 0.95)
toy_test_pred <- augment(toy_lm, newdata = toy_test, # 予測区間
                        interval = "prediction", conf.level = 0.95)
#!/ 信頼区間, 予測区間の水準の既定値は 0.95
```

R : モデルからの予測

- 東京の気候データによる例

```

#' 9,10月のデータでモデルを構築し, 8,11月のデータを予測
#' データの整理
tw_data <- read_csv("data/tokyo_weather.csv")
tw_train <- tw_data |> # モデル推定用データ
  filter(month %in% c(9,10)) # %in% は集合に含むかどうかを判定
tw_test <- tw_data |> # 予測用データ
  filter(month %in% c(8,11))
#' モデルの構築
tw_model <- temp ~ solar + press # モデルの定義
tw_lm <- lm(tw_model, data = tw_train) # モデルの推定
tidy(tw_lm) # 回帰係数の評価
glance(tw_lm) # モデルの評価
#' あてはめ値と予測値の計算
tw_train_fitted <- augment(tw_lm, newdata = tw_train) # あてはめ値
tw_test_fitted <- augment(tw_lm, newdata = tw_test) # 予測値

```

- グラフ表示の例

```

#' 予測結果を図示
bind_rows(tw_train_fitted, tw_test_fitted) |> # 2つのデータフレームを結合
mutate(month = as_factor(month)) |> # 月を因子化して表示に利用
ggplot(aes(x = .fitted, y = temp)) +
  geom_point(aes(colour = month, shape = month)) + # 月ごとに色と形を変える
  geom_abline(slope = 1, intercept = 0, # 予測が完全に正しい場合のガイド線
             colour = "gray") +
  labs(x = "fitted", y = "observed")

```

R : 区間表示のための関数

- 関数 `ggplot2::geom_errorbar()` : 区間の表示

```

geom_errorbar(
  mapping = NULL,
  data = NULL,
  stat = "identity",
  position = "identity",
  ...,
  na.rm = FALSE,
  orientation = NA,
  show.legend = NA,
  inherit.aes = TRUE
)
#' mapping: 区間を表すために xmin,xmax または ymin,ymax を与える
#' data: データフレーム
#' ...: その他の描画オプション
#' orientation: 特別な場合に指定 (一般に向きは mapping で自動的決定)
#' 詳細は '?ggplot2::geom_errorbar' を参照

```

- 関数 `broom::augment()` の場合は 'lower/upper' を用いる
- 関数 `stats::predict()` の場合は 'lwr/upr' を用いる

練習問題

- 東京の気候データを用いて以下の実験を試みなさい
 - 8月のデータで回帰式を推定する
 - 上記のモデルで9月のデータを予測する

```

#' 特定の月のデータを取り出すには, 例えば以下のようにすればよい
tw_data <- read_csv("data/tokyo_weather.csv")
tw_train <- tw_data |> filter(month == 8) # 単一の数字と比較

```

```
tw_test <- tw_data |> filter(month %in% c(9,10)) # 集合と比較
```

発展的なモデル

非線形性を含むモデル

- 目的変数 y
- 説明変数 x_1, \dots, x_p
- 説明変数の追加で対応可能
 - 交互作用 (交差項): $x_i x_j$ のような説明変数の積
 - 非線形変換: $\log(x_k)$ のような関数による変換

カテゴリカル変数を含むモデル

- 数値ではないデータ
 - 悪性良性
 - 血液型
- 適切な方法で数値に変換して対応:
 - 2 値の場合は 1,0 (真, 偽) を割り当てる
 - * 悪性: 1
 - * 良性: 0
 - 3 値以上の場合は **ダミー変数** を利用する (カテゴリ数-1 個)
 - * A 型: (1,0,0)
 - * B 型: (0,1,0)
 - * O 型: (0,0,1)
 - * AB 型: (0,0,0)

実習

R: 線形でないモデル式の書き方

- 交互作用を記述するためには特殊な記法がある
- 非線形変換はそのまま関数を記述すればよい
- 1 つの変数の多項式は関数 $I()$ を用いる

```
#' 目的変数 Y, 説明変数 X1, X2, X3
#' 交互作用を含む式 (formula) の書き方
Y ~ X1 + X1:X2      # X1 + X1*X2
Y ~ X1 * X2         # X1 + X2 + X1*X2
Y ~ (X1 + X2 + X3)^2 # X1 + X2 + X3 + X1*X2 + X2*X3 + X3*X1
#' 非線形変換を含む式 (formula) の書き方
Y ~ f(X1)           # f(X1) (fは任意の関数)
Y ~ X1 + I(X2^2)    # X1 + X2^2
```

R: カテゴリカル変数の取り扱い

- 何も宣言しなくても通常は適切に対応してくれる
- 陽に扱う場合は関数 `factor()` を利用する

```

#' factor属性の与え方
X <- c("A", "S", "A", "B", "D")
Y <- c(85, 100, 80, 70, 30)
toy_data1 <- tibble(X, Y)
toy_data2 <- toy_data1 |> # 因子化
  mutate(X2 = factor(X)) # 関数 as_factor() を用いてもよい
glimpse(toy_data2) # 作成したデータフレームの素性を見る (pillar::glimpse())
toy_data3 <- toy_data2 |> # 順序付き (levels) の因子化
  mutate(X3 = factor(X, levels=c("S", "A", "B", "C", "D")))
glimpse(toy_data3) # toy_data2 とは factor の順序が異なる
toy_data4 <- toy_data2 |>
  mutate(Y2 = factor(Y > 60)) # 条件による因子化
glimpse(toy_data4) # 条件の真偽で 2 値に類別される

```

練習問題

- 東京の気候データ (9-11 月) を用いて気温を回帰する以下のモデルを検討しなさい
 - 日射量, 気圧, 湿度の線形回帰モデル
 - 湿度の対数を考えた線形回帰モデル
 - 最初のモデルにそれぞれの交互作用を加えたモデル
- 東京の気候データ (1 年分) を用いて気温を回帰する以下のモデルを検討しなさい
 - 降水の有無を表すカテゴリカル変数を用いたモデル (雨が降ると気温が変化することを検証する)
 - 上記に月をカテゴリカル変数として加えたモデル (月毎の気温の差を考慮する)

補足

R : モデルの探索

- 変数が増えるとモデルの比較が困難
- 関数 `stats::step()` で自動化することができる

```

#' モデルの探索
adv_data <- read_csv('https://www.statlearning.com/s/Advertising.csv')
summary(lm(sales ~ radio, data = adv_data))
summary(lm(sales ~ TV + radio, data = adv_data))
summary(lm(sales ~ TV + radio + newspaper, data = adv_data))
summary(adv_init <- lm(sales ~ TV * radio * newspaper, data = adv_data))
adv_opt <- step(adv_init) # 最大のモデルから削減増加による探索
summary(adv_opt) # 探索された (準) 最適なモデルの確認

```

- 全探索ではないので最適とは限らないことに注意は必要

R : package::car

- 回帰モデルの評価
 - 与えられたデータの再現
 - 新しいデータの予測
- モデルの再構築のための視覚化
 - **residual plots**: 説明変数・予測値と残差の関係
 - **marginal-model plots**: 説明変数と目的変数・モデルの関係
 - **added-variable plots**: 説明変数・目的変数をその他の変数で回帰したときの残差の関係

- **component+residual plots**: 説明変数とそれ以外の説明変数による残差の関係などが用意されている

例題

- これまでに用いたデータでモデルを更新して評価してみよう
 - 変数間の線形回帰の関係について仮説を立てる
 - モデルのあてはめを行い評価する
 - * 説明力があるのか? (F 統計量, t 統計量, 決定係数)
 - * 残差に偏りはないか? (様々な診断プロット)
 - * 変数間の線形関係は妥当か? (様々な診断プロット)
 - 検討結果を踏まえてモデルを更新する (評価の繰り返し)

次回の予定

- 第1回: 主成分分析の考え方
- 第2回: 分析の評価と視覚化