2023年3月22日@京都大学+Zoom

KUDH Basics
統計ソフトウェア「R」ワークショップ

3限目

分散分析と線形モデル

講師：山崎 大暉

（立命館大学・日本学術振興会）

Daiki Yamasaki

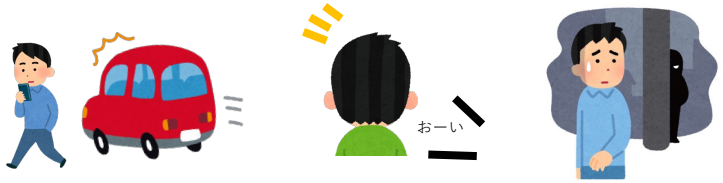
Kyoto University (Prof. Ashida Lab)

- Audiovisual interaction in spatial perception
- Perception of distance and approach motion

2021– Postdoc

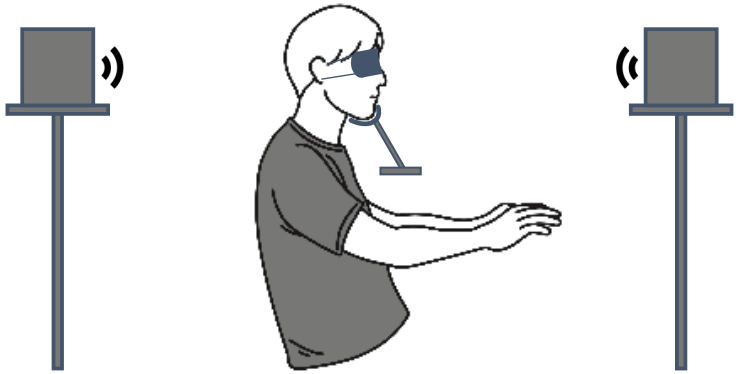
Ritsumeikan University (Prof. Nagai Lab)

- Influence of sounds around the body on vision, attention, and behavior
- Spatial perception in interpersonal communication



Recent works

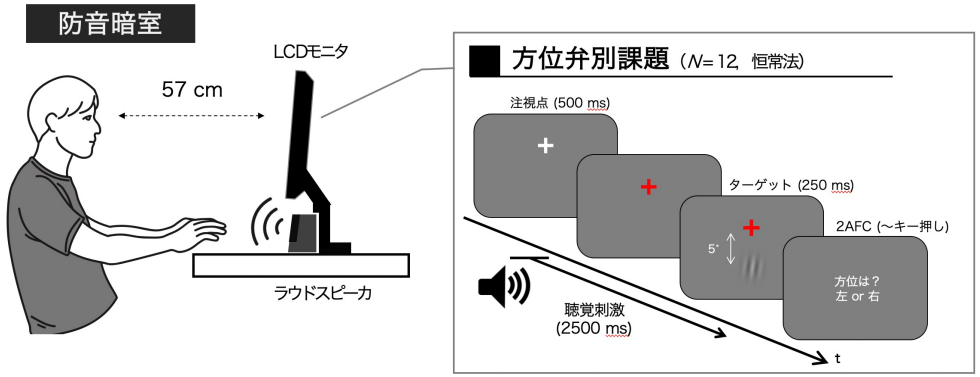
身体の前後への注意



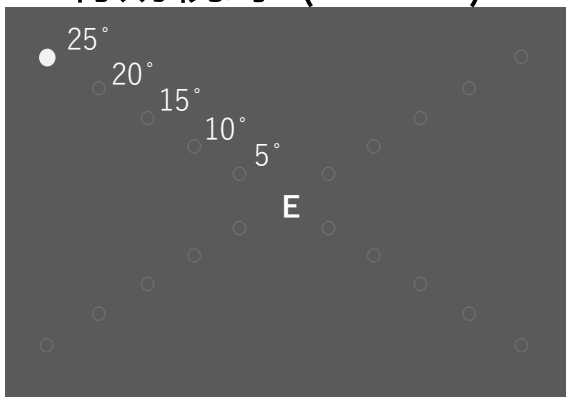
ASDとパーソナルスペース，音の距離知覚

音の身体性

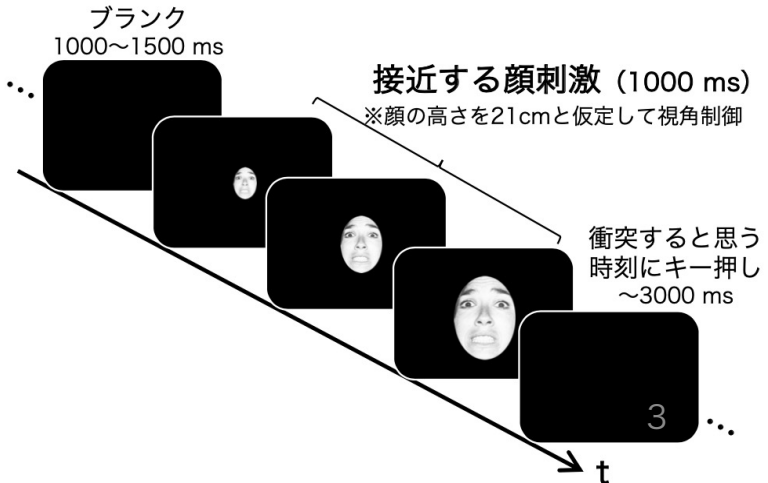
接近音の初期視覚促進



有効視野 (UFOV)



衝突時間推定



午後の流れと到達目標

1. lm()関数で線形モデルを使おう

- 量的変数を説明変数とした線形モデル：回帰
- 質的変数を説明変数とした線形モデル：分散分析，t検定
- 量的変数と質的変数を説明変数とした線形モデル

2. glm()関数で一般化線形モデルを使おう

- ポアソン分布，二項分布にしたがうデータを分析する
- 確率分布とリンク関数の指定
- 結果の解釈

Rで線形モデル

パッケージのインストールと読み込み

- インストールして使うパッケージと関数
  - ✓tidyverse
    - データフレーム操作やグラフ描画に色々な関数を使います（詳しくは前回のWS）
  - ✓car
    - 主効果・交互作用の検定を行うAnova()関数を使います
  - ✓multcomp
    - 多重比較を行うglht()関数を使います
  - ✓emmeans
    - 多重比較を行うemmeans()関数を使います
  - ✓faraway
    - ポアソン回帰のためのデータフレームgalaを使います

- パッケージをインストールせずに使えるデフォルト関数
  - ✓lm()関数：線形モデルを記述してパラメータの推定をします
  - ✓glm()関数：同じく一般化線形モデルを扱います
  - ✓summary()関数：データフレームやモデルを要約します
  - ✓View()関数，head()関数：データフレームを表示します

応答変数と説明変数

- データフレームiris
  - ✓3種のアヤメ50個体ずつの花びらとガクの長さと幅を測定したデータ
  - ✓150行，5列
  - ✓種（Species）が質的変数
  - ✓その他は量的変数

```
View(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa

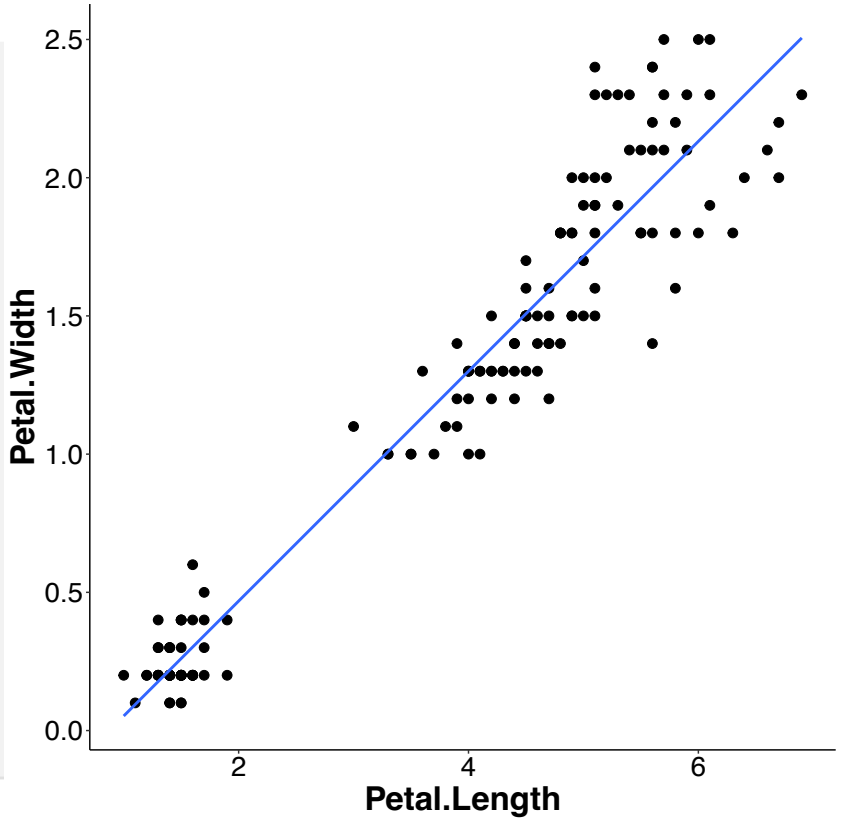
説明変数が量的変数1つの線形モデル

Petal.Lengthをx軸，Petal.Widthをy軸にした散布図と回帰直線

```r
library(ggplot2)

ggplot(iris, aes(x = Petal.Length, y = Petal.Width))+
  geom_point(size = 1.2)+
  geom_smooth(method="lm",se=F)+

  theme_classic()+
  theme(axis.title = element_text(size = 18, face = "bold"),
        axis.text = element_text(size =  15, color = "black"),
        strip.background = element_blank(),
        aspect.ratio = 1
        )
```

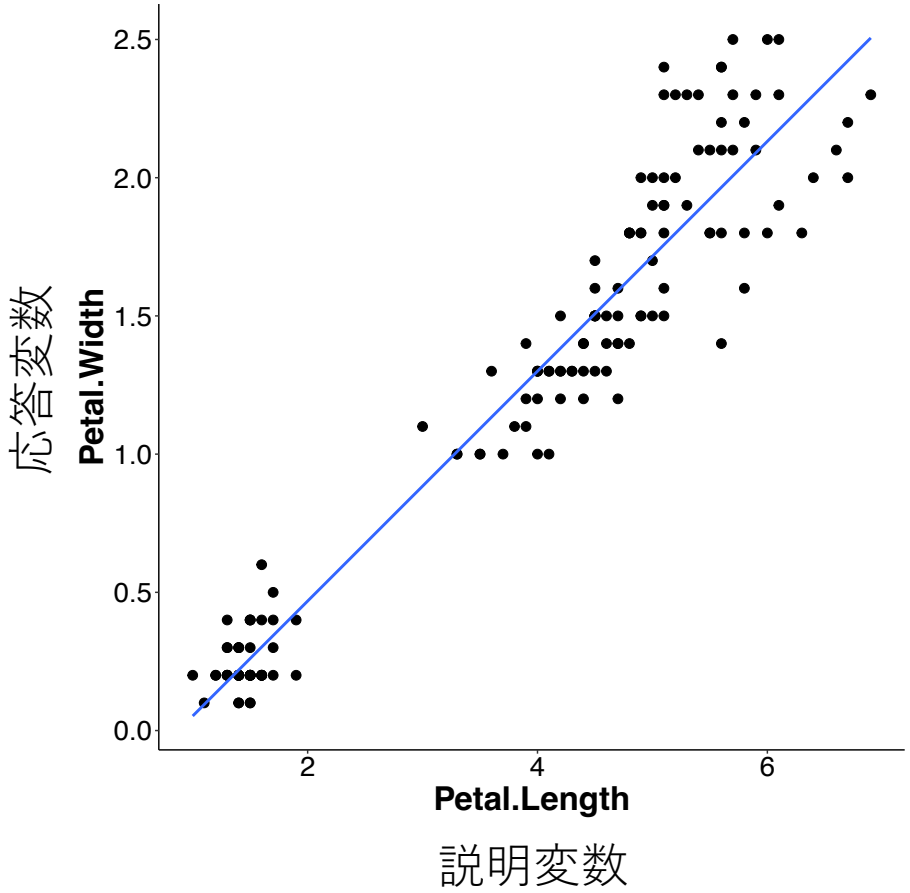


応答変数と説明変数

●応答変数
- 変動に興味のある変数
- Y軸：Petal.Width

●説明変数
- その値の変動とともに応答変数が変動することが想定される変数
- X軸：Petal.Length

説明変数(x)の変動に伴う
応答変数(y)の変動をグラフで表現する

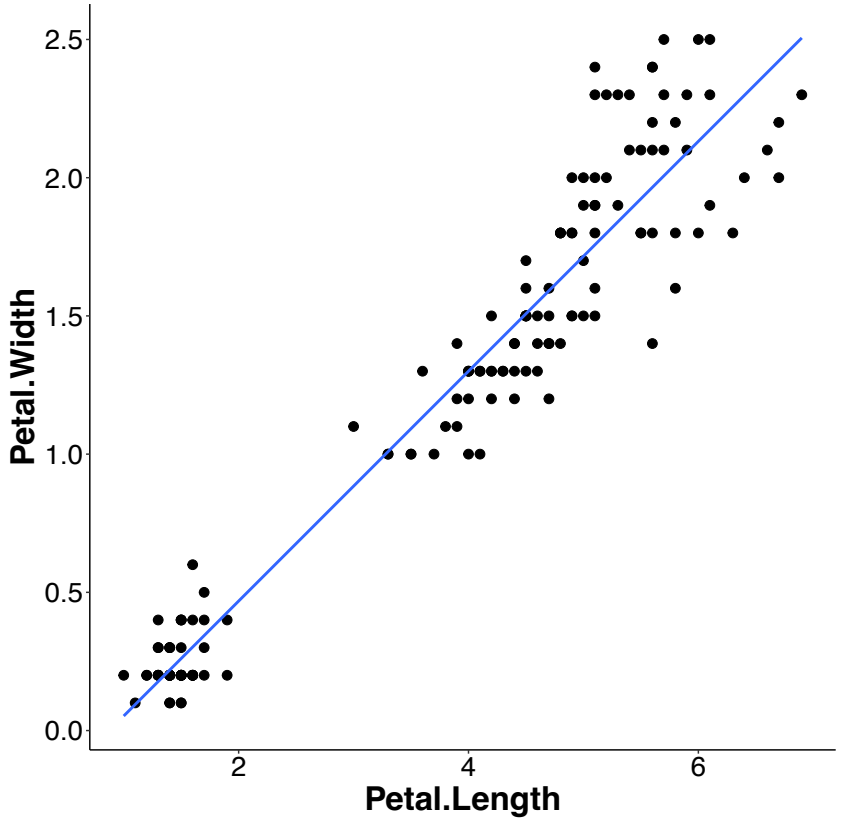


回帰直線をモデルで表現

●線形モデル

$$y = \beta_0 + \beta_1 x$$

- y：予測値
  - ✓xがある値をとる時のyの値
- β_0：切片
  - ✓xが0のときの予測値
- β_1：傾き
  - ✓xが1大きくなった時の予測値の変化量



回帰直線をモデルで表現

●線形モデル

$$y = \beta_0 + \beta_1 x$$

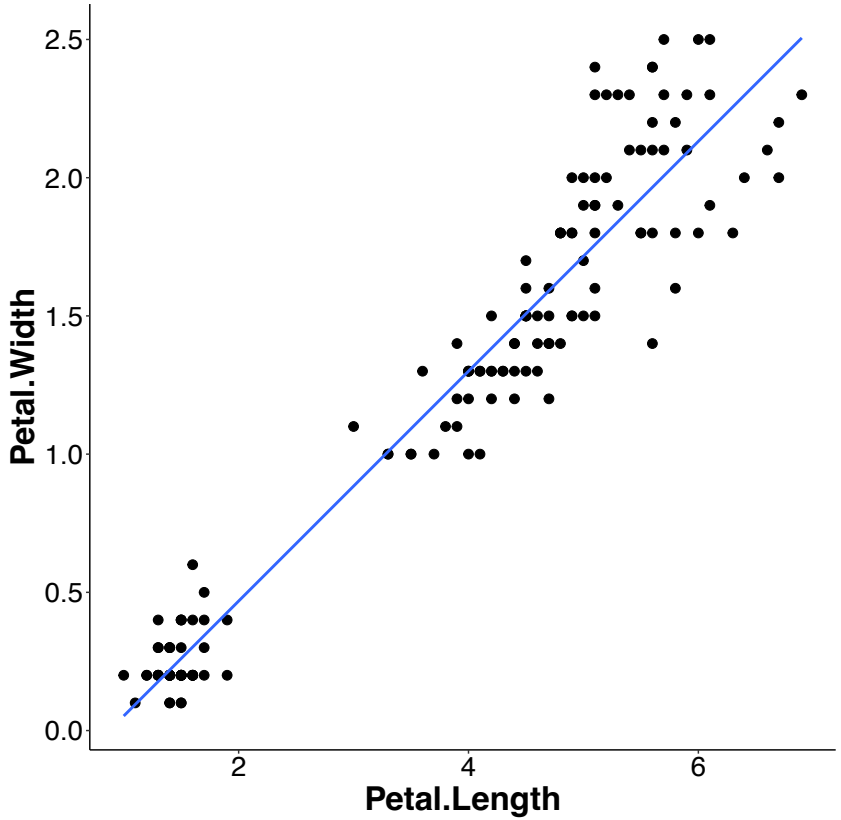
●Rのlm()関数を使った記法

```
lm(y ~ x, data = df)
```

応答変数 ~ 説明変数

データフレームの名前

```
lm(formula = Petal.Width ~ Petal.Length, data = iris)
```



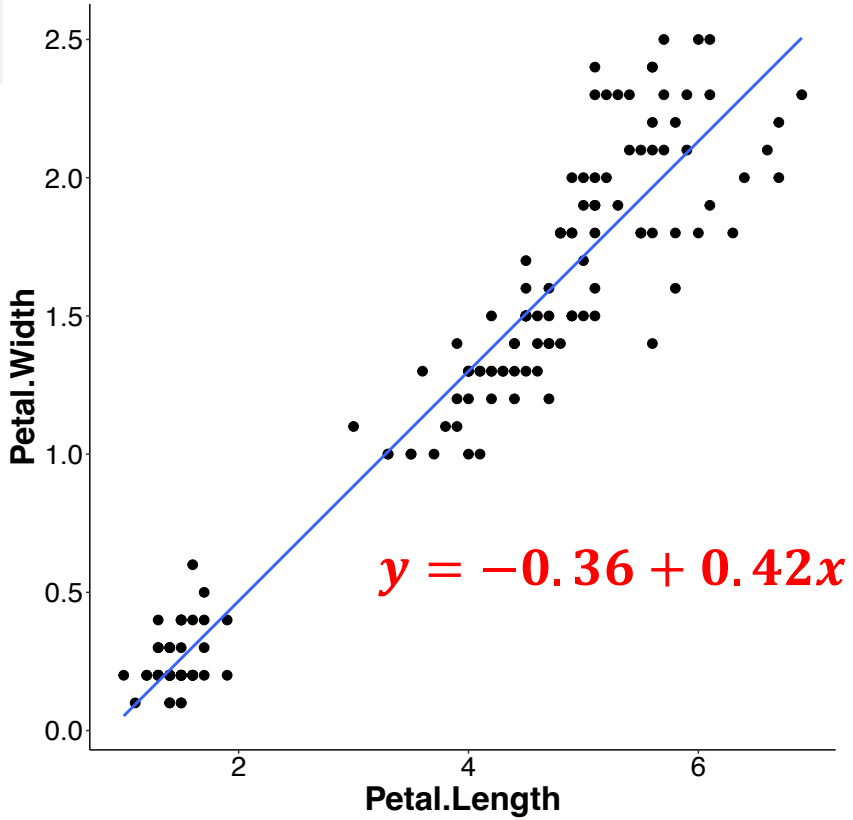
モデルのパラメータを推定する

```
m <- lm(Petal.Width ~ Petal.Length, data = iris)
m
```

これをRStudioのConsoleに打つと↓

```
Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

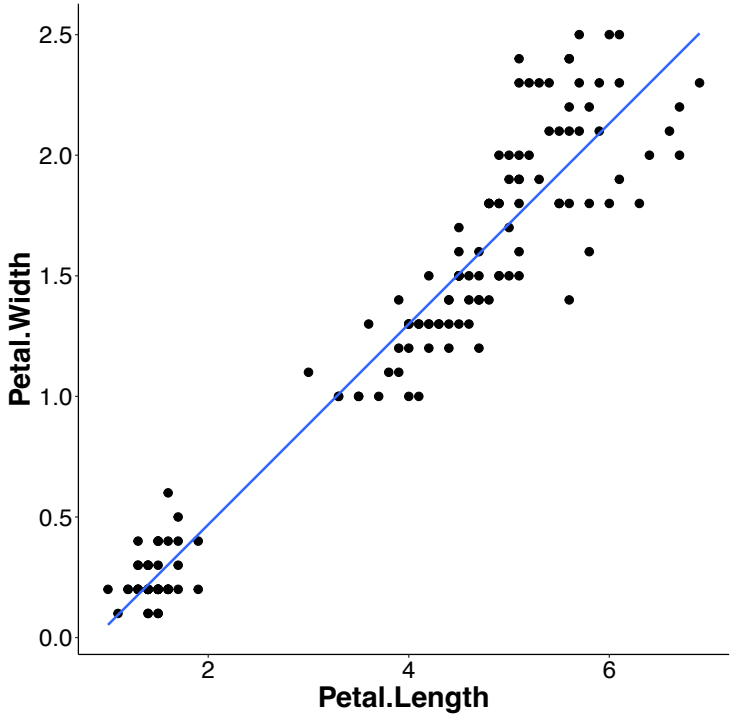
Coefficients:
 (Intercept)  Petal.Length
     -0.3631         0.4158
```



同じデータに複数の統計モデルを想定できる

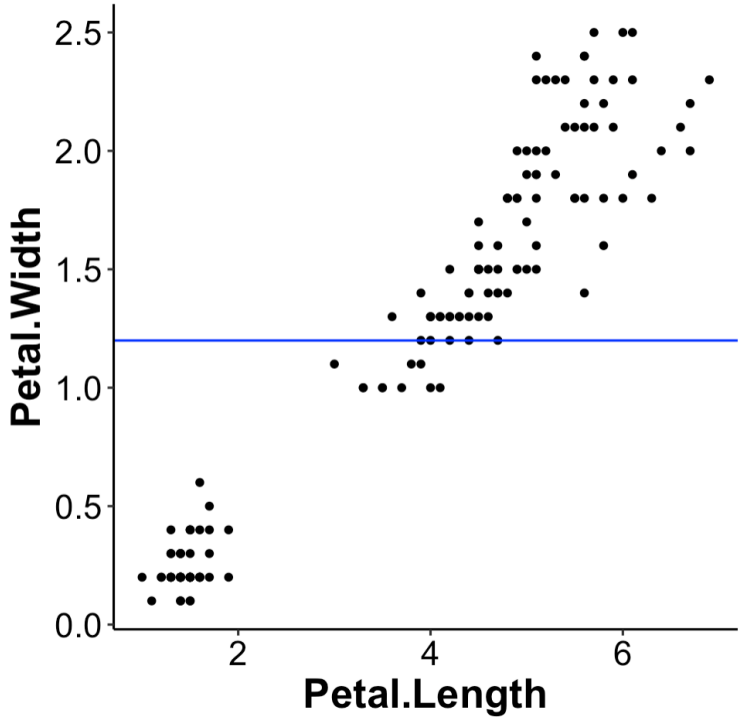
●統計モデル：応答変数の変動を数理モデルで表現したもの

✓どのモデルがデータをよく説明する？



✓ xモデル

- xの値に関わらず予測値は一定



✓ 切片モデル

- xの値によらず予測値は一定

統計モデルの関係：ネスト

●切片モデルは，xモデルの特殊ケース
- ✓xモデルにおいて，$\boldsymbol{\beta_1 = 0}$となる場合
- ✓切片モデルはxモデルにネストされている

xモデル：$y = \beta_0 + \beta_1 x$

切片モデル： $y = \beta'_0$

- ✓回帰直線を表す数式
  - xモデル：$y = \beta_0 + \beta_1 x$
  - 切片モデル： $y = \beta'_0$

```
m <- lm(Petal.Width ~ Petal.Length, data = iris)
m0 <- lm(Sepal.Width ~ 1, data = iris)
```

- ✓パラメータの個数
  - xモデル：2つ
  - 切片モデル：1つ（1はintercept）

モデルの評価：仮説検定とモデル選択

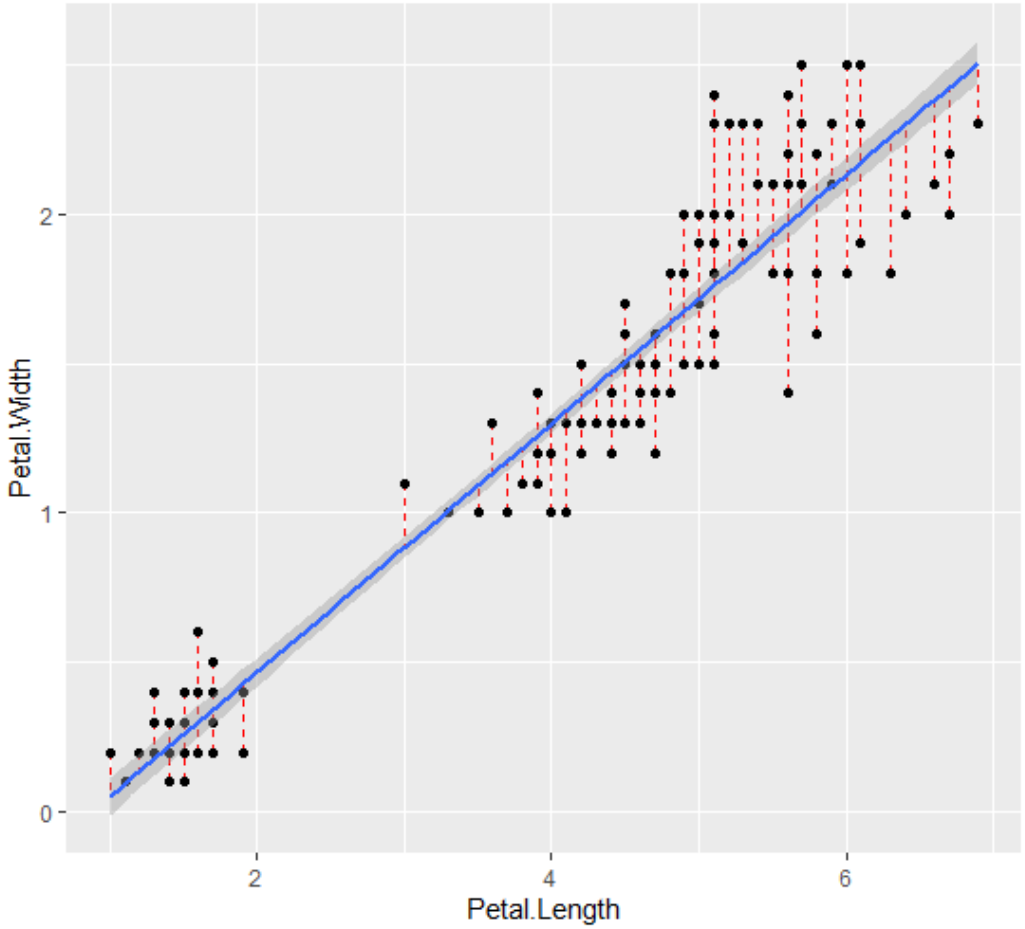
●**モデル選択**（今回は扱いません）
- ✓「**モデルの予測の良さ**」を評価する
  - AIC（赤池情報量規準）など**モデル選択基準**をもとに評価

●**仮説検定**
- ✓説明変数を追加することによる「**あてはまりの良さの改善**」を評価する
  - **p値**をもとに評価
  - あてはまりの良さの指標：残差平方和，尤度など

最小二乗法によるパラメータ推定

- 残差平方和（残差の二乗の総和）が最小になるような切片・傾きを推定
  - ✓残差：モデル予測値と実測値の差

- 残差平方和が小さくなるモデルがデータによく当てはまっていそう
  - ✓説明変数を加えたモデルは，元のモデルよりも必ず残差平方和が小さくなる

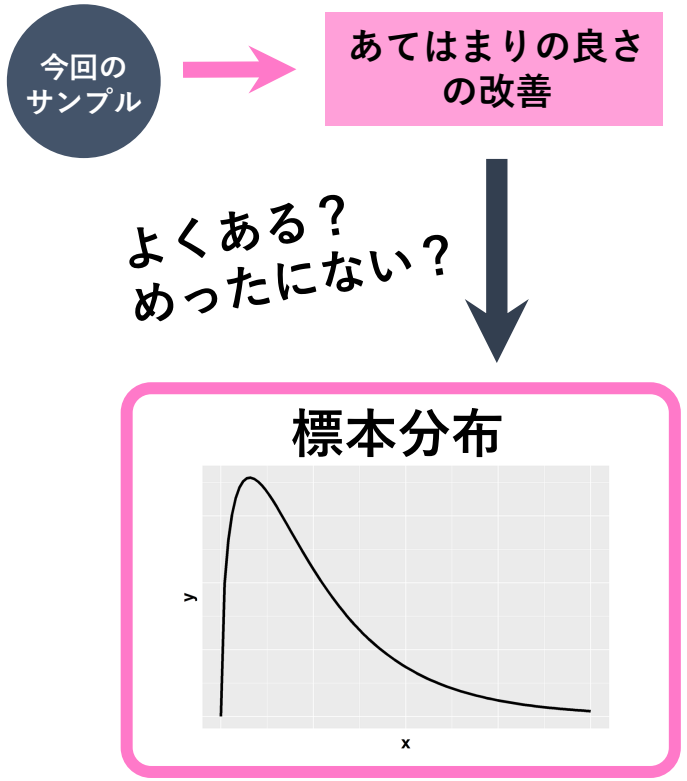


あてはまりの良さの改善を評価する

- 説明変数を加えると残差平方和は小さくなる
  - ✓説明変数を加えることによる残差平方和の減少が，帰無仮説が正しい場合に滅多に見られないものなのかどうかを評価

	モデル1	モデル2
説明変数の個数	3	2
パラメータの個数	6	4
残差平方和	10.68	12.19

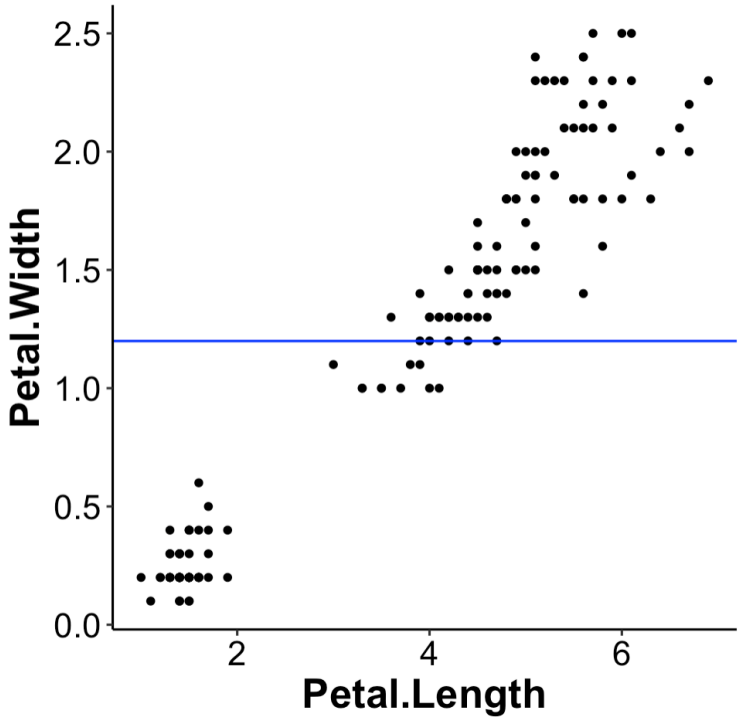
説明変数を1つ追加したことによる
残差平方和の減少
（あてはまりの良さの改善）



●説明変数を加えたxモデルの残差平方和は，切片モデルよりも小さくなるか？

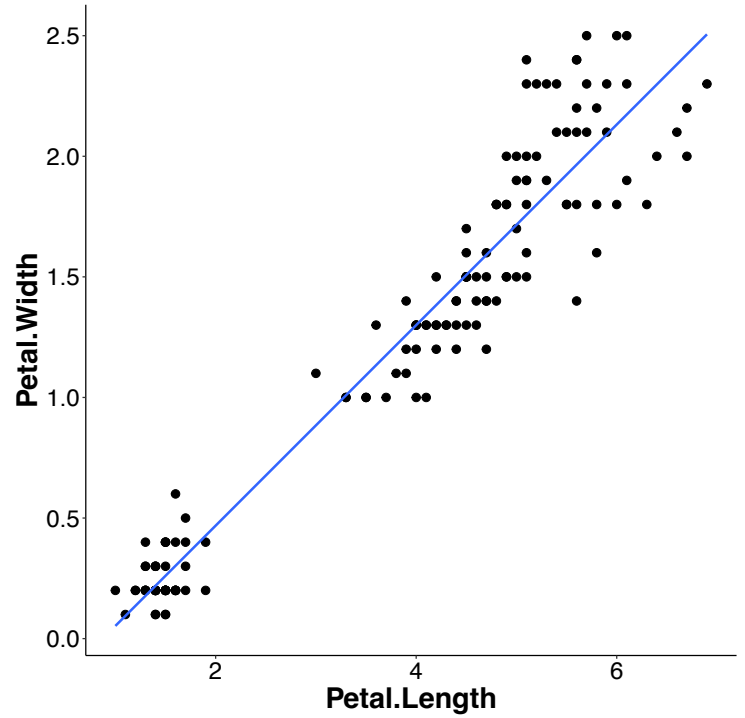
✓F検定

✓帰無仮説「説明変数の追加によって残差平方和は減少しない」



✓ 切片モデル

- xの値によらず予測値は一定



✓ xモデル

- xの値に関わらず予測値は一定

単回帰における仮説検定：F検定

- carパッケージのAnova()関数
  - ✓第一引数：説明変数を加えたxモデルのfittedオブジェクト
  - ✓第二引数：検定統計量であるF値を指定（test = “F”）

```
library(car) #パッケージを使う前にlibrary()関数で読み込み

Anova(m, test = "F")
```

```
Anova Table (Type II tests)

Response: Petal.Width
              Sum Sq  Df F value    Pr(>F)
Petal.Length  80.26   1  1882.5 < 2.2e-16 ***
Residuals      6.31 148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anova()関数のデフォルトはタイプ II 平方和

説明変数を加えることで減った残差平方和

モデルの残差平方和

モデルの決定係数

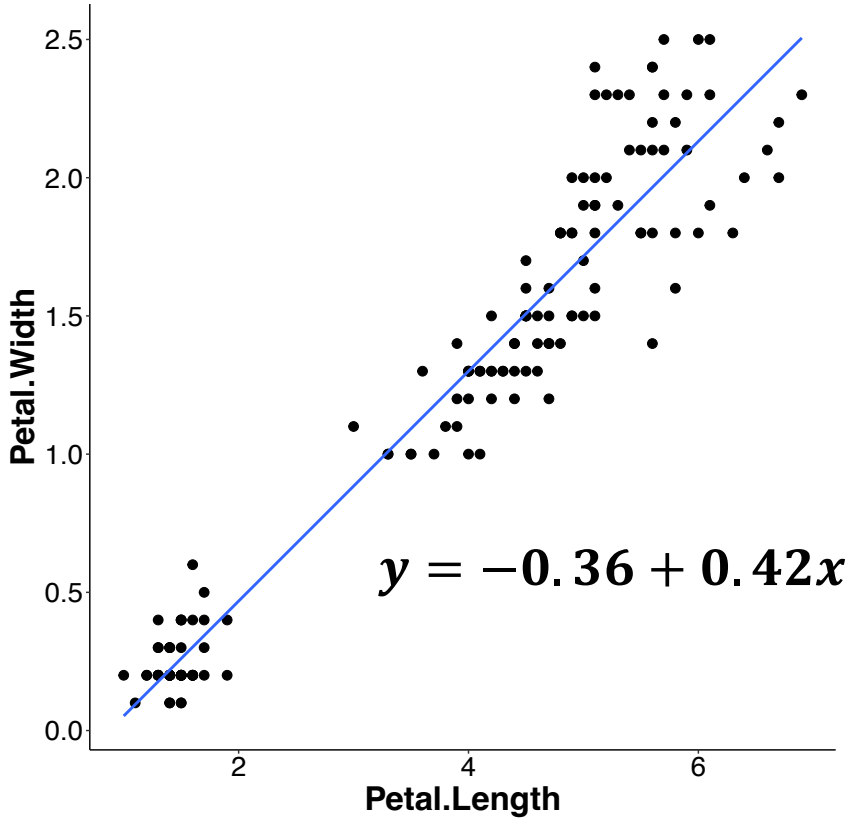
```
summary(m)
```

```
Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```



モデルの決定係数：切片モデルと比較して，残差平方和が減少した割合

単回帰における仮説検定：F検定

- carパッケージのAnova()関数
  - ✓第一引数：説明変数を加えたxモデルのfittedオブジェクト
  - ✓第二引数：検定統計量であるF値を指定（test = “F”）

```
library(car) #パッケージを使う前にlibrary()関数で読み込み

Anova(m, test = "F")
```

```
Anova Table (Type II tests)

Response: Petal.Width
             Sum Sq  Df F value    Pr(>F)
Petal.Length  80.26   1  1882.5 < 2.2e-16 ***
Residuals      6.31 148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

自由度，F値，p値

単回帰における仮説検定：F検定

```
library(car) #パッケージを使う前にlibrary()関数で読み込み

Anova(m, test = "F")
```

```
Anova Table (Type II tests)

Response: Petal.Width
              Sum Sq  Df F value    Pr(>F)
Petal.Length  80.26   1  1882.5 < 2.2e-16 ***
Residuals      6.31 148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

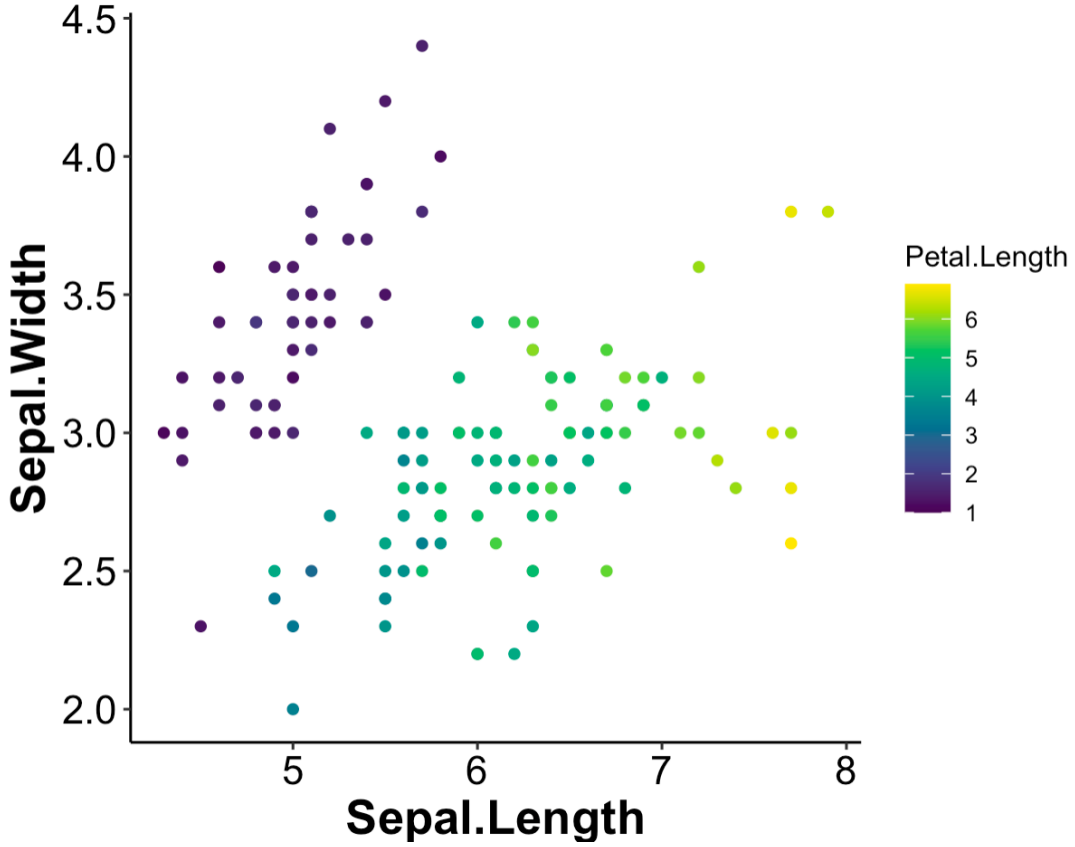
- F検定の結果，説明変数Petal.Lengthの値によって応答変数Petal.Widthの値が有意に変化した（$F(1, 148) = 1882.5$, $p < .001$）

説明変数が量的変数2つの線形モデル

重回帰分析を線形モデルで表現

- 重回帰分析（Multiple regression）
  - ✓説明変数が複数の連続変数である場合の線形回帰
  - ✓Sepal.Widthの値はSepal.LengthとPetal.Lengthで変わるか？

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Sepal.Width, col = Petal.Length))+
  geom_point()+
  scale_color_viridis_c()+
  theme_classic()+
  theme(axis.title = element_text(size = 18, face = "bold"),
        axis.text = element_text(size =  15, color = "black"),
        aspect.ratio = 1
        )
```



重回帰分析を線形モデルで表現

```
m6 <- lm(Sepal.Width ~ Sepal.Length*Petal.Length, data = iris)
summary(m6)
```

```
Call:
lm(formula = Sepal.Width ~ Sepal.Length * Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86960 -0.19846  0.00743  0.20704  0.72871

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.51011    0.64336   2.347 0.020257 *
Sepal.Length               0.46940    0.12954   3.624 0.000400 ***
Petal.Length              -0.42907    0.11832  -3.626 0.000397 ***
Sepal.Length:Petal.Length  0.01795    0.02186   0.821 0.413063
---
```

$$y = 1.51 + 0.469 * \textbf{Sepal. Length} - 0.429 * \textbf{Petal. Length} + 0.018 * \textbf{Sepal. Length} * \textbf{Petal. Length}$$

主効果の項

交互作用の項

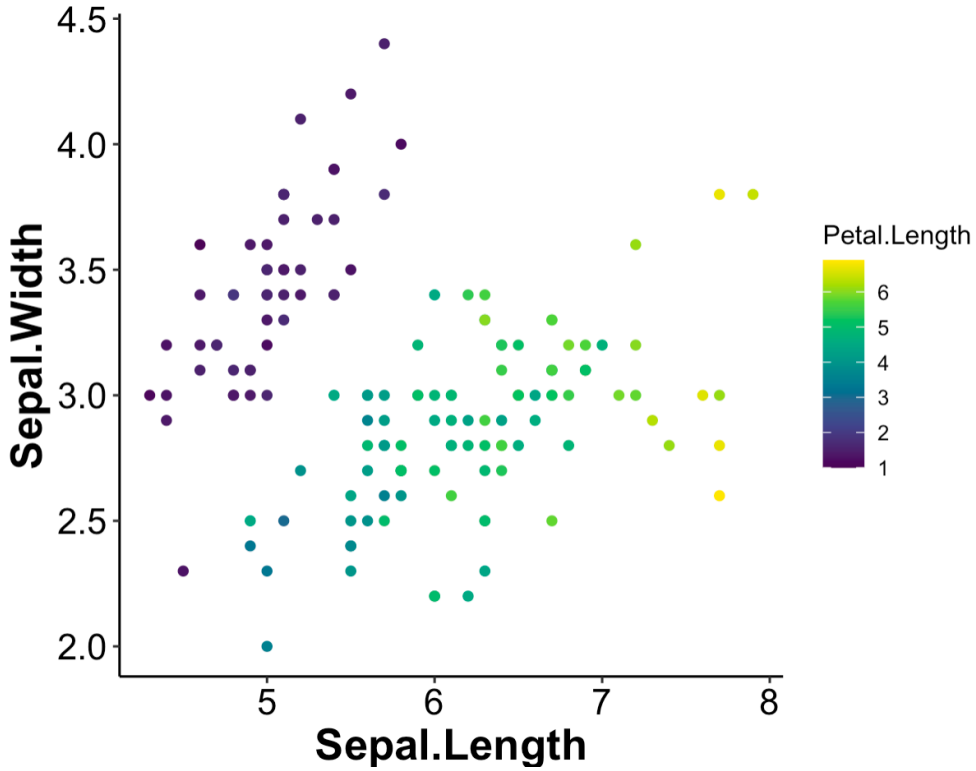
重回帰分析を線形モデルで表現

●線形モデルをcar::Anova()関数に渡して同様に検定ができる

```
Anova(m6)
```

```
Anova Table (Type II tests)

Response: Sepal.Width
                          Sum Sq  Df  F value    Pr(>F)
Sepal.Length              7.7237   1  73.6233 1.276e-14 ***
Petal.Length             12.5284   1 119.4233 < 2.2e-16 ***
Sepal.Length:Petal.Length  0.0707   1   0.6738    0.4131
Residuals                15.3165 146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

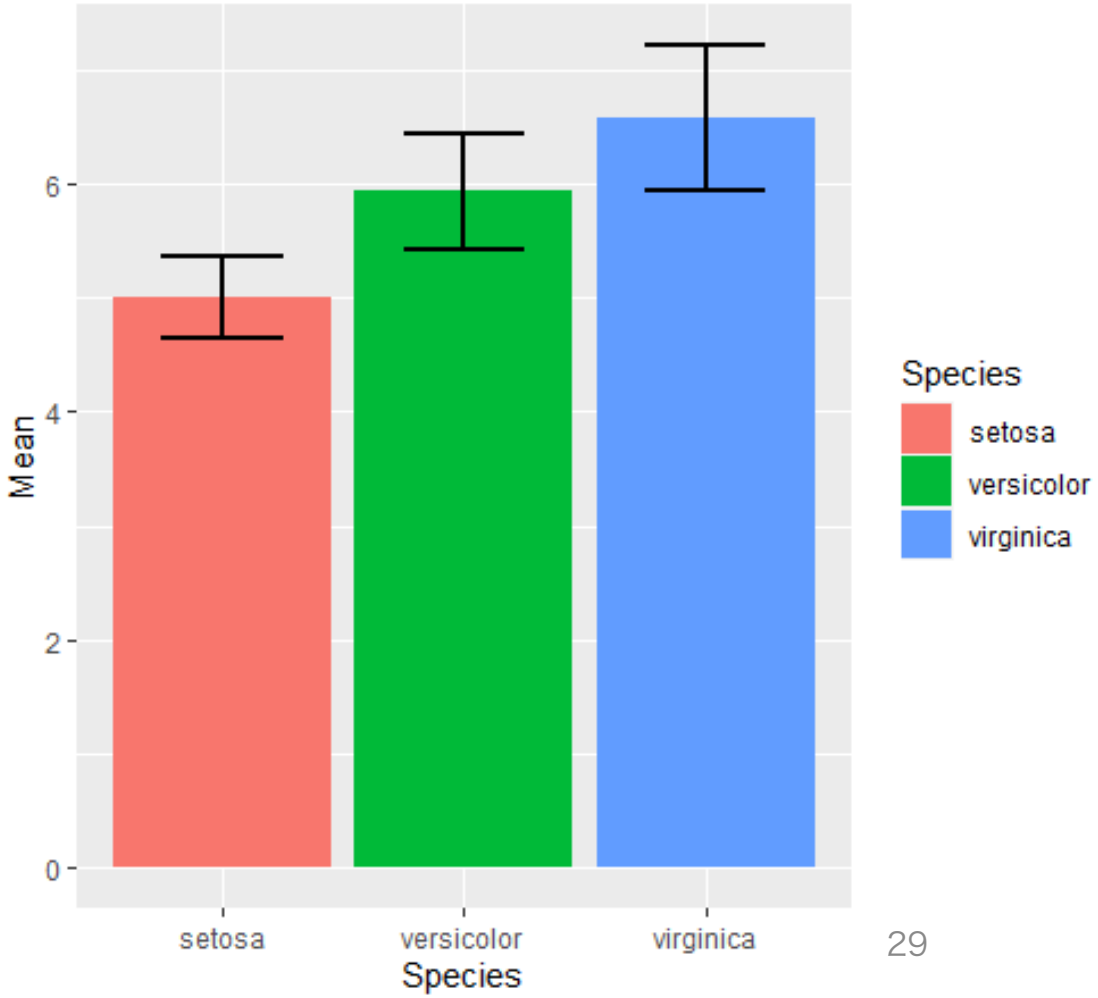


- ✓ Sepal.Lengthの主効果が有意（$F(1, 146) = 73.623$, $p < .001$）
- ✓ Petal.Lengthの主効果が有意（$F(1, 146) = 119.423$, $p < .001$）
- ✓ 交互作用は非有意（$F(1, 146) = 0.674$, $p = .41$）

説明変数が質的変数1つの線形モデル

分散分析を線形モデルで考える

- irisのSepal.Lengthの平均値を種間比較
  - ✓一要因参加者間計画の分散分析
  - ✓説明変数Speciesは質的変数
    - 3水準：setosa，versicolor，virginica



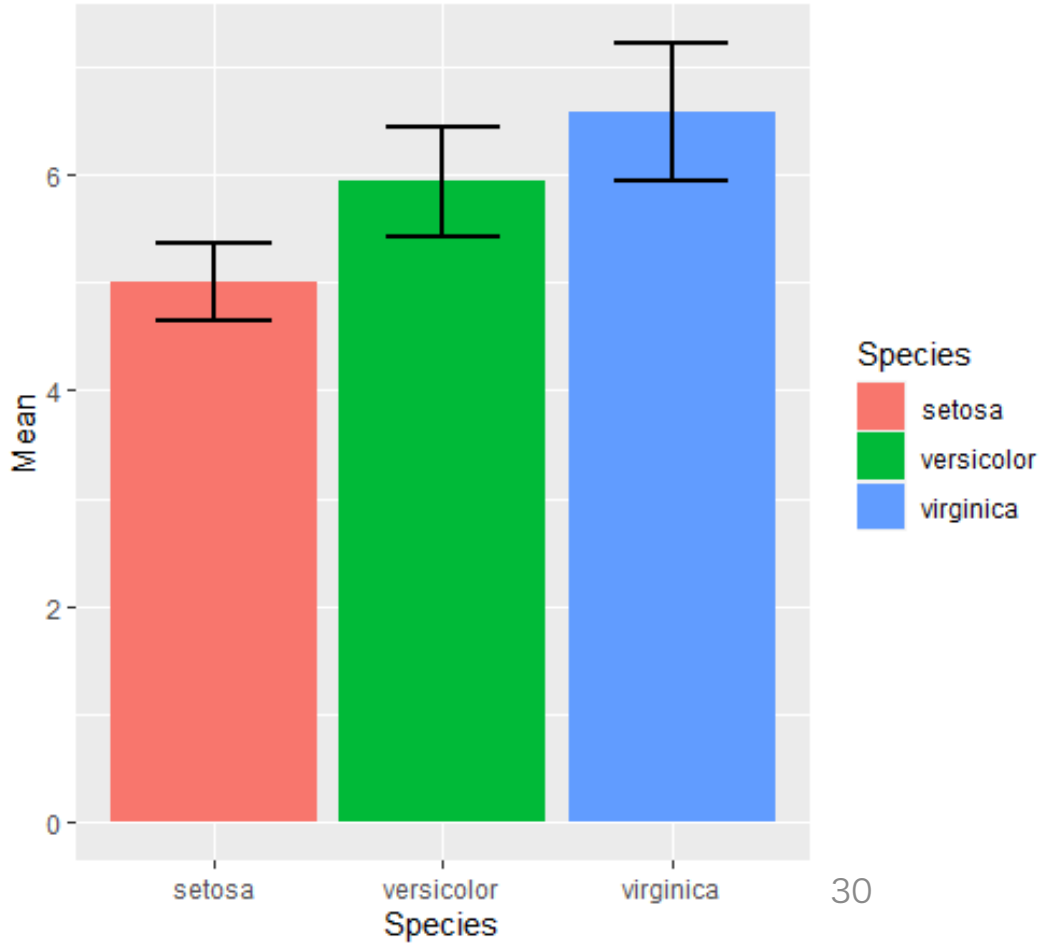
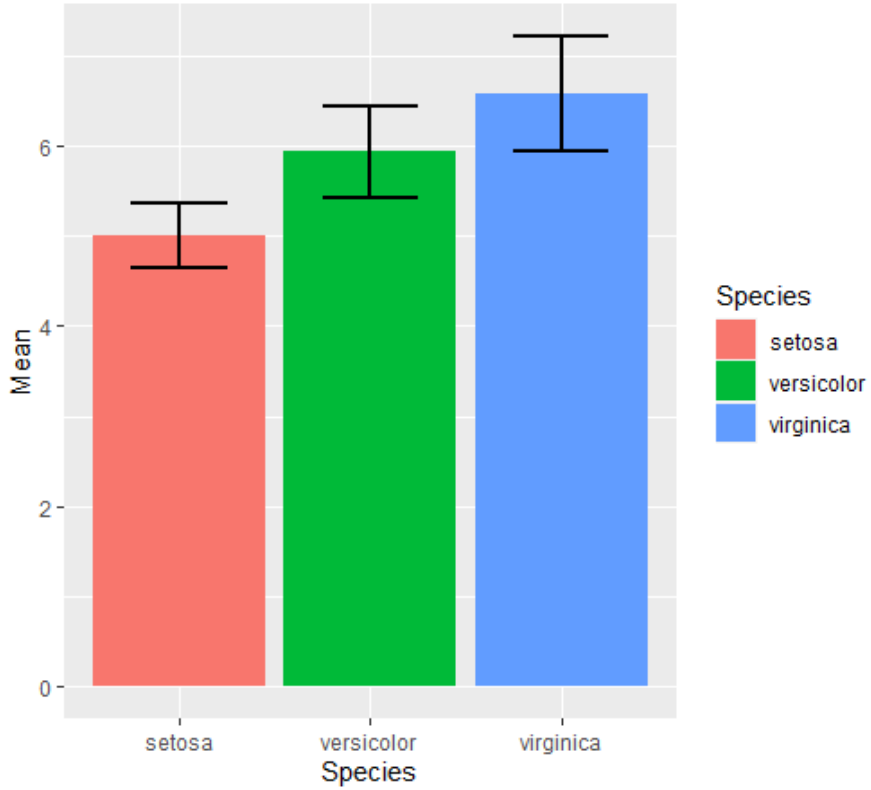
分散分析を線形モデルで考える

- summary()関数でパラメータの推定値を確認
  - ✓質的変数による予測値の変動をどう数式で表現する？

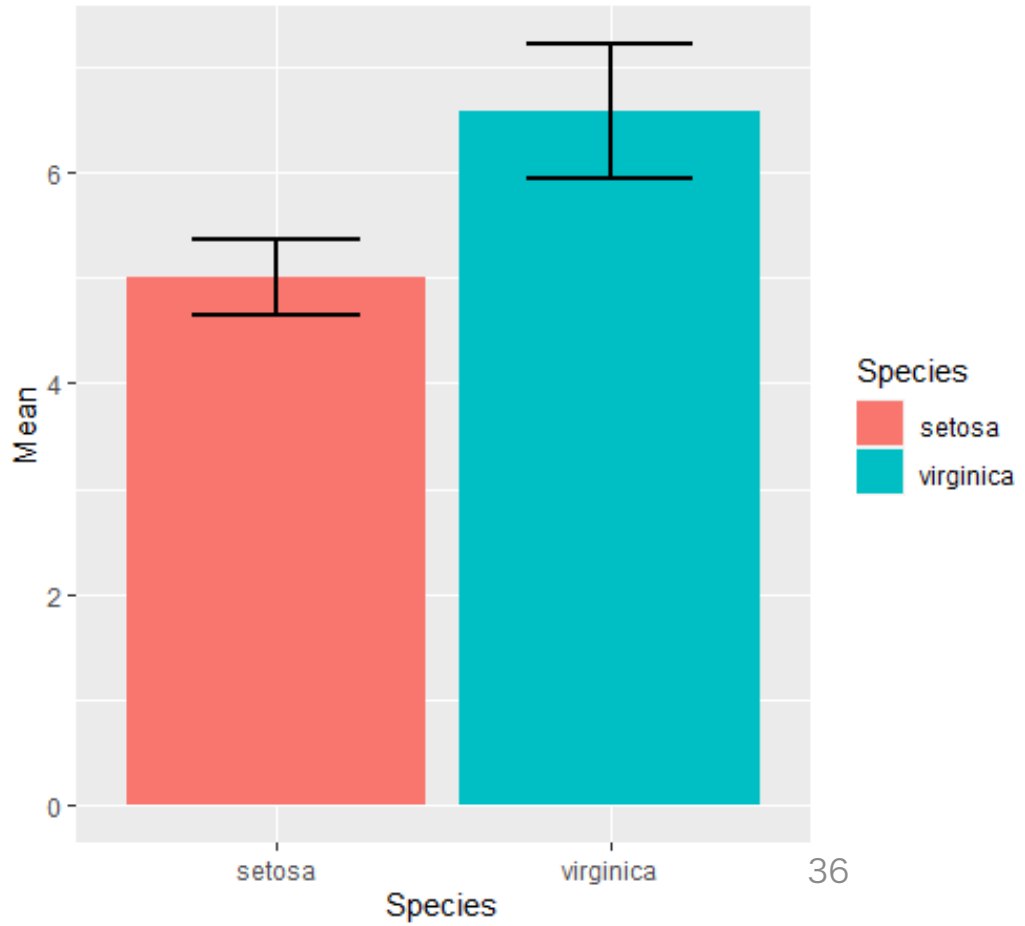
```
```{r}
summary(m2)
```
```

```
Call:
lm(formula = Sepal.Length ~ Species, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6880 -0.3285 -0.0060  0.3120  1.3120

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.0060     0.0728  68.762  < 2e-16 ***
Speciesversicolor  0.9300     0.1030   9.033 8.77e-16 ***
Speciesvirginica   1.5820     0.1030  15.366  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5148 on 147 degrees of freedom
Multiple R-squared:  0.6187,    Adjusted R-squared:  0.6135
F-statistic: 119.3 on 2 and 147 DF,  p-value: < 2.2e-16
```

切片

傾き



分散分析を線形モデルで考える

$$\mathbf{y} = \mathbf{5.01} + \mathbf{0.93} * d_{versicolor} + \mathbf{1.58} * d_{virginica}$$

●ダミー変数を用いて各水準の予測値を表現

✓$d_{versicolor}$
- 観測値がversicolorなら1，それ以外なら0

✓$d_{virginica}$
- 観測値がvirginicaなら1，それ以外なら0

✓観測値がsetosaなら，どちらのダミー変数も0



※最も若い名前の水準の予測値が切片になる
- ✓irisのSpeciesでは，setosaが一番若い
- ✓setosaを基準として，水準ごとにどう変わるかを表現

説明変数が質的変数の場合のsummary()出力

$$y = 5.01 + 0.93 * d_{versicolor} + 1.58 * d_{virginica}$$

●ダミー変数によって，説明変数が**質的変数**の場合の予測値も数式で表現できる

- Speciesがsetosaのとき：

$$y = 5.01 + 0.93 * 0 + 1.58 * 0 = 5.01$$

- Speciesがversicolorのとき：

$$y = 5.01 + 0.93 * 1 + 1.58 * 0 = 5.94$$

- Speciesがvirginicaのとき：

$$y = 5.01 + 0.93 * 0 + 1.58 * 1 = 6.59$$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
Speciesvirginica    1.5820     0.1030  15.366  < 2e-16 ***
```

説明変数が質的変数の場合のsummary()出力

$y = 5.01 + 0.93 * d_{versicolor} + 1.58 * d_{virginica}$

- summary()出力を読み解くコツ
  - ✓まず，切片に相当する水準を考える：最も名前が若い水準
  - ✓**傾き**は，切片の水準から別の水準に変化したときの予測値の変化量
    - 切片の水準はsetosa：5.01
    - 水準がsetosaからversicolorに変化すると，5.01+0.93
    - 水準setosaからvirginicaに変化すると，5.01+1.58

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
Speciesvirginica    1.5820     0.1030  15.366  < 2e-16 ***
```

分散分析を線形モデルで考える

- lm()関数とcar::Anova()関数で，分散分析と同一の分析ができる
  - ✓lm()関数でfittedオブジェクトを作成
  - ✓carパッケージのAnova()関数に投入

```
m2 <- lm(Sepal.Length ~ Species, data = iris)
Anova(m2, test = "F")

```

```
Anova Table (Type II tests)

Response: Sepal.Length
           Sum Sq  Df F value    Pr(>F)
Species   63.212   2  119.26 < 2.2e-16 ***
Residuals 38.956 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

t検定を線形モデルで考える

- irisのsetosaとvirginicaのSepal.Lengthの平均値を比較
  - ✓1要因2水準の参加者間計画 ⇒ t検定のかたち



t検定を線形モデルで考える

```
library(tidyverse)

iris2 <- iris %>%
  filter(Species != "versicolor")

m3 <- lm(Sepal.Length ~ Species, data = iris2)
summary(m3)
```

- tidyverseパッケージを読み込み
- tidyverseのfilter()関数と%>%（パイプ演算子）を使って，irisからSpeciesがversicolorでないものだけを抽出してiris2というデータフレームに格納
- lm()関数で線形モデルを記述
- モデルの要約

```
Call:
lm(formula = Sepal.Length ~ Species, data = iris2)

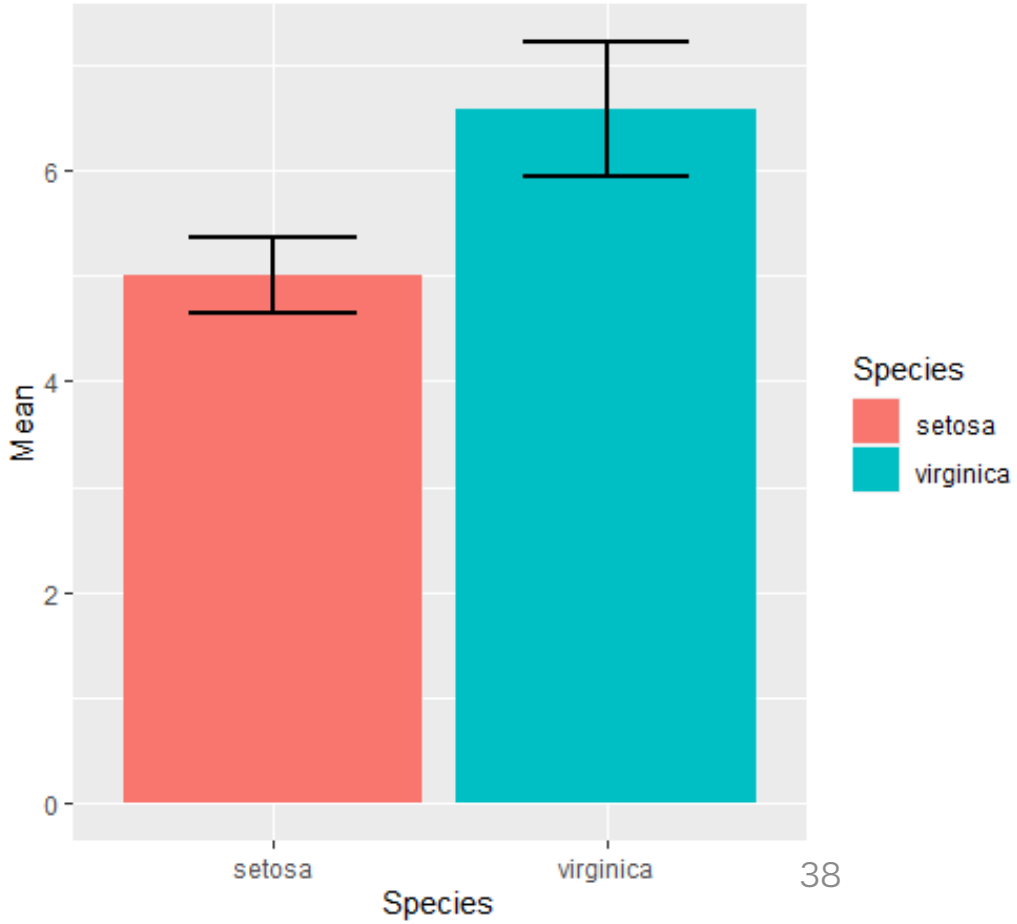
Residuals:
    Min      1Q  Median      3Q     Max
-1.6880 -0.2880 -0.0060  0.2985  1.3120

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.0060     0.0727   68.85   <2e-16 ***
Speciesvirginica   1.5820     0.1028   15.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5141 on 98 degrees of freedom
Multiple R-squared:  0.7072,    Adjusted R-squared:  0.7042
F-statistic: 236.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

t検定を線形モデルで考える

- irisのsetosaとvirginicaのSepal.Lengthの平均値を比較
  - ✓1要因2水準の参加者間計画 ⇒ t検定のかたち

✓Anova()関数でF検定

- あくまで線形モデルの枠組み
- t値ではなく，F値が検定統計量

```
```{r}
Anova(m3)
```
```

```
Anova Table (Type II tests)

Response: Sepal.Length
             Sum Sq Df F value    Pr(>F)
Species      62.568  1  236.74 < 2.2e-16 ***
Residuals 25.901 98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

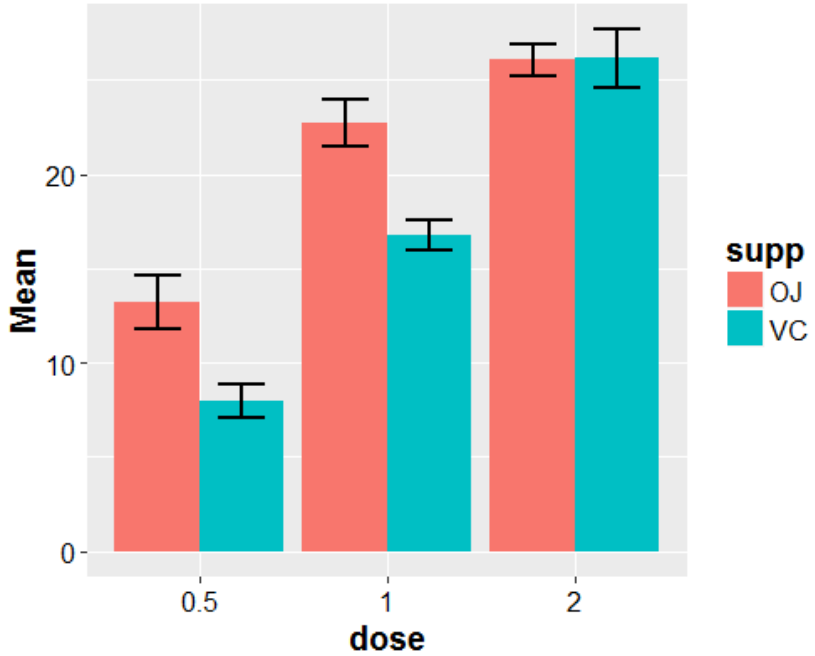


説明変数が質的変数2つの線形モデル

ToothGrowth：モルモットの歯の長さデータ

- モルモット60個体のデータ
  - ✓len ：歯の長さ
  - ✓supp：ビタミンC（VC），オレンジジュース（OJ）の2水準
  - ✓dose：0.5，1.0，2.0 mgの3水準

- lenの平均値をsuppおよびdoseで比較



	len	supp	dose
1	4.2	VC	0.5
2	11.5	VC	0.5
3	7.3	VC	0.5
4	5.8	VC	0.5
5	6.4	VC	0.5
6	10.0	VC	0.5
7	11.2	VC	0.5
8	11.2	VC	0.5
9	5.2	VC	0.5
10	7.0	VC	0.5
11	16.5	VC	1.0
12	16.5	VC	1.0
13	15.2	VC	1.0

二要因の線形モデル：主効果+交互作用

●lenの平均値をsuppおよびdoseで比較

```
TG_chr <- ToothGrowth %>%
  mutate(dose_c = as.character(dose))

m4 <- lm(len ~ supp * dose_c, data = TG_chr)
m4 <- lm(len ~ supp + dose_c + supp:dose_c, data = TG_chr)

summary(m4)
```

- ToothGrowthのdoseを文字列型に変換してdose_cとしたデータフレーム”TG_chr”を作成
- TG_chrを使った線形モデルをlm()で記述
  - ✓ 二要因の主効果と交互作用
- モデルの要約

● **主効果と交互作用の記法**

✓ len ~ supp ——— suppの主効果

✓ len ~ supp + dose ——— suppとdoseの主効果

✓ len ~ supp:dose ——— suppとdoseの交互作用

✓ len ~ supp+dose+supp:dose
✓ len ~ supp*dose
} suppの主効果，doseの主効果，交互作用

二要因の線形モデル：主効果+交互作用

```
summary(m4)
```

```
Call:
lm(formula = len ~ supp + dose_c + supp:dose_c, data = TG_chr)

Residuals:
   Min     1Q Median     3Q    Max
 -8.20  -2.72  -0.27   2.65   8.27

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.230      1.148  11.521 3.60e-16 ***
suppVC           -5.250      1.624  -3.233  0.00209 **
dose_c1           9.470      1.624   5.831 3.18e-07 ***
dose_c2          12.830      1.624   7.900 1.43e-10 ***
suppVC:dose_c1   -0.680      2.297  -0.296  0.76831
suppVC:dose_c2    5.330      2.297   2.321  0.02411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.631 on 54 degrees of freedom
Multiple R-squared:  0.7937,    Adjusted R-squared:  0.7746
F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

$$y = 13.23 - 5.25 * d_{supp_{VC}} + 9.47 * d_{dose_1} + 12.83 * d_{dose_2} - 0.68 * d_{supp_{VC}} * d_{dose_1} + 5.33 * d_{supp_{VC}} * d_{dose_2}$$

- ●y：応答変数lenの予測値
- ●切片の水準：supp=OJかつdose=0.5
- ●全水準を表現するためのダミー変数
  - $supp_{VC}$：suppがOJ->VCのとき1
  - $dose_1$：doseが0.5->1のとき1
  - $dose_2$：doseが0.5->2のとき1

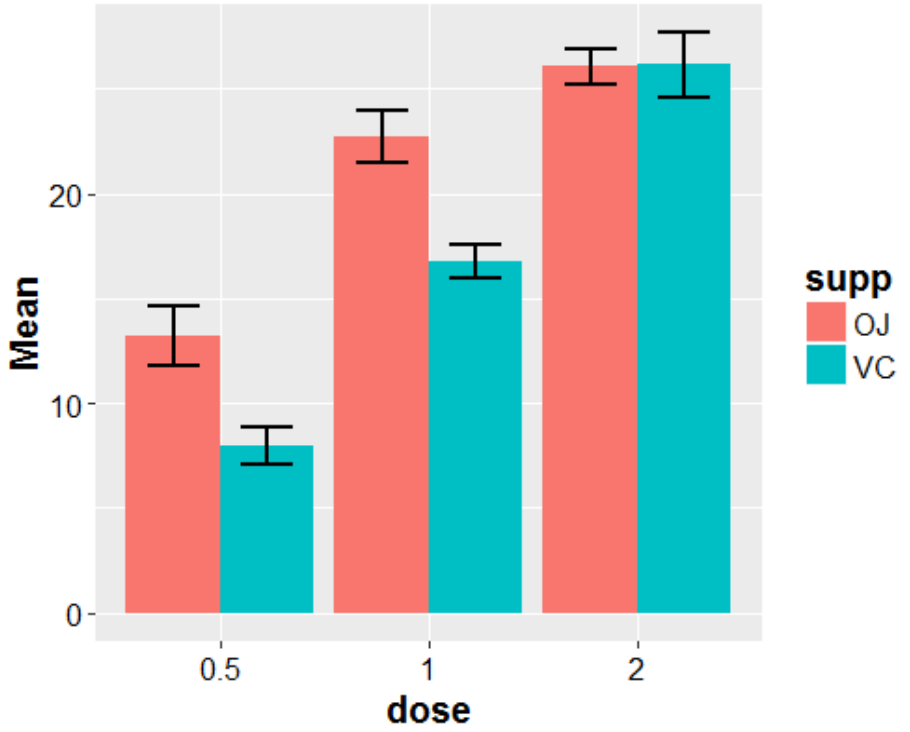
ToothGrowth：モルモットの歯の長さデータ

- lenの平均値をsuppおよびdoseで比較
  - ✓lm()で作ったモデルをAnova()関数に投入
  - ✓suppの主効果，doseの主効果，supp:doseの交互作用

```
Anova(m4)
```

```
Anova Table (Type II tests)

Response: len
             Sum Sq Df F value    Pr(>F)
supp         205.35  1  15.572 0.0002312 ***
dose_c      2426.43  2  92.000 < 2.2e-16 ***
supp:dose_c  108.32  2   4.107 0.0218603 *
Residuals    712.11 54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

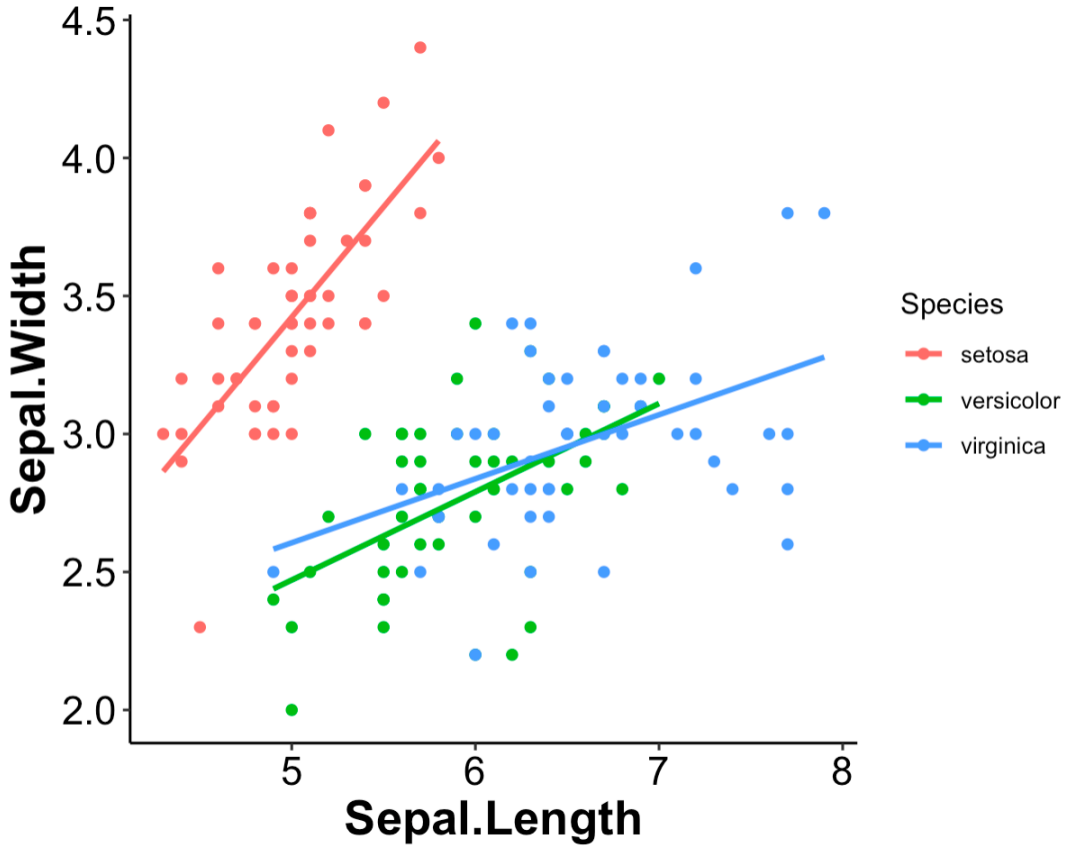


説明変数が
質的変数と量的変数の線形モデル

質的変数+量的変数：共分散分析

- irisのSepal.Widthの値は，SpeciesとSepal.Lengthによって変わるか？
  - ✓Species：質的変数（3水準），Sepal.Length：量的変数（共変量）
  - ✓散布図と，Speciesごとに層別化した回帰直線

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Sepal.Width, col = Species))+
  geom_point()+
  geom_smooth(method = "lm", se=F)+
  theme_classic()+
  theme(axis.title = element_text(size = 18, face = "bold"),
        axis.text = element_text(size =  15, color = "black"),
        aspect.ratio = 1
        )
```



共分散分析を線形モデルで表現する

```
m5 <- lm(Sepal.Width ~ Sepal.Length*Species, data = iris)
summary(m5)
```

```
Call:
lm(formula = Sepal.Width ~ Sepal.Length * Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72394 -0.16327 -0.00289  0.16457  0.60954

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.5694     0.5539  -1.028 0.305622
Sepal.Length                  0.7985     0.1104   7.235 2.55e-11 ***
Speciesversicolor             1.4416     0.7130   2.022 0.045056 *
Speciesvirginica              2.0157     0.6861   2.938 0.003848 **
Sepal.Length:Speciesversicolor -0.4788    0.1337  -3.582 0.000465 ***
Sepal.Length:Speciesvirginica -0.5666     0.1262  -4.490 1.45e-05 ***
---
```

$$
\begin{aligned}
y = & -0.57 \\
& +0.80 * x \\
& +1.44 * d_{versicolor} \\
& +2.02 * d_{virginica} \\
& -0.48 * x * d_{versicolor} \\
& -0.57 * x * d_{virginica}
\end{aligned}
$$

共分散分析を線形モデルで表現する

$$y = -0.57 + 0.80 * x + 1.44 * d_{versicolor} + 2.02 * d_{virginica} - 0.48 * x * d_{versicolor} - 0.57 * x * d_{virginica}$$

- Species = setosa のとき
  - ✓ダミー変数は$d_{versicolor}$も$d_{virginica}$も0をとる

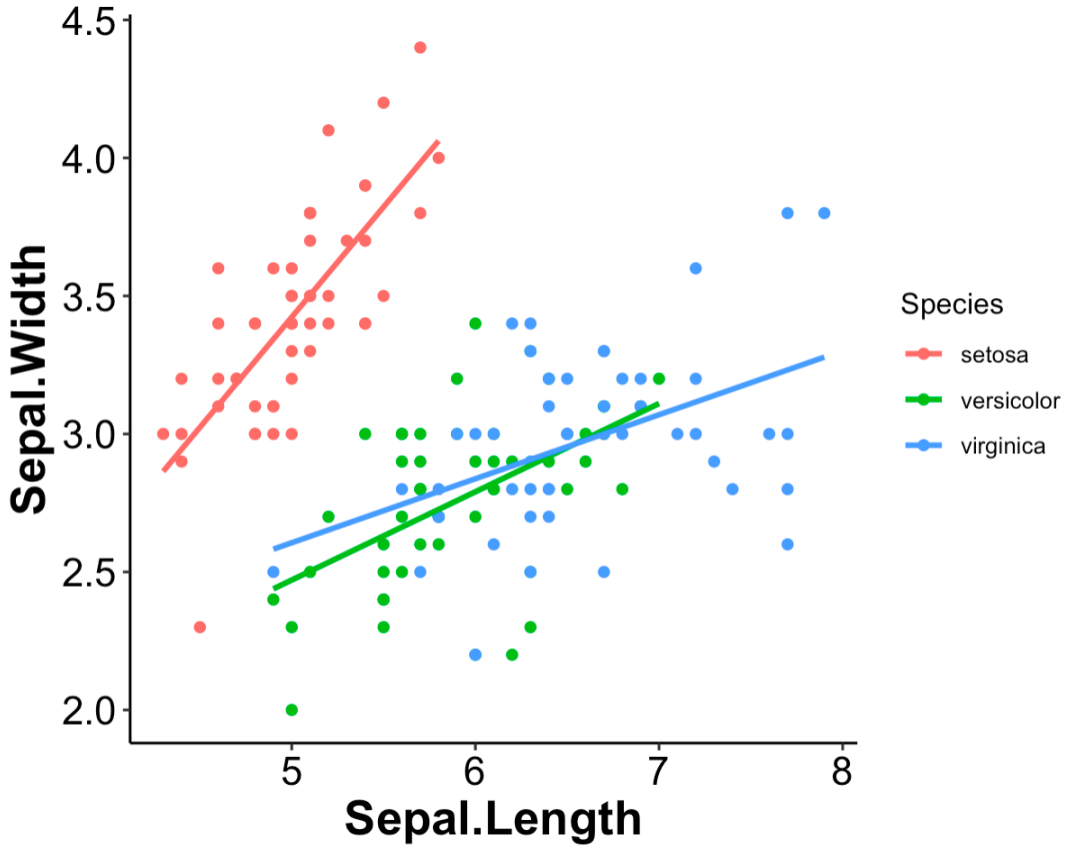
$$y = \mathbf{-0.57 + 0.80x}$$

- Species = versicolor のとき

$$y = -0.57 + 0.80 * x + 1.44 - 0.48 * x = \mathbf{0.87 + 0.31x}$$

- Species = virginica のとき

$$y = -0.57 + 0.80 * x + 2.02 - 0.57 * x = \mathbf{1.45 + 0.23x}$$



共分散分析を線形モデルで表現する

$$y = -0.57 + 0.80 * x + 1.44 * d_{versicolor} + 2.02 * d_{virginica} - 0.48 * x * d_{versicolor} - 0.57 * x * d_{virginica}$$

- ●交互作用：ある要因の影響の大きさが，他の要因の水準で変化する
  - ✓Speciesによって切片・傾き（xの係数）が異なることが表現されている

- Species = setosa のとき

$$y = -0.57 + \mathbf{0.80 * x} + 1.44 * 0 + 2.02 * 0 - 0.48 * x * 0 - 0.57 * x * 0$$
$$= -0.57 + \mathbf{0.80} * x$$

- Species = versicolor のとき

$$y = -0.57 + \mathbf{0.80 * x} + 1.44 * 1 + 2.02 * 0 \mathbf{- 0.48 * x * 1} - 0.57 * x * 0$$
$$= 0.87 + \mathbf{0.31} * x$$

- Species = virginica のとき

$$y = -0.57 + \mathbf{0.80 * x} + 1.44 * 0 + 2.02 * 1 - 0.48 * x * 0 \mathbf{- 0.57 * x * 1}$$
$$= 1.45 \mathbf{+ 0.23} * x$$

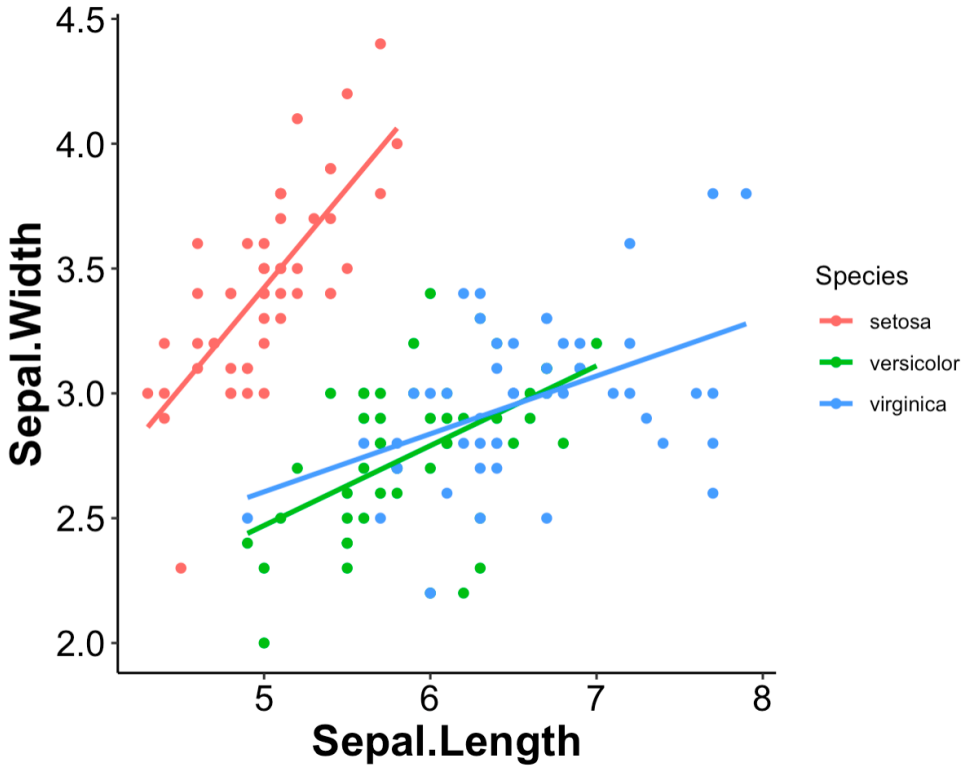
質的変数+量的変数：共分散分析

●モデルをcar::Anova()関数に渡して検定

```
Anova(m5)
```

```
Anova Table (Type II tests)

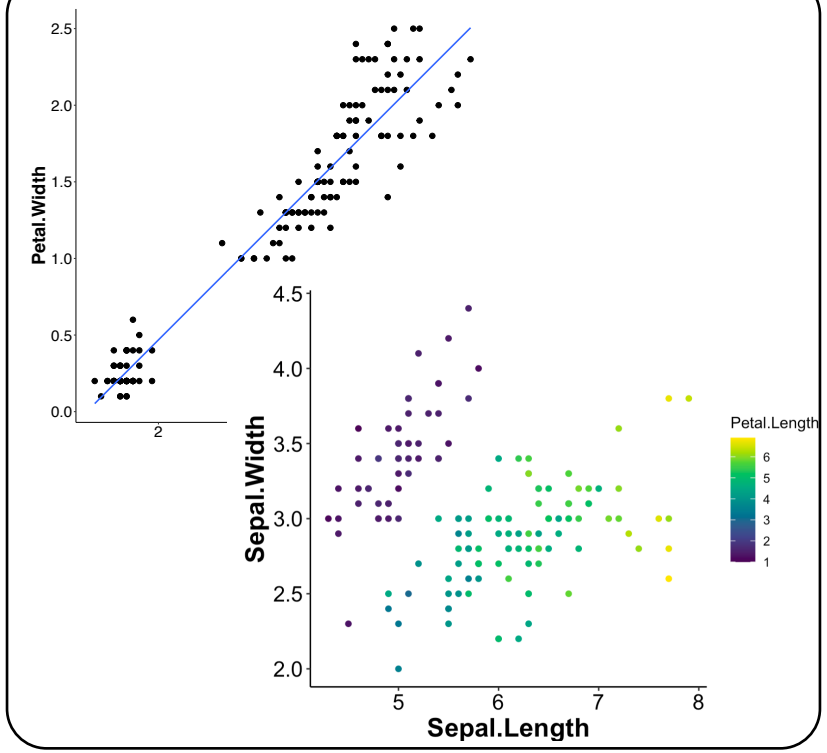
Response: Sepal.Width
                     Sum Sq  Df F value    Pr(>F)
Sepal.Length          4.7689   1  64.299 3.368e-13 ***
Species              15.7225   2 105.995 < 2.2e-16 ***
Sepal.Length:Species  1.5132   2  10.201 7.190e-05 ***
Residuals            10.6800 144
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



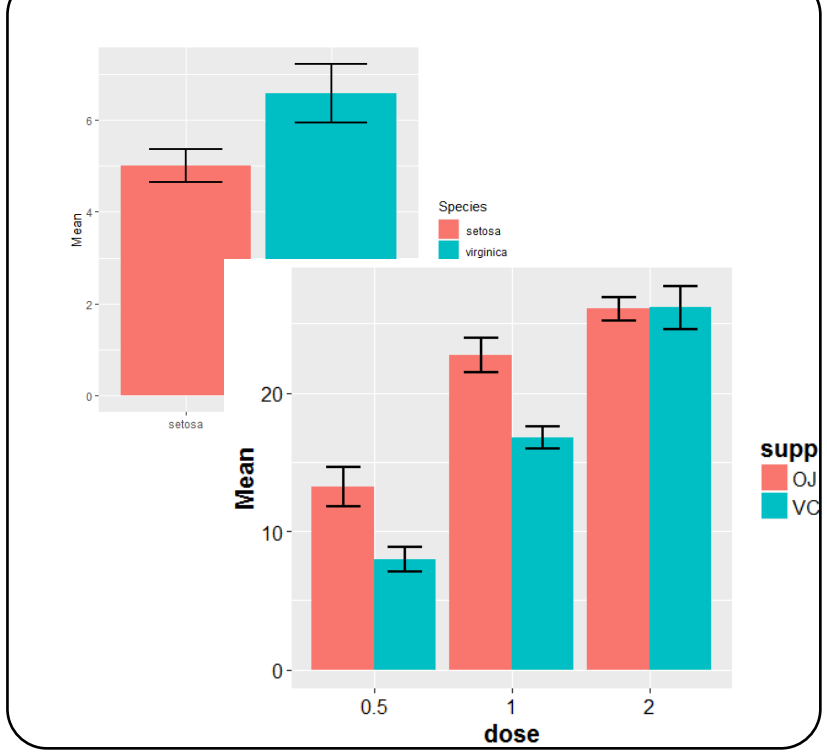
✓ Sepal.Lengthの主効果が有意（$F(1, 144) = 64.299$, $p < .001$）
✓ Speciesの主効果が有意（$F(2, 144) = 105.995$, $p < .001$）
✓ 交互作用が有意（$F(2, 144) = 10.201$, $p < .001$）

●各種分析は，線形モデルの枠組みで統一的に扱うことができる

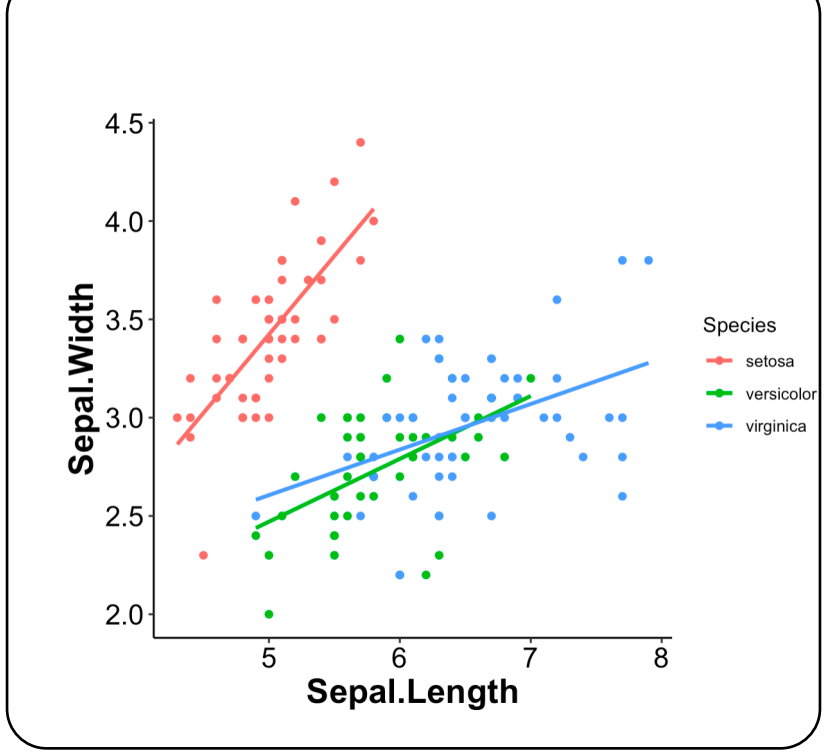
回帰分析



分散分析・t検定



共分散分析



説明変数が質的変数，量的変数，その両方であっても，

$y = \beta_0 + \beta_1 * x + \beta_2 * d + \cdots$ のかたちで予測値の変動を表現できる

質的変数の水準間の多重比較

線形モデルでも多重比較がしたい

- multcompパッケージのglht()関数で多重比較が実行できる
  - ✓パッケージは使う前に読み込み

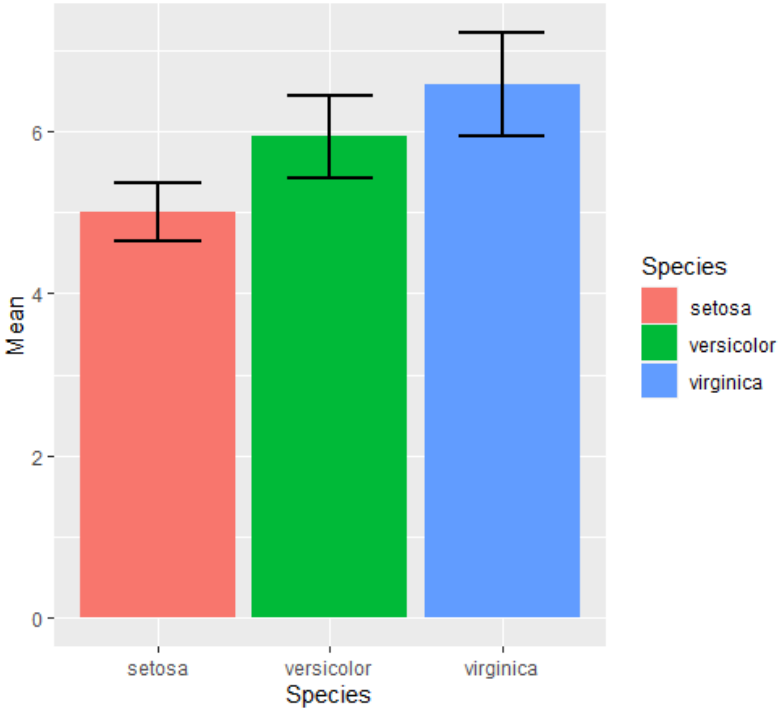
```
library(multcomp)
```

- さっき作ったモデル「m2」を使ってみよう

```
m2
```

```
Call:
lm(formula = Sepal.Length ~ Species, data = iris)
```

- ✓応答変数：iris$Sepal.Length
- ✓説明変数：iris$Species（3水準）



- ✓Anova()関数でF検定を実施した結果，Speciesの効果が有意だった
  ($F(2, 147) = 119.26$, $p < .001$)

●3水準はどんな関係？

✓multcomp::glht()：Tukeyの方法で総当たりの一対比較

```
m2_mp <- glht(m2, linfct = mcp(Species = "Tukey"))
```

- 第一引数：fittedオブジェクト（モデル）
- 第二引数：linfct = mcp(multiple comparison)関数で，比較したい変数と手法を指定

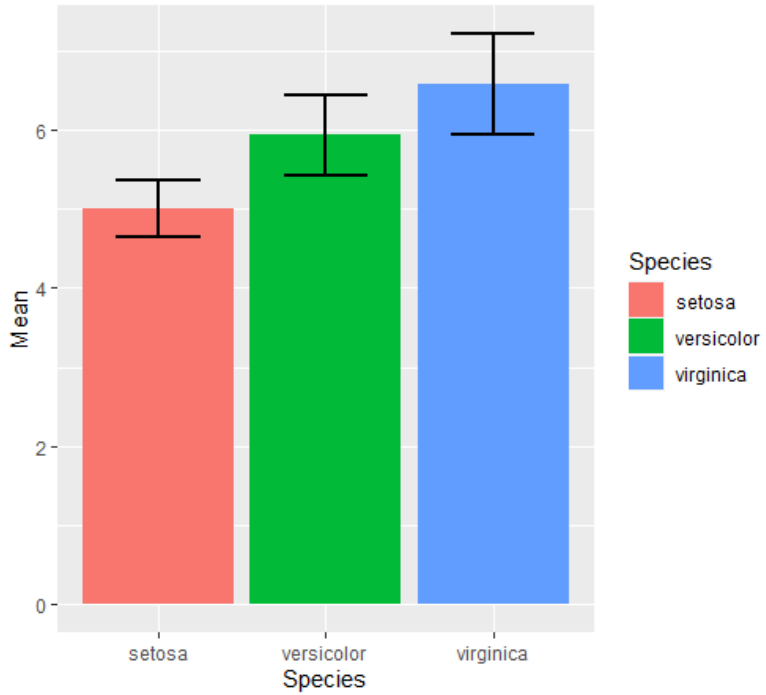
```
summary(m2_mp)
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Sepal.Length ~ Species, data = iris)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
versicolor - setosa == 0       0.930      0.103   9.033   <1e-08 ***
virginica - setosa == 0        1.582      0.103  15.366   <1e-08 ***
virginica - versicolor == 0    0.652      0.103   6.333   <1e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```



線形モデルでも多重比較がしたい

- emmeansパッケージのemmeans()関数でもできる

```
library(emmeans)
```

✓emmeans()関数にfittedオブジェクトを投入

✓第二引数specs = pairwise ~ 比較したい変数

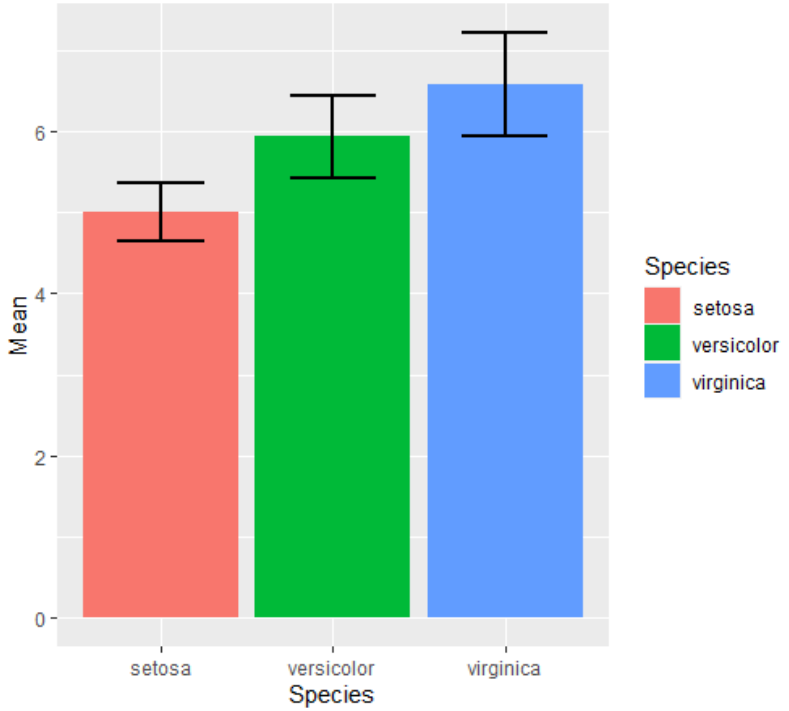
```
emmeans(m2, specs = pairwise ~ Species)
```

```
$emmeans
 Species    emmean     SE  df lower.CL upper.CL
 setosa       5.01 0.0728 147     4.86     5.15
 versicolor   5.94 0.0728 147     5.79     6.08
 virginica    6.59 0.0728 147     6.44     6.73

Confidence level used: 0.95

$contrasts
 contrast               estimate    SE  df t.ratio p.value
 setosa - versicolor      -0.930 0.103 147  -9.033  <.0001
 setosa - virginica       -1.582 0.103 147 -15.366  <.0001
 versicolor - virginica   -0.652 0.103 147  -6.333  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```



●この場合はモデルを使わず，pairwise.t.test()関数でもできる

✓「データフレーム名$列名」で，応答変数，説明変数を指定
✓P値の補正方法を p.adjust.method = “holm”としてHolm法を適用

```
pairwise.t.test(iris$Sepal.Length, iris$Species, p.adjust.method = "holm")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  iris$Sepal.Length and iris$Species

           setosa  versicolor
versicolor 1.8e-15 -
virginica  < 2e-16 2.8e-09

P value adjustment method: holm
```

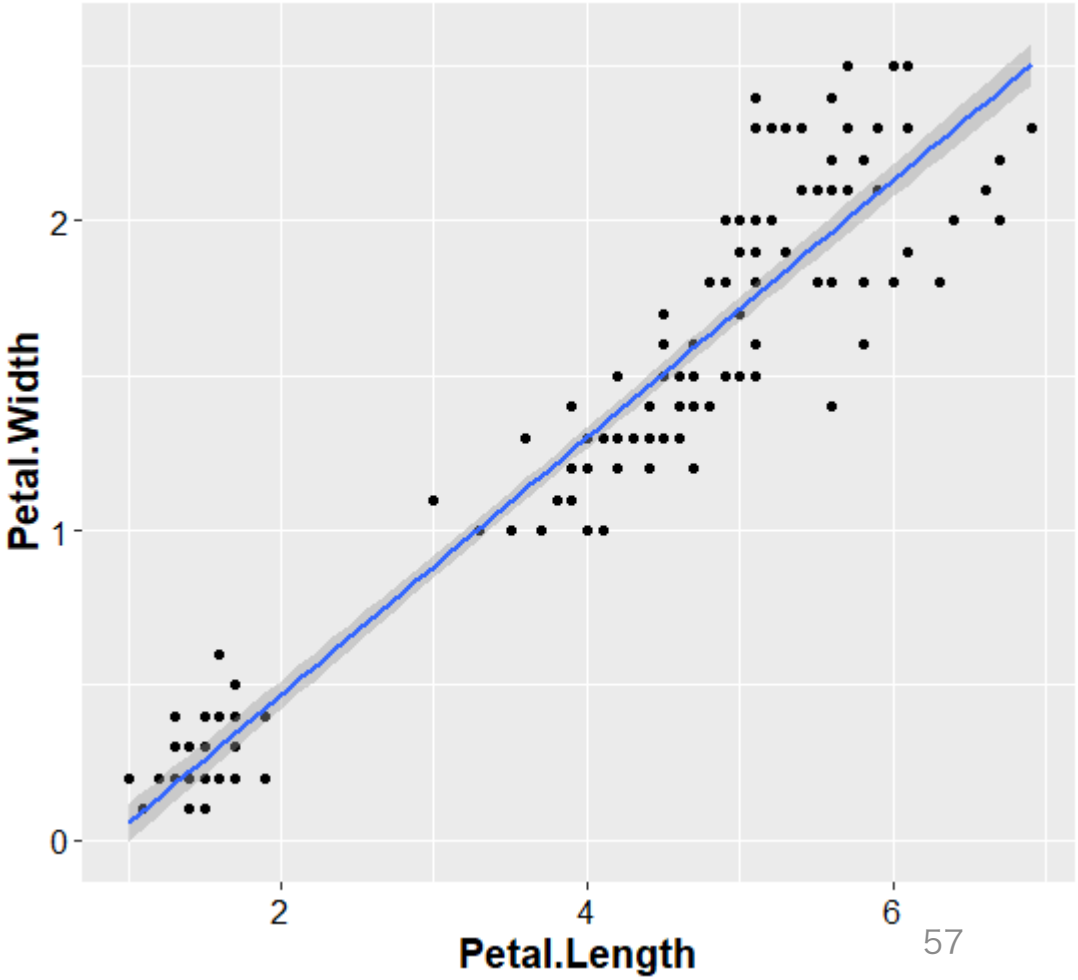
正規線形モデルから一般化線形モデルへ

統計モデルは確率論的過程を仮定

●統計モデルは，決定論的過程と確率論的過程を仮定している

✓**決定論的過程**
- 切片・傾きの推定値をもとに，
  数式で表現できる $(y = \beta_0 + \beta_1 x)$
- 説明変数の値が決まれば
  予測値も一意に決まる

✓**確率論的過程**
- 予測値付近での観測値のばらつき方
- 予測値の値が決まっても，
  具体的な観測値は決定できない
- **確率分布**を用いて表現される



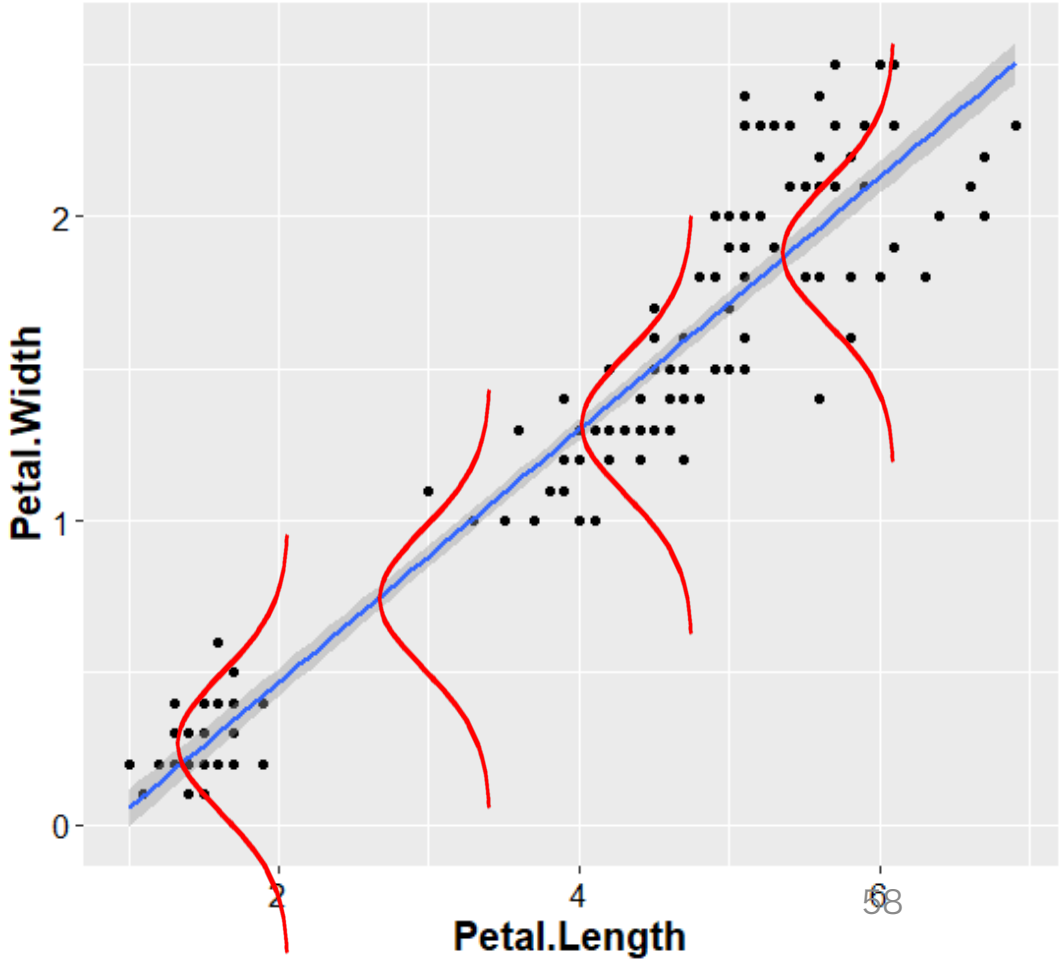
観測値が正規分布から得られると仮定

●これまでの統計モデルは確率分布に正規分布を仮定（正規線形モデル）

✓確率論的過程

- 観測値は正規分布に従う
  - ✓説明変数の値における予測値が平均
  - ✓正規分布の標準偏差は一定

- 応答変数の値は，正規分布に従って確率的に変動する



正規線形モデル

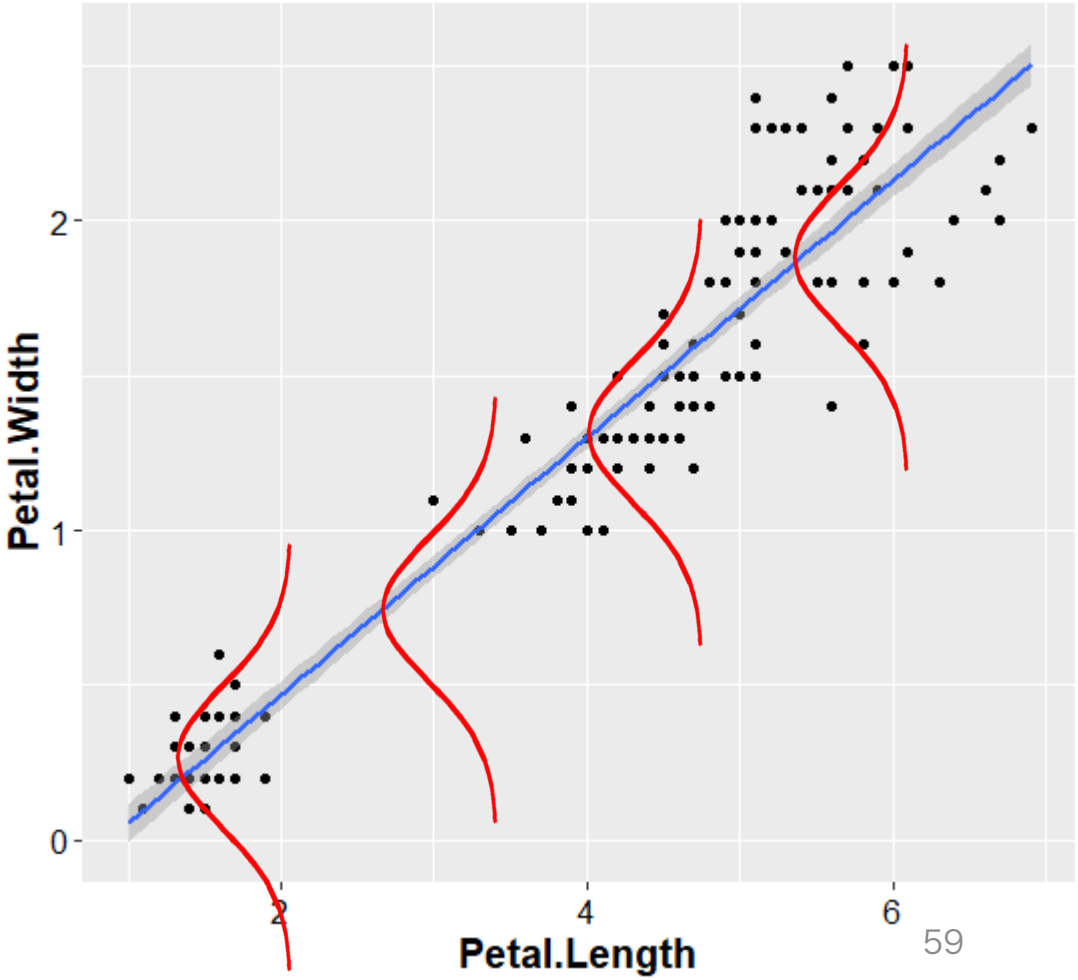
●確率分布に正規分布を仮定した統計モデル

$$Y \sim Normal(\mu, \sigma)$$ **確率分布**

$$\mu = \beta_0 + \beta_1 x + \beta_2 d + \cdots$$ **線形予測子**

✓「～」は確率論的過程
✓「＝」は決定論的過程

- Y：観測値
- μ：正規分布の平均・予測値
- σ：正規分布の標準偏差
- β_0：切片
- β_1など：傾き



正規線形モデル

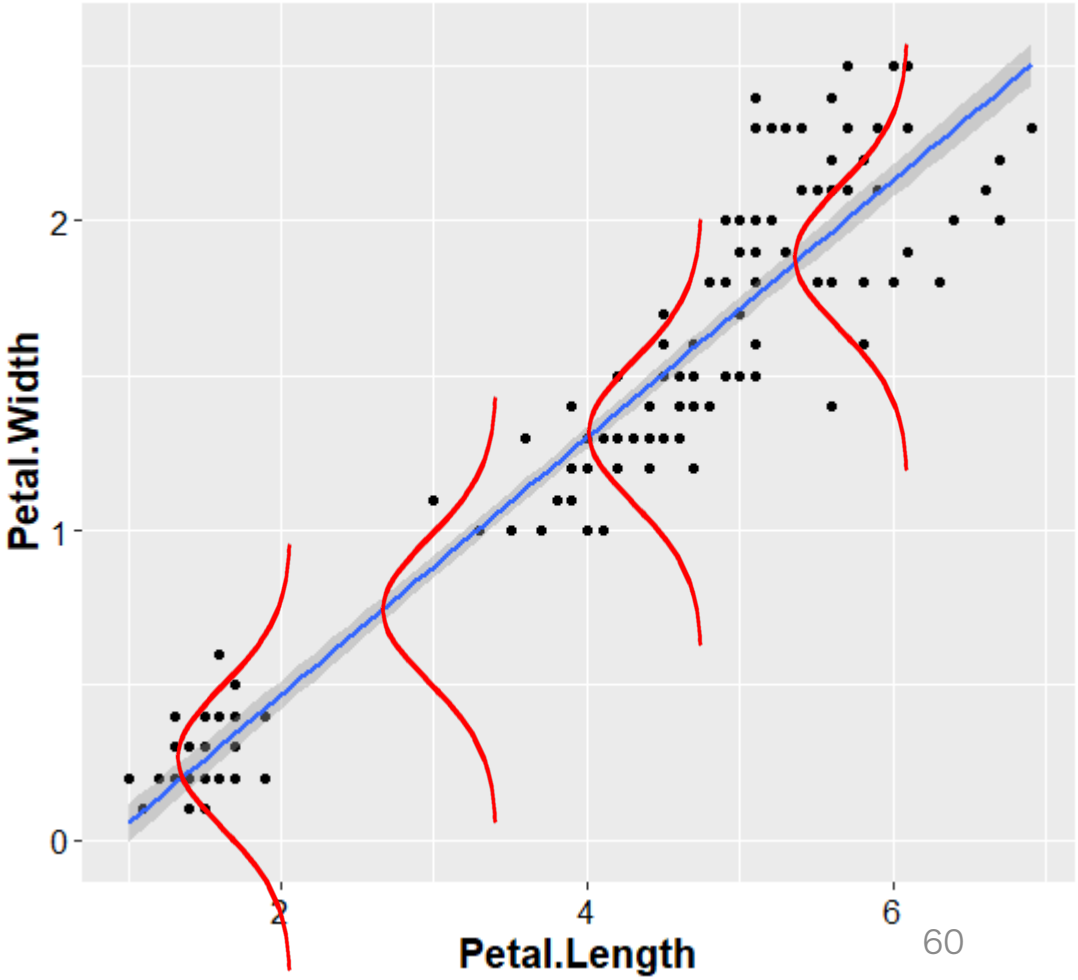
●確率分布に正規分布を仮定した統計モデル

$$Y \sim Normal(\mu, \sigma)$$ **確率分布**

$$\mu = \beta_0 + \beta_1 x + \beta_2 d + \cdots$$

線形予測子

●観測値が正規分布から得られると仮定
- 正規分布の標準偏差は一定と仮定



正規線形モデルの限界

- **正規分布を想定できないデータ**
  - ✓**上限のないカウントデータ**
    - 例：釣りで釣れた魚の数
    - 実測値は**非負の整数**
    - 分布が左右非対称
  - ✓**上限のあるカウントデータ**
    - 例：コイントス10回中で表が出る回数
    - 実測値は**上限のある非負の整数**
    - 上限・下限に近いとき，分布が左右非対称

カウントデータなどでは，
正規分布を仮定することは不適切