# جلسه ۳ – بخش ۴: همبستگی و رگرسیون خطی

🎯 هدف بخش: در این بخش پژوهشگر یاد می‌گیرد رابطه بین دو یا چند متغیر کمی را در داده‌های پژوهشی شیلات، آبزیان و میگو بررسی کند؛ ابتدا با همبستگی شدت و جهت رابطه را بسنجد و سپس با رگرسیون خطی مقدار یک متغیر پاسخ را بر اساس یک یا چند متغیر پیش‌بین مدل‌سازی کند.

---

## تفاوت همبستگی و رگرسیون

| مفهوم | هدف | خروجی اصلی | مثال در شیلات |
|---|---|---|---|
| همبستگی | بررسی شدت و جهت رابطه بین دو متغیر | ضریب همبستگی و `p-value` | رابطه بین شوری آب و نرخ رشد میگو |
| رگرسیون خطی ساده | پیش‌بینی یک متغیر پاسخ با یک متغیر پیش‌بین | معادله خط، ضرایب، `R-squared` و `p-value` | پیش‌بینی وزن ماهی بر اساس طول |
| رگرسیون خطی چندگانه | پیش‌بینی پاسخ با چند متغیر پیش‌بین | اثر هم‌زمان چند عامل | پیش‌بینی رشد میگو با دما، شوری و اکسیژن |

📌 نکته مهم: همبستگی فقط می‌گوید دو متغیر با هم تغییر می‌کنند یا نه؛ اما رگرسیون یک مدل پیش‌بینی و تفسیر اثر ارائه می‌دهد.

---

## داده‌های نمونه برای بخش

در این مثال، داده‌های رشد میگو در یک دوره پرورش فرضی بررسی می‌شود. متغیرهای اندازه‌گیری‌شده شامل دمای آب، شوری، اکسیژن محلول، طول و وزن نهایی میگو هستند.

```
shrimp_growth <- data.frame(
  temperature = c(27.1, 28.0, 28.5, 29.2, 30.0, 30.5, 31.0, 31.5, 32.0, 32.5,
                  27.8, 28.6, 29.4, 30.2, 31.1, 31.8, 32.3, 29.8, 30.7, 28.9),
  salinity = c(22, 23, 24, 25, 26, 27, 28, 29, 30, 31,
               23, 24, 25, 26, 27, 28, 29, 25, 27, 24),
  dissolved_oxygen = c(6.8, 6.7, 6.5, 6.4, 6.2, 6.1, 5.9, 5.8, 5.7, 5.5,
                       6.6, 6.5, 6.3, 6.1, 5.9, 5.8, 5.6, 6.2, 6.0, 6.4),
  length = c(8.1, 8.4, 8.7, 9.0, 9.2, 9.5, 9.7, 9.9, 10.0, 10.2,
             8.3, 8.6, 8.9, 9.3, 9.6, 9.8, 10.1, 9.1, 9.4, 8.8),
  weight = c(9.8, 10.5, 11.2, 12.1, 12.8, 13.4, 14.0, 14.5, 14.9, 15.3,
             10.3, 11.0, 11.9, 12.9, 13.8, 14.4, 15.1, 12.4, 13.2, 11.5)
)

head(shrimp_growth)
```

```
  temperature salinity dissolved_oxygen length weight
1        27.1       22              6.8    8.1    9.8
2        28.0       23              6.7    8.4   10.5
3        28.5       24              6.5    8.7   11.2
4        29.2       25              6.4    9.0   12.1
5        30.0       26              6.2    9.2   12.8
6        30.5       27              6.1    9.5   13.4
```

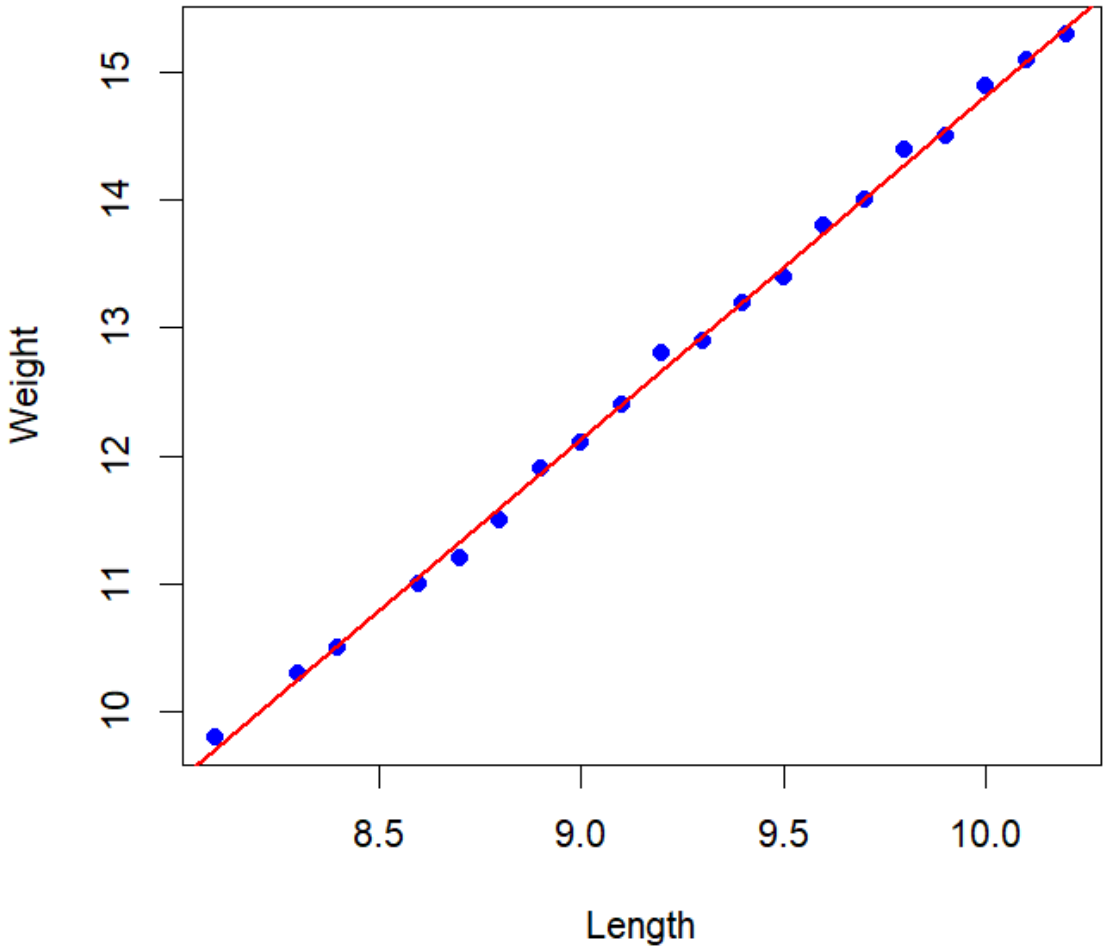
---

## نمودار پراکنش برای بررسی اولیه رابطه

پیش از اجرای آزمون همبستگی یا رگرسیون، باید رابطه بین دو متغیر با نمودار پراکنش بررسی شود.

```
plot(shrimp_growth$length, shrimp_growth$weight,
     xlab = "Length",
     ylab = "Weight",
```

```
     pch = 19,
     col = "blue")

abline(lm(weight ~ length, data = shrimp_growth),
       col = "red",
       lwd = 2)
```




📌 تفسیر نمودار: اگر نقاط به‌صورت تقریبی در اطراف یک خط صعودی قرار گرفته باشند، رابطه مثبت خطی بین طول و وزن وجود دارد. اگر روند نزولی باشد، رابطه منفی است. اگر الگوی مشخصی دیده نشود، رابطه خطی ضعیف یا ناموجود است.

---

## همبستگی پیرسون

همبستگی پیرسون زمانی استفاده می‌شود که هر دو متغیر کمی باشند و رابطه بین آن‌ها تقریباً خطی باشد.

### کاربرد

📍 مثال: آیا بین طول و وزن نهایی میگو رابطه معنی‌دار وجود دارد؟

### فرض‌ها

- هر دو متغیر کمی باشند.
- رابطه بین دو متغیر تقریباً خطی باشد.
- داده‌ها دارای پرت شدید نباشند.
- توزیع دو متغیر یا حداقل الگوی رابطه برای تحلیل پیرسون مناسب باشد.

### فرض صفر و فرض مقابل

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

در اینجا $\rho$ ضریب همبستگی جامعه است.

## اجرای آزمون در R

```
cor.test(shrimp_growth$length, shrimp_growth$weight,
         method = "pearson")
```

```
Pearson's product-moment correlation

data:  shrimp_growth$length and shrimp_growth$weight
t = 101.26, df = 18, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9977332 0.9996611
sample estimates:
      cor
0.9991234
```

📌 نتیجه‌گیری: چون `p-value < 0.05` است، فرض صفر رد می‌شود. بین طول و وزن میگو رابطه خطی معنی‌دار وجود دارد. مقدار مثبت و نزدیک به ۱ ضریب همبستگی نشان می‌دهد رابطه بسیار قوی و مستقیم است.

---

## تفسیر ضریب همبستگی

ضریب همبستگی پیرسون با $r$ نمایش داده می‌شود و مقدار آن بین ۱- و ۱+ است.

$$-1 \leq r \leq 1$$

| مقدار تقریبی `r` | شدت رابطه | تفسیر |
|---|---|---|
| نزدیک به `0` | بسیار ضعیف | رابطه خطی مشخص وجود ندارد |
| `0.1` تا `0.3` | ضعیف | رابطه کم |
| `0.3` تا `0.5` | متوسط | رابطه قابل توجه |
| `0.5` تا `0.7` | نسبتاً قوی | رابطه واضح |
| بالاتر از `0.7` | قوی | رابطه بسیار مشخص |
| نزدیک به `1` یا `-1` | بسیار قوی | رابطه تقریباً خطی کامل |

📌 نکته پژوهشی: معنی‌دار بودن آماری الزاماً به معنی اهمیت زیستی یا کاربردی نیست. در پژوهش‌های شیلات، مقدار اثر و تفسیر زیستی باید همراه با `p-value` گزارش شود.

---

## همبستگی اسپیرمن

اگر داده‌ها نرمال نباشند، رابطه خطی نباشد اما روند یکنواخت داشته باشد، یا داده‌ها رتبه‌ای باشند، از همبستگی اسپیرمن استفاده می‌شود.

### کاربرد

📍 مثال: آیا با افزایش شوری، وزن نهایی میگو به‌صورت یکنواخت افزایش می‌یابد؟

### فرض‌ها

- متغیرها حداقل رتبه‌ای یا کمی باشند.
- رابطه بین دو متغیر یکنواخت باشد.
- نیاز جدی به نرمال بودن داده‌ها ندارد.
- نسبت به داده‌های پرت مقاوم‌تر از پیرسون است.

## فرض صفر و فرض مقابل

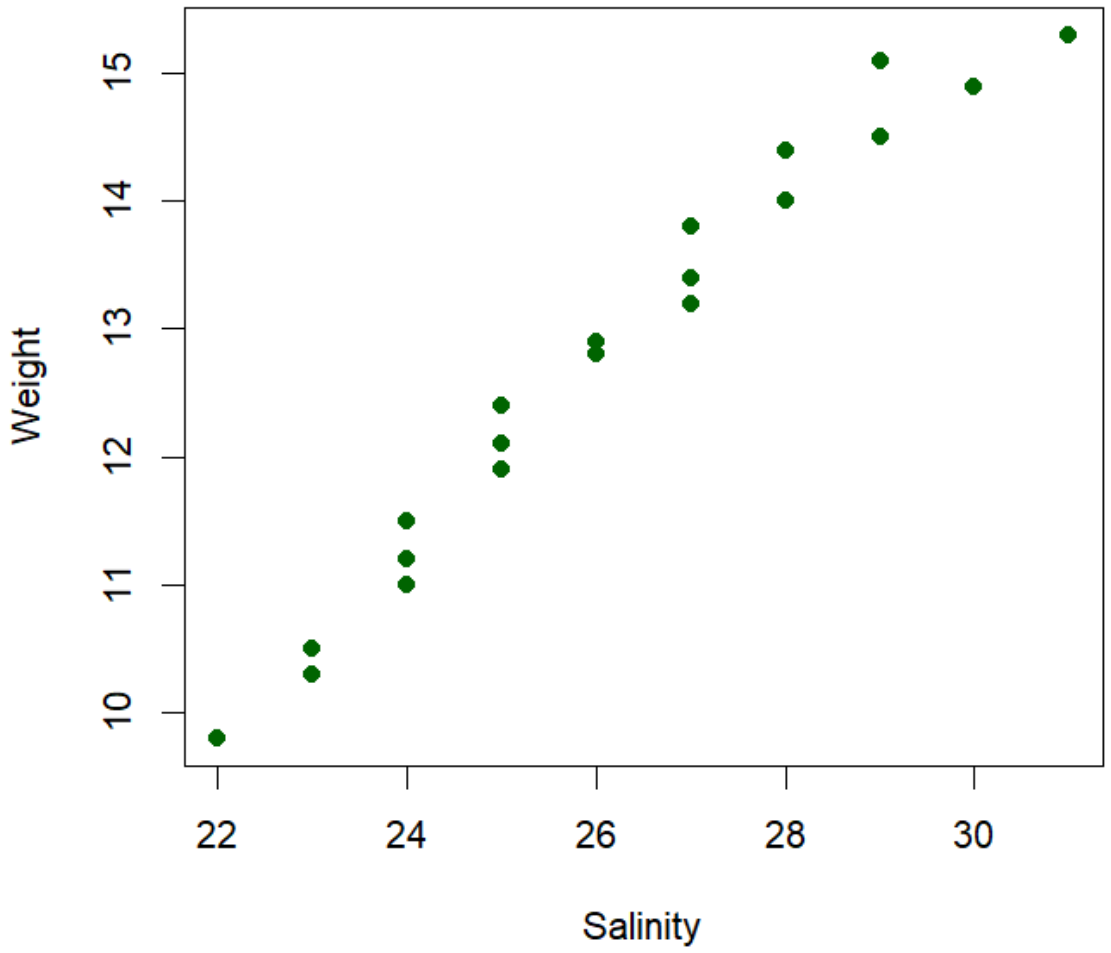
$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

در اینجا $\rho_s$ ضریب همبستگی رتبه‌ای اسپیرمن است.

## نمودار مناسب

```
plot(shrimp_growth$salinity, shrimp_growth$weight,
     xlab = "Salinity",
     ylab = "Weight",
     pch = 19,
     col = "darkgreen")
```




## اجرای آزمون در R

```
cor.test(shrimp_growth$salinity, shrimp_growth$weight,
         method = "spearman")
```

خروجی خلاصه:

```
Spearman's rank correlation rho

data:  shrimp_growth$salinity and shrimp_growth$weight
S = 11.042, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9916974
```

📌 نتیجه‌گیری: چون `p-value < 0.05` است، رابطه رتبه‌ای معنی‌دار بین شوری و وزن میگو وجود دارد. مقدار مثبت ضریب اسپیرمن نشان می‌دهد با افزایش شوری، وزن نهایی نیز در این داده‌ها افزایش یافته است.

---

## ماتریس همبستگی برای چند متغیر

وقتی چند متغیر کمی وجود دارد، می‌توان ماتریس همبستگی را برای بررسی اولیه روابط استفاده کرد.

```
cor(shrimp_growth)
```

```
                 temperature   salinity dissolved_oxygen     length     weight
temperature        1.0000000  0.9770534       -0.9940614  0.9943990  0.9958072
salinity           0.9770534  1.0000000       -0.9788717  0.9826908  0.9822243
dissolved_oxygen  -0.9940614 -0.9788717        1.0000000 -0.9889552 -0.9898355
length             0.9943990  0.9826908       -0.9889552  1.0000000  0.9991234
weight             0.9958072  0.9822243       -0.9898355  0.9991234  1.0000000
```

برای گرد کردن خروجی:

```
round(cor(shrimp_growth), 2)
```

```
                 temperature salinity dissolved_oxygen length weight
temperature             1.00     0.98            -0.99   0.99   1.00
salinity                0.98     1.00            -0.98   0.98   0.98
dissolved_oxygen       -0.99    -0.98             1.00  -0.99  -0.99
length                  0.99     0.98            -0.99   1.00   1.00
weight                  1.00     0.98            -0.99   1.00   1.00
```

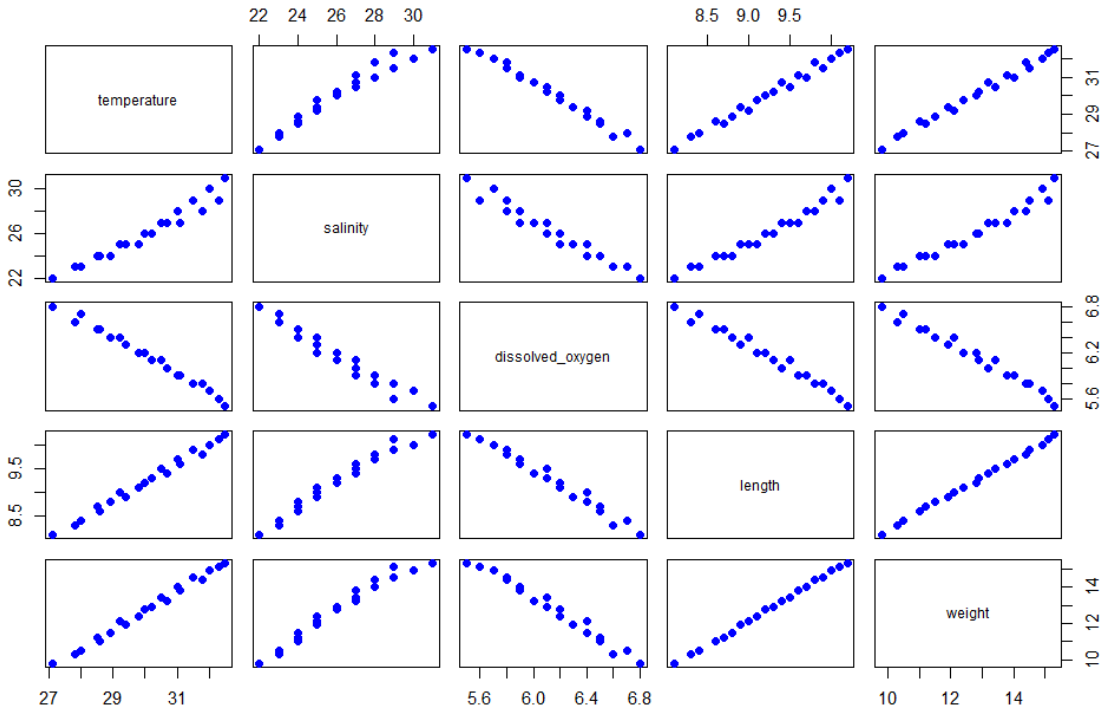
📌 تفسیر: در این داده‌ها، وزن میگو با طول، دما و شوری رابطه مثبت دارد، اما با اکسیژن محلول رابطه منفی دیده می‌شود. این تفسیر باید با دانش زیستی و طراحی آزمایش بررسی شود، زیرا همبستگی بالا همیشه به معنی رابطه علی نیست.

---

## نمودار ماتریس پراکنش

برای مشاهده هم‌زمان روابط بین چند متغیر کمی از نمودار ماتریس پراکنش استفاده می‌شود.

```
pairs(shrimp_growth,
      pch = 19,
      col = "blue")
```




📌 کاربرد نمودار: این نمودار کمک می‌کند رابطه‌های خطی، غیرخطی، داده‌های پرت و خوشه‌بندی احتمالی بین متغیرها بررسی شود.

---

## رگرسیون خطی ساده

رگرسیون خطی ساده زمانی استفاده می‌شود که بخواهیم یک متغیر پاسخ کمی را بر اساس یک متغیر پیش‌بین کمی مدل‌سازی یا پیش‌بینی کنیم.

📍 مثال: پیش‌بینی وزن نهایی میگو بر اساس طول بدن.

مدل رگرسیون خطی ساده به‌صورت زیر است:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

در این مدل:

| نماد | مفهوم |
|---|---|
| $Y_i$ | متغیر پاسخ؛ وزن میگو |
| $X_i$ | متغیر پیش‌بین؛ طول میگو |
| $\beta_0$ | عرض از مبدأ |
| $\beta_1$ | شیب خط رگرسیون |
| $\varepsilon_i$ | خطای تصادفی |

---

## اجرای مدل رگرسیون خطی ساده

```
model_simple <- lm(weight ~ length, data = shrimp_growth)

summary(model_simple)
```

خروجی خلاصه:

```
Call:
lm(formula = weight ~ length, data = shrimp_growth)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.12225 -0.05337 -0.01102  0.05378  0.13082 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.1144     0.2461  -49.23   <2e-16 ***
length        2.6939     0.0266  101.26   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07238 on 18 degrees of freedom
Multiple R-squared:  0.9982,	Adjusted R-squared:  0.9982 
F-statistic: 1.025e+04 on 1 and 18 DF,  p-value: < 2.2e-16
```

### معادله مدل

بر اساس خروجی بالا:

$$\widehat{Weight} = -12.1144 + 2.6939 \times Length$$

- تفسیر شیب: به‌ازای افزایش یک واحد در طول میگو، وزن نهایی به‌طور متوسط حدود `2.6939` واحد افزایش می‌یابد.
- تفسیر `p-value` مربوط به شیب: چون `p-value < 0.05` است، اثر طول بر وزن معنی‌دار است.
- تفسیر `R-squared` : مقدار `R-squared = 0.9982` یعنی حدود ۹۹.۸۲٪ از تغییرات وزن میگو توسط طول توضیح داده می‌شود.

---

## فرضیه‌ها در رگرسیون خطی ساده

برای بررسی معنی‌داری اثر متغیر پیش‌بین، فرضیه مربوط به شیب مدل بررسی می‌شود.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

اگر $\beta_1 = 0$ باشد، یعنی متغیر پیش‌بین رابطه خطی معنی‌داری با پاسخ ندارد.

- قاعده تصمیم: اگر `p-value` مربوط به متغیر `length` کمتر از ۰.۰۵ باشد، فرض صفر رد می‌شود و اثر طول بر وزن معنی‌دار گزارش می‌شود.

---

## بررسی مفروضات رگرسیون خطی

رگرسیون خطی فقط با معنی‌دار بودن `p-value` کافی نیست. باید مفروضات مدل نیز بررسی شوند.

### مفروضات اصلی

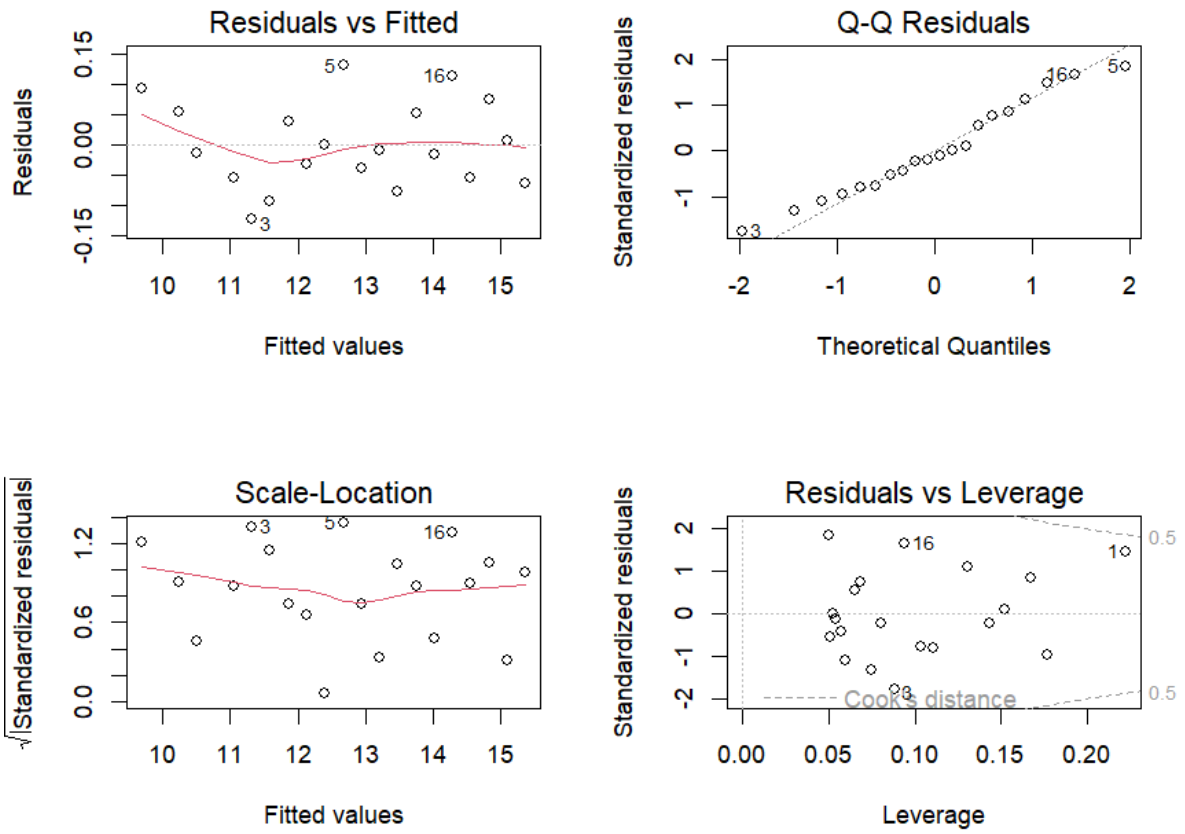
| مفروضه | مفهوم | روش بررسی |
|---|---|---|
| خطی بودن | رابطه پاسخ و پیش‌بین خطی باشد | نمودار پراکنش و `Residuals vs Fitted` |
| نرمال بودن باقی‌مانده‌ها | خطاها تقریباً نرمال باشند | Shapiro–Wilk و Q–Q plot |
| همگنی واریانس | پراکندگی خطاها ثابت باشد | `Residuals vs Fitted` و `Scale-Location` |
| استقلال خطاها | مشاهدات مستقل باشند | طراحی آزمایش و ترتیب زمانی |
| نبود داده پرت اثرگذار | یک مشاهده مدل را بیش از حد تحت تأثیر قرار ندهد | Cook's distance و `Residuals vs Leverage` |

---

## نمودارهای تشخیصی مدل

```r
par(mfrow = c(2, 2))
plot(model_simple)
```

```
par(mfrow = c(1, 1))
```




| نمودار | کاربرد | وضعیت مطلوب |
|---|---|---|
| `Residuals vs Fitted` | بررسی خطی بودن و همگنی واریانس | پراکندگی تصادفی بدون الگوی مشخص |
| `Normal Q-Q` | بررسی نرمال بودن باقی‌مانده‌ها | قرار گرفتن نقاط نزدیک خط |
| `Scale-Location` | بررسی ثبات واریانس | پراکندگی تقریباً یکنواخت |
| `Residuals vs Leverage` | شناسایی نقاط اثرگذار | نبود نقاط با اهرم و Cook بالا |

📌 **نکته: اگر در نمودار `Residuals vs Fitted` الگوی خمیده دیده شود، مدل خطی ساده ممکن است مناسب نباشد. اگر پراکندگی باقی‌مانده‌ها با افزایش مقدار برازش‌شده زیاد شود، مشکل ناهمگنی واریانس مطرح است.**

---

## آزمون نرمال بودن باقی‌مانده‌ها

```
shapiro.test(residuals(model_simple))
```

**خروجی خلاصه:**

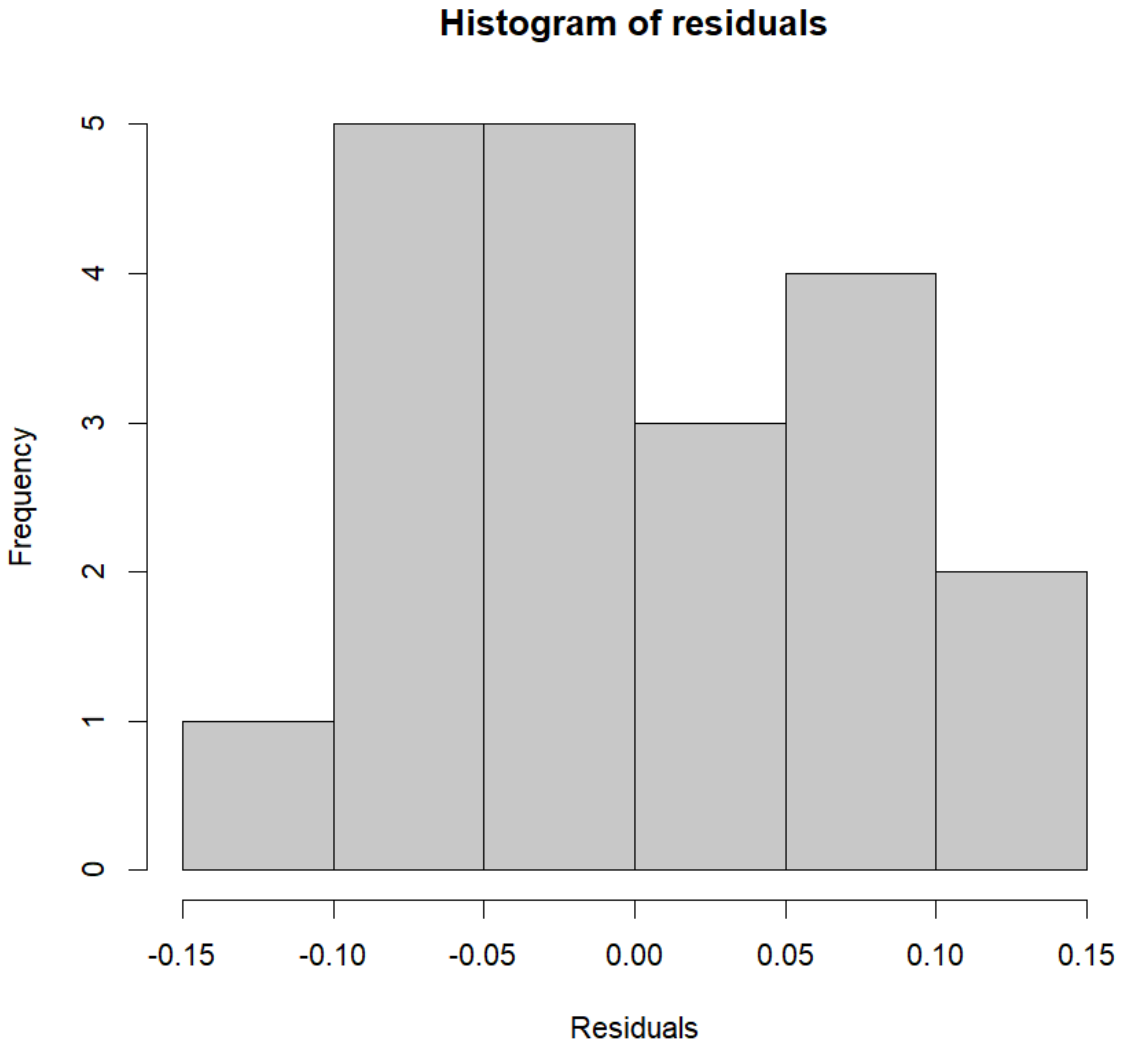
```
Shapiro-Wilk normality test

data:  residuals(model_simple)
W = 0.97274, p-value = 0.8114
```

📌 **نتیجه‌گیری: چون `p-value > 0.05` است، شواهد کافی علیه نرمال بودن باقی‌مانده‌ها وجود ندارد. بنابراین مفروضه نرمال بودن باقی‌مانده‌ها قابل قبول است.**

---

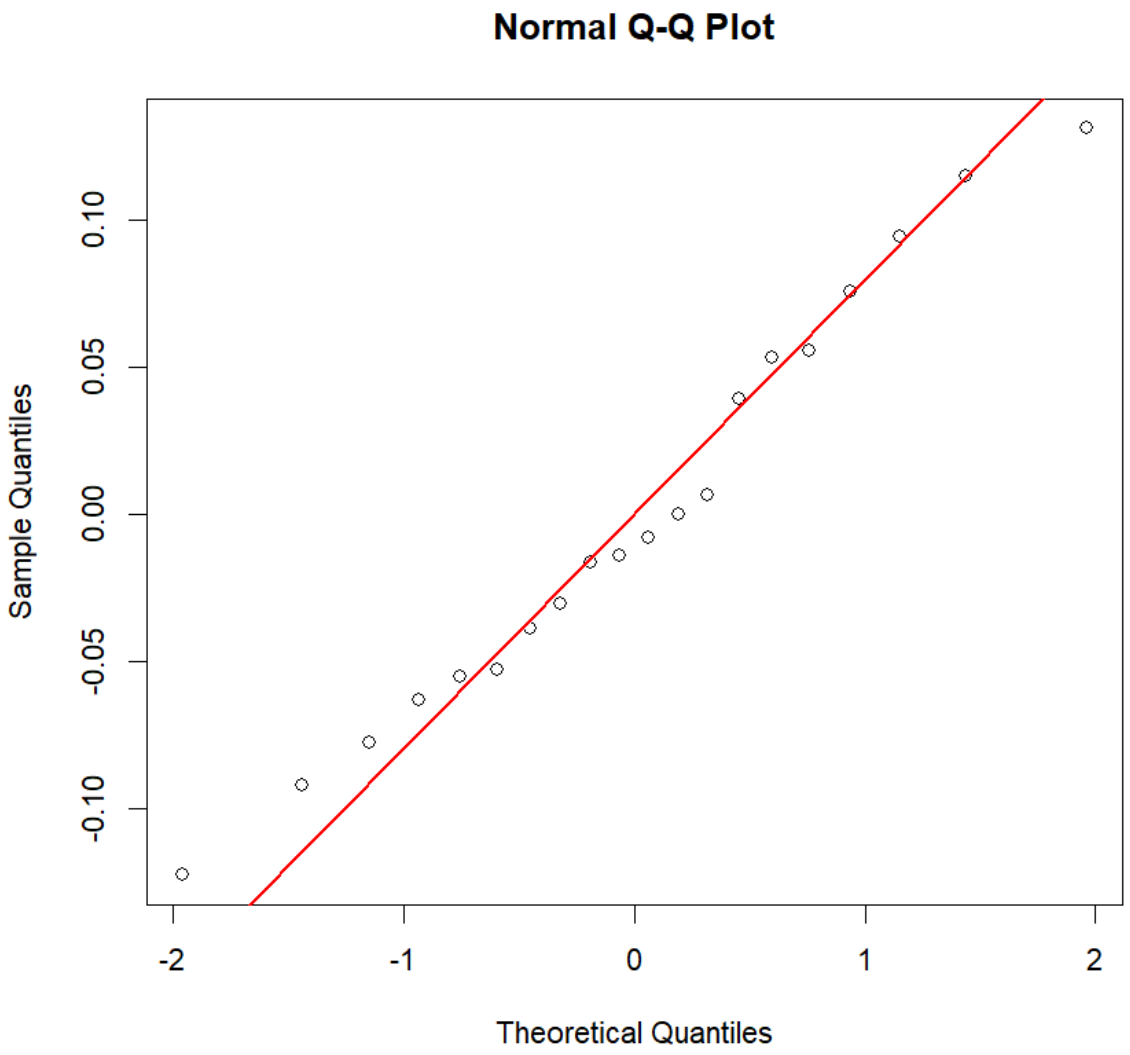
## بررسی باقی‌مانده‌ها با نمودار

```r
hist(residuals(model_simple),
     xlab = "Residuals",
     main = "Histogram of residuals",
     col = "gray80")
```




📌 تفسیر: هیستوگرام باید تقریباً زنگوله‌ای باشد و در نمودار Q-Q نقاط باید نزدیک خط قرمز قرار گیرند.

```r
qqnorm(residuals(model_simple))
qqline(residuals(model_simple), col = "red", lwd = 2)
```




📌 تفسیر: نمودار Q–Q نقاط باید نزدیک خط قرمز قرار گیرند.

---

## رگرسیون خطی چندگانه

رگرسیون خطی چندگانه زمانی استفاده می‌شود که یک متغیر پاسخ کمی با چند متغیر پیش‌بین کمی یا کیفی مدل‌سازی شود.

📍 مثال: پیش‌بینی وزن نهایی میگو بر اساس دما، شوری و اکسیژن محلول.

مدل کلی:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

در این مثال:

$$Weight_i = \beta_0 + \beta_1 Temperature_i + \beta_2 Salinity_i + \beta_3 DissolvedOxygen_i + \varepsilon_i$$

---

## اجرای رگرسیون خطی چندگانه

```
model_multiple <- lm(weight ~ temperature + salinity + dissolved_oxygen,
                     data = shrimp_growth)

summary(model_multiple)
```

```
Call:
lm(formula = weight ~ temperature + salinity + dissolved_oxygen,
    data = shrimp_growth)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.24113 -0.12118  0.04483  0.08884  0.19919 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -24.99835   11.06300  -2.260 0.038148 *  
temperature        0.98054    0.19577   5.009 0.000129 ***
salinity           0.15493    0.06608   2.345 0.032281 *  
dissolved_oxygen   0.68889    0.86508   0.796 0.437499    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1447 on 16 degrees of freedom
Multiple R-squared:  0.9938,	Adjusted R-squared:  0.9926 
F-statistic:   851 on 3 and 16 DF,  p-value: < 2.2e-16
```

📌 تفسیر کاربردی: اگر `p-value` متغیر `temperature` کمتر از ۰.۰۵ باشد، اثر دما بر وزن نهایی میگو با کنترل اثر شوری و اکسیژن معنی‌دار است.

---

## فرضیه‌ها در رگرسیون چندگانه

### آزمون کلی مدل

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \text{حداقل یکی از ضرایب برابر صفر نیست}$$

📌 اگر `p-value` کلی مدل کمتر از ۰.۰۵ باشد، مدل رگرسیونی از نظر آماری معنی‌دار است.

### آزمون هر ضریب

برای هر متغیر پیش‌بین:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

📌 اگر `p-value` مربوط به یک متغیر کمتر از ۰.۰۵ باشد، آن متغیر با ثابت بودن سایر متغیرها اثر معنی‌دار بر پاسخ دارد.

---

## پیش‌بینی با مدل رگرسیون

پس از برازش مدل، می‌توان برای مشاهدات جدید پیش‌بینی انجام داد.

📍 مثال: پیش‌بینی وزن نهایی میگو با طول `9.5` .

```r
new_shrimp <- data.frame(
  length = c(9.5)
)

predict(model_simple, newdata = new_shrimp)
```

خروجی خلاصه:

```
       1 
13.47734 
```

📌 تفسیر: بر اساس مدل خطی ساده، وزن مورد انتظار برای میگویی با طول `9.5` حدود ۱۳.۴۷` واحد است.

### پیش‌بینی با بازه اطمینان

```
predict(model_simple, newdata = new_shrimp,
        interval = "confidence")
```

خروجی خلاصه:

```
     fit      lwr      upr
13.47734   13.44014   13.51455
```

### پیش‌بینی با بازه پیش‌بینی

```
predict(model_simple, newdata = new_shrimp,
        interval = "prediction")
```

```
       fit      lwr      upr
1 13.47734 13.32079 13.63389
```

📌 تفسیر: بازه پیش‌بینی معمولاً از بازه اطمینان پهن‌تر است، زیرا عدم قطعیت مربوط به یک مشاهده جدید را نیز در نظر می‌گیرد.

---

## مقایسه مدل‌ها

گاهی پژوهشگر می‌خواهد بداند آیا افزودن متغیر جدید باعث بهبود معنی‌دار مدل می‌شود یا نه.

```
model_temp <- lm(weight ~ temperature, data = shrimp_growth)

model_temp_sal <- lm(weight ~ temperature + salinity, data = shrimp_growth)

anova(model_temp, model_temp_sal)
```

خروجی خلاصه:

```
Analysis of Variance Table

Model 1: weight ~ temperature
Model 2: weight ~ temperature + salinity
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     18 0.45029
2     17 0.34842  1   0.10187 4.9705 0.03956 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

📌 نتیجه‌گیری: چون `p-value < 0.05` است، افزودن متغیر شوری باعث بهبود معنی‌دار مدل نسبت به مدل فقط با دما شده است.

---

## گزارش‌نویسی نتایج همبستگی

برای گزارش همبستگی در مقاله یا پایان‌نامه، باید نوع آزمون، مقدار ضریب همبستگی، جهت رابطه و `p-value` ذکر شود.

### نمونه گزارش

🔬 بین طول و وزن نهایی میگو رابطه مثبت و معنی‌دار مشاهده شد. نتیجه آزمون همبستگی پیرسون نشان داد که این رابطه بسیار قوی است ( `r = 0.998` , `p-value < 0.001` ). بنابراین، با افزایش طول بدن، وزن نهایی میگو نیز افزایش می‌یابد.

---

# گزارش‌نویسی نتایج رگرسیون

برای گزارش رگرسیون باید معادله مدل، معنی‌داری ضرایب، مقدار `R-squared` و وضعیت مفروضات بیان شود.

## نمونه گزارش

🔬 مدل رگرسیون خطی ساده برای پیش‌بینی وزن نهایی میگو بر اساس طول بدن معنی‌دار بود ( `p-value < 0.001` ). مقدار `R-squared = 0.9989` نشان داد که حدود ۹۹.۸۹٪ تغییرات وزن توسط طول بدن توضیح داده می‌شود. شیب مدل مثبت و معنی‌دار بود، بنابراین افزایش طول بدن با افزایش وزن نهایی همراه است.