Statistical hw2

全金

2025-03-16

3

(a)

iii

$$\begin{aligned}\text{Salary} = 50 + 20 \cdot \text{GPA} + 0.07 \cdot \text{IQ} + 35 \cdot \text{Level} \\ + 0.01 \cdot (\text{GPA} \times \text{IQ}) - 10 \cdot (\text{GPA} \times \text{Level})\end{aligned}$$

- College

$$\begin{aligned}\text{Salary}_{\text{College}} &= 50 + 20\text{GPA} + 0.07\text{IQ} + 35(1) \\ &\quad + 0.01(\text{GPA} \times \text{IQ}) - 10(\text{GPA} \times 1) \\ &= 85 + 10\text{GPA} + 0.07\text{IQ} + 0.01(\text{GPA} \times \text{IQ})\end{aligned}$$

- High School

$$\text{Salary}_{\text{HS}} = 50 + 20\text{GPA} + 0.07\text{IQ} + 0.01(\text{GPA} \times \text{IQ})$$

$$\text{Salary}_{\text{College}} - \text{Salary}_{\text{HS}} = 35 - 10\text{GPA}$$

- When $35 - 10\text{GPA} > 0 \implies \text{GPA} < 3.5$, college graduates earn higher salaries.
- When $\text{GPA} > 3.5$, high school graduates earn higher salaries.

(b)

```
beta0 <- 50
beta1 <- 20
beta2 <- 0.07
beta3 <- 35
beta4 <- 0.01
beta5 <- -10
```

```r
GPA <- 4.0
IQ <- 110
Level <- 1

salary <- beta0 +
          beta1 * GPA +
          beta2 * IQ +
          beta3 * Level +
          beta4 * (GPA * IQ) +
          beta5 * (GPA * Level)

cat("Salary:", round(salary, 1), "kUSD")
```

```
## Salary: 137.1 kUSD
```

(c)

不对。系数大小不能直接推断统计显著性，需看其标准误和 p 值。

4

(a)

训练：三次回归小于线性回归三次模型包含更多参数，过拟合训练数据，训练 RSS 会更低。

(b)

测试：线性小于三次。三次回归可能过拟合。

(c)

三次回归更低，参数更多更灵活。

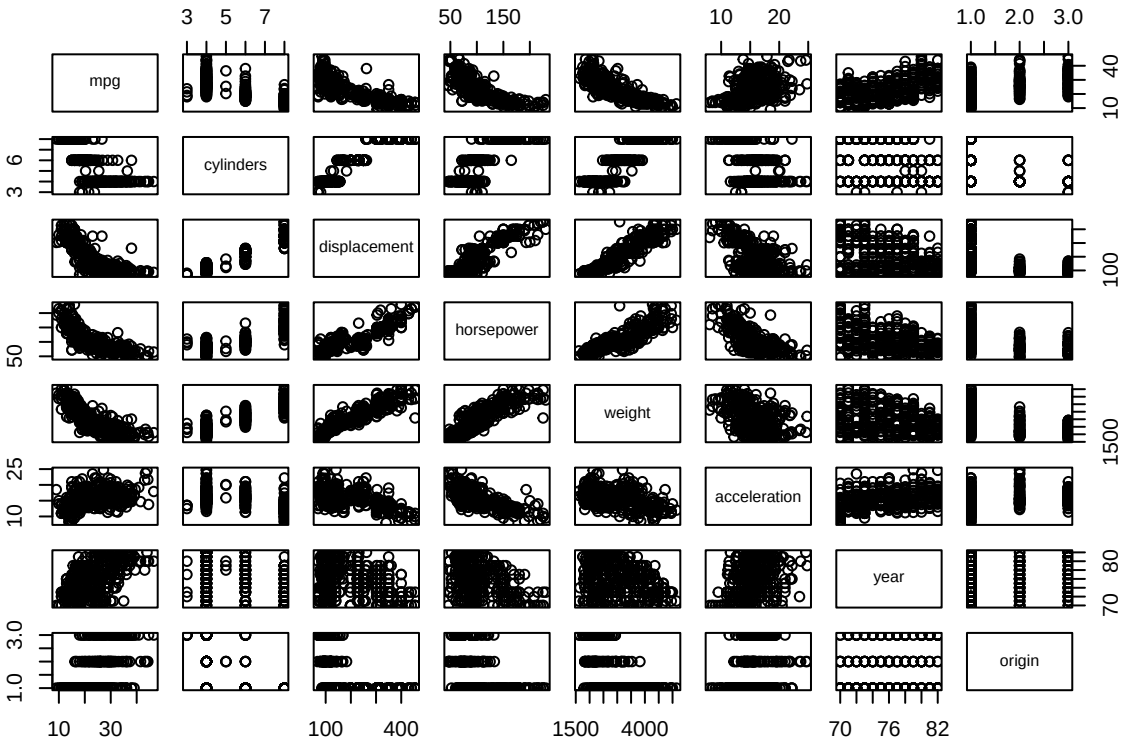
(d)

无法判断。若更接近线性则线性更小，若高度非线性则三次更小。

9

(a)

```
library(ISLR2)
data(Auto)
pairs(Auto[, -9])
```



(b)

```
cor_matrix <- cor(Auto[, -9])
print(cor_matrix)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
```

```
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c)

```
model <- lm(mpg ~ . - name, data = Auto)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i: 是的，预测变量整体与响应变量 mpg 之间存在显著关系。

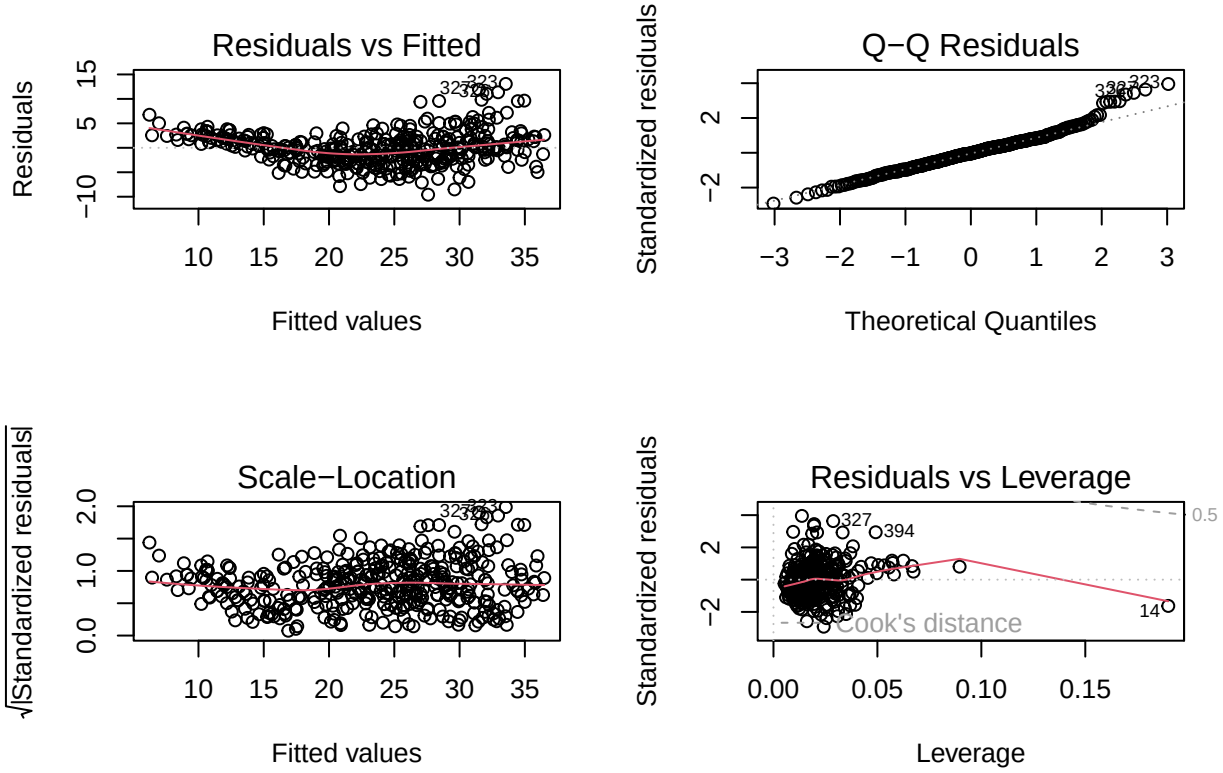
ii: 在显著性水平 =0.05 下，以下预测变量的 p 值小于 0.05，具有统计显著性：displacement，weight，year，origin。

iii: 表示在其他变量不变的情况下，汽车的生产年份每增加 1 年，mpg 平均增加约 0.75 英里每加仑，说明

随着时间推移，汽车的燃油效率有显著提升。

(d)

```
par(mfrow = c(2, 2))
plot(model)
```



如图，残差图有明显离群点，杠杆图有异常高杠杆作用点。

(e)

```
model_interaction <- lm(mpg ~ weight * year, data = Auto)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = mpg ~ weight * year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339, Adjusted R-squared:  0.8326
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

p-value < 2.2e-16 远远小于 0.05，说明交互项 weight:year 具有统计显著性。

(f)

```r
model_log <- lm(mpg ~ log(weight) + sqrt(horsepower), data = Auto)
summary(model_log)
```

```
##
## Call:
## lm(formula = mpg ~ log(weight) + sqrt(horsepower), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1029  -2.5380  -0.4015   2.1391  15.6049
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      167.7882     9.6088  17.462  < 2e-16 ***
## log(weight)      -16.5530     1.4473 -11.437  < 2e-16 ***
## sqrt(horsepower)  -1.2514     0.2277  -5.496 7.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.041 on 389 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.732
## F-statistic:   535 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
model_log <- lm(mpg ~ sqrt(weight) + displacement + year, data = Auto)
summary(model_log)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(weight) + displacement + year, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.909 -2.092 -0.128  1.898 14.027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.678073   4.258782   1.568    0.118
## sqrt(weight) -0.804857   0.058644 -13.724   <2e-16 ***
## displacement  0.004840   0.004411   1.097    0.273
## year          0.780624   0.048766  16.008   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.273 on 388 degrees of freedom
## Multiple R-squared:  0.8255, Adjusted R-squared:  0.8242
## F-statistic:   612 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
model_log <- lm(mpg ~ (weight)^2 + displacement + year, data = Auto)
summary(model_log)
```

```
##
## Call:
## lm(formula = mpg ~ (weight)^2 + displacement + year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8400 -2.2917 -0.1177  2.0420 14.3559
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.436e+01  4.021e+00  -3.572 0.000398 ***
## weight       -6.664e-03  5.710e-04 -11.670  < 2e-16 ***
## displacement  2.835e-04  4.744e-03   0.060 0.952382
## year          7.580e-01  5.100e-02  14.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.432 on 388 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8067
## F-statistic: 544.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

p-value 均小于 2.2e-16

13

(a)-(c)

```
set.seed(1)
x <- rnorm(100)
eps <- rnorm(100, sd = sqrt(0.25))
y <- -1 + 0.5 * x + eps
```

y 的长度为 100，$\beta_0 = -1$，$\beta_1 = 0.5$。

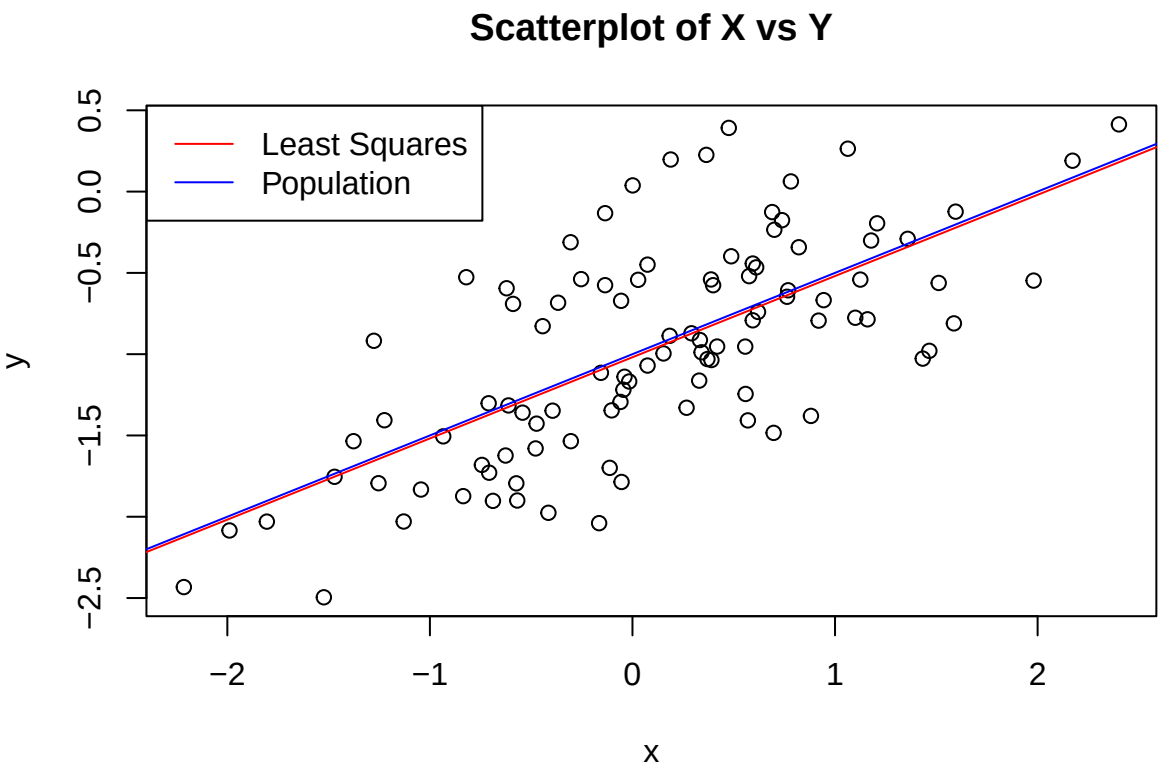
(d)-(f)

```
plot(x, y, main = "Scatterplot of X vs Y")
model_linear <- lm(y ~ x)
summary(model_linear)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
```

```
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

```
abline(model_linear, col = "red")
abline(a = -1, b = 0.5, col = "blue")
legend("topleft", legend = c("Least Squares", "Population"), col = c("red", "blue"), lty = 1)
```



(g)

```
model_quad <- lm(y ~ x + I(x^2))
summary(model_quad)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
```

```
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

不能二次项系数 p 值 = 0.164，说明二次项预测作用不显著，且未提升模型解释力或减少残差，因此未提高拟合度。

(h)

```r
eps_low <- rnorm(100, 0, sqrt(0.1))  # sd = 0.316
y_low <- -1 + 0.5 * x + eps_low
model_low <- lm(y_low ~ x)
summary(model_low)
```

```
##
## Call:
## lm(formula = y_low ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92152 -0.15252 -0.01433  0.20531  0.83534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.99135    0.03311  -29.94   <2e-16 ***
## x            0.50669    0.03678   13.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3287 on 98 degrees of freedom
## Multiple R-squared:  0.6595, Adjusted R-squared:  0.656
## F-statistic: 189.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

R^2 提高，模型解释力提高，残差波动减小。

(i)

```r
eps_high <- rnorm(100, 0, sqrt(0.5))  # sd = 0.707
y_high <- -1 + 0.5 * x + eps_high
model_high <- lm(y_high ~ x)
summary(model_high)
```

```
##
## Call:
## lm(formula = y_high ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7793 -0.3856 -0.0267  0.4758  1.3286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.95922    0.07091 -13.527  < 2e-16 ***
## x            0.46062    0.07876   5.848 6.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7039 on 98 degrees of freedom
## Multiple R-squared:  0.2587, Adjusted R-squared:  0.2512
## F-statistic:  34.2 on 1 and 98 DF,  p-value: 6.553e-08
```

R^2 下降，模型解释力降低，残差波动增大。

(j)

```r
confint(model_linear)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```r
confint(model_low)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.0570515 -0.9256389
## x            0.4337114  0.5796757
```

```r
confint(model_high)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.0999424 -0.8185064
## x            0.3043238  0.6169242
```

噪声水平直接影响置信区间的宽度和模型的解释能力，但不会引入估计偏差。在所有噪声条件下，线性回归均能有效识别 X 与 Y 的显著关系，体现了模型对噪声的鲁棒性。