810198358

امیرمهدی انصاری پور

فصل 5

1)

Flag variable : two-sample Z-test

Multinomial variable: χ^2-test

Continuous variable: two-sample t-test

2) two-sample t-test:

$H_0: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2$

$$t_{data} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.1}{\sqrt{\frac{(5.2)^2}{2000} + \frac{(4.9)^2}{600}}} = \frac{0.1}{\sqrt{0.01352 + 0.04001}}$$

$$\frac{0.1}{\sqrt{0.053}} = 0.43$$

$$\text{P-value} = 2P(t_{599} > 0.43) = 0.66$$

با مقدار بالای p-value (0.1 << 0.66) ، میزان شباهت زیادی

H_0 تقریبا برقرار است و train ، test از یکدیگر مستقل ، در نتیجه partition مورد valid است.

s.a.m

H_0 : $P_{married\ train} = P_{married\ test}$
$P_{single\ train} = P_{single\ test}$
$P_{other\ train} = P_{other\ test}$

H_a : حداقل یک ادعای H_0 درست نیست.

(4

Expected :

	Married	Single	others
Trained	$\frac{1040 \times 2000}{2600}$	$\frac{1000 \times 2000}{2600}$	$\frac{560 \times 2000}{2600}$
Test	$\frac{1040 \times 600}{2600}$	$\frac{1000 \times 600}{2600}$	$\frac{560 \times 600}{2600}$

E :

	M	S	O
Train	800	769	431
Test	240	231	129

O :

	M	S	O
Train	800	750	450
Test	240	250	110

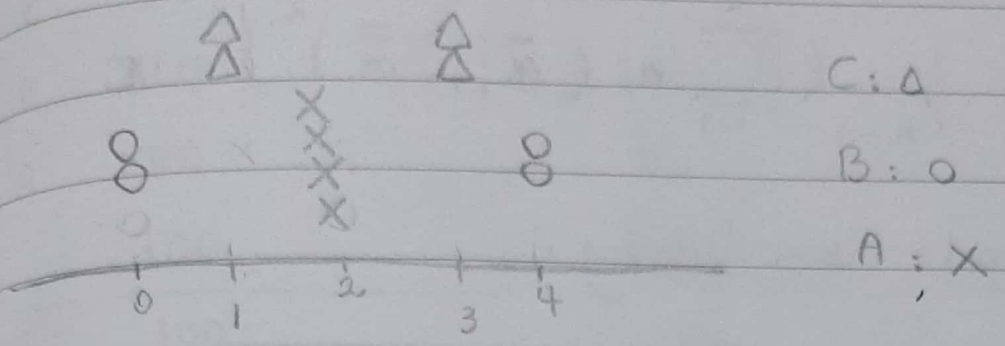
$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 0 + 0 + \frac{1}{769} + \frac{(19)^2}{231} + \frac{(19)^2}{431} + \frac{(19)^2}{129} = 5.2$$

(1.56, 0.84, 2.8)

$$P\text{-value} = P(X^2_2 > 5.2) = 0.07$$

فرض صفر (یا همان استقلال train, test در subsetها) s.a.m
با توجه به مقدار p-value ($0.07 < \alpha = 0.1$) ، تصمیم به [illegible] valid / cs. نیست

6) a) dotplot زیر را رسم کنید



همانطور که دیده میشود، میانگین هر 3 رده با یکدیگر برابر است (2)،

اما هیچ overlap میان گروه‌ها نیست. بنابراین دیاگرام dotplot نشان

فرض H_0 (مستقل بودن 3 نمونه) تأیید وجود دارد. چرا

شود که 3 نمونه مستقل نیستند چون توزیع‌ها همگی دارند. اگرچه میانگین

همگی یکی است، اما واریانس هر subset متفاوت است.

b) تفاوت میانگین گروه‌ها: Between-sample variability

واریانس داخل هر گروه (نسبت

به میانگین آن گروه) Within-sample variability:

Between-sample variability: (c

$$MSTR = \frac{1}{k-1}\left(\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2\right)$$

Within-Sample variability:

$$MSE = \frac{1}{n_t - K}\left(\sum_{i=1}^{k}(n_i - 1)s_i^2\right)$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

d) با توجه آماره مجبور ما زمانیکه MSTR / $F_{data} = \frac{MSTR}{MSE}$

از MSE، چیا، بزرگتر باشد، یعنی تفاوت در میانگین از تفاوت در واریانس

چیا، بیشتر باشد، آنگاه فرض صفر (برابری میانگین‌ها) و استقلال) رد میشود.

H_0 : $\mu_{\text{credit card}} = \mu_{\text{Debit card}} = \mu_{\text{check}}$

H_a : H_0 is wrong

Credit Card	Debit Card	Check
100	80	50
110	120	70
90	90	80
100	110	80
$\bar{x}$ 100	100	70

$\Rightarrow \bar{\bar{x}} = 90$

$$MSTR = \frac{1}{k-1} \sum_i (n_i (\bar{x}_i - \bar{\bar{x}})^2) = 2\,(10^2 + 10^2 + 20^2)$$

$$= \boxed{1200}$$

$$MSE = \frac{1}{n_t - k} \left(\sum_{i=1}^{k} (n_i - 1) \sum_{j=1}^{n_i} \frac{1}{(n_i - 1)} (x_{ij} - \bar{x}_i)^2 \right)$$

$$= \frac{1}{12-3} \left[(3 \times \tfrac{1}{3} \times (0 + 10^2 + 10^2 + 0)) + (3 \times \tfrac{1}{3} \times (20^2 + 20^2 + 10^2 + 10^2)) + (3 \times \tfrac{1}{3} \times (20^2 + 0^2 + 10^2 + 10^2)) \right] =$$

$$\frac{1}{9} [200 + 1000 + 600] = \frac{1}{9} [1800] = \boxed{200}$$

$$\Rightarrow F = \frac{MSTR}{MSE} = 6$$

$$\text{P-value} = P(F_{(2,9)} > 6) = 0.02$$

بدلیل کوچکتر بودن p-value از α (0.05)، H_0 را رد

میکنیم (میانگین ها برابر نیستند).

(10

$\widehat{weight}_i = -180 + 5 \times height_i$

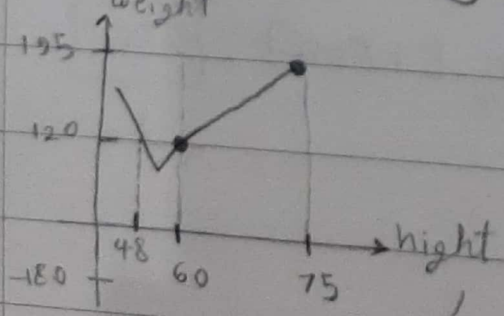
a) $5 \times 3 = 15$ pound

b) $\widehat{weight}_i = -180 + 5(65) = 145$ pound

c) $\widehat{weight}_i = -180 + 5(48) = 60$ pound

در اینجا باید مراقب extrapolation باشیم، زیرا مدل رگرسیون ما

برای height های [60 ,75] اینچ عمل prediction انجام میدهد



در واقع ما از رابطه weight - hight

در خارج این بازه اطلاعی نداریم، ممکن است تابع شکل دیگری

داشته باشد.

s.a.m

d) با هر inch افزایش قد، به متوسط وزن 5 pound افزوده

میشود. (دست مرد $60 \leq height \leq 75$ است)

e) عرض از مبدأ 180- در اینجا معنوی ندارد. (زیرا نمی‌تواند

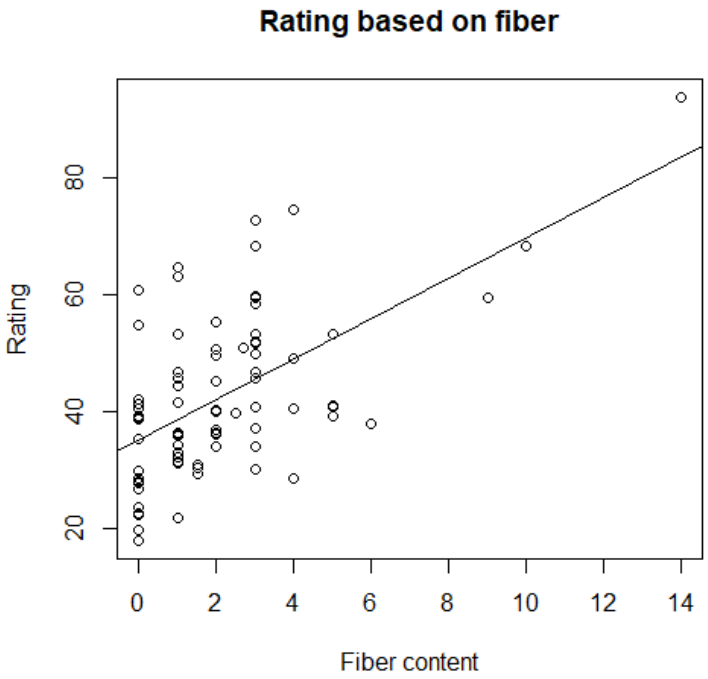
شخصی با $height = 0$, $weight < 0$ داشت.

(11

با اجرای کد زیر داریم: (با توجه به coefficient ها)

```
cereals = read.csv(file.choose(), header = TRUE);
cereals[1:5, ];

fibers = cereals$fiber;
ratings = cereals$rating;
lm.out = lm(ratings~fibers)
plot(
  fibers,
  ratings,
  main = "Rating based on fiber",
  xlab = "Fiber content",
  ylab = "Rating"
)
abline(lm.out)
summary(lm.out)
```



```
Call:
lm(formula = ratings ~ fibers)

Residuals:
    Min      1Q  Median      3Q     Max
-20.436  -8.159  -2.037   6.491  27.216

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.2566     1.7674  19.948  < 2e-16 ***
fibers        3.4430     0.5524   6.233 2.45e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.48 on 75 degrees of freedom
Multiple R-squared:  0.3412,    Adjusted R-squared:  0.3325
F-statistic: 38.85 on 1 and 75 DF,  p-value: 2.445e-08
```

The regression equation is:

rating[i] = 35.2566 + 3.4430 * fibers[i]

(13
مقدار عرض از مبدا (در اینجا 35.2566) در این مثال معنی میدهد. معنای آن این است که اگر یک محصولی اصلا fiber نداشته باشد، rating آن 35.2566% می باشد.

(15
با توجه به مقدار R-squared در summary، که برابر 0.3 است، نمیتوان گفت که مدل ما بخوبی fit شده روی داده train.

(16
با جاگذاری مقدار fiber = 3 در معادله بالا: (point estimation)
rating [3] = 35.2566 + 3.4430 * 3 = 45.5856

مقدار فوق، برآورد نقطه ای میانگین rating یک محصول با 3 gram fiber است.

(17

با اجرای کد زیر داریم:

```
> predict(lm.out, data.frame(fibers = 3), interval = "confidence")
       fit      lwr      upr
1 45.58553 42.81792 48.35314
>
```

بازه اطمینان برای میانگین rating یک محصول با fiber = 3gram برابر [48.35314, 42.81792] می باشد. (با مرکز 45.58553، همان برآورد نقطه ای سوال قبل)

(18

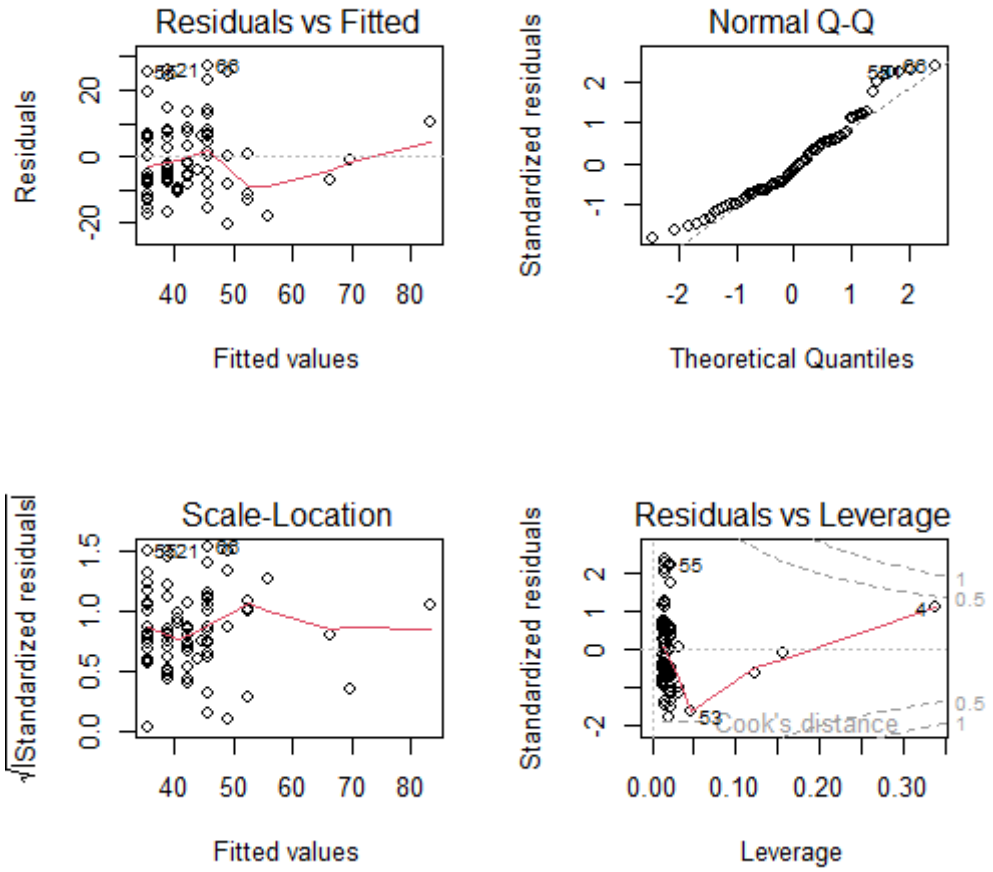
با اجرای کد زیر داریم:

```
> predict(lm.out, data.frame(fibers = 3), interval = "prediction")
       fit      lwr      upr
1 45.58553 22.55513 68.61593
>
```

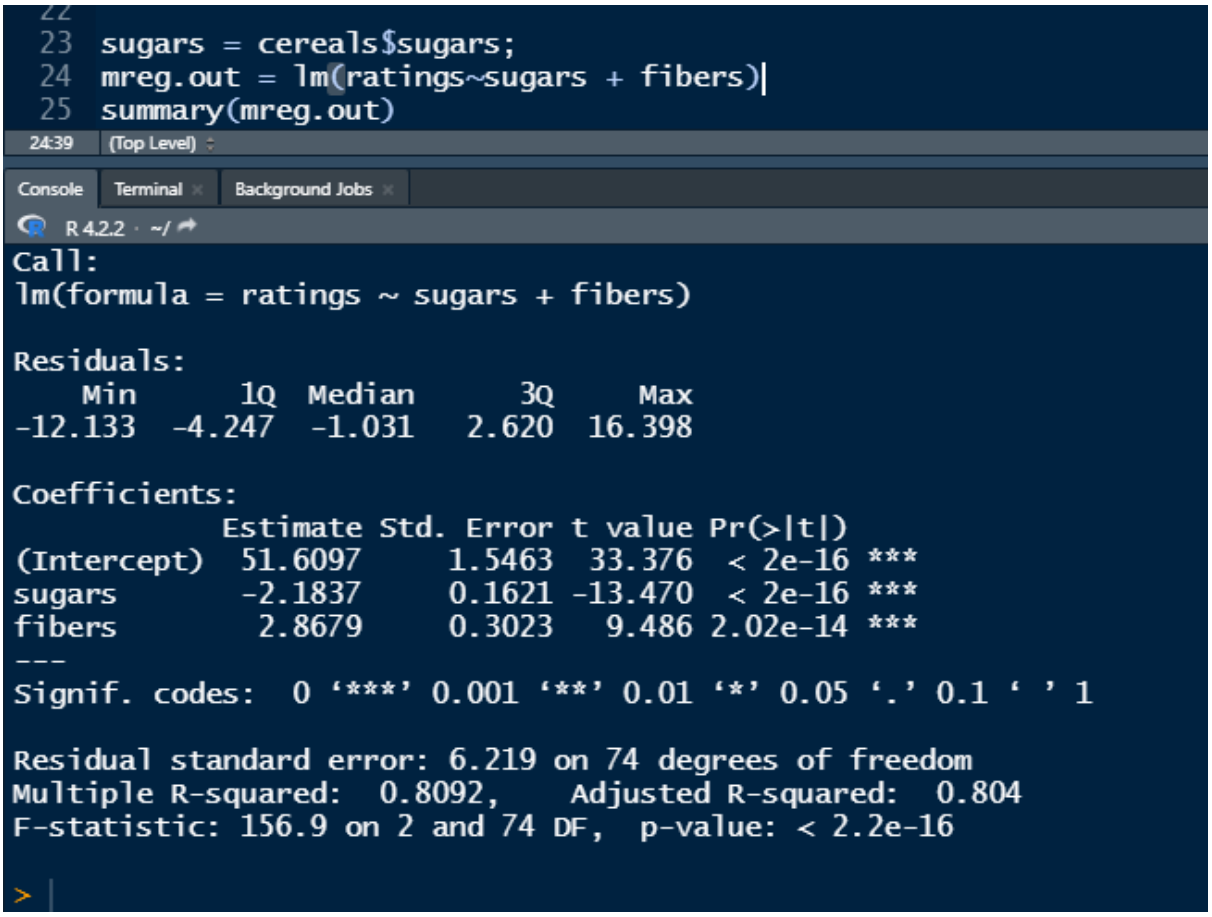
بازه پیش بینی rating یک محصول با fiber = 3gram (دقت شود که در اینجا میانگین برآورد نمیشود و مقدار متغیر هدف یک random record برآورد میشود) برابر [68.61593, 22.55513] می باشد.

(19

با توجه به نمودار های residual بر حسب xi، دیده میشود که متغیر هدف (rating) کمی به راست چوله است (نمودار QQ). اما میتوان از خطی بودن مدل و همگنی واریانس ها اطمینان حاصل کرد (نمودار residuals vs fitted)



(20
با ران کردن کد زیر داریم:



The estimated multiple regression equation is:

rating[i] = 51.6097 + 2.8679 * fibers[i] + -2.1837 * sugars[i]

(21
تفسیر ضریب fiber در معادله سوال قبل:

با ثابت نگه داشتن مقدار sugar، میانگین rating برای یک محصول با هر گرم افزایش fiber، 2.8679 درصد افزایش می یابد.

(22
مقدار r-squared در حالت multiple regression حدود 0.8 می باشد که نسبت به حالت simple regression با مقدار 0.3 بسیار بزرگتر است. بدلیل استفاده از داده بیشتر در مدل، رگرسیون حاصل بهتر توانسته record ها را برآورد کند.