```
x           0.31634    0.22809   1.387   0.1712
z1        -24.24981   31.51425  -0.769   0.4450
z2        -34.10414   34.74505  -0.982   0.3307
x:z1        0.08521    0.31483   0.271   0.7877
x:z2        0.18995    0.35103   0.541   0.5907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.741 on 54 degrees of freedom
Multiple R-squared:  0.4591,    Adjusted R-squared:  0.409
F-statistic: 9.166 on 5 and 54 DF,  p-value: 2.329e-06
```

همان‌گونه که در خروجی بالا می‌توان دید، فرض صفر بودن تمامی ضریب‌های مدل رگرسیون برازش داده شده غیر از عرض از مبدأ در سطح خطای ۰/۰۵ پذیرفته شده است. علت این مسأله وجود همخطی چندگانه‌ی شدید بین ضریب‌های مدل رگرسیون با اثرهای متقابل است.

برای دیدن مقدارهای VIF در این مدل، تابع vif(.) در بسته‌ی نرم‌افزاری car را به صورت زیر به‌کار می‌بریم.

```
> library(car)
>vif(fit)
         x         z1         z2        x:z1        x:z2
  2.869837 139.542183 169.620140 139.148762 166.479597
```

در خروجی بالا می‌بینید، به‌طور تقریبی همه‌ی مقدارهای VIF از ۱۰۰ بزرگ‌تر هستند، که این نشان‌دهنده‌ی همخطی چندگانه‌ی شدید در مدل برازش داده شده است.

۱۱.۴ تمرین‌ها

۱. بر اساس داده‌های $(x_1, y_1), \cdots, (x_n, y_n)$، اطلاعات زیر به‌دست آمده است

$$\sum_{i=1}^{n} y_i = ۱۱۰۰, \sum_{i=1}^{n} x_i y_i = ۶۱۸۰۰, (\mathbf{X}'\mathbf{X})^{-1} = \frac{۱}{۳۴۰۰۰} \begin{bmatrix} ۲۸۴۰۰ & -۵۰۰ \\ -۵۰۰ & ۱۰ \end{bmatrix},$$

که در آن $\mathbf{X}' = \begin{bmatrix} ۱ & \cdots & ۱ \\ x_۱ & \cdots & x_n \end{bmatrix}$ است. مقدار $\hat{\beta}$ را به دست آورید.

۲. مقدارهای $\beta_۰$ و $\beta_۱$ را به دست آورید.

$$\begin{pmatrix} ۱ & ۰ \\ ۰ & ۱ \\ ۲ & ۰ \end{pmatrix} \begin{pmatrix} \beta_۰ \\ \beta_۱ \end{pmatrix} = \begin{pmatrix} ۱۰ \\ ۰ \\ ۵ \end{pmatrix}$$

۳. در مدل رگرسیون خطی چندگانه به صورت زیر

$$y_i = \beta_۰ + \beta_۱ x_{i۱} + \cdots + \beta_{p-۱} x_{i(p-۱)} + \epsilon_i,\ i = ۱, \ldots, n$$

با فرض $\epsilon_i \overset{iid}{\sim} N(۰, \sigma^۲)$، نشان دهید به ازای هر $j \neq j'$، $\hat{\beta}_j$ و $\hat{\beta}_{j'}$ از هم مستقل هستند، اگر و تنها اگر

$$\sum_{i=۱}^{n} x_{ij} x_{ij'} = ۰,\ j \neq j' (= ۱, \ldots, p-۱)$$

۴. در مدل رگرسیونی $\boldsymbol{y} = \mathbf{X}_{n \times p} \beta + \epsilon$ با $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\beta}$، نشان دهید $\mathrm{Var}(\hat{\mathbf{y}}) = \sigma^۲ \mathbf{H}$ و در نتیجه

$$\sum_{i=۱}^{n} \mathrm{Var}(\hat{y}_i) = p\sigma^۲$$

۵. برای مدل رگرسیون خطی چندگانه، با فرض همگنی واریانس خطاها، نشان دهید

$$\mathrm{Var}(\bar{y}) \leq \mathrm{Var}(\hat{y}_i) \leq \mathrm{Var}(y_i),\ i = ۱, \ldots, n$$

۶. میانگین $m \geq ۱$ مقدار مشاهده شده‌ی مستقل از متغیر پاسخ برای یک سطح از متغیر پیشگو (x^*) یعنی $\bar{y}^*_m = \frac{۱}{m} \sum_{j=۱}^{m} y_{jx^*}$ را در نظر بگیرید.

الف. متغیر تصادفی $Z^* = \bar{y}_m^* - \hat{y}_{x^*}$ را در نظر بگیرید. نشان دهید

$$\mathrm{Var}(Z^*) = \sigma^2 \left[\frac{1}{m} + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2} \right]$$

ب. یک بازه‌ی پیش‌بینی $100(1-\alpha)\%$ با دم‌های برابر برای $\bar{y}_m^*$ به دست آورید.

۷. نشان دهید در مدل رگرسیون خطی ساده، درایه‌های ماتریس $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ به صورت زیر هستند.

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}, \; i, j = 1, \ldots, n$$

۸. فرض کنید می‌خواهیم براورد ضریب‌های β در مدل رگرسیونی $y = \mathbf{X}\beta + \epsilon$ را با در نظر گرفتن $\mathbf{R}\beta = \mathbf{r}$ به دست آوریم. با تشکیل تابع لاگرانژ برای حداقل کردن مجموع توان دوم خطا با در نظر گرفتن قید داده شده به صورت زیر

$$\psi(\beta, \lambda) = (y - \mathbf{X}\beta)'(y - \mathbf{X}\beta) - 2\lambda'(\mathbf{R}\beta - r)$$

نشان دهید که براورد کم‌ترین توان دوم مقید β برابر است با

$$\hat{\beta}_{\mathbf{R}} = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(r - \mathbf{R}\hat{\beta})$$

۹. برای مدل زیر هر یک از فرض‌های خطی زیر را به فرم کلی $\mathbf{R}\beta = 0$ با مشخص کردن ماتریس $\mathbf{R}$ بنویسید

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

الف. $H_0 : \beta_1 = \beta_2 = \beta_4$؛

ب. $H_0 : \beta_1 = \beta_2, \; \beta_3 = \beta_4, \; \beta_0 = 0$؛

ج. $H_0 : \beta_1 - 2\beta_5 = 3\beta_4, \; \beta_1 + 2\beta_3 = 0$.

۱۰. داده‌های جدول ۹.۴ حجم بازدمی (y) ۱۹ بیمار مبتلا به آسم را به همراه سن (x_1)، جنسیت (x_2)، قد (x_3) و وزن (x_4) آن‌ها ارائه می‌دهد (بتل و همکاران، ۱۹۸۵).

جدول ۹.۴: داده‌های بیماران مبتلا به آسم (۰ = زن ,۱ = مرد)

ردیف	حجم بازدمی(y)	سن(x_1)	جنسیت(x_2)	قد(x_3)	وزن(x_4)
۱	۴/۷	۲۴	۱	۱۷۵	۷۸/۰
۲	۴/۳	۳۶	۱	۱۷۳	۶۷/۶
۳	۳/۵	۲۸	۰	۱۷۱	۹۸/۰
۴	۴/۰	۲۵	۱	۱۶۶	۶۵/۵
۵	۳/۲	۲۶	۰	۱۶۶	۶۵/۰
۶	۴/۷	۲۲	۱	۱۷۶	۶۵/۵
۷	۴/۳	۲۷	۱	۱۸۵	۸۵/۵
۸	۴/۷	۲۷	۱	۱۷۱	۷۶/۳
۹	۵/۲	۳۶	۱	۱۸۵	۷۹/۰
۱۰	۴/۲	۲۴	۱	۱۸۲	۸۸/۲
۱۱	۳/۵	۲۶	۱	۱۸۰	۷۰/۵
۱۲	۳/۲	۲۹	۱	۱۶۳	۷۵/۰
۱۳	۲/۶	۳۳	۰	۱۸۰	۶۸/۰
۱۴	۲/۰	۳۱	۱	۱۸۰	۶۵/۰
۱۵	۴/۰	۳۰	۱	۱۸۰	۷۰/۴
۱۶	۳/۹	۲۲	۱	۱۶۸	۶۳/۰
۱۷	۳/۰	۲۷	۱	۱۶۸	۹۱/۲
۱۸	۴/۵	۴۶	۱	۱۷۸	۶۷/۵
۱۹	۲/۴	۳۶	۱	۱۷۳	۶۲/۰

الف. یک مدل رگرسیون خطی چندگانه به این داده‌ها برازش دهید.

ب. آزمون فرض معنی‌داری ضریب‌های تک‌تک متغیرهای پیشگو را انجام دهید.

ج. مقدار R^2 و R^2_{adj} را محاسبه کنید و درباره‌ی نیکویی برازش مدل به داده‌ها اظهارنظر کنید.

۱۱. برای داده‌های تمرین ۱۰:

الف. عامل‌های تورم واریانس (VIF) ضریب‌های رگرسیونی را حساب کنید و در مورد همبستگی خطی متغیرهای پیشگوی $x_1, \ldots, x_4$ اظهارنظر کنید.

ب. نمودارهای متغیرهای اضافه شده را رسم کنید و درباره‌ی خطی بودن مدل رگرسیون خطی چندگانه اظهارنظر کنید.

ج. نمودارهای مانده‌ها را رسم کنید و درباره‌ی درستی فرض‌های اساسی اظهارنظر کنید.

۱۲. فرض کنید h_{ii} برابر iامین عضو قطر اصلی ماتریس $\mathbf{H}$ باشد. ثابت کنید در یک رگرسیون خطی چندگانه، $\frac{1}{n} \leq h_{ii} \leq 1$ (از خودتوان بودن ماتریس $\mathbf{H}$ استفاده کنید).

۱۳. می‌دانیم که برای یک بردار تصادفی y (با n مؤلفه) با میانگین μ و ماتریس واریانس ـ کواریانس $\sigma^2\mathbf{A}$، که در آن $\mathbf{A}$ یک ماتریس معین مثبت $n \times n$ است،

$$E\left(\mathbf{y}'\mathbf{A}\mathbf{y}\right) = \sigma^2 tr\left(\mathbf{A}\right) + \mu'\mathbf{A}\mu,$$

الف. نشان دهید

$$E\left(S_e^2\right) = \sigma^2$$

ب. در مدل رگرسیونی $y = \mathbf{x}\beta + \epsilon$ با $\mathrm{Cov}(\epsilon) = \sigma^2\mathbf{I}_n$ نشان دهید

$$E\left(\mathbf{y}'\mathbf{y}\right) = n\sigma^2 + \beta'\mathbf{X}'\mathbf{X}\beta.$$

۱۴. برای پیش‌بینی برخی مشخصه‌های گونه‌های خاص از درختان یک جنگل مانند تعداد و یا متوسط قطر آن درختان می‌توان برخی مشخصه‌های دیگر آن‌ها را به‌کارگرفت. در جدول ۱۰.۴ برخی مشخصه‌های درختان کاج در یک ناحیه‌ی جنگلی مانند سن درختان (Age) ، متوسط ارتفاع درختان شاخص (HD) ، تعداد درختان در یک جریب (N) و متوسط قطر درختان (MDBH) داده شده است (مایرز، ۱۹۹۰). این داده‌ها برای پیش‌بینی MDBH به عنوان متغیر پاسخ به کار گرفته می‌شوند و متغیرهای مستقل در آن عبارتند از $x_1 = \mathrm{HD}$، $x_2 = \mathrm{Age} \times \mathrm{N}$ و $x_3 = \frac{\mathrm{HD}}{\mathrm{N}}$.

الف. مدل رگرسیون خطی چندگانه برازش دهید.

ب. ضریب تعیین و ضریب تعیین تعدیل شده را به‌دست آورده و تفسیر کنید.

ج. جدول تحلیل واریانس را به‌دست آورید و آزمون معنی‌داری مدل را در سطح خطای ۵ درصد انجام دهید.

د. براورد واریانس مدل را به‌دست آورده و به کمک آن واریانس پارامترهای مدل را محاسبه کنید.

ه. آزمون $H_0 : \beta_3 = 0$ در برابر $H_1 : \beta_3 \neq 0$ را انجام دهید.

و. بردار مقدارهای برازش یافته را به دست آورید.

جدول ۱۰.۴: داده‌های درختان کاج

Age	HD	N	MDBH	Age	HD	N	MDBH
۱۹	۵۱/۵	۵۰۰	۷	۱۳	۳۷/۳	۸۰۰	۵/۴
۱۴	۴۱/۳	۹۰۰	۵	۲۱	۵۴/۲	۶۵۰	۶/۴
۱۱	۳۶/۷	۶۵۰	۶/۲	۱۱	۳۲/۵	۵۳۰	۵/۴
۱۳	۳۲/۲	۴۸۰	۵/۲	۱۹	۵۶/۳	۶۸۰	۶/۷
۱۳	۳۹	۵۲۰	۶/۲	۱۷	۵۲/۸	۶۲۰	۶/۷
۱۲	۲۹/۸	۶۱۰	۵/۲	۱۵	۴۷	۹۰۰	۵/۹
۱۸	۵۱/۲	۷۰۰	۶/۲	۱۶	۵۳	۶۲۰	۶/۹
۱۴	۴۶/۸	۷۶۰	۶/۴	۱۶	۵۰/۳	۷۳۰	۶/۹
۲۰	۶۱/۸	۹۳۰	۶/۴	۱۴	۵۰/۵	۶۸۰	۶/۹
۱۷	۵۵/۸	۶۹۰	۶/۴	۲۲	۵۷/۷	۴۸۰	۷/۹

ز. آزمون فرض زیر را در سطح خطای ۵ درصد انجام دهید.

$$H_0: \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad H_1: \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

ح. بردار مانده‌ها را با به‌کارگیری رابطه‌ی (۱۷.۴) به‌دست آورید و سپس براورد واریانس مدل را محاسبه کنید.

۱۵. بر اساس یک مجموعه داده با دو متغیر پیشگوی کمی و متغیر پاسخ متناظر، خلاصه اطلاعات زیر داده شده‌اند.

$$n = 25, \quad \mathbf{y}'\mathbf{y} = 18310/6290, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 559/60 \\ 7375/44 \\ 337072 \end{bmatrix},$$

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} 0/11132151 & -0/0044486 & -0/0000836 \\ & 0/0027437 & -0/0000478 \\ & & 0/0000012 \end{bmatrix}$$

مطلوب است

الف. براورد $\hat{\beta}$ ب. مقدار آماره‌ی آزمون $H_0 : \beta_2 = 0$

ج. مقدار R^2_{adj} د. براورد واریانس خطای مدل

۱۶. فرض کنید در یک مدل رگرسیونی با ۴ متغیر مستقل و $n = 17$ جدول تحلیل واریانس به صورت جدول ۱۱.۴ باشد.

الف. جدول تحلیل واریانس را کامل کنید. ب. ضریب تعیین و S^2_e را به‌دست آورید.

جدول ۱۱.۴: جدول تحلیل واریانس تمرین ۱۶

منبع تغییرات	درجه‌ی آزادی	مجموع توان دوم	میانگین توان دوم	آماره‌ی آزمون
مدل				۸۵۱/۷۲
خطا		۲۶۲/۰۷۲		
کل				

۱۷. در یک فرایند پاکسازی ذغال‌سنگ میزان ذرات معلق بر حسب میلی گرم در لیتر به عنوان متغیر وابسته و مقدار PH فرایند و نوع پلیمر به‌کار رفته برای انجام واکنش به عنوان متغیرهای مستقل در نظر گرفته شده است. داده‌ها در جدول ۱۲.۴ داده شده است (مایرز، ۱۹۹۰). برای این داده‌ها مدل رگرسیونی را به‌دست آورده و تأثیر PH بر متوسط ذرات معلق در هر نوع پلیمر را با هم مقایسه کنید.

جدول ۱۲.۴: داده‌های پاک سازی ذغال سنگ

نوع پلیمر	x_1 (PH)	y (mg/l)	نوع پلیمر	x_1 (PH)	y (mg/l)
p۲	۷/۹	۲۹۷	p۱	۶/۵	۲۹۲
p۲	۸/۷	۳۶۴	p۱	۶/۹	۳۲۹
p۲	۹/۲	۳۷۵	p۱	۷/۸	۳۵۲
p۲	۶/۶	۱۶۷	p۱	۸/۴	۳۷۸
p۲	۷/۰	۲۲۵	p۱	۸/۸	۳۹۲
p۲	۷/۲	۲۴۷	p۱	۹/۲	۴۱۰
p۲	۷/۶	۲۶۸	p۱	۶/۷	۱۹۸
p۲	۸/۷	۲۸۸	p۱	۶/۹	۲۲۷
p۲	۹/۲	۳۴۲	p۱	۷/۵	۲۷۷

۱۸. برای یک مجموعه داده با در نظر گرفتن دو مدل رگرسیونی

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \quad y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

جدول تحلیل واریانس به صورت جدول‌های ۱۳.۴ و ۱۴.۴ به‌دست آمده است. با به‌کارگیری این جدول‌ها، آزمون $H_0 : \beta_3 = \beta_4 = 0$ را در سطح خطای ۵ درصد انجام دهید.

جدول ۱۳.۴: جدول تحلیل واریانس مدل اول تمرین ۱۸

منبع پراکندگی	مجموع توان دوم	درجه‌ی آزادی	میانگین توان دوم	آماره‌ی آزمون	pـ مقدار
رگرسیون	۱۴۲۵۰/۴۶۰	۴	۳۵۶۲/۶۱۵	۳۰/۲۴۳	۰/۰۰۰[a]
خطا	۱۷۴۳۴/۱۰۸	۱۴۸	۱۱۷/۷۹۸		
کل	۳۱۶۸۴/۵۷	۱۵۲			

جدول ۱۴.۴: جدول تحلیل واریانس مدل دوم تمرین ۱۸

منبع تغییرات	مجموع توان دوم	درجه‌ی آزادی	میانگین توان دوم	آماره‌ی آزمون	pـ مقدار
رگرسیون	۱۳۸۶۹/۸۲۵	۲	۶۹۳۴/۹۱۳	۵۸/۳۹۱	۰/۰۰۰[a]
خطا	۱۷۸۱۴/۷۴	۱۵۰	۱۱۸/۷۶۵		
کل	۳۱۶۸۴/۵۷	۱۵۲			

۱۹. بر اساس یک مجموعه داده با $\sum_{i=1}^{۳۲} y_i = ۶۴۲/۹$، خلاصه اطلاعات زیر به‌دست آمده است.

$$\mathbf{y}'\mathbf{y} = ۱۴۰۴۲/۳۱, \quad \mathbf{x}'\mathbf{y} = \begin{bmatrix} ۶۴۲/۹ \\ ۲۳۸۰/۲۷ \\ ۱۹۰۹/۷۳ \end{bmatrix},$$

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} ۵/۷۶۸۰ & -۱/۱۱۱۱ & -۰/۵۴۰۹ \\ & ۰/۲۲۹۱ & ۰/۰۸۹۲ \\ & & ۰/۰۶۸۴ \end{bmatrix}$$

با توجه به اطلاعات داده شده به پرسش‌های زیر پاسخ دهید.

الف. براورد بردار ضریب‌های رگرسیونی را به‌دست آورید.

ب. مقدارهای S_{yy}، SSR و SSE را به‌دست آورده و سپس جدول تحلیل واریانس را تشکیل دهید و به کمک آن آزمون معنی‌داری مدل رگرسیونی را در سطح خطای ۵ درصد انجام دهید.

ج. ضریب تعیین مدل را به‌دست آورید و تفسیر آن را بنویسید.

د. براورد ماتریس واریانس‌ـ کواریانس و براورد بردار ضریب‌های رگرسیونی را به‌دست آورید.

ه. براورد واریانس مقدار برازش یافته‌ی متغیر پاسخ به ازای $\mathbf{X}_t = (۱, ۳/۹۰, ۲/۶۲)$ را به‌دست آورید.

و: یک بازه‌ی اطمینان ۹۵ درصد با دم‌های برابر برای میانگین متغیر پاسخ به ازای $\mathbf{X}_t$ در قسمت "ه" به‌دست آورید.

۲۰. در یک مدل رگرسیون خطی چندگانه به صورت $\mathbf{y} = \mathbf{X}\beta + \epsilon$ با $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'$، نشان دهید

$$R^۲ = ۱ - \frac{y'(\mathrm{I} - \mathrm{H})y}{y'y - n\bar{y}^۲},$$

که در آن n تعداد داده‌ها است.

۲۱. در یک مدل رگرسیون خطی با $p - ۱$ متغیر پیشگو و خطاهای $\epsilon_۱, \ldots, \epsilon_n \overset{iid}{\sim} N(۰, \sigma^۲)$، نشان دهید

$$\frac{(\hat{\beta} - \beta)'(\mathbf{X}'\mathbf{X})(\hat{\beta} - \beta)}{\sigma^۲} \sim \chi^۲_p$$

۲۲. اگر در یک مدل رگرسیون خطی با $p - ۱$ متغیر پیشگو، ضریب تعیین را با $R^۲$ و ضریب تعیین مدل کاهش‌یافته با در نظر گرفتن فرض H را با $R^۲_H$ نمایش دهیم، نشان دهید آماره‌ی آزمون خطی کلی برای آزمون H برابر است با

$$F_* = \frac{(n-p)(R^۲ - R^۲_H)}{۱ - R^۲}$$

۲۳. مدل رگرسیون خطی $y_i = \beta_۱ x_{i۱} + \beta_۲ x_{۲i} + \epsilon_i$، $i = ۱, \ldots, n$، را در نظر بگیرید، که در آن $\epsilon_۱, \ldots, \epsilon_n \overset{iid}{\sim} N(۰, \sigma^۲)$ و $\sigma^۲$ نامعلوم باشد. بر اساس داده‌ها، خلاصه اطلاعات زیر به‌دست

آمده است

$$\sum_{i=1}^{15} x_{i1}^2 = 20,\ \sum_{i=1}^{15} x_{i1}x_{i2} = 5,\ \sum_{i=1}^{15} x_{i2}^2 = 10,\ \sum_{i=1}^{15} x_{i1}y_i = 20,$$

$$\sum_{i=1}^{15} y_i x_{i2} = 10,\ \sum_{i=1}^{15} y_i^2 = 25$$

الف. براورد کم‌ترین توان دوم $\beta_1 + \beta_2$ را بیابید.

ب. براورد کم‌ترین توان دوم $\beta_1 + \beta_2$ را با درنظرگرفتن فرض $\beta_1 = \beta_2 + 0/25$ بیابید.

ج. واریانس $\hat{\beta}_2$ در قسمت "ب" را بیابید و نشان دهید که از واریانس $\tilde{\beta}_2 = \frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2) - \frac{1}{8}$، که در آن $\hat{\beta}_1 + \hat{\beta}_2$ براوردگر کم‌ترین توان دوم $\beta_1 + \beta_2$ در قسمت "الف" است، کم‌تر است.