توجه شود که چنانچه متغیر رسته‌ای بین متغیرهای توضیحی داشته باشیم ، تابع lm برای آن متغیر به تعداد یکی کمتر از تعداد رسته‌هایش ، متغیر نشانگر در نظر می‌گیرد.

```
> auto = read.table("C:/..... /auto.txt", header =T)

> head(auto)

> Summary(lm(Price ~ style + numDoor, auto))
```

سطوح مربوط به

حال اگر سطوحی از متغیرهای نشانگر حضورشان در مدل معنادار نبود ، لازم است آن متغیرها با سطحی که برای آن متغیر نشانگر در نظر گرفته نشده ، ترکیب شود و دوباره مدل برازش داده شود.

در این مثال سطح hardtop معنادار نیست و لازم است با converti که برای آن متغیر نشانگر تعریف نشده یکی شود.

```
> auto$style[which(auto$style == "hardtop")] = "converti"

> summary(lm(Price ~ style + numDoor , auto))
```

البته این مطلب در بحث انتخاب مدل و انتخاب متغیر باید مورد توجه قرار گیرد.

هم‌خطی چندگانه Multicolinearity

چنانچه متغیرهای توضیحی موجود در مدل، همبستگی خطی بالایی داشته باشند، مشکل هم‌خطی چندگانه به وجود می‌آید. در واقع در این حالت واریانس برآورد ضرایب رگرسیونی بزرگ شده، و این امر باعث می‌شود که دقت فاصله اطمینان مرتبط با ضرایب رگرسیونی کم شود. همچنین آماره‌های آزمون T مربوط به آزمون‌های صفر بودن ضرایب رگرسیون جزئی، کوچک‌تر می‌شوند و فرض H_0 به اشتباه رد می‌شود و بنابراین ممکن است متغیرهایی که حضورشان در مدل مهم است به دلیل آنکه با متغیرهای توضیحی دیگر مدل همبستگی دارند، از مدل حذف شوند.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \qquad T = \frac{\hat{\beta}_j - 0}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \qquad \widehat{Var}(\hat{\beta}_j) = MSE\, C_{jj}$$

$$C = (X'X)^{-1}$$

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-p) \sqrt{\widehat{Var}(\hat{\beta}_j)}$$

به منظور بررسی این مسئله می‌توان ضریب همبستگی خطی پیرسون را برای تمام زوج متغیرهای توضیحی موجود در مدل محاسبه کرد. چنانچه این ضریب بزرگ باشد نشان‌دهنده همبستگی خطی بالایی بین این دو متغیر توضیحی وجود دارد.

بررسی ضریب همبستگی خطی پیرسون جهت تشخیص هم‌خطی چندگانه، کافی نیست. ممکن است متغیرهای توضیحی دو به دو همبستگی کمی داشته باشند ولی همبستگی چندگانه زیادی بین آنها برقرار باشد. برای بررسی این مطلب می‌توان عامل تورم واریانس یعنی VIF را محاسبه کرد.

$$VIF_j = \frac{1}{1-R_j^2}$$

R_j^2 ضریب تشخیص مدل رو به رو $\rightarrow$ $X_j = \alpha_0 + \underbrace{\alpha_1 X_1 + \dots + \alpha_k X_k}_{X_j \text{ اینجا قرار ندارد}} + e$

$R_j = \sqrt{R_j^2}$

ضریب همبستگی چندگانه

هم‌خطی یعنی اطلاعات موجود در X_j بر اساس سایر متغیرهای توضیحی قابل بیان است.

VIF های بزرگتر از ۵ یا ۱۰ نشانگر همخطی چندگانه قوی هستند.

```
> plot(Highway1[c(2,6,7,9)])
```
→ نمودار پراکنش دو به دوی متغیرها.

```
> cor(Highway1[c(2,6,7,9)])
```

```
> library(car)
```
تابع های مربوط به متغیرهای توضیحی مدل

```
> vif(mymodel)
```

چنانچه همخطی چندگانه قابل توجهی بین متغیرهای توضیحی وجود داشت. ساده‌ترین راه که البته راه مناسبی نیست، حذف متغیر توضیحی است که بالاترین VIF را دارد و سپس برازش مجدد مدل.

* البته در مدل های چندجمله‌ای یا مدل هایی که اثر متقابل دارند، مرکزی کردن متغیرها می‌تواند به میزان زیاد همخطی چندگانه را کاهش دهد.

مثال:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \longrightarrow$$

وقتی دامنه تغییرات x کم باشد همخطی چندگانه زیادی بین x و x^2 وجود دارد.

$$Y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_2 (x_i - \bar{x})^2 + \epsilon_i$$

→ همخطی چندگانه کاهش می‌یابد.

بررسی فرضیات مدل:

الف- نرمال بودن توزیع خطا:

$$\epsilon_i \sim Normal$$

به منظور بررسی این مسئله لازم است بررسی کنیم که آیا توزیع $\hat{\epsilon}_i$ ها نرمال هست یا نه.

$$\epsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})$$

$$\hat{\epsilon}_i = Y_i - \widehat{E(Y_i)} = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik})$$

از آنجا که $\hat{\epsilon}_i$ها قابل مشاهده هستند، $\hat{\epsilon}_i$ها به عنوان نماینده‌هایی از ϵ_i در نظر گرفته می‌شوند.

توجه شود که این فرض معادل فرض Normal بودن Y ها است ولی در این فرض منظور توزیع Y ها در ترازهای مشخص X یا X های از متغیرهای توضیحی است که معمولاً از هر ترازی تعداد بسیار محدودی مشاهده داریم (معمولاً یک مشاهده) و بنابراین نمی‌توان این فرض را با بررسی توزیع Y ها بررسی کرد.

```
> e = resid(mymodel)
> par(mfrow = c(1,2))
> qqnorm(e)
> qqline(e, col = "blue")
> hist(e, freq = F, main = "Distribution of Residuals")
> xfit <- seq(min(e), max(e), length=40)
> yfit <- dnorm(xfit)
> lines(xfit, yfit, col = "red", lwd = 2)
> shapiro.test(e)
> ks.test(e, "pnorm")
> library(MASS)
> r <- studres(mymodel)
> d <- stdres(mymodel)
```

$$d_i = \frac{e_i}{\sqrt{MSE}}$$

Standardized Residuals

$$\bar{d} = 0 , \quad S_d^2 = \frac{n-p}{n-1}$$

$$r_i = \frac{e_i}{\sqrt{Var(e_i)}} = \frac{e_i}{\sqrt{MSE\left[1-\left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}}$$

Studentized Residuals