به نام خداوند بخشنده مهربان

عنوان اسلایدها:

مروری بر رگرسیون خطی ۱

تهیه‌کننده: محراب عتیقی

دانشگاه اراک-بهمن‌ماه

رگرسیون چیست؟

- رگرسیون یک نوع تحلیل آماری برای بررسی وابستگی یا ارتباط بین متغییرها می‌باشد. و به سوالاتی دررابطه با ارتباط بین متغییرها پاسخ می‌دهد:

۱) آیا بین متغییرها ارتباطی وجود دارد؟

۲) یک متغییرخاص تحت تاثیر کدام متغییر یا متغییرهای دیگر است؟

انواع متغیر ها در تحلیل رگرسیون

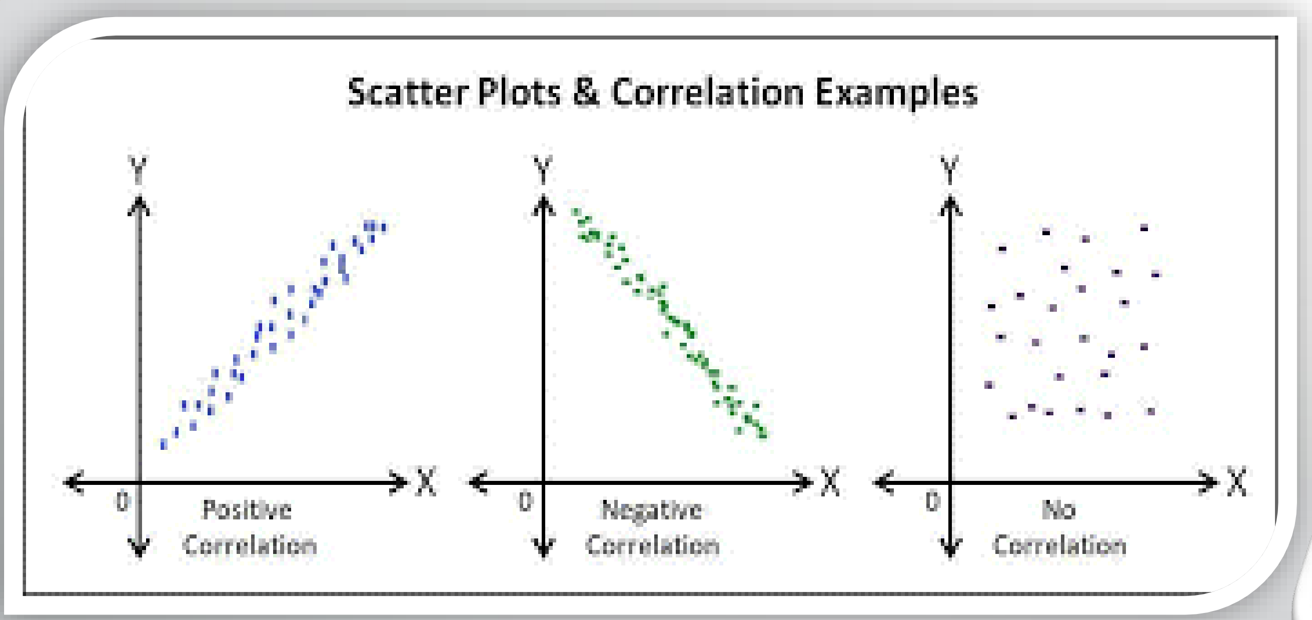
متغییر مستقل(x):

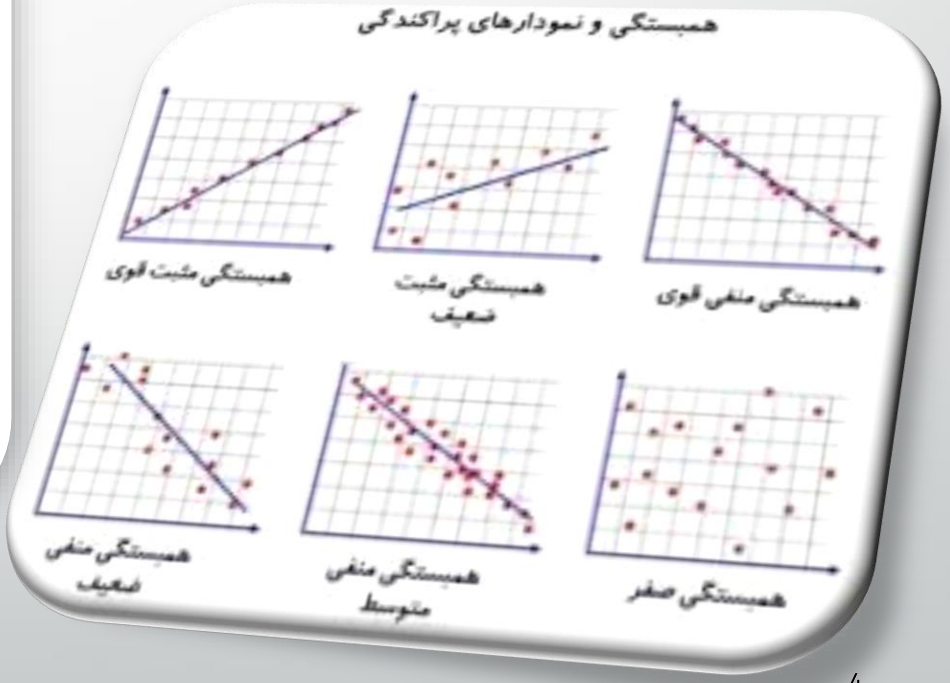
متغییری که قابل تغییراست و بر متغییر وابسته اثرگذار است،که به آن متغییر پیشگو و یا پیش‌بین نیز گفته می‌شود.

متغییر وابسته(y):

متغییری که تحت تاثیر متغیر و یا متغییرهای مستقل است و با تغییر متغییر مستقل نیز تغییر می‌کند.

انواع رابطه متغییر مستقل و متغییر وابسته به کمک نمودار پراکنش:





مدل رگرسیون خطی ساده

مدل رگرسیون خطی ساده مدلی، شامل یک متغییر مستقل و یک متغییر وابسته می‌باشد.

این مدل به صورت زیر نوشته می‌شود:

عرض ازمبدا جامعه

متغییر مستقل (پیشگو)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

متغییر پاسخ
(وابسته)

شیب جامعه

خطای تصادفی

که درآن شیب و عرض از مبدا، پارامترهای ثابت ولی مجهول هستند.

تعاریف:

متغییر: صفتی است که از یک فرد به فرد دیگر تغییر می‌کند.

متغییر کمی: صفتی است که قابل سنجش و اندازه‌گیری است.

متغییر کیفی: صفتی است که قابل سنجش و اندازه گیری نیست.

پارامتر: مقادیری مربوط به جامعه که مجهول اما ثابت هستند.

برآورد: مقادیری بدست‌آمده از نمونه‌تصادفی که معلوم و تصادفی هستند.

خطا چیست؟

خطا یا باقی‌مانده همان فاصله نقاط پیش‌بینی شده تا خط رگرسیونی جامعه می‌باشد.

توزیع متغییر تصادفی خطا

- خطاها متغییرهای تصادفی هستند.
- $\varepsilon \sim N(0,\sigma^2)$
- خطاها ناهمبسته هستند، یعنی مقدار یک خطا به مقدار خطای دیگری وابسته نیست.
- واریانس σ^2، پارامتری نامعلوم است.
- $\sum \varepsilon_i = \sum e_i = 0$

توزیع متغییر تصادفی پاسخ

- متغییر پاسخ y، یک متغییر تصادفی هست.
- پاسخ‌ها نیز ناهمبسته هستند.(چون خطاها ناهمبسته هستند.)
- متغییر پاسخ y، به ازای هرمقدار x یک توزیع احتمال نرمال دارد.(فقط میانگین توزیع متغییر پاسخ ما تغییر می‌کند)
- $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

ضرایب مدل رگرسیون خطی ساده و تفسیر آنها

- **عرض از مبدا ($\boldsymbol{\beta_0}$):** برابر است با میانگین توزیع متغییر پاسخ y، هنگامی که x=0 باشد. به شرط اینکه مقدار صفر در دامنه مقادیر x وجود داشته باشد، درغیر این‌صورت تفسیر ندارد.
- **شیب($\boldsymbol{\beta_1}$):** به ازای یک واحد تغییر در مقدار x، میانگین توزیع متغییر پاسخy ، به میزان $\boldsymbol{\beta_1}$ تغییر می‌کند.

برآورد حداقل مربعات ضرایب رگرسیونی

- برآورد شیب:

$$\widehat{\boldsymbol{\beta_1}} = \frac{\boldsymbol{s_{xy}}}{\boldsymbol{s_{xx}}}$$

که در آن: $s_{xx} = \sum (x_i - \bar{x})^2$, $s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

- برآورد عرض از مبدا:

- $\widehat{\boldsymbol{\beta_0}} = \overline{\boldsymbol{y}} - \widehat{\boldsymbol{\beta_1}}\, \overline{\boldsymbol{x}}$

ویژگی‌های برآوردگرهای حداقل مربعات ضرایب رگرسیون

$\widehat{\boldsymbol{\beta_1}}$, $\widehat{\boldsymbol{\beta_2}}$ ناااریب هستند و دارای کمترین واریانس هستند.

ولذا طبق قضیه گوس-مارکوف، این برآوردگرهای حداقل مربعات برای مدل ما بهترین برازش را دارند.

برآورد واریانس:

$$\widehat{\sigma^2} = \frac{SSE}{n-2} = MSE$$

که در آن $SSE = \sum e_i^2$

ویژگی‌های برآوردگر واریانس:

- میانگین مربعات خطا، MSE برآوردگری نااریب از σ^2 است.
- برآوردگر σ^2 ، وابسته به مدل است.

توزیع برآوردگرهای ضرایب خط رگرسیونی

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right),$$

$$\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right)$$

آزمون فرض درباره عرض ازمبداء خط رگرسیونی:

$$\begin{cases} H_0 : \beta_0 = \beta_{00} \\ H_1 : \beta_0 \neq \beta_{00} \end{cases} \quad , t_0 = \frac{\widehat{\beta_0} - \beta_{00}}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} \quad , |t_0| > t_{\left(n-2, 1-\frac{\alpha}{2}\right)}$$

آزمون فرض درباره شیب خط رگرسیونی:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases} \quad , t_0 = \frac{\widehat{\beta_1} - 0}{\sqrt{MSE/s_{xx}}} \qquad , |t_0| > t_{(n-2, 1-\frac{\alpha}{2})}$$

$$, Z_0 = \frac{\widehat{\beta_1} - 0}{\sqrt{\sigma^2/s_{xx}}}$$

مجموع مربعات کل:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad ,df = n - 1$$

زیر یکدرجه آزادی مربوط، روی انحرافات ازدست رفته است.

مجموع مربعات رگرسیون:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad ,df = 1$$

مجموع مربعات خطا:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad ,df = n - 2$$

همیشه داریم:

$$SST = SSR + SSE$$

آزمون آنالیز واریانس درباره شیب خط رگرسیونی:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases} ,F_0 = \frac{MSR}{MSE} \qquad ,F_0 > F_{(1-\alpha,1,n-1)}$$

$$,MSR = \frac{SSR}{df_{MSR}} ,MSE = \frac{SSE}{df_{MSE}}$$

فواصل اطمینان برای شیب و عرض از مبداء و واریانس:

- فاصله اطمینان شیب خط رگرسیون:

$$\frac{\widehat{\beta_1}-\beta_1}{\sqrt{MSE/s_{xx}}}\sim t_{(n-2)} \rightarrow \left(\widehat{\beta_1}-t_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{\frac{MSE}{S_{xx}}}<\beta_1<\widehat{\beta_1}+t_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{\frac{MSE}{S_{xx}}}\right)$$

- فاصله اطمینان عرض از مبداء خط رگرسیون:

$$\frac{\widehat{\beta_0}-\beta_0}{\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}}\sim t_{(n-2)} \rightarrow \left(\widehat{\beta_0}-t_{\left(n-2,1-\frac{\alpha}{2}\right)}\frac{\widehat{\beta_0}-\beta_0}{\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}}<\beta_0<\widehat{\beta_0}+t_{\left(n-2,1-\frac{\alpha}{2}\right)}\frac{\widehat{\beta_0}-\beta_0}{\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}}\right)$$

- فاصله اطمینان برای واریانس:

$$\frac{(n-2)MSE}{\sigma^2}\sim X^2_{(n-2)} \rightarrow \left(\frac{(n-2)MSE}{X^2_{\left(\frac{\alpha}{2},n-2\right)}}<\sigma^2<\frac{(n-2)MSE}{X^2_{\left(1-\frac{\alpha}{2},n-2\right)}}\right)$$

فواصل اطمینان برای میانگین پاسخ به ازای x_0 و پیش‌بینی m مشاده جدید y_0.

- فاصله اطمینان برای میانگین پاسخ به ازای x_0:

$$\widehat{y_0} \sim N\left(E[y|x_0], \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)\right)$$

$$\left(\widehat{y_0}-\mathrm{t}_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)} < E[y|x_0] < \widehat{y_0}+\mathrm{t}_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right)$$

- فاصله اطمینان برای پیش‌بینی m مشاهده جدید y_0:

$$\left(\widehat{y_0}-\mathrm{t}_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{MSE\left(\frac{1}{m}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)} < y_0 < \widehat{y_0}+\mathrm{t}_{\left(n-2,1-\frac{\alpha}{2}\right)}\sqrt{MSE\left(\frac{1}{m}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right)$$

ضریب تعیین:

- نسبت تغییرات تبیین شده متغییر پاسخ به وسیله متغییر رگرسیونی x

$$R^2 = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}} \qquad 0 \leq R^2 \leq 1$$

- نکته: اگر ρ، ضریب همبستگی بین دومتغییر پاسخ و پیشگو ما باشد، آنگاه:

$$R^2 = \rho^2$$

معیارهای مناسبت مدل:

- باقی‌مانده‌های ما پرت و دورافتاده نباشند.
- توزیع باقی مانده ها نرمال باشد.
- رسم نمودارهای باقی‌مانده استیودنت شده و استاندارد.
- رسم انواع نمودارهای باقی‌مانده ها دربرابر متغییرهای پیشگو و پاسخ مدل.
- بررسی فرض ثبات واریانس خطاها.
- آزمون فقدان برازش خطا.

آزمون فرض بررسی فقدان برازش خط

- این آزمون تنها زمانی قابل اجراست که:

به ازای حداقل یک سطح x، مشاهدات تکراری برای y وجود داشته باشد.

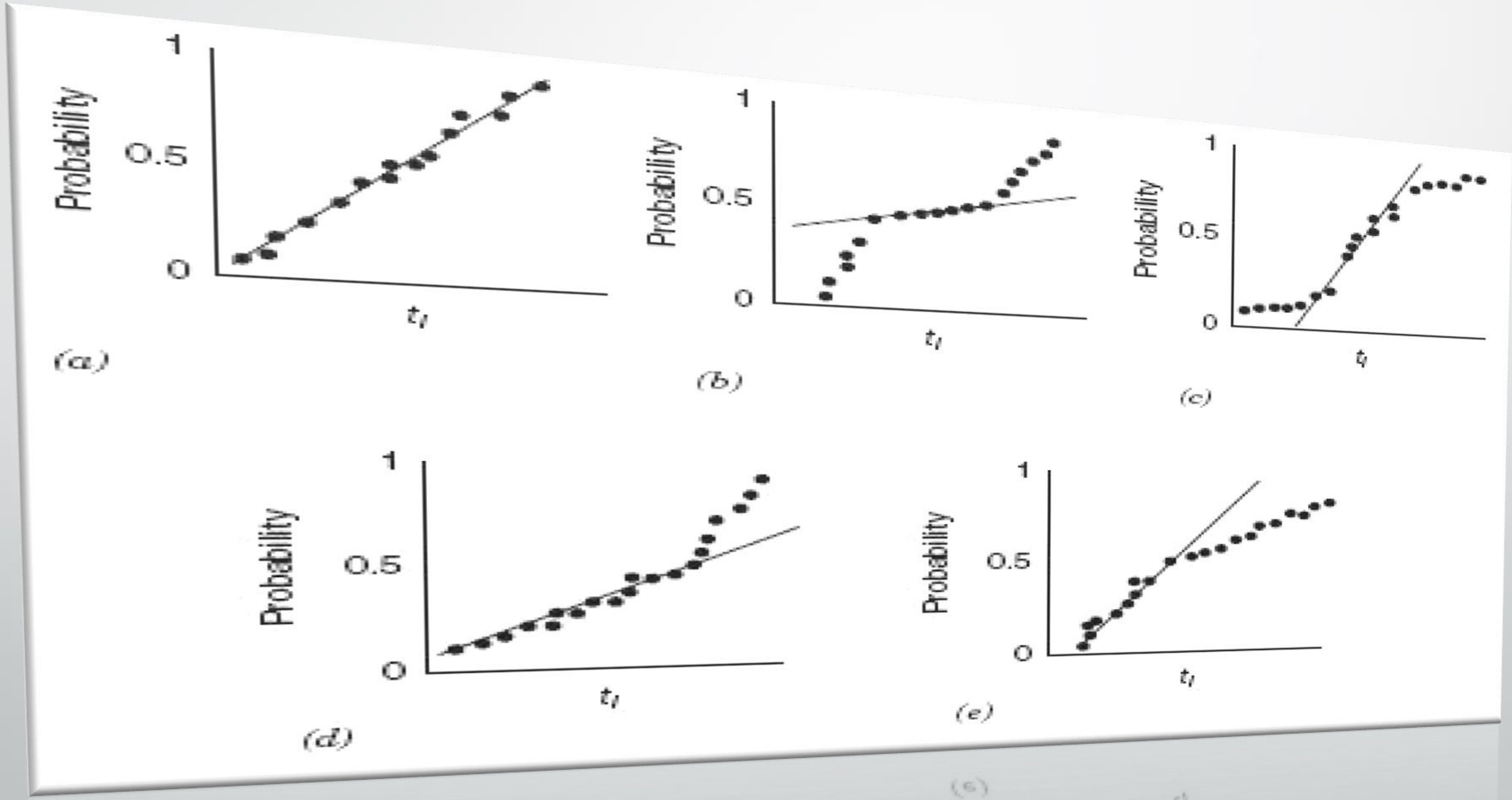
$$SSE = SS_{PE} + SS_{LOF}$$

$$(y_{ij} - \hat{y}_i) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

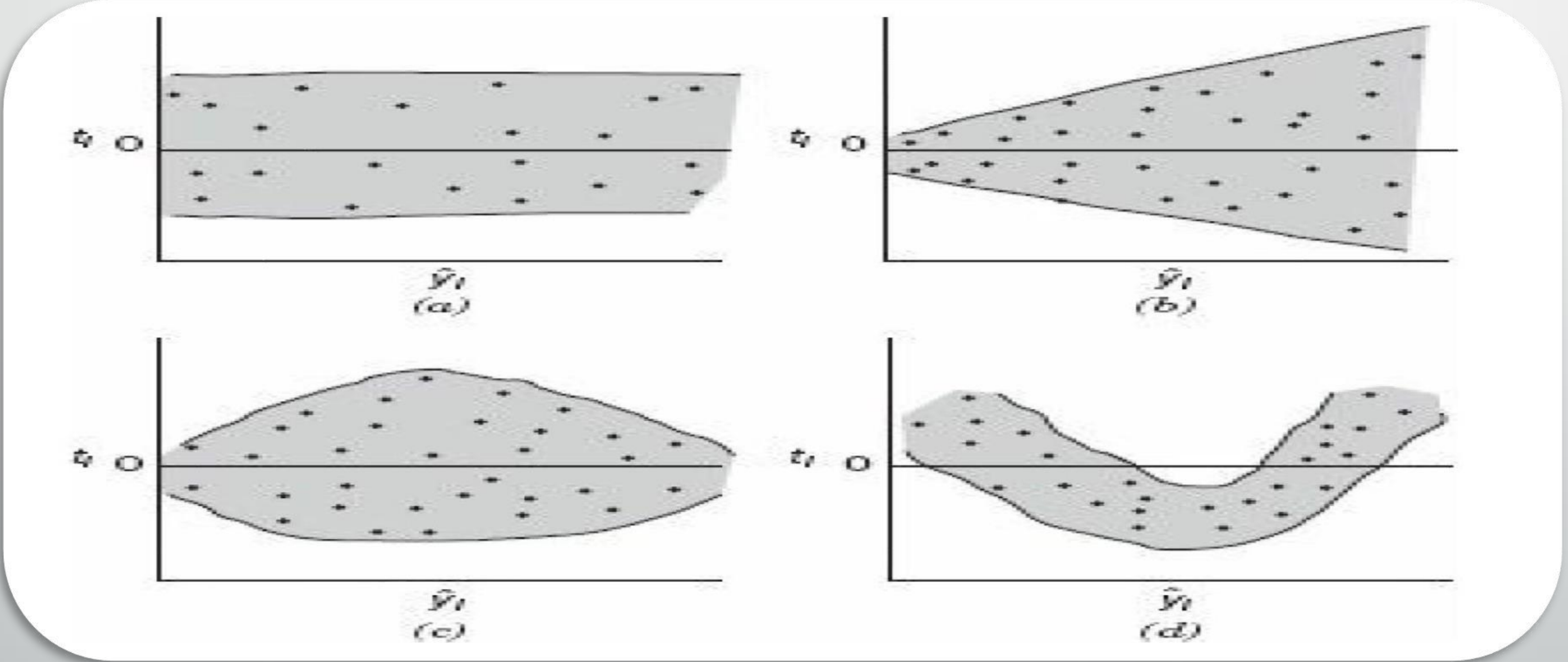
$$\sum_{i=1}^{m}\sum_{j=1}^{h_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}, F_0 = \frac{\frac{SS_{LOF}}{m-2}}{\frac{SS_{PE}}{n-m}} = \frac{MS_{LOF}}{MS_{PE}} \qquad , F_0 > F_{(\alpha,(m-2),(n-m))}$$

بررسی فرض نرمال بودن خطاها



بررسی ثبات واریانس خطاها



انواع باقی‌مانده ها:

- باقی‌مانده‌ها:

$$e_i = y_i - \widehat{y}_i \quad , e_i \sim N(0, \sigma^2)$$

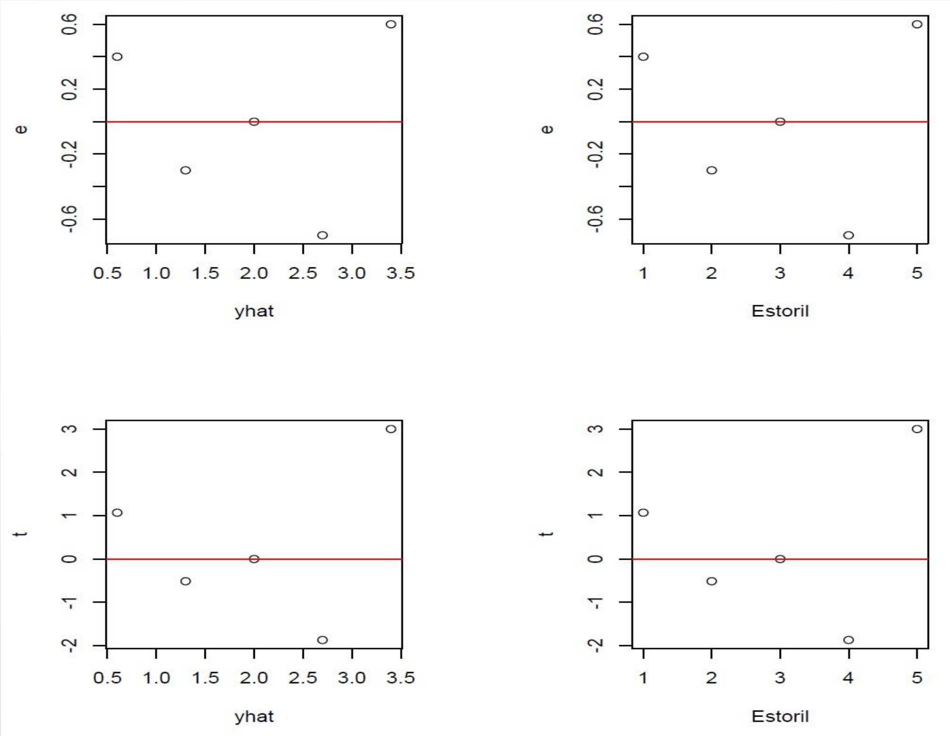
- باقی‌مانده‌های استاندارد شده:

$$d_i = \frac{e_i - 0}{\sqrt{(MSE)}} \quad , d_i \sim N(0,1)$$

- باقی‌مانده‌های استیودنت شده:

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

بررسی داده‌های پرت و مشاهدات دورافتاده



کدهای R برای مسایل مطرح شده:

- برای رسم نمودار پراکنش:

```
Plot(x,y)
```

- برای ساختن یک مدل رگرسیونی یگانه داریم:

```
Fit<-lm(y~x)
```

- برای رسم خط رگرسیونی داریم:

```
plot(x,y);abline(fiit,col="red")
```

- برای بدست آوردن شیب خط و عرض از مبداء داریم:

```
Coefficients(fit) or coef(fit)
```

- برای بدست آوردن مقدار و پیشبینی به صورت فاصله اطمینان داریم:

```
New<-data.frame(x=c(….));predict(fit,newdata=New , interval="prediction",level=0.95)
```

- برای بدست آوردن متوسط y درنقطه جدید یا قدیمی‌ها بصورت فاصله اطمینان داریم:

```
Predict(fit,newdata=New ,interval="confidence" , level=0.95)
```

- برای رسم نمودار یک برآورد فاصله‌ای یا متوسط آنها و مقادیر یشبینی شده بصورت یکجا داریم:

```
matplot(x,cbind(a,b),type="L")
```

- برای دیدن نتیجه آزمون وجود شیب خط رگرسیون داریم:

```
Summary(fit)    and   anova(fit)
```

در دستور `summary` می‌توانیم به ضریب تعیین، SSها و درجه های آزادی و... پی ببریم .

- برای بدست آوردن فاصله اطمینان برای شیب و عرض از مبداء داریم:

```
Confit(fit,level=0.95)
```

- برای بدست آوردن ضریب همبستگی داریم:

```
Cor.test(x,y,method="pearson")
```

- برای مشاهده باقی‌مانده‌های اصلی،استاندارد و استیودنت شده داریم:

```
e<-residuals(fit);d<-rstandard(fit);t<-rstudent(fit)
```

- برای مشاهده مقادیر $\hat{y}$، داریم:

```
Yhat<-fitted(fit)
```

- برای رسم نمودار احتمال نرمال داریم:

```
qqnorm(e)
```

- برای بررسی آزمون فقدان برازش خط داریم:

```
fit1<-lm(y~x);fit2<-aov(y~factor(x));anova(fit1,fit2)
```

در ادامه به بررسی مثالی از مباحث مطرح شده می‌پردازیم:

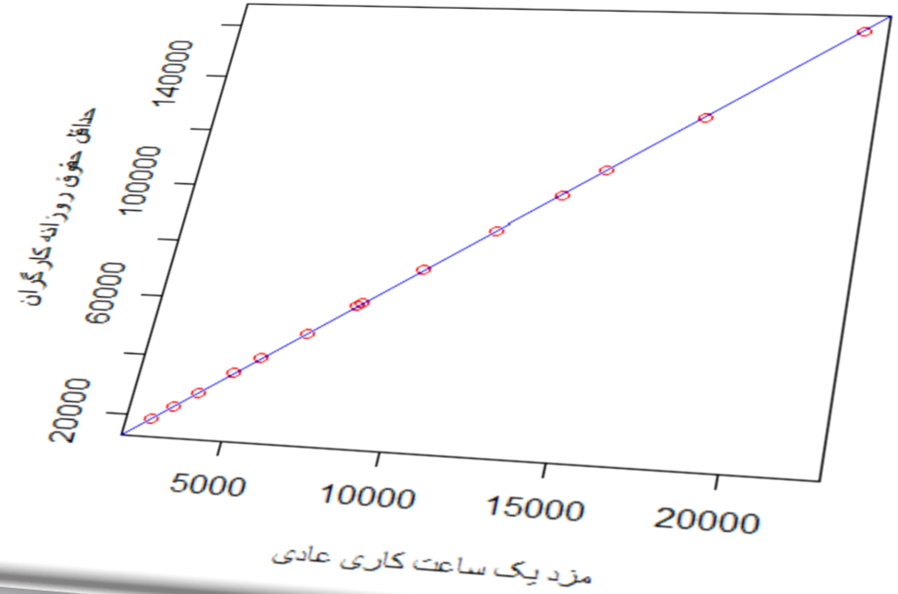
در این قسمت می‌خواهیم به بررسی مزد یک ساعت کاری عادی کارگران مشمول قانون کار، بروی حداقل حقوق آنها بپردازیم:

- ابتدا داده‌های خود را به شکل زیر فراخوانی می‌کنیم:

```
> data
   min_hoghogh mozd_per_h
1       162370      22152
2       129900      17720
3       110100      15020
4       101000      13777
5        87840      11984
6        73200       9986
7        61000       8322
8        60000       8186
9        50000       6821
10       40864       5575
11       35534       4848
12       28446       3881
13       23282       3175
14       18930       2583
```

حال برای رسم نمودار پراکنش و خطرگرسیونی بین این دو متغییر مستقل و پیشگو داریم:

```
>fit1=lm(min_hoghogh~mozd_per_h)
> plot(mozd_per_h,min_hoghogh,xlab="مزد یک ساعت کاری عادی ",ylab="حداقل حقوق روزانه کارگران",col="REd")
> abline(fit1,col="Blue")
```



برای بدست آوردن مقادیر شیب خط و عرض از مبداء آن داریم:

```
> coef(fit1)
 (Intercept) mozd_per_h
-0.4451111   7.3302412
```

برای بدست آوردن خلاصه‌ای از مدل یگانه خود داریم:

```
>summary(fit1)

 Call:
lm(formula = min_hoghogh ~ mozd_per_h)
Residuals: Min 1Q Median 3Q Max
-9.0585 -3.3171 -1.7360 0.8163 11.7117
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4451111 3.2300262 -0.138 0.893
 mozd_per_h 7.3302412 0.0002903 25246.590 <2e-16 ***
--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 6.156 on 12 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
 F-statistic: 6.374e+08 on 1 and 12 DF, p-value: < 2.2e-16
```

برای بدست آوردن یک فاصله اطمینان ۹۵٪ برای پیش‌بینی و متوسط مقدار متغییر وابسته خود داریم:

```
>predict.lm(fit1,interval="confidence")
     fit         lwr         upr
1    162379.06   162370.33   162387.79
2    129891.43   129885.15   129897.71
3    110099.78   110094.81   110104.75
4    100988.29   100983.83   100992.75
5    87845.17    87841.27    87849.06
6    73199.34    73195.75    73202.94
7    61001.82    60998.15    61005.49
8    60004.91    60001.22    60008.60
9    49999.13    49995.15    50003.12
10   40865.65    40861.26    40870.04
11   35536.56    35531.90    35541.23
12   28448.22    28443.14    28453.30
13   23273.07    23267.66    23278.48
14   18933.57    18927.88    18939.26
```

```
>predict.lm(fit1,interval = "prediction")
     fit       lwr       upr
1  162379.06 162363.06 162395.06
2  129891.43 129876.62 129906.24
3  110099.78 110085.47 110114.08
4  100988.29 100974.15 101002.42
5  87845.17  87831.20  87859.13
6  73199.34  73185.46  73213.23
7  61001.82  60987.92  61015.73
8  60004.91  59991.00  60018.82
9  49999.13  49985.14  50013.12
10 40865.65  40851.54  40879.76
11 35536.56  35522.36  35550.77
12 28448.22  28433.88  28462.56
13 23273.07  23258.61  23287.53
14 18933.57  18919.00  18948.14
```

برای بدست آوردن یک فاصله اطمینان ۹۵٪ برای پیش‌بینی و متوسط مقدار متغییر وابسته خود درنقاطی جدید داریم:

```
> New<-data.frame(c(22000,19000,18000,16000,18888,17980,13900,16578,15890,21000,17800,18000,15600,12780))
```

```
>predict.lm(fit1,interval="confidence",newdata = New)
    fit       lwr       upr
1  162379.06 162370.33 162387.79
2  129891.43 129885.15 129897.71
3  110099.78 110094.81 110104.75
4  100988.29 100983.83 100992.75
5  87845.17  87841.27  87849.06
6  73199.34  73195.75  73202.94
7  61001.82  60998.15  61005.49
8  60004.91  60001.22  60008.60
9  49999.13  49995.15  50003.12
10 40865.65  40861.26  40870.04
11 35536.56  35531.90  35541.23
12 28448.22  28443.14  28453.30
13 23273.07  23267.66  23278.48
14 18933.57  18927.88  18939.26
```

```
>predict.lm(fit1,interval = "prediction",newdata = New)
 fit lwr upr
 1 162379.06 162363.06 162395.06
 2 129891.43 129876.62 129906.24
 3 110099.78 110085.47 110114.08
 4 100988.29 100974.15 101002.42
 5 87845.17 87831.20 87859.13
 6 73199.34 73185.46 73213.23
 7 61001.82 60987.92 61015.73
 8 60004.91 59991.00 60018.82
 9 49999.13 49985.14 50013.12
 10 40865.65 40851.54 40879.76
 11 35536.56 35522.36 35550.77
 12 28448.22 28433.88 28462.56
 13 23273.07 23258.61 23287.53
 14 18933.57 18919.00 18948.14
```

برای بدست آوردن ضریب همبستگی متغییر مستقل و متغییر پاسخ خود داریم:

```
>cor.test(min_hoghogh,mozd_per_h,method = "pearson")
                Pearson's product-moment correlation
data: min_hoghogh and mozd_per_h t = 25247, df = 12, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 1 1 sample estimates: cor 1
```

برای انجام آزمون آنوا یا آنالیز واریانس داریم:

```
>anova(fit1)

Analysis of Variance Table Response:
 min_hoghogh Df Sum Sq     Mean Sq    F value Pr(>F)
 mozd_per_h   1  2.4153e+10 2.4153e+10 637390311 < 2.2e-16 ***
 Residuals   12  4.5500e+02 3.8000e+01
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

برای انجام آزمون فقدان برازش داریم:

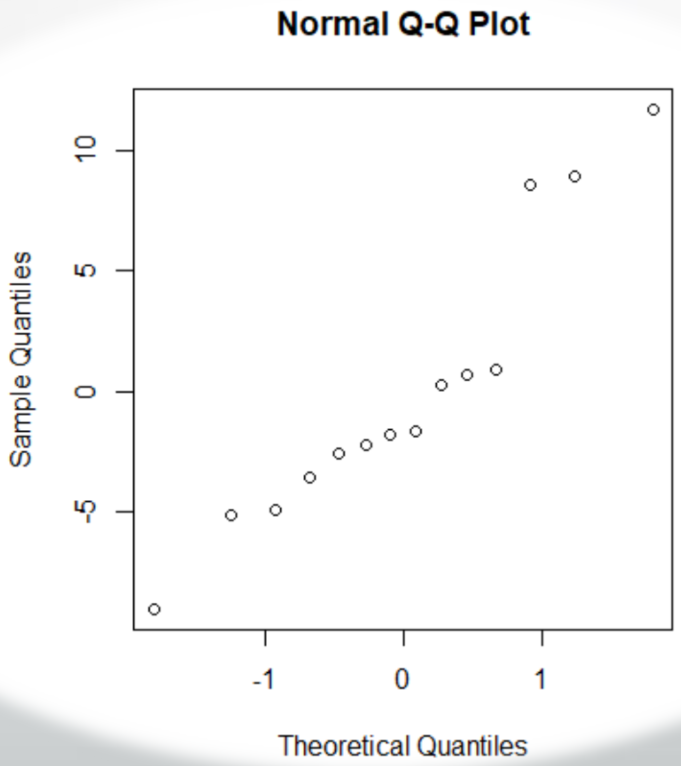
```
>model.reg<-lm(min_hoghogh~mozd_per_h)
>model.aov<-aov(min_hoghogh~factor(mozd_per_h))
```

```
anova(model.reg,model.aov)
 Analysis of Variance Table Model
1: min_hoghogh ~ mozd_per_h Model
2: min_hoghogh ~ factor(mozd_per_h)
   Res.Df RSS     Df    Sum of Sq F Pr(>F)
1  12     454.71
2  0      0.00    12    454.71         0
```

```
> #the pure error value is:
> anova(model.reg,model.aov)[2,2]
 [1] 0
> #the pure error df is :
> anova(model.reg,model.aov)[2,1]
 [1] 0
> #the lack of fit value is:
> anova(model.reg,model.aov)[2,4]
 [1] 454.7149
> #the lack of fit df is:
> anova(model.reg,model.aov)[2,3]
 [1] 12
```

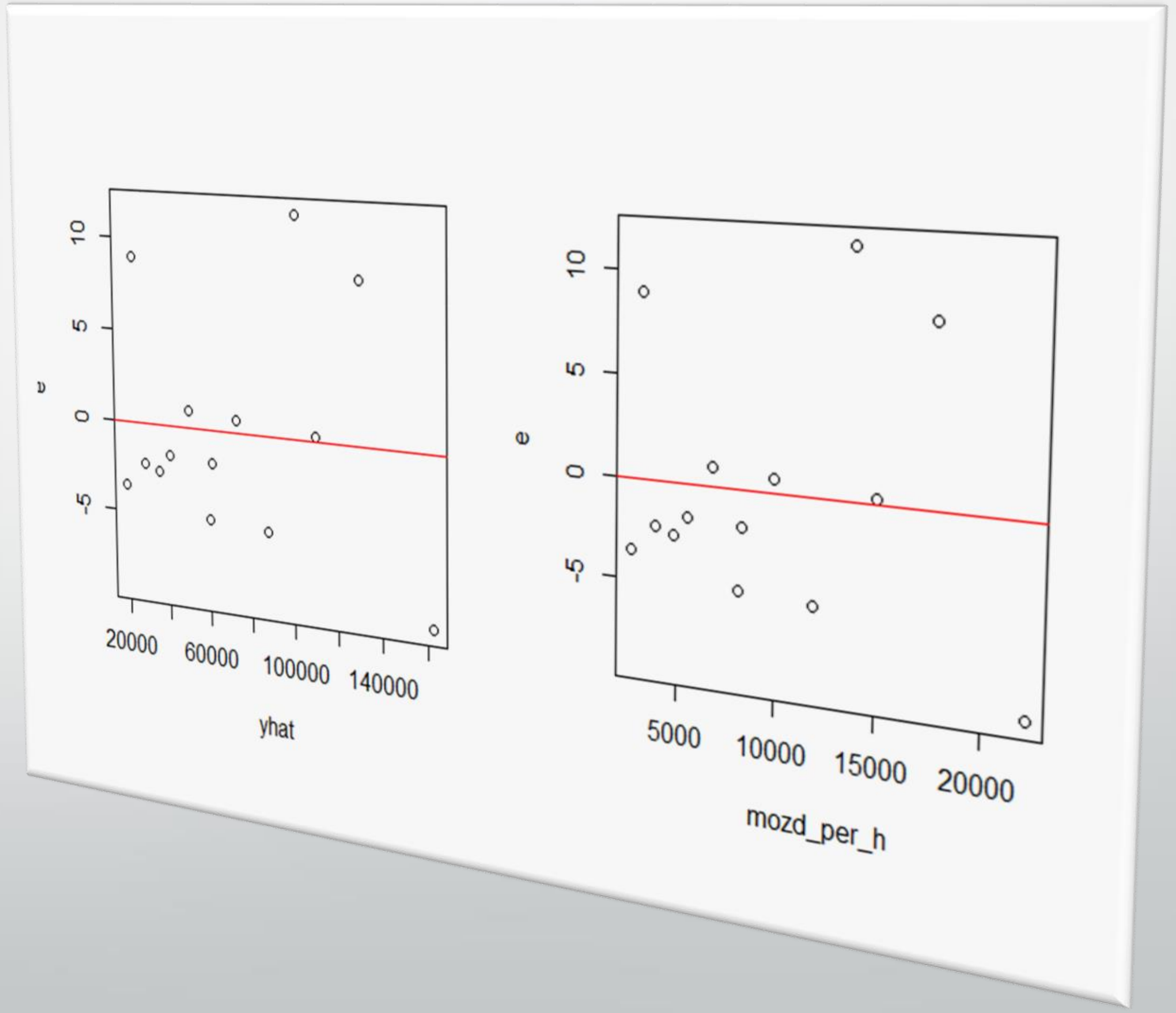
برای تعریف انواع باقی‌مانده‌ها و رسم نمودار نرمال احتمال باقی‌مانده‌ها داریم:

```
yhat<-fitted(fit1)
> s<-rstandard(fit1)
> t<-rstudent(fit1)
> e<-residuals(fit1)
> qqnorm(e)
```



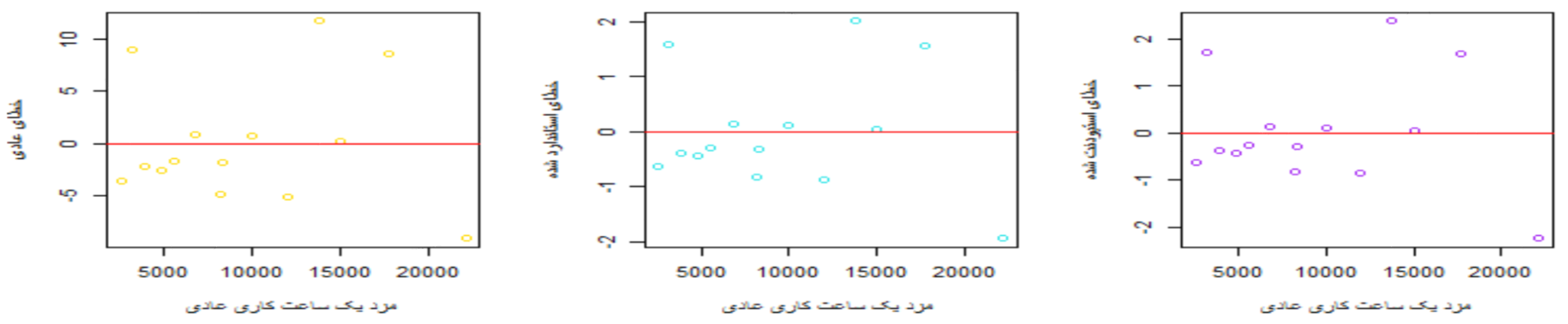
برای رسم نمودار، باقی‌مانده‌ها در برابر متغییر مستقب و متغییر پیشگو داریم:

```
> par(mfrow=c(1,2),pty="s")
> plot(yhat,e)
> abline(h=0,col="red")
> plot(mozd_per_h,e)
> abline(h=0,col="red")
```



برای رسم نمودارهای انواع باقی‌مانده‌ها داریم:

```
> par(mfrow=c(1,3),pty="s")
> plot(mozd_per_h,e,xlab="مزد یک ساعت کاری عادی ",ylab="خطای عادی",col="Gold")
> abline(h=0,col="red")
> plot(mozd_per_h,s,xlab="مزد یک ساعت کاری عادی ",ylab="خطای استاندارد شده",col=85)
> abline(h=0,col="red")
> plot(mozd_per_h,t,xlab="مزد یک ساعت کاری عادی ",ylab="خطای استیودنت شده",col="purple")
> abline(h=0,col="red")
```



با تشکر از همراهی شما عزیزان