МАТФ
Универзитет у Београду
Математички факултет

Биостатистика и анализа података

Марија Цупарић

Десети час

Регресиони модели

- ⋆ Проучавамо како се може моделовати зависност једне променљиве у односу на другу.
- ⋆ Регресионо моделирање се бави проналажењем адекватног модела којим се описује утицај такозваних независних променљивих (предиктора) X на зависну променљиву (променљиву одговора) Y.
- ⋆ Регресиони модел: $Y = f(X) + \varepsilon$, где је $f(X)$ регресиона функција, а ε случајна променљива која не зависи од X (и најчешће има $\mathcal{N}(0, \sigma^2)$ расподелу).
- ⋆ Вредности независне променљиве x (тачке) често можемо сами да бирамо и онда узимамо узорак за Y у тим тачкама.
- ⋆ Уколико сами бирамо x, то је планирани експеримент, а ако не, онда имамо посматрани експеримент.

- ⋆ Најчешћа сврха регресионог проучавања је предвиђање.
- ⋆ Одређујемо једначину уз помоћ које ће вредност од Y бити превиђена довољно добро на основу других променљивих које се јављају у једначини.

Пример

(1) Испитујемо раст биљака (Y) у односу на количину светлости (x)

(2) Предвиђање масе јединки (Y) на основу више фактора (x_1 количина хране, x_2 температура средине, . . .)

(3) Испитујемо везу између концентрације загађујуће супстанце (x) и броја угинулих јединки (Y)

Сврха: разумевање утицаја фактора средине, издвајање значајних фактора и поређење различитих услова или третмана.

Линеарна регресија

Наш задатак је да на основу посматраних података нађемо линеарну једначину (график је права линија) која изражава средњу вредност од Y преко x.

Ако знамо да независна променљива има конкретну вредност x, који је најбољи избор за предвиђање вредности Y?

Пример Ако је концентрација загађујуће супстанце у води 50 $\mathrm{mg/l}$, колики је очекивани број угинулих јединки?
Дакле, тражимо $Y|x = 50$ и то је случајна величина јер ће, чак и при истој концентрацији загађења, број угинулих јединки варирати између различитих узорака или популација. Логично предвиђање за Y када је $x = 50$ је средња вредност Y-а за конкретну вредност x.

Дефиниција: Нека је x нека променљива и нека је Y случајна величина. Регресиона крива Y на x је график функције средње вредности Y за различите вредности x, тј. график функције $\mu_{Y|x}$. За регресиону криву Y на x каже се да је линеарна ако је $\mu_{Y|x} = \alpha + \beta x$ за неке реалне бројеве α и β. Број β назива се нагибом линеарне регресије.

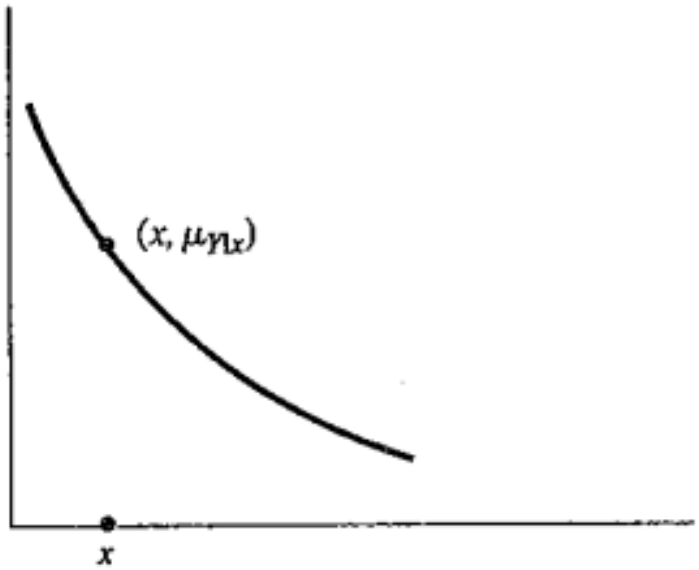
Дакле, разматрамо

просту - постоји само једна независна променљива

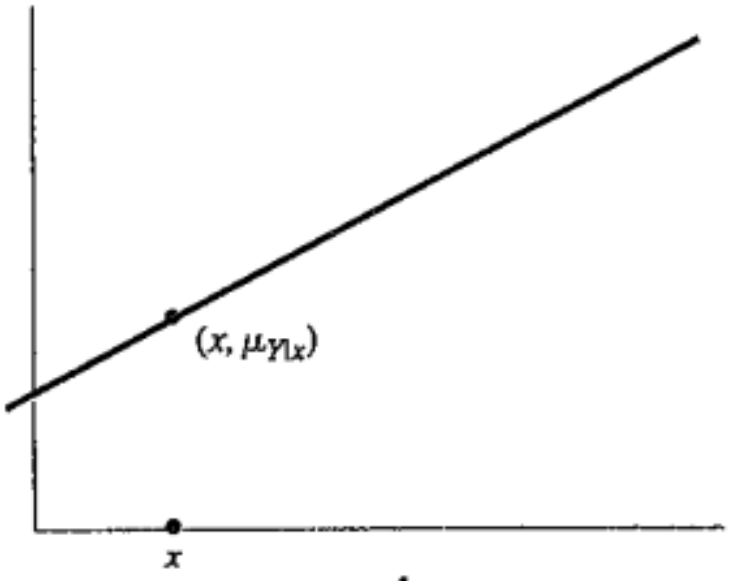
линеарну - регресиона крива је права линија

регресију

Задатак: оценити α и β на основу x и Y



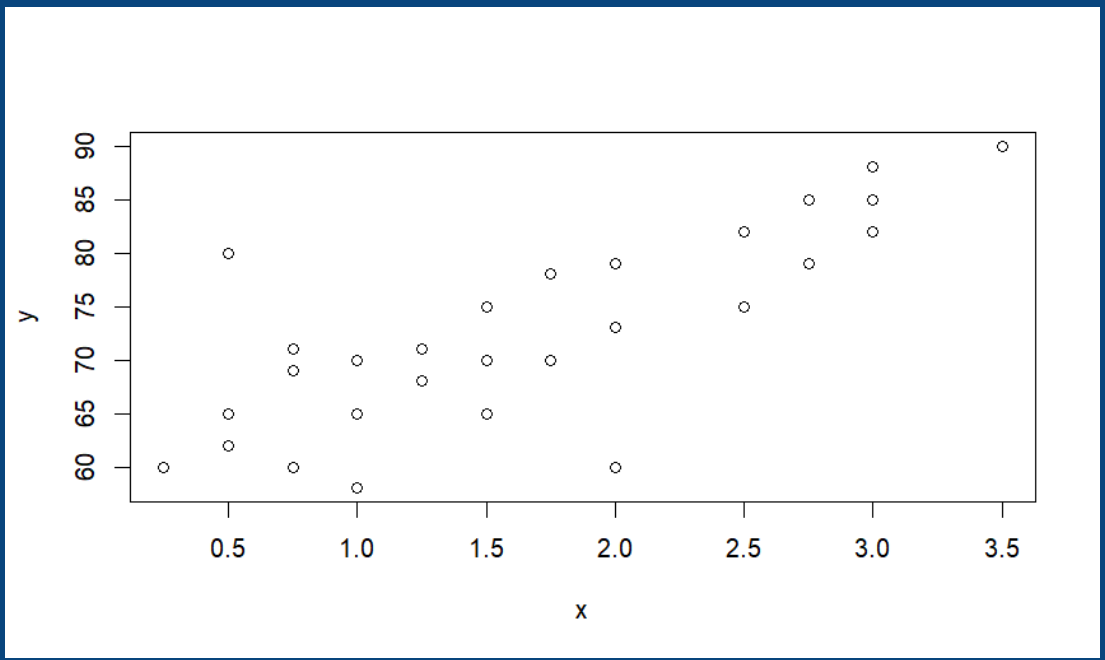
нелинаерна регресиона крива



линарана регресиона крива

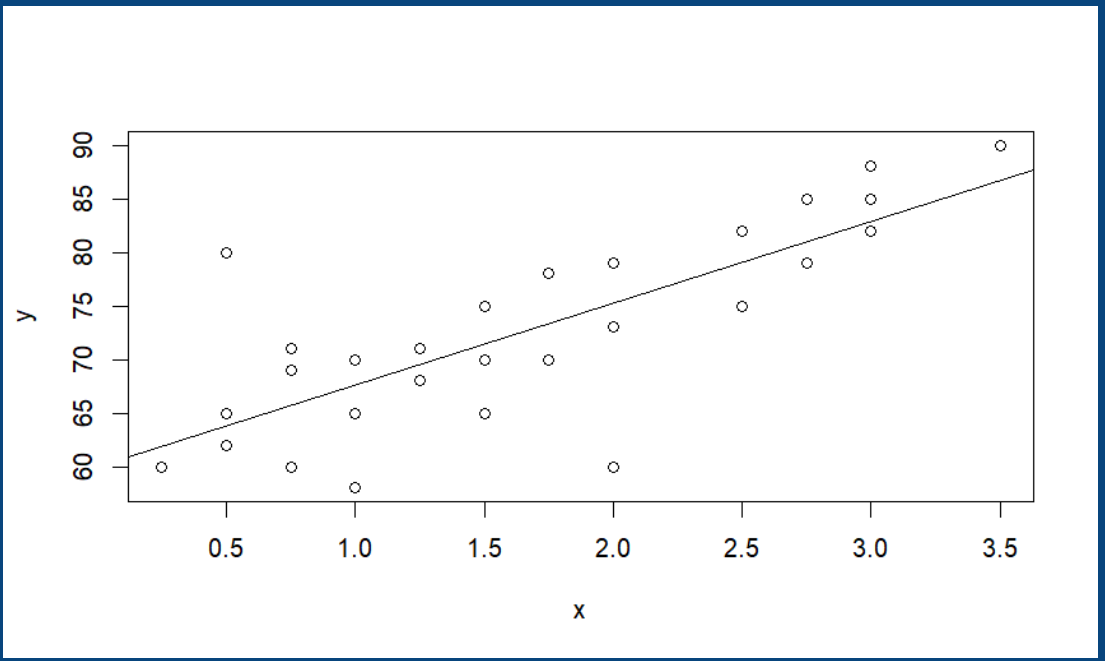
Пре било какве анализе, треба утврдити да се подаци заиста могу моделовати линараном функцијом. У ту сврху црта се график тачака (x_i, y_i).

Пример. Нека x представља број сати које је студент провео у учењу за тест из статистике, а Y број поена коју је добио на тесту. На узорку од 28 студената добијени су резултати:
(0.25,60), (0.75,60), (1.0,65), (1.5,70), (1.75,70), (2.5,75), (3.0,85),
(0.5,62), (0.75,69), (1.0,70), (1.5,75), (2.0,73), (2.5,82), (3.0,82),
(0.5,65), (0.75,71), (1.25,68), (1.5,65), (2.0,79), (2.75,85), (3.0,88),
(0.5,80), (1.0,58), (1.25,71), (1.75,78), (2.0,60), (2.75,79), (3.5,90)



Затим треба одредити једначину праве линеарне регресије на основу случајног узорка $(x_1, Y_1), \ldots, (x_n, Y_n)$. Реализација овог узорка је $(x_1, y_1), \ldots, (x_n, y_n)$.

Пример (наставак).



Метод најмањих квадрата

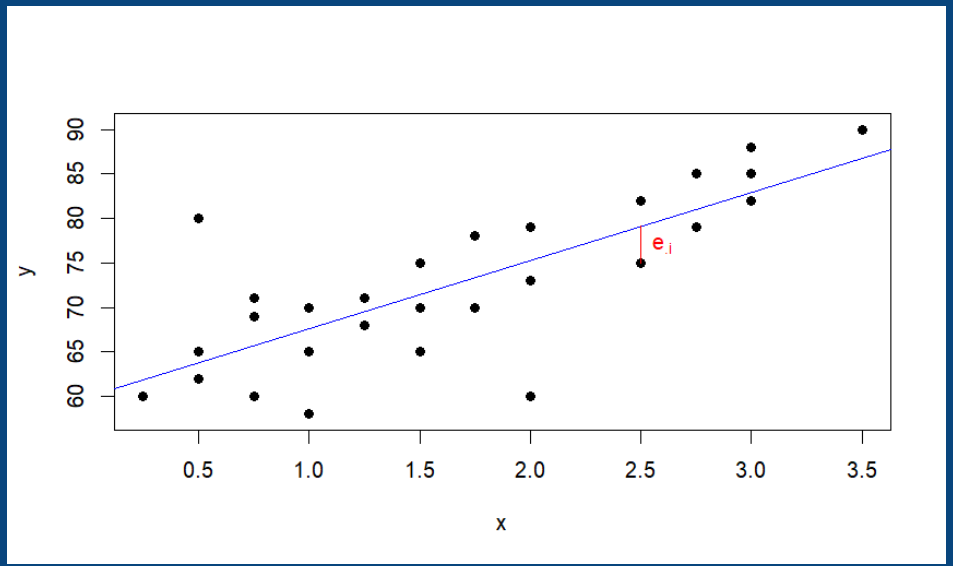
Како добити оцену праве? Узимамо ону праву којој су тачке графика најближе. Меримо вертикална растојања тачака од графика, такозване резидуале

$$e_i = y_i - (a + bx_i),$$

а оцена праве биће за оно a и b за које је збир квадрата резидуала најмањи.

$$SSE = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Пример (наставак).



⋆ Оцена регресионе праве је $\widehat{\mu}_{Y|x} = \widehat{\alpha} + \widehat{\beta}x$, где је

$$\widehat{\beta} = B = \frac{n\sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i x_i - \overline{x}\,\overline{Y}}{\frac{1}{n}\sum x^2 - \overline{x}^2}, \quad \widehat{\alpha} = A = \overline{Y}_n - \widehat{\beta}\overline{x}_n$$

Пример (наставак). У нашем случају $\overline{x} = 1.66$, $\overline{y} = 72.68$, $\sum xy = 3556.75$ и $\sum x^2 = 100.375$. Према томе,

$$b = 7.69, \quad a = 59.91$$

и $\widehat{\mu}_{Y|x} = 59.91 + 7.69x$. Дакле, ако неко учи 2.2 сата предвиђамо да ће освојити $\widehat{\mu}_{Y|2.2} = 59.91 + 7.69 \cdot 2.2 = 76.828$.

- ⋆ Предвиђање помоћу регресионе линије важи само тамо где су подаци. Изван овог опсега, немамо евиденцију да је веза и даље линеарна па се не сме користити, а ако бисмо је користили често бисмо добили бесмислене или чак немогуће вредности.
- ⋆ Методом најмањих квадрата у ствари тачкасто оцењујемо μ_Y (или Y) за сваку вредност x_0. Али за интервалне оцене и тестирање треба нам бољи модел који поред средње вредности описује и одступања од ње.

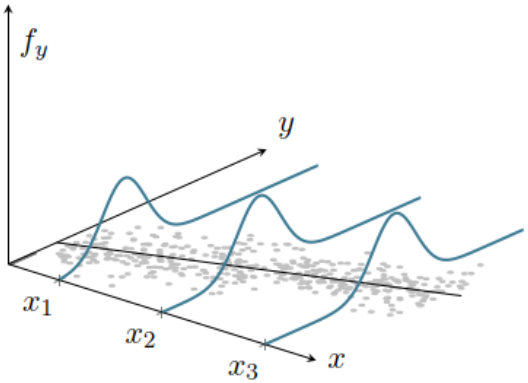
Проста линеарна регресија

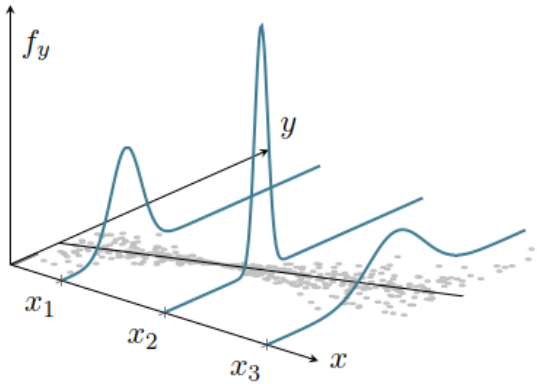
- ⋆ Претпостављамо да је средња вредност $\mu_{Y|x_i} = \alpha + \beta x_i$ за свако i
- ⋆ Одступање вредности Y_i од њене средње вредности, $\alpha + \beta x_i$, називамо грешком регресије и обележавамо са ε_i
- ⋆ Претпостављамо да свако ε_i има нормалну расподелу са средњом вредношћу 0 и неком дисперзијом σ^2 и да су међусобно независни

Теорема: Прост линеарни регресиони модел је

$$Y|x_i = (\alpha + \beta x_i) + \varepsilon_i,$$

где су ε_i независне случајне величине са нормалном $\mathcal{N}(0, \sigma^2)$ расподелом.





- ⋆ Из модела, Y_i зависи од две ствари, $\alpha + \beta x_i$, средње вредности, и ε_i, необјашњене грешке
- ⋆ Модел има три непозната параметра α, β и σ^2
- ⋆ α и β оцењујемо методом најмањих квадрата
- ⋆ Оцена за σ^2 је

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (A + Bx))^2 = \frac{SSE}{n-2}$$

Треба имати у виду да је у уопштем случају тешко израчунати SSE директно из дефиниције. Уводимо „рачунске пречице”:

$$S_{yy} = \sum(y-\overline{y})^2 = \frac{1}{n}(n\sum y^2 - (\sum y)^2) = \sum y^2 - n\overline{y}^2$$

$$S_{xx} = \sum(x-\overline{x})^2 = \frac{1}{n}(n\sum x^2 - (\sum x)^2) = \sum x^2 - n\overline{x}^2$$

$$S_{xy} = \sum(x-\overline{x})(y-\overline{y}) = \frac{1}{n}(n\sum xy - \sum x\sum y) = \sum xy - n\overline{xy}$$

$$\Longrightarrow B = \frac{S_{xy}}{S_{xx}}, \quad SSE = S_{yy} - BS_{xy}$$

Пример (наставак).

```
x <- c(0.25, 0.75, 1.0, 1.5, 1.75, 2.5, 3.0, 0.5, 0.75, 1.0, 1.5, 2.0,
   2.5, 3.0, 0.5, 0.75, 1.25, 1.5, 2.0, 2.75, 3.0, 0.5, 1.0, 1.25, 1.75,
   2.0, 2.75, 3.5)
y <- c(60, 60, 65, 70, 70, 75, 85, 62, 69, 70, 75, 73, 82, 82, 65, 71, 68,
   65, 79, 85, 88, 80, 58, 71, 78, 60, 79, 90)
model<-lm(y~ x)
summary(model)
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2754  -2.3619   0.1381   3.3567  16.2052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.968      2.289   26.20  < 2e-16 ***
x              7.654      1.209    6.33 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.817 on 26 degrees of freedom
Multiple R-squared:  0.6065,    Adjusted R-squared:  0.5914
F-statistic: 40.07 on 1 and 26 DF,  p-value: 1.056e-06
```

Интервал поверења за $\mu_{Y|x}$

- Желимо да формирамо интервал поверења за средњу вредност од Y када је $x = x_0$.
- Као и до сада, треба нам нека случајна величина са познатом расподелом која је повезана на неки начин са параметром

Теорема: Ако је $\widehat{\sigma}^2 = \frac{SSE}{n-2}$, онда случајна променљива

$$\frac{\widehat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\widehat{\sigma}_n \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}}} \in t_{n-2}.$$

$100\beta\%$ интервал поверења за $\mu_{Y|x_0}$ је

$$\left(\widehat{\mu}_{Y|x_0} - c \cdot \widehat{\sigma}_n \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}}, \widehat{\mu}_{Y|x_0} + c \cdot \widehat{\sigma}_n \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}} \right),$$

где је c одређено тако да $P\{T > c\} = \frac{1-\beta}{2}$ где $T \in t_{n-2}$.

Пример (наставак). Желимо да одредимо интервал поверења за средњи број поена на тесту, ако је учено 2.2 сата. Видели смо да је $\overline{x} = 1.66$, $b = 7.69$ и $\widehat{\mu}_{Y|2.2} = 76.828$. Израчунајмо још шта је остало:

$$S_{yy} = \sum (y - \overline{y})^2 = 2236.107$$
$$S_{xx} = \sum (x - \overline{x})^2 = 23.15179$$
$$S_{xy} = \sum (x - \overline{x})(y - \overline{y}) = 177.1964$$
$$\Longrightarrow SSE = S_{yy} - bS_{xy} = 873.4667,\ \widehat{\sigma}^2 = \frac{SSE}{n-2} = 33.59487$$

Према томе, 95% интервал поверења је

$$\left(\widehat{\mu}_{Y|x_0} - c \cdot \widehat{\sigma}_n \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}}, \widehat{\mu}_{Y|x_0} + c \cdot \widehat{\sigma}_n \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}} \right) = (74.20, 79.45),$$

где је $c = F_{t_{26}}^{-1}(0.975) = 2.055$. Дакле, верујемо са поверењем од 95% да ће студент који учи 2.2 сата имати између 74 и 79 поена.

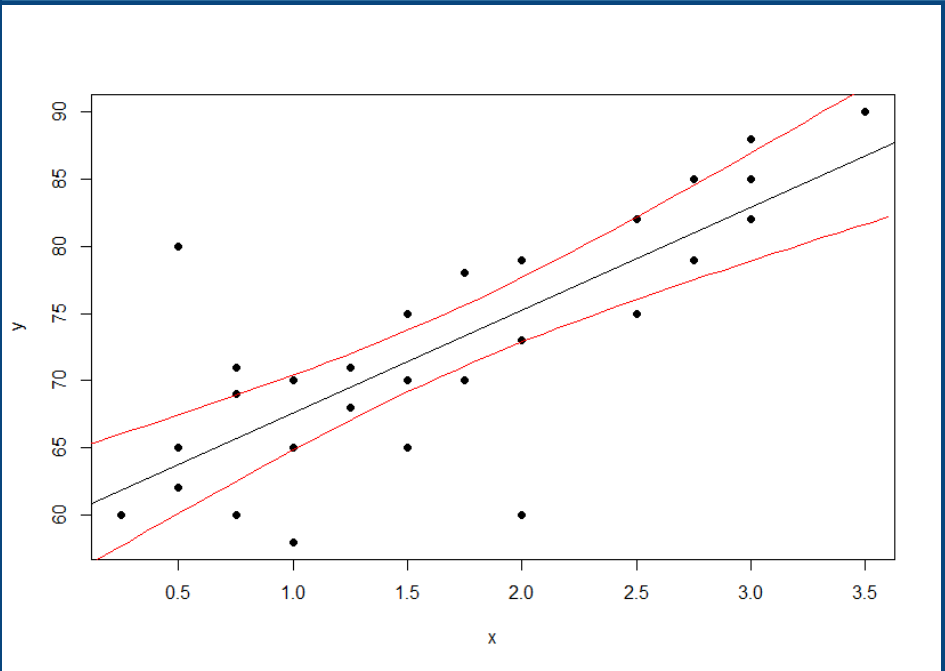
R
```
predict(model,data.frame(x=2.2),interval="confidence")
```

```
> predict(model,data.frame(x=2.2),interval="confidence")
       fit      lwr      upr
1 76.80609 74.17873 79.43346
```

Може се формирати и трака поверења регресионе линије тако што се за свако изабрано x одреди интервал поверења а затим се ти интервали споје глатком кривом.

Пример (наставак).

```
yy<-predict(model,interval="confidence",
    data.frame(x=seq(0,3.6,length=51)))
plot(x,y,pch=19,xlab="x", ylab="y")
abline(model)
lines(seq(0,3.6,length=51),yy[,3],col="red")
lines(seq(0,3.6,length=51),yy[,2],col="red")
```



Интервал предвиђања за $Y|x_0$

- ⋆ Желимо да формирамо интервал који ће, априори, обухватати посматрану вредности $Y|x_0$ са одређеним нивоом поверења
- ⋆ У овом случају желимо да „ухватимо" будућу вредност случајне променљиве, односно да предвидимо интервалом вредност од $Y|x$

Теорема: Ако је $\widehat{\sigma}^2 = \frac{SSE}{n-2}$, онда случајна променљива

$$\frac{\widehat{Y}|x_0 - Y|x_0}{\widehat{\sigma}_n\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x}_n)^2}{S_{xx}}}} \in t_{n-2}.$$

$100\beta\%$ интервал предвиђања за $\mu_{Y|x_0}$ је

$$\left(\widehat{Y}|x_0 - c\cdot\widehat{\sigma}_n\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x}_n)^2}{S_{xx}}},\ \widehat{Y}|x_0 + c\cdot\widehat{\sigma}_n\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x}_n)^2}{S_{xx}}}\right),$$

где је c одређено тако да $P\{T > c\} = \frac{1-\beta}{2}$ где $T \in t_{n-2}$.

Пример (наставак). Желимо да одредимо интервал предвиђања за број поена на тесту, ако је учено 2.2 сата. Користећи претходно добијене резултате, 95% интервал предвиђања је

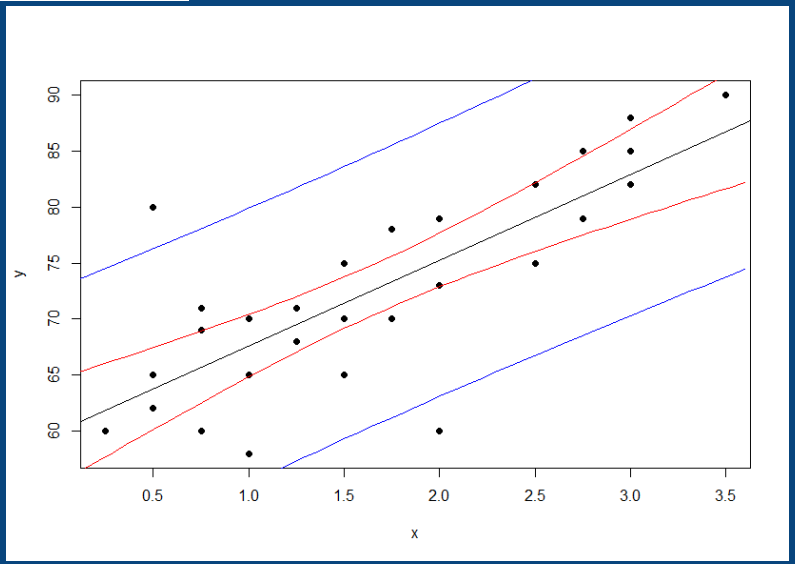
$$\left(\widehat{Y}|x_0 - c\cdot\widehat{\sigma}_n\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x}_n)^2}{S_{xx}}}, \widehat{Y}|x_0 + c\cdot\widehat{\sigma}_n\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x}_n)^2}{S_{xx}}}\right) = (64.58, 89.07),$$

где је $c = F^{-1}_{t_{26}}(0.975) = 2.055$. Дакле, верујемо са поверењем од 95% да ће студент који учи 2.2 сата имати између 64.6 и 89 поена.

R
```
predict(model,data.frame(x=2.2), interval="prediction")
```

```
> predict(model, data.frame(x = 2.2), interval = "prediction")
       fit      lwr      upr
1 76.80609 64.56298 89.04921
```

R
```
yy2<-predict(model,interval="-
    prediction", data.frame(x
    =seq(0,3.6,length=51)))
lines(seq(0,3.6,length=51),
    yy2[,3],col="blue")
lines(seq(0,3.6,length=51),
    yy2[,2],col="blue")
```



Тестирање хипотеза у вези са параметрима регресије

- ⋆ Видели смо да је први корак у регресионој анализи нацртати график података и, ако је линеарна регресија адекватан избор, онда треба да постоји линеарни тренд са $\beta \neq 0$
- ⋆ Ако нисмо сигурни да ли је линеарни модел адекватан, можемо то тестирати

хипотезе: $H_0(\beta = 0)$ против $H_1(\beta \neq 0)$, односно Y је исто за свако x и регресиони модел је непотребан против алтернативе да је линеарни регресиони модел користан за предвиђање Y на основу x

тест статистика: $T_0 = \frac{B}{\widehat{\sigma}_n}\sqrt{S_{xx}}$ при H_0 има t_{n-2}

***p*-вредност теста:** двострука површина лево од вредности t_0, ако је $t_0 < 0$ или двострука површина десно од вредности t_0, ако је $t_0 > 0$, где је t_0 реализована вредност од T_0.

Пример (наставак). Тестирамо да ли нам је регресиони модел број поена на тесту у односу на број сати учења довољно добар да бисмо могли на основу њега да вршимо предвиђања.

хипотезе: $H_0(\beta = 0)$ против $H_1(\beta \neq 0)$

тест статистика: $T_0 = \frac{B}{\widehat{\sigma}_n}\sqrt{S_{xx}}$ при H_0 има t_{n-2}

реализована вредност тест статистике: $T_0 = \frac{7.69}{\sqrt{33.84238}}\sqrt{23.15179} = 6.36$

***p*-вредност теста:** $2P\{T_0 > 6.36\} \approx 0$

Дакле, одбацујемо нулту хипотезу, односно линеарни модел је користан за предвиђање.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2754  -2.3619   0.1381   3.3567  16.2052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.968      2.289   26.20  < 2e-16 ***
x              7.654      1.209    6.33 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.817 on 26 degrees of freedom
Multiple R-squared:  0.6065,    Adjusted R-squared:  0.5914
F-statistic: 40.07 on 1 and 26 DF,  p-value: 1.056e-06
```

Вишеструка линеарна регресија

- У пракси се често дешава да је потребно испитати утицај више предиктора на зависну променљиву

Пример Разматрамо масу јединке (Y) на основу количину хране (x_1), температуре (x_2), старост јединке (x_3).

- Вишеструки линеарни модел, који претпоставља да Y зависи од променљивих $x_1, x_2, ..., x_p$, је $Y|x_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon_i$, где су ε_i независне случајне величине са нормалном $\mathcal{N}(0, \sigma^2)$ расподелом
- Циљ је моделирање условне расподеле зависне променљиве
- На основу узорка $(x_{11}, ..., x_{p1}, Y_1), ..., (x_{1n}, ..., x_{pn}, Y_n)$ оцењујемо параметре $\beta_0, ..., \beta_p$ методом најмањих квадрата
- Поред стандардних тачкастих и интервалних оцена, једна од најважнијих ствари у вишеструкој регресији је избор модела, тј. одредити које од x_j треба да постоје у формули регресије
- Пошто није могуће да цртамо вишедимензионе променљиве, морамо да добијемо одговор тестирањем

Пример. Размотримо базу података о биљним врстама на Галапагосу. База се састоји од 30 редова које представљају острва и 7 променљивих:

`Species` – број различитих биљних врста на острву

`Area` – површина острва (у km^2)

`Elevation` – максимална надморска висина острва (у метрима)

`Nearest` – удаљеност до најближег острва (у km)

`Scruz` – удаљеност до острва Санта Круз (у km)

`Adjacent` – површина суседних острва (у km^2)

Променљиву `Species` узећемо за зависну, а последњих 5 као потенцијалне предикторе (независне променљиве).

Пре учитавања базе треба инсталирати пакет `faraway`.

```
library(faraway)
data(gala)
```

Направимо прво модел са свим променљивим.

```
model<-lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data=gala)
summary(model)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

коефицијент β_j уз x_j представља за колико се повећава μ_Y када се x_j повећа за 1, а остали $x-$еви остану непромењени

Проблем је што се често мењањем једног x_j у пракси мењају и други, тј. предиктори нису у општем случају независни.

Дијагостика модела

Оцењивање и закључивање на основу модела зависе од неколико претпоставки:

грешке претпоставља се да су грешке ε_i нормално расподељене, са истом дисперзијом и независне

модел претпостављамо да је структура модела $\mu_{Y|x} = x\beta$ добра

неуобичајне тачке неколико тачака које не одговарају моделу могу значајно променити модел

Бавићемо се само провером претпоставки у вези са грешкама. Како проверавамо?

нормална расподељеност $QQ-$дијаграм резидуала

једнака дисперзија график тачкастих оцена $\widehat{y}_i$ и ε_i

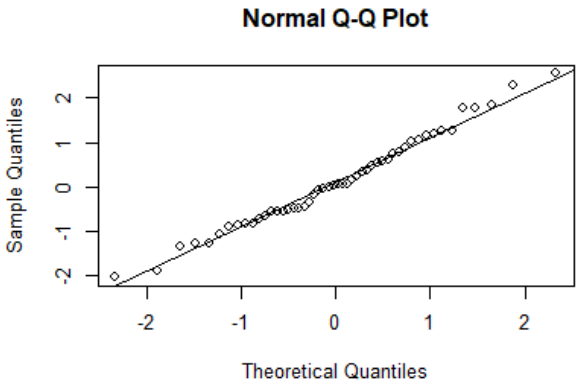
независност проверава се зависност суседних резидуала (ово има смисла ако су Y_i дати неким редоследом, нпр. временски следе један другог)

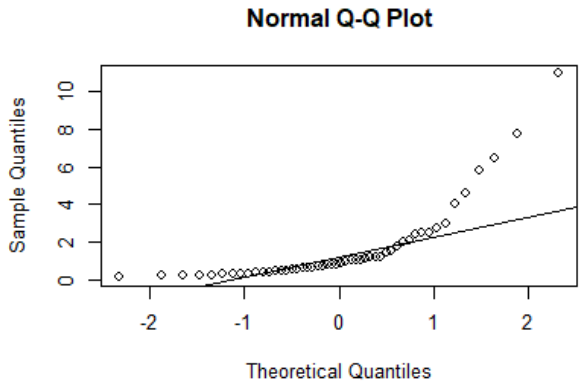
$QQ-$дијаграм

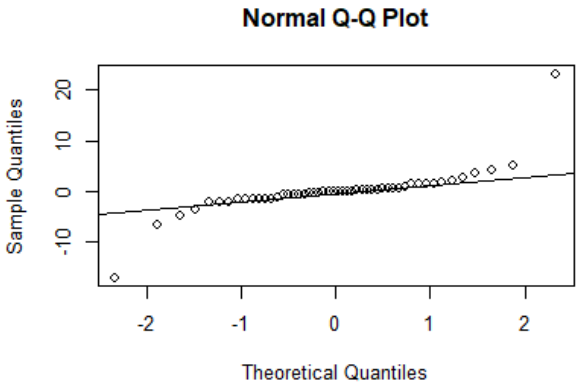
- QQ-дијаграм служи да се испита да ли се за неки узорак може сматрати да је из нормалне расподеле
- На овом дијаграму на $x-$оси налазе се теоријски квантили стандардне нормалне расподеле, а на $y-$оси узорачки квантили, односно упоређујемо резидуале са „идеално"нормалним вредностима
- Користи се чињиница да је свака нормална случајна променљива линеарна функција стандардне нормалне, па и квантили случајног узорка се претпоставља да неће много одступати од праве
- Уколико тачке дијаграма не одступају много од праве, можемо сматрати да је расподела нормална, у супротном не можемо

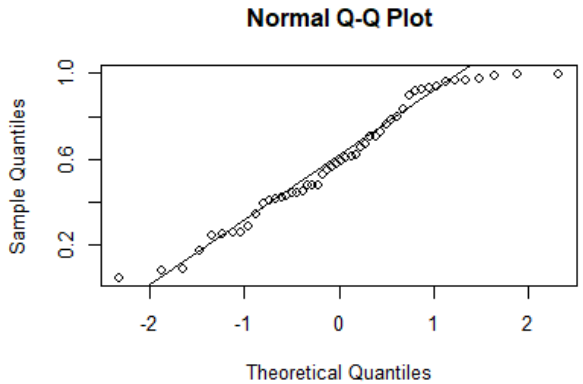
нормална расподела грешке
$\Longrightarrow$ може се користити линеарни модел

несиметрична расподела грешке
$\Longrightarrow$ препоручује се трансформација квадратним кореном или логаритмом









симетрична расподела „тешких репова”
$\Longrightarrow$ не може се користити линеарни модел

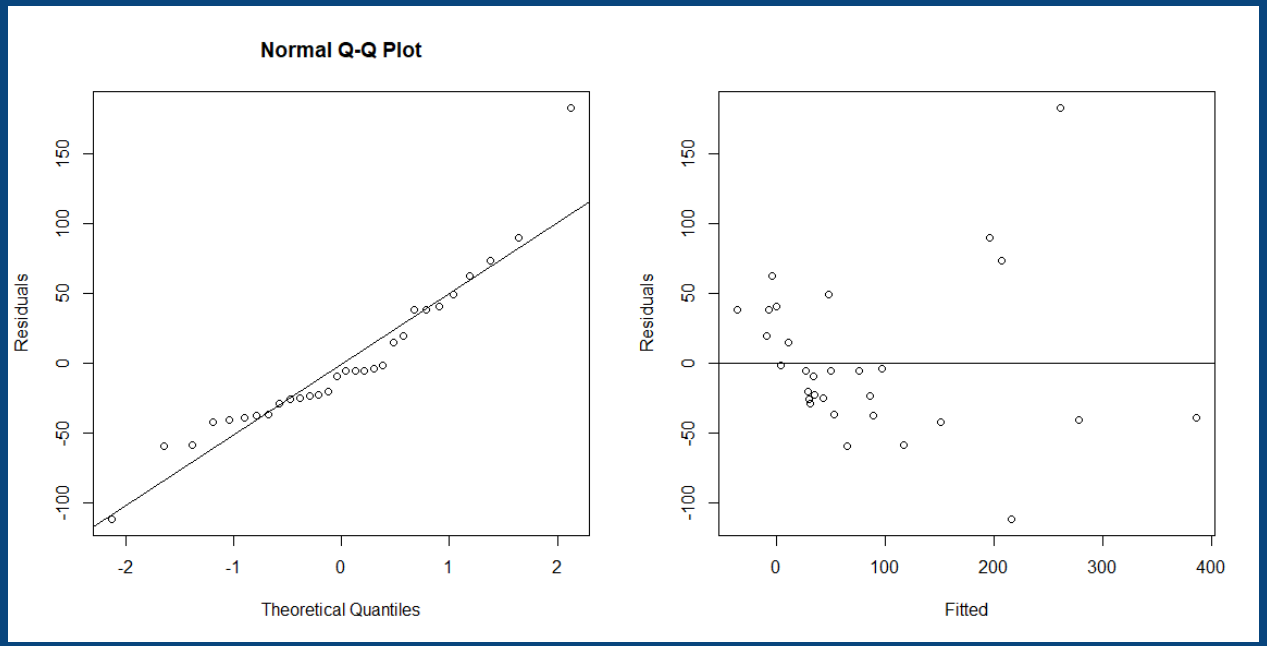
ограничена симетрична расподела грешке
$\Longrightarrow$ може се користити линеарни модел

График тачкастих оцена и резидуала

- ⋆ Проверавамо да ли је дисперзија у резидуалима повезана са неком другом величином
- ⋆ Цртамо график $\widehat{y}$ и $\widehat{\varepsilon}$
- ⋆ Ако су дисперзије једнаке, требало би да се види константност у вертикалном правцу, а тачке треба да буду симетричне око хоризонталне линије у 0

Пример (наставак).

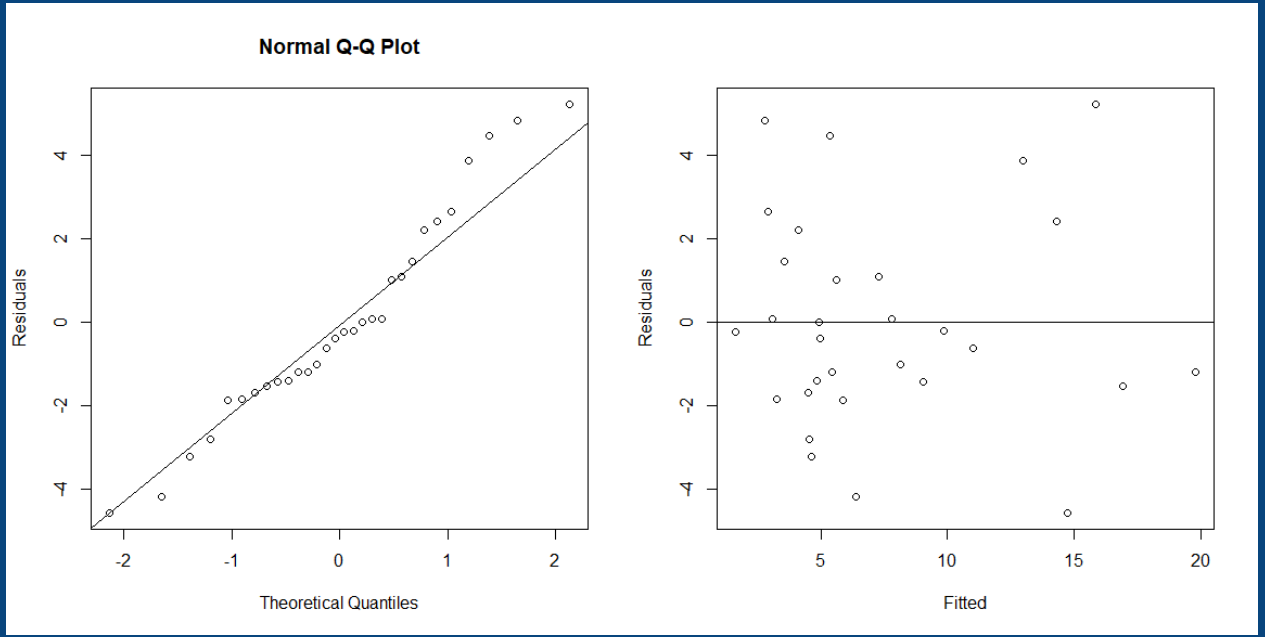
```
qqnorm(residuals(model),ylab="Residuals")
qqline(residuals(model))
plot(fitted(model), residuals(model),xlab="Fitted",ylab="Residuals")
```



Оба дијаграма нам говоре да наше претпоставке нису лоше, али можемо пробати да побољшамо модел трансформацијом

Пример (наставак).

```
model.koren <- lm(sqrt(Species)~ Area + Elevation + Scruz + Nearest +
    Adjacent, gala)
```



Можемо видети да су сада претпоставке боље задовољене, међутим нови модел је обично тежи за интерпретацију

Избор модела

- Циљ нам је доћи до што једноставнијег модела (тј. са што мање променљивих) који је једнако добар као и модел са свим предикторима
- Најчешће кренемо од модела са свим потенцијалним предикторима па избацујемо један по један
- Избацивање једног предиктора мења модел (вредности коефицијената и њихову значајност), тако да се може догодити да предиктор који није био значајан у ширем моделу сада то буде, и обрнуто
- За упоређивање два модела користимо Фишеров $F-$тест
  - Претпоставимо да имамо два модела (ω се може добити од Ω стављањем да су неки параметри једнаки нули) и нека Ω има p непознатих параметара, а ω q непознатих параметара, $q < p$.
  - Желимо да тестирамо H_0 да су оба модела једнако добра против алтенативе да су модели једнако добри, односно $H_0 : \beta_{q+1} = 0, ..., \beta_p = 0$ против алтернативе $H_1 : \exists i\ \beta_i \neq 0$.
  - Тест статистика: $F_0 = \frac{SSE_\omega - SSE_\Omega}{SSE_\Omega} \frac{n-p}{p-q}$ при H_0 има $F_{p-q,n-p}$, где су SSE_Ω и SSE_ω редом збирови квадрата грешака модела
  - $p-$вредност теста је површина десно од вредности f_0 коју F_0 узме у узорку

Пример (наставак).

```
model2 <- lm(Species~ Elevation + Adjacent, data=gala)
```

```
> summary(model)

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

```
> summary(model2)

Call:
lm(formula = Species ~ Elevation + Adjacent, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-103.41  -34.33  -11.43   22.57  203.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43287   15.02469   0.095 0.924727
Elevation    0.27657    0.03176   8.707 2.53e-09 ***
Adjacent    -0.06889    0.01549  -4.447 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.86 on 27 degrees of freedom
Multiple R-squared:  0.7376,    Adjusted R-squared:  0.7181
F-statistic: 37.94 on 2 and 27 DF,  p-value: 1.434e-08
```

```
anova(model,model2)
```

```
Analysis of Variance Table

Model 1: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
Model 2: Species ~ Elevation + Adjacent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     24  89231
2     27 100003 -3    -10772 0.9657  0.425
```

Видимо да је $p-$вредност велика па сматрамо да су модели једнако добри, а пошто је други модел једноставнији определићемо се за њега

Да ли можемо још смањити број променљивих?

```
> model3 <- lm(Species ~ Elevation, data=gala)
> anova(model2,model3)
> model3 <- lm(Species ~ Adjacent, data=gala)
> anova(model2,model4)
```

Добијају се $p-$вредности $F-$теста 0.0001344 и 2.533e-09, редом, што значи да у оба случаја одбацујемо H_0.

Квалитет модела

- ⋆ Колико добро наш модел описује зависност између предиктора и зависне променљиве?
- ⋆ Удео објашњеног варијабилитета R^2 у моделу назива се коефицијент детерминације
- ⋆ Бољем моделу одговара већи коефицијент детерминације

Пример (наставак).

```
> summary(model)

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

```
> summary(model2)

Call:
lm(formula = Species ~ Elevation + Adjacent, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-103.41  -34.33  -11.43   22.57  203.65

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43287   15.02469   0.095 0.924727
Elevation    0.27657    0.03176   8.707 2.53e-09 ***
Adjacent    -0.06889    0.01549  -4.447 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.86 on 27 degrees of freedom
Multiple R-squared:  0.7376,    Adjusted R-squared:  0.7181
F-statistic: 37.94 on 2 and 27 DF,  p-value: 1.434e-08
```

Видимо да се врло мало разликују вредности.

Пример (наставак).

```
> summary(model3)

Call:
lm(formula = Species ~ Elevation, data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-218.319  -30.721  -14.690    4.634  259.180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.33511   19.20529   0.590     0.56
Elevation    0.20079    0.03465   5.795 3.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454,    Adjusted R-squared:  0.5291
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

```
> summary(model4)

Call:
lm(formula = Species ~ Adjacent, data = gala)

Residuals:
   Min     1Q Median     3Q    Max
-85.46 -71.41 -42.60   7.62 359.67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.32743   22.27411   3.786 0.000744 ***
Adjacent     0.00347    0.02505   0.139 0.890831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 116.6 on 28 degrees of freedom
Multiple R-squared:  0.0006847, Adjusted R-squared:  -0.03501
F-statistic: 0.01918 on 1 and 28 DF,  p-value: 0.8908
```

На основу коефицијента детерминације видимо да је модел 4 потпуно бесмислен, а модел 3 доста лошији него када имамо оба предиктора.

```
Call:
lm(formula = Species ~ Area, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-99.495 -53.431 -29.045   3.423 306.137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.78286   17.52442   3.640 0.001094 **
Area         0.08196    0.01971   4.158 0.000275 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.73 on 28 degrees of freedom
Multiple R-squared:  0.3817,    Adjusted R-squared:  0.3596
F-statistic: 17.29 on 1 and 28 DF,  p-value: 0.0002748
```

```
> anova(model6,model2)
Analysis of Variance Table

Model 1: Species ~ Area + Elevation + Adjacent
Model 2: Species ~ Elevation + Adjacent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     26  96776
2     27 100003 -1   -3227.3 0.8671 0.3603
```

Видимо да је променљива *Area* значајна у простом моделу, али када је додамо у модел 2 показује се да она не доноси нове информације у модел.