SAP - Treća auditorna vježba

Case study *bike sharing data*: Linarna regresija

Stjepan Begušić, David Bojanić, Andro Merćep, Tessa Bauman, Tomislav Kovačević

15.12.2021.

Linearna regresija

Linearna regresija korisna je u raznim istraživačkim i praktičnim situacijama, a daje odgovore na nekoliko bitnih pitanja:

- Postoji li veza između ulazne varijable (ili više ulaznih varijabli) - regresora, i izlazne varijable (reakcije)?
- Koliko je jaka ta veza?
- Koje ulazne varijable najviše utječu na izlaznu varijablu i koliko je jak taj efekt?
- Možemo li predvidjeti izlaz za neke nove vrijednosti ulaznih varijabli i s kojom točnošću?

Model linearne regresije i estimacija parametara

Model linearne regresije pretpostavlja linearnu vezu između ulaznih i izlaznih varijabli:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

Pretpostavke modela:

- linearnost veze X i Y
- pogreške nezavisne, homogene i normalno distribuirane s $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Iz podataka je moguće dobiti procjenu modela:

$$\hat{Y} = b_0 + \sum_{j=1}^{p} b_j x_j + e,$$

odnosno:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

u matričnom zapisu.

Procjena je zasnovana na metodi najmanjih kvadrata, tj. minimizaciji tzv. “sum of squared errors”:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Derivacijom se dobije:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Da bi se ova jednadžba mogla riješiti potrebno je invertirati matricu $\mathbf{X}^T\mathbf{X} \in \mathrm{R}^{p\times p}$ (složenost $O(n^3)$), uz pretpostavku da je matrica **punog ranga**.

Estimacija parametara linearne regresije u R-u, kao i statistički testovi vezani uz parametre i estimirani model dostupni su u funkciji `lm` u paketu `stats`.

Bike sharing data

Podatci za analizu su dani u datoteci `bike.sharing`, te sadrže informacije o vremenskim prilikama i broju bicikla koje je određena bike-sharing agencija iznajmila taj dan. Skup podataka dostupan je na: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset - a tamo se nalazi i popis varijabli.

```
bike.sharing.data = read.table('bikesharing.csv',header = T,sep = ",")
summary(bike.sharing.data)
```

```
##     instant         dteday             season           yr
##  Min.   :  1.0   Length:731         Min.   :1.000   Min.   :0.0000
##  1st Qu.:183.5   Class :character   1st Qu.:2.000   1st Qu.:0.0000
##  Median :366.0   Mode  :character   Median :3.000   Median :1.0000
##  Mean   :366.0                      Mean   :2.497   Mean   :0.5007
##  3rd Qu.:548.5                      3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :731.0                      Max.   :4.000   Max.   :1.0000
##       mnth          holiday           weekday        workingday
##  Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000
##  Median : 7.00   Median :0.00000   Median :3.000   Median :1.000
##  Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684
##  3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000
##  Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000
##    weathersit         temp             atemp              hum
##  Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200
##  Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267
##  Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279
##  3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302
##  Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725
##    windspeed          casual         registered        cnt
##  Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22
##  1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152
##  Median :0.18097   Median : 713.0   Median :3662   Median :4548
##  Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504
##  3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956
##  Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714
```

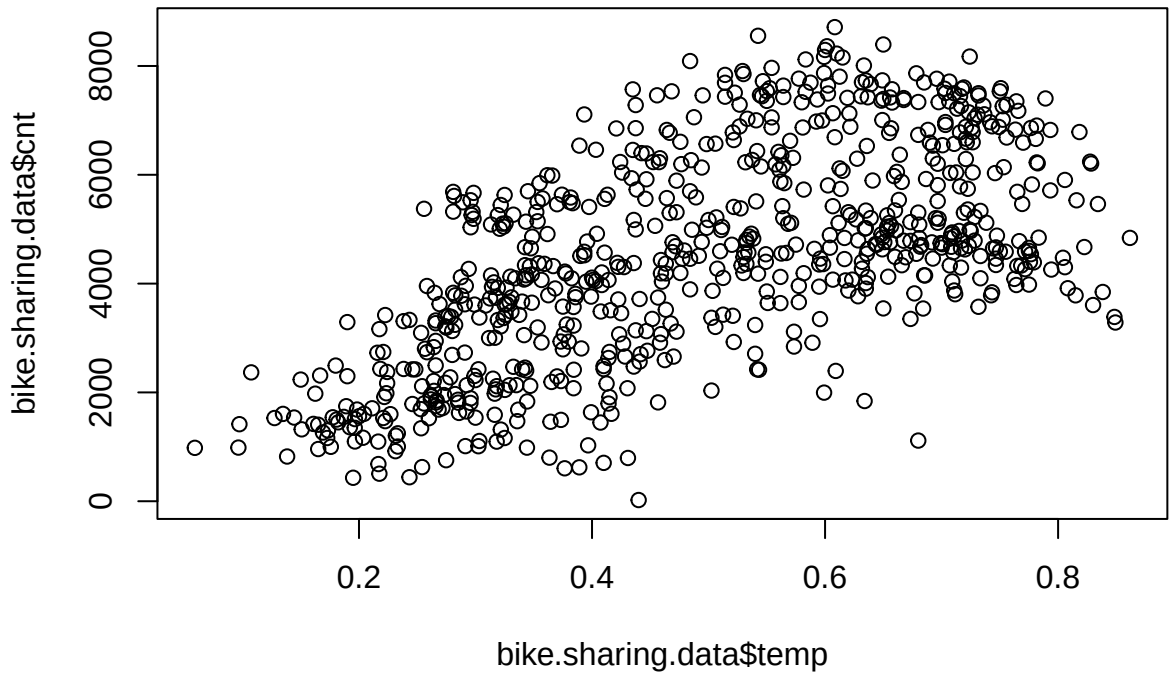
```
bike.sharing.data$dteday <- as.Date(bike.sharing.data$dteday,format("%Y-%m-%d")) # datetime formatting
```

Kako bi znali predvidjeti potrebu za biciklima, možemo ispitati različite varijable koje bi mogle utjecati na broj iznajmljenih bicikala:

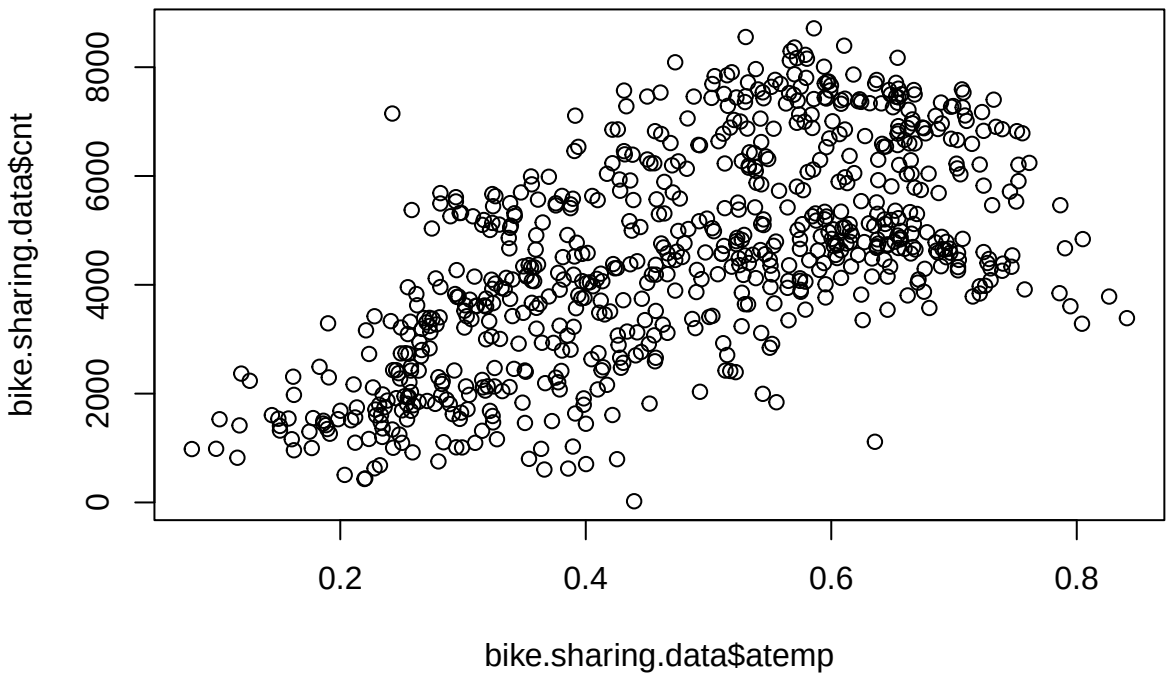
- Prosječna temperatura
- Vlažnost
- Brzina vjetra

Kad promatramo utjecaj samo jedne nezavisne varijable X na neku zavisnu varijablu Y, grafički je moguće dobiti jako dobar dojam o njihovom odnosu - tu je najčešće od pomoći scatter plot.

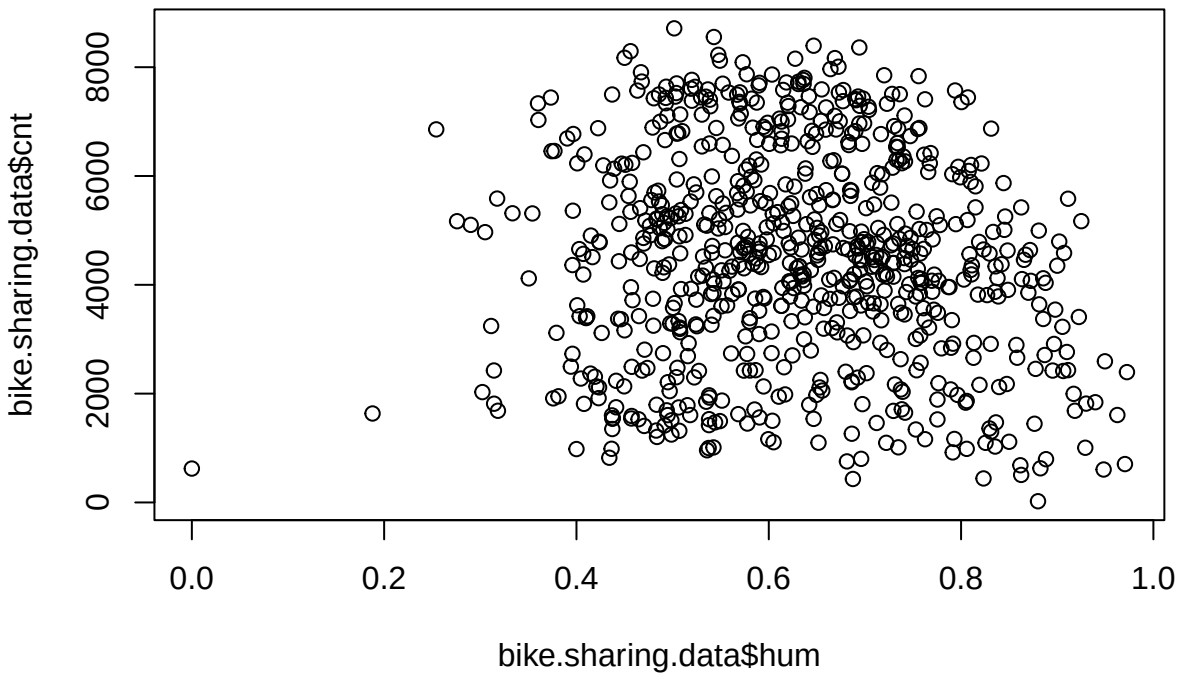
```
plot(bike.sharing.data$temp,bike.sharing.data$cnt) #prosjecna temp vs broj iznajmljenih
```



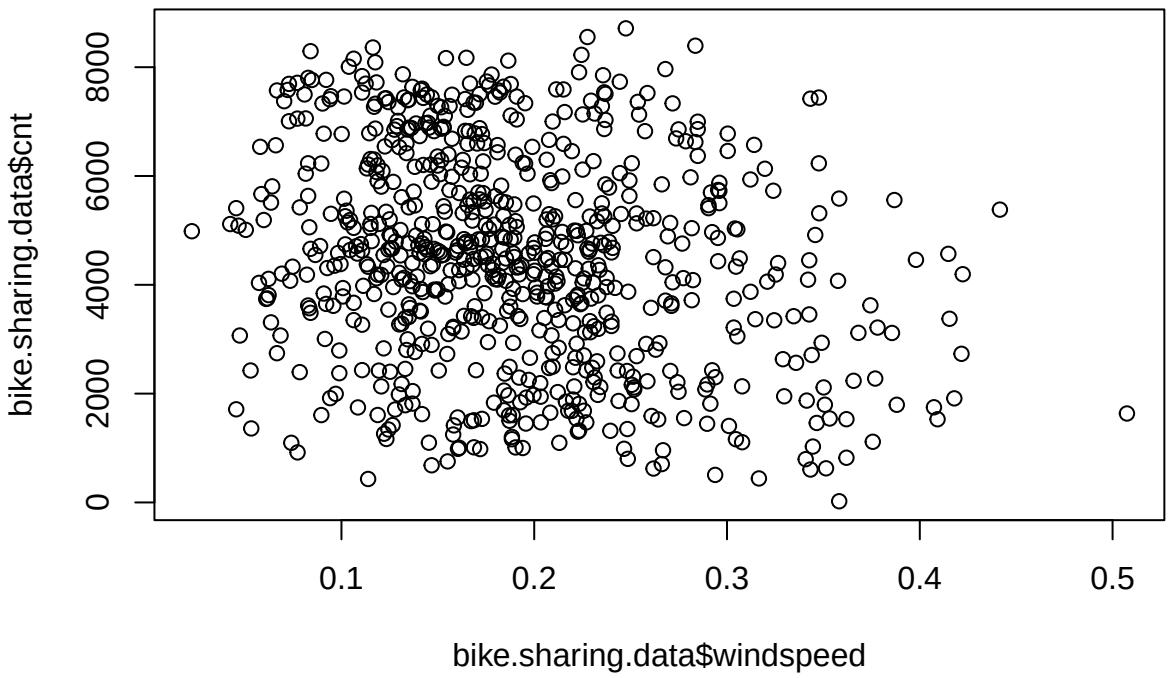

```
plot(bike.sharing.data$atemp,bike.sharing.data$cnt) #prosjecni dojam temp vs broj iznajmljenih
```




```
plot(bike.sharing.data$hum,bike.sharing.data$cnt) #prosjecna vlaznost vs broj iznajmljenih
```




```
plot(bike.sharing.data$windspeed,bike.sharing.data$cnt) #prosjecna brzina vjetra vs broj iznajmljenih
```

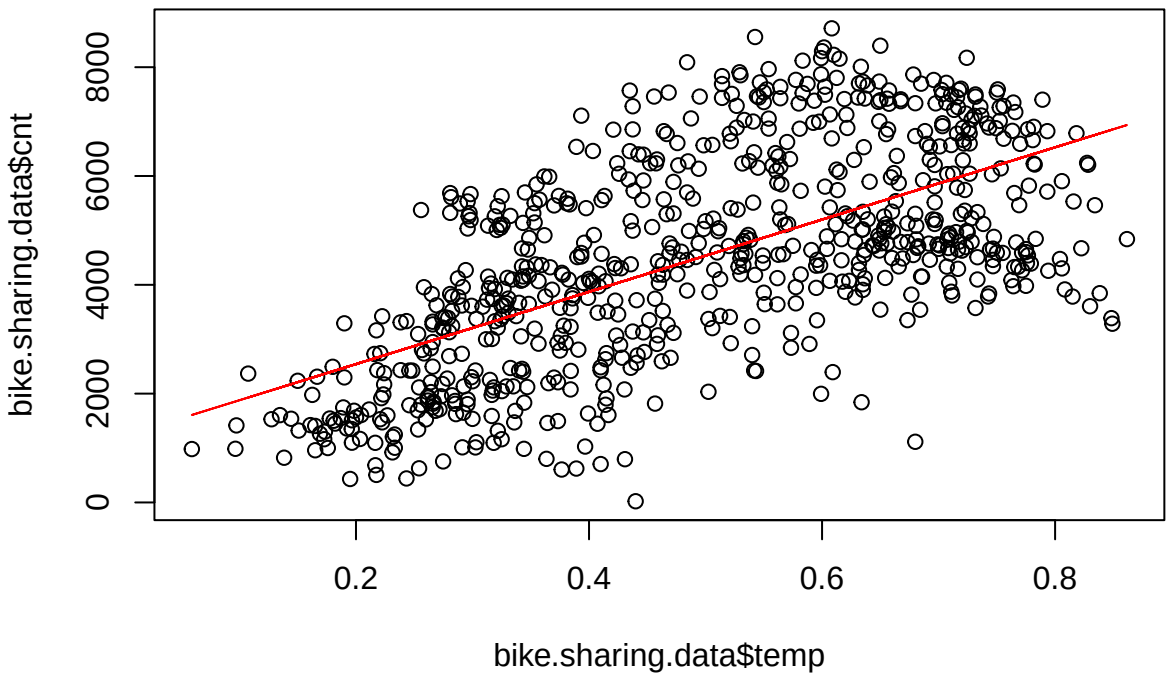


Očito je da temperatura (i prosječna dnevna temperatura i dojam temperature) ima izražen (i to pozitivan) utjecaj na izlaznu varijablu. S druge strane - vlažnost i brzina vjetra izgledaju kao puno slabiji kandidati za modeliranje broja iznajmljenih bicikala (uz neke naznake negativnog utjecaja).

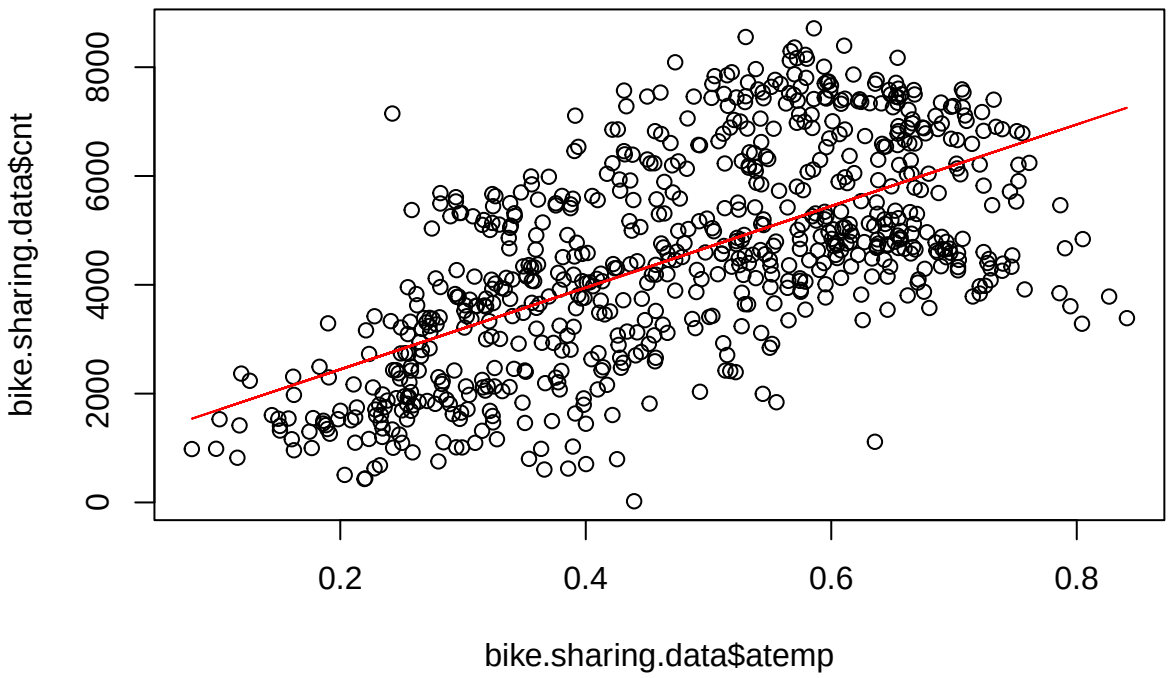
Kako bi ispitali pojedinačni utjecaj ovih varijabli, procijenit ćemo model jednostavne regresije - po jedan za svaku nezavisnu varijablu (uz cnt - broj iznajmljenih bicikala - kao zavisnu varijablu).

Regresijski model procjenjuje se funkcijom lm() koja kao parametre prima zavisne i nezavisne varijable, odnosno data.frame sa svim varijablama i definiciju varijabli u modelu.

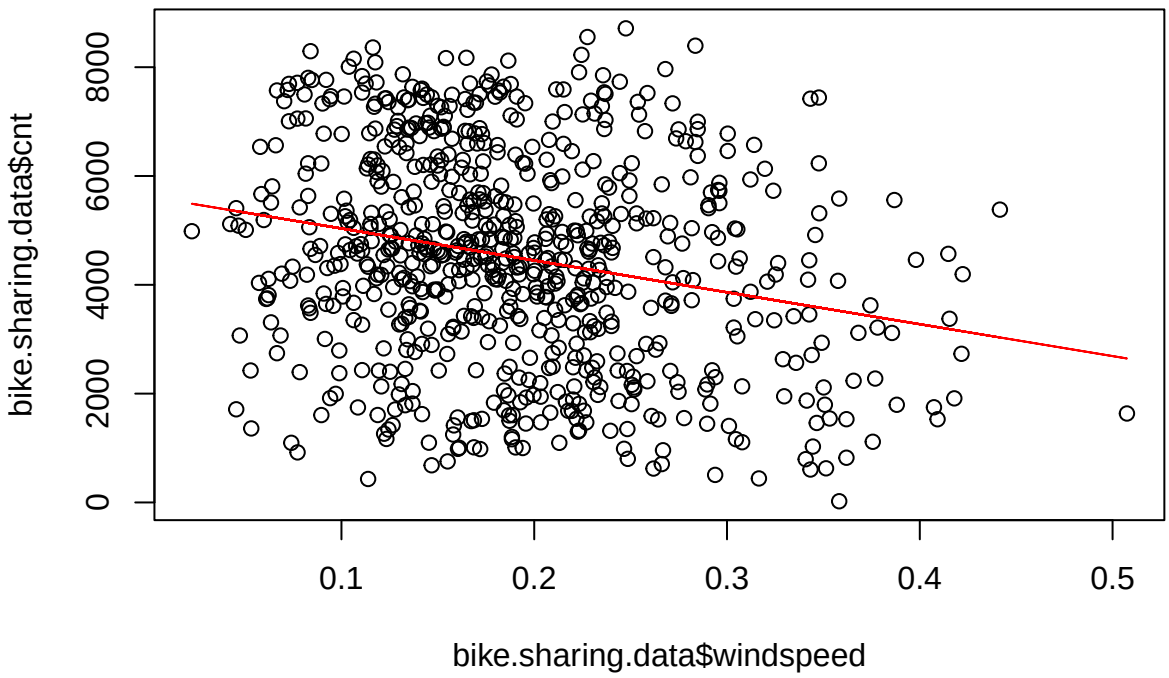
```
fit.temp = lm(cnt~temp,data=bike.sharing.data) #linearni model broja iznajmljenih bicikla (cnt) i tempe

fit.atemp = lm(cnt~atemp,data=bike.sharing.data) #linearni model broja iznajmljenih bicikla (cnt) i tem

fit.hum = lm(cnt~hum,data=bike.sharing.data) #linearni model broja iznajmljenih bicikla (cnt) i tempera

fit.windspeed = lm(cnt~windspeed,data=bike.sharing.data) #linearni model broja iznajmljenih bicikla (cn


plot(bike.sharing.data$temp,bike.sharing.data$cnt) #graficki prikaz podataka
lines(bike.sharing.data$temp,fit.temp$fitted.values,col='red') #graficki prikaz procijenjenih vrijednos
```



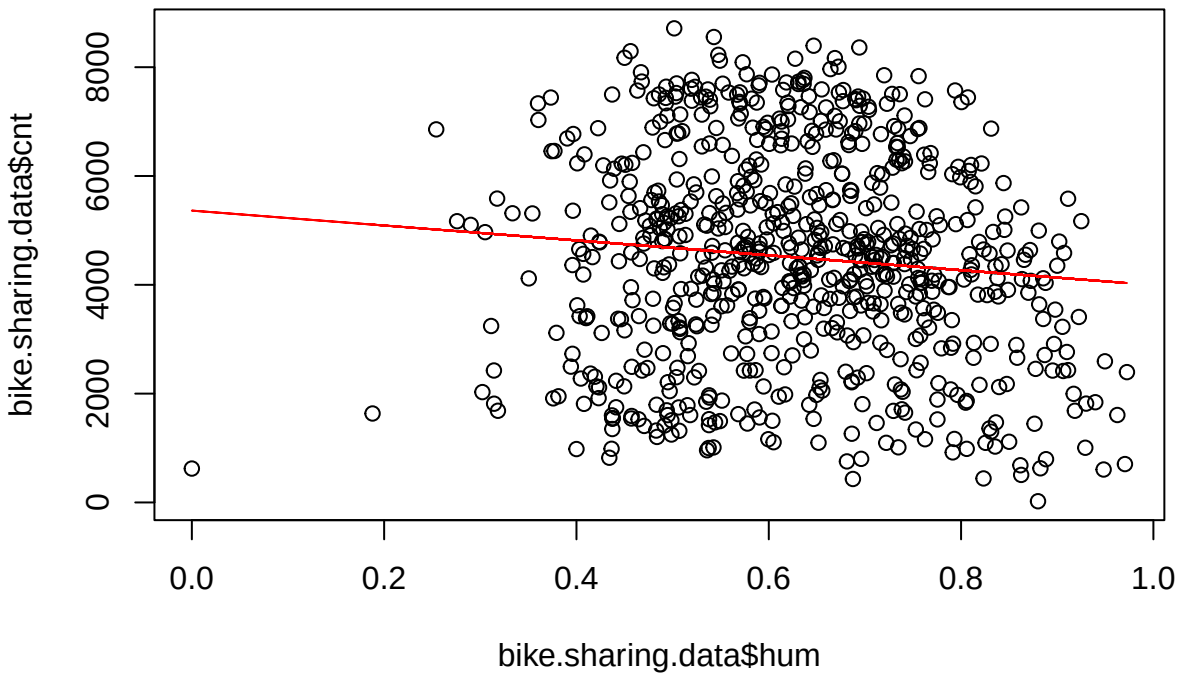

```
plot(bike.sharing.data$atemp,bike.sharing.data$cnt) #graficki prikaz podataka
lines(bike.sharing.data$atemp,fit.atemp$fitted.values,col='red') #graficki prikaz procijenjenih vrijedn
```




```
plot(bike.sharing.data$windspeed,bike.sharing.data$cnt) #graficki prikaz podataka
lines(bike.sharing.data$windspeed,fit.windspeed$fitted.values,col='red') #graficki prikaz procijenjenih
```




```
plot(bike.sharing.data$hum,bike.sharing.data$cnt) #graficki prikaz podataka
lines(bike.sharing.data$hum,fit.hum$fitted.values,col='red') #graficki prikaz procijenjenih vrijednosti
```

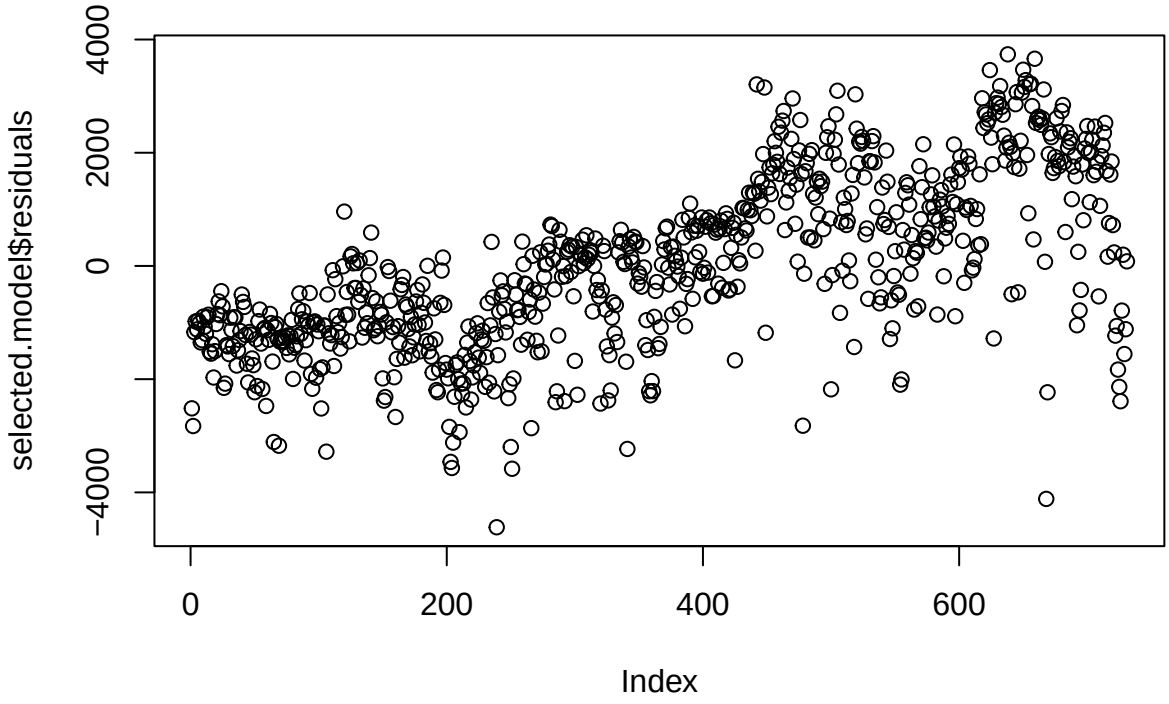


Nagibi pravaca linearne regresije potvrđuju tvrdnje o efektima pojedinih razmatranih varijabli na izlaznu varijablu. Kako bi se dobiveni modeli analizirali i usporedili, prvo je potrebno provjeriti da pretpostavke modela nisu (jako) narušene. Pritom su najbitnije pretpostavke o regresorima (u multivarijatnoj regresiji regresori ne smiju biti međusobno jako korelirani) i o rezidualima (normalnost reziduala i homogenost varijance).

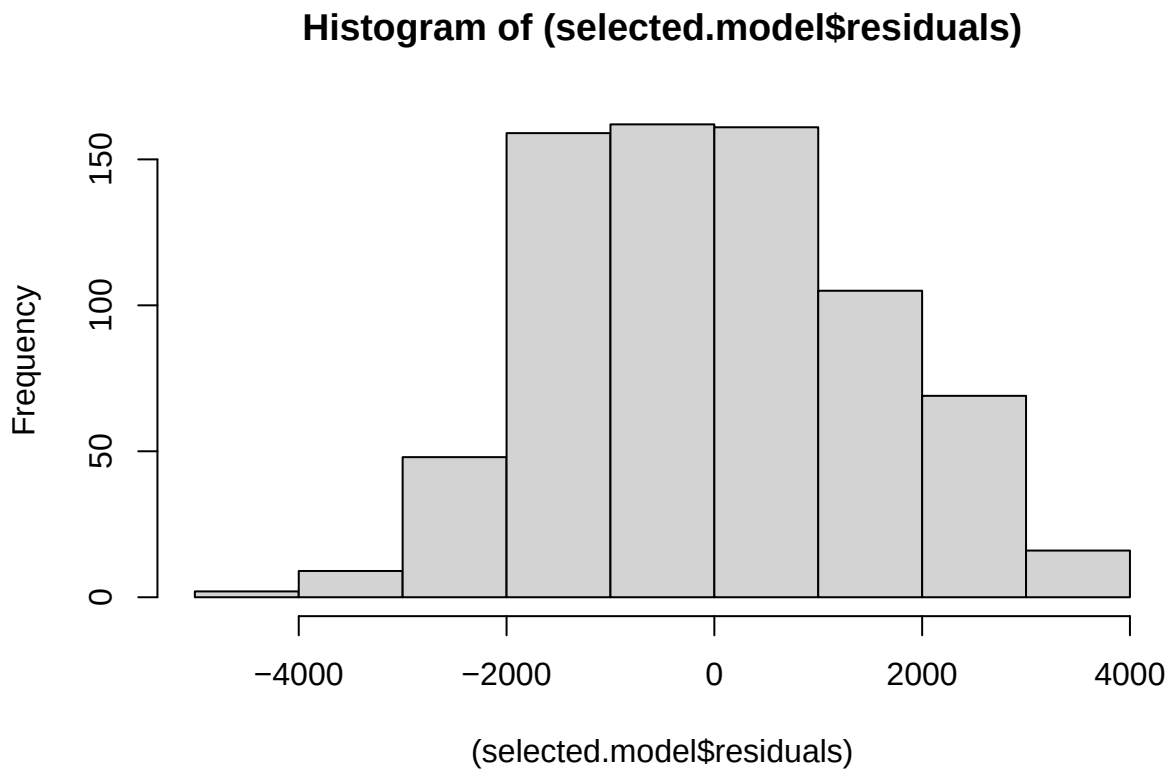
Normalnost reziduala i homogenost varijance

Normalnost reziduala moguće je provjeriti grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Kolmogorov-Smirnovljevog testa.

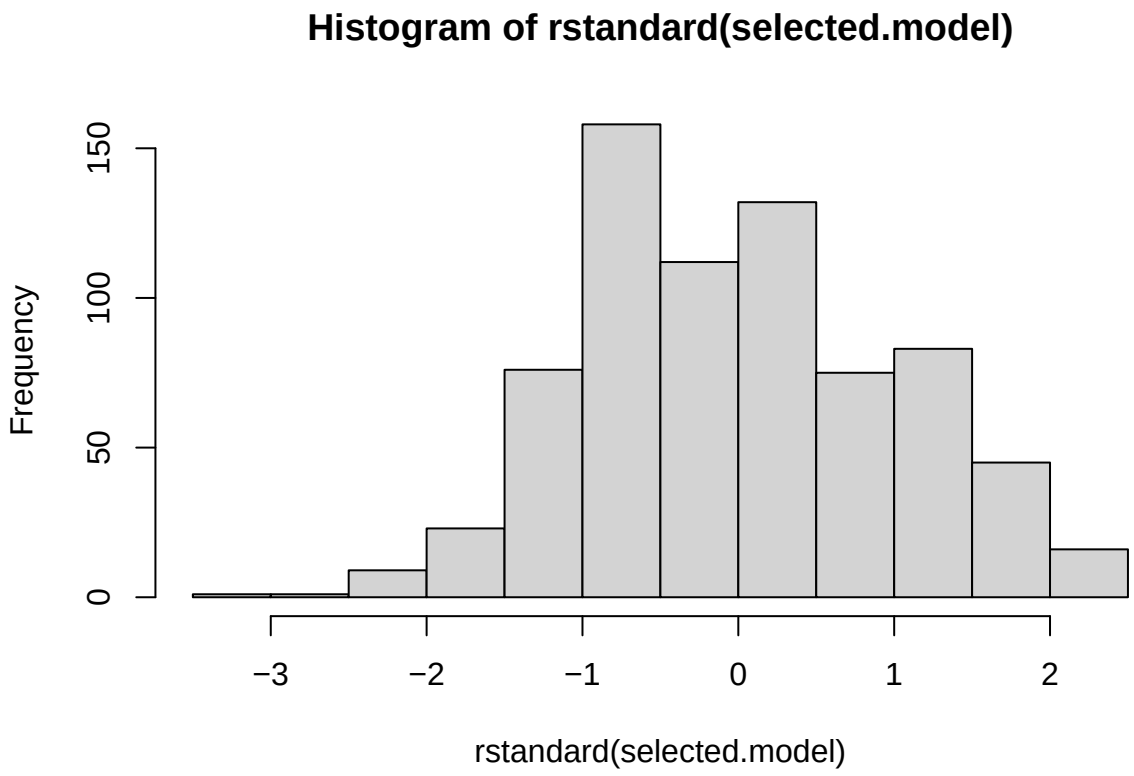
```
selected.model = fit.temp

plot(selected.model$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```



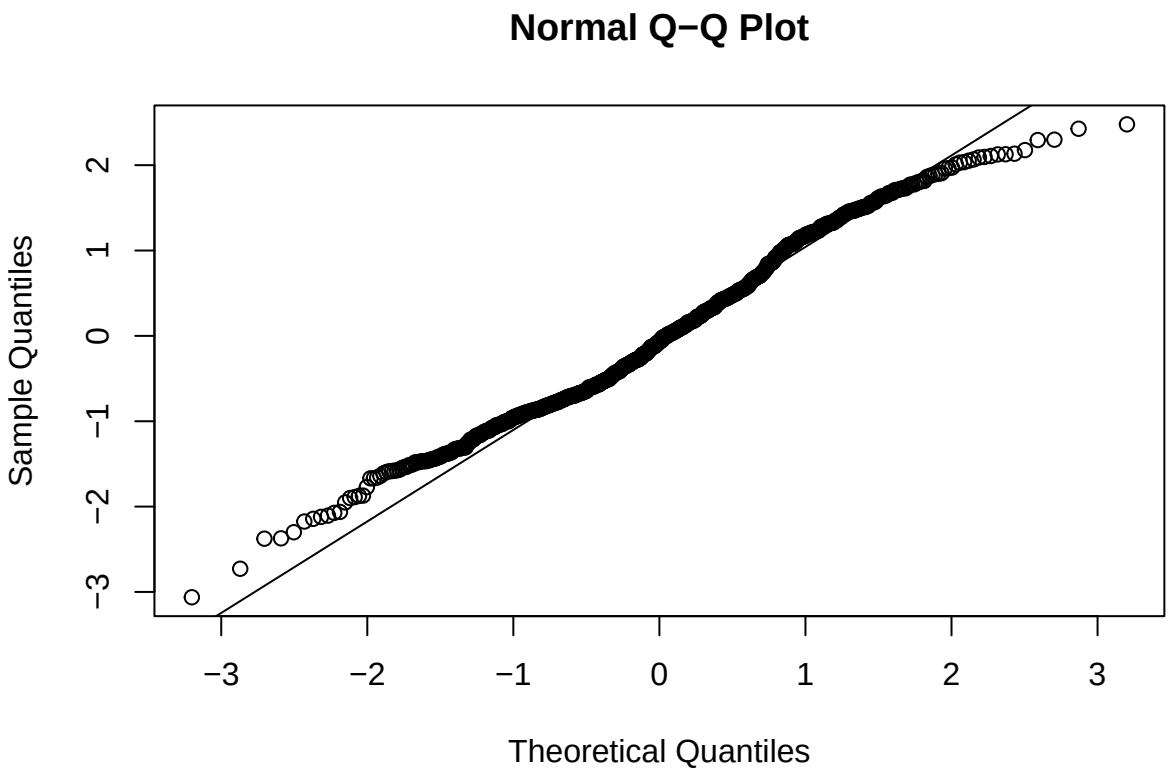

```
#histogram je vrlo interpretativan
hist((selected.model$residuals))
```



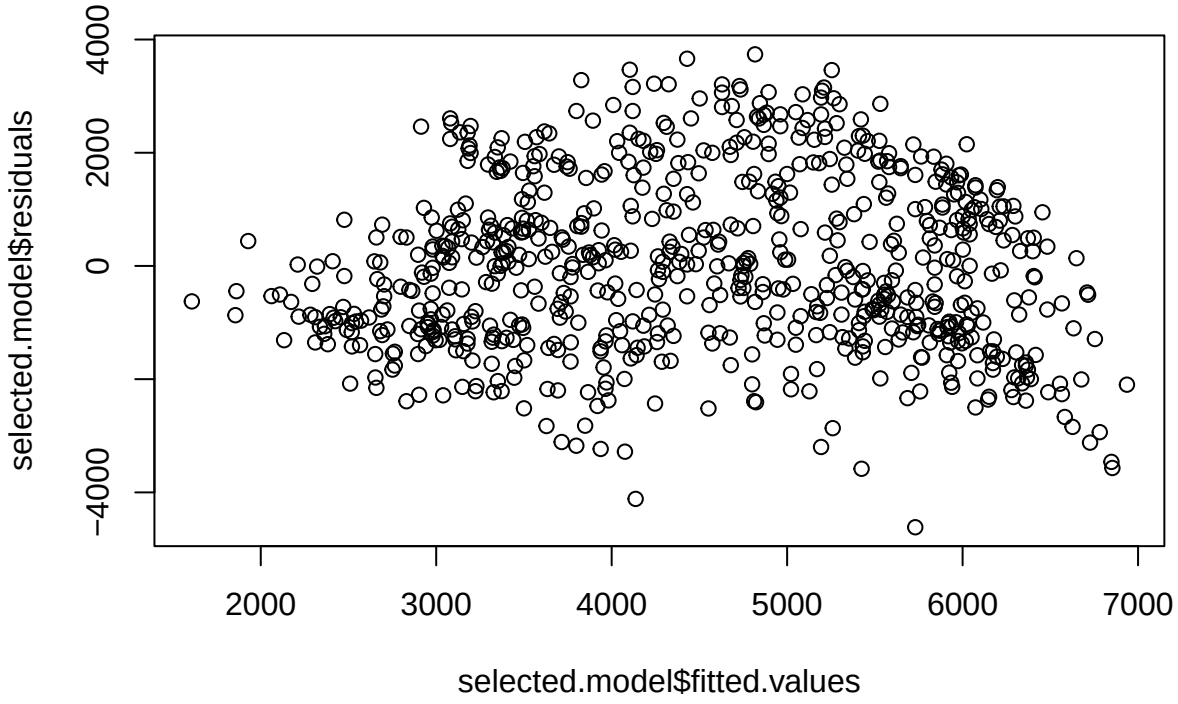

```
hist(rstandard(selected.model))
```



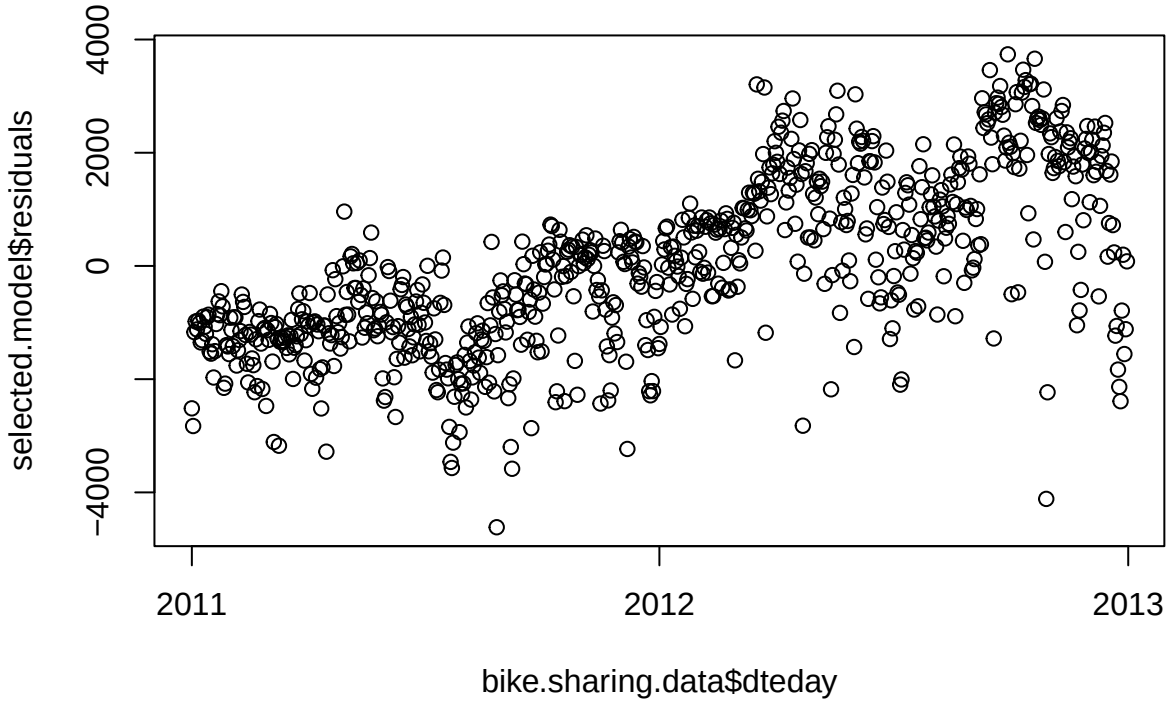

```
#q-q plot reziduala s linijom normalne distribucije
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```




```
plot(selected.model$fitted.values,selected.model$residuals) #reziduale je dobro prikazati u ovisnosti o
```




```
plot(bike.sharing.data$dteday,selected.model$residuals) #a ponekad i u ovisnosti o nekim drugim varijab
```




```
#KS test na normalnost
ks.test(rstandard(fit.windspeed),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.windspeed)
## D = 0.044968, p-value = 0.104
## alternative hypothesis: two-sided
```

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(rstandard(fit.windspeed))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.windspeed)
## D = 0.044822, p-value = 0.0014
```

- Grafički prikaz reziduala samo po indeksu po kojem su dani u podatcima rijetko kad može dati potpunu sliku o njihovoj prirodi - doduše, u ovom slučaju su podatci poredani kronološki, pa taj grafički prikaz odgovara i onom po datumima - koji svjedoči o određenoj vremenskoj zavisnosti podataka.

- Histogram je vrlo lako čitljiv i interpretativan način prikazivanja ovakvih varijabli, te se lako može zaključiti nešto o općenitom obliku distribucije reziduala - u ovom slučaju, ta distribucija donekle nalikuje normalnoj (što otprilike pokazuje i q-q plot), te nije previše zakrivljena.
- Također je jako bitno da u ovisnosti o predviđanjima modela sami reziduali ne pokazuju heterogenost varijance (ne "šire" se s povećanjem $\hat{y}$). No, u ovisnosti o datumu postoji određena dinamika reziduala (ne "izgledaju" potpuno slučajno) koju model ne objašnjava. Takve vremenske zavisnosti se najčešće modeliraju tzv. autoregresivnim modelima (ARMA, ARIMA, ARIMAX, itd.) koji nisu predmet ovog kolegija.
- Statistički testovi se razlikuju u rezultatima (iako se preporuča korištenje Lillieforsove korekcije, u praksi se još uvijek često koristi i K-S test a i druge inačice). No, budući da reziduali ne pokazuju preveliko odstupanje od normalnosti (u smislu zakrivljenosti ili drugih razlika u distribuciji) te je poznato da je t-test robustan na (ne)normalnost - u analizi podataka se u ovakvim slučajevima i dalje mogu donositi statistički zaključci iz regresijskih modela.

Ocjena kvalitete modela i statističko zaključivanje o procijenjenom modelu

Ako pretpostavke modela nisu (neprihvatljivo) prekršene, moguće je primijeniti različite statističke testove o procijenjenim koeficijentima i modelu.

t-test koeficijenata modela

Budući da vrijedi $B_i \sim N(\mu_{B_i}, \sigma_{B_i})$, $\mu_{B_i} = \beta_i$, statistika

$$T = \frac{B_i - \beta_i}{SE(B_i)}$$

ima t-distribuciju s $n - k - 1$ stupnjeva slobode, gdje je k broj parametara. Većina programskih paketa, pa tako i R, pri estimiranju koeficijenata linearne regresije automatski testira $\beta_i = 0$. One koeficijente za koje možemo odbaciti $H_0 : \beta_i = 0$ u korist $H_1 : \beta_i \neq 0$ zovemo **značajni koeficijenti**.

Mjere kvalitete prilagodbe modela podatcima

SSE

Mjera koju minimiziramo estimiranjem parametara modela ("fitanjem na podatke") je SSE:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$\mathbf{R^2}$

Vrlo česta mjera kvalitete prilagodbe modela je koeficijent deteminacije, definiran kao:

$$R^2 = 1 - \frac{SSE}{SST},$$

gdje je: $SST = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$ tzv. "total corrected sum of squares". Koeficijent determinacije R^2 je za linearne modele po definiciji $R^2 \in [0, 1]$ i opisuje koji postotak varijance u izlaznoj varijabli Y je estimirani linearni model objasnio/opisao.

Adjusted R^2

Prilagođeni koeficijent determinacije penalizira dodatne parametre u modelu:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

F-test

Za ispitivanje signifikantnosti čitavog modela koristi se F-statistika:

$$f = \frac{SSR/k}{SSE/(n-k-1)},$$

gdje je $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Sve navedene mjere se mogu vidjeti pozivanjem summary() funkcije nad objektom koji vraća lm().

```
summary(fit.temp)
```

```
##
## Call:
## lm(formula = cnt ~ temp, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4615.3 -1134.9  -104.4  1044.3  3737.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1214.6      161.2   7.537 1.43e-13 ***
## temp          6640.7      305.2  21.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1509 on 729 degrees of freedom
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.3929
## F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

```
summary(fit.atemp)
```

```
##
## Call:
## lm(formula = cnt ~ atemp, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4598.7 -1091.6   -91.8  1072.0  4383.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    945.8      171.3   5.522 4.67e-08 ***
## atemp         7501.8      341.5  21.965  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1504 on 729 degrees of freedom
## Multiple R-squared:  0.3982, Adjusted R-squared:  0.3974
## F-statistic: 482.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

```
summary(fit.hum)
```

```
##
## Call:
## lm(formula = cnt ~ hum, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4741.0 -1386.9    50.3  1439.3  4036.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5364.0      322.7  16.623  < 2e-16 ***
## hum          -1369.1      501.2  -2.732  0.00645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1929 on 729 degrees of freedom
## Multiple R-squared:  0.01013,    Adjusted R-squared:  0.008774
## F-statistic: 7.462 on 1 and 729 DF,  p-value: 0.006454
```

```
summary(fit.windspeed)
```

```
##
## Call:
## lm(formula = cnt ~ windspeed, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4522.7 -1374.7   -74.6  1461.8  4544.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5621.2      185.1  30.374  < 2e-16 ***
## windspeed    -5862.9      900.0  -6.514 1.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1884 on 729 degrees of freedom
## Multiple R-squared:  0.05501,    Adjusted R-squared:  0.05372
## F-statistic: 42.44 on 1 and 729 DF,  p-value: 1.36e-10
```

Kao što je bilo vidljivo iz inicijalnih grafičkioh prikaza, temperatura kao varijabla ima vrlo jak efekt na broj iznajmljenih bicikla i objašnjava najveći postotak varijance (što se očituje u najvećim vrijednostima R^2). Također, iako nisu svi modeli jednako kvalitetni, u svim slučajevima su koeficijenti uz zavisnu varijablu značajni, te F-testovi upućuju na to i da su svi modeli značajni (objašnjavaju značajno više varijance od nul modela). Očito čak i varijable hum i windspeed nisu suvišne u modeliranju broja iznajmljenih bicikala, iako je možda njihova vrijednost nešto manja od temp ili atemp.

Korelacijski koeficijent i veza s linearnim modelom

Korelacijski koeficijent je vrlo često korišten koncept zasnovan na linearnoj regresiji, te opisuje smjer i prirodu veze dviju varijabli. Pearsonov korelacijski koeficijent definiran je kao:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

gdje je $S_{xx} = \sum(x_i - \bar{x})^2$,$S_{yy} = \sum(y_i - \bar{y})^2$, a $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$.

Korelacijski koeficijent direktno je određen linearnom regresijom i koeficijentom determinacije R^2 i iznosi $r = \sqrt{R^2}$.

```
cor(bike.sharing.data$hum,bike.sharing.data$cnt)
```

```
## [1] -0.1006586
```

```
cor.test(bike.sharing.data$hum,bike.sharing.data$cnt)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bike.sharing.data$hum and bike.sharing.data$cnt
## t = -2.7317, df = 729, p-value = 0.006454
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17191732 -0.02835191
## sample estimates:
##        cor
## -0.1006586
```

```
summary(fit.hum)
```

```
##
## Call:
## lm(formula = cnt ~ hum, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4741.0 -1386.9    50.3  1439.3  4036.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5364.0      322.7  16.623  < 2e-16 ***
## hum          -1369.1      501.2  -2.732  0.00645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1929 on 729 degrees of freedom
## Multiple R-squared:  0.01013,    Adjusted R-squared:  0.008774
## F-statistic: 7.462 on 1 and 729 DF,  p-value: 0.006454
```

Višestruka regresija

Prije procjene modela višestruke regresije potrebno je provjeriti da pojedini parovi varijabli nisu (previše) korelirani. U principu je određena korelacija između varijabli neizbježna, ali varijable s vrlo visokom korelacijom će uzrokovati probleme u interpretaciji regresijskih rezultata.

```
fit.temps = lm(cnt ~ atemp + temp, bike.sharing.data) #regresija s jako koreliranim varijablama
summary(fit.temps)
```

```
##
## Call:
## lm(formula = cnt ~ atemp + temp, data = bike.sharing.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4604  -1089    -92   1069   3865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    981.0      188.4   5.208 2.48e-07 ***
## atemp         6314.1     2658.1   2.375   0.0178 *
## temp          1066.2     2366.3   0.451   0.6524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1505 on 728 degrees of freedom
## Multiple R-squared:  0.3984, Adjusted R-squared:  0.3968
## F-statistic: 241.1 on 2 and 728 DF,  p-value: < 2.2e-16
```

```
cor(bike.sharing.data$temp,bike.sharing.data$atemp)
```

```
## [1] 0.9917016
```

Regresija s jako koreliranim ulaznim varijablama će uglavnom dati neke rezultate, ali na temelju njih ne možemo donositi nikakve zaključke. U slučaju savršene linearne zavisnosti ili koreliranosti ulaznih varijabli, procjena regresijskog modela će biti nestabilna i barem jedan koeficijent će biti NA.

Stoga je potrebo odabrati onaj podskup varijabli za koje smatramo da objašnjavaju različite efekte u podatcima i nisu međusobno (previše) korelirane.

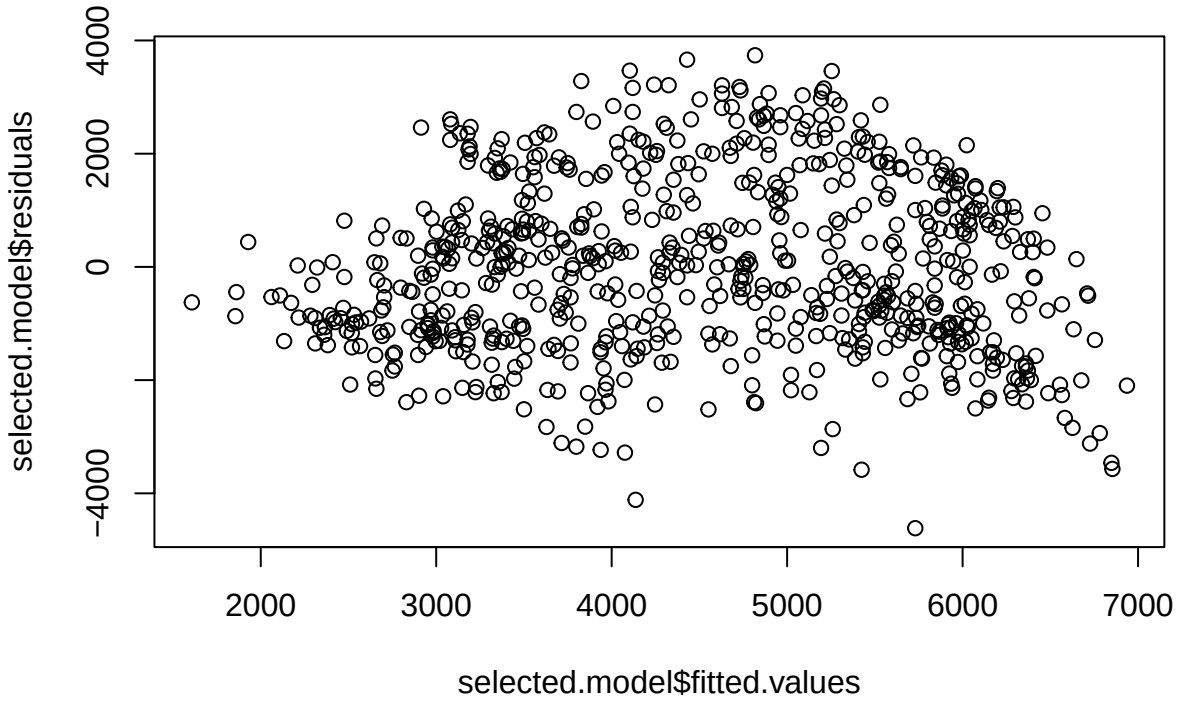
```
cor(cbind(bike.sharing.data$temp,bike.sharing.data$atemp,bike.sharing.data$hum,bike.sharing.data$windsp
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  1.0000000  0.9917016  0.1269629 -0.1579441
## [2,]  0.9917016  1.0000000  0.1399881 -0.1836430
## [3,]  0.1269629  0.1399881  1.0000000 -0.2484891
## [4,] -0.1579441 -0.1836430 -0.2484891  1.0000000
```

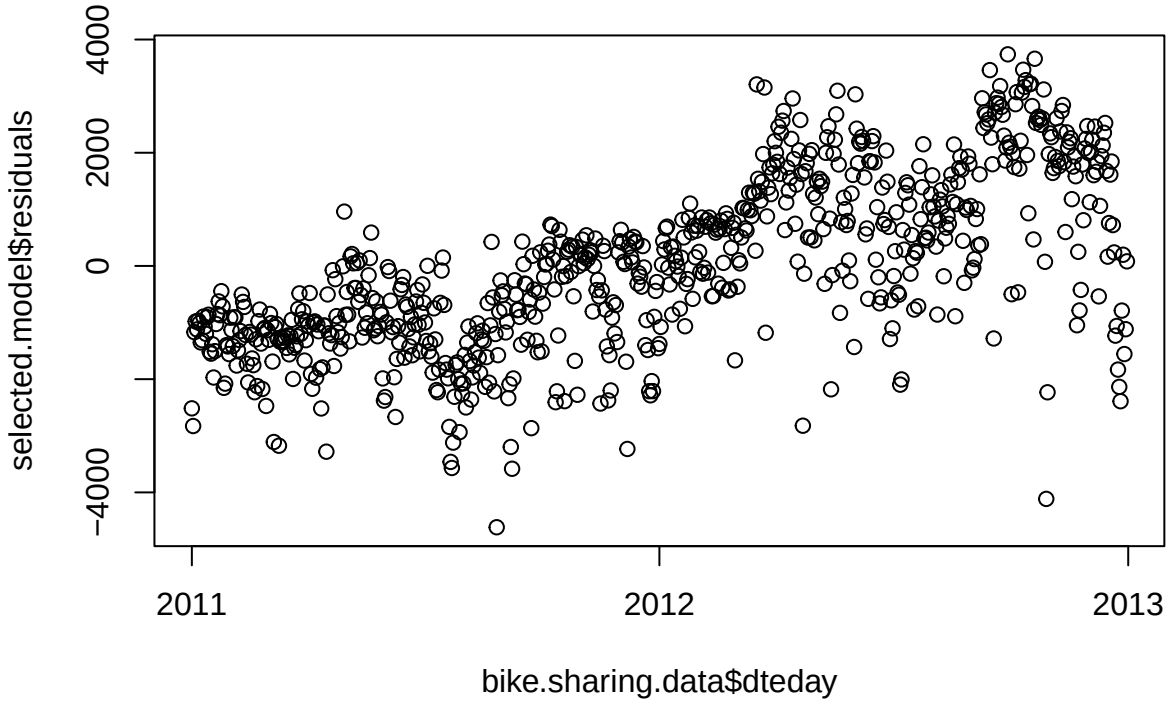
```
fit.multi = lm(cnt ~ atemp + hum + windspeed, bike.sharing.data)
summary(fit.multi)
```

```
##
## Call:
## lm(formula = cnt ~ atemp + hum + windspeed, data = bike.sharing.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4890  -1043    -82   1072   4384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3774.0      342.9  11.007  < 2e-16 ***
## atemp         7504.1      330.2  22.723  < 2e-16 ***
## hum          -3167.5      383.4  -8.261 6.84e-16 ***
## windspeed    -4411.7      709.8  -6.215 8.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1422 on 727 degrees of freedom
## Multiple R-squared:  0.4632, Adjusted R-squared:  0.461
## F-statistic: 209.1 on 3 and 727 DF,  p-value: < 2.2e-16
```

```
plot(selected.model$fitted.values,selected.model$residuals) #reziduali u ovisnosti o procjenama modela
```




```
plot(bike.sharing.data$dteday,selected.model$residuals) #reziduali u ovisnosti o datumu
```




```
#KS test na normalnost
ks.test(rstandard(fit.windspeed),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.windspeed)
## D = 0.044968, p-value = 0.104
## alternative hypothesis: two-sided
```

```
#Lillieforsov test na normalnost
require(nortest)
lillie.test(rstandard(fit.windspeed))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.windspeed)
## D = 0.044822, p-value = 0.0014
```

Zašto su sad varijable hum i windspeed toliko "značajnije" nego kad ih koristimo same? Često se dogodi i obrnut slučaj - uključivanjem dodatnih varijabli pojedine varijable mogu "prestati" biti značajne. U višestrukoj regresiji interakcije (korelacije) varijabli međusobno i sa zavisnom varijablom dolaze do izražaja - moguće su različite interpretacije. Može se tvrditi da je uključivanje varijable temp dodatno "očistilo" reziduale modela

u kojem bi se koristio samo hum ili windspeed i time je dio varijance koji objašnjavaju ove dvije varijable došao do izražaja. U slučaju da su temp i hum ili windspeed objašnjavali iste efekte u podatcima, očekivali bismo da će uključivanje temp uzrokovati da hum ili windspeed "prestanu" biti značajni.

Ove interakcije su uzrok različitih fenomena u statistici, a jedan od poznatijih je i Simpsonov paradoks (https://en.wikipedia.org/wiki/Simpson%27s_paradox).

Model višestruke regresije koji smo ovako dobili objašnjava cca. 46% varijance u podatcima - generalno je teško govoriti koliki je R^2 "dovoljan" za kakve podatke budući da to upravo najviše ovisi o samom području primjene - za razne društvene i ekonomske studije (bilo što vezano uz ljudsko ponašanje) će već 30% biti zadovoljavajući rezultat, dok za neke fizikalne procese ni 80% nije dovoljno dobar model. U konkretnom slučaju, budući da se ipak radi o nečem vezanom uz ljudsko ponašanje, ovaj rezultat se čini dobar, ali kao što se vidi u analizi reziduala (grafički prikaz u odnosu na izlaz modela i u odnosu na datum) - postoje još neki efekti u podatcima koji ovaj model ne uspjeva objasniti.

Kategorijske nezavisne varijable

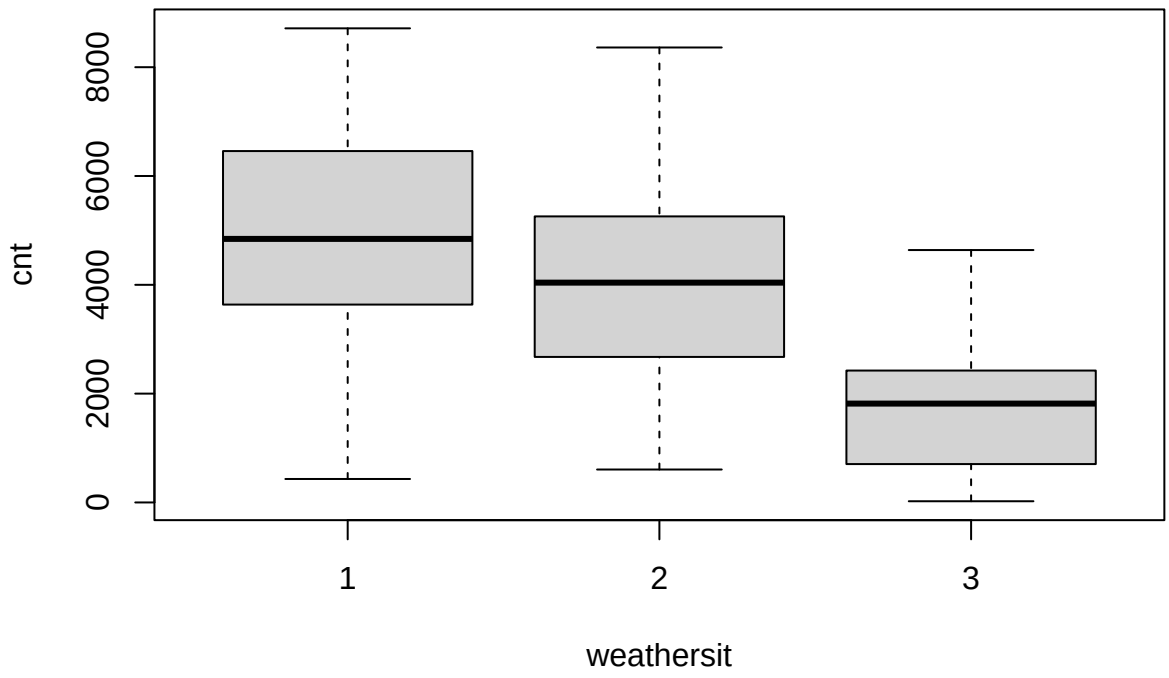
U skupu podataka raspolažemo s nekim kategorijskim varijablama, npr. season (godišnje doba), mnth (mjesec), holiday (indikator radi li se o prazniku taj dan), weekday (dan u tjednu), weathersit (vremenska situacija). Kategorijske varijable se mogu uključiti kao regresori u analizu, ali je prethodno potrebno provjeriti nekoliko stvari:

- radi li se o varijabli na nominalnoj ili ordinalnoj skali,
- ima li varijabla linearan efekt na izlaznu varijablu,
- predstavlja li određena kategorijska varijabla nešto što je određenom metričkom varijablom već predstavljeno.

U konkretnom slučaju, varijabla season je samo varijabla nešto grublje granulacije od varijable mnth, a za obje bismo očekivali da objašnjavaju sličan efekt u podatcima kao i varijabla temp. Varijable holiday i weathersit bi mogle biti korisne i zanimljive.

Korištenje kategorijskih varijabli s više od dvije kategorije kao int vrijednosti u regresiji se ne preporuča za nominalne varijable, iako u tom obliku mogu izgledati korisne u modelima.

```
boxplot(cnt~weathersit,data=bike.sharing.data) #kvadratni dijagram se moze koristiti za graficki provje
```




```
fit.multi.1 = lm(cnt ~ atemp + hum + windspeed + weathersit, bike.sharing.data)
summary(fit.multi.1)
```

```
##
## Call:
## lm(formula = cnt ~ atemp + hum + windspeed + weathersit, data = bike.sharing.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4269.2 -1066.3  -111.5  1083.0  4178.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3716.0      340.3  10.921  < 2e-16 ***
## atemp         7220.3      336.3  21.473  < 2e-16 ***
## hum          -1996.5      494.6  -4.036 6.01e-05 ***
## windspeed    -3857.4      719.5  -5.361 1.11e-07 ***
## weathersit    -464.6      125.6  -3.700 0.000232 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1410 on 726 degrees of freedom
## Multiple R-squared:  0.4731, Adjusted R-squared:  0.4702
## F-statistic:   163 on 4 and 726 DF,  p-value: < 2.2e-16
```

Rezultati upućuju na to da je ovako predstavljena varijabla season značajna u modelu, no model je vrlo

vjerojatno samo uhvatio efekt vrlo malih vrijednosti izlaza za zimu (season = 1) i ne može objasniti efekt smanjenih vrijednosti za jesen (season = 4) u odnosu na proljeće i ljeto.

Za predstavljanje kategorijskih varijabli kao ulaz regresijskog modela postoje različite tehnike, a jedna od najjednostavnijih i najčešće korištenih su tzv. dummy varijable. Svaka kategorija u kategorijskoj varijabli predstavljena je svojom vlastitom indikatorskom varijablom koja poprima vrijednost 1 u slučaju da originalna kategorijska varijabla poprima vrijednost te kategorije, a 0 inače. Jednostavno generiranje dummy varijabli dostupno je u paketu fastDummies.

```
require(fastDummies)
```

```
## Loading required package: fastDummies
```

```
bike.sharing.data.d = dummy_cols(bike.sharing.data,select_columns='weathersit')

#procjena modela s dummy varijablama
fit.multi.d = lm(cnt ~ atemp + hum + windspeed + holiday + weathersit_1 + weathersit_2, bike.sharing.da
summary(fit.multi.d)
```

```
##
## Call:
## lm(formula = cnt ~ atemp + hum + windspeed + holiday + weathersit_1 +
##     weathersit_2, data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4233.9 -1072.6   -94.9  1061.2  4184.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1395.9      582.1   2.398   0.0167 *
## atemp          7215.4      333.4  21.640  < 2e-16 ***
## hum           -2001.3      490.1  -4.083 4.94e-05 ***
## windspeed     -3592.5      716.7  -5.012 6.77e-07 ***
## holiday        -648.3      309.8  -2.093   0.0367 *
## weathersit_1   1791.9      351.0   5.105 4.23e-07 ***
## weathersit_2   1510.6      328.8   4.594 5.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1397 on 724 degrees of freedom
## Multiple R-squared:  0.4844, Adjusted R-squared:  0.4801
## F-statistic: 113.3 on 6 and 724 DF,  p-value: < 2.2e-16
```

Dummy varijable će uvijek biti linearno zavisne ako ih sve koristimo u regresijskim modelima (objašnjenje: ako znamo da vrijednost kategorijske varijable nije ni jedna od 3 kategorije, onda sigurno znamo da je 4. kategorija) - stoga je uvijek potrebno isključiti jednu od dummy varijabli iz modela. Bez obzira na to koje varijable uključili, ukupni model će biti isti, ali samo zaključivanje o pojedinim dummy varijablama u slučajevima kad imamo više od dvije kategorije će biti nešto kompliciranije.

Vremenski zavisne varijable

U nekim slučajevima znamo da izlazna varijabla ima i izraženu vremensku zavisnost - u konkretnom slučaju možemo biti sigurni da, ukoliko znamo današnji broj iznajmljenih bicikala, mala je vjerojatnost da će sutrašnji

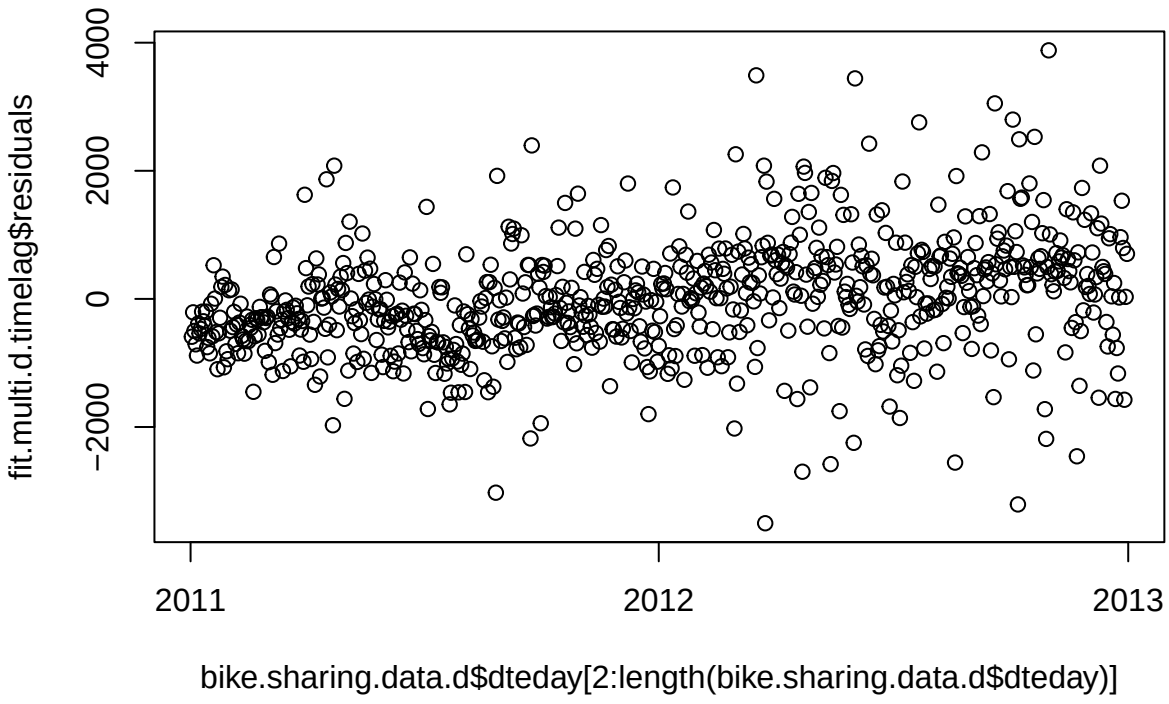
biti previše različit, čak i kad modeliramo efekte vremena, temperature itd. To je uostalom vidljivo i u grafičkim prikazima reziduala u ovisnosti u datumu za gore navedene modele.

Postoji jednostavan način na koji se neki od ovih vremenskih efekata mogu modelirati bez upotrebe složenijih modela - u regresiju se kao ulazna varijabla uključi vremenski pomaknuta izlazna varijabla koja označava "prošlost" koja je u trenutku modeliranje uglavnom poznata. Konkretno, uz sve navedene varijable, za modelirati varijablu cnt u trenutku t možemo uključiti i samu varijablu cnt u trenutku $t-1$.

```
#vremenski pomak varijable cnt
bike.sharing.data.d$lag.cnt = c(NA,bike.sharing.data.d$cnt[1:length(bike.sharing.data.d$cnt)-1])

#procjena modela s vremenski pomaknutom varijablom cnt na ulazu
fit.multi.d.timelag = lm(cnt ~ lag.cnt + atemp + hum + windspeed + holiday + weathersit_1 + weathersit_
summary(fit.multi.d.timelag)
```

```
##
## Call:
## lm(formula = cnt ~ lag.cnt + atemp + hum + windspeed + holiday +
##     weathersit_1 + weathersit_2, data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3502.5  -517.2   -20.6   498.6  3880.5
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.173e+02  3.761e+02  -0.844   0.3991
## lag.cnt       7.136e-01  2.212e-02  32.260  < 2e-16 ***
## atemp         1.856e+03  2.701e+02   6.870 1.39e-11 ***
## hum          -7.897e+02  3.158e+02  -2.501   0.0126 *
## windspeed    -2.170e+03  4.606e+02  -4.712 2.95e-06 ***
## holiday      -3.873e+02  1.983e+02  -1.953   0.0512 .
## weathersit_1  1.837e+03  2.245e+02   8.183 1.25e-15 ***
## weathersit_2  1.439e+03  2.104e+02   6.839 1.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 893.4 on 722 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7887, Adjusted R-squared:  0.7866
## F-statistic: 384.9 on 7 and 722 DF,  p-value: < 2.2e-16
```

```
plot(bike.sharing.data.d$dteday[2:length(bike.sharing.data.d$dteday)],fit.multi.d.timelag$residuals) #r
```



Transformacije podataka, dodavanje interkacijskih članova

U nekim situacijama, u svrhu izgradnje boljeg modela poželjno je nad ulaznim ili izlaznim varijablama primjeniti transformacije, najčešće $f(x) = \log x$ ili $f(x) = e^x$. Također, moguće je u model regresije dodavati tzv. interakcijske članove ili kvadrate, kubove, ... itd. ulaznih varijabli, npr. x_1^2, $x_1 x_2$, x_2^2.

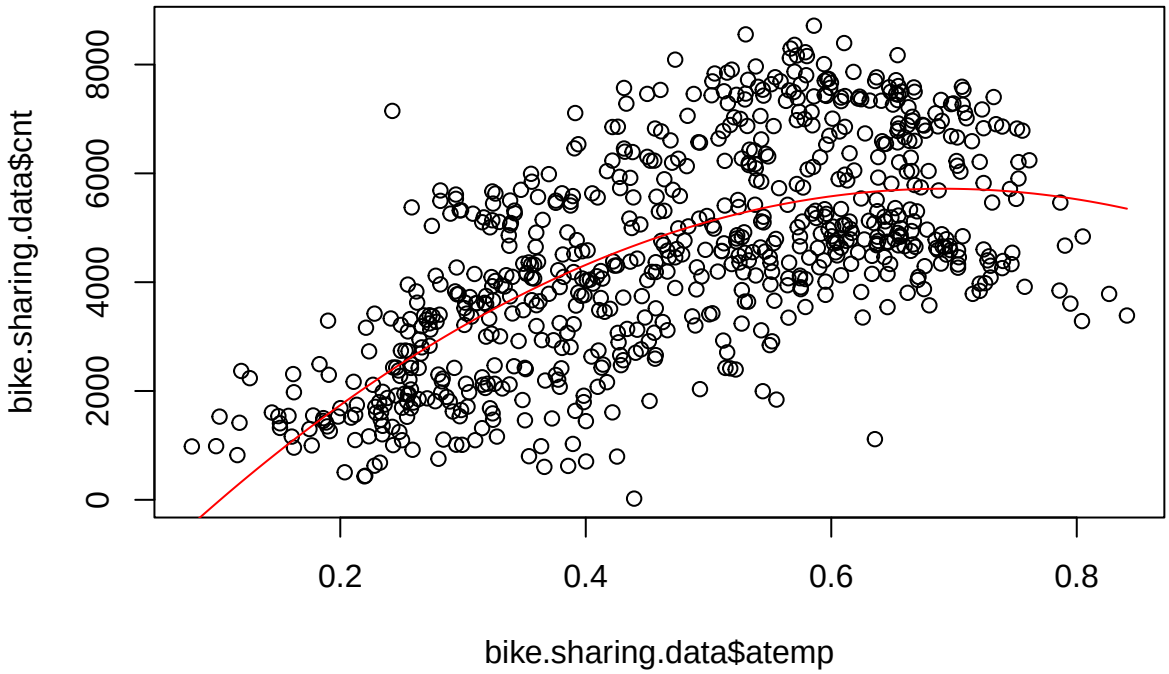
U oba slučaja modifikacije se primjenjuju na temelju pretpostavki o prirodi interakcije i modelu. Na primjeru temperature, u jednom od prvih grafova se mogao vidjeti potencijalno nelinearan efekt temperature - na najvišim temperaturama broj iznajmljenih bicikala se ipak smanjivao (što ima smisla).

```
# moguce je provjeriti gore navedenu tvrdnju prvo na primjeru samo temperature
fit.atemp.sq = lm(cnt ~ atemp + I(atemp^2),bike.sharing.data.d)
summary(fit.atemp.sq)
```

```
##
## Call:
## lm(formula = cnt ~ atemp + I(atemp^2), data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4648.8 -1042.4  -130.1  1148.7  4751.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2155        413  -5.218 2.36e-07 ***
```

```
## atemp          22767       1894  12.017  < 2e-16 ***
## I(atemp^2)    -16460       2012  -8.180 1.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1440 on 728 degrees of freedom
## Multiple R-squared:  0.4489, Adjusted R-squared:  0.4474
## F-statistic: 296.5 on 2 and 728 DF,  p-value: < 2.2e-16
```

```
#jednostavan nacin za graficki prikazati nelinearne krivulje
f = function(x, coeffs)
  return(coeffs[[1]] + coeffs[[2]] * x + coeffs[[3]] * x^2)
plot(bike.sharing.data$atemp,bike.sharing.data$cnt)
curve(f(x, fit.atemp.sq$coefficients), add = TRUE, col = "red")
```



Uključivanjem ovako transformiranih varijabli moguće je dodatno poboljšati ukupni model višestruke regresije.

```
#model regresije sa svim varijablama
fit.multi.d.timelag.sq = lm(cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed + holiday + weathersit
summary(fit.multi.d.timelag.sq)
```

```
##
## Call:
## lm(formula = cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed +
```

```
##     holiday + weathersit_1 + weathersit_2, data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3437.7  -533.6   -17.7   449.6  3610.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.747e+03  4.089e+02  -4.272 2.20e-05 ***
## lag.cnt       6.636e-01  2.232e-02  29.730  < 2e-16 ***
## atemp         1.135e+04  1.285e+03   8.830  < 2e-16 ***
## I(atemp^2)   -9.782e+03  1.297e+03  -7.543 1.39e-13 ***
## hum          -1.319e+03  3.122e+02  -4.225 2.70e-05 ***
## windspeed    -2.408e+03  4.448e+02  -5.412 8.49e-08 ***
## holiday      -3.569e+02  1.911e+02  -1.868   0.0622 .
## weathersit_1  1.824e+03  2.163e+02   8.432  < 2e-16 ***
## weathersit_2  1.434e+03  2.027e+02   7.077 3.51e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 860.7 on 721 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8041, Adjusted R-squared:  0.802
## F-statistic:   370 on 8 and 721 DF,  p-value: < 2.2e-16
```

Odabir parametara modela

U odabiru konačnog modela koji biste preporučili tvrtci za iznajmljivanje bicikala, potrebno je voditi se i principom jednostavnosti - jednostavniji model je uglavnom preferiraniji ukoliko je jednako dobar kao i neki alternativni složeniji model. Budući da će modeli s više varijabli u pravilu uvijek objašnjavati veći udio varijance od modela s manjim podskupom istih varijabli, nije moguće usporediti modela s različitim brojem varijabli gledajući samo njihove greške.

Pri odabiru modela u odnosu za velik broj razmatranih varijabli moguće je koristiti različite tehnike (tzv. model selection) koje nisu dio ovog kolegija. No, kao jedan od jednostavnijih alata za usporedbu modela različitih broja parametara moguće je koristiti i prilagođeni koeficijent determinacije R^2_{adj}, koji penalizira dodatne parametre u modelu.

U ovom slučaju, varijabla holiday potencijalno nije toliko korisna u modelu i možda se može izbaciti.

```
#model s varijablom holiday
fit.multi.d.timelag.sq = lm(cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed + holiday + weathersit
summary(fit.multi.d.timelag.sq)
```

```
##
## Call:
## lm(formula = cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed +
##     holiday + weathersit_1 + weathersit_2, data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3437.7  -533.6   -17.7   449.6  3610.1
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.747e+03  4.089e+02  -4.272 2.20e-05 ***
## lag.cnt       6.636e-01  2.232e-02  29.730  < 2e-16 ***
## atemp         1.135e+04  1.285e+03   8.830  < 2e-16 ***
## I(atemp^2)   -9.782e+03  1.297e+03  -7.543 1.39e-13 ***
## hum          -1.319e+03  3.122e+02  -4.225 2.70e-05 ***
## windspeed    -2.408e+03  4.448e+02  -5.412 8.49e-08 ***
## holiday      -3.569e+02  1.911e+02  -1.868   0.0622 .
## weathersit_1  1.824e+03  2.163e+02   8.432  < 2e-16 ***
## weathersit_2  1.434e+03  2.027e+02   7.077 3.51e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 860.7 on 721 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8041, Adjusted R-squared:  0.802
## F-statistic:   370 on 8 and 721 DF,  p-value: < 2.2e-16
```

```
#model bez varijable holiday
fit.multi.d.timelag.sq.final = lm(cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed + weathersit_1 +
summary(fit.multi.d.timelag.sq.final)
```

```
##
## Call:
## lm(formula = cnt ~ lag.cnt + atemp + I(atemp^2) + hum + windspeed +
##     weathersit_1 + weathersit_2, data = bike.sharing.data.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3434.3  -533.2   -18.9   452.4  3622.5
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.755e+03  4.095e+02  -4.284 2.08e-05 ***
## lag.cnt       6.650e-01  2.235e-02  29.757  < 2e-16 ***
## atemp         1.140e+04  1.287e+03   8.858  < 2e-16 ***
## I(atemp^2)   -9.833e+03  1.299e+03  -7.571 1.14e-13 ***
## hum          -1.329e+03  3.127e+02  -4.249 2.43e-05 ***
## windspeed    -2.412e+03  4.456e+02  -5.412 8.48e-08 ***
## weathersit_1  1.808e+03  2.165e+02   8.351 3.44e-16 ***
## weathersit_2  1.423e+03  2.029e+02   7.013 5.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 862.2 on 722 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8032, Adjusted R-squared:  0.8013
## F-statistic: 420.9 on 7 and 722 DF,  p-value: < 2.2e-16
```

Rezultati upućuju na to da varijabla holiday ipak daje određenu korisnu informaciju u modelu, čak i kad koristimo R^2_{adj}.

Druga često korištena metoda je jednostavno izbacivanje onih regresora koji nemaju značajne koeficijente - no zbog interakcija među regresorima u multivarijatnoj regresiji to nije uvijek pouzdana metoda. Također, u slučaju jako velikog broja varijabli se mogu javiti i problemi ponovoljenih usporedbi (višestrukog testiranja).

Zaključak

Konačan model sadržava relevantne varijable koje objašnjavaju čak preko 80% varijance broja iznajmljenih bicikala dnevno. Osim metričkih varijabli temperature zraka, vlažnosti i vjetra, uključen je i kvadrat temperature zraka (zbog nelinearnog efekta), kategorijska varijabla koja ukazuje na praznike, dummy varijable za kategoriju vremenske situacije, te prethodna ("jučerašnja") vrijednost broja iznajmljenih bicikala.

Sve navedene varijable osim holiday su značajne na razini 0.01, kao i sam model, na što upućuju rezultati t-testova pojedinih koeficijenata i F-testa čitavog modela.