Domača naloga 1

Linearni modeli

Alen Kahteran

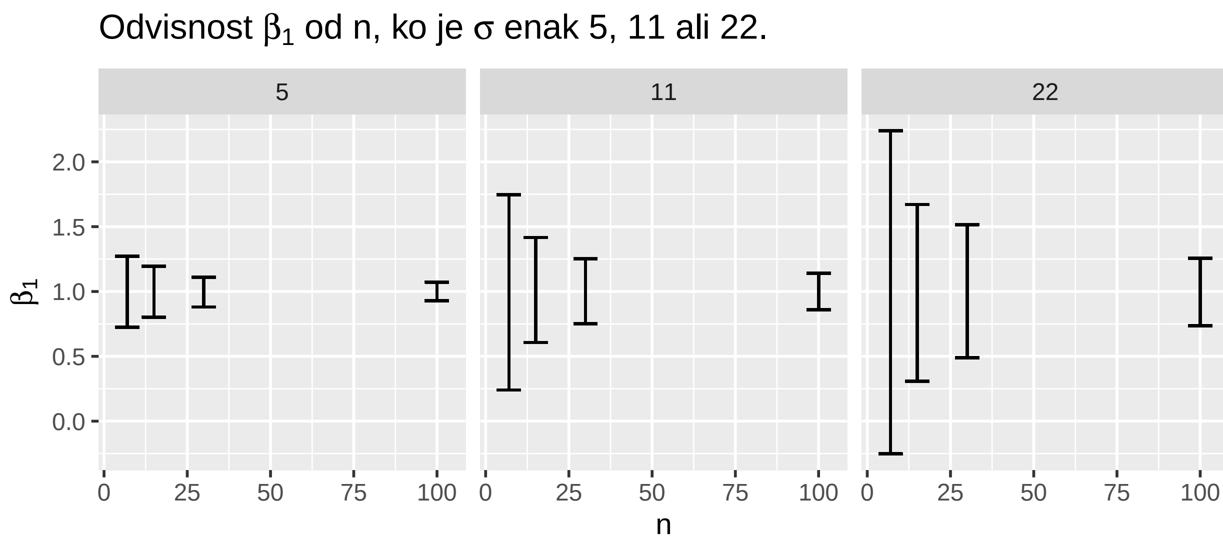
22 11 2020

Simulacija za enostavno linearno regresijo

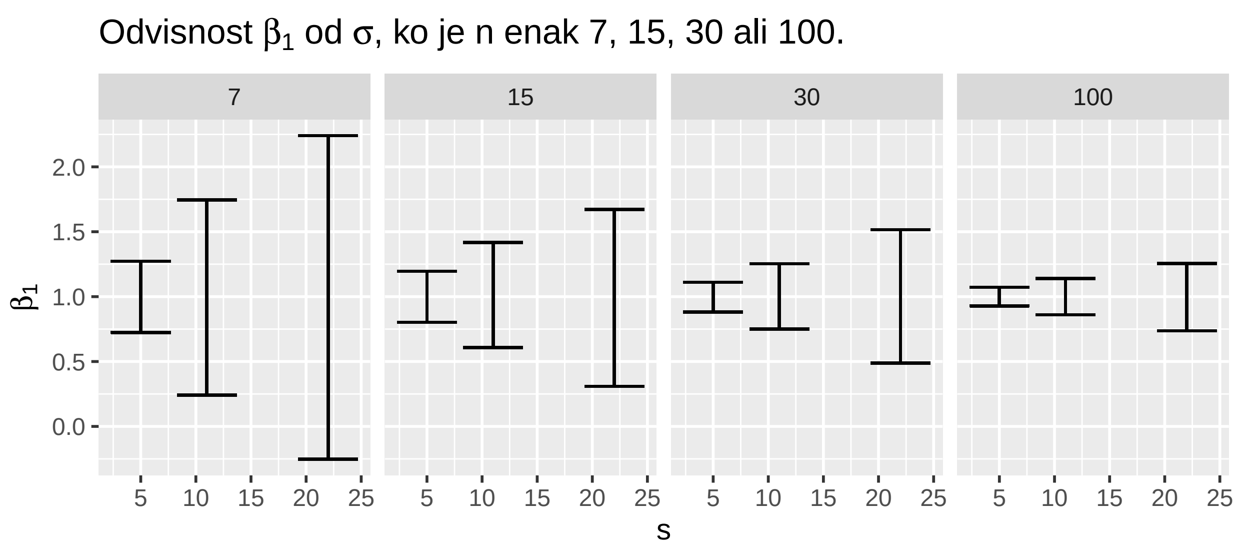
```r
# dolocitev beta0, beta1 in stevilo simulacij
b_0 <- 100
b_1 <- 1
N_sim <- 1000
# dolocitev n in sigma
n_sizes <- c(7, 15, 30, 100)
s_sizes <- c(5, 11, 22)

# vse mozne kombinacije
combinations <- expand.grid(n_sizes, s_sizes)
# preimenujemo
names(combinations) <- c("n", "s")
# list za zhranjevanje
storage <- list()

# funkcija
reg.sim <- function(x, beta0, beta1, sigma, Nsim) {
    # pripravimo prazne vektorje za rezultate simulacij, cenilki parametrov b0 in b1,
    # p-vrednost za testiranje domneve beta1=0,
    # spodnjo in zgornjo mejo intervala zaupanja za beta1
    b0 <- numeric(Nsim)
    b1 <- numeric(Nsim)
    p <- numeric(Nsim)
    sp.meja.b1 <- numeric(Nsim)
    zg.meja.b1 <- numeric(Nsim)
    n <- length(x)

    for (i in 1:Nsim) {
        epsilon <- rnorm(n, mean=0, sd=sigma)
        y <- beta0 + beta1 * x + epsilon
        mod <- lm(y ~ x)
        b0[i] <- coef(mod)[1]
        b1[i] <- coef(mod)[2]
        p[i] <- coefficients(summary(mod))[2, 4]
        sp.meja.b1[i] <- confint(mod)[2, 1]
        zg.meja.b1[i] <- confint(mod)[2, 2]
    }
```

```r
  return(data.frame(b0, b1, p, sp.meja.b1, zg.meja.b1))
}

# dolocimo seme za ponovljivost
set.seed(8)

# zanka cez vse n-je in sigme
for (row in 1:nrow(combinations)){
    # za lepsi izpis
    n_ <- combinations[row, "n"]
    s_ <- combinations[row, "s"]
    # shranimo v storage
    storage[paste0("n", n_, "s", s_)] <- list(reg.sim(sample(15:70, n_, replace=TRUE),
                                                     b_0,
                                                     b_1,
                                                     s_,
                                                     N_sim) %>%
                                                 # dodamo n in s vrednosti za izris
                                                 mutate(n = n_,
                                                        s = s_))
}

# pretvorimo v korekten format
full_data <- unnest(tibble(storage), cols = c(storage))

# izracunamo stvari po skupinah za izris
full_data <- full_data %>%
    group_by(n, s) %>%
    mutate(mean_b1 = mean(b1),
           mean_sp_b1 = mean(sp.meja.b1),
           mean_zg_b1 = mean(zg.meja.b1),
           test_power = 1-sum(p>0.05)/N_sim) %>%
    ungroup()
```

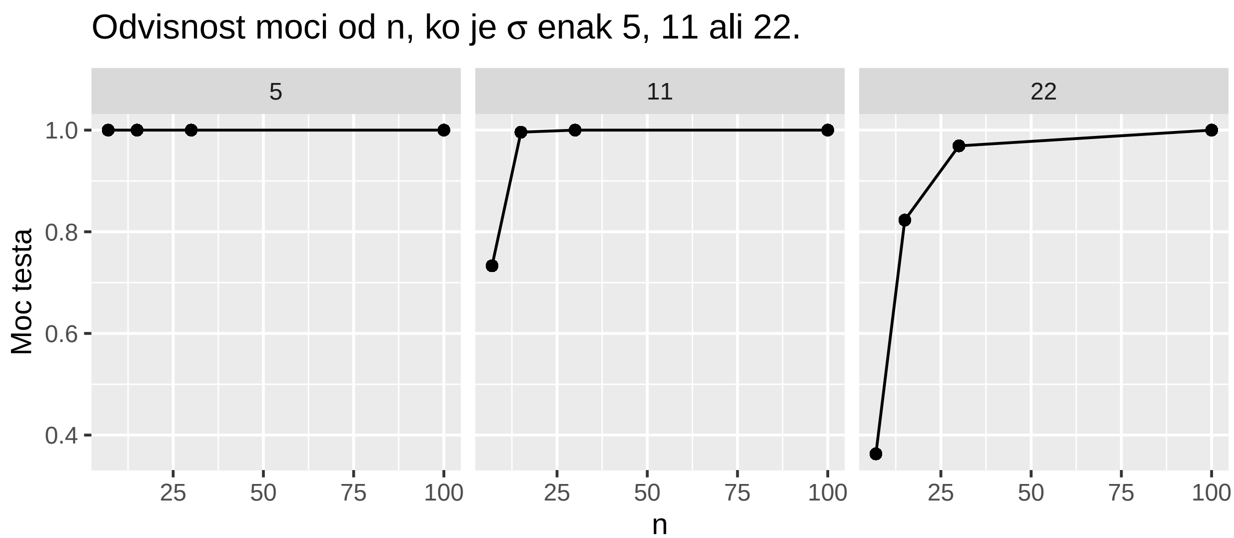
```r
# izris beta1 od n, za razlicne sigma
ggplot(full_data, aes(x=n, y=mean_b1)) +
    geom_errorbar(aes(ymin=mean_sp_b1, ymax=mean_zg_b1)) +
    facet_grid(cols=vars(s)) +
    labs(y = TeX("$\\beta_1$"),
         title = TeX("Odvisnost $\\beta_1$ od $n$, ko je $\\sigma$ enak 5, 11 ali 22."))
```



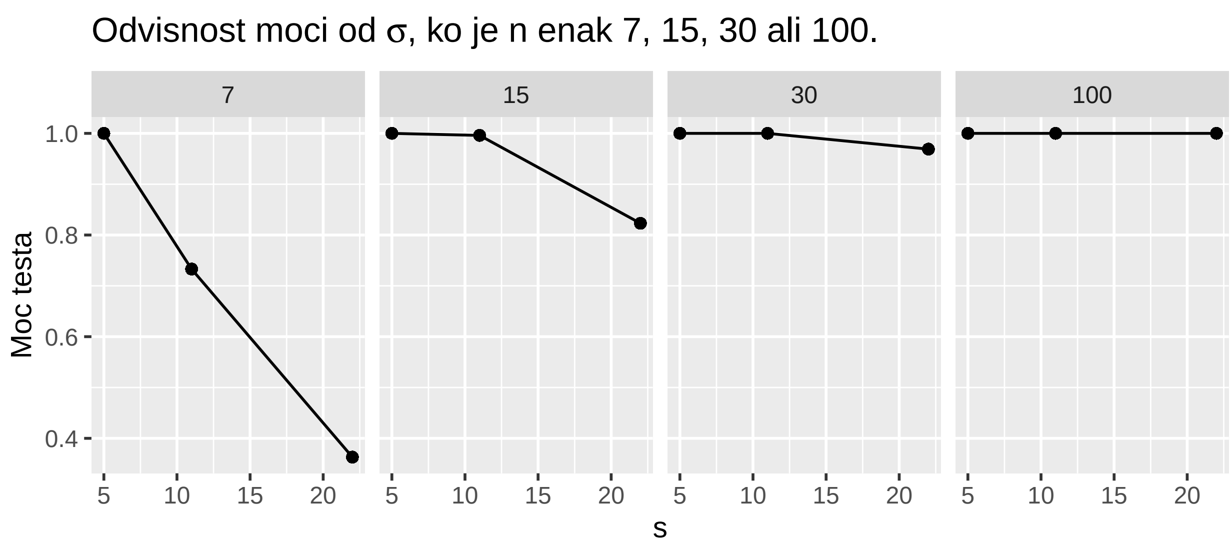

```
# izris beta1 od sigma, za razlicne n
ggplot(full_data, aes(x=s, y=mean_b1)) +
    geom_errorbar(aes(ymin=mean_sp_b1, ymax=mean_zg_b1)) +
    facet_grid(cols=vars(n)) +
    labs(y = TeX("$\\beta_1$"),
             title = TeX("Odvisnost $\\beta_1$ od $\\sigma$, ko je $n$ enak 7, 15, 30 ali 100."))
```




```
# izris moci testa od n, za razlicne sigma
ggplot(full_data, aes(x=n, y=test_power)) +
    geom_line() +
    geom_point() +
    facet_grid(cols=vars(s))  +
    labs(y = "Moč testa",
        title = TeX("Odvisnost moci od $n$, ko je $\\sigma$ enak 5, 11 ali 22."))
```




```r
# izris moci testa od sigma, za razlicne n
ggplot(full_data, aes(x=s, y=test_power)) +
    geom_line() +
    geom_point() +
    facet_grid(cols=vars(n))  +
    labs(y = "Moč testa",
         title = TeX("Odvisnost moci od $\\sigma$, ko je $n$ enak 7, 15, 30 ali 100."))
```



Iz prvega grafa je razvidno, da ko večamo velikost vzorca se pri enakem σ interval zaupanja manjša, kar je smiselno, saj iz več podatkov bolje napovemo β_1. Vidno je tudi da je interval zaupanja glede na σ večji, saj σ predstavlja kolikšen bo mogoč odklon pri izračunu. Kar je pravzaprav še bolje vidno na drugem grafu, kjer je n fiksiran. Poleg tega je videti da se za večje n vrednosti interval zaupanja manjša, kar smo videli tudi že na prvem grafu.

Pri grafih moči je videti, da se zna kdaj zgoditi, da pri večjih *sigma* in pri manjših n-jih, da ničelne hipoteze ne zavrnemo. To je zelo dobro razvidno iz zadnjega grafa, kjer je videti da se pri majhnem n večkrat zgodi da ničelne hipoteze ne zavrnemo. Poleg tega je videti da pri majhnih n vrednostih je pomembno da je tudi σ majhen, če želimo ničelno hipotezo zavrniti, kar pri velikih n-jih ne velja, ker je praktično vseeno kolikšen je σ (za naše primere).