Lab1

Jakub Kucinski

3/22/2022

Task 1

The results of experiments conducted for task 1 can be found in table below:

```
n = 1000
X = matrix(rnorm(950000, 0, 1.0/sqrt(1000)), n, 950)
e = rnorm(n)
beta = c(3, 3, 3, 3, 3, rep(0, 945))
Y = X %*% beta + e
significance = 0.1
nvars = c(10, 100, 500, 950)
nmodels = length(nvars)
n_experiments = 500

task1 = data.frame(matrix(ncol=10, nrow=0))
colnames(task1) = c("ncols", "nsignificant", "avgstd", "avglength", "truedisc", "falsedisc", "truebonf"

for (i in 1:nmodels){
  k = nvars[i]
  Xi = X[, 1:k]
  reg = lm(Y~Xi -1, x = TRUE)

  betahat = reg$coefficients
  betastds = sqrt(diag(solve(t(reg$x) %*% reg$x)))
  statistic = betahat / betastds
  pvals = 2*(1-pnorm(abs(statistic)))

  n_significant = sum(pvals < significance)

  intervallength = 2*qnorm(1-significance/2)*betastds
  avgstd = sum(betastds[1:k]) / k
  avglength = sum(intervallength) / k

  truedisc = sum(pvals[1:5] < significance)
  falsedisc = sum(pvals[6:k] < significance)
  truebonf = sum(pvals[1:5] < significance/k)
  falsebonf = sum(pvals[6:k] < significance/k)

  sorted_pvals = sort(pvals[1:k] , index.return=TRUE)
  significance_levels = rep(significance/k, k) * seq(1, k)
  largest_i = 0
```

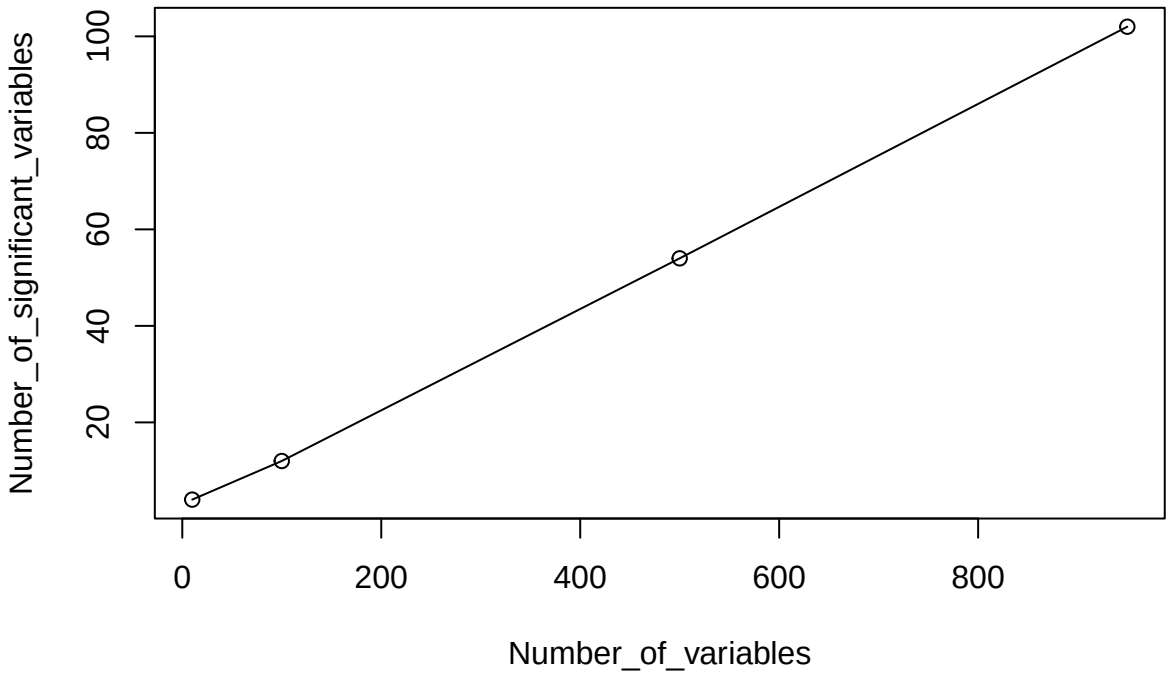
```
  for (idx in k:1){
    if (sorted_pvals$x[idx] < significance_levels[idx]){
      largest_i = idx
      break
    }
  }
  truebh = sum(sorted_pvals$ix[1:largest_i] <= 5)
  falsebh = sum(sorted_pvals$ix[1:largest_i] > 5)
  p_BH = p.adjust(pvals, method="BH")
  truebh = sum(p_BH[1:5] <= significance)
  falsebh = sum(p_BH[6:k] <= significance)

  task1[nrow(task1) + 1,] = c(k, n_significant, avgstd, avglength, truedisc, falsedisc, truebonf, falseb
}
task1
```

```
##   ncols nsignificant   avgstd avglength truedisc falsedisc truebonf falsebonf
## 1    10            4 1.006105  3.309791        4         0        3         0
## 2   100           12 1.054198  3.468003        4         8        1         0
## 3   500           54 1.410895  4.641431        3        51        0         0
## 4   950          102 4.576202 15.054364        0       102        0         0
##   truebh falsebh
## 1      4       0
## 2      1       0
## 3      0       0
## 4      0       0
```

a)

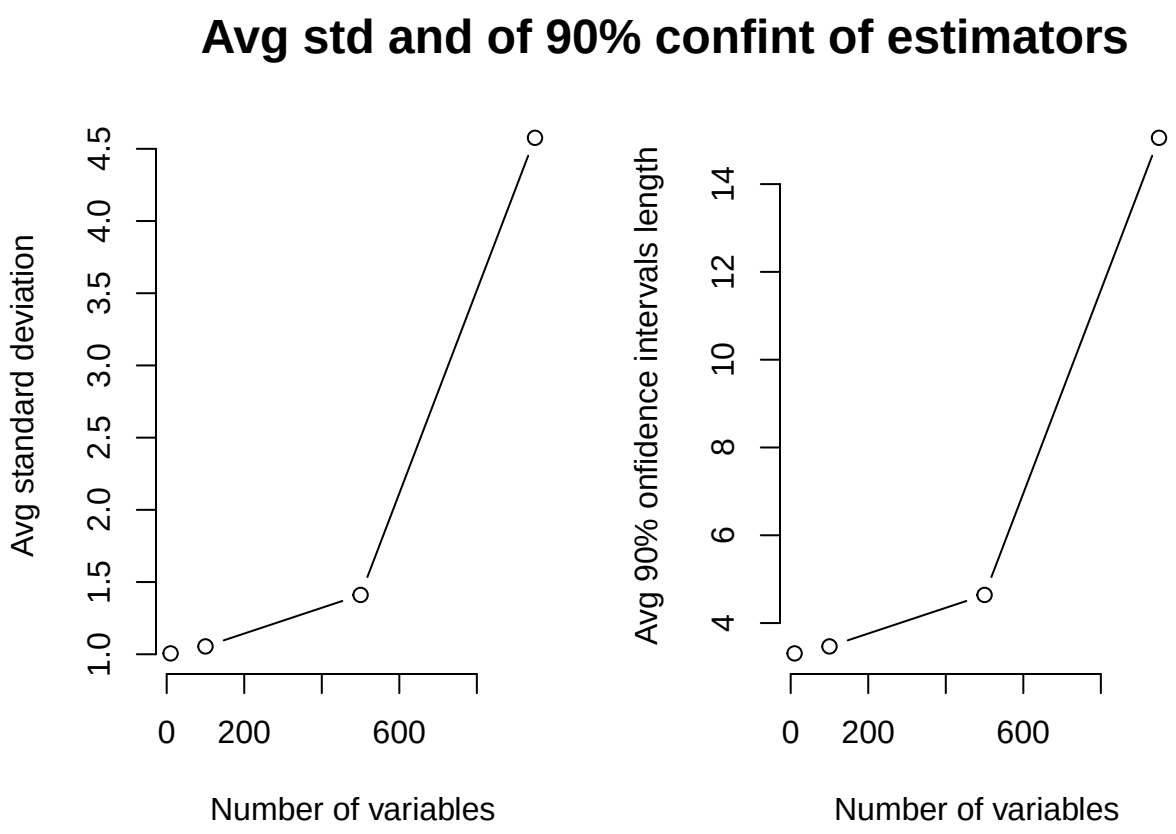
Below we can find a visualization of number number of found signifant variables with respect to number of



variables in the model.
From how the data was generated we know that we have exactly 5 true non-zero coefficients (5 first ones). Column "nsignificant" in the table corresponds to number of coefficients which we found significant after performing test statistic. We can see that after testing significance of model coefficients for model with 10 variables we are getting more or less 5 significant variables (which is on par with number of true non-zero coefficients). For greater models we get more significant variables - around 10% of number of variables, as all except first 5 are true zeros.

b)

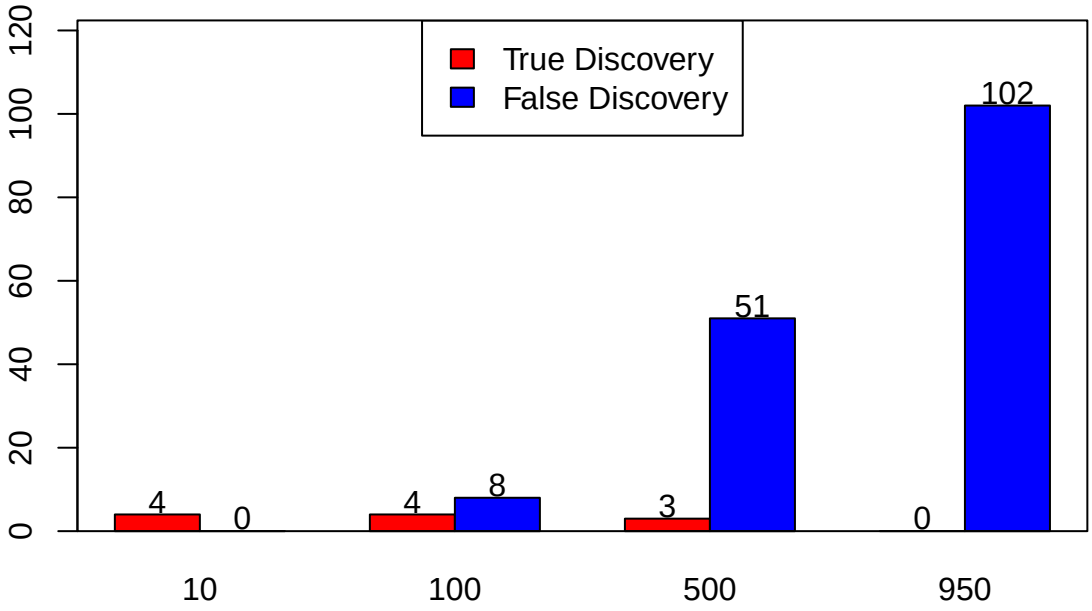
Plots below show the relations of average standard deviations and average lengths of the respective 90% confidence intervals of the estimators of individual regression coefficients compared to the number of variables in the model.



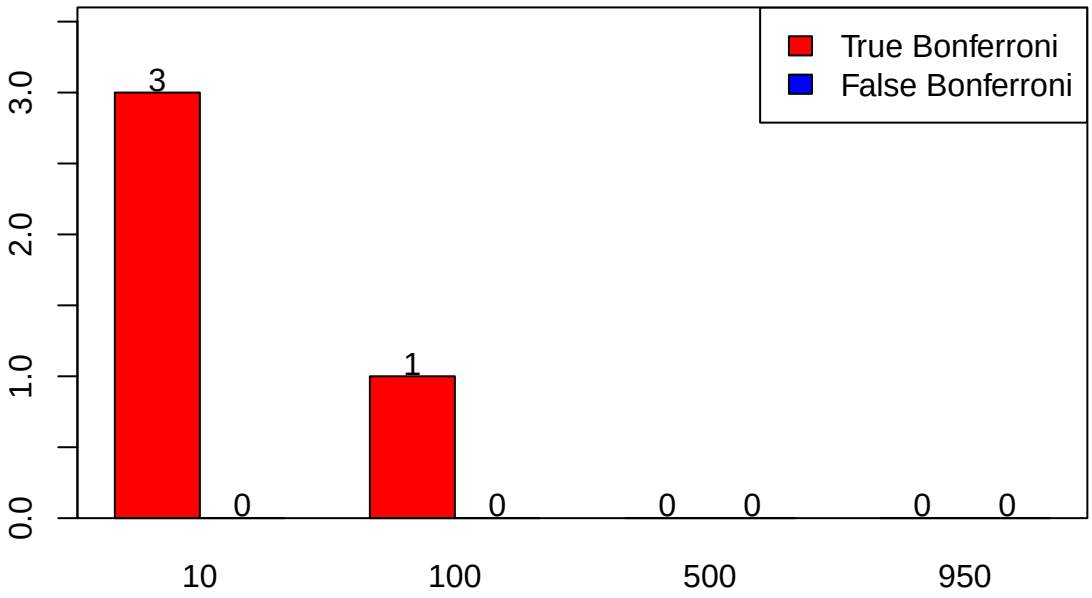
From table we can see, that average standard deviation of the estimators of individual regression coefficients increases with number of variables in the model. While for model with 10 and 100 coefficients it is around 1.0, for the model with 500 the uncertainty of coefficients estimators increases to around 1.4 and for 950 variables it becomes much larger - more than 4. That means that model becomes very uncertain about the coefficients values'. This make sense, as we have only 1000 data samples, so very little more than the number of variables, so the signal becomes very low. The same conclusions stands for average length of the respective 90% confidence intervals, as they are just the standard deviations multiplied by some positive constant.

c)

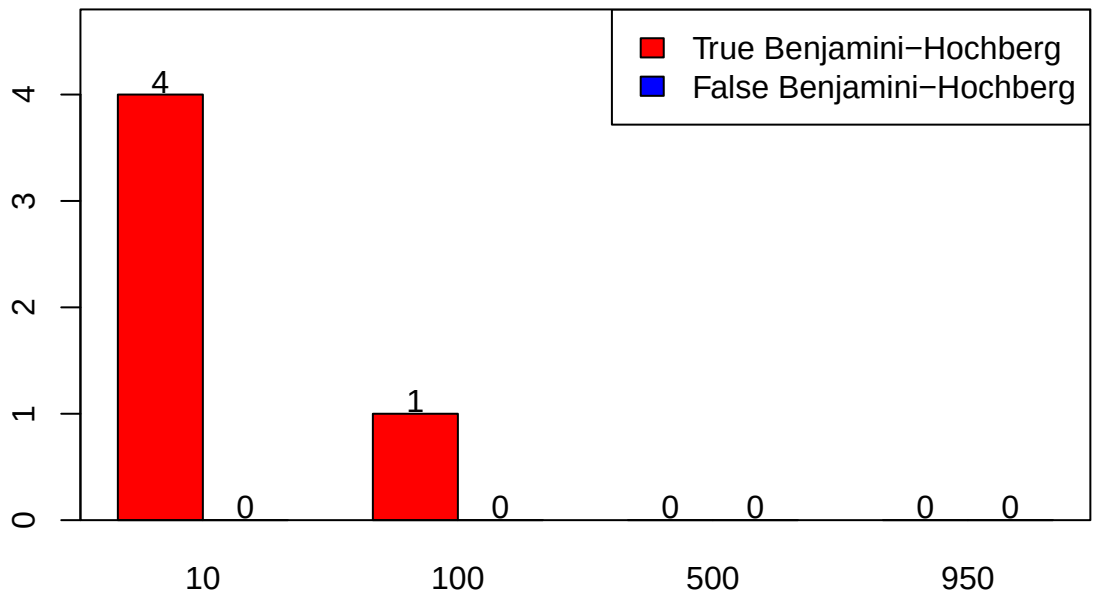
Plots below show numbers of true and false discoveries for different models sizes without adjusting for multiple testing, using Bonferroni correction and using Benjamini-Hochberg correction.



We can see that without adjusting for multiple testing the number of true discoveries decreases with number of variables and number of false discoveries increases (more or less 10% of model variables) as for about 10% of true zero coefficient the null hypothesis should be rejected under significance level 0.1.



Using Bonferroni correction the number of false discoveries is close to zero for all models as it is very restrictive and rejects the null hypothesis if p-value is smaller than significance level divided by number of variables. Number of true positives decreases with model size.



For models sing Benjamini-Hochberg correction the number of false discoveries is close to zero for all models as it is quite restrictive. Number of true positives decreases with model size.

Task 2

```
fdr = function (truedisc, falsedisc){
  all_disc = truedisc + falsedisc
  all_disc[all_disc == 0] = 1
  fdp = falsedisc / all_disc
  mean(fdp)
}

fwer = function (truedisc, falsedisc){
  if_found_falsedisc = falsedisc > 0
  mean(if_found_falsedisc)
}

power = function (true_beta, true_sigma_beta, significance){
  c = qnorm(1-significance/2)
  1 - pnorm(c - true_beta / true_sigma_beta) + pnorm(- c - true_beta / true_sigma_beta)
}
```

```
n = 1000
X = matrix(rnorm(950000, 0, 1.0/sqrt(1000)), n, 950)
e = rnorm(n)
```

```
beta = c(3, 3, 3, 3, 3, rep(0, 945))
Y = X %*% beta + e
significance = 0.1
nvars = c(10, 100, 500, 950)
nmodels = length(nvars)
n_experiments = 500

fwerfdr = data.frame(matrix(ncol=13, nrow=0))
colnames(fwerfdr) = c("ncols", "TD", "FD", "FWER", "FDR", "TD_BF", "FD_BF", "FWER_BF", "FDR_BF", "TD_BH

betas_variances =  matrix(0, nmodels, 950)
intervals = matrix(0, nmodels, 950)

for (i in 1:nmodels){
  k = nvars[i]
  betas = matrix(0, k, n_experiments)
  truedisc = matrix(0, n_experiments)
  falsedisc = matrix(0, n_experiments)
  truebonf = matrix(0, n_experiments)
  falsebonf = matrix(0, n_experiments)
  truebh = matrix(0, n_experiments)
  falsebh = matrix(0, n_experiments)
  # total_true_discoveries = 0
  # total_false_discoveries = 0
  # total_true_bonf = 0
  # total_false_discoveries = 0
  # total_true_discoveries = 0
  # total_false_discoveries = 0
  for (i_exp in 1:n_experiments){
    e = rnorm(n)
    Y = X %*% beta + e

    pvals = matrix(0, k)
    intervallength = matrix(0, k)

    Xi = X[, 1:k]
    reg = lm(Y~Xi -1, x = TRUE)

    betahat = reg$coefficients
    betas[1:k, i_exp] = betahat

    betastds = sqrt(diag(solve(t(reg$x) %*% reg$x)))
    statistic = betahat / betastds
    pvals[1:k] = 2*(1-pnorm(abs(statistic)))


    truedisc[i_exp] = sum(pvals[1:5] < significance)
    falsedisc[i_exp] = sum(pvals[6:k] < significance)
    truebonf[i_exp] = sum(pvals[1:5] < significance/k)
    falsebonf[i_exp] = sum(pvals[6:k] < significance/k)

    p_BH = p.adjust(pvals[1:k], method="BH")
    truebh[i_exp] = sum(p_BH[1:5] <= significance)
```

```
    falsebh[i_exp] = sum(p_BH[6:k] <= significance)
  }
  fwerfdr[nrow(fwerfdr) + 1,] = c(k, mean(truedisc), mean(falsedisc), fwer(truedisc, falsedisc), fdr(tru
                                  mean(truebonf), mean(falsebonf), fwer(truebonf, falsebonf), fdr(trueb
                                  mean(truebh), mean(falsebh), fwer(truebh, falsebh), fdr(truebh, falseb
  variances = apply(betas, 1, var)
  betas_variances[i, 1:k] = variances
  intervallength = 2*qnorm(1-significance/2)*sqrt(variances)
  intervals[i, 1:k] = intervallength
}
fwerfdr
```

```
##   ncols    TD     FD  FWER        FDR TD_BF FD_BF FWER_BF      FDR_BF TD_BH
## 1    10 4.596  0.466 0.386 0.08087619 3.270 0.040   0.040 0.009133333 4.158
## 2   100 4.372  9.572 1.000 0.67284310 1.742 0.094   0.090 0.041966667 2.342
## 3   500 3.428 49.768 1.000 0.93424938 0.256 0.134   0.122 0.108333333 0.346
## 4   950 0.834 94.106 1.000 0.99086313 0.002 0.082   0.062 0.062000000 0.008
##   FD_BH FWER_BH     FDR_BH
## 1 0.268   0.240 0.04964762
## 2 0.390   0.284 0.09968095
## 3 0.244   0.186 0.13633333
## 4 0.224   0.080 0.07885714
```

```
theoretical_df = data.frame(matrix(ncol=9, nrow=0))
colnames(theoretical_df) = c("ncols", "avg_variances", "theoretical_vars", "avg_intervals", "theoretical
                             "theoretical_mFDR", "theoretical_mFDR_BF")
for (i in 1:nmodels){
  k = nvars[i]
  avg_variances = sum(betas_variances[i, 1:k]) / k
  avg_intervals = sum(intervals[i, 1:k]) / k
  e = rnorm(n)
  Y = X %*% beta + e
  Xi = X[, 1:k]
  reg = lm(Y~Xi -1, x = TRUE)
  theoretical_betas_vars = diag(solve(t(reg$x) %*% reg$x))
  theoretical_vars = mean(theoretical_betas_vars)
  theoretical_interval = 2*qnorm(1-significance/2)*mean(sqrt(theoretical_betas_vars))
  theoretical_FWER = 1 - (1-significance)^(k-5)
  theoretical_FWER_BF = 1 - (1-significance/k)^(k-5)

  theoretical_mFDR = (significance * (k-5)) / (significance * (k-5) + sum(power(3, sqrt(theoretical_beta
  theoretical_mFDR_BF = (significance/k * (k-5)) / (significance/k * (k-5) + sum(power(3, sqrt(theoretic

  theoretical_df[nrow(theoretical_df) + 1,] = c(k, avg_variances, theoretical_vars, avg_intervals, theor
                                                theoretical_mFDR, theoretical_mFDR_BF)

}
theoretical_df
```

```
##   ncols avg_variances theoretical_vars avg_intervals theoretical_intervals
## 1    10      1.007491         1.009170      3.301163              3.304528
## 2   100      1.121469         1.109822      3.481074              3.464784
## 3   500      1.998479         1.991408      4.645933              4.640178
```

```
## 4   950     19.768880        19.807260     14.550645             14.570719
##   theoretical_FWER theoretical_FWER_BF theoretical_mFDR theoretical_mFDR_BF
## 1         0.409510          0.04900995       0.09880036          0.01483551
## 2         0.999955          0.09067029       0.68105114          0.05326795
## 3         1.000000          0.09426626       0.93507578          0.25928858
## 4         1.000000          0.09469097       0.99095949          0.96839379
```

a)

From table above we can see, that the average variance of the estimators of individual regression coefficients is (as expected) very close to the theoretical estimate (avg_variances and theoretical_vars).

b)

From table above we can see, that he average length of the 90% confidence interval is (as expected) very close to the theoretical estimate (avg_intervals and theoretical_intervals).

c)

In first of the tables in task 2 we can see the average number of true and false discoveries and the estimations of FWER and FDR for models of different sizes without adjusting for multiple testing, using Bonferroni correction and using Benjamini-Hochberg correction. The average number of TD decreases with growing model size and for Bonferroni and Benjamini-Hochberg corrections becomes almost zero for model with 950 variables. Number of FD without adjusting for multiple testing is around 10% of no. true zero variables. For Bonferroni and Benjamini-Hochberg corrections is below 0.5 for all model sizes. Estimated FWER and FDR are very high when not using adjusting. When using corrections they become much smaller. The theoretical values of FWER and FDR can be found in last table. The theoretical values of FWER are close to ones estimated. Calculating theoretical values for FDR is difficult so instead I calculated theoretical mFDR. It was a little bit bigger than the estimated FDR without adjusting. When using Bonferroni correction differences between theoretical mFDR and estimated FDR become large.