List 1 - Multiple Regression and Multiple Testing

Statistical Learning

Paulina Podgórska

Let us generate the design matrix $X_{n\times p} = X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Then we generate the vector of values of the response variable $Y = X\beta + \epsilon$, where $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$ and $\epsilon \sim N(0, I)$.
We will perfmorm analysis using 4 models containing only:

- first 10 columns of X (p = 10)
- first 100 columns of X (p = 100)
- first 500 columns of X (p = 500)
- all 950 columns (p = 950)

Least square estimator of β

The formula for least square estimator of β is:

$$\hat{\beta}_{LS} = argmin_{\beta\in\mathbb{R}^1}||Y - X\beta||^2 = (X'X)^{-1}X'Y.$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^3(X'X)^{-1})$$

Table 1: First 10 least squares estimators for each model

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
X_{10}	3.242	2.344	3.141	3.831	2.314	0.873	0.776	-0.327	0.424	-0.339
X_{100}	3.384	2.613	3.040	3.504	2.301	1.282	1.381	-0.178	0.193	-0.021
X_{500}	4.900	4.644	3.550	3.909	2.836	2.017	0.124	-0.912	0.458	0.267
X_{950}	8.186	0.653	0.068	5.251	1.503	5.036	6.066	4.389	3.175	4.069

Next step is performing tests of significance of individual regression coefficients at the significance level of $\alpha = 0.1$.

$H_{0i} : \beta_i = 0$ $H_{1i} : \beta_i \neq 0$

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)},$$

where $s^2(\hat{\beta}_i) = s^2(X'X)^{-1}[i, i]$. In order to perform tests in R we compare p-values to our α value.

Table 2: p-values for 10 first columns

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
X_{10}	0.001	0.021	0.002	0	0.029	0.410	0.44	0.748	0.679	0.748
X_{100}	0.001	0.015	0.004	0.001	0.043	0.248	0.194	0.869	0.857	0.985
X_{500}	0	0.001	0.016	0.004	0.05	0.155	0.928	0.516	0.749	0.855
X_{950}	0.025	0.858	0.985	0.133	0.743	0.197	0.077	0.288	0.460	0.397

As we can see, for the first four models, the correct coefficients tested as significant. Model with 950 columns performs much worse.

Standard deviation and length of confidence intervals

Let's examine the average standard deviation and the average length of 90% confidence intervals of our estimators.

Table 3: Average standard deviation and average length of 90% confidence intervals

	X_{10}	X_{100}	X_{500}	X_{950}
Avg. sd	1.019	1.074	1.422	4.054
Avg. 90% CI	3.355	3.538	4.685	13.588

Above values increase with the number of variables in a model. The model with 950 columns performs significantly worse.

True and false discoveries

The next step is to examine the number of false and true discoveries for different models.

Bonferroni correction Its purpose is to minimize the number of Type I errors. It consists in reducing the nominal significance level of each of the set of related tests in direct proportion to their total number. We reject H_0 if $p_i \leqslant \frac{\alpha}{p}$, for $i = 1, \ldots, p$.

Benjamini - Hochberg correction It consists in sorting the p_i values in descending order and finding the i_0 index such that $i_0 = \max\{i : p_i \leqslant \frac{i}{n}\alpha\}$. Then we discard $H_{(i)} : i \leqslant i_0$.

Table 4: True and false discoveries for different models

	X_{10}	X_{100}	X_{500}	X_{950}
TD	5	5	5	1
FD	1	7	60	129
TD - Bonf.	3	1	0	0
FD - Bonf	0	0	0	0
TD - B.H.	5	2	0	0
FD - B.H.	0	0	0	0

The greater the number of variables, the fewer true discoveries are detected and the more false discoveries. After applying the Bonferroni correction and the Benjamini-Hochberg correction, the number of false discoveries is zero. There are fewer true discoveries than in the case of no correction.

500 repeats

In this section we will repeat above experiments 500 times on the basis of which we will draw conclusions about the estimators.

The average variance of the estimators

Inverse Wishart distribution As we know, the formula for variance of the estimators of regression coefficients is as follows:

$$s^2(\hat{\beta}_i) = s^2(X'X)^{-1}[i,i].$$

If elements of X are iid from $N(0, \frac{1}{\sqrt{n}})$, then $X'X$ has a Wishart distribution.

$$X'X \sim W_p(n, diag(\frac{1}{n}, ..., \frac{1}{n})),$$

$$(X'X)^{-1} \sim W_p^{-1}(n, diag(n, ...n)).$$

The $(X'X)^{-1}$ has the inverse Wishart distribution. The expected values on the diagonal are equal to $\frac{n}{n-p-1}$. Therefore, in our case the theoritical value of variance of the estimators can be calculated with $\hat{\sigma}^2 = \frac{n}{n-p-1}$.

Table 5: Average variance and the theoritical value

	X_{10}	X_{100}	X_{500}	X_{950}
Avg. variance	1.007	1.106	1.988	19.752
Theoretical value	1.011	1.112	2.004	20.408

We can observe several times higher average variance for the model with p = 950. We can conclude from this that in the case of a very large number of variables (>500) the estimation of parameters is more difficult. In this situation, the parameter estimators are highly varied and dispersed around the mean. Our calculated values are close to the theoretical ones.

The average length of the 90% interval

$$P_{H_0}(-q_{t(1-\alpha/2)}(n-p) \leqslant t_i \leqslant q_{t(1-\alpha/2)}(n-p)) = 1-\alpha$$

$$\hat{\beta}_i - q \cdot s(\hat{\beta}_i) \leqslant \beta_i \leqslant \hat{\beta}_i + q \cdot s(\hat{\beta}_i)$$

Therefore, we can calculate the width of the interval with:

$$2q \cdot s(\hat{\beta}_i).$$

Table 6: Average length of the 90% confidence interval and the theoretical value

	X_{10}	X_{100}	X_{500}	X_{950}
Avg. CI	3.305	3.464	4.647	14.897
Theoretical value	3.311	3.473	4.666	15.142

Similar to the variance of the β estimators, the average width of the 90% confidence interval multiplies for p = 950, making it difficult to correctly estimate the parameters for multivariate models. Again the calculated values are close to the theoretical ones.

True and false discoveries, FWER and FDR

Theoretical number of False Discoveries:

- no correction: $n_0\alpha = (p-k)\alpha$,
- Bonferroni correction: $n_0\frac{\alpha}{p} = (p-k)\frac{\alpha}{p}$,

where $k = 5$.

FWER estimator - Probability of one or more false positives when tested repeatedly. FWER = P(Number of false discoveries>0)

Theoretical value for:

- no correction: $1-(1-\alpha)^{(p-k)}$,
- Bonferroni correction: $1-(1-\alpha/p)^{(p-k)}$.

FDR estimator - average fraction of the number of false discoveries in the group of all discoveries FDR = E[(Number of false discoveries)/max(1,Number of discoveries)]

Table 7: Without adjusting to multiple testing

	TD	FD	Theo. FD	FWER	Theo. FWER	FDR
X_{10}	4.516	0.488	0.5	0.386	0.41	0.083
X_{100}	4.428	9.594	9.5	0.998	1.00	0.670
X_{500}	3.430	49.520	49.5	1.000	1.00	0.933
X_{950}	0.900	95.600	94.5	1.000	1.00	0.990

Table 8: Bonferroni correction

	TD	FD	Theo. FD	FWER	Theo. FWER	FDR
X_{10}	3.286	0.056	0.050	0.054	0.049	0.015
X_{100}	1.642	0.112	0.095	0.102	0.091	0.051
X_{500}	0.244	0.130	0.099	0.122	0.094	0.111
X_{950}	0.000	0.112	0.099	0.060	0.095	0.060

Table 9: Benjamini-Hochberg correction

	TD	FD	FWER	FDR
X_{10}	4.168	0.280	0.238	0.052
X_{100}	2.244	0.392	0.298	0.100
X_{500}	0.308	0.224	0.348	0.126
X_{950}	0.004	0.614	0.386	0.072

We get the lowest number of false discoveries and FWER using Bonferroni correction. This method also gives us the lowest number of true discoveries. Because of that we can lose valuable information about data, unless our biggest concern is the amount of false discoveries. As expected, without adjusting to multiple testing we get the worst results – FWER and FDR are really high. Benjamini-Hochberg method is a choice in a situation when we want to achieve a low number of false discoveries, not at a cost of true discoveries – FWER and FDR are higher than with Bonferroni correction but TD is also higher. Overall, regardless of the method we choose, the more the number of coefficients the harder it is to correctly predict their significance.