Teoria e Pratica dei Modelli Statistici

Homework 5

Edoardo Ricca

1. *Write the variance inflation factor (VIF) and explain why it is a useful diagnostic tool.*

The **variance inflation factor**, or VIF, is a diagnostic tool for measuring the correlation of a covariate x_j with the other regressors and the impact its degree of collinearity has on the variance of its estimated parameter $\hat{\beta}_j$. The VIF is defined as:

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

Knowing that:

$$\mathbb{V}ar[\hat{\beta}_j] = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

it's easy to see that:

$$\mathbb{V}ar[\hat{\beta}_j] = \frac{1}{1 - R_j^2} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} = \text{VIF} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

As the correlation of x_j with the other covariates, measured by R_j^2, increases, the VIF increases as well and consequently the variance of $\hat{\beta}_j$ (with infinite variance in case of perfectly linear dependence of covariates i.e $R_j^2 = 1$). For this reason, the VIF becomes a useful tool to understand if there exists a collinearity problem. As a rule of thumb, if $\text{VIF}_j > 10$ then it's considered a serious problem.

2. *Write the ridge estimator and compare its properties with those of the least squares estimator in terms of unbiasedness, variance and mean squared error.*

The differences between the OLS and **ridge estimator** are illustrated in this table:

Estimator	OLS	RIDGE
$\hat{\boldsymbol{\beta}}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$
$\mathbb{E}[\hat{\boldsymbol{\beta}}]$	$\boldsymbol{\beta}$	$\boldsymbol{\beta} - \lambda(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}$
$\mathbb{C}ov[\hat{\boldsymbol{\beta}}]$	$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$= \sigma^2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$

where we obtained the expected value of the ridge estimator by:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}_{RIDGE}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] = \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}] = \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}[\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}-\lambda\mathbf{I}]\boldsymbol{\beta} = \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})\boldsymbol{\beta}-\lambda\mathbf{I}\boldsymbol{\beta}] = \\ &= \boldsymbol{\beta}-\lambda(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\boldsymbol{\beta}\end{aligned}$$

indicating that the RIDGE estimator, differently from the OLS, is biased with a bias that tends to underestimate the true value of beta.

The covariance matrix for the RIDGE estimator was obtained, alongside its proof, from this site.

It can be shown that for every $\lambda > 0$:

$$\mathbb{C}ov[\hat{\boldsymbol{\beta}}_{\text{OLS}}] - \mathbb{C}ov[\hat{\boldsymbol{\beta}}_{\text{RIDGE}}]$$

is always positive definite which implies that:

$$\mathbb{V}ar[\hat{\boldsymbol{\beta}}_{\text{OLS}}] > \mathbb{V}ar[\hat{\boldsymbol{\beta}}_{\text{RIDGE}}] \quad \forall j \in \mathbb{N}^+$$

As a consequence, given that the MSE of an estimator is the the sum of its variance (or its trace) and its biased squared, while the RIDGE estimator has a bias that contributes to the MSE it is possible to show that there always exists a λ such that:

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) > \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{RIDGE}})$$

given by the lower variance of the RIDGE estimator.

This becomes particularly useful in the case of collinearity among covariates: not only because it assures that $\mathbf{X}'\mathbf{X}$ doesn't become singular and therefore not invertible but also because, as previously demonstrated, the variance of the OLS estimator increases as the degree of collinearity does, making the RIDGE estimator the one with a lower MSE.

3. *Consider the standardized residuals and the studentized residuals*

a. *Define each symbol*

- **standardized residuals**:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

  - $\hat{\varepsilon}_i$: ith residual (difference between observed value and model prediction)
  - $\hat{\sigma}\sqrt{1-h_{ii}}$: estimated standard deviation of the ith residual

- $\hat{\sigma}$: estimated error standard deviation
- h_{ii}: ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ also interpreted as the leverage term of the ith data point.

- **studentized residuals**:

$$r*_i = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \mathbf{x}'_{(i)}(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}}$$

The subscript (i) indicates that the variables lack the ith row (one observation is removed) and that estimators are computed without considering the ith observation. This is done to ensure that $\hat{\varepsilon}_{(i)} \perp\!\!\!\perp \hat{\sigma}_{(i)}$ giving us a proper distribution. The other symbols are the same as introduced earlier (only with the leave-one-out variation).

b. *Explain how those residuals are used*

The standardized residuals are used to check model assumptions. Since it is possible to have heteroscedastic residuals with homoscedastic errors, to check for the homoscedasticity assumption we compute the standardized residuals. If the assumptions are correct, then the standardized residuals will be homoscedastic. Standardized residuals can also be used to test distributional assumptions of normality by comparing the estimated and theoretical quantiles through the usage of Q-Q plots or other normality tests (e.g. Jarque-Bera).
Lastly, standardized residuals can be used to detect nonlinear effects of covariates by plotting them as a function of the estimated values.

Studentized residuals can be used for the above purpose as well. They can also be used, like standardized residuals, to check for model assumptions. Another important use is in the detection of outliers.

c. *Explain the advantage of studentized residuals*

Studentized residuals allow us to use its known and specific distribution (t_{n-p-1}) to test whether or not an observation is an outlier (the one left out) for a certain significance level α.

d. *Explain why even studentized residuals may not be able to discover multiple outliers*

When testing whether or not an observation belongs to the data-generating process of the standardized residuals, we are applying statistical hypothesis testing. In this case, we are testing the null hypothesis that multiple observations are not outliers. When multiple hypotheses are tested, the probability of observing a rare event increases, and therefore, the probability of making a type I error (rejecting the null when it's true) increases. For this reason, using studentized residuals may not be sufficient.

e. *Explain why an outlier does not necessarily influence the estimates of the regression coefficients*

There are mainly two reasons. First the leverage of the outlier: an observation with a low leverage may have a very small impact on the estimation while an observation with high leverage can alter

it drastically. Then, a very small leverage observation can make the alteration almost null. Second, there may be symmetrically opposed outliers that can potentially cancel each other.

4. *Explain why simultaneously testing all residuals raises a problem about the significance level and mention a solution.*

As mentioned before, testing multiple hypotheses automatically decreases the statistical power of the test by raising the probability of type I error (given a certain level of significance α).
In order to properly execute the test, it is useful to adjust the significance level for each hypothesis/observation. A possible approach is the *Bonferroni correction* which adjusts the significance level by transforming it into α/n where n is the number of observations. The problem with this approach is that it tends to be quite conservative increasing the likelihood of type II error. A more practical solution is to visualize the studentized residuals on a scatter plot.

5. *Write the leverage. Which is the point with minimum leverage?*

The leverage is represented by the diagonal elements h_{ii} of the hat matrix $\mathbf{H}$.
The leverage in the case of one covariate x_i is given by:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

while in matrix notation is expressed as:

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})$$

Where $\tilde{\mathbf{x}}$ is the design matrix centered (by subtracting from each observation of each covariate its mean). Given those two functions, we can see how the minimum leverage achievable is $1/n$, which is possible when $x_i = \bar{x}$, which cancels the second addendum.

6. *Interpret Cook distance's formula in terms of influence, leverage and outlyingness. In particular, explain why a point with high leverage is not always a point with high influence*

Cook's distance is defined by the equation:

$$D_i = \frac{r_i^2 h_{ii}}{p(1 - h_i i)}$$

To talk about its interpretation from the influential point of view, it is easier to see it when the formula is expressed as:

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{p \cdot \hat{\sigma}^2}$$

here we can observe how the numerator is the Euclidian distance between $\hat{\mathbf{y}}_{(i)}$ and $\hat{\mathbf{y}}$. The first one is an estimate of $\mathbb{E}[\mathbf{y}|\mathbf{X}]$ computed by leaving out the ith observation. By computing the distance between this estimator and the one estimated while taking into account all observations, we can

see how influential the ith observation was on the estimated conditional mean. If the Euclidian distance is large, it means that the ith observation had a large influence on the estimate. Otherwise, the opposite is true. The denominator standardizes the result.

Looking instead at the first equation, we notice how the leverage of the observation plays a vital role. The higher the leverage, ceteris paribus, the higher Cook's distance will be.
Even though there is no rigorous definition of an outlier, we can see from the equation that an observation with both a large standardised residual and a high leverage will create a high distance between $\hat{\mathbf{y}}_{(i)}$ and $\hat{\mathbf{y}}$ indicating the possibility of an outlier; since it is only one observation that shifts severely the conditional mean.

As mentioned, it should be noticed that to obtain a large Cook's distance (i.e. a large influence on the estimate) an observation with a high leverage is not a sufficient condition. If the observation has a high leverage but a small enough residual, the overall impact will be small as a result. A combination of both high leverage and large standardized residual is necessary for a high influence on the estimate.

1 Application: model selection

Last time we selected the following model:

testscr	Coefficient	S. e.	t	P>|t|	LB[95%]	UB[95%]
meal_pct	-.3967566	.0277678	-14.29	0.000	-.4513409	-.3421724
el_pct	-.2143334	.0317123	-6.76	0.000	-.2766716	-.1519953
avginc1	4.396992	.6239015	7.05	0.000	3.170565	5.62342
gr_span_dum	-3.526186	1.188313	-2.97	0.003	-5.862097	-1.190274
avginc3	-1.086443	.4164218	-2.61	0.009	-1.905019	-.2678667
expn_stu	.0012145	.0007308	1.66	0.097	-.0002221	.0026512
comp_stu	12.59799	6.716825	1.88	0.061	-.6055355	25.80151
_cons	670.1228	3.940668	170.05	0.000	662.3765	677.8691

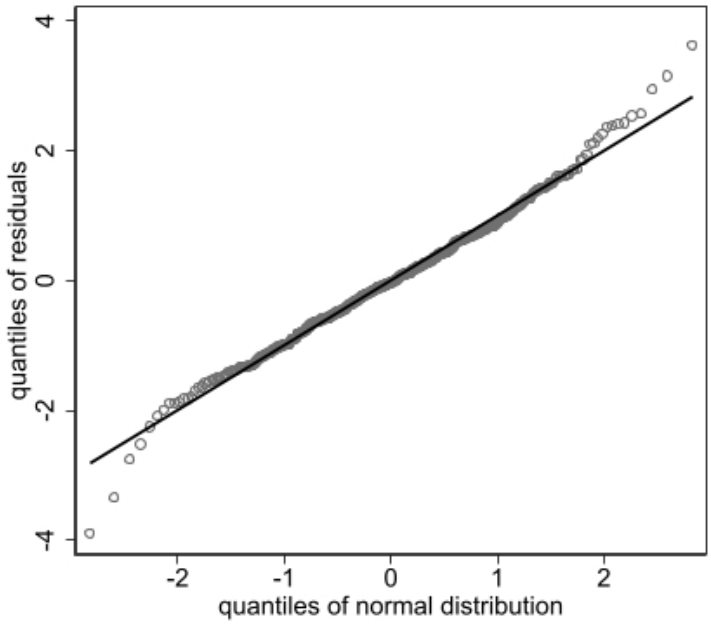
Before proceeding with the exercise, I noticed some possible undetected non-linear effects from both the average income of the district and the number of computers per student. Therefore, I increased the order of the polynomial for the variable `avginc` and built orthogonal polynomials for the variable `comp_stu`.

As a result, I estimated a new model (using the leap-and-bound algorithm) and obtained lower global choice criteria than the previous model. This new model was also tested using 5,10,20-fold and leave-one-out cross-validation, obtaining a lower average squared error than the previous model.

Therefore, we will adopt the new model:

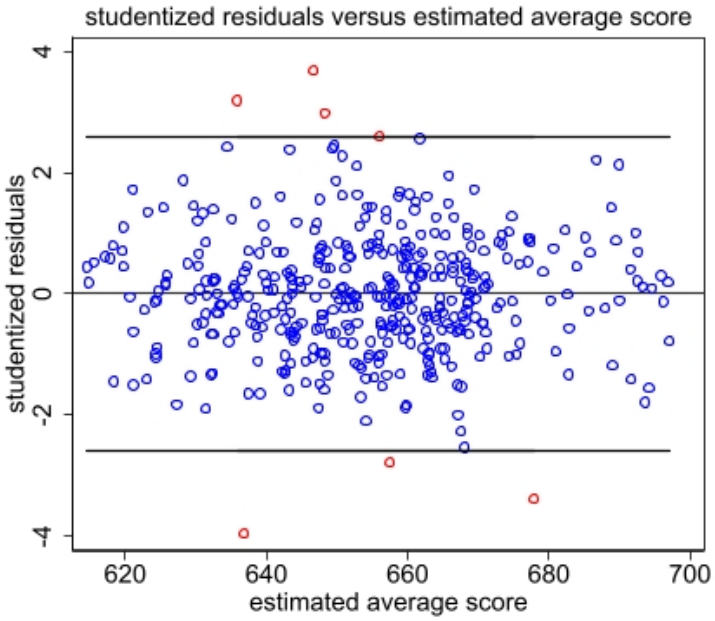
testscr	Coefficient	S. e.	t	P>|t|	LB[95%]	UB[95%]
meal_pct	-.3627607	.035346	-10.26	0.000	-.4322426	-.2932787
el_pct	-.2342861	.0330047	-7.10	0.000	-.2991656	-.1694065
avginc1	4.352524	.6193003	7.03	0.000	3.135124	5.569924
gr_span_dum	-3.864208	1.181192	-3.27	0.001	-6.186156	-1.542259
avginc3	-1.136889	.4113154	-2.76	0.006	-1.945439	-.3283387
c3	1.043022	.4023155	2.59	0.010	.2521637	1.833881
avginc5	.8624724	.4094078	2.11	0.036	.0576722	1.667273
expn_stu	.0018817	.0007102	2.65	0.008	.0004855	.0032778
calw_pct	-.1041841	.055719	-1.87	0.062	-.2137147	.0053465
_cons	668.7549	3.947088	169.43	0.000	660.9959	676.514

1. *QQ plot of the standardized residuals to assess normality*



The Q-Q plot confronts the empirical quantiles of the standardized residuals with the theoretical quantiles of a standard normal. It seems that the quantiles strongly correlate except for some deviation for the higher quantiles. The residuals seem to have slightly fatter tails than a normal distribution.

2. *Plot of the studentized residuals against the predicted values, with confidence bands using the t distribution (alpha=0.01)*



From the plot, it appears that only 7 observations (red) fall outside the 99% interval of the t_{410} which is expected since it represents approximately 0.017% of the total observations. We can conclude from the Q-Q plot and the above plot that the model meets its assumptions regarding the residuals.

3. *List observations with very large residuals*

The 7 observations with studentized residuals that fall below the quantile of order 0.005 and above the order 0.995 of the t_{410} distribution are the following:

# observation	rstudent
6	-3.958439
367	3.68679
180	-3.384595
262	3.187091
419	2.97556
77	-2.776629
371	2.594314

more information about the observations are available on using the .do file attached.

4. *Compute the leverage and list observations above the conventional threshold*

After computing the threshold equal to $2p/n = 0.04285714$, the observations above are 41. The list can be found on the .do file attached.

5. *Compute Cook's distance and list observations above 0.5*

There are only two observations with a Cook's distance above 0.5. The two are listed below:

# observation	Cook's distance
414	0.0616705
6	0.0520275

6. *Remove the 10 observations with largest Cook's distance and refit the model; compare with the estimates obtained with the full sample*

By removing the 10 observations indicated, we obtain the following results:

testscr	Coefficient	S. e.	t	P>|t|	LB[95%]	UB[95%]
meal_pct	-.3916423	.0347384	-11.27	0.000	-.4599349	-.3233497
el_pct	-.2300539	.0313611	-7.34	0.000	-.2917071	-.1684007
avginc1	3.456809	.6596017	5.24	0.000	2.16009	4.753528
gr_span_dum	-4.867547	1.117394	-4.36	0.000	-7.064246	-2.670847
avginc3	-1.407487	.4086574	-3.44	0.001	-2.210872	-.6041027
c3	1.228485	.387493	3.17	0.002	.4667079	1.990262
avginc5	1.460594	.4571768	3.19	0.002	.5618243	2.359364
expn_stu	.0015043	.000672	2.24	0.026	.0001832	.0028254
calw_pct	-.0653352	.0562783	1.16	0.246	-.1759735	.045303
_cons	672.1322	3.760832	178.72	0.000	664.7387	679.5256

The new fitted model has different coefficients and generally lower standard errors and shorter confidence intervals (except for some polynomial terms). The p-values are especially lower except for the variable `calw_pct` which becomes rejectable under the hypothesis $\beta \neq 0$.
More relevantly, the big difference between models is, as expected, in their model choice criteria

since:

MCC	Original Model	New model
AIC	2967.755	2830.701
BIC	3008.158	2870.863
CV 5-fold (Avg. Root MSE)	8.31687	7.66019
CV 10-fold (Avg. Root MSE)	8.29796	7.64673
CV 20-fold (Avg. Root MSE)	8.28918	7.63823
CV LOO (Root MSE)	8.2829	7.6333

The new model clearly outperforms the old one in matters of mean square error since we removed the observations with the highest Euclidian distance with the estimated $\hat{\mathbf{y}}$ of the model previously fitted.