KATHOLIEKE UNIVERSITEIT

LEUVEN

Statistical Modelling

Exam project

Prof. dr. Gerda Claeskens

Daniel Izquierdo Juncàs (r0654210)

5th June 2017

PART A: Simulation study about post-selection inference

For this simulation study about post-selection inference we generate data from a linear regression model $Y = X\theta + \varepsilon$ where Y is a vector with length n = 200 and X is a 200 × 5 design matrix. The first column of X is the intercept and the other four are generated with a multivariate normal distribution with mean vector 0 and covariance matrix $\begin{pmatrix}1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1\end{pmatrix}$. The parameter vector is $\theta = (\theta_1, \theta_2, \theta_3, 0,0)^t$ = $(3,3,3,0,0)^t$ and the error vector ε with length 200 contains i.i.d. elements from a standard normal distribution *N (0,1)*.

A.2

We focus on the model M that contains the parameters $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, 0)^t$ (M = 2) and we base the model selection on the AIC. We present the results for the off-diagonal elements of the covariance matrix $\rho = 0.25$ and $\rho = 0.75$. In Table A1 the simulated coverage probabilities for each of the four parameters is presented, that is how many times the true parameter lies inside the constructed confidence interval (CI) over the 1000 simulations. Ideally, when the number of samples is large these coverages should be 0.95 for the 95% CI's. This is approximately the case of the 3 parameters that have non-zero values in the real model for the naïve CI when $\rho = 0.25$; however, the coverage of this CI for $\hat{\theta}_4$ is very poor. The PoSI CI coverages are around 99%, higher than it could be expected for the traditional CI's and even for $\hat{\theta}_4$ the PoSI CI it is good. For $\rho = 0.75$ the coverages of the naïve CI's are slightly worse than for $\rho = 0.25$ but this is not the case for the PoSI CI's.

		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
$\boldsymbol{\rho = 0.25}$	**Naive CI**	0.934	0.950	0.932	0.690
	PoSI CI	0.988	0.994	0.989	0.955
$\boldsymbol{\rho = 0.75}$	**Naive CI**	0.937	0.897	0.876	0.678
	PoSI CI	0.990	0.987	0.984	0.972

Table A1: Simulated coverage probabilities for $\rho = 0.25$ and $\rho = 0.75$. AIC selection and focus on M = 2.

		$\boldsymbol{\rho = 0.25}$		$\boldsymbol{\rho = 0.75}$	
		2.5%	**97.5%**	**2.5%**	**97.5%**
Naive	$\hat{\theta}_1$	2.8595	3.1366	2.8592	3.1363
	$\hat{\theta}_2$	2.8520	3.1452	2.7644	3.2308
	$\hat{\theta}_3$	2.8519	3.1456	2.7632	3.2293
	$\hat{\theta}_4$	-0.1435	0.1504	-0.2275	0.2389
PoSI	$\hat{\theta}_1$	2.8062	3.1898	2.7961	3.1994
	$\hat{\theta}_2$	2.7957	3.2015	2.6582	3.3370
	$\hat{\theta}_3$	2.7955	3.2021	2.6570	3.3354
	$\hat{\theta}_4$	-0.1999	0.2069	-0.3337	0.3451

Table A2: Average 95% confidence intervals for each parameter for the naïve and post-selection approaches for $\rho = 0.25$ and $\rho = 0.75$. AIC selection and focus on M = 2.

	$\boldsymbol{\rho = 0.25}$		$\boldsymbol{\rho = 0.75}$	
	2.5%	**97.5%**	**2.5%**	**97.5%**
$\hat{\theta}_1$	2.8464	3.1417	2.8476	3.1416
$\hat{\theta}_2$	2.8541	3.1468	2.7408	3.2942
$\hat{\theta}_3$	2.8394	3.1647	2.7193	3.2875
$\hat{\theta}_4$	-0.2003	0.1898	-0.3202	0.3028

Table A3: Empirical 95% confidence intervals by using the sample quantiles of the 1000 estimates of each parameter for $\rho = 0.25$ and $\rho = 0.75$. AIC selection and focus on M = 2.

In Table A2 we show the average confidence interval for each parameter calculated with both methods (naïve and post-selection). The average naïve CI for $\hat{\theta}_1$ is the same regardless of ρ as this parameter acts as the intercept; the PoSI CI for $\hat{\theta}_1$ is also very similar for both ρ. All the other parameters present wider CI's (naïve and PoSI) for higher ρ's. The average PoSI CI's are around 0.10 wider than the average naïve CI's for $\rho = 0.25$ and around 0.20 wider for $\rho = 0.75$ (except for $\hat{\theta}_1$), so they are more conservative.

In Table A3 the empirical 95% confidence intervals are presented. For $\hat{\theta}_4$ (parameter with zero value in the real model) these CI's are very close to the PoSI CI's of Table A2. For the other 3 parameters with non-zero values in the real model, the empirical CI's are narrower and are more similar to the naïve CI's of Table A2.

A.3

We focus on the model M that contains the parameters $\theta = (\theta_1, \theta_2, \theta_3, 0,0)^t$ (M = 4), which is also how the real model looks like, and we base the model selection on the BIC, that has a preference for simpler models. In Table A4 we see that this time the naïve CI coverage probabilities are very close to 95% for the 3 parameters. The PoSI CI is more conservative as it has coverages of more than 99% in all cases.

		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
$\boldsymbol{\rho = 0.25}$	**Naive CI**	0.942	0.957	0.951
	PoSI CI	0.995	0.993	0.994
$\boldsymbol{\rho = 0.75}$	**Naive CI**	0.942	0.954	0.951
	PoSI CI	0.997	0.997	0.998

Table A4: Simulated coverage probabilities for $\rho = 0.25$ and $\rho = 0.75$. BIC selection and focus on M = 4.

From Table A5, the naïve and PoSI average CI's for $\rho = 0.25$ are very similar to the ones in Table A2. For $\rho = 0.75$ the naïve and PoSI CI's are narrower than their counterparts in Table A2, more in the case of the PoSI ones. Again, we see that the PoSI average CI's are larger in all cases than their naïve counterpart. In Table A6 we observe very similar empirical CI's for $\rho = 0.25$ compared with the empirical CI's of the case where we focus in M = 2 and selection by AIC (Table A3). However, for $\rho = 0.75$ only $\hat{\theta}_1$ CI is similar to the one obtained before; for $\hat{\theta}_2$ and $\hat{\theta}_3$ when we use the BIC and focus in M = 4 the empirical CI's get narrower. This is because in this case the generated datasets by *conddatagen* are the ones that select the real model.

		$\boldsymbol{\rho = 0.25}$		$\boldsymbol{\rho = 0.75}$	
		2.5%	**97.5%**	**2.5%**	**97.5%**
Naive	$\hat{\theta}_1$	2.8610	3.1388	2.8612	3.1390
	$\hat{\theta}_2$	2.8545	3.1428	2.7862	3.2083
	$\hat{\theta}_3$	2.8575	3.1455	2.7915	3.2133
PoSI	$\hat{\theta}_1$	2.8088	3.1910	2.7979	3.2023
	$\hat{\theta}_2$	2.8003	3.1970	2.6900	3.3045
	$\hat{\theta}_3$	2.8034	3.1997	2.6955	3.3093

Table A5: Average confidence intervals for each parameter for the naïve and post-selection approaches for $\rho = 0.25$ and $\rho = 0.75$. BIC selection and focus on M = 4.

	$\boldsymbol{\rho = 0.25}$		$\boldsymbol{\rho = 0.75}$	
	2.5%	**97.5%**	**2.5%**	**97.5%**
$\hat{\theta}_1$	2.8578	3.1428	2.8578	3.1428
$\hat{\theta}_2$	2.8578	3.1343	2.7981	3.2017
$\hat{\theta}_3$	2.8573	3.1423	2.7911	3.2083

Table A6: Empirical 95% confidence intervals by using the sample quantiles of the 1000 estimates of each parameter for $\rho = 0.25$ and $\rho = 0.75$. BIC selection and focus on M = 4.

PART B: Modelling data about car accidents

B.1

	Year	*Safety*	*Speed*	*Power*	*Airbag*	*Temp*	*Wind*	*Severity*
mean	2004.360000	4.1020	87.1933	111.1433	53.7167	25.6233	9.6033	58.9633
std.dev	4.608818	3.5489	36.6705	16.2467	24.2134	4.9918	9.4214	18.7590
median	2004.000000	3.3000	91.0000	107.0000	44.0000	25.0000	7.0000	62.5000
min	1995.000000	0.0000	15.0000	80.0000	36.0000	11.0000	0.0000	7.0000
max	2015.000000	10.6000	157.0000	147.0000	160.0000	38.0000	52.0000	91.0000

Table B1: Exploratory analysis of the variables in the dataset.

In Table B1 we show some basic statistics and see that some variables have a non-zero minimum. For instance, taking Year = 0 or taking Speed or Power of the car = 0 would make no sense in a car accident. Hence, to ease the interpretation of the intercept of the following models we decide to mean centre the variables Year, Speed, Power and Airbag. In Figure B1 the histogram for each variable and the correlation matrix are shown where we see that the only relevant correlation is between Airbag and Safety. We see that the response Severity is left skewed and that its correlation with other variables is not highly informative, although we can see some curvature trends (e.g. for Year).

We assume that the errors are normally distributed and we fit a linear regression model with Severity as the response and the other variables as the regressors. First, a multicollinearity diagnostic is performed (Table B2) and it is observed that the Variance Inflation Factors (VIF) are all smaller than 5 so the data does not present multicollinearity. We start fitting the full model for only main effects. In the full model, the variables Safety, Temp and Wind are not significant at a 5% level so model selection is performed. Considering only the models with main effects there is a total of 2^7 = 128 possible subsets. As this number is relatively low, we decide to perform all subsets model selection (Table B3). Basing our decision in the AIC, the model with the lowest value is the one that includes all covariates except Temp and Wind. However, there are 6 other models with AIC differences of 2 or less with the best one so we can also support our choice with the BIC and the adjusted R^2 values. The first, prefers simpler models as it chooses the one that also excludes Safety; the adjusted R^2 suggests more complicated models like the full one. We decide to choose the model suggested by the BIC as it is simpler and in the full model Safety is not significant.

	VIF
Year	1.110
Safety	3.365
Speed	1.330
Power	1.041
Airbag	2.761
Temp	1.004
Wind	1.015

Table B2: VIFs.

(Intercept)	*Safety*	*Airbag.c*	*Power.c*	*Speed.c*	*Temp*	*Wind*	*Year.c*	*BIC*	*adjR^2*	*AIC*
61.81	-0.6935	-0.2782	0.22030	0.2287	*NA*	*NA*	-2.489	2404	0.55901	2378
62.59	-0.7205	-0.2776	0.22242	0.2300	*NA*	-0.0693806	-2.492	2409	0.56021	2379
58.96	*NA*	-0.3508	0.21236	0.2089	*NA*	*NA*	-2.426	2402	0.55385	2379
61.55	-0.6924	-0.2784	0.22040	0.2287	0.009990	*NA*	-2.489	2409	0.55902	2380
59.51	*NA*	-0.3526	0.21386	0.2093	*NA*	-0.0573795	-2.426	2407	0.55468	2381
62.30	-0.7193	-0.2777	0.22253	0.2300	0.010892	-0.0694208	-2.492	2414	0.56022	2381
58.40	*NA*	-0.3508	0.21261	0.2089	0.022075	*NA*	-2.426	2407	0.55389	2381

Table B3: First 7 models ordered by AIC from all subsets regression. *NA* stands for variables not included in the model.

	Estimate	*Std. Error*	*t-value*	*Pr(>|t|)*
(Intercept)	58.9633	0.7284	80.95	<2e-16
Airbag.c	-0.3508	0.0310	-11.33	<2e-16
Power.c	0.2124	0.0455	4.66	0.0000047
Speed.c	0.2089	0.0202	10.34	<2e-16
Year.c	-2.4258	0.1632	-14.86	<2e-16

Table B4: estimates of the chosen model.

In Table B4 the estimates of the chosen model are presented with their standard errors and tests of significance. In this model $R^2 = 0.554$, so the model explains 55.4 % of the variance of the response. The intercept is the expected Severity of an accident with mean Airbag, Power, Speed and Year: an accident in the Year ~ 2004 with a car of 111.1 horsepowers, airbag volume of 53.7 liters and speed 87.19 km/h would have an expected severity of 58.96. The other four estimates are the slopes of each regressor and they represent the change in the expected response when the regressor is incremented in one unit having the other variables fixed. For instance, if the power of the car is increased by 1 horsepower the expected increase in the Severity is 0.2124.

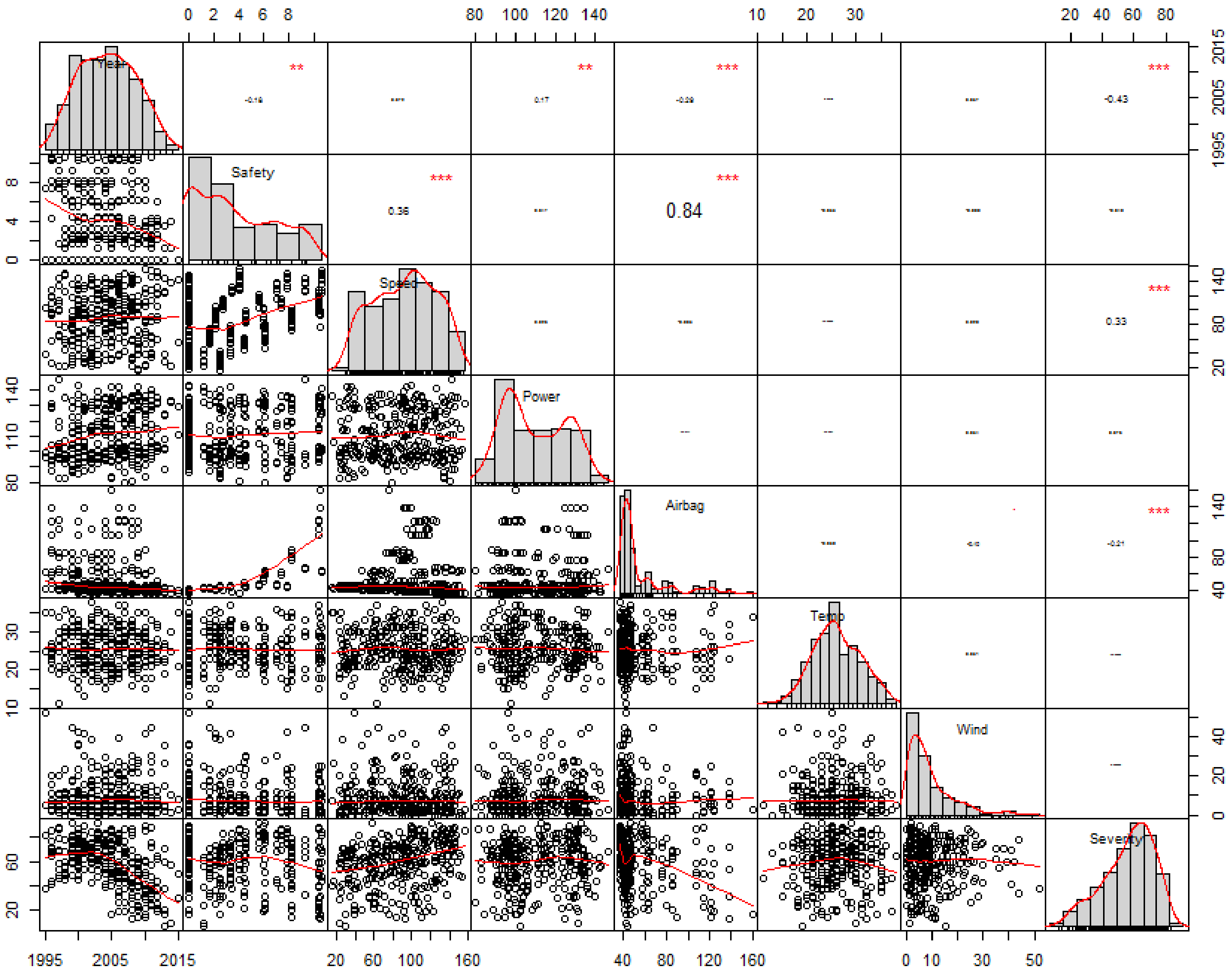

Figure B1: correlation matrix with scatterplots and histograms.

B.2

Working with the original dataset (non-centered variables) we consider a full model that includes the main, quadratic, cubic and two-way interaction effects and compute a lasso estimator with the tuning parameter λ chosen by 10-fold cross-validation. The cross-validation mean squared prediction error (CV Error) depends on λ and the objective is to find the parameter value that minimizes it. Parameters very close to zero give the same results as ordinary least squares estimator (OLS) and large values of λ increase the bias too much and set all coefficients equal to zero. A grid of λ's is created starting from 0.001 and increasing its value by 0.001 until 0.750. This range of λ's is considered appropriate, although if λ is very close to zero it may cause convergence problems in the *glmnet* function. In Figure B2 we plot the CV Error against λ and see how for very small values of λ the CV Error is high, then it rapidly decreases and from λ ≈ 0.35 it increases slowly again. The value that minimizes the CV Error is λ = 0.250 but λ's between 0.15 and 0.40 give similar results for the CV Error. Taking different seeds usually gives results inside this range.

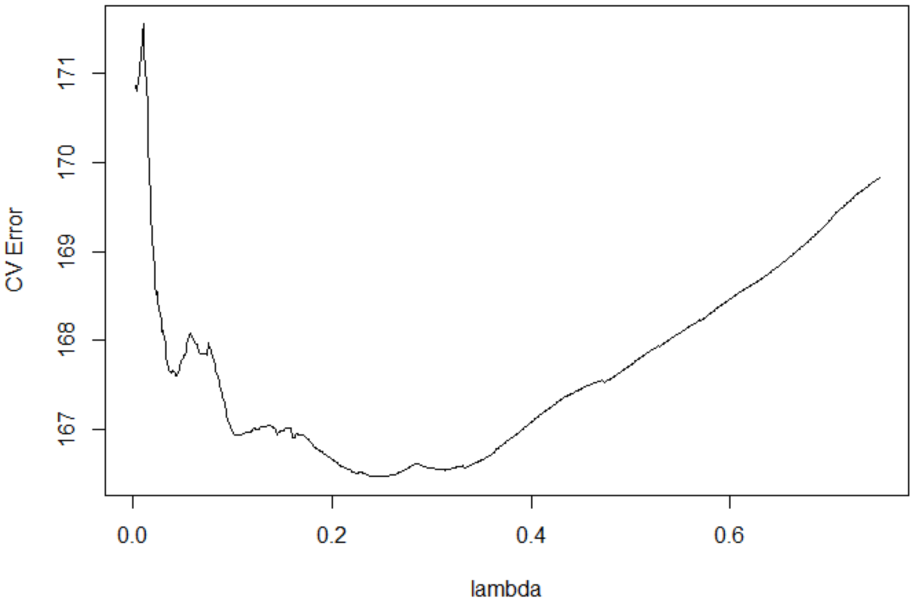

Figure B2: CV Error against λ.

We fit a lasso model with λ = 0.250 and obtain the model

$$\begin{aligned}\hat{y}_i = 2117.17 &+ 0.20\,Power + 5.8 \cdot 10^{-5}\ Year\!:\!Speed - 0.00011\,Year\!:\!Airbag - 0.0074\,Safety\!:\!Speed \\ &+ 0.00097\,Speed\!:\!Wind + 2.1 \cdot 10^{-5}\,Power\!:\!Temp - 0.00053\,Airbag\!:\!Temp \\ &- 0.00033\,Year^2 + 0.00069\,Speed^2 - 9.6 \cdot 10^{-8}\,Year^3 - 1.2 \cdot 10^{-6}\,Airbag^3 \\ &- 0.00012\,Wind^3\end{aligned}$$

There are 12 terms selected, the other 30 coefficients are shrunk toward zero and their terms not included in the model. Only one main effect is included (Power) but all the variables are present in the form of some interactions (there are 6 interaction terms), two quadratic and three cubic effects are also present. We conclude that the main effects model fitted with OLS in B1 can be improved as it missed some higher order terms regarded important by lasso. It is remarkable that lasso estimator can include interactions or higher order effects without the need of including also their main effects. In addition, we have to consider that lasso estimator does not perform very well in non-sparse settings and that it has a tendency for overselection. Finally, in this case the intercept does not have a direct interpretation because as stated before it is not possible to set all variable values to zero.

B.3

(1)

We construct a flexible additive model with Severity as the response and all the other variables as regressors:

$$Y_i = \beta_0 + \sum_{j=1}^{7} f_j(x_{ji}) + \varepsilon_i$$

where Y_i is the response for the observation *i*, β_0 the intercept, ε_i the error *i* and the functions f_j are smooth functions for each of the 7 regressors. The cubic thin plate splines, which are the default basis functions of *spm*, are used. As in the following part we will perform model selection and restricted maximum likelihood (REML) gives some problems to compare models with different mean structures, we use maximum likelihood (ML) as the method for automatic smoothing parameter selection.

	df	*spar*	*knots*
f(Year)	3.056	16.70	4
f(Safety)	1.000	13670.00	4
f(Speed)	1.000	547600.00	30
f(Power)	1.000	36260.00	14
f(Airbag)	1.000	114100.00	7
f(Temp)	1.858	46.15	5
f(Wind)	1.000	23070.00	8

Table B5: Summary for non-linear components of the full flexible additive model.

In Table B5 the summary for the non-linear components (all of them) is exposed. It is seen that the degrees of freedom for Safety, Speed, Power, Airbag and Wind are 1, suggesting that a linear fit for these variables might be appropriate. Similarly, a quadratic fit for Temp and a cubic fit for year might also be appropriate. The second column shows the smoothing parameter (spar), which is very large in this case because the data is very disperse (Figure B1). The 3rd column is the number of knots taken for the non-parametric fit of every variable. In Figure B3 the non-parametric fits are plotted; we see that Safety, Speed, Power, Airbag and Wind are linear and only Year and Temp show curvature. The standard error bands for Safety, Temp and Wind are the widest indicating that the relation between these variables and the response might not be significant.

(2)

Model selection is performed manually by the AIC. Starting with the full model (AIC = 2362) we drop one variable at each step until the minimum AIC is reached. The model selected has AIC = 2358 and is the one that includes Year, Safety, Speed, Power and Airbag as covariates (Table B6). During this stepwise procedure two models presented singularities so the AIC for them could not be computed; the second is the one that discards also the variable Safety so it might be another good candidate model.

	df	*spar*	*Knots*
f(Year)	2.955	17.88	4
f(Safety)	1.000	12990.00	4
f(Speed)	1.000	70320.00	30
f(Power)	1.000	49620.00	14
f(Airbag)	1.000	106300.00	7

Table B6: summary for non-linear components of the final flexible additive model.

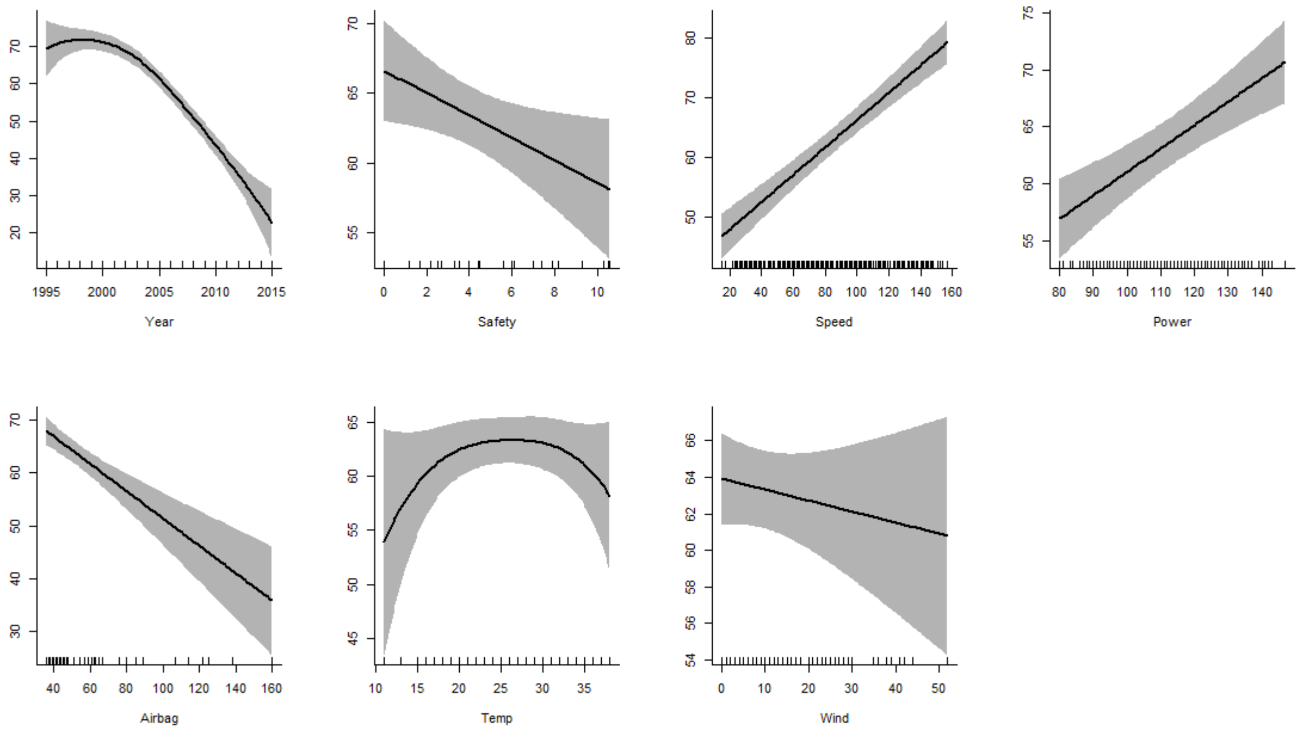

Figure B3: non-parametric fits with shaded standard error bands.

(3)

Looking at Table B6 we see that the degrees of freedom suggest fitting a parametric model with Safety, Speed, Power and Airbag as linear terms and Year as a cubic term. In Table B7 the estimates, standard errors and tests of significance of the polynomial model are shown. We used the function *poly* to produce orthogonal polynomials in order to not have correlated variables that produce singularity problems. All the terms are significant at a 5% level except the cubic term for Year. The AIC for this model is 2347, so it is considered better than the linear model of B1 and than the non-parametric fit of B3 (1). The adjusted R^2 is 0.597.

	Estimate	*Std. Error*	*t-value*	*Pr(>|t|)*
(Intercept)	33.1431	5.4981	6.03	5e-9
poly(Year,3)1	-197.5856	12.5470	-15.75	<2e-16
poly(Year,3)2	-69.3010	11.9160	-5.82	1.6e-8
poly(Year,3)3	16.2647	12.1522	1.34	0.182
Safety	-0.7390	0.3544	-2.09	0.038
Speed	0.2287	0.0216	10.59	<2e-16
Power	0.2086	0.0432	4.83	2.2e-6
Airbag	-0.2658	0.0475	-5.59	5.2e-8

Table B7: estimates of the polynomial model.

The residual against fitted values plot shows a curved band, which indicates that our assumptions might not be correct (Figure B4). The residuals are approximately normally distributed and we also note the presence of some influential observations.

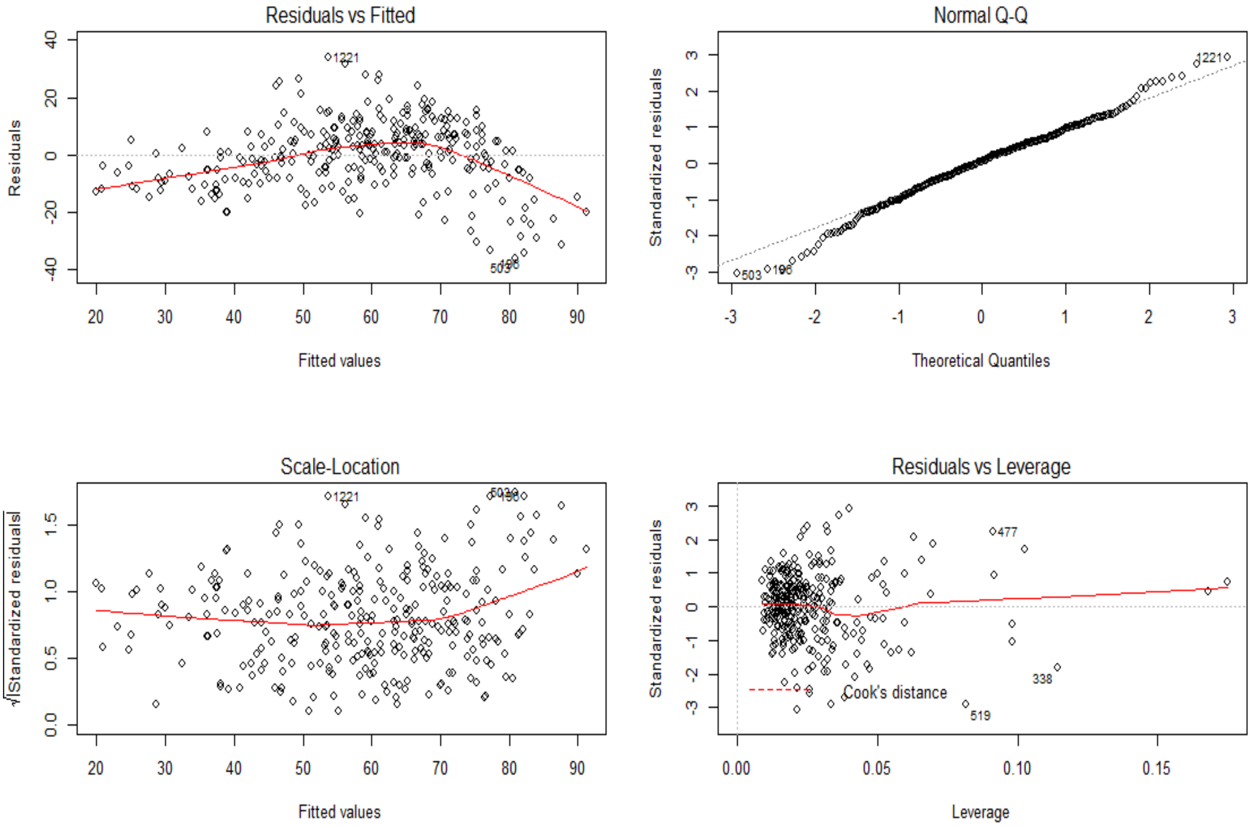

Figure B4: regression diagnostic plot of the polynomial model.

Finally, we compare this polynomial model with the one obtained with lasso in B2. Lasso only selected one main effect (Power) whereas with this approach we selected 5 main effects. The only higher order terms in this model are the quadratic and cubic (non-significant) effects of Year, which are also included in the lasso model. Lasso also included the quadratic effect of speed and the cubic effects of airbag and wind which the non-parametric fit did not detect. In addition, we can not compare directly the estimates of both models because in B3 we followed an additive approach so we did not consider the interaction terms as we did in B2.