Ανάλυση Δεδομένων

Νεντίδης Δημήτριος 1η Εργασία

26 Απριλίου 2025

Περιεχόμενα

1 Εισαγωγή

Στα πλαίσια της πρώτης εργασίας μελετάται η εξάρτηση του προσδόκιμου όριου ζωής από 17 διαφορετικές μεταβλητές μεταξύ των οποίων είναι η κατανάλωση αλκόολ, το επίπεδο μόρφωσης, ή ακόμα και ο δείκτης σωματικής μάζας. Τα δεδομένα δίνονται για διάφορες χώρες και χρονιές. Στην παρούσα ανάλυση δεν λαμβάνονται υπόψη τα δεδομένα για τις χρονιές 2001 και 2007.

2 Απλή γραμμική παλινδρόμηση

Αρχικά εκτελείται μια απλή γραμμική παλινδρόμηση για κάθε μια από τις υποψήφιες προβλεπτικές μεταβλητές, με εξαίρεση την χώρα προέλευσης των δεδομένων και την κατάσταση της χώρας, εάν είναι αναπτυγμένη ή αναπτυσόμενη. Μηδενική υπόθεση για όλες τις προβλεπτικές μεταβλητές είναι η απουσία στατιστικής σχέσης μεταξύ της μεταβλητής και της απόκρισης. Η υπόθεση αυτή εξετάζεται για επίπεδο σημαντικότητας 0.01.

Για την αξιολόγηση των αποτελεσμάτων τα μοντέλα συνοψίζονται σε δύο πίνακες. Στον πρώτο φαίνεται η εξίσωση που περιγράφει την σχέση της απόκρισης Y με την εκάστοτε μεταβλητή, ενώ στον δεύτερο φαίνονται οι συντελεστές προσδιορισμού, ή με άλλα λόγια τι ποσοστό της μεταβλητότητας εξηγήται από το μοντέλο, και το p-value. Μπορεί να παρατηρηθεί ότι τα μοντέλα για τον πληθυσμό μιας χώρας καθώς και για το έτος από όπου προέρχονται τα δεδομένα, δεν είναι στατιστικά σημαντικά βάση του κριτηρίου που ορίστηκε για την απόρριψη της μηδενικής υπόθεσης. Επιπλέον, οι μεταβλητές "Percentage Expenditure", "Measles", "GDP" έχουν πολύ χαμηλό συντελεστή κλίσης β. Αξιοσημείωτο είναι βέβαια ότι οι διάφορες μεταβλητές δεν θεωρούνται κανονικοποιημένες και συνεπώς η συμβολή τους δεν μπορεί να κριθεί εύκολα. Με άλλα λόγια, συντελεστής κοντά στο μηδέν δεν καθιστά απαραίτητα ανάξια λόγου την μεταβλητή. Ωστόσο, λαμβάνοντας υπόψη το μεγάλο μέγεθος του δείγματος και τους συντελεστές αυτών τιμών σε συνδυασμό με τα p-values και τα R^2 τους, πρακτικά είναι εμφανές ότι δεν συμμετέχουν με ουσιαστικό τρόπο στην διακύμανση των τιμών παρά την στατιστικά σημαντική συσχέτιση.

Οι προβλεπτικές μεταβλητές που θεωρούνται ότι εξηγούν την δια-

Προβλεπτική μεταβλητή	Εξίσωση
Year	$Y = -211.026 + 0.140 \cdot X$
Adult Mortality	$Y = 77.788 - 0.050 \cdot X$
Alcohol	$Y = 65.345 + 0.887 \cdot X$
Percentage Expenditure	$Y = 67.862 + 0.002 \cdot X$
Hepatitis B	$Y = 63.499 + 0.073 \cdot X$
Measles	$Y = 69.471 - 0.000072 \cdot X$
BMI	$Y = 60.153 + 0.240 \cdot X$
Under-five Deaths	$Y = 69.774 - 0.010 \cdot X$
Polio	$Y = 58.283 + 0.132 \cdot X$
Total Expenditure	$Y = 65.269 + 0.675 \cdot X$
Diphtheria	$Y = 57.577 + 0.140 \cdot X$
HIV/AIDS	$Y = 71.016 - 0.887 \cdot X$
GDP	$Y = 67.400 + 0.000334 \cdot X$
Population	$Y = 69.351 - 2.775 \times 10^{-9} \cdot X$
Thinness 5-9 Years	$Y = 73.546 - 0.860 \cdot X$
Income Composition of Resources	$Y = 47.318 + 34.804 \cdot X$
Schooling	$Y = 41.410 + 2.298 \cdot X$

Πίνακας 1: Εξισώσεις απλής γραμμικής παλινδρόμησης.

κύμανση των τιμών, και άρα έχουν σημαντική συσχέτιση με την απόκριση, είναι πρώτα πρώτα τα σχολικά χρόνια που ολοκληρώνονται (Schooling), το Income composition of resources, η θνησιμότητα ενηλίκων (Adult Mortality), όλα με $R^2 > 0.5$. Τα μοντέλα για τους θανάτους που προκύπτουν από HIV/AIDS σε ηλικίες 0-4 έτη (HIV/AIDS) και τυο σωματικού δείκτη μάζας (BMI), με συντελεστή προσδιορισμού 0.345 και 0.293 αντίστοιχα, δείχνουν και να εξηγούν ένα μέρος της διακύμανσης, όχι όμως εξίσου ικανοποιητικά.

3 Πολλαπλή γραμμική παλινδρόμηση

Ενώ προηγουμένως το μοντέλο περιείχε μόνο μια μεταβλητή κάθε φορά, σε αυτή την ενότητα εξετάζεται η μοντελοποίηση της απόκρισης συναρτήσει όλων των προβλεπτικών μεταβλητών. Η μηδενική υπόθεση εδώ είναι ότι όλοι συντελεστές των μεταβλητών πρόβλεψης είναι ίσοι με το μηδέν. Πιο συγκεκριμένα, το μοντέλο που προέκυψε έχει συντε-

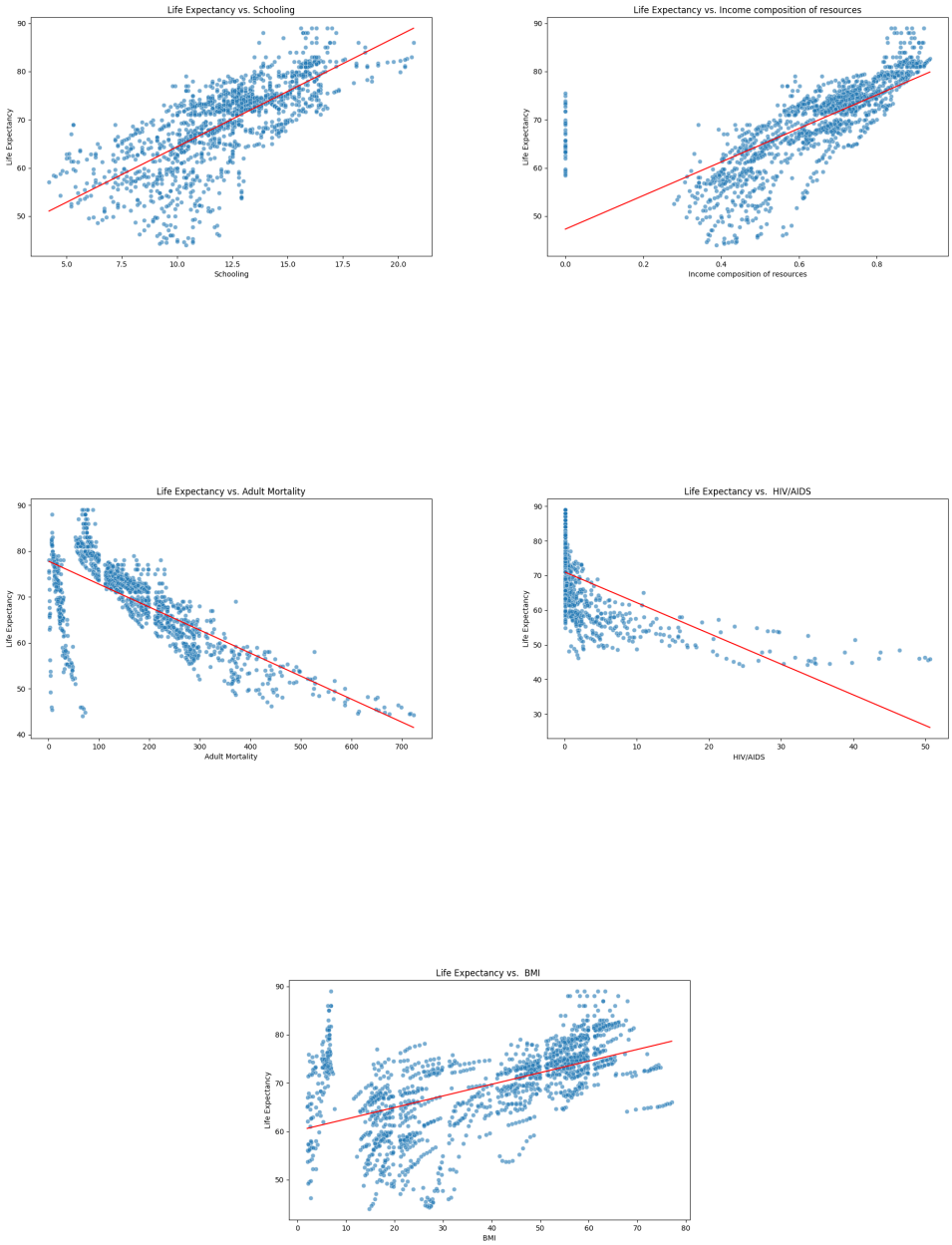
Προβλεπτική μεταβλητή	R^2	p-value τάξη μεγέθους
Year	0.004	10^{-2}
Adult Mortality	0.505	10^{-225}
Alcohol	0.165	10^{-59}
Percentage Expenditure	0.173	10^{-62}
Hepatitis B	0.045	10^{-16}
Measles	0.007	10^{-3}
BMI	0.293	10^{-112}
Under-five Deaths	0.037	10^{-14}
Polio	0.116	10^{-41}
Total Expenditure	0.032	10^{-12}
Diphtheria	0.120	10^{-42}
HIV/AIDS	0.345	10^{-136}
GDP	0.201	10^{-73}
Population	0.000	10^{-1}
Thinness 5-9 Years	0.208	10^{-76}
Income Composition of Resources	0.527	10^{-239}
Schooling	0.536	10^{-245}

Πίνακας 2: Αξιολόγηση απλής γραμμικής παλινδρόμησης.

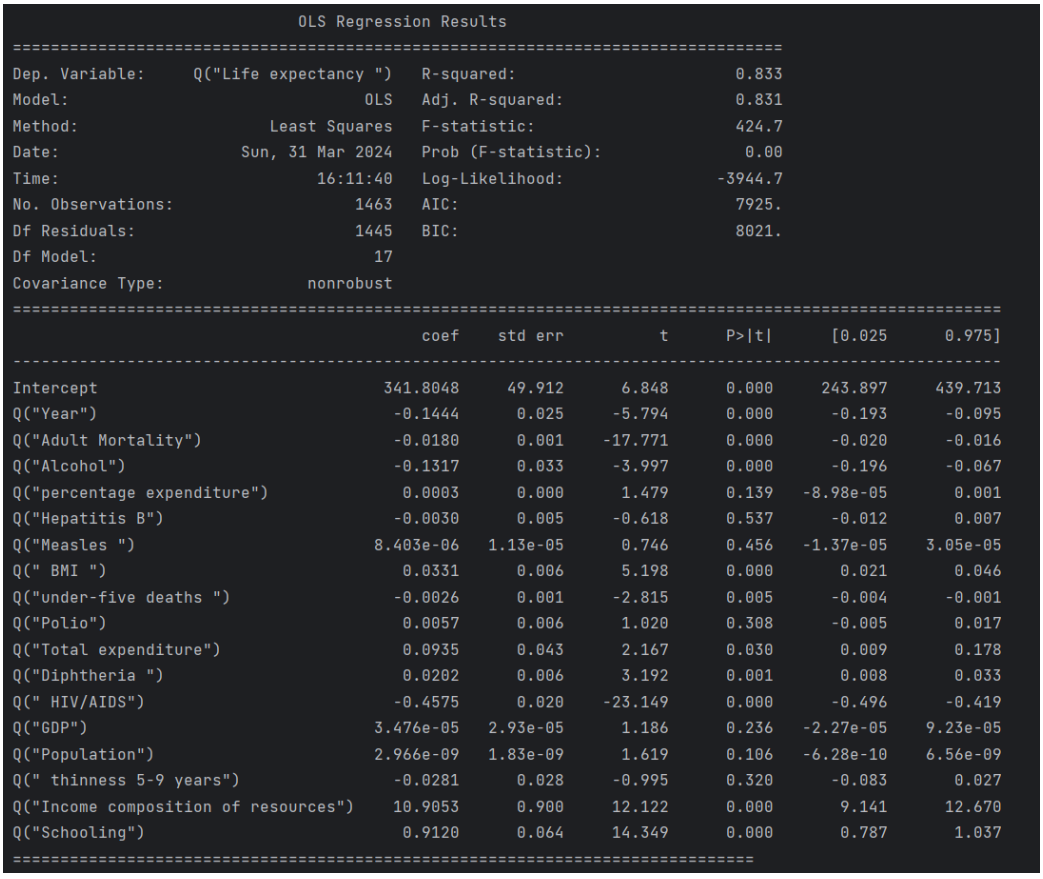
λεστή προσδιορισμού ίσο με 0.83 δηλαδή το μοντέλο εξηγεί το 83% της μεταβλητότητας που παρατηρήται στα δεδομένα.

Η θνησιμότητα ενηλίκων (Adult Mortality), η κατανάλωση αλκόολ (Alcohol), ο δείκτης σωματικής μάζας (BMI), οι περιπτώσεις HIV/AIDS σε ηλικίες 0-4 ετών (HIV/AIDS), το Income Composition of Resources, και τα ολοκληρωμένα χρόνια εκπαίδευσης θεωρούνται οι μεταβλητές για τις οποίες απορρίπτεται η μηδενική υπόθεση. Για όλα το p-value είναι μικρότερο του 0.001, κάτι που υποδηλώνει ότι είναι πρακτικά αδύνατο να εμφανιζόντουσαν τέτοια αποτελέσματα εάν ίσχυε η μηδενική υπόθεση. Για τις υπόλοιπες μεταβλητές η απόρριψη της μηδενικής υπόθεσης δεν μπορεί να γίνει με τα διαθέσιμα δεδομένα.

Κοιτώντας τους συντελεστές μπορεί να εξαχθεί ότι, παραδείγματος χάρη, ένας επιπλέον χρόνος εκπαίδευσης οδηγεί σε αύξηση του προσδόκιμου ορίου ζωής κατά 0.9 χρόνια. Για κάποιες μεταβλητές, όπως το ICOR, δεν είναι εύκολη η ανάγνωση των αποτελεσμάτων εξαιτίας της φύσης της μεταβλητής.



Σχήμα 1: Διαγράμματα διασποράς



OLS Regression Results

Dep. Variable:	Q("Life expectancy ")	R-squared:	0.833
Model:	OLS	Adj. R-squared:	0.831
Method:	Least Squares	F-statistic:	424.7
Date:	Sun, 31 Mar 2024	Prob (F-statistic):	0.00
Time:	16:11:40	Log-Likelihood:	-3944.7
No. Observations:	1463	AIC:	7925.
Df Residuals:	1445	BIC:	8021.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	341.8048	49.912	6.848	0.000	243.897	439.713
Q("Year")	-0.1444	0.025	-5.794	0.000	-0.193	-0.095
Q("Adult Mortality")	-0.0180	0.001	-17.771	0.000	-0.020	-0.016
Q("Alcohol")	-0.1317	0.033	-3.997	0.000	-0.196	-0.067
Q("percentage expenditure")	0.0003	0.000	1.479	0.139	-8.98e-05	0.001
Q("Hepatitis B")	-0.0030	0.005	-0.618	0.537	-0.012	0.007
Q("Measles ")	8.403e-06	1.13e-05	0.746	0.456	-1.37e-05	3.05e-05
Q(" BMI ")	0.0331	0.006	5.198	0.000	0.021	0.046
Q("under-five deaths ")	-0.0026	0.001	-2.815	0.005	-0.004	-0.001
Q("Polio")	0.0057	0.006	1.020	0.308	-0.005	0.017
Q("Total expenditure")	0.0935	0.043	2.167	0.030	0.009	0.178
Q("Diphtheria ")	0.0202	0.006	3.192	0.001	0.008	0.033
Q(" HIV/AIDS")	-0.4575	0.020	-23.149	0.000	-0.496	-0.419
Q("GDP")	3.476e-05	2.93e-05	1.186	0.236	-2.27e-05	9.23e-05
Q("Population")	2.966e-09	1.83e-09	1.619	0.106	-6.28e-10	6.56e-09
Q(" thinness 5-9 years")	-0.0281	0.028	-0.995	0.320	-0.083	0.027
Q("Income composition of resources")	10.9053	0.900	12.122	0.000	9.141	12.670
Q("Schooling")	0.9120	0.064	14.349	0.000	0.787	1.037

Σχήμα 2: Σύνοψη ανάλυσης πολλαπλής γραμμικής παλινδρόμησης.

4 Forward Selection

Σε αυτό το σημείο εξετάζεται το κατά πόσο η επιλογή μόνο συγκεκριμένων μεταβλητών πρόβλεψης στην απόκριση θα μπορούσε να διατηρήσει την ικανότητα του μοντέλου να εξηγήσει την μεταβλητότητα των τιμών, κρατώντας το R^2 σε υψηλές τιμές. Για την επιλογή των καταλληλότερων μεταβλητών χρησιμοποιήται η μέθοδος της πρόσω επιλογής. Αρχικά υπάρχει ένα κενό μοντέλο στο οποίο εξετάζεται η προσθήκη μεταβλητών που περνάν από ένα κριτήριο, στην προκειμένη περίπτωση p-value < 0.001. Η διαδικασία ξεκινάει από την μεταβλητή με την μεγαλύτερη συσχέτιση με την απόκριση. Αφού έχει εισαχθεί μια μεταβλητή στο μοντέλο δεν μπορεί να αφαιρεθεί στην συνέχεια. Το κριτήριο που επιλέγεται εδώ, για την προσθήκη μιας μεταβλητής στο μοντέλο, είναι το R^2. Εάν η προσθήκη μιας μεταβλητής το βελτιώνει σημαντικά τότε η μεταβλητή κρίνεται σκόπιμο να προστεθεί. Τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα, όποια μεταβλητή δεν εμφανίζεται έχει συντελεστή μηδέν. Συνολικά το R^2 είναι ίσο με 0.835 για αυτό το μοντέλο, με p-value < 0.001 καθιστώντας το στατιστικά σημαντικό. Όπως φαίνεται διατηρήθηκε η "προβλεπτική ικανότητα" του μοντέλου με μείωση των μεταβλητών από 17 σε 9.

Προβλεπτική μεταβλητή	**Συντελεστής**
Σταθερά	321.7432
Schooling	1.0530
Adult Mortality	-0.0192
HIV/AIDS	-0.4258
Income Composition of Resources	9.9186
Percentage Expenditure	0.0005
BMI	0.0341
Year	-0.1347
Diphtheria	0.0225
Alcohol	-0.1269

Πίνακας 3: Συντελεστές πολλαπλής γραμμικής παλινδρόμησης με πρόσω επιλογή.

5 Μη γραμμική παλινδρόμηση

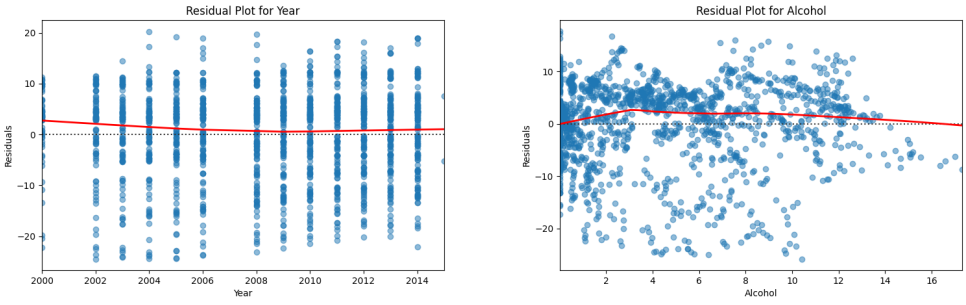
Παρατηρώντας τα residual plots για τις ξητούμενες μεταβλητές χρησιμοποιώντας τα γραμμικά μοντέλα που αναπτύχθηκαν για το πρώτο ερώτημα φαίνεται ότι για το Schooling και το BMI υπάρχει εμφανής μη γραμμική συμπεριφορά, ενώ για το Year και το Alcohol λιγότερο. Όπως αναφέρθηκε και νωρίτερα, δεν υπάρχει κάποια ένδειξη για εξάρτηση του Life expectancy από την χρονία που προκύπτουν τα δεδομένα.

Ποβλεπτική μεταβλητή	Εξίσωση
Year	$Y = 0.109 + 72.739X - 0.0725X^2 + 1.806 \times 10^{-5}X^3$
Alcohol	$Y = 65.638 + 0.423X + 0.0901X^2 - 0.0044X^3$
BMI	$Y = 73.809 - 1.353X + 0.0437X^2 - 0.0003X^3$
Schooling	$Y = 81.052 - 8.596X + 0.936X^2 - 0.0254X^3$

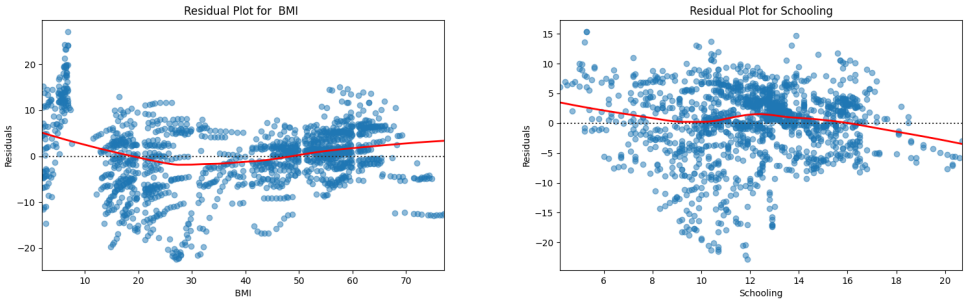
Πίνακας 4: Εξισώσεις μη γραμμικής παλινδρόμησης.

Προβλεπτική μεταβλητή	R^2	p-value τάξη μεγέθους
Year	0.007	0.00347
Alcohol	0.163	10^{-63}
BMI	0.430	10^{-200}
Schooling	0.551	10^{-285}

Πίνακας 5: Αξιολόγηση πολυονυμικής παλινδόμησης.

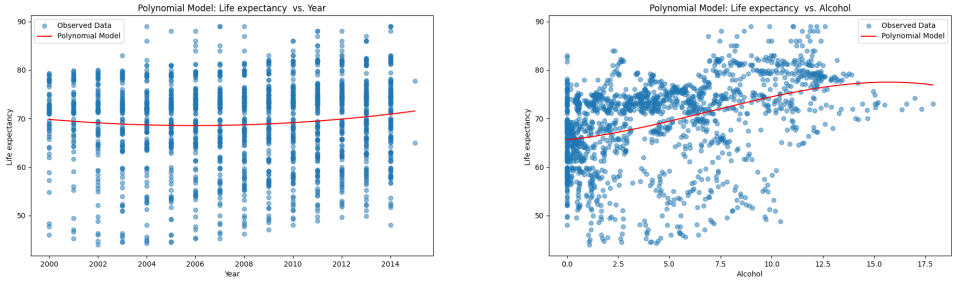


Σχήμα 3: Residual plots.

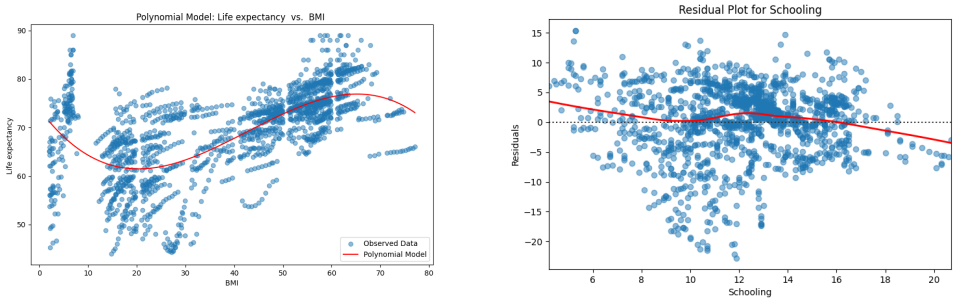


Σχήμα 4: Residual plots συνέχεια.

Όλα τα μοντέλα είναι στατιστικά σημαντικά, με p-values σαφώς μικρότερα από 0.001, με εξαίρεση την προβλεπτική μεταβλητή Year, το μοντέλο της οποίας είναι αρκετά σημαντικό στατιστικά. Σχηματίζοντας τα πολυονυμικά μοντέλα μπορεί να φανεί ότι υπάρχει μια βελτίωση στο fit. Όσον αφορά την προβλεπτική μεταβλητή Year, όπως και με το γραμμικό μοντέλο δεν φαίνεται να υπάρχει κάποια ουσιαστική συσχέτιση με την απόκριση. Η κατανάλωση αλκόολ εξηγεί 16.3% ελαφρώς λιγότερο από ότι το γραμμικό μοντέλο, ενώ το μοντέλο των ολοκληρωμένων σχολικών χρόνων εμφανίζει R^2 ίσο με 55.1% ελαφρώς αυξημένο από το γραμμικό μοντέλο. Ουσιαστική διαφορά εμφανίζεται στο BMI, όπου το πολυωνυμικό μοντέλο είναι βελτιωμένο κατά περίπου 50% από $R^2 = 0.29$ στο γραμμικό, σε 43% εδώ. Κρίνοντας από το γεγονός ότι η αύξηση της τάξης του μοντέλο αυξάνει σημαντικά την πολυπλοκότητα του μοντέλου, κρίνεται σκόπιμη η υιοθέτηση μόνο του μοντέλου για το BMI. Έχοντας πει αυτό βέβαια, το πολυωνυμικό μοντέλο για το Schooling είναι επίσης καλύτερο. Αξιοσημείωτο είναι ότι για όλα τα πολυονυμικά μοντέλα με εξαίρεση το Schooling, ο όρος στην 3ης τάξης είναι σαφώς μικρότερος από αυτόν 2ης τάξης.



Σχήμα 5: Μη γραμμικά μοντέλα.



Σχήμα 6: Συνέχεια μη γραμμικών μοντέλων.