Regression Analysis - Project

Vincent Buekers - r0754046

15th January 2019

Contents

1 Hospital Data

For the first analysis, data on hospitals in the United States are examined. A sample of 113 hospitals has been collected regarding 9 variables. Only the following variables are included in the analysis based on variable selection procedures discussed in section 1.1:

- lstay: Length of stay at the hospital
- irisk: Average estimated probability of acquiring infection in hospital (in percent)
- cultratio: Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection x 100
- xrayratio: Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia x 100
- facil: Percent of 35 potential facilities and services that are provided by the hospital

The objective of this analysis is to explain the probability of acquiring an infection in a hospital based on the variables included in this data set. Based on training data, which consists of 93 out of the 113 observations, various variable selection methods are used to retain the relevant variables. An appropriate linear model is then constructed, based on which predictions are made for the 20 left out observations in the validation data. Finally, a ridge regression is also conducted to compare its predictive performance with that of the linear model.

1.1 Variable Selection

As mentioned above, the first step in this analysis is to identify the relevant variables with respect to the response variable irisk. Backward and forward elimination based on the AIC both selected the variables lstay, cultratio, xrayratio and facil. The obtained AIC for each is -8.14. Stepwise selection has exactly the same outcome, regardless of starting from the full model (including all predictors) or the null model (intercept only). However, backward and forward elimination based on the F-statistic did not include the variable *xraytratio*. With this procedure, the included variables are thus *lstay*, *cultratio* and *facil*. These procedures translate into the following model equations:

Model 1:

$$irisk = \beta_0 + \beta_1 lstay + \beta_2 cultratio + \beta_3 xrayratio + \beta_4 facil + \epsilon \tag{1}$$

Model 2:

$$irisk = \beta_0 + \beta_1 lstay + \beta_2 cultratio + \beta_3 facil + \epsilon \tag{2}$$

When comparing these two models (table 1.1), model 1 just turns out slightly better than model 2 in terms of PRESS/n and MSE, while model 2 has a slightly lower MSEP. However, these differences are very minor. Hence, the added value of *xrayratio* is questionable. Moreover, it was also found individually insignificant based on preliminary regression results (P=0.124).

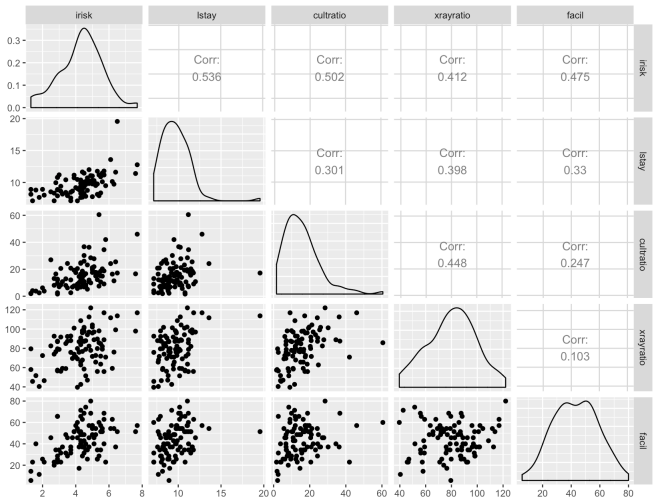
Table 1.1: Comparative Measures

	Model 1	Model 2
PRESS/n	0.933	0.938
MSE	0.870	0.883
MSEP	2.586	2.576

1.2 Exploratory Analysis

When looking at the plot provided in figure 1.1, some preliminary ideas can already be taken away. For instance, both lstay and cultratio have heavily right skewed distributions. They do however, seem to be a linearly related to the response variable irisk based on their scatter plots. As was assumed in the preceding section, the response irisk does not seem to be very dependent on the variable xrayratio as the data points are quite scattered without a clear pattern. This strengthens the preference for model (2) as opposed to model (1). Furthermore, the correlations between the predictors are very acceptable so there ought to be no multicollinearity issues in the subsequent regression modelling.

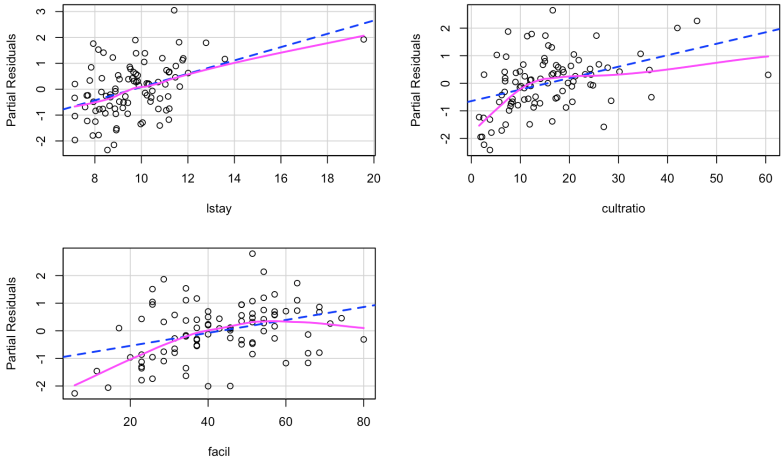
Figure 1.1: Exploratory Plot



1.3 Linear Regression

When conducting a linear regression based on modem equation (2), the residual plots (Appendix 1.1) show that a higher order term could improve the fit since curvature is detected. To now determine whether or not add higher order terms or transformations of the aforementioned predictors, partial residual plots can offer guidance. Figure 1.2 displays these partial residuals, along with their least squares slope estimates.

Figure 1.2: Partial Residual Plots



Applying a log-transformation to *cultratio* seems suitable, while the partial residual plot for *facil* suggests adding a quadratic term. However, the variable *facil* is a proportion (in percentages). Hence, a logit transformation might be more appropriate as power transformations are not very helpful for proportions. This corresponds to the following transformation $log(facil/(1-facil)$. The variable *lstay* does not need to undergo a transformation at first glance.

Hence, the resulting linear model is as follows:

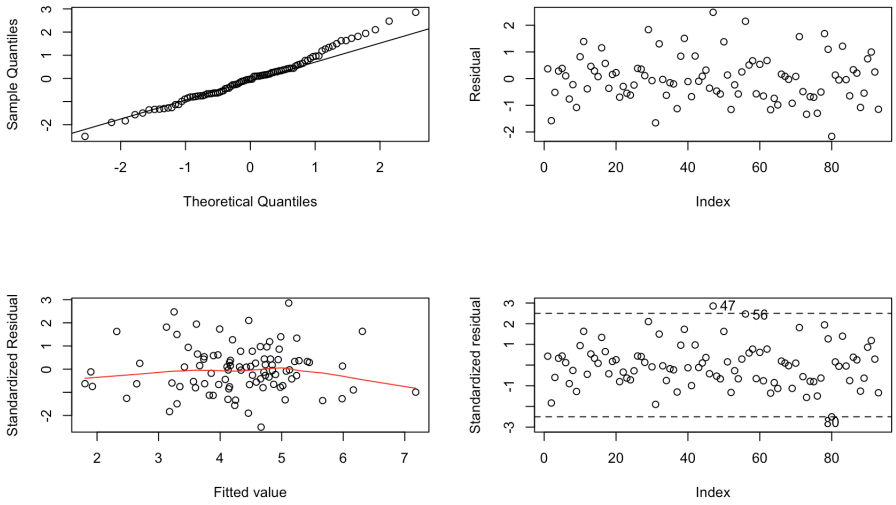
$$irisk = \beta_0 + \beta_1 lstay + \beta_2 log(cultratio) + \beta_3 log(facil/1 - facil) + \epsilon \tag{3}$$

1.4 Model Validation

In order to reliably infer from the obtained regression results, it needs to be verified if the Gauss-Markov conditions are upheld. To that end, the residual plots in figure 1.3 are examined.

- As shown in the QQ-plot, there is some deviation from normality among the standardized residuals at the right tail of the distribution. The magnitude of these deviations is quite contained however, so this should be acceptable. Indeed, the Shapiro-Wilk normality test did not reject normality (p = 0.4696).
- Moreover, the residuals also seem to be uncorrelated based on the plot of the residuals versus their index as there is no clear pattern.
- While there remains some slight curvature, the plot of the standardized residuals versus their fitted values does seem to approximately satisfy the assumption which states that the residuals should be symmetric around zero with constant variance. Nevertheless, a weighted least squares regression has been considered in order to possibly obtain a more satisfactory residual plot. This did not yield any improvement whatsoever.
- There is one clear outlier, namely observation 47. Observations 56 and 80 are on the verge of exceeding the theoretical bounds [-2,5; 2,5]. Yet, these outliers should not call for concern.

Figure 1.3: Model Diagnostics



1.5 Inference

Now that a supposedly valid model has been obtained, the next step in this analysis is to interpret the results obtained from the regression model based on (3). The model output is summarized in table 1.2 below. Clearly, all predictors are considered significant both individually (based on the t-values) and jointly (based on the F-value). About 55% of the total variability in the data is accounted or by this regression model. The penalization inherent in the adjusted R-squared is very mild as the amount of predictors is very reasonable. Based on the coefficient estimates the following conclusions can be drawn. Note that since *irisk* is expressed as a percentage, the numbers below are also expressed as such.

- On average, and all else equal, the probability of acquiring an infection increases as the length of stay increases. A unit increase (so staying at the hospital for an additional day), changes the expected infection probability by +.24%.
- On average, and all else equal, the probability of acquiring an infection increases as the amount of cultures performed on a patient increases. For instance, if the amount of cultures performed would be doubled, the infection probability would be expected to increase by about 0.14*log(2)% = 0.1%
- On average, and all else equal, the probability of acquiring an infection increases as more facilities are used during the hospital stay. The exact influence is somewhat difficult to express due to the logit transformation.

Table 1.2: Model Output

	Estimate	Standard error	t-value	Pr($>\|t\|$)
Intercept	0.14731	0.62320	0.236	0.81368
lstay	0.24989	0.05746	4.349	3.63e-05 ***
log(cultratio)	0.14163	0.033	5.260	9.84e-07 ***
logit(facil)	0.14440	0.024	2.862	0.00524 **

$R^2 = 0.5461$	$R^2_a = 0.5308$	F-statistic = 35.69	Pr(>F): 3.071e-15

1.6 Prediction

To evaluate the predictive performance of model (3), predictions are made for the 20 left out observations in the validation data set. Based on the general prediction procedure, an R^2 value of about 0.51% is obtained. The mean squared error of prediction (MSEP) is about 1.190928.

1.7 Ridge Regression

A ridge regression is considered as a possible alternative to the previously evaluated linear model. On one hand, a ridge regression with the predictors obtained by the variable selection procedure in 1.1 is evaluated. This model thus consists of the predictors lstay, cultratio and facil and is referred to as "Ridge Reduced". On the other hand, all the other variables from the hospital data are added, and is referred to as "Ridge Full". The penalty parameter is selected by means of cross-validation. The predictive performance of both these ridge regressions are summarized in table 1.3. Clearly both the difference in R^2 and MSEP is very minor. Hence, the additional variables do not offer much incremental predictive power with regards to infection probability.

Table 1.3: Predictive Performance - Ridge

	Ridge Reduced	Ridge Full
R^2	0.490	0.515
MSEP	0.907	0.864

Lastly, these results can also be compared with the linear model (3) discussed in the previous sections. In terms of the R^2 the differences are once again very minor. Yet, the predictive performance as measured by the MSEP, is noticeably better for both the ridge regressions.

2 Fuel Data

For the second analysis, observations on motor fuel consumption across all 51 states of the US are studied. These data were recorded in 2001 with the inclusion of the following variables:

- Income: Per capita personal income/1000
- Miles: Miles of Federal-aid highway miles in the state
- Tax: Gasoline state tax rate in cents per gallon
- Dlic: Ratio of licenced drivers over the population of the state x 1000
- Fuel: Gasoline sold for road use in 1000s of gallons

The objective of this analysis to examine how fuel consumption is to the remaining variables, which will serve as predictors in the subsequent analysis. To do so, both classical and robust techniques are used to obtain an appropriate model and to detect possible outliers with each of these methods.

2.1 Classical Analysis

The variables measured in the fuel data all differ substantially in magnitude as they are all measured in different units. As classical outlier detection methods take into account distances, it would seem appropriate to standardize the data. Consequently, the correlation transformation is applied to the aforementioned variables. A standardized linear model with *Fuel* as the response and the remaining variables as predictors corresponds to the following model equation:

$$Fuel' = \beta_1' Income' + \beta_2' Miles' + \beta_3' Tax' + \beta_4' Dlic' + \epsilon' \tag{4}$$

Yielding the following output:

Table 2.1: Model Output

	Estimate	Standard error	t-value	Pr(> |t|)
Income	-0.3572	0.1092	-3.271	0.002009 **
Miles	0.2392	0.1103	2.168	0.035287 *
Tax	-0.2102	0.1065	-1.974	0.054326 .
Dlic	0.4384	0.1113	3.940	0.000269 ***

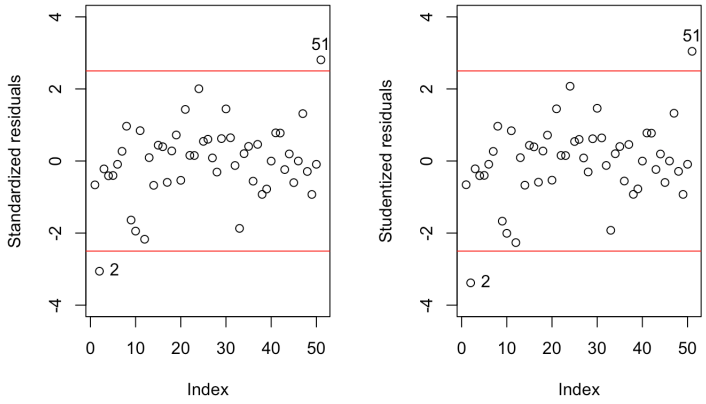
$R^2 = 0.4755$	$R_a^2 = 0.4309$	F-statistic = 10.65	Pr(>F): 3.153e-06

This standardized linear model explains about 47% of the variability in the fuel data as indicated by the R^2. Moreover, the model is considered globally significant by the F-statistic. Both income and tax are inversely related to fuel consumption, while the ratio of licensed drivers and the miles of Federal-aid highway seems to increase the fuel consumption in a given state. Note that the predictor Tax is only significant at the 10% level.

2.1.1 Vertical Outliers

Since the least squares estimator is heavily influenced by vertical outliers, it is important to detect and classify these observations as such. To that end, both the standardized and the deleted residuals are examined. Figure 2.1 illustrates that there are two observations exceeding the bounds [-2,5; 2,5], namely 2 and 51. The studentized residuals, computed based on the deleted residuals, identify the same two observations as vertical outliers.

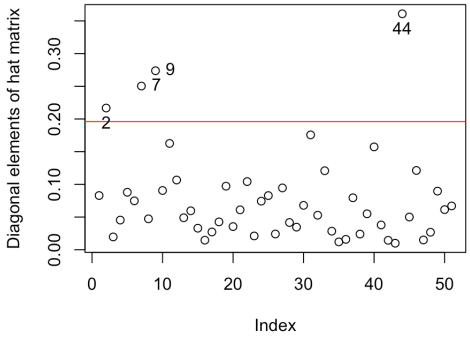
Figure 2.1: Standardized and Studentized LS residuals



2.1.2 Leverage Points

Residual plots are unable to distinguish between leverage points and vertical outliers. With that in mind, the diagonal elements of the hat matrix are analyzed. The diagonal elements of such matrices measure both the effect of the i-th observation on its own prediction and the distance to the center of the data points in the predictor-space. These elements are plotted in Figure 2.2, from which leverage points 2, 7, 9 and 44 can be identified. With this classical approach, it is not possible to seperate the good leverage points from the bad ones.

Figure 2.2: Diagonal elements of Hat matrix



2.1.3 Single Case Diagnostics

Now that the outlying observations with respect to the response and predictor space are possibly identified, the influence of these observations on the regression fit can still be assessed. To this end, the single case diagnostics DFFITS, DFBETAS and Cook's distance are considered.

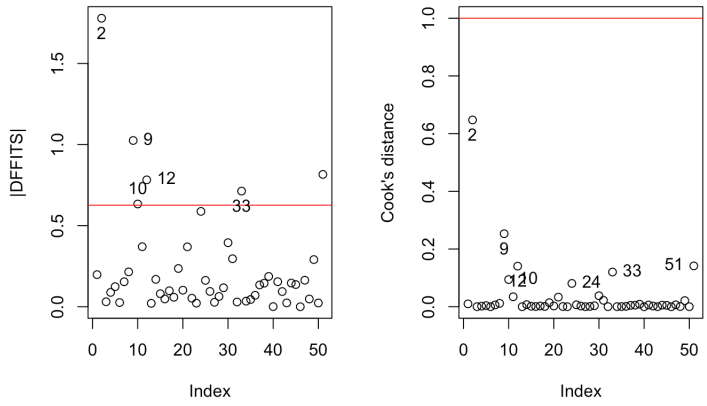
Firstly, the DFBETAS measures the influence of an observation on the regression coefficients. The observations that are considered influential for each regression coefficient are summarized in table 2.2 below. It is clear that observations 2, 9, 10, 12, 33 and 51 are consistently pointed out as influential observations in terms of the regression coefficients. Observation 11 is only an outlier with respect to Tax, while observation 24 is only indicated as influential with respect to the variable Dlic.

Table 2.2: DFBETAS

	Influential observations
Income	9, 33
Miles	2, 9, 10, 12, 51
Tax	2, 10, 11, 12, 51
Dlic	2, 9, 10, 12, 24, 33

Secondly, The DFFITS diagnostic and Cook's distance can be examined visually. Figure 2.3 shows that according to both these measures, observations 2, 9, 10, 12, 33 and 51 can be considered influential observations with respect to the fitted values. Observation 24 is only detected based on Cook's distance, yet is is also a borderline case in the DFFITS plot.

Figure 2.3: DFFITS and Cook's Distance



2.2 Robust Analysis

As the least squares (LS) estimator is quite sensitive to vertical outliers and bad leverage points, a reweighted least trimmed squares (LTS) estimator is conducted with the variables from (4) in order to identify the outliers in a robust manner. Note that a breakdown value of 25% has been chosen. This should guarantee proper resistance of the estimator towards outliers while also being reliably efficient. Furthermore, The variables have also been standardized as with the classical approach.

Table 2.3: Model Output

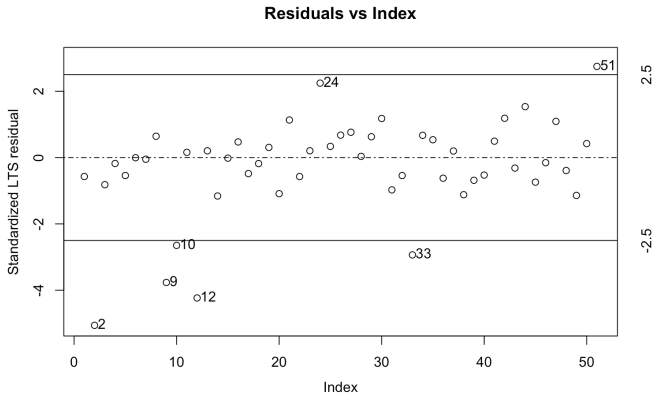
	Estimate	Standard error	t-value	Pr($>\|t\|$)
Income	-0.28486	0.06462	-4.408	7.63e-05 ***
Miles	0.10502	0.07055	1.488	0.144481
Tax	-0.37623	0.06987	-5.384	3.45e-06 ***
Dlic	0.35358	0.08394	4.213	0.000139 ***

$R^2 = 0.6367$	$R^2_a = 0.6095$	F-statistic = 23.37	Pr(>F): 6.657e-09

The model output given above is largely consistent with the results of the least squares regression. Yet, a reweighted least squares regression can account for an amount of variability that is substantially larger. Whereas earlier, the predictor Tax was only considered significant at the 10% level, now the variable Miles is considered insignificant by any means. Apart from this difference, the coefficient estimates are very comparable with the least squares estimates.

To examine the ability of the LTS estimator to detect vertical outliers, a residual plot (Figure 2.2) with the standardized LTS residuals versus their index is analyzed. As is shown, the LTS estimators is able to identify the following observations as vertical outliers: 2, 9, 10, 12, 33 and 51.

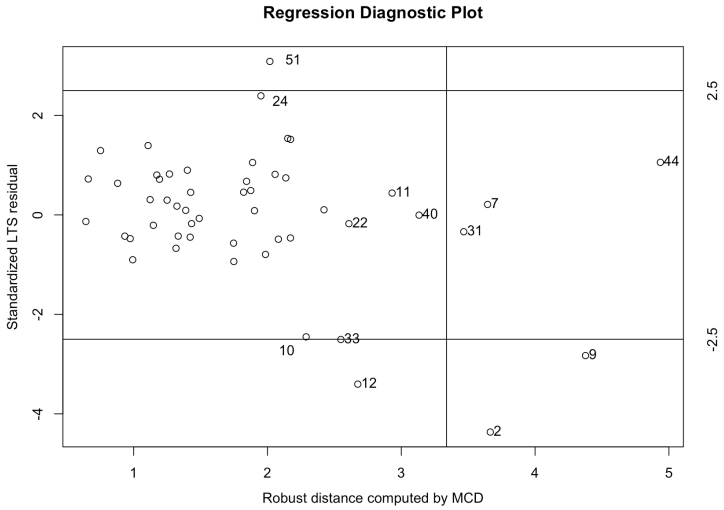
Figure 2.2: Standardized LTS residuals versus Index



However, based on Figure 2.2 it is not possible to distinguish between leverage points and vertical outliers. Since the covariance structure of the data is not taken into account yet, the Mahalanobis distance needs to be considered. When comparing these Mahalanobis distances to the robust distances in a diagnostic plot, different types of outliers can be identified. Such a diagnostic plot is given in figure 2.3.

- Vertical outliers: 10, 12, 33 and 51
- Bad leverage points: 2 and 9
- Good leverage points: 7, 31 and 44

Figure 2.3: Diagnostic Plot



In the end, the only outliers masked in the classical analysis are observation 31 and 44, which are good outliers. However, the benefit of using the LTS estimator is that one can classify the types of outliers in the data. Moreover, it is able to account for a larger variability in the and due to its resistance towards outliers also the coefficient estimates should be more consistent.

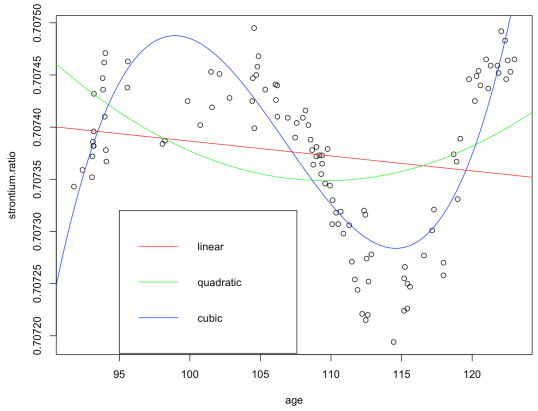
3 Fossil Data

For the third analysis, the given data contains 106 observations on fossils. Two variables have been measured, namely the ratio of strontium isotopes and the age (in million years) of the fossils. The objective of this analysis is to model the effect of age on strontium ratios. Both parametric and nonparametric regression are considered to obtain an adequately fitting model.

3.1 Parametric Regression

To determine whether a linear, a quadratic or a cubic parametric regression has to be considered, the different model fits have been visually summarized in figure 4.1. A cubic fit is clearly the fit that most accurately follows the structure of the data. However, it still does not capture all the fluctuations in the relationship between age and strontium ratios. To that end, nonparametric alternatives are considered in the following section. Although the decision for a nonparametric fit is not ambiguous, an approximate F-test can also offer a numerical justification. When comparing a nonparametric model (section 3.2) to either a linear, a quadratic, or a cubic parametric regression model, a P-value of zero is obtained in each of these cases. Moreover, the residual plots for a cubic parametric regression model (provided in appendix 4.3.1) also show clear violations of the Gauss-Markov conditions. So regardless of the cubic model being able to approximate the data structure, it can not be considered valid.

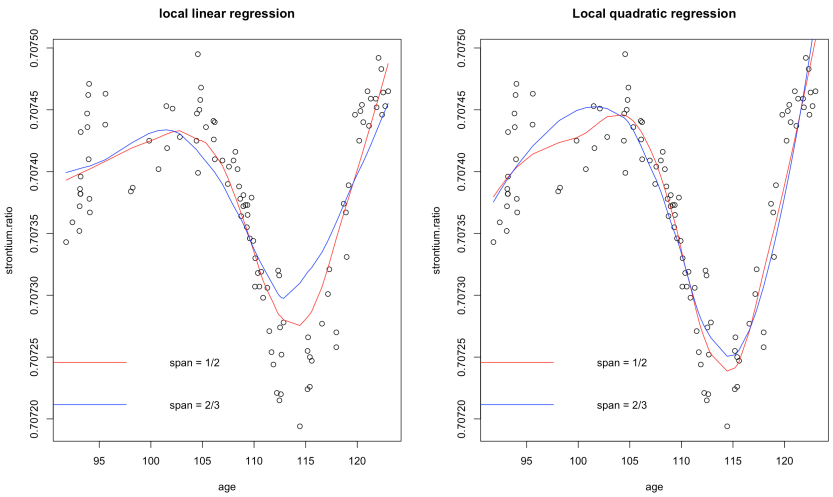
Figure 3.1: Parametric Slopes



3.2 Nonparametric Regression

Since the linear model is not appropriate nor validated, nonparamtric techniques have to be implemented. To do so, the loess method is used to obtain both a locally weighted linear fit and a locally weighted quadratic fit. Figure 3.2 provides a graphical overview of how the span and degree affect the nonparametric fit. A span of s = 1/2 and s = 2/3 are evaluated for both the linear and quadratic function. From these locally smoothed functions, it is clear that depending on the span and the degree, one obtains very different results. A local linear fit still does quite poorly in approximating the data structure. Yet, the local quadratic fit seems very appropriate. Moreover, a local quadratic fit with a span of 2/3 has smoother results in comparison to a span of s = 1/2. Note that a quadratic smoother with a span of s = 2/3 is equivalent with 4.89 parameters, which is about the same as a fourth-degree polynomial regression.

Figure 3.2: Smoothing functions



3.3 Model Validation

Next, the normality of the residuals as well as their constant variance need to be verified. The residual plots added in appendix (4.3.2) show that the residuals are uncorrelated and have a constant variance fluctuating around zero. Furthermore, the normality assumption seems to be satisfied as well, except for an acceptable minor deviation at the left tail of the distribution.

3.4 Model Fit & Interpretation

The chosen method with regards to the fossil data is a quadratic smoother with a span of s = 2/3. Figure 3.3 visualizes the relationship between the age of a fossil and its strontium ratio. Between the age of 105 and 115 there seems to be a very steep decline in strontium ratios. For the consecutive interval between 115 and 125, strontium ratios increase rapidly. The fluctuation seems more contained for fossils that are less old, say between the age of 90 and 105.

Figure 3.3: Nonparametric Model Fit