Essentials for Normal Error Linear Regression Model

Denis Ostroushko

2022-12-25

Contents

1 Introduction

This document is intended as a, hopefully, detailed guide to regression analysis in R. In particular, I present a step by step guide to develop a Normal Error Regression Model (NERM). Other people may call it a Gaussian regression model, a multiple regression model, or any other number of names. I intend to include as munch theory and intuition as possible in each section. There are three main reason to do so for me

1. This document will serve as study guide for me. I took a regression based course three times by now, twice as an undergraduate student in Fall 2016 and Spring 2019 at the University of Minnesota Morris campus, and once in Fall 2022 as a graduate student at the University of Minnesota Twin Cities campus. Here, I combine all accumulated methods and knowledge I collected over the years. There are certain methods I always have to look up, or google when I work with regression models, and hopefully a guide written by me for me can be the best reference.
2. As a guide, I intend to use this file when I prepare for my preliminary exam in May of 2023, after I finish the first year of the MS program.
3. While writing this guide I push myself to use `git` as much as possible, something I intended to do for a while.

Please refer to a table of context to find of topic of interest. Each section should have the following parts:

- If an R package is used, I introduce the package and document functions that I used. Will follow an informal format
- An intuitive explanation of the method, and a formal one, if it is applicable. The level of rigor is at the level of an MS - level regression course. NAME A BOOK THAT IS USED
- An application of the method with comments

2 Exploratory Analysis

I obtained this data set as a part of PUBH 7405 Course. We used this data set for a few homework assignments. A full summary of the data set is given in Table 1. This data set contains 86 observations, which is a perfect size for an example.

Table 1:

Variable	Description	Variable Type	Unique Values
age	Age of a patient	numeric	40
gender	Gender of a Patient: 1 = female, 2 = male	numeric	2
cpd	Cigarettes Per Day consumed	numeric	18
carbon_monoxide	Carbon Monoxide measurement	numeric	24
cotinine	A derivative of Nicotine	numeric	85
nnal	a derivative of NNN, a toxin only comes from tobacco products	numeric	83

2.1 Analysis statement

For the purpose of this exercise we will use NNAL measurements as a response variable and all other variables as potential predictors. Through this exercise we will evaluate all of this variables for their predictive power, change their scale, consider higher order powers (non-linear curves), and might throw away some predictors due to low predictive power.

2.2 Univariate Analysis: Distributions

It is important to assess the shape of the distribution of predictors. There are many implications that we need to consider:

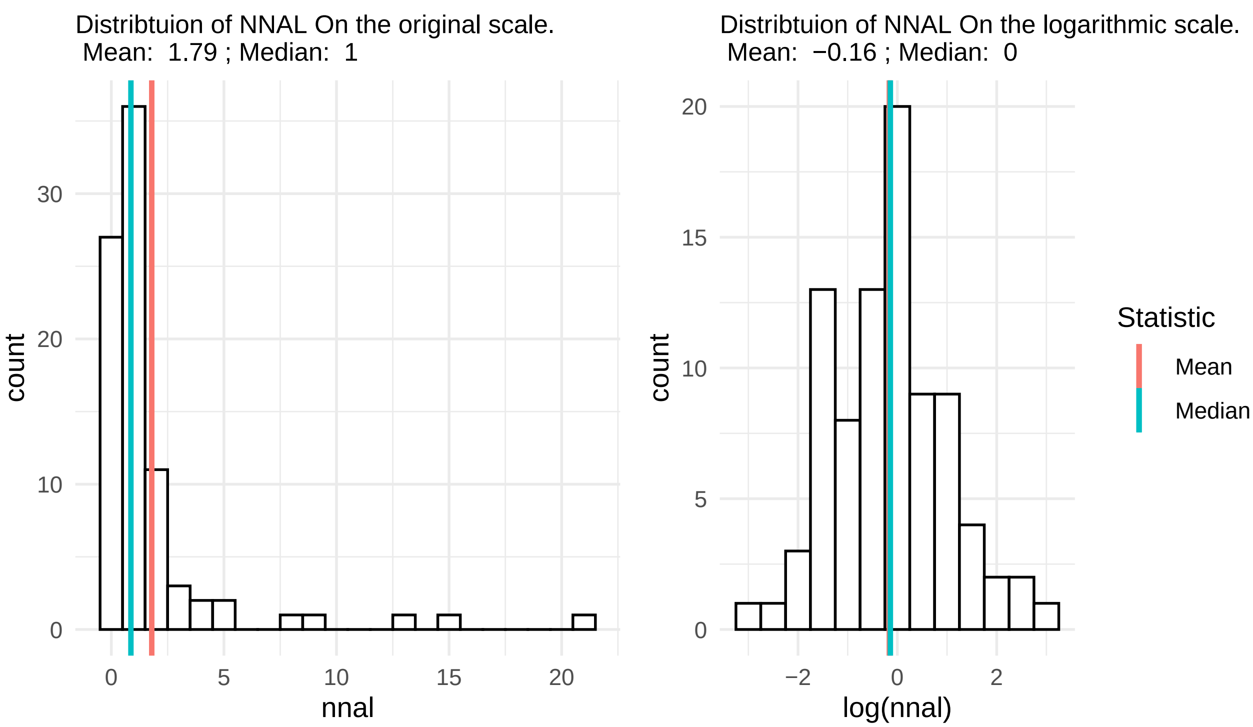
1. The distribution shape matters. Of course, in the regression context, the relationship with the response variable is far more important. However, having a distribution with a heavy tail, clusters of variables, extended tail(s), etc. Many extreme values and outliers gives you data might not fit the linear regression model well.

2. The range of values that are available to us matters. When we compute the standard errors for the regression coefficients, a part of the formula includes $\Sigma(X_i - \bar{X})^2$ in the denominator. Therefore, a high range of observations of a predictor X_i around its mean will result in a larger value of the summation term. This will make the standard error smaller.

3. We need to know the range of predictors when we try to make predictions using a developed regression model. Stepping outside of these ranges for each selected predictor constitutes extrapolation. The further we go outside of the scope that we used for model development, the more we extrapolate.

We will also see that making predictions using values of X_i further away from means of each X_i results in higher prediction error.

Response variable - NNAL - distribution



Cotinine follows similar but less extreme distribution

Predictor - CPD

We consider CPD (Cigarettes per Day) as a predictor. Looking for outliers is one of the reasons we want to visually assess the data. As we can see on Figure 1 there are some extreme outliers in the data. These values can be potentially influential on the model fit, coefficients, and other metrics/parameters we are estimating. We will keep the presence of this outlier in mind, and return to a statistical/informal evaluation in the later sections.

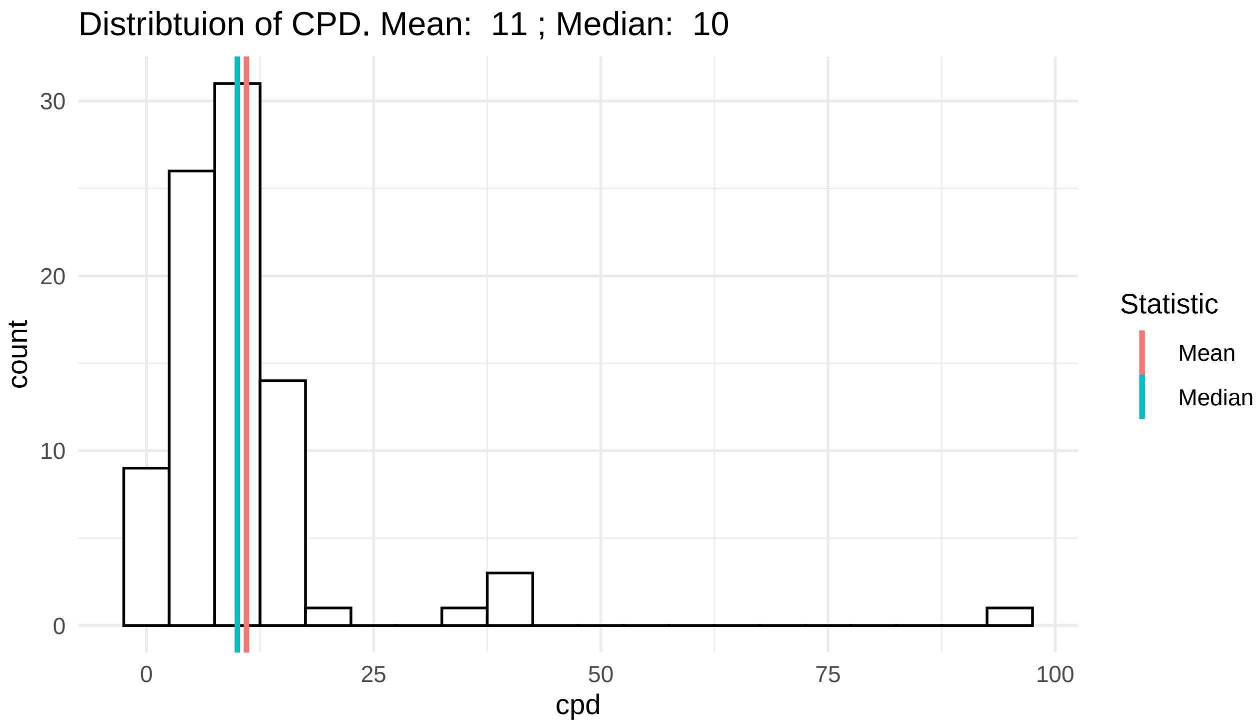
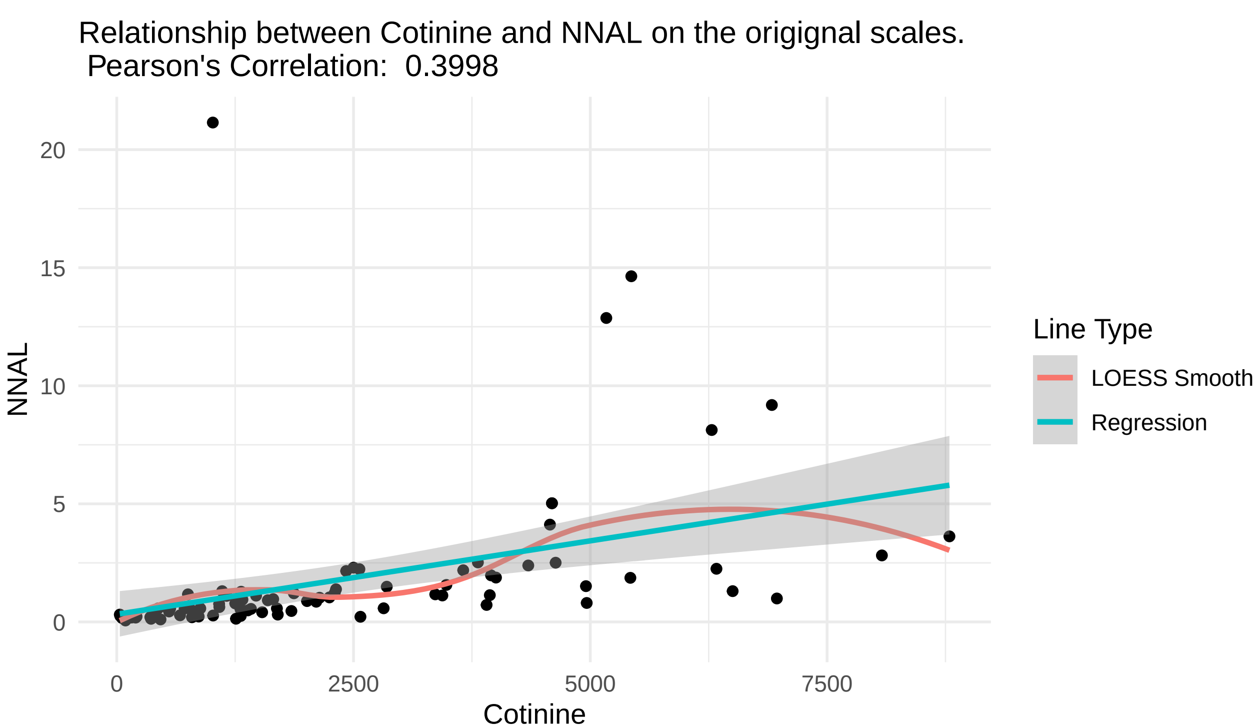


Figure 1: Distribution of CPD

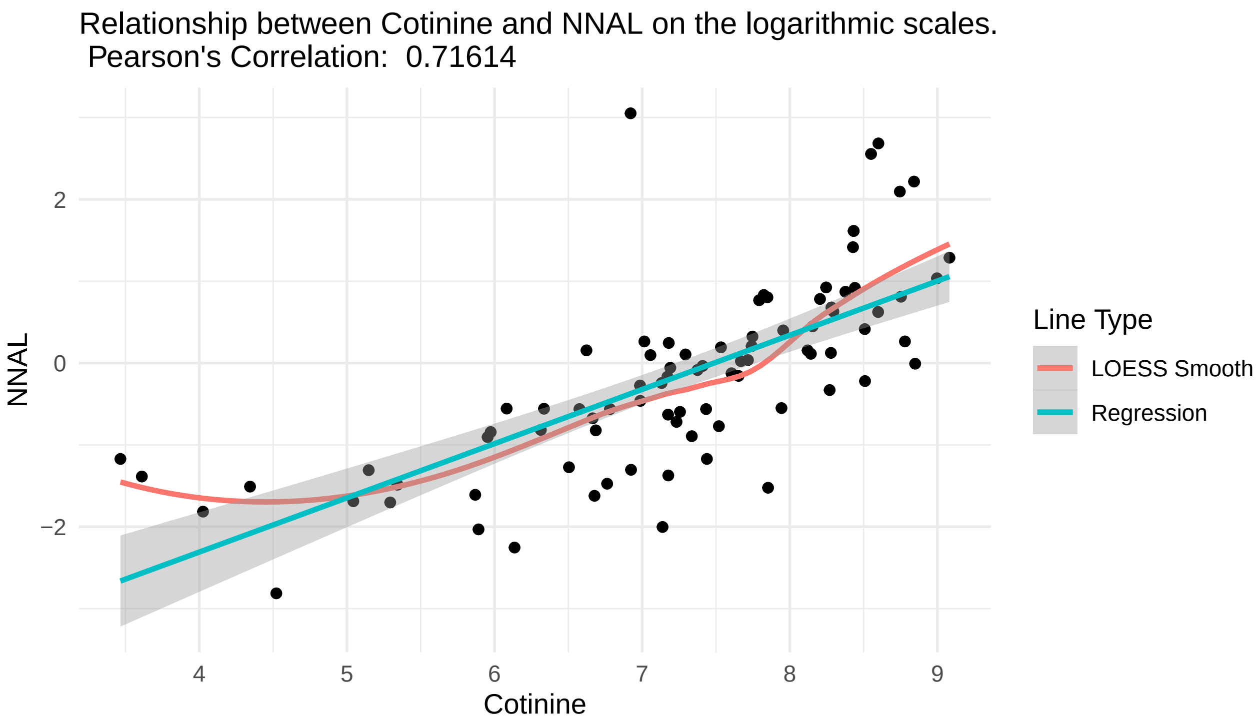
2.3 Relationship Type

Cotinine - NNAL

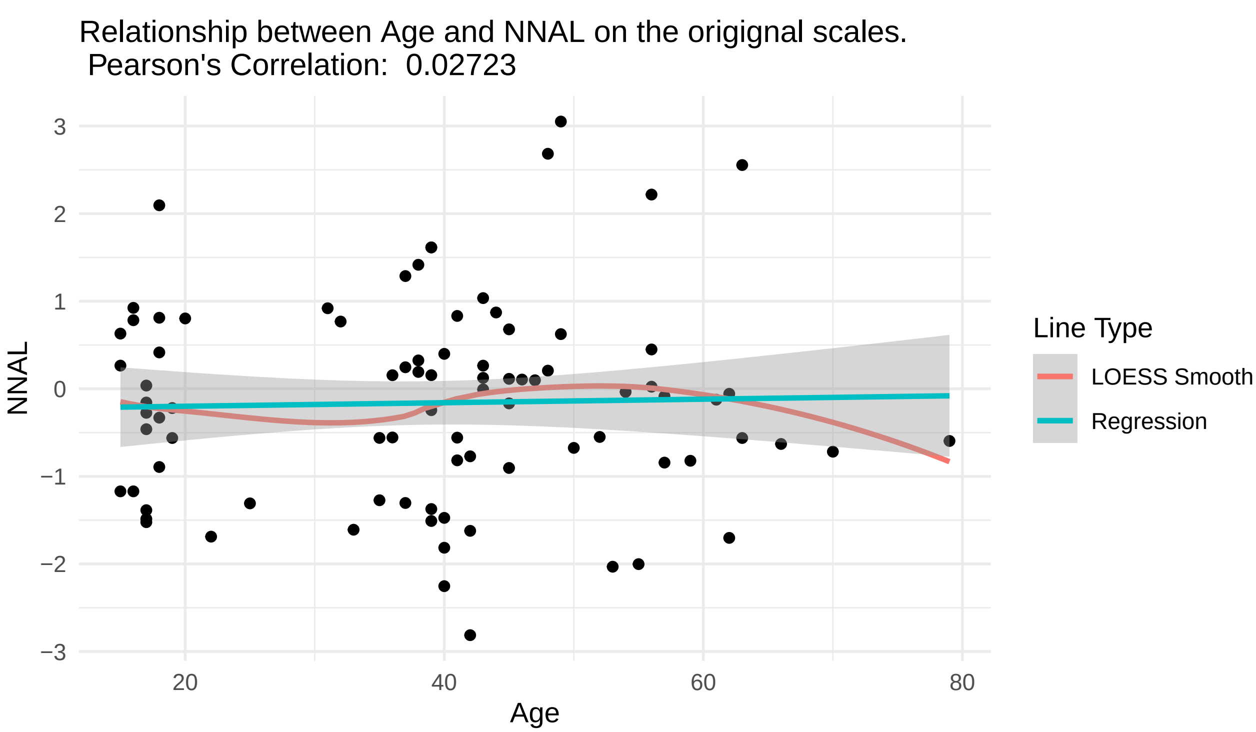
Poor fit, many outliers



Better fit, higher correlation, perhaps, better use quadratic function here. Investigate

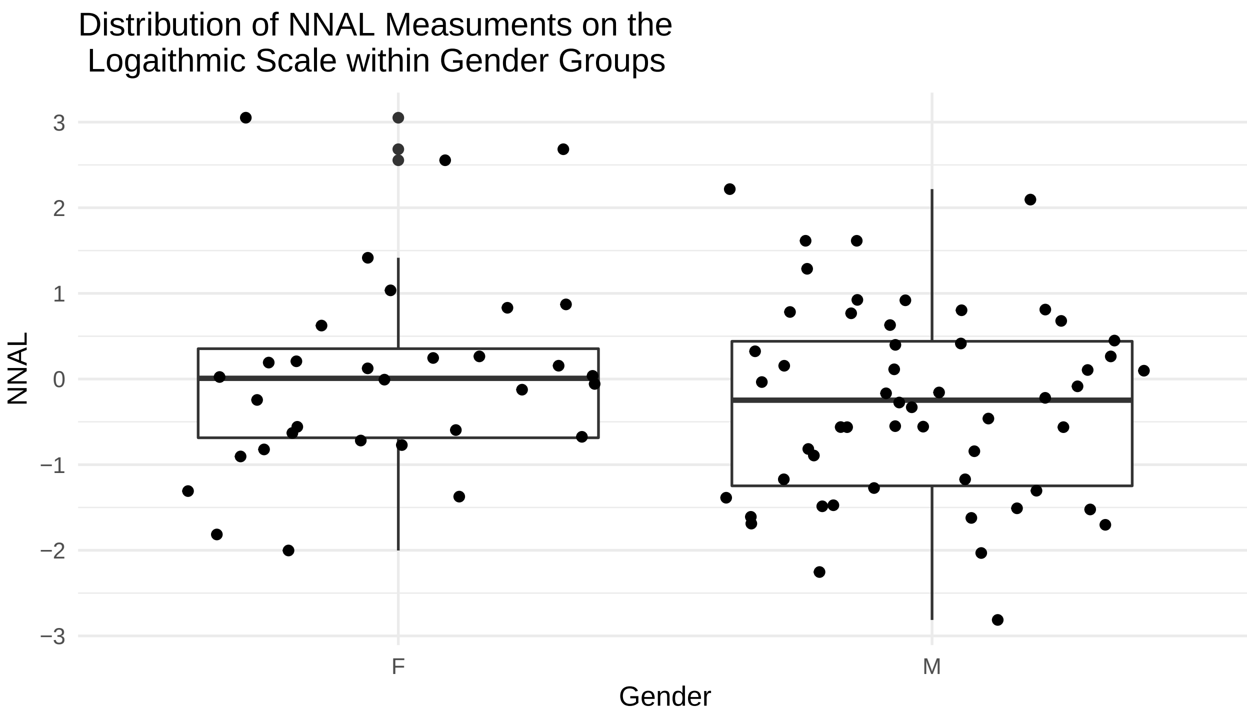


age - log NNAL



Since age is a useless predictor, dichotomize it into 10 buckets, we will need it for an example of a concept later.

gender - log NNAL

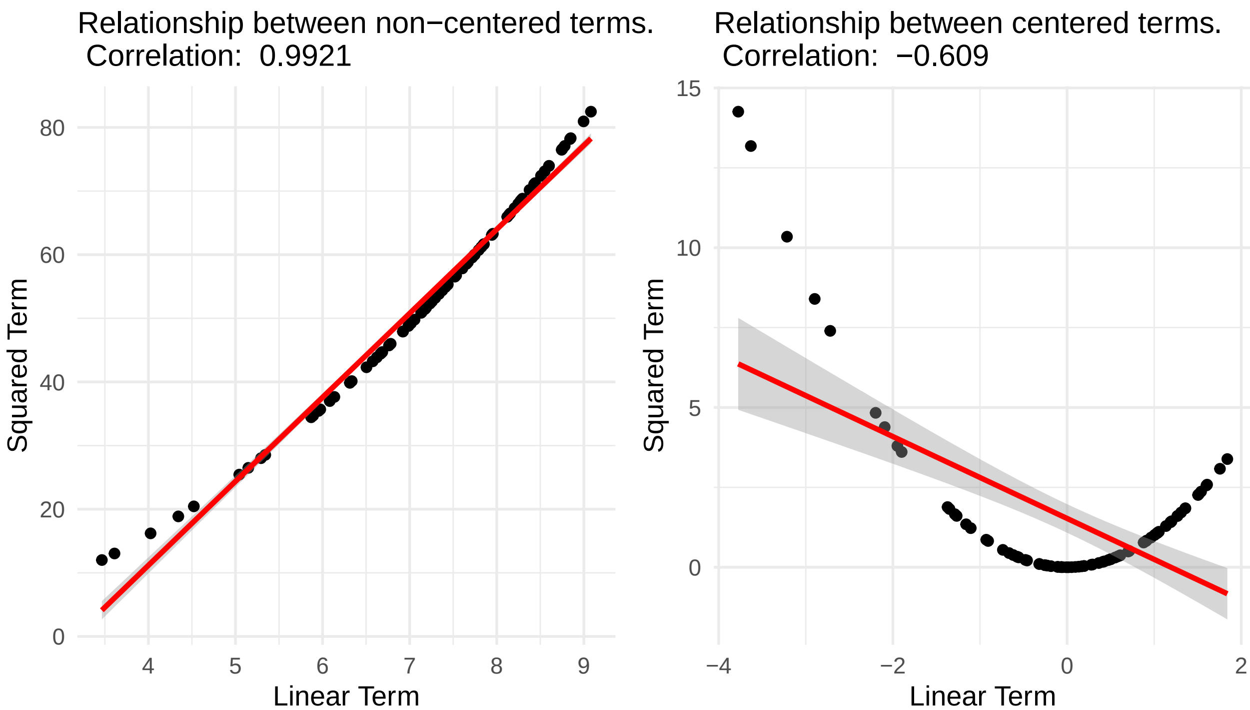


2.4 Higher Order Terms Implications

When we include a higher order term in a model our equation becomes:

$$E[Y_i] = \cdots + \hat{\beta}_i * X_i + \hat{\beta}_{i+1} * X_i^2 + \ldots$$

So, we have introduced a high degree of correlation between the two predictors now, which should increase the standard error of the $\hat{\beta}_k$ estimates. Thus, we consider a linear transformation of X_i, called centering. For example, in our problem we want to consider cotinine levels as a predictor, and we concluded that we might want to use a higher order term for this variable.



So, we consider the following steps:

1. Perform a scale transformation: we use a natural logarithm in our case. Also may consider square root, other log bases, etc..
2. Perform centering by subtracting the mean
3. Now we can include a squared term to the linear equation without multicollinearity implications

Note that in our case collinearity was not reduced drastically. Location of the mean affects this phenomena. More central mean location forces a more balanced distribution of negative and positive values of a centered variable. When we have a distribution that is skewed towards one of the signs (positive or negative), the effect of centering on correlation reduces. More skewed distribution implies lesser effect of centering on correlation between the two terms.

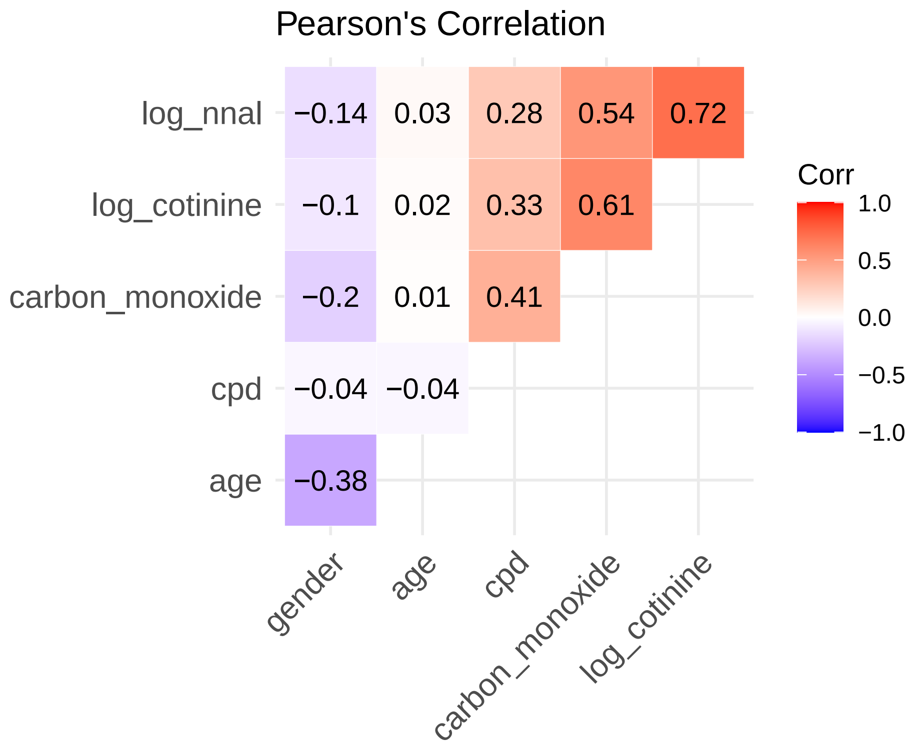
In the process of evaluating the model we will consider a number of configurations of scales and centering.

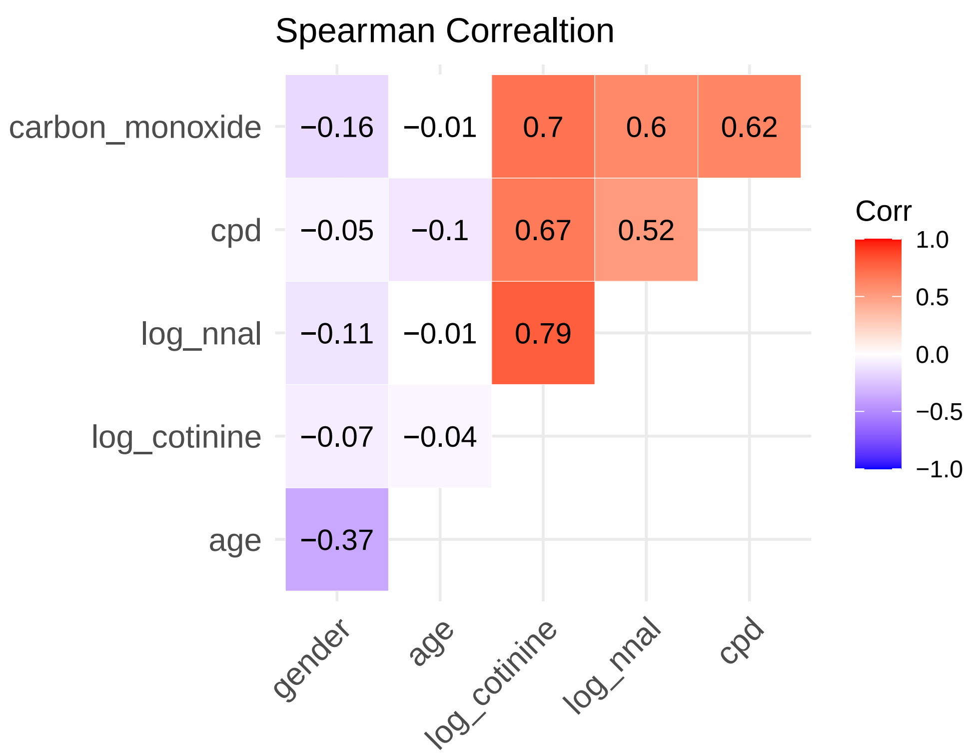
2.5 Correlation

2.5.1 Multicollinearility Issue

1. Inflates Standard Errors
2. Effect of correlated variables is split between the two variables
3. Effects that are split are not a unique solution

2.5.2 Types of Correlation Metrics





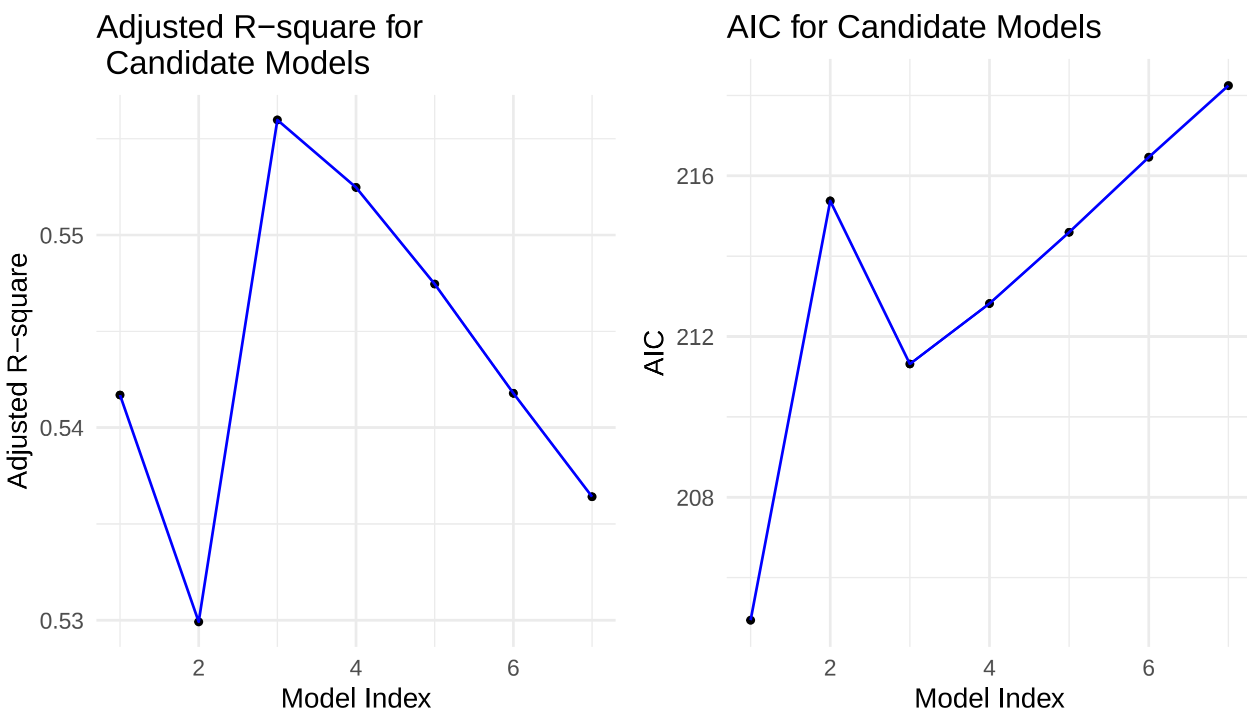
3 Model Selection

say how many possible we have

3.1 Metric Driven Approach

Table 2: Best Candidate Models

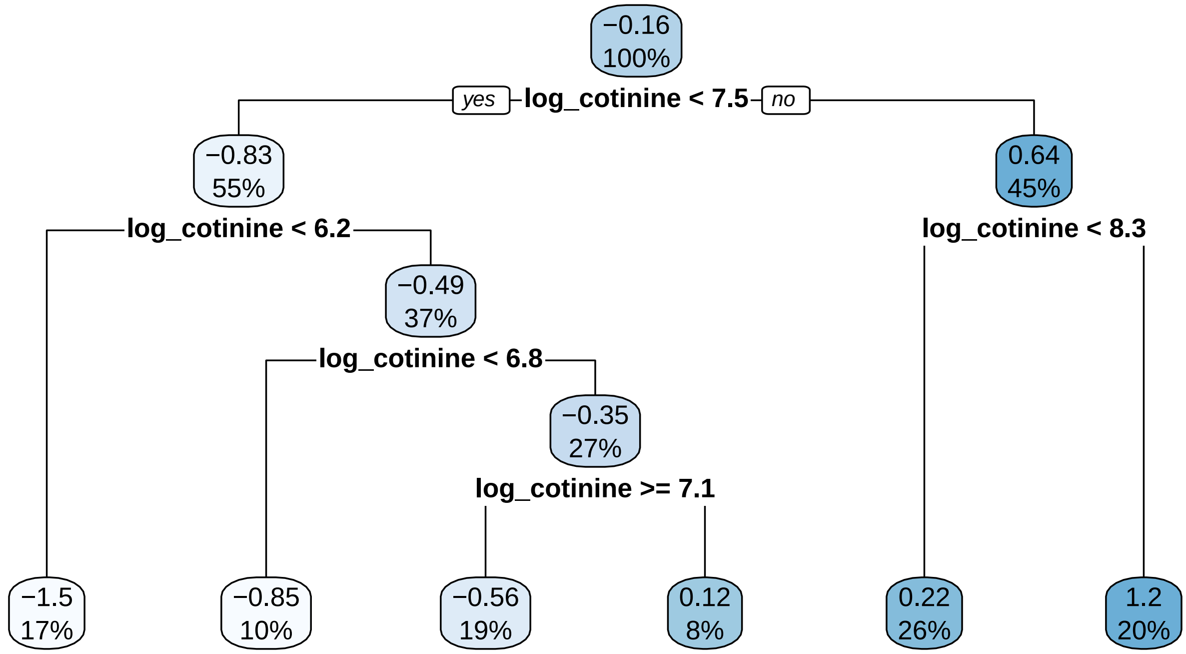
	n	predictors
6	1	I(log_cotinine^2)
12	2	age_buckets I(log_cotinine^2)
41	3	age_buckets log_cotinine I(log_cotinine^2)
71	4	age_buckets gender log_cotinine I(log_cotinine^2)
108	5	age_buckets gender log_cotinine I(log_cotinine^2) cpd:carbon_monoxide
124	6	age_buckets gender carbon_monoxide log_cotinine I(log_cotinine^2) cpd:carbon_monoxide
127	7	age_buckets gender cpd carbon_monoxide log_cotinine I(log_cotinine^2) cpd:carbon_monoxide

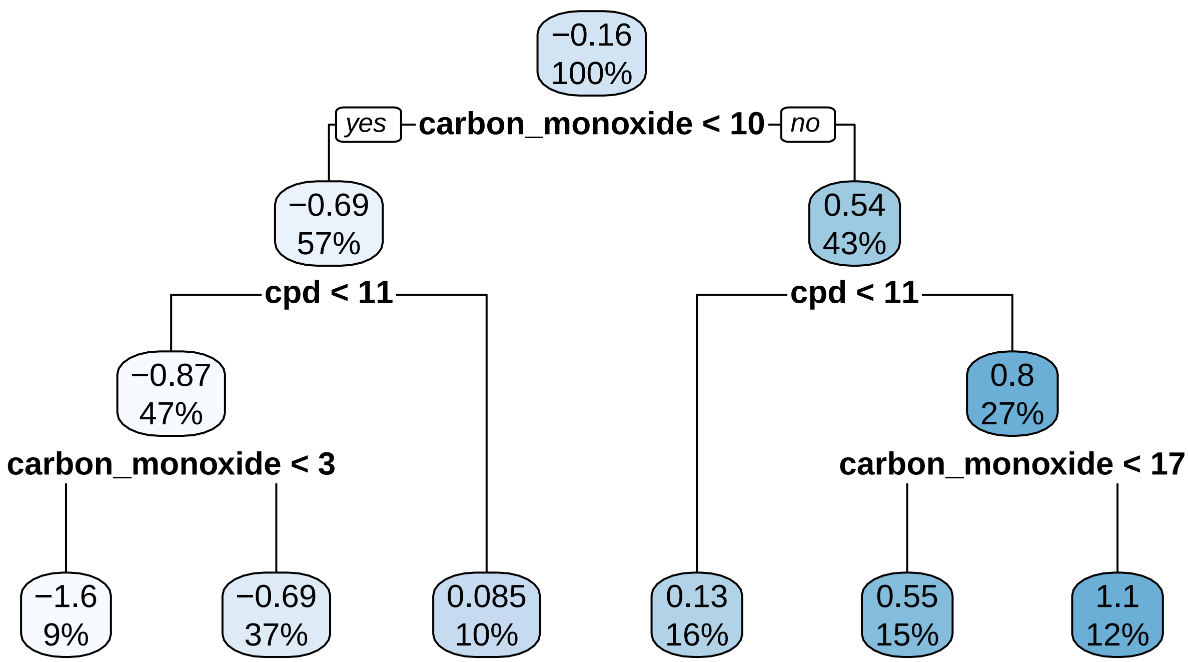


3.1.1 R-squared and Adjusted R-squared

3.1.2 AIC

3.2 Regression Trees





4 Model Evaluation

4.1 Overall F Test

explain

```
## Analysis of Variance Table
```

```
##
## Model 1: log_nnal ~ 1
## Model 2: log_nnal ~ age_buckets + gender + cpd * carbon_monoxide + log_cotinine +
##     I(log_cotinine^2)
##   Res.Df    RSS Df Sum of Sq      F         Pr(>F)
## 1     85 112.37
## 2     70  42.90 15    69.468 7.5568 0.000000001272 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

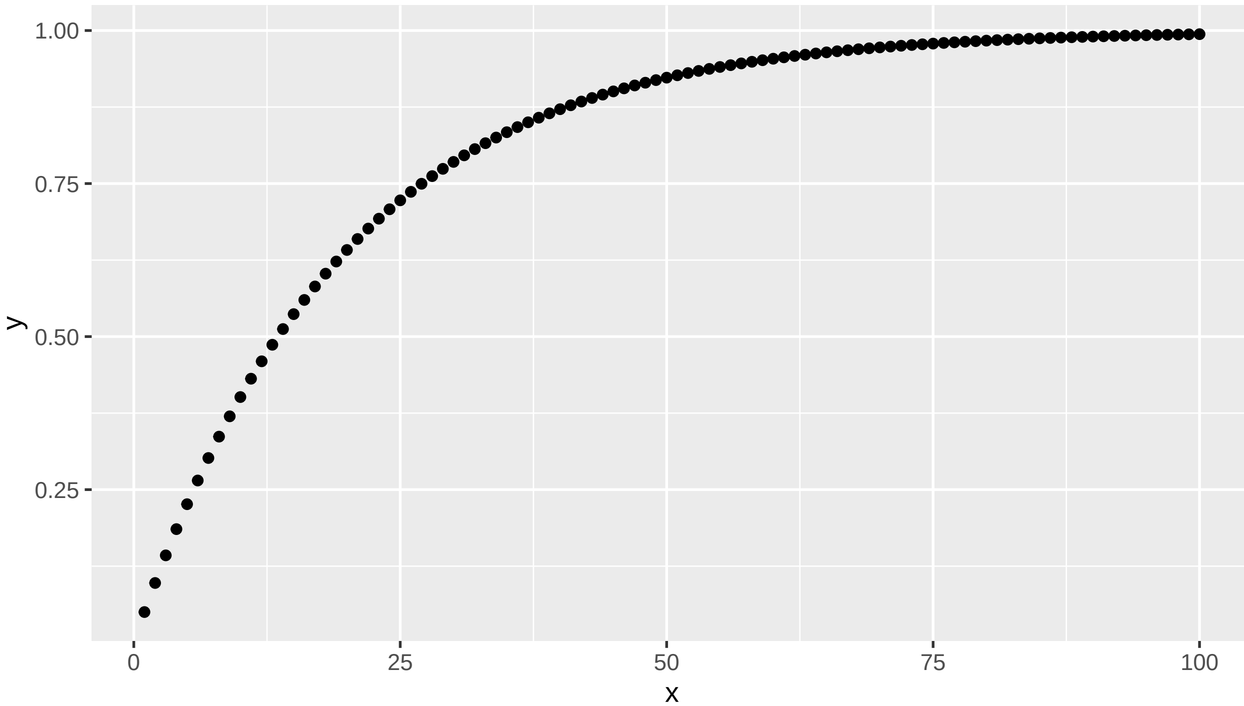
4.2 Single Predictor T Test

Predictor	Estiamte	Standard Error	Z Value	P value	Significant
(Intercept)	1.649919	2.079230	0.793524	0.430154	
age_buckets(17,18]	-0.094977	0.432586	-0.219556	0.826856	
age_buckets(18,34]	0.245754	0.357860	0.686732	0.494520	
age_buckets(34,38]	0.279175	0.346954	0.804647	0.423748	
age_buckets(38,40]	0.178167	0.338451	0.526417	0.600262	
age_buckets(40,42]	-0.047840	0.405741	-0.117909	0.906478	
age_buckets(42,45]	0.065360	0.367178	0.178006	0.859232	
age_buckets(45,52]	0.868400	0.378398	2.294938	0.024739	*
age_buckets(52,58]	0.121215	0.362724	0.334180	0.739243	
age_buckets(58, Inf]	0.191440	0.373077	0.513138	0.609470	
gender	-0.140253	0.202698	-0.691931	0.491269	
cpd	0.004291	0.010199	0.420761	0.675219	
carbon_monoxide	0.010602	0.023815	0.445194	0.657553	
log_cotinine	-1.415333	0.656184	-2.156916	0.034448	*
I(log_cotinine^2)	0.155778	0.051606	3.018583	0.003541	*
cpd:carbon_monoxide	-0.000533	0.000780	-0.683928	0.496279	

formula for probability of at least one false positive

compare single predictor t test with drop one approach F test

4.2.1 Why adjust



4.2.2 Bonferroni Adjustments

Explain how we calculate the number of predictors with the dictomized variable

Predictor	P value	Significant at Adj. Level
(Intercept)	0.430154	
age_buckets(17,18]	0.826856	
age_buckets(18,34]	0.494520	
age_buckets(34,38]	0.423748	
age_buckets(38,40]	0.600262	
age_buckets(40,42]	0.906478	
age_buckets(42,45]	0.859232	
age_buckets(45,52]	0.024739	
age_buckets(52,58]	0.739243	
age_buckets(58, Inf]	0.609470	
gender	0.491269	
cpd	0.675219	
carbon_monoxide	0.657553	
log_cotinine	0.034448	
I(log_cotinine^2)	0.003541	*
cpd:carbon_monoxide	0.496279	

4.2.3 Hochberg Adjustments

Predictor	P value	Comparison P-value	Significant at Adj. Level
age_buckets(40,42]	0.906478	0.0500000	
age_buckets(42,45]	0.859232	0.0250000	
age_buckets(17,18]	0.826856	0.0166667	
age_buckets(52,58]	0.739243	0.0125000	
cpd	0.675219	0.0100000	
carbon_monoxide	0.657553	0.0083333	
age_buckets(58, Inf]	0.609470	0.0071429	
age_buckets(38,40]	0.600262	0.0062500	
cpd:carbon_monoxide	0.496279	0.0055556	
age_buckets(18,34]	0.494520	0.0050000	
gender	0.491269	0.0045455	
(Intercept)	0.430154	0.0041667	
age_buckets(34,38]	0.423748	0.0038462	
log_cotinine	0.034448	0.0035714	
age_buckets(45,52]	0.024739	0.0033333	
I(log_cotinine^2)	0.003541	0.0031250	

4.2.4 Holm Adjustments

Predictor	P value	Comparison P-value	Significant at Adj. Level
I(log_cotinine^2)	0.003541	0.0035714	*
age_buckets(45,52]	0.024739	0.0038462	
log_cotinine	0.034448	0.0038462	
age_buckets(34,38]	0.423748	0.0038462	
(Intercept)	0.430154	0.0038462	
gender	0.491269	0.0038462	
age_buckets(18,34]	0.494520	0.0038462	
cpd:carbon_monoxide	0.496279	0.0038462	
age_buckets(38,40]	0.600262	0.0038462	
age_buckets(58, Inf]	0.609470	0.0038462	
carbon_monoxide	0.657553	0.0038462	
cpd	0.675219	0.0038462	
age_buckets(52,58]	0.739243	0.0038462	
age_buckets(17,18]	0.826856	0.0038462	
age_buckets(42,45]	0.859232	0.0038462	
age_buckets(40,42]	0.906478	0.0038462	

4.3 Multiple Predictors F Test

```
## Analysis of Variance Table
##
## Model 1: log_nnal ~ age_buckets + log_cotinine + I(log_cotinine^2)
## Model 2: log_nnal ~ age_buckets + gender + cpd * carbon_monoxide + log_cotinine +
##     I(log_cotinine^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     74 43.437
## 2     70 42.900  4   0.53739 0.2192 0.9269
```

```
## Analysis of Variance Table
##
## Model 1: log_nnal ~ age_buckets + log_cotinine
## Model 2: log_nnal ~ age_buckets + log_cotinine + I(log_cotinine^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     75 50.590
## 2     74 43.437  1    7.1525 12.185 0.0008158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log_nnal ~ log_cotinine + I(log_cotinine^2)
## Model 2: log_nnal ~ age_buckets + log_cotinine + I(log_cotinine^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

```
## 1     83 48.660
## 2     74 43.437  9    5.2233 0.9887 0.4567
```

4.4 Extra Sum of Squares

4.5 Type I Sum of Squares

Sequential Sum of Squares

```
## Analysis of Variance Table
##
## Response: log_nnal
##                   Df Sum Sq Mean Sq F value        Pr(>F)
## age_buckets        9 15.235  1.6928  2.7830      0.007500 **
## gender             1  2.108  2.1080  3.4657      0.066795 .
## cpd                1  5.430  5.4298  8.9268      0.003856 **
## carbon_monoxide    1 19.297 19.2973 31.7256 0.00000033309 ***
## log_cotinine       1 21.805 21.8055 35.8490 0.00000007928 ***
## I(log_cotinine^2)  1  5.306  5.3059  8.7231      0.004258 **
## Residuals         71 43.186  0.6083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: log_nnal
##                   Df Sum Sq Mean Sq F value              Pr(>F)
## log_cotinine       1 57.629  57.629 94.7440 0.00000000000001051 ***
## I(log_cotinine^2)  1  6.079   6.079  9.9933            0.002311 **
## age_buckets        9  5.223   0.580  0.9541            0.484905
## gender             1  0.249   0.249  0.4093            0.524361
## cpd                1  0.000   0.000  0.0001            0.994084
## carbon_monoxide    1  0.002   0.002  0.0028            0.958003
## Residuals         71 43.186   0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.6 Type II Sum of Squares

4.6.1 Partial R-squared

page 269

```
## Anova Table (Type II tests)
##
## Response: log_nnal
##                   Sum Sq Df F value   Pr(>F)
## log_cotinine       2.594  1  4.2642 0.042576 *
## I(log_cotinine^2)  5.306  1  8.7231 0.004258 **
## age_buckets        4.591  9  0.8386 0.583243
## gender             0.249  1  0.4092 0.524450
## cpd                0.000  1  0.0003 0.986682
## carbon_monoxide    0.002  1  0.0028 0.958003
## Residuals         43.186 71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.7 Type III Sum of Squares

```
## Anova Table (Type III tests)
##
## Response: log_nnal
##                      Sum Sq Df F value   Pr(>F)
## (Intercept)           0.386  1  0.6297 0.430154
## log_cotinine          2.851  1  4.6523 0.034448 *
## I(log_cotinine^2)     5.584  1  9.1118 0.003541 **
## age_buckets           4.791  9  0.8686 0.557144
## gender                0.293  1  0.4788 0.491269
## cpd                   0.108  1  0.1770 0.675219
## carbon_monoxide       0.121  1  0.1982 0.657553
## cpd:carbon_monoxide   0.287  1  0.4678 0.496279
## Residuals            42.900 70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 Diagnostics

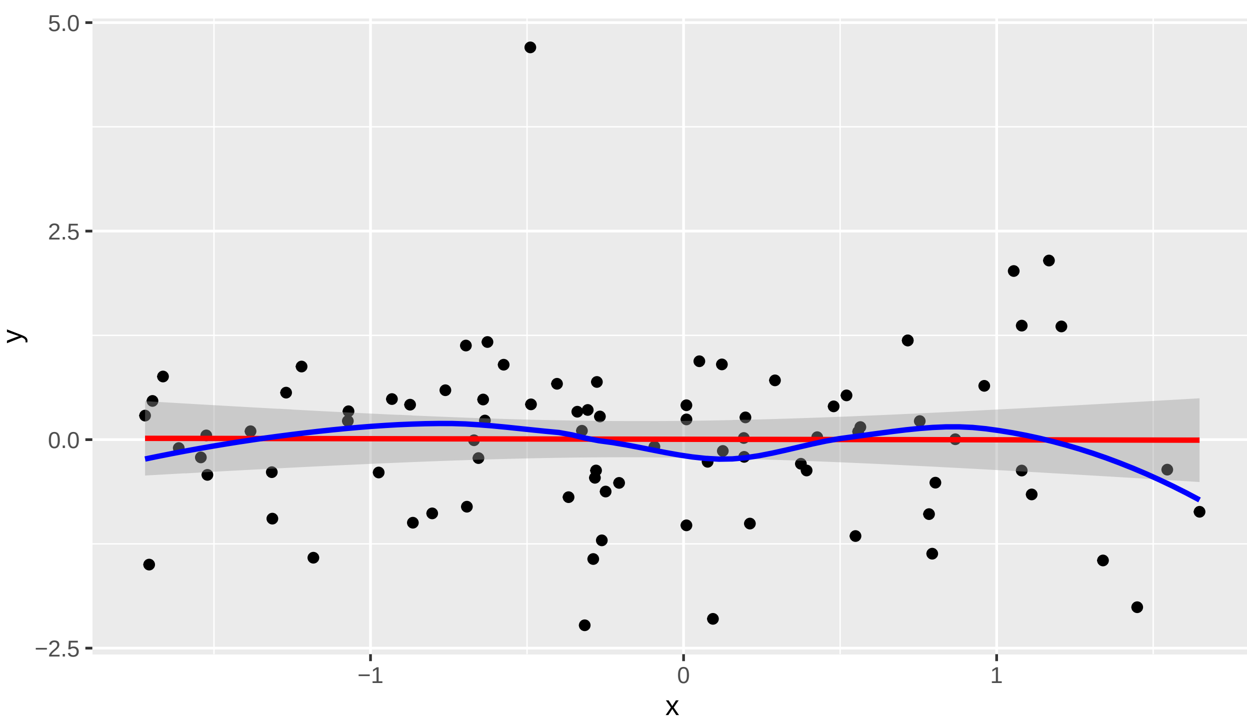
5.1 Variable Related

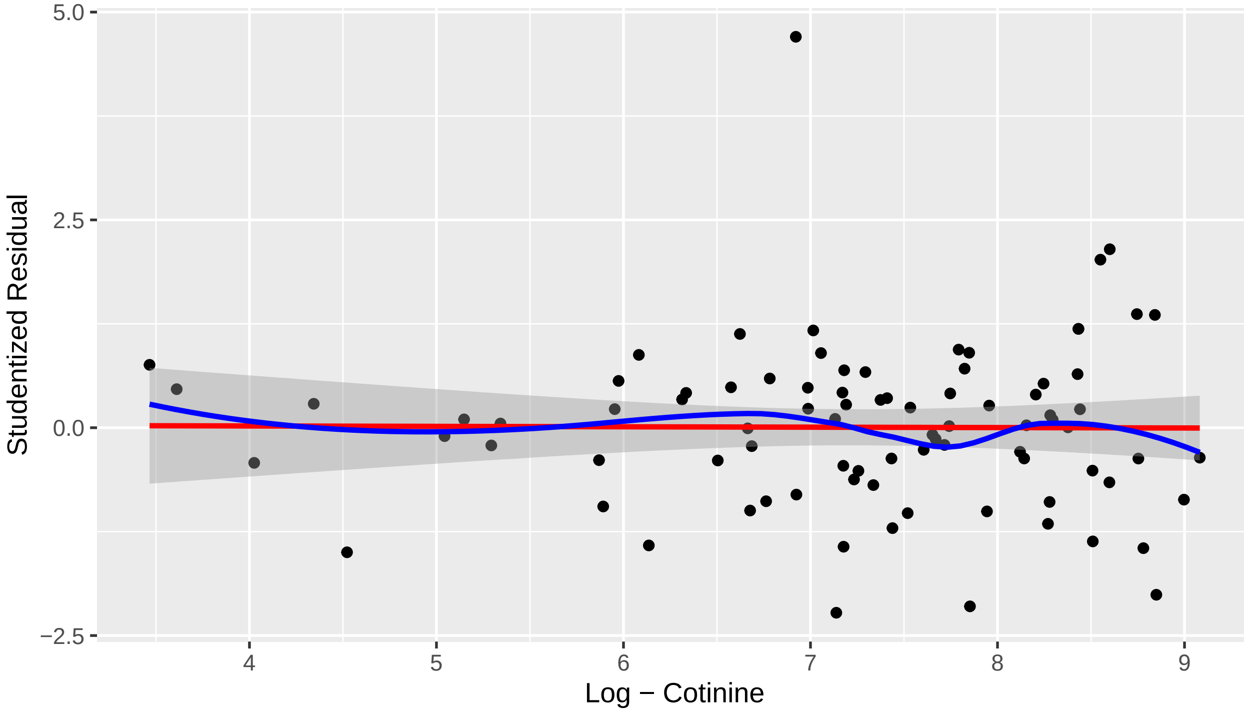
5.1.1 Assumtions to Verify

Take from HW 3

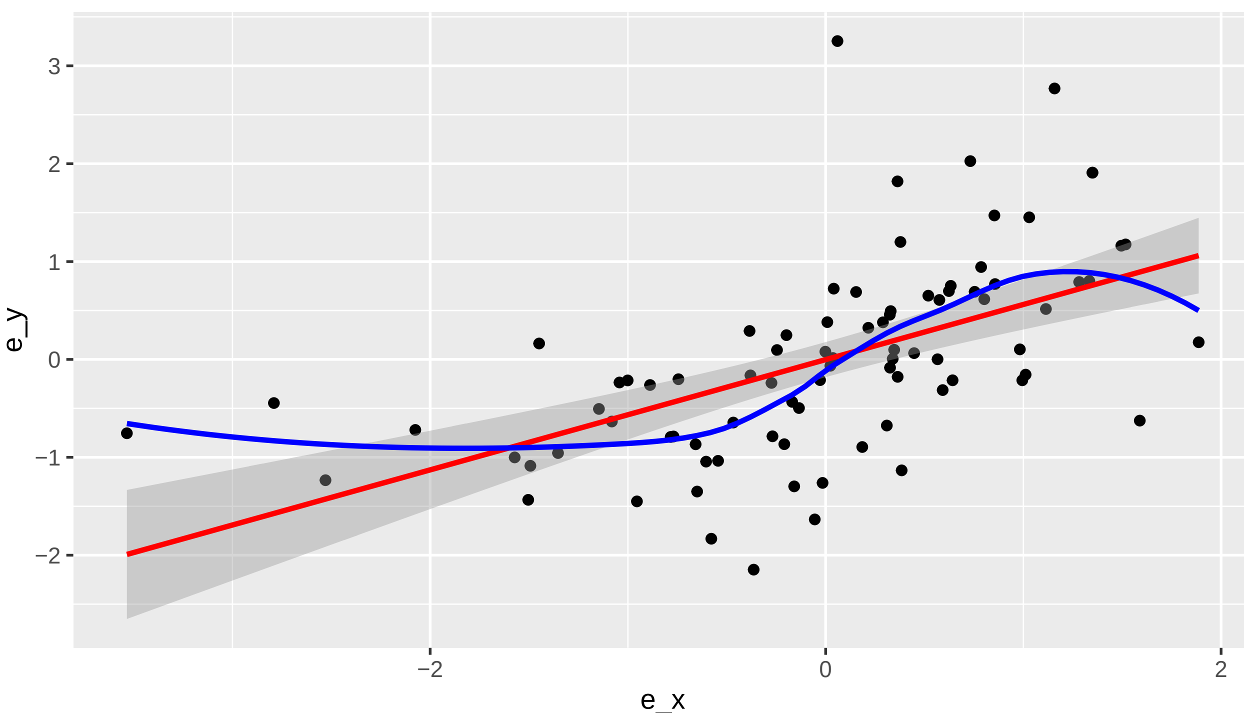
1. Constant Variance
2. Independence of Predictors and Residuals
3. Normality of Residuals

5.1.2 Residual Plots

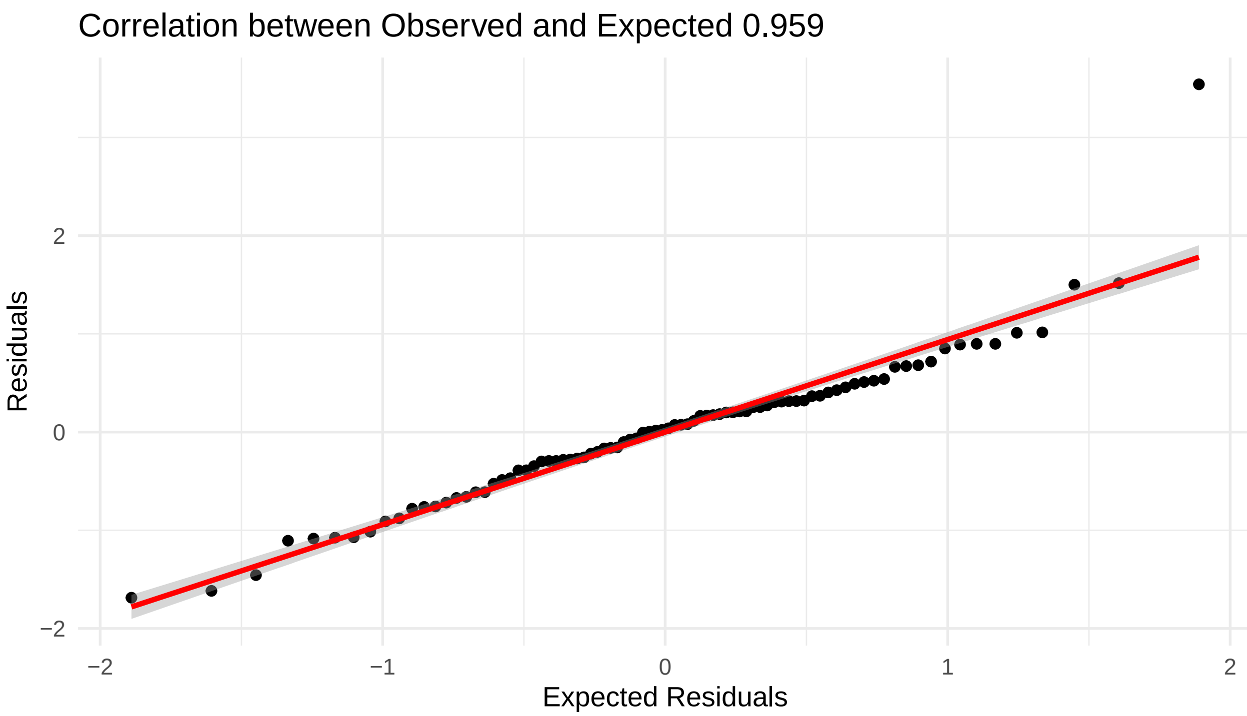




5.1.3 Added Variable Plot



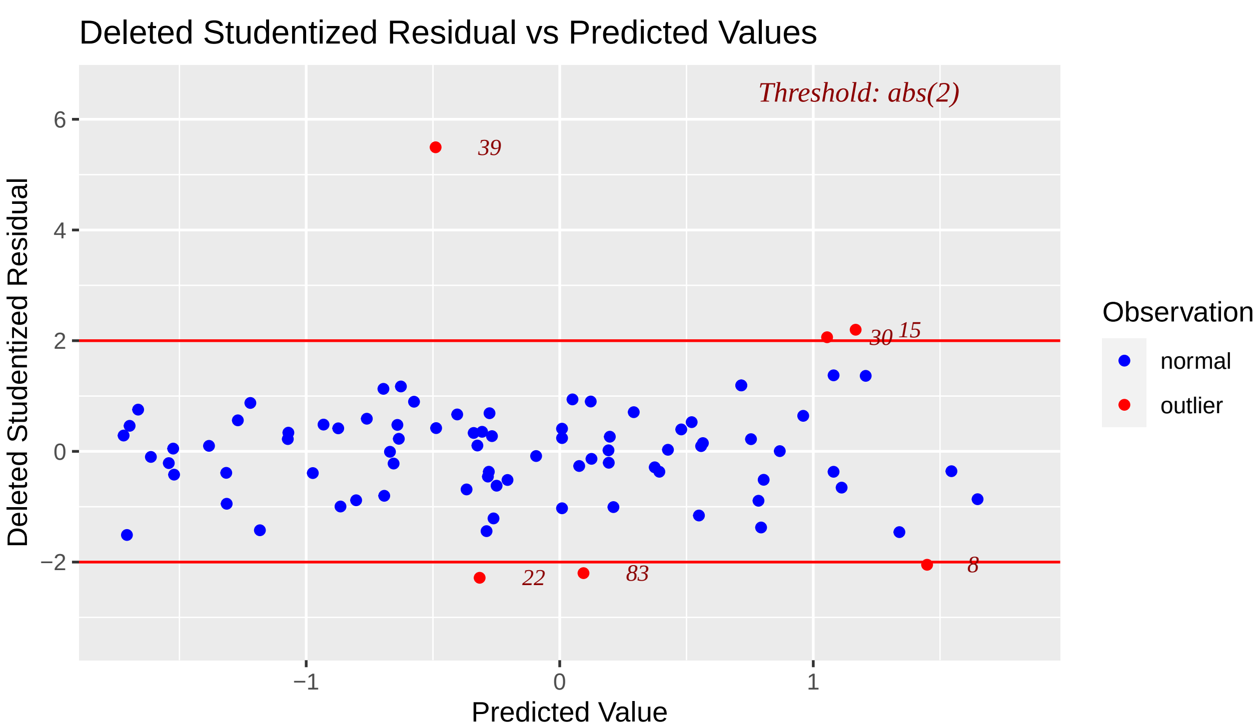
5.1.4 Residual Normality



5.2 Outliers - Observation Related

5.2.1 Deleted Studentized Residuals

Book page 395-396



5.2.2 Cook's Distance

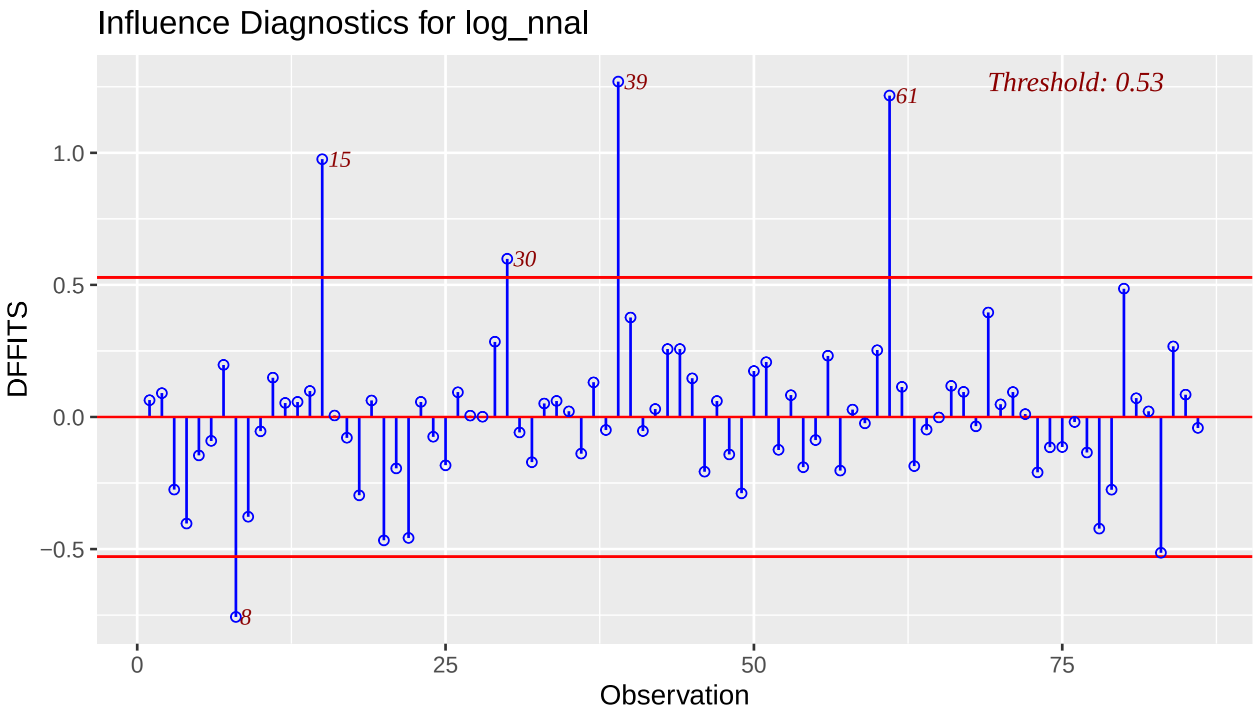
5.2.3 Leverage Values from the Hat Matrix

page 399

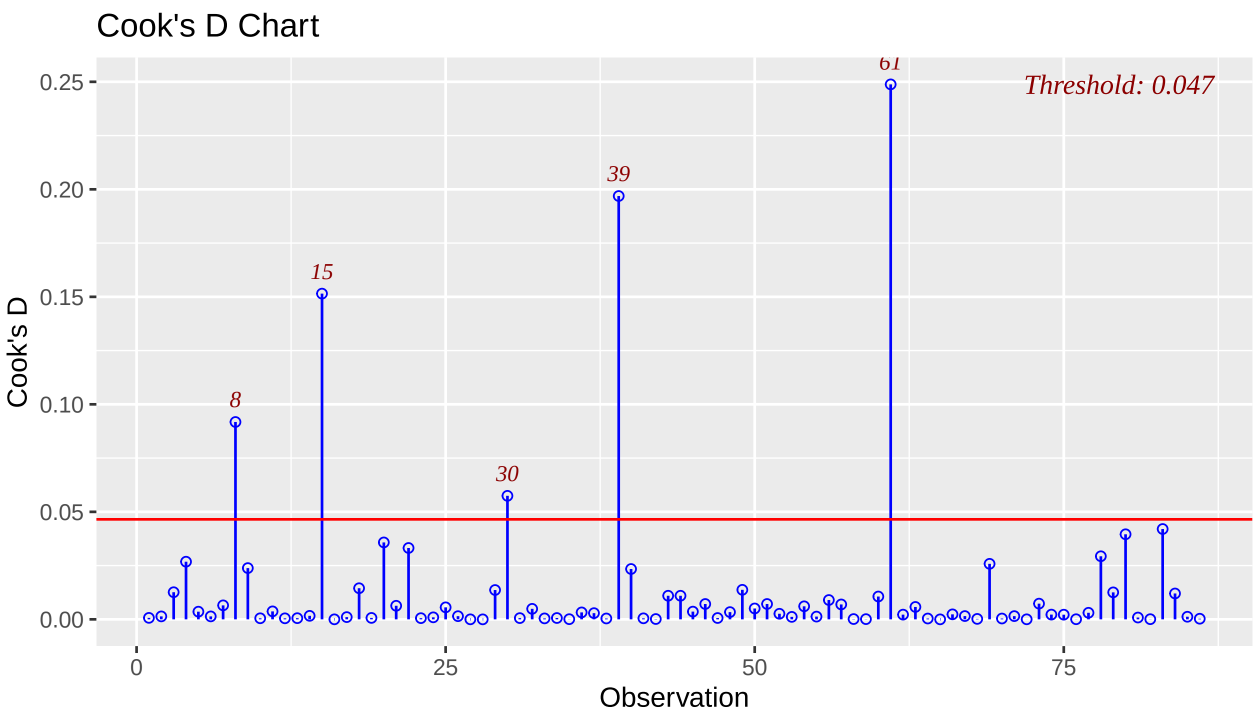
```
##   5   9  15  29  57  61  80  84
##   5   9  15  29  57  61  80  84
```

5.2.4 DFFITS

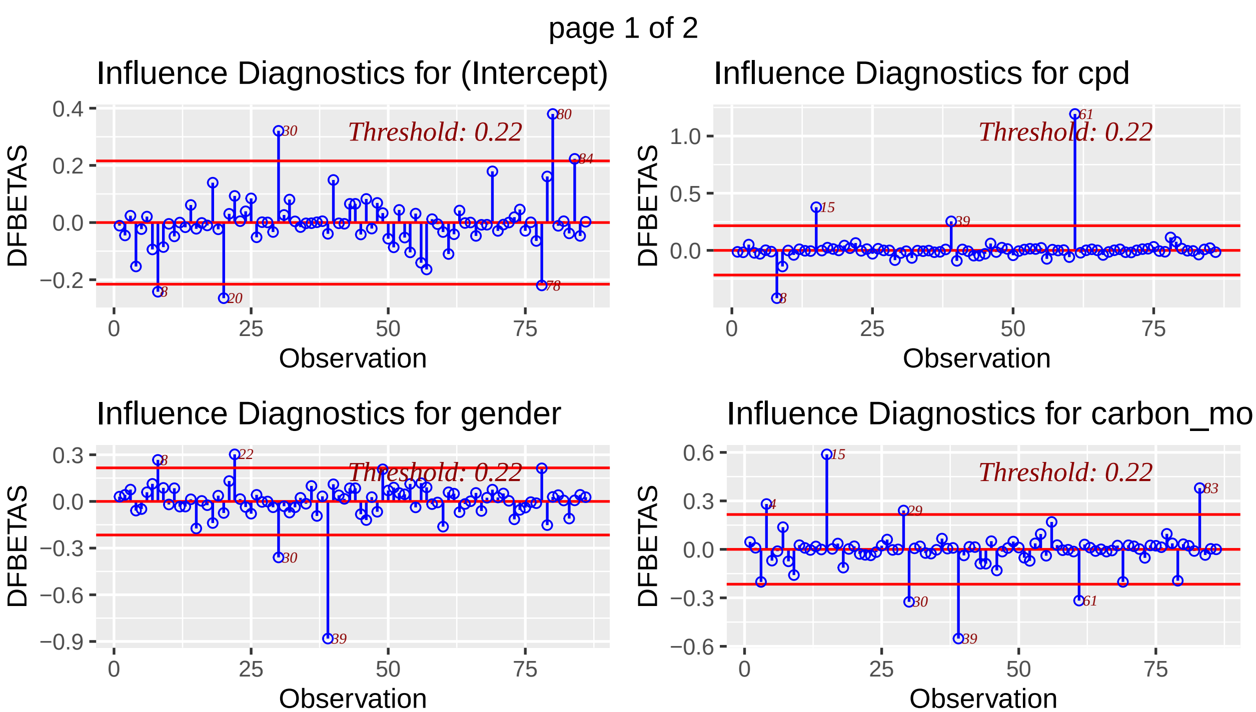
page 401

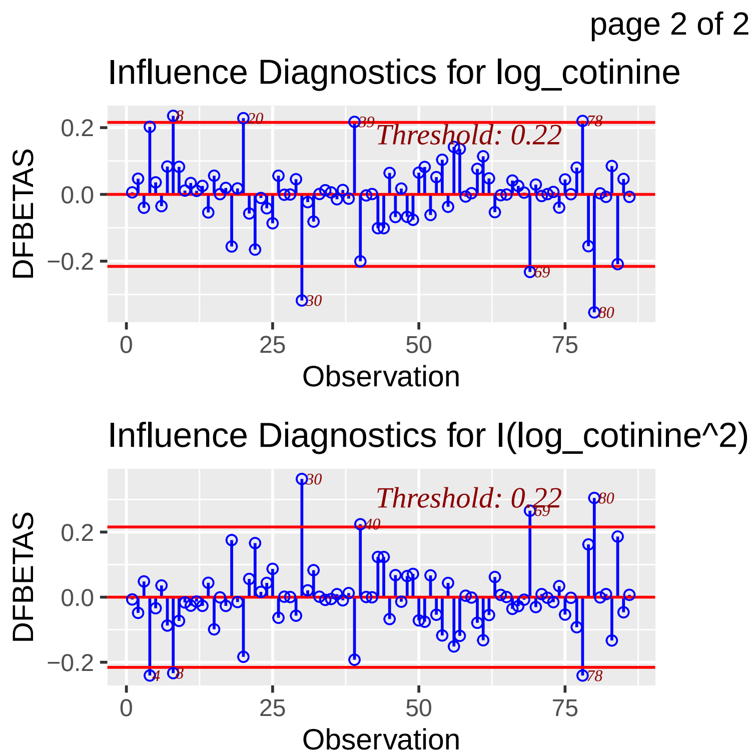


5.2.5 Cook's Distance



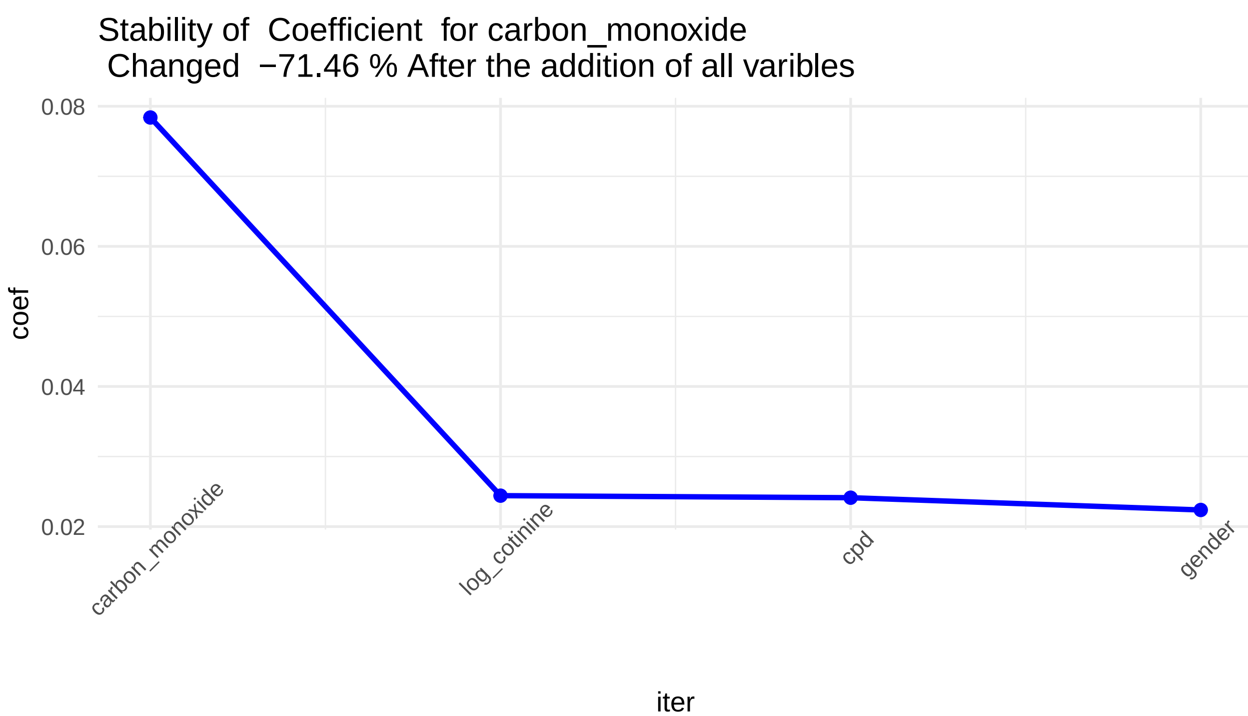
5.2.6 DFBETAS

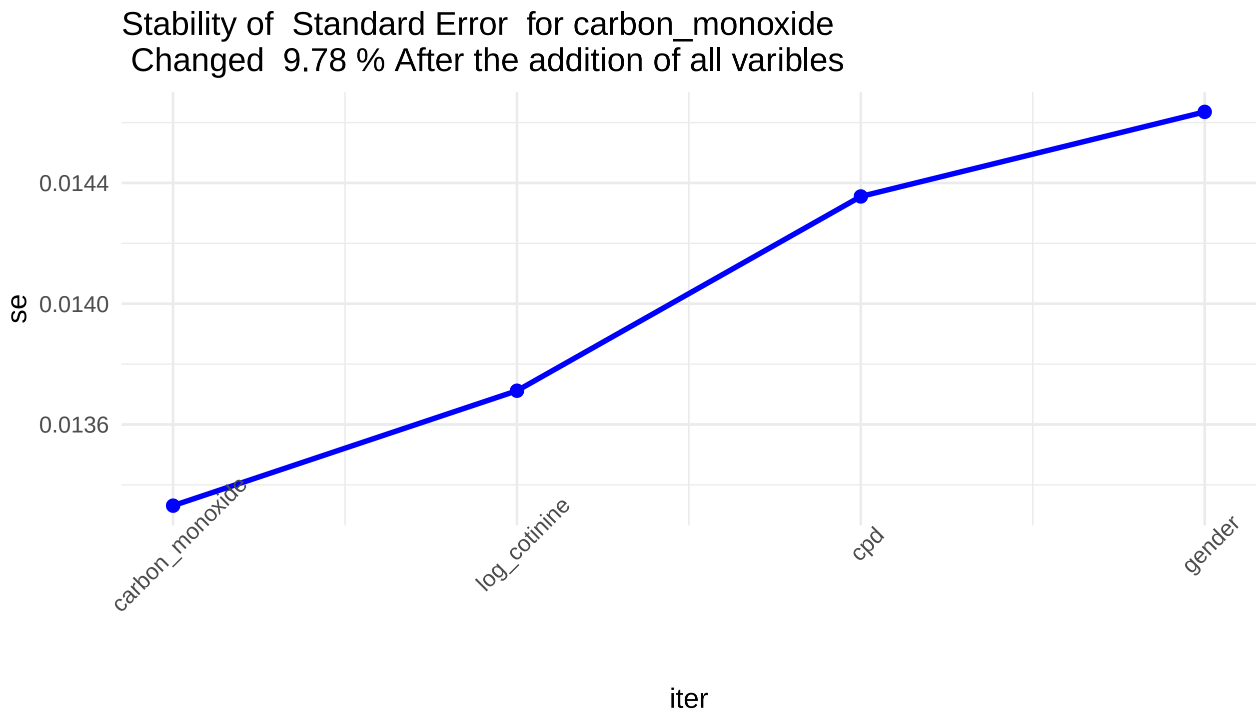




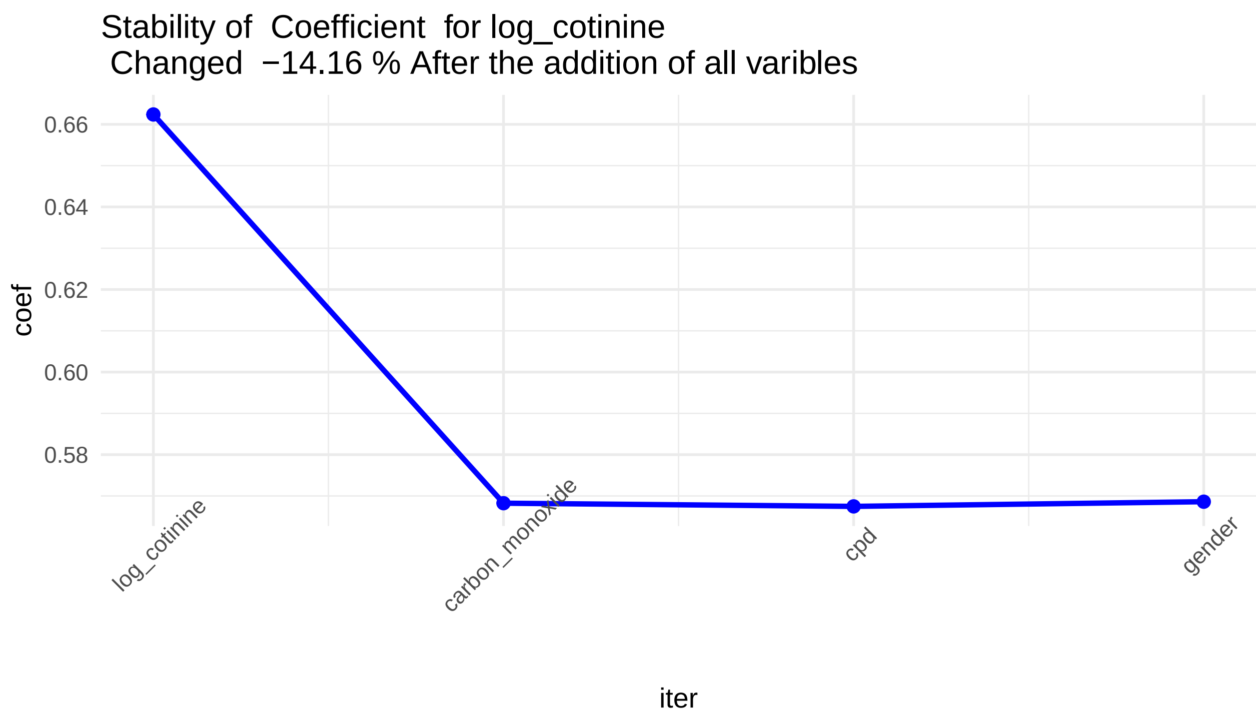
5.3 Informal Diagnostics

5.3.1 Coefficient Stability and Standard Error Inflation

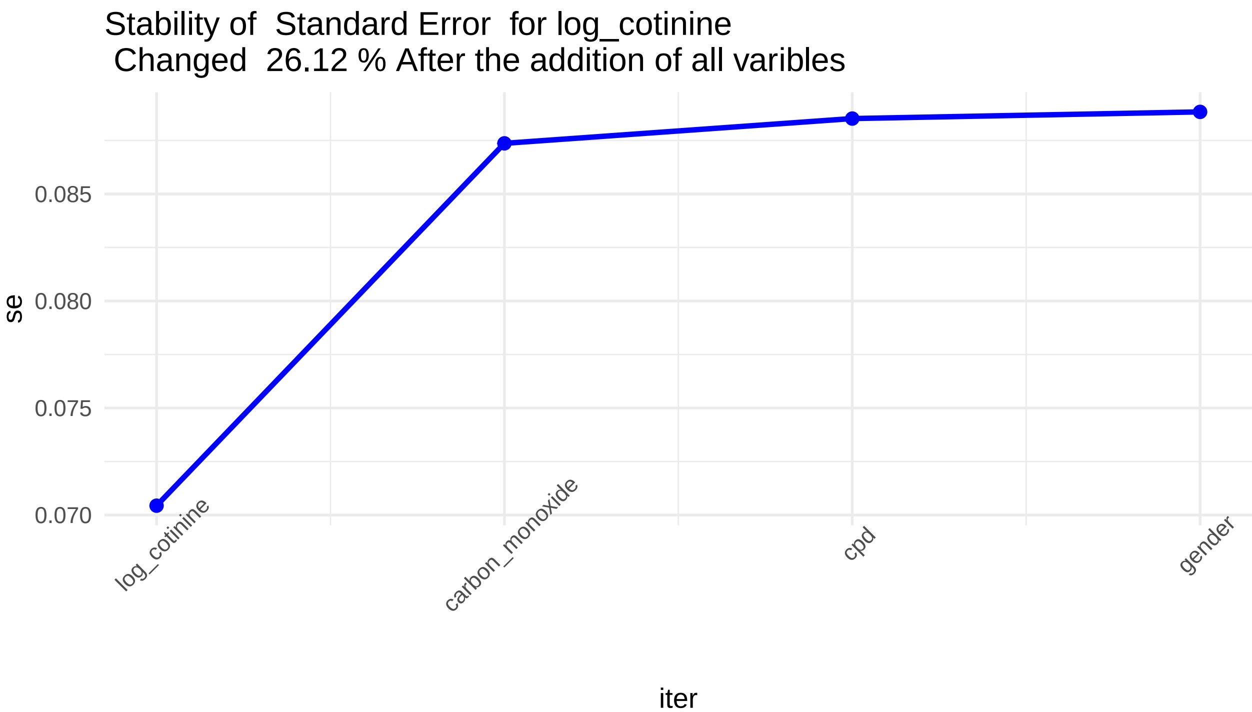




Stability of Standard Error for carbon_monoxide
Changed 9.78 % After the addition of all varibles
0.0144
0.0140
0.0136
se
carbon_monoxide
log_cotinine
cpd
gender
iter



Stability of Coefficient for log_cotinine
Changed −14.16 % After the addition of all varibles
0.66
0.64
0.62
0.60
0.58
coef
log_cotinine
carbon_monoxide
cpd
gender
iter



5.3.2 Variance Inflation Factor

```
##                           predictor         se        VIF
## log_cotinine           log_cotinine 0.58851477 -76.171702
## carbon_monoxide     carbon_monoxide 0.01585260  -2.243734
## cpd                             cpd 0.00776754  -1.222221
## gender                       gender 0.17844727  -1.071357
## I(log_cotinine^2) I(log_cotinine^2) 0.04619842 -83.118731
```

```
##                       predictor          se       VIF
## log_cotinine       log_cotinine 0.088834419 -1.601387
## carbon_monoxide carbon_monoxide 0.014635391 -1.764551
## cpd                         cpd 0.008084738 -1.221714
## gender                   gender 0.183309892 -1.043136
```

6 Summary of Diagnostrics and Final Model For Inference

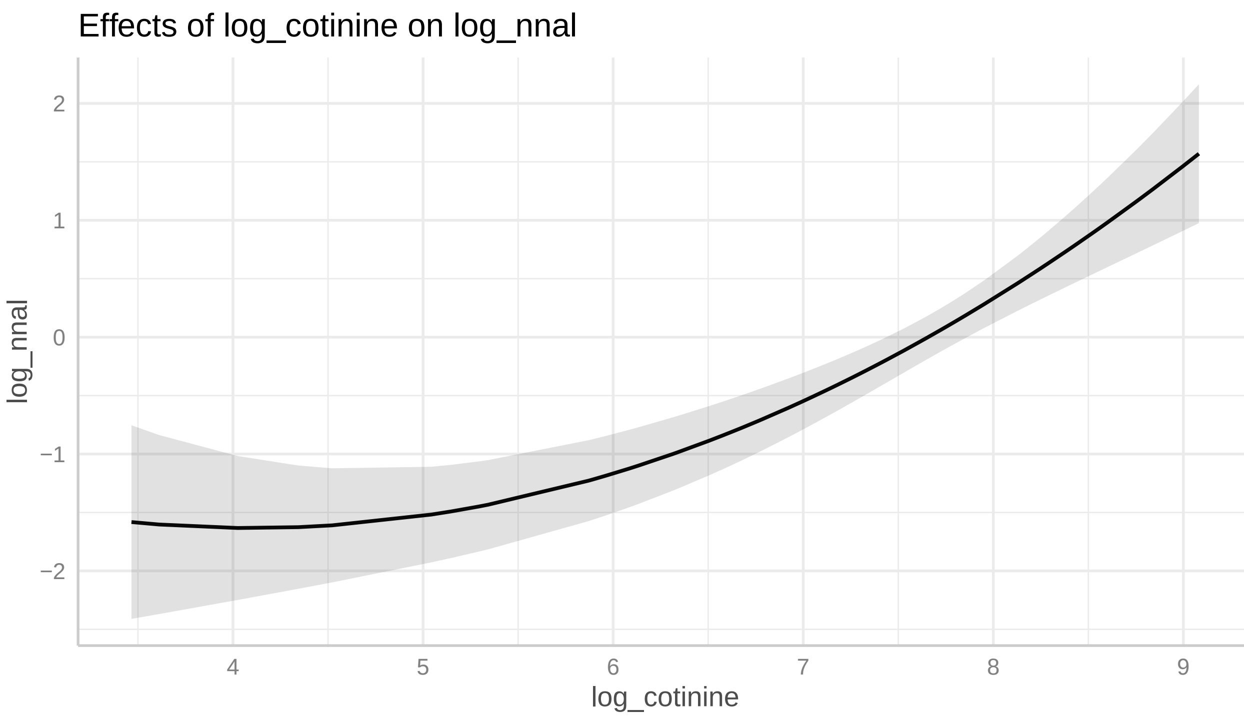
filter out outliers, other shit and fit the final model

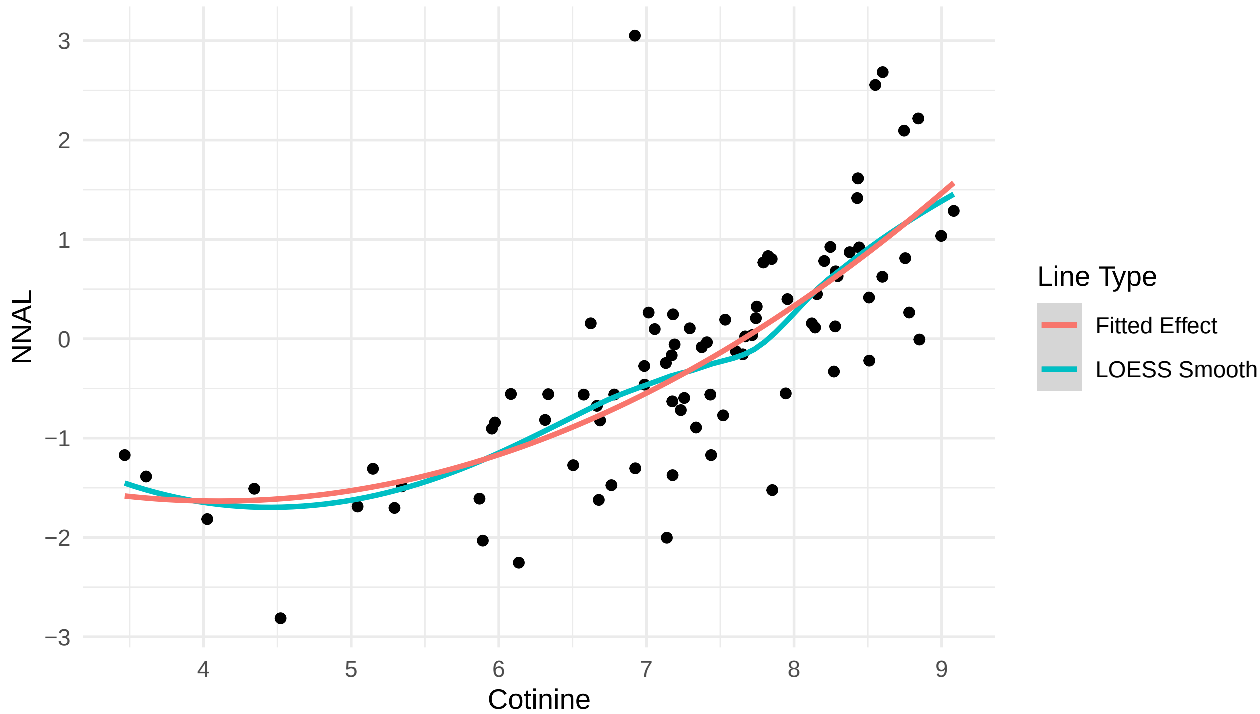
7 Inference

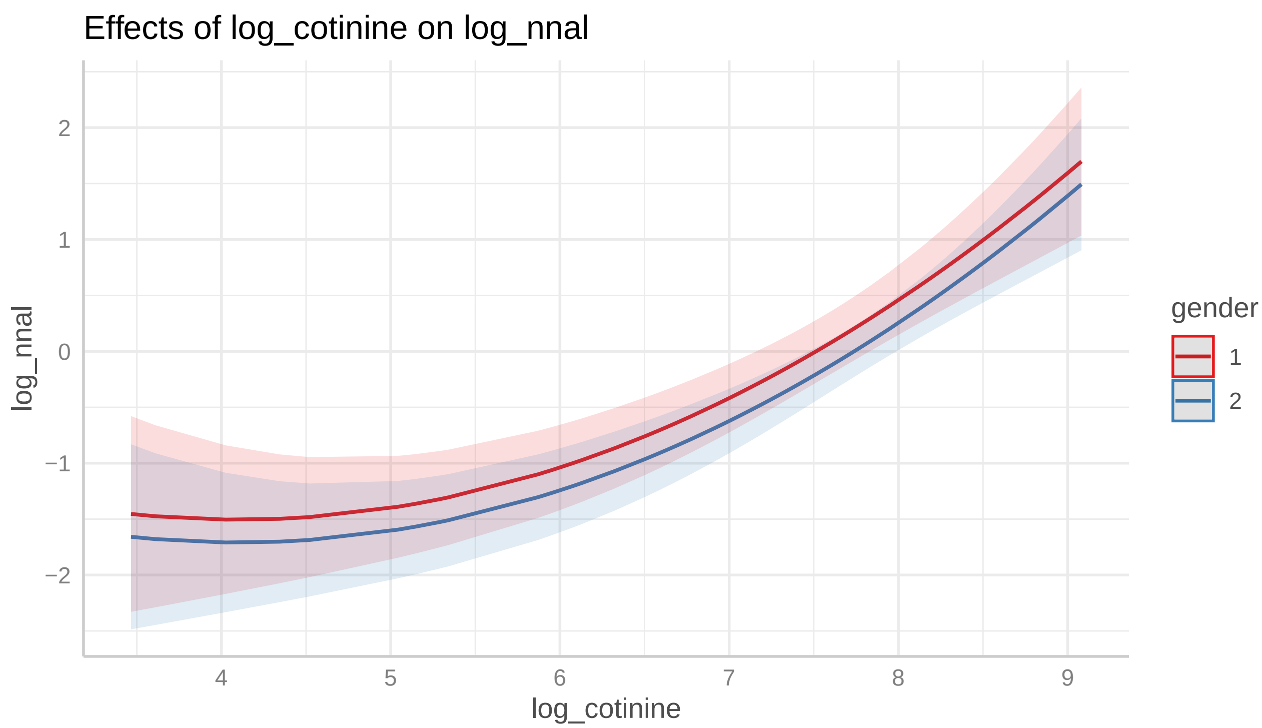
7.1 Coefficient Inference

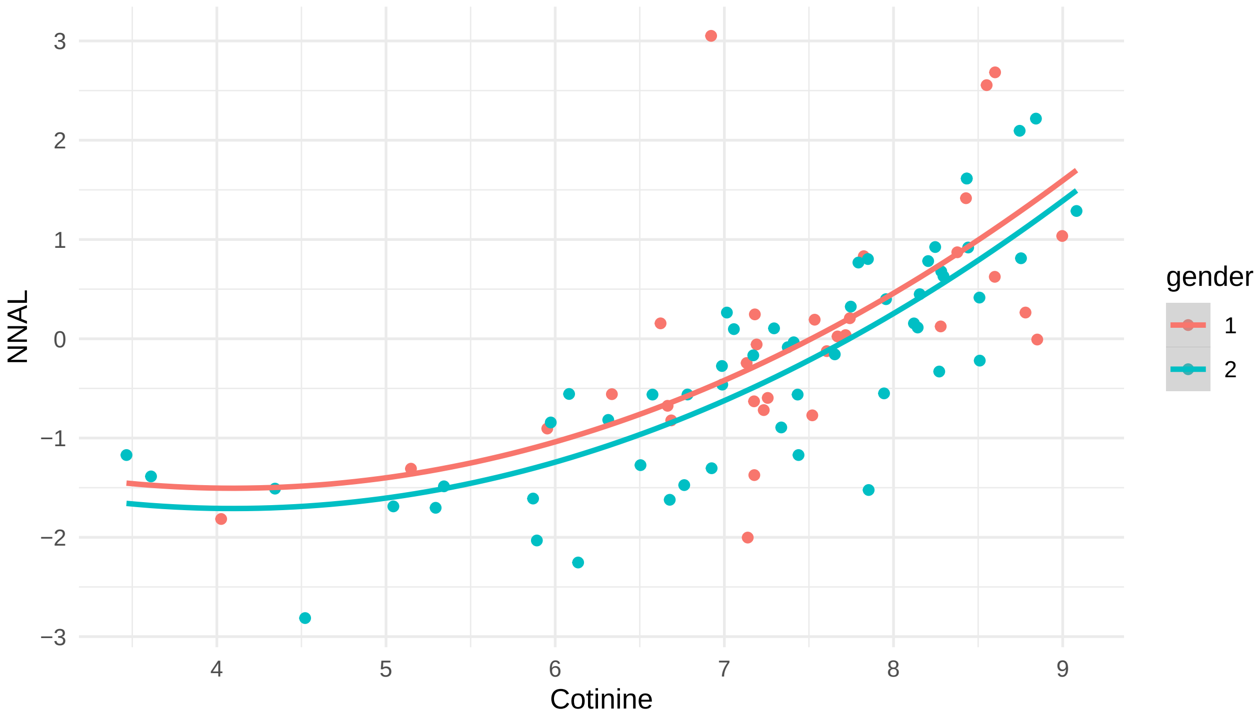
Predictor	Estiamte	Standard Error	Z Value	P value	Significant
(Intercept)	0.839865	1.946893	0.431387	0.667347	
gender	-0.204383	0.178447	-1.145341	0.255483	
cpd	0.000382	0.007768	0.049191	0.960889	
carbon_monoxide	0.001927	0.015853	0.121576	0.903540	
log_cotinine	-1.056302	0.588515	-1.794861	0.076453	
I(log_cotinine^2)	0.128918	0.046198	2.790520	0.006577	*

7.2 Effect Plots




NNAL
Cotinine
Line Type
Fitted Effect
LOESS Smooth


Effects of log_cotinine on log_nnal
log_nnal
log_cotinine
gender
1
2



7.3 Estimating Effects and Predictions

Now we can link visual effects and with the fitted effects

```
##          x predicted std.error  conf.low  conf.high
## 1 3.465736 -1.582351   0.42253 -2.410495 -0.7542075
```

8 More on Predictions: Deeper dive into the estimates

8.1 Average Response Level C.I.

Page 58

8.2 Single Observation C.I.

8.3 N Observations C.I.

9 Summary

10 Appendix - Code