

Analysis of USDA feed grains database to predict the price of corn

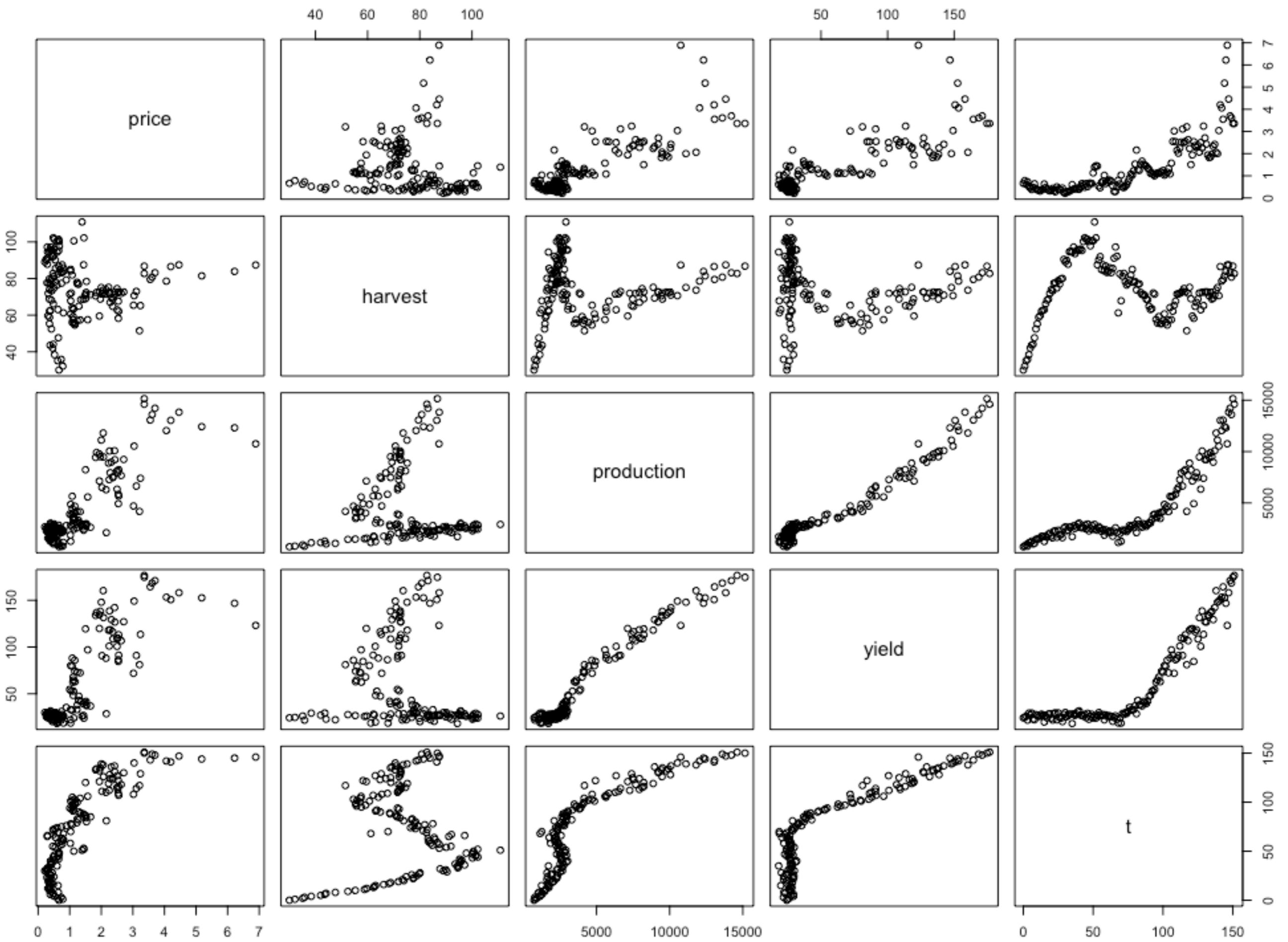


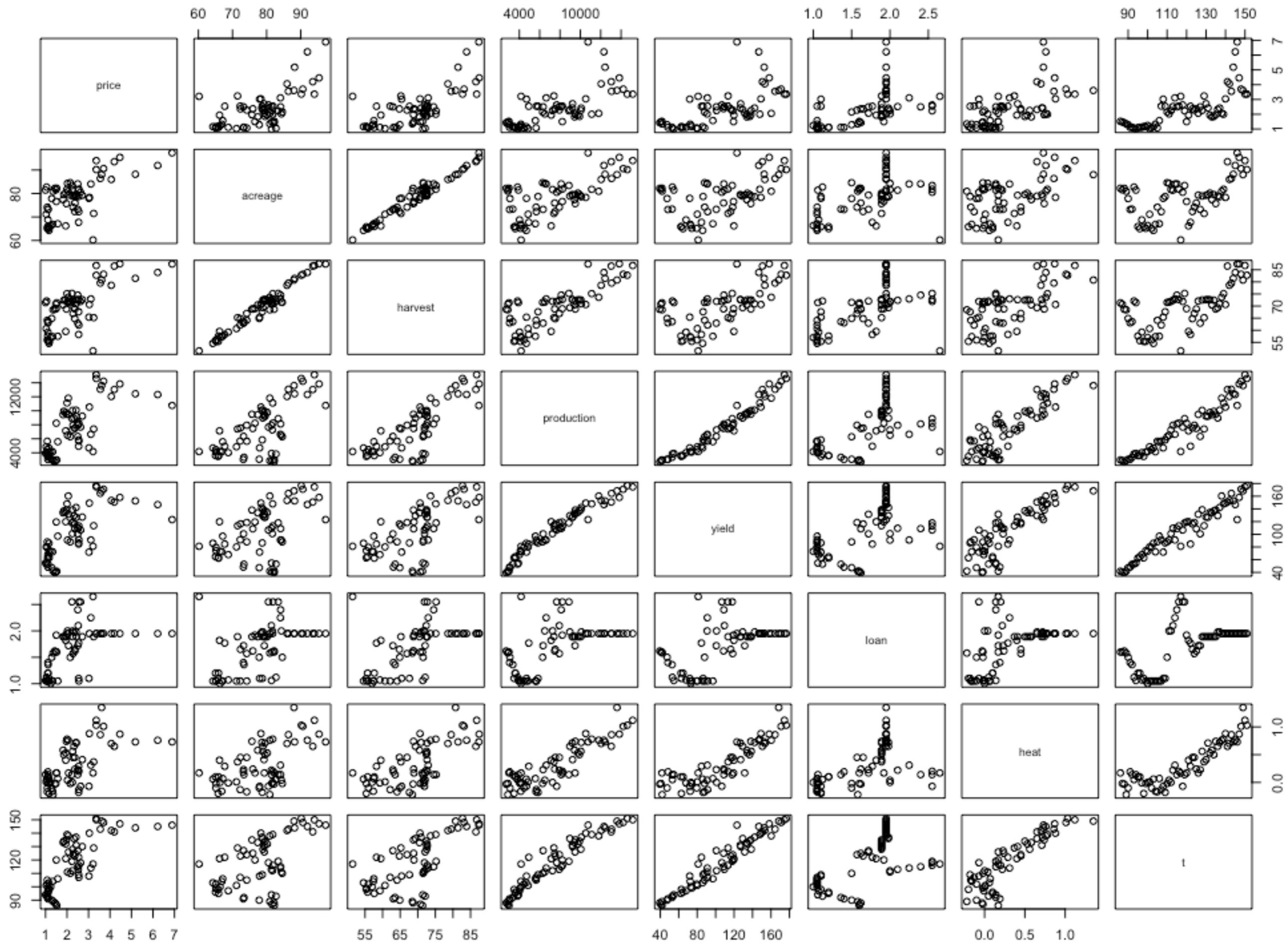
Zane Billings
MATH 375
8 May, 2019

Introduction

- USDA Feed Grains database records a lot of data on corn, wheat, sorghum, and oats.
- All of the data is time series data, but not all of the different variables recorded start at the same year.
- Used price received by farmers as the response.







Stepwise Regression

$$\text{price} = \beta_0 + \beta_1(\text{time}) + \beta_2(\text{acreage}) + \beta_3(\text{harvest}) + \beta_4(\text{production}) + \beta_5(\text{yield}) + \beta_6(\text{loan}) + \beta_7(\text{heat})$$

VS.

$$\text{price} = \beta_0$$

Forward:

Residuals:

	Min	1Q	Median	3Q	Max
	-0.65876	-0.24846	-0.08864	0.19155	1.05809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.547122	0.954369	-4.765	1.86e-05 ***
t	0.065546	0.013239	4.951	9.93e-06 ***
loan	0.609355	0.142711	4.270	9.43e-05 ***
yield	-0.018167	0.005991	-3.032	0.00394 **
heat	-0.943429	0.337792	-2.793	0.00753 **

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.3948 on 47 degrees of freedom

Multiple R-squared: 0.6757, Adjusted R-squared: 0.6481

F-statistic: 24.48 on 4 and 47 DF, p-value: 5.44e-11

Reverse:

Residuals:

	Min	1Q	Median	3Q	Max
	-0.64112	-0.27448	-0.03946	0.22948	0.95797

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.8769347	1.5093377	-4.556	3.83e-05 ***
t	0.0674920	0.0122647	5.503	1.60e-06 ***
acreage	0.0285022	0.0114826	2.482	0.016769 *
production	-0.0002656	0.0000763	-3.481	0.001107 **
loan	0.5585764	0.1568908	3.560	0.000874 ***
heat	-0.8922684	0.3314695	-2.692	0.009876 **

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.3873 on 46 degrees of freedom

Multiple R-squared: 0.6945, Adjusted R-squared: 0.6613

F-statistic: 20.91 on 5 and 46 DF, p-value: 7.647e-11

Stepwise Regression

price = $\beta_0 + \beta_1(\text{time}) + \beta_2(\text{acreage}) + \beta_3(\text{harvest}) + \beta_4(\text{production}) + \beta_5(\text{yield}) + \beta_6(\text{loan}) + \beta_7(\text{heat})$

VS.

price = β_0

Forward:

Residuals:

Min	1Q	Median	3Q	Max
-0.65876	-0.24846	-0.08864	0.19155	1.05809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.547122	0.954369	-4.765	1.86e-05 ***
t	0.065546	0.013239	4.951	9.93e-06 ***
loan	0.609355	0.142711	4.270	9.43e-05 ***
yield	-0.018167	0.005991	-3.032	0.00394 **
heat	-0.943429	0.337792	-2.793	0.00753 **

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 0.3948 on 47 degrees of freedom

Multiple R-squared: 0.6757, Adjusted R-squared: 0.6481

F-statistic: 24.48 on 4 and 47 DF, p-value: 5.44e-11

Reverse:

Residuals:

Min	1Q	Median	3Q	Max
-0.64112	-0.27448	-0.03946	0.22948	0.95797

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.8769347	1.5093377	-4.556	3.83e-05 ***
t	0.0674920	0.0122647	5.503	1.60e-06 ***
acreage	0.0285022	0.0114826	2.482	0.016769 *
production	-0.0002656	0.0000763	-3.481	0.001107 **
loan	0.5585764	0.1568908	3.560	0.000874 ***
heat	-0.8922684	0.3314695	-2.692	0.009876 **

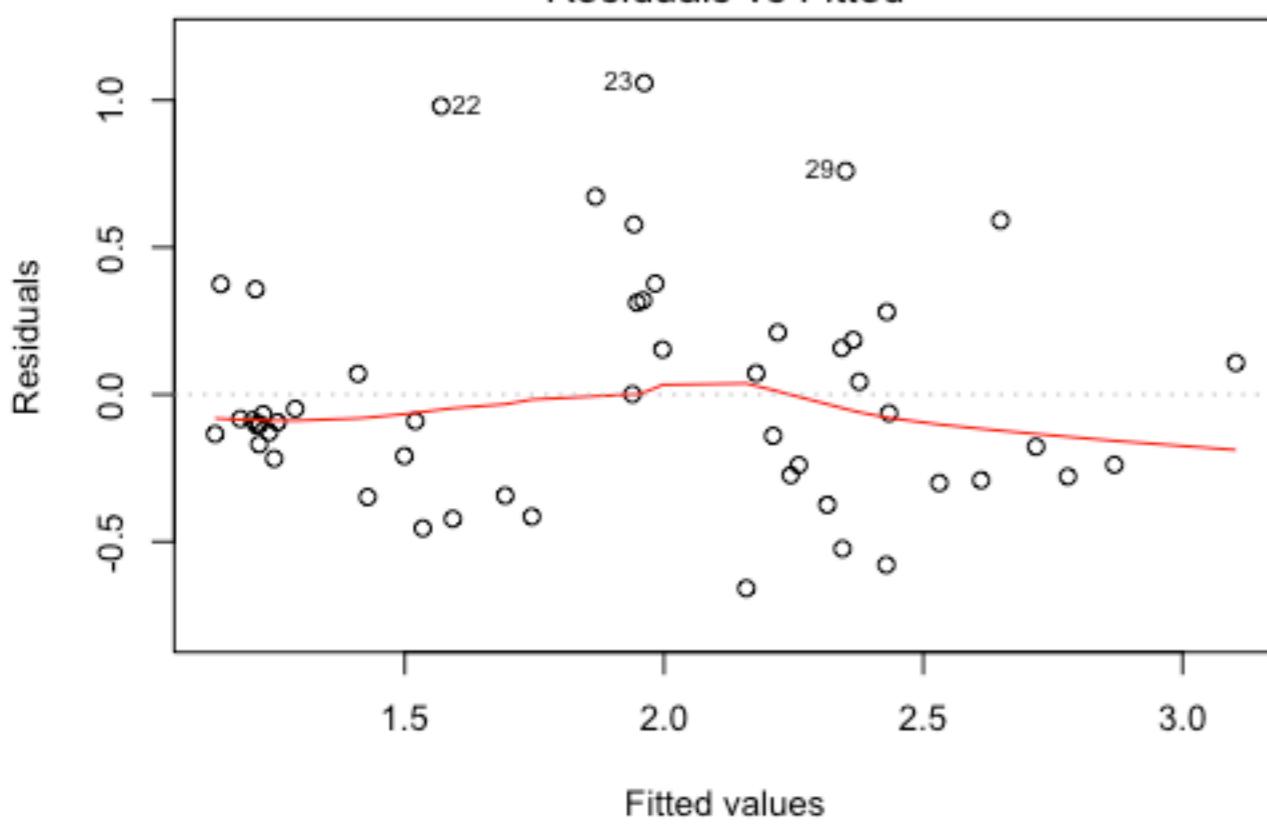
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 0.3873 on 46 degrees of freedom

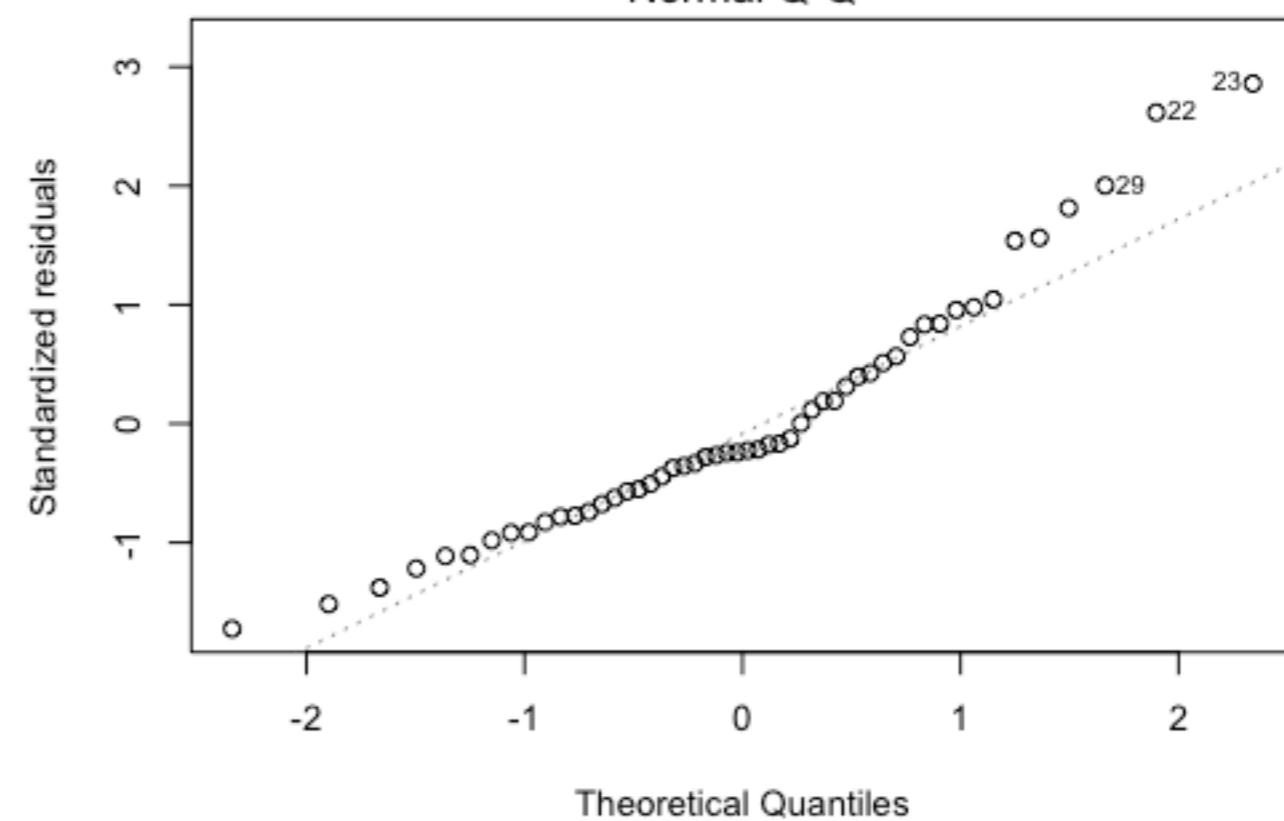
Multiple R-squared: 0.6945, Adjusted R-squared: 0.6613

F-statistic: 20.91 on 5 and 46 DF, p-value: 7.647e-11

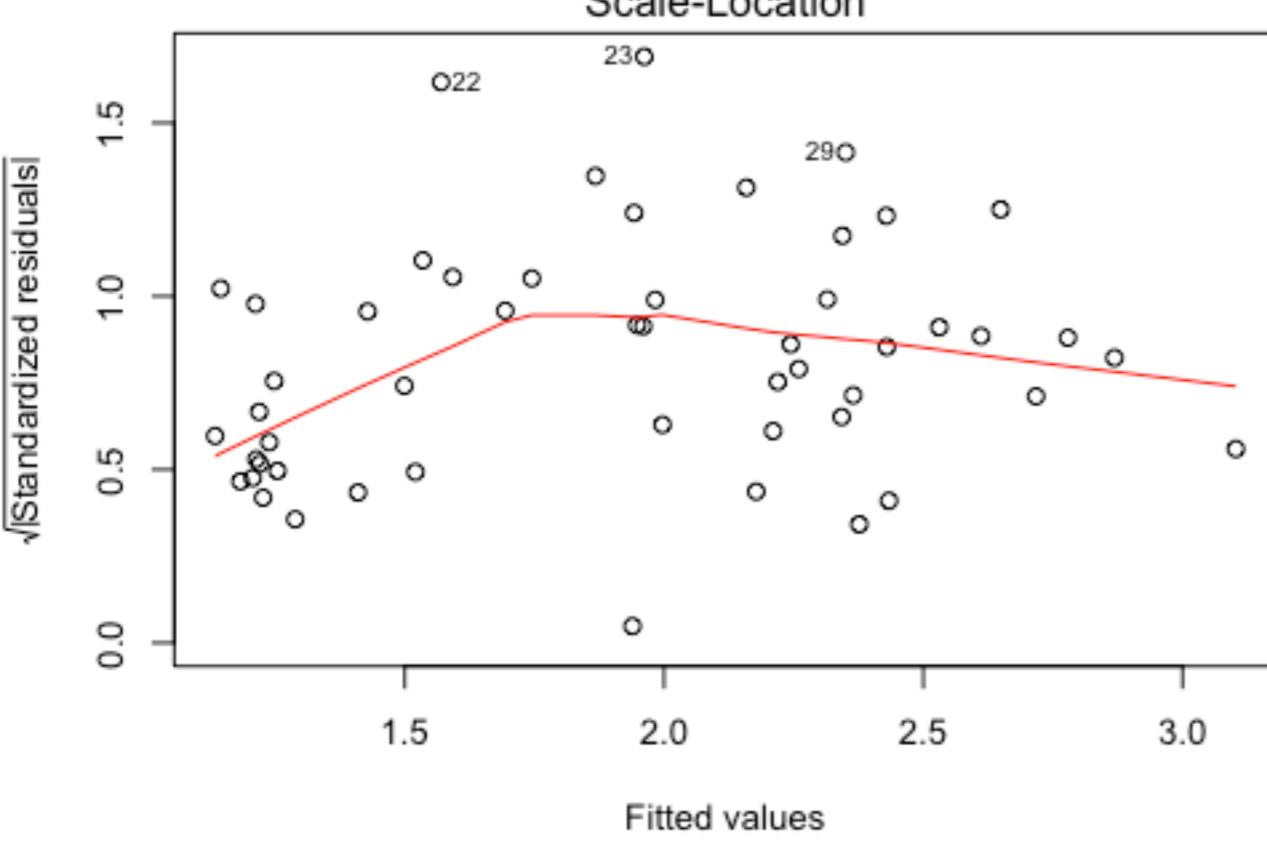
Residuals vs Fitted



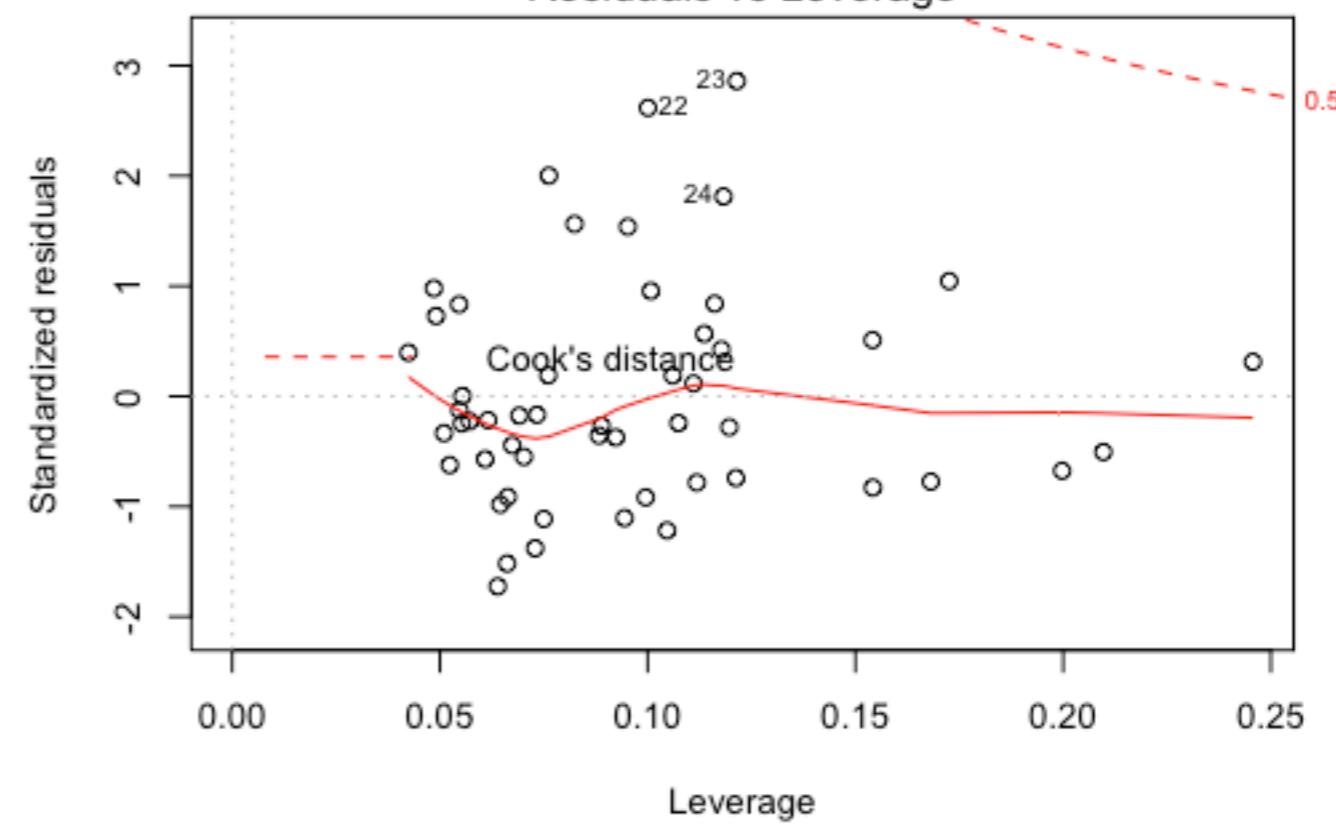
Normal Q-Q



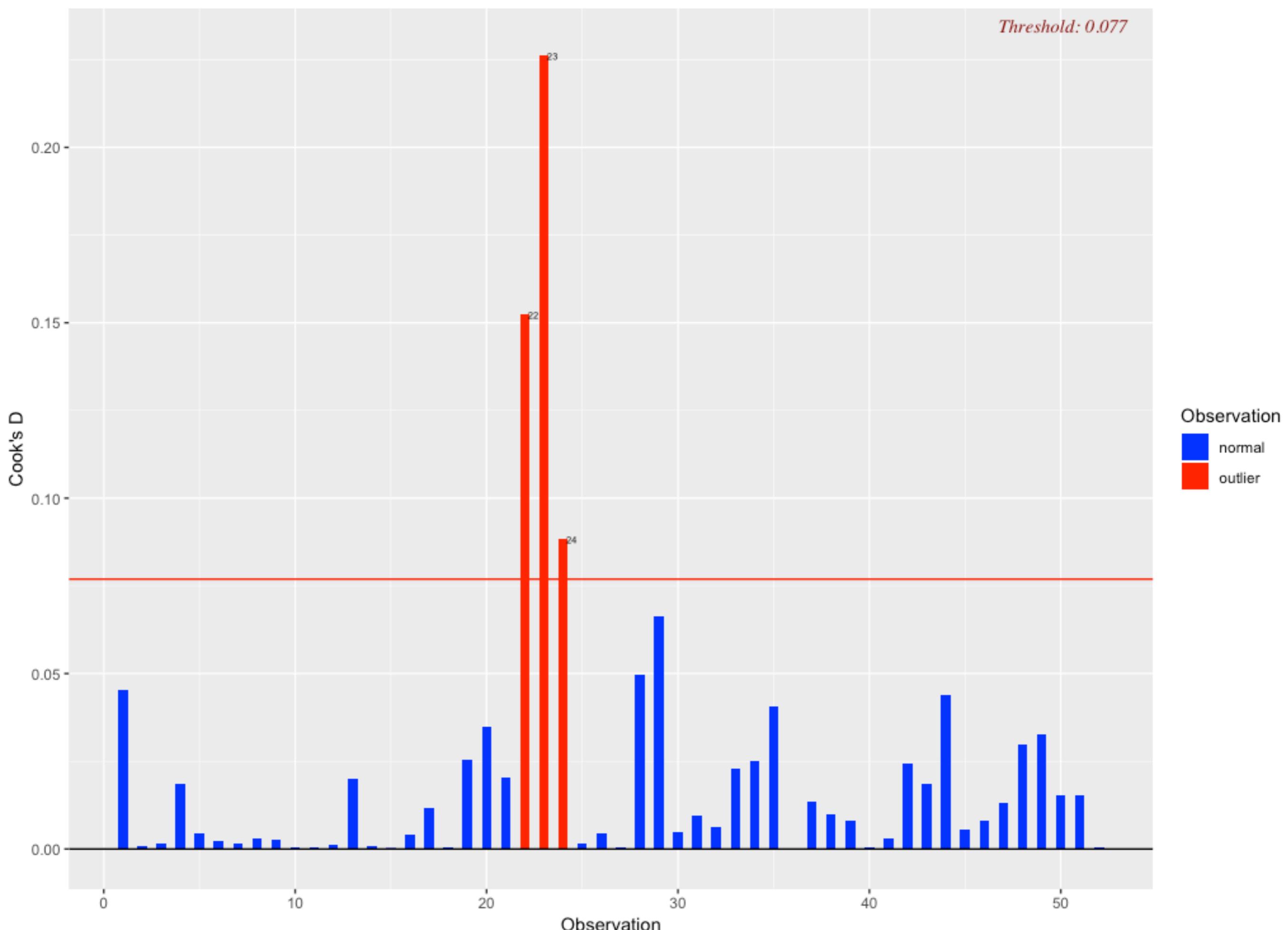
Scale-Location



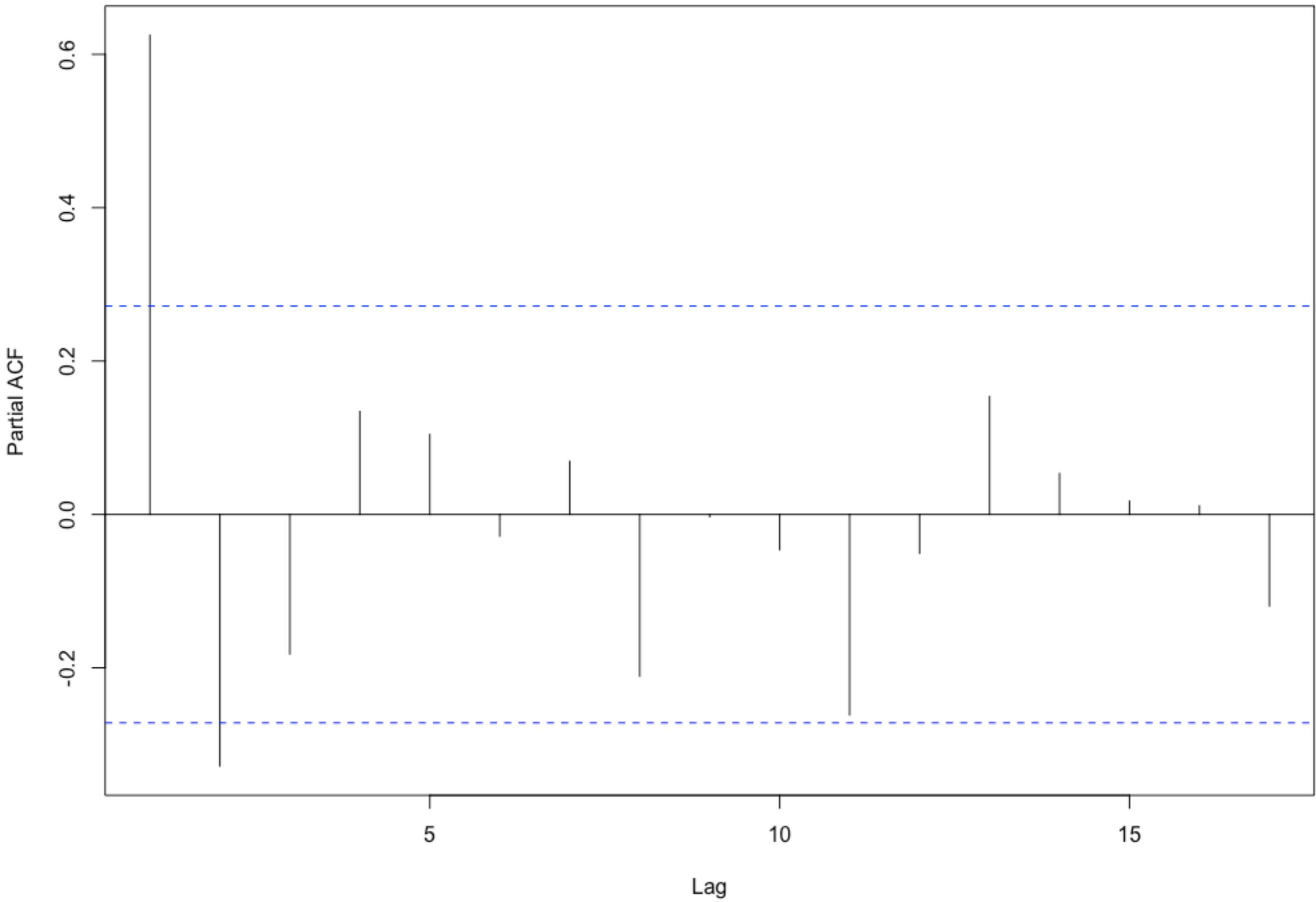
Residuals vs Leverage



Cook's D Bar Plot



Residual PACF



GLS accommodates correlated residuals

1.

Generalized least squares fit by REML
Model: price ~ t + loan + yield + heat
Data: Corn1952Training
AIC BIC logLik
40.31833 53.26936 -13.15916

Correlation Structure: AR(1)
Formula: ~t
Parameter estimate(s):
Phi
0.866462

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	-2.7952969	1.5638059	-1.787496	0.0803
t	0.0514150	0.0148595	3.460085	0.0012
loan	0.3543629	0.2544055	1.392906	0.1702
yield	-0.0178266	0.0028937	-6.160440	0.0000
heat	0.1910170	0.2683956	0.711700	0.4802

2.

Generalized least squares fit by REML
Model: price ~ t + loan + yield
Data: Corn1952Training
AIC BIC logLik
38.01165 49.23886 -13.00583

Correlation Structure: AR(1)
Formula: ~t
Parameter estimate(s):
Phi
0.8521398

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	-2.9787540	1.4387223	-2.070416	0.0438
t	0.0527077	0.0138949	3.793305	0.0004
loan	0.3773481	0.2495705	1.511990	0.1371
yield	-0.0173434	0.0028113	-6.169096	0.0000

3.

Generalized least squares fit by REML
Model: price ~ t + yield
Data: Corn1952Training
AIC BIC logLik
37.20752 46.66662 -13.60376

Correlation Structure: AR(1)
Formula: ~t
Parameter estimate(s):
Phi
0.8879908

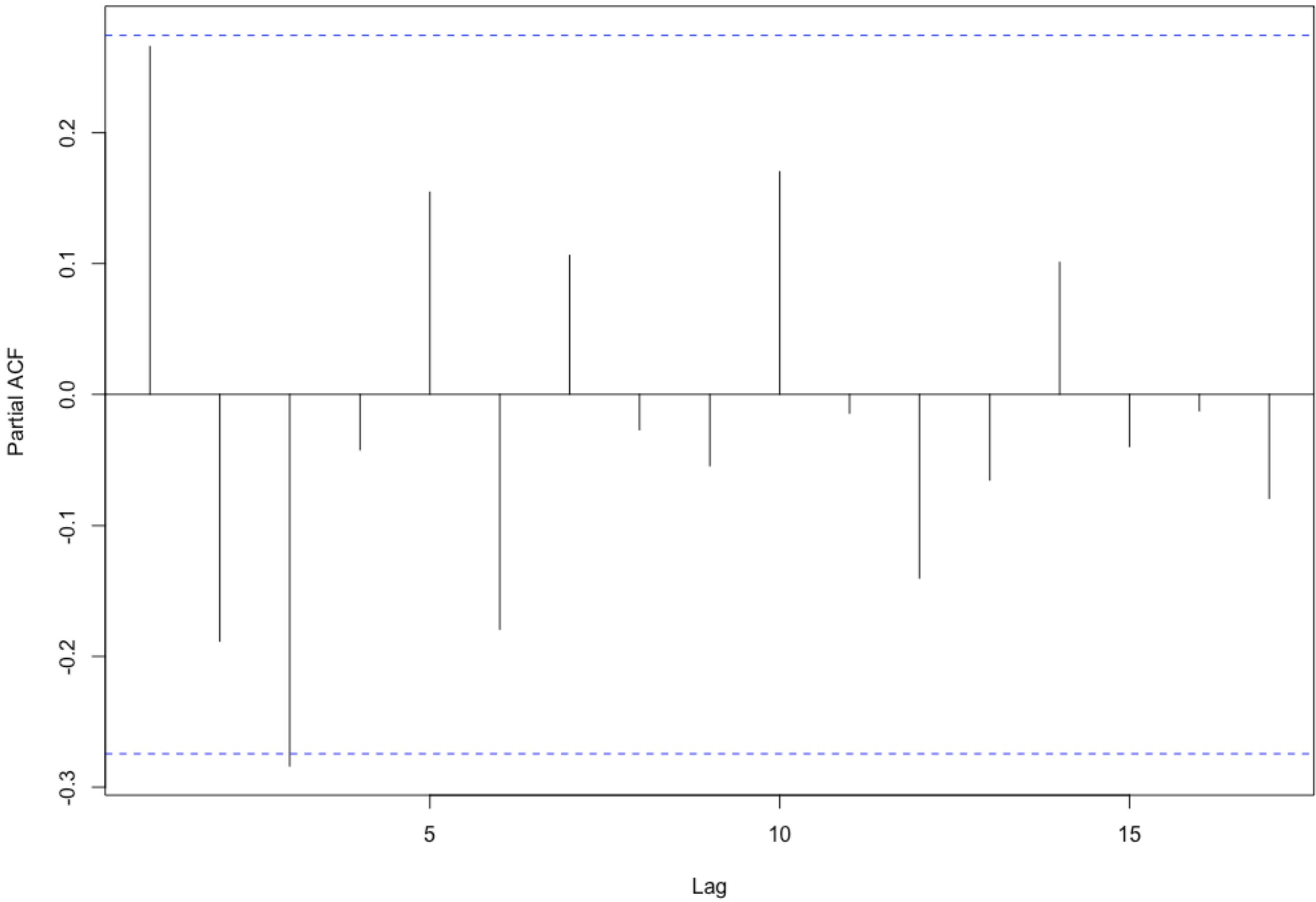
Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	-2.7917639	1.7423442	-1.602303	0.1155
t	0.0566660	0.0160393	3.532945	0.0009
yield	-0.0173597	0.0028087	-6.180676	0.0000

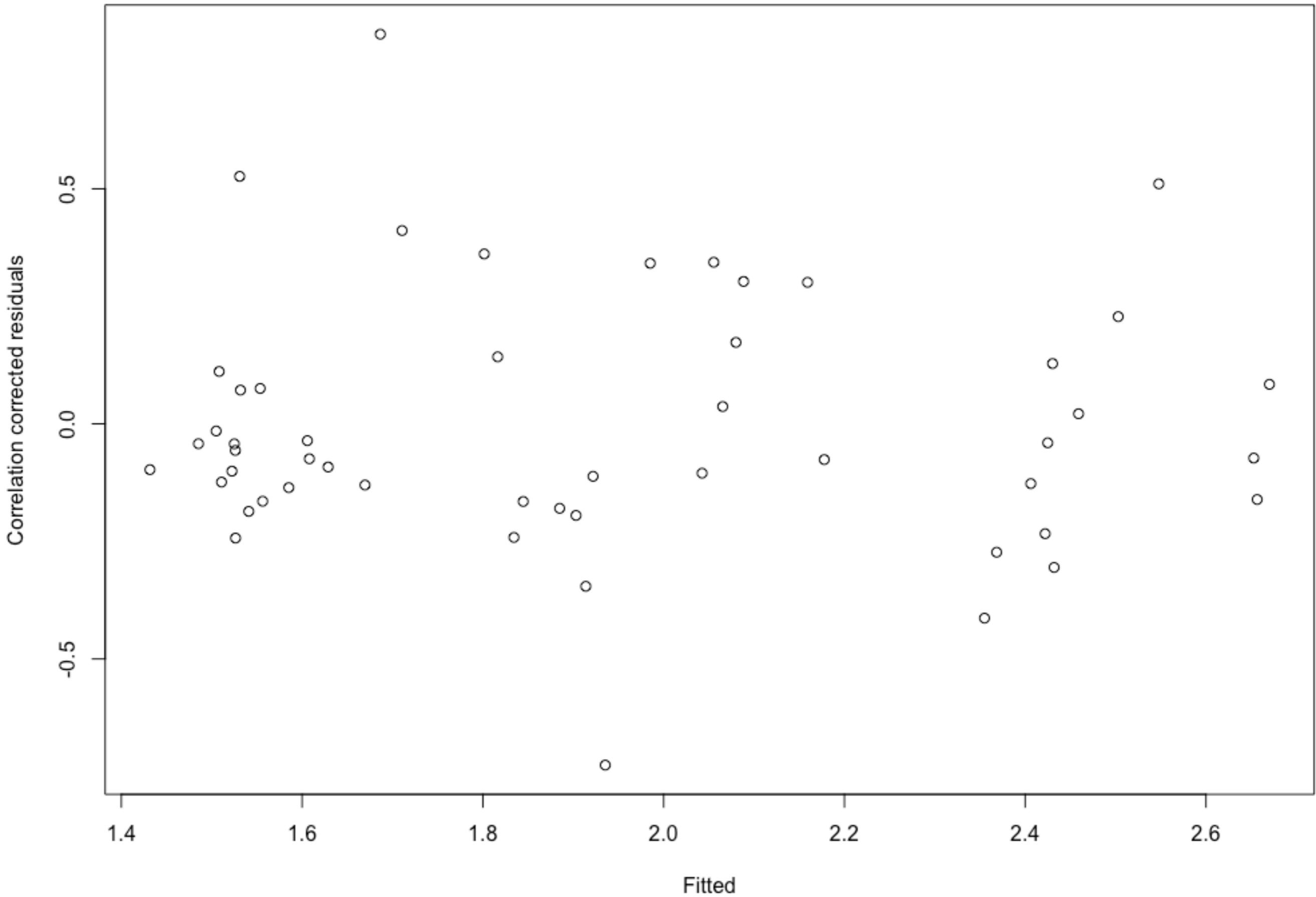
Given that residuals violate OLS assumptions, generate new model.
In the new model, only two parameters are significant.

$$\text{price} = \beta_0 + \beta_1 t + \beta_2(\text{yield})$$

Correlation corrected residuals PACF



GLS Residuals v. Fitted



Testing Data

Generalized least squares fit by REML

Model: price ~ t + yield

Data: Corn1952Testing

AIC BIC logLik
43.2713 45.26078 -16.63565

Correlation Structure: AR(1)

Formula: ~t

Parameter estimate(s):

Phi

0.9999689

Coefficients:

	Value	Std. Error	t-value	p-value
(Intercept)	-12.469298	91.13226	-0.136826	0.8936
t	0.172401	0.19087	0.903260	0.3857
yield	-0.057778	0.01387	-4.166236	0.0016

Correlation:

(Intr) t

t -0.300

yield 0.002 -0.091

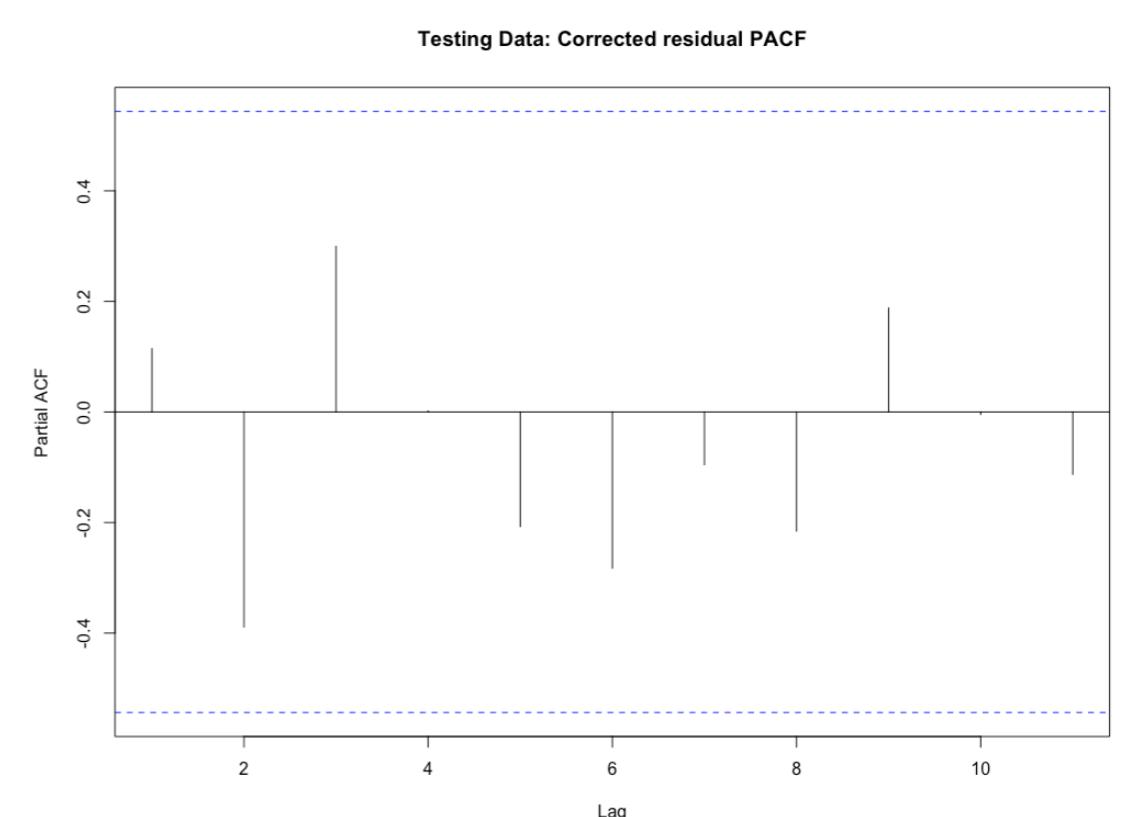
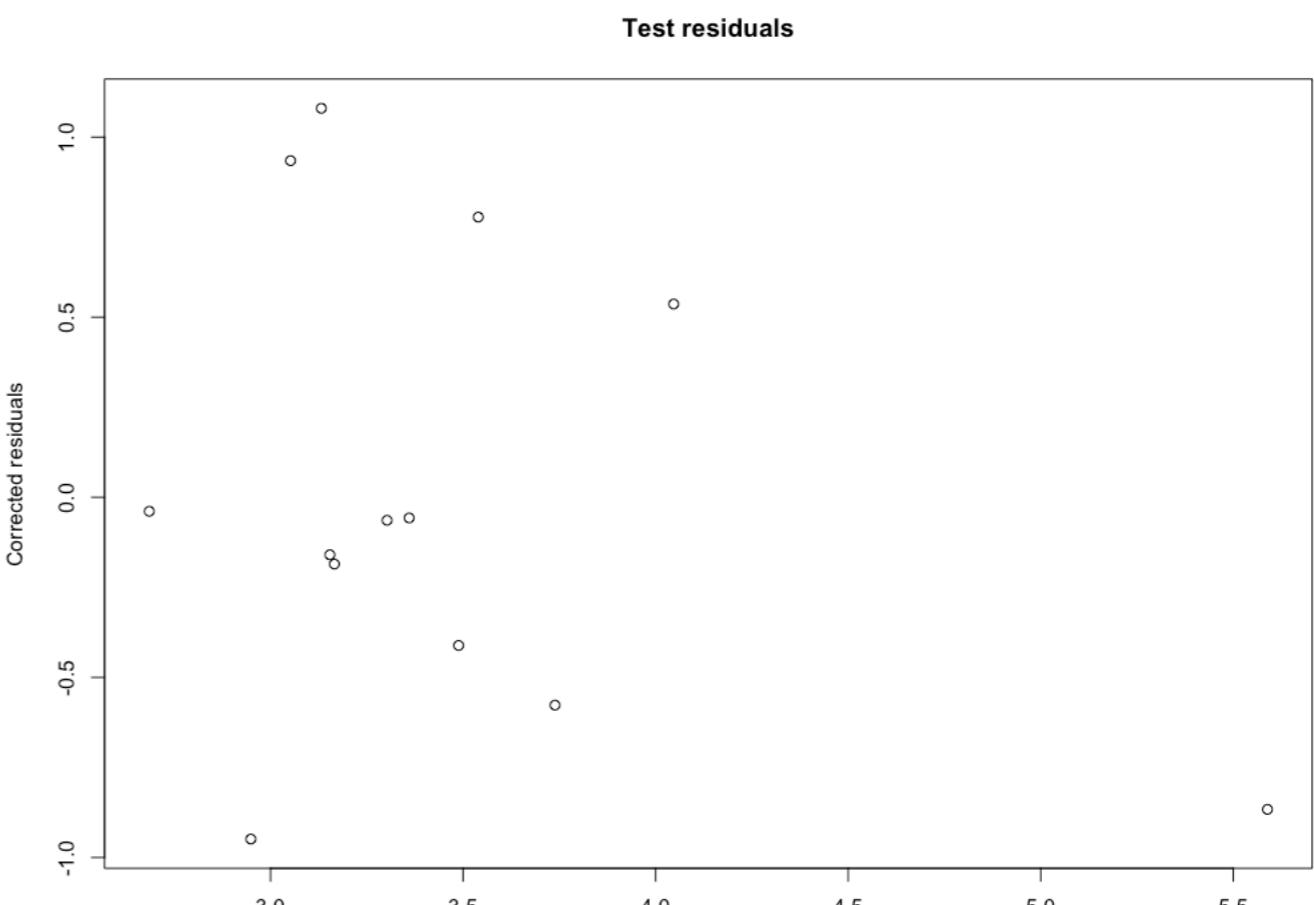
Standardized residuals:

Min	Q1	Med	Q3	Max
-0.0109199437	0.0001636806	0.0072220332	0.0118228856	0.0250053830

Residual standard error: 86.9036

Degrees of freedom: 14 total; 11 residual

Low n...can't really trust this very much.



Conclusions

- Further considerations could be made to correct minimal NCV, such as taking $\log(\text{price})$ and using WGLS.
- Since only yield/time were significant, refit using 1866 data set to see if still significant. This would give higher n and possibly better predictions.
- Assuming the United States exists in its current form in several hundred years, repeating analysis with more data points would be useful.