

Inverkehrsetzungen in der Schweiz

Code ▼

Markus Ilg

(mailto:#)markus.ilg@markus-ilg.at (mailto:markus.ilg@markus-ilg.at)

Überblick

Begleitendes Beispiel (Ilg / Baumeister 2019).

Datenherkunft

Schweizerische Eidgenossenschaft - Bundesamt für Statistik.

Abgerufen aus STAT-TAB am 24.8.2019:

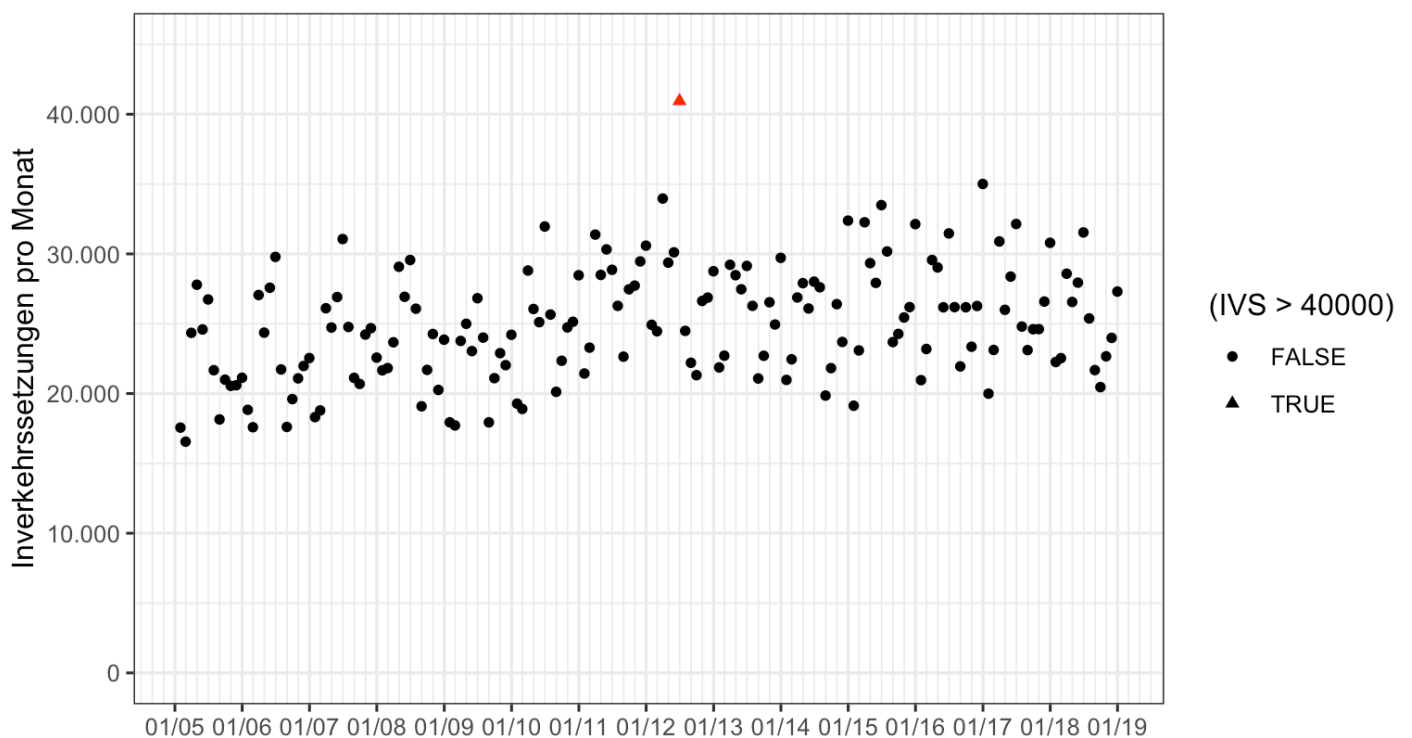
https://www.pxweb.bfs.admin.ch/pxweb/de/px-x-1103020200_120/-/px-x-1103020200_120.px

(https://www.pxweb.bfs.admin.ch/pxweb/de/px-x-1103020200_120/-/px-x-1103020200_120.px)

Datenaufbereitung

Die Daten wurden aus STAT-TAB als csv-Datei exportiert.

monatliche Inverkehrsetzungen von Personenkraftwagen
Schweiz, 2005 - 2018



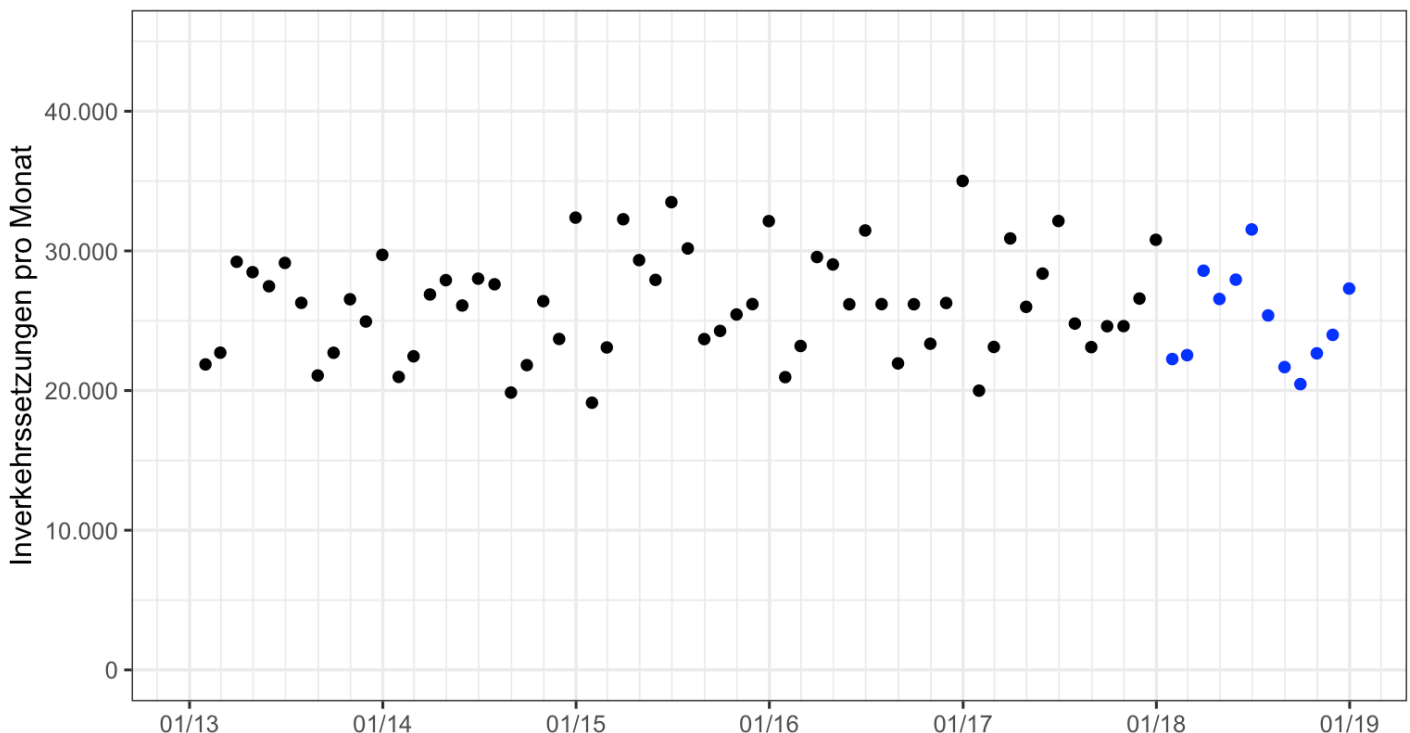
Sofort fällt der Peak im Juni 2012 bei Inverkehrsetzungen auf - verursacht durch eine Änderung der CO2-Emissionsvorschriften ab Juli 2012 (<https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/fuer-medienschaffende/medienmitteilungen.assetdetail.36344.html> (<https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/fuer-medienschaffende/medienmitteilungen.assetdetail.36344.html>)).

Für den Zweck der Analyse - eine Prognose der Entwicklung des Automobilmarktes für ein Jahr, ist dieser Ausreißer eher störend. Sinnvoller ist es auf Daten ab 2013 zurückzugreifen. Ein erster Blick auf die Daten mit `summary` zeigt, dass sich die Inverkehrsetzungen im Intervall von 19.129 bis 35.008 bewegen, im Durchschnitt sind es 26.064 Inverkehrsetzungen. Es werden keine fehlenden Werte angezeigt.

Datum	IVS
Min. :2013-01-31	Min. :19129
1st Qu.:2014-07-23	1st Qu.:23105
Median :2016-01-15	Median :26184
Mean :2016-01-14	Mean :26064
3rd Qu.:2017-07-07	3rd Qu.:28502
Max. :2018-12-31	Max. :35008

In der folgenden Abbildung sind die Daten aus dem Berichtsjahr 2018 blau gekennzeichnet, das sie in der Folge nicht zur Schätzung sondern zur Validierung des Modells herangezogen werden, um dessen Prognosefähigkeit beurteilen zu können. Konnte man beim Analysezeitraum 2005 - 2018 noch einen positiven linearen Trend bei Inverkehrsetzungen vermuten, so ist dies in der verkürzten Analyseperiode nicht erkennbar. Die Werte schwanken zudem erheblich, wobei die meisten (90 %) der Werte zwischen 21.700 und 31.400 Inverkehrsetzungen liegen.

monatliche Inverkehrsetzungen von Personenkraftwagen
Schweiz, 2013 - 2018



Modellierung

Lineares Grundmodell

Zahlreiche Modelle werden in der Zeitreihenanalyse verwendet (vgl. einführend Backhaus et al. (2018) (<https://www.springer.com/de/book/9783662566541>), S. 126 ff). Beim *linearen Modell* erfolgt in der Grundform additive Zerlegung $Y = A + K + S + u$ mit Y als zu erklärender Variable, A als Trendkomponente, K als Konjunkturkomponente (lange zyklische Komponente), S als Saisonkomponente und u als Störgröße oder zufällige Komponente. Im einfachsten Fall ohne Saison- und Konjunkturkomponenten reduziert sich das Modell auf $Y = A + u$. Mit der Spezifikation der Trendkomponente A entsteht das lineare Trendmodell:

$$Y = \alpha + \beta \cdot t + u$$

Auch zahlreiche nichtlineare Modelle können über geeignete Transformationen im linearem Kontext analysiert werden. Zu nennen sind bspw. das Quadratwurzelmodell $Y = \alpha + \beta\sqrt{t} + u$ oder das logarithmische Modell $Y = \alpha + \beta \cdot \ln(t) + u$, die beide Sättigungseffekte abbilden können. Für einen Überblick über weitere Modelle und geeignete Linearisierungen vgl. Backhaus et al. (2018) (<https://www.springer.com/de/book/9783662566541>), S 148 ff.

Call:

```
lm(formula = IVS ~ Datum, data = ivs_simple)
```

Residuals:

Min	1Q	Median	3Q	Max
-6917.4	-2967.7	110.1	2465.7	8925.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.520e+04	1.160e+04	2.172	0.0332 *
Datum	5.133e-02	6.894e-01	0.074	0.9409

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3701 on 70 degrees of freedom

Multiple R-squared: 7.918e-05, Adjusted R-squared: -0.01421

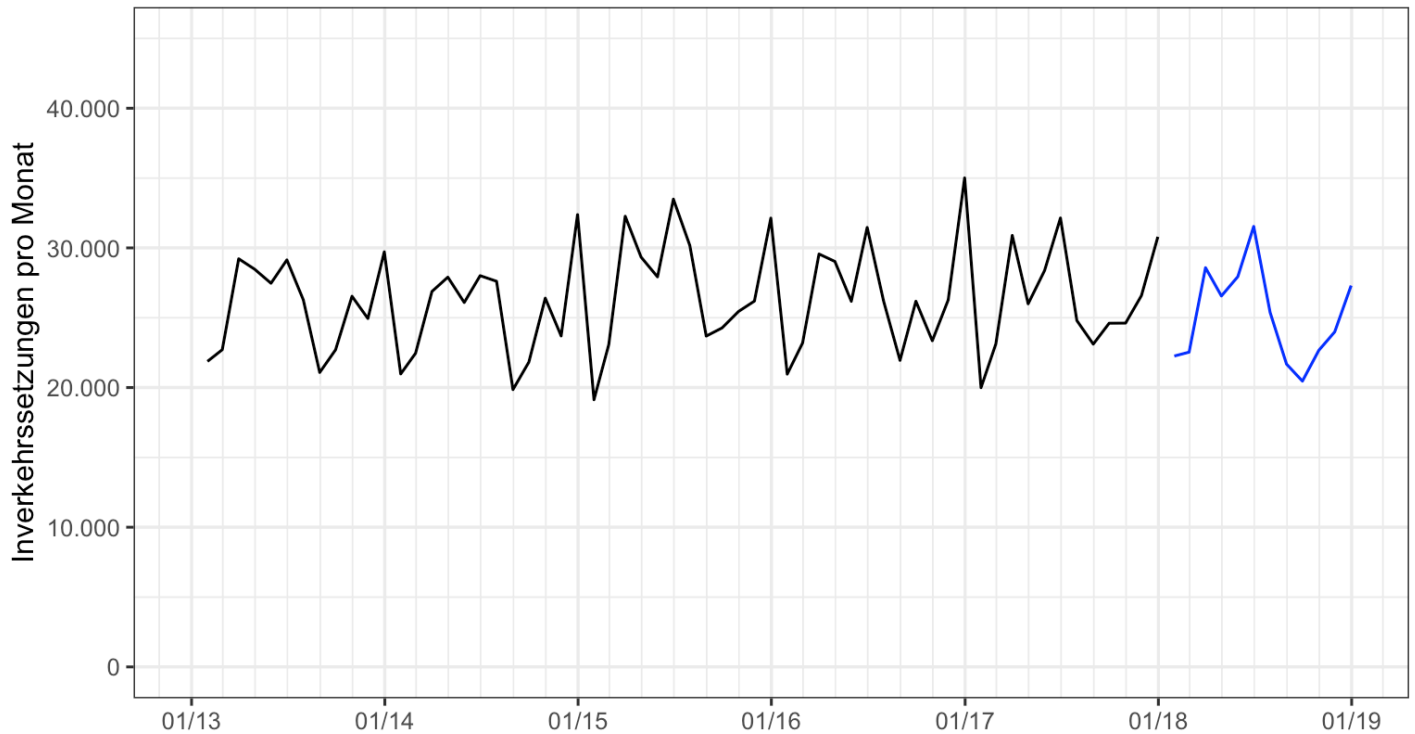
F-statistic: 0.005543 on 1 and 70 DF, p-value: 0.9409

Das Bestimmtheitsmaß R^2 ist praktisch 0, das Modell trägt nicht zur Erklärung der Streuung der Zeitreihe bei. Sowohl das Modell insgesamt (F-Statistik) als auch die Parameterschätzung für den Regressor *Datum* sind nicht signifikant.

Berücksichtigung saisonaler Schwankungen

Das lineare Trendmodell liefert keinen Erklärungsbeitrag. Werden die Datenpunkte als Liniendiagramm dargestellt, wird eine zyklische Schwankung deutlich. Für die Modellbildung werden nur die Daten von 2013 bis 2017 verwendet, 2018 wird für die Beurteilung der Modellperformance zurückgehalten.

monatliche Inverkehrsetzungen von Personenkraftwagen Schweiz, 2013 - 2018



Modelliert man die monatlichen Schwankungen, erhöht sich der Erklärungsgehalt des Modells deutlich. Dazu werden die Monate als zweistellige Zeichenketten modelliert ("01" bis "12"). R erzeugt daraus automatisch 11 Dummyvariablen zur Abbildung der Monate im Modell.

Call:

```
lm(formula = IVS ~ Datum + Monat, data = ivs_simple_train)
```

Residuals:

Min	1Q	Median	3Q	Max
-2676.5	-1014.5	170.3	724.4	3164.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.120e+04	6.427e+03	1.742	0.08801	.
Datum	5.702e-01	3.880e-01	1.469	0.14839	
Monat02	2.311e+03	9.821e+02	2.353	0.02284	*
Monat03	9.143e+03	9.823e+02	9.308	3.08e-12	***
Monat04	7.511e+03	9.827e+02	7.644	8.71e-10	***
Monat05	6.552e+03	9.832e+02	6.664	2.64e-08	***
Monat06	1.018e+04	9.838e+02	10.345	1.06e-13	***
Monat07	6.321e+03	9.846e+02	6.420	6.22e-08	***
Monat08	1.228e+03	9.855e+02	1.246	0.21899	
Monat09	3.193e+03	9.866e+02	3.236	0.00222	**
Monat10	4.528e+03	9.878e+02	4.584	3.38e-05	***
Monat11	4.778e+03	9.891e+02	4.831	1.49e-05	***
Monat12	1.123e+04	9.906e+02	11.337	4.79e-15	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1553 on 47 degrees of freedom

Multiple R-squared: 0.862, Adjusted R-squared: 0.8267

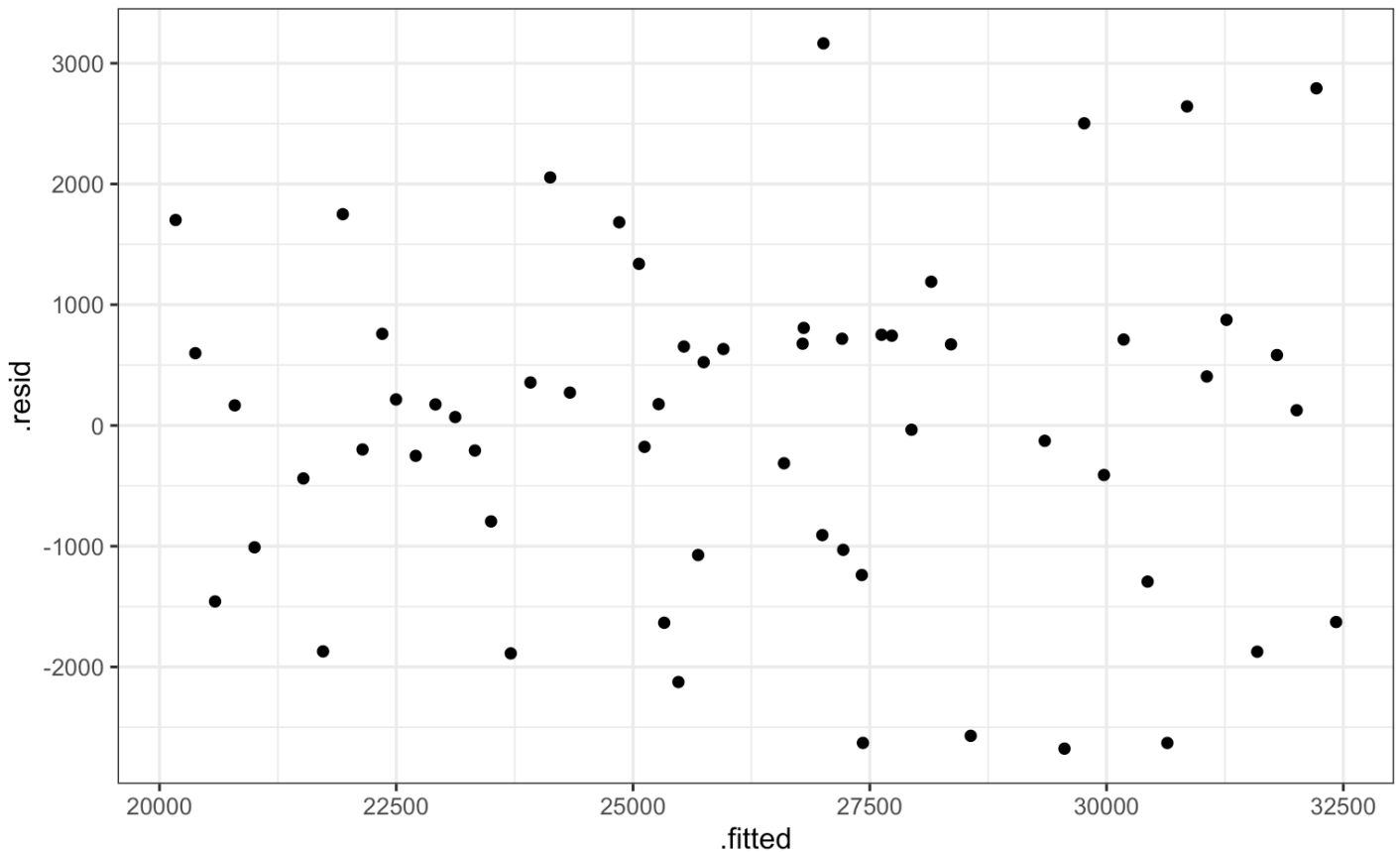
F-statistic: 24.46 on 12 and 47 DF, p-value: 3.282e-16

Der Anteil der erklärten Varianz erhöht sich nun auf über 86 Prozent. Die meisten der Saison-Dummies haben statistisch signifikante Koeffizienten, gleiches gilt für das Modell im Gesamten (vgl. F-Statistik).

Prüfung der Modellierungsvoraussetzungen

Wir prüfen die Modellierungsvoraussetzungen analog zu Backhaus et al. (2018) (<https://www.springer.com/de/book/9783662566541>), S. 98 ff).

1. Nitlinearität: die graphische Darstellung lässt - unter Berücksichtigung der Saisonkomponente einen darunter liegenden linearen Trend vermuten.
2. Erwartungswert der Störgröße = 0: ist nicht von Bedeutung, wirkt sich ggf. auf den Koeffizienten b_0 aus.
3. Falsche Regressorenauswahl: Multikollinearität falsch spezifizierter Regressoren verzerrt die Parameterschätzer, bei fehlender Multikollinearität wirken sie sich zumindest wieder auf den Ordinatenabschnitt b_0 aus. Mit *Datum* und *Monat* sind hier allerdings keine überflüssige Regressoren erkennbar.
4. Heteroskedastizität: Der Plot der Residuen gegen die erklärte Variable lässt keine für Heteroskedastizität typischen Verläufe erkennen.



Heteroskedastizität kann auch mit dem Goldfeld-Quandt-Test überprüft werden. Die Nullhypothese (Homoskedastizität) kann nicht verworfen werden.

Goldfeld-Quandt test

```
data: mod
GQ = 0.85735, df1 = 17, df2 = 17, p-value = 0.6227
alternative hypothesis: variance increases from segment 1 to 2
```

- Autokorrelation: Insbesondere bei Zeitreihen besteht die Gefahr der Verzerrung des Standardfehlers der Regressionskoeffizienten durch autokorrelierte Residuen. Im Ergebnis ergeben sich dann verzerrte Konfidenzintervalle für die Regressionskoeffizienten. Eine Visuelle Prüfung erfolgt wie bei der Heterskedastizität im Plot der Residuen gegen die geschätzten Werte. Ebenso ist die Prüfung mit Durbin-Watson-Test möglich, dessen Nullhypothese (keine Autokorrelation) nicht verworfen werden kann.

Durbin-Watson test

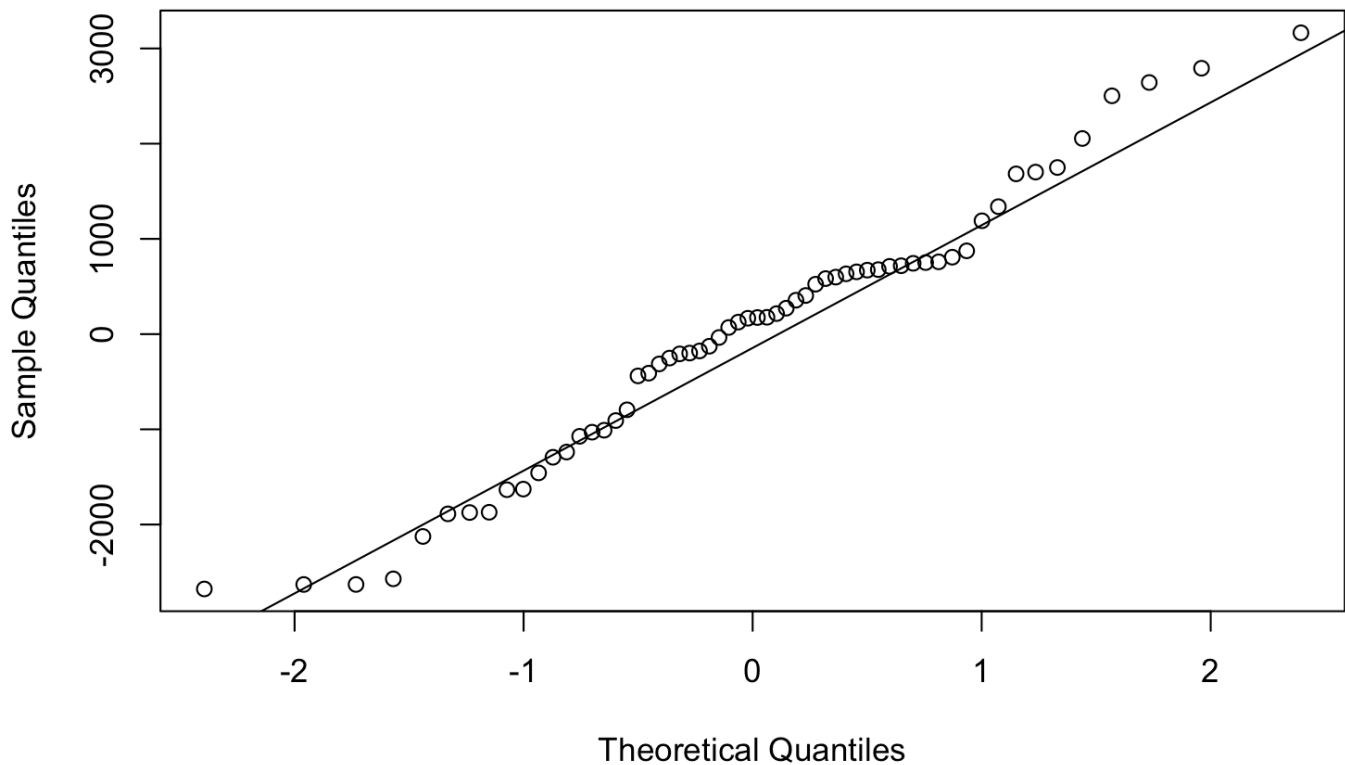
```
data: mod
DW = 1.9995, p-value = 0.4883
alternative hypothesis: true autocorrelation is greater than 0
```

- Multikollinearität: Schätzungen für Regressionskoeffizienten werden unzuverlässiger, wenn sich die

Koeffizienten aus anderen Modellkoeffizienten errechnen lassen. Paarweise Multikollinearität ist vorhanden, wenn der Betrag der paarweisen Korrelationen nahe 1 ist. Abhängigkeiten von mehreren anderen Variablen kann durch die Toleranz oder deren Kehrwert, den Varianz-Inflation-Factor (VIF) ermittelt werden. VIF-Werte im Bereich von 10 oder größer sind auf jeden Fall problematisch. Im Beispiel ist die Multikollinearität mit Datum bzw. Monat als Regressoren nicht relevant.

7. Normalverteilung der Residuen: Bei großer Stichprobe ist eine Abweichung von der Normalverteilungsannahme nicht so tragisch. Der QQ-Plot sieht allerdings nicht perfekt aus.

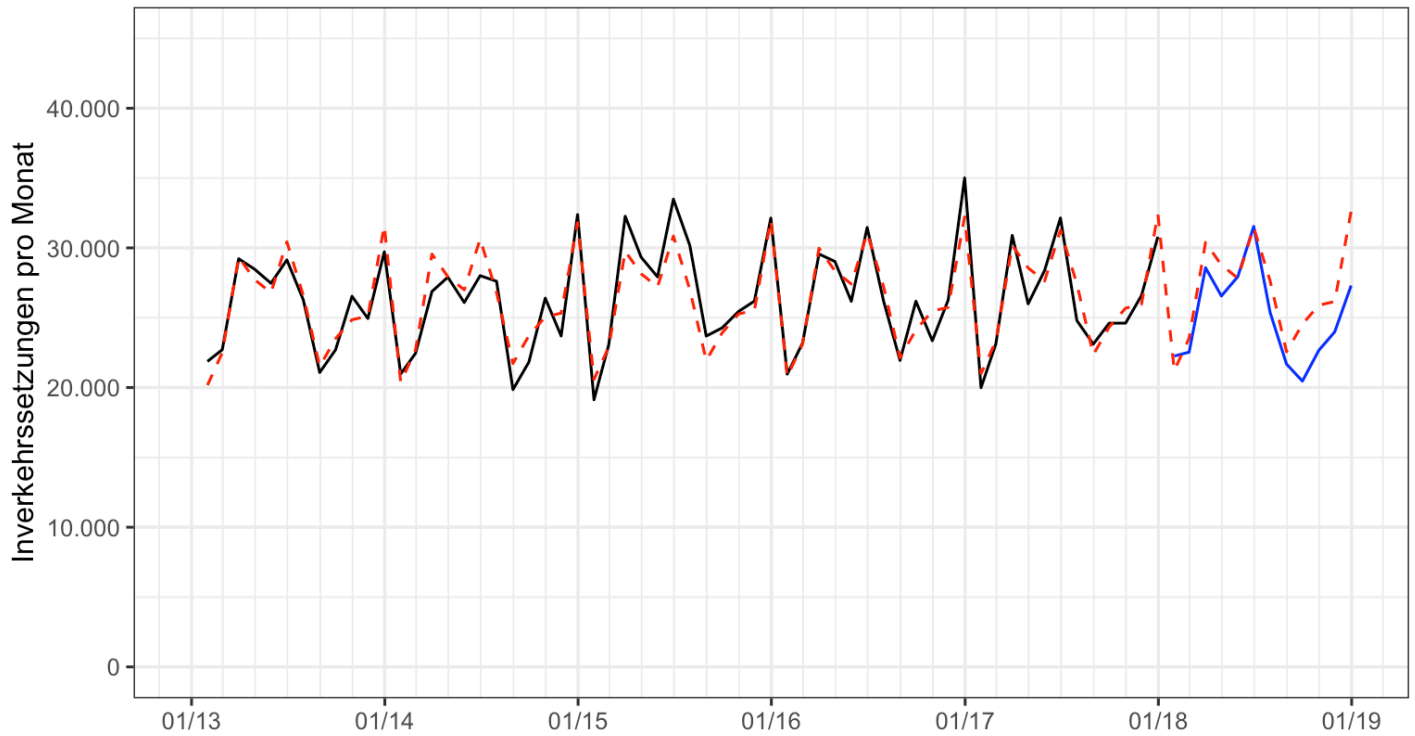
Normal Q-Q Plot



Erstellung von Prognosen

Das berechnete Modell erlaubt die Berechnung von Prognosewerten. Durch den Vergleich der Prognosewerte für die Testdaten mit den Ist-Werten der Testdaten (2018) können Performance-Maße berechnet werden, die den Vergleich unterschiedlicher Modelle erlauben.

monatliche Inverkehrsetzungen von Personenkraftwagen Schweiz, 2013 - 2018



mae <dbl>	mape <dbl>	rmse <dbl>	mse <dbl>
2016.205	0.08344398	2522.631	6363668

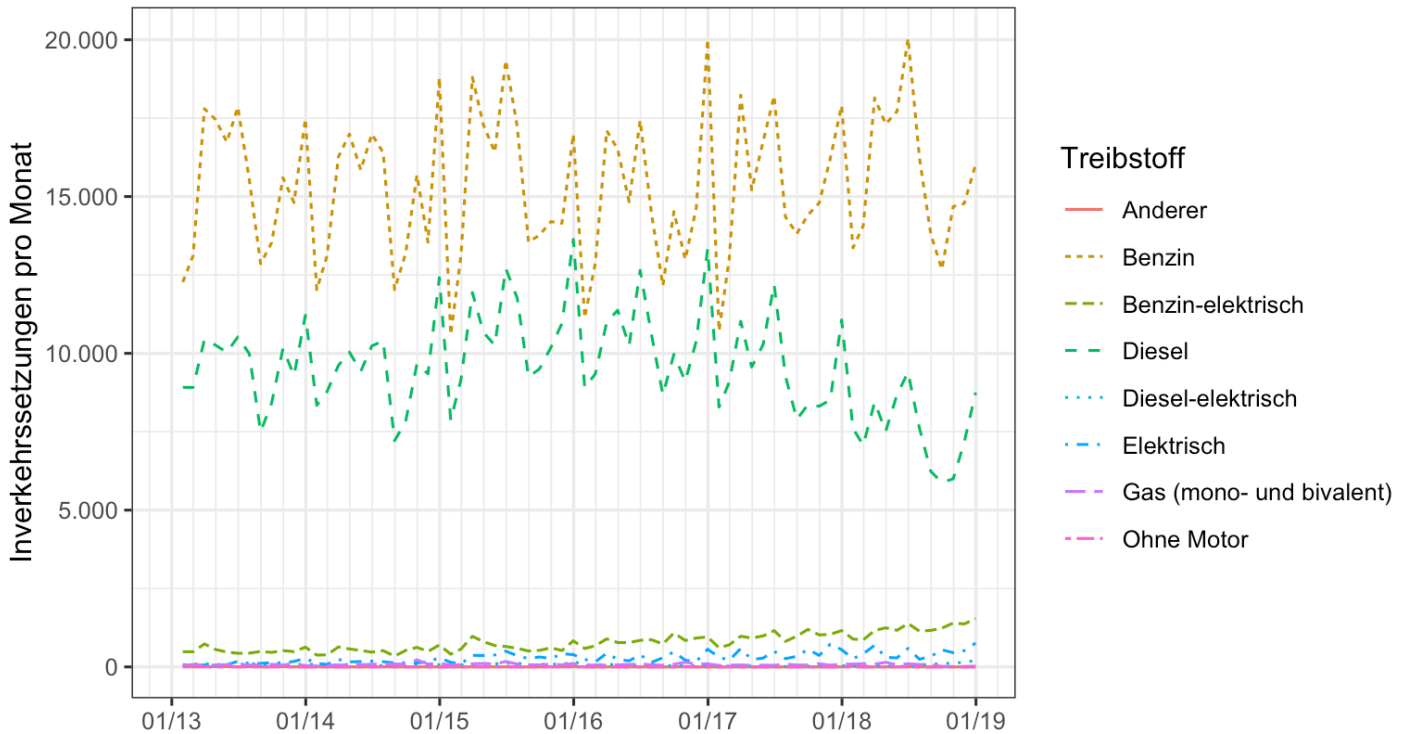
1 row

Differenzierung der Antriebsart

Die Wiederholung der Analyse mit dem zusätzlichen Regressor zeigt ein völlig anderes Bild.

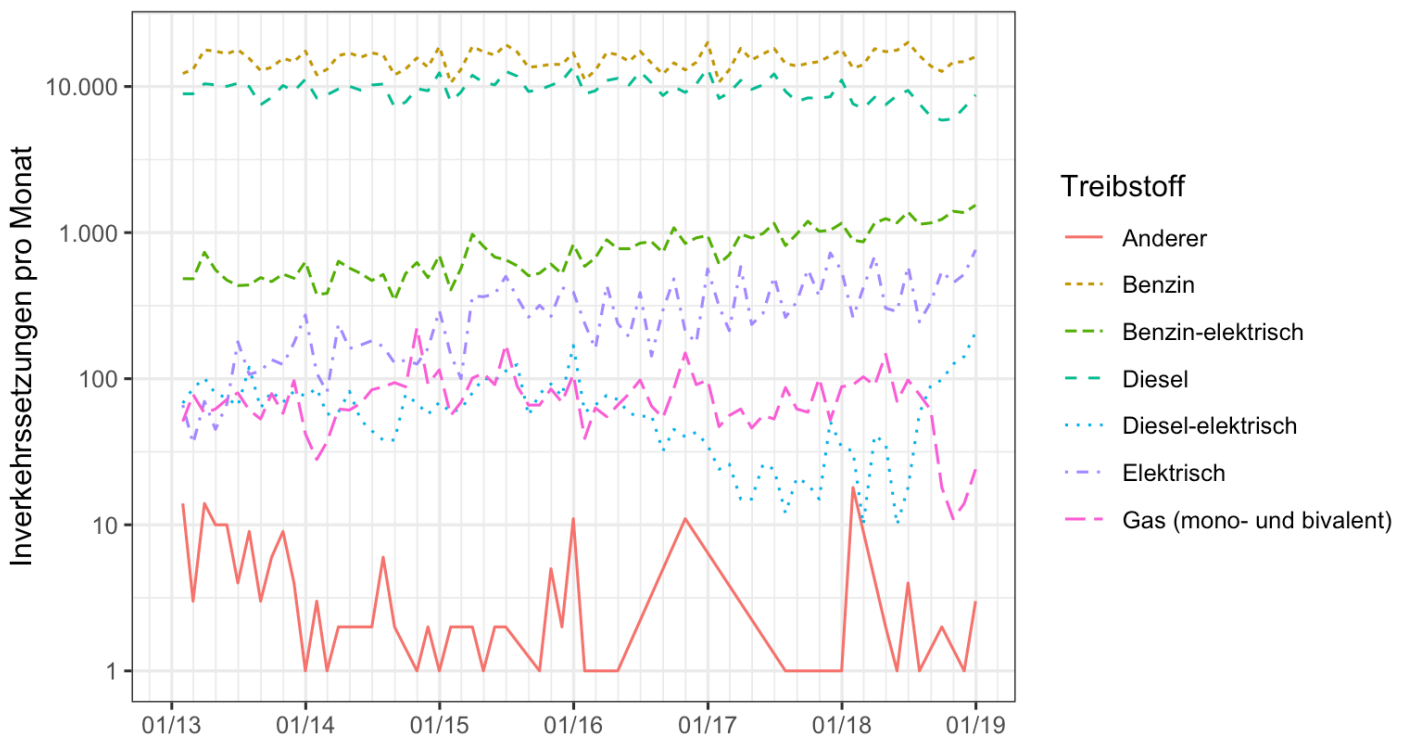
Inverkehrsetzungen PKW, nach Treibstoffart

Schweiz, 2013 - 2018



Inverkehrsetzungen PKW, nach Treibstoffart

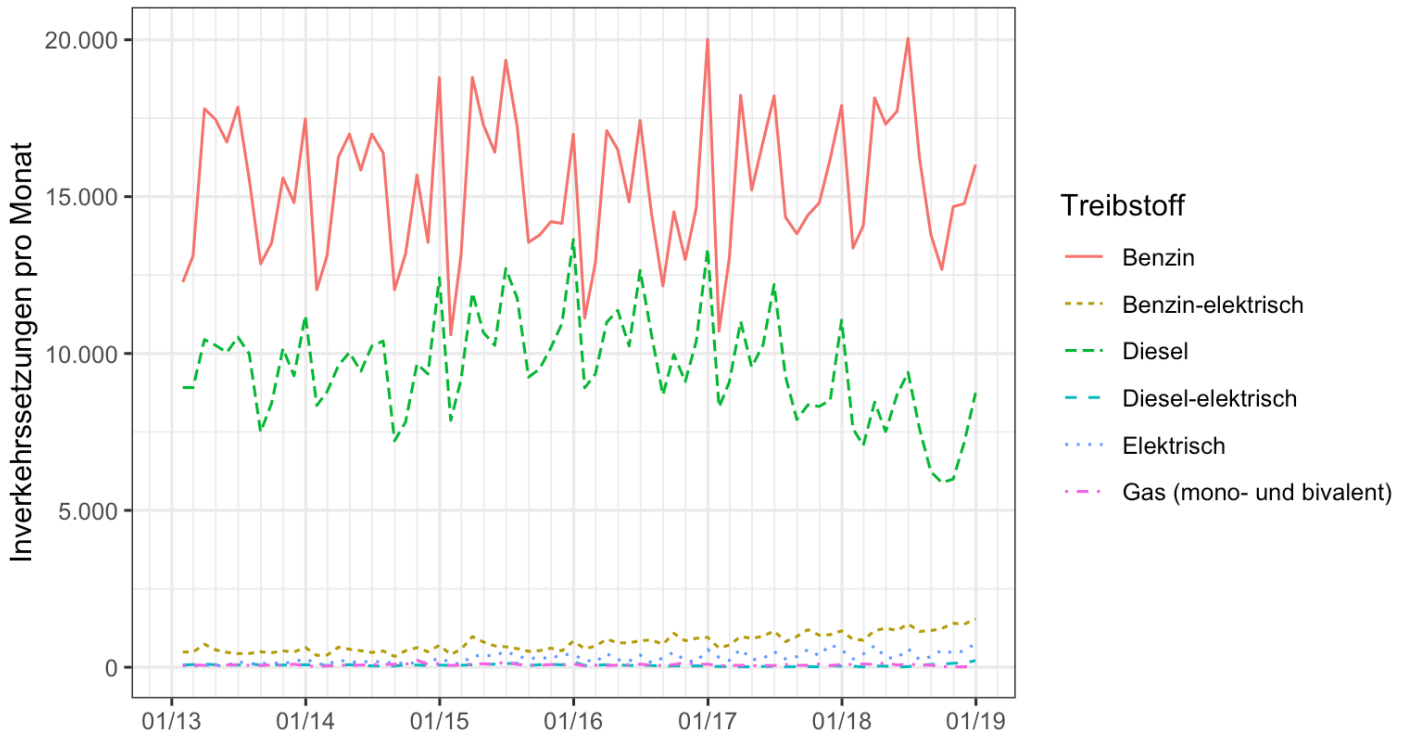
Schweiz, 2013 - 2018



Die Treibstoffarten “Anderer” und “Ohne Motor” spielen in der Analyse keine besondere Rolle. Zur besseren Übersichtlichkeit werden sie ausgeschlossen.

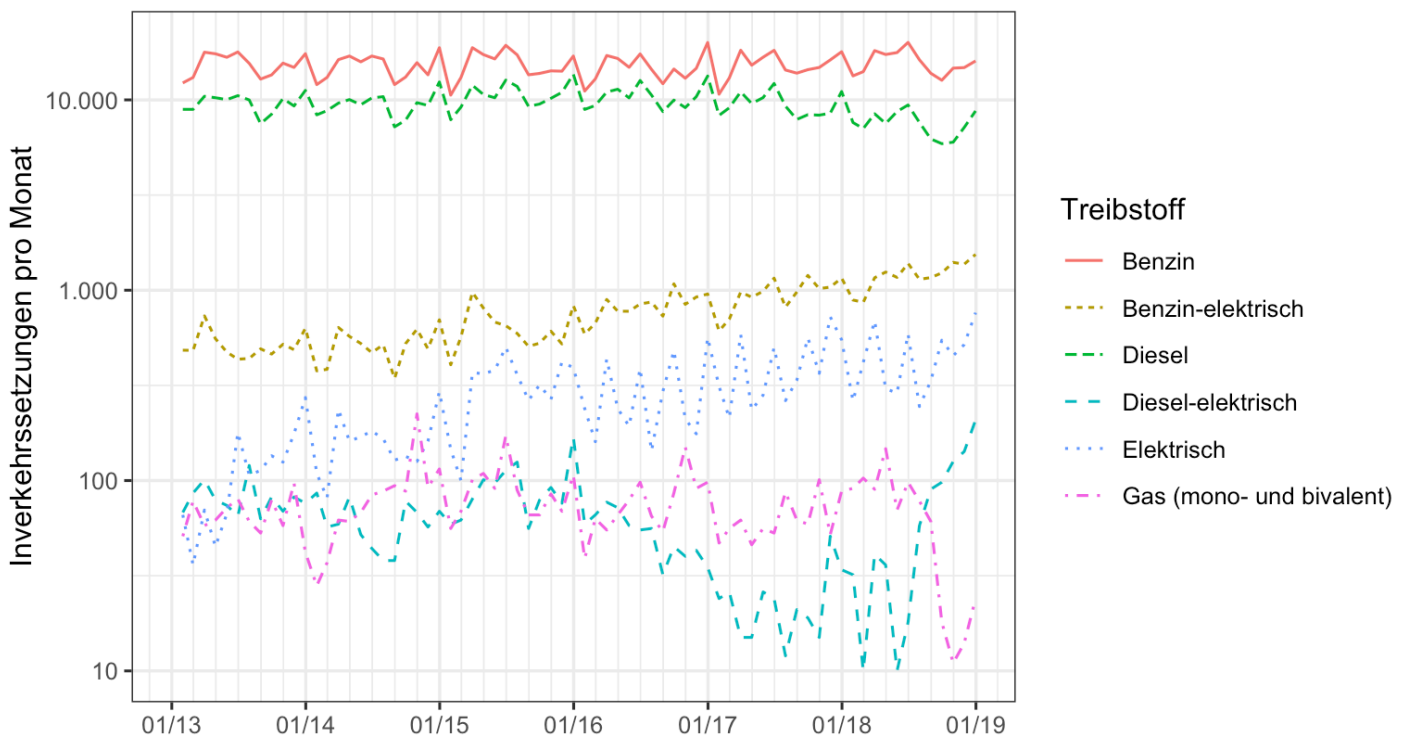
Inverkehrsetzungen PKW, nach Treibstoffart

Schweiz, 2013 - 2018



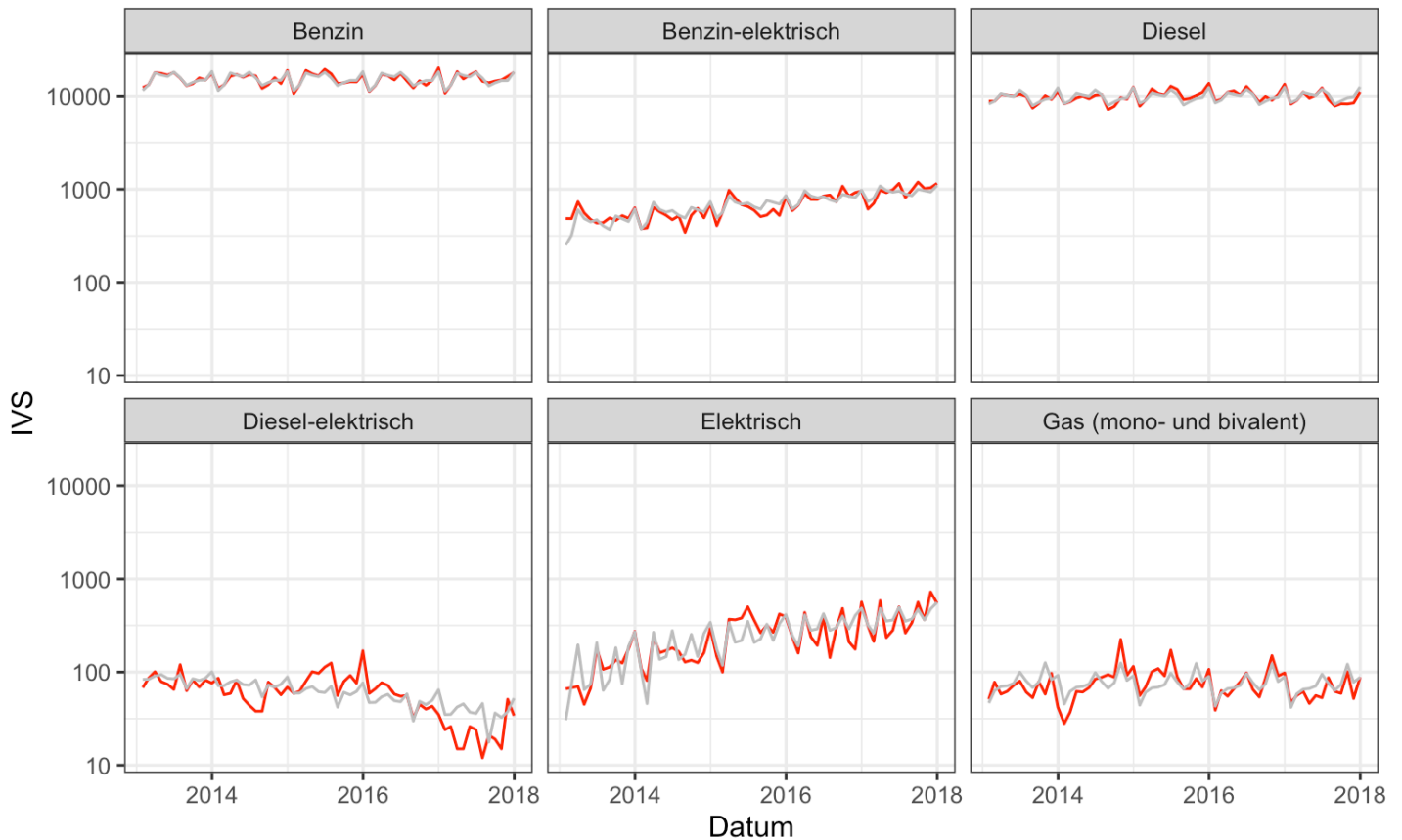
Inverkehrsetzungen PKW, nach Treibstoffart

Schweiz, 2013 - 2018



Modelle je Treibstoffart

Modelle berechnen und plotten

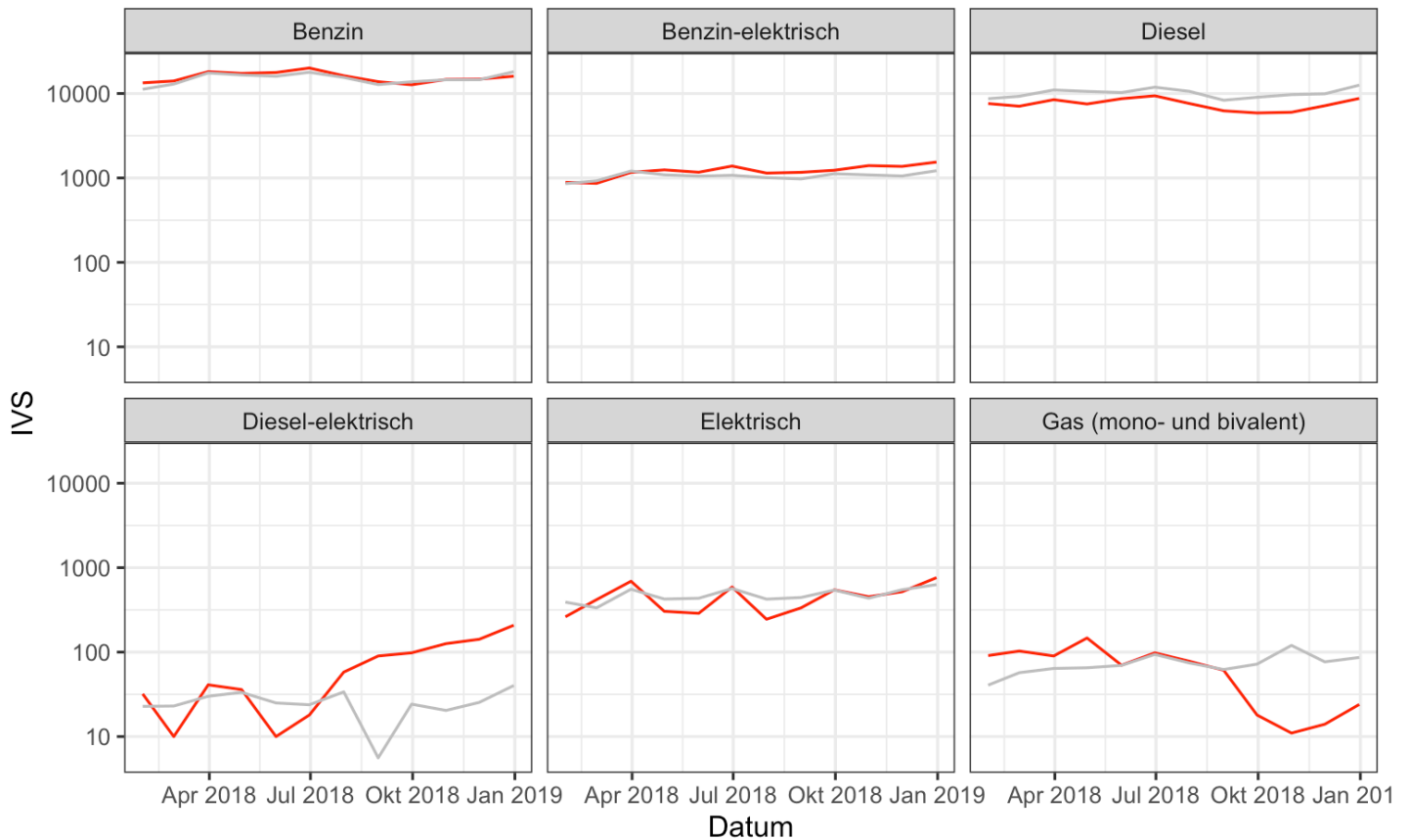


Treibstoff <chr>	model <list>	r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>
Gas (mono- und bivalent)	<S3: lm>	0.3690224	0.2079218	28.42973	2.290633
Diesel-elektrisch	<S3: lm>	0.3968938	0.2429092	26.49653	2.577491
Elektrisch	<S3: lm>	0.7245803	0.6542603	92.08874	10.304053
Diesel	<S3: lm>	0.7328046	0.6645845	831.86707	10.741769
Benzin-elektrisch	<S3: lm>	0.8037181	0.7536035	110.86649	16.037624
Benzin	<S3: lm>	0.8699177	0.8367052	919.96558	26.192473

6 rows | 1-8 of 13 columns

Performanceberechnung

Je Treibstoffart werden für das jeweilige Modell die Modellkennzahlen ermittelt.



Quellen

Backhaus et al. (2018): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 15 Aufl. Berlin. Heidelberg. Springer-Verlag (<https://www.springer.com/de/book/9783662566541>).

Ilg, Markus; Baumeister, Alexander (2019): Analytics im Marketing-Controlling. In: Die Digitalisierung des Controllings: Anwendungsbeispiele aus Theorie und Praxis, hrsg. von Imke Keimer und Ulrich Egle. Springer.

Ruppert, David; Matteson, David S. (2015): Statistics and Data Analysis for Financial Engineering Springer-Verlag (<https://www.springer.com/de/book/9781493926138>)