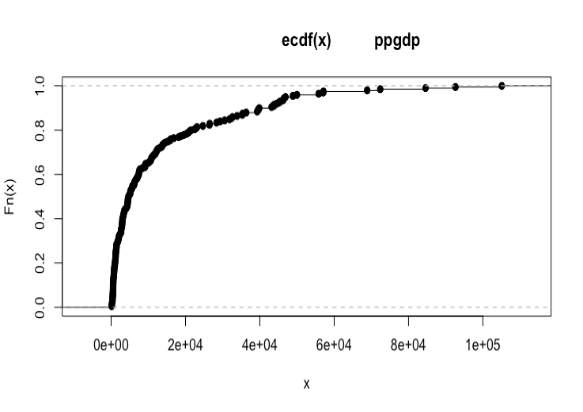
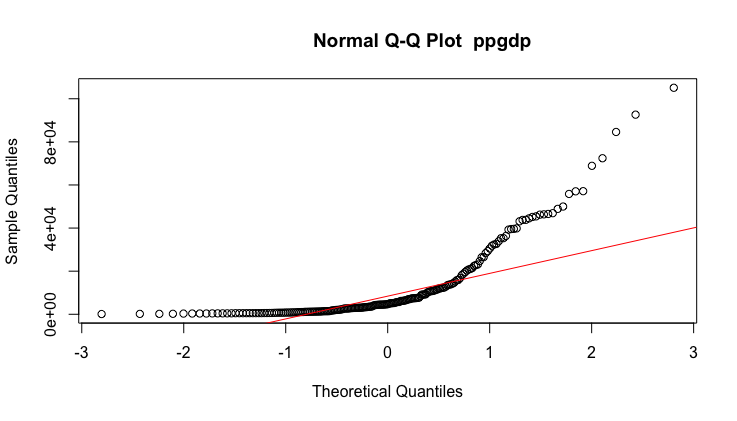
**HÜ4**

1.



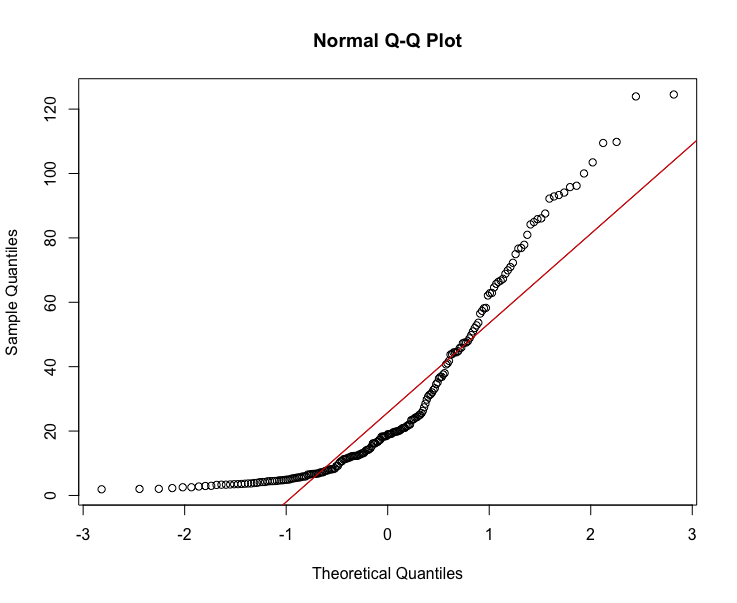
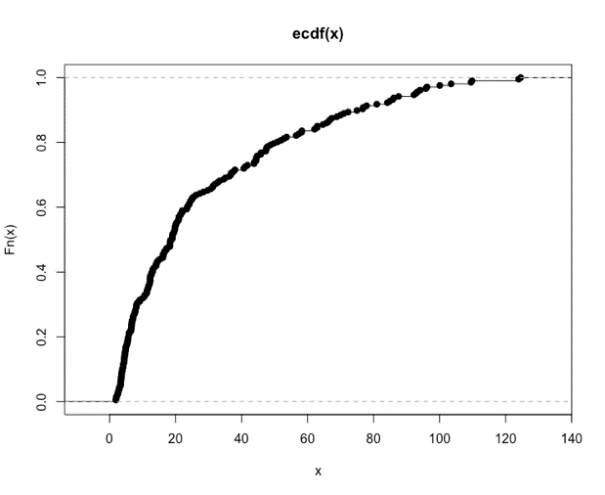
Charakteristika von ppgdp:

Unimodale Datenverteilung (ersichtlich am Histogramm). Verteilung und Dichtekurve weisen auf eine Datenverteilung mit einer Schiefe nach rechts hin. Am QQPlot ersichtlich ist ein schwerer rechter Rand (Punkte entfernen sich stärker von der roten Linie hinweg) und ein leichter Rand links.

> summary(UN$ppgdp)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

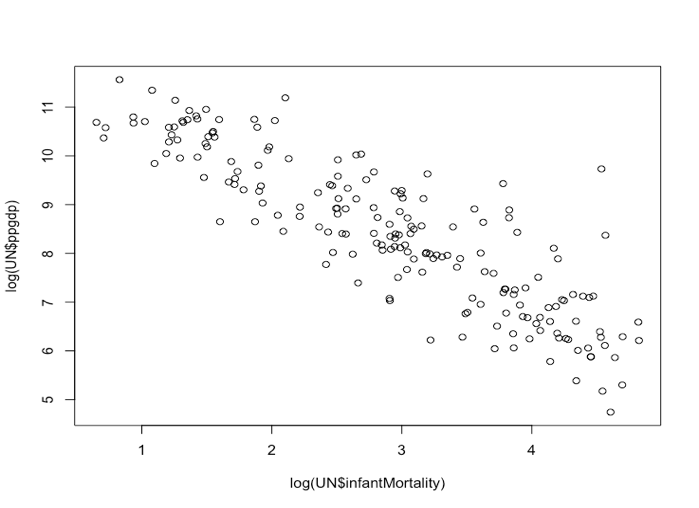
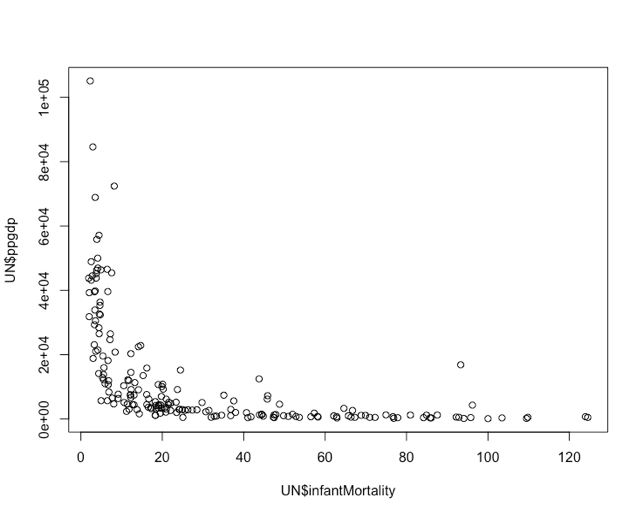
114.8 1283.0 4684.5 13012.0 15520.5 105095.4 14



Charakteristika zu Infant mortality:   
Vergleichbar mir ppgdp- Unimodale Datenverteilung. Verteilung und Dichtekurve weisen auf eine Datenverteilung mit einer Schiefe nach rechts auf. QQPlot zeigt einen schweren rechten Rand und einen leichten linken Rand.

|  |
| --- |
| > summary(UN$infantMortality) |
| Min. 1st Qu. Median Mean 3rd Qu. Max. NA's | |
| 1.916 7.019 19.007 29.440 44.477 124.535 6 | |

|  |
| --- |
|  |



Beim Scatterplott wurden beiden Daten aneinandergelegt und miteinander verglichen. Dabei wurde festgestellt, dass diese Daten miteinander im Zusammenhang stehen. Dieser Zusammenhang ist jedoch nicht linear (keine Gerade).

Auf dem rechten Bild wurden die Werte logarithmiert und abermals via Scatterplot dargestellt. Zwar ist ein Streuungseffekt gegeben, jedoch ist eine Abhängigkeit zu erkennen. Daher wird ein Test auf Korrelation durchgeführt.

Hypothesenaufstellung:

H0: Korrelation ist gleich null.

HA: Korrelation ist nicht gleich null.

> cor.test(log(UN$infantMortality), log(UN$ppgdp))

Pearson's product-moment correlation

data: log(UN$infantMortality) and log(UN$ppgdp)

t = -25.018, df = 191, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9047107 -0.8376375

sample estimates:

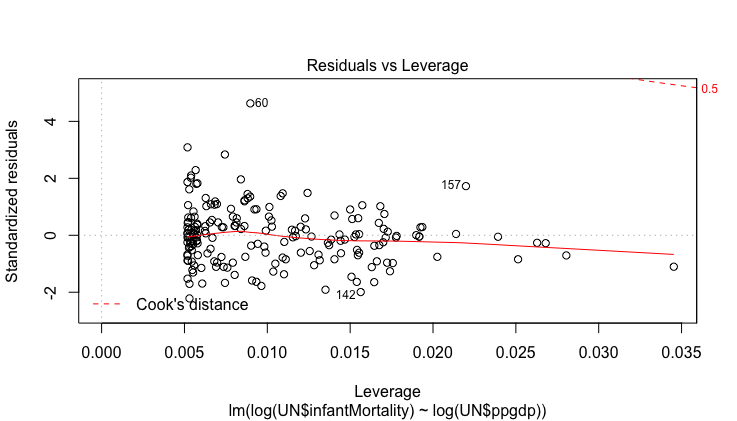
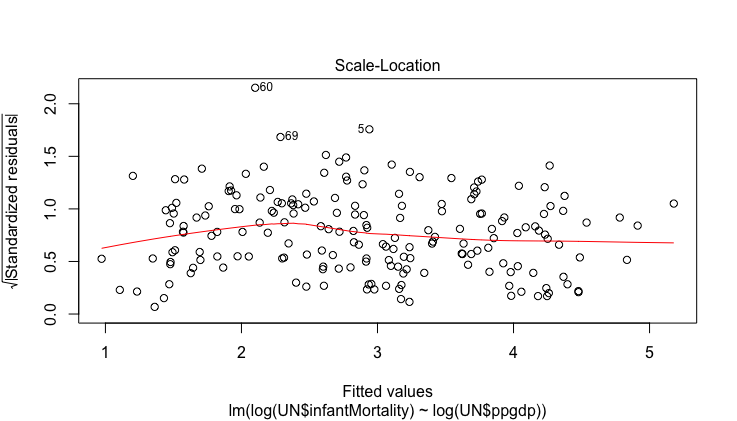
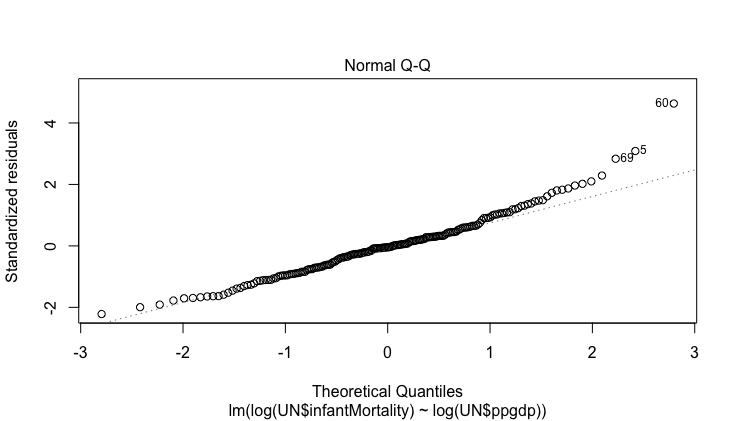
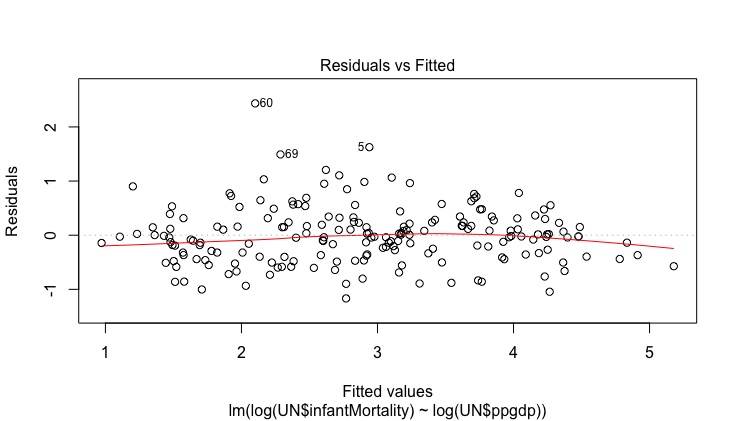
cor

-0.8753203

t steht für die Teststatistik und liegt bei 25,018; Df gibt die Freiheitsgrade an. Dadurch lässt sich die Anzahl der Stichproben ermitteln. Df=191(+2) = 193 Gesamtproben.

Der p-Wert (2,2e-16) ist < 0.05 und daher hoch signifikant. Dieser gibt Auskunft darüber, dass die H0 verworfen werden kann. Der gesuchte Wert liegt zu 95% in einem Intervall zwischen -0.9047107 und -0.8376375. Der geschätzte Korrelationswert (-0.8753203) gibt nun an, dass eine negative Korrelation vorliegt. Dies bedeutet, dass eine Erhöhung des Bruttoinlandsproduktes zu einer Verringerung der Kindersterblichkeit führt.

**Ist das Modell gültig?**



1. Bei dem Plot Residuals vs. Fitted werden die Residuen gegen die vom Model vorhergesagten Werte aufgetragen. Die Schwankungsbreite der Daten erstreckt sich von -1 bis +2. Die Linie weist Schwankungen im negativen Skalenbereich auf, diese Schwankung ist jedoch nur schwach ausgeprägt. Sowohl xi wie auch xj sind unabhängig und unkorelliert und haben keinen linearen Zusammenhang. Dadurch ist die erste Bedingung für die lineare Korrelation erfüllt. Der Erwartungswert der Residuen verläuft bei 0, wodurch die zweite Voraussetzung erfüllt ist. Es sind 3 von 193 Werten erkennbar, die außerhalb liegen- diese Anzahl ist jedoch gerechtfertigt. Damit ist auch die 3. Voraussetzung (iid ist gleichmäßig verteilt) erfüllt. Es sind 3 ausreißende Daten zu erkennen (5, 60, 69)
2. Im Quantil-Quantil Plot werden die standardisierten Residuen gegen die Normalverteilung aufgetragen. Es liegt links ein leichter Rand vor und rechts ein schwerer Rand. Dies spricht gegen eine Normalverteilung der Residuen.
3. Der Scale Location Plot zeigt den Verlauf der Standardabweichung der Residuen über den dargestellten Bereich. der es ermöglicht zu analysieren, ob die Standardabweichung der Residuen über den Bereich der erklärten Variable gleich bleibt. Der Verlauf zeigt keine Trichterbildung. Dies deuetet auf eine relative Homogenität der Daten hin. Daher ist Homoskedastizität gegeben und somit die 4. Voraussetzung erfüllt.
4. Die Cook`s Distanz analysiert die Datenpunkte und gewichtet diese. Wenn sich Datenpunkte im Datensatz befinden, welche signifikant zu einem anderen Ergebnis führen würden, werden diese in Form einer Distanz angezeigt. Werte mit einer höheren Distanz als „1“ werden als besorgniserregend bezeichnet, da sie zu einem „aushebeln“ bzw. Kippen der Gerade führe könnte (Hebelwirkung). Keiner der Datenpunkte in dem gezeigten Plot weist eine Distanz von ±1 auf.

**Wie gut ist das Modell?**

> summary(lmUN)

Call:

lm(formula = log(UN$infantMortality) ~ log(UN$ppgdp))

Residuals:

Min 1Q Median 3Q Max

-1.16789 -0.36738 -0.02351 0.24544 2.43503

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.10377 0.21087 38.43 <2e-16 \*\*\*

log(UN$ppgdp) -0.61680 0.02465 -25.02 <2e-16 \*\*\*

---

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

Residual standard error: 0.5281 on 191 degrees of freedom

(20 observations deleted due to missingness)

Multiple R-squared: 0.7662, Adjusted R-squared: 0.765

F-statistic: 625.9 on 1 and 191 DF, p-value: < 2.2e-16

Die Fragestellung der H0 bezieht sich darauf, ob ein linearer Zusammenhang besteht.

H0: es besteht kein linearer Zusammenhang

HA: es ist ein linearer Zusammenhang gegeben

Mit den Residuen werden die Abstände der einzelnen Punkte senkrecht zu der Geraden angegeben. Der Median liegt bei -0.02351. Dieser Wert liegt nahe an 0, wodurch die Schwankung sehr gering ist. Auch der Erwartungswert der Residuen liegt nahe an 0. Der Median wird mit Q1 und Q3 in Relation gesetzt und sollte nahe bei 0 liegen (Voraussetzung: IE[IE]=0).

Die Gerade schneidet bei 8.10377 die y-Achse (Intercept Alpha). Die Standardabweichung liegt bei 0.211. Dies ist ein geringer Wert, wodurch die Streuung der Daten als gering angenommen werden kann. Der P Wert von 2e^-16 liegt unter der 5% Signifikanzniveau und ist damit signifikant (\*\*\*). H0 kann somit verworfen werden. Der geschätzte Koeffizient von -0.61680 für die ppgdp bezieht sich auf den Beta Wert. Der Standardfehler von log(ppgdp) beträgt 0.002465. Der P-Wert für log(ppgdp) liegt bei 2e^-16 und ist signifikant(\*\*\*).

Der Residual standard error (0.02465) gibt die Gesamtfläche unter allen Fehlerquadraten an. Die Anzahl der Beobachtungen wird durch die Anzahl der Freiheitsgrade +2 (für jede Zeile +1, intercept + ppgdp) angegeben. Daher wurden insgesamt 193 Beobachtungen berücksichtigt. Multiple R-Squared (Bestimmtheitsmaß) gibt den Anteil der Gesamtstreuung an, welcher durch die Regression berücksichtig wird- je näher sich dieser Wert „1“ nähert, desto besser wird das Modell erklärt. Adjusted R squared bezieht zusätzlich noch die Summe der Freiheitsgrade mit ein. Im Fall von diesem Modell hat Multiple R-squared einen Wert von 0.7662 und Adjusted R-Squared einen Wert von 0.765. Das heißt, zwischen log(ppgdp) und log(infantMortality) ist ein mittlerer bis starker positiver linearer Zusammenhang.

Mit F-statistic (hier: 625,9) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 191 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. Der p- Wert der F- Statistik ist kleiner als 2.2e^(-16), das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Regressionsgerade: log(InfantMortality)= 8,10-0,6168\*log(ppgdp).

2.

Die Ergebnisse aus der HÜ2 Bsp1 haben veranschaulicht, dass Education die Korrelation mit Fertility und Agriculture am niedrigsten ist. Das hat zu einer Schlussfolgerung geführt und zwar je höher die Ausbildung der Menschen ist, umso niedriger ist die Fertilitätsrate und umso weniger sind sie Landwirtschaft tätig. Aus diesem Grund werden Fertility-Education und Agriculture-Education Variablen an zwei unterschiedlichen Modellen untersucht.

Modell 1: Education-Fertility

Hypothesentestung:

H0: es gibt keinen linearen Zusammenhang.

HA: es gibt einen linearen Zusammenhang.

**Wie gut ist das Modell?**

> lmswiss<-lm(Education~Fertility, data = swiss)

> summary(lmswiss)

Call:

lm(formula = Education ~ Fertility, data = swiss)

Residuals:

Min 1Q Median 3Q Max

-11.4895 -5.3618 -0.9056 4.0360 24.0653

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.81788 6.11244 7.659 1.08e-09 \*\*\*

Fertility -0.51095 0.08582 -5.954 3.66e-07 \*\*\*

---

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

Residual standard error: 7.271 on 45 degrees of freedom

Multiple R-squared: 0.4406, Adjusted R-squared: 0.4282

F-statistic: 35.45 on 1 and 45 DF, p-value: 3.659e-07

Residuen sind Abweichungen der Punkte von der Gerade. Das heißt die senkrechte Entfernung der schwarzen Punkte (Daten) von der Gerade. Median ist -0.9056 und liegt aber in Relation zu Min bis Max sehr nahe bei null. Der Erwartungswert der Residuen liegt nahe bei null. Median muss im Vergleich zu 1Q. und 3. Q nahe bei null liegen (Voraussetzung: IE [Ei] = 0).

Intercept (alpha) ist der Wert, wo die Gerade die y-Achse schneidet (hier: 46,817). Standradfehler/Standardabweichung liegt bei 6,112. Grundsätzlich gilt: je größer die Standardabweichung, desto größer die Gefahr, dass die null mit drinnen liegt. P-Wert von 1.08e-09 liegt unterhalb von 5% des Signifikantsniveau. H0 kann verworfen werden, der Schätzer für das Interzept die geforderte Signifikanz erreichtGeschätzter Koefficient Fertility (-0.51095) ist bezogen auf ß-Wert. Standardfehler von Fertility ist 0,0858 (sehr niedrig). Sprich das ist die Wahrscheinlichkeit, dass die null mit drinnen ist, ist 3.66e-07 (signifikant).

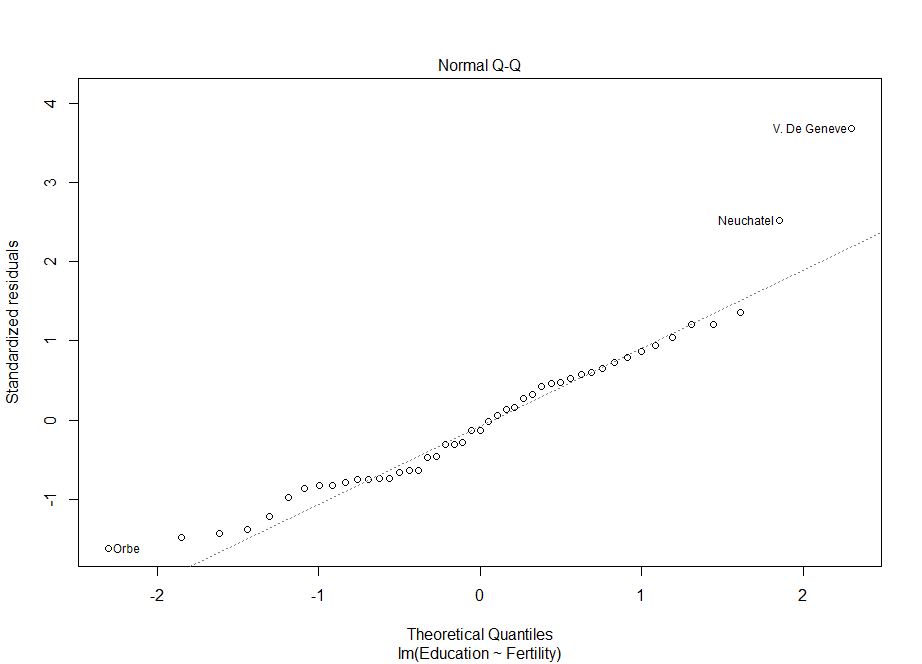
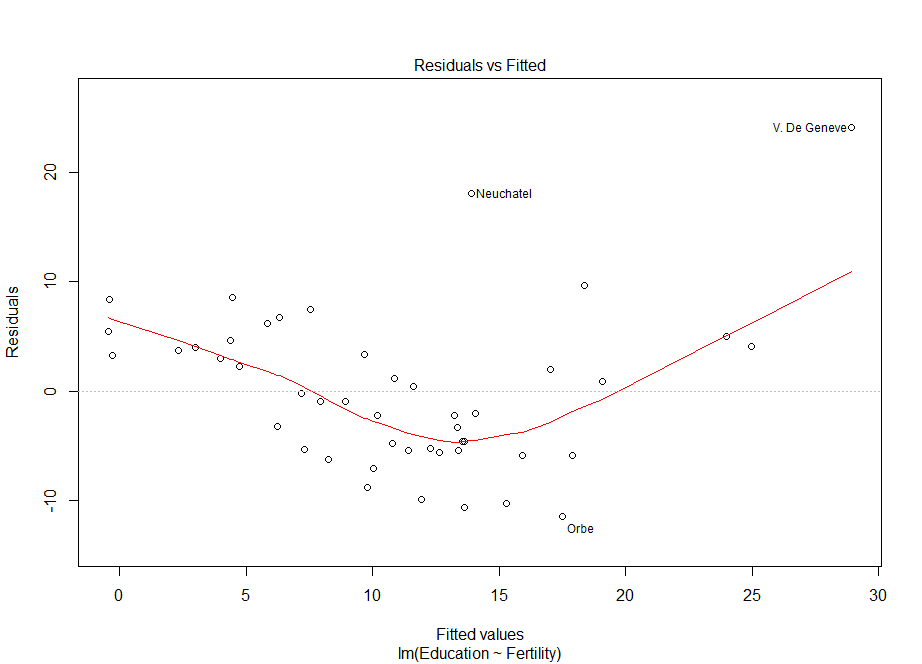
Residual standard error (7.271) ist die Fläche der Fehlerquadrate Residuen-Standardfehler. 45 degrees of freedom ist die Anzahl der Beobachtungen. 45+2=47 ist die Gesamtgröße der Stichprobe. Multiple R-squared= Bestimmtheitsmaß gibt an, welcher den Anteil der Gesamtstreuung durch die Regression erklärt wird. Je höher der Anteil, desto besser das Modell (1=bestmöglich). In diesen Fall ist R-squared 0.4406 und der angepasste (adjusted) R-squared ist 0.4282. Adjusted R squared ist angepasster Wert durch Einbeziehen der Freiheitsgrade der Summe der Quadrate. Das heißt, zwischen Education-Fertility ist ein schwach positiver linearer Zusammenhang.

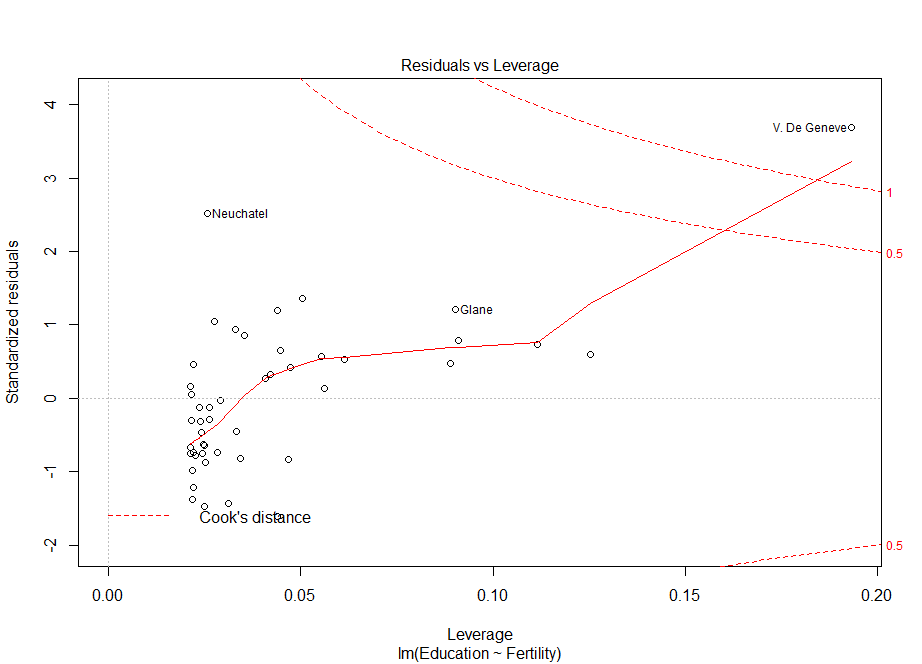
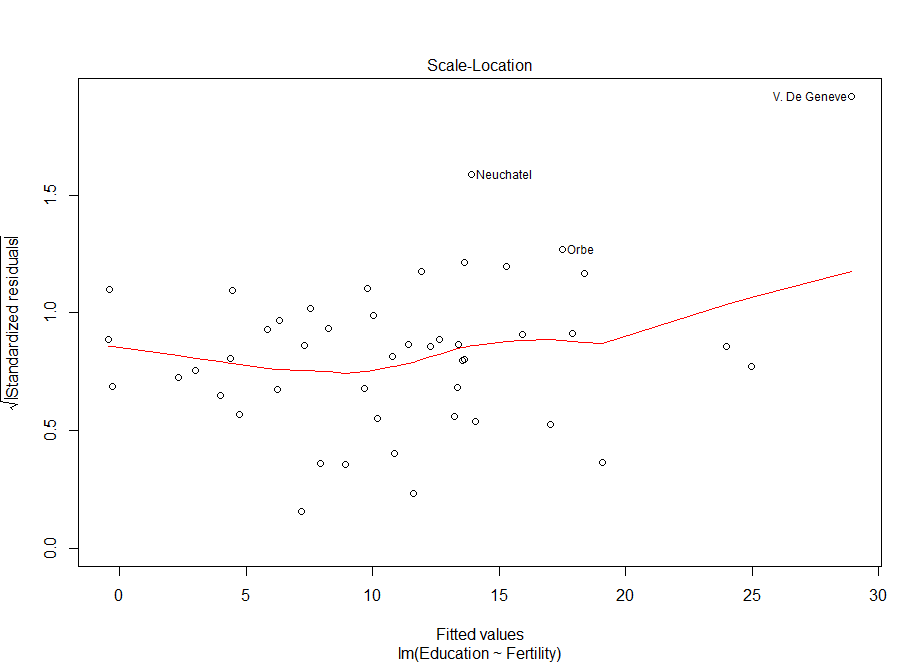
Mit F-statistic (hier: 35,45) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 45 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. F-Statistik von 35,45 ist kleiner als 3,659e-7, das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Trotz der Ungenauigkeit des Models gibt die F- Statistik mit 35.45 an, dass die Varianz der Daten mit dem Model um das ca. 35-fache besser erklärt wird als wenn man versuchen würden, die Varianz der Daten ausschließlich mit Mittelwerten zu erklären.

Regressionsgerade: Education= 46,82-0,51\*Fertility.

**Ist das Modell gültig?**





Residuals vs Fitted-Plot zeigt, dass die meisten Daten/Fehler im Bereich von Null bis -10 schwanken. Die andere Gruppe der Daten/Fehler schwankt zwischen Null und +10. Nur wenige Werte schwanken um Null. Städte wie V.De Geneva, Orbe und Neuchatel sind als Ausreißer definiert. Zusammenfassend kann man sagen, dass die Daten nicht gleichmäßig verteilt sind, aber es keine periodische und cluster Bildungen. Die 1.Voraussetzung für die lineare Regression ist erfüllt, da xi, xj unabhängig, unkorreliert sind und haben keinen linearen Zusammenhang. Die 2.Voraussetzung ist nicht erfüllt, da IE [Ei] ≠ Null ist und somit ein systematischer Fehler vorhanden ist. Die 3.Voraussetzung ist nicht erfüllt da, Eiid zwar unabhängig aber nicht gleichmäßig (3 von 47 Werten draußen) verteilt. Die 4.Voraussetzung ist erfüllt, da Var (Ei) = σ2 weil die Varianzen gleich und somit homoskedastisch sind.

Der QQ-Plot zeigt, dass die meisten Residauls auf der Geradelinie liegen. Die Daten von den Städten V.De Geneva, Orbe und Neuchatel liegen nicht auf der Geradelinie bzw. sie verursachen schwere Ränder. Da sie als Ausreißer angenommen werden und wird hier Normalverteilung angenommen. Die Voraussetzung 5 ist somit erfüllt. Scale-Location Plot weist daraufhin, dass es die meisten standardisierte Residuals-Daten im Bereich zwischen 0.5 und 1.0 liegen. All diese Indizien weisen auf homoskedastizität hin. Cooks Distance Plot zeigt, dass die Gerade gehebelt werden kann.

Modell 2: Education-Agriculture

Hypothesentestung:

H0: es gibt keinen linearen Zusammenhang.

HA: es gibt einen linearen Zusammenhang.

**Wie gut ist das Modell?**

> lmswiss<-lm(Education~Agriculture, data = swiss)

> summary(lmswiss)

Call:

lm(formula = Education ~ Agriculture, data = swiss)

Residuals:

Min 1Q Median 3Q Max

-12.632 -4.270 0.543 2.280 28.630

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.69527 2.68890 9.184 6.98e-12 \*\*\*

Agriculture -0.27076 0.04852 -5.580 1.30e-06 \*\*\*

---

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

Residual standard error: 7.474 on 45 degrees of freedom

Multiple R-squared: 0.409, Adjusted R-squared: 0.3959

F-statistic: 31.14 on 1 and 45 DF, p-value: 1.305e-06

Residuen sind Abweichungen der Punkte von der Gerade. Das heißt die senkrechte Entfernung der schwarzen Punkte (Daten) von der Gerade. Median ist 0.543 und liegt aber in Relation zu Min bis Max sehr nahe bei null. Der Erwartungswert der Residuen liegt nahe bei null. Median muss im Vergleich zu 1Q. und 3. Q nahe bei null liegen (Voraussetzung: IE [Ei] = 0).

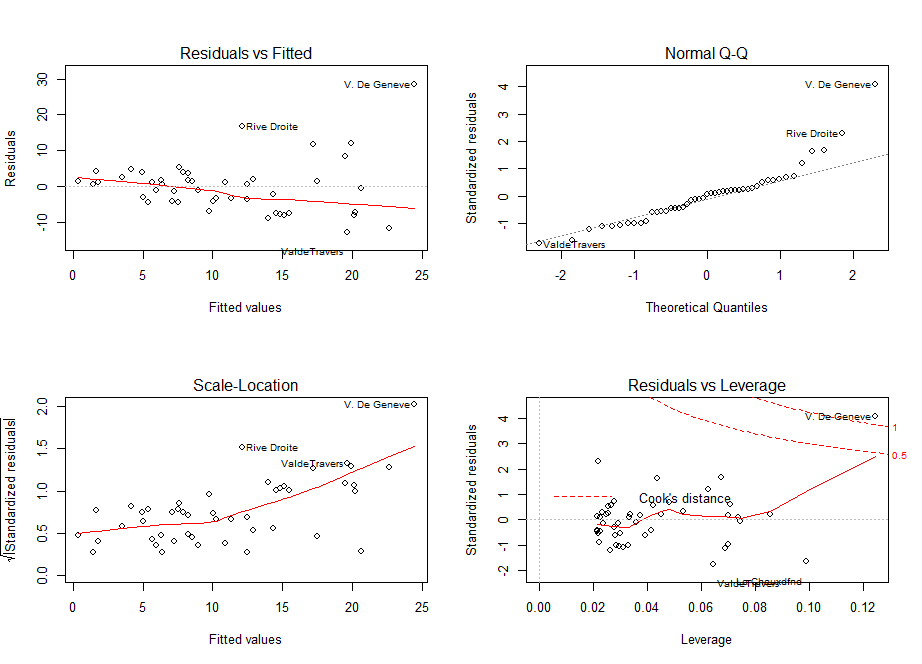
Intercept (alpha) ist der Wert, wo die Gerade die y-Achse schneidet (hier: 24,695). Standradfehler/Standardabweichung liegt bei 2,689. Grundsätzlich gilt: je größer die Standardabweichung, desto größer die Gefahr, dass die null mit drinnen liegt. P-Wert von 6,98e-12 liegt unterhalb von 5% des Signifikantsniveau. H0 kann verworfen werden, da Achsenabschnitt hoch signifikant ist. Geschätzter Koefficient Agriculture (-0.27076) ist bezogen auf ß-Wert. Standardfehler von Agriculture ist 0,0485 (sehr niedrig). Sprich das ist die Wahrscheinlichkeit, dass die null mit drinnen ist, ist 1,30e-06 (hoch signifikant).

Residaul standard error (7.474) ist die Fläche der Fehlerquadrate Residuen-Standardfehler. 45 degrees of freedome ist die Anzahl der Beobachtungen. 45+2=47 ist die Gesamtgröße der Stichprobe. Multiple R-squared= Bestimmtheitsmaß gibt an, welcher den Anteil der Gesamtstreuung durch die Regression erklärt wird. Je höher der Anteil, desto besser das Modell (1=bestmöglich). In diesen Fall ist R-squared 0.409 und der angepasste (adjusted) R-squared ist 0.3959. Adjusted R squared ist angepasster Wert durch Einbeziehen der Freiheitsgrade der Summe der Quadrate. Das heißt, zwischen Education-Agriculture ist ein schwach positiver linearer Zusammenhang.

Mit F-statistic (hier: 31,14) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 45 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. F-Statistik von 31,14 ist kleiner als 1,305e-06, das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Regressionsgerade: Education= 24,70-0,27\*Agriculture.

**Ist das Modell gültig?**



Residuals vs Fitted-Plot zeigt, dass die meisten Daten/Fehler im Bereich von Null bis -10 schwanken. Die andere Gruppe der Daten/Fehler schwankt zwischen Null und +10. Nur wenige Werte schwanken um Null. Städte wie V.De Geneva, Rive Drolle und Valde Travers sind als Ausreißer definiert. Zusammenfassend kann man sagen, dass die Daten nicht gleichmäßig verteilt sind, aber es keine periodische und cluster Bildungen. Die 1.Voraussetzung für die lineare Regression ist erfüllt, da xi, xj unabhängig, unkorreliert sind und haben keinen linearen Zusammenhang. Die 2.Voraussetzung ist nicht erfüllt, da IE [Ei] ≠ Null ist und somit ein systematischer Fehler vorhanden ist. Die 3.Voraussetzung ist nicht erfüllt da, Eiid zwar unabhängig aber nicht gleichmäßig verteilt. Die 4.Voraussetzung ist nicht erfüllt, da Var (Ei) ≠ σ2 weil die Varianzen unterschiedlich und somit heteroskedastisch sind.

Der QQ-Plot zeigt schwere Ränder und daher keine Normalverteilung. Die Voraussetzung 5 ist nicht erfüllt. Scale-Location Plot weist daraufhin, dass es die meisten standardisierte Residuals-Daten im Bereich zwischen 0.5 und 1.0 liegen. All diese Indizien weisen auf heteroskedastizität hin. Cooks Distance Plot zeigt, dass die Gerade gehebelt wird

3.

Die Ergebnisse aus der HÜ2 Bsp2 haben veranschaulicht, Eine höhere Korrelation hingegen ist gegeben zwischen Population und Murder, Income und Life Exp, Income und HS Grad (starke Korrelation), Illiteracy und Murder (starke Korrelation) und zwischen Life Exp. und HS Grad. Daraus lässt sich rückschließen, dass ein höheres Einkommen die durchschnittliche Lebenserwartung steigert, ein höheres Einkommen mit dem Bildungsgrad zusammenhängt und die Mordrate mit Analphabetismus in Verbidnung steht. Aus diesem Grund werden Murder-Illiteracy und Murder-Population Variablen an zwei unterschiedlichen Modellen untersucht.

Modell 1: Vergleich Murder mit Illiteracy.

H0: es gibt einen linearen Zusammenhang

HA: Kein linearer Zusammenhang gegeben

Call:

lm(formula = Murder ~ Population, data = state)

Residuals:

Min 1Q Median 3Q Max

-4.9855 -3.0119 -0.3128 2.4986 7.9014

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.1713934 0.6869410 8.984 7.49e-12 \*\*\*

Population 0.0002841 0.0001121 2.535 0.0146 \*

---

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

Residual standard error: 3.503 on 48 degrees of freedom

Multiple R-squared: 0.1181, Adjusted R-squared: 0.09972

F-statistic: 6.427 on 1 and 48 DF, p-value: 0.01455

Mit den Residuen werden die Abstände der einzelnen Punkte senkrecht zu der Geraden angegeben. Der Median liegt bei -0.3128. Dieser Wert liegt nahe an 0, wodurch die Schwankung eher gering ist. Auch der Erwartungswert der Residuen liegt nahe an 0. Der Median wird mit Q1 und Q3 in Relation gesetzt und sollte nahe bei 0 liegen (Voraussetzung: IE[IE]=0).

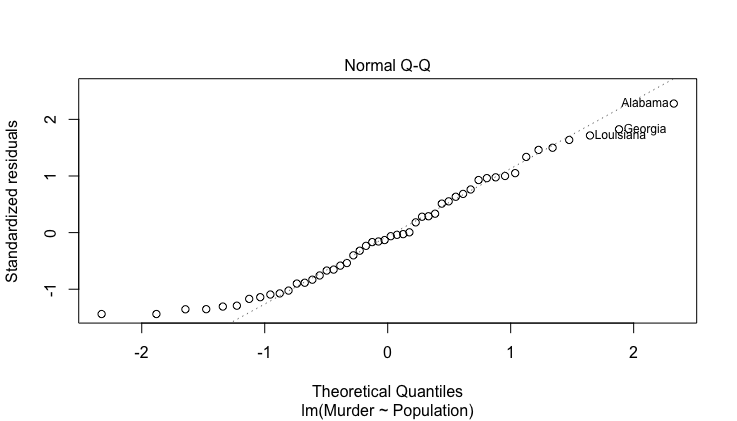
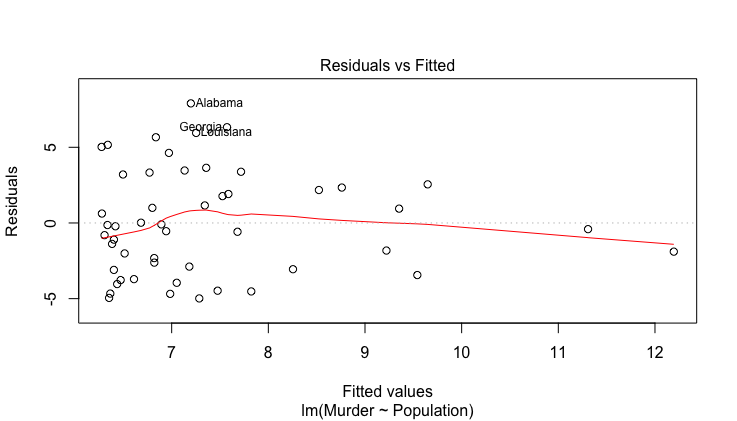
Die Gerade schneidet bei 6.1713934 die y-Achse (Intercept Alpha). Die Standardabweichung liegt bei 0.687. Dies ist ein geringer Wert, wodurch die Streuung der Daten als gering angenommen werden kann. Der P Wert von 7.49e^-12 liegt unter der 5% Signifikanzniveau und ist damit hoch signifikant (\*\*\*). H0 kann somit verworfen werden. Der geschätzte Koeffizient von 0.0002841 für die Population bezieht sich auf den Beta Wert. Der Standardfehler von Population beträgt 0.00011. Der P-Wert für Population liegt bei 0.0146 und ist schwach signifikant.

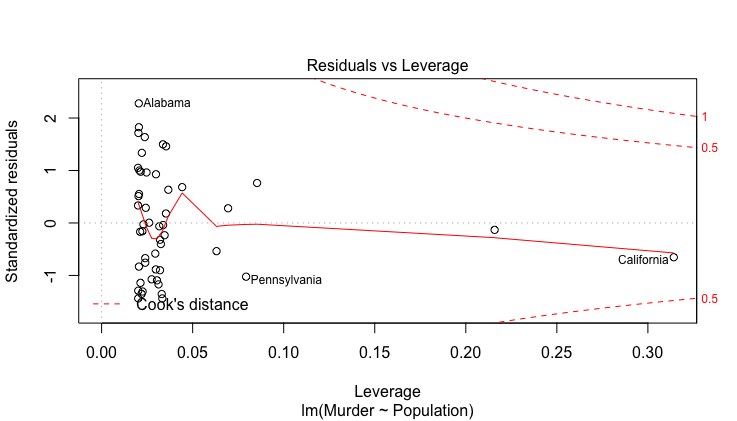
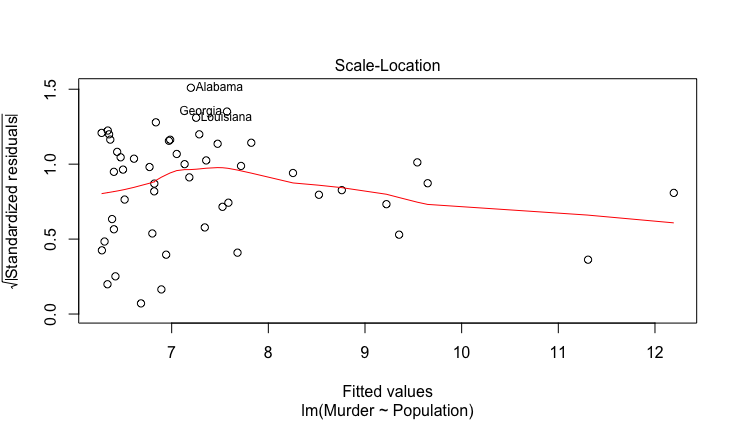
Der Residual standard error (3.503) gibt die Gesamtfläche unter allen Fehlerquadraten an. Die Anzahl der Beobachtungen wird durch die Anzahl der Freiheitsgrade +2 (für jede Zeile +1, intercept + Population) angegeben. Daher wurden insgesamt 50 Beobachtungen berücksichtigt. Multiple R-Squared (Bestimmtheitsmaß) gibt den Anteil der Gesamtstreuung an, welcher durch die Regression berücksichtig wird- je näher sich dieser Wert „1“ nähert, desto besser wird das Modell erklärt. Adjusted R squared bezieht zusätzlich noch die Summe der Freiheitsgrade mit ein. Im Fall von diesem Modell hat Multiple R-squared einen Wert von 0.1181 und Adjusted R-Squared einen Wert von 0.09972. Damit ist die Streuung zwischen Murder und Population sehr hoch ist.

Mit F-statistic (hier 6,427) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 48 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. F-Statistik von 6,427 ist kleiner als 0,01455, das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Regressionsgerade: Murder= 6,17 +0.0002841\*Population.

**Ist das Modell gültig?**





Bei dem Plot Residuals vs. Fitted werden die Residuen gegen die vom Model vorhergesagten Werte aufgetragen. Die Schwankungsbreite der Daten erstreckt sich von -5 bis +5. Die meisten Daten schwanken nicht um 0. Die Linie weist Schwankungen im positiven und negativen Skalenbereich auf, was auf eine Inhomogenität der Daten schließen lässt. Sowohl xi wie auch xj sind unabhängig und unkorelliert und haben keinen linearen Zusammenhang. Dadurch ist die erste Bedingung für die lineare Korrelation erfüllt. Der Erwartungswert der Residuen verläuft bei 0, wodurch die zweite Voraussetzung erfüllt ist. Es sind 3 von 50 Werten erkennbar, die außerhalb liegen- diese Anzahl ist jedoch noch gerechtfertigt. Damit ist auch die 3. Voraussetzung (iid ist gleichmäßig verteilt) erfüllt. Es sind 3 ausreißende Daten zu erkennen (Alabama, Georgia und Louisiana)

Im Quantil-Quantil Plot werden die standardisierten Residuen gegen die Normalverteilung aufgetragen. Die meisten Punkte liegen auf der Gerade, dennoch gibt es sowohl links wie auch rechts leichte Ränder. Dies spricht für eine systematische Annährung einer Normalverteilung der Daten.

Der Scale Location Plot zeigt den Verlauf der Standardabweichung der Residuen über den dargestellten Bereich. Der es ermöglicht zu analysieren, ob die Standardabweichung der Residuen über den Bereich der erklärten Variable gleich bleibt. Der Verlauf zeigt eine Trichterbildung. Dies deuetet auf eine Inhomogenität der Daten hin. Daher ist keine Homoskedastizität gegeben und somit die 4. Voraussetzung nicht erfüllt.

Die Cook`s Distanz analysiert die Datenpunkte und gewichtet diese. Wenn sich Datenpunkte im Datensatz befinden, welche signifikant zu einem anderen Ergebnis führen würden, werden diese in Form einer Distanz angezeigt. Werte mit einer höheren Distanz als „1“ werden als besorgniserregend bezeichnet, da sie zu einem „aushebeln“ bzw. Kippen der Gerade führe könnte (Hebelwirkung). Keiner der Datenpunkte in dem gezeigten Plot weist eine Distanz von ±1 auf.

Murder- Illiteracy

Call:

lm(formula = Murder ~ Illiteracy, data = state)

Residuals:

Min 1Q Median 3Q Max

-5.5315 -2.0602 -0.2503 1.6916 6.9745

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3968 0.8184 2.928 0.0052 \*\*

Illiteracy 4.2575 0.6217 6.848 1.26e-08 \*\*\*

---

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

Residual standard error: 2.653 on 48 degrees of freedom

Multiple R-squared: 0.4942, Adjusted R-squared: 0.4836

F-statistic: 46.89 on 1 and 48 DF, p-value: 1.258e-08

Mit den Residuen werden die Abstände der einzelnen Punkte senkrecht zu der Geraden angegeben. Der Median liegt bei -0.2503. Dieser Wert liegt- in Relation gesetzt mit Min. (-5.5315) und Max. Wert 6.9745)- nahe an 0, wodurch die Schwankung eher gering ist. Der Erwartungswert der Residuen liegt jedoch nicht nahe an 0. Der Median wird mit Q1 und Q3 in Relation gesetzt und sollte nahe bei 0 liegen (Voraussetzung: IE[IE]=0).

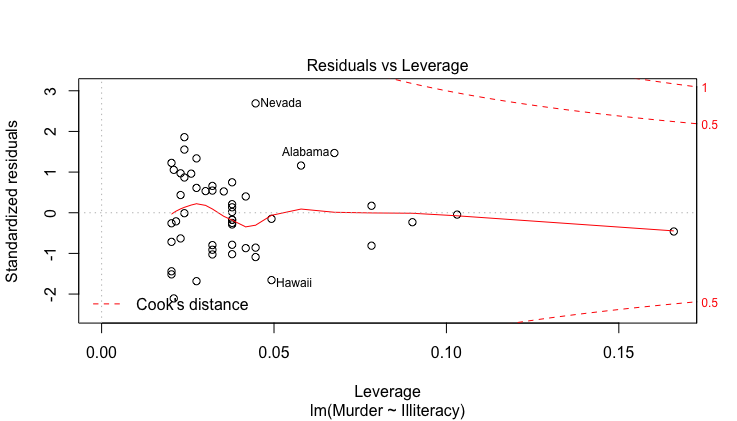
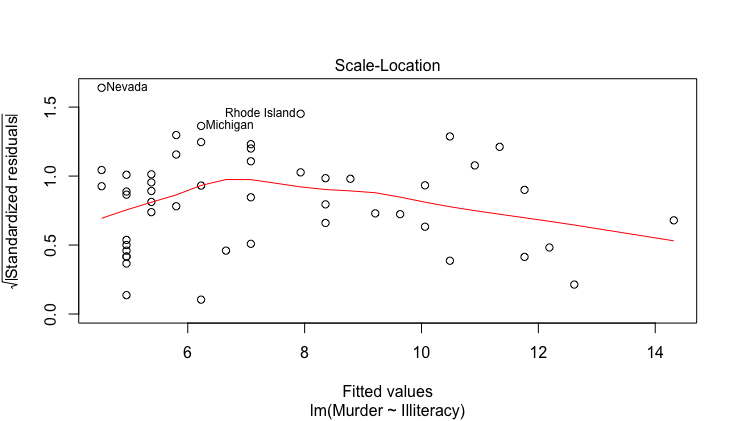
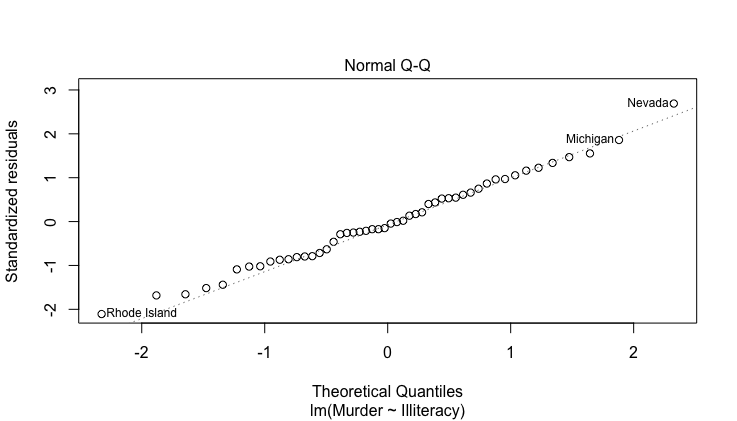
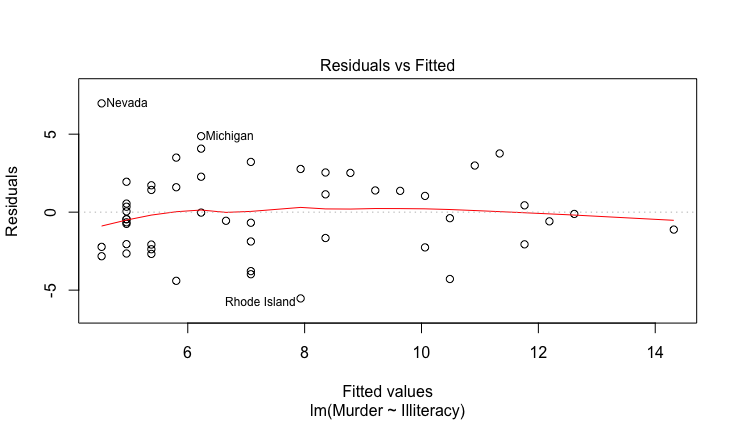
Die Gerade schneidet bei 2.3968 die y-Achse (Intercept Alpha). Die Standardabweichung liegt bei 0.818. Dies ist ein geringer Wert, wodurch die Streuung der Daten als gering angenommen werden kann. Der P Wert von 0.0052 liegt unter dem 1% Signifikanzniveau und ist damit signifikant (\*\*). H0 kann somit verworfen werden. Der geschätzte Koeffizient von 4.2575 für Illiteracy bezieht sich auf den Beta Wert. Der Standardfehler von Illiteracy beträgt 0.6217. Der hoch signifikante P-Wert von 1.26e^-0.8 weist auf einen deutlich linearen Zusammenhang der Daten hin. H0 kann somit verworfen werden.

Der Residual standard error (2.653) gibt die Gesamtfläche unter allen Fehlerquadraten an. Die Anzahl der Beobachtungen wird durch die Anzahl der Freiheitsgrade +2 (für jede Zeile +1, intercept + Illiteracy) angegeben. Daher wurden insgesamt 50 Beobachtungen berücksichtigt. Multiple R-Squared (Bestimmtheitsmaß) gibt den Anteil der Gesamtstreuung an, welcher durch die Regression berücksichtig wird- je näher sich dieser Wert „1“ nähert, desto besser wird das Modell erklärt. Adjusted R squared bezieht zusätzlich noch die Summe der Freiheitsgrade mit ein. Im Fall von diesem Modell hat Multiple R-squared einen Wert von 0.4942 und Adjusted R-Squared einen Wert von 0.4836. Das heißt, zwischen Murder und Illiteracy ist ein schwach positiver linearer Zusammenhang.

Mit F-statistic (hier: 46,89) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 48 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. F-Statistik ist kleiner als 1,258e-08, das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Regressionsgerade: Murder= 2,3968+4.2575\*Illiteracy.

**Ist das Modell gültig?**



1. Bei dem Plot Residuals vs. Fitted werden die Residuen gegen die vom Model vorhergesagten Werte aufgetragen. Die Linie weist keine große Sprünge auf, was auf eine Homogenität der Daten schließen lässt. Sowohl xi wie auch xj sind unabhängig und unkorreliert und haben keinen linearen Zusammenhang. Dadurch ist die erste Bedingung für die lineare Korrelation erfüllt. Der Erwartungswert der Residuen verläuft bei 0, wodurch die zweite Voraussetzung erfüllt ist. Die Werte sind annähernd gleichmäßig um die Mitte verteilt. Es sind 3 von 50 Werten erkennbar, die außerhalb liegen- diese Anzahl ist jedoch nicht gerechtfertigt. Damit ist auch die 3. Voraussetzung (iid ist gleichmäßig verteilt) erfüllt.
2. Im Quantil-Quantil Plot werden die standardisierten Residuen gegen die Normalverteilung aufgetragen. Die meisten Punkte liegen auf der Gerade, weshalb Normalverteilungshypothese nicht verletzt sein dürfte.
3. Der Scale Location Plot zeigt den Verlauf der Standardabweichung der Residuen über den dargestellten Bereich. , der es ermöglicht zu analysieren, ob die Standardabweichung der Residuen über den Bereich der erklärten Variable gleich bleibt. Der Verlauf der zeigt keine Anzeichen dafür, dass die Varianzen nicht einheitlich verlaufen würden. , Daher ist Homoskedastizität gegeben und somit auch die 4. Voraussetzung erfüllt. Die Staaten Rhode Island, Michigan und Nevada sind im QQPlot als Ausreißer markiert, alle andere Staaten liegen jedoch auf der Gerade, weßhalb von einer Normalverteilung ausgegangen werden kann. Diese erfüllt Voraussetzung 5.
4. Die Cook`s Distanz analysiert die Datenpunkte und gewichtet diese. Wenn sich Datenpunkte im Datensatz befinden, welche signifikant zu einem anderen Ergebnis führen würden, werden diese in Form einer Distanz angezeigt. Werte mit einer höheren Distanz als „1“ werden als besorgniserregend bezeichnet, da sie zu einem „aushebeln“ bzw. Kippen der Gerade führe könnte (Hebelwirkung). Keiner der Datenpunkte in dem gezeigten Plot weist eine Distanz von ±1 auf.

4.

Hypothesentestung:

H0: es gibt keinen linearen Zusammenhang.

HA: es gibt einen linearen Zusammenhang.

**Wie gut ist das Modell?**

> model<- lm(lakehurondata$x~lakehuron$year, data = LakeHuron)

> summary(model)

Call:

lm(formula = lakehurondata$x ~ lakehuron$year, data = LakeHuron)

Residuals:

Min 1Q Median 3Q Max

-2.50997 -0.72726 0.00083 0.74402 2.53565

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 625.554918 7.764293 80.568 < 2e-16 \*\*\*

lakehuron$year -0.024201 0.004036 -5.996 3.55e-08 \*\*\*

---

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

Residual standard error: 1.13 on 96 degrees of freedom

Multiple R-squared: 0.2725, Adjusted R-squared: 0.2649

F-statistic: 35.95 on 1 and 96 DF, p-value: 3.545e-08

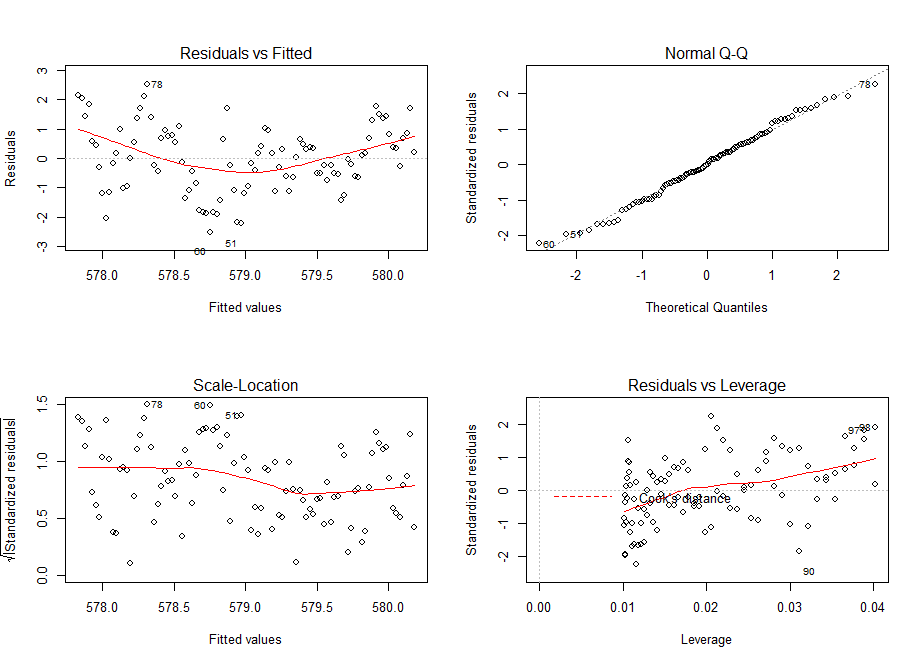
Residuen sind Abweichungen der Punkte von der Gerade. Das heißt die senkrechte Entfernung der schwarzen Punkte (Daten) von der Gerade. Median ist 0,00083 und liegt aber in Relation zu Min bis Max sehr nahe bei null. Der Erwartungswert der Residuen liegt nahe bei null. Median muss im Vergleich zu 1Q. und 3. Q nahe bei null liegen (Voraussetzung: IE [Ei] = 0).

Intercept (alpha) ist der Wert, wo die Gerade die y-Achse schneidet (hier: 625,555). Standradfehler/Standardabweichung liegt bei 7,7643. Grundsätzlich gilt: je größer die Standardabweichung, desto größer die Gefahr, dass die null mit drinnen liegt. P-Wert von 2e-16 liegt unterhalb von 5% des Signifikantsniveau. H0 kann verworfen werden, da Achsenabschnitt hoch signifikant ist. Geschätzter Koefficient year (-0.024201) ist bezogen auf ß-Wert. Standardfehler von Fertility ist 7,7643 (sehr niedrig). Sprich das ist die Wahrscheinlichkeit, dass die null mit drinnen ist, ist 3.55e-08 (hoch signifikant).

Residaul standard error (1,13) ist die Fläche der Fehlerquadrate Residuen-Standardfehler. 96 degrees of freedome ist die Anzahl der Beobachtungen. 96+2=98 ist die Gesamtgröße der Stichprobe. Multiple R-squared= Bestimmtheitsmaß gibt an, welcher den Anteil der Gesamtstreuung durch die Regression erklärt wird. Je höher der Anteil, desto besser das Modell (1=bestmöglich). In diesen Fall ist R-squared 0,2725und der angepasste (adjusted) R-squared ist 0,2649. Adjusted R squared ist angepasster Wert durch Einbeziehen der Freiheitsgrade der Summe der Quadrate. Das heißt, zwischen Wasserstand und Jahr ist ein schwach positiver linearer Zusammenhang.

Mit F-statistic (hier: 35,45) wird goodness of fit, also wie gut die Daten passen, angeschaut. Es wird getestet ob man ein lineares Modell braucht oder ob es reicht mit Mittelwert zu testen. Hier wird das Modell mit 96 Freiheitsgraden mit einem Modell mit einem einzigen Freiheitsgrad, sprich ein einziger MW wird hineingelegt. F-Statistik von 35,45 ist kleiner als 3,545e-08, das heißt lineares Modell ist besser als wenn man nur einen Gesamtmittelwert reinlegt.

Regressionsgerade: Wasserstand= 625,55-0,21\*Zeitmesspunkt.

**Ist das Modell gültig?** 

1. Bei dem Plot Residuals vs. Fitted werden die Residuen gegen die vom Model vorhergesagten Werte aufgetragen. Die Linie weist keine große Sprünge auf, was auf eine Homogenität der Daten schließen lässt. Sowohl xi wie auch xj sind unabhängig und unkorreliert und haben keinen linearen Zusammenhang. Dadurch ist die erste Bedingung für die lineare Korrelation erfüllt. Der Erwartungswert der Residuen verläuft bei 0, wodurch die zweite Voraussetzung erfüllt ist. Die Werte sind annähernd gleichmäßig um die Mitte verteilt. Damit ist auch die 3. Voraussetzung (iid ist gleichmäßig verteilt) erfüllt.
2. Im Quantil-Quantil Plot werden die standardisierten Residuen gegen die Normalverteilung aufgetragen. Die meisten Punkte liegen auf der Gerade, weshalb Normalverteilungshypothese nicht verletzt sein dürfte.
3. Der Scale Location Plot zeigt den Verlauf der Standardabweichung der Residuen über den dargestellten Bereich, der es ermöglicht zu analysieren, ob die Standardabweichung der Residuen über den Bereich der erklärten Variable gleich bleibt. Der Verlauf der zeigt keine Anzeichen dafür, dass die Varianzen nicht einheitlich verlaufen würden. Daher ist Homoskedastizität gegeben und somit auch die 4. Voraussetzung erfüllt. Die Staaten Rhode Island, Michigan und Nevada sind im QQ-Plot als Ausreißer markiert, alle andere Staaten liegen jedoch auf der Gerade, weshalb von einer Normalverteilung ausgegangen werden kann. Diese erfüllt Voraussetzung 5.
4. Die Cook`s Distanz analysiert die Datenpunkte und gewichtet diese. Wenn sich Datenpunkte im Datensatz befinden, welche signifikant zu einem anderen Ergebnis führen würden, werden diese in Form einer Distanz angezeigt. Werte mit einer höheren Distanz als „1“ werden als besorgniserregend bezeichnet, da sie zu einem „aushebeln“ bzw. Kippen der Gerade führe könnte (Hebelwirkung). Keiner der Datenpunkte in dem gezeigten Plot weist eine Distanz von ±1 auf.