**HÜ4**

1.

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).

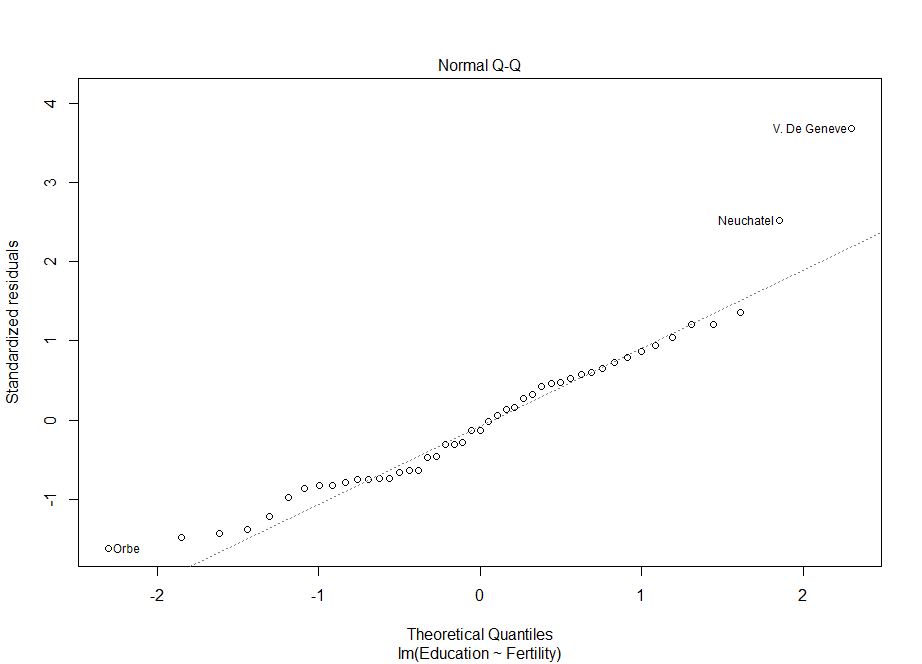
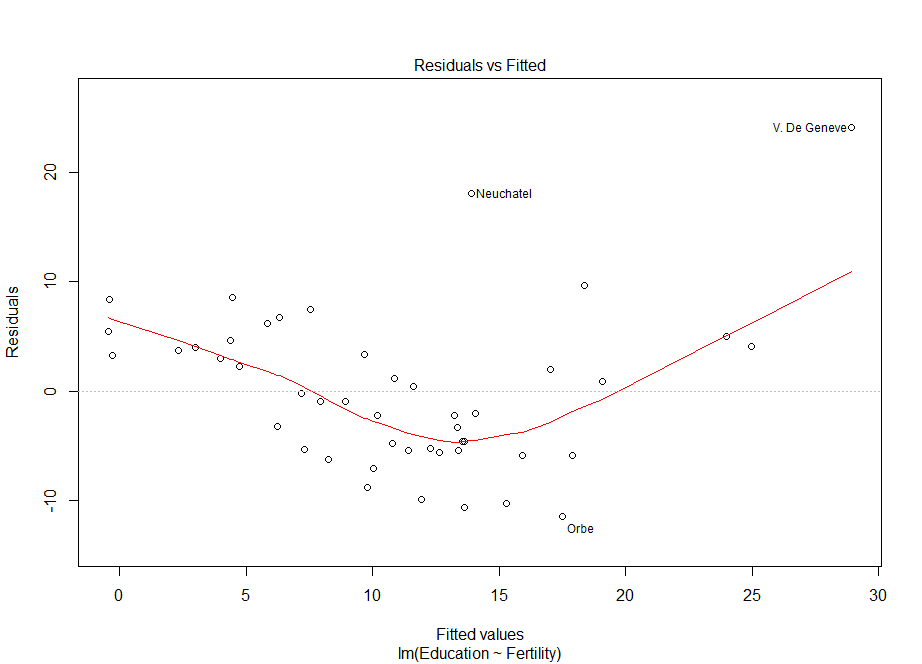
Intercept (alpha) ist der Wert, wo die Gerade die y-Achse schneidet (hier: 46,817). Standradfehler/Standardabweichung liegt bei 6,112. Grundsätzlich gilt: je größer die Standardabweichung, desto größer die Gefahr, dass die null mit drinnen liegt. P-Wert von 1.08e-09 liegt unterhalb von 5% des Signifikantsniveau. H0 kann verworfen werden, da Achsenabschnitt hoch signifikant ist. Geschätzter 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 (hoch signifikant).

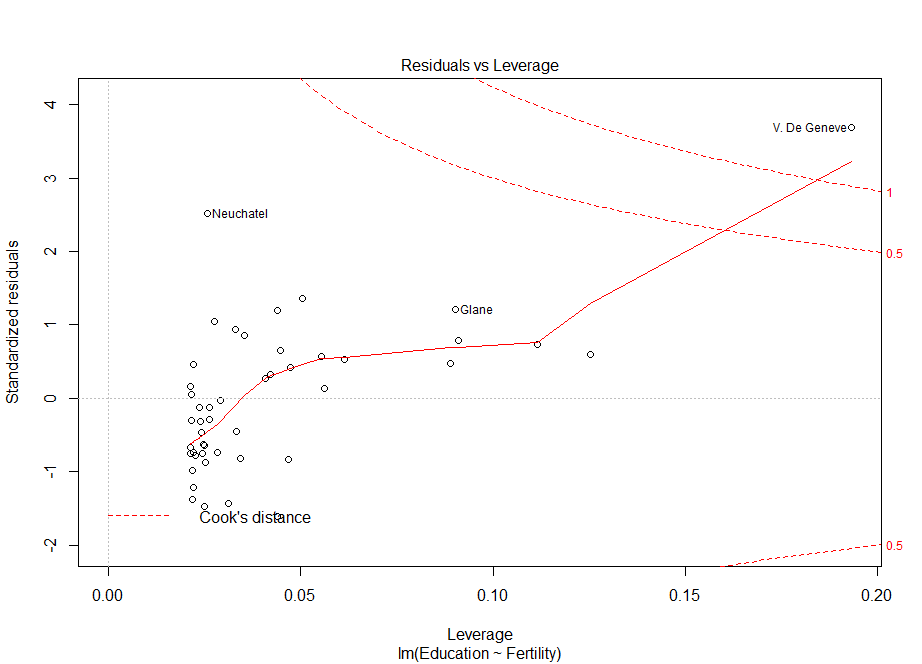
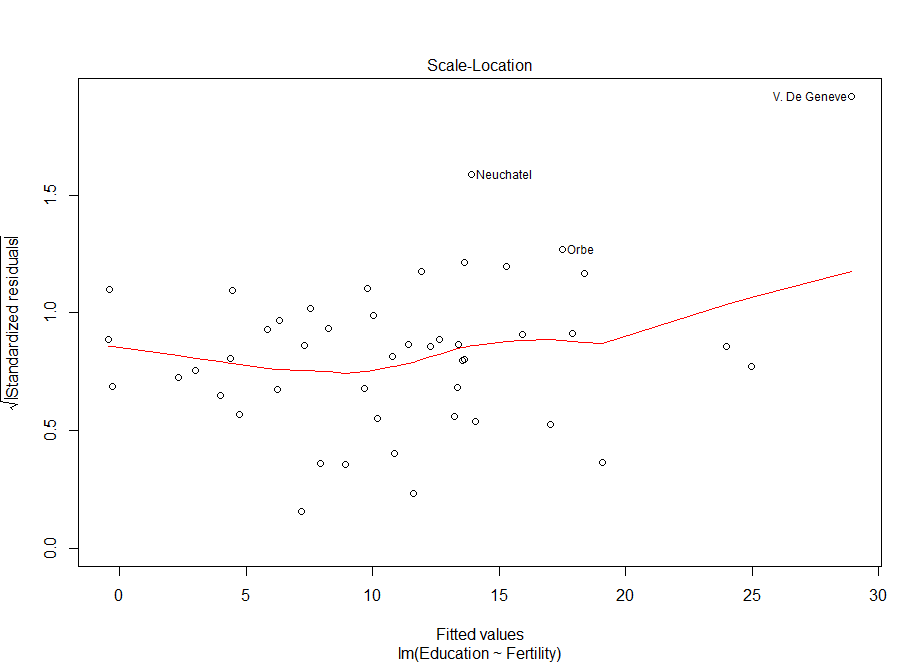
Residaul standard error (7.271) ist die Fläche der Fehlerquadrate Residuen-Standardfehler. 45 degrees of freedome ist die Anzahl der Beobachtungen. 45+2=47 ist die Gesamtgröße der Stichprobe. Multiple R-squared= Bestimmtheitsmaß gibt an, welcher den Anteil der Gesamtstreuung durch die Regression erklärt wird. Je höher der Anteil, desto besser das Modell (1=bestmöglich). In diesen Fall ist R-squared 0.4406 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, dass die Streuung zwischen Education-Fertility sehr hoch ist.

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.

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 nicht gehebelt werden kann obwohl es drei Ausreißer gibt.

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).

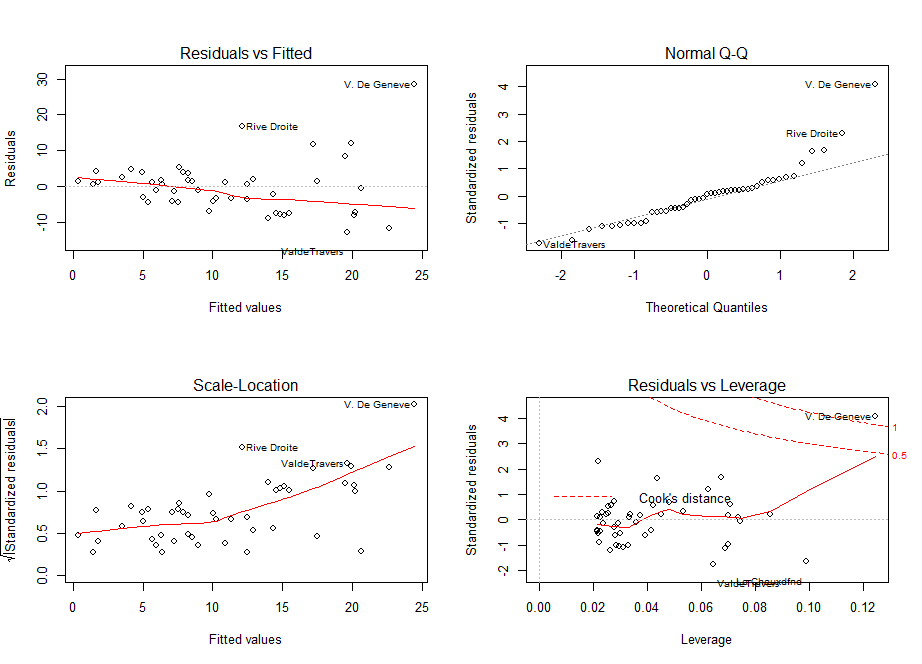
Intercept (alpha) ist der Wert, wo die Gerade die y-Achse schneidet (hier: 24,695). Standradfehler/Standardabweichung liegt bei 2,689. Grundsätzlich gilt: je größer die Standardabweichung, desto größer die Gefahr, dass die null mit drinnen liegt. P-Wert von 6,98e-12 liegt unterhalb von 5% des Signifikantsniveau. 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, dass die Streuung zwischen Education-Fertility sehr hoch ist.

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 nicht gehebelt werden.

3.

4.