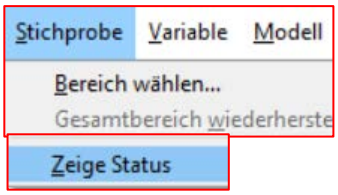
Übung 6: Rauchen und Schwangerschaft
Musterlösungen

1. Analyse der Daten:



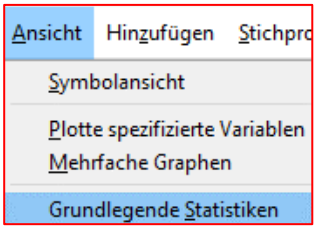
i. Wie viele Frauen sind in der Stichprobe enthalten?

1388 Beobachtungen

iii. Wie hoch ist der durchschnittliche Zigarettenkonsum pro Tag? Ist dieser Durchschnittswert repräsentativ für die typische Frau aus der Stichprobe?

	arith. Mittel	Median	Minimum	Maximum
faminc	29,027	27,500	0,50000	65,000
motheduc	12,936	12,000	2,0000	18,000
cigs	2,0872	0,00000	0,00000	50,000

	Std. Abw.	Var'koeff.	Schiefe	Überwölbung
faminc	18,739	0,64559	0,61762	-0,52660
motheduc	2,3767	0,18373	-0,032120	0,64824
cigs	5,9727	2,8616	3,5604	14,934



gretl: Ansicht / Grundlegende Statistiken/ Variablen wählen

Der durchschnittliche Zigarettenkonsum beträgt 2.09 und beinhaltet auch die 1176 Nicht-Raucherinnen in der Stichprobe → In diesem Fall kann man sagen, dass die typische Frau während der Schwangerschaft nicht raucht.

iv. Wie viele Frauen Rauchen während der Schwangerschaft? Was ist der Anteil von Raucherinnen in der Stichprobe?

gretl Hauptfenster: Stichprobe/Restringiere durch Bedingung/ Boolsche Bedingung: cigs > 0

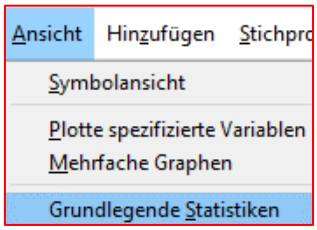
→ Teilmenge der Raucherinnen → **212** (=1388 -1176) Frauen haben einen positiven Zigarettenkonsum (cigs > 0) was 15% der Frauen in der Stichprobe entspricht.

v. Wie hoch ist der durchschnittliche Zigarettenkonsum pro Tag unter den Raucherinnen?

Der durchschnittliche Zigarettenkonsum unter den Raucherinnen beträgt 13.7, was wesentlich höher als der Durchschnitt über die gesamte Stichprobe ist.

	arith. Mittel	Median	Minimum	Maximum
faminc	20,917	18,500	0,50000	65,000
motheduc	11,637	12,000	6,0000	18,000
cigs	13,665	10,000	1,0000	50,000

	Std. Abw.	Var'koeff.	Schiefe	Überwölbung
faminc	15,142	0,72392	1,0458	0,95217
motheduc	1,7753	0,15256	0,15604	1,6180
cigs	8,6909	0,63599	1,3020	2,5502



vi. Wie hoch ist der durchschnittliche Familieneinkommen? Vergleichen Sie zwischen der Stichprobe und Teilmenge der Raucherinnen

Das durchschnittliche Familieneinkommen beträgt $29‘027 über die gesamte Stichprobe. Interessanterweise beträgt es unter den Raucherinnen nur $20‘917.

vii. Wie viele Neugeborene sind in der Stichprobe weiss?

299 Beobachtungen wurden entfernt → 1089 = (1388 -299) Neugeborene sind weiss, was 78.45% der Stichprobe darstellt.

2. Welchen Einfluss erwarten Sie für die Variablen cigs und faminc (Familieneinkommen) auf das Geburtsgewicht des Neugeborenen (Vorzeichen für β_2 und β_3)? Begründen Sie Ihre Antwort.

   $\beta_{cigs} < 0 \rightarrow$ je mehr geraucht wird, desto geringer das Geburtsgewicht des Neugeborenen

   $\beta_{faminc} > 0 \rightarrow$ reichere Familien werden sich mehr Gesundheitsvorsorge für das Ungeborene leisten können und dadurch das Geburtsgewicht positiv beeinflusst

3. Schätzen Sie das Modell 1: *bwght* = $\beta_1 + \beta_2$ *cigs* + u

```
             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       119,772        0,572341      209,3       0,0000   ***
  cigs         -0,513772     0,0904909      -5,678     1,66e-08 ***

Mittel d. abh. Var.      118,6996   Stdabw. d. abh. Var.    20,35396
Summe d. quad. Res.      561551,3   Stdfehler d. Regress.   20,12858
R-Quadrat                0,022729   Korrigiertes R-Quadrat  0,022024
```

4. Welche Korrelation erwarten Sie zwischen den Variablen *cigs* und *faminc* (Familieneinkommen)? Erklären Sie, warum die Korrelation positiv oder negativ sein könnte.

   Das Vorzeichen der Korrelation zwischen *cigs* und *faminc* ist a priori nicht eindeutig:

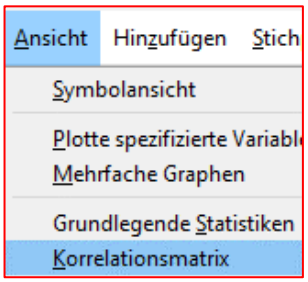
   Positive Korrelation: Reichere haben mehr Geld für Luxusgüter wie Zigaretten übrig

   Negative Korrelation: Reichere haben ein höheres Gesundheitsbewusstsein (weswegen reichere Mütter in der Schwangerschaft eher nicht oder weniger rauchen)

5. Analysieren Sie die Korrelationsstruktur zwischen den Variablen *bwght*, *cigs* und *faminc.*

   gretl Hauptfenster: Ansicht/Korrelationsmatrix →Variablen bwght, cigs und faminc auswählen

```
 bwght        faminc         cigs
1,0000        0,1089       -0,1508   bwght
              1,0000       -0,1730   faminc
                            1,0000   cigs
```



   Die Korrelation zwischen *cigs* und *faminc* ist negativ → Reichere Frauen haben im Durchschnitt ein höheres Gesundheitsbewusstsein und rauchen weniger.

6. Ermitteln Sie die Korrelation zwischen *cigs* und *faminc* (Familieneinkommen) mittels Regression. Einmal für die gesamte Stichprobe, einmal für die Gruppe der Raucherinnen. Wie ändert sich diese Korrelation für diese Teilmenge aus der Stichprobe?

   Regression für die gesamte Stichprobe

```
Abhängige Variable: faminc

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       30,1598        0,524988      57,45       0,0000   ***
  cigs        -0,542928      0,0830042     -6,541      8,58e-011 ***

Mittel d. abh. Var.      29,02666   Stdabw. d. abh. Var.    18,73928
Summe d. quad. Res.      472475,2   Stdfehler d. Regress.   18,46324
R-Quadrat                0,029945   Korrigiertes R-Quadrat  0,029245
```

Korrelation wird aufgrund der Korrelationsmatrix aus Frage 5 negativ geschätzt:

$\rho = -\sqrt{0.03} \cong -0.173 \rightarrow$ entspricht der Korrelation aus Korrelationsmatrix (-0.173 gretl)

Regression für die Gruppe der Raucherinnen
gretl: Stichprobe/Restringiere durch Bedingung/cigs > 0

```
Abhängige Variable: faminc

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       23,8077       1,93141        12,33      1,26e-026 ***
  cigs        -0,211509     0,119344       -1,772     0,0778    *

Mittel d. abh. Var.     20,91745   Stdabw. d. abh. Var.    15,14250
Summe d. quad. Res.     47668,34   Stdfehler d. Regress.   15,06626
R-Quadrat               0,014736   Korrigiertes R-Quadrat  0,010045
```

Korrelation wird negativ geschätzt: $\rho = -\sqrt{0.0147} \cong -0.121$

Korrelation beträgt -0.121 (gretl)

→gretl: Ansicht/Korrelationsmatrix, da sich der Stichprobenbereich auf die Raucherinnen reduziert hat.

```
 bwght        faminc          cigs
1,0000        0,1361       -0,1006   bwght
              1,0000       -0,1214   faminc
                            1,0000   cigs
```

Die Korrelation hat sich nur leicht reduziert, wenn nur die Teilprobe der Raucherinnen analysiert wurde.

Hinweis: Die Berechnung der Korrelation über "Wurzel aus R-Quadrat" funktioniert nur bei einer Einfachregression (nur eine erklärende Variable, hier cigs).

7. Welchen Effekt hat vermutlich die Hinzunahme von *faminc* auf den geschätzten Regressionskoeffizienten b_{cigs} ?

Hinweis: Benutzen Sie Ihr Ergebnis aus Frage 6

ρ_{12}: Korrelation zwischen Variablen x_1 und x_2

$\tilde{b}_i$: unterspezifiziertes Modell: Modell 1ohne *faminc*

	$\rho_{12} > 0$	$\rho_{12} < 0$
$b_2 > 0$	$E(\tilde{b}_2) > b_2$	$E(\tilde{b}_2) < b_2$
$b_2 < 0$	$E(\tilde{b}_2) < b_2$	$E(\tilde{b}_2) > b_2$

Unterspezifiziertes Modell: $bwght = \tilde{b}_1 + \tilde{b}_2 cigs$

„Korrekteres“ Modell: $bwght = b_1 + b_2 cigs + b_3 faminc$

ρ_{12} = Korrelation(*cigs*, *faminc*) < 0 und $b_{cigs} < 0$ (Modell 1)

⇒ Im unterspezifizierten Modell ist damit zu rechnen, dass $\tilde{b}_2$ zu klein geschätzt wird. Bei Berücksichtigung von *faminc* wird sich $\tilde{b}_2$ voraussichtlich vergrössern → $b_2 > \tilde{b}_2$.

Standardfehler: Kein grosser Effekt, da die Variablen (mit ρ_{12} = -0.173) nur schwach korrelieren.

8. Schätzen Sie das Modell 2: *bwght* = $\beta_1 + \beta_2$ *cigs* + β_3 *faminc* + *u*

```
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      116,974        1,04898       111,5       0,0000   ***
  cigs        -0,463408     0,0915768      -5,060     4,75e-07 ***
  faminc       0,0927647    0,0291879       3,178     0,0015   ***

Mittel d. abh. Var.      118,6996   Stdabw. d. abh. Var.     20,35396
Summe d. quad. Res.      557485,5   Stdfehler d. Regress.    20,06282
R-Quadrat                0,029805   Korrigiertes R-Quadrat   0,028404
F(2, 1385)               21,27392   P-Wert(F)                7,94e-10
Log-Likelihood          -6130,414   Akaike-Kriterium         12266,83
Schwarz-Kriterium        12282,54   Hannan-Quinn-Kriterium   12272,70
```

Unterspezifiziertes Modell: $\widehat{bwght}$ = 119.772 **– 0.513***cigs*

„Korrekteres Modell“: $\widehat{bwght}$ = 116.97 **– 0.463** *cigs* + 0.0927*faminc*

$\tilde{b}_2$ = -0.513 **<** b_2 =-0.463 → $\tilde{b}_2$ hat sich vergrössert!

Die Vermutung hat sich bestätigt: b_{cigs} steigt von −0.51 auf −0.46 (das ist ein Anstieg!) Auch: Der Standardfehler hat sich leicht erhöht: 0.0904 → 0.0915.

Es soll nun die Dummy-Variable *male* als zusätzlicher Regressor hinzugefügt werden (Wert 1, wenn das Neugeborene männlich ist, 0 für weiblich).

9. Vermuten Sie, dass die Berücksichtigung dieser Dummy-Variable einen deutlichen Effekt auf b_{cigs} und b_{faminc} oder deren Standardfehler hat? Warum bzw. warum nicht? Überprüfen Sie Ihre Vermutung anschliessend.

Das Geschlecht des Kindes, legt die Natur normalerweise ‘rein zufällig’ bei der Zeugung fest. Es sollte weder mit dem Zigarettenkonsum der Mutter in der Schwangerschaft *cigs* noch mit dem Familieneinkommen *faminc* korrelieren noch mit sonstigen ökonomischen Grössen (dies könnte sich jedoch ändern bei ” künstlicher Befruchtung“, welche sich fast nur einkommensstarke Familien leisten können) ⇒ Dummy-Variable *male* ist keine ausgelassene Variable und sollte fast keinen Effekt auf b_{cigs}, b_{faminc} haben.

10. Schätzen Sie das Modell 3: bwght = $\beta_1 + \beta_2$cigs + $\beta_3$faminc + $\beta_4$male + u

```
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      115,228        1,20788       95,40       0,0000   ***
  cigs        -0,461046     0,0913378     -5,048      5,07e-07 ***
  faminc       0,0968798    0,0291453      3,324      0,0009   ***
  male         3,11397      1,07640        2,893      0,0039   ***

Mittel d. abh. Var.      118,6996   Stdabw. d. abh. Var.     20,35396
Summe d. quad. Res.      554134,6   Stdfehler d. Regress.    20,00965
R-Quadrat                0,035636   Korrigiertes R-Quadrat   0,033546
F(3, 1384)               17,04780   P-Wert(F)                7,10e-11
Log-Likelihood          -6126,230   Akaike-Kriterium         12260,46
Schwarz-Kriterium        12281,40   Hannan-Quinn-Kriterium   12268,29
```

$\widehat{bwght}$ = 115.228 – 0. 461*cigs* + 0.09687*faminc* + *3.114male*

	b_{cigs}	se()	b_{faminc}	se()
Modell 2	-0.4634	0.0915	0.0927	0.0913
Modell 3	-0.4610	0.0292	0.0968	0.0291

⇒ Das Hinzufügen von *male* hat fast keinen Effekt auf deren Koeffizienten und Standardfehler gehabt.

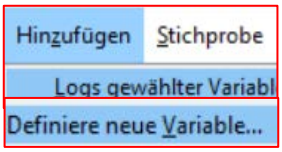
11. Interpretieren Sie b_{faminc} im Modell 3.

b_{faminc} = 0.096880 ≈ +0.1 Unzen

Bei zusätzlichem Familieneinkommen um $1000 ist eine Gewichtszunahme des Neugeborenen um 0.1 Unzen (= 0.1 x 28.35gr = 2.8 Gramm) zu erwarten, ceteris paribus (geschätzt auf Basis der Stichprobe).

12. Schätzen Sie das Modell 4 mit dem Geburtsgewicht des Neugeborenen in Gramm ausgedrückt.

Modell 4: $bwghtgr = \beta_1^* + \beta_2^* cigs + \beta_3^* fa\min c + \beta_4^* male + u$



Hinweis: 1 Unze = 28.35 Gramm → Variable bwghtgr = bwght x 28.35

gretl Hauptfenster: Hinzufügen / Definiere neue Variable/ bwghtgr = bwght x 28.35

```
Abhängige Variable: bwghtgr

            Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      3266,71        34,2434        95,40      0,0000   ***
  cigs        -13,0706       2,58943       -5,048     5,07e-07 ***
  faminc        2,74654      0,826268       3,324     0,0009   ***
  male         88,2810      30,5158         2,893     0,0039   ***

Mittel d. abh. Var.     3365,133   Stdabw. d. abh. Var.     577,0349
Summe d. quad. Res.     4,45e+08   Stdfehler d. Regress.    567,2737
R-Quadrat               0,035636   Korrigiertes R-Quadrat   0,033546
F(3, 1384)              17,04780   P-Wert(F)                7,10e-11
```

13. Wie ist die Beziehung zwischen den Koeffizienten aus Modell 3 und 4.

Beziehung: $b_i^* = b_i \text{ x } 28.35$

Beispiel: b^*_{cigs} *= -13.07 =* b_{cigs} *x 28.35 = -0.461 x 28.35*

14. Interpretieren Sie den Koeffizienten b_{faminc}

Bei 1000$ mehr Familieneinkommen ist eine Gewichtszunahme des Neugeborenen um ca. 2.74 Gramm (= 0.1 Unzen) zu erwarten (geschätzt auf Basis der Stichprobe, ceteris paribus)

15. Folgende Modelle wurden geschätzt. Interpretieren Sie jeweils den Koeffizienten b_3.

i. *bwght* = 112.138 − 0.465*cigs* + 1.927 ln(*faminc*) + 3.096 *male*

lin-log Spezifikation: Steigt das Familieneinkommen (*faminc*) um 1%, erhöht sich das Geburtsgewicht im Durchschnitt um 0.01 x 1.927 = 0.02 Unzen, ceteris paribus

Interpretation einer lin-log Spezifikation: Eine Zunahme von x um 1% führt c.p. zu einer Änderung von y um 0.01 x b_3 *Einheiten.*

ii. *ln(bwght = 4.703 − 0.00406cigs + 0.0169 ln(faminc) + 0.0258 male*

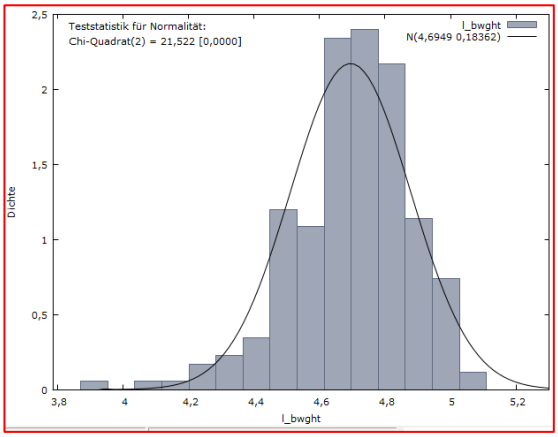
log-log Spezifikation: Steigt das Familieneinkommen (*faminc*) um 1%, erhöht sich das Geburtsgewicht (bwght) im Durchschnitt um 0.017%, ceteris paribus.

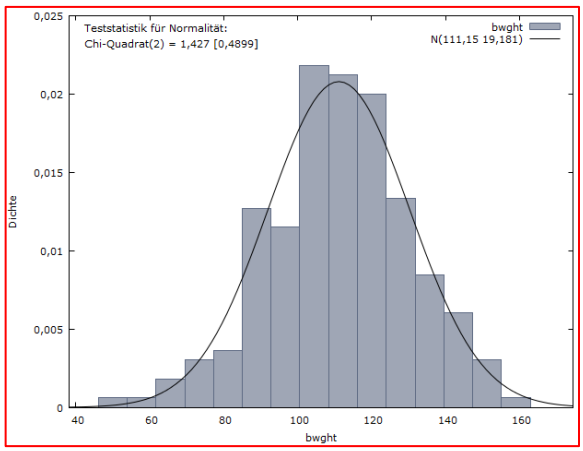
iii. *ln(bwght)* = 4.729 – 0.0401*cigs* + 0.000878 *faminc* + 0.0259 *male*

log-lin Spezifikation: Steigt das Familieneinkommen (*faminc)* um $1000 (=1 Einheit), erhöht sich das Geburtsgewicht (bwght) im Durchschnitt um 0.0009 × 100% = 0.09%, ceteris paribus.

*Interpretation einer log-lin Spezifikation: Eine Zunahme von **x** um 1 Einheit führt c.p. zu einer Änderung von **y** um 100% x b_3 Einheiten*

16. Erstellen Sie ein Histogramm von ln(bwght) und bwght. Welcher Unterschied ist zu vermerken?





Die Logarithmierung des Geburtsgewichtes reduziert die Normalität der Daten

17. Der Regressor *faminc* wurde durch *fatheduc* (Ausbildungsdauer des Vaters gemessen in Jahren) ersetzt. Interpretieren Sie jeweils den Koeffizienten b_3:

i. bwght = 113.260 – 0.571cigs + 0.411 fatheduc + 3.568 male lin-lin Modell

Steigt die Ausbildungsdauer des Vaters (*fatheduc)* um 1 Jahr, erhöht sich das Geburtsgewicht (*bwght)* im Durchschnitt um 0.4 Unzen (= 11.3 Gramm cp.)

ii. bwght = 106.528 – 0.574cigs + 4.772 ln(fatheduc) + 3.524 male lin-log Modell

Steigt die Ausbildungsdauer des Vaters (*fatheduc)* um 1%, erhöht sich *bwght* im Durchschnitt um ca 4.77/100 = 0.048 Unzen (=1.36 Gramm) cp.

iii. ln(bwght) = 4.664 – 0.005cigs + 0.0372 ln(fatheduc) + 0.0313 male log-log Modell

Steigt die Ausbildungsdauer des Vaters (*fatheduc)* um 1%, erhöht sich das Geburtsgewicht (*bwght)* im Durchschnitt um 0.037%, cp.

iv. ln(bwght) = 4.716 –0.0049cigs + 0.0033 fatheduc + 0.0317 male log-lin Modell

Steigt die Ausbildungsdauer des Vaters (*fatheduc)* um 1 Jahr, erhöht sich das Geburtsgewicht (*bwght)* im Durchschnitt um 0.003×100% = 0.3%, cp.

18. Schätzen Sie das Modell 5

Modell 5: bwght = $\beta_1 + \beta_2$cigs + $\beta_3$parity + $\beta_4$faminc + $\beta_5$motheduc + $\beta_6$fatheduc + u

Die Variable *parity* stellt die Reihenfolge des Neugeborenen unter den Familienkindern dar. Wenn *parity* = 3 bedeutet es, dass das erfasste Neugeborene das dritte Kind der Frau ist.

```
Fehlende oder unvollständige Beobachtungen entfernt: 197
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      114,524        3,72845       30,72      6,87e-153 ***
  cigs        -0,595936     0,110348      -5,401     8,02e-08  ***
  parity       1,78760      0,659406       2,711     0,0068    ***
  faminc       0,0560414    0,0365616      1,533     0,1256
  motheduc    -0,370450     0,319855      -1,158     0,2470
  fatheduc     0,472394     0,282643       1,671     0,0949    *

Mittel d. abh. Var.     119,5298   Stdabw. d. abh. Var.     20,14124
Summe d. quad. Res.     464041,1   Stdfehler d. Regress.    19,78878
R-Quadrat               0,038748   Korrigiertes R-Quadrat   0,034692
F(5, 1185)              9,553500   P-Wert(F)                5,99e-09
Log-Likelihood         -5242,220   Akaike-Kriterium         10496,44
Schwarz-Kriterium       10526,94   Hannan-Quinn-Kriterium   10507,93
```

i. Interpretieren Sie den Wert *parity* = 3.

Das bedeutet, dass das erfasste Neugeborene das dritte Kind der Frau ist.

ii. Warum reduziert gretl hier jeweils die Zahl der einbezogenen Familien? Könnte das Konsequenzen über die Repräsentativität der "selektierten" Familien haben?

In der Stichprobe gibt es Beobachtungen ohne Angaben über die Ausbildung des Vaters. Das ist 197/1388 = 14.2% der Stichprobe, was nicht gravierend ist.

Es ist hier unklar warum diese Angaben über Ausbildungsjahre des Vaters nicht vorhanden sind. Eine Möglichkeit wäre, dass die Identität des Vaters unbekannt ist!

Wenn wir davon ausgehen, dass Mütter mit tieferem Ausbildungsniveau eher davon betroffen werden, gibt es eine Verzerrung für die Repräsentativität der selektierten Familien.

iii. Spielt die Reihenfolge des Neugeborenen eine Rolle für das Geburtsgewicht? Interpretieren Sie den Koeffizienten b_{parity}.

Der Koeffizient von *parity* ist statistisch signifikant auf dem 1%-Signifikanzniveau. Die Reihenfolge des Neugeborenen spielt eine Rolle zur Bestimmung des Geburtsgewichtes.

Das Geburtsgewicht erhöht sich um 1.78 Unzen (= 1.78 x 28.35 = 50.5gr) pro zusätzliches Kind, ceteris paribus.

iv. Sind alle Steigungskoeffizienten gemeinsam signifikant (Modell 5)? Wie lautet die Nullhypothese?

Nullhypothese: H_0: $\beta_2 = \ldots = \beta_6 = 0$

Der p-Wert von F-Test ist gleich 0 $\Rightarrow$ die Nullhypothese wird abgelehnt.

Mindestens eine erklärende Variable ist von null verschieden. Die ausgewählten Regressoren erklären gemeinsam einen Teil der Varianz von *bwght*.

19. Testen Sie die Nullhypothese im Modell 5, dass die Elternausbildung keinen Effekt auf das Geburtsgewicht des Neugeborenen hat.

i. Mittels gretl Test

gretl: Tests / Variable weglassen → Schätze reduziertes Modell → interpretieren Sie den p-Wert.

H_0: $\beta_5 = \beta_6 = 0 \Leftrightarrow \beta_{motheduc} = \beta_{fatheduc} = 0$

```
Nullhypothese: Die Regressionskoeffizienten sind Null für die Variablen
  motheduc, fatheduc
Teststatistik: F(2, 1185) = 1,43727, p-Wert 0,23799
```

```
Modell 14: KQ, benutze die Beobachtungen 1-1191
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      115,470        1,65590        69,73      0,0000   ***
  cigs        -0,597852     0,108770       -5,496     4,74e-08 ***
  parity       1,83227      0,657540        2,787     0,0054   ***
  faminc       0,0670618    0,0323938       2,070     0,0386   **

Mittel d. abh. Var.      119,5298   Stdabw. d. abh. Var.     20,14124
Summe d. quad. Res.      465166,8   Stdfehler d. Regress.    19,79607
R-Quadrat                0,036416   Korrigiertes R-Quadrat   0,033981
F(3, 1187)               14,95330   P-Wert(F)                1,47e-09
Log-Likelihood          -5243,663   Akaike-Kriterium         10495,33
Schwarz-Kriterium        10515,66   Hannan-Quinn-Kriterium   10502,99
```

Schlussfolgerung: p-Wert = 0.23 > α = 5% H_0 kann nicht verworfen werden.

Beide Koeffizienten sind simultan gleich null → die Elternausbildung leisten gemeinsam kaum einen Erklärungsbeitrag für das Geburtsgewicht!

ii. Bestimmen Sie den kritischen Wert F_c mittels gretl. Was ist Ihre Schlussfolgerung?

gretl Hauptfenster: Werkzeuge / Statistische Tabellen / F / rechtsseitige Wahrscheinlichkeit = 0.05

Nenner-FG = 2 =# Restriktionen *Zähler-FG = N-K = 1191-6 =1185, K = 6*

Kritischer Wert $F_c(0.95, 2, 1185) = 3$

Das unrestringierte Modell wurde mit N =1388 -197 = 1'191 Daten geschätzt, da nicht alle Beobachtungen Informationen über *fatheduc* enthalten.

$F_e = 1.43 < F_c \Rightarrow H_0$ kann nicht verworfen werden → *motheduc* und *fatheduc* sind gemeinsam nicht von null verschieden! Sie leisten gemeinsam kaum einen Erklärungsbeitrag für das Geburtsgewicht!

p-Wert = 0.23 > α = 5% $\Rightarrow H_0$ nicht verwerfen

iii. Berechnen Sie den F-Wert mittels Bestimmtheitsmass R^2 durch eigene Schätzung des restringierten Modells.

Da die Ausbildungsangaben für die Mütter immer vorhanden sind, selektieren Sie im Hauptfenster die Variable *fatheduc* und dann das Menü auswählen: *Stichprobe / Entferne Beobachtungen mit Fehlwerten* (nicht dauerhaft).

R^2 = 0.0387 (Modell 5) und R_r^2 = 0.0364 (restringiertes Modell)

$$F = \frac{\left(R^2 - R_r^2\right)/L}{\left(1-R^2\right)/(N-K)} = \frac{(0.0387-0.0364)}{(1-0.0387)}\frac{1191-6}{2} = 1.43$$

```
Modell 17: KQ, benutze die Beobachtungen 1-1191
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      114,524        3,72845       30,72       6,87e-153 ***
  cigs        -0,595936     0,110348      -5,401      8,02e-08  ***
  parity       1,78760      0,659406       2,711      0,0068    ***
  faminc       0,0560414    0,0365616      1,533      0,1256
  motheduc    -0,370450     0,319855      -1,158      0,2470
  fatheduc     0,472394     0,282643       1,671      0,0949    *

Mittel d. abh. Var.      119,5298   Stdabw. d. abh. Var.    20,14124
Summe d. quad. Res.      464041,1   Stdfehler d. Regress.   19,78878
R-Quadrat                0,038748   Korrigiertes R-Quadrat  0,034692
```

Unter Berücksichtigung aller Beobachtungen ist das R^2 anders! Deshalb ist die Benutzung der Proxy-Variable wichtig.

```
Modell 15: KQ, benutze die Beobachtungen 1-1388
Abhängige Variable: bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      114,214        1,46930       77,73       0,0000    ***
  cigs        -0,477154     0,0915180     -5,214      2,13e-07  ***
  parity       1,61637      0,603955       2,676      0,0075    ***
  faminc       0,0979201    0,0291868      3,355      0,0008    ***

Mittel d. abh. Var.      118,6996   Stdabw. d. abh. Var.    20,35396
Summe d. quad. Res.      554615,2   Stdfehler d. Regress.   20,01833
R-Quadrat                0,034800   Korrigiertes R-Quadrat  0,032708
F(3, 1384)               16,63327   P-Wert(F)               1,28e-10
Log-Likelihood          -6126,832   Akaike-Kriterium        12261,66
Schwarz-Kriterium        12282,61   Hannan-Quinn-Kriterium  12269,50
```

$\widehat{bwght} = 114.214 - 0.477 cigs + 1.6163 parity + 0.0979 faminc$

20. Schätzen Sie das Modell 6:

$\ln(bwght) = \beta_1 + \beta_2 cigs + \beta_3 \ln(faminc) + \beta_4 parity + \beta_5 male + \beta_6 white + u$

```
Modell 14: KQ, benutze die Beobachtungen 1-1388
Abhängige Variable: l_bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       4,65771      0,0221653     210,1       0,0000    ***
  cigs       -0,00435015   0,000851842    -5,107     3,73e-07  ***
  l_faminc    0,00927740   0,00593081      1,564     0,1180
  parity      0,0159828    0,00563877      2,834     0,0047    ***
  male        0,0265458    0,0100295       2,647     0,0082    ***
  white       0,0547875    0,0130518       4,198     2,87e-05  ***

Mittel d. abh. Var.      4,760031   Stdabw. d. abh. Var.    0,190662
Summe d. quad. Res.      48,04116   Stdfehler d. Regress.   0,186446
R-Quadrat                0,047187   Korrigiertes R-Quadrat  0,043740
F(5, 1382)               13,68835   P-Wert(F)               4,58e-13
Log-Likelihood           364,8246   Akaike-Kriterium       -717,6492
Schwarz-Kriterium       -686,2355   Hannan-Quinn-Kriterium -705,9010
```

i. Was ist der Effekt auf das Geburtsgewicht, wenn die Mutter 10 Zigaretten pro Tag mehr raucht?

Δcigs = 10

Δl_bwght = -0.00435(10) = -0.0435 →ca. 4.4% weniger Geburtsgewicht

ii. Wie viel mehr Geburtsgewicht weist ein männliches Neugeborenes gegenüber einem Weiblichen auf, ceteris paribus? Ist der Koeffizient β_5 signifikant auf 5%-Niveau?

Ein männliches Neugeborenes wiegt ca. 2.6% mehr gegenüber der Referenzgruppe (=weibliches Neugeborenes), ceteris paribus.

Faustregel: t-Quotient > 2 → H_0 →verwerfen Koeffizient ist statistisch signifikant!

p-Wert = 0 →Diese Dummy-Variable ist statistisch signifikant

iii. Wie viel mehr Geburtsgewicht weist ein weisses Neugeborenes gegenüber der Referenzgruppe auf, ceteris paribus? Ist der Koeffizient β_6 signifikant auf 5%-Niveau?

Ein weisses Neugeborenes wiegt ca. 5.5% mehr gegenüber der Referenzgruppe (= nicht weisses Neugeborenes), ceteris paribus.

p-Wert = 0 →Diese Dummy-Variable ist statistisch signifikant

21. Schätzen Sie das Modell 7:

ln(bwght) = β_1 + $\beta_2$cigs + β_3 ln(faminc) + $\beta_4$parity + $\beta_5$male + $\beta_6$white + $\beta_7$motheduc + $\beta_8$fatheduc + u

```
Modell 13: KQ, benutze die Beobachtungen 1-1388 (n = 1191)
Fehlende oder unvollständige Beobachtungen entfernt: 197
Abhängige Variable: l_bwght

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       4,65267      0,0381545      121,9       0,0000   ***
  cigs       -0,00521438   0,00102675      -5,079     4,42e-07 ***
  l_faminc    0,0110315    0,00854044       1,292     0,1967
  parity      0,0172014    0,00613350       2,804     0,0051   ***
  male        0,0341430    0,0107022        3,190     0,0015   ***
  white       0,0453991    0,0150870        3,009     0,0027   ***
  motheduc   -0,00297633   0,00297307      -1,001     0,3170
  fatheduc    0,00327634   0,00260843       1,256     0,2093

Mittel d. abh. Var.      4,767536   Stdabw. d. abh. Var.     0,188013
Summe d. quad. Res.      39,99114   Stdfehler d. Regress.    0,183861
R-Quadrat                0,049303   Korrigiertes R-Quadrat   0,043678
F(7, 1183)               8,764331   P-Wert(F)                1,55e-10
Log-Likelihood           331,1061   Akaike-Kriterium        -646,2122
Schwarz-Kriterium       -605,5518   Hannan-Quinn-Kriterium  -630,8901
```

Gretl entfernt automatisch Einträge ohne Angaben für motheduc oder fatheduc

i. Was ist die Auswirkung eines zusätzlichen Ausbildungsjahres der Mutter auf das Geburtsgewicht?

Wenn die Mutter ein zusätzliches Ausbildungsjahr hat, wiegt das Neugeborene etwa 100(-0.00297) ≅ 0.3% weniger, ceteris paribus.

Hinweis: Diese Interpretation ist mit Vorsicht zu geniessen, da der Koeffizient von null nicht verschieden ist.

22. Schätzen Sie das Modell 8:

bwght = β_1 + $\beta_2$cigs + β_3 ln(faminc) + $\beta_4$parity + $\beta_5$male + $\beta_6$white + $\beta_7$motheduc + $\beta_8$fatheduc + u

```
Fehlende oder unvollständige Beobachtungen entfernt: 197
Abhängige Variable: bwght

            Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const      106,538        4,07630       26,14      3,14e-119 ***
  cigs        -0,597376     0,109695      -5,446     6,27e-08  ***
  l_faminc     1,22061      0,912434       1,338     0,1812
  parity       1,91752      0,655284       2,926     0,0035    ***
  male         3,82465      1,14339        3,345     0,0008    ***
  white        4,63746      1,61185        2,877     0,0041    ***
  motheduc    -0,336755     0,317634      -1,060     0,2893
  fatheduc     0,415149     0,278676       1,490     0,1366

Mittel d. abh. Var.      119,5298   Stdabw. d. abh. Var.     20,14124
Summe d. quad. Res.      456463,7   Stdfehler d. Regress.    19,64313
R-Quadrat                0,054445   Korrigiertes R-Quadrat   0,048850
F(7, 1183)               9,730940   P-Wert(F)                7,99e-12
Log-Likelihood          -5232,416   Akaike-Kriterium         10480,83
Schwarz-Kriterium        10521,49   Hannan-Quinn-Kriterium   10496,15
```

i. Wie viel mehr Geburtsgewicht weist ein männliches Neugeborenes gegenüber der Referenzgruppe auf, ceteris paribus? Ist der Koeffizient b_5 signifikant auf dem 5%-Niveau?

Ein männliches Neugeborenes wiegt ca. 3.82 Unzen (= ca 108.4 Gramm) mehr gegenüber der Referenzgruppe (=weibliches Neugeborenes), ceteris paribus → die anderen Variablen sind gleich!

23. Antworten Sie auf diese Fragen mittels einer Regression.

i. Wie viel wiegt ein weibliches Neugeborenes im Durchschnitt in Kg?

```
Abhängige Variable: bwght

            Koeffizient   Std.-fehler   t-Quotient   p-Wert
  ----------------------------------------------------------
  const       117,167       0,787514      148,8      0,0000 ***
  male          2,94235     1,09115         2,697    0,0071 ***

Mittel d. abh. Var.      118,6996   Stdabw. d. abh. Var.     20,35396
Summe d. quad. Res.      571612,8   Stdfehler d. Regress.    20,30810
R-Quadrat                0,005219   Korrigiertes R-Quadrat   0,004501
F(1, 1386)               7,271438   P-Wert(F)                0,007091
Log-Likelihood          -6147,782   Akaike-Kriterium         12299,56
Schwarz-Kriterium        12310,03   Hannan-Quinn-Kriterium   12303,48
```

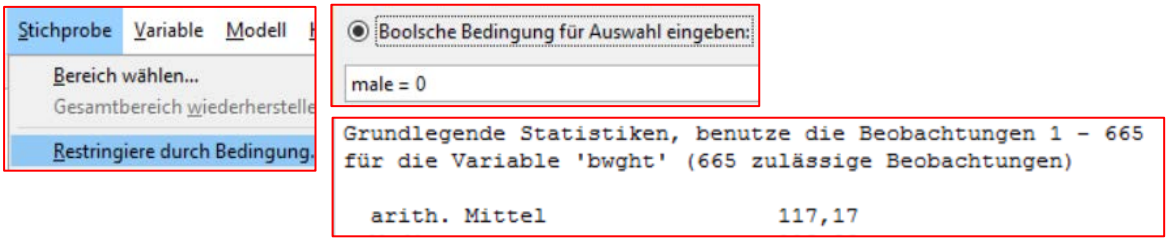
Ein weibliches Neugeborenes wiegt im Durchschnitt 117.16 Unzen (= ca 3.321 Kg) →Interzept = Geburtsgewicht der Referenzgruppe (=weibliches Neugeborenes).

ii. Wie viel mehr Geburtsgewicht in Gramm weist ein männliches Neugeborenes gegenüber einem Weiblichen auf?

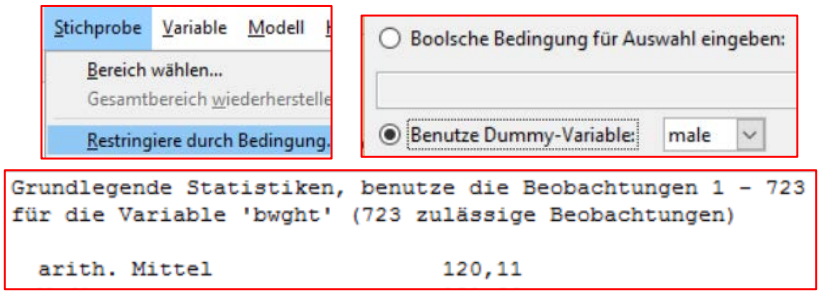
Ein männliches Neugeborenes wiegt ca. 2.94 Unzen (= ca. 83.35 Gramm) mehr gegenüber der Referenzgruppe (=weibliches Neugeborenes) was 120.109 (=117.16+ 2.942) Unzen entspricht→ ein männliches Neugeborenes wiegt im Durchschnitt 3.405 Kg (=120.109 x 28.35 gr).

iii. Bestätigen Sie Ihre Ergebnisse durch das Menü „ Grundlegende Statistiken“ für die entsprechenden Teilmengen.

Durchschnittliches Geburtsgewicht für die Teilmenge von weiblichen Neugeborenen. Die Stichprobe wurde durch die Bedingung male = 0 restringiert. Dieses Ergebnis untermauert die Antwort ii)



Durchschnittliches Geburtsgewicht für die Teilmenge von weiblichen Neugeborenen. Die Stichprobe wurde durch folgende Bedingung. Dieses Ergebnis untermauert die Antwort ii)



iv. Warum ist der Steigungskoeffizient kleiner als β_{male} im Modell 8?

Im Modell 8 werden die anderen erklärenden Variablen (Ausbildungsniveau der Eltern, parity) kontrolliert, was diesen Unterschied erklärt.

24. Welches Modell würden Sie vorziehen? Begründen Sie ihre Antwort.

Zusammenstellung der zu vergleichenden Modelle

Modell 2: $\widehat{bwght} = 116.97 - 0.463cigs + 0.0927faminc$

Modell 3: $\widehat{bwght} = 115.228 - 0.461cigs + 0.09687faminc + 3.114male$

Modell 5: $\widehat{bwght} = 114.524 - 0.596cigs + 1.787parity + 0.0560faminc - 0.37motheduc + 0.472fatheduc$

Modell 6: $\widehat{lnbwght} = 4.657 - 0.00435cigs + 0.00927lnfaminc + 0.0159parity + 0.0265male + 0.0547white$

Modell 7: $\widehat{lnbwght} = 4.657 - 0.00521cigs + 0.0172parity + 0.0117lnfaminc + 0.0341male + 0.045white - 0.0029motheduc + 0.00327fatheduc$

Modell 8: $\widehat{bwght} = 106.53 - 0.5973cigs + 1.917parity + 1.22lnfaminc + 3.82male + 4.63white - 0.336motheduc + 0.415fatheduc$

Modell	2	3	5	6	7	8

Abh. Variable	bwght	bwght	bwght	lnbwght	lnbwght	bwght
#Regressoren	*3*	*4*	*6*	*6*	*8*	*8*
$\overline{R}^2$	*0.028*	*0.0327*	*0.0346*	*0.0437*	*0.0436*	*0.0488*
Akaike	*12266*	*12261.6*	*10496*	*-717.64*	*-646.21*	*10480*
SIC	*12282*		*10526*	*-686.2*	*-605.5*	*10521*

Modelle mit der abhängigen Variable *bwght*: 2, 3, 5 und 8

Unter diesen konkurrierenden Modellen weist das Regressionsmodell **8** den geringsten Wert für das Akaike und SIC-Informationskriterium und das höchste $\overline{R}^2$ auf.

Achtung: Die adjustierten R^2 können nur zwischen Modellen verglichen werden, in denen die abhängige Variable y identisch ist → nicht vergleichbar zwischen log-lin und lin-lin oder lin-log Modellen!

Wie bei dem adjustierten Bestimmtheitsmass R^2 können die Informationskriterien nur zwischen Modellen verglichen werden, welche die gleiche abhängige Variable y besitzen. Deshalb müssen die Modelle 6 und 7 mit *l_bwght* separat betrachtet werden.

Zwischen den beiden Modellen 6 und 7 (abhängige Variable = ln_bwght) weist das Modell 6 das geringste Informationskriterium auf. Modell 6 berücksichtigt das Ausbildungsniveau der Eltern nicht, welches auch nicht signifikant ist. Es gibt sicherlich viele andere möglichen Regressionsmodelle, welche zusätzlichen erklärenden Variablen enthalten.

Diese Regressionsmodelle können einen Grossteil der Varianz des Geburtsgewichtes nicht gut erklären, da andere physiologischen Faktoren wie z.B. Gewicht und Grösse der Frau eine bedeutende Rolle spielen.

Modell 6 (log-lin) und Modell 8 (lin) sind Kandidaten. Welches ist zu bervorzugen? Der Vergleich erfolgt über R^2 (beide haben gleich viele Variablen). Für das log-lin Modell 6 ist die Berechnung von R^2 über den Stichproben-Korrelationskoeffizienten!