

**Aufgabe 3** (6 + 4 + 8 + 10 + 4 + 4 + 9 + 3 + 4 = 52 Punkte)

*Hinweis: Beachten Sie die Tabellen mit Quantilen am Ende der Aufgabenstellung!*

Mit Hilfe eines multiplen linearen Regressionsmodells soll das Geburtsgewicht `bweight` von Säuglingen (in [g]) mit Hilfe verschiedener Regressoren erklärt werden.

Man vermutet zunächst, dass das Gewicht `mweight` der Mutter (in [kg]) unmittelbar vor der Schwangerschaft sowie die Frage, ob die Mutter Raucherin ist (Dummyvariable `msmoke`, die den Wert 1 annimmt, wenn die Mutter Raucherin ist, 0 sonst), einen Einfluss auf das Geburtsgewicht haben. Daher wird das Modell

$$\text{bweight}_i = \beta_0 + \beta_1 \text{mweight}_i + \beta_2 \text{msmoke}_i + u_i$$

( $i = 1, \dots, n$ ) mit Hilfe der KQ-Methode unter Annahme homoskedastischer Störgrößen mit folgendem Ergebnis geschätzt:

Call:

```
lm(formula = bweight ~ mweight + msmoke)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max     |
|----------|---------|--------|--------|---------|
| -2030.90 | -445.69 | 29.16  | 521.76 | 1967.76 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 2501.125 | 230.836    | 10.835  | <2e-16 *** |
| mweight     | 9.340    | 3.726      | ???     | 0.0130 *   |
| msmoke      | -272.081 | ???        | -2.577  | 0.0107 *   |

---

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

Residual standard error: 707.8 on 186 degrees of freedom

Multiple R-squared: 0.06777, Adjusted R-squared: ???

F-statistic: 6.761 on 2 and 186 DF, p-value: 0.001464

Gehen Sie zunächst von der korrekten Spezifikation des Modells aus.

- Bestimmen Sie die drei fehlenden Werte (im Output mit ??? gekennzeichnet) und notieren Sie diese (mit den Bezeichnungen  $t_{\text{mweight}}$ ,  $\hat{\sigma}_{\text{msmoke}}$  und  $\bar{R}^2$ ) in Ihr Klausurheft.
- Hat das Gewicht der Mutter unmittelbar vor der Schwangerschaft einen signifikant positiven Einfluss auf das Geburtsgewicht (Signifikanzniveau  $\alpha = 0.01$ )? Ist das Geburtsgewicht signifikant niedriger, wenn die Mutter Raucherin ist, verglichen damit, dass die Mutter Nichtraucherin ist (Signifikanzniveau  $\alpha = 0.01$ )? Begründen Sie jeweils Ihre Antwort.
- Testen Sie zum Signifikanzniveau  $\alpha = 0.05$ , ob die erwartete Zunahme beim Geburtsgewicht je Kilogramm Körpergewicht der Mutter unmittelbar vor der Schwangerschaft mehr als 3 Gramm beträgt.

Geben Sie hierzu die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.

- (d) Aus der Regressormatrix
- $\mathbf{X}$
- berechnet man:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.1063482 & -0.0016447 & -0.0107551 \\ -0.0016447 & 0.0000277 & 0.0000347 \\ -0.0107551 & 0.0000347 & 0.0222526 \end{pmatrix}$$

Geben Sie ein Prognoseintervall ( $1 - \alpha = 0.95$ ) für das erwartete Geburtsgewicht (in [g]) an, wenn die zukünftige Mutter Nichtraucherin ist und unmittelbar vor der Schwangerschaft 75 Kilogramm wiegt.

- (e) Als Output eines Breusch-Pagan-Tests (nach Koenker) erhalten Sie das folgende Resultat:

studentized Breusch-Pagan test

```
data:  lm(bweight ~ mweight + msmove)
BP = 1.2116, df = 2, p-value = 0.5456
```

Was wird mit diesem Test untersucht und wie lautet das Ergebnis der Untersuchung, wenn man ein Signifikanzniveau von  $\alpha = 0.05$  zu Grunde legt?

- (f) Man vermutet nun, dass auch das Alter
- `mage`
- der Mutter (in Jahren) einen Einfluss auf das Geburtsgewicht hat und schätzt daher (unter Annahme homoskedastischer Störgrößen) das Modell

$$\text{bweight}_i = \beta_0 + \beta_1 \text{mweight}_i + \beta_2 \text{msmove}_i + \beta_3 \text{mage}_i + u_i$$

( $i = 1, \dots, n$ ) mit dem Ergebnis:

Call:

```
lm(formula = bweight ~ mweight + msmove + mage)
```

Residuals:

|  | Min      | 1Q      | Median | 3Q     | Max     |
|--|----------|---------|--------|--------|---------|
|  | -2070.97 | -431.56 | 12.95  | 521.70 | 1811.67 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 2362.495 | 300.683    | 7.857   | 3.12e-13 *** |
| mweight     | 8.853    | 3.791      | 2.335   | 0.0206 *     |
| msmove      | -269.257 | 105.801    | -2.545  | 0.0117 *     |
| mage        | 7.154    | 9.925      | 0.721   | 0.4719       |

---

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

Residual standard error: 708.8 on 185 degrees of freedom

Multiple R-squared: 0.07039, Adjusted R-squared: 0.05531

F-statistic: 4.669 on 3 and 185 DF, p-value: 0.003606

Können Sie die oben geäußerte Vermutung bei einem Signifikanzniveau von  $\alpha = 0.10$  bestätigen?

Würden Sie auf Grundlage der folgenden Varianzinflationsfaktoren von einem Multikollinearitätsproblem bei der obigen Modellschätzung ausgehen?

| mweight  | msmove   | mage     |
|----------|----------|----------|
| 1.034917 | 1.003332 | 1.034932 |

- (g) In Erweiterung des ursprünglichen Modells vermutet man nun, dass die Ethnie der Mutter (im Datensatz vorhandene Ausprägungen: **white**, **black**, **other**) einen Einfluss auf das Geburtsgewicht hat und fügt daher die beiden Dummyvariablen **mwhite** bzw. **mblack** ein, die jeweils den Wert 1 annehmen, wenn die Ethnie die Ausprägung **white** bzw. **black** hat, und 0 sonst. Man erhält das folgende (unwesentlich gekürzte) Ergebnis der Modellschätzung (unter Annahme homoskedastischer Störgrößen):

Call:

```
lm(formula = bweight ~ mweight + msmove + mwhite + mblack)
```

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 2404.305 | 223.773    | 10.744  | < 2e-16  | *** |
| mweight     | 8.681    | 3.759      | 2.310   | 0.022024 | *   |
| msmove      | -399.772 | 108.514    | -3.684  | 0.000302 | *** |
| mwhite      | 394.979  | 117.834    | 3.352   | 0.000974 | *** |
| mblack      | -108.935 | 165.919    | -0.657  | 0.512288 |     |

---

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

Residual standard error: 680.3 on 184 degrees of freedom

Multiple R-squared: 0.1481, Adjusted R-squared: 0.1296

F-statistic: 7.998 on 4 and 184 DF, p-value: 5.757e-06

Untersuchen Sie mit einem geeigneten Test, ob die ethnische Herkunft der Mutter insgesamt einen signifikanten ( $\alpha = 0.01$ ) Einfluss auf das Geburtsgewicht hat.

Geben Sie hierzu insbesondere die Hypothesen, die Teststatistik mit ihrer Verteilung unter  $H_0$ , den kritischen Bereich, die realisierte Teststatistik sowie die Testentscheidung an. Beantworten Sie auch explizit die oben formulierte Fragestellung.

- (h) Erläutern Sie kurz, warum es problematisch wäre, im Modell aus Teil (g) zusätzlich eine Dummyvariable **other** (mit Wert 1, falls die Ausprägung der Ethnie der Mutter **other** ist, 0 sonst) aufzunehmen.
- (i) Geben Sie auf Basis des in Teil (g) formulierten und geschätzten Modells eine Punktprognose für das erwartete Geburtsgewicht (in [g]) an, wenn die zukünftige Mutter die ethnische Herkunft **white** hat, Nichtraucherin ist und unmittelbar vor der Schwangerschaft 75 Kilogramm wiegt.

*Hinweis: Verwenden Sie die folgende Tabelle mit Quantilen einiger  $t(n)$ -Verteilungen*

| $n \backslash p$ | 0.85  | 0.90  | 0.95  | 0.975 | 0.99  | 0.995 | 0.9995 |
|------------------|-------|-------|-------|-------|-------|-------|--------|
| 184              | 1.039 | 1.286 | 1.653 | 1.973 | 2.347 | 2.603 | 3.344  |
| 186              | 1.039 | 1.286 | 1.653 | 1.973 | 2.347 | 2.603 | 3.344  |

*sowie die folgende Tabelle mit 0.99-Quantilen einiger  $F(m, n)$ -Verteilungen:*

| $n \backslash m$ | 2      | 3      | 184    | 185    | 186    |
|------------------|--------|--------|--------|--------|--------|
| 2                | 99.000 | 99.166 | 99.494 | 99.494 | 99.494 |
| 3                | 30.817 | 29.457 | 26.188 | 26.188 | 26.187 |
| 184              | 4.722  | 3.890  | 1.411  | 1.410  | 1.410  |
| 185              | 4.722  | 3.889  | 1.410  | 1.410  | 1.409  |
| 186              | 4.721  | 3.889  | 1.409  | 1.409  | 1.408  |

**Aufgabe 4** (6 + 6 + 5 + 2 + 7 + 4 = 30 Punkte)

Mit einem multiplen Regressionsmodell der Gestalt

$$\begin{aligned} \text{wage}_i = & \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{blue}_i + \beta_3 \text{unio}_i + \beta_4 \text{educ}_i \text{blue}_i \\ & + \beta_5 \text{educ}_i \text{unio}_i + \beta_6 \text{blue}_i \text{unio}_i + \beta_7 \text{educ}_i \text{blue}_i \text{unio}_i + u_i \end{aligned}$$

( $i \in \{1, \dots, n\}$ ) soll das Einkommen **wage** in Abhängigkeit der Anzahl von Ausbildungsjahren **educ**, den Dummyvariablen

- **blue** mit Wert 1 für Arbeiter und 0 für Angestellte sowie
- **unio** mit Wert 1 bei Bezahlung nach Tarifvertrag und 0 sonst

sowie diversen Interaktionstermen erklärt werden. Dazu wurde das Modell zunächst unter Annahme unabhängig identisch normalverteilter Störgrößen wie folgt geschätzt:

Call:

```
lm(formula = wage ~ educ + blue + unio + I(educ * blue) + I(educ * unio) + I(blue * unio) + I(educ * blue * unio))
```

Residuals:

| Min     | 1Q     | Median | 3Q    | Max    |
|---------|--------|--------|-------|--------|
| -1179.7 | -281.4 | -37.7  | 204.6 | 3462.3 |

Coefficients:

|                       | Estimate | Std. Error | t value | Pr(> t )     |
|-----------------------|----------|------------|---------|--------------|
| (Intercept)           | -257.11  | 205.78     | -1.249  | 0.211987     |
| educ                  | 111.46   | 13.78      | 8.090   | 3.44e-15 *** |
| blue                  | 514.61   | 281.47     | 1.828   | 0.068009 .   |
| unio                  | 1231.94  | 423.39     | 2.910   | 0.003754 **  |
| I(educ * blue)        | -61.24   | 21.88      | -2.799  | 0.005295 **  |
| I(educ * unio)        | -98.54   | 29.14      | -3.382  | 0.000768 *** |
| I(blue * unio)        | -606.57  | 511.10     | -1.187  | 0.235784     |
| I(educ * blue * unio) | 65.57    | 38.62      | 1.698   | 0.090128 .   |

---

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

Residual standard error: 455 on 587 degrees of freedom

Multiple R-squared: 0.2747, Adjusted R-squared: 0.266

F-statistic: 31.75 on 7 and 587 DF, p-value: < 2.2e-16

- (a) Wie groß ist gemäß der obigen Modellschätzung der Effekt eines zusätzlichen Ausbildungsjahres für
- nach Tarifvertrag bezahlte Arbeiter,
  - außertariflich bezahlte Angestellte?
- (b) Welches Einkommen prognostiziert (Punktprognose) die obige Modellschätzung für
- einen Arbeiter mit 12 Ausbildungsjahren, der nach Tarifvertrag bezahlt wird,
  - einen Angestellten mit 15 Ausbildungsjahren, der außertariflich bezahlt wird?

- (c) Man vermutet, dass die bei der ersten Modellschätzung getroffene Annahme homoskedastischer Störgrößen nicht gerechtfertigt ist. Um dies zu überprüfen, wird zunächst eine Hilfsregression mit den quadrierten Residuen  $\text{uhat2}_i := \hat{u}_i^2$  der ursprünglichen Modellschätzung mit folgendem Ergebnis durchgeführt:

Call:

```
lm(formula = uhat2 ~ educ + blue + unio + I(educ * blue) + I(educ * unio) + I(blue * unio) + I(educ * blue * unio))
```

Residuals:

| Min     | 1Q      | Median | 3Q    | Max      |
|---------|---------|--------|-------|----------|
| -502034 | -124076 | -63761 | 28002 | 11484361 |

Coefficients:

|                       | Estimate | Std. Error | t value | Pr(> t )  |
|-----------------------|----------|------------|---------|-----------|
| (Intercept)           | -494141  | 283661     | -1.742  | 0.0820 .  |
| educ                  | 58682    | 18992      | 3.090   | 0.0021 ** |
| blue                  | 520614   | 388005     | 1.342   | 0.1802    |
| unio                  | 759687   | 583637     | 1.302   | 0.1935    |
| I(educ * blue)        | -51627   | 30163      | -1.712  | 0.0875 .  |
| I(educ * unio)        | -70406   | 40167      | -1.753  | 0.0802 .  |
| I(blue * unio)        | -710348  | 704548     | -1.008  | 0.3138    |
| I(educ * blue * unio) | 63436    | 53244      | 1.191   | 0.2340    |

---

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

Residual standard error: 627200 on 587 degrees of freedom

Multiple R-squared: 0.0622, Adjusted R-squared: 0.05102

F-statistic: 5.562 on 7 and 587 DF, p-value: 3.225e-06

Führen Sie einen Breusch-Pagan-Test in der Variante nach Koenker durch, um das Vorliegen von Heteroskedastie in den Störgrößen zum Signifikanzniveau  $\alpha = 0.01$  zu überprüfen.

Geben Sie hierzu nur den kritischen Bereich, die realisierte Teststatistik sowie die Antwort auf die Frage, ob von Heteroskedastie ausgegangen werden sollte, an.

Hinweis: Verwenden Sie die folgende Tabelle mit Quantilen einiger  $\chi^2(n)$ -Verteilungen

| $n \backslash p$ | 0.01  | 0.025 | 0.05  | 0.50  | 0.90   | 0.95   | 0.975  | 0.99   |
|------------------|-------|-------|-------|-------|--------|--------|--------|--------|
| 1                | 0.000 | 0.001 | 0.004 | 0.455 | 2.706  | 3.841  | 5.024  | 6.635  |
| 2                | 0.020 | 0.051 | 0.103 | 1.386 | 4.605  | 5.991  | 7.378  | 9.210  |
| 3                | 0.115 | 0.216 | 0.352 | 2.366 | 6.251  | 7.815  | 9.348  | 11.345 |
| 4                | 0.297 | 0.484 | 0.711 | 3.357 | 7.779  | 9.488  | 11.143 | 13.277 |
| 5                | 0.554 | 0.831 | 1.145 | 4.351 | 9.236  | 11.070 | 12.833 | 15.086 |
| 6                | 0.872 | 1.237 | 1.635 | 5.348 | 10.645 | 12.592 | 14.449 | 16.812 |
| 7                | 1.239 | 1.690 | 2.167 | 6.346 | 12.017 | 14.067 | 16.013 | 18.475 |
| 8                | 1.646 | 2.180 | 2.733 | 7.344 | 13.362 | 15.507 | 17.535 | 20.090 |
| 9                | 2.088 | 2.700 | 3.325 | 8.343 | 14.684 | 16.919 | 19.023 | 21.666 |
| 10               | 2.558 | 3.247 | 3.940 | 9.342 | 15.987 | 18.307 | 20.483 | 23.209 |

- (d) Welche der Ergebnisse aus Teil (a) und (b) können sich ändern, wenn statt der ursprünglichen Modellschätzung eine Schätzung mit robusten Standardfehlern, beispielsweise unter Verwendung von  $\hat{V}_{\text{hcl}}(\hat{\beta})$ , durchgeführt wird?

- (e) Unter Zulassung von Heteroskedastie in den Störgrößen soll nun ein Konfidenzintervall für den Effekt eines zusätzlichen Ausbildungsjahres für nach Tarifvertrag bezahlte Angestellte konstruiert werden.
- (i) Bestimmen Sie zunächst den marginalen Effekt eines zusätzlichen Ausbildungsjahres für nach Tarifvertrag bezahlte Angestellte in Abhängigkeit der Regressionsparameter.
- (ii) Aus der heteroskedastie-konsistenten Schätzung  $\hat{V}_{\text{hcl}}(\hat{\beta})$  der Varianz-Kovarianzmatrix von  $\hat{\beta}$  wurde bereits das Zwischenergebnis

$$\mathbf{a}'\hat{V}_{\text{hcl}}(\hat{\beta})\mathbf{a} = 427.311 \quad \text{mit} \quad \mathbf{a}' = (0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0)$$

berechnet. Bestimmen Sie damit das realisierte Konfidenzintervall zum Konfidenzniveau  $1 - \alpha = 0.95$  für den Effekt eines zusätzlichen Ausbildungsjahres für nach Tarifvertrag bezahlte Angestellte.

*Hinweis:* Verwenden Sie  $t_{587,0.975} = 1.964$ .

- (f) Verwenden Sie die nachstehende Grafik mit zwei Konfidenzellipsen zum Konfidenzniveau  $1 - \alpha = 0.99$  für die beiden Parameter  $\beta_6$  und  $\beta_7$ , um eine Entscheidung zwischen

$$H_0 : \begin{pmatrix} \beta_6 \\ \beta_7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{gegen} \quad H_1 : \begin{pmatrix} \beta_6 \\ \beta_7 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

zum Signifikanzniveau  $\alpha = 0.01$  zunächst unter der Annahme homoskedastischer Störgrößen zu treffen.

Wie fällt die Entscheidung aus, wenn Sie von heteroskedastischen Störgrößen ausgehen (und dies bei der Durchführung des Tests auch entsprechend berücksichtigen)?

