1. Lösungen zu Kapitel 7

Übungsaufgabe 7.1

Um zu testen ob die Störterme ε_i eine konstante Varianz haben, sprich die Homogenitätsannahme erfüllt ist, sind der *Breusch-Pagan-Test* und der *White-Test* zwei verbreitete Tests.

Beide testen die Nullhypothese, das Vorliegen von Homoskedastizität ($H_0 : Var(\varepsilon|\boldsymbol{X}) = \sigma^2$), gegen die Alternativhypothese dass Heteroskedastizität vorliegt ($H_1 : Var(\varepsilon|\boldsymbol{X}) = \sigma_i^2$).

Für den *Breusch-Pagan-Test* wird im ersten Schritt das lineare Regressionsmodell geschätzt:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon. \tag{1.1}$$

Um zu testen, ob die Residuen mit den erklärenden Variablen korreliert sind, muss eine Annahme über die Beziehung zwischen ε und X_k getroffen werden. Der Breusch-Pagan-Test geht von einer linearen Beziehung aus. Daher werden im zweiten Schritt die quadrierten Residuen des geschätzten Modells auf die erklärenden Variablen regressiert:

$$\hat{\varepsilon}^2 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_K X_K + \nu. \tag{1.2}$$

Die Nullhypothese des Vorliegens von Homoskedastizität besagt, dass die erklärenden Variablen gemeinsam keinen signifikanten Einfluss auf die Residuen ausüben ($H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_K = 0$). Dies kann mit Hilfe des F-Tests oder LM-Tests überprüft werden. So hat die F-Statistik für die Nullhypothese folgende Form:

$$F_{BP} = \frac{R^2_{\hat{\varepsilon}^2}/K}{(1 - R^2_{\hat{\varepsilon}^2})/(N - (K+1))},$$

wobei $R^2_{\hat{\varepsilon}^2}$ das Bestimmtheitsmaß aus Gleichung (1.2) ist. Ist F_{BP} größer als der kritische Wert der F-Verteilung, kann die Nullhypothese homoskedastischer Fehler abgelehnt werden.[1]

[1] Die entsprechende LM-Statistik lautet: $LM_{BP} = N \cdot R^2_{\hat{\varepsilon}^2}$. Ist bei vorgegebener Irrtumswahrscheinlichkeit LM_{BP} größer als der kritische Wert der χ^2-Verteilung mit K Freiheitsgraden, kann die Nullhypothese homoskedastischer Fehler verworfen werden.

Alternativ lässt sich die Homoskedastizitätsannahme mit dem *White-Test* testen, bei dem *a priori* keine Annahmen über die funktionale Beziehung zwischen der Fehlervarianz und den Regressoren benötigt werden. Darüber hinaus reagiert er im Gegensatz zum Breusch-Pagan-Test nicht sensitiv auf eine Verletzung der Annahme normalverteilter Fehler.

Wie beim Breusch-Pagan-Test wird im ersten Schritt Gleichung (1.1) geschätzt.

Im zweiten Schritt werden die geschätzten Residuen nicht nur auf die einfachen Regressoren, sondern auch auf deren Quadrat und deren Kreuzprodukte regressiert:

$$\begin{aligned}\hat{\varepsilon}^2 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_K X_K \\ + \gamma_{K+1} X_1^2 + \gamma_{K+2} X_2^2 + \ldots + \gamma_{2K} X_K^2 \\ + \gamma_{2K+1} X_1 X_2 + \ldots \gamma_{K+[0,5K(K+1)]} X_{K-1} X_K + \nu. \end{aligned} \tag{1.3}$$

Wie zuvor beim Breusch-Pagan-Test kann nun mit Hilfe eines F- oder LM-Tests die Hypothese getestet werden, dass alle Regressoren mit Ausnahme der Konstante gleichzeitig Null sind.

Übungsaufgabe 7.2

a) Es ist zu vermuten, dass der Bierkonsum mit steigendem Einkommen stärker variiert.
Während bei Haushalten mit einem geringen Einkommen nach Erfüllung der Grundbedürfnisse (Miete, Nahrung, etc.) nur wenig Varianz im Bierkonsum möglich ist, können Haushalte mit einem hohen Einkommen mehr Geld für ihren Bierkonsum ausgeben. Gleichzeigt ist es aber auch möglich, dass Haushalte mit einem hohen Einkommen viel sparen oder Bier durch andere Konsumgüter substituieren und somit einen unterdurchschnittlichen Bierkonsum haben. Dies führt zu einer höheren Varianz der Bierkonsumausgaben in Haushalten mit einem höheren Einkommen.
Diese Vermutung ließe sich durch die in Aufgabe **7.1** beschriebenen Tests testen.

b) Sind Informationen über die funktionale Form der Heteroskedastizität vorhanden, können diese für eine Transformation des Regressionsmodells verwendet werden.
Hat die Varianz des Störterms die Form $\sigma_i^2 = \sigma^2 E^2$, erfolgt die Transformation durch Division aller Variablen durch E^2:

$$\frac{\ln(B)}{E} = \frac{\beta_0}{E} + \beta_1 \frac{\ln(E)}{E} + \beta_2 \frac{S}{E} + \beta_3 \frac{A}{E} + \frac{\varepsilon}{E}.$$

c)

$$\ln(B) * E = \beta_0 * E + \beta_1 \ln(E) * E + \beta_2 S * E + \beta_3 A * E + \varepsilon * E$$

d)

$$\frac{\ln(B)}{\sqrt{E}} = \frac{\beta_0}{\sqrt{E}} + \beta_1 \frac{\ln(E)}{\sqrt{E}} + \beta_2 \frac{S}{\sqrt{E}} + \beta_3 \frac{A}{\sqrt{E}} + \frac{\varepsilon}{\sqrt{E}}$$

Übungsaufgabe 7.3

a) Das Modell schätzt die Anzahl der täglich gerauchten Zigaretten einer Person in Abhängigkeit vom Preis der Zigaretten, des Einkommens, der Schuljahre, des Alters sowie des Geschlechts.
Weder der Zigarettenpreis noch das Einkommen haben einen statistisch signifikanten Einfluss auf die Anzahl der täglich gerauchten Zigaretten. Die Teststatistiken für den zweiseitigen t-Test sehen folgendermaßen aus:

$$|t_{Zigarettenpreis}| = \left|\frac{-0,069}{0,207}\right| = 0,333$$
$$|t_{Einkommen}| = \left|\frac{0,012}{0,026}\right| = 0,462.$$

Beide Werte sind kleiner als der kritische Wert der t-Verteilung sowohl bei einer Irrtumswahrscheinlichkeit von einem Prozent ($t_{krit} = 2,576$), fünf Prozent (1,960) als auch bei einer Irrtumswahrscheinlichkeit von zehn Prozent (1,645).
Ein weiteres Schuljahr reduziert die Anzahl der täglich gerauchten Zigaretten um 0,029. Dieser Effekt ist statistisch signifikant von Null verschieden.
Das Alter hat einen positiven aber abnehmenden Effekt auf den Zigarettenkonsum. Sowohl der Koeffizient des Alters als auch der Koeffizient des Alters zum Quadrat sind signifikant. Ab einem Alter von 76,9 Jahren reduziert ein weiteres Lebensjahr den Zigarettenkonsum.
Der Unterschied zwischen Männern und Frauen im Zigarettenkonsum ist nicht signifikant.
Insgesamt erklären die Variablen 6,2% der gesamten Varianz im Zigarettenkonsum.

b)

$$\begin{aligned}\text{Zigaretten} = &-0,069 * 67,44 + 0,012 * 6.500 - 0,029 * 16 \\ &+0,020 * 77 - 0,00026 * 77^2 - 0,026 * 1 + 0,656 \\ &= 73,5\end{aligned}$$

Die durchschnittliche Anzahl von Zigaretten pro Tag für eine Person mit den vorgegebenen Eigenschaften liegt bei 73,5.

c) Bei der Berechnung robuster Standardfehlern wird versucht, die korrekte Varianz entsprechend der Gleichung $V(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$ zu schätzen.

Die Varianz der Koeffizienten eines bivariaten Regressionsmodells bei Heteroskedastizität hat die Form:

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2\sigma_i^2}{[\sum_{i=1}^{N}(X_i - \bar{X})^2]^2}. \tag{1.4}$$

White (1980) hat gezeigt, dass Gleichung (1.4) geschätzt werden kann, indem σ_i durch die quadrierten Residuen aus der OLS-Regression $\hat{\varepsilon}^2$ ersetzt werden:

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2\hat{\varepsilon}_i^2}{[\sum_{i=1}^{N}(X_i - \bar{X})^2]^2}. \tag{1.5}$$

Diese leicht zu berechnende Anpassung der Standardfehler ist hilfreich, da man unabhängig von der aktuellen Form der Heteroskedastizität *robuste Standardfehler* erhält.
In dem vorliegenden Modell gibt es keine wichtigen Unterschiede zwischen den normalen Standardfehlern und den robusten Standardfehlern. Bei einigen Koeffizienten ist der robuste Standardfehler größer als der normale Standardfehler (z.B. log(Zigarettenpreis)) und bei anderen Koeffizienten kleiner (z.B. Alter).
Jedoch führt im vorliegenden Beispiel die Betrachtung der robusten Standardfehler anstelle der normalen Standardfehler der Koeffizienten in keinem Fall zum Verwerfen (Nicht-Verwerfen) einer zuvor nicht-verworfenen (verworfenen) Nullhypothese beim t-Test.

d) Um zu testen, ob die Variablen *Alter* und $Alter^2$ gemeinsam statistisch signifikant von Null verschieden sind, kann man den F-Test oder den LM-Test verwenden. Mit diesen kann die Nullhypothese $H_0 : \beta_{Alter} = \beta_{Alter^2} = 0$ getestet werden.
Bei beiden Vorgehen wird im ersten Schritt das restringierten Modells, d.h. des Modells ohne die Variablen *Alter* und $Alter^2$, geschätzt.
Für den F-Test wird zusätzlich das unrestringierte Modell geschätzt und die Teststatistik berechnet sich aus dem R^2 des restringierten und des unrestringierten Modells,

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(N - (K + 1))}.$$

Der kritische Wert kann bei gegebener Irrtumswahrscheinlichkeit der F-Verteilung entnommen werden. Ist die Teststatistik größer als der kritische Wert, so kann die Nullhypothese, dass das Alter und das $Alter^2$ keinen signifikanten Beitrag zum Erklären des Zigarettenkonsums leisten, verworfen werden.
Beim LM-Test das das Schätzen des unrestringierten Modells nicht notwendig. Im zweiten Schritt werden die geschätzten Residuen aus dem restringierten Modell auf die nicht dort nicht berücksichtigen Variablen, sprich Alter und $Alter^2$ regressiert,

$$\hat{\varepsilon} = \gamma + \gamma_1 Alter + \gamma_2 Alter^2 + \nu. \tag{1.6}$$

Die Testgrößer berechnet sich als $N \cdot R^2_{\hat{\varepsilon}}$, wobei N die Anzahl an Beobachtungen ist und $R^2_{\hat{\varepsilon}}$ das Bestimmtheitsmaß aus Regression (1.6). Der kritische Wert wird der χ^2_q-Verteilung entnommen, wobei q die Anzahl der Restriktionen (in diesem Fall $q = 2$) ist. Auch hier wird die Nullhypothese abgelehnt, wenn die LM-Statistik größer als der kritische Wert ist.

e) Um zu testen, ob der Effekt der Schuljahre auf das Rauchverhalten für Frauen und Männer verschieden ist, kann das Modell um einen Interaktionsterm zwischen der Anzahl der Schuljahre und einer Dummy-Variablen für Frauen erweitert werden.
Mit Hilfe des t-Tests kann getestet werden, ob dieser Koeffizient signifikant verschieden von Null ist.

Übungsaufgabe 7.4

a) Heteroskedastizität tritt automatisch bei gruppierten Daten auf. Im vorliegenden Beispiel liegen die Daten in Form von Durchschnittswerten auf Industrieebene vor. Der Störterm des Modells ε_i ist ebenfalls ein Mittelwert.
Die Varianz des Störterms ist gegeben durch $Var(\bar{\varepsilon_i}) = \sigma^2/N_i$, wobei N_i die Anzahl der Unternehmen in der jeweiligen Industrie beschreibt. Da N_i zwischen den Industrien variiert, nimmt auch die Varianz des Fehlerterms einen unterschiedlichen Wert an.

b) Während der Breusch-Pagan-Test eine lineare Beziehung zwischen der Fehlervarianz und den Regressoren annimmt, testet der Glesjer-Test die folgende Beziehung:

$$|\hat{\varepsilon}| = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k + \nu.$$

Beim Glesjer-Test wird folglich der Betrag der geschätzten Residuen $\hat{\varepsilon}$ und nicht die quadrierten geschätzten Residuen auf die Konstante und die erklärenden Variablen regressiert.

c) Der White-Test hat den Vorteil, dass keinerlei *a-priori*-Informationen über die Beziehung zwischen Fehlervarianz und den für die Heteroskedastizität verantwortlichen Regressoren notwendig sind.
Im ersten Schritt werden die Umsätze in Abhängigkeit von dem durchschnittlichen Budget für Werbung aller Unternehmen in der Industrie geschätzt.

$$U_i = \beta_0 + \beta_1 W_i + \varepsilon_i$$

Im zweiten Schritt werden die quadrierten geschätzten Residuen auf die erklärenden Variablen und deren Quadrat regressiert (im Fall von mehreren erklärenden Variablen auch deren Kreuzprodukt):

$$\hat{\varepsilon}^2 = \gamma_0 + \gamma_1 W_i + \gamma_2 W_i^2 + \nu.$$

Mit Hilfe des F- oder LM-Tests kann nun die Hypothese getestet werden, dass alle Regressoren mit Ausnahme der Konstante gleichzeitig Null sind. Mit dem vorliegenden R^2 der Hilfsregression von 0,474 ergibt sich eine Teststatistik für den LM-Test von 57 $(0,474 \cdot 120)$.
Die Teststatistik ist asymptotisch χ_2^2-verteilt und der kritische Wert bei einer Irrtumswahrscheinlichkeit von 5% ist 5,99. Da die LM-Statistik größer als der kritische Wert ist, wird die Nullhypothese (Homoskedastizität) abgelehnt.

d) Bei dem *Generalized Least Squares* werden vorliegende Informationen genutzt um heteroskedastische Fehler explizit zu modellieren. Hierzu werden die Variablen des Modells so transformiert, dass die Residuen homoskedastisch sind.
Im vorliegenden Fall ist die funktionale Form der Heteroskedastizität bekannt. Der Standardfehler des Störterms hat die Form $\sigma_i/\sqrt{N_i}$, wobei N_i die Anzahl der Unternehmen in der jeweiligen Industrie beschreibt.
Das Modell kann folglich durch Multiplikation mit $\sqrt{N_i}$ transformiert werden:

$$U_i * \sqrt{N_i} = \beta_0 * \sqrt{N_i} + \beta_1 W_i * \sqrt{N_i} + \varepsilon_i * \sqrt{N_i}.$$

Die Varianz des Fehlerterms hat nun die Form:

$$Var(\varepsilon_i * \sqrt{N_i}) = Var(\varepsilon_i) * N_i = \sigma^2.$$

Die Fehler des transformierten Modells sind homoskedastisch und die Schätzung des transformierten Regressionsmodells liefert konsistente, unverzerrte und insbesondere effiziente Schätzer.

Übungsaufgabe 7.5

a) Bis auf die Koeffizienten der Größe der Hochschule sowie der Dummy-Variablen, die anzeigt ob es sich um eine Fachhochschule handelt, sind alle Koeffizienten für sich genommen signifikant von Null verschieden.
Je höher die Reputation einer Hochschule desto höher sind die jährlichen Studiengebühren. Ein Anstieg des Index um eine Kategorie führt zu einer Erhöhung der Gebühren um 3.985,20 €. Eine private Hochschule erhebt im Schnitt um 8.406,79 € höhere Gebühren. Schließlich erheben Hochschulen in den neuen Bundesländern um 2.376,51 € geringere Gebühren. Das R^2 zeigt, dass die Regression 72% der gesamten Variation der Studiengebühren erklärt.

b)

$$\begin{aligned}\text{Studiengebühren} &= 7.311,17 + 3.985,20 \cdot 4,5 - 0,20 \cdot 1.500 \\ &\quad +8.406,79 \cdot 1 - 416,38 \cdot 0 - 2.376,51 \cdot 0 \\ &= 33.351,36\end{aligned}$$

Die vorhergesagten Kosten für eine Universität mit den gegebenen Charakteristika entsprechen 33.351,36 €.

c) Die F-Statistik errechnet sich aus dem R^2 des unrestringierten Modells, d.h. dem Modell welches die Variablen Größe und Fachhochschule enthält, und dem restringierten Modell ohne diese erklärenden Variablen. Man schätzt das folgende restringierte Modell:

$$\text{Gebühren} = \beta_0 + \beta_1 \text{Reputation} + \beta_2 \text{Privat} + \beta_3 \text{neue Bundesländer} + u$$

und behält das R_r^2 dieser Regression.
Die Formel für die F-Statistik lautet:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(N - (K+1))}$$

Der kritische Wert im vorliegenden Beispiel bei einer Irrtumswahrscheinlichkeit von 5% ist bei der F-Verteilugn für zwei Zählerfreiheitsgrade und 100 Nennerfreiheitsgrade $F_{krit} = 3,09$.
Da die errechnete F-Statistik von 1,23 kleiner als der kritische Wert ist, kann die Nullhypothese, dass Größe und Fachhochschule gleichzeitig Null sind, nicht abgelehnt werden.

d) In diesem Fall wird die Nullhypothese $H_0 : \beta_{\text{Größe}} = 0$ gegen die Alternativhypothese $H_1 : \beta_{\text{Größe}} < 0$ getestet. Die t-Statistik lautet:

$$t = \frac{-0,20 - 0}{0,07} = -2,857.$$

Bei einer Irrtumswahrscheinlichkeit von 5% ist der kritische t-Wert mit mit 100 Freiheitsgraden gleich -1,660. Da die t-Statistik kleiner als der kritische t-Wert ist, wird die Nullhypothese abgelehnt.
Die Huber-White-Standardfehler führen nicht zum Verwerfen der Nullhypothese während die unkorrigierten Standardfehler zum Verwerfen der Nullhypothese führen. Der Grund hierfür ist, dass bei heteroskedastischen Fehlern das OLS-Modell falsche Standardfehler liefert, was wiederum dazu führt, dass t-Tests und F-Tests irreführende Ergebnisse liefern. Man erhält mit der OLS-Methode zwar weiterhin erwartungstreue und konsistente Schätzer, jedoch sind diese nicht mehr effizient.
Der t-Test ist folglich nicht zuverlässig und sollte nicht verwendet werden.