Übung 1 „Regressionsanalyse“

"Energiemodelle und energiepolitische Analysen"

(SS 2014)

Abgabe bis 15.4. als pdf an hartner@eeg.tuwien.ac.at

Beispiel 1:
Ökonometrisches Nachfragemodell

1.1 Aufgabenstellung

Für zwei Länder ist vergleichsweise zu untersuchen, welche Parameter den STROMverbrauch im Sektor **Private Haushalte** beeinflussen. Dazu sind im Wesentlichen Preis- und Einkommenselastizitäten zu schätzen. Ausgangsdaten sind Zeitreihen dieser Länder für:

- Stromverbrauch Haushalte
- Strompreis
- Bruttoinlandsprodukt
- Index Real/nominal

Diese Daten werden zur Verfügung gestellt, siehe File attached.
Basierend auf den Ergebnissen der Regression sind **vier Szenarien** bis 2030 zu rechnen:

* Hochpreis + hohes GDP-Wachstum
* Hochpreis + niedriges GDP-Wachstum
* Moderater Preis + hohes GDP-Wachstum
* Moderater Preis + niedriges GDP-Wachstum

Annahmen:
Hochpreis: +4%/Jahr; + Moderater Preis +2%/Jahr; hohes GDP-Wachstum: +3%/Jahr; niedriges GDP-Wachstum + 1 %/Jahr;

1.2 Theoretische Grundlagen

Die Beschreibung der Entwicklung des Energieverbrauchs mit Hilfe von Preisen und Einkommen und möglichen zusätzlichen Parametern bzw. Verzögerungen mit Hilfe des Ansatzes einer sogenannten Produktionsfunktion ist eine wichtige Methode zur Erstellung von Energieprognosen.

Ausgangsmodell

Der einfachste Ansatz dazu lautet:

$$E_t = K \cdot p_t^{\alpha} \cdot Y_t^{\beta}$$

mit:

K	Konstante
E_t	Stromnachfrage im Jahr t
p_t	Haushaltsstrompreis im Jahr t
Y_t	Einkommen (z.B. Bruttoinlandsprodukt (BIP)
α	Preiselastizität
β	Einkommenselastizität

oder in der für die Abbildung der Elastizitäten (in einer linearen Regressionsanalyse) notwendigen logarithmischen Form:

$$\ln(E_t) = C + \alpha \cdot \ln(p_t) + \beta \cdot \ln(Y_t) \qquad \text{(Modell A)}$$

Durch das Logarithmieren können nun die linearen Koeffizienten in einer linearen Regressionsanalyse berechnet werden. Ausgehend von dieser Grundgleichung ist es nun möglich, verschiedene weitere Effekte zu untersuchen. In dieser Gleichung ist E_t die abhängige (*dependent*) Variable, p_t und Y_t sind die unabhängigen (*independent*) Variablen.

1.3 Vorgehensweise

- Überprüfen der Daten für die untersuchten Länder
- Umrechnen in reale Preise und reales Einkommen.
- Führen Sie mit Hilfe von Matlab, Excel, oder einem anderen Programm Ihrer Wahl Regressionsanalysen **für zumindest drei verschiedene Modelle** nach Kap. 5.5 bzw. 5.6 im Skriptum durch. (Basismodell, Modell mit Trend, Modell mit Lag). In Excel benötigen Sie das Add-In "Analyse-Funktionen" bzw. "Data analysis" um die Analyse-Funktion "Regression" ausführen zu können. In Matlab stehen die Funktionen „fitlm“, "regress" oder "regstats" zur Verfügung.

Die Dokumentation der Ergebnisse soll für den Gesamtenergieverbrauch in etwa wie in Tab. 1 beschrieben ausschauen:

Tabelle 1:

Ergebnisse der Schätzungen des Energieverbrauchs im Verkehr in Italien 1970 - 1995 mit verschiedenen Modellen. (Die Werte der T-Statistik in den eckigen Klammern)

	Modell A	**Modell Y**	**Modell Z**
C	5.38 [6.12]	4.29 [5.14]	4.09 [4.26]
α	-0.35 [-1.90]	-0.30 [-1.98]	-0.20 [-2.09]
β	0.40 [2.98]	0.35 [3.45]	0.30 [4.89]
θ	- -	0.02 [6.18]	- -
λ	- -	- -	0.50 [4.80]
δ	- -	0.08 [0.45]	- -
γ	- -	0.52 [3.46]	- -
ζ	- -	- -	0.12 [3.18]
SSE	0.338	0.256	0.233
R^2 (korrigiert)	0.91	0.94	0.97
F	123.4	129.5	160.7

- Dokumentieren Sie weiters die Gleichungen, welche Modelle Sie geschätzt haben und die Ergebnisse. Wichtig: Schätzen Sie zumindest ein Modell in dem ein Trend und ein Modell in dem ein Lag berücksichtigt wird!
- Für die Berechnung der **Szenarien** sollten Sie das Modell verwenden, das Ihrer Meinung nach **die besten Ergebnisse erzielt.** Hier ist vor allem auf die Plausibilität der Zusammenhänge zu achten!!!

- Stellen Sie die Szenarien graphisch dar! Vergleichen Sie in einer Grafik auch die **wahren historischen Werte mit Ihren Modellergebnissen!**
- **Zusatzfrage:** Welche weiteren Einflussfaktoren sollten Ihrer Meinung nach in das Modell aufgenommen werden um die Ergebnisse zu verbessern?

1.4 Kommentar zu statistischen Tests

Üblicherweise wird der kritische Wert für die T-Statistik aus Tabellen abgelesen und hängt vor allem von der Anzahl der Beobachtungen, den Freiheitsgraden des Modells und dem gewählten Signifikanzniveau ab. Sie können für Ihre Auswertung einen Standardwert für einen kritischen T-Wert von 1,96 annehmen. Liegt der Betrag der T-Statistik des Koeffizienten über dem kritischen Wert so kann dieser als Signifikant angenommen werden. Dies gilt auch für die F-Statistik. Diese Testgröße testet ebenfalls ob alle geschätzten Koeffizienten gemeinsam in diesem Modell einen signifikanten Einfluss auf die abhängige Variable haben. Der kritische Wert für F kann so wie der kritische T-Wert aus Tabellen abgelesen werden. Liegt die Prüfgröße F über diesem Wert, ist die Hypothese, dass die Parameter keinen Einfluss auf die abhängige Variable nehmen abzulehnen. Für die Übung reicht es die Prüfgröße F anzugeben (sie liegt in allen Modellen über dem kritischen Wert) und die Werte für T-Statistiken der einzelnen Parameter zu berücksichtigen, um die Qualität des Modells zu bewerten. Bei Anwendungen in der Praxis müsste das Modell genauer geprüft werden, wozu noch weitere Prüfgrößen mit einzubeziehen wären.

Beispiel 2:

Modellierung der Wärmenachfrage eines Fernwärmenetzes mit linearer Regression - einfache Modellansätze

Gegeben ist die stündliche Nachfrage nach Wärme (Raumwärme und Warmwasser) in einem Fernwärmenetz (gemessene Leistungsmittelwerte der Einspeisung - stündlich) und die dazugehörige Umgebungstemperatur. Die Nachfrager sind hauptsächlich Haushalte, zum Teil aber auch Gewerbebetriebe. Vergleichen und interpretieren Sie unterschiedliche Modellansätze zur Abschätzung der Nachfrage in Abhängigkeit von der Temperatur und der Tageszeit:

2.1) Modell 1: Einfache Abschätzung über Temperatur

$$y_t = \beta_0 + \beta_1 \cdot T_t$$

In diesem Modellansatz hängt die Nachfrage nur von der Umgebungsemperatur ab. Die Koeffizienten β_i (i=0,1) ergeben sich jeweils aus der linearen Regression, wobei die Funktion fitlm(…) zu verwenden ist. (In der Matlab Ausgabe unter der Spalte „Estimate“ bzw. in der Variable *„Modellname“*.Coefficients.Estimate(*index*).) Der Koeffizient β_0 entspricht jeweils der Konstanten (=Intercept).

a) Geben Sie die Koeffizienten und die dazugehörige t-Statistik an. Wie interpretieren Sie die Ergebnisse?

b) Vergleichen Sie die modellierten Werte für die Beobachtungen von t=1680:1775 bzw. t=6480:6575. Erstellen Sie dazu eine Grafik. Wie interpretieren Sie die Abweichungen? Wodurch unterscheiden sich die Abweichungen in den beiden Beobachtungsperioden?

2.2) Modell 2: Versuch der zusätzlichen Modellierung des Tagesverlaufs

$$y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot h_t + \beta_3 \cdot {h_t}^2 + \beta_4 \cdot {h_t}^3$$

Hier wird versucht, den typischen Tagesverlauf der Nachfrage (der nicht von der Temperatur abhängt) in das Modell zu integrieren. Die Variable h entspricht dabei der Spalte „Stunde“ im Datenblatt (bzw. data_heat.Stunde im Mat-File), weist also jeder Beobachtung, die dazugehörige Stunde zu. Die Stunden gehen hier zusätzlich zur Temperatur als Polynom 3.Grades in das Modell ein (Die Daten müssen also dementsprechend aufbereitet werden bevor die Regression durchgeführt wird). Die jeweiligen Koeffizienten werden wiederum über die lineare Regression mit der Funktion fitlm(…) geschätzt.

a) Geben Sie die Koeffizienten und die dazugehörige t-Statistik an. Wie interpretieren Sie die Ergebnisse?
b) Vergleichen Sie erneut die modellierten Werte für die Beobachtungen von t=1680:1775 bzw. t=6480:6575 und erstellen Sie eine Grafik. Was beobachten Sie? Wieso kann der gewählte Modellansatz den charakteristischen Tagesverlauf der Beobachtungen nicht wiedergeben?

2.3) Modell 3: Modellierung der Nachfrage getrennt für einzelne Stunden

$$y_t^j = \beta_0 + \beta_1 \cdot T_t^j$$

Der Modellansatz entspricht dem Ansatz aus Modell 1. Allerdings werden nicht mehr alle Beobachtungen in das Modell aufgenommen. Es wird jeweils eine Regression nur für jene Beobachtungen, die zur jeweiligen Stunde j aufgezeichnet wurden in das Modell aufgenommen. Führen Sie diesen Ansatz für die Stunde 7 und für die Stunde 23 durch. Bevor Sie die Regression durchführen, müssen Sie die Daten nach den jeweiligen Tagesstunden filtern (z.B. mit data=data(data.Stunde=j,:)). Im beigefügten Matlab Skript wird beispielhaft gezeigt, wie Sie aus dem Dataset Data_heat die Daten für eine bestimmte Stunde auswählen können.

a) Vergleichen Sie die beiden Konstanten (β_0) sowie die Koeffizienten des Temperatureinflusses (β_1) aus den Ergebnissen der Regression für Stunde 7 und Stunde 23. Wie interpretieren Sie diese und wie interpretieren Sie die Unterschiede zwischen den beiden Stunden?

b) Vergleichen Sie das jeweilige Bestimmtheitsmaß (R2) aus den beiden Modellen für Stunde 7 und 23 nach Modellansatz 3 mit dem Bestimmtheitsmaß aus Modell 1. Wie würden Sie die Qualität der beiden Modellansätze beurteilen? Woraus ergibt sich der Unterschied?

c) Vergleichen Sie die modellierten Werte für die Beobachtungen für Stunde 7 und für Stunde 23 mit den gemessenen Werten der jeweiligen Stunde über alle 366 Tage. Erstellen Sie dazu eine Grafik. Was beobachten Sie? Wieso schwanken zu einer bestimmten Zeit im Jahr die modellierten Werte um die relativ konstante gemessene Nachfrage? Vergleichen Sie dazu den modellierten Zusammenhang zwischen Temperatur und Nachfrage (Skizze) mit dem Scatterplot Temperatur vs. Nachfrage aus der Angabe. (Scatterplot siehe Matlab Skript oder Folien zur Übungsangabe).

2.4) Zusatzfrage: Verbesserungen des Modells

*(Hier gibt es nur zusätzliche Punkte für die Übung aber **keine Abzüge**, wenn Sie die Frage nicht beantworten. Maximale Punkteanzahl bleibt allerdings 12,5 Punkte auf die gesamte Übung.)*

a) Wie würden Sie vorgehen, um das Problem, das im letzten Punkt von Beispiel 2.3.c beobachtet wurde zu beheben und damit bessere Vorhersagen für diesen Zeitraum zu erhalten?
b) Welche Verbesserungsvorschläge für weitere Modellansätze fallen Ihnen ein? Gehen Sie dabei allerdings weiter davon aus, dass Ihnen nur die gegebenen Daten bzw. allgemein zugängliche Daten zur Verfügung stehen. Formulieren Sie wenn möglich einen verbesserten Modellansatz mathematisch.

Anhang:

Hier wird nur kurz die allgemeine Form eines Regressionsmodells angedeutet und Definitionen für Prüfgrößen des Modells beschrieben. Sie benötigen diese Angaben **nicht** für die Lösung der Übungsaufgabe!!! Sie dienen nur als Hintergrundinformation zu den Funktionen und Ausgabewerten der Regressionsergebnisse in Matlab bzw. anderer Software. Für Interessierte hier noch ein Link zu einem Skriptum des IHS Wien in dem die wichtigsten Begriffe und Herleitungen beschrieben werden:
http://homepage.univie.ac.at/robert.kunst/emwi.pdf

Zudem finden Sie in zahlreichen Online-Tutorials weitere Hintergrundinformationen.

Kurze Formelsammlung:

$$y_t = \beta_1 + \beta_2 \cdot x_{2\,t} + \ldots\ldots + \beta_k \cdot x_{kt} + u_t$$

y_t ..... *Beobachtungen der abhängigen Variable*
β_i ..... *Koeffizienten die geschätzt werden*
x_{jt} ..... *Beobachtungen der Einflussparameter*
u_t ..... *zufälliger Fehler*

$y'y = \hat{y}'\hat{y} + e'e$ (*Quadratsummenzerlegung*)

y ..... *Vektor der beobachteten abhängigen Variable*

$\hat{y}$ ..... *Vektor der durch das Modell geschätzen abhängigen Variable*

e ..... *Fehler des Modells* $\left(y - \hat{y}\right)$
$e'e = RSS$ (*residual sum of squares*)

$y'y = TSS$ (*total sum of squares*)

$\hat{y}'\hat{y} = ESS$(*explainded sum of squares*)

$$F = \frac{\frac{\hat{y}'\hat{y}}{K-1}}{\frac{e'e}{T-K}}$$

F ..... *Prüfgröße F*
K ..... *Anzahl der Freiheitsgrade*(*Koeffizienten inklusive Konstante*)
T ..... *Anzahl der Beobachtungen*

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{e'e}{y'y}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{(T-1)}{(T-K)}$$

R^2 .....*Bestimmheitsmaß*

R^2_{adj} .....*adjustiertes Bestimmtheitsmaß*

Auffinden der Koeffizienten:

Die Lösung für die Koeffizienten β_i ergibt sich aus der Minimierung der Fehlerquadrate. Sie ist hier angedeutet. Für die Herleitung verweisen wir auf die umfangreiche Literatur zu Ökonometrie bzw. Regressionsanalyse.

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1\,k} \\ 1 & x_{22} & \dots & x_{2\,k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix}$$

In Matrixschreibweise:

$$Y = X\beta + u$$

Lösung für Koeffizienten die zu minimalen Fehlerquadraten führen:

$$\hat{\beta} = (X'X)^{-1}X'Y$$