# CAS Datenanalyse April 2016

## Nullserie für Regressions- und Zeitreihenanalyse

Prüfungsdauer: 60 Minuten　　　　　　　　　　　Total Punkte: 60

Prüfungsnote = 1 + (#Punkte/60)5

### Aufgabe 1: Terminologie

1. Erklären Sie, was die **Schiefe** ist. Charakterisieren Sie eine rechtsschiefe Verteilung mittels Mittelwert und Median? (___/4P)

   Die Schiefe gibt die Richtung und Stärke der Asymmetrie einer Wahrscheinlichkeitsverteilung an. Sie zeig, ob und wie stark die Verteilung nach rechts (positive Schiefe) oder nach links (negative Schiefe) geneigt ist.
   Positive Schiefe: der **Median** ist kleiner als das arithmetische Mittel → Median < $\bar{x}$

2. Erklären Sie, was ein **Random Walk** ist. Ist dieser Prozess stationär? Begründen Sie Ihre Antwort. (___/4P)

   Eine Zufallsvariable $X_t$ folgt einem Random Walk wenn

   $X_t = X_{t-1} + U_t$ 　　　　mit $E(u_t) = 0$ und $var(U_t)$ ist konstant

   Der Random Walk ist nicht stationär, da sich der Zufallspfad von seinem Mittelwert beliebig weit entfernen kann.

## Aufgabe 2: Bekanntes Fallbeispiel (___/22P)

Sie wollen den Einfluss der Unternehmensperformance auf das Gehalt der CEOs untersuchen. Sie haben folgende Variablen zur Verfügung:

- SALARY: jährlicher Gehalt in Tausend dollar
- MKTVAL: Börsenkapitalisierung in Mio dollar
- PROFITS: Reingewinn des Unternehmens in Mio dollar
- SALES: Umsatz des Unternehmens in Mio dollar
- CEOTEN: Firmenzugehörigkeit in Jahren im Unternehmen (als CEO und nicht-CEO)
- COMTEN: Firmenzugehörigkeit in Jahren
- Profmarg: Gewinnmarge

1. Die **Standardabweichung** der Variable *SALES* beträgt 6‘088. Interpretieren Sie diese Zahl. (___/2P)

   Die durchschnittliche Abweichung vom Mittelwert beträgt USD 6‘088.7

   Folgendes Regressionsmodell wurde für Sie geschätzt:

   Modell 1: ln(salary) = $\beta_1 + \beta_2$ ln(sales) + $\beta_3$ ln(mktval) + u

```
Abhängige Variable: l_SALARY

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  ------------------------------------------------------------
  const       4,62092       0,254408       18,16      4,95e-042 ***
  l_SALES     0,162128      0,0396703       4,087     6,67e-05  ***
  l_MKTVAL    0,106708      0,0501240       2,129     0,0347    **

Mittel d. abh. Var.      6,582848   Stdabw. d. abh. Var.    0,606059
Summe d. quad. Res.      45,30965   Stdfehler d. Regress.   0,510294
R-Quadrat                0,299114   Korrigiertes R-Quadrat  0,291057
F(2, 174)                37,12853   P-Wert(F)               3,73e-14
Log-Likelihood          -130,5594   Akaike-Kriterium        267,1188
Schwarz-Kriterium        276,6472   Hannan-Quinn-Kriterium  270,9832
```

2. Interpretieren Sie die drei geschätzten Regressionskoeffizienten. (___/6P)

   Das zu erwartende *l_salary* eines Unternehmens mit \$1. Mio. Umsatz und einer Börsenkapitalisierung von \$1 Mio. liegt bei 4.62.

   $e^{b1}$ = 101.49 ≈ 100 ist dann (ungefähr) das zu erwartende CEO-Gehalt bei einem solchen Unternehmen.

   Interpretation: Das CEO-Gehalt eines Unternehmens mit \$1 Mio. Jahresumsatz und einem Marktwert von \$1 Mio. liegt bei ca. \$100‘000.

   $b_2$ = 0.16: Mit einer Umsatzerhöhung von 1% ist eine durchschnittliche Erhöhung des CEO-Gehaltes um ca. 0.16% zu erwarten, ceteris paribus.

   $b_3$ = 0.11: Eine Erhöhung des Unternehmensmarktwertes (Börsenkapitalisierung) um 1% bewirkt eine durchschnittliche Erhöhung des CEO-Gehaltes um ca. 0.11%, ceteris paribus.

Folgendes Regressionsmodell wurde für Sie geschätzt.

Modell 2: $\ln(salary) = \beta_1 + \beta_2 \ln(sales) + \beta_3 \ln(mktval) + \beta_4 profits + u$

```
Abhängige Variable: l_SALARY

             Koeffizient    Std.-fehler    t-Quotient    p-Wert
  -------------------------------------------------------------
  const       4,68692       0,379729        12,34       1,65e-025 ***
  l_SALES     0,161368      0,0399101
  l_MKTVAL    0,0975286     0,0636886        1,531      0,1275
  PROFITS     3,56601e-05   0,000151960      0,2347     0,8147

Mittel d. abh. Var.      6,582848   Stdabw. d. abh. Var.    0,606059
Summe d. quad. Res.     45,29524   Stdfehler d. Regress.   0,511686
R-Quadrat                0,299337   Korrigiertes R-Quadrat  0,287186
F(3, 173)               24,63629   P-Wert(F)               2,53e-13
Log-Likelihood        -130,5312   Akaike-Kriterium        269,0625
Schwarz-Kriterium      281,7671   Hannan-Quinn-Kriterium  274,2150
```

3. Sind die Variablen *profits* und *l_mktval* individuell signifikant auf dem 5%-Signifikanzniveau? Begründen Sie Ihre Antwort mittels **p-Wert**. (___/2P)

   Die erklärenden Variablen *mktval* und *profits* sind nicht signifikant auf dem 5%-Signifikanzniveau da die p-Werte > 5% $\rightarrow H_0$ kann nicht verworfen werden.

4. Ist die Variable *l_sales* statistisch signifikant auf dem 5%-Signifikanzniveau? Leider wurden der t-Quotient und p-Wert gelöscht. Führen Sie einen t-Test durch und wenden Sie die Faustregel ($t_c = 2$) an. (___/2P)

   t-Wert = 0.1613/0.0399 = 4.043 > 2 $\rightarrow H_0$ kann verworfen werden

   → *l_sales* ist statistisch signifikant!

5. Interpretieren Sie den geschätzten Koeffizienten von *profits*. (___/2P)

   Interpretation als Semi-Elastizität: Mit einer Gewinnerhöhung um $1 Mio. (=1 Einheit) steigt das durchschnittliche CEO-Gehalt um $3.566 \cdot 10^{-5} \times 100 = 0.356\%$, ceteris paribus.

6. Warum könnte es dennoch Sinn machen, beide Variablen *mktval* und *profits* in die Regression aufzunehmen? (___/6P)

   **Kontrollvariablen-Aspekt**: Falls man der Effekt von *profits* analysieren möchte, sollte die Börsenkapitalisierung (*mktval*) kontrolliert werden, d.h. es soll der vom Effekt von *mktval* bereinigte Effekt von *profits* auf *l_salary* ermittelt werden.

   **Omitted-Variable Aspekt**: In der Regression ohne *profits* ist im Koeffizienten von *mktval* eigentlich der indirekte Effekt von *profits* vorhanden. Die Variable *profits* ist einerseits mit *mktval* stark positiv korreliert und hat andererseits für sich genommen einen positiven Effekt auf die Variable *salary* (Gehalt).

   **Variablengruppe-Aspekt**: Beide Variablen *mktval* und *profits* können als Variablengruppe betrachtet werden, die für "kapitalmarktorientierte Performance-Masse" steht. Beide Variablen messen zwar Ähnliches, aber jede für sich genommen hebt doch andere Aspekte hervor: Die Börsenkapitalisierung ist zukunftsorientiert und hängt sehr stark vom antizipierten zukünftigen Wachstumspotenzial des Unternehmens ab. Der Reingewinn hingegen ist vergangenheitsorientiert und spiegelt die kurzfristige vergangene Entwicklung des Unternehmens wieder.

7. Welches Regressionsmodell würden Sie vorziehen? Begründen Sie Ihre Antwort. (___/4P)

**Zusammenstellung der Modelle:**

Modell 1: ln(salary) = 4.621 + 0.162 ln(sales) + 0.107 ln(mktval)

Modell 2: ln(salary) = 4.687 + 0.161 ln(sales) + 0.0975 ln(mktval) + 0.0000357 profits

Modell 3: ln(salary) = 4.558 + 0.162 ln(sales) + 0.1018 ln(mktval) + 0.000029 profits + 0.0117ceoten

Modell 4: ln(salary) = 4.441 + 0.164 ln(sales) + 0.0984 ln(mktval) + 0.000039 profits + 0.0452ceoten -0.00121ceoten$^2$

Modell 5: ln(salary) = 4.621 + 0.158 ln(sales) + 0.112 ln(mktval) – 0.00226 profmarg

Modell 6: ln(salary) = 4.438 + 0.187 ln(sales) + 0.1013 ln(mktval) – 0.0026 profmarg + 0.048ceoten - 0.00114ceoten$^2$ – 0.008498 comten

Modell 7: ln(salary) = 4.424 + 0.186 ln(sales) + 0.1018 ln(mktval) – 0.0026 profmarg + 0.0477ceoten - 0.00112ceoten$^2$ – 0.006063 comten – 0.000054 comten$^2$

| | Modell 1 | Modell 2 | Modell 3 | Modell 4 | Modell 5 | **Modell 6** | Modell 7 |
|---|---|---|---|---|---|---|---|
| # Regressor | 3 | 4 | 5 | 6 | 4 | **7** | 8 |
| adj. $R^2$ | 0.291 | 0.2872 | 0.302 | 0.324 | 0.291 | **0.3522** | 0.3486 |
| Akaike | 267.12 | 269.06 | 266.21 | 261.61 | 268.01 | **255.03** | 256.98 |

Modell 6 weist das höchste adjustierte $R^2$ und das kleinste Akaike-Informationskriterium auf.
→ Modell 6 ist vorzuziehen.

Bei diesem Modell sind alle Koeffizienten ausser *profmarg* individuell statistisch signifikant. Es ist dennoch sinnvoll eine Rentabilitätskennzahl wie die Gewinnmarge aufzunehmen, wegen dem sogenannten Omitted-Variable- und Kontrollaspekt.

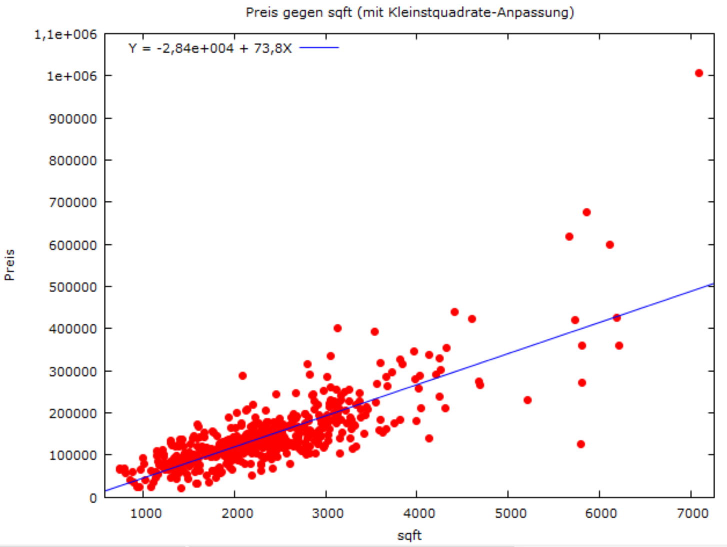
**Aufgabe 3: Unbekanntes Fallbeispiel** (___/18P)

Sie wollen die Bestimmungsfaktoren für den Hauspreis statistisch analysieren. Sie sammeln Daten über den Verkaufspreis, die Fläche der Häuser in Quadratfuss (square feet) und deren Alter.

Folgende Variablen stehen zur Verfügung:

- Preis: Hauspreis
- sqft: Wohnfläche in Quadratfuss
- age: Alter des Hauses

1. Betrachten Sie folgendes Streudiagramm vom Hauspreis gegen Hausfläche (SQFT) für traditionelle Häuser. Was stellen Sie fest?




Bei zunehmender Wohnfläche steigt die Streuung des Hauspreises. Das stellt eine Verletzung gegen die Konstanz der Varianz dar.

Folgendes Modell wurde für Sie geschätzt. Preis = $\beta_1 + \beta_2 SQFT + \beta_3 AGE + u$

```
Modell 4: KQ, benutze die Beobachtungen 1-1080
Abhängige Variable: Preis

             Koeffizient   Std.-fehler   t-Quotient    p-Wert
  -------------------------------------------------------------
  const      -41947,7      6989,64         -6,001      2,67e-09  ***
  sqft           90,9698      2,40310      37,86       4,26e-200 ***
  age          -755,041     140,894        -5,359      1,02e-07  ***

Mittel d. abh. Var.      154863,2   Stdabw. d. abh. Var.     122912,8
Summe d. quad. Res.      6,69e+12   Stdfehler d. Regress.    78814,86
R-Quadrat                0,589592   Korrigiertes R-Quadrat   0,588829
F(2, 1077)               773,6077   P-Wert(F)                5,2e-209
Log-Likelihood          -13707,80   Akaike-Kriterium         27421,59
Schwarz-Kriterium        27436,55   Hannan-Quinn-Kriterium   27427,26
```

2. Interpretieren Sie die Regressionskoeffizienten $b_2$ und $b_3$

   $b_2 = 90.97$ → Für ein gegebenes Hausalter bewirkt eine Flächenerhöhung um 1 Quadratfuss einen Anstieg des Hauspreises um USD90.97, ceteris paribus.

   $b_3 = -755$ → Für eine gegebene Wohnfläche bewirkt eine Erhöhung des Hausalters um 1 Jahr eine durchschnittliche Preisreduktion um USD 755, ceteris paribus.

3. Erstellen Sie ein 95%-Konfidenzintervall für den Parameter $b_2$. Der kritische t-Wert beträgt $t_c(0.975, 1077) = 1.96$. Wie viele Beobachtungen liegen zugrunde?

   95%-Konfidenzintervall für $b_2 = 90.97$

   Anzahl Beobachtungen = N      N-K =1077 → N = 1077 + 3 = 1080

   95%-Konfidenzintervall: $b_2 \pm se(b_2) = 90.97 \pm 1.96 \times 2.4 = [86.26, 95.67]$

4. Interpretieren Sie **konkret** Ihr 95%-Konfidenzintervall

   Bei Wiederholung des Experimentes, würde in 95% der Fälle der wahre Parameter $\beta_2$ vom berechneten Konfidenzintervall überdeckt werden.

5. Testen Sie folgende Vermutung: Wenn ein Haus um 1 Jahr altert, sinkt dessen Preis (P) um weniger als USD 1000.
   Wie lautet Ihre Schlussfolgerung mittels p-Wert?
   Der kritische Wert für das 5%-Signifikanzniveau beträgt $t_c(0.95, 1077) = 1.65$

   Die Vermutung gehört zur Alternativhypothese

   $H_0$: $b_3 \leq -1000$      $H_1$: $b_3 > -1000$

   c = -1000 = Konstante, die wir testen wollen

$$t_e = \frac{b_3 - c}{se(b_3)} = \frac{-755.04 + 1000}{140.89} = 1.74$$

   *Hinweis: Es reicht, wenn 2 Kommastellen für die Berechnung genommen werden! Runden Sie Ihre Endergebnisse auf.*

   $t_e > 1.65 \rightarrow H_0$ kann verworfen werden.

   Interpretation: Wenn das Haus um 1 Jahr altert, reduziert sich der Hauspreis um weniger als \$1000.

## Aufgabe 4: Zeitreihenanalyse (___/12P)

Das Holt-Verfahren mit $\alpha = 0.4$ und $\gamma = 0.6$ wurde angewandt.

Niveaugleichung: $L_t = \alpha y_t + (1-\alpha)(L_{t-1} + b_{t-1})$

Trendgleichung: $b_t = \gamma(L_t - L_{t-1}) + (1-\gamma)b_{t-1}$

$\hat{y}_{19}(18)$ : Prognosewert zum Zeitpunkt 18 für Periode 19

Sie bekommen folgende Werte:

| $y_{20} - \hat{y}_{20}(19)$ | $\hat{y}_{19}(18)$ | $b_{18}$ | $L_{19}$ |
|---|---|---|---|
| 24 | 1980 | 1200 | 700 |

1. Welche Gleichung beschreibt am besten den zugrunde liegenden, datenerzeugenden Prozess nach dem Holt-Verfahren?

   Datenerzeugender Prozess: $y_t = \beta_1 + \beta t + e$ Zeitreihe mit Trend!

   mit $\beta_1$ und $\beta_2$ sich langsam verändernden Parameter.

2. Bestimmen Sie den Wert von **$L_{18}$**?

   $L_{18} = \alpha y_{18} + (1-\alpha)(L_{17} + b_{17})$

   $\hat{y}_{19}(18) = L_{18} + b_{18}$ $\Leftrightarrow$ **$L_{18}$** $= \hat{y}_{19}(18) - b_{18} = 1980 – 1200 = 780$

3. Bestimmen Sie den Wert von **$b_{19}$**

   **$b_{19}$** $= 0.6(L_{19} - L_{18}) + 0.4b_{18} = 0.6(700 – 780) + 0.4(1200) = 432$

4. Bestimmen Sie den Wert von **$L_{20}$**

   $y_{20} - \hat{y}_{20}(19) = y_{20} - (L_{19} + b_{19}) = 24$

   $y_{20} = 24 + L_{19} + b_{19} = 24 + 700 + 432 = 1'156$

   **$L_{20}$** $= \alpha y_{20} + (1-\alpha)(L_{19} + b_{19}) = 0.4(1156) + 0.6(700 + 432) = 1141.6$

5. Schreiben Sie die Gleichung der einfachen exponentiellen Glättung in Fehlerkorrekturform.

   Exponentielle Glättung: $\hat{y}_{t+1}(\mathrm{t}) = \alpha y_t + (1-\alpha)\hat{\mathrm{y}}_{\mathrm{t}}(\mathrm{t}-1)$

   Fehlerkorrekturform: $\hat{y}_{t+1}(\mathrm{t}) = \hat{\mathrm{y}}_{\mathrm{t}}(t-1) + \alpha e_t$ →wobei $e_t = y_{t+1} - \hat{\mathrm{y}}_{\mathrm{t}}(t-1)$

6. Was ist im Allgemeinen die Folge, wenn die einfache exponentielle Glättung für Daten mit einem abnehmenden Trend angewendet wird?

   Der Prognosewert wird die laufende Beobachtung im Allgemeinen überschätzen.

   im Allgemeinen $e_t = y_{t+1} - \hat{\mathrm{y}}_{\mathrm{t}}(t-1) < 0$