

SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2024-12-09

Skrivtid: kl 10-15 (5 timmar)

Godkända hjälpmedel: Miniräknare utan lagrade formler och text.

Bifogade hjälpmedel: Formel- och tabellsamling för statistik och dataanalys II, 15 hp.

Tentamen består av 5 uppgifter uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas för fulla poäng.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
- Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan ge poängavdrag.
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningar i de efterföljande uppgifterna.
- I beräkningar från R-utskrifter får du utgå från det som är givet.

Tentamen kan ge maximalt 100 poäng. För godkänt resultat krävs minst 50 poäng.

Betygsgränser:

A	90-100
B	80-89
C	70-79
D	60-69
E	50-59
Fx	40-49
F	0-40

Obs! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Lösningsförslag läggs ut på Athena efter att tentamenstiden är över.

Lycka till!

Uppgift 1 (20p)

Datafilen **valresultat** visar väljarstödet i Sveriges 290 kommuner vid ett riksdagsval. Följande resultat erhålles vid en analys av det procentuella väljarstödet för Moderaterna (M) som responsvariabel, och med kommunens medelinkomst i kronor, om kommunen är en högskoleort (1 = ja, 0 = nej) samt interaktionen mellan medelinkomst och högskoleort som förklaringsvariabler. Observera att vissa delar av utskriften censurerats.

Call:

```
lm(formula = M ~ Inkomst + Högskoleort + Inkomst * Högskoleort,  
    data = valresultat)
```

Residuals:

Min	1Q	Median	3Q	Max
-21.7345	-3.0859	-0.0866	3.2658	17.7304

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.407e+01	2.934e+00	-11.609	<2e-16 ***
Inkomst	2.224e-04	1.199e-05		<2e-16 ***
Högskoleort	3.581e+01	1.776e+01	2.016	0.0447 *
Inkomst:Högskoleort	-1.405e-04	7.084e-05	-1.983	0.0483 *

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

Residual standard error: 4.834 on 286 degrees of freedom

Multiple R-squared: 0.5508,

F-statistic: 116.9 on 3 and 286 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: M

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Inkomst	1	8091.5	8091.5	346.2615	< 2e-16 ***
Högskoleort	1	10.3	10.3	0.4387	0.50828
Inkomst:Högskoleort	1	91.9	91.9	3.9336	0.04829 *
Residuals	286	6683.3	23.4		

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

- a) Finns det ett signifikant samband mellan inkomst och väljarstöd för M? Formulera hypoteser, beräkna testvariabeln och presentera din slutsats. Välj 5% signifikansnivå. (5p)

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$t = \frac{2.224e - 04}{1.199e - 05} = 18.5$$

$$p < 2e - 16$$

- b) Tolka med ord effekten av interaktionstermen på väljarstödet för M. (5p)

Ej högskoleort: $M = -3.407 + 2.224e - 04 \cdot Inkomst + 3.581 \cdot 0 - 1.405e - 04 \cdot Inkomst \cdot 0 = -3.407 + 2.224e - 04$

Högskoleort: $M = -3.407 + 2.224e - 04 \cdot Inkomst + 3.581 \cdot 1 - 1.405e - 04 \cdot Inkomst \cdot 1 = (-3.407 + 3.581) + (2.224e - 04 - 1.405e - 04) \cdot Inkomst$

- c) Beräkna modellens justerade förklaringsgrad och redogör för vad den används till. (5p)

$$\begin{aligned} R_{adj}^2 &= 1 - \frac{n - 1}{n - (p + 1)} \cdot \frac{SSE}{SST} \\ &= 1 - \frac{290 - 1}{290 - (3 + 1)} \cdot \frac{6683.3}{8091.5 + 10.3 + 91.9 + 6683.3} = 0.546 \end{aligned}$$

- d) Beräkna en punktskattning av det förväntade väljarstödet för M i en högskoleort med medelinkomst 300 000 kr. (5p)

$$\hat{M} = -3.407 + 2.224e - 04 \cdot 300000 + 3.581 \cdot 1 - 1.405e - 04 \cdot 300000 \cdot 1 = 26.31\%$$

Uppgift 2 (20p)

Vi fortsätter att arbeta med samma data som i uppgift 1, och lägger dessutom till andelen pensionärer (i procent) och andelen arbetslösa (i procent) som förklaringsvariabler.

Residuals:

Min	1Q	Median	3Q	Max
-21.4581	-2.8152	-0.0333	2.7301	18.4017

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.406e+01	6.890e+00	-2.041	0.04220	*
Inkomst	1.686e-04	1.834e-05	9.190	< 2e-16	***
Högskoleort	4.484e+01	1.728e+01	2.594	0.00998	**
Pensionärer	-6.804e-02	9.250e-02	-0.736	0.46260	
Arbetslöshet	-9.802e-01	2.027e-01	-4.836	2.18e-06	***
Inkomst:Högskoleort	-1.751e-04	6.878e-05	-2.545	0.01145	*

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

Residual standard error: 4.661 on 284 degrees of freedom

Multiple R-squared: 0.5853, Adjusted R-squared: 0.578

F-statistic: 80.17 on 5 and 284 DF, p-value: < 2.2e-16

- a) Jämför denna modell med modellen från uppgift 1 med F-test. Formulera hypoteser, beräkna testvariabeln och presentera din slutsats. Välj 5% signifikansnivå. (10p)

$$H_0: \beta_3 = \beta_4 = 0$$
$$H_a: \text{ej både } \beta_3 \text{ och } \beta_4 \text{ är } 0$$

$$F = \frac{(R_{FM}^2 - R_{RM}^2)/(p - q)}{(1 - R_{FM}^2)/(n - p - 1)} = \frac{(0.5853 - 0.5508)/(5 - 3)}{(1 - 0.5853)/(290 - 5 - 1)} = 11.81$$

$$F_{\alpha=0.05}(p - q = 5 - 3 = 2, n - p - 1 = 290 - 5 - 1 = 284) = 3.03$$

H_0 förkastas

- b) Om du skulle förbättra modellen från uppgift 2, vilket är det första steg du skulle ta? Beskriv med ord. (5p)

Variabeln *Pensionärer* är ej signifikant. Träna om modellen utan denna variabel och bedöm modellanpassningen.

- c) Vilka antaganden vilar den linjära regressionsmodellen på? (5p)

- Konstant varians bland modellens feltermmer (homoscedasticitet)
- Normalfördelade feltermmer
- Oberoende feltermmer

Uppgift 3 (20p)

Vi fortsätter arbeta med datafilen **valresultat** och definierar nu som responsvariabel om det råder borgerlig majoritet i kommunen (kodat som 1) eller ej (kodat som 0).
Förklaringsvariabler är kommunens medelinkomst i kronor och andelen pensionärer i procent.

Call:

```
glm(formula = borgerlig_majoritet ~ Inkomst + Pensionärer, family = binomial,
     data = valresultat)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.727e+01	1.232e+01	-4.647	3.37e-06	***
Inkomst	1.527e-04	3.095e-05	4.935	8.01e-07	***
Pensionärer	5.194e-01	1.542e-01	3.368	0.000758	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 145.554 on 289 degrees of freedom
Residual deviance: 57.378 on 287 degrees of freedom
AIC: 63.378

Number of Fisher Scoring iterations: 8

Total Observations in Table: 290

	valresultat\$predicerade		
valresultat\$borgerlig_majoritet	0	1	Row Total
0	266	4	270
1	9	11	20
Column Total	275	15	290

- a) Tolka parameterskattningen för **Inkomst** i termer av en oddskvot. (5p)

$$e^{1.527e-04} = 1.0002$$

För varje kronas ökning i medelinkomst i kommunen, så ökar oddset för borgerlig majoritet med 0.02 procentenheter.

- b) Beräkna modellens accuracy. (3p)

$$\text{accuracy} = (\text{sann positiv} + \text{sann negativ}) / \text{stickprovsstorlek} = (11 + 266) / 290 = 0.955$$

c) Beräkna modellens sensitivitet. (3p)

$$\textbf{sensitivitet} = (\text{sann positiv}) / (\text{sann positiv} + \text{falsk negativ}) = 11 / (11 + 9) = 0.550$$

d) Beräkna modellens specificitet. (3p)

$$\textbf{specificitet} = (\text{sann negativ}) / (\text{falsk positiv} + \text{sann negativ}) = 266 / (4 + 266) = 0.985$$

e) Predicera sannolikheten för borgerlig majoritet i en kommun med medelinkomst 300 000 kronor och där andelen pensionärer är 25%. (6p)

$$P(y = 1) = \frac{1}{1 + e^{-(-5.727e+01 + (1.527e-04) * 300000 + (5.194e-01) * 25)}} = 0.821$$

Uppgift 4 (20p)

I kontrollerad miljö studeras tillväxten hos en viss fiskart över tid. I följande analys är responsvariabel fiskens längd i millimeter (**length**) och som förklaringsvariabel fiskens ålder i år (**age**) samt kvadraten på åldern (**age_sq**).

Call:

```
lm(formula = length ~ age + age_sq, data = bluegills)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.846	-8.321	-1.137	6.698	22.098

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.622	11.016	1.237	0.22
age	54.049	6.489	8.330	2.81e-12 ***
age_sq	-4.719	0.944	-4.999	3.67e-06 ***

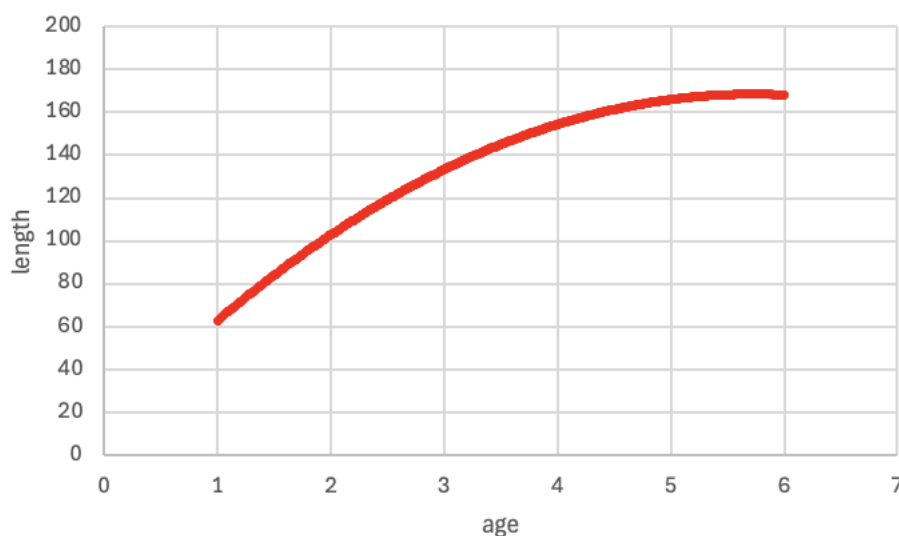
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.91 on 75 degrees of freedom

Multiple R-squared: 0.8011, Adjusted R-squared: 0.7958

F-statistic: 151.1 on 2 and 75 DF, p-value: < 2.2e-16

- a) Skissa, med hjälp av parameterskattningarna, formen på den anpassade regressionslinjen. (5p)



- b) Tolka effekten av fiskens ålder på dess längd. (5p)

Uttrycket för fiskens ålder på dess längd ges av

$$\frac{dy}{dx} = 54.049 + 2 \cdot (-4.719) \cdot x$$

- c) Beräkna en punktskattning av den förväntade längden hos en tre år gammal fisk. (5p)

$$\hat{y} = 13.622 + 54.049 \cdot 3 - 4.719 \cdot 3^2 = 133.298$$

- d) Vi noterar att förklaringsgraden är förhållandevis hög. Beskriv vad överanpassning innebär och redogör för en strategi för att hantera detta problem. (5p)

Överanpassning inträffar när en statistisk modell anpassas så väl till träningsdata att den fångar upp inte bara de underliggande mönstren utan även slumpmässigt brus eller irrelevanta detaljer. Detta leder till att modellen presterar väldigt bra på träningsdata men generaliserar dåligt till nya, osedda data.

Uppgift 5 (20p)

Vi studerar det månatliga genomsnittliga bensinpriset i kronor per liter från januari 2010 till och med september 2024. En regressionsmodell med tidsindex (som hade värdet 1 i januari 2010 och 177 i september 2024) och indikatorvariabler för månader (med december som referenskategori) anpassas till data:

Residuals:

Min	1Q	Median	3Q	Max
-4.2713	-1.2052	0.0707	0.9960	5.6523

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.831382	0.497814	23.767	<2e-16 ***
index	0.034016	0.002457	13.843	<2e-16 ***
månadapr	0.821842	0.620143	1.325	0.187
månadaug	0.625777	0.620143	1.009	0.314
månadfeb	0.449875	0.620202	0.725	0.469
månadjan	0.283891	0.620246	0.458	0.648
månadjul	0.826460	0.620129	1.333	0.184
månadjun	0.913810	0.620124	1.474	0.143
månadmaj	0.841159	0.620129	1.356	0.177
månadmar	0.675859	0.620168	1.090	0.277
månadnov	0.319731	0.630730	0.507	0.613
månadokt	0.553747	0.630744	0.878	0.381
månadsep	0.465094	0.620168	0.750	0.454

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

Residual standard error: 1.669 on 164 degrees of freedom

Multiple R-squared: 0.5454, Adjusted R-squared: 0.5121

F-statistic: 16.4 on 12 and 164 DF, p-value: < 2.2e-16

lag Autocorrelation D-W Statistic p-value

1	0.935803	0.1226884	0
---	----------	-----------	---

Alternative hypothesis: rho != 0

a) Beräkna en punktskattning av bensinpriset i december 2024. (5p)

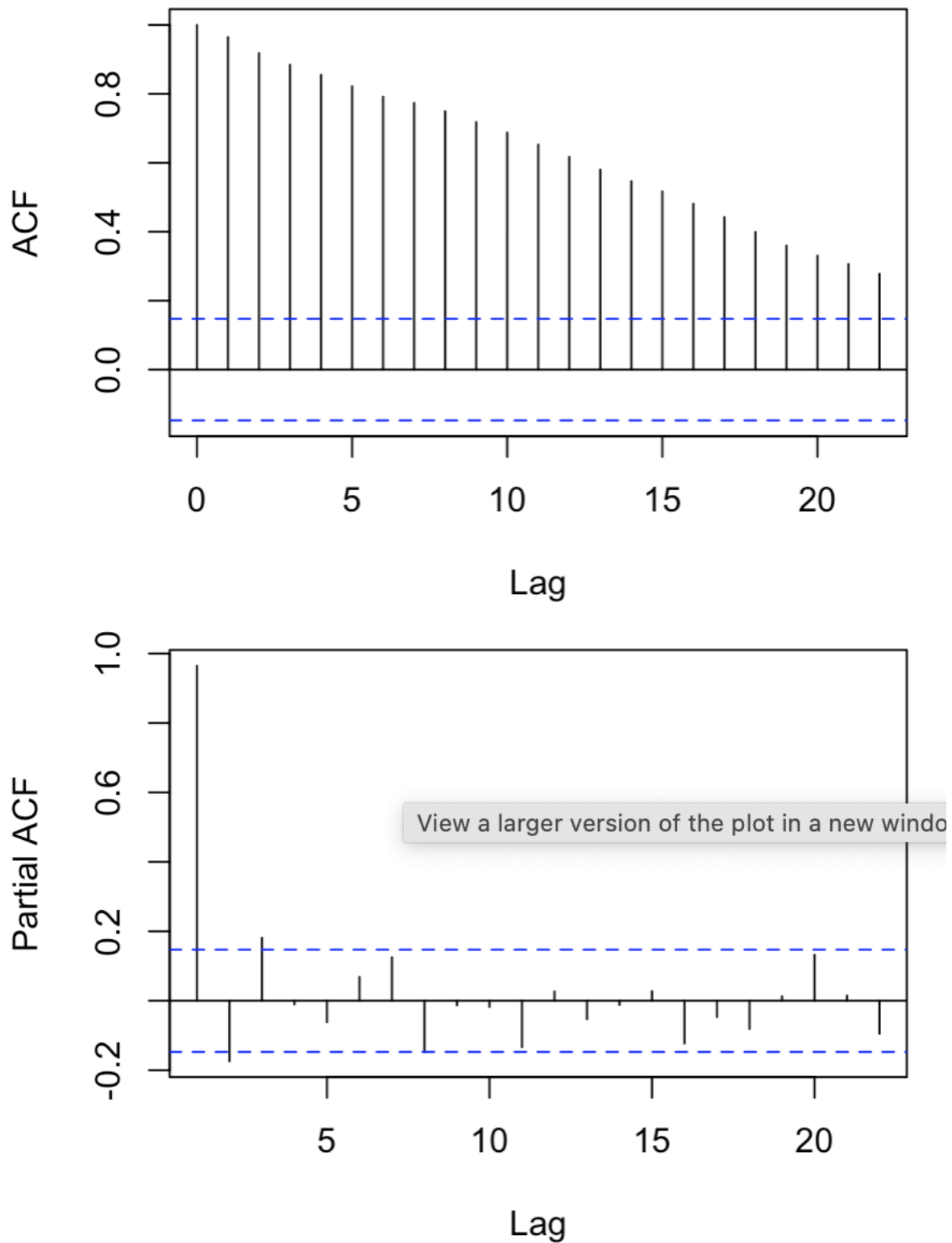
December 2024 innebär tidsindex 180.

$$\hat{y} = 11.831382 + 0.034016 \cdot 180 = 17.95$$

- b) Föreligger autokorrelation bland modellens residualer? Presentera relevant information från utskriften och tolka. (5p)

Ja, Durbin-Watsonstatistikan är under 1 vilket indikerar positiv autokorrelation bland modellens residualer.

Följande figurer visar ACF och PACF för literpriset bensin.



c) Förefaller tidsserien stationär? Motivera. (5p)

Nej, mönstret i ACF är alldeles för långsamt avtagande.

Slutligen följer resultatet av en ARIMA-modellering av bensinpriset:

```
Series: bensinpris$kr_per_liter  
ARIMA(2,1,2)
```

Coefficients:

	ar1	ar2	ma1	ma2
	-0.3459	-0.8115	0.5261	0.7523
s.e.	0.1013	0.1963	0.0934	0.2290

```
sigma^2 = 0.319: log likelihood = -147.38  
AIC=304.75 AICc=305.11 BIC=320.6
```

d) Vad innebär termerna 2, 1 och 2 i den modell som valts? (5p)

Att data differentierats i ett steg och att modellen är en AR av andra ordningen och MA av andra ordningen.