

SDAI (ST1201), Tentamen 2, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2025-01-23

Skrivtid: 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)

Vi studerar som responsvariabel resultatet på en tentamen (antal poäng). Förklaringsvariabler är det genomsnittliga antalet studietimmar per vecka och vid vilket campus studenten varit registrerad (Västerås och Eskilstuna), samt interaktionen mellan studietid och campus. Följande något censurerade resultat erhålles.

Residuals:

Min	1Q	Median	3Q	Max
-19.9167	-5.0441	0.2709	4.3048	23.4532

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.08067	1.28161	1.623	0.105
timmar_per_vecka	0.73034	0.02615	27.931	<2e-16 ***
campusVästerås	0.25757	1.77594	0.145	0.885
timmar_per_vecka:campusVästerås	-0.04027	0.03706	-1.087	0.278

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

Residual standard error: 7.162 on 494 degrees of freedom

Multiple R-squared: Adjusted R-squared: 0.7501

F-statistic: 498.4 on 3 and 494 DF, p-value:

Analysis of Variance Table

Response: resultat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
timmar_per_vecka	1	76329	76329	1488.2558	< 2e-16 ***
campus	1	294	294	5.7412	0.01694 *
timmar_per_vecka:campus	1	61	61	1.1809	0.27770
Residuals	494	25336	51		

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

- a) Testa om samtliga modellens förklaringsvariabler saknar effekt på responsvariabeln. Formulera hypoteser, ange testvariabeln, ange kritiskt värde och slutsats. Välj 5% signifikansnivå. (5p)

$$H_0: \beta_1 = \beta_2 = \beta_3$$
$$H_a: \text{ej alla } \beta \text{ är noll}$$

$$F = \frac{MSR}{MSE} = \frac{(76329 + 294 + 61)/3}{25336/494} = 498.4$$

$$F_{\alpha=0.05}(p=3, n-p-1=494) = 2.63$$

H_0 förkastas

b) Beräkna modellens förklaringsgrad. (5p)

$$R^2 = \frac{SSR}{SST} = \frac{76329 + 294 + 61}{76329 + 294 + 61 + 25336} = 0.75$$

c) Formulera regressionsekvationen för en student från Västerås. (5p)

$$\begin{aligned} resultat &= 2.08067 + 0.73034 \cdot timmar + 0.25757 \cdot 1 + (-0.04027) \cdot timmar \cdot 1 \rightarrow \\ resultat &= 2.33824 + 0.69007 \cdot timmar \end{aligned}$$

d) Predicera det förväntade tentamensresultatet för en Eskilstunastudent som studerar 38 timmar i veckan. (5p)

$$\widehat{resultat} = 2.08067 + 0.73034 \cdot 38 + 0.25757 \cdot 0 + (-0.04027) \cdot timmar \cdot 0 = 29.83$$

Uppgift 2 (20p)

Vi fortsätter arbeta med samma data och bygger en ny modell med responsvariabel tentamensresultat, och som förklaringsvariabler studietid samt ålder.

Residuals:

Min	1Q	Median	3Q	Max
-19.1648	-4.6104	0.1407	4.6580	23.5409

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.95023	2.42857	5.744	1.61e-08 ***
timmar_per_vecka	0.67012	0.01990	33.680	< 2e-16 ***
ålder	-0.36460	0.06975	-5.227	***

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

Residual standard error: 7.013 on 495 degrees of freedom

Multiple R-squared: 0.7613, Adjusted R-squared: 0.7604

F-statistic: 789.6 on 2 and 495 DF, p-value: < 2.2e-16

```
> vif(modell2)
```

timmar_per_vecka	ålder
1.207622	1.207622

a) Här följer fem påståenden baserat på regressionsutskriften. Vilket av dem är sant? (5p)

A	Koefficienten för timmar_per_vecka är signifikant positiv med en uppskattad effekt på 0.67012, och p-värdet är mindre än 2×10^{-16}
B	Variabeln ålder har en positiv koefficient, vilket innebär att när åldern ökar, ökar också den beroende variabelns värde
C	Det justerade R^2 -värdet är 0.8614, vilket tyder på att modellen förklarar 86% av variationen i data
D	F-värdet är 45.67, vilket innebär att modellen inte är särskilt signifikant
E	Residualernas standardfel är 3.501, vilket tyder på mycket liten spridning i residualerna

Svar A

b) Tolka effekten av studentens ålder på tentamensresultatet med ord. (5p)

För varje ytterligare år äldre studenten är, så förväntas tentamensresultatet vara i genomsnitt 0.36 poäng lägre.

- c) Finns några tecken på multikollinearitet? Motivera. (5p)

VIF-värdena indikerar inga tecken på multikollinearitet.

- d) Förklara skillnaden på förklaringsgrad och justerad förklaringsgrad, och när respektive mått används. (5p)

Förklaringsgraden anger hur stor andel av variationen i responsvariabeln som modellen förklarar. Den justerade förklaringsgraden innehåller en bestraffning för modellens komplexitet och lämpar sig för modelljämförelse.

Uppgift 3 (20p)

Vi definierar en ny responsvariabel som 1 om studenten klarade tentamen och 0 annars. Följande resultat erhålles vid analys med studietid per vecka och studentens ålder som förklaringsvariabler.

Call:

```
glm(formula = passed ~ timmar_per_vecka + ålder, family = binomial,
     data = studietid_och_resultat)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.69900	1.10964	-2.432	0.015	*
timmar_per_vecka	0.15803	0.01405	11.252	< 2e-16	***
ålder	-0.15514	0.03555	-4.364	1.28e-05	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 690.25 on 497 degrees of freedom
Residual deviance: 295.93 on 495 degrees of freedom
AIC: 301.93

Number of Fisher Scoring iterations: 6

- a) Tolka koefficienten för variabeln *timmar_per_vecka* som en oddskvot. (5p)

$\exp(0.15803) = 1.17$. För varje ytterligare timme per vecka man studerar ökar oddset för att klara tentamen med 1.17.

- b) Beräkna sannolikheten för att en student som studerar 38 timmar per vecka och är 25 år gammal ska klara tentamen. (5p)

$$P(y = 1) = \frac{1}{1 + e^{-(-2.6990 + 0.15803 \cdot 38 - 0.15514 \cdot 25)}} = 0.36$$

- c) Beskriv en strategi för att utvärdera hur välanpassad en binär logistisk regressionsmodell är. (5p)

Skapa confusion matrix och beräkna accuracy, sensitivitet och specificitet.

- d) Ange vilken typ av test som lämpar sig för att jämföra två logistiska regressionsmodeller, och förklara kortfattat hur testet fungerar. (5p)

Likelihoodkvottestet för att jämföra logistiska regressionsmodeller används för att testa om en mer komplex modell förbättrar förklaringskraften jämfört med en enklare modell.

Uppgift 4 (20p)

Vid ett universitetskafé genomförs en studie av hur priset på en kopp kaffe påverkar efterfrågan. En elasticitetsmodell ansätts, där Q är konsumtionen av kaffe i hundratals koppar och P är priset i kronor. $\log$ i modellutskriften avser tiologaritmen. Stickprovsstorleken är 14.

Call:

```
lm(formula = log_Q ~ log_P, data = kaffe)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.47450	-0.15041	-0.04136	0.38502	0.61558

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.0266	0.9721	4.142	0.00116	**
log_P	-2.2442	0.7045	-3.185	0.00717	**

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

Residual standard error: 0.5146 on 13 degrees of freedom

(5 observations deleted due to missingness)

Multiple R-squared: 0.4384, Adjusted R-squared: 0.3952

F-statistic: 10.15 on 1 and 13 DF, p-value: 0.007165

- a) Uttryck ekvationen för den modell som anpassats innan den logaritmerades. (5p)

$$Q = A \cdot P^{E_P}$$

- b) Tolka elasticitetsskattningen med ord. (5p)

Priselastisiteten för kaffe är -2.2. När priselastisiteten är mindre än -1 betraktas en vara som priskänslig.

- c) Pröva på 5% signifikansnivå om priselastisiteten för kaffe är signifikant högre än -1. Formulera hypoteser, ange testvariabel och kritiskt värde samt din slutsats. Välj 5% signifikansnivå. (10p)

$$H_0: E_p \leq -1$$

$$H_a: E_p > -1$$

$$t = \frac{-2.2442 - (-1)}{0.7045} = -1.77$$

$t_{14-2=12;0.95} = 2.179$, H_0 kan ej förkastas

Uppgift 5 (20p)

a) Vilka fyra komponenter kan en tidsserie bestå av? (5p)

A	Median, medelvärde, standardavvikelse, percentil
B	Primärdata, sekundärdata, kvalitativ data, kvantitativ data
C	Trend, säsong, cykel och slump
D	Data, frekvens, variation, intervall
E	Observationer, kategorier, trender, fördelningar

Svar C

b) Vilken av följande metoder lämpar sig för att undersöka om feltermerna i en tidsseriemodell är behäftade med autokorrelation? (5p)

A	Likelihoodkvottest
B	Walds test
C	Whites test
D	Durbin-Watson's test
E	Partiellt F-test

Svar D

c) Vilka tre komponenter ingår i en ARIMA-modell? (5p)

A	Trend, säsong, slumpmässig komponent
B	Medelvärde, varians, standardavvikelse
C	Datajustering, utjämning, modellvalidering
D	Regression, prediktion, anpassning
E	Autoregression, integrering, glidande medelvärde

Svar E

d) Baserat på följande utskrift visar resultatet av en analys på 40 observationer av kvartalsvisa data. Variabeln tid avser ett tidsindex med värde 1 på den första observationen och 40 på den sista. Beräkna en punktskattning av tidsseriens värde i tidpunkt 42, som är kvartal 2. (5p)

Call:

```
lm(formula = obs ~ tid + kvartal, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.08441	-0.55087	-0.05885	0.48159	1.84187

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.27446	0.38983	26.356	< 2e-16	***
tid	0.46367	0.04902	9.459	3.56e-11	***
kvartal2	2.49732	0.63153	3.954	0.000356	***
kvartal3	-0.77253	1.05812	-0.730	0.470189	
kvartal4	2.33740	1.52346	1.534	0.133954	

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

Lösning

$$\hat{y} = 10.27446 + 0.46367 \cdot 42 + 2.49732 \cdot 1 = 32.25$$