

Tentamen Econometrie 1, 4 april 2006, 9.30-11.30 uur

Dit tentamen duurt 2 uur! Toiletbezoek is niet toegestaan.

De uitslag komt uiterlijk na 15 werkdagen op Blackboard. Desgewenst kunt u daarna uw werk inzien bij de docent. Vul bovenaan op ieder antwoordenblad de gevraagde gegevens in. Bij dit tentamen mag u gebruik maken van één handgeschreven formuleblad op A4-formaat. U mag ook de uitgereikte tabellen gebruiken. Inlevering van uw uitwerkingen en van de uitgereikte tabellen is verplicht.

In elk van de opgaven dient u ervan uit te gaan dat er voldaan is aan de basisveronderstellingen van het lineaire regressiemodel. Eventuele afwijkingen blijken uit de opgaven. Tussen haakjes worden onder de regressiecoëfficiënten hun standaardfouten vermeld. Voer eventuele toetsen uit met een significantieniveau van 5%.

MOTIVEER UW ANTWOORDEN!

Opgave 1 (a=5, b=5, c=9, d=7, e=13, f=4, g=9, h=10 totaal=62)

Gelijke behandeling van mannen en vrouwen was in Noord Amerika vaak een reden voor discussie. Een empirisch onderzoek moet duidelijkheid brengen. Er zijn gegevens verzameld van 100 willekeurige wetenschappelijk medewerkers van universiteiten in Noord Amerika. Deze gegevens betreffen de volgende variabelen.

SALARY = jaarsalaris (in US dollars)

YEARS = aantal jaren gewerkt sinds het behalen van de eerste graad (bachelor)

PHD = 1 als gepromoveerd, 0 als niet gepromoveerd

EVALUATION = gemiddelde score bij onderwijsevaluaties

ARTICLES = aantal gepubliceerde wetenschappelijke artikelen

GENDER = 1 als man, 0 als vrouw

Het volgende model wordt opgesteld:

$$SALARY = \beta_1 + \beta_2 YEARS + \beta_3 PHD + \beta_4 EVALUATION + \beta_5 ARTICLES + \beta_6 GENDER + \varepsilon$$

Op de volgende bladzijden staan enkele Eviews outputs (RESID01 = OLS-residuen van het model).

- Interpreteer de geschatte coëfficiënten van de constante term, *YEARS* en *GENDER*.
- Stel dat de term $\beta_7 GENDER * YEARS$ aan het model wordt toegevoegd. Wanneer is het zinvol om dit te doen? Kan men de geschatte coëfficiënt van de variabele *GENDER* dan nog op soortgelijke wijze interpreteren? Leg uit!
- Bepaal een 95% betrouwbaarheidsinterval voor het verschil in salaris tussen mannen en vrouwen. Is dit verschil significant?
- Stel dat de variabele *PHD* in werkelijkheid niet relevant is (dwz. $\beta_3 = 0$). Welke gevolgen heeft het onterecht opnemen van *PHD* als verklarende variabele dan voor de verwachte waarde van het middelpunt van het onder c bedoelde betrouwbaarheidsinterval? Wat zijn de gevolgen voor de verwachte lengte van dit betrouwbaarheidsinterval?
- Toets of de prestaties op het gebied van onderwijs en onderzoek even belangrijk zijn voor het salaris. Neem aan dat beide prestatiegebieden even belangrijk zijn, indien geldt $\beta_4 s_{EVALUATION} = \beta_5 s_{ARTICLES}$ ($s_{EVALUATION}$ en $s_{ARTICLES}$ zijn de steekproefstandaarddeviaties).
- Is het zinvol om te toetsen of de storingsterm normaal verdeeld is? Waarom wel/niet?
- Toets het model op heteroscedasticiteit. Geef o.a. een duidelijke specificatie van de hypothesen.
- De steekproef bevat 2 personen die een Nobelprijs hebben ontvangen (de 11^e en de 97^e waarneming). Toets of voor deze personen hetzelfde model van toepassing is als op de andere personen in de steekproef. Specificeer o.a. het veronderstelde model (zonder restricties) en de restricties die binnen dit model worden getoetst.

	SALARY	YEARS	PHD	EVALUATION	ARTICLES	GENDER
Mean	44889.79	23.70000	0.850000	5.349600	12.16000	0.630000
Median	43718.00	24.00000	1.000000	5.335000	11.00000	1.000000
Maximum	70073.00	40.00000	1.000000	6.830000	26.00000	1.000000
Minimum	19943.00	8.000000	0.000000	3.930000	1.000000	0.000000
Std. Dev.	12906.33	9.541510	0.358870	0.541420	5.928394	0.485237
Skewness	0.056141	-0.029423	-1.960392	-0.038176	0.137178	-0.538520
Kurtosis	2.058159	1.725197	4.843137	2.922945	2.101198	1.290004
Jarque-Bera	3.748632	6.785778	78.20710	0.049030	3.679652	17.01709
Probability	0.153460	0.033611	0.000000	0.975783	0.158845	0.000202
Sum	4488979.	2370.000	85.00000	534.9600	1216.000	63.00000
Sum Sq. Dev.	1.65E+10	9013.000	12.75000	29.02038	3479.440	23.31000
Observations	100	100	100	100	100	100

Dependent Variable: SALARY

Method: Least Squares

Date: 03/29/06 Time: 16:19

Sample: 1 100

Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5916.189	3140.847	-1.883629	0.0627
YEARS	1021.696	48.92556	20.88266	0.0000
PHD	725.7481	961.5237	0.754790	0.4523
EVALUATION	3728.939	619.8220	6.016145	0.0000
ARTICLES	439.1485	80.69258	5.442241	0.0000
GENDER	1089.720	631.9795	1.724296	0.0879
R-squared	0.948186	Mean dependent var		44889.79
Adjusted R-squared	0.945430	S.D. dependent var		12906.33
S.E. of regression	3014.947	Akaike info criterion		18.91868
Sum squared resid	8.54E+08	Schwarz criterion		19.07499
Log likelihood	-939.9338	F-statistic		344.0367
Durbin-Watson stat	2.080184	Prob(F-statistic)		0.000000

Estimated Covariance Matrix of the Coefficients (bij de bovenstaande regressie)

	C	YEARS	PHD	EVALUATION	ARTICLES	GENDER
C	9864920.	-25629.31	-438246.4	-1791572.	69146.42	-80516.66
YEARS	-25629.31	2393.710	19775.77	-2433.492	-2819.838	-958.0802
PHD	-438246.4	19775.77	924527.8	-75823.25	-29607.20	-80381.43
EVALUATION	-1791572.	-2433.492	-75823.25	384179.3	-10695.58	-18177.90
ARTICLES	69146.42	-2819.838	-29607.20	-10695.58	6511.292	1412.093
GENDER	-80516.66	-958.0802	-80381.43	-18177.90	1412.093	399398.1

Dependent Variable: ABS(RESID01)

Method: Least Squares

Date: 03/29/06 Time: 16:40

Sample: 1 100

Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2297.637	429.1797	5.353554	0.0000
YEARS	6.788571	16.81029	0.403834	0.6872

R-squared	0.001661	Mean dependent var	2458.526
Adjusted R-squared	-0.008526	S.D. dependent var	1589.156
S.E. of regression	1595.916	Akaike info criterion	17.60808
Sum squared resid	2.50E+08	Schwarz criterion	17.66018
Log likelihood	-878.4040	F-statistic	0.163082
Durbin-Watson stat	2.182931	Prob(F-statistic)	0.687215

Dependent Variable: SALARY

Method: Least Squares

Date: 04/03/06 Time: 11:07

Sample: 1 10 12 96 98 100

Included observations: 98

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5986.939	3091.985	-1.936277	0.0559
YEARS	1023.958	47.62945	21.49842	0.0000
PHD	1136.917	955.7405	1.189567	0.2373
EVALUATION	3704.146	605.9601	6.112854	0.0000
ARTICLES	411.3708	79.23381	5.191860	0.0000
GENDER	1123.744	619.2940	1.814556	0.0729

R-squared	0.949659	Mean dependent var	44510.35
Adjusted R-squared	0.946923	S.D. dependent var	12730.47
S.E. of regression	2932.904	Akaike info criterion	18.86464
Sum squared resid	7.91E+08	Schwarz criterion	19.02291
Log likelihood	-918.3675	F-statistic	347.1065
Durbin-Watson stat	2.066160	Prob(F-statistic)	0.000000

Opgave 2 (a=8, b=10, c1=10, c2=10, totaal=38)

Beschouw het model $y = X\beta + \varepsilon$. Hierin is X een *stochastische* $n \times k$ matrix, y en ε zijn $n \times 1$ en β is $k \times 1$. De eerste kolom van de matrix X bestaat uit louter énen. Definieer $\iota = (1 \dots 1)'$.

- a) Toon aan dat $\bar{y}\iota$ op te vatten is als een loodrechte projectie door te laten zien dat te schrijven valt $\bar{y}\iota = Ay$, waarbij de matrix A precies de vorm van een projectiematrix heeft. Waar wordt in dit geval op geprojecteerd? Leidt ook de projectiematrix af, waarmee een vector omgezet kan worden in een vector in afwijking van het gemiddelde.
- b) Voor $k=3$ kan het model geschreven worden als $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$. Leid voor $k=3$ op basis van de stelling van Frisch-Waugh de volgende formule voor de OLS-schatter van β_2 af:
$$b_2 = \frac{\text{Var}(x_3)\text{Cov}(x_2, y) - \text{Cov}(x_2, x_3)\text{Cov}(x_3, y)}{\text{Var}(x_2)\text{Var}(x_3) - [\text{Cov}(x_2, x_3)]^2}.$$
- c) Beschouw de “kwadratische schatter” van σ^2 van de vorm $\hat{\sigma}^2 = y' A_X y$, waarin A_X een symmetrische positief semi definitie $n \times n$ matrix is, die alleen van X afhangt en niet van y .
- (c1) Toon aan dat $\hat{\sigma}^2$ zuiver is onder de voorwaarden $A_X X = 0$ en $\text{trace}(A_X) = 1$. Ga er hierbij vanuit dat X stochastisch is.
- (c2) Beschouw de schatter $s^2 = e'e/(n-k)$, waarin de vector e het OLS-residu voorstelt. Werk dit uit totdat u de vorm $s^2 = y' A_X y$ verkrijgt, waarbij de matrix A_X alleen van X afhangt (en niet van y). Toon aan dat in dit specifieke geval de matrix A_X voldoet aan de voorwaarden $A_X X = 0$ en $\text{trace}(A_X) = 1$ voor zuiverheid.

Uitwerking tentamen Econometrie 1, 4 april 2006

Opgave 1

- 1a) De constante heeft geen zinvolle interpretatie, want bv. EVALUATION=0 komt niet voor. De coëfficiënt van YEARS geeft aan dat een jaar extra werkervaring leidt tot een extra salaris van ongeveer 1022 euro bij constant houden van de overige variabelen. De coëfficiënt van GENDER geeft aan dat een man ongeveer 1090 euro meer verdient dan een vrouw die in een vergelijkbare situatie verkeert.
- 1b) Dit is zinvol, als de salarisgroei bij toename van de werkervaring verschillend is voor mannen en vrouwen, dus als het effect van YEARS op SALARY is verschillend voor mannen (GENDER=1) en vrouwen (GENDER=0). De coëfficiënt van GENDER kan dan niet meer op dezelfde wijze geïnterpreteerd worden, want het geschatte salarisverschil tussen een man en een vrouw is dan niet b_6 , maar $b_6 + b_7YEARS$, dus afhankelijk van het aantal jaren ervaring.
- 1c) $1089.720 \pm 1.99 * 631.9795$, waarbij $t_{0.025,94} \approx 1.99$, oftewel $(-167.9; 2347.4)$. Dit verschil is niet significant, want het bevat de waarde 0 niet (zie ook $t=1.724296 < 1.99$ en $Prob.=0.0879 > 0.05$).
- 1d) Als een irrelevante verklarende variabele (PHD) in het model wordt opgenomen, dan zijn de OLS-schatters nog wel zuiver, dus er zijn geen gevolgen voor de verwachte waarde van het middelpunt van het betrouwbaarheidsinterval. Echter, de OLS-schatters worden wel inefficiënt, m.a.w. de variantie van het middelpunt wordt groter. Daardoor wordt de standaardfout van het middelpunt naar verwachting groter, dus het betrouwbaarheidsinterval wordt naar verwachting langer.
- 1e) Toets de restrictie $R\beta = 0$ met $R = [0 \ 0 \ 0 \ 0.54142 \ -5.928394 \ 0]$. Toetsingsgrootheid:
 $F = (Rb)'[R\{s^2(X'X)^{-1}\}R']^{-1}(Rb)/J \sim F[J=1, n-k=94]$. Kritiek gebied: $F \geq F_{0.05,1,94} \approx 3.94$.
Uitkomst: $Rb = 0.54142 * 3728.939 - 5.928394 * 439.1485 = -584.523$,
 $R\{s^2(X'X)^{-1}\} = [\dots \ 271410 \ -44392.3 \ \dots]$, $R\{s^2(X'X)^{-1}\}R' = 410122$,
 $F = (-584.523)[410122]^{-1}(-584.523)/1 = 0.833 < 3.94$ dus de restrictie niet verwerpen. We kunnen niet concluderen dat er een verschil in belangrijkheid is.
NB. De toets kan ook gedaan worden via $t = -584.523 / \sqrt{410122} = -0.913 < t_{0.025,94} = 1.99$.
- 1f) Toetsen op normaliteit is overbodig, want de steekproef is groot genoeg om de asymptotische toepasbaarheid van de t en F verdelingen te kunnen gebruiken.
- 1g) Binnen de regressie $\sigma_i = \lambda_1 + \lambda_2YEARS_i + \eta_i$, waarbij σ_i wordt geschat door de absolute waarde van het OLS-residu, toetsen we $H_0 : \lambda_2 = 0$ tegen $H_1 : \lambda_2 \neq 0$. Glejser test: $t=0.403834$ met $Prob. = 0.6872 > 0.05$ dus er is geen significante heteroscedasticiteit.
- 1h) Definieer $D1=1$ voor de 11^e waarneming en 0 voor de overige waarnemingen en definieer $D2=1$ voor de 97^e waarneming en 0 voor de overige waarnemingen. Het model zonder restricties is
 $SALARY = \beta_1 + \beta_2YEARS + \beta_3PHD + \beta_4EVALUATION + \beta_5ARTICLES + \beta_6GENDER$
 $+ \beta_7D1 + \beta_8D2 + \varepsilon$
We toetsen $H_0 : (\beta_7, \beta_8) = (0, 0)$ tegen $H_1 : (\beta_7, \beta_8) \neq (0, 0)$. Toetsingsgrootheid:
 $F = \frac{(RSS_R - RSS_U)/2}{RSS_U/92} \sim F(2, 92)$. Uitkomst: $F = \frac{(8.54 * 10^8 - 7.91 * 10^8)/2}{7.91 * 10^8 / 92} = 3.66 > F_{0.05,2,92} \approx 3.10$.
Conclusie: op de nobelprijswinnaars is niet hetzelfde model van toepassing als op de rest.

Opgave 2

2a) $\bar{y}_t = \left(\frac{1}{n} \sum_{i=1}^n y_i \right) t = \frac{1}{t't} (t'y) t = t(t't)^{-1} t'y = Ay$ met $A = t(t't)^{-1} t'$. Hierbij is A de projectiematrix bij

loodrechte projectie op de vectorruimte van de vector t .

De vector y inafwijking van het gemiddelde is $\tilde{y} = y - \bar{y}_t = I_n y - t(t't)^{-1} t'y = [I_n - t(t't)^{-1} t'] y = M_t y$, waarin $M_t = I_n - t(t't)^{-1} t'$ de projectiematrix mbt. het orthogonale complement van de vector t is.

2b) Partitioneer $X = [t \quad X_B]$ en soortgelijk $\beta = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}$. Definieer $M_t = I_n - t(t't)^{-1} t'$, dan geldt volgens

Frisch-Waugh $b_B = [X_B' M_t X_B]^{-1} X_B' M_t y = [(M_t X_B)' (M_t X_B)]^{-1} (M_t X_B)' (M_t y)$. In termen van variabelen in afwijking van het gemiddelde $\tilde{y} = M_t y$ en $\tilde{X}_B = M_t X_B$ is dit $b_B = (\tilde{X}_B' \tilde{X}_B)^{-1} \tilde{X}_B' \tilde{y} \Rightarrow$

$$b_B = \begin{bmatrix} \sum \tilde{x}_{i2}^2 & \sum \tilde{x}_{i2} \tilde{x}_{i3} \\ \sum \tilde{x}_{i2} \tilde{x}_{i3} & \sum \tilde{x}_{i3}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum \tilde{x}_{i2} \tilde{y}_i \\ \sum \tilde{x}_{i3} \tilde{y}_i \end{bmatrix} = \begin{bmatrix} n \text{Var}(x_2) & n \text{Cov}(x_2, x_3) \\ n \text{Cov}(x_2, x_3) & n \text{Var}(x_3) \end{bmatrix}^{-1} \begin{bmatrix} n \text{Cov}(x_2, y) \\ n \text{Cov}(x_3, y) \end{bmatrix}$$

$$= \begin{bmatrix} \text{Var}(x_2) & \text{Cov}(x_2, x_3) \\ \text{Cov}(x_2, x_3) & \text{Var}(x_3) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(x_2, y) \\ \text{Cov}(x_3, y) \end{bmatrix}$$

$$= \frac{1}{\text{Var}(x_2) \text{Var}(x_3) - [\text{Cov}(x_2, x_3)]^2} \begin{bmatrix} \text{Var}(x_3) & -\text{Cov}(x_2, x_3) \\ -\text{Cov}(x_2, x_3) & \text{Var}(x_2) \end{bmatrix} \begin{bmatrix} \text{Cov}(x_2, y) \\ \text{Cov}(x_3, y) \end{bmatrix}$$

Het 1^e element van b_B is $b_2 = \frac{\text{Var}(x_3) \text{Cov}(x_2, y) - \text{Cov}(x_2, x_3) \text{Cov}(x_3, y)}{\text{Var}(x_2) \text{Var}(x_3) - [\text{Cov}(x_2, x_3)]^2}$.

2c1) $E[\hat{\sigma}^2 | X] = E[y' A_X y | X] = E[(X\beta + \varepsilon)' A_X (X\beta + \varepsilon) | X] =$
 $= E[\beta' X' A_X X \beta + \beta' X' A_X \varepsilon + \varepsilon' A_X X \beta + \varepsilon' A_X \varepsilon | X]$
 $= \beta' X' A_X X \beta + \beta' X' A_X E[\varepsilon | X] + E[\varepsilon | X]' A_X X \beta + \text{trace}(E[\varepsilon \varepsilon' | X] A_X)$
 $= \beta' X' A_X X \beta + 0 + 0 + \text{trace}(\sigma^2 I_n A_X) = \beta' X' A_X X \beta + \sigma^2 \text{trace}(A_X)$
 $= \sigma^2$ als $A_X X = 0$ en $\text{trace}(A_X) = 1$.

Dan geldt onder dezelfde voorwaarden $E[\hat{\sigma}^2] = E[E[\hat{\sigma}^2 | X]] = E[\sigma^2] = \sigma^2$, dus $\hat{\sigma}^2$ is zuiver.

2c2) Definieer $M = I_n - X(X'X)^{-1} X'$, dan geldt

$$s^2 = \frac{e'e}{n-k} = \frac{(My)'(My)}{n-k} = \frac{y'My}{n-k} = y' \left(\frac{1}{n-k} M \right) y = y' A_X y \quad \text{met } A_X = \frac{1}{n-k} M.$$

Deze $A_X = \frac{1}{n-k} M$ hangt alleen van X af en niet van y . Verder geldt:

$$A_X X = \frac{1}{n-k} M X = \frac{1}{n-k} [I_n - X(X'X)^{-1} X'] X = \frac{1}{n-k} [X - X(X'X)^{-1} X'X] = 0,$$

$$\text{trace}(A_X) = \text{trace} \left(\frac{1}{n-k} M \right) = \frac{1}{n-k} \text{trace}[I_n - X(X'X)^{-1} X']$$

$$= \frac{1}{n-k} [\text{trace}(I_n) - \text{trace}\{X(X'X)^{-1} X'\}] = \frac{1}{n-k} [n - \text{trace}\{(X'X)^{-1} X'X\}] = \frac{1}{n-k} [n - k] = 1$$

dus aan de voorwaarden voor zuiverheid is voldaan.