Inferens

Økonometri A

Bertel Schjerning

Program

Fordelingen af OLS estimatoren (W4.1)

Hypotesetestning med en parameter (W4.2-W4.3)

- t-test
- p-værdier
- Konfidensintervaller

Test af lineære kombinationer (W4.4)

Multiple lineær hypoteser (W4.5)

- F-test

Fordeling af teststørrelser

Motivation

Hvad ved vi om OLS indtil videre?

OLS estimatoren er en *estimator* ("estimations maskine"):

Input (Data)		**Output (Estimater)**
Sample 1: $\{(y_1, X_1), ..., (y_n, X_n)\}$		$(\hat{\beta}_0, \hat{\beta}_1, ...)_1$
$\vdots$	$\rightarrow$ **OLS** $\rightarrow$	$\vdots$
Sample k: $\{(y_1, X_1), ..., (y_n, X_n)\}$		$(\hat{\beta}_0, \hat{\beta}_1, ...)_k$

OLS estimaterne er med andre ord stokastiske, og vi har vist at

$$\hat{\beta}_j \sim F\left(\beta_j, \frac{\sigma^2}{SSR_j}\right)$$

$$\hat{\boldsymbol{\beta}} \sim F\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

Men vi ved stadig ikke hvordan $\hat{\beta}_j$ er fordelt.

Motivation: Skal vi bruge fordelingen af OLS estimaterne?

Typisk vil vi gerne bruge estimationer til at *teste hypoteser*:

- **Fx:** Er der et positivt afkast af DatØk uddannelsen?
- **Fx:** Er afkastet højere end for Datalogi?

Indtil videre: Vi har estimeret effekten af fx uddannelse på lønnen.

Men: Vi vil også gerne vide, om det estimat, vi får, lige så vel kunne skyldes *stikprøveusikkerheden*.

For at kunne svare på sådanne spørgsmål, er det *ikke* nok at kende:

- **Middelværdi** af $\hat{\beta}_j$
- **Varians** af $\hat{\beta}_j$

Vi skal også kende fordelingen!

Motivation: Eksempler på hypoteser

Vi kunne eksempelvis være interesseret i at undersøge om:

- $\beta_j = 0$ *(ingen effekt af variabel)*
- $\beta_j = a$ *(parameter lig en given værdi)*
- $\beta_j \leq a$ *(parameter mindre eller lig en værdi)*
- $\beta_j = \beta_h$ *(fx: $\beta_{DatØk} = \beta_{Datalogi}$)*
- $k_j\beta_j + k_h\beta_k = a$ *(lineær restriktion på β)*
- $f(\beta) = a$, hvor $f(.)$ er en funktion af parametrene

Testteori: Recap

Procedure for hypotesetestning:

1. **Definer nulhypotesen H_0 og alternativhypotesen H_1.**

   Vi opretholder hypotesen H_0, indtil den er modbevist.

2. **Udregn teststørrelsen i din stikprøve.**

   Vi har brug for en teststørrelse ("test statistic"), som vi kender den teoretiske fordeling af, forudsat at H_0 er sandt.

3. **Beregn sandsynligheden for at observere teststørrelsen**, *hvis H_0 er sand.*

4. **Afvis** H_0, hvis sandsynligheden er tilstrækkelig lille.

   Det er os, der vælger, hvad vi mener med "lille". Dette vender vi tilbage til.

Fordelingen af OLS estimatoren

Fordelingen af OLS: Normalitet

Ny antagelse:

MLR.6 Fejlleddene u er uafhængige af de forklarende variable X og er normalfordelt med middelværdi 0 og varians σ^2:

$$u \sim N(0, \sigma^2)$$

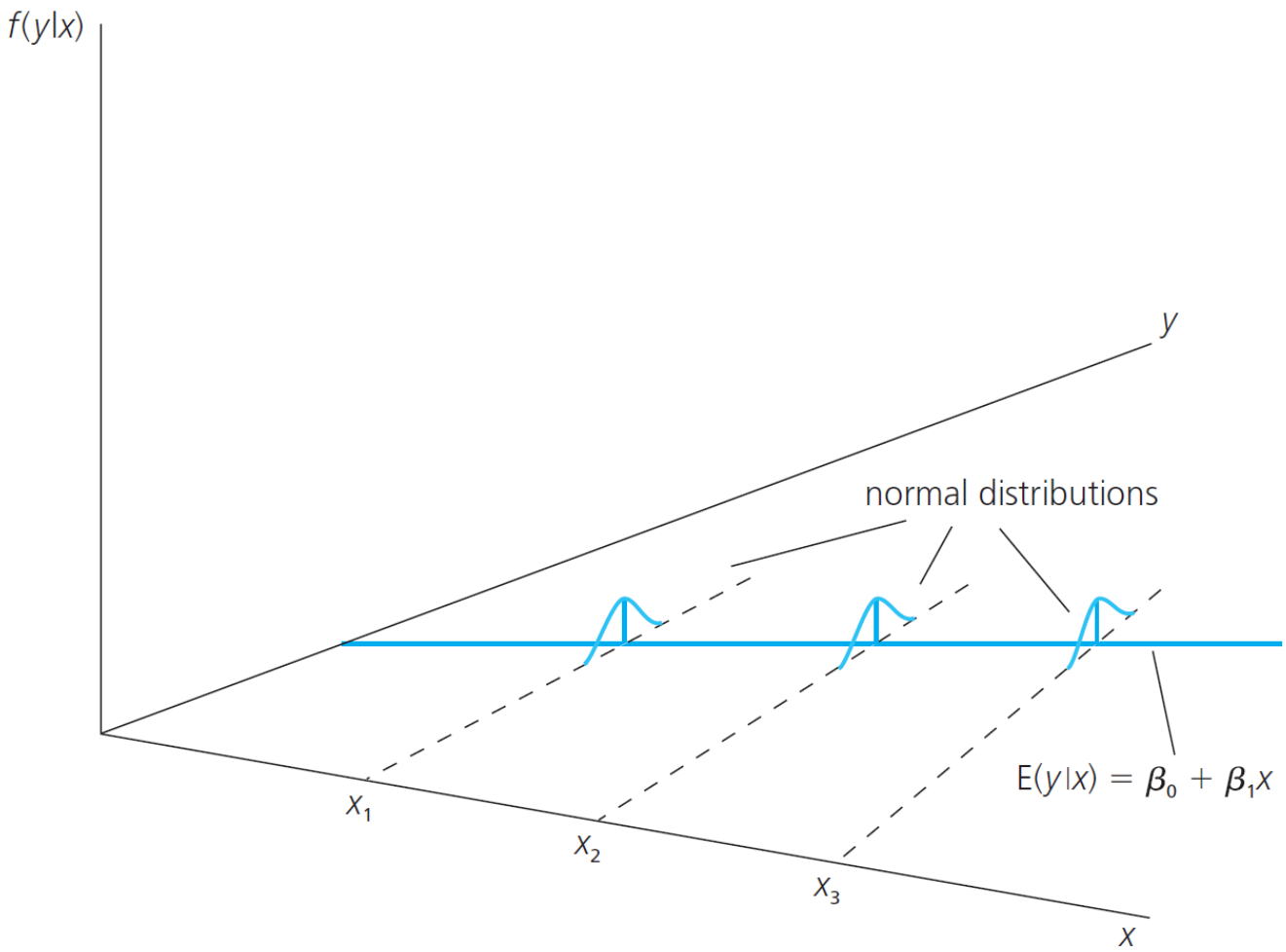
MLR.1-MLR.6 kaldes de klassiske lineære model antagelser (CLM).

Under MLR.1-MLR.6 er:

$$y|\mathbf{x} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2)$$

Fordelingen af OLS: Normalitet

FIGURE 4.1 The homoskedastic normal distribution with a single explanatory variable.



Fordelingen af OLS: Normalitet

Er antagelse MLR.6 rimelig?

- **Det afhænger af problemstillingen:**
  For nogle modeller er antagelsen om normalitet i fejlleddene realistisk, mens den i andre er problematisk.
- **Normaltfordelingen kan være en god antagelse, når:**
  u er summen af mange små, uafhængige faktorer *(Centrale grænseværdi sætning)*.
- **Urealistisk, når:** y kun kan antage et begrænset antal værdier (som binære eller kategoriske variable).
- **Ultimativt et empirisk spørgsmål** (afhænger af data/kontekst).
- **Testbar antagelse:** Antagelsen om normalfordeling kan testes.

Vigtigt: MLR.6 er kun nødvendig, når vi arbejder med små stikprøver (mere om det i W.5)

Fordelingen af OLS: Normalitet

Teorem 4.1: OLS estimator med **kendt** σ^2 er normalfordelt

Under antagelse af MLR.1–MLR.6 er OLS estimatoren normalt fordelt:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

For den enkelte β_j har vi:

$$\hat{\beta}_j \sim N\left(\beta_j, Var(\hat{\beta}_j)\right)$$

og

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sigma / \sqrt{SST_j(1 - R_j^2)}} \sim N(0,1)$$

Fordelingen af OLS: t-fordelingen

Variansen af $\hat{\beta}$ indeholder den ukendte parameter σ^2, som vi erstatter med $\hat{\sigma}^2$. Dette fører til:

Teorem 4.2: OLS estimator med **ukendt** σ^2 er t-fordelt

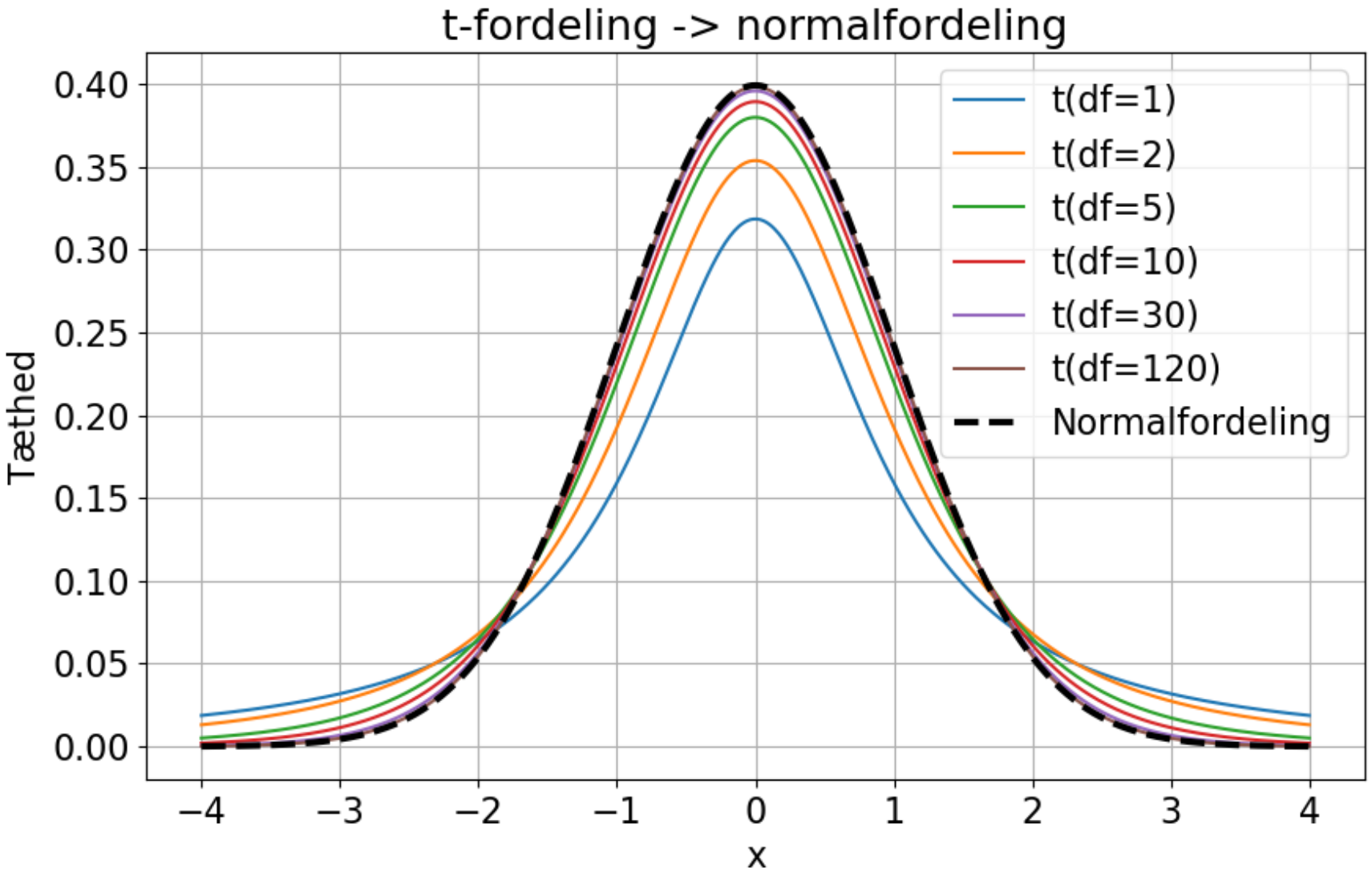
Under CLM antagelserne (MLR.1-MLR.6) gælder:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}/\sqrt{SST_j(1-R_j^2)}} \sim t_{n-k-1}$$

- $n-k-1$: Frihedsgrader (**df**, degrees of freedom).
- $k+1$: Antallet af β-parametre, som estimeres.

Bemærk: t-fordelingen konvergerer mod normalfordelingen, når antallet af frihedsgrader er "stor" ($n-k-1 > 120$).

Fordelingen af OLS: t-fordelingen og frihedsgrader



Hypotesetestning med en parameter

Hypotesetestning med en parameter

Med Teorem 4.2 kan vi teste hypoteser om en parameter i populationsregressionsmodellen:

$$H_0 : \beta_j = a$$

hvor a er en konstant. Vi anvender følgende **t-teststørrelse**:

$$t = \frac{\hat{\beta}_j - a}{se(\hat{\beta}_j)}$$

- Nulhypotesen postulerer, at parameteren er lig en bestemt værdi.
- Hvis H_0 er sand, følger teststørrelsen en t-fordeling.

Spørgsmål: Teststørrelsen er 0 når $\hat{\beta}_j = a$, men hvornår er t-teststørrelsen *"stor nok"* til at afvise H_0 til fordel for en alternativhypotese?

Valg af alternativhypotese

Vi kan fastlægge forskellige alternative hypoteser:

- **Ensidet alternativ:** $H_1 : \beta_j > a$ eller $H_1 : \beta_j < a$
- **Dobbeltsidet alternativ:** $H_1 : \beta_j \neq a$

Det relevante alternativ afhænger af problemstillingen.

Eksempel: I lønregressionen kan vi være interesseret i at undersøge:

- $H_0 : \beta_{edu} = 0$
- $H_1 : \beta_{edu} > 0$ eller $H_1 : \beta_{edu} \neq 0$

De fleste Python-pakker, såsom `statsmodels`, beregner som standard en t-test for hver parameter, altså so test for hypotesen

$$H_0 : \beta_j = 0 \text{ og } H_1 : \beta_j \neq 0$$

Lønregression: Python eksempel

```
# Code: Se 04_inference/04_inference_examples.ipynb

OLS Regression Results for Dependent Variable: lwage
==============================================================================
Number of Observations: 1078
Degrees of Freedom: 1073 (Residual), 5 (Model)
R-squared: 0.2058
TSS: 111.2507, RSS: 88.3507, ESS: 22.9000
==============================================================================
Variable   Coefficient   Std. Error          t     P>|t|    95% Conf. Interval
------------------------------------------------------------------------------
const           4.2699      0.0466     91.7066    0.0000   [  4.1786, 4.3613  ]
educ            0.0270      0.0027     10.1735    0.0000   [  0.0218, 0.0322  ]
experience      0.0428      0.0102      4.1983    0.0000   [  0.0228, 0.0628  ]
experience2    -0.0018      0.0007     -2.3782    0.0176   [ -0.0032, -0.0003 ]
experience3     0.0286      0.0154      1.8543    0.0640   [ -0.0017, 0.0588  ]
==============================================================================
```

Kritisk niveau for t-teststørrelsen

Når vi fastlægger, hvor stor t-teststørrelsen skal være, for at vi afviser H_0, balancerer vi to typer af fejl:

- Type I (falsk positiv): Vi afviser H_0 selvom H_0 er sand.
- Type II (falsk negativ): Vi afviser ikke H_0 selvom H_0 er falsk.

I økonometri fastlægger vi typisk det "kritiske niveau" for t-teststørrelsen, hvor vi afviser H_0 ved at svare på:

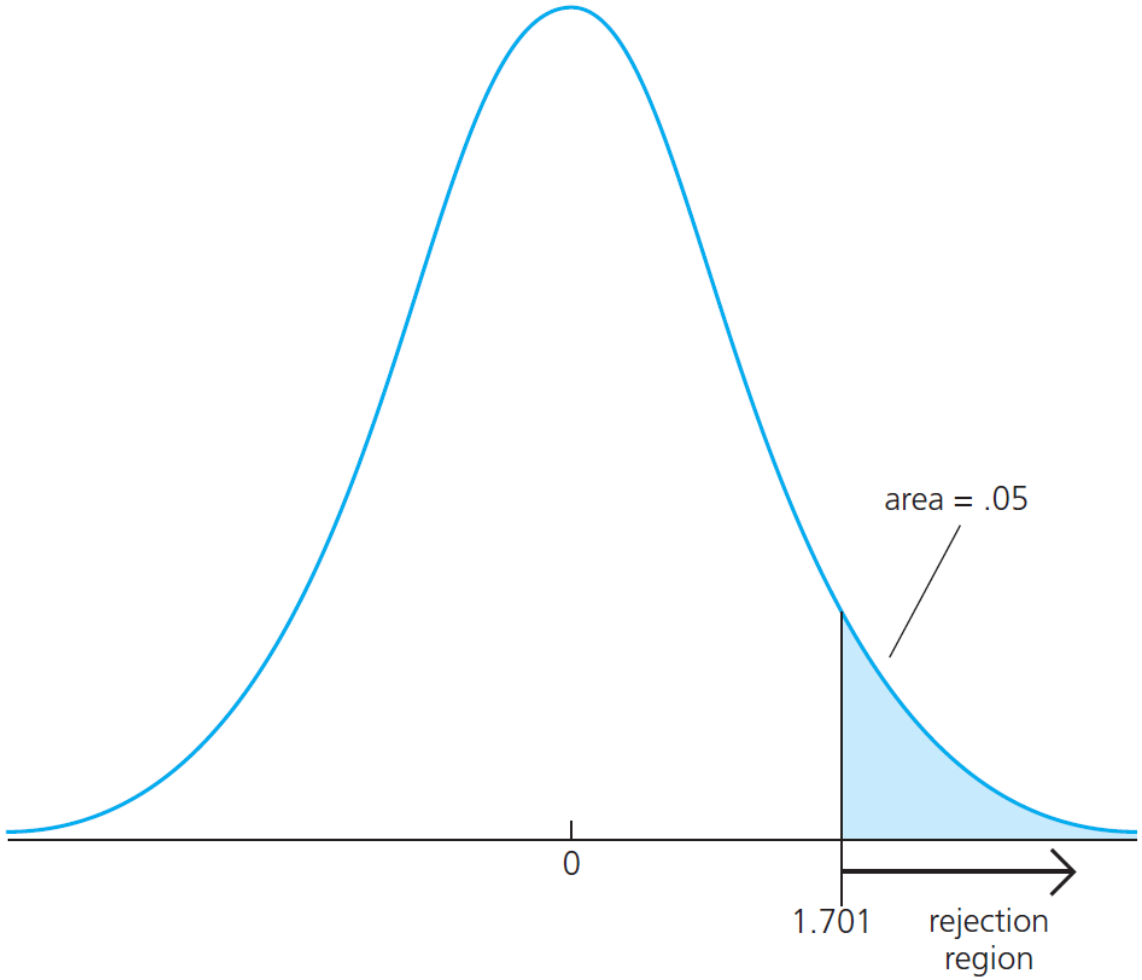
- Med hvilken sandsynlighed vil vi riskere at afvise en sand H_0?
- Dvs. sandsynligheden for type I fejl.

Typisk tillader vi en sandsynlighed for type I fejl på 10%, 5% eller 1%.

Kritisk niveau for t-teststørrelsen: $H_1 : \beta_j > 0$

Enkeltsidet alternativ, signifikansniveau 5%, 28 frihedsgrader:

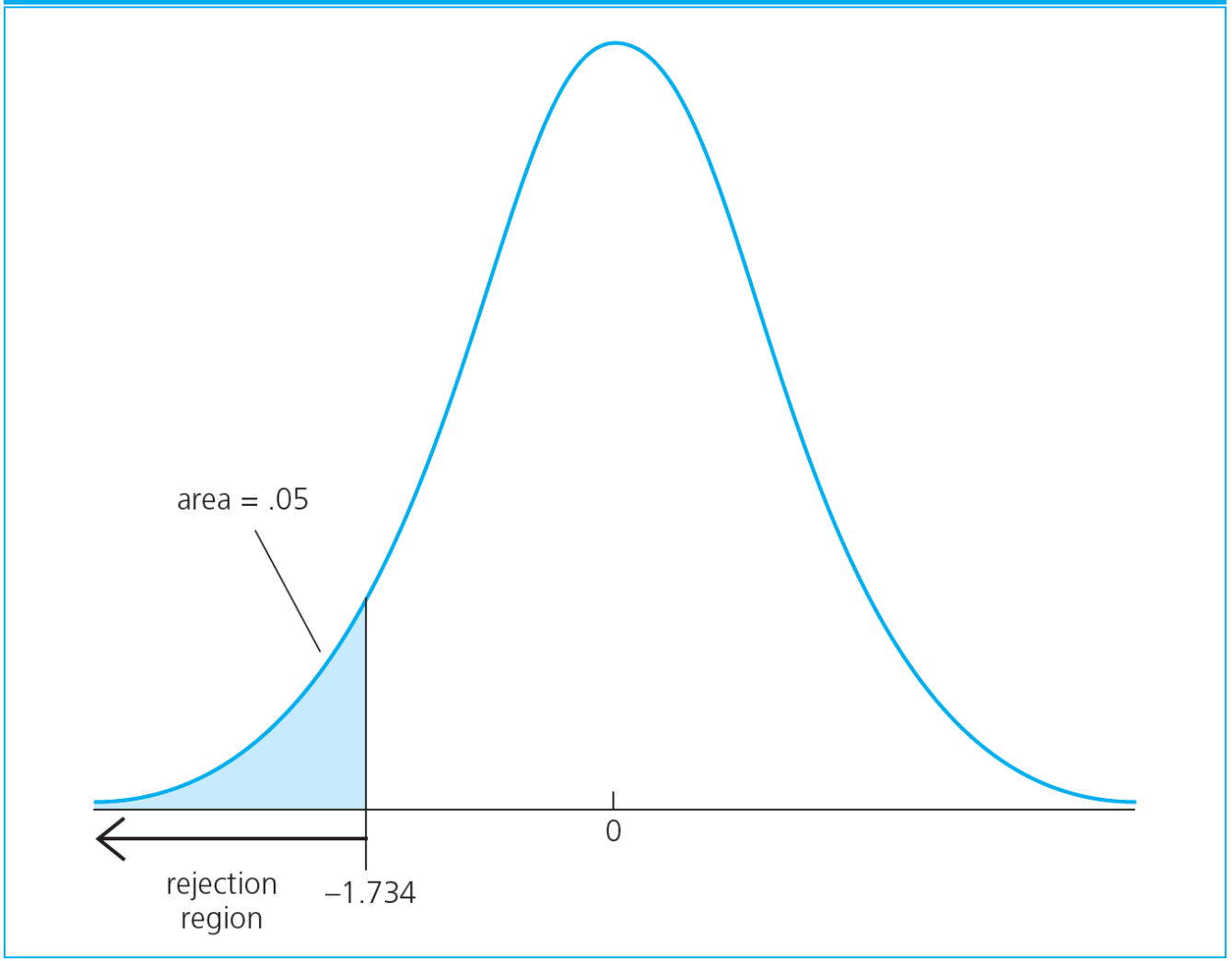
FIGURE 4.2 5% rejection rule for the alternative $H_1: \beta_j > 0$ with 28 *df*.



Kritisk niveau for t-teststørrelsen: $H_1 : \beta_j < 0$

Enkeltsidet alternativ, signifikansniveau 5%, 18 frihedsgrader:

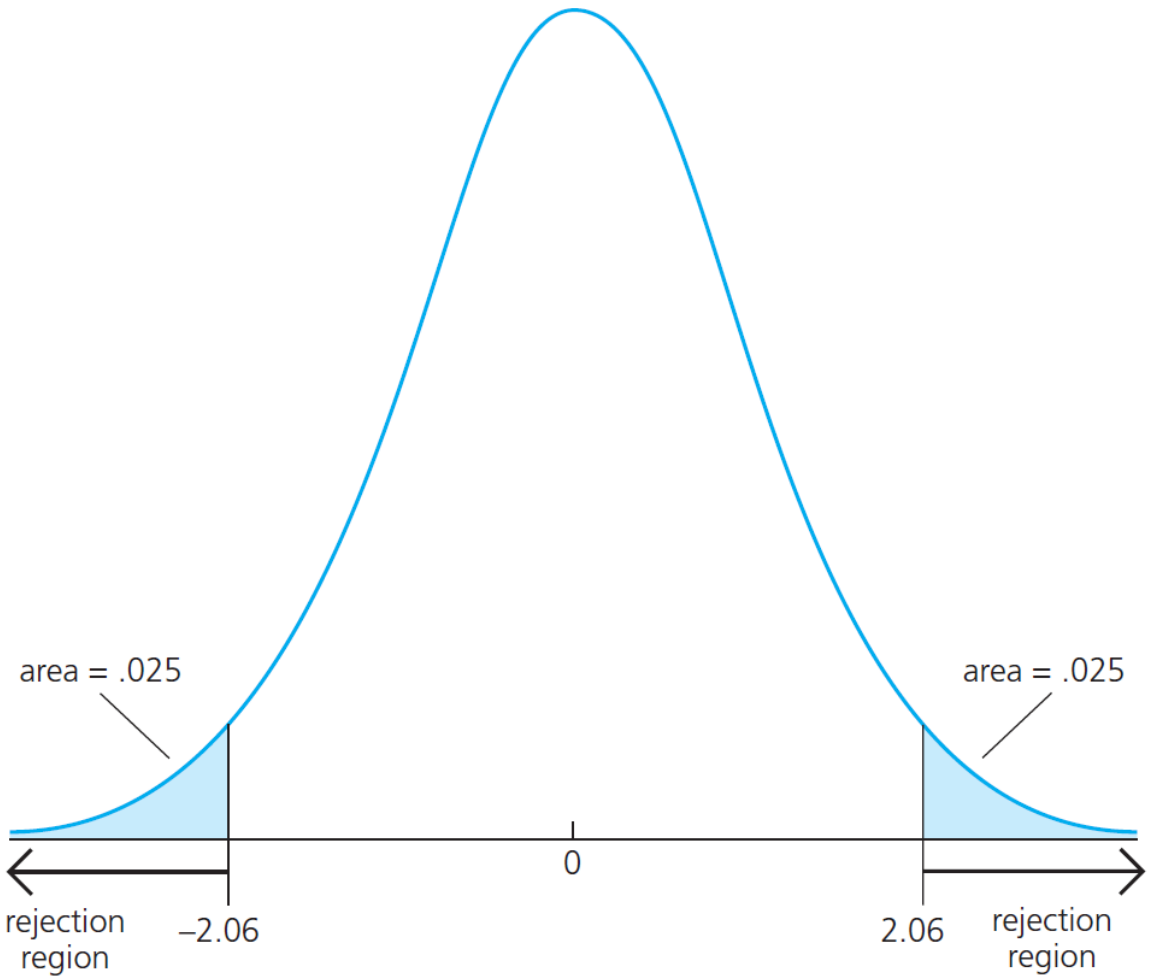
FIGURE 4.3 5% rejection rule for the alternative $H_1: \beta_j < 0$ with 18 *df*.



Kritisk niveau for t-teststørrelsen: $H_1 : \beta_j \neq 0$

Dobbeltsidet alternativ, signifikansniveau 5% (2.5% i hver side):

FIGURE 4.4 5% rejection rule for the alternative $H_1: \beta_j \neq 0$ with 25 *df*.



Lønregression: Python eksempel (igen)

Er t-teststørrelsen for experience3 stor nok til at afvise H0?
Kritisk værdi for t-tests på 5% signifikansniveau: +/- 1.96

```
OLS Regression Results for Dependent Variable: lwage
==============================================================================
Number of Observations: 1078
Degrees of Freedom: 1073 (Residual), 5 (Model)
R-squared: 0.2058
TSS: 111.2507, RSS: 88.3507, ESS: 22.9000
==============================================================================
Variable   Coefficient   Std. Error      t
------------------------------------------------------------------------------
const           4.2699      0.0466     91.7066
educ            0.0270      0.0027     10.1735
experience      0.0428      0.0102      4.1983
experience2    -0.0018      0.0007     -2.3782
experience3     0.0286      0.0154      1.8543
==============================================================================
```

p-værdier

I stedet for at lægge os fast på et kritisk niveau for t-teststørrelsen, kan vi også spørge:

- Givet H_0, hvad er sandsynligheden for at observere en teststørrelse (T), som er mindst lige så stor som den vi faktisk observerer (t)?

Svaret giver **p-værdien**. Definition:

- Dobbelt sidet alternativ $H_1 : \beta_j \neq a$:

$$P(|T| > |t|) = 2P(T > |t|)$$

- Enkeltsidet alternativ $H_1 : \beta_j > a$

$$P(T > t)$$

p-værdien viser på hvilket niveau vi kan afvise H_0.

Beregning af p-værdi i dobbelt-sidet test

Vi tester $H_0 : \beta_j = a$ mod $H_1 : \beta_j \neq a$. Under H_0 er sandsynligheden for, at teststørrelsen T ligger uden for intervallet $[-|t|, |t|]$:

$$P(|T| \geq |t|) = P(T \geq |t|) + P(T \leq -|t|)$$

Da t-fordelingen er **symmetrisk**, har vi $P(T \leq -|t|) = P(T \geq |t|)$, så vi kan nøjes med at beregne den ene sandsynlighed og gange med 2:

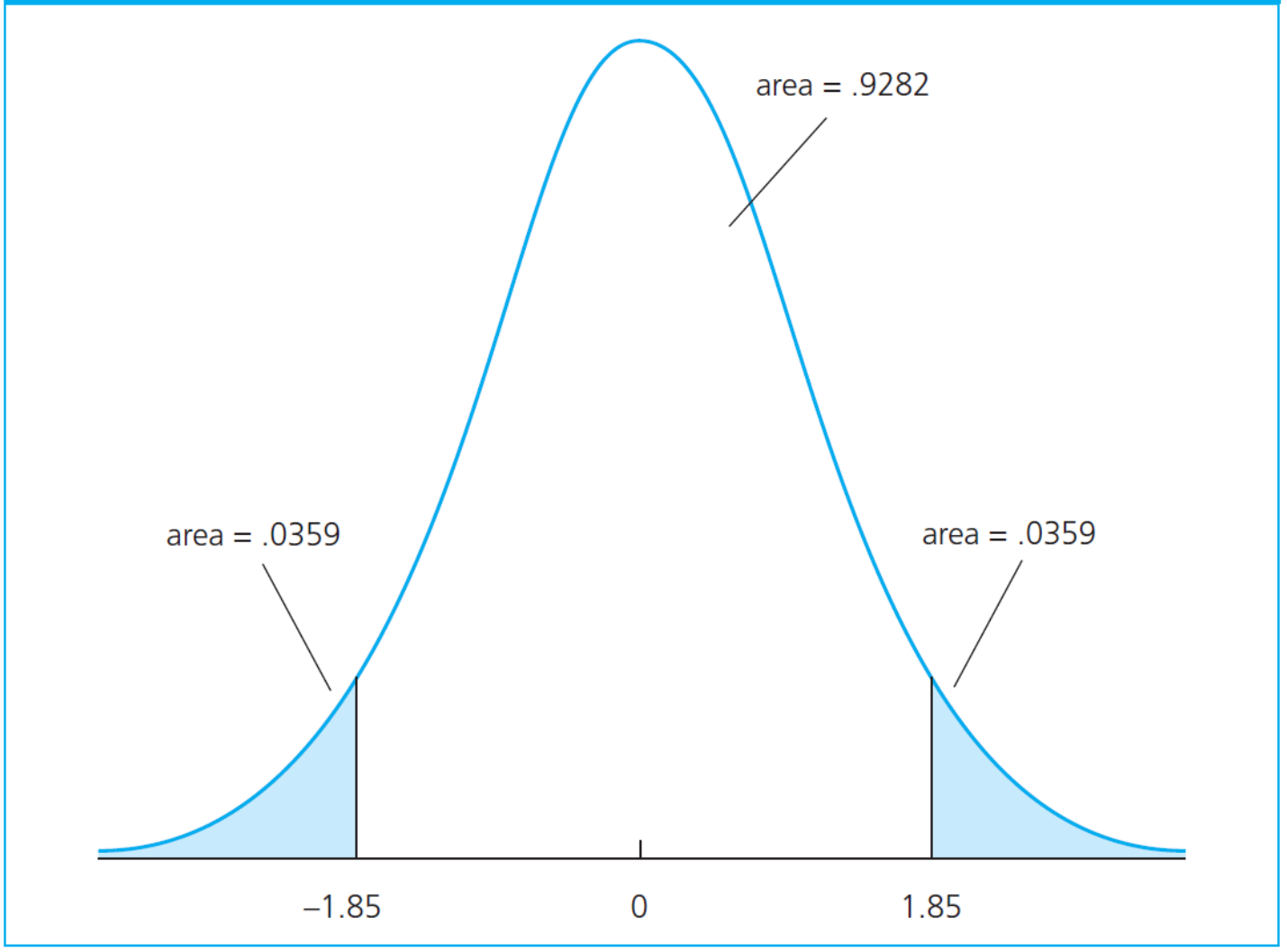
$$\text{p-værdi} = 2 \cdot P(T \geq |t|) = 2 \cdot (1 - \text{CDF}(|t|; n - k - 1))$$

hvor $\text{CDF}(t; n - k - 1) = P(T \leq t)$ er den kumulative fordelingsfunktion (CDF) for t-fordelingen med $n - k - 1$ frihedsgrader.

Bemærk: Symmetrien gør, at vi kun behøver at beregne sandsynligheden for én hale.

p-værdier for $H_1 : \beta_j \neq 0$

FIGURE 4.6 Obtaining the *p*-value against a two-sided alternative, when $t = 1.85$ and $df = 40$.



Lønregression: Python eksempel (igen)

p-værdi for t-tests kan beregnes ved opslag i t-fordeling

```
  # Dobbelt sidede p-værdier for t-test kan beregnes som:
  p_values = 2 * (1 - stats.t.cdf(np.abs(t_stat), df=n - k))
```

```
OLS Regression Results for Dependent Variable: lwage
==========================================================================
Number of Observations: 1078
Degrees of Freedom: 1073 (Residual), 5 (Model)
R-squared: 0.2058
TSS: 111.2507, RSS: 88.3507, ESS: 22.9000
==========================================================================
Variable   Coefficient   Std. Error          t     P>|t|
--------------------------------------------------------------------------
const           4.2699      0.0466     91.7066    0.0000
educ            0.0270      0.0027     10.1735    0.0000
experience      0.0428      0.0102      4.1983    0.0000
experience2    -0.0018      0.0007     -2.3782    0.0176
experience3     0.0286      0.0154      1.8543    0.0640
==========================================================================
```

p-værdier

Quiz: Antag at vi har følgende regressionsmodel

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

og vi ønsker at teste hypotesen $H_0 : \beta_2 = 0$ mod alternativhypotesen $H_1 : \beta_2 \neq 0$. Vi finder at

$$\hat{\beta}_2 = 0.56 \text{ og p-value} = 0.086$$

Hvad er p-værdien, hvis vi havde testet mod $H_1 : \beta_2 > 0$?

1. p-værdi=0.086
2. p-værdi=0.043
3. p-værdi=0.172

Konfidensintervaller

Definition:

Konfidenstintervallet for β_j er $[\underline{b}_j; \bar{b}_j]$, hvor

$$\begin{aligned} \underline{b}_j &= \hat{\beta}_j - c \cdot se(\hat{\beta}_j) \\ \bar{b}_j &= \hat{\beta}_j + c \cdot se(\hat{\beta}_j), \end{aligned}$$

hvor c fx er 97.5 fraktilen i en t-fordeling med $n - k - 1$ frihedsgrad.

Fortolkning:

Hvis vi kunne gentage estimationen med forskellige datasæt udtrukket fra samme population, så ville den sande parameter i 95% af tilfældene ligge indenfor konfidensintervallet.

- Vi siger at den sande værdi ag β med 95% sikkerhed ligger indenfor konfidensintervallet.

Lønregression: Python eksempel (igen)

```
# Code: Se 04_inference/04_inference_examples.ipynb

OLS Regression Results for Dependent Variable: lwage
==============================================================================
Number of Observations: 1078
Degrees of Freedom: 1073 (Residual), 5 (Model)
R-squared: 0.2058
TSS: 111.2507, RSS: 88.3507, ESS: 22.9000
==============================================================================
Variable   Coefficient   Std. Error          t     P>|t|    95% Conf. Interval
------------------------------------------------------------------------------
const           4.2699      0.0466     91.7066    0.0000   [  4.1786, 4.3613  ]
educ            0.0270      0.0027     10.1735    0.0000   [  0.0218, 0.0322  ]
experience      0.0428      0.0102      4.1983    0.0000   [  0.0228, 0.0628  ]
experience2    -0.0018      0.0007     -2.3782    0.0176   [ -0.0032, -0.0003 ]
experience3     0.0286      0.0154      1.8543    0.0640   [ -0.0017, 0.0588  ]
==============================================================================
```

Let's Code!

- **Jupyter Notebook:** `04_inference_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 1:** t-test, p-værdier og konfidensintervaller
- Hypotesetestning i model for timeløn, uddannelse og erfaring

Las os alle teste

Statistisk signifikans vs økonomisk signifikans

Statistisk signifikans er ikke ensbetydende med at estimatet er økonomisk relevant

Nogle gange er et resultat statistisk signifikant, men koefficienten er så lille, at den ikke har nogen praktisk signifikans.

Økonomisk signifikans skal vurderes ud fra parameterestimatets størrelse relativt til:

- Den gennemsnitlige værdi af y.
- I forhold til tidligere studier.
- Andre relevante sammenligninger (f.eks. afkast af uddannelse sammenlignet med afkast af erfaring).

Husk:

- Statistisk signifikans er vigtige, men glem ikke at fortolk $\hat{\beta}$.

Statistisk signifikans vs økonomisk signifikans

4 eksempler:

1. $\hat{\beta}_1 = 0.005,\ se(\hat{\beta}_1) = 0.005,\ t = 1,\ CI = [-0.005; 0.015]$.
2. $\hat{\beta}_1 = 0.005,\ se(\hat{\beta}_1) = 0.001,\ t = 5,\ CI = [\ \ 0.003; 0.007]$.
3. $\hat{\beta}_1 = 0.150,\ se(\hat{\beta}_1) = 0.150,\ t = 1,\ CI = [-0.150; 0.450]$.
4. $\hat{\beta}_1 = 0.150,\ se(\hat{\beta}_1) = 0.030,\ t = 5,\ CI = [\ \ 0.090; 0.210]$.

Statistisk signifikans vs økonomisk signifikans

4 eksempler:

1. $\hat{\beta}_1 = 0.005$, $se(\hat{\beta}_1) = 0.005$, $t = 1$, $CI = [-0.005; 0.015]$.
2. $\hat{\beta}_1 = 0.005$, $se(\hat{\beta}_1) = 0.001$, $t = 5$, $CI = [\quad 0.003; 0.007]$.
3. $\hat{\beta}_1 = 0.150$, $se(\hat{\beta}_1) = 0.150$, $t = 1$, $CI = [-0.150; 0.450]$.
4. $\hat{\beta}_1 = 0.150$, $se(\hat{\beta}_1) = 0.030$, $t = 5$, $CI = [\quad 0.090; 0.210]$.

2 og 4 er statistisk *signifikante*, men 2 er (nok) for lille til at have praktisk økonomisk betydning.

1 og 3 er *insignifikante*, men 1 er stadig brugbart. Her kan vi med rimelig sikkerhed afvise at effekten er stor. (skarpt 0).

Med 3 kan vi ikke afvise ret meget (upræcist 0)

Test af lineære kombinationer

Lineære kombinationer af parametre: Eksempel

Vi kan også teste hypoteser vedrørende lineære kombinationer af parametrene:

- $H_0 : \beta_1 = \beta_2$
- $H_0 : \beta_1 + \beta_2 + \beta_3 = 2$
- $H_0 : \beta_1 - \beta_2 = \beta_3$

Disse hypoteser involverer flere parametre, men kun **en restriktion**.

Eksempel: Cobb-Douglas produktionsfunktionen

$$
\begin{aligned}
Y_i &= A L_i^{\alpha} K_i^{\beta} U_i \\
\text{Tag log:} \quad \Rightarrow \log Y_i &= a_0 + \alpha \log L_i + \beta \log K_i + u_i .
\end{aligned}
$$

Hypotese: $H_0 : \alpha + \beta = 1$ (konstant skalaafkast)

Kan man bruge t-testet?

Lineære kombinationer af parametre: Eksempel

Det er let at udregne $\hat{\alpha} + \hat{\beta}$.

Det er lidt mere besværligt at udregne $se(\hat{\alpha} + \hat{\beta})$, da

$$var(\hat{\alpha} + \hat{\beta}) = var(\hat{\alpha}) + var(\hat{\beta}) + 2cov(\hat{\alpha}, \hat{\beta})$$

Vi kan dog snyde OLS til at gøre det for os ved at omskrive modellen:

$$\begin{aligned} \log Y_i &= a_0 + \alpha \log L_i + \beta \log K_i + u_i \\ &= a_0 + (\alpha + \beta) \log L_i + \beta(\log K_i - \log L_i) + u_i \\ &= a_0 + \theta \log L_i + \beta \log(K_i/L_i) + u_i \end{aligned}$$

Vi kan regressere $\log Y$ på $\log L$ og $\log(K/L)$.

Vi er tilbage i hypotesetestning med en parameter med $H_0 : \theta = 1$ (konstant skalaafkast).

Let's Code!

- **Jupyter Notebook:** `04_inference_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 2:** Returns to scale in French manufacturing
- Hypotese test med lineære restriktioner på en eller flere parametre

Las os alle teste

Multiple lineær hypoteser

Test af multiple lineær hypoteser

Ved tests af hypoteser, som indeholder mere end 1 restriktion, kan vi ikke "snyde" OLS til at bruge t-testet.

Model med 4 forklarende variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u. \tag{1}$$

Eksempler på multiple hypoteser:

1. $H_0 : \beta_3 = 1$ (en restriktion).
2. $H_0 : \beta_3 = 1, \beta_2 = \beta_1$ (to restriktioner).
3. $H_0 : \beta_1 + \beta_2 + \beta_3 = 1, \beta_4 = 0$ (to restriktioner).
4. $H_0 : \beta_1 = 0, \beta_2 = 2, \beta_3 = 1, \beta_4 = 0$ (fire restriktioner).

H_0: restriktionerne holder. Alternativhypotesen: H_0 er ikke opfyldt.

Test af multiple lineær hypoteser

Under H_0 er den sande model en **restrikteret** version af ligning (1).

Med eksemplerne på forrige slide er den restrikterede model:

1. $y - x_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + u.$
2. $y - x_3 = \beta_0 + \beta_1 (x_1 + x_2) + \beta_4 x_4 + u.$
3. $y - x_1 = \beta_0 + \beta_2 (x_2 - x_1) + \beta_3 (x_3 - x_1) + u.$
4. $y - 2x_2 - x_3 = \beta_0.$

Bemærk: Den restrikterede model er stadig lineær i parametrene.

Test af multiple lineær hypoteser

Under H_0 er den sande model en **restrikteret** version af ligning (1).

Med eksemplerne på forrige slide er den restrikterede model:

1. $y - x_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + u.$
2. $y - x_3 = \beta_0 + \beta_1 (x_1 + x_2) + \beta_4 x_4 + u.$
3. $y - x_1 = \beta_0 + \beta_2 (x_2 - x_1) + \beta_3 (x_3 - x_1) + u.$
4. $y - 2x_2 - x_3 = \beta_0.$

Bemærk: Den restrikterede model er stadig lineær i parametrene.

Ide til test:

- Under H_0 vil den restrikterede model forklarer data ligeså godt som den urestrikterede.
- Hvis H_0 ikke er sand, vil den restrikterede model forklarer data dårligere. Derved vil SSR stige og R^2 falde.

Test af multiple lineær hypoteser

Teststørrelsen for multiple hypoteser

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)}$$

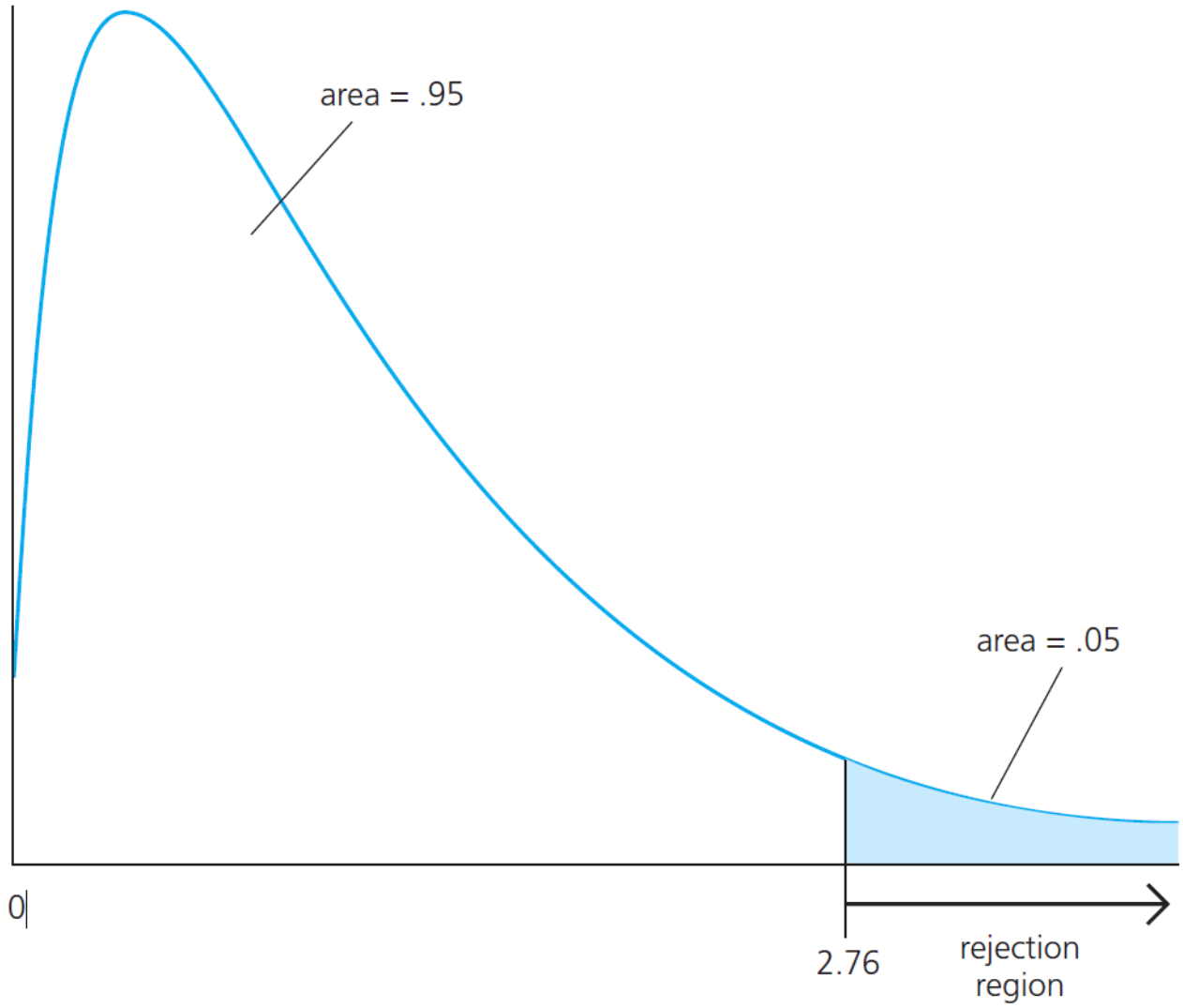
Husk: $R^2 = 1 - SSR/SST$

- Tælleren er altid positiv. Variationen i residualerne er størst, når modellen er restrikteret.
- q antal restriktioner.
- $n-k-1$ antal frihedsgrader i den urestrikterede model.
- Hvis MLR.1-MLR.6 er opfyldt, så er

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F(q, n-k-1).$$

F-fordelingen

FIGURE 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



Test af multiple lineær hypoteser

Afsluttende bemærkninger om F-testet:

- Vi kan sagtens bruge F-testet til også at enkelte parametre og lineære kombinationer. Det er det Stata gør med *test* kommandoen.
- Ved test af enkelte parameter når den dobbeltsidede t-test og F-testet ens konklusion.
- Stata og Statsmodels beregner som standard F-testet for $H_0 : \beta_1 = 0, ..., \beta_k = 0$
- Nogle gange kan ovenstående H_0 ikke afvises, selvom enkelte β er signifikant forskellige fra 0.
    - Ved et 5% signifikansniveau vil $\hat{\beta}$ i gennemsnit være signifikant selvom $H_0 : \beta = 0$ er sand i...

Lønregression: Stata eksempel

```
reg lwage educ experience experience2 experience3

      Source |       SS           df       MS      Number of obs   =     1,078
-------------+----------------------------------   F(4, 1073)      =     69.53
       Model |  22.8999718         4  5.72499296   Prob > F        =    0.0000
    Residual |  88.3507171     1,073  .082339904   R-squared       =    0.2058
-------------+----------------------------------   Adj R-squared   =    0.2029
       Total |  111.250689     1,077  .103296833   Root MSE        =    .28695

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0270326   .0026572    10.17   0.000     .0218188    .0322464
  experience |    .042783   .0101905     4.20   0.000     .0227874    .0627786
 experience2 |  -.0017684   .0007436    -2.38   0.018    -.0032276   -.0003093
 experience3 |   .0000286   .0000154     1.85   0.064    -1.66e-06    .0000588
       _cons |   4.269921   .0465607    91.71   0.000     4.178561    4.361282
------------------------------------------------------------------------------
```

Lønregression: Stata eksempel

```
test educ
 ( 1)  educ = 0
       F(  1,  1073) =  103.50
            Prob > F =    0.0000

test experience3
 ( 1)  experience3 = 0
       F(  1,  1073) =    3.44
            Prob > F =    0.0640

test experience+experience2+experience3=0
 ( 1)  experience + experience2 + experience3 = 0
       F(  1,  1073) =   18.73
            Prob > F =    0.0000

test experience experience2 experience3
 ( 1)  experience = 0
 ( 2)  experience2 = 0
 ( 3)  experience3 = 0
       F(  3,  1073) =   54.65
            Prob > F =    0.0000
```

Let's Code!

- **Jupyter Notebook:** `04_inference_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 3:** Test af multiple lineære hypoteser (F-test)

Las os alle teste

Fordeling af teststørrelser

Fra normal til χ^2, t og F-fordelinger

Idé: Under normalitetsantagelsen (MLR.6) kan vi udlede fordelingen af teststørrelser.

- **Normal $\rightarrow$ χ^2:** Summen af kvadrerede standard normalfordelte variable er χ^2-fordelt.
  $\Rightarrow$ Residualkvadratsum $SSR/\sigma^2 \sim \chi^2_{n-p}$.
- $\chi^2 \rightarrow t$**:** Koefficienttest bygger på

$$t = \frac{\hat{\beta}_j - a}{\hat{\sigma} \cdot se(\hat{\beta}_j)} = \frac{Z}{\sqrt{W/(n-p)}} \sim t_{n-p},$$

  med $Z \sim N(0,1),\ W \sim \chi^2_{n-p}$.

- $\chi^2 \rightarrow F$**:** Flere restriktioner testes via

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-p)} \sim F_{q,n-p}.$$

Fra normal til χ^2 i regression (og SSR)

Antag (MLR.6) normalitet: $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- Hvis $Z_1, \ldots, Z_m \overset{iid}{\sim} N(0,1)$, så $\sum_{i=1}^{m} Z_i^2 \sim \chi^2_m$.
- I OLS: residualer $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ og $SSR = \sum_{i=1}^{n} \hat{u}_i^2$.
  Der er kun $n - p$ uafhængige residualinformationer (med $p = k + 1$ inkl. konstant). Formelt: $\operatorname{rank}(\mathbf{M}_X) = n - p$ hvor $\mathbf{M}_X = \mathbf{I} - \mathbf{P}_X$.
- Derfor

$$\boxed{\frac{SSR}{\sigma^2} \sim \chi^2_{n-p}} \quad \Rightarrow \quad \hat{\sigma}^2 \equiv \frac{SSR}{n-p} \text{ er unbiased for } \sigma^2,$$

  og

$$(n-p)\,\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Frihedsgraderne $n - p$ skyldes, at estimering af p parametre "forbruger" p grader af frihed.

t-testet: hvorfor t-fordelt, og hvilke frihedsgrader?

Enkelt koefficient j og $H_0 : \beta_j = a$. Fra OLS under normalitet:

$$\hat{\beta}_j \sim N\big(\beta_j,\ \sigma^2 v_j\big), \qquad v_j = \big[(\mathbf{X}'\mathbf{X})^{-1}\big]_{jj}.$$

Trin:

1. **Standardisering (kendt σ):** $Z = \dfrac{\hat{\beta}_j - a}{\sigma\sqrt{v_j}} \sim N(0,1)$.
2. **Erstat σ med residualvarians:** $W = \dfrac{SSR}{\sigma^2} \sim \chi^2_{n-p}$ og $\hat{\sigma}^2 = \dfrac{SSR}{n-p}$, så $\dfrac{W}{n-p} = \dfrac{\hat{\sigma}^2}{\sigma^2}$.
3. **Uafhængighed:** $Z \perp W$ (normalitet + orthogonalitet).

Sammensætning:

$$t = \frac{\hat{\beta}_j - a}{\hat{\sigma}\sqrt{v_j}} = \frac{Z}{\sqrt{W/(n-p)}} \sim t_{n-p}$$

F-testet: hvorfor $F_{q,\ n-p}$ og hvorfor division med q og $n-p$?

Test $H_0:\ \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ med $\mathrm{rank}(\mathbf{R}) = q$ restriktioner. Lad SSR_r og SSR_{ur} være residualkvadratsummer i hhv. restriktet og urestriktet model.

Kernefakta under H_0 og normalitet:

$$\frac{SSR_r - SSR_{ur}}{\sigma^2} \sim \chi^2_q, \qquad \frac{SSR_{ur}}{\sigma^2} \sim \chi^2_{n-p},$$

og disse er uafhængige (ortogonale projektioner).

Definition af F-statistik (middelkvadrater):

$$\boxed{F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-p)} \sim F_{q,\ n-p}}$$

Hvorfor q og $n-p$? Tælleren frigør q parametre $\Rightarrow \chi^2_q$; nævneren estimerer σ^2 med $n-p$ frihedsgrader. Divisionen giver to *middelkvadrater*, hvis forhold er F.

Opsumming

Opsumming

OLS estimaterne er stokastiske variable.

- Forskellige stikprøver giver forskellige estimater

Inferens handler om, hvor sikrer vi kan være på at OLS estimaterne afspejler de sande parameter.

- Fx om estimaterne er signifikant forskellige fra vores 0-hypotese.

Overordnet strategi for hypotesetestning:

- Opskriv en teststørrelse, som burde være 0 under H_0.
- Anerkend at teststørrelse kan være forskellige fra nul pga. stikprøveusikkerhed.
- Find fordelingen af teststørrelsen givet H_0, stikprøvestørrelse mv.
- Find kritiske værdier, p-værdier givet teststørrelse og fordelingen.