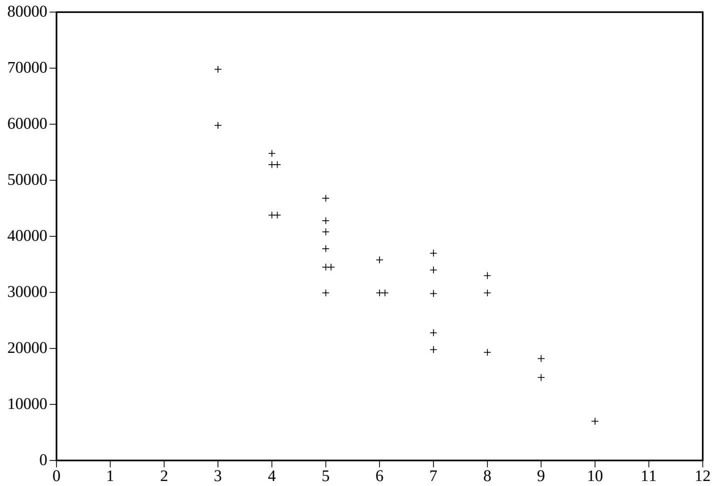
Kapitel 8

REGRESSIONSANALYSE

EKSEMPEL 8.1. I september 1992 var en os nærtstående universitetslærer T i den lykkelige situation, at han havde råd til at købe en brugt bil. Af forskellige grunde havde han besluttet, at det skulle være en Toyota Hiace. For at få et overblik over hvad sådan nogle koster, købte T et eksemplar af Den Blå Avis, og fandt desuden frem fra avisbunken en Blå Avis fra oktober året før. For de i alt 28 Toyota Hiacer som var sat til salg i disse to aviser med bemærkning om, at de ikke umiddelbart skulle til syn, nedskrev han (blandt andre ting, som vi vil vende tilbage til i et senere kapitel) bilens alder i år på annonceringstidspunktet og prisen i kr. Det gav anledning til følgende datasæt:

Alder	Pris	Alder	Pris	Alder	Pris	Alder	Pris
9	14800	10	7000	4	52800	4	43800
8	33000	7	19800	5	42800	4	43800
5	29900	6	29900	5	46800	4	52800
7	29800	7	22800	3	69800	8	29900
5	34500	5	34500	6	35800	6	29900
7	37000	3	59800	9	18200	5	37800
5	40800	4	54800	7	34000	8	19300

Man bliver ikke meget klogere af at kikke på denne tabel. Et meget bedre overblik får man ved at se på en tegning, hvor hver bil repræsenteres ved et punkt, idet bilens alder afsættes ud af x–aksen og dens pris op ad y–aksen:



Pris mod alder for 28 Toyota Hiacer

Ikke forbavsende ser man at prisen aftager med alderen. Det ser ud som om prisen som funktion af alderen er nogenlunde beskrevet ved en linie, der skærer den lodrette akse i højden ca. 80.000 (svarende til at en ny bil skulle koste 80.000 kr), og aftager således at en 10–11 år gammel bil skulle være helt værdiløs.

Med henblik på at bestemme en sådan linie på en lidt mere kvalificeret og objektiv måde, og samtidig få styr på hvor nøjagtigt den er bestemt og hvor stor variation der er omkring den, opstiller vi følgende statistiske model. Lad y_i og x_i betegne henholdsvis pris og alder for bil nr. i. Vi opfatter priserne som observerede værdier af uafhængige stokastiske variable Y_i, der er normalfordelte med samme varians σ^2 og med middelværdier af formen

$$\mathrm{E}Y_i = \alpha + \beta x_i.$$

Bemærk at bilernes aldre (x'erne) i denne forbindelse opfattes som givne tal.

8.1. Den simple regressionsmodel.

Ovennævnte model er et eksempel på en *lineær regressionsmodel*. Vi kalder den *simpel* fordi der kun optræder én *forklarende variabel* x i udtrykket for $\mathrm{E}Y_i$. Ordet *regression* betyder på latin noget i retning af "tilbageskriden". Navnets oprindelse har noget at gøre med genetiske anvendelser, hvor sådanne modeller blev brugt til at beskrive i hvor høj grad nedarvede egenskaber er tilbøjlige til at "skride tilbage" mod et populationsgennemsnit. Men den nutidige brug af ordet har ikke meget med dette at gøre. At analysere en model af denne type kaldes at udføre en regressionsanalyse. Formålet er at forklare *responsvariablen* eller *den afhængige variabel* y ved *den uafhængige variabel* x. I en lidt indforstået statistikerjargon kalder man også dette "at udføre en regression af y'er på x'er" eller — endnu mere kortfattet — "at regressere y'er på x'er".

Generelt ser den simple regressionsmodel sådan ud: Observationerne består af n samhørende værdier af x'er og y'er, altså n talpar (x_i, y_i), $i = 1, \ldots, n$. Modellen går ud på at $y_1, \ldots, y_n$ er observationer af uafhængige normalfordelte stokastiske variable med samme varians σ^2 og middelværdier

$$\mu_i = \mathrm{E}Y_i = \alpha + \beta x_i.$$

Likelihoodfunktionen bliver således

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - (\alpha + \beta x_i)\right)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2\right), \end{aligned}$$

og log–likelihooden (på nær en additiv konstant)

$$l(\alpha, \beta, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2\right).$$

8.2. Estimation.

Maksimeringen af dette udtryk foretages lettest ved først at maksimere m.h.t. α og β for fast σ^2. Maksimaliseringsestimatorerne for α og β får vi åbenbart ved at minimere kvadratsummen

$$K(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

af observationernes afvigelser fra deres middelværdier under modellen. Geometrisk har dette den simple fortolkning, at den linie man lægger ind på tegningen er bestemt ved, at kvadratsummen af de lodrette afstande fra punkterne (x_i, y_i) til linien gøres så lille som muligt.

Den simple regressionsmodel er et eksempel på en *lineær normalfordelingsmodel.* Vi skal senere indse, at maksimaliseringsestimatorerne i sådanne modeller er veldefinerede og bestemt som løsning til likelihoodligningerne, som i denne forbindelse også kaldes *normalligningerne.* Her vil vi nøjes med at opstille disse ligninger og løse dem. Vi begynder med at indføre nogle betegnelser for mellemregningsstørrelser:

$$\mathrm{S}_x = \sum x_i, \qquad \bar{x} = \frac{1}{n}\mathrm{S}_x, \qquad \mathrm{SS}_x = \sum x_i^2,$$

$$\mathrm{SSD}_x = \sum (x_i - \bar{x})^2 = \mathrm{SS}_x - \frac{1}{n}\mathrm{S}_x^2,$$

$$\mathrm{S}_y = \sum y_i, \qquad \bar{y} = \frac{1}{n}\mathrm{S}_y, \qquad \mathrm{SS}_y = \sum y_i^2,$$

$$\mathrm{SSD}_y = \sum (y_i - \bar{y})^2 = \mathrm{SS}_y - \frac{1}{n}\mathrm{S}_y^2,$$

$$\mathrm{SP}_{xy} = \sum x_i y_i, \qquad \mathrm{SPD}_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \mathrm{SP}_{xy} - \frac{1}{n}\mathrm{S}_x\mathrm{S}_y.$$

Her står SP for "**S**um of **P**roducts", SPD for "**S**um of **P**roducts of **D**eviations", og den sidste formel bevises på omtrent samme måde som formlen for SSD'erne udtrykt ved SS'er og S'er.

Med disse betegnelser får vi følgende udtryk for differentialkvotienterne m.h.t. α og β af den kvadratsum vi skal minimere:

$$\frac{d}{d\alpha}K(\alpha, \beta) = -2\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i)) = -2\left(\mathrm{S}_y - n\alpha - \mathrm{S}_x\beta\right),$$

$$\frac{d}{d\beta}K(\alpha, \beta) = -2\sum_{i=1}^{n}x_i(y_i - (\alpha + \beta x_i)) = -2\left(\mathrm{SP}_{xy} - \mathrm{S}_x\alpha - \mathrm{SS}_x\beta\right).$$

Normalligningerne, som sætter disse to differentialkvotienter til 0, kan så skrives

$$\mathrm{S}_y = n\alpha + \mathrm{S}_x\beta,$$
$$\mathrm{SP}_{xy} = \mathrm{S}_x\alpha + \mathrm{SS}_x\beta.$$

Ved at gange den første ligning med S_x/n og trække den fra den anden får vi

$$\mathrm{SP}_{xy} - \frac{1}{n}\mathrm{S}_x\mathrm{S}_y = 0\alpha + \left(\mathrm{SS}_x - \frac{1}{n}\mathrm{S}_x^2\right)\beta,$$

eller

$$\mathrm{SPD}_{xy} = \mathrm{SSD}_x\beta.$$

Maksimaliseringsestimatoren for hældningen β bliver således

$$\hat{\beta} = \frac{\mathrm{SPD}_{xy}}{\mathrm{SSD}_x}.$$

Ved at indsætte dette i den første ligning og løse den m.h.t. α får vi estimatet

$$\hat{\alpha} = \frac{1}{n}\left(S_y - S_x\frac{\mathrm{SPD}_{xy}}{\mathrm{SSD}_x}\right) = \bar{y} - \bar{x}\hat{\beta}$$

for liniens afskæring på y–aksen. Bemærk, som en umiddelbar konsekvens af den sidste ligning, at den således estimerede regressionslinie (altså linien med hældning $\hat{\beta}$ og afskæring $\hat{\alpha}$) går gennem "punktskyens" tyngdepunkt $(\bar{x}, \bar{y})$, hvilket forekommer intuitivt rimeligt.

Ved at indsætte $\hat{\alpha}$ og $\hat{\beta}$ i log–likelihooden får vi nu den delvis maksimerede log–likelihood

$$l(\hat{\alpha}, \hat{\beta}, \sigma^2) = -\frac{1}{2}\left(n\log\sigma^2 + \frac{1}{\sigma^2}\mathrm{SSD}_{\mathrm{res}}\right).$$

hvor

$$\mathrm{SSD}_{\mathrm{res}} = \sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2$$

er *residualkvadratsummen,* altså kvadratsummen af observationernes afvigelser fra deres estimerede middelværdier (de fittede værdier).

Da den delvis maksimerede log–likelihood har præcis samme udseende som den vi udledte i kapitel 7 i forbindelse med modellen for identisk fordelte variable (der hed residualkvadratsummen bare SSD_y), får vi umiddelbart at den antager sin maksimale værdi for

$$\sigma^2 = \sigma^2_{ML} = \frac{\mathrm{SSD}_{\mathrm{res}}}{n}.$$

Ligesom i kapitel 7 benytter man imidlertid et korrigeret variansestimat, i dette tilfælde

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}}{n-2}.$$

Som begrundelse for, at der her skal divideres med $n-2$ i stedet for med n, kan vi her kun henvise til at modellen har to middelværdiparametre (α og β). Forklaring følger senere.

Residualkvadratsummen kan udtrykkes ved de mellemregningsstørrelser vi har indført på følgende måde:

$$\begin{aligned}
\mathrm{SSD}_{\mathrm{res}} &= \sum \big(y_i - (\hat\alpha + \hat\beta x_i)\big)^2 = \sum \big((y_i-\bar y) - \hat\beta(x_i-\bar x)\big)^2 \\
&= \sum (y_i-\bar y)^2 + \hat\beta^2 \sum (x_i-\bar x)^2 - 2\hat\beta \sum (y_i-\bar y)(x_i-\bar x) \\
&= \mathrm{SSD}_y + \hat\beta^2 \mathrm{SSD}_x - 2\hat\beta \mathrm{SPD}_{xy} \\
&= \mathrm{SSD}_y + \left(\frac{\mathrm{SPD}_{xy}}{\mathrm{SSD}_x}\right)^2 \mathrm{SSD}_x - 2\left(\frac{\mathrm{SPD}_{xy}}{\mathrm{SSD}_x}\right)\mathrm{SPD}_{xy} \\
&= \mathrm{SSD}_y - \frac{\mathrm{SPD}_{xy}^2}{\mathrm{SSD}_x} = \mathrm{SSD}_y - \hat\beta^2 \mathrm{SSD}_x.
\end{aligned}$$

EKSEMPEL 8.1 fortsat. I Toyota–Hiace datasættet er alle priser delelige med 100. Det er derfor oplagt at man under beregningerne kan undlade at indtaste de to sidste nuller, svarende til at alle priser regnes i enheden 100 kr. Det gør ikke nogen forskel for modellen — blot skal man selvfølgelig huske det når man bagefter fortolker parameterestimaterne.

Følgende størrelser udregnes:

$$\mathrm{S}_x = 166, \qquad \bar x = 166/28 = 5.9286, \qquad \mathrm{SS}_x = 1080,$$

$$\mathrm{SSD}_x = 1080 - 166^2/28 = 95.8571,$$

$$\mathrm{S}_y = 10059, \qquad \bar y = 10059/28 = 359.25, \qquad \mathrm{SS}_y = 4160311,$$

$$\mathrm{SSD}_y = 4160311 - 10059^2/28 = 546615.25,$$

$$\mathrm{SP}_{xy} = 53183, \qquad \mathrm{SPD}_{xy} = 53183 - 166 \times 10059/28 = -6452.5.$$

$$\mathrm{SSD}_{\mathrm{res}} = 546615.25 - \frac{(-6452.5)^2}{95.8571} = 112273.336$$

Ved tilpas snedig brug af en programmerbar lommeregner kan man udregne alle disse størrelser uden at indtaste observationerne mere end én gang. I hvert trin indtastes en x–værdi og en y–værdi, og derefter opdateres ved hjælp af et passende program fem registre, som skal ende med at indeholde S_x, S_y, SS_x, SS_y og SP_{xy}. SSD'er og SPD'en udregnes bagefter, ved hjælp af ovenstående formler. Mange lommeregnere har i øvrigt et sådant regressionsprogram indbygget.

Vi kan nu udregne maksimaliseringsestimaterne:

$$\hat\beta = \frac{-6452.5}{95.8571} = -67.31374,$$

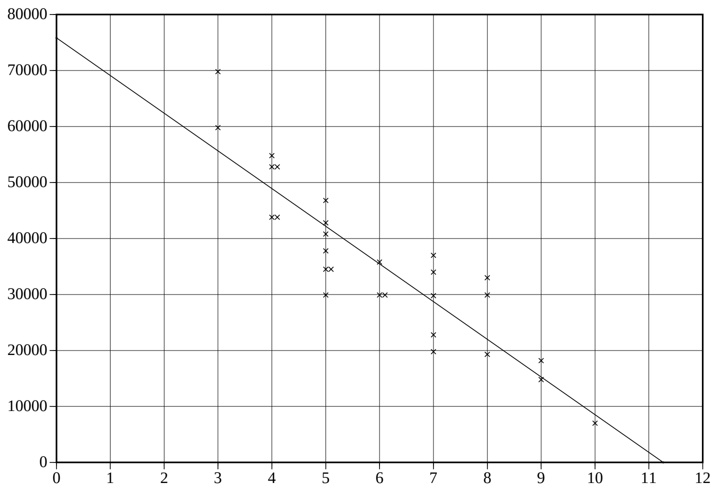
$$\hat{\alpha} = 359.25 + 5.9286 \times 67.31374 = 758.3262$$
$$\hat{\sigma}^2 = \frac{112273.336}{28-2} = 4318.20,$$
$$\hat{\sigma} = \sqrt{4318.20} = 65.71.$$

Konklusion (i passende afrundede tal, og idet vi nu går tilbage til den gængse pengeenhed "1 kr."): De angivne priser på Toyota Hiace'er af alder 3–10 år falder lineært med alderen, svarende til en nyværdi på 76.000 kr. og et prisfald pr. år på 6.700 kr. Prisformlen kan, med angivelse af den variation der findes omkring regressionslinien, skrives

$$\text{PRIS} = (76000 - \text{ALDER} \times 6700)\text{kr.} \pm 13000\text{kr.}$$

Ved angivelse af variationen har vi her benyttet konventionen "plus/minus dobbelt standardafvigelse". Angivelsen er en ret grov vurdering, idet den kun tager hensyn til den enkelte observations variation omkring sin middelværdi, men ikke tager hensyn til at denne middelværdi i sig selv er estimeret med en vis fejl.

Nedenstående tegning viser datasættet med den estimerede regressionslinie indtegnet. Denne tegning (eller selve prisformlen) kan man så bruge til vurdering af, om et forelagt tilbud er rimeligt. Denne vurdering skal selvfølgelig i praksis sammenvejes med forskellige andre ting, men med dette forbehold giver analysen et solidt kvantitativt udgangspunkt for vurdering og prisforhandling. I bilbranchen har man i øvrigt tabeller til den slags. Om de er baseret på regressionsanalyse vides ikke.



Toyota Hiace data med regressionslinien indtegnet

Bemærk: Afskæringen på 76000 kr. skal selvfølgelig ikke opfattes bogstaveligt som prisen på en ny bil. Den er formodentlig dyrere. At konkludere dette ville være *ekstrapolation,* og det kan være farligt. Det vi kan sige er kun, at i aldersintervallet 3–10 år er prisen som funktion af alderen rimeligt godt beskrevet ved en linie.

8.3. Estimatorernes fordeling.

Af resultater vedrørende den generelle lineære normalfordelingsmodel, som vi senere skal bevise, følger det at residualkvadratsummen $\mathrm{SSD}_{\mathrm{res}}$ er χ^2–fordelt med skalaparameter σ^2 og et antal frihedsgrader, der kan udregnes som antallet af observationer minus antallet af parametre i modellen; i dette tilfælde altså med $n-2$ frihedsgrader. Dette er én af begrundelserne for korrektionen af variansestimatet, idet man herved åbenbart får en central estimator. Endvidere gælder det, at hele sættet af residualer (og dermed også residualkvadratsummen og $\hat{\sigma}^2$) er stokastisk uafhængig af estimaterne for middelværdiparametrene — i dette tilfælde $\hat{\alpha}$ og $\hat{\beta}$.

Af omskrivningen

$$\hat{\beta} = \frac{1}{\mathrm{SSD}_x} \sum (x_i - \bar{x})(Y_i - \bar{Y}) = \frac{1}{\mathrm{SSD}_x} \sum (x_i - \bar{x}) Y_i$$

følger det umiddelbart, at $\hat{\beta}$ (som linearkombination af de oprindelige normalfordelte variable) er normalfordelt, med middelværdi

$$\begin{aligned} \mathrm{E}\hat{\beta} &= \frac{1}{\mathrm{SSD}_x} \sum (x_i - \bar{x}) \mathrm{E}Y_i = \frac{1}{\mathrm{SSD}_x} \sum (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \frac{1}{\mathrm{SSD}_x} \Big(0\alpha + \beta \sum (x_i - \bar{x})(x_i - \bar{x}) \Big) = \beta. \end{aligned}$$

Estimatoren $\hat{\beta}$ er altså central — hvilket igen er et specialtilfælde af et generelt resultat for lineære normalfordelingsmodeller, som siger at maksimum–likelihood estimaterne for middelværdiparametrene er centrale. Variansen på $\hat{\beta}$ kan udregnes således:

$$\mathrm{var}(\hat{\beta}) = \left(\frac{1}{\mathrm{SSD}_x} \right)^2 \sum (x_i - \bar{x})^2 \mathrm{var}(Y_i) = \frac{1}{\mathrm{SSD}_x} \sigma^2.$$

Betragt nu størrelsen

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\frac{1}{\mathrm{SSD}_x} \hat{\sigma}^2}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{1}{\mathrm{SSD}_x} \frac{\mathrm{SSD}_{\mathrm{res}}}{n-2}}}$$

I denne brøk er, ifølge tidligere bemærkninger, tæller og nævner stokastisk uafhængige. Tælleren er normalfordelt med middelværdi 0 og varians $\frac{1}{\mathrm{SSD}_x}\sigma^2$, medens nævneren er kvadratroden af en stokastisk variabel som er χ^2–fordelt med $n-2$ frihedsgrader og skalaparameter $\frac{1}{n-2}\frac{1}{\mathrm{SSD}_x}\sigma^2$. Heraf følger umiddelbart ved passende normering (idet vi refererer til T–fordelingens fortolkning som fordelingen af forholdet mellem en normeret normalfordelt variabel, og kvadratroden af en deraf uafhængig χ^2–fordelt variabel divideret med sit frihedsgradsantal, se Ssr. side 127) at T følger en T–fordeling med $n-2$ frihedsgrader. Hvis vi med $t_{n-2}(97.5)$ betegner 97.5%–fraktilen i denne T–fordeling har vi således med sandsynlighed 0.95

$$-t_{n-2}(97.5) \leq \frac{\beta - \hat{\beta}}{\sqrt{\frac{1}{\mathrm{SSD}_x}\hat{\sigma}^2}} \leq t_{n-2}(97.5)$$

eller

$$\hat{\beta} - t_{n-2}(97.5)\sqrt{\frac{1}{\mathrm{SSD}_x}\hat{\sigma}^2} \leq \beta \leq \hat{\beta} + t_{n-2}(97.5)\sqrt{\frac{1}{\mathrm{SSD}_x}\hat{\sigma}^2}.$$

Hermed har vi indset, at eksakte sikkerhedsgrænser for β kan udledes på præcis samme måde som vi gjorde det for parameteren μ i kapitel 7. Metoden, som viser sig at være generelt anvendelig for middelværdiparametre i lineære normalfordelingsmodeller, kan beskrive således:

Differensen mellem maksimum–likelihood estimatoren $\hat{\beta}$ og parameteren β følger en normalfordeling med middelværdi 0 og en varians af formen $c\sigma^2$, hvor c er en konstant som afhænger af modellen (i det konkrete tilfælde $c = 1/\mathrm{SSD}_x$). I en model hvor variansen σ^2 var kendt ville dette umiddelbart føre til 95% konfidensgrænser af formen $\hat{\beta} \pm 1.96\sqrt{c\sigma^2}$ for parameteren β. Hvis man her erstatter σ^2 med $\hat{\sigma}^2$, og samtidig erstatter 1.96 med 95%–fraktilen i den relevante T–fordeling, får man eksakte konfidensgrænser for β i modellen med ukendt varians. Antal frihedsgrader for T–fordelingen er det samme som det der optræder i residualkvadratsummens χ^2–fordeling — altså antal observationer minus antal parametre.

På tilsvarende måde kan man naturligvis angive sikkerhedsgrænser for parameteren α i den simple regressionsanalyse. Vi vil her se på et mere generelt problem, som består i for en given værdi x_0 at angive sikkerhedsgrænser for parameterfunktionen $\mu_0 = \alpha + \beta x_0$, som er middelværdien af en hypotetisk respons Y der svarer til værdien x_0 af den uafhængige variable. Ved at udregne sikkerhedsgrænser for denne parameter får vi, som biprodukt, sikkerhedsgrænser for α (ved at sætte $x_0 = 0$). Middelværdien af $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}x_0$ udregnes til

$$\begin{aligned}\mathrm{E}(\hat{\alpha} + \hat{\beta}x_0) &= \mathrm{E}((\bar{Y} - \bar{x}\hat{\beta}) + \hat{\beta}x_0) = \frac{1}{n}\sum(\alpha + \beta x_i) + (x_0 - \bar{x})\mathrm{E}\hat{\beta} \\ &= \alpha + \beta\bar{x} + (x_0 - \bar{x})\beta = \alpha + \beta x_0.\end{aligned}$$

Estimatoren $\hat{\mu}_0$ er altså også central, i overenstemmelse med det generelle resultat vi tidligere har nævnt. Variansen bliver

$$\mathrm{var}(\hat{\mu}_0) = \mathrm{var}\left(\bar{Y} + \hat{\beta}(x_0 - \bar{x})\right)$$

$$= \operatorname{var}(\bar{Y}) + (x_0 - \bar{x})^2 \operatorname{var}(\hat{\beta}) + 2(x_0 - \bar{x}) \operatorname{cov}(\bar{Y}, \hat{\beta})$$
$$= \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSD}_x} \right) \sigma^2.$$

I den sidste omskrivning har vi benyttet at $\operatorname{cov}(\bar{Y}, \hat{\beta}) = 0$, hvilket følger af at $\hat{\beta}$ er linearkombination af de stokastiske variable $Y_i - \bar{Y}$, der alle er ukorrelede med $\bar{Y}$ (som man let ser ved en direkte udregning). For parameteren $\mu_0 = \alpha + \beta x_0$ har vi således 95% konfidensgrænserne

$$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2}(97.5) \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSD}_x} \right) \hat{\sigma}^2}$$

hvilket specielt for $x_0 = 0$ giver konfidensgrænserne

$$\hat{\alpha} \pm t_{n-2}(97.5) \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SSD}_x} \right) \hat{\sigma}^2}$$

for afskæringen α.

Eksempel 8.1 fortsat. T bliver ringet op af en mand M, og følgende dialog udspiller sig:

M: Ahem, høh, jeg hedder M, jeg ringer fordi jeg hørte at du går og kikker efter en gammel Hiace. Jeg har en fra 85 — nå ja, den er faktisk nærmest fra 86, det er en 85–model jeg købte som ny lige efter årsskiftet. Den er synet for et halvt år siden. Er det noget du er interesseret i?

T: Mnøh, hvad havde du tænkt dig at den skulle koste?

M: Nå, sådan en 38–40.000 kr. havde jeg regnet med.

T (efter følgende lynhurtige overvejelse): Den er fra 85, dvs. 7 år gammel, men vi siger $6\frac{1}{2}$ på grund af det han sagde. Gennemsnitsprisen for sådan en er ifølge tegningen ca. 32.000): Desværre, jeg er ude efter noget i nabolaget af sådan en 20–25.000, mere kan jeg ikke give i øjeblikket.

Vi afbryder denne kedsommelige samtale. Den illustrerer hvad regressionsanalyse kan bruges til i praksis, men hvordan den ender kan sådan set være ligegyldigt. Eksemplet har den svaghed, at det er hentet fra en privatøkonomi, hvor man under normale omstændigheder næppe ville begynde at hyre statistiske konsulenter eller lignende for at købe en brugt bil. Men det kræver ikke megen fantasi at forestille sig tilsvarende modeller anvendt under lidt større forhold, for eksempel i forbindelse med ejendomshandel. Faktisk anvender vurderingsmyndighederne en lignende model (men naturligvis med flere forklarende variable) til fastsættelse af ejendomsværdier.

Til illustration af beregningen af konfidensgrænser kan man, i relation til dette eksempel, stille et mere præcist spørgsmål: Hvad er estimatet for den forventede værdi af prisen for en en 6.5 år gammel bil ifølge modellen, og hvor stor usikkerhed er der på dette estimat?

For parameteren $\alpha + \beta \times 6.5$ har vi ifølge tidligere udregninger 95% konfidensgrænserne

$$75832 - 6731 \times 6.5 \pm 2.056\sqrt{\left(\frac{1}{28} + \frac{(6.5 - 5.93)^2}{95.86}\right) \times 43182000}$$

eller

$$32080 \pm 2672.$$

Vi kan altså med rimelig sikkerhed gå ud fra, at den gennemsnitlige markedspris for en $6\frac{1}{2}$ år gammel bil ligger (i runde tal) mellem 29400 og 34750 kr. Så det er korrekt af T at vurdere den opgivne pris på 38–40.000 til at være i overkanten. Samtidig skal dog bemærkes at den kun ligger ca. $(38000 - 32080)/6571 = 0.9$ standardafvigelser over middelværdien, så isoleret set er prisen ikke ekstremt høj. Hvis M ikke er til at rokke, men bilen i øvrigt viser sig at have egenskaber som kan begrunde en merpris på ca. 6000 i forhold til gennemsnittet, kan tilbuddet være værd at overveje.

8.4. Test for hældning = 0.

Den mest interessante hypotese i forbindelse med den simple regressionsanalysemodel er hypotesen om, at regressionslinien har hældning 0. I så fald er vi tilbage i modellen for identisk fordelte normale observationer (med $\mu = \alpha$). Vi kender derfor, fra kapitel 7, den maksimale værdi af log–likelihooden under denne hypotese. Ved udregninger som er helt analoge til dem vi foretog på side 65 i kapitel 7 — og som vi derfor vil springe over — ser man, at kvotienttestet for denne hypotese er ækvivalent med en tosidet vurdering af teststørrelsen

$$t = \frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\mathrm{SSD}_x}}}$$

i en T–fordeling med $n-2$ frihedsgrader. Sammenhængen mellem testet og opstillingen af konfidensgrænser er den samme som vi har set et par gange før.

Eksempel 8.1 fortsat. I Toyota Hiace eksemplet er det ikke så interessant at teste hypotesen $\beta = 0$, da tegningen tydeligt viser at der er sammenhæng mellem alder og pris, og hypotesen iøvrigt er komplet idiotisk ud fra almindelig sund fornuft. Men vi gør det alligevel, for øvelsens skyld. Vi får

$$t = \frac{-6731}{\sqrt{\frac{43182000}{95.86}}} = -10.029$$

som ligger et godt stykke til venstre for 0.01% fraktilen i T–fordelingen med 26 frihedsgrader. Den eksakte P–værdi er faktisk af størrelsesorden 10^{-10}. Så der *er* sammenhæng mellem pris og alder.

OPGAVE 8.4.1. Nedenstående data er taget fra Peter Tryfos: Methods for Business Analysis and Forecasting, Wiley 1998. Data vedrører flyruter fra "city A" (formodentlig i USA, hvis datamaterialet overhovedet er autentisk) til 17 destinationer. De fem oplysninger i hver linie er

Destination (1 , . . . , 17).
Pris for enkeltbillet i $.
Afstand i miles.
Antal passager (sandsynligvis på årsbasis).
Destinationsbyens indbyggerantal.

1	360	1463	85830	444000
2	360	1448	68780	529000
3	207	681	83050	224000
4	111	270	482430	2798000
5	93	190	247540	626000
6	141	393	37020	499000
7	291	1102	29920	151000
8	183	578	20210	112000
9	309	1204	23340	130000
10	300	1138	33640	132000
11	90	184	45840	154000
12	162	502	71150	112000
13	477	1828	18130	208000
14	84	179	46890	266000
15	231	818	117830	570000
16	54	90	11150	296000
17	429	1813	146570	1137000

Opstil en simpel regressionsmodel der beskriver billetprisen som funktion af afstanden (de to sidste variable ignorerer vi i øjeblikket). Estimer parametrene og fortolk dem økonomisk.

OPGAVE 8.4.2. Analyser passagertallet som funktion af destinationsbyens indbyggerantal i data fra opgave 8.4.1 ved hjælp af en simpel regressionsmodel. Dette er et eksempel hvor hypotesen "afskæring=0" har en rimelig fortolkning (forklar, og udfør testet).