Stockholms universitet

Analys av diamantpriser med multipel linjär regression

Oskar Söderlund

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:3
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:3**
http://www.math.su.se/matstat

Analys av diamantpriser med multipel linjär regression

Oskar Söderlund*

Juni 2016

Sammanfattning

I det här examensarbetet undersöker vi vilka variabler som påverkar priset på diamanter. Målet med arbetet är att hitta en linjär modell, med hjälp av multipel linjär regression, som på bästa sätt förklarar variationen av priset på diamanterna. Det visar sig att sambandet mellan diamanternas pris och förklarande variabler ej kan uttryckas linjärt och vi använder oss istället av en multiplikativ modell härledd från en logaritmtransformation av responsvariabeln. Grundmodellen bestod av sju förklarande variabler som vi tror kan ha en påverkan på priset. Efter vi upptäckt att flertalet variabler haft stark korrelation har vi slagit samman dessa tre variabler till ett gemensamt medelvärde. Vi har även behövt använda en kvadratisk term av diamanternas carat för att få en bättre modellanpassning. Undersökning av modellens residualer plottade mot det predikterade värdet visade på att små och stora diamanter ej hade ett kontinuerligt samband vilket resulterade i en uppdelning av observationerna och totalt fyra stycken modeller testades för att finna de två mest lämpliga. De resulterande modellerna pekade mot att vikten var den variabel som mest förklarade prisets variation. Klarheten var mer inflytelserik än färgen för små diamanter och vice versa för stora. Diamanternas certifiering visade sig ha betydelse på priset även fast det inte troddes ha det.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: Oskar_soderlund@hotmail.com. Handledare: Jan-Olov Persson, Gudrun Brattström.

Sammanfattning

I det här examensarbetet undersöker vi vilka variabler som påverkar priset på diamanter. Målet med arbetet är att hitta en linjär modell, med hjälp av multipel linjär regression, som på bästa sätt förklarar variationen av priset på diamanterna. Det visar sig att sambandet mellan diamanternas pris och förklarande variabler ej kan uttryckas linjärt och vi använder oss istället av en multiplikativ modell härledd från en logaritmtransformation av responsvariabeln. Grundmodellen bestod av sju förklarande variabler som vi tror kan ha en påverkan på priset. Efter vi upptäckt att flertalet variabler haft stark korrelation har vi slagit samman dessa tre variabler till ett gemensamt medelvärde. Vi har även behövt använda en kvadratisk term av diamanternas carat för att få en bättre modellanpassning. Undersökning av modellens residualer plottade mot det predikterade värdet visade på att små och stora diamanter ej hade ett kontinuerligt samband vilket resulterade i en uppdelning av observationerna och totalt fyra stycken modeller testades för att finna de två mest lämpliga. De resulterande modellerna pekade mot att vikten var den variabel som mest förklarade prisets variation. Klarheten var mer inflytelserik än färgen för små diamanter och vice versa för stora. Diamanternas certifiering visade sig ha betydelse på priset även fast det inte troddes ha det.

Abstract

In this bachelor thesis, we examine the variables that affect the price of diamonds. The purpose of this thesis is to find a linear model, using multiple linear regression, which best explains the variation in the price of diamonds. It turns out that the relationship between the diamond price and explanatory variables can not be expressed linearly, we instead use a multiplicative model derived from a logarithm transformation of the response variable. The basic model consisted of seven explanatory variables that we believed had an impact on the price. After we discovered that most of the variables had a strong correlation, we combined these three variables to a common average. We also had to use a quadratic term of the diamond carat to get a better adaptation of the model. Examination of the models residuals plotted against the predicted value showed that small and big diamonds did not contain a continuous relationship which led to a division of the observations and a total of four models were tested to find the two most appropriate. The resulting models indicated that a diamonds weight was the variable that explained the variation of the price the most. The clarity was more influential than the color of small diamonds, and vice versa for the large diamonds. A diamonds certificate proved to be a significant factor for the price even though it was believed that it would not be.

Förord och tack

Denna kandidatuppsats i matematisk statistik omfattar 15 högskolepoäng och är utförd vid matematiska institutionen på Stockholms universitet.

Jag vill rikta ett tack till mina två handledare Gudrun Brattström och Jan-Olov Persson för en värdefull handledning med god kommunikation. Jag vill även tacka de kurskamrater som bidragit med synpunkter, tips och stöd.
Ni vet vilka ni är.

Innehåll

1 Introduktion

Diamanter är en ädelsten som består av hårt pressad kol och är den hårdast mineralen som förekommer i naturen. Diamanter kan säljas lösa, men monteras oftast på någon form av smycke innan försäljning. Vi kommer i det här arbetet endast undersöka lösa, runt slipade diamanter. Diamanter kan antingen vara konstgjorda eller skapade av naturen självt. Skillnaden är mycket svår att upptäcka, men generellt sett är konstgjorda diamanter något mer gulaktiga i färgen samt mindre till storleken[1]. Diamanterna som kommer analyseras i detta arbete är hämtade från Brilliance.com där de garanterar att diamanterna ej är konstgjorda. Brilliance är en världsledande återförsäljare av diamanter som samarbetar tillsammans med hundratals diamantslipningsföretag och gemmologer världen över. Uppfattningen hos allmänheten är att större diamant betyder dyrare diamant, detta behöver inte alls vara sant då diamanter har mängder av karaktäristiska drag som både kan höja och sänka diamantens värde. Vanligtvis brukar man tala om "The four C's", dvs. colour, cut, clarity och carat (färg, skärning, klarhet och vikt), som de styrande faktorerna till priset. Vi kommer i detta arbete undersöka hur väl det stämmer.

1.1 Mål och syfte

Målet med det här arbetet är att ta fram en modell som på bästa sätt förklarar priset på diamanter samt få förståelse för vilka variabler som inverkar på priset och hur de gör det. Syftet med arbetet är att ta fram en modell som skulle kunna användas i praktiken genom t. ex. värdering, men även att fördjupa mina kunskaper inom regression och processen av ett projektorienterat arbete.

De frågor som väntas bli besvarade genom arbetet:
- Hur kan priset på en diamant förklaras bäst?
- Vilka variabler bidrar mest/minst till ett ökat pris?
- Har variablerna samma effekt på små samt stora diamanter?

2 Teori

I detta avsnitt kommer vi gå igenom den allmänna teorin för multipel linjär regression och de metoder som kommer användas genom modelleringens gång.

2.1 Linjär regression

Linjär regression används till att försöka finna ett linjärt samband mellan en, eller flera, förklarande variabler och dess responsvariabel. I arbetet kommer undersökning av flertalet förklarande variabler ske för att se om de tillsammans kan beskriva priset på en diamant.

2.1.1 Multipel linjär regression

Den allmänna modellen för multipel linjär regression beskrivs nedan som:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \epsilon_i, \quad i = 1, \ldots, n$$

där ϵ_i är det oberoende och normalfördelade försöksfelet med väntevärde 0 och varians σ^2, $\epsilon_i \sim N(0, \sigma^2)$.
Multipel linjär regression kan också skrivas på matrisformen $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ som:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Det är sedan vektorn $\boldsymbol{\beta}$ samt variansen av vektorn $\boldsymbol{\epsilon}$ vi kommer skatta med hjälp av minstakvadrat-metoden, vilket förhoppningsvis leder till att responsvariabeln kan beskrivas av de förklarande variablerna med ett linjärt samband utan för stora avvikelser.
Vi kommer att gå in djupare på minstakvadrat-metoden i kapitel (2.1.3).

2.1.2 Hypotestest

Hypotesen vi kommer testa är om de förklarande variablerna har någon inverkan på responsvariabeln, dvs:

$$H_0 : \beta_i = 0,$$

mot alternativhypotes,

$$H_a : \beta_i \neq 0.$$

Hypotesten testas m.h.a. teststatistikan

$$t = \frac{\hat{\beta}_i - \beta_{H_0}}{SE_{\hat{\beta}_i}} \sim t(n-k), \tag{1}$$

där β_{H_0} är β_i under nollhypotesen, dvs. 0, och $(n-k)$ är antalet frihetsgrader, där k är antalet variabler i modellen (exklusive interceptet).

2.1.3 Parameterskattning

För att skatta våra parametrar kan vi använda oss av minstakvadrat-metoden, vilken går ut på att minimera avståndet mellan våra observationer och den skattade regressionsvektorn, alltså att göra våra residualer så små som möjligt genom beskrivningen av en regressionsvektor. Skattningarna för alla β_i skapar då en modell med minsta möjliga residualkvadratsumma. Minstakvadratskattningen, $\hat{\beta}$, tas fram med hjälp av formeln:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

För att sedan minimera residualkvadratsumman använder vi våra $\hat{\beta}$ i regressionsmodellen och får ut de skattade värdena för våra observationer, $\hat{y} = X\hat{\beta}$. Det minsta kvadrerade avståndet mellan våra faktiska och skattade värden på observationerna ger då residualkvadratsumman:

$$RSS = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n \hat{\epsilon}_i^2.$$

2.2 Om modellval

Det är många faktorer som spelar in när det kommer till att välja en lämplig modell för sin data. Syftet med en modell är ofta att prediktera framtida värden och att samtidigt göra det effektivt. Det är självklart viktigt att ens modell är väl anpassad till data, men den får samtidigt inte vara för komplicerad då det bör finnas en bakomliggande teoretisk motivering för användandet av de förklarande variablerna samt att svåra beräkningar tar tid, vilket i längden kan bli kostsamt för användaren. Av samma anledning bör modellen inte heller innehålla för många förklarande variabler.

"Förenkla så mycket som möjligt, men inte mer än så". -Albert Einstein.

De faktorer och "verktyg" jag kommer använda i detta arbete är listat i de följande del-kapitlen.

2.2.1 R^2 och R^2_{adj}

Förklaringsgraden R^2 definieras som den andel av den totala variationen som modellen "förklarar" [4, s.69]. Förklaringsgradens värde varierar mellan 0 och 1, där ett högre värde talar för att en modell har bättre anpassning till observationerna. Förklaringsgradens värde i sig kan för det mesta inte användas för att säga om en modell är bra eller dålig, utan används mer frekvent i bemärkelsen att jämföra olika modeller med varandra.
Formeln för R^2 är följande:

$$R^2 = \frac{Kvs(regression)}{Kvs(total)} = 1 - \frac{Kvs(residual)}{Kvs(total)}$$

När vi tillför en ytterligare förklarande variabel så ökar alltid R^2, även fast den förklarande variabeln i sig inte är särskilt förklarande. Därför används ofta istället det justerade förklaringsgrad-måttet, R^2_{adj}, som tar hänsyn till antalet förklarande variabler i modellen och hur stor variansreduktion vi får. R^2_{adj} har följande samband med R^2:

$$1 - R^2_{adj} = (1 - R^2)\frac{df_{total}}{df_{residual}}. \tag{2}$$

2.2.2 Korrelation, multikollinearitet och variansinflationsfaktorn

Korrelation mellan en eller flera variabler innebär att ett linjärt samband mellan variablerna existerar. T. ex. kan variabler som *"antal steg per dag"* och *"Andel fett av kroppsvikt"* vara negativt korrelerade, dvs. Den ena variabeln påverkar den andre. På motsatt sätt kan variabler som *"Genomsnittsbetyg"* och *"Studietid per dag"* vara positivt korrelerade. Riktningen och styrkan för korrelationen mäts vanligtvis med "Pearsons korrelationskoefficient";

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

där $\rho_{X,Y}$ går mellan -1 och 1 för variablerna X och Y. Ett $\rho_{X,Y} \geq 0.5$ eller $\rho_{X,Y} \leq -0.5$ visar på ett tydligt linjärt samband mellan variablerna [2]. Hög korrelation behöver inte vara någonting negativt men kan användas för att förenkla modeller antingen genom reduktion eller sammanslagning av variablerna.

En liknande statistik term är multikollinearitet, vilket förekommer när två eller flera förklarande variabler har ett linjärt samband som kan uttryckas med linjärkombinationer av varandra. D.v.s. de beskriver responsvariabeln på ett liknande sätt. Ett problem med multikollinearitet är att två kollineära variabler, som var för sig har ett signifikant samband med responsvariabeln, tillsammans kan visa sig vara insignifikanta, men också vice versa.
För att lösa ett kollinearitetsproblem kan man till exempel sätta samman de kollineära variablerna med dess medelvärde eller helt enkelt ta bort en av de förklarande variablerna. Variansinflationfaktorn (VIF), är ett statistiskt mått som beräknar hur mycket större variansen av regressionskoefficienten $\hat{\beta}_i$ för en variabel har i kombination av de andra förklarande variablerna. Om en variabel är multikollinjär får den ett högt VIF-värde, och på så vis kan VIF påvisa om multikollinearitet förekommer. Formeln för VIF är definierad som:

$$VIF = \frac{1}{1 - R^2_i}, \tag{3}$$

där ett VIF-värde över 5 eller 10 tyder på att mulltikollinearitet råder mellan de förklarande variablerna. R^2_i är den förklaringsgrad som talar för hur mycket variationen i x_i som kan förklaras av alla de övriga förklarande variablerna.

2.2.3 Cooks avstånd

Cooks avstånd,

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^T S(\hat{\beta}_{(i)} - \hat{\beta})/m\hat{\sigma}^2, \tag{4}$$

är ett mått på en observations inflytande på regressionen som mäter effekten på $\hat{\beta}$ av att utesluta observationen i fråga[4, s.78]. I formeln är $\hat{\beta}_{(i)}$ skattningen när observation i hålls utanför datamaterialet, m-antal förklarande variabler och $S = X^T X$. Cooks avstånd har ett tröskelvärde där observationer över detta tröskelvärde anses vara väldigt inflytelserika på skattningen. Vanligtvis befinner sig en stor andel av observationerna ovanför detta tröskelvärde, det är därför mest användbart att bara kontrollera de observationer med högst värden på Cooks avstånd. En kontroll utförs för att undersöka ifall observationen kan anses felaktig eller för extrem i något anseende, man bör då fundera över ifall man ska behålla observationen eller ej.

2.2.4 Stegvis variabelselektion

När vi vill bestämma vilka av våra variabler som ska användas i modellen tittar vi främst på p-värdet som genererats utav vårt t-test(1). Förhåller sig skattningens p-värde under önskvärd signifikansgrad, oftast 5%, så väljer vi att inkludera variabeln i modellen. Detta kan göra automatiskt med hjälp av olika programpaket och det är främst tre metoder olika metoder som används. *Stepwise regression* är en metod som går ut på att stegvis addera signifikanta variabler till modellen för att där efter analysera den aktuella modellen i jakt efter att reducera insignifikanta variabler. *Stepwise regression* är en kombination av de två andra vanligt förekommande metoderna, *backward elimination* och *forward selection*, som går ut på att antingen börja med en full modell, innehållande alla valda variabler, och stegvist reducera de insignifikanta variabler, eller att börja med en tom modell för att sedan stegvist addera de variabler som är mest signifikanta för modellen. I detta arbete kommer vi att använda oss av *Stepwise regression*.

2.2.5 Residualer

Residualerna uttrycks som $\hat{\epsilon}_i = y_i - \hat{y}_i$, där antagandet för regressionen är att feltermerna är oberoende och normalfördelade med samma varians och väntevärde 0, som nämnt i kapitel 2.1.1. En viktig plot att undersöka om man vill kontrollera sina residualer är när residualerna plottas mot de predikterade värdena $\hat{y}$. Residualerna visualiseras då horisontellt jämtmed väntevärdet och gör det lätt att upptäcka ifall residualerna beror på x-axelns värden eller ej. Antas residualerna följa någon slags form kan transformationer av de förklarande variablerna komma till pass, mer om det i kapitel 2.3. Ett annat residualrelaterat problem man kan tänkas stöta på är *"heteroskedasticitet"*, vilket innebär att residualernas storlek antingen ökar eller minskar när $\hat{y}$ ökar, residualerna antar en kon-form. Antagandet av lika fördelade residualer uppfylls då inte, vilket

kan resultera i felaktiga slutsatser från regressionsmodellen. Det vi strävar efter är *"homoskedasticitet"*, vilket innebär att residualerna har samma varians.

2.3 Variabler och transformationer

När man arbetar med "riktig" data är det sällsynt att datan lätt kan matas in och på en gång köras i diverse programpaket för att sedan få fram en fungerande modell. Det krävs oftast mycket arbete med variablerna i form av transformationer och omvandlingar av de värden variablerna kan anta. I följande två delavsnitt går jag igenom några av de metoder som vi senare kommer tillämpa.

2.3.1 Dummy-variabler

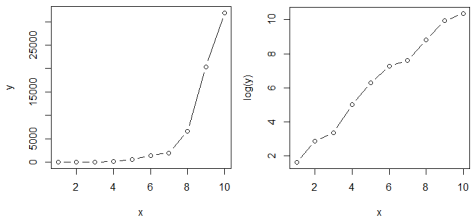
En dummy-variabel omvandlar både kategoriska- och numeriska variablers värden till antingen 1 eller 0. En modell innehållande en dummy-variabel kan se ut som följande:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 D_{1i} + \epsilon_i, \quad i = 1, ..., n. \tag{5}$$

Där x_{1i} är en numerisk variabel och D_{1i} är en dummy-variabel. Exempelvis kan dummy-variabeln för *"Kön"* anta värdet 1 för kvinnor och 0 för män. Med dummy-variabeln förklarar koefficienten för variabeln *"Kön"* effekten på modellen för "egenskapen" kvinna. När dummy-variabeln för *"Kön"* är 0 motsvarar det en man vilket tas med i skattningen för interceptet. Man kan även använda dummy-variabler för numeriska variabler då t. ex. personer över åldern 50 ansätts till värdet 1 och personer under 50 år ansätts till värdet 0. Dummy-variabler är alltså ett bra sätt att jämföra skillnader mellan olika grupper eller att förenkla en variabel där man vet att det finns stora skillnader mellan värdena på variabeln.

2.3.2 Logaritmtransformationer

Vi strävar ett linjärt samband mellan de förklarande variablerna och responsvariabeln. När en responsvariabel har ett exponentiellt samband till de förklarande variablerna, eller vice versa, så kan en logaritmtransformering normalisera den skeva fördelningen för variabeln och göra sambandet mellan respons- och förklarande variabeln mer linjär. I 'Figur 1' nedan plottas responsvariabeln Y mot variabeln X, då Y har ett exponentiellt samband till X, och hur en logaritmering av Y skapar ett mer linjärt förhållande mellan dem.



(a) Relationen mellan Y och X.

(b) Relationen mellan $Log(Y)$ och X.

Figur 1: Punktplottar mellan Y och X före och efter logtransformering av Y.

3 Data

Datamängden är manuellt hämtad från Brilliance.com[3] som är en hemsida där diamantmäklare med, enligt dem själva, väldigt litet prispåslag köper in diamanter från över 250 olika diamantslipningsföretag och sedan lägger ut dem till försäljning för den stora massan. Data består av 1350 stycken slumpvist utvalda runda diamanter från ett större urval på upp emot 84 000 stycken diamanter. Diamanterna är till största delen likformigt insamlade efter dess carat med ca. 30 diamanter inom varje delintervall om 0.1. Alltså 30 diamanter på 0.2ct, 30 diamanter på 0.3ct osv. upp till 5.0ct. Diamanternas övriga variabler, som pris, färg etc. har inte tagits i åtanke under insamlingen och antar därför den fördelning som de naturligen besitter. Efter diamanternas slipning skickas de till olika laboratorier där diamanterna får sin certifiering efter att varje variabel betygsatts.
I delkapitel *3.1* listas alla variabler upp tillsammans med beskrivningar.

3.1 Ursprunglig data

Pris- Diamanternas pris till försäljning som bestäms av Brilliance.

Carat- Diamanternas vikt, uttryckt i enheten carat(ct), där 0.2 gram motsvarar 1.0 carat.

Färg- Diamanternas färg på en 10-grading skala D-M, där D är färglös och anses vara mest värdefull och M är svagt gulaktig och då minst värdefull.

Symmetri- Hur symmetrisk diamanten är, bestäms på en 5-gradig skala där ”Ideal” är mest symmetrisk och ”Fair” är minst symmetrisk.

Skärning- Hur diamanten är skuren, en bra skärning får diamanten att reflektera ljuset på ett önskvärt vis samt vara stryktålig. Bestäms på en 5-gradig skala där ”Super ideal” är bäst och ”Good” är sämst.

Polering- Hur lent putsad ytan på diamanten är, undersöks med mikroskåp. Bestäms på en 5-gradig skala där ”Ideal” är bäst polerad och ”Fair” är sämst polerad.

Klarhet- Om diamanten har invärtes och/eller ytliga missfärgningar, repor, grumlig nyans etc. Bestäms på en 10-gradig skala där ”FL” (Flawless) är mest klar och ”I1” (Included 1) är minst klar.

Certifiering- Vilket laboratorium som har betygsatt ovanstående variabler. Fem olika certifikat där certifikaten ej är rangordnade.

I 'Tabell 1' nedan listas alla variabler upp med de värden som de kan anta. Variablernas värden är rangordnade efter hur bra dem anses, med sämsta värden högst upp i kolumnen och bäst längst ner. Det enda undantaget för rangordningen är variabeln *Certifiering* då certifikaten, enligt hemsidan, ska anses likvärdiga.

Pris	*Vikt(Carat)*	*Färg*	*Symmetri*	*Skärning*	*Polering*	*Klarhet*	*Certifiering*
\$394	0.2 ct.	M	Fair	Good	Fair	I1	GIA
⋮	⋮	L	Good	Very good	Good	SI3	AGS
\$537,964	5.0 ct.	K	Very good	Excellent	Very good	SI2	EGL
		J	Excellent	Ideal	Excellent	SI1	HRD
		I	Ideal	Super ideal	Ideal	VS2	IGI
		H				VS1	
		G				VVS2	
		F				VVS1	
		E				IF	
		D				FL	

Tabell 1: Tabell över variabler

3.2 Transformationer och behandling av variabler

Eftersom sex av de sju förklarande variablerna är kategoriska så kommer vi använda någon form av numerisk omskrivning för de variablerna. De kategoriska variablerna *Färg, Symmetri, Skärning, Putsning* och *Klarhet* har alla en ordinalskala, dvs. variablernas värden kan rankas kvalitativt, därför är det lämpligt att omvandla de kategoriska variablerna till numeriska. Vi vill att våra estimat sedan ska vara lättolkade, därför omvandlar vi de kategoriska variablerna till numeriska med värden mellan 0 och 1 på följande vis:

Symmetri	**Numeriskt värde**
Fair	$1/5 = 0.2$
Good	$2/5 = 0.4$
Very good	$3/5 = 0.6$
Excellent	$4/5 = 0.8$
Ideal	$5/5 = 1.0$

T. ex. Har då färgen "D" numera värdet 1.0 och färgen "M" har värdet 0.1.
Eftersom värdena på variabeln *Certifiering* ska anses vara jämlika och att majoriteten av alla diamanter är certifierade av *"GIA"* så omvandlar vi variabeln till en dummyvariabel(5) med värdet 1 när diamanterna är certifierade av *"GIA"* och 0 annars. Vi gör detta för att verkligen se om certifikatet saknar betydelse.

Nu när vi har omvandlat våra variabler så undersöker vi vidare om några av variablerna korrelerar. I 'Figur 2' nedan ser vi "Pearson Correlation Coefficients" mellan alla förklarande variabler.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Carat	Farg	Polering	Symmetri	Skarning	klarhet	GIA
Carat	1.00000 1351	-0.26197 <.0001 1351	0.10602 <.0001 1350	0.14248 <.0001 1350	0.13555 <.0001 1351	-0.17781 <.0001 1351	-0.00746 0.7841 1351
Farg	-0.26197 <.0001 1351	1.00000 1351	0.04075 0.1345 1350	0.01125 0.6797 1350	0.02785 0.3063 1351	0.16762 <.0001 1351	0.08087 0.0029 1351
Polering	0.10602 <.0001 1350	0.04075 0.1345 1350	1.00000 1350	0.63571 <.0001 1350	0.63519 <.0001 1350	0.14549 <.0001 1350	0.11504 <.0001 1350
Symmetri	0.14248 <.0001 1350	0.01125 0.6797 1350	0.63571 <.0001 1350	1.00000 1350	0.72037 <.0001 1350	0.10155 0.0002 1350	0.16092 <.0001 1350
Skarning	0.13555 <.0001 1351	0.02785 0.3063 1351	0.63519 <.0001 1350	0.72037 <.0001 1350	1.00000 1351	0.14594 <.0001 1351	0.19864 <.0001 1351
klarhet	-0.17781 <.0001 1351	0.16762 <.0001 1351	0.14549 <.0001 1350	0.10155 0.0002 1350	0.14594 <.0001 1351	1.00000 1351	0.11536 <.0001 1351
GIA	-0.00746 0.7841 1351	0.08087 0.0029 1351	0.11504 <.0001 1350	0.16092 <.0001 1350	0.19864 <.0001 1351	0.11536 <.0001 1351	1.00000 1351

Figur 2: Variablernas korrelation

Vi ser att framför allt tre variabler är starkt positiv korrelerade med varandra (markerade med rött). Variablerna *Polering*, *Symmetri* och *Skärning*, som är de korrelerande variablerna, är alla variabler som har med ytbehandlingen att göra, alltså hur väl diamantsliparna arbetat på diamanten. Vi skapar därför en ny variabel, som vi kallar *"Ytbehandling"* som består av medelvärdet av variablerna *Polering*, *Symmetri* och *Skärning* tillsammans:

$$Ytbehandling = \frac{Polering + Symmetri + Skärning}{3}.$$

'Figur 3' nedan visar korrelationerna efter omvandlingen av variablerna:

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	Carat	Farg	ytbehandling	klarhet	GIA
Carat	1.00000 1351	-0.26197 <.0001 1351	0.14599 <.0001 1350	-0.17781 <.0001 1351	-0.00746 0.7841 1351
Farg	-0.26197 <.0001 1351	1.00000 1351	0.02875 0.2911 1350	0.16762 <.0001 1351	0.08087 0.0029 1351
ytbehandling	0.14599 <.0001 1350	0.02875 0.2911 1350	1.00000 1350	0.14981 <.0001 1350	0.18793 <.0001 1350
klarhet	-0.17781 <.0001 1351	0.16762 <.0001 1351	0.14981 <.0001 1350	1.00000 1351	0.11536 <.0001 1351
GIA	-0.00746 0.7841 1351	0.08087 0.0029 1351	0.18793 <.0001 1350	0.11536 <.0001 1351	1.00000 1351

Figur 3: Variablernas korrelation

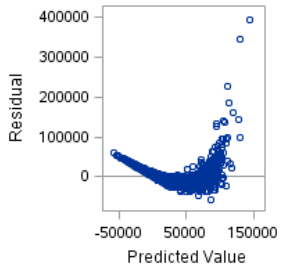
Efter omvandlingen har vi inte längre några variabler som är starkt korrelerade, vilket är önskvärt.

Vidare har vi vår viktigaste variabel, responsvariabeln *Pris*, som måste undersökas. Vi utför en multipel linjär regression på vår modell efter de tidigare transformationerna:

$$Pris = \alpha + \beta_1 Carat + \beta_2 Färg + \beta_3 Klarhet + \beta_4 ytbehandling$$

$$+\beta_5(Dummyvariabel\ för\ GIA) + \epsilon,$$

och undersöker residualerna plottade mot det predikterade värdet av *Pris*, vilket visas i 'Figur 4' nedan.



Figur 4: Residualer plottade mot det predikterade värdet av Priset.

Plottens punkter i 'Figur 4' följer ett exponentiellt mönster vilket motiverar en logaritmering av responsvariabeln *Pris*. Vi transformerar därför *Pris* till *log(Pris)* vilket kommer göra det lättare att anpassa en linjär modell.

3.3 Slutgiltiga variabler

Våra slutgiltiga variabler vi kommer arbeta vidare med i *Kapitel 4* beskrivs nedan i 'Tabell 2'.

Variabel	**Typ**	**Min.Värde**	**Max.Värde**
log(Pris)	Responsvariabel	5.976	13.196
Carat	Numerisk	0.2	5.0
Färg	Numerisk	0.1	1.0
Klarhet	Numerisk	0.1	1.0
Ytbehandling	Numerisk	0.2	1.0
Certifikat	Dummy-variabel	0	1

Tabell 2: Slutgiltiga variabler.

4 Statistisk modellering

Med hjälp av teorin i 'Avsnitt 2' samt vår behandling av data i 'Avsnitt 3' kommer vi i följande delavsnitt arbeta oss fram till en modell som bäst förklarar priset på diamanterna utifrån deras variabler och sedan analysera resultatet.

4.1 VIF och residualer.

Vi startar modelleringen med att undersöka samtliga variablers VIF-värden(3) för att ta reda på om multikollinearitet råder. Resultaten visas i 'Tabell 3' nedan.

Variabel	VIF
Carat	1.134
Färg	1.099
Klarhet	1.088
Ytbehandling	1.091
Certifikatdummy	1.049

Tabell 3: Tabell över variablernas VIF-värden.

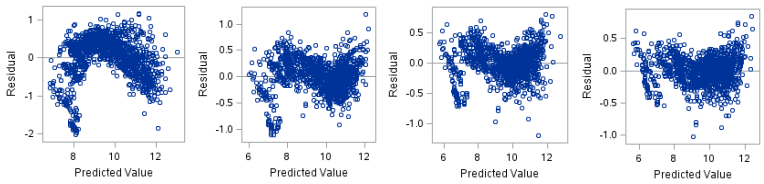
Eftersom inget av variablernas VIF-värde överstiger 5 så kan vi för tillfället utesluta multikollinearitet i modellen. Skulle införandet av transformationer senare behövas kan vi behöva undersöka variablernas VIF-värden igen.

Vår grundmodell är uttryckt som:

$$log(Pris) = \alpha + \beta_1 Carat + \beta_2 Färg + \beta_3 Klarhet + \beta_4 Ytbehandling$$

$$+\beta_5 Certifikatdummy + \epsilon,$$

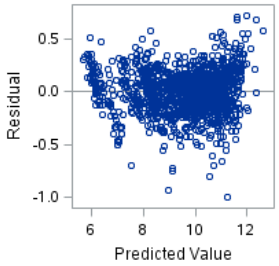
och vill i början av modelleringen främst fokusera på att undersöka modellens residualer plottade mot det predikterade värdet. Detta för att finna vilka transformationer som kan anses lämpliga för modellen. En residualplott av en multipel linjär regression för $log(Pris)$ med samtliga variabler syns i '(a)Figur 5' nedan. Det syns tydligt att vår residualplott följer en böjd form och är inte alls homoskedastisk, därför bör vi lägga till en kvadratisk term i modellen. Undersöker vi även residualerna plottade mot de förklarande variablerna,('Figur 14' i Appendix), så står det klart att det är variabeln $Carat^2$ vi behöver ha med i modellen. En residualplott med variabeln $Carat^2$ visas i '(b)Figur 5' nedan. Vi testar även hur residualerna ser ut när vi använder polynom upp till grad 4.



(a) Med $Carat$ (b) Med $Carat^2$ (c) Med $Carat^3$ (d) Med $Carat^4$

Figur 5: Residualplottar med alla variabler där carat har polynom upp till grad 4.

De plottade residualerna slutar att följa ett kvadratiskt samband efter införandet av $Carat^2$ men gynnas inte märkbart av polynom av högre grad. Dock är residualerna fortfarande inte homoskedastiska utan antar istället en lustig form som tyder på att variablerna beter sig annorlunda för olika intervall.
Vi prövar därför även att log-transformera Carat-variablerna upp till polynom grad 2. Residualerna för denna modell visas i 'Figur 6' nedan:



Figur 6: Residualplot med $log(Carat)$ och $log(Carat)^2$

Residualplotten ser nu något bättre ut, dock finns det fortfarande något slags trendbrott för de tidigare observationerna. En blick på $Carat$-variabelns residualer i modellen med $Carat^2$, som visas i 'Figur 15' i Appendix, visar på att diamanter mellan 0.2ct och ungefär 1.0ct uppträder märkbart annorlunda mot övriga diamanter.

4.2 Uppdelning av observationer.

Utför vi multipel linjär regression på endast $Carat + Carat^2$ samt $log(Carat) + log(Carat)^2$ mot responsvariabeln $log(Pris)$ får modellerna en förklaringsgrad på över 0.87 vardera, det är alltså tydligt att $Carat$ är den variabeln som förklarar variationen i priset mest. Vi väljer, på grund av "trendbrottet", därför att dela upp våra observationer i två grupper och får då en modell som beskriver priset på diamanter i storlekarna ≤ 1.0 ct, innehållande 277 observationer, och en modell som beskriver priset på diamanter i storlekarna > 1.0 ct, innehållande

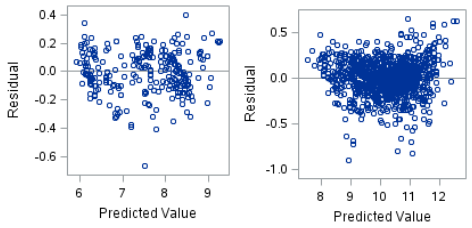
1073 observationer. Detta görs för modellen med och utan logaritmerade *carat*-variabler. En specifikation över modellerna ges i följande 'Tabell 4'.

Modell	**Carat**	**Funktion**
Små.ct Stora.ct	$\leq 1ct.$ $> 1ct.$	$log(Pris) = \alpha + \beta_1 Carat + \beta_2 Carat^2 + \cdots + \epsilon$
Små.log.ct Stora.log.ct	$\leq 1ct.$ $> 1ct.$	$log(Pris) = \alpha + \beta_1 log(Carat) + \beta_2 (log(Carat))^2 + \cdots + \epsilon$

Tabell 4: Tabell över våra fyra framtagna modeller.

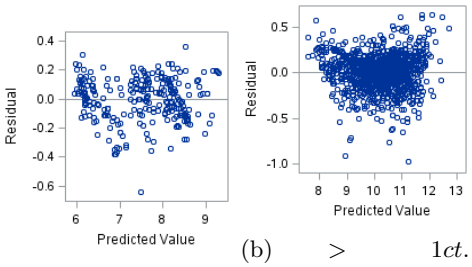
Modellerna *Små.ct* och *Stora.ct* är alltså de modeller som saknar logaritmerade *Carat*-variabler och modellerna *Små.log.ct* och *Stora.log.ct* är de som har logaritmerade *Carat*-variabler.

'Figur 7' och 'Figur 8' nedan visar residualerna plottade mot det predikterade värdena för alla modellerna. Som vi kan se är nu samtliga residualer tillräckligt homoskedastiska för att vi ska kunna gå vidare och se om vi kan reducera någonting i modellerna och till slut bestämma oss för vilken som anses bäst på att förklara variationen av priset.



(a) $\leq 1ct.$ Små.ct (b) $> 1ct.$ Stora.ct

Figur 7: Uppdelade modeller utan logaritmerade förklaringsvariabler.



(a) $\leq 1ct.$ Små.log.ct (b) $> 1ct.$ Stora.log.ct

Figur 8: Uppdelade modeller med logaritmerade förklaringsvariabler.

En ytterligare kontroll av variablernas nuvarande VIF-värden, efter införandet av transformationer, ges här nedan i 'Tabell 5'.

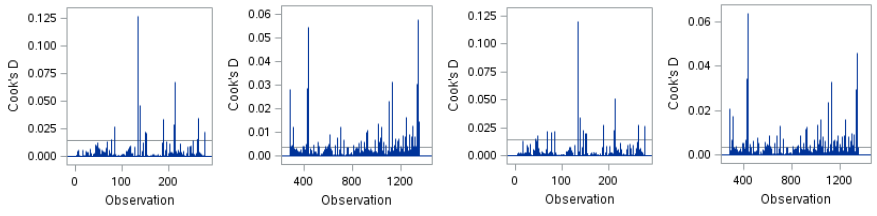
	VIF			VIF	
Variabel	Modell Små.ct	Modell Stora.ct	Variabel	Modell Små.log.ct	Modell Stora.log.ct
Carat	**31.846**	**35.295**	$log(Carat)$	**7.052**	**20.717**
$Carat^2$	**31.241**	**35.368**	$(log(Carat))^2$	**6.731**	**20.742**
Färg	1.137	1.026	*Färg*	1.138	1.027
Klarhet	1.309	1.050	*Klarhet*	1.304	1.050
Ytbehandling	1.268	1.058	*Ytbehandling*	1.275	1.058
Certifikatdummy	1.097	1.070	*Certifikatdummy*	1.100	1.068

Tabell 5: Tabell över variablernas VIF-värden.

Att VIF-värdet ökar för en variabel när vi lägger till en kvadrerad term av samma variabel kan ses förståeligt då dem beskriver sambandet till responsvariabeln på liknande vis. Vi ser att alla variabler innehållande *carat* i tabellen får ett högt VIF-värde p.g.a. detta. Ett sätt att komma runt problemet med högt VIF-värde för *carat*-variablerna är att centrera variablernas värden med hjälp av dess medelvärde. Man kan alltså omvandla de förklarande variablerna $Carat$ och $Carat^2$ till $(Carat - \overline{Carat})$ respektive $(Carat - \overline{Carat})^2$ för att sänka deras VIF-värden. Eftersom detta "knep" enbart sänker VIF-värdena utan att förbättra vare sig residualplottar eller R^2_{Adj} väljer vi bort det alternativet eftersom det bara skulle leda till en mer komplicerad modell. Utöver det behöver vi dessa kvadrerade variabler för homoskedasticitetens skull och väljer därför att låta dem vara. Övriga variabler håller sig fortfarande under VIF-värdet 5.

4.3 Cooks avstånd.

En åskådning av residualplottarna i 'Figur 7' och 'Figur 8' visar på existerande outliers. Vi använder då Cooks avstånd(4) för att finna och identifiera dessa outliers. Nedan, i 'Figur 9', ser vi modellernas observationer och deras inflytande på skattningarna genom Cooks avstånd.



(a) Små.ct (b) Stora.ct (c) Små.log.ct (d) Stora.log.ct

Figur 9: Cooks avstånd för samtliga observationer i de olika modellerna.

I figurerna ser vi att många observationer överskrider det streckade tröskelvärdet för en "inflytelserik observation" och att en del observationer överskrider det mycket mer än andra. I 'Tabell 6' tar vi då en närmre titt på de extremaste observationerna för att se om de ska tas bort från datamängden eller ej.

Obs.	Pris	$\hat{y}_{(.ct)}$	$\hat{y}_{(.log.ct)}$	Carat	Färg	Klarhet	Ytbehandling	Certifikat
134	\$944	\$1,841.8	\$1,824.8	0.6	1.0	0.3	0.767	EGL
214	\$925	\$1,347.1	\$1,304.5	0.8	0.1	0.1	0.767	GIA
433	\$3,080	\$7,532.6	\$7,685.6	1.5	1.0	0.1	0.467	EGL
1345	\$28,366	\$64,828.5	\$75,274.7	5.0	0.8	0.2	0.833	EGL

Tabell 6: Observationer med högt "Cooks avstånd"-värde, där $\hat{y}_{(\ )}$ anger det predikterade värdet från angivna modeller.

Vi ser att samtliga av de extremaste observationerna har ett mycket lägre pris än motsvarande predikterade värden för diamanterna även fast de flesta av diamanterna inte alls är särskilt "dåliga". Däremot ser vi ingenting som skulle kunna tyda på felaktigeter med dessa observationer, därför kan vi inte riktigt ta bort dem utan vidare. Det är snarare så att vår modell inte lyckas "fånga upp" dessa enstaka observationer. Därför väljer vi att behålla observationerna.

4.4 Stegvis variabelselektion.

Vi utför nu en stegvis variabelselektion(avsnitt 2.2.4) i våra fyra modeller för att se om vi kan reducera insignifikanta variabler. Modellernas $R^2_{adj}(2)$ och de återstående variablernas p-värden visas i 'Tabell 7' nedan.

Små.ct		R^2_{adj} : 0.9715	**Små.log.ct**		R^2_{adj} : 0.9729
Varibel $\leq 1ct$	Estimat	P-värde	Varibel $\leq 1ct$	Estimat	P-värde
Intercept	3.207	$< .0001$	Intercept	7.181	$< .0001$
Carat	6.234	$< .0001$	$log(Carat)$	2.292	$< .0001$
$Carat^2$	-2.246	$< .0001$	$log(Carat)^2$	0.299	$< .0001$
Färg	0.854	$< .0001$	Färg	0.858	$< .0001$
Klarhet	1.073	$< .0001$	Klarhet	1.091	$< .0001$
Ytbehandling	0.266	$< .0001$	Ytbehandling	0.307	$< .0001$
Certifikatdummy	0.053	$< .0148$	Certifikatdummy	0.059	$< .0055$
Stora.ct		R^2_{adj} : 0.9576	**Stora.log.ct**		R^2_{adj} : 0.9569
Varibel $> 1ct$	Estimat	P-värde	Varibel $> 1ct$	Estimat	P-värde
Intercept	5.074	$< .0001$	Intercept	6.386	$< .0001$
Carat	1.616	$< .0001$	Carat	1.945	$< .0001$
$Carat^2$	-0.151	$< .0001$	$log(Carat)^2$	—	—
Färg	1.506	$< .0001$	Färg	1.503	$< .0001$
Klarhet	1.338	$< .0001$	Klarhet	1.335	$< .0001$
Ytbehandling	0.276	$< .0001$	Ytbehandling	0.293	$< .0001$
Certifikatdummy	0.259	$< .0001$	Certifikatdummy	0.263	$< .0001$

Tabell 7: Resultatet efter stegvis variabelselektion.

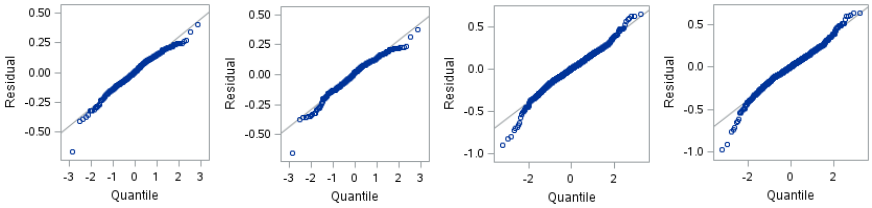
Samtliga variabler visade sig vara signifikanta för alla modeller utom $log(Carat)^2$ i *Modell Stora.log.ct* som hade ett p-värde på **0.6219**. Modellens residualplott förändrades minimalt efter borttagandet av $log(carat)^2$, vilket kan ses i 'Figur 16' i appendix.

4.5 Modellval.

Vi vill nu jämföra modellerna med varandra och ställer då de två modellerna för diamanter mellan 0.2 och 1 carat mot varandra och modellerna över 1 carat mot varandra. Vi tar hänsyn till residualplottarna samt R^2_{adj} för att avgöra vilken modell som är bäst. I 'Tabell 8' nedan ställer vi upp modell *Små.ct* mot modell *Små.log.ct* och modell *Stora.ct* mot modell *Stora.log.ct*, och i 'Figur 10' tittar vi på QQ-plottarna av residualerna mot den teoretiska normalfördelningen.

Modell	R^2_{adj}
Små.ct	0.9715
Små.log.ct	**0.9729**
Stora.ct	**0.9576**
Stora.log.ct	0.9569

Tabell 8: Jämföring av modeller.



(a) Små.ct (b) Små.log.ct (c) Stora.ct (d) Stora.log.ct

Figur 10: Modellernas QQ-plottar.

Eftersom residualplottarna mellan de modeller som har logaritmerade *Carat*-variabler och de som inte har de ser så pass lika ut bedömmer vi främst modellerna utefter deras R^2_{adj}. För de mindre diamanterna ser vi att *Små.log.ct* är bättre i det avseendet än *Små.ct* medan för de större diamanterna är modell *Stora.ct* bättre på att förklara priset än *Stora.log.ct*. Vi väljer därför ut modellerna *Små.log.ct* och *Stora.ct* till våra slutgiltiga modeller för att beskriva diamantpriser.

4.6 Jämförelse av modeller

De valda modellernas skattningar med standardavvikelser visas i figur 11 nedan:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.18096	0.05092	141.02	<.0001
lcaratsmall	1	2.29168	0.06128	37.40	<.0001
lcarat2	1	0.29868	0.03829	7.80	<.0001
Farg	1	0.85826	0.04291	20.00	<.0001
klarhet	1	1.09073	0.04306	25.33	<.0001
ytbehandling	1	0.30707	0.06437	4.77	<.0001
GIA	1	0.05884	0.02102	2.80	0.0055

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.07384	0.05485	92.51	<.0001
caratbig	1	1.61615	0.03242	49.85	<.0001
Carat2	1	-0.15112	0.00550	-27.46	<.0001
Farg	1	1.50616	0.02618	57.54	<.0001
klarhet	1	1.33773	0.03175	42.13	<.0001
ytbehandling	1	0.27649	0.04036	6.85	<.0001
GIA	1	0.25927	0.01475	17.58	<.0001

Figur 11: Parameterestimat och standardavvikelser för modellerna *Små.log.ct* och *Stora.ct*.

För att kanske finna en förståelse för varför diamanter under 1ct och diamanter över 1ct uppträder olika vi vill nu testa om de mindre och större diamanternas variabler påverkar priset olika. Det gör vi genom det tvåsidiga testet:

$$Z = \frac{\beta_{i(B)} - \beta_{j(S)}}{\sqrt{Var(\beta_{i(B)} - \beta_{j(S)})}} = \frac{\beta_{i(B)} - \beta_{j(S)}}{\sqrt{Var(\beta_{i(B)}) + Var(\beta_{j(S)}) - 2Cov(\beta_{i(B)}, \beta_{j(S)})}}$$

$$= \frac{\beta_{i(B)} - \beta_{j(S)}}{\sqrt{Var(\beta_{i(B)}) + Var(\beta_{j(S)})}},$$

där $Cov(\beta_{i(B)}, \beta_{j(S)}) = 0$ på grund av oberoende, $B =$ "Big" står för parametrarna i modellen med stora diamanter och $S =$ "$Small$" står för parametrarna i modellen med små diamanter. index i och j förklarar vilken variabel som testet avser. Med $H_0: \ \beta_{i(B)} - \beta_{j(S)} = 0$, mot
$H_a: \ \beta_{i(B)} \neq \beta_{j(S)}$.
Om $(Z_{\lambda_{0.025}} = -1.96) > Z$, eller om $Z > (Z_{\lambda_{0.975}} = 1.96)$ så förkastar vi nollhypotesen på signifikansnivån 5%.

Vi testar samtliga förklarande variabler förutom de innehållande $Carat$ eftersom modellerna redan är skilda i det avseendet. De test vi utför är:
Om (1) $\beta_{Färg(B)} = \beta_{Färg(S)}$,
om (2) $\beta_{Klarhet(B)} = \beta_{Klarhet(S)}$,
om (3) $\beta_{Ytbehandling(B)} = \beta_{Ytbehandling(S)}$
samt om (4) $\beta_{GIA(B)} = \beta_{GIA(S)}$.

Resultaten av testen blir då:

(1) $\beta_{Färg(B)} = \beta_{Färg(S)}$, $Z = \{\ 12.89 > 1.96$.
(2) $\beta_{Klarhet(B)} = \beta_{Klarhet(S)}$, $Z = \{\ 3.738 > 1.96$.

(3) $\beta_{Ytbehandling(B)} = \beta_{Ytbehandling(S)},\ Z = \{\ -0.44 < 1.96.$
(4) $\beta_{GIA(B)} = \beta_{GIA(S)},\ Z = \{\ 7.789 > 1.96.$ Nästan alla parameterskattningar skiljer sig, vissa mer än andra, mellan de små och stora diamanterna, undantaget är ytbehandlingen som inte har någon signifikant skillnad. Vi kan alltså förkasta nollhypotesten med 5% signifikansnivå för test (1), (2) och (4), men ej för test (3).

5 Resultat

De modeller vi har arbetat med i de tidigare avsnitten har alla haft en logaritmerad responsvariabel. För att gå tillbaka till en icke-logaritmerad responsvariabel gör vi båda sidor av funktionen till potenser av e.
Sammanfattningsvis har vi då tagit fram följande modeller för diamantpriset, tillsammans med koefficientskattningarna från 'Tabell 7', på diamanter mellan $0.2ct$ och $1.0ct$:

$$Pris = exp\{7.181 + 2.292 \cdot log(Carat) + 0.299 \cdot log(Carat)^2 + 0.858 \cdot Färg$$

$$+1.091 \cdot Klarhet + 0.307 \cdot Ytbehandling + 0.059 \cdot Certifikatdummy + \epsilon\},$$

samt diamanter över $1.0ct$ till och med $5.0ct$:

$$Pris = exp\{5.074 + 1.616 \cdot Carat - 0.151 \cdot Carat^2 + 1.506 \cdot Färg$$

$$+1.338 \cdot Klarhet + 0.276 \cdot Ytbehandling + 0.259 \cdot Certifikatdummy + \epsilon\}.$$

Antagandet från början var att feltermen ϵ var normalfördelad med väntevärde 0 och varians σ^2. Efter att feltermen gjorts till en potens av e blir den omvandlade feltermen, $\epsilon' = e^{\epsilon}$, istället log-normalfördelad med väntevärde μ' och varians $(\sigma')^2$.
Modellernas skattade varianser, med $log(Pris)$ blev 0.14745^2 för modellen med mindre diamanter samt $0.19252^2 \approx 0.03706$ för modellen med de större diamanterna. Efter omvandlingen till responsvariabeln $Pris$ blir väntevärde och varians istället:

$$\mu' = e^{\mu} = e^0 = 1,$$

$$(\sigma')^2 = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} = \begin{cases} 0.1498755^2, & \textit{för de små diamanterna}, \\ 0.1979524^2, & \textit{för de stora diamanterna}. \end{cases}$$

Det finns alltså en större osäkerhet av priset för de större diamanterna än för de små. En standardavvikelse från väntevärdet ger alltså en prisändring på ca. 15% på de små och ca. 20% på de stora diamanterna.

I modellen för de små diamanterna ser vi till exempel att koefficientskattningen för *Färg* är 0.858, vilket innebär att när variabeln *Färg* är på sin högsta nivå, dvs. =1, så har *Färg* en multiplikativ ökning av diamantens pris med $e^{0.858*1} \approx 2.358$. Detta kan jämföras med när *Färg* är på sin lägsta nivå, =0.1, som då har en multiplikativ inverkan på priset med $e^{0.858*0.1} \approx 1.090$. Samtliga variabler har efter omvandlingen från $log(Pris)$ till $Pris$ en multiplikativ inverkan på diamantpriset med faktorn $e^{\beta_i * Variabelvärde}$.
I 'Tabell 9' nedan har vi tolkat alla koefficientskattningars multiplikativa inverkan på priset av de förklarande variablerna när de är på sin högsta respektive lägsta nivå. Vi har även tagit fram kvoten Högsta/Lägsta -nivå uttryckt i procent vilket visar på den procentuella förändring av priset en variabeln har när den går från sin lägsta nivå till sin högsta nivå.

		Multiplikativ inverkan på Priset	
Variabel	**Nivå**	**Små diamanter**	**Stora diamanter**
Färg	Högsta(=1)	2.358	4.509
	Lägsta(=0.1)	1.090	1.163
	Högsta/Lägsta(%)	**116.3**%	**287.7**%
Klarhet	Högsta(=1)	2.977	3.811
	Lägsta(=0.1)	1.115	1.143
	Högsta/Lägsta(%)	**167**%	**233.4**%
Ytbehandling	Högsta(=1)	1.360	1.318
	Lägsta(=0.2)	1.063	1.057
	Högsta/Lägsta(%)	**27.9**%	**24.7**%
Certifikatdummy	Högsta(=1)	1.061	1.296
	Lägsta(=0)	1	1
	Högsta/Lägsta(%)	**6.1**%	**29.6**%

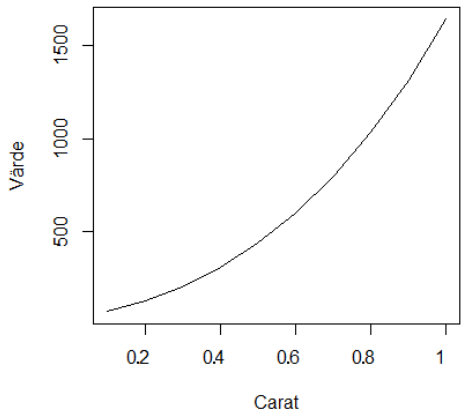
Tabell 9: Tolkning av koefficientskattningar.

Det vi kan utläsa från 'Tabell 9', förutom att få en förståelse för hur varje variabel påverkar priset, är att *Klarhet* är den variabeln som har störst inverkan på priset för de små diamanterna och att *Färg* är den variabel som har störst inverkan på priset för de stora diamanterna, (bortsett från diamantens vikt). Som vårt test i *avsnitt 4.6* visade, ökade alla variablers, utom ytbehandlingens, parameterskattningar när diamanterna blev större, vilket också kan ses i tabellen. Störst procentuell förändring hade dummy variabeln för *Certifikat* som gick från 6.1% ökat pris när diamanten undersökts av institutet *GIA* till en 29.6%-ig ökning av priset. Detta jämfört med en 0%-ig ökning av priset när en diamant validerats av något annat institut.

Variabeln $Carat$, som beskriver den största delen av priset, har olika transformationer i de två framtagna modellerna. I båda modellerna har $Carat$ en multiplikativ relation till de andra förklarande variablerna fast på olika vis.
I modellen för de mindre diamanterna beskrivs priset med hjälp av $Carat$ på följande sätt:

$$e^{2.292 \cdot log(Carat)} * e^{0.299 \cdot log(Carat)^2}.$$

Notera att för $Carat < 1$ blir den multiplikativa faktorn mindre än 1, dvs. minskning av priset sker. Detta hänger ihop med den höga skattningen av interceptet, ($e^{7.181} = \$1314.22$), som tillsammans med låga värden på $Carat$ minskar det skattade priset. För att förtydliga: Om $Pris$ endast beskrevs av interceptet och $Carat$-variablerna så antar $Pris$ värdet \$1314.22 när $Carat = 1$ och $e^{7.181} * e^{2.292 \cdot log(0.2)} * e^{0.299 \cdot log(0.2)^2} = \$306.45$ när $Carat = 0.2$.
I figuren nedan visas hur priset på de små diamanterna beskrivs med endast intercept och $Carat$-variabler.

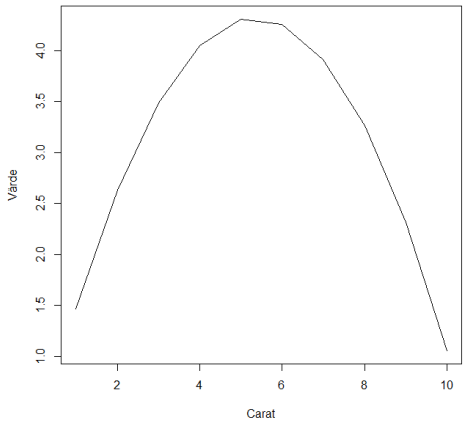


Figur 12: Värden för $e^{7.181} * e^{2.292 \cdot log(Carat)} * e^{0.299 \cdot log(Carat)^2}$.

I modellen för de större diamanterna beskrivs priset med hjälp av $Carat$ på följande sätt:

$$e^{1.616 \cdot Carat - 0.151 \cdot Carat^2}.$$

Värdena, som exponenten i uttrycket ovan antar, bildar en konkav kurva när $Carat$ går mellan 1ct och 10ct. En plot över detta visas i 'Figur 13' nedan:



Figur 13: Värden för $(1.616 \cdot Carat - 0.151 \cdot Carat^2)$.

Uttrycket $e^{1.616 \cdot Carat - 0.151 \cdot Carat^2}$ antar sitt maximum vid 5.351ct,$(= e^{4.3236})$, och avtar sedan, vilket syns i 'Figur 11'. Det är därför ett rimligt antagande att modellen endast lämpar sig för diamanter upp till ca 5ct eftersom det vore orimligt att diamanters pris skulle avta när diamanterna blir större än så.

6 Diskussion

Målet med det här arbetet var att undersöka vilka variabler som påverkar priset på diamanter samt hur dem gör det. Inte helt överraskande förklarades prisets variation till största delen av diamanternas vikt, men även andra av diamanternas attribut hjälpte till att få en bra förklarande modell som vi sedan kunde analysera.

Datamängden vi använde i analysen var manuellt insamlad vilket dels ger en ökad risk för mänskliga fel och dels kan datamängden variera väldigt mycket beroende på hur data har samlats in. I det här arbetet använde vi oss av 1350st observationer från ett urval på ca 84 000st vilket motsvarar ungefär 1.6%, vilket betyder att insamlingsmetoden kan vara väldigt avgörande för resultatet. Därför kan vi inte vara helt säkra på att våra skattningar verkligen stämmer överens med den totala poolen av diamanter särskilt bra. En insamlingsmetod som hade förbättrat precisionen av modellen hade varit om alla variabler samlats in ortogonalt mot varandra, dvs. att det för varje variabel finns lika andelar av de andra variablernas värden. På så vis kan man undvika korrelationer mellan variabler som kanske inte existerar i verkligheten. Detta var dock väldigt svårt att utföra med manuell insamling av data.

Observationernas trendbrott vid $1ct$ är ett exempel på någonting som kan bero på hur datan är insamlad. Om den är felaktigt insamlad eller ej är oklart, däremot är fenomenet väldigt underligt och är någonting som man skulle behöva undersöka ytterligare.

En variabel som visade sig vara relevant var $Certifikat$. Vart diamanterna certifierats ska, enligt brilliance.com, inte påverka diamanternas attribut, dvs. diamanternas andra variabler ska vara opartiskt betygsatta. Varför kommer det då sig att våra resultat visar på att diamanter certifierade hos GIA är dyrare än diamanter som certifierats hos de andra instituten?
Eftersom GIA är det mest erkända institutet över hela världen blir en certifiering från GIA någon slags säkerhet på att diamanten är rekorderlig. Eftersom detta är allmänt känt bland kunder och återförsäljare kan det vara så att brilliance.com passar på att ta ett större prispåslag för diamanterna certifierade hos GIA. Enligt våra resultat har ett certifikat från GIA större prispåverkan på stora diamanter än de små. Anledningen till det kan vara att köpet av en stor diamant oftast leder till en större investering än köpet av en liten diamant. Det blir då extra viktigt att diamanten är certifierad hos ett världskänt institut då det säkerställer diamantens värde men även gör diamanten mer lättsåld i framtiden. Detta utnyttjar då möjligvis brilliance.com i sin prissättning.

Vidare visade våra resultat på att klarheten hos en diamant var den viktigaste egenskapen för priset för de små diamanterna, bortsett från storleken, med färgen som den andra mest prispåverkande variabeln. För stora diamanter var det dock tvärt om, där färgen hade en större prispåverkan än klarheten. Det kan vara så att både klarheten och färgen blir viktigare när man investerar i en större diamant då de båda egenskaperna blir mer tydliga för ögat, men

att diamantens färg träder fram något mer än diamantens klarhet och värderas därmed högre.

Diamanternas egenskaper består inte endast av de vi använt oss av i detta arbete, förutom att diamanter finns i flertalet olika former har de även uppmätta drag som bredd, djup och även vilket sken de avger när de utsätts för UV-ljus. Om dessa variabler har någon större betydelse för diamantens värde låter jag vara osagt, det är däremot någonting man kan undersöka ytterligare.

Det vore även intressant att se hur resultaten skulle påveras av en ännu större mängd observationer och kanske med en annan insamlingsteknik.

I arbetet gjorde vi av en uppdelning av våra observationer och tillämpade två modeller för hela datamängden. Detta är inte alltig nödvändigt, användning av splines hade varit en metod, av flera, för att lösa problemet och man hade då kunnat uttrycka hela datamängden med hjälp av en modell, men det är nästan mer intressant att fördjupa sig i varför denna uppdelning behövdes. Med andra transformationer, fler variabler, fler observationer och annan insamlingsteknik hade man kanske kunnat förklara variationen av priset på ett simplare vis.

Målet med arbetet var främst att ta fram en modell som beskrev diamanternas pris utifrån dess karaktärsdrag. Modellen vi tagit fram skulle kunna användas i praktiken för att förutspå/bestämma priser på diamanter som ännu inte blivit prissatta. Vi hade dock behövt gå in djupare på modellens prediktionsförmåga samt skapa prediktionsintervall för att undersöka hur väl modellen passar för det ändamålet. Det är någonting som ett fortsatt arbete kunnat innehålla.

Referenser

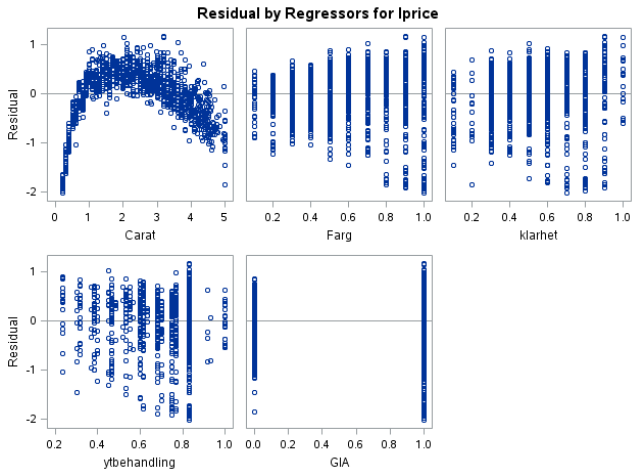
[1] 2016. URL: https://sv.wikipedia.org/wiki/Diamant.

[2] 2016. URL: http://www.dummies.com/how-to/content/how-to-interpret-a-correlation-coefficient-r.html.

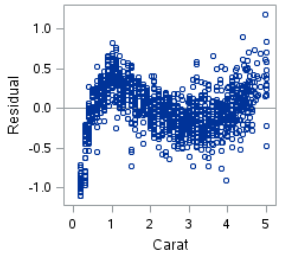
[3] 2016. URL: http://www.brilliance.com.

[4] Rolf Sundberg. *Lineära Statistiska Modeller*. Matematiska institutionen, Stockholms universitet, 2015.

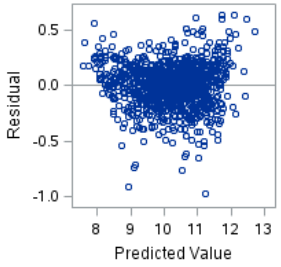
Appendix



Figur 14: Variablernas residualplottar.



Figur 15: Carats residualer



Figur 16: Modell Stora.log.ct's residualer