Quantilsregression: Eine Einführung (fast ohne* Formeln), Version 1.0, Holger Sennhenn-Reulen (NW-FVA)

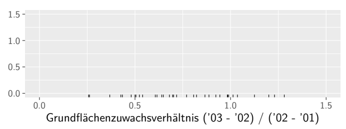
Warum Quantilsregression?

Quantilsregression ist eine statistische Regressionsverfahren mit dem nicht nur die 'mittlere' Lagebeziehung zwischen abhängigen und unabhängigen Variablen, sondern auch die Beziehung in verschiedenen Wahrscheinlichkeitsbereichen der Verteilung – den *Quantile* – der abhängigen Variablen modelliert werden kann. Während die gewöhnliche (lineare) Regression den bedingten Erwartungswert ('Mittelwert') als wesentliche Größe modelliert, ermöglicht die Quantilsregression die Schätzung von Regressionszusammenhängen für verschiedene Quantile, wie z.B. das 25%, 50% (Median) oder 90% Quantil. Damit kann ein umfassenderes Bild der Beziehung erhalten werden, da, falls vorhanden, die Variation in unterschiedlichen Teilen der Verteilung berücksichtigt wird.

Kästen 1 bis 3 zeigen wie wir implizit auch alle Quantile modellieren wenn wir gewöhnliche lineare Regression betreiben – oder eben jedes andere parametrische Verteilungsmodell nutzen. In den weiteren Kästen wird dann 'richtige' Quantilsregression veranschaulicht.

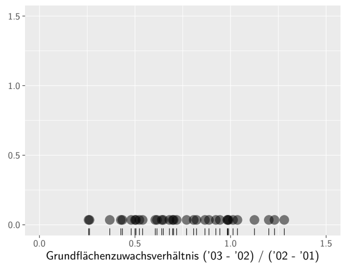
1) Verteilungsmodell

Darstellung der Verteilung der Variable *Grundflächenzuwachsverhältnis* als ein sogenannter Rug-Plot: der Wert der Variable wird mit je einem Strich pro Beobachtungseinheit auf der x-Achse dargestellt. Die y-Achse ist hier erstmal bedeutungslos.

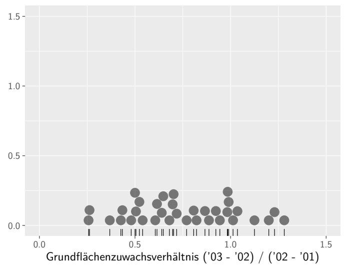


Datenquelle: United Nations Economic Commission for Europe (UNECE, 2006): The Condition of Forests in Europe. *Executive Report*.

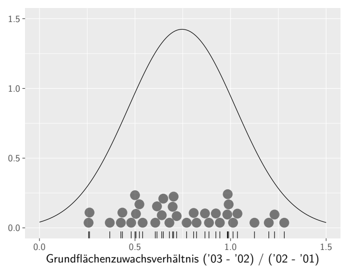
Diese Darstellung wird nun zusätzlich mit je einem Punkt pro Beobachtungseinheit ergänzt.



Und hier werden diese Punkte nun 'hochgestapelt' dargestellt. Daraus ergibt sich eine erste Idee über die Dichte bzw. Verteilung der Werte dieser beobachteten bzw. beobachtbaren variablen.

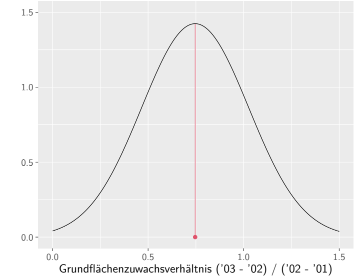


Wenn wir dafür nun eine Normalverteilung als statistisches Modell annehmen, dann können wir die beiden parameter dieser Verteilung mit den Daten schätzen, und die Dichte dieser Verteilung als schwarze Linie in der Grafik hinzufügen.

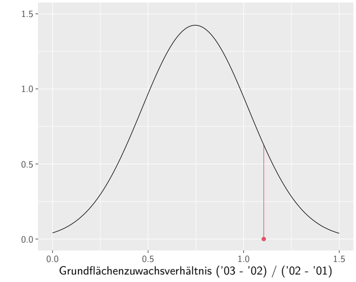


2) Verteilungsmodell

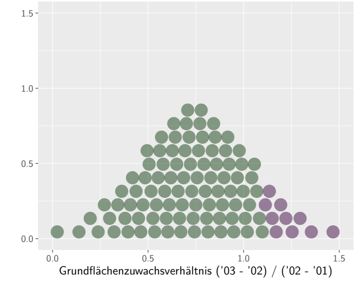
In dieser Dichtedarstellung zeichnen wir nun die Lage des Erwartungswertes als roten Punkt und zusätzlich als vertikale rote Linie ein.



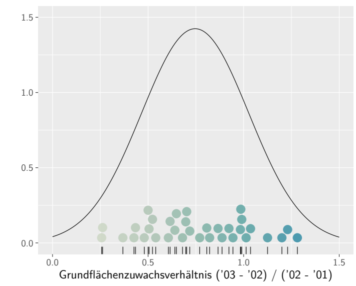
Auf dieselbe Art und Weise können wir nun aber auch z.B. das 95% Quantil einzeichnen, welches wir ja auch bereits geschätzt haben, da dieses ja mit den beiden Verteilungsparametern und der Dichtefunktion der Normalverteilung komplett definiert ist – wie jedes andere Quantil der Verteilung auch.



Alternativ können wir dieses Quantil auch wieder mit Punkten in einer 'Sampling-Darstellung' abbilden. Jeder Punkt bildet hier eine hypothetische Beobachtungseinheit ab – oder jeder Punkt kann auch für eine immer gleich große Gruppe von Beobachtungseinheiten stehen, auch wieder 'hoch gestapelt' dargestellt. Die 5% größten Werte sind blau markiert.

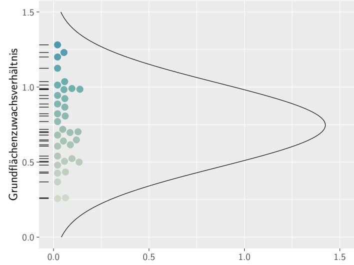


Zurück zur empirischen Verteilung unserer gemessenen Werte, nur Punkte jetzt farblich nach den aufsteigenden Werten markiert.

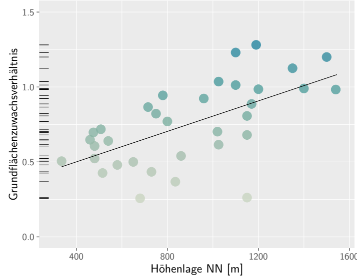


3) Verteilungsmodell

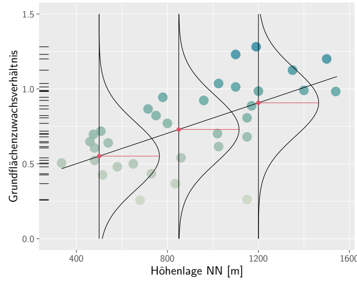
Und diese Darstellung können wir nun auch 'umkippen'.



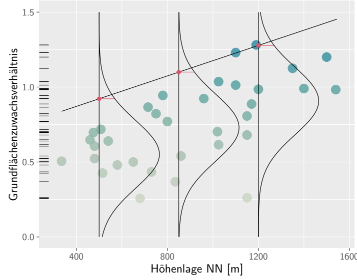
Auf der x-Achse nun die 'Einflussgröße' *Höhenlage NN* und mit 'Regressionsgerade' als scharze Linie:



Diese Regressionsgerade modelliert den bedingten Erwartungswert, wobei durch die Normalverteilungsannahme an die Residuen auch an jedem Wert der Einflußgröße eine vollständige Dichte der Zielvariablen definiert ist.



Wir können damit auch jedes bedingte Quantil der Verteilung einzeichnen und mit einer Geraden verbinden – hier wieder das 95% Quantil:

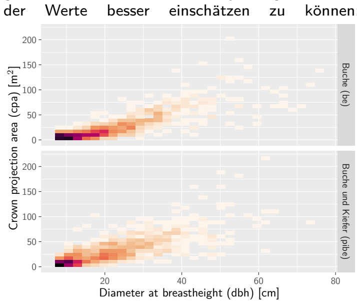


4) Datenbeispiel

Heym, Ruíz-Peinado, Del Río et al. (2017): EuMIXFOR empirical forest mensuration and ring width data from pure and mixed stands of Scots pine (Pinus sylvestris L.) and European beech (Fagus sylvatica L.) through Europe. *Annals of Forest Science* 74, 63. https://doi.org/10.1007/s13595-017-0660-z.

„[...] allometric relationships between diameter at breast height and crown projection area (cpa), calculated based on mean squared crown radii, cpa $= a \cdot \text{dbh}^b$ [...]"

Darstellung der Werte, wobei statt eines Streudiagramms ein bivariates Histogramm genutzt wird um die jeweilige Dichte der Werte besser einschätzen zu können:



5) Quantilsregression

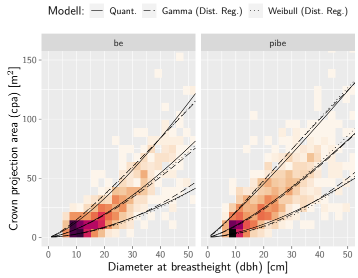
Allometrische Funktion:

$$\text{cpa}_i = a_i \cdot \text{dbh}_i^{b_i}$$

Effekte des Bestandestyps *Plot* (*be* oder *pibe*) auf beide nicht-linearen Parameter a und b.

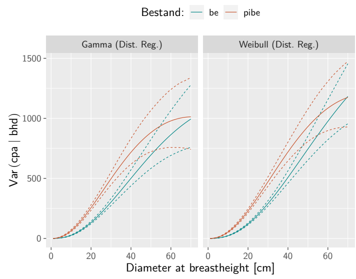
Quantilsregression für 10%, 50% (Median) und 90% Quantile.

Als Vergleichsmodelle Gamma- und Weibullregression mit Erwartungswert wie in Quantilsregression, und zweiten Verteilungsparametern in Abhängigkeit von Plot und dbh.



6) Verteilungsregression

Darstellung der geschätzten Varianz als Skalenmaß der kompletten Verteilung:



7) Fazit

Quantilsregression vs. Verteilungsregression:

- Quantilsregression passt sich flexibel den einzelnen Quantilen an: Das kann durch Quantile-Crossing aber auch in nicht immer sinnvolle Ergebnisse münden.
- Quantilsregression kann – über den Median – nur für wahr zugrundeliegende symmetrische Verteilungen den bedingten Erwartungswert modellieren.
- Verteilungsregression kann weitere bedingten Momente (Varianz, Schiefe, ...) direkt modellieren.
- Verteilungsregression – wenn passend – effizienter parametrisiert.
- Quantilsregression bietet in speziellen Anwendungen direktere Modellierung der Fragestellung, so ist z.B. von der WHO kindliches Übergewicht definiert als *liegt über dem 90% Quantil der Alters-Referenzpopulation*.

*Ergab sich als Take-Away aus *Intro to Bayes* das mal zu versuchen ...