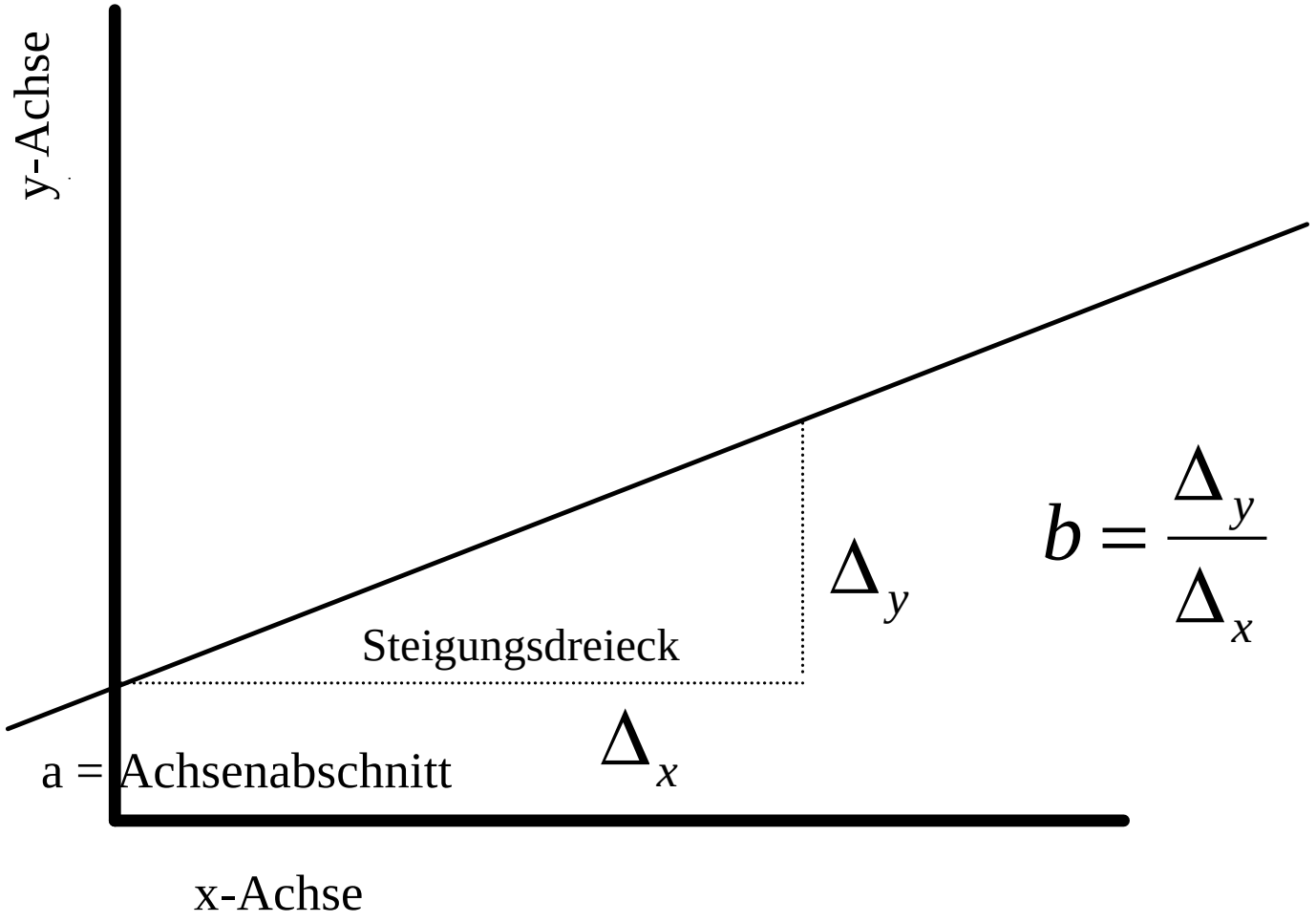
Grundgedanke der Regressionsanalyse

- Bisher wurden durch Koeffizienten die Stärke von Zusammenhängen beschrieben
- Mit der Regressionsrechnung können für intervallskalierte Variablen darüber hinaus *Modelle* geschätzt werden
- Ein Modell ist der (mathematisch) formalisierte Kern einer Theorie
- Modelle ermöglichen es, die abhängige Variable durch die unabhängige Variable vorherzusagen
- Die im Kurs verwendeten Modelle spezifizieren lineare Zusammenhänge:
  - je mehr x, desto mehr (oder weniger) y
  - die Beziehung x ➔ y kann durch eine einfache Gerade veranschaulicht werden
  - kompliziertere Modelle sind denkbar

Das mathematische Modell

- Im Regressionsmodell wird eine abhängige Variable y auf eine unabhängige Variable x zurückgeführt (regrediert)
- Beide Variablen sind mindestens intervallskaliert
- Diese Beziehung wird durch die Gleichung y=a+b*x beschrieben
- In diesem Modell denkt man sich den Wert der abhängigen Variablen zusammengesetzt aus:
  - einer Konstanten a
  - dem mit einem Faktor b multiplizierten Wert der unabhängigen Variablen x
- Die Beziehung kann durch eine Gerade veranschaulicht werden
  - a ist der Achsenabschnitt, d.h. der Wert, den y annimmt, wenn x=0
  - b ist die Steigung der Geraden, d.h. die Veränderung von y, wenn x um eine Einheit zunimmt

Regressionsgerade



Bestimmung der Regressionsgleichung

- Welche Parameter für a und b sollen in die Schätzgleichung eingesetzt werden? $\hat{y} = a + b \times x_i$
- Die beste Schätzung erhält man, wenn die Abstände zwischen der Regressionsgeraden und den empirischen Meßpunkten minimiert werden
- Die einfachen Abstände zur Geraden sind ungeeignet, weil es für jede Wolke von Meßpunkten unendlich viele Geraden gibt, für die sich die einfachen Abweichungen zu null addieren
- Deshalb werden die quadrierten Abweichungen verwendet, wodurch große Abweichungen stärker gewichtet werden (OLS-Schätzung)
- Minimiert wird die quadrierte Abweichung in y-Richtung

Berechnung von b

- Gesucht wird ein Wert, der die SAQ_y minimiert
- Durch Betrachtung der Ableitungen kommt man zu der Formel

$$b_{yx} = \frac{SAP}{SAQ_x}$$

- D.h., es müssen wie bei der Berechnung von r die Abweichungsprodukte und die Abweichungsquadrate (für x) bestimmt werden
- Die auf diese Weise geschätzte Gerade läuft durch den Punkt $(\bar{x}; \bar{y})$ (Eigenschaft der Kleinste-Quadrate-Schätzung)
- Deshalb kann ein Steigungsdreieck konstruiert werden, mit dessen Hilfe sich a bestimmen läßt

Berechnung von a

- Für x=0 ist y=a (Achsenabschnitt)
- Damit ist klar, daß die Gerade durch die Punkte (0;a) und $(\bar{x};\bar{y})$ gehen muß
- Die Steigung des dadurch definierten Dreiecks ist b. Durch Umformen läßt sich a ermitteln:

$$b_{yx} = \frac{\Delta_y}{\Delta_x} = \frac{\bar{y} - a_{yx}}{\bar{x} - 0} = \frac{\bar{y} - a_{yx}}{\bar{x}}$$

$$a_{yx} = \bar{y} - b_{yx} \times \bar{x}$$

Qualität der Regression

- Wird durch r^2 beschrieben. Andere Bezeichnungen R^2, Determinationskoeffizient, Varianzaufklärung
- Ist ebenfalls ein PRE-Maß
- Bester y-Prognosewert für einen beliebigen Fall wäre ohne weitere Zusatzinformation der Durchschnitt (quadrierte Abweichungen minimal)
- Abweichung vom Durchschnitt = Vorhersagefehler = SAQ bzw. Varianz

r^2

- Gesamtabweichung (Vorhersagefehler): $y_i - \bar{y}$
- Einen Teil der Abweichung vom Durchschnitt „erklärt“ das Regressionsmodell: $\hat{y}_i - \bar{y}$
- Die Abweichung zwischen vorhergesagtem Wert und tatsächlichem Wert wird durch das Modell nicht erklärt: $y_i - \hat{y}_i$
- Die Gesamtabweichung läßt sich zerlegen:
  Gesamtabweichung = nicht-erklärte Abweichung + erklärte Abweichung: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$
- Auch hier müssen wieder die quadrierten Abweichungen, d.h., die Varianzen betrachtet werden, da die Summe aller Abweichungen gleich null ist.
- Auch die Varianzen können (über alle Fälle hinweg) zerlegt werden

r^2 II

- r^2: Alle drei Abweichungen quadrieren und über alle Meßwerte aufsummieren

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SAQ_y(erklärt)}{SAQ_y(gesamt)}$$

- r^2 nimmt Werte zwischen 0 (keine Varianzaufklärung) und 1 (totale Varianzaufklärung an.
- Die Wurzel aus r^2 ist mit dem Korrelationskoeffizienten r identisch, allerdings vorzeichenlos