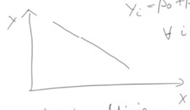
Einfache Regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \forall\, i = 1, \dots, n$$



globales Minimum
der Abweichungen Y_i und $\hat{y}_i$

$$Q((\beta_0, \beta_1)) = \frac{1}{n} \sum \hat{\varepsilon}_i^2 = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2$$
$$= \frac{1}{n} \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

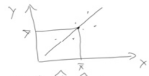
$Q(\beta_0, \beta_1)$ differenzieren und Null setzen!

$$\frac{d\, Q((\beta_0, \beta_1))}{d \beta_0} \stackrel{!}{=} 0 \qquad \frac{d\, Q(\beta_0, \beta_1)}{d \beta_1} \stackrel{!}{=} 0$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

gerade geht durch den Schwerpunkt der Daten:



Es gilt mit $\hat{\beta}_0, \hat{\beta}_1$:
Fehler heben sich auf!

$$\sum \hat{\varepsilon}_i = 0$$

Das sind alles Eigenschaften des Kleinsten-Quadrate-Schätzers

aufg. 6

Modell 1:

$$\frac{d\, \hat{mpg}}{d\, \text{horsepower}} = -0{,}0553032 \approx -0{,}055$$

$$\boxed{\begin{aligned} \hat{mpg} = 46{,}365 &- 0{,}147 \text{ cylinders} \\ &+ 0{,}001 \text{ displacement} \\ &- 0{,}055 \text{ horsepower} \\ &- 0{,}006 \text{ weight} \end{aligned}}$$

Wenn sich horsepower um eine Einheit erhöht, so ändert sich ceteris paribus (c.p.) die Variable mpg im Durchschnitt um −0,055 Einheiten.

Modell 3:

$$\hat{mpg} = 58{,}495 - 0{,}488 \text{ horsepower} + 0{,}0013 \text{ horsepower}^2$$

$$\boxed{\frac{d\, \hat{mpg}}{d\, \text{horsepower}} = -0{,}488 + 2 \cdot 0{,}0013 \cdot \text{horsepower}}$$

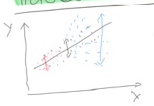
Der Ø Effekt einer c.p. Erhöhung von horsepower um eine Einheit ist auch abhängig vom Niveau von horsepower.

aufg. 7

Die beiden Modelle sind hinsichtlich der Güte nicht!! miteinander zu vergleichen, da sie als zu erklärende Variablen zwei unterschiedliche Variablen haben $Y \neq \log(Y)$ und damit auch $SQT_Y \neq SQT_{\log(Y)}$

ABER: Transformation der Daten führt zu linearer Beziehung in den Daten! Das ist gut für lineares Modell.

Heteroskedastizität



$$\varepsilon_i \sim N(0, \sigma^2) \to \sigma_K^2, \ \sigma_G^2$$

Problem:
- Schätzer nicht effizient
  ↳ Std. Fehler der Schätzer ↗
- Standardfehler nicht konsistent:
  Aussagen von Hypothesentest nicht mehr verlässlich!!

aufg. 8

FORWARD Selection

auswahl der Variablen um salary zu erklären.

① $Y = \beta_0 + \beta_1 x_j$

$j = 1, \dots, p$

- p Modelle geschätzt
- Modell mit geringsten SQR oder höchstem R^2
  → j gewählt.

② $y = \beta_0 + \beta_1 x_j + \beta_2 x_k$

$k = 1, \dots, p-1, \ k \neq j$

- (p−1) Modelle schätzen
- beste Modell wählen
  → k gewählt

③ ④ u.s.w. $\boxed{R^2}$ STOP-Kriterium

u.s.w.

Finaler Schritt: aus den Modellen mit 1 Variablen bis hin zu p Variablen wird das beste Modell mit korr. R^2 ausgewählt.