Arbeitsblatt lineare Regression

Formeln:

Arithmetisches Mittel $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (Gleiches Vorgehen für die Y-Werte)

Varianz $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (Gleiches Vorgehen für die Y-Werte)

Standardabweichung $s = \sqrt{s^2}$

Kovarianz $\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$

Lineare Regression: $b = cov/s^2_x$ $a = \bar{y} - b * \bar{x}$

Aufgaben:

1) Gegeben sind die folgenden vier Punkte:
   x-Werte 0, 2, 3, 5
   y-Werte 8, 3, 1, -2

a) Bestimme die Mittelwerte von x und y, die Varianzen und die Standardabweichung s, die Kovarianz und die Gleichung der Regressionsgeraden $\hat{Y} = a + b * x$.

   $\bar{x} = 2,5$ $\bar{y} = 2,5$
   $s^2_x = 3,25$ $s^2_y = 13,25$
   $s_x = 1,8$ $s_y = 3,64$
   $cov = -6,5$

   $\hat{Y} = 7,15 + (-2) * x$

b) Beschreibe die Gerade mithilfe der slope und des intercept.
   Die Gerade schneidet die Y-Achse an der Stelle Y = 7,15 (a = 7,15) und hat eine Steigung von -2 (b = -2). Wenn X um eine Einheit größer wird, verändert sich Y um -2 Einheiten.

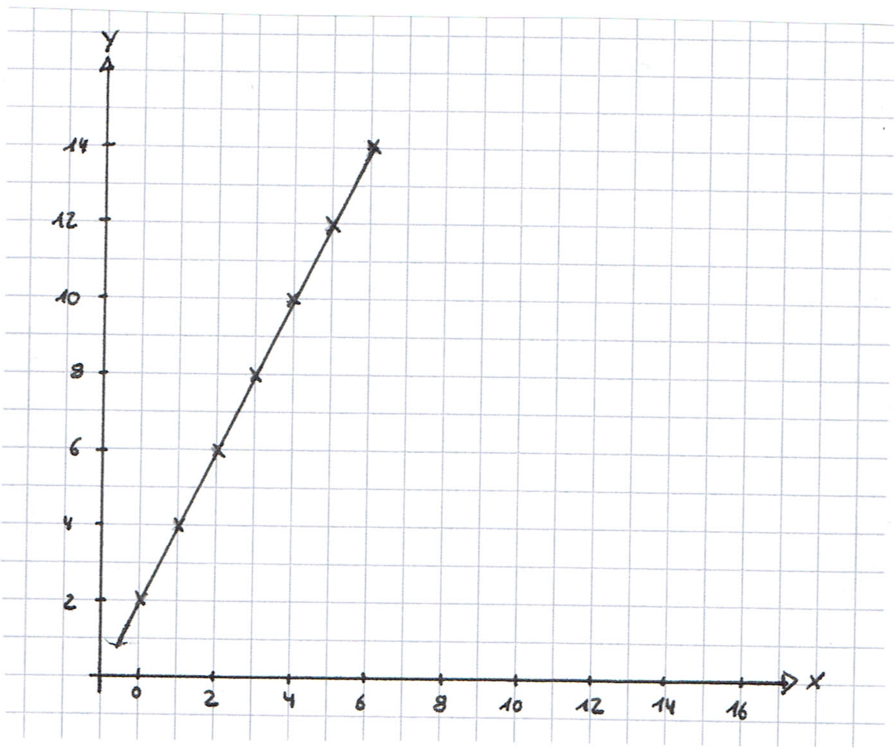
2) Gegeben sei die Regressionsgleichung $\hat{Y} = 4 - 0,5X$. Welche Aussage kann auf dieser Grundlage für die Korrelation zwischen X und Y getroffen werden?
   Die Korrelation zwischen X und Y ist negativ, da die Gerade eine negative Steigung (b = -0,5) aufweist.

   Gegeben sei die Regressionsgleichung $\hat{Y} = 2 + 4X$. Welche Aussage kann auf dieser Grundlage für die Korrelation zwischen X und Y getroffen werden?

   Die Korrelation zwischen X und Y ist positiv, da die Gerade eine positive Steigung (b = 4) aufweist.

3) Gegeben sei eine lineare Regression mit a = 12 und b = -1. Wie lautet der vorausgesagte Wert für $\hat{Y}$ an der Stelle X = 5?
   $\hat{Y} = 12 + (-1) * 5 = 7$

4) Zeichne die Regressionsgerade $\hat{Y} = 2 + 2X$ in ein Koordinatensystem, in dem der Wert 0 auf der X-Achse <u>nicht</u> gleichzeitig die Y-Achse darstellt.



5) Beschreibe in eigenen Worten die Funktionsweise der linearen Regression und erkläre das Prinzip der „Ordinary-Least-Squares“.

- *Vorhersage von Merkmalsausprägungen einer abhängigen Variable Y auf der Basis einer oder mehrerer unabhängiger Variablen X*
- *Es wird angenommen, dass der Zusammenhang zwischen X und Y linear ist, d.h. dass die Stärke und die Richtung des Zusammenhangs in jedem beliebigen Werteintervall auf der Variablen X gleich ist*
- *Y (Explanandum) soll durch X (Explanans) erklärt/vorhergesagt werden*
- *Korrelationen immer mit Unsicherheiten behaftet, wenn r nicht gleich 1 ist*
  → Schätzung aller Werte so, dass der Vorhersagefehler über alle Werte hinweg möglichst gering bleibt
- *Gleichung für die vorhergesagten Werte lautet* $\hat{Y} = a + b * X$
  → Für jeden X-Wert gibt es einen beobachteten und einen vohergesagten Y-Wert
  → Differenz zwischen Y und Y ist das Regressionsresiduum e
- *„Ordinary-Least-Squares“: Regressionsgerade soll so durch den Punkteschwarm gelegt werden, dass die Summe der quadrierten Regressionsresiduen minimal ist; Kriterium der kleinsten Quadrate; Minimierungsvorschrift ist erfüllt, wenn Regressionsparameter wie folgt bestimmt werden:*
  $b = cov/s^2_x$ $a = \bar{y} - b * \bar{x}$