Aufgabe 1:

Betrachten Sie wieder die *Galton*-Daten aus der Vorlesung. Für mittlere Größe der Eltern (X) und Größe der Kinder (Y) werde ein linearer Zusammenhang unterstellt. Das Streudiagramm in Abb. 1 zeigt die gemessenen Werte von 928 Eltern-Kind Paaren.

Darüber hinaus seien Ihnen folgende Größen bekannt:

$$n = 928,\ \bar{x} = 68.3082,\ \bar{y} = 68.0885,\ \hat{\Sigma} = \begin{bmatrix} 6.34 & 2.0646 \\ 2.0646 & 3.1946 \end{bmatrix},$$

wobei $\hat{\Sigma}$ die empirische Varianz-Kovarianzmatrix für (X, Y) bezeichne. Die Berechnungen sind ohne statistische Software durchzuführen.

(a) Zeichnen Sie mit bloßem Auge die Regressionsgerade in das Streudiagramm, die Ihrer Meinung nach den Zusammenhang bestmöglich beschreibt.

(b) Berechnen Sie $\hat{\beta}_0$ und $\hat{\beta}_1$. Tragen Sie die entsprechende Gerade ebenfalls in das Streudiagramm ein und vergleichen Sie die beiden Geraden.

(c) Berechnen und interpretieren Sie das Bestimmtheitsmaß R^2.

(d) Berechnen Sie $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ und bestimmen Sie ein 95%-Konfidenzintervall für β_1.

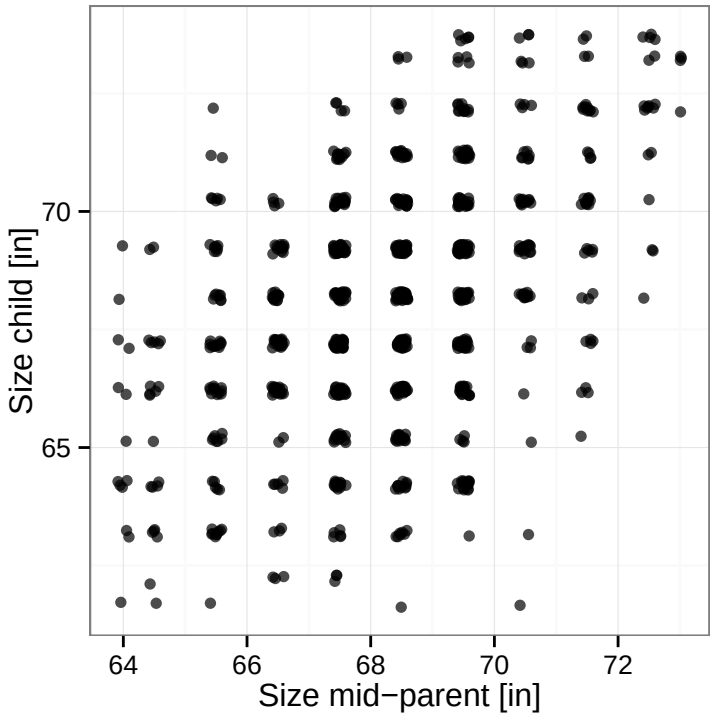

Abbildung 1: Streudiagramm Größe Kind in Abhängigkeit der mittleren Größe der Eltern

Aufgabe 2: Regression durch den Ursprung

Aus inhaltlichen Überlegungen heraus kann man in manchen Situationen davon ausgehen, dass die Regressionsgerade durch den Ursprung verläuft ($\beta_0 = 0$).
Das entsprechende Regressionsmodell ist

$$Y_i = \beta_1 x_i + \epsilon_i;\ i = 1, \ldots, n \tag{1}$$

mit den Annahmen (4) bis (7) aus Blatt1, Aufgabe 3.

(a) Berechnen Sie den KQ-Schätzer $\hat{\beta}_1$ für β_1. Ist $\hat{\beta}_1$ auch ML-Schätzer?

(b) Zeigen Sie, dass sich die Residuen im Allgemeinen nicht zu Null aufsummieren. Warum tun sie dies im Gegensatz dazu im einfachen Regressionsmodell mit Intercept β_0? Erklären Sie den Unterschied.

Aufgabe 3: Konfidenz- und Prognoseintervalle

Datensatz 2[1] auf der Übungshomepage enthält Informationen über die Fläche (in km^2, 2. Spalte) und die Einwohnerzahl (in Tausend, 3. Spalte) aller 190 Staaten der Erde (Stand 1993). Man interessiert sich für den Zusammenhang zwischen Einwohnerzahl und Fläche.

(a) Lesen Sie die Daten ein und untersuchen Sie explorativ die Verteilung der Daten. Aus welchem Grund sollte man darüber nachdenken, die Daten zu transformieren? Welche Transformation erscheint hier sinnvoll?

(b) Berechnen Sie die Schätzungen des von Ihnen gewählten Regressionsmodells und interpretieren Sie diese sowohl für die transformierten Daten als auch für die ursprünglichen Daten.

(c) Berechnen Sie 99 %-Konfidenzintervalle für die beiden Regressionskoeffizienten und interpretieren Sie diese. Untersuchen Sie, ob die Bevölkerungsdichte von der Größe der Länder abhängt.

(d) Lassen Sie nun Deutschland weg und schätzen Sie den so reduzierten Datensatz. Berechnen Sie ein 95 %-Prognoseintervall für die Einwohnerzahl Deutschlands. Liegt der wahre Wert in diesem Intervall?

Hinweis: Achten Sie darauf, wie sich ein Prognoseintervall für Y_{n+1} von einem Konfidenzintervall für $E[Y_i],\ i = 1, \ldots, n$, unterscheidet.

[1] Aus Riedwyl, Hans: „Lineare Regression und Verwandtes“, Birkhäuser Verlag: 1997