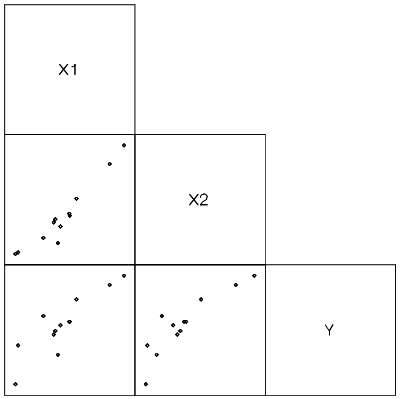
Lösungsskizze Serie 3

1. R-Output

```
> r.catheter <- lm(y~x1+x2,catheter)
> summary(r.catheter)

Call:
lm(formula = y ~ x1 + x2, data = catheter)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0497 -1.2753 -0.2595  1.9095  6.9933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.08527    8.77037   2.404   0.0396 *
x1           0.07681    0.14412   0.533   0.6070
x2           0.42752    0.36810   1.161   0.2753
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.94 on 9 degrees of freedom
Multiple R-Squared: 0.8056,     Adjusted R-squared: 0.7624
F-statistic: 18.65 on 2 and 9 degrees of freedom,       p-value: 0.0006301
```

Der F-Test testet die Nullhypothese $H_0 : \beta_1 = \beta_2 = 0$. Da die Teststatistik `F-Ratio`$= 18.65$ einem P-Wert von 0.001 entspricht, wird die Nullhypothese auf dem 5%-Niveau verworfen. Es können also nicht beide Variable gleichzeitig aus dem Modell entfernt werden, obwohl beide für sich allein betrachtet P-Werte aufweisen, welche viel grösser als 0.05 sind! Man könnte ohne Verlust eine der Erklärenden weglassen, nicht aber beide.
Erläuterung des Phänomens: Der Koeffizient β_1 der Grösse drückt die Änderung im Erwartungswert der Katheterlänge aus, wenn die Grösse um einen cm länger ist, das Gewicht aber konstant bleibt. Weil es jedoch zwischen den beiden erklärenden Grössen ebenfalls einen starken linearen Zusammenhang gibt, ist dieser Koeffizient nicht sehr gut bestimmt. (Analoge Erklärung für β_2).
Der folgende Scatterplot illustriert diesen Sachverhalt.



2. a) Nach Serie 1, Aufgabe 1b) gilt:

$$\begin{aligned} Y - \hat{Y} &= Y - \mathcal{E}[Y] - \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}(X_1 - \mathcal{E}[X_1]) \\ X_2 - \hat{X}_2 &= X_2 - \mathcal{E}[X_2] - \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)}(X_1 - \mathcal{E}[X_1]) \end{aligned}$$

Daraus ergibt sich

$$\begin{aligned} \mathrm{Cov}(Y - \hat{Y}, X_2 - \hat{X}_2) &= \mathcal{E}[(Y - \hat{Y})(X_2 - \hat{X}_2)] = \mathrm{Cov}(X_2, Y) - \frac{\mathrm{Cov}(X_1, X_2) \cdot \mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)} \\ &= \sqrt{\mathrm{Var}(X_2) \cdot \mathrm{Var}(Y)}(\rho(X_2, Y) - \rho(X_1, X_2) \cdot \rho(X_1, Y)) \\ \mathrm{Var}(Y - \hat{Y}) &= \mathrm{Var}(Y) - \frac{\mathrm{Cov}(X_1, Y)^2}{Var(X_1)} = \mathrm{Var}(Y)(1 - \rho(X_1, Y)^2) \\ \mathrm{Var}(X_2 - \hat{X}_2) &= \mathrm{Var}(X_2)(1 - \rho(X_1, X_2)^2) \end{aligned}$$

Somit ist die Korrelation zwischen $Y - \hat{Y}$ und $X_2 - \hat{X}_2$ gegeben durch

$$\frac{\mathrm{Cov}(Y - \hat{Y}, X_2 - \hat{X}_2)}{\sqrt{\mathrm{Var}(Y - \hat{Y}) \cdot \mathrm{Var}(X_2 - \hat{X}_2)}} = \frac{\rho(X_2, Y) - \rho(X_1, X_2) \cdot \rho(X_1, Y)}{\sqrt{(1 - \rho(X_1, Y)^2)(1 - \rho(X_1, X_2)^2)}}$$

Dies ist gerade die Formel für die partielle Korrelation zwischen X_2 und Y.

b) Definiere neue Zufallsvariablen

$$\begin{aligned} \bar{X_1} &:= X_1 - \mathcal{E}[X_1] \\ \bar{X_2} &:= X_2 - \hat{X}_2 = X_2 - \mathcal{E}[X_2] - \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)}(X_1 - \mathcal{E}[X_1]) \end{aligned}$$

Der entscheidende Punkt ist, dass $\bar{X_1}$ und $\bar{X_2}$ unkorreliert sind ($\mathrm{Cov}(\bar{X_1}, \bar{X_2}) = 0$), wie man leicht nachrechnet. Für die beste lineare Prognose $\bar{Y} = \eta + \lambda\bar{X_1} + \mu\bar{X_2}$ vereinfachen sich durch die Unkorreliertheit die Formeln für die Berechnung der Koeffizienten aus Serie 1, Aufgabe 1b):

$$\eta = \mathcal{E}(Y), \quad \lambda = \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}, \quad \mu = \frac{\mathrm{Cov}(\bar{X_2}, Y)}{\mathrm{Var}(\bar{X_2})}$$

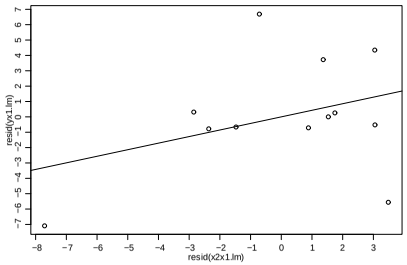
Somit ergibt sich:

$$\bar{Y} = \hat{Y} + \frac{\mathrm{Cov}(\bar{X_2}, Y)}{\mathrm{Var}(\bar{X_2})}(X_2 - \hat{X}_2)$$

Dies ist gerade die gesuchte Form.

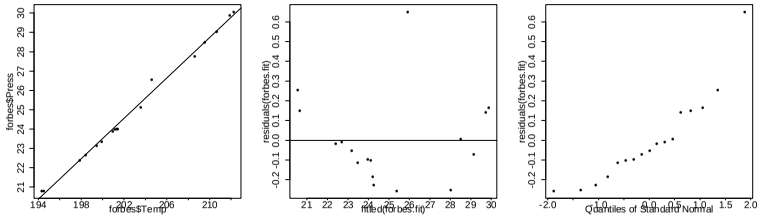
c) Der R-Code zeigt, wie man den Koeffizienten von `x2` aus Aufgabe 1 mit Hilfe von einfachen Regressionen bestimmt und den Plot erzeugt:

```
catheter <- read.table.url("http://stat.ethz.ch/Teaching/Datasets/catheter.dat", header=T)
yx1.lm <- lm(y~x1,catheter)                    # Regression von  y auf x1
x2x1.lm <- lm(x2~x1,catheter)                  # Regression von x2 auf x1

ryrx2.lm <- lm(resid(yx1.lm)~resid(x2x1.lm)-1)  # Regression durch den Nullpunkt
ryrx2.lm$coef     # = 0.4275231; gerade der Koeff. von x2 in der multiplen Regr.

plot(resid(x2x1.lm),resid(yx1.lm))             # Partial Residual Plot
abline(final.lm)                               # Gerade in Plot einzeichnen
```



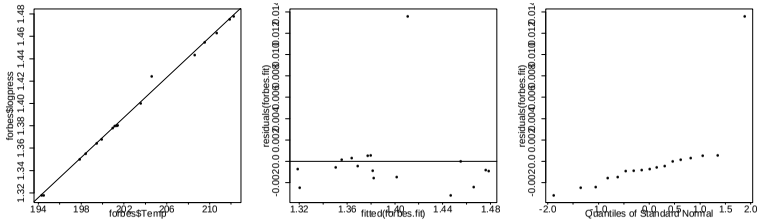
Die Steigung der eingezeichneten Geraden im "Partial Residual Plot" ist gerade der Koeffizient von `x2` aus der multiplen Regression.

3. **a)** Originalmodell:



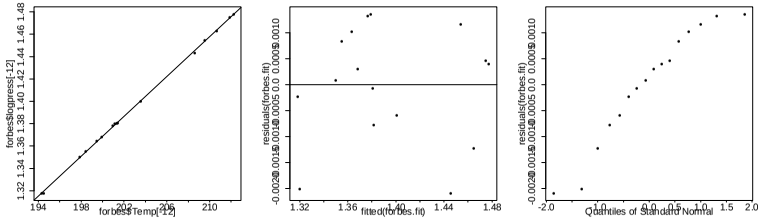
Wir sehen im Tukey-Anscombe-Plot eindeutig eine Struktur der Fehler (Parabelform), mit Ausnahme eines vermutlichen Ausreissers. Im Normalplot fällt nur der vermutliche Ausreisser auf.

b) Logarithmisches Modell:



Sowohl im Tukey-Anscombe-Plot wie im Normalplot fällt nun der Ausreisser deutlich auf.

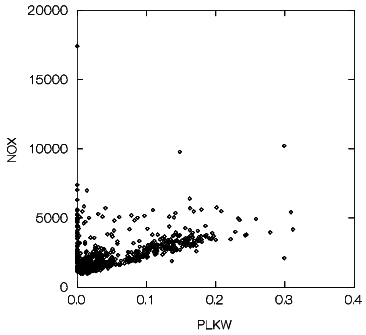
c) Logarithmisches Modell ohne 12. Beobachtung:



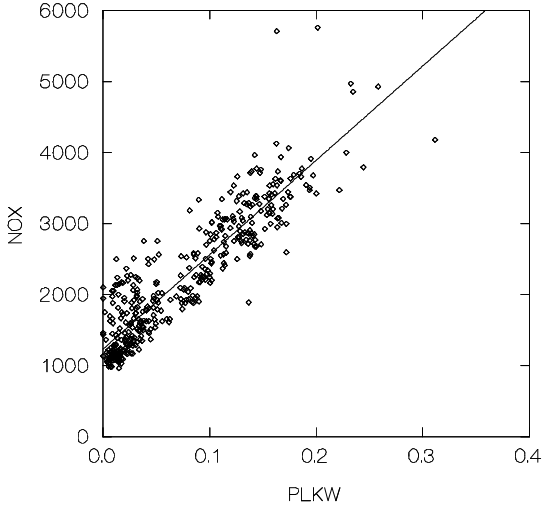
Der Tukey-Anscombe-Plot zeigt keine Auffälligkeiten mehr, und der Normalplot scheint auch in Ordnung zu sein (kleiner Tip: man skaliere das Grafikfenster jeweils so, dass die "Gerade" im Normalplot ungefähr eine Steigung 1 hat, da das menschliche Auge am sensibelsten auf Abweichungen von dieser Geraden reagiert.)

4. **a)** Der Emmissionsfaktor von Nicht-Lastwagen ist α, derjenige von Lastwagen $\alpha + \beta$.

b) Im Streudiagramm `NOX` gegen `PLKW` fallen einige Punkte mit grossen NOx-Werten auf (Ausreisser). Bei diesen Beobachtungen war die mittlere Emission pro Fahrzeug viel grösser als üblich.



c) Im Streudiagramm `NOX` gegen `PLKW` ohne die Punkte mit `VLUFT` < 5 gibt es nur noch wenige Ausreisser. Der Zusammenhang zwischen `NOX` und `PLKW` ist linear.
Eine mögliche Erklärung für die noch verbleibenden Ausreisser liegt in der Vermutung, dass schlecht gewartete Einzelfahrzeuge solche Spitzenwerte verursachen.



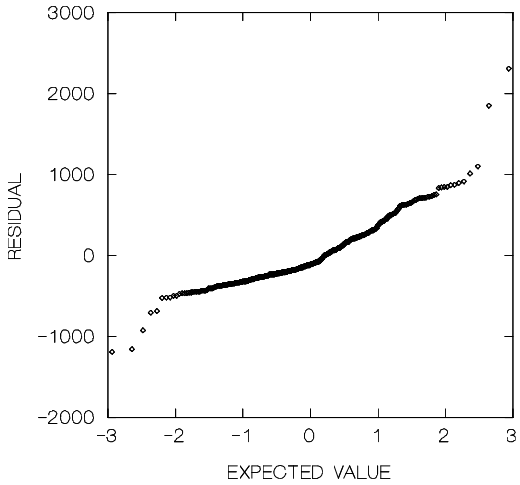
d) Setzt man im Modell $\texttt{NOX}_i = \alpha + \beta \cdot \texttt{PLKW}_i + E_i$ den Anteil der Lastwagen (`PLKW`) auf **0**, so erhält man den mittleren Emissionsfaktor für Nicht-Lastwagen; dieser entspricht dem Koeffizienten α. Für einen Lastwagen-Anteil von 1 ergibt $\alpha + \beta$ den mittleren Emissionsfaktor für Lastwagen.

Kategorie	mittlerer Emissionsfaktor (in mg/km)
Nicht-Lastwagen	$\widehat{\alpha} =$ 1231.885
Lastwagen	$\widehat{\alpha} + \widehat{\beta} =$ 1231.885 + 13286.470 = 14518.355

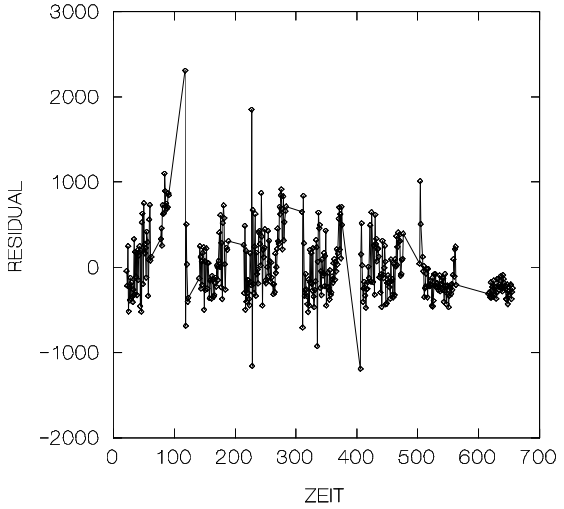
R-Code

```
> vluft <- !(is.na(gubrist[,'VLUFT']))&(gubrist[,'VLUFT'] >= 5)
> gubrist2 <- gubrist[vluft,]
> plot(gubrist2[,'PLKW'], gubrist2[,'NOX'])
> gubrist.lm <- lm(NOX ~ PLKW, gubrist2)
> abline(gubrist.lm)
> summary(gubrist.lm)

Call:
lm(formula = NOX ~ PLKW, data = gubrist2)

Residuals:
    Min      1Q  Median      3Q     Max
-1193.3  -253.7  -113.1   215.6  2309.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1231.89      30.14   40.87   <2e-16 ***
PLKW        13286.47     310.57   42.78   <2e-16 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 384.9 on 406 degrees of freedom
Multiple R-Squared: 0.8184,     Adjusted R-squared: 0.818
F-statistic:  1830 on 1 and 406 degrees of freedom,     p-value:     0
```

e) Der Normal-Plot zeigt, dass die Fehler nicht normalverteilt sind. Es ist deutlich die Schiefe der Verteilung zu erkennen.
Man kann sich vorstellen, dass eine entsprechende Zusammensetzung des Fahrzeugparks bezüglich der Emissionsfaktoren der einzelnen Fahrzeuge zu einer solchen Erscheinung führt.



Im Zeitreihenplot der Residuen sind Muster zu sehen. Dies deutet darauf hin, dass die Voraussetzung der Unabhängigkeit der Fehler ϵ_i verletzt ist.
Die Schadstoff-Messungen bilden eine Zeitreihe. In solchen Situationen ist zu erwarten, dass aufeinanderfolgende zufällige Abweichungen $\epsilon_i, \epsilon_{i+1}$ im Modell stochastisch abhängig sind.



Bemerkung: Diese Korrelationen haben insbesondere auf die Schätzung der Koeffizienten α und β und deren Standardfehler einen Einfluss. Üblicherweise werden die Standardfehler zu klein geschätzt, wenn man die Korrelationen nicht berücksichtigt und die Korrelation positiv ist.

f) Aufgrund der Schiefe der Fehlerverteilung müsste man eigentlich eine Transformation durchführen. Normalerweise würde man die Zielgrösse transformieren (z.B. logarithmieren) und dann nochmals eine lineare Regression rechnen. Dies ist im vorliegenden Fall aber nicht zulässig, da $\alpha + \beta \cdot \texttt{PLKW}_i$ eine "Bilanz-Gleichung" ausdrückt und die interessierenden Grössen α und β die Koeffizienten aus dem ursprünglichen Modell sind.
Ein Möglichkeit wäre, das Modell

$$\log(\texttt{NOX}_i) = \log(\alpha + \beta \cdot \texttt{PLKW}_i) + \epsilon_i \qquad (1)$$

anzupassen.
Modell (1) ist ein nichtlineares Regressionsmodell; nichtlinear, weil die Parameter α und β in nichtlinearer Weise ins Modell eingehen.