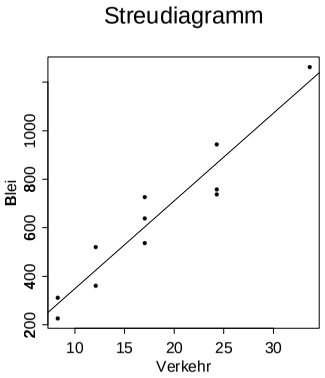
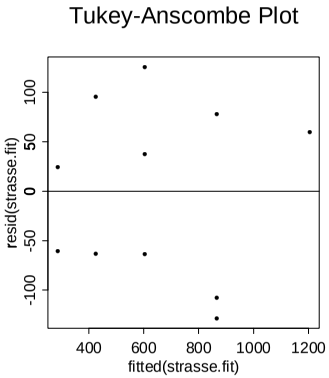
Lösungsskizze Serie 2

1. a) Computer-Output






```
> strasse.lm <- lm(blei ~ verkehr, data = strasse)
> summary(strasse.lm)

Call:
lm(formula = blei ~ verkehr, data = strasse)

Residuals:
    Min      1Q  Median      3Q     Max
-128.43  -63.13   24.52   69.32  125.72

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.842     72.143  -0.178    0.863
verkehr       36.184      3.693   9.798 4.24e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92.19 on 9 degrees of freedom
Multiple R-Squared: 0.9143,     Adjusted R-squared: 0.9048
F-statistic: 96.01 on 1 and 9 degrees of freedom,       p-value: 4.239e-06
```

b) Die 4.Modellvoraussetzung ist eigentlich verletzt, da die x_i immer auf 1000 Autos genau gerundet sind. Von der Grössenordnung her, kann man aber diesen Fehler vernachlässigen. Insgesamt scheint die Lösung, wie aus dem Streudiagramm ersichlich, sinnvoll zu sein.

c) Der p-Wert zum zweiseitugen Test $H_o : \beta = 0$ gegen $H_A : \beta \neq 0$ ist $4.24 \cdot 10^{-6} < 0.01$. Somit wird H_o klar verworfen.
Um den p-Wert zum einseitigen Test $H_o : \beta = 0$ gegen $H_A : \beta > 0$ zu bekommen, muss man den p-Wert des Outputs noch halbieren. Der zugehörige p-Wert ist dann $2.12 \cdot 10^{-6} < 0.01$. Damit wird auch hier H_o verworfen.
Während im zweiseitigen Test nur überprüft wird, ob ein Zusammenhang zwischen der Verkehrs- und der Bleibelastung besteht, interessiert uns im einseitigen Test, ob die Bleibelastung mit dem Verkehr zunimmt.

d) Aus dem Output: $\widehat{\alpha} = -12.842$.
Der p-Wert für diesen einseitigen Test ist $0.863/2 = 0.432$. Somit kann H_o nicht verworfen werden. Müssten wir die Nullhypothese verwerfen, würde dies bedeuten, dass der Achsenabschnitt und damit in einem gewissen Bereich auch der Bleigehalt sicher negativ wäre - eine unsinnige Aussage, weswegen wir unser Modell kritisch überprüfen müssten.

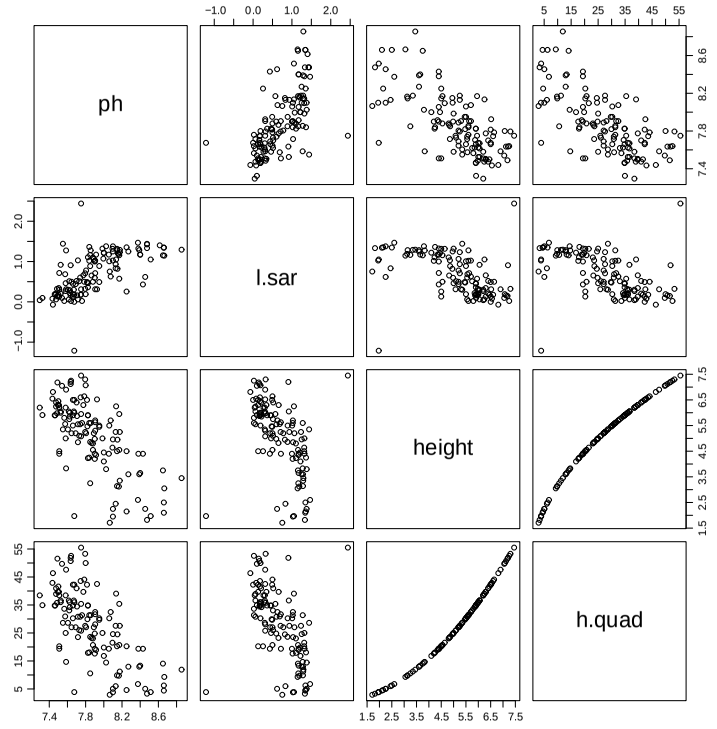
e) 95% KI(α) $= -12.84 \pm t_{9;0.975} \cdot 72.14 = (-175.88, 150.2)$
95% KI(β) $= 36.18 \pm t_{9;0.975} \cdot 3.69 = (27.84, 44.52)$ $\quad (t_{9;0.975} = 2.262)$
Ja, der Zusammenhang zu den obigen Tests ist über den Dualitätssatz (vgl. Einführungsvorlesung) gegeben. Wenn man das Konfidenzintervall kennt, dann kennt man das Testergebnis für alle $\beta = konst.$ auf dem Niveau 5%.

f) 25 und $50\mu g/g$ liegen nicht mehr im Vertrauensintervall (vgl. e)). $40\mu g/g$ liegt klar drin. Damit ist nur dieser Wert mit unseren Daten verträglich (d.h. wir können ihm nicht widersprechen).

2. a) Resultat von R:

```
> pairs(basisch)
```

Wir sehen, dass zwischen der quadrierten Höhe `h.quad` und dem `ph` Wert ein negativ linearer Zusammenhang besteht. Die ph-Werte liegen alle über 7 also im basischen Bereich. Der Einfluss der zusätzlichen Variablen l.sar ist nicht klar ersichtlich. Im weiteren fallen zwei Ausreisser auf.



b) Resultat von R:

```
> basisch.lm <- lm(h.quad ~ ph + l.sar, data=basisch)
> summary(basisch.lm)

Call:
lm(formula = h.quad ~ ph + l.sar, data = basisch)

Residuals:
      Min        1Q    Median        3Q       Max
-35.02890  -5.98680   0.04063   4.96733  30.69947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  221.239     25.810   8.572 4.15e-14 ***
ph           -24.288      3.388  -7.168 6.72e-11 ***
l.sar         -3.363      2.120  -1.586    0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.471 on 120 degrees of freedom
Multiple R-Squared: 0.4764,     Adjusted R-squared: 0.4677
F-statistic: 54.59 on 2 and 120 degrees of freedom,     p-value:     0
```

c) Auf dem 5%-Niveau wird $H_0 : \beta_2 = 0$ nicht verworfen (P-Wert=0.1153). Somit bringt der Einbezug von `l.sar` ins Modell nichts.

d) Resultat von R:

```
> new.point <- data.frame(ph=8, l.sar=1)                    # neue Versuchsbedingung
> conf.interval <- predict(basisch.lm, new=new.point, int="confidence",level=.95)
> pred.interval <- predict(basisch.lm, new=new.point, int="prediction",level=.95)


> conf.interval
         fit      lwr      upr
[1,] 23.57443 21.58269 25.56616
> pred.interval
         fit      lwr      upr
[1,] 23.57443 4.717598 42.43126
```

`fit`=angepasster Wert, `lwr`=untere Intervallgrenze, `upr`=obere Intervallgrenze.

Das Prognoseintervall für die Höhe ist einfach zu berechnen: man zieht einfach die Wurzel aus den Werten des Prognoseintervalls für die quadrierte Höhe. Dies ist ersichtlich aus 1.5.1 (f). Man löst nach y_0 auf und kann dann auf beiden Seiten die Wurzel ziehen. Für das Vertrauensintervall der erwarteten Höhe darf man dies aber nicht tun. Dies sieht man, wenn man 1.5.1 (e) nach $E[y_0]$ auflöst. Wir brauchen $E[\sqrt{y_0}]$, würden aber $\sqrt{E[y_0]}$ bekommen. Es gilt aber $\sqrt{E[y_0]} \neq E[\sqrt{y_0}]$.
Eine Regression mit `height` als Zielvariable und anschliessender Berechnung des Vertrauensintervalls ist auch nicht möglich, da gewisse Modellvoraussetzungen *nicht* erfüllt sind (Stichworte: Q-Q-Plot, Tukey-Anscombe-Plot), d.h. dass Modell passt nicht.
Was kann man also tun? Nichts! Wir können kein Vertrauensintervall für die uns ursprünglich interessierende Grösse `height` angeben. Dies ist ein Nachteil von Transformationen. Da aber $E[\sqrt{y_0}]$ und $\sqrt{E[y_0]}$ in erster Ordnung gleich sind, würde man in der Praxis trotzdem einfach die Wurzel ziehen. Eine andere Möglichkeit wäre noch, das Intervall durch Simulation zu bestimmen.

3. Das volle Modell (drei verschiedene Geraden) kann wie folgt geschrieben werden:

$$X = \begin{bmatrix} 1 & u_1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_l & 0 & 0 & \vdots & \vdots \\ 0 & 0 & 1 & v_1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & v_m & 0 & 0 \\ \vdots & \vdots & 0 & 0 & 1 & w_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & w_n \end{bmatrix} \qquad Y = \begin{bmatrix} x_1 \\ \vdots \\ x_l \\ y_1 \\ \vdots \\ y_m \\ z_1 \\ \vdots \\ z_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \alpha_3 \\ \beta_3 \end{bmatrix}$$

Bemerkung:
$(\widehat{\alpha}_1, \widehat{\beta}_1)$ werden nur aus $(x_1, u_1), \ldots, (x_l, u_l)$ geschätzt. Analoges gilt für $(\widehat{\alpha}_2, \widehat{\beta}_2)$ und $(\widehat{\alpha}_3, \widehat{\beta}_3)$. Die Spalten sind blockweise orthogonal (jede Gerade bildet einen Block.).

Das reduzierte Modell (Nullhypothese) $\beta_1 = \beta_2 = \beta_3 = \beta$ schreibt man mit

$$B\theta = 0 \qquad B = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Zur Berechnung der Quadratsumme SS_{ε_0} verwenden wir Y wie oben und für X und θ neu

$$X_0 = \begin{bmatrix} 1 & 0 & 0 & u_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \vdots & u_l \\ 0 & 1 & \vdots & v_1 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & 0 & v_m \\ \vdots & 0 & 1 & w_1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & w_n \end{bmatrix} \qquad \theta_0 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{bmatrix}$$

Die Teststatistik ist dann $(N = l + m + n)$

$$F = \frac{(SS_{\varepsilon_0} - SS_\varepsilon)/2}{SS_\varepsilon/(N-6)} \sim F_{2,N-6;1-\alpha}$$

4. Da $E[\epsilon_i^4] < \infty$, ist die erste Bedingung bereits erfüllt. Wir überprüfen jeweils zuerst die zweite Bedingung und dann die dritte.

a)

$$\sum_{i=1}^{n} x_i x_i^T = \sum_{i=1}^{n} f(i)^2$$
$$= \sum_{i=1}^{n} i^{2\gamma}$$

$\lambda_{\min,n} = \sum_{i=1}^{n} i^{2\gamma} \to \infty$ für $2\gamma \geq -1 \Leftrightarrow \gamma \geq -\frac{1}{2}$.
Für $\gamma < \frac{1}{2}$ ist die zweite Bedingung nicht erfüllt. Die erklärende Variable geht dann zu schnell gegen Null, d.h. sehr viele Werte x_i liegen sehr nahe bei Null und nur sehr wenige in grösseren Abständen von Null entfernt und daher hat man zuwenig Information, um das lineare Modell zu bestimmen.

$$\begin{aligned} \max_{1\leq j\leq n} x_j^T (\sum_{i=1}^{n} x_i x_i^T)^{-1} x_j &= \max_{1\leq j\leq n} \frac{f(j)^2}{\sum_{i=1}^{n} f(i)^2} \\ &= \max_{1\leq j\leq n} \frac{j^{2\gamma}}{\sum_{i=1}^{n} i^{2\gamma}} \end{aligned}$$

Für $\gamma > 0$ ist:

$$\max_{1\leq j\leq n} \frac{j^{2\gamma}}{\sum_{i=1}^{n} i^{2\gamma}} = \frac{n^{2\gamma}}{\sum_{i=1}^{n} i^{2\gamma}} \sim \frac{n^{2\gamma}}{n^{2\gamma+1}} \to 0 \text{ für } n \to \infty$$

Für $\gamma < 0$ ist:

$$\max_{1\leq j\leq n} \frac{j^{2\gamma}}{\sum_{i=1}^{n} i^{2\gamma}} = \frac{1}{\sum_{i=1}^{n} i^{2\gamma}} \to 0 \text{ für } n \to \infty$$

b)

$$\begin{aligned} \sum_{i=1}^{n} x_i x_i^T &= \sum_{i=1}^{n} f(i)^2 \\ &= \sum_{i=1}^{n} 2^{2i} \\ &= \sum_{i=1}^{n} 4^i \\ &= 4 \cdot \frac{4^n - 1}{4 - 1} \sim \frac{4}{3} 4^n \to \infty \text{ für } n \to \infty \end{aligned}$$

$$\begin{aligned} \max_{1\leq j\leq n} x_j^T (\sum_{i=1}^{n} x_i x_i^T)^{-1} x_j &= \max_{1\leq j\leq n} \frac{f(j)^2}{\sum_{i=1}^{n} f(i)^2} \\ &= \max_{1\leq j\leq n} \frac{2^{2j}}{\sum_{i=1}^{n} 2^{2i}} \\ &= \frac{4^n}{\sum_{i=1}^{n} 4^i} \\ &= \frac{4^n}{4 \cdot \frac{4^n - 1}{4-1}} \sim \frac{3}{4} \cdot \frac{4^n}{4^n} \not\to 0 \text{ für } n \to \infty \end{aligned}$$