Lösung zu Aufgabe 29

a) a1) Für die quadratische lineare Regression ist der Ausdruck

$\sum_{i=1}^{n}(y_i - (a + bv_i + cv_i^2))^2$

zu minimieren. Dabei hat jeder Summand die Form

$(y_i - (a + bv_i + cv_i^2))^2 =$
$y_i^2 + a^2 + b^2v_i^2 + c^2v_i^4 - 2ay_i - 2by_iv_i - 2cy_iv_i^2 + 2abv_i + 2acv_i^2 + 2bcv_i^3$

Für die Saldierung sind daher die Summen, Quadratsummen und Produktsummen zu bilden:

Versuch	1	v	y	v^2	v^3	v^4	y^2	yv	yv^2
1	1	20	2.0	400	8000	160000	4.00	40	800
2	1	40	5.5	1600	64000	2560000	30.25	220	8800
3	1	60	12.0	3600	216000	12960000	144.00	720	43200
4	1	80	26.0	6400	512000	40960000	676.00	2080	166400
5	1	100	59.0	10000	1000000	100000000	3481.00	5900	590000
6	1	120	95.0	14400	1728000	207360000	9025.00	11400	1368000
Summe	6	420	199.5	36400	3528000	364000000	13360.25	20360	2177200
Var	a^2	$2ab$	$-2a$	b^2	$2bc$	c^2	1	$-2b$	$-2c$

Die zu minimierende Funktion ist also

$\sum_{i=1}^{n}(y_i - (a + bv_i + cv_i^2))^2 = 13360.25 + 6a^2 + 36400b^2 + 364000000c^2 - 399a - 40720b - 4354400c + 840ab + 72800ac + 7056000bc$

Der Ausdruck wird partiell nach a, b, c abgeleitet und gleich Null gesetzt:

$$\begin{array}{rrrcr} 12a & +840b & +72800c & = & 399 \\ 840a & +72800b & +7056000c & = & 40720 \\ 72800a & +7056000b & +728000000c & = & 4354400 \end{array}$$

Lösung mit dem Taschenrechner oder mit R ist

$a = 14.05, b \approx -0.76455, c \approx 0.011987$

Die quadratische Regression liefert daher die Funktion

$y = 14.05 - 0.76455v + 0.011987v^2$

a2) $Y = X\beta + \sigma V$, dabei enthält X zeilenweise die Einträge $1, v_i, v_i^2$ (der quadratische Term bekommt eine eigene Spalte) und $\beta = (\beta_0, \beta_1, \beta_2)^T = (a, b, c)^2$ (in der Notation von Ansatz 1). Es ist also

$$X = \begin{pmatrix} 1 & 20 & 400 \\ 1 & 40 & 1600 \\ 1 & 60 & 3600 \\ 1 & 80 & 6400 \\ 1 & 100 & 10000 \\ 1 & 120 & 14400 \end{pmatrix}, Y = \begin{pmatrix} 2 \\ 5,5 \\ 12 \\ 26 \\ 59 \\ 95 \end{pmatrix}$$

Der KQ-Schätzer $\hat{\beta}$ ist Lösung von $(X^TX)\beta = X^TY$. Dabei ist

* $X^TX = \begin{pmatrix} 6 & 420 & 36400 \\ 420 & 36400 & 3528000 \\ 36400 & 3528000 & 364000000 \end{pmatrix}$, $X^TY = \begin{pmatrix} 199,5 \\ 20360 \\ 2177200 \end{pmatrix}$ (Das sind die Summen aus dem ersten Ansatz) Das LGS ist (bis auf den Faktor 2) dasselbe, dieser Faktor kommt durchs Ableiten im ersten Ansatz.

* Lösung ist $\hat{\beta} = (X^TX)^{-1}X^TY$. Die inverse Matrix kann z.B. mit dem TR bestimmt werden und ergibt

$$(X^TX)^{-1} = \begin{pmatrix} 3.20000000 & -0.09750000 & 0.00062500 \\ -0.09750000 & 0.00342411 & -0.00002344 \\ 0.00062500 & -0.00002344 & 0.00000017 \end{pmatrix}$$

und

$$(X^TX)^{-1} = \begin{pmatrix} 14,05 \\ -0,76455 \\ 0,011987 \end{pmatrix}$$

b)
- $Y^TY = 13360,25$ (siehe erster Ansatz)
- $\hat{\beta}^T(X^TY) = (14,05; -0,76455; 0,011987) \begin{pmatrix} 199,5 \\ 20360 \\ 2177200 \end{pmatrix} \approx 13333,91$
- $SS_{Res} = 13360,25 - 13333,91 = 26,34$
- $\hat{\sigma}^2 = SS_{Res}/(6-3) = 8,78 (\sigma = 2,963)$

c) Tabelle der notwendigen Werte und Summen:

	v	y	y^2	$\hat{y}_{lin}$	$(y-\hat{y}_{lin})^2$	$\hat{y}_{quad}$	$(y-\hat{y}_{quad})^2$
1	20.00	2.00	4.00	-12.43	208.18	3.55	2.41
2	40.00	5.50	30.25	5.84	0.12	2.65	8.14

Daraus folgt für den quadratischen Ansatz

c1) $SS_T = 6726.625$ wie oben

c2) $SS_{res} = 26.43$

c3) $R^2 = 1 - \frac{SS_{res}}{SS_T} = 1 - \frac{26.43}{6726.625} \approx 0.9961$

c4) $\tilde{R}^2 = 1 - \frac{SS_{res}/(6-2-1)}{SS_T/(6-1)} = 1 - \frac{26.43/3}{6726.625/5} \approx 0.9935$

d) Für den linearen Ansatz:

$F_0 = \frac{SS_R/k}{SS_{res}/(n-k-1)} = \frac{(SS_T - SS_{res})/k}{SS_{res}/(n-k-1)} = \frac{(6726.625-884.59)/1}{884.59/4} \approx 26.4169$

Für den quadratischen Ansatz:

$F_0 = \frac{SS_R/k}{SS_{res}/(n-k-1)} = \frac{(SS_T - SS_{res})/k}{SS_{res}/(n-k-1)} = \frac{(6726.625-26.34)/2}{26.34/3} \approx 381.56$

e) Zunächst der lineare Ansatz: Nochmal die Parameterschätzer:

$X^TX = \begin{pmatrix} 6 & 420 \\ 420 & 36400 \end{pmatrix}$,

$C = (X^TX)^{-1} = \frac{1}{6 \times 36400 - 420^2} \begin{pmatrix} 36400 & -420 \\ -420 & 6 \end{pmatrix} = \begin{pmatrix} \frac{13}{15} & -\frac{1}{100} \\ -\frac{1}{100} & \frac{1}{7000} \end{pmatrix}$

$X^T y = \begin{pmatrix} \frac{399}{2} \\ 20360 \end{pmatrix}$, $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -30.7 \\ \frac{1279}{1400} \end{pmatrix} = \begin{pmatrix} -30.7 \\ 0.91357 \end{pmatrix}$

Schätzer für σ^2: $\hat{\sigma}^2 = \frac{SS_{res}}{n-k-1} = \frac{884.59}{4} = 221.1475$, Ergibt $\hat{\sigma} \approx 14.871$

Standardfehler für $\hat{\beta}_0$: $se(\hat{\beta}_0) = \hat{\sigma}\sqrt{C_{11}} = 14.871 \times \sqrt{\frac{13}{15}} \approx 13.8442$. Daraus die t-Statistik $\hat{\beta}_0 / se(\hat{\beta}_0) \approx 2.218$

Standardfehler für $\hat{\beta}_1$: $se(\hat{\beta}_1) = \hat{\sigma}\sqrt{C_{22}} = 14.871 \times \sqrt{\frac{1}{7000}} \approx 0.1777$. Daraus die t-Statistik $\hat{\beta}_1 / se(\hat{\beta}_1) \approx 5.14$

Jetzt der quadratische Ansatz:

$$X^T X = \begin{pmatrix} 6 & 420 & 36400 \\ 420 & 36400 & 3528000 \\ 36400 & 3528000 & 364000000 \end{pmatrix}$$

$$C = (X^T X)^{-1} = \begin{pmatrix} 3.20 & -0.0975 & 0.000625 \\ -0.0975 & 0.0034241071 & -0.0000234375 \\ 0.000625 & -0.0000234375 & 0.0000001674107 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 399/2 \\ 23360 \\ 2177200 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T y = \begin{pmatrix} 14.05000000 \\ -0.76455357 \\ 0.01198661 \end{pmatrix}$$

Schätzer für σ^2: $\hat{\sigma}^2 = \frac{SS_{res}}{n-k-1} = \frac{26.43}{3} = 8.81$. Ergibt $\hat{\sigma} \approx 2.963$

Standardfehler für $\hat{\beta}_0$: $se(\hat{\beta}_0) = \hat{\sigma}\sqrt{C_{11}} = 2.963 \times \sqrt{3.2} \approx 5.301$. Daraus die t-Statistik $\hat{\beta}_0 / se(\hat{\beta}_0) \approx 2,65$

Standardfehler für $\hat{\beta}_1$: $se(\hat{\beta}_1) = \hat{\sigma}\sqrt{C_{22}} = 2.963 \times \sqrt{0.034241071} \approx 0.173404$. Daraus die t-Statistik $\hat{\beta}_0 / se(\hat{\beta}_0) \approx -4.409$

Standardfehler für $\hat{\beta}_2$: $se(\hat{\beta}_2) = \hat{\sigma}\sqrt{C_{33}} = 2.963 \times \sqrt{0.0000001674107} \approx 0.001212$. Daraus die t-Statistik $\hat{\beta}_0 / se(\hat{\beta}_0) \approx 9.886$

f) Formel für die 95%KI:

$\hat{\beta}_j - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}}$

Für den linearen Fall: $t_{0.975,4} = 2.776445$

$-30.7 - 2.776445 \times 14.871\sqrt{13/15} \leq \beta_0 \leq -30.7 + 2.776445 \times 14.871\sqrt{13/15}$, d.h. $-69.14 \leq \beta_0 \leq 7.73$

$0.91357 - 2.776445 \times 14.871\sqrt{1/7000} \leq \beta_1 \leq 0.91357 + 2.776445 \times 14.871\sqrt{1/7000}$, d.h. $0.42007 \leq \beta_1 \leq 1.40706$

Für den quadratischen Fall $t_{0.975,3} = 3.182446$

$14.05 - 3.182446 \times 2.963\sqrt{3.2} \leq \beta_0 \leq 14.05 + 3.182446 \times 2.963\sqrt{3.2}$, d.h. $-2.818 \leq \beta_0 \leq 30.918$

$-0.76455-3.182446\times 2.963\sqrt{0.00342} \le \beta_0 \le -0.76455+3.182446\times 2.963\sqrt{0.00342}$, d.h. $-1.316 \le \beta_1 \le -0.213$

$0.01199 - 3.182446 \times 2.963\sqrt{0.00000016741} \le \beta_0 \le 0.01199 + 3.182446 \times 2.963\sqrt{0.00000016741}$, d.h. $0.0081 \le \beta_2 \le 0.0158$

g) KI für den erwarteten Wert:

$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\vec{x}_0^T(\vec{X}^T\vec{X})^{-1}\vec{x}_0} \le E(y|\vec{x}_0) \le \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\dots}$

Zunächst das lineare Modell:

Beispiel: Datensatz für $v = 20$:

$x_0^T(X^TX)^{-1}x_0 = (1,20)\begin{pmatrix} \frac{13}{15} & -\frac{1}{100} \\ -\frac{1}{100} & \frac{1}{7000} \end{pmatrix}\begin{pmatrix} 1 \\ 20 \end{pmatrix} = \frac{13}{15} + 2\times 20\times 1\times(-\frac{1}{100}) + \frac{20^2}{7000} = \frac{11}{21}$

daraus dann $-12.43 - 2.776445 \times 14.871 \times \sqrt{\frac{11}{21}} \le E(y|x_0) \le -12.43 + \cdots$, d.h. $-42.31 \le E(y|x_0) \le 17.45$

Insgesamt ergibt sich folgende Tabelle (mit R erzeugt)

	fit	lwr	upr
1	-12.43	-42.31	17.45
2	5.84	-16.59	28.28
3	24.11	6.55	41.68
4	42.39	24.82	59.95
5	60.66	38.22	83.09
6	78.93	49.05	108.81

Jetzt das quadratische Modell. Wieder das Beispiel $v = 20$

$$\begin{aligned} x_0^T(X^TX)^{-1}x_0 &= \begin{pmatrix} 1 \\ 20 \\ 400 \end{pmatrix}^T \begin{pmatrix} 3.20 & -0.0975 & 0.000625 \\ -0.0975 & 0.0034241071 & -0.0000234375 \\ 0.000625 & -0.0000234375 & 0.0000001674107 \end{pmatrix}\begin{pmatrix} 1 \\ 20 \\ 400 \end{pmatrix} \\ &= 0.8214286 \end{aligned}$$

Daraus dann

$3.55-3.182446\times 2.963\times\sqrt{0.8214286} \le E(y|x_0) \le 3.55+\cdots$, d.h. $-4.99 \le E(y|x_0) \le 12.1$

Insgesamt ergibt sich folgende Tabelle (mit R erzeugt)

	fit	lwr	upr
1	3.55	-4.99	12.10
2	2.65	-2.58	7.87
3	11.33	5.58	17.08
4	29.60	23.85	35.35
5	57.46	52.23	62.69
6	94.91	86.36	103.46

h) Prognoseintervall für den Bremsweg

$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1+\vec{x}_0^T(\vec{X}^T\vec{X})^{-1}\vec{x}_0)} \le y_0 \le \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}$

Unterschied zum Konfidenzintervall: „1+“unter der Wurzel.

Beispielsweise im linearen Modell für $v = 50$: Als Vorhersage erhält man $\hat{y}_{lin} = 30.7 + 0.9136 \times 50 \approx 14.98$. Weiter dann für das Prognoseintervall

$1 + x_0^T (X^T X)^{-1} x_0 = 1 + (1, 50) \begin{pmatrix} \frac{13}{15} & -\frac{1}{100} \\ -\frac{1}{100} & \frac{1}{7000} \end{pmatrix} \begin{pmatrix} 1 \\ 50 \end{pmatrix} = 1 + \frac{13}{15} + 2 \times 50 \times 1 \times (-\frac{1}{100}) + \frac{50^2}{7000} = 1 + \frac{47}{210} = \frac{257}{210}$

daraus dann $14.98 - 2.776445 \times 14.871 \times \sqrt{\frac{257}{210}} \leq y_0 \leq 14.98 + \cdots$, d.h. $-30.70 \leq E(y|x_0) \leq 60.56$

Die Tabelle der Prognoseintervalle lautet dann (mit R erzeugt)

v	fit	lwr	upr
50	14.98	-30.70	60.65
130	88.06	34.53	141.60

und im quadratischen Modell

v	fit	lwr	upr
50	5.79	-5.06	16.63
130	117.23	101.82	132.65

Lösung zu Aufgabe 30

Ansatz:

- $F_0 = \dfrac{SS_R/k}{SS_{res}/(n-k-1)} = \frac{n-k-1}{k} \dfrac{SS_R}{SS_{res}}$, also ist $\frac{SS_R}{SS_{res}} = \frac{k}{n-k-1} F_0$
- $R^2 = \dfrac{SS_R}{SS_T} = \dfrac{SS_R}{SS_R + SS_{res}} = \dfrac{1}{1 + \frac{SS_{res}}{SS_R}}$

Setzt man beides zusammen, so bekommt man:

$$R^2 = \frac{1}{1 + \frac{n-k-1}{k} \frac{1}{F_0}} = \frac{F_0}{F_0 + \frac{n-k-1}{k}}$$

Im konkreten Fall gilt $R^2 = \frac{52.75}{52.75 + \frac{234}{15}} \approx 0.77$

Fazit: Zwischen der F-Statistik für die Modellgüte und dem Bestimmtheitsmaß gibt es eine 1-1-Beziehung. Grundsätzlich müsste man nur einen der Werte kennen, um den anderen bestimmen zu können.