LÖSUNG 9B

a.

- Man kann erwarten, dass der Absatz mit steigendem Preis abnimmt, mit höherer Anzahl der Außendienstmitarbeiter sowie mit erhöhten Werbeausgaben steigt. Insofern besteht die Erwartung, dass in der geschätzten Regressionsgleichung der Regressionskoeffizient der Variable PREIS ein negatives Vorzeichen und die Regressionskoeffizienten der Variablen ADM sowie WERBUNG positive Vorzeichen haben.

- Mit "Analysieren", "Regression", "Linear" wird die Dialogbox "Lineare Regression" aufgerufen. Die Variable ABSATZ wird in das Eingabefeld „abhängige Variable“ und die unabhängigen Variablen PREIS, ADM (Außendienstmitarbeiter) und WERBUNG in das entsprechende Eingabefeld übertragen. Als Methode wird "Einschluss" gewählt.

- Das Bestimmtheitsmaß beträgt $R^2 = 0{,}919$. Knapp 92 Prozent der Varianz von GEHALT wird durch die Erklärungsvariablen PREIS, ADM und WERBUNG vorhergesagt (erklärt). Es handelt sich daher um ein Regressionsmodell mit sehr starker Vorhersage(Erklärungs-)kraft.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,959[a]	,919	,899	4546,245

a. Einflußvariablen : (Konstante), Werbebudget, Anzahl der Außendienstmitarbeiter, Verkaufspreis

- Die erwarteten positiven Vorzeichen der (Grundgesamtheits-) Regressionskoeffizienten β_i von ADM (Außendienstmitarbeiter) und WERBUNG sowie das negative Vorzeichen von PREIS werden empirisch bestätigt. Insofern gibt es keinen Widerspruch zwischen theoretischer Erwartung und empirischer Schätzung. Die geschätzte Regressionsgleichung lautet:

geschätzter ABSATZ = 210.159,444 - 3.832,503*PREIS + 6.723,478*ADM + 0,069*WERBUNG

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressions-koeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	210159,444	23729,909		8,856	,000
	Verkaufspreis	-3832,503	444,013	-,725	-8,632	,000
	Anzahl der Außendienstmitarbeiter	6723,478	840,997	,665	7,995	,000
	Werbebudget	,069	,024	,242	2,903	,013

a. Abhängige Variable: Absatzmenge

- Zur Prüfung, ob die Regressionskoeffizienten sich signifikant von 0 unterscheiden, wird ein einseitiger t-Test durchgeführt. Der Test soll zur Entscheidung führen, ob die H_0-Hypothese $\beta_i = 0$ oder die H_1-Hypothese (Alternativhypothese) $\beta_i < 0$ bzw. $\beta_i > 0$ angenommen werden soll (Annahme der Hypothese H_0 bedeutet, dass der angenommene Einfluss der Erklärungsvariable auf den Absatz negiert wird). Die Prüfvariable $t = \frac{b_i - \beta_i}{s_b}$ (bzw. unter Annahme der H_0-Hypo-

these $\beta_i = 0$: $t = \frac{b_i}{s_b}$) hat eine t-Verteilung mit n-m-1 Freiheitsgraden [n = Anzahl der Beobachtungen (Fälle), m = Anzahl der erklärenden Variablen]. Bei einem angenommenen Signifikanzniveau von $\alpha = 0{,}05$ (5 %) und Freiheitsgraden = n-m-1 = 16 - 3 - 1 = 12 ergibt sich aus einer t-Tabelle (sie kann von den Internetseiten zum Buch heruntergeladen werden) ein kritischer Wert für t in Höhe von 1,782. Da der empirische t-Wert der Prüfgröße für die Variable PREIS, ADM und WERBUNG absolut (ohne Berücksichtigung des Vorzeichens) größer ist als der kritische Wert, wird die H_0-Hypothese ($\beta_i = 0$) abgelehnt und die Alternativhypothese ($\beta_i < 0$ bzw. $\beta_i > 0$) angenommen. Der Preis, die Anzahl der Außendienstmitarbeiter sowie der Werbeumfang haben wie erwartet einen signifikanten Einfluss auf die Höhe des Absatzes. Zu dieser Schlussfolgerung kommt man auch, wenn man die in der Ausgabe angeführten Werte von "Sig." mit dem Signifikanzniveau $\alpha = 0{,}05$ vergleicht. Da für jede der Erklärungsvariablen der Wert von "Sig." kleiner ist als 0,05, wird die H_0-Hypothese abgelehnt und die H_1-Hypothese angenommen.

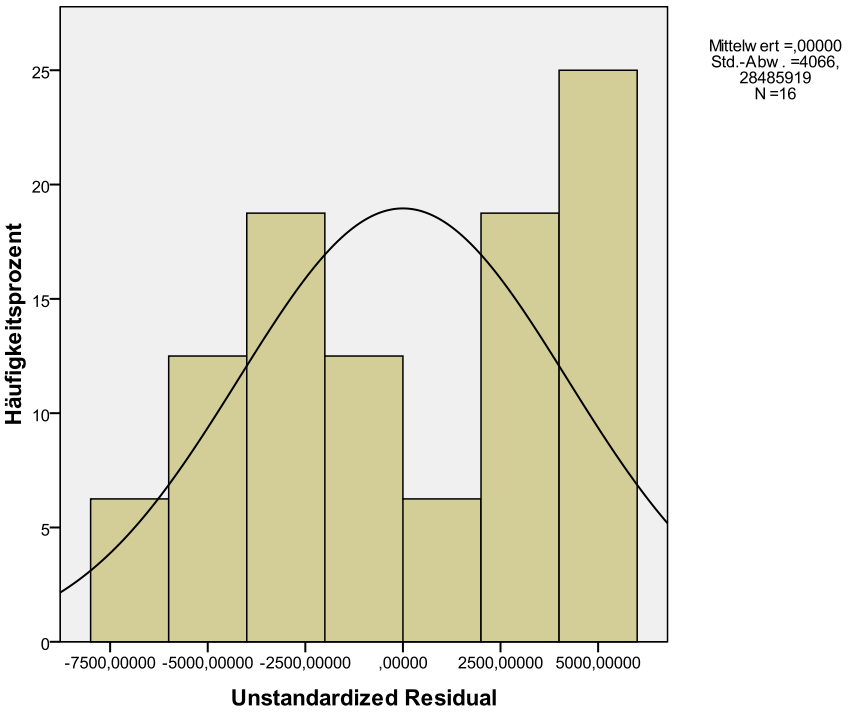
b.

- Für die Prüfungen auf Normalverteilung und Homoskedastizität der Residualvariable muss man auf gespeicherte Residual- und Vorhersagewerte zugreifen, d.h. man muss die Schätzung der Regressionsgleichung eventuell wiederholen wenn diese nicht vorliegen.

  Zum Speichern der Residual- und Vorhersagewerte öffnet man durch Klicken auf die Schaltfläche "Speichern" in der Dialogbox "Lineare Regression" die Unterdialogbox "Lineare Regression: Speichern" und wählt in den Feldern "Vorhersagte Werte" und "Residuen" die Option "Nicht standardisiert". Die Vorhersage- und die Residualwerte werden als PRE_1 und RES_1 den Variablen im Dateneditor hinzugefügt. Außerdem wird eine Tabelle mit dem Mittelwert und der Standardabweichung und weitere Daten für die Residualwerte und Vorhersagewerte (standardisiert und nicht standardisiert) erstellt.

- Ein Histogramm zum Prüfen auf Normalverteilung (zum Vorgehen s. Lösung Aufgabe 9A) der Residualvariable zeigt, dass es erhebliche Abweichungen von der Normalverteilung gibt.

  Da dieses Anwendungsbeispiel auf nur 16 Datenfällen beruht, kann man auch keine Normalverteilung erwarten. Grundsätzlich sollte man eine Regressionsanalyse mit höheren Fallzahlen durchführen.



- Nur zur Demonstration haben wir auch einen Test auf Normalverteilung der Residualvariable durchgeführt (Zur Durchführung s. Lösung Aufgabe 9A). Obwohl es – wie aus dem Histogramm ersichtlich – erhebliche Abweichungen zur Normalverteilung gibt, weisen die Tests keine signifikanten Abweichungen zur Normalverteilung aus. Das liegt natürlich am kleinen Stichprobenumfang. Hier sieht man noch einmal, dass die Tests nicht sinnvoll sind (s. Lösung zu Aufgabe 3A).

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
RES_1 Unstandardized Residual	,163	16	,200*	,923	16	,191

a. Signifikanzkorrektur nach Lilliefors

*. Dies ist eine untere Grenze der echten Signifikanz.

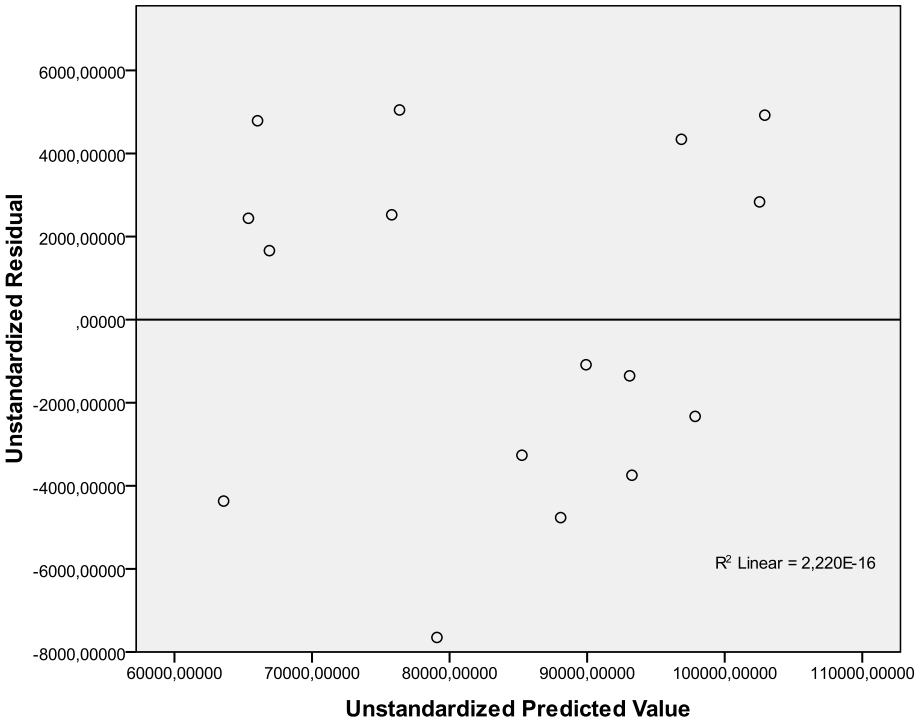
- Heteroskedastizität bedeutet, dass σ_ε^2, die Streuung der Residualvariable ε des (Grundgesamtheits-) Regressionsmodells, sich mit der Höhe einer (oder mehrerer) Erklärungsvariablen verändert (s. Kapitel 18.4.2). Um Grafiken entsprechend den in Abb. 18.7 im Buch zu erzeugen, muss man die Residualwerte e_i ($= y_i - \hat{y}_i$), d. h. die Differenzen zwischen tatsächlichem und per Regressionsgleichung geschätztem Absatz) zunächst speichern und dann im nächsten Schritt in einem Streudiagramm zusammen mit den Erklärungsvariablen darstellen. In der Praxis verzichtet man häufig darauf, für jede der Erklärungsvariablen ein Streudiagramm mit e_i auf der senkrechten Achse zu erzeugen. Man erzeugt dann nur ein Diagramm mit e_i auf der senkrechten und $\hat{y}_i$ (dem per Regressionsgleichung geschätztem Absatz, also den Vorhersagewerten von Absatz) auf der waagerechten Achse.

- Ein einfaches Streudiagramm bietet das gewünschte Diagramm ("Grafik", "Diagrammerstellung…", Auswählen von „Streu-/Punktdiagramm“. Durch Doppelklicken auf das Symbol für ein „Einfaches Streudiagramm" dieses in die Diagrammvorschau übertragen. Ziehen von RES_1 auf

„Y-Achse?“ und PRE_1 auf „X-Achse?“). Für die unten angeführte Darstellung wurde im Diagramm-Editor eine Regressionsgerade in die Punktwolke des Diagramms eingefügt.

Es gibt eine zweite Möglichkeit, eine derartige Grafik zu erstellen. Im Unterschied zu dieser Grafik werden auf den Achsen die Vorhersagewerte und die Residualwerte als standardisierte Z-Werte (s. Lösung zu Aufgabe 2g) dargestellt. Um ein derartiges Streudiagramm zu erstellen, wird in der Dialogbox "Lineare Regression" die Schaltfläche "Diagramme" geklickt. In der Unterdialogbox "Lineare Regression: Diagramme" wird in das Eingabefeld "Y:" des Feldes "Streudiagramm 1 von 1" die Variable ZRESID (standardisierte Residualwerte) und in das Eingabefeld für "X:" die Variable ZPRED (standardisierte Vorhersagewerte) übertragen.

- Aus dem Streudiagramm ist deutlich zu erkennen, dass die Streuung der Residualwerte von der Höhe der Vorhersagewerte unabhängig ist. Die Streuung der Residualwerte nimmt mit Zunahme der Vorhersagewerte weder zu noch ab. Man kann insofern davon ausgehen, dass das Regressionsmodell frei von Heteroskedastizität ist, also Homoskedastizität vorliegt. Diese Modellvoraussetzung der klassischen linearen Regression ist demnach erfüllt.



c.

- Die Daten der Datei ABSATZ.SAV beziehen sich auf Verkaufsbezirke eines Unternehmens. Da es sich bei den Variablen nicht um Zeitreihen handelt, kann es keine Autokorrelation (auch serielle Korrelation genannt) geben.

d.

- Die Hypothese lautet, dass mit höherer Anzahl von Mailings der Absatz steigt. Das erwartete Vorzeichen des Regressionskoeffizienten für die Erklärungsvariable MAILING ist also positiv.

- Die Regressionsanalyse wird wie oben durchgeführt unter Hinzunehmen von MAILING als weitere unabhängige Variable. Um die partiellen Diagramme anzufordern, die als Fallbeschriftung die Verkaufsbezirksnummern enthalten sollen, wird in der Dialogbox "Lineare Regression" in

das Eingabefeld von "Fallbeschriftungen" die Variable BEZIRK übertragen. Danach wird die Schaltfläche "Diagramme" zur Öffnung der Unterdialogbox "Lineare Regression: Diagramme" geklickt. In dieser wird "Alle partiellen Diagramme erzeugen" angefordert.

- Der Regressionskoeffizient der Variable MAILING hat erwartungsgemäß ein positives Vorzeichen. Die Werte der Regressionskoeffizienten der anderen Erklärungsvariablen verändern sich. Der Wert des Regressionskoeffizienten von ADM verändert sich außergewöhnlich stark (von 6.723,478 auf -1372,882). Sogar das Vorzeichen des Regressionskoeffizienten von ADM wechselt. Es wird negativ und widerspricht damit der Vorzeichenerwartung. Diese starke Veränderung in der Höhe des Regressionskoeffizienten von ADM bei Einschluss der Variable MAILING in das Regressionsmodell ist ein Indikator für eine starke Korrelation von ADM und MAILING. Durch Einschluss der Variable MAILING kommt es zur Multikollinearität (s. Kapitel 18.4.4). Es macht keinen Sinn, beide Variablen gleichzeitig in das Erklärungsmodell aufzunehmen, da – bedingt durch eine hohe Korrelation von ADM und MAILING – beide Variablen sich in ihrem Einfluss auf ABSATZ nicht unterscheiden.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressions-koeffizient B	Standard-fehler	Beta		
1	(Konstante)	203009,252	23379,266		8,683	,000
	Verkaufspreis	-3711,737	435,531	-,702	-8,522	,000
	Anzahl der Außendienstmitarbeiter	-1372,882	5815,151	-,136	-,236	,818
	Werbebudget	,078	,024	,274	3,288	,007
	Anzahl der Mailings	7,949	5,654	,809	1,406	,187

a. Abhängige Variable: Absatzmenge

- Der Regressionskoeffizient der Variable MAILING ist nicht signifikant von 0 verschieden. Bei n-m-1 = 16 - 4 - 1 = 11 Freiheitsgraden und einem unterstellten Signifikanzniveau von $\alpha = 0{,}05$ ergibt sich aus einer tabellierten t-Verteilung (sie kann von den Internetseiten zum Buch herunter geladen werden) ein kritischer Wert von 1,796. Der empirische t-Wert unterschreitet mit t = 1,406 den kritischen Wert aus der tabellierten t-Verteilung. Die H_0-Hypothese (kein Einfluss von MAILING auf ABSATZ) wird demgemäß angenommen.

- Aber auch der Regressionskoeffizient von ADM ist in diesem Regressionsmodell – im Unterschied zum vorherigen Modell – nicht signifikant von 0 verschieden. Die Multikollinearität des Modells führt zu widersprüchlichen Ergebnissen. Durch die hohe Korrelation von ADM und MAILING ist es nicht möglich, den separaten Einfluss der beiden Variablen auf den Absatz zu messen. Man sollte nur eine der beiden Variablen in das Modell aufnehmen und bei der Modellevaluation berücksichtigen, dass die einbezogene Variable den Einfluss beider Variablen widerspiegelt.

- Zur Prüfung eines Modells auf Multikollinearität bietet SPSS statistische Kennziffern für eine Kollinearitätsdiagnose an (s. Kapitel 18.2.2). In der Dialogbox "Lineare Regression" klickt man die Schaltfläche "Statistiken…" zur Öffnung der Unterdialogbox "Lineare Regression: Statistiken". Hier fordert man "Kollinearitätsdiagnose" an. An die Tabelle mit den Regressionskoeffizienten werden Kennziffern zur Diagnose von Multikollinearität angehängt: "Toleranz" und "VIF". Der Wert von "Toleranz" einer Erklärungsvariable gibt an, wie hoch der Varianzanteil dieser Variable ist, der durch die anderen unabhängigen Variablen in der Gleichung nicht erklärt wird. Der Wert von "VIF" (Variance Inflation Factor) ist der Kehrwert von "Toleranz". Eine Va-

riable mit kleiner Toleranz (und damit hohem VIF) trägt wenig zur Vorhersage der abhängigen Variable bei. Mit abnehmender "Toleranz" (d. h. zunehmendem VIF) steigt auch die Varianz des Regressionskoeffizienten, wodurch er zu einer instabilen Schätzung wird (s. die starke Veränderung des Regressionskoeffizienten von ADM). Die sehr geringen Toleranzwerte (bzw. hohe VIF) von ADM und MAILING zeigen an, dass das Regressionsmodell kein gutes Modell ist, da es mit Multikollinearität behaftet ist.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	Kollinearitätsstatistik	
		Regressions-koeffizient B	Standard-fehler	Beta			Toleranz	VIF
1	(Konstante)	203009,252	23379,266		8,683	,000		
	Verkaufspreis	-3711,737	435,531	-,702	-8,522	,000	,921	1,086
	Anzahl der Außendienstmitarbeiter	-1372,882	5815,151	-,136	-,236	,818	,019	52,968
	Werbebudget	,078	,024	,274	3,288	,007	,899	1,113
	Anzahl der Mailings	7,949	5,654	,809	1,406	,187	,019	52,962

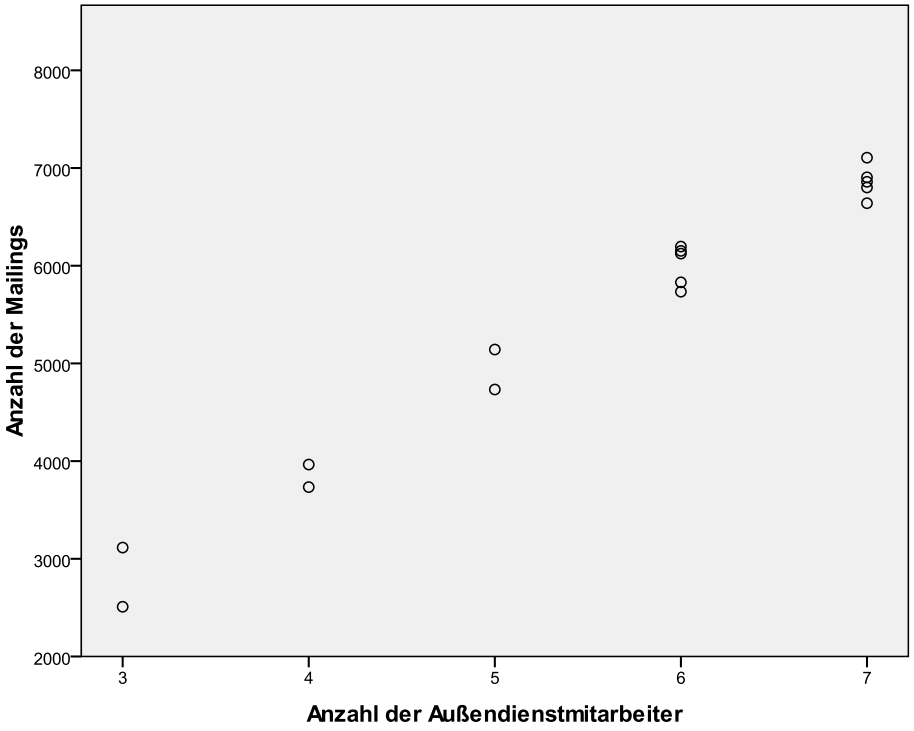
a. Abhängige Variable: Absatzmenge

- Der Korrelationskoeffizient von ADM und MAILING liegt mit 0,989 sehr nahe bei 1. Auch das Streudiagramm zeigt den starken linearen Zusammenhang zwischen diesen beiden Variablen.

Korrelationen

		Anzahl der Außendienstmitarbeiter	Anzahl der Mailings
Anzahl der Außendienstmitarbeiter	Korrelation nach Pearson	1	,989**
	Signifikanz (1-seitig)		,000
	N	16	16
Anzahl der Mailings	Korrelation nach Pearson	,989**	1
	Signifikanz (1-seitig)	,000	
	N	16	16

**. Die Korrelation ist auf dem Niveau von 0,01 (1-seitig) signifikant.

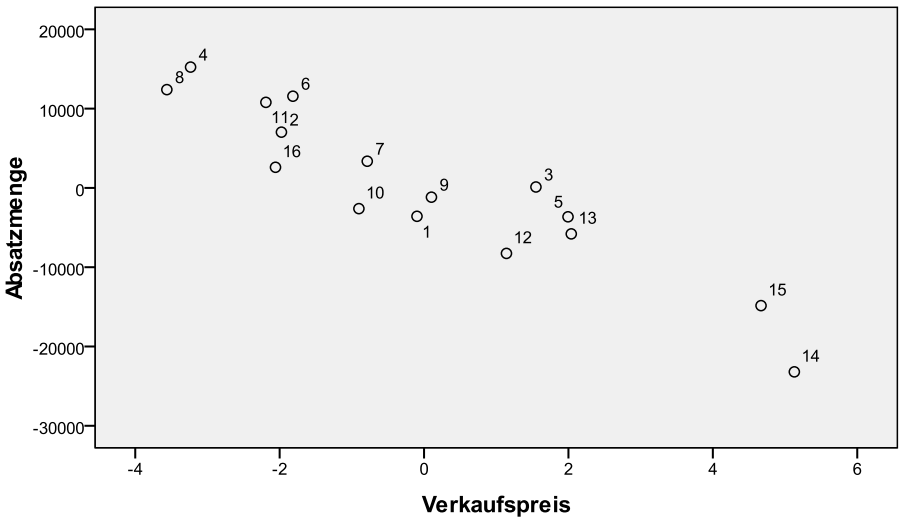


- Im Folgenden sind die 4 partiellen Regressionsdiagramme zu sehen. Um Sie zu erhalten, markiert man in der unterdialogobox „Lineare Regression: Diagramme“ das Auswahlkästchen „Alle partiellen Diagramme erzeugen“. Es handelt sich dabei um 4 Streudiagramme. Die Variablenwerte, die auf den beiden Achsen der jeweiligen Diagramme abgetragen sind, werden wie im Fall der Berechnung von partiellen Korrelationskoeffizienten berechnet (s. Lösung zu Aufgabe 8). Der lineare Einfluss aller anderen Erklärungsvariablen ist aus beiden im Streudiagramm abgebildeten Variablen herausgerechnet.

- Im Diagramm zur Darstellung des partiellen Zusammenhangs zwischen ABSATZ und PREIS wird deutlich sichtbar, dass ein negativer starker linearer Zusammenhang besteht. Die Ziffern geben die Verkaufsbezirke an. Der partielle Zusammenhang zwischen ABSATZ und WERBUNG ist positiv, aber weniger stark ausgeprägt. Zwischen ABSATZ und ADM sowie zwischen ABSATZ und MAILING bestehen keine (partiellen) korrelative Zusammenhänge. Dieses liegt daran, dass ADM und MAILING sehr hoch korrelieren, d. h. praktisch den gleichen Einfluss auf ABSATZ haben. Wird aus ABSATZ und aus ADM der Einfluss von MAILING herausgerechnet, so verschwindet auch der Einfluss von ADM auf ABSATZ. Analog gilt: Wird der Einfluss von ADM aus ABSATZ und aus MAILING herausgerechnet, so verschwindet damit auch der Einfluss von MAILING auf ABSATZ.

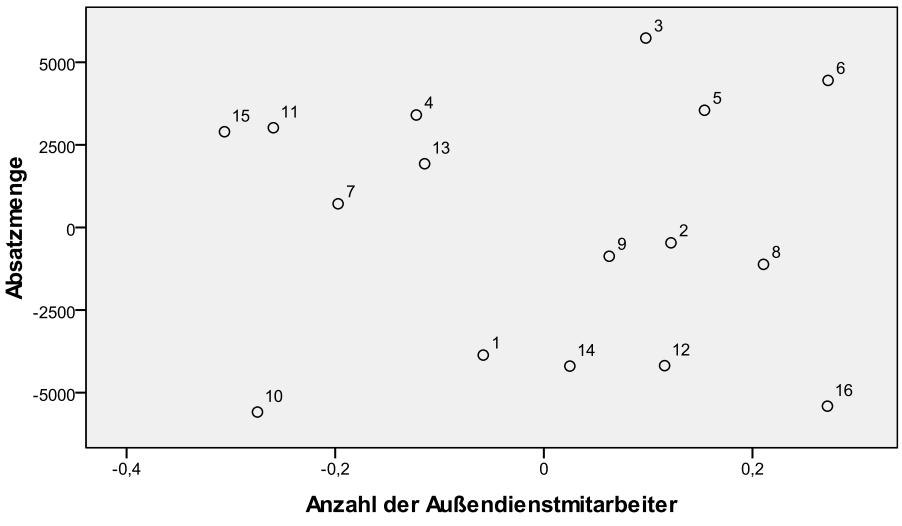
Partielles Regressionsdiagramm

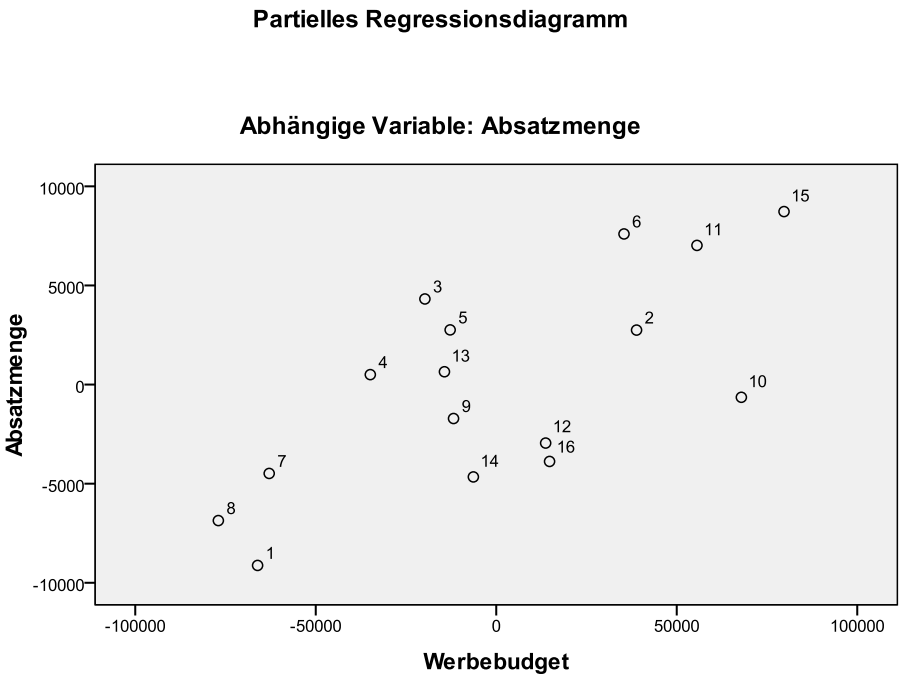
Abhängige Variable: Absatzmenge

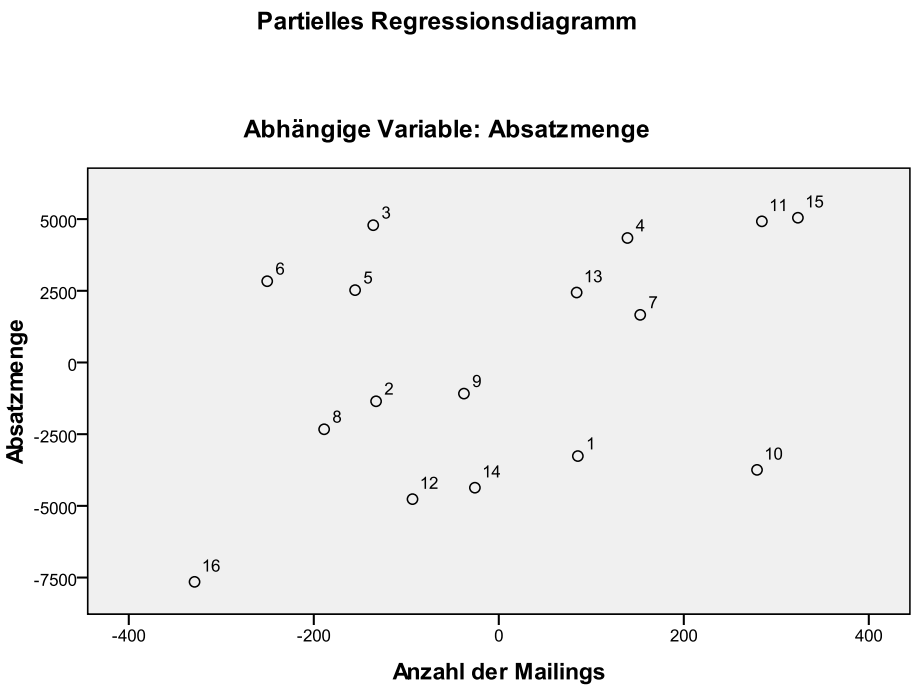


Partielles Regressionsdiagramm

Abhängige Variable: Absatzmenge







- Im Modell bei Einschluss der Variable MAILING erhöht sich das Bestimmtheitsmaß R^2 nur wenig von 0,919 auf 0,931. Entscheidender für die Modellbewertung aber ist, dass das korrigierte Bestimmtheitsmaß sich erhöht und das Maß Standardfehler des Schätzers wesentlich kleiner wird. Das Modell hat insofern schon eine höhere Erklärungskraft.

  Es kommt aber auf die Modellbewertung insgesamt an. Das Modell ist mit starker Multikollinearität behaftet und wird daher als mangelhaft abgelehnt.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,965[a]	,931	,906	4371,811

a. Einflußvariablen : (Konstante), Anzahl der Mailings, Werbebudget, Verkaufspreis, Anzahl der Außendienstmitarbeiter

b. Abhängige Variable: Absatzmenge