LÖSUNG 9A

a.

- Das Regressionsmodell soll nur für Büropersonal angewendet werden Management- und Bewachungspersonal (MIND =0) soll nicht einbezogen werden. Mit "Daten", "Fälle auswählen", "Falls Bedingung zutrifft" wird die Bedingung für die Auswahl der Bankangestellten definiert: TAETIG = 1 & MIND = 0.

- Man kann plausibel davon ausgehen, dass mit längerer Ausbildungszeit sowie mit längerer Dauer der Beschäftigung und mit längerer Berufserfahrung das Gehalt größer sein wird. Für die Regressionskoeffizienten werden daher positive Vorzeichen erwartet.

- Mit „Analysieren“, "Regression", "Linear" wird die Dialogbox "Lineare Regression" aufgerufen. Die Variable GEHALT wird in das Eingabefeld „abhängige Variable“ und die unabhängigen Variablen AUSBILD, DAUER und ERFAHR werden in das entsprechende Eingabefeld übertragen. Als Methode wird "Einschluss" gewählt.

- Das Bestimmtheitsmaß beträgt R^2 = 0,313. In sozialwissenschaftlichen Untersuchungen zählt man das zu schwachen bis mittelstarken Zusammenhängen. Nur weniger als ein Drittel der Varianz von GEHALT wird durch die Erklärungsvariablen vorhergesagt (erklärt). Es handelt sich also um ein Regressionsmodell mit schwacher Vorhersage(Erklärungs-)kraft.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,560[a]	,313	,306	6.661,227

a. Einflußvariablen : (Konstante), Berufserfahrung in Monaten, Beschäftigungsdauer, Ausbildung (in Jahren)

- Die aufgrund von Plausibilitätsüberlegungen erwarteten positiven Vorzeichen der Regressionskoeffizienten b_i von Ausbildungs- und Beschäftigungsdauer werden empirisch gestützt.

  Das Vorzeichen des Regressionskoeffizienten von Berufserfahrung ist negativ und entspricht somit nicht der Erwartung. Variable mit falschem Vorzeichen sollte man i.d.R. nicht in das Modell aufnehmen.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressions-koeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	-2261,288	3887,336		-,582	,561
	Ausbildung (in Jahren)	1693,601	176,157	,500	9,614	,000
	Beschäftigungsdauer	117,307	39,738	,149	2,952	,003
	Berufserfahrung in Monaten	-7,158	4,340	-,086	-1,649	,100

a. Abhängige Variable: Gehalt

- Zur Prüfung, ob die Regressionskoeffizienten β_i sich signifikant von 0 unterscheiden, wird ein einseitiger t-Test durchgeführt. Der Test soll zur Entscheidung führen, ob die H_0-Hypothese β_i

= 0 oder die H_1-Hypothese (Alternativhypothese) $\beta_i > 0$ angenommen werden soll. Die Prüfvariable

$$t = \frac{b_i - \beta_i}{s_b}$$

(s_b = geschätzte Standardabweichung des geschätzten Regressionskoeffizienten).

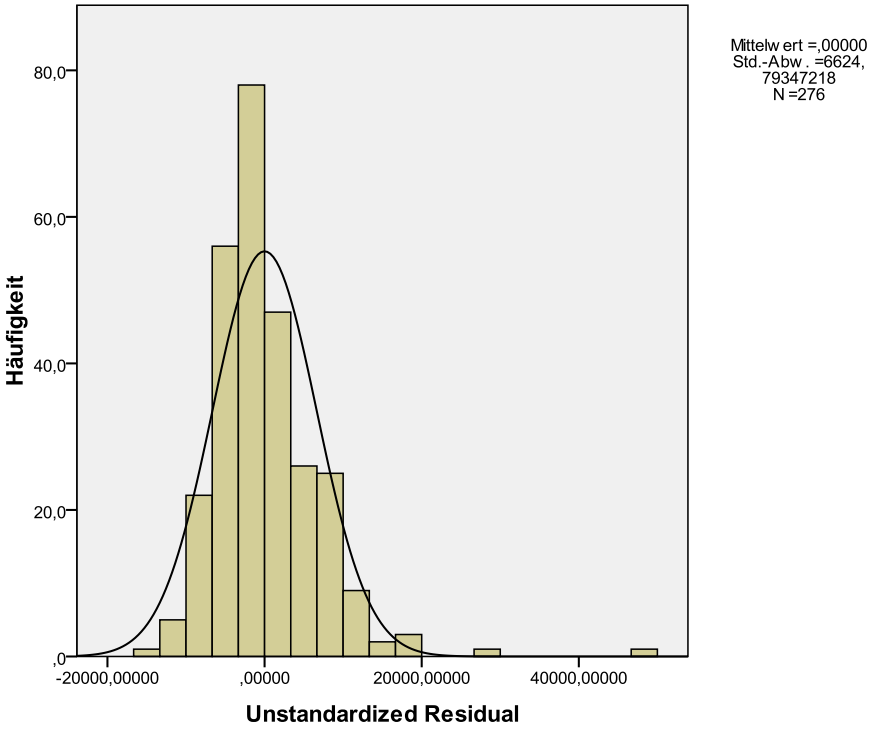
bzw. unter Annahme der H_0-Hypothese $\beta_i = 0$: $t = \frac{b_i}{s_b}$ hat eine t-Verteilung mit n-m-1 Freiheitsgraden [n = Anzahl der Beobachtungen (Fälle), m = Anzahl der erklärenden Variablen]. Bei einem angenommenen Signifikanzniveau von $\alpha = 0{,}025$ und Freiheitsgraden = n-m-1= 276 - 3 - 1 = 272 ergibt sich ein kritischer Wert für t in Höhe von 1,969 (berechnet mit der Funktion IDF.T(0.975,272). Für n > 30 kann die t-Verteilung durch eine Standardnormalverteilung approximiert werden. Aus der Standardnormalverteilungstabelle ergibt sich ein kritischer Wert von 1,96. Da der empirische t-Wert der Prüfgröße für AUSBILD und DAUER größer ist als der kritische, wird die H_0-Hypothese ($\beta_i = 0$) abgelehnt und die Alternativhypothese angenommen. Ausbildungs- und Beschäftigungsdauer haben demgemäß einen signifikanten Einfluss auf die Höhe des Gehalts. Dies ergibt sich auch aus „Sig." $< \alpha = 0{,}025$.

Der Regressionskoeffizient von Berufserfahrung hingegen erweist sich als nicht signifikant („Sig" = 0,10 > α = 0,05). Dies ist ein zweiter Grund, die Variable wieder aus dem Modell zu entfernen.

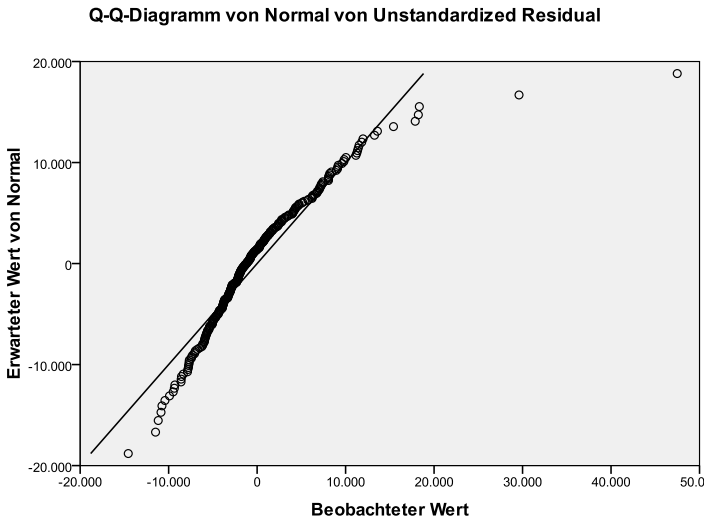
b.

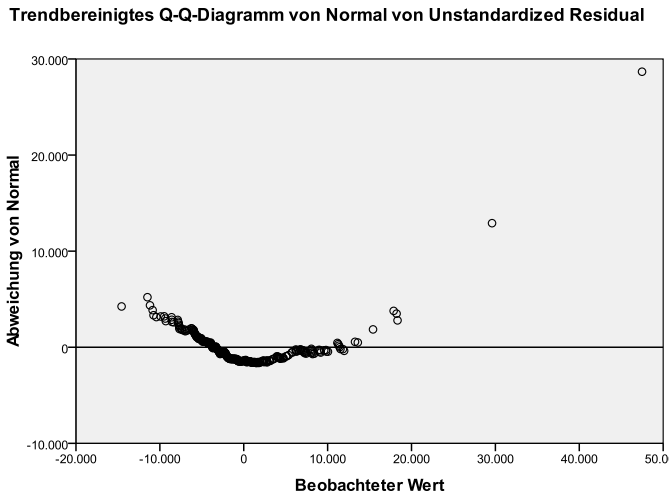
- Um die Prüfungen vornehmen zu können, muss man die Vorhersage- und die Residualwerte speichern. Dazu öffnet man durch Klicken auf die Schaltfläche "Speichern" in der Dialogbox "Lineare Regression" die Dialogbox "Lineare Regression: Speichern" und wählt in den Feldern "Vorhersagte Werte" und "Residuen" die Option "Nicht standardisiert". Die Vorhersage- und die Residualwerte werden als PRE_1 und RES_1 den Variablen im Dateneditor hinzugefügt. Außerdem wird eine Ausgabetabelle mit dem Mittelwert und der Standardabweichung und weitere Daten für die Residualwerte und Vorhersagewerte (standardisiert und nicht standardisiert) erstellt.
- Eine Möglichkeit zu prüfen, ob die Residualwerte annähernd normalverteilt sind, bietet das Erstellen eines Histogramms der Residualwerte ("Grafik", „Diagrammerstellung…", Wählen von "Histogramm", Doppelklicken auf das Symbol für „Einfaches Histogramm", Ziehen von RES_1 auf „X-Achse?". In der Unterdialogbox „Elementeigenschaften" (öffnen durch Klicken auf die Schaltfläche „Elementeigenschaften") in „Eigenschaften bearbeiten von" für „Balken1" das Auswahlkästchen "Normalverteilungskurve anzeigen" markieren).

  Aus dem Histogramm wird deutlich, dass die Verteilung der Residualwerte linkssteil und spitzer ist als eine Normalverteilung. Die Abweichung von der Normalverteilung ist aber nicht gravierend.



- Ergänzend kann man ein QQ-Diagramm der Residualvariable erstellen. („Analysieren“, „Deskriptive Statistiken“, „QQ-Diagramme…“, „RES_1“ in „Variable(n):“ übertragen, als „Testverteilung“ „Normalverteilung“ wählen). Auch in dieser Darstellung werden Abweichungen von der Normalverteilung sichtbar.





- Auch eine Bewertung der Schiefe- und der Wölbungmaße kann hilfreich sein (s. Lösung zu Aufgabe 3a).

- Mit der Prozedur "Explorative Datenanalyse" kann man auch Tests zur Prüfung auf Normalverteilung durchführen. Dabei sollte man den Shapiro-Wilk-Test bevorzugen. Auf das Testergebnis, das eine signifikante Abweichung der Residualwerte von der Normalverteilung anzeigt, sollte man aber wegen der relativ hohen Fallzahl nicht vertrauen.

  Besser ist es, sich auf deskriptive Analysen zu stützen (s. Lösung zu Aufgabe 3a).

Tests auf Normalverteilung

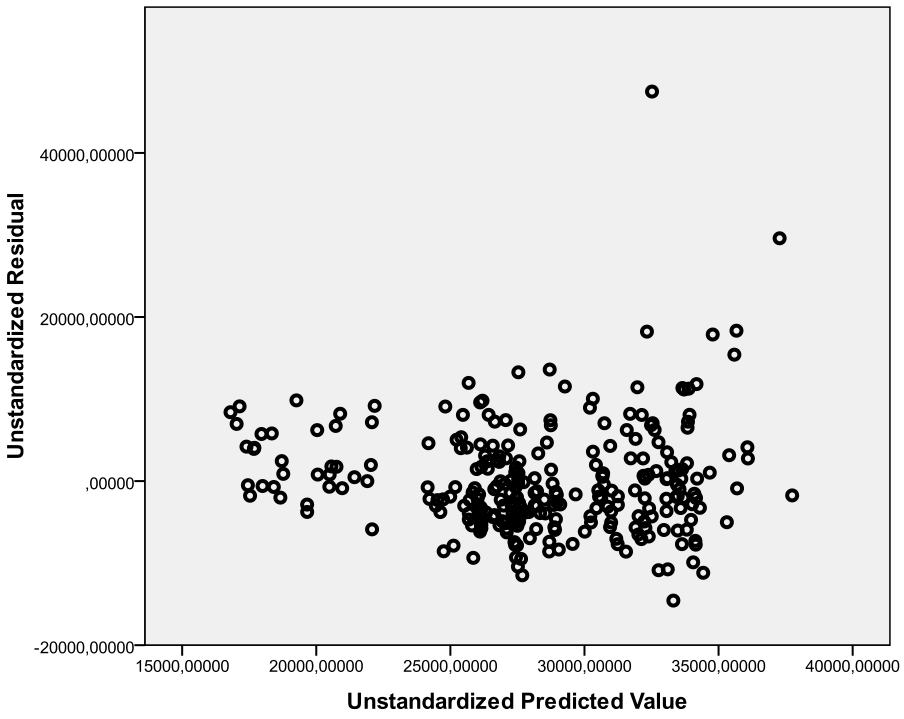
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
RES_1 Unstandardized Residual	,095	276	,000	,883	276	,000

a. Signifikanzkorrektur nach Lilliefors

- Wird die Bedingung der Normalverteilung der Residualwerte verletzt, dann sind die t-Tests zur Prüfung auf Signifikanz der Regressionskoeffizienten unzuverlässig.

  Obwohl die Abweichung der Residualwerte von der Normalverteilung nicht gravierend ist, könnte man prüfen, ob eventuell eine logarithmische Transformation von GEHALT zu einem besseren Regressionsmodell führt.

- Zum Prüfen der Frage, ob die Residualwerte sich mit der Höhe der Vorhersagewerte verändern, wird ein einfaches Streudiagramm erstellt ("Grafik", "Diagrammerstellung…" Auswählen von „Streu-/Punktdiagramm“. Durch Doppelklicken auf das Symbol für „Einfaches Streudiagramm" dieses in die Diagrammvorschau übertragen. Ziehen von RES_1 auf „Y-Achse?“ und PRE_1 auf „X-Achse?“).



- Aus dem Streudiagramm ist tendenziell zu erkennen, dass die Streuung der Residualwerte mit der Höhe der Vorhersagewerte zunimmt. Eine derartige Zunahme der Streuung der Residualwerte bedeutet eine Verletzung einer Modellvoraussetzung der klassischen linearen Regression. Man spricht von Heteroskedastizität der Residualwerte. Erforderlich ist Homoskedastizität (die Streuung der Residualwerte sollte mit größer werdenden Vorhersagewerten konstant bleiben).

  Bei Vorliegen von Heteroskedastizität sind die Signifikanztests zur Prüfung der Regressionskoeffizienten nicht zuverlässig und können zu falschen Testergebnissen führen. Daher sollte man versuchen, ein besseres Regressionsmodell zu entwickeln. Als Lösung bietet es sich an, die abhängige Variable zu transformieren. So könnte z.B. geprüft werden, ob eine Logarithmierung oder ein Quadratwurzelziehen von GEHALT ein besseres Modell ermöglicht.

c.

- Damit die nominalskalierte Variable GESCHL als eine erklärende Variable in ein Regressionsmodell aufgenommen werden kann, muss sie als Dummy-Variable (als 0/1-Variable) kodiert sein (s. Kapitel 18.3). Die Variable GESCHL ist eine Stringvariable mit den Variablenwerten m (männlich) und w (weiblich) und kann daher für die Regressionsanalyse zunächst nicht als erklärende Variable genutzt werden. Sie wird daher in die Variable GESCHL1 umkodiert: aus den alten Variablenwerten m und w von GESCHL werden die neuen Werte von GESCHL1 1 für m und 0 für w ("Transformieren", "Umkodieren in andere Variable", "Eingabevar:" GESCHL; "Ausgabevar:" GESCHL1, „Ändern“. Schaltfläche "Alte und neue Werte", öffnet eine Unterdialogbox. Dort "Alter Wert": m , "Neuer Wert": 1 eintragen und "Hinzufügen" klicken sowie "Alter Wert": w, "Neuer Wert": 0 eintragen und "Hinzufügen". Mit „Weiter“ und „OK“ ausführen.).

- Durchführen der Analyse wie oben (aber ohne ERFAHR) und unter Hinzunehmen von GESCHL1 als weitere unabhängige Variable.

- Bei Ergänzung der Erklärungsvariablen DAUER und AUSBILD um GESCHL1 erhöht sich das Bestimmtheitsmaß R^2 von 0,306 auf 0,374.

  Entscheidender für die Modellbewertung aber ist, dass sich das korrigierte Bestimmtheitsmaß erhöht (von 0,301 auf 0,367) und das Maß Standardfehler des Schätzers wesentlich kleiner wird (von 6682,174 auf 6362,485). Das Modell hat eine höhere Erklärungskraft. Diese ist aber nach wie vor gering bis mittel.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,611[a]	,374	,367	6.362,485

a. Einflußvariablen : (Konstante), geschl1, Beschäftigungsdauer, Ausbildung (in Jahren)

Zum Vergleich: „Modellzusammenfassung“ für Erklärungsvariable DAUER und AUSBILD

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,554[a]	,306	,301	6.682,174

a. Einflußvariablen : (Konstante), Ausbildung (in Jahren), Beschäftigungsdauer

- Alle Regressionskoeffizienten haben das erwartete positive Vorzeichen. Ausbildungs- und Beschäftigungsdauer erhöhen das Gehalt. Dieses gilt auch für die Variable GESCHL1, da man für die Männer ein höheres Einkommen bei gleicher Ausbildungs- und Beschäftigungsdauer erwartet. Es gibt also keinen Widerspruch zum erwarteten Ergebnis.

- Alle Regressionskoeffizienten sind signifikant (s. ausführlicher oben).

- Der Regressionskoeffizient der Dummy-Variable GESCHL1 gibt an, um wie viel das Gehalt der Männer im Durchschnitt höher ist als bei den Frauen bei gleicher Ausbildungs- und Beschäftigungsdauer (um 4.537,46 $).

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressions-koeffizient B	Standard-fehler	Beta	T	Sig.
1	(Konstante)	-489,524	3650,262		-,134	,893
	Beschäftigungsdauer	108,285	37,996	,137	2,850	,005
	Ausbildung (in Jahren)	1427,294	174,555	,422	8,177	,000
	geschl1	4537,459	840,797	,278	5,397	,000

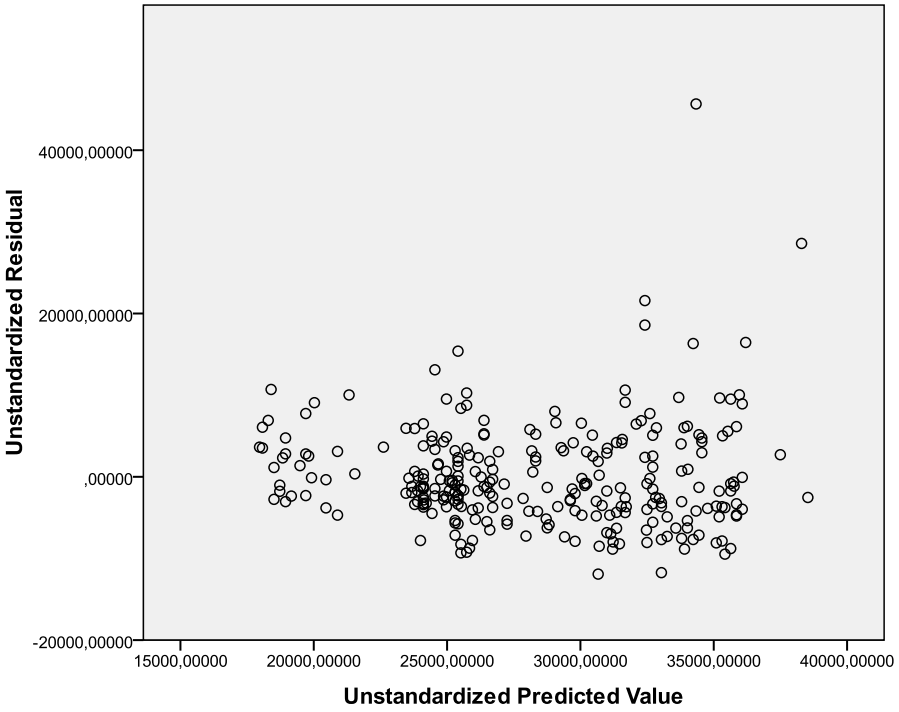
a. Abhängige Variable: Gehalt

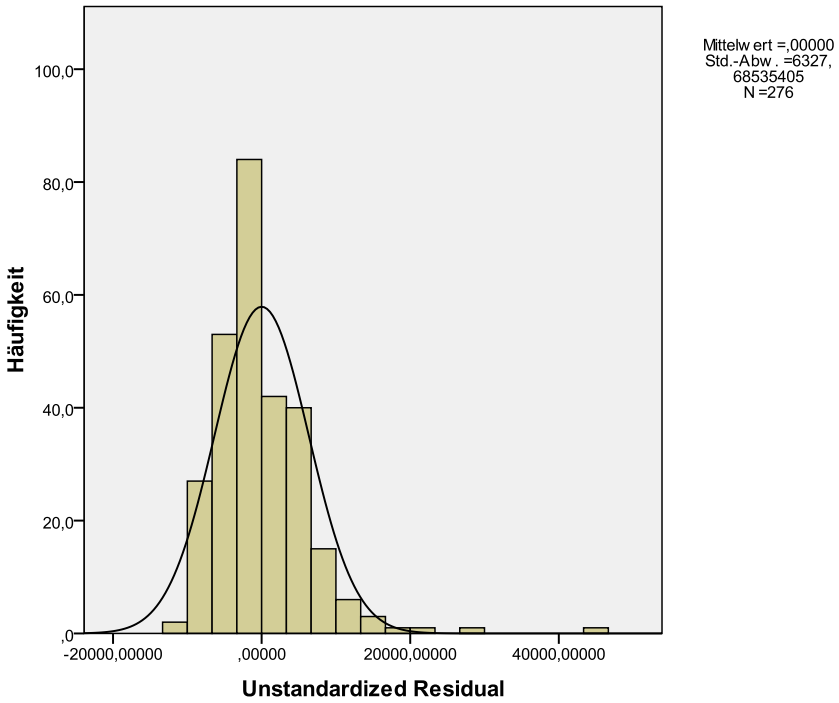
- Das Streudiagramm mit den Residualwerten des neuen Modells auf der Y-Achse und den Vorhersagewerten auf der X-Achse zeigt, dass auch in diesem Modell Heteroskedastizität vorliegt.

  Auch für die Residualvariable gibt es weiterhin – wenn auch keine gravierenden – Abweichungen von der Normalverteilung.

  Es war auch nicht zu erwarten, dass hinsichtlich dieser Kriterien mit dem Einschluss von GESCHL1 eine Modellverbesserung eintritt. Oben wurde als eventueller Lösungsweg zur Verbesserung des Modells eine Transformation von GEHALT vorgeschlagen.

  Insgesamt kann man hinsichtlich einer Verbesserung des Modells also nur auf eine etwas höhere Erklärungskraft des Modells verweisen; das Modell hat einen etwas besseren "Fit".





d.

- Das Vorzeichen des Regressionskoeffizienten von GESCHL1 wechselt von positiv zu negativ, wenn Frauen mit 1 und Männer mit 0 kodiert werden. Der Regressionskoeffizient gibt nun an, um wie viel im Durchschnitt das Gehalt der Frauen kleiner ist als das der Männer bei gleicher Länge der Ausbildungs- und Beschäftigungsdauer.