Lösungsnotizen Aufgabe 40 ../../Aufgabensammlung/Loesungen/L_2_16.tex

a) **Dummykodierung:**
Idee: Erstelle aus der Variable Lesen mit drei Ausprägungen zwei neue Variablen mit je zwei Ausprägungen (sog. Dummyvariablen), die dritte Ausprägung (sog. Referenzkategorie) ergibt sich, wenn die beiden Dummyvariablen den Wert 0 annehmen.

$$\texttt{Lesenoft} = \begin{cases} 1, & \text{falls ja,} \\ 0, & \text{falls nein} \end{cases}$$

$$\texttt{Lesenselten} = \begin{cases} 1, & \text{falls ja,} \\ 0, & \text{falls nein} \end{cases}$$

Referenzkategorie Lesen = mittel, trifft zu, wenn sowohl `Lesenoft`, als auch `Lesenselten` = 0 sind.

b) **Interpretation der Schätzungen der Regressionskoeffizienten**

`(Intercept)`	mittlere 'Basis-Fehlerzahl' (alle Einflussgrößen haben den Wert 0), d.h. die Fehlerzahl für Schüler ohne Leseförderung, aus der 4. Klasse, weiblich, die 'mittel' außerhalb der Schule lesen (Referenzkategorie, da Dummyvariablen nur für andere beiden Ausprägungen im Modell), beträgt 14.8.

Wie verändert sich die Zielgröße, wenn sich jeweils *eine* Einflussgröße um eine Einheit ändert, die anderen Einflussgrößen aber *gleich* bleiben? **(wichtig!)**

`Lesezeitmin`	pro Minute Leseförderzeit in der Schule sinkt die Fehlerzahl um 0.087 (wenn alle anderen Einflussgrößen gleich bleiben!) Förderzeiten in Studie sind 15, 30, 60, 70 Minuten (über `table(Lesezeitmin)`). Bei Schülern mit 30 Min Förderzeit sinkt die Fehleranzahl im Schnitt um $30 \cdot 0.087 = 2.61$.
`Jahrgang`	Schüler in der 3. Klasse machen im Schnitt 5.9 Fehler mehr als Schüler in der 4. Klasse.
`Geschlecht`	Jungs machen im Schnitt 3.2 Fehler mehr als Mädchen. Kann das daran liegen, dass Mädchen einfach öfter lesen? Nein, das haben wir ja auch im Modell, s.u.
`Lesenoft`	Schüler, die öfters lesen, machen im Vergleich zu Schülern, die mitteloft lesen, im Schnitt 2.9 Fehler weniger.
`Lesenselten`	Schüler, die selten lesen, machen im Vergleich zu Schülern, die mitteloft lesen, im Schnitt 4.9 Fehler mehr. D.h. der Unterschied zwischen selten und oft Lesen beträgt $2.93325 + 4.94820 = 7.88145$ Fehler! Mehr noch als der Klassenstufen-Unterschied...

c) Alle Einflussgrößen signifikant ($\alpha = 0.05$), da p-Werte (in Spalte `Pr(>|t|)`) alle kleiner 0.05.

d) KI ($\gamma = 0.99$!) für $\hat\beta_{\texttt{Geschlecht}}$:

$$\left[\hat\beta_{\texttt{Geschlecht}} \pm \hat\sigma_{\hat\beta_{\texttt{Geschlecht}}} \cdot t^{(n-p-1)}_{\frac{1+\gamma}{2}}\right] = \left[3.16466 \pm 1.22758 \cdot t^{(174)}_{0.995}\right] \qquad \left(t^{(174)}_{0.995} \approx z_{0.995} = 2.575829\right)$$
$$= [0.0026,\ 6.3267]$$

Überdeckt 0 nicht trotz $\gamma = 0.99$!

Widerspruch zum p-Wert 0.01077 liegt an Approximation $t^{(174)}_{0.995} \approx z_{0.995}$:

$$t^{(174)}_{0.995} = 2.604379$$
$$z_{0.995} = 2.575829$$

KI mit $t^{(174)}_{0.995}$ ist $[-0.032,\ 6.3617]$ und passt also zum p-Wert. Vorsicht also mit der Approximation bei „knappen“ Entscheidungen!