SKETCH of the SOLUTIONS
STK4900/9900 - 2021

Problem 1

a

- Proportion of smokers among the controls: $\hat{p}_0 = 134/203 = 0.66$;
- Proportion of smokers among the cases: $\hat{p}_1 = 172/194 = 0.89$;
- 95% confidence interval for p_0: $0.66 \pm 1.96 \cdot \sqrt{\frac{0.66(1-0.66)}{203}} = (0.59; 0.72)$;
- 95% confidence interval for p_1: $0.89 \pm 1.96 \cdot \sqrt{\frac{0.89(1-0.89)}{194}} = (0.83; 0.92)$.

The proportion of smokers among the cases are larger than the proportion of smokers among the controls, making us suspect that there is a relationship between smoking and experiencing oral cancer. A proper statistical test should be performed to evaluate this impression.

b

Left-out values:

- A $= 91 \cdot 203/397 = 46.53$;
- B $= 91 \cdot 194/397 = 44.47$;
- C $= 306 \cdot 203/397 = 156.47$;
- D $= 306 \cdot 194/397 = 149.53$.

Statistical test:

- Test statistic:
  $\chi^2_{obs} = \frac{(69-46.53)^2}{46.53} + \frac{(22-44.47)^2}{44.47} + \frac{(134-156.47)^2}{156.47} + \frac{(172-149.53)^2}{149.53} = 28.81$
  (note: using R one may obtain 27.54 due to numerical approximations)
- p-value $= 1 - Pr(\chi^2_1 \leqslant \chi^2_{obs}) \approx 0$.
- The null-hypothesis (independence) is rejected.

c

The missing value is: $z_{\text{obs}} = \frac{\hat{\beta}_{\text{smoker}}}{\hat{\sigma}_{\text{smoker}}} = \frac{1.3927}{0.2706} = 5.147$.
Being a smoker increases the risk of experiencing oral cancer by 1.39 in terms of log-odds.
This means that the odds-ratio is $e^{1.39} = 4.01$, i.e., that the odds of experiencing oral cancer among the smokers is 4 to 1 with respect to the non-smokers. Since 4.01 is larger than 1, we can conclude that it is more likely to find persons with oral cancer among the smokers than among the non-smokers.

d

The value 0.053624 is the expected change in log-odds for a one-unit increase in cigarettes smoked per day.
The variable `smoking` is constructed by dichotomising the variable `cigs`, and in this process part of the information is lost. Knowing how many cigarettes a person smokes is indeed more informative than only knowing if the person smokes or not. As a consequence, the model that uses `smoking` as a explanatory variable can explain less of the total deviance, resulting in a larger residual deviance.

e

The expected change in log-odds is now 0.035480. It is smaller than that computed in the model fitted at the point before as an effect of the presence of other covariates in the model. Part of the increase in the risk of experiencing oral cancer is now explained by other variables, correlated with `cigs`.

f

Statistical test:

- Test statistic: $G_{obs} = D_{\text{restr}} - D_{\text{unrestr}} = 444.27 - 443.84 = 0.43$;
- Degrees of freedom: $393 - 392 = 1$;
- p-value $= 1 - Pr(\chi_1^2 \leqslant G_{obs}) = 0.51$.

The null-hypothesis is not rejected, so the decrease in terms of residual deviance is not statistically significant and the data do not support the presence of the variable `age` in the model. We would go with the latter (restricted) model.

Problem 2

a

Left-out values:

- D $= 4 - 1 = 3$;
- F1 $= \frac{\text{Sum Sq}_{\text{bodystyle}}/\text{Df}_{\text{bodystyle}}}{\text{Sum Sq}_{\text{Residuals}}/\text{Df}_{\text{Residuals}}} = \frac{69.14/3}{1015.67/186} = 4.22$;
- F2 $= \frac{\text{Sum Sq}_{\text{drive.wheels}}/\text{Df}_{\text{drive.wheels}}}{\text{Sum Sq}_{\text{Residuals}}/\text{Df}_{\text{Residuals}}} = \frac{448.71/1}{1015.67/186} = 82.17$;
- F3 $= \frac{\text{Sum Sq}_{\text{interaction}}/\text{Df}_{\text{interaction}}}{\text{Sum Sq}_{\text{Residuals}}/\text{Df}_{\text{Residuals}}} = \frac{13.83/3}{1015.67/186} = 0.84$;
- P1 $= 1 - Pr(F_{3,186} \leqslant 4.22) \approx 0.01$;
- P1 $= 1 - Pr(F_{1,186} \leqslant 82.17) \approx 0$;
- P1 $= 1 - Pr(F_{3,186} \leqslant 0.84) \approx 0.47$.

We can conclude that the car body style and the location of the drive wheels is associated with the distance travelled with one litre of fuel, but there is no interaction among the two covariates. Basically, the effect of car body style on the travelled distance is not influenced by the location of the drive wheels, and the other way around.

b

Both the engine size and the type of fuel have an effect on the distance that the model of car travels with one litre. An increase of 1 cubic decimetre in the engine corresponds to an expected decrease of about 2.77 km per litre. Obviously, a larger engine needs more fuel and the distance travelled decreases. The expected difference in km travelled with 1 litre of fuel between a gas and a diesel car is 2.76, with a gas car that is less economic (less distance travelled with 1 litre) than a diesel car. The intercept here does not have a real meaning, as it is outside the range of engine size considered: it would mean that a diesel car with a 0 dm^3 engine would run 19.03 km with one litre of gas, that makes little sense (no 0 dm^3 engine exists).

Confidence intervals:

- Intercept: $\hat{\beta}_{\text{intercept}} \pm t_{191;0.975} \cdot \text{se}_{\text{intercept}} = (17.79; 20.27)$;
- `engine.size`: $\hat{\beta}_{\text{engine.size}} \pm t_{191;0.975} \cdot \text{se}_{\text{engine.size}} = (-3.17; -2.37)$;
- `fuel`: $\hat{\beta}_{\text{fuel}} \pm t_{191;0.975} \cdot \text{se}_{\text{fuel}} = (-3.67; -1.84)$.

c

The term `fit` is the point estimate for the km travelled with 1 litre by the new model of car. It is calculated as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 19.035 - 2.767 * 2.3 - 2.76 * 0 = 12.67$$

since the new car has engine size $x_1 = 2.3$ cubic decimeter and $x_2 = 0$ for a diesel engine.

The two intervals provide an interval estimate for two different quantities, a mean response in the case of the confidence interval, a prediction of a new value in the case of the prediction interval. As the latter needs to also incorporate the variability related to the new observation (estimated by $s^2_{\mathrm{y|x}}$), it is in general wider.

d

The three plots provide a graphical evaluation of three assumptions of the linear model: linearity, normality and homoscedasticity (equal variance).

Here:

- from the first plot, it seems that the assumption of linearity is violated, as there is a clear path among the residuals, which are not randomly spread like white noise;
- the second plot shows a relatively strong departure from the normality, as many points do not lie on the diagonal;
- the third (but it was also clear from the first) plot shows us that the assumption of homoscedasticity is also violated, as the variance clearly increases by moving toward the right part of the plot(s).