Solution Quiz:

Which of the following statement(s) is/are correct?

(a) In which scenarios are inherently interpretable models usually much harder to interpret?

⇒ E.g. linear models with many features and interactions or decision trees with deep trees are not easy to interpret.

(b) Why does usually interpretability become worse or more difficult if the generalization performance of the model improves?

⇒ Methods become more complex.

(c) Should we always prefer interpretable models? Explain and describe for which use cases interpretable models would be inconvenient?

⇒ If the performance of more complex models is much better than the one of an interpretable model.

(d) In the linear model, the effect and importance of a feature can be inferred from the estimated β-coefficients. Is this statement true or false. Explain!

⇒ **Wrong**, for the importance of a feature in a linear model one has to calculate other statistical quantities such as the t-statistic or the p-value.

(e) What is so special about LASSO compared to a LM with regards to interpretability? Would you always prefer LASSO over a LM?

⇒ Penalty leads to feature selection, is probably often preferable but maybe not always (optimization more difficult, has hyperparameters to tune, inference more difficult → keyword: post-selection inference!)

(f) Do the beta-coefficients of GLM always provide simple explanations with respect to the target outcome to be predicted?

⇒ No, only for GLM with Gaussian link, for logistic regression e.g. interpretations are w.r.t. log-odds which is not understandable for everyone

(g) Explain the feature importance provided by model-based boosting. What is the difference to the (Gini) feature importance from decision trees?

(h) How can we use inherently interpretable models to provide insights whether two features are dependent?

⇒ Model x_1 on x_2 (linear or non-linear) and look at the goodness of fit measures like R^2

(i) What are the disadvantages of CART? What methods address them and how?

⇒ Two problems:

1. Selection bias towards high-cardinal/continuous features
2. Does not consider significant improvements when splitting ($\rightsquigarrow$ overfitting)

Solution provided by unbiased recursive partitioning via conditional inference trees (`ctree`) or model-based recursive partitioning (`mob`): Separate selection of feature used for splitting and split point AND hypothesis test as stopping criteria

Solution 1:

	LM			GAM		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	0.38	0.12 – 0.65	**8.851e-03**	0.38	0.35 – 0.42	**3.196e-07**
x1	-0.01	-0.42 – 0.41	9.749e-01			
s(x1)						**2.542e-05**
Observations	11			11		
R^2 / R^2 adjusted	0.000 / -0.111			0.988		

The R^2-value for the GAM model is the adjusted one.

What is the **Adjusted R^2**?

Problem with R^2: R^2 rises each time a feature is added, no matter if the added feature improves the fit or not.

The adjusted R^2 considers the number of terms in a model and only increases after adding a feature, if the fit becomes better:

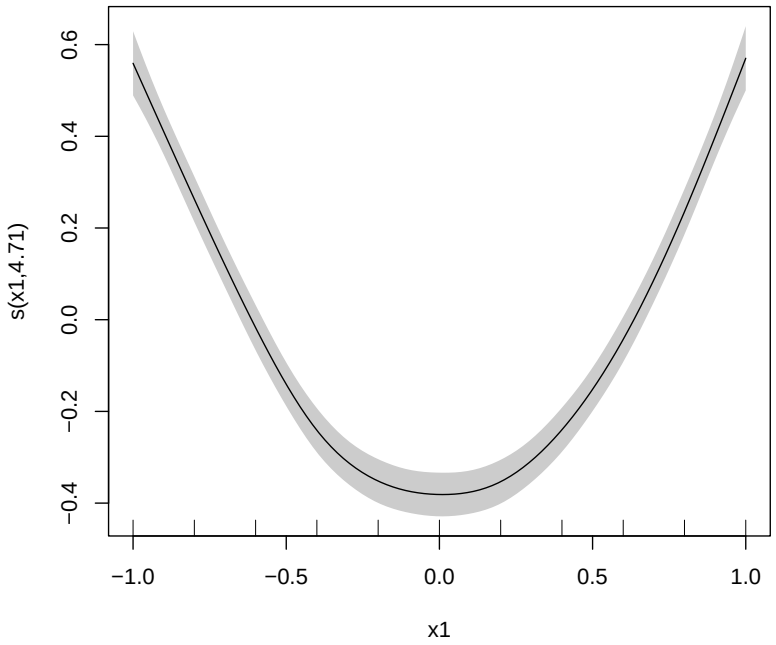
$$\text{adj. } R^2 = 1 - \frac{SSE_{LM} \,/\, (n-p-1)}{SSE_c \,/\, (n-1)},$$

where n is the sample size and p is the number of features.

Hence, it is the proportion of explained variance using unbiased estimates.

a) LM: Since the β-values are not scaled, they are not well interpretable. From the high p-value and $R^2 = 0$ it can be inferred that there is no linear relationship. (Remember: $R^2 = \rho^2$.)

b) GAM: The p-value of 2.54×10^{-5} (but also the adjusted R^2 value) reveals that there is a strong relationship between x_1 and x_2. Hence, generalized additive models (GAM) are able to detect non-linear relationships. The p-value does not reveal the shape of this relationship.



Solution 2:

	WINTER	SPRING	SUMMER	FALL	Σ
y=0	174.00	111.00	98.00	128.00	511.00
y=1	7.00	73.00	90.00	50.00	220.00
Σ	181.00	184.00	188.00	178.00	731.00

a) Odds for "high number of bike rentals" vs. "low to medium number of bike rentals" in winter:

$$\text{odds} = \frac{P(y = 1 \,|\, \texttt{season} = \text{WINTER})}{P(y = 0 \,|\, \texttt{season} = \text{WINTER})} = \frac{7}{174} = 0.04$$

Interpretation: In winter the occurrence of `cnt` > 5531 (y=1) is 0.04 times as likely as `cnt` ≤ 5531 (y=0).

b) Odds Ratio:

$$\begin{aligned}\text{odds ratio} &= \frac{P(y = 1 \,|\, \texttt{season} = \text{SPRING}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{SPRING})}{P(y = 1 \,|\, \texttt{season} = \text{WINTER}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{WINTER})} \\ &= \frac{73/111}{7/174} = 16.35\end{aligned}$$

Interpretation: there is a 16.35 times higher chance of having "high bike rentals" in season SPRING compared to the reference category (WINTER).

c) Table:

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-3.2131	0.3854	0.0000
seasonSPRING	2.7941	0.4138	0.0000
seasonSUMMER	3.1280	0.4121	0.0000
seasonFALL	2.2731	0.4199	0.0000

The intercept gives the odds for "high number of bike rentals" vs. "low to medium number of bike rentals" in winter: $\exp(-3.2131) = 0.04$. Interpretation as in a).

Regarding the estimate of seasonSPRING: odds ratio (when season changes from winter to spring) $= \exp(2.7941) = 16.35$. Interpretation as in b).

d) Table:

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-8.5176	1.2066	0.0000
seasonSPRING	1.7427	0.5977	0.0035
seasonSUMMER	-0.8566	0.7660	0.2635
seasonFALL	-0.6417	0.5543	0.2470
temp	0.2902	0.0391	0.0000
hum	-0.0627	0.0124	0.0000
windspeed	-0.0925	0.0305	0.0024
days_since_2011	0.0166	0.0014	0.0000

If all features are considered in the model, the β-value for the intercept is higher in absolute terms, but the odds changes to $\exp(-8.5176) = 0.0002$, i.e. the probability of "high number of bike rentals" is even less in winter when considering the full model compared to the one only containing feature `season`. Also the higher chance of having "high bike rentals" in season SPRING compared to WINTER declined to $\exp(1.7427) = 5.71$ (vs. 16.35 in the smaller model).

Solution 3:

iteration	baselearner	risk_reduction
1	days_since_2011	140 782.94
2	temp	135 986.35
3	days_since_2011	110 314.74
4	temp	106 854.21
5	temp	86 551.91

In each iteration the difference in risk reduction is given. The feature importance can be calculated by summing up this difference in each iteration to the corresponding feature (which is equivalent to the baselearner). The feature importance looks like this:

feature	risk_reduction
days_since_2011	251 097.68
temp	329 392.47