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.

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

Problem with normal R^2: It 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:

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

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

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

Interpretation: The very low values for R^2 for the linear model show that no variance can be explained by a linear model, so there is no linear relationship, whereas the high adjusted R^2 for the GAM shows that there is a strong relationship found by the GAM.

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

c) GAM: The p-value of 2.54×10^{-5} (as 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, but we can visualize the relationship learned by the GAM incl. the confidence band, see figure 1.

Solution 2:

Use this as recap exercise in case this is forgotten

Solution 3:

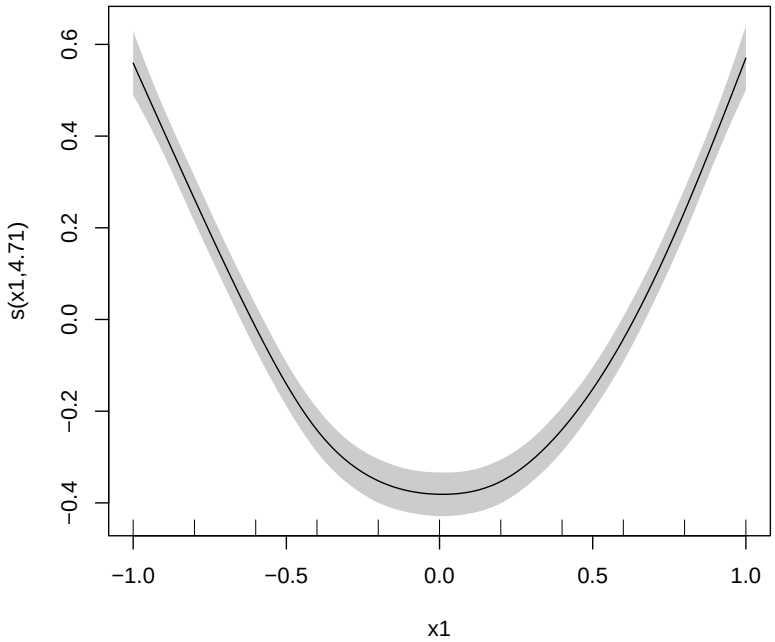


Figure 1: Visualization of the GAM output over a certain range, showing the relationship learned by the GAM

	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 ($y = 0$), the odds are $1 : 25$, which means "$y = 0$" is 25 times as likely than "$y = 1$".

b) Odds Ratio of spring vs. winter:

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

For summer we get

$$\text{odds ratio} = \frac{P(y = 1 \,|\, \texttt{season} = \text{SUMMER}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{SUMMER})}{P(y = 1 \,|\, \texttt{season} = \text{WINTER}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{WINTER})} = \frac{90/98}{7/174} = 22.83,$$

and for fall:

$$\text{odds ratio} = \frac{P(y = 1 \,|\, \texttt{season} = \text{FALL}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{FALL})}{P(y = 1 \,|\, \texttt{season} = \text{WINTER}) \,/\, P(y = 0 \,|\, \texttt{season} = \text{WINTER})} = \frac{50/128}{7/174} = 9.71.$$

Interpretation: The chances (the odds) of having "high bike rentals" are 16.35 times higher in season SPRING compared to the reference category (WINTER). As in winter the odds of "$y = 0$" are $25 : 1$, this means in spring they are roughly $(25 : 16) : 1$, which means roughly $5 : 3$. Similarly, in summer the odds are 22.83 times higher that in winter, which means that in summer they are close to $1 : 1$ (since in winter they were $1 : 25$), so the chances in summer are roughly 50-50.

c) Table:

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

	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

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

d) (i) **Offset.** $\delta = (-0.0627)(62.79) + (-0.0925)(12.76) + (0.0166)(365.00) = \boxed{0.94}$.

(ii) **Probabilities.**

Effective intercept: $\beta_0 + \delta = -7.58 \qquad \eta(x_1) = -7.58 + 0.29\,x_1, \qquad p(x_1) = \sigma\big(\eta(x_1)\big).$

x_1 (°C)	$\eta(x_1)$	$p(x_1)$
10	-4.677	0.009
15	-3.227	0.038
20	-1.777	0.144
25	-0.327	0.419
30	1.123	0.755
35	2.573	0.929

(iii) **Derivation of the marginal effect.**

Goal: Compute the instantaneous change in the predicted probability when temperature x_1 increases, holding all other features constant:

$$\frac{\partial p}{\partial x_1} \;=\; \frac{\partial}{\partial x_1}\sigma\big(\eta\big), \qquad \eta = \beta_0 + \delta + \beta_1 x_1,$$

where $p = \sigma(\eta) = \big(1 + e^{-\eta}\big)^{-1}$ and δ collects the fixed contributions of the remaining covariates.

i. **Applying the chain rule**

$$\frac{\partial p}{\partial x_1} \;=\; \frac{\partial \sigma(\eta)}{\partial \eta}\,\frac{\partial \eta}{\partial x_1}.$$

ii. **Derivative of the sigmoid**

We begin with the definition of the sigmoid function:

$$\sigma(\eta) = \frac{1}{1 + e^{-\eta}}.$$

We aim to compute its derivative with respect to η:

$$\frac{d}{d\eta}\left(\frac{1}{1 + e^{-\eta}}\right).$$

Rewriting using the power rule:

$$\sigma(\eta) = (1 + e^{-\eta})^{-1},$$

so by the chain rule:

$$\frac{d\sigma}{d\eta} = -1 \cdot (1 + e^{-\eta})^{-2} \cdot \frac{d}{d\eta}(1 + e^{-\eta}) = \frac{e^{-\eta}}{(1 + e^{-\eta})^2}.$$

To express this in terms of $\sigma(\eta)$, note:

$$\sigma(\eta) = \frac{1}{1 + e^{-\eta}}, \quad \text{and} \quad 1 - \sigma(\eta) = \frac{e^{-\eta}}{1 + e^{-\eta}}.$$

Hence:

$$\sigma(\eta)\,(1 - \sigma(\eta)) = \frac{1}{1 + e^{-\eta}} \cdot \frac{e^{-\eta}}{1 + e^{-\eta}} = \frac{e^{-\eta}}{(1 + e^{-\eta})^2}.$$

So the derivative simplifies to:

$$\frac{\partial p}{\partial \eta} = \frac{d\sigma}{d\eta} = \sigma(\eta)\,(1 - \sigma(\eta)) = p\,(1 - p).$$

iii. **Derivative of the linear predictor**

$$\frac{\partial \eta}{\partial x_1} = \beta_1, \qquad \text{because } \eta = \beta_0 + \delta + \beta_1 x_1.$$

iv. **Combining the two results**

$$\boxed{\frac{\partial p}{\partial x_1} = p\,(1-p)\,\beta_1}.$$

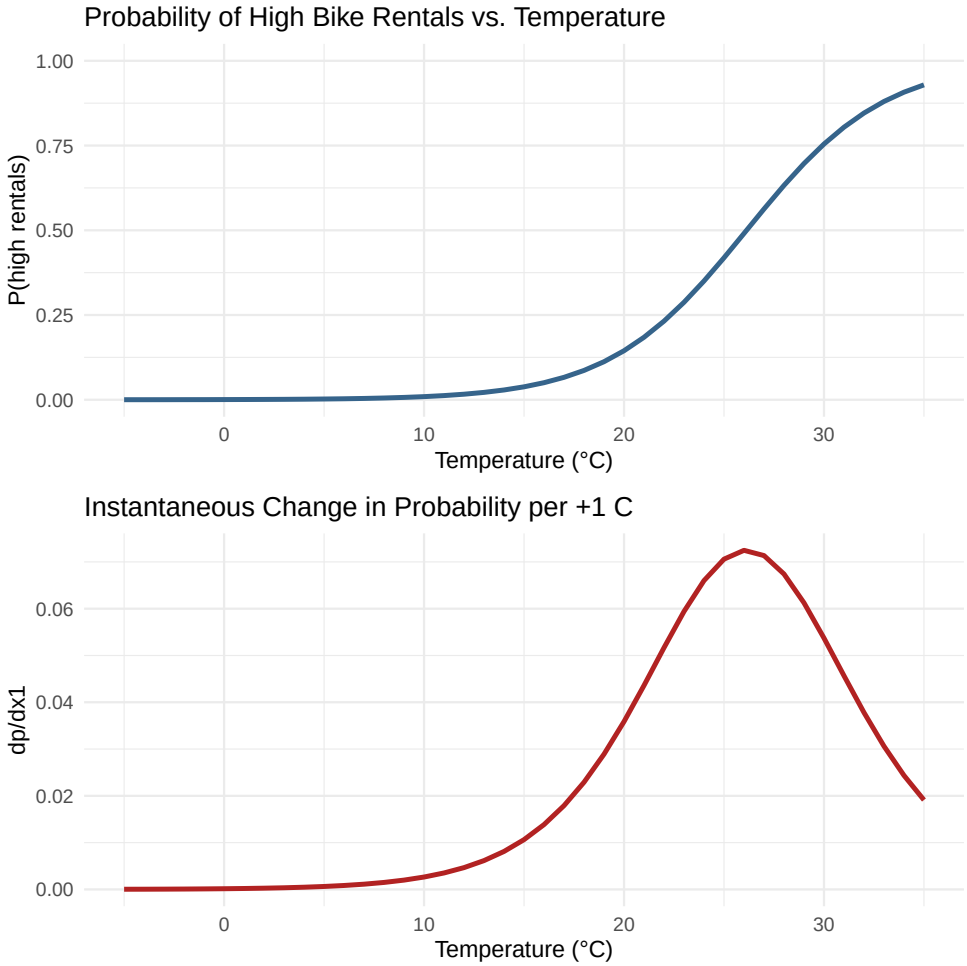
v. **Interpretation**
The factor $p(1-p)$ is maximal at $p = 0.5$ and vanishes as $p \to 0$ or $p \to 1$, illustrating that x_1 has the largest marginal effect when the model is most uncertain and a negligible effect in the extreme–probability regions, when the model is extremely confident. The coefficient β_1 scales this intrinsic sensitivity, so the overall effect is *context-dependent*: It varies with the current value of the linear predictor through p.

The scaling term $p(1-p) \leq 0.25$ peaks at $p = 0.5$, explaining why the same coefficient $\beta_1 = 0.29$ produces the largest probability change near the mid-range and almost none in the tails.

(iv) **Numeric marginal effects:** $dp/dx_1 = p(1-p)\beta_1 = 0.29\,p(1-p)$.
$\Rightarrow$ Largest effect occurs near $x_1 \approx 30°\text{C}$ where $p \approx 0.5$.

x_1	p	dp/dx_1
15 °C	0.038	0.011
30 °C	0.755	**0.054**
35 °C	0.929	0.019

Visualization of the marginal effect:



Side note: The curve for the probability is actually a PDP (see later in chapter on feature effects), because it shows the marginal effect of a single feature when all other features are averaged. If other fixed values are taken for the other features (e.g. values of some fixed single data point), the offset changes, in which case the sigmoid curve always looks the same and is only shifted depending on the other values. This is the difference between a PDP and an ICE plot, which we will discuss later in more detail.

(v) **Classification** at threshold 0.5: Predicted "high-rental" for temperatures with $p \geq 0.5$: $\boxed{30°\text{C}, 35°\text{C}}$.

e) 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

Again, look at your interpretations of the β-coefficients and of the effects of single features from parts c) and d). What changes now in the full model?

If all features are considered in the model, the β-value for the intercept is almost the same as in the second model and much higher in absolute terms than in the first model. Compared to part c), the odds change to $\exp(-8.5176) = 0.0002$, i.e. the probability of "high number of bike rentals" is (on average!) even less in winter when considering the full model compared to the one only containing feature `season`. Also, the average higher chance / higher odds of having "high bike rentals" in season SPRING compared to WINTER declined to $\exp(1.7427) = 5.71$ (vs. 16.35 in part b) in the smaller model).

Comparing to part d), one can observe that the estimates for all the coefficient stay almost the same, and additionally including the feature `season` seems to have no effect on the overall model. We can also observe very high p-values for the estimates of the different `season` categories, and low p-values for all the other features and the intercept. Hence, when considering all features, the data does not really provide evidence for a significant effect of the season on the number of bike rentals, and one should evaluate whether the feature `season` should be included in the model at all.

This is very different conclusion compared to part c) maybe has to do with the fact that the season is correlated with the temperature and the other weather-related features, so replacing the feature `season` with these other features is almost already enough for the model, and the remaining effect of the season alone is very small.

The interpretation of the other features can also be done in terms of the odds ratio, analogously to part c), or in terms of the marginal feature effect or feature effect change, as in part d), although these interpretations actually don't change at all compared to d), since the corresponding coefficient estimates also did not change at all.

Solution 4:

TO DO: Add pseudocode of naive algo and of solution here

The implementations of the solution in R or in python can be found on Moodle in the files "*CART_sol.R*" and "*CART_sol.py*". As described in the exercise, the main idea is to not compute the sums for the two potential child nodes in every step. In the following, we present two ways to work around this.

The first one is to compute the cumulative sums for the target vector, that is, compute a vector which at position k contains the entry $\sum_{i=1}^{k} y_i$ (or y_i^2 respectively), and then for each split point find the corresponding index of the feature of interest, and look up the respective sums in the precomputed vector. This solution is implemented as "Solution 1" in the code.

The other, more direct way of implementation is to maintain two sums, one for the potential left child of the current split point and one for the potential right child. These two sums are then updated in every step, so that the left sum always increases and the right sum always decreases. This method is implemented as "Solution 2" in the code.

Both these solutions in principle achieve the desired result, namely that the whole algorithm of finding the optimal split point has a computational complexity of $O(dn)$ instead of $O(dn^2)$. Nevertheless, the two implementations differ in another point, which may be practically even more relevant. Namely, the "Solution 1" relies on vectorized operations, acting on the whole vector of observations or potential split points in parallel. These vectorized operations in theory require an effort of order $O(n)$ each (linear in the length of the vector), because they perform one operation for every element in the vector, but they are much faster in practice due to parallelization.

The "Solution 2" does not use such operations at all and therefore strictly fulfills the theoretical runtime requirement, whereas "Solution 1" strictly speaking does not. Nevertheless, at least for the implementation in R, "Solution 1" actually runs faster in practice.

The solution also contains the calculation of a deeper tree reflecting the structure of the DGP used.

NOTE: This exercise also serves a preparation for the FAST algorithm introduced in chapter 3.