Solution Quiz:

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

(a) Interpretation methods are *only* used ot explain the global behavior of a model.

⇒ **Wrong**, there are several needs for interpretability. (Gain global and local insights how the IML model works, better control, improve and debug the IML model, justify decisions)

(b) If a model-agnostic and a model-specific interpretation method are applied on the same ML model, the output of the two methods will always be the same.

⇒ **Wrong**, as the methods work different they will probably give a divergent output.

(c) While feature effects methods show the influence of a feature on the target, feature importance methods focus on a feature's impact on the model performance.

⇒ **Correct**.

(d) In IML we distinguish between global IML methods, which explain the behavior of the model over the entire feature space, and local IML methods, which only explain the prediction of individual observations.

⇒ **Correct**.

(e) Technically, Pearson correlation is a measure of *linear* statistical dependence.

⇒ **Correct**.

(f) All in the lecture mentioned measures for correlation and dependencies are limited to continuous random variables.

⇒ **Wrong**, mutual information is not limited to continuous random variables.

(g) A feature interaction between two features x_j and x_k is apparent if a change in x_j influences the impact of x_k on the target.

⇒ **Correct**.

Solution 1:

a) Calculation of Pearson correlation coefficient of x_1 and x_2

$$\rho(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)(x_2^{(i)} - \overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_2^{(i)} - \overline{x}_2)^2}}$$

given the dataset

	1	2	3	4	5	6	7	8	9	$\sum_{i=1}^{n}$
y	-7.79	-5.37	-4.08	-1.97	0.02	2.05	1.93	2.16	2.13	-10.92
x_1	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	0
x_2	0.95	0.57	0.29	-0.03	0.02	0.08	0.23	0.54	0.98	3.63

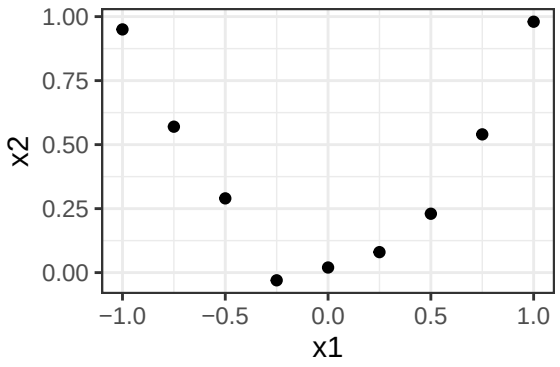
The individual differences to the means are

	1	2	3	4	5	6	7	8	9
$x_1^{(i)} - \overline{x}_1$	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
$x_2^{(i)} - \overline{x}_2$	0.55	0.17	-0.11	-0.43	-0.38	-0.32	-0.17	0.14	0.58

$$\rho(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)(x_2^{(i)} - \overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_2^{(i)} - \overline{x}_2)^2}}$$
$$= \frac{-0.574 + -0.125 + 0.057 + 0.108 + 0 + -0.081 + -0.087 + 0.103 + 0.577}{2.086} = \frac{0.05}{2.086} = 0.002$$

The Pearson correlation coefficient is close to 0 $\Rightarrow$ there is **no linear** relationship between x_1 and x_2.

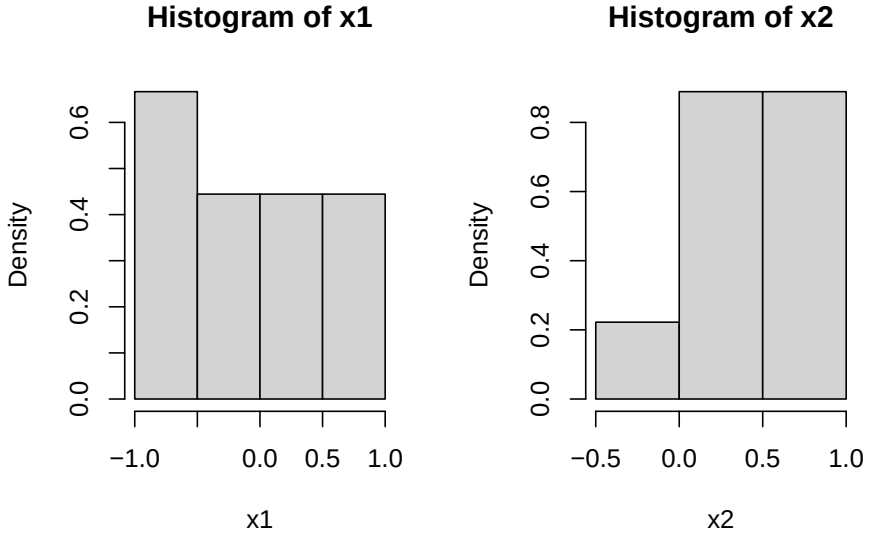
b) The scatter plot reveals that there is a strong non-linear/quadratic relationship between x_1 and x_2. The Pearson correlation coefficients is not suitable for detecting non-linear relationships.



$\Rightarrow$ More suitable: **Mutual Information (MI)**

$$MI(x_1; x_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right)\right] = \sum_{x_1}\sum_{x_2} p(x_1, x_2) log\left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right)$$

Problem: distribution needed.
Solution: e.g. histograms with Gaussian kernel:



Now taking the mean values as replacement for the values in x_1 and x_2:

	1	2	3	4	5	6	7	8	9
$x_1^\star$	-0.75	-0.75	-0.75	-0.25	-0.25	0.25	0.25	0.75	0.75
$x_2^\star$	0.75	0.75	0.25	-0.25	0.25	0.25	0.25	0.75	0.75

Table with joint and marginal distribution:

$x_1^\star$ / $x_2^\star$	-0.25	0.25	0.75	p_{x_1}
-0.75	0.00	0.11	0.22	0.33
-0.25	0.11	0.11	0.00	0.22
0.25	0.00	0.22	0.00	0.22
0.75	0.00	0.00	0.22	0.22
p_{x_2}	0.11	0.44	0.44	1.00

Now we can calculate the approximate MI:

$$\begin{aligned}
MI(x_1^\star; x_2^\star) =& \sum_{x_1^\star} \sum_{x_2^\star} p(x_1^\star, x_2^\star)\, log\left(\frac{p(x_1^\star, x_2^\star)}{p(x_1^\star)p(x_2^\star)}\right) \\
=& 0\, log\left(\frac{0}{0.33 \cdot 0.11}\right) + 0.11\, log\left(\frac{0.11}{0.33 \cdot 0.44}\right) + 0.22\, log\left(\frac{0.22}{0.33 \cdot 0.44}\right) \\
&+ 0.11\, log\left(\frac{0.11}{0.22 \cdot 0.11}\right) + 0.11\, log\left(\frac{0.11}{0.22 \cdot 0.44}\right) + 0\, log\left(\frac{0}{0.22 \cdot 0.44}\right) \\
&+ 0\, log\left(\frac{0}{0.22 \cdot 0.11}\right) + 0.22\, log\left(\frac{0.22}{0.22 \cdot 0.44}\right) + 0\, log\left(\frac{0}{0.22 \cdot 0.44}\right) \\
&+ 0\, log\left(\frac{0}{0.22 \cdot 0.11}\right) + 0\, log\left(\frac{0}{0.22 \cdot 0.44}\right) + 0.22\, log\left(\frac{0.22}{0.22 \cdot 0.44}\right) \\
=& 0.603
\end{aligned}$$

$\Rightarrow$ MI shows that there is a dependency.

Solution 2:

First, recall that the formula for the coefficient of determination R^2 is:

$$R^2 = 1 - \frac{SSE_{LM}}{SSE_c}$$

where $SSE_{LM} = \sum_{i=1}^{n}(y^{(i)} - \hat{f}_{LM}(x^{(i)}))^2$ is the sum of squares due to regression and $SSE_c = \sum_{i=1}^{n}(y^{(i)} - \bar{y})^2$ is the total sum of squares. The formula for the Pearson correlation coefficient ρ is:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x^{(i)} - \bar{x}) \cdot (y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}}.$$

We have:

$$\begin{aligned}
R^2 &= 1 - \frac{SSE_{LM}}{SSE_c} \\
&= 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{f}_{LM}(x^{(i)}))^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}
\end{aligned}$$

Similarly, we can write:

$$\begin{aligned}
\rho^2 &= \left(\frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}\right)^2 \\
&= \left(\frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^n (y^{(i)} - \bar{y})^2}}\right)^2 \\
&= \frac{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})\right)^2}{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})^2\right)\left(\sum_{i=1}^n (y^{(i)} - \bar{y})^2\right)} \\
&= \frac{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})\right)^2}{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})^2\right)\left(\sum_{i=1}^n (y^{(i)} - \bar{y})^2\right)} \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2} \\
&= \left(\frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}\right)^2 \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2}
\end{aligned}$$

Now, note that

$$\sum_{i=1}^n (y^{(i)} - \bar{y})^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2.$$

Proof:

$$\begin{aligned}
\sum_{i=1}^n (y^{(i)} - \bar{y})^2 &= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)} + \hat{y}^{(i)} - \bar{y})^2 \\
&= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + (\hat{y}^{(i)} - \bar{y})^2 + 2(y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y}) \\
&= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 + 2\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y})
\end{aligned}$$

It remains to show that

$$\begin{aligned}
2\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y}) &= 0 \\
\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})\hat{y}^{(i)} - \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})\bar{y} &= 0 \\
\bar{y}\sum_{i=1}^n y^{(i)} - \hat{y}^{(i)} &= 0 \\
\sum_{i=1}^n y^{(i)} - \hat{y}^{(i)} &= 0
\end{aligned}$$

where we have used the fact that $\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})\hat{y}^{(i)} = 0$ as the residuals $(y^{(i)} - \hat{y}^{(i)})$ and $\hat{y}^{(i)}$ are not correlated. Substituting these results into the expression for R^2, we obtain:

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2 - \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2 - \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x^{(i)} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x^{(i)} - \bar{x})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \\
&= \rho^2
\end{aligned}
$$

Hence, we have shown that $R^2 = \rho^2$, which completes the proof. Note that this result is valid only for simple linear regression, where there is only one independent variable. For multiple regression, the coefficient of determination is defined differently and does not necessarily equal the square of the Pearson correlation coefficient.

Solution 3:

Problem: The function $f(\mathbf{x}) = 2x_1 + 3x_2 - x_1|x_2|$ is not differentiable for $x_2 = 0$. Hence, different cases need to be considered:

$$
\begin{gathered}
\text{Case 1: } x_2 > 0 \\
\text{Case 2: } x_2 < 0 \\
\text{Case 3: } x_2 = 0
\end{gathered}
$$

Case 1: $x_2 > 0$

$$
\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ (2x_1 + 3x_2 - x_1 x_2)\right)^2 = \left(\frac{\partial}{\partial x_2}\ (2 - x_2)\right)^2 = (-1)^2 = 1 > 0
$$

Case 2: $x_2 < 0$

$$
\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ (2x_1 + 3x_2 - x_1(-x_2))\right)^2 = \left(\frac{\partial}{\partial x_2}\ (2 + x_2)\right)^2 = 1^2 = 1 > 0
$$

Case 3: $x_2 = 0$
Not considered, as analysis of interactions via definition requires the consideration of intervals. The examination of single points does not make sense.

$\Rightarrow$ x_1 and x_2 interact with each other.