• If predictor has two level => dummy variable => "one hot-encoding".

Extension of linear model

we had $Y = B_0 + B_1 X_1 + B_2 X_2 + \epsilon$ → include ~~predictor~~ interaction term

→ $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_1 X_2 + \epsilon$

$= B_0 + (B_1 + B_3 X_2) X_1 + B_2 X_2 + \epsilon$

$= B_0 + \tilde{B}_1 X_1 + B_2 X_2 + \epsilon$

→ the association between X_1 and Y is no longer constant.

→ A change in X_2 will change the association between X_1 and Y

→ " " " X_1 " " " " X_2 " Y

→ we can interpret B_3 as the ~~increase~~ change of impact of X_2 ~~with~~ associated with one unit increase of X_1 (and vice versa).

The hierarchical principle states that: if we include an interaction in a model, we should also include the main effects, even if the p-value associated with their coefficient are not significant

→ to accomodate non-linear relationship we can use polynomial regression. (ch 7).

Potential problems

① Non-linearity of the response-predictor relationship: plot residuals versus the predicted (or fitted) values $\hat{y}_i$ → if discernible pattern → non-linearity in the data → solution is to use non linear transformations such as $\log X$, $\sqrt{X}$, X^2

② correlation of Error term: if correlation → ϵ_i inform on ϵ_{i+1}: tracking. estimated standard error lower than the true one, p-value lower than they should be → look at plot residual time.

③ Non-constant variance of Error term: we suppose that is constant $Var(\epsilon_i) = \sigma^2$. we can plot non-constant (heteroscedaticity) by a funnel shape in the residual plot. solution => transform Y using a concave function such as $\log Y$ or $\sqrt{Y}$

Sometimes, we have a good idea of the variance of each response. For example, the ith response could be an average of n_i raw observations. if each of the is uncorrelated with variance $\sigma_i^2 = \sigma^2/n$, then we can use weighted least squares, with weights proportional to the inverse variances.

④ Outliers: residuals plot can be used to identify outliers. But we can also plot the studentized residuals, computed by dividing each residual e_i by its estimated standard error (if > 3 => outlier)

⑤ High leverage points: it is cause for concern if the least squares line is heavily affected by just a couple of observations. we compute the leverage statistics $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}$

⑥ Collinearity: it refers to the situation in which two or more predictor variables are closely related to one another => can be difficult to separate out the individual effects of collinear variables on the response.

• Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the SE for $\hat{B}_j$ to grow (because t-statistic is $\hat{B}_j / SE(\hat{B}_j)$ and t-statistic ↓) and we may fail to reject $H_0: B_j = 0$) => the probability of ~~rejecting~~ detecting a non-zero coefficient is reduced by collinearity).

• We can use correlation matrix of predictors => but won't show multicollinearity

This is why better to use the variance inflation factor: the ratio of the variance of $\hat{B}_j$ when fitting the full model divided by the variance of $\hat{B}_j$ if fit on its own. If VIF close to 1: non-collinearity (should be < 5 or 10)

Solution: suppress one of them or merge them.

Comparaison of linear regression with KNN

- The parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f.
- if the relation is linear: it is hard for NP to compete with LR: a NP incurs a cost in variance that is not offset by a reduction in bias.
- Spreading 50 observations over $p=20$ dimension => no nearby neighbors => curse of dimensionality. That is the k observations that are nearest to a given test observation x_0 may be very far away from x_0 in p-dimensional space when p is large, leading to poor prediction of $f(x_0)$
- if the MSE of KNN is only slightly lower to LN, we prefer LN for interpretability.

Marketing plan

- Is there relationship between Y and X -> fit Y on X -> F-statistic -> if p-value of F small, then at least one impact.
- How strong the relationship -> look at RSE and R^2
- which x_i in X are associated with Y => p-value associated with each predictor t-statistic
- How large is the association between x_i and Y: look at interval confidence $\beta_j \pm 2\,SE(\beta_j)$
if large interval check for a large collinearity (if VIF close to 1, then no problem)
- How accurately can we predict f. -> if we want $Y = f(x)+\epsilon$ we use prediction interval
-> if we want Y in average -> we use confidence interval
prediction > than confidence because has ϵ
- is the relationship linear?
-> plot residual -> if no pattern yes
- Is there synergy among x_i => if small p-value of interaction term indicate yes (then the model is non additive as it needs interactions term). The R^2 should ↗.