Chapter 4 Interpretable Models

일부 모델의 경우에는 그 자체로 해석이 쉬운 것들이 있다.

- Linear Regression
- Logistic Regression
- Decision Tree

용어 설명

- 선형성은 feature와 target의 값이 선형적인 관계인 것
- 단조성은 feature와 target의 값이 한 방향으로만 일관되게 진행되는 것
- interaction은 복수 개의 feature사이의 관계가 모델링되는 것 (ex. (방 개수, 집 크기) => 집값 예측)

or classification (class):

선형과 단조성은 해석에 유리

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

Linear Regression

정의 및 구성

- target의 값이 feature 값의 가중합으로 예측되는 것
- weight for each feature, bias, error term following Gaussian dist
- weight esitmation coms from LSME
- weight estimation come with confidence interval
  - 95% interval = 95 out of 100 try, the confidience interval include true weight

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

선형 모델이 가정하는 데이터 특징

- Linearity
  - prediction as a linear combination of features
  - each term is additive (easy of seperation of the effect)
- Normality
  - target follows a normal distribution
  - If not, use Generalized Linear Model (GLM)
- Homoscedasticity (constant variable)
  - variance of error be constant over entire feature space
  - ex) 방크기 feature로 집값 예측시, 작은 방크기에서의 예측 오류와 큰 방 크기에서의 예측 오류가 다르면 안됨
- Indepondece
  - each instance is indep

    - ex) 동일한 환자에서 연속 피뽑은 샘플 => not independent
    - If not, use GEE
- Fiexed features
    - feature is not random variable,
    - it is just constant, free of measurement errors
    - but unrealistic assumption
- Absence of multicollinearity
    - It not, use interaction term

선형 모델의 해석

- For numeric feature
    - feature의 값을 1 unit 변화시키면, target 값이 weight만큼 변한다.
- Binary feature
    - reference category로부터 해당 category로 변화시키면 weight만큼 변한다.
    - ex) 날씨 = {Sunny, Cloudy}일 때 sunny가 reference
- Categorical feature
    - L-1 one-hot encoding
    - same as binary feature
- Intercept
    - meaningless case : all numeric features set to zero, all categorical feature to references
    - meaningful case : when all numeric featuees are standadised(mean of zero, deviation of one)
- R-squared
    - how much of the target variance is explained by the model
    - higher R-squared, better model explains the data

$$R^2 = 1 - SSE/SST$$

- Feature Importance
    - absolute value of t-statistic (estimated weight scaled with std)
    - the more variance of estimated weight => we are not sure of correct importance => make less feature importance

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Example (prediction of # of rented bikes)

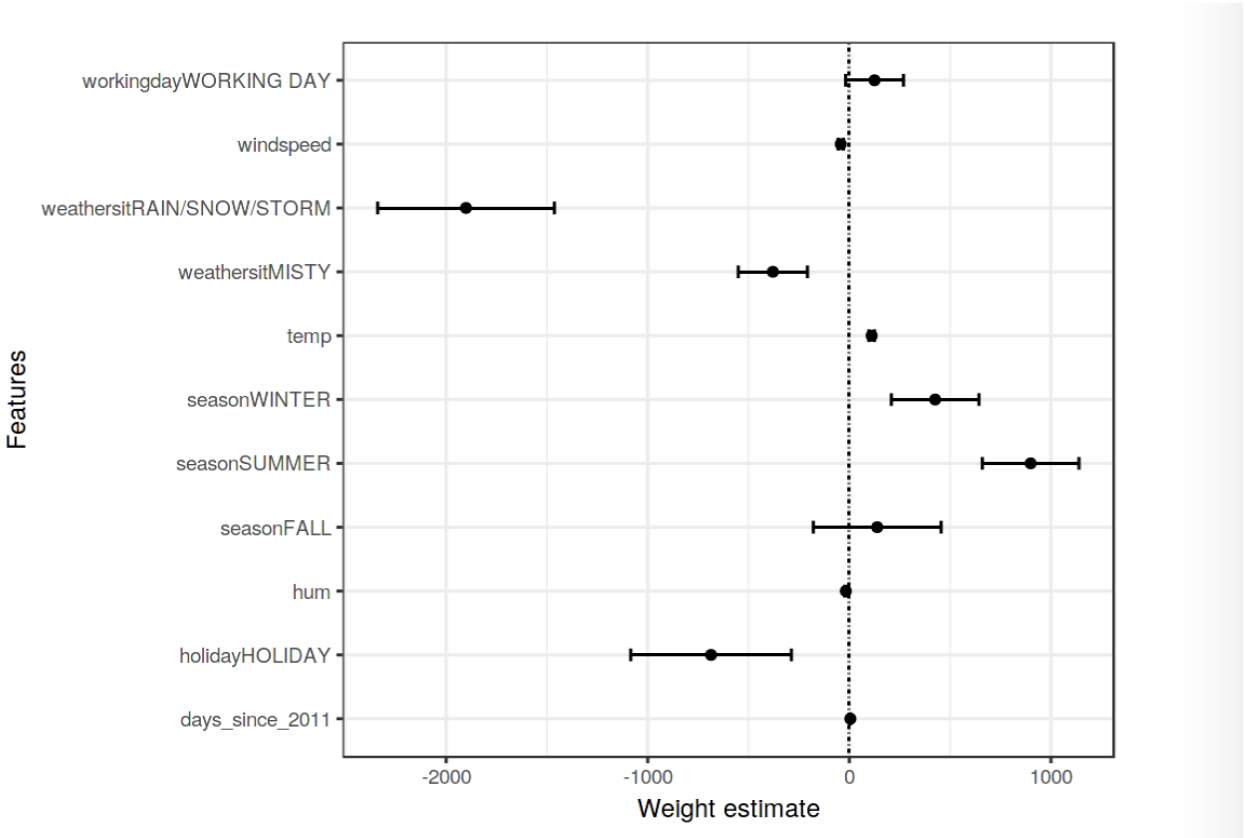
	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

- ex) 1도 올라가면 110개의 자전거가 추가로 랜트된다.
- ex) 좋은 날씨 대비 비가 오면 랜트수가 1901개 줄어든다.
- 모든 해석은 다른 feature가 고정되었을 때를 가정한다.
- 하나의 feature unit만 증가시키는 것을 때로는 unrealistic inout이 되고 만다.
  - ex) increase of # of room without increasing size of room => 현실적으로 비존재할 수도

Visual Interpretation

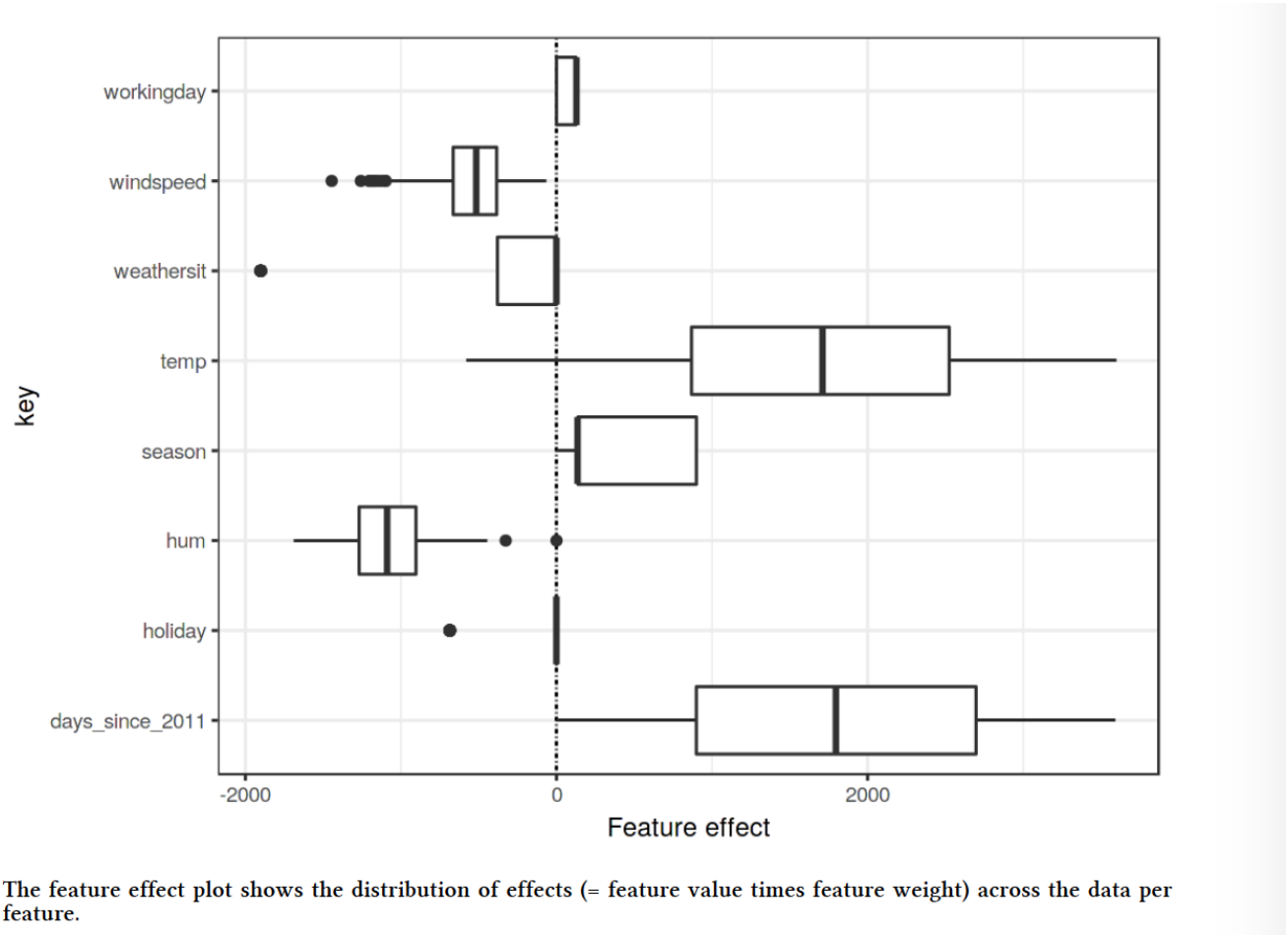
Weight Plot

- median값, interval의 zero 포함, interval의 길이 등을 고려
- 각 feature마다 scale이 다르다는 점이 문제이다.
  - 정확히 비교할려면 모든 feature의 scale을 맞추어야 한다.



Effect Plot

- effect = weight * feature value
- weight은 feature scale에 따라 차이가 나지만, effect는 scale-independent하다.
- 카테고리 feature는 하나의 boxplot으로 summarized되서 표시



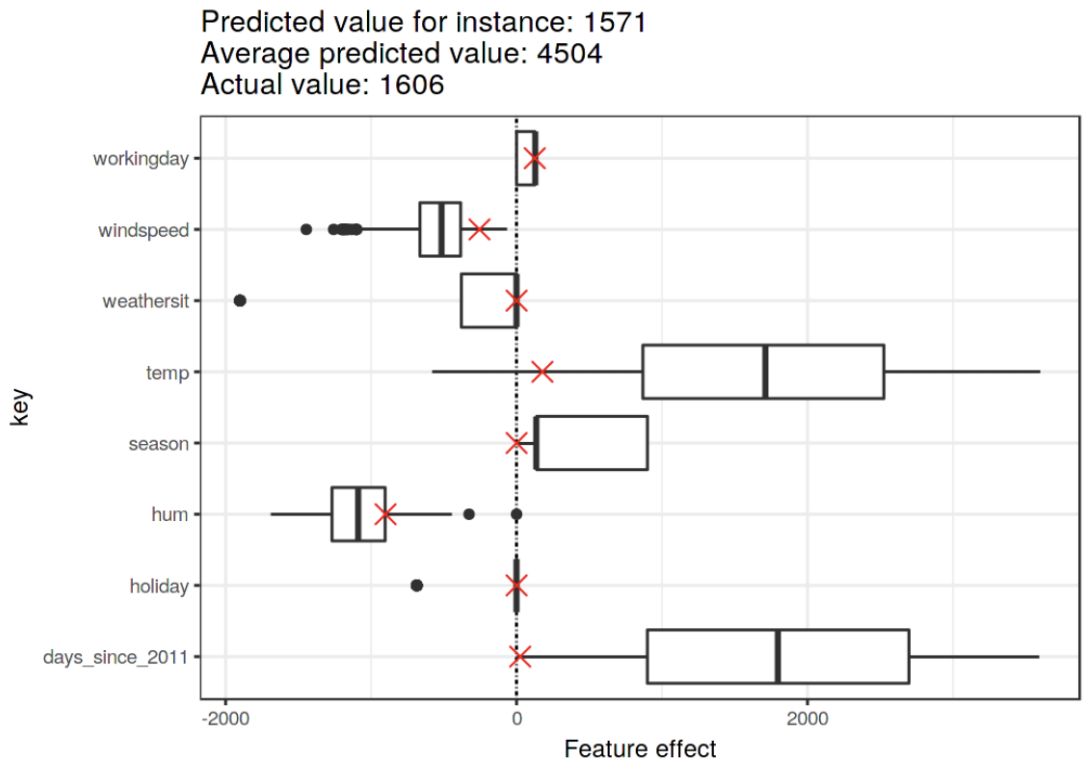
The feature effect plot shows the distribution of effects (= feature value times feature weight) across the data per feature.

Explain Individual Predictions

- 예를 들어서 다음의 feature를 갖는 하나의 data point를 설명해보자

Feature	Value
season	SPRING
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5

- red cross가 각 feature별 예측 기여도를 나타냄
- 대조적인 설명
  - 평균치보다 낮은 주요 원인 - 낮은 온도, 2011년의 초반



The effect plot for one instance shows the effect distribution and highlighting the effects of the instance of interest.

Do Linear Models Create Good Explanations?

No because

- Contrastive한 설명을 할 수는 있지만, reference instance라는 것이 artificial, meaningless하다.
  - 다만 모든 numerical feature가 standarized되고 모든 category feature가 effect coding된다면 reference instance는 일종의 평균 instance가 된다.
  - 여전히 현실적이지 않을 수는 있지만, 상대적으로 more meaningful하다.
- selective한 설명이지 않다.
  - 모든 feature를 다 동원해서 설명하므로,
  - 대안으로 sparse linear model

Sparse Linear Model

seletive한 설명을 위해 sparsity를 linear model 안으로 도입

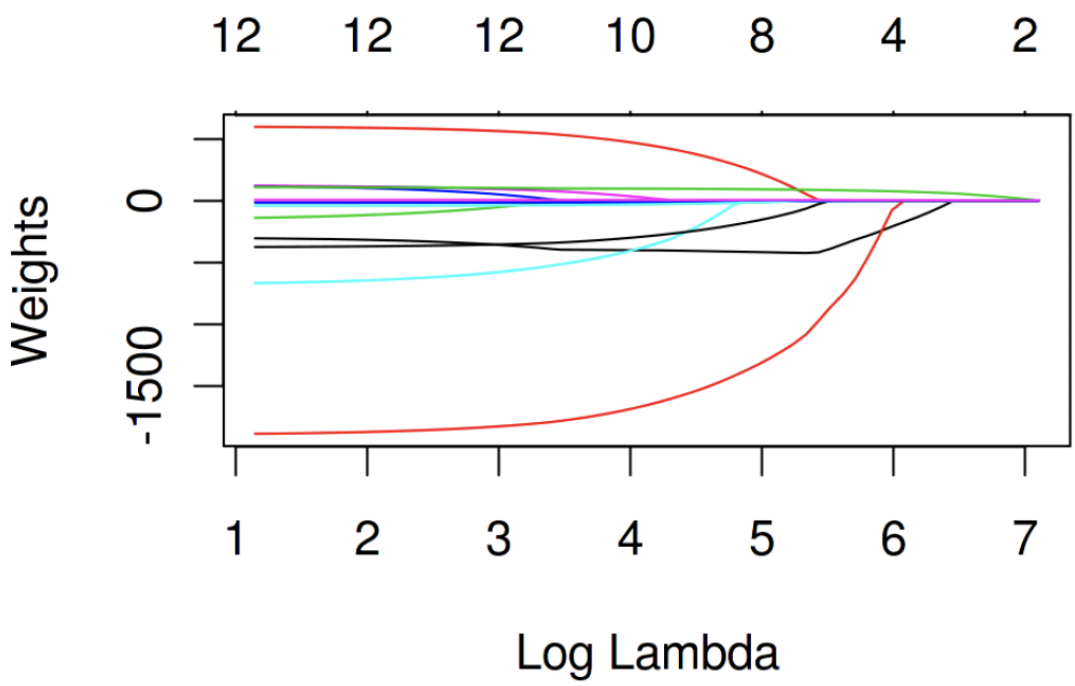
LASSO (least absolute shrinkage selection operator)

최적화시에 큰 weight를 억제시키는 L1-norm 추가

$$min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-x_i^T\beta)^2+\lambda||\beta||_1\right)$$

lambda가 크면 많은 feature weight이 0으로 된다.

- CV를 통해서 적절한 lambda를 정하자



Example with Lasso

강한 lambda를 사용해서 2개의 feature만 살아남게 학습한 결과

	Weight
seasonSPRING	0.00
seasonSUMMER	0.00
seasonFALL	0.00
seasonWINTER	0.00
holidayHOLIDAY	0.00
workingdayWORKING DAY	0.00
weathersitMISTY	0.00
weathersitRAIN/SNOW/STORM	0.00
temp	52.33
hum	0.00
windspeed	0.00
days_since_2011	2.15

Other Methods for Sparsity in Linear Models

Preprocessing을 통한 방법들

- manually selection of features by expert knowledge
- univariate selection like high correlation coefficient

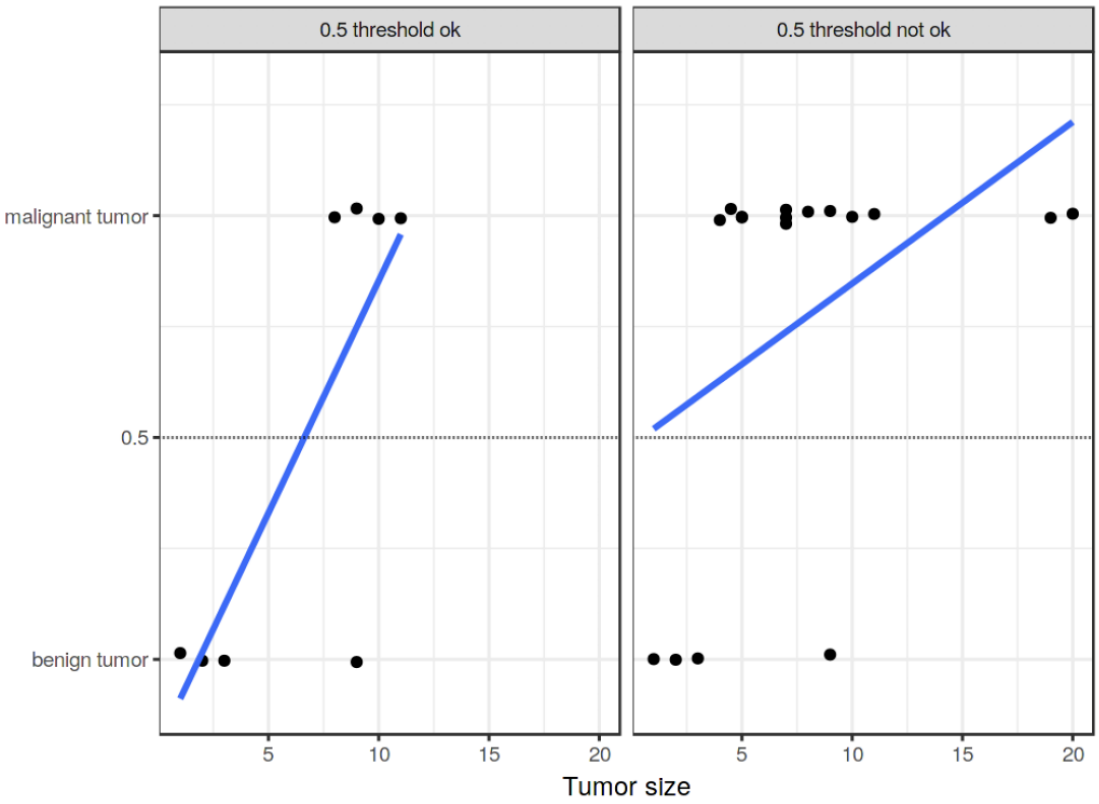
Step-wise method

- forward selection : 소수의 feature로부터 시작해서 하나씩 feature를 추가해가면서
- backward selection : 모든 feature를 다 쓴 것부터 시작해서 하나씩 빼가면서

Logistic Regression

What is Wrong with Linear Regression for Classification?

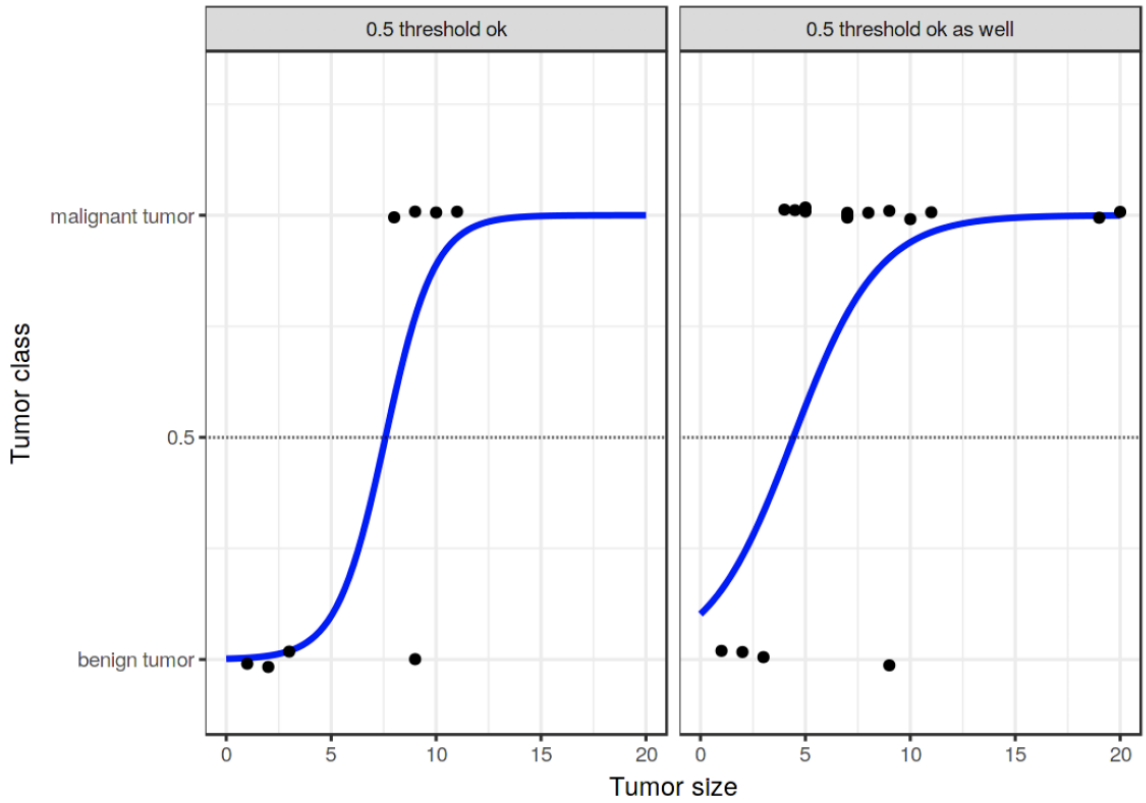
- linear model은 output probability를 출력해주지 못한다.
- predicted class를 일종의 숫자로 취급하여 이들을 구분할 hyperplane을 학습하는 것
  - multiple calss로 확장이 안된다. label = 1,2,3
- linear model은 0보다 작은 값, 1보다 큰 값도 외삽한다.



A linear model classifies tumors as malignant (1) or benign (0) given their size. The lines show the prediction of the linear model. For the data on the left, we can use 0.5 as classification threshold. After introducing a few more malignant tumor cases, the regression line shifts and a threshold of 0.5 no longer separates the classes. Points are slightly jittered to reduce over-plotting.

Theory

- logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.



The logistic regression model finds the correct decision boundary between malignant and benign depending on tumor size. The blue line is the logistic function shifted and squeezed to fit the data.

Interpretation

Linear model for the log odds

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

log odds

$$log\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

A change in a feature by one unit changes the odds ratio (multiplicative) by a factor of exp(beta).

$$\frac{odds_{x_j+1}}{odds} = \frac{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j(x_j+1) + \ldots + \beta_p x_p\right)}{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p\right)}$$

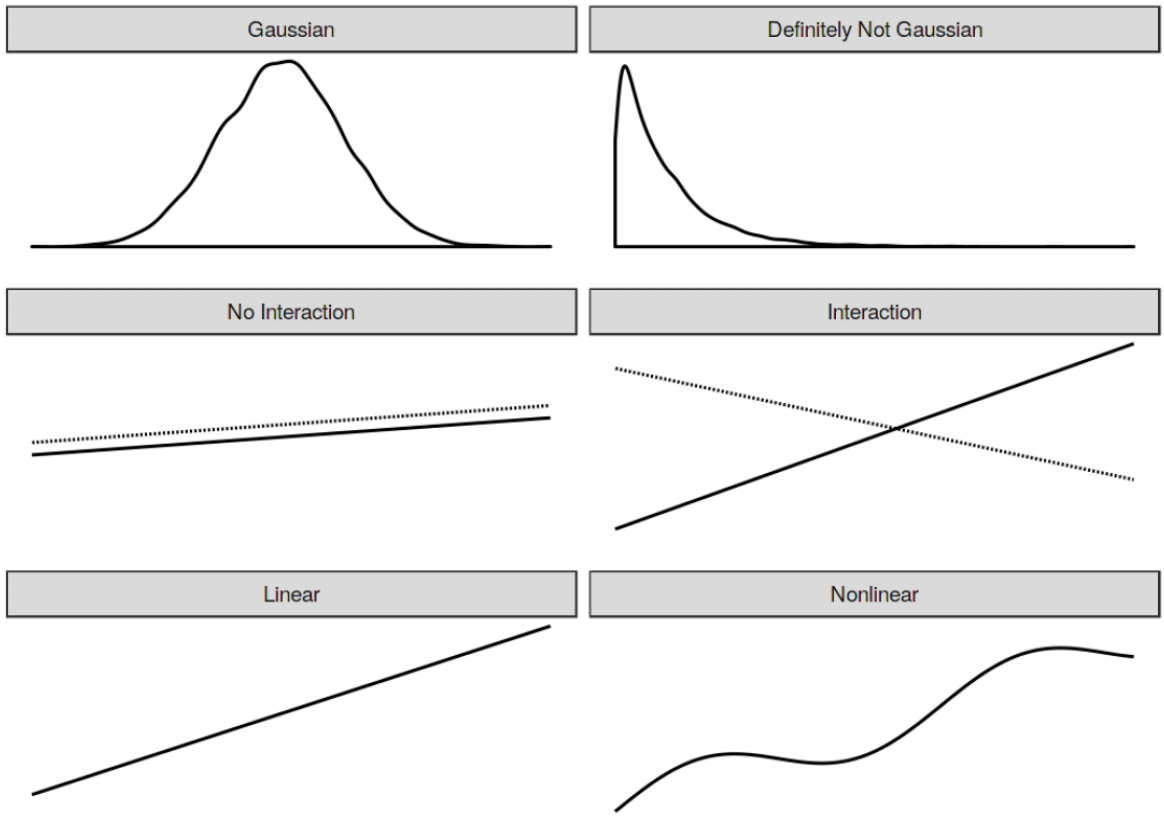
$$\frac{odds_{x_j+1}}{odds} = exp\left(\beta_j(x_j+1) - \beta_j x_j\right) = exp\left(\beta_j\right)$$

GLM, GAM and more

linear model의 가정들이 안맞을 때

- error term does not follow gaussian => GLM
- there are interaction terms => Add interation terms

- relationship is non-linear => GAM



Three assumptions of the linear model (left side): Gaussian distribution of the outcome given the features, additivity (= no interactions) and linear relationship. Reality usually does not adhere to those assumptions (right side): Outcomes might have non-Gaussian distributions, features might interact and the relationship might be nonlinear.

Non-Gaussian Outcomes - GLMs

outcome이 gaussian이 아닌 경우

- a category (cancer v.s. healthy)
- a count
- very skewed
- multi-modal

GLM

- Keep the weighted sum of the features,
- but allow non-Gaussian outcome distributions (from exponential family)
- and connect the expected mean of this distribution and the weighted sum through a possibly nonlinear function. (link function)

$$g(E_Y(y|x)) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$$

- ex) count of somthing
  - use of Poission dist, log as link

$$ln(E_Y(y|x) = x^T\beta$$

- ex) logistic regression
  - use of bernoulli dist, logistic ft as link

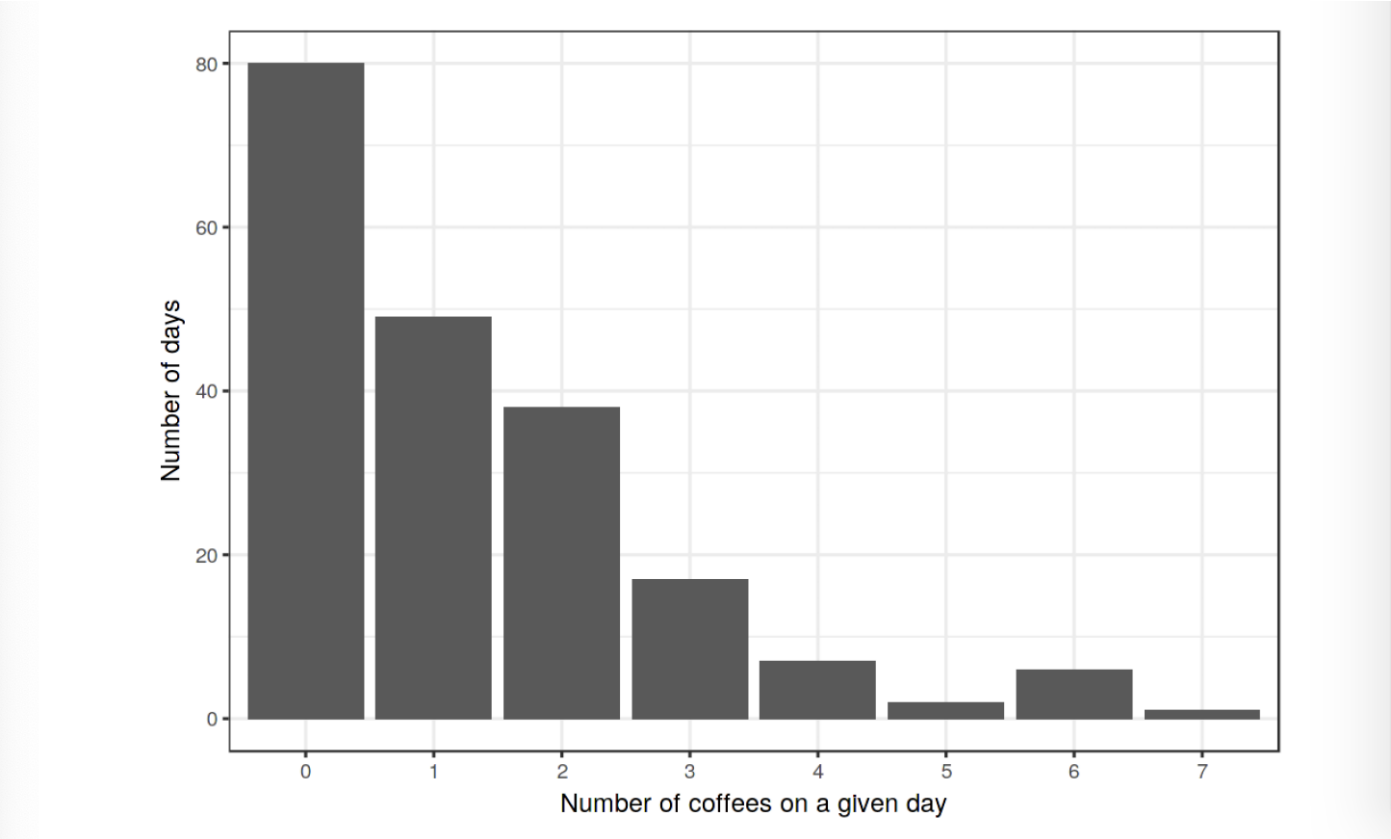
$$x^T\beta = ln\left(\frac{E_Y(y|x)}{1-E_Y(y|x)}\right) = ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)$$

Ex) # of cup of coffee for a day estimation

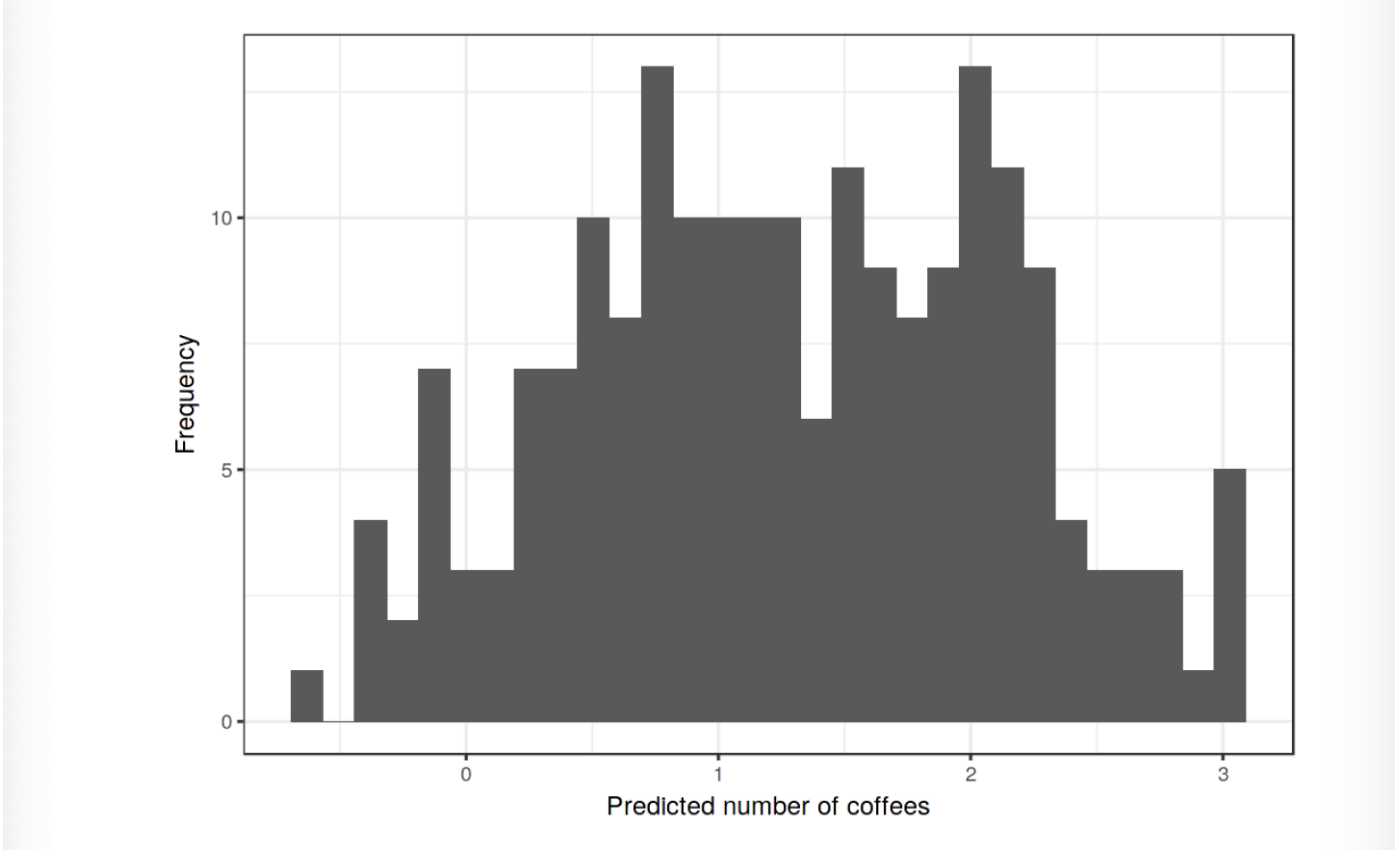
input features

- stress level (1-10)
- how well slept night before (1-10)
- working day or not

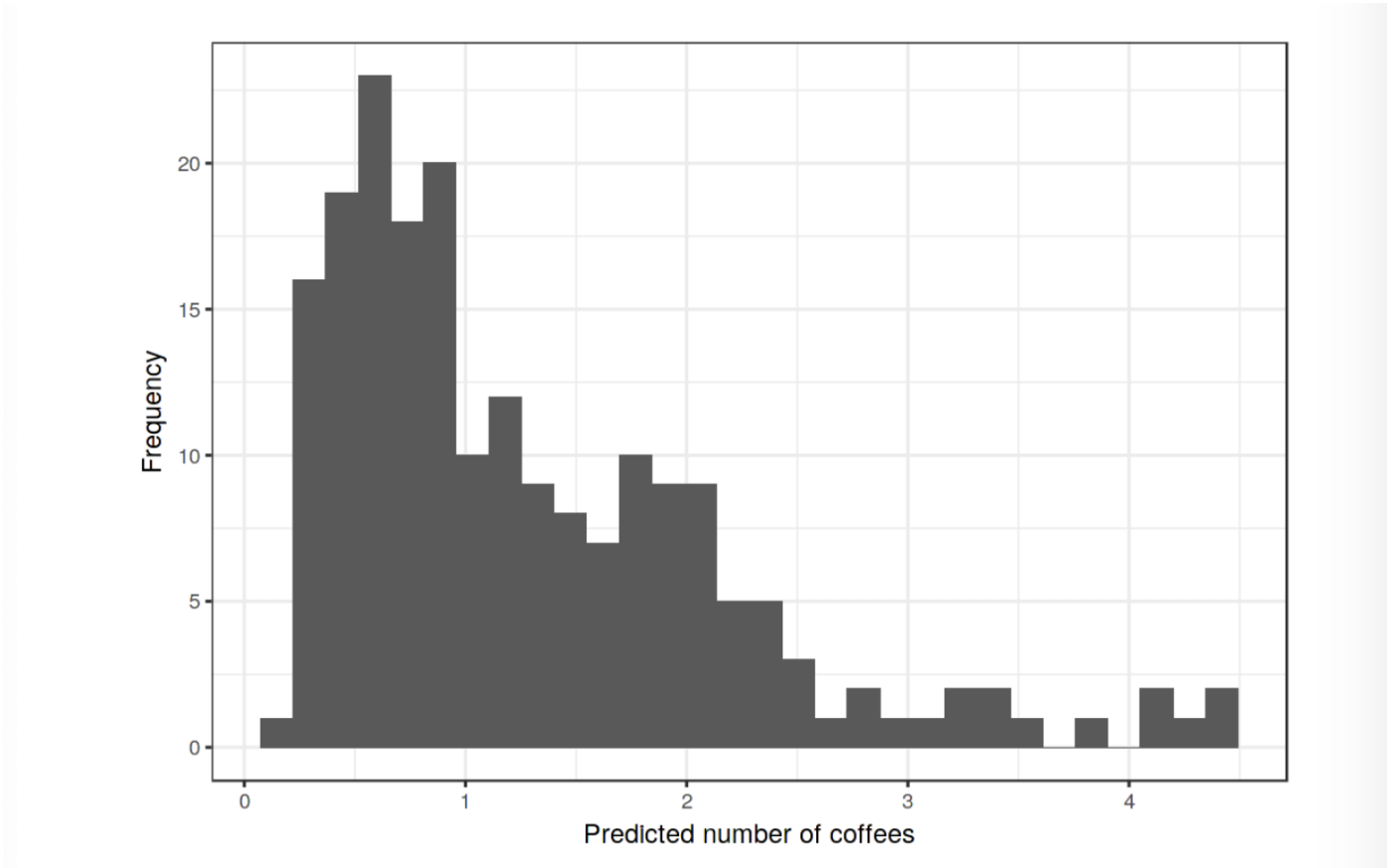
Actual output (true observation) is not gaussian



But, linear model prediction negative cup of coffee



By GLM (use of possion dist, and log link)



Intepretation

effect is multiplicative

$$ln(E(\text{coffees}|\text{stress}, \text{sleep}, \text{work})) = \beta_0 + \beta_{\text{stress}}x_{\text{stress}} + \beta_{\text{sleep}}x_{\text{sleep}} + \beta_{\text{work}}x_{\text{work}}$$

$$E(\text{coffees}|\text{stress}, \text{sleep}, \text{work}) = exp(\beta_0 + \beta_{\text{stress}}x_{\text{stress}} + \beta_{\text{sleep}}x_{\text{sleep}} + \beta_{\text{work}}x_{\text{work}})$$

Increase of stress by one unit => increase by factor of 1.11

number of coffeef on a work day is on 2.42 times number of coffees on a day off

	weight	exp(weight) [2.5%, 97.5%]
(Intercept)	-0.12	0.89 [0.56, 1.38]
stress	0.11	1.11 [1.06, 1.17]
sleep	-0.16	0.85 [0.81, 0.89]
workYES	0.88	2.42 [1.87, 3.16]

Interactions

ex) 방 크기와 방 개수

Add interaction between category and numeric

- rental bicycle 문제에서 휴일여부와 기온

Intercept	workY	temp	workY.temp
1	1	25	25
1	0	12	0
1	0	30	0
1	1	5	5

Add interaction between category and category

work	wthr
Y	Good
N	Bad
N	Ok
Y	Good

Next, we include interaction terms:

Intercept	workY	wthrGood	wthrOk	workY.wthrGood	workY.wthrOk
1	1	1	0	1	0
1	0	0	0	0	0
1	0	0	1	0	0
1	1	1	0	1	0

Add interaction between two numerics => trivial

Interpretation of interation

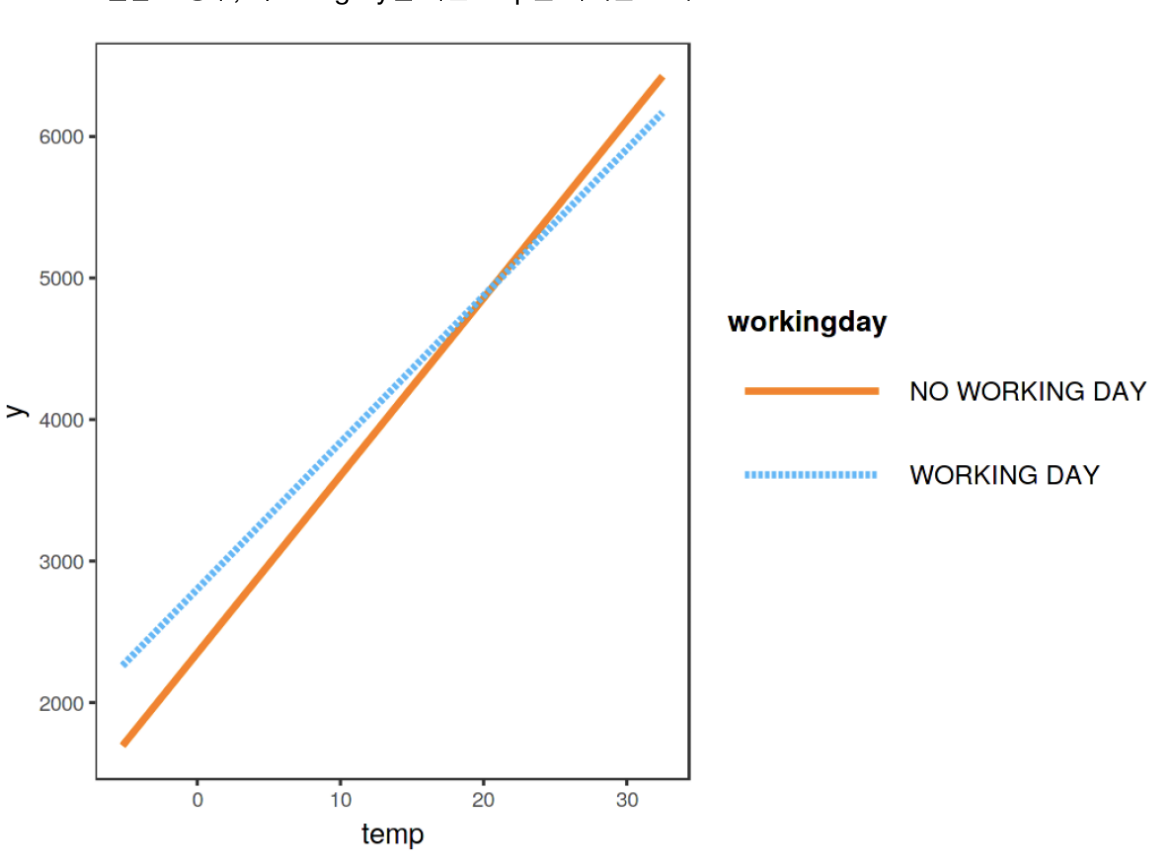
somewhat complicated

- Does the temperature have a negative effect given it is a working? The answer is no,
- temparature 1도 상승시 영향 = temp(NoWork) + workingDay:temp = 125.4 - 21.8

	Weight	Std. Error	2.5%	97.5%
(Intercept)	2185.8	250.2	1694.6	2677.1
seasonSUMMER	893.8	121.8	654.7	1132.9
seasonFALL	137.1	161.0	-179.0	453.2
seasonWINTER	426.5	110.3	209.9	643.2
holidayHOLIDAY	-674.4	202.5	-1071.9	-276.9
workingdayWORKING DAY	451.9	141.7	173.7	730.1
weathersitMISTY	-382.1	87.2	-553.3	-211.0
weathersitRAIN/SNOW/STORM	-1898.2	222.7	-2335.4	-1461.0
temp NoWork	125.4	8.9	108.0	142.9
hum	-17.5	3.2	-23.7	-11.3
windspeed	-42.1	6.9	-55.5	-28.6
days_since_2011	4.9	0.2	4.6	5.3
workingdayWORKING DAY:temp	-21.8	8.1	-37.7	-5.9

Interaction visualization

category + numeric 결합인 경우, 각 category별 다른 slop를 가지는 그래프



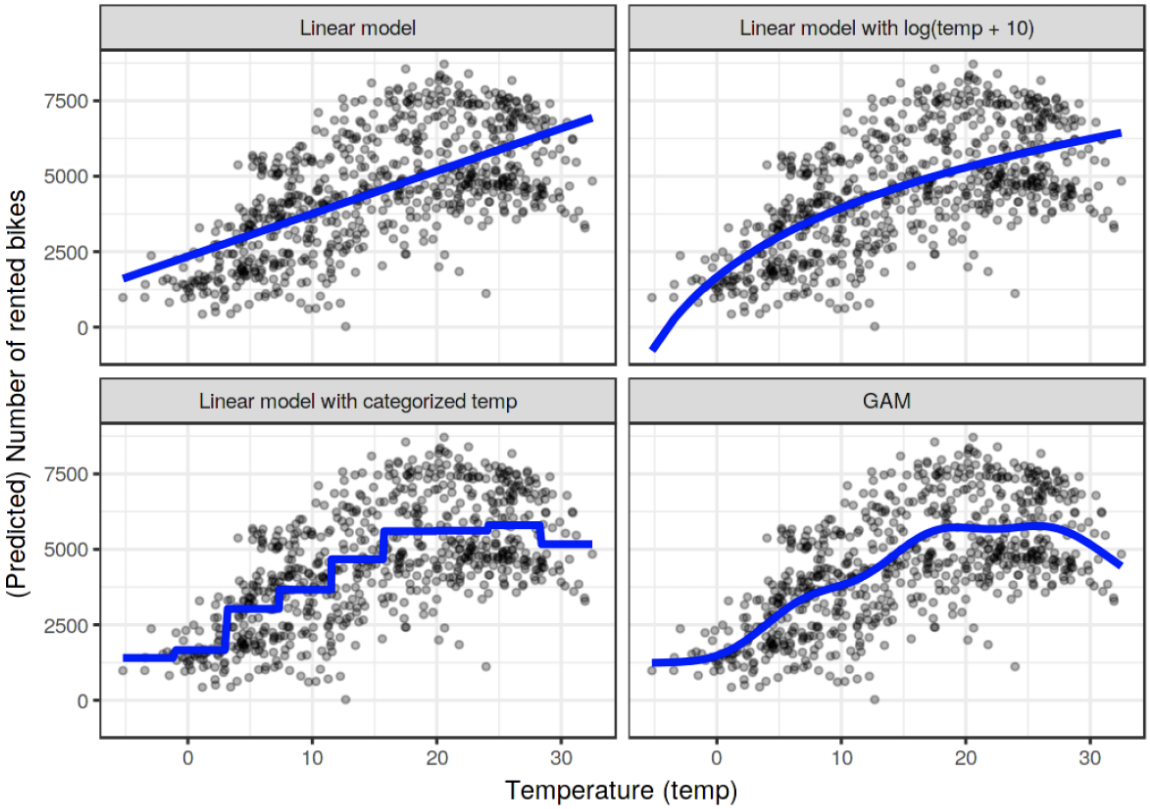
The effect (including interaction) of temperature and working day on the predicted number of bikes for a linear model. Effectively, we get two slopes for the temperature, one for each category of the working day feature.

Non-linear effects - GAM

x,y사이가 선형 관계가 아닐 때

- simple transform like logarithm
- categorization of feature
- generalized additive model (GAM)

with temperature treated as categorical feature and using regression splines (GAM).



Predicting the number of rented bicycles using only the temperature feature. A linear model (top left) does not fit the data well. One solution is to transform the feature with e.g. the logarithm (top right), categorize it (bottom left), which is usually a bad decision or use Generalized Additive Models that can automatically fit a smooth curve for temperature (bottom right).

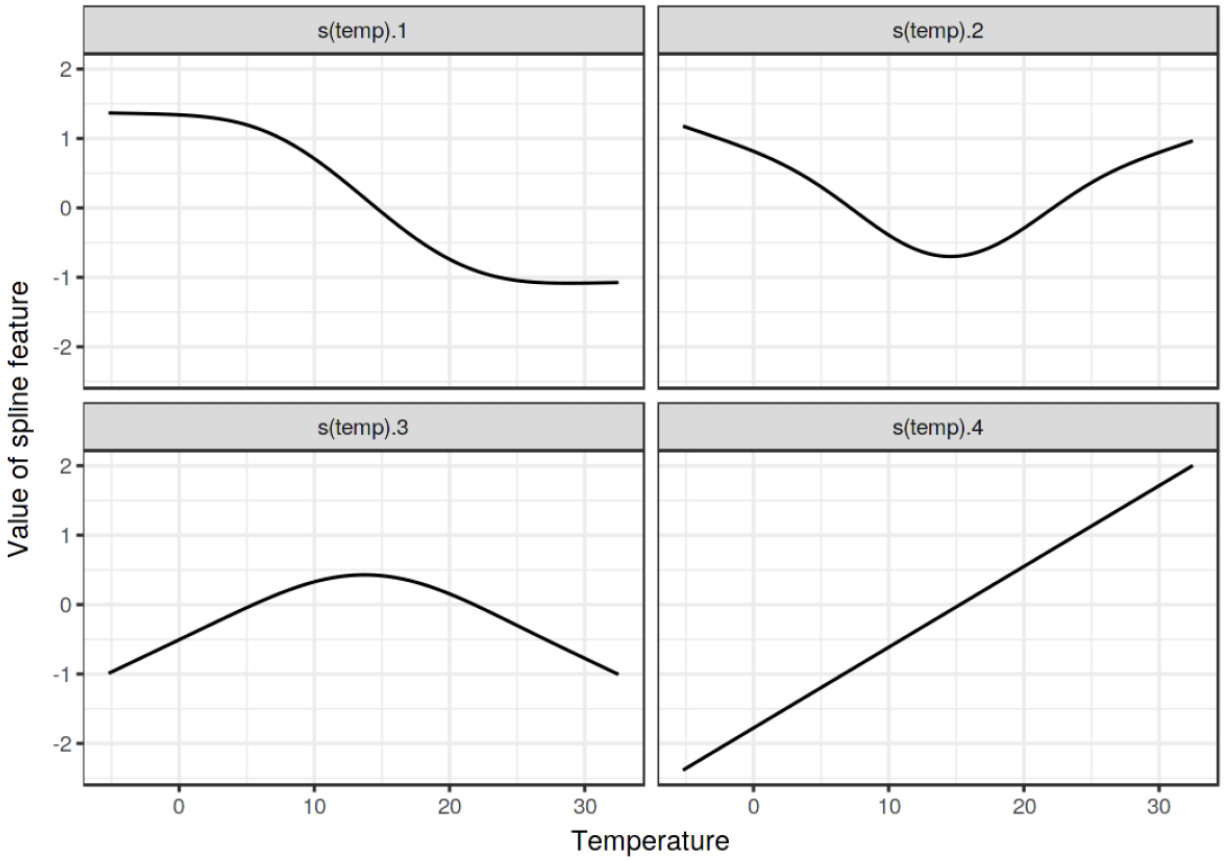
Generalized Additive Models (GAMs)

GLM과 다른 점은 각 feature가 output과 nonlinear ft f로 매핑

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

어떻게 non-linear function을 배울 것인가? => Spline

- ex) temperature와 rented bicycle간의 비선형 관계를 4개의 spline combination으로 fit



To smoothly model the temperature effect, we use 4 spline functions. Each temperature value is mapped to (here) 4 spline values. If an instance has a temperature of 30 C, the value for the first spline feature is -1, for the second 0.7, for the third -0.8 and for the 4th 1.7.

Interpretation of GAM

no longer easy interpretable

In [ ]: