STA130F24 WEEK 10

Multiple Linear Regression Review
Decision Tree

How to Assess your Linear Regression Model?

Evidence-based

- Using p-values, Hypothesis Test, and Confidence Interval on the estimated coefficient.
- In Multiple Linear Regression, this can help us to build our model by adding or removing predictor variables based on the evidence present in the data (Hypothesis Test)

Performance-based

- More "machine learning" way to determine the "best" model using **train-test** framework.
- "Good" model should be able to predict the test data as good as it is able to predict the train data.
- Often measured by metrics like R-squared or Confusion Matrix

Hypothesis Test and CI on Coefficients

- Assess the evidence of a linear association in the data based on a **null hypothesis** that the **slope** (the "on average" change in Yi per "single unit" change in xi is zero

$$H_0 : \beta_1 = 0 \quad \text{(there is no linear assocation between } Y_i \text{ and } x_i \text{ "on average")}$$
$$H_A : H_0 \text{ is false}$$

p-value **95% CI**

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5361	0.155	3.465	0.003	0.211	0.861
Q("Bird Flu Cases")	0.0023	0.000	21.480	0.000	0.002	0.003

How about other numbers?

Dep. Variable:	Q("Shuttlecock Price")	**R-squared:**	0.962
Model:	OLS	**Adj. R-squared:**	0.960
Method:	Least Squares	**F-statistic:**	461.4
Date:	Thu, 24 Oct 2024	**Prob (F-statistic):**	2.80e-14
Time:	16:22:52	**Log-Likelihood:**	15.352
No. Observations:	20	**AIC:**	-26.70
Df Residuals:	18	**BIC:**	-24.71
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.5361	0.155	3.465	0.003	0.211	0.861
Q("Bird Flu Cases")	0.0023	0.000	21.480	0.000	0.002	0.003

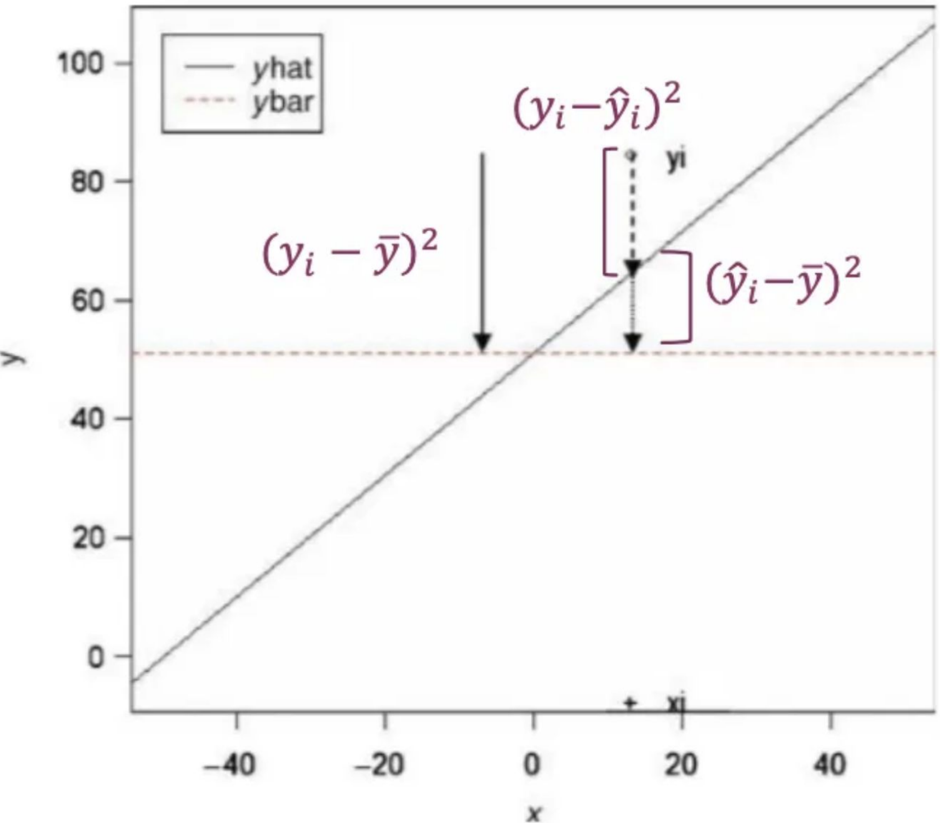
Coefficient of Determination

- Proportion of variation in the response that has been explained by the model
- 0 <= R-squared <= 1

P-value

- Identify the existence of a linear relationship (multiple linear regression)
- Testing “all slopes are zero” vs “at least one slope is not zero”

The idea behind R-squared



Total amount of variation prior to fitting the model

$$SST = \sum (y_i - \bar{y})^2,$$

Unexplained variation from fitting the model

$$RSS = \sum (y_i - \hat{y}_i)^2,$$

Proportion of variation that is explained by the model

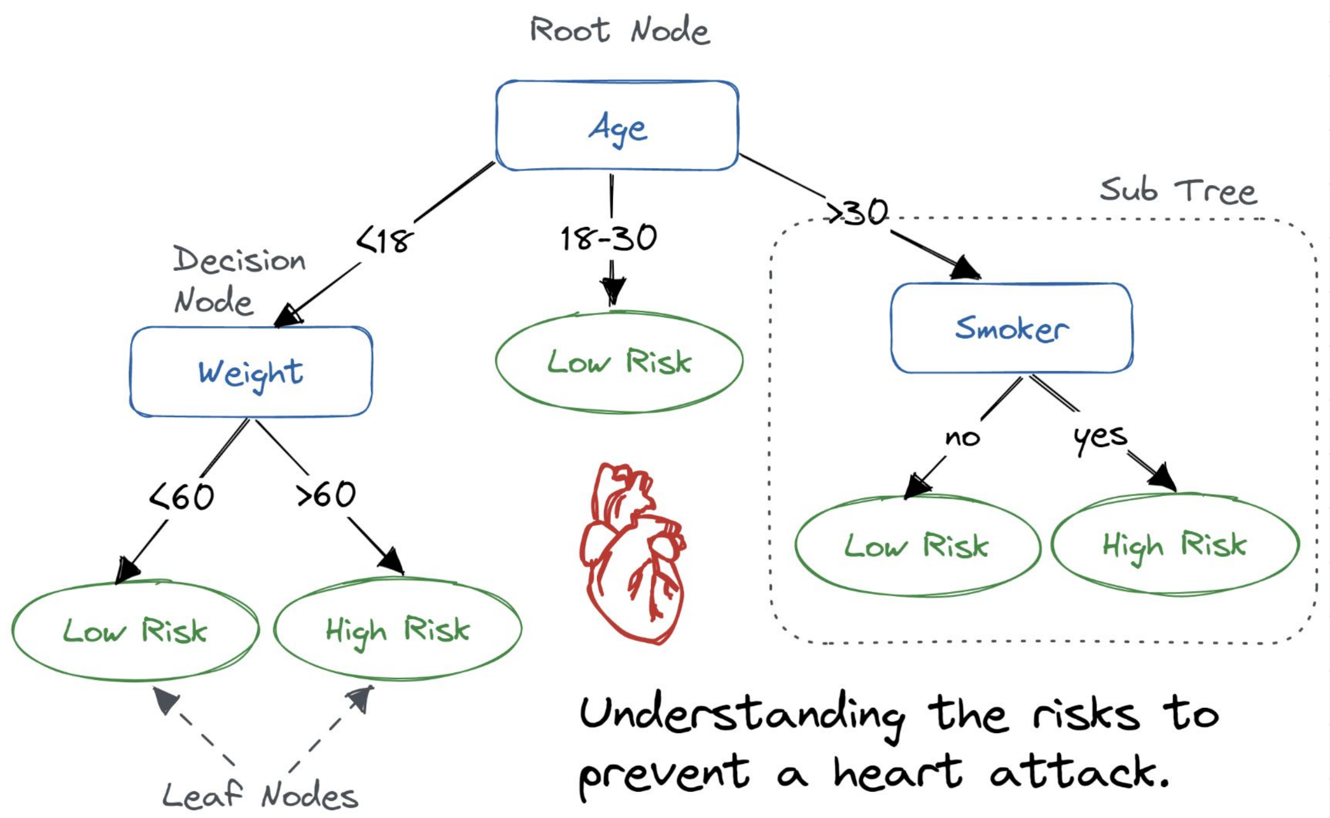
$$R^2 = 1 - \frac{RSS}{SST}$$

Can R-squared get worse with additional predictors?

- Bigger value of R-squared means your model can capture more variation from your data.
- Adding additional predictors to the model, will at least maintain or increase that explanatory power, and thus R-squared can't get worse.
- **Adjusted R-squared:** take into consideration the number of predictors, and will penalize the addition of predictors (since it adds complexity to our model)

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

Decision Tree predict the value of an outcome based on the sequential application of rules based on predictor variables.



Classification vs Regression

- Prediction of numeric outcomes is referred to as **regression (Simple and Multiple Linear Regression**.
- Prediction of categorical outcomes (binary or multi-class classification) is referred to as **classification (Decision Tree).**
- **Note:** Logistic Regression is a classification methodology.

Confusion Matrix

	Predicted "Negative"	**Predicted "Positive"**
Actually "Negative"	**True** *Negative* (TN)	**False** *Positive* (FP)
Actually "Positive"	**False** *Negative* (FN)	**True** *Positive* (TP)