Chapter 3: Linear Regression

Name	Content
TYPE	cheat sheet
BOOK	An Introduction to Statistical Learning
AUTHORS	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
PUBLISHER	Springer

Linear Regression is a type of supervised learning

- Useful for predicting a quantitative response
- Used as a basis for more advanced methods.

Some important questions to ask:

1. Is there a relationship between X and Y?
2. How strong is the relationship between X and Y?
3. Which predictor (X) contributes to the response, Y?
4. How accurately can we estimate the effect of each predictor X on response, Y?
5. How accurately can we predict future response Y?
6. Is the relationship between X and Y linear?
7. Is there an interaction effect between the various predictors X?

3.1 & 3.2 Simple & Multiple Linear Regression

- Predicting a quantitative response based on a single predictor.
- Assumes there is a *linear* relationship between the predictor and response.
- Basic form: $Y \approx \beta_0 + \beta_1 X + \epsilon$
  - β_0 (intercept) and β_1 (slope) are unknown constants representing the model's coefficients and parameters.
  - ϵ is the error term
  - Training data is used to find estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Goal of linear regression is to get the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ as close to the data points we have.
- Most common method: *least squares* which tries to minimize the RSS
- *Unbiased* a model that does not *systematically* over or under estimate the true parameters.

Name	What it does	Looking for	Based on
RSS = Residual Sum of Squares	A measure of the amount of variance **not** explained by the model	Lower number means the model is explaining more of the variation	
RSE = Residual Standard Error	Uses the RSS to determine how well the model fits the data.	Lower the better, but 0 means the model is likely overfit.	RSS
SE = Standard Error	how far off the estimate is from the true population mean.	In small sample sizes, smaller SE might show a relationship between the predictor and response. In large sample sizes, larger SE might be needed to show a relationship between the predictor and response.	
TSS = Total Sum of Squares	Measures the total variance in the response		
t-statistic	measures the number of standard deviations the estimated slope is from 0.	The larger the *t* the more evidence your predictor is significantly different than the average.	SE
p-value	probability of getting values at least as extreme as yours by chance	A small *p-value* indicates the relationship we see between the predictor and response is not due to chance.Standard p-values are 0.05 and 0.01 (5 and 1%, respectively.) A high t-statistic with a low p-value means there is likely a relationship between the predictor and response.	

Name	What it does	Looking for	Based on
R^2 = R-Squared	Proportion of variance explained by the model	Always lies between 0 and 1. 1 = all the variation is explained by the model, 0 = none of the variation is explained by the model	RSS and TSS
F-Statistic	Tells you if a group of variables are significant together	if the f-statistic is greater than the f-critical value that means the results are significant.	
Leverage statistic	tells you how much a single observation is affecting model fit	The bigger the number, the more leverage the observation has.	

- A multiple linear regression uses multiple predictors to predict the response.

Additive Models & Interaction Terms

- Any one predictors effect on the response is independent of the other predictors.
- Interaction effects are when one of more variables have a greater effect on the response than either of them have on it alone.
    - If you include an interaction in the model, always include the main effects, too.

Polynomial Regression

- Used when the relationship between the predictors and response is non-linear.
- Can be achieved by transforming the predictors in the model (squaring, cubing, logging, etc)

Potential Problems

Name	Problem	How do you know?	Solution
Non-linearity	the relationship between the predictor and response is nonlinear	Residual plots	transform the predictors

Name	Problem	How do you know?	Solution
Correlation of Error Terms	If the error terms are correlated you will think some parameters are statistically significant when they are not.	Plot the model as a function of time. No pattern = no correlation	Usually a problem of unit conversion or time series data
Heteroscedasticity, or nonconstant variance of error terms	error terms are nonconstant (an assumption made in linear regression models)	Plot the residuals. Funnel shape = heteroscedasticity	transform the response.
Outliers	observations with values far away from other observations which can affect the RSE (used to calculate confidence intervals)	Plot the data	you can delete them if you want, but be careful that it doesn't change the model fit
Leverage	whether a point influences the fit more than other points	Can be calculated using the leverage statistic. High number = high leverage	
Collinearity	When two or more predictors are closely related to each other	Correlation matrix	remove one of the predictors
Multicollinearity	Three or more predictors are highly correlated	Variance Inflation Factor (VIF). VIF = 1 means no collinearity	Drop a problematic predictor