Chapter 3

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

Linear Regression is a simple approach to supervised learning.

- useful for predicting a quantitative response.
- used as basis for more advanced statistical learning 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 Simple Linear Regression:

Simple linear regression: refers to predicting a quantitative response, Y, based on a single predictor, X.

- it assumes there is a linear relationship between X and Y.
- its equation is: $Y \approx \beta_0 + \beta_1 X$.
- where $\approx$ means "is approximately modeled as."
- also referred to as: *regressing Y on X*
- *e.g.* **partisanship** $\approx \beta_0 + \beta_1 \times education)$
- β_0 & β_1 are unknown constants. They represent the model's *coefficients* or *parameters.*
- β_0 is the intercept. and β_1 is the slope.
- Training data is used to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

- The $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are used to predict partisanship based on a particular level of education.
- This gives us the equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. (aka the *least squares line*)
- $\hat{y}$ is the prediction of Y on the basis of X = x.
- The ^ symbol is used to denote an estimate for an unknown parameter, coefficient, or the predicted value of the response.

3.1.1 Estimating the coefficients

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown, we use data to estimate the coefficients.

- $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ are n number of observation pairs that contain observations for X & Y.
- the goal of linear regression is to get the intercept $(\hat{\beta}_0)$ and slope $(\hat{\beta}_1)$ to be as close as possible to the data points we have.
- There are lots of ways to determine how close our model fits the data, but the most common is to minimize the *least squares.*
- $\hat{y} = \beta_0 + \beta_1 x_i$ is the prediction for Y based on the ith value of X.
- $e_i = y_i - \hat{y}$ is the residual.
- *residual:* difference between the observed value and the predicted value of the ith observation.
- *Residual Sum of Squares* = $RSS = e_1^2 + e_2^2 + ... + e_n^2$
  - $RSS = (y_1 - \hat{\beta}_0 - \hat{\beta_1 x}_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta_1 x}_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta_1 x}_n)^2$
- *least squares* chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS
- *least squares coefficient estimates* for simple linear regression:
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$
  - where: $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{2} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{2} x_i$

3.1.2 Assessing the Accuracy of the Coefficient Estimates

If f is a linear function: $Y = \beta_0 + \beta_1 X + \epsilon$

- This is the *population regression line*
- β_0 is the intercept (Y when X is 0)
- β_1 is the average increase in Y associated with a one-unit increase in X.
- ϵ is the error. Typically this is assumed to be independent of X.
- We can find the least squares line, but rarely know the true values of the population regression line. *i.e.,* we can only estimate the values for the population using the sample.
- *Unbiased* estimators do not **systematically** over or under estimate the true parameters
- *Standard error* is used to determine how far off a single estimate of the population mean is.
  - $Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$

        - μ is the population mean
        - σ is the standard deviation of each y_i of some variable, Y
        - Standard error tells us the average amount that the estimate, $\hat{\mu}$, differs from the actual value of μ
        - Since it is divided by the number of observations, n, our standard error gets smaller as the number of observations gets bigger.
    - We can also calculate the standard error for β_0 and β_1
        - $SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x^2}}{\sum_{i=1}^{2}(x_1 - \bar{x})^2}]$
            - If $\bar{x} = 0$,this is the same as the $SE\hat{\mu}$ equation. This would also mean that $\hat{\beta}_0$ is equal to $\bar{y}$
        - $SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{2}(x_1 - \bar{x})^2}$
            - $SE(\hat{\beta}_1)$ will be smaller when x_i is more spread out. This gives us more *leverage*: how far away the independent variable values are from the other observations.
        - Where $\sigma^2 = Var(\epsilon)$
        - Generally, σ^2 is not known, but can be estimated from the data.
- *Residual Standard Error*: Tells us how well our model fits the data. If it is 0, it's likely that your model is overfit.
    - $RSE = \sqrt{RSS/(n-2)}$
    - This can be used for multiple regressions, too.
- We can use standard errors to compute *confidence intervals:* a range of values for the unknown parameter. Essentially, it is how confident we can be that the estimated interval will contain the true value of the parameter.
    - Range is the upper and lower limits computed from the sample of data.
    - 95% confidence interval for β_1: $\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$
    - For β_0 it is $\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$
- Standard errors are also useful in *hypothesis testing.* Most commonly, they're used to test the *null hypothesis.*
- *Null Hypothesis:*
    - H_0: There is no relationship between X and Y.
    - Mathematically: $H_0 : \beta_1 = 0$
- *Alternative Hypothesis:*
    - H_a: There is some relationship between X and Y.
    - Mathematically: $H_a : \beta_1 \neq 0$
- Standard errors help us determine whether $\hat{\beta}_1$ is sufficiently far away from zero to be confident that β_1 is non-zero.
    - if β_1 is not zero, then there is a relationship between X and Y.
- If $SE(\hat{\beta}_1)$ is small, then relatively small values of $\hat{\beta}_1$ may provide evidence that $\beta_0 \neq 0$
- if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order to reject the null hypothesis.

- *t-statistic:* measures the number of standard deviations that $\hat{\beta}_1$ is from 0.
  - $t = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_1)}$
  - If there is no relationship between X and Y, the *t-statistic* will have a *t*-distribution with *n-2* degrees of freedom.
  - For values of n that are $\approx$ 30, the *t*-distribution will resemble a normal bell curve.
- *p-value:* probability of observing any number equal to |*t*| or larger in absolute value, assuming $\beta_1 = 0$
  - A small p-value indicates that it's unlikely that the association we see between the predictor, X, and the response, Y is due to chance, in the absence of any real association between the predictor and the response.
  - Thus, small p-value means there's likely an association between the predictor, X, and the response, Y.
  - Meaning we can *reject the null hypothesis.*
  - Usually, the p-value should be 0.05 or 0.01 (5 and 1% respectively) to reject the null hypothesis.
  - p-hacking is where researchers try a bunch of different models and only report those with significant p-values. It is bad.

3.1.3 Assessing the Accuracy of the Model

When we reject the null hypothesis (and accept the alternative hypothesis) we want to know *how well* the model fits the data.

- to do so, we use the *residual standard error* (described above) and the R^2 statistic.

Residual Standard Error:

RSE is an estimate of the standard deviation of ϵ. Essentially, it is the average amount that the response, Y, will deviate from the true regression line.

- $RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_1 - \bar{y_i})^2}$
- The RSE gives us a number of how far off we are from the true Y is on the basis of X (*e.g.:* a prediction of partisanship on the basis of education will be off by the RSE amount.)
- RSE is a measure of *lack of fit* of the model and will be small if the model fits the data well.
- It is an absolute measure of lack of fit.
- It is also measured on the units of Y, so it's not always clear what a "good" RSE is.

R^2 Statistic

The R^2 statistic is an alternative measure of fit.

- it the *proportion* of the variance explained. *i.e.* how much of the variance is explained by our model.
- It always lies between 0 and 1.
- It is independent of the scale of Y.
- It can also be used to measure fit for multiple regression models.
- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$
  - $TSS = \sum(y_1 - \bar{y})^2$
  - TSS is the total sum of squares.
    - TSS measures the total variance in the response, Y. Essentially, it's the variance inherent in the response before the regression is performed.
  - the RSS is the amount of variability that's left unexplained after performing the regression.
  - So TSS - RSS measures the amount of variability in the response that is explained by the regression.
- R^2 measures the proportion of variability in Y that can be explained using X. (*e.g.,* how much of a person's partisanship can be explained by their education level.)
- An R^2 closer to 1 indicates a large proportion of the variability in the response has been explained by the regression.
- An R^2 closer to 0 means the regression didn't explain much of the variability in the response.
  - could be because the model is wrong or the inherent error σ^2 is high
- Although easier to interpret than the RSE, it is still difficult to determine what a good R^2 statistic is.
- the R^2 is the same as the squared correlation (r^2) which is the correlation between X and Y and a measure of the linear relationship between the two.
  - $r^2 = Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
  - R^2 and r^2 are identical only in the simple linear regression setting.
  - For multiple regression models. $R^2 = Cor(Y, \hat{Y})^2$
    - This is the square of the correlation between the response and the fitted linear model
    - R^2 will always increase when more variables are added to the model.

3.2 Multiple Linear Regression

Multiple Linear Regression: is using multiple predictors to predict the response.

- The first step in multiple regression models is to compute the *F-statistic* and examine its *p-value.*
- each predictor (p) is given a separate slope coefficient (β).
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$
  - Where p is a stand-in for the number of predictors there are.

3.2.1 Estimating the Regression Coefficients

Like linear regression, the regression coefficients $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ are unknown.

- We must estimate them using
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$
- Least squares is also used to estimate the parameters
  - $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
    - $= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{x2} - ... - \hat{\beta}_p x_{ip})^2$
  - These: $\hat{\beta}_0 + \hat{\beta}_1 + ... + \hat{\beta}_p$ are the multiple least squares coefficient estimates
  - Each coefficient is interpreted as holding the other coefficients as fixed. For example, if you have three predictors education, salary, and age. Let's say that you increase someone's salary by 1,000 dollars. In a multiple regression, this represents the average effect of increasing a person's salary by 1,000 dollars while holding education and age as fixed.
- If your correlation coefficient for salary is significant when running a simple linear regression, but isn't significant when running a multiple regression model it demonstrates that your salary variable may not actually be predictive of your partisanship.

3.2.2 Some Important Questions

1. Is there a relationship between the response, Y, and predictors, X?
   - For multiple regressions, to determine if there is a relationship between the predictors and the response, we have to determine whether all of the regression coefficients ($\beta_1 = \beta_2 = ... = \beta_p$) are equal to zero.
   - Null hypothesis for multiple regression:
     - H_0 : $\beta_1 = \beta_2 = ... = \beta_p = 0$
   - Alternative hypothesis for multiple regression:
     - H_a : at least one β_p is non-zero
   - This is tested using the *F-statistic:*
     - $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$
     - If the linear model assumptions are correct:
       - $E\{ RSS/(n-p-1) \}$ = σ^2
         - E = expectation
       - $E\{ (TSS - RSS)/p \}$ = σ^2
       - This is a test of H_0 if all the coefficients are zero.
     - When there is no relationship between the response and predictors, the *F-statistic* should be close to 1 (accept H_0)
     - When there is a relationship between the response and predictors, the *F-statistic* should be greater than 1 (reject H_0 / accept H_a)
     - How big should the *F-statistic* be?
       - This depends on the values of n and p.

        - A large n may only require a *F-statistic* that is slightly greater than 0.
        - A larger *F-statistic* is needed to reject H_0 if n is small.
    - When ϵ_1 have a normal distribution, then the *F-statistic* follows an F-distribution.
    - Statistical software (like R) will calculate the p-value based on the F-statistic. The p-value will helps us determine whether we can reject the null hypothesis, H_0
  - To test if some of the coefficients are zero:
    - $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0$
    - The *F-statistic* for testing some of the coefficients is:
      - $F = \frac{(RSS_0 - RSS)/q}{RSS/n-p-1}$
      - This reports the partial effect of adding that specific variable to the model
  - Using individual *t-statistics* and their *p-values* to determine if there's a relationship between the variables and the response, there's a good chance that we will *incorrectly* conclude that there is a relationship
    - *F-statistic* does not have this problem because it adjusts for the number of predictors.
    - *F-statistic* can only be used when the number of predictors is less than the total number of observations.
2. Deciding on Important Variables

After determining that there is a relationship between at least one of the predictors is associated with the response, then we need to determine which predictor it is. This leads to *variable selection.* Instead of trying out all the models (which would be time consuming and almost impossible for any model that doesn't have a very small number of predictors *i.e.,* 1 or 2 max), there are three ways to select variables:

- *Forward selection:*
  - Start with the *null model:* the model that contains the intercept but no predictors.
  - Fit p simple linear regressions
  - Add to the null model the variable that results in the lowest RSS.
  - Add the variable that results in the lowers RSS for the new two-variable model.
  - Stop when a predetermined stopping rule is met.
- *Backward selection:*
  - Start with all the variables in the model.
  - Remove the variable with the largest *p-value, i.e.,* the least statistically significant.
  - Refit the new $(p-1)$ model and repeat the previous step.
  - Stop when a predetermined stopping rule is met.
- *Mixed selection:*
  - Start with no variables in the model.
  - Add in the variables one by one.
  - If the *p-value* for one of the added variables rises above a certain predetermined threshold, remove it.

    - Repeat until they have a sufficiently low p-value.

3. Model fit
    - R^2 can be used to assess multiple regression model fit in the same way that it can be used for simple linear models (described above).
    - The RSE is different. It is:
        - $RSE = \sqrt{\frac{1}{n-p-1}RSS}$
    - Graphic the model is useful to see patterns that cannot be observed in numbers alone.
        - Sometimes graphing a model will show that there is *synergy* or *interaction* effects between the predictors that mean that increasing one (*e.g.* education) also increases the other (*e.g.* income) which have a double effect on the response, (*e.g.* partisanship)
4. Predictions
    - We can use the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$ to use the fitted multiple regression model for prediction.
    - There are three types of uncertainty present in multiple regression models:
        - The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, ... \beta_p$
        - that means the *least squares plane:* $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p x_p$ is only an estimate for the *true population regression plane:* $f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$
        - This means that there is some *reducible error* that can be improved with better models. We can compute a *confidence interval* to determine how close $\hat{Y}$ is to $f(X)$
    - Since the model is only an approximation for reality, it necessarily contains *model bias.* This is ignored because there's no way to model reality.
    - There is also *irreducible error* because of random error, ϵ.
        - To determine how far off $\hat{Y}$ is from fX we use *prediction intervals.* They are always wider than confidence intervals.
        - *prediction intervals* are used to quantify the uncertainty surrounding the predictor for a particular unit. (*e.g.* you would use a prediction interval to quantify the uncertainty of **education** for a particular individual.)
        - *confidence intervals* are used to quantify the average uncertainty for predictor has on a large number of units. (*e.g.* you would use a confidence interval to quantify the uncertainty surrounding *average* **education** over the population.)

3.3 Other Considerations in the Regression Model

3.3.1 Qualitative Predictors

Predictors with Only Two Levels

A qualitative predictor (*factor*) with only two levels can only take two possible responses. (*e.g.,* Male v. Female,** pregnant v. Not-pregnant)

- They are coded as *dummy variables* with a 1 or 0.

$$x = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person is male} \end{cases}$$

- using the dummy variable as a predictor in the regression equation:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i \text{ th person is female} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th person is male.} \end{cases}$$

- coding the variables as 1 or 0 is arbitrary, it just impacts how you interpret the regression output.

Qualitative Predictors With More Than Two Levels

Since dummy variables only take one of two numerical values, predictors with multiple variables (*e.g.,* ethnicity) need to be coded differently.

- One method is to code multiple dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{th person is Asian} \\ 0 & \text{if the } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if the } i\text{th person is Caucasian} \\ 0 & \text{if the } i\text{th person is not Caucasian} \end{cases}$$

- these are then used in the regression equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i \text{ th person is Caucsian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if the } i \text{ th person is Asian} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th person is African American.} \end{cases}$$

- β_0 is the average for African Americans, β_1 is the difference in the average for Asian & African Americans, and β_2 is the difference in average between Caucasian and African Americans.
- There will always be one fewer dummy variables than number of levels (which is why there's no "Not African American" category.) This is known as the *baseline.*

** This is problematic, but in quantitative science gender or biological sex is usually coded as either / or.

3.3.2 Extensions of the Linear Model

The linear model makes two assumptions that are often not reflected in practice.

- Linear models assume that the relationship between the predictors and response are:
  1. *additive:* This means that any one predictor's effect on the response is independent of the other predictors.
     - For example, looking at the effects of age, income, and education on a person's partisanship, the linear model will assume the changes in income are independent of the changes in age or education on a person's partisanship.
  2. *linear:* The change in the response due to a one-unit change in the predictor is constant, regardless of the value of the predictor.
     - So for every one unit change in income the effect on partisanship is constant, regardless of whether that one unit change is making 50,000 to 60,000 or 50,000 to 100,000.

These assumptions can be violated (or relaxed) in a variety of ways.

Removing the Additive Assumption

An *interaction effect* is when one or more variables have a greater effect on the response than either of them have on it alone.

- For example, a person's income and ethnicity both have an effect on their partisanship. However, looking at these two predictors independently does not tell the whole story. Rich white men will be more conservative than rich African American men because the interaction of income and ethnicity.

To include an interaction term, the model looks like this:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$
- It computes the product of X_1 and X_2
- The resulting equation is:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{1}$$

- In this equation $\tilde{\beta_1}$ changes with X_2, so the effect of X_1 on Y is not constant (thus breaking the additive assumption).
- For the partisanship example, the equation would be:
  - $partisanship = \beta_0 + (\beta_1 + \beta_3 \times income) \times education + \beta_2 \times income + \epsilon$
  - β_3 can be interpreted as the increase in impact that $education$ has for a one unit increase in $income$ (or vice versa).

- Models that **do not** contain an interaction term are called *main effects* terms.
- *Hierarchical Principle:* if we include an interaction in a model, we should also include the main effects, even if the p-values associated with the main effect's coefficients are not significant.
  - Essentially, always include the main effects if you're using an interaction.
- Interaction effects can be applied to quantitative, qualitative, or a mix of both types of variables.

Non-linear Relationships

Polynomial Regression is used then the relationship between the response and predictors is non-linear.

- one method to incorporate non-linear associations in a linear model is to transform the predictors in the model.
- the model is still linear, it just has X_1 as a variable and X_2 as the same variable squared.
  - $partisanship = \beta_0 + \beta_1 \times education + \beta_2 \times education^2 + \epsilon$

Potential Problems

1. Non-linearity of the data
   - To identify if you have a non-linear relationship you can use: *Residual Plots*
     - For residual plots you plot the residuals $e_i = y_1 - \hat{y}$, vs the predictor x_i
     - For multiple regression models, you plot the residuals vs. the predicted values of $\hat{y}_1$.
   - Ideally, this plot will show no pattern.
   - If there is a pattern, you will need to perform a transformation on the predictors.
     - *e.g.* log X, $\sqrt{X}$, and X^2
2. Correlation of Error Terms
   - Linear regressions assume that the error terms ($\epsilon_1, \epsilon_2, ...\epsilon_n$) are uncorrelated.
   - If the error terms are correlated, then you will underestimate the true standard errors. This means that you will likely think that some parameters are statistically significant when they're not.
   - Correlated error terms are often found in *time-series data* which are measurements about predictors at different points in time (*e.g.* income at age 20, 30, 40, 50 for one individual)
     - to see if the error terms are correlated, you can plot your model as a function of time.
     - if there is no pattern, then the error terms are uncorrelated.
     - if the error terms are positively correlated, then there will be *tracking:* where the adjacent residuals may have similar values.
3. Non-Constant Variance of Error Terms
   - Linear regression models also assume that the error terms have a constant variance, $Var(\epsilon_1) = \sigma^2$
   - Standard errors, confidence intervals, and confidence testing rely on this assumption.
   - *Heteroscedasticity,* or non-constant variance, shows when the residuals are plotted and produce a funnel shape.

        - if there is Heteroscedasticity, you can transform the response variable, Y, by using either log Y or $\sqrt{Y}$
        - if you know certain responses will have a non-linear variance, you can use *weighted least squares.* This is where you weight certain observations.
4. Outliers
    - *Outliers:* are points where the predicted value of the response is far from the actual response value (y_i).
    - Outliers may not affect the regression fit, but can affect the RSE which is used to determine confidence intervals, your interpretation might be off.
    - *Residual plots* are used to identify outliers, but sometimes it's hard to tell how far off a point needs to be before it's considered an outlier.
        - A solution is to use *studentized residuals:* which is dividing each residual e_i by its estimated standard error.
            - if the absolute value of the studentized residual is greater than 3, the observation is a possible outlier.
    - You can simply remove outliers if you think they were caused by errors in data collection or recording. But, be careful because outliers may also indicate a bad model.
5. High Leverage Points
    - *High leverage points* are observations with an unusual value for x_i.
    - These observations have a large impact on the estimated regression line which is a problem because a few high leverage points can change the entire model fit.
    - For simple linear regression models, just look for observations with values that fall outside the normal range.
    - For multiple linear regression models, use the *leverage statistic.*
        - The bigger the number, the more leverage the observation has.
        - $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$
        - h_i will increase as the distance from x_i and $\bar{x}$ increases.
6. Collinearity
    - *Collinearity:* is when two or more predictor variables are closely related to one another.
    - If two predictors are *collinear* it's difficult to separate out which one is actually associated with the response.
        - *e.g.* if education and income are *collinear* and we want to know how these affect partisanship, it will be difficult to determine whether it is income that increases partisanship or whether it is the effect of education.
    - Collinearity reduces the accuracy of the estimates of the regression coefficients, causing the standard error to grow. In turn, since the *f-statistic* is based on the standard error, collinearity decreases the *t-statistic.*

        - This reduces the *power* of the hypothesis test (i.e. the probability of correctly detecting a non-zero coefficient and rejecting the null hypothesis) is reduced.
    - You should look for collinearity when fitting the model.
        - Look at a correlation matrix of the predictors.
            - An element that's high in absolute value shows there's a pair of highly correlated variables.
            - This will not be useful when detecting *multicollinearity.*
    - *Multicollinearity:* is when three or more variables are highly correlated
        - The *Variance Inflation Factor (VIF)* is used to determine if there's multicollinearity. It is the ratio of the variance when fitting the full model divided by the variance if fit on its own.
        - The smallest number is 1 which indicates a complete absence of collinearity
        - If the VIF is greater than 5 or 10 then there is evidence of multicollinearity
        - $VIF(\hat{\beta}_j = \frac{1}{1- R^2_{X_j|X_{-j}}})$
            - $R^2_{X_j|X_{-j}}$ is R^2 from a regression of X_j onto all the other predictors.
    - To solve collinearity, drop one of the problematic variables from the regression (since one of the variables is providing redundant information). Of combine the collinear variables into one predictor. (*e.g.,* take the average standardized versions of $income$ and $education$ to create a new variable called $class$)

3.5 Comparison of Linear Regression with *K*-Nearest Neighbors

One of the best non-parametric methods is *K-Nearest Neighbors Regression (KNN regression).*

- Given a value for K and a prediction point for x_0, *KNN:*
    - First, *KNN* identifies the K training observations that are closes to x_0, represented by $\mathcal{N}_0$
    - Second, it estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$
        - $\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \epsilon \mathcal{N}_0} y_i$
- The best value for K will depend on the *bias-variance tradeoff*
    - (From section 2.2.2): the *bias-variance tradeoff* requires a low variance as well as a low squared bias.
- A small K value is more flexible which will have low bias but high variance.
- A high K value will have a smother and less variable fit, because changing one observation will have a smaller effect. However, this might result in high bias because it will mask the structure of $f(X)$
- adding dimensions generally causes *KNN* to perform worse, because it reduces the sample size.

- Essentially, in order to find a nearest neighbor, *KNN* would have to have more data points on which to find a neighbor thus reducing the sample size.
- In highly dimensional data it would results in an observation having no *nearby neighbor,* which is the *curse of dimensionality.*

Using a *parametric* approach is better than a *non-parametric* approach if the *parametric* form that has been selected is close to the true form of $f(X)$.

- *parametric* models will do better than *non-parametric* methods when there's a small number of observations per predictor.
- Linear models are also easier to interpret, so we might give up some fit for interpretability.

///////// All the knowledge, very little math /////////

Basic linear equation: $Y = \beta_0 + \beta_1 X + \epsilon$

Step one: Is there a relationship between the predictor and the response?

1. Linear models are all about estimating values for $\hat{\beta}_0$ (intercept) and $\hat{\beta}_1$ (slope) so that the model fits the predictors as close as possible.
2. β_1 is the average increase in the response, Y, associated with a one-unit increase in X.
   - *e.x.* $partisanship = \beta_0 + \beta_1 \times education + \epsilon$
   - Here, for every one unit increase in $education$ we would see a β_1 increase in $partisanship$.
3. One the model is fit, we want to know how well it's mirroring reality by looking closely at the errors: ϵ
   - For linear models, ϵ is assumed to be independent of X
4. We can calculate the *standard error* (SE) to determine the average amount that the population mean ($\hat{\mu}$) is from reality.
   - *standard error* can also be used on β_0 and β_1
   - *standard error* will always become smaller as the number of observations gets bigger.
   - *standard error* also tells us if $\hat{\beta}_1$ is sufficiently far away from 0 to say that there is a relationship between the predictors (X) and the response (Y).
5. *t-statistic* measures the number of standard deviations that $\hat{\beta}_1$ is from 0.
   - the *t-statistic* is used with the *p-value.*
   - it is based on the standard error SE
   - the greater the *t-statistic,* the more evidence your results are significantly different from the average.
   - the smaller the *t-statistic,* the more evidence that your results are not significantly different from the average.
6. *p-value* is the odds that the result you observed happened by chance.

- the smaller the *p-value* means there's likely a relationship between the predictor and response (since it's unlikely to have occurred by chance)
- Usually a *p-value* of 0.05 or 0.01 (5 and 1 %, respectively) mean we can reject the *null hypothesis.*

Step two: How well does the model fit the data?

7. *Residual standard error,* RSE, tells us how well the model fits the data.
    - It is based on the RSS.
    - If it's 0, your model is probably overfit.
    - RSE is also used to compute *confidence intervals.*
        - *Confidence intervals* tell us how confident we can be that the estimated interval contains the parameter's real value.
    - *e.g.* a prediction of $partisanship$ on the basis of $education$ will be off by the RSE amount.
    - It is measured in Y units, so it's hard to know what a good RSE is.
8. R^2 is the proportion of the variance explained by the model.
    - *e.g.* how much of a person's $partisanship$ can be explained by their $education$ level.
    - It always lies between 0 and 1.
    - The closer to 1 the better.
    - It is independent of Y.
    - Uses the *residual sum of squares*, RSS and the *total sum of squares,* TSS
9. *Residual sum of squares,* RSS, is the amount of variability that's left unexplained after performing the regression.
10. *Total sum of squares,* TSS, measures the total variance in the response (Y).