# SHEET - An Introduction to Statistical Learning
# Chapter 2 - Statistical Learning

PLAYE Nicolas

1 December 2023

## 0.1 Answers of Exercises

**Question 1: Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio , and newspaper , rather than in terms of the coefficients of the linear model..**

1-Intercept: The null hypothesis is that the intercept is equal to zero. In other words, it suggests that if no money is spent on TV, radio, or newspaper advertising, sales would be zero. The very low p-value (¡ 0.0001) allows us to reject this null hypothesis, indicating that the intercept is significantly different from zero. This means there are some baseline sales even without advertising. 2-TV: The null hypothesis is that TV advertising has no effect on sales, meaning there is no relationship between the amount spent on TV ads and sales. The very low p-value (¡ 0.0001) strongly rejects this hypothesis, suggesting a significant and positive relationship between TV advertising and sales. 3-Radio: The null hypothesis is that radio advertising has no effect on sales, implying no relationship between the radio ad budget and sales. The p-value (¡ 0.0001) is also very small, which allows us to reject this null hypothesis. This shows there is a significant and positive relationship between radio advertising and sales. 4-Newspaper: The null hypothesis is that newspaper advertising has no effect on sales. With a p-value of 0.8599, we fail to reject this hypothesis, indicating that newspaper advertising does not have a statistically significant effect on sales in this model.

Conclusions: -TV and radio advertising have a significant and positive impact on sales. -Newspaper advertising does not have a statistically significant effect on sales. -The model predicts a level of sales that is significantly different from zero, even when no advertising money is spent.

**Question 2: Carefully explain the differences between the KNN classifier and KNN regression methods**

The KNN regression method is closely related to the KNN classifier. The key difference between the KNN (k-nearest neighbors) classifier and regression methods lies in the type of output they generate and the nature of the problem they are used to solve. when KNN classification predicts categories (classes), KNN regression predicts continuous numerical values. The output is computed: - In classification, the prediction is made based on a majority vote among the k neighbors, -In regression, the prediction is made by averaging the target values of the k neighbors.

**Question 3: Suppose we have a data set with five predictors X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get**

$$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$$

**(a)Which answer is correct, and why?**
**i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.**
**ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates. 128 3. Linear Regression**
**iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.**
**iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.**
We have

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$$

When IQ and GPA are fixed value we have

$$Y = Cte + 35X_3 - 10X_5$$

Reformulated we Have

$$Salary = Cte + 35Level - 10Level * GPA$$

We are searching the response by comparaison of level. For college level we have

$$Salary_{college} = Cte + 35 - 10GPA$$

For High School level we have

$$Salary_{hs} = Cte$$

Then

$$Salary_{college} \geq Salary_{hs}$$

$$Cte + 35 - 10 * GPA \geq Cte$$

$$35 - 10 * GPA \geq 0$$

$$3.5 \geq GPA$$

The result shows that knowing that college graduates earn more on average than high school graduates depends on GPA. In fact if GPA is lower than 3.5, then College graduates earn more on average than high school graduates. So if GPA is high enough (over 3.5) then high school graduates earns more one average than college graduates. Which is answer iii.

**(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.** We have

$$Y = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA * IQ - 10 * GPA * Level$$

Then

$$Y = 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 4.0 * 110 - 10 * 4.0 * 1 = 137.1$$

**(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer** The size of the coefficient for an interaction term in this case,

$$\hat{\beta}_4 = 0.01$$

alone does not provide enough information to conclude whether or not there is little evidence of an interaction effect. A small coefficient means that the effect of the interaction between GPA and IQ on the response (starting salary) is small in magnitude, but it does not directly indicate whether the interaction is statistically significant.To determine if there is little evidence of an interaction effect, we would need to examine: - The p-value associated with the interaction term. A high p-value would indicate weak evidence against the null hypothesis (no interaction effect). - The confidence interval for the interaction term. If the confidence interval includes zero, it would suggest that the interaction effect might not be significant. Thus, the small size of the coefficient suggests a minor practical effect, but we would need more information (such as statistical significance) to conclude whether the interaction effect is supported by the data.

**Question 4: I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e.**

$$\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{X} + \beta_\mathbf{2}\mathbf{X^2} + \beta_\mathbf{3}\mathbf{X^3} + \varepsilon$$

**(a) Suppose that the true relationship between X and Y is linear, i.e.**

$$\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{X} + \varepsilon$$

**Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.**

If the true relationship between $X$ and $Y$ is linear, i.e.,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

then we would expect the training residual sum of squares (RSS) for the cubic regression to be **lower** than the RSS for the linear regression. - The linear model only includes terms for $X$ and the intercept, fitting the true underlying relationship between $X$ and $Y$. This would result in a good fit with the appropriate level of complexity for the data. - The **cubic model** includes additional terms for $X^2$ and $X^3$, which means it has more flexibility. Even though the true relationship is linear, the cubic regression can still fit the data more closely because it has more parameters, allowing it to "bend" and potentially overfit the noise in the training data. This additional flexibility leads to a lower training RSS because the model can adapt more to the specific points in the training data. Therefore, even though the true relationship is linear, the cubic model will likely have a lower training RSS due to its greater capacity to fit the data points closely, even if some of that extra fit is to random noise. However, this lower training RSS does not imply that the cubic model is a better representation of the true relationship—it might perform worse on new data due to overfitting.

**(b) Answer (a) using test rather than training RSS.**
When considering the test residual sum of squares (RSS) instead of the training RSS, we would expect the linear regression model to have a lower test RSS than the cubic regression model, assuming the true relationship between $X$ and $Y$ is linear. - The linear model reflects the true underlying relationship between $X$ and $Y$ (i.e., $Y = \beta_0 + \beta_1 X + \varepsilon$). Therefore, it is expected to generalize well to unseen data, leading to a lower test RSS. - The cubic model has more flexibility and includes unnecessary higher-order terms ($X^2$ and $X^3$) that do not represent the true relationship. While it can fit the training data better (lower training RSS), this extra flexibility can lead to overfitting. Overfitting occurs when the model captures not just the underlying pattern but also the random noise in the training data. This results in worse performance on new (test) data because the model is too complex for the true linear relationship. Therefore, in the test set, where the goal is to generalize to new data, the linear model should have a lower test RSS than the cubic model. The cubic model's overfitting on the training data will typically result in a higher test RSS.

**(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.**
If the true relationship between XX and YY is not linear, but we do not know how far it is from linear, we would expect the training RSS for the cubic regression to be lower than the training RSS for the linear regression. -The linear regression model assumes a simple linear relationship between XX and YY. If the true relationship is not linear, the linear model will not be flexible enough to capture the more complex relationship, leading to a higher residual sum of squares (RSS) on the training data. - The cubic regression model is more flexible because it includes higher-order terms (X2X2 and X3X3). This extra flexibility allows the cubic model to better fit the training data, even if the true relationship is non-linear. The cubic regression will be able to capture more complex patterns in the data, including non-linear trends, resulting in a lower training RSS compared to the linear regression model. - Since the cubic regression model can fit the data more closely by including additional terms, it will almost always have a lower training RSS than the linear model, regardless of how non-linear the true relationship is. This is because adding more terms to the model increases its ability to fit the data, even if the higher-order terms are not necessary or beneficial for generalization. In summary, the cubic model will likely have a lower training RSS due to its greater flexibility, regardless of how far the true relationship is from linear. However, this lower training RSS does not necessarily mean the cubic model is better—it could be overfitting the training data.

**(d) Answer (c) using test rather than training RSS.**

If the true relationship between X and Y is not linear, and we consider the test RSS rather than the training RSS, the situation changes: - Linear Regression Model: This model assumes a linear relationship between X and Y. If the true relationship is not linear, the linear model might not fit the training data perfectly, leading to a higher training RSS. However, because it is simpler, it may generalize better to new, unseen data. - Cubic Regression Model: This model includes additional terms ($X_2$ and $X_3$) and can fit a more complex relationship. It will likely have a lower training RSS because it can adjust more closely to the specific training data, including any non-linear patterns. Lower Test RSS for Linear Model: If the true relationship is not extremely complex or non-linear, the linear model might generalize better to new data. The cubic model, while fitting the training data well, may overfit and not perform as well on the test data. This overfitting can lead to a higher test RSS for the cubic model because it captures not only the underlying pattern but also the noise in the training data. Potential for Lower Test RSS for Cubic Model: If the true relationship is significantly non-linear and the cubic model is appropriately capturing the underlying pattern, it might have a lower test RSS compared to the linear model. This would be the case if the additional complexity in the cubic model is genuinely useful and not just capturing noise. Finally, without knowing the exact nature of the non-linearity, we generally expect that if the non-linearity is moderate or if the cubic model overfits, the linear regression model will likely have a lower test RSS. This is because it avoids overfitting and maintains better generalization to unseen data. However, if the non-linearity is significant and the cubic model is correctly capturing the true relationship, it might have a lower test RSS. In practice, the cubic model tends to have a lower training RSS but may not always outperform the linear model on test data due to potential overfitting.

**Question 4: Consider the fitted values that result from performing linear regression without an intercept. In this setting, the $i$-th fitted value takes the form**

$$\hat{y}_i = x_i\hat{\beta},$$

**where**

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}.$$

**Show that we can write**

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

**What is $a_{i'}$?**

**(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.**

- Type of Problem: Regression (CEO salary is a continuous variable).
- Interest: Inference (understand relationship between variables and CEO salary).
- n (number of observations): 500 (one for each firm).
- p (number of predictors): 3 (profit, number of employees, industry).

**(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.**

- Type of Problem: Classification (categorical outcome success/failure).
- Interest: Prediction (forecasting the success or failure of a new product).
- n (number of observations): 20 (for each previously launched similar product).
- p (number of predictors): 13 (price charged, marketing budget, competition price, and ten other variables).

**(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.**

-Type of Problem: Regression (predicting a continuous outcome - the % change in the USD/Euro exchange rate).

-Interest: Prediction (forecasting the future value of the exchange rate based on changes in stock markets).

- n (number of observations): The number of weeks in 2012. Since 2012 was a leap year, there would be 52 weeks.
- p (number of predictors): 3 (% change in the US market, % change in the British market, and % change in the German market).

**Question 3: We now revisit the bias-variance decomposition.**

**Definitions:**

In statistical learning, **bias-variance decomposition** is a fundamental concept used to understand and improve the performance of machine learning models. It involves breaking down the error of a model into two main components: bias and variance, along with an irreducible error term.

- **Bias**: This refers to the error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to miss relevant relations between features and target outputs (underfitting). It's essentially the difference between the average prediction of our model and the correct value which we are trying to predict.
- **Variance**: This is the error due to too much complexity in the learning algorithm. High variance can cause the model to model the random noise in the training data, rather than the intended outputs (overfitting). It's the variability of model prediction for a given data point.
- **Irreducible Error**: This is the error inherent in the problem itself, often due to randomness or noise in the data. It can't be reduced by any model.

As previously mentioned and demonstrated, we have:

$$E[(y-\hat{f})^2] = \mathrm{Var}(\hat{f}) + \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\varepsilon)$$

**(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.**

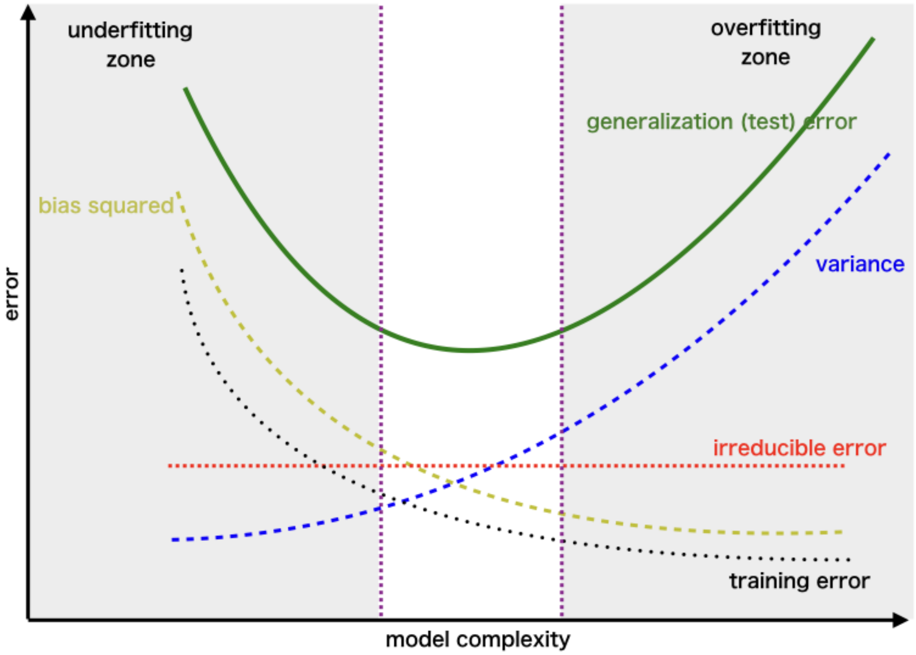

Figure 1: Biais-Variance Tradeoff

**(b) Explain why each of the five curves has the shape displayed in part (a).**

**Bias Squared (yellow curve)**: As model complexity increases, the model has more capacity to fit the data, and thus bias decreases. This is why the curve starts high and slopes downwards.

**Variance (blue curve)**: High variance can cause over-fitting. As model complexity increases, a model's sensitivity to data fluctuations increases, leading to higher variance. Hence, the curve starts low and slopes upwards.

**Training Error (yellow line)**: This is the error on the training set, which typically decreases as the model becomes more complex, because a more complex model can fit the training data better. Therefore, the curve trends downward. It goes down the irreducible error because it learns the noise at a certain level of complexity.

**Generalization (Test) Error (green curve)**: This is the total error of the model on unseen data. It's the sum of bias squared, variance, and irreducible error. Initially, as complexity increases, the model learns better, and the error decreases. However, after a certain point, increasing complexity only fits to noise, increasing the variance without reducing bias, causing the error to rise again. This creates a U-shaped curve.

**Irreducible Error (black dotted line)**: This is the error that cannot be reduced regardless of how good the model is, often due to noise in the data itself. It's constant, hence the flat line; it doesn't change with model complexity.

**Question 4: You will now think of some real-life applications for statistical learning.**

**(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer**

- **Medical Diagnosis**
  - Response: Diagnosis (e.g., presence or absence of a disease).
  - Predictors: Patient symptoms, medical test results, age, gender, etc.
  - Goal: prediction – accurately diagnose the presence or absence of a disease in new patients. However, inference can also be important, for understanding which factors are most indicative of certain diseases, which can guide treatment and prevention strategies.
- **Credit Scoring in Finance**
  - Response: Creditworthiness (e.g., high risk or low risk).
  - Predictors: Credit history, debts, income, employment status, age.
  - Goal: prediction - determine the likelihood that a person will repay their debts. This helps financial institutions decide whether to grant a loan. Inference can be secondary, useful for understanding which factors most strongly predict default risk.
- **Customer Churn Prediction in Telecommunications**
  - Response: Churn (whether a customer will leave or quite the service).
  - Predictors: patterns, customer interactions, billing history, plans.
  - Goal: prediction – anticipate which customers are at risk of leaving so that the company can take proactive steps to retain them. While prediction is the primary goal, inference can help understand the reasons behind churn, aiding in developing better customer retention strategies.

**(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.**

- **Real Estate Pricing:**
  - Response: The price of a house.
  - Predictors: Location, size (square footage), number of bedrooms, age of the house, proximity to amenities, etc.
  - Goal: prediction. Real estate agents and buyers use regression to predict the market value of a house based on its characteristics. While some inference about the importance of each predictor (like how much value a bedroom adds) might be made, the primary focus is on accurately predicting prices.
- **Healthcare - Patient Outcomes Prediction:**
  - Response: Patient recovery time or treatment effectiveness.
  - Predictors: Age, gender, pre-existing health conditions, type of treatment, lifestyle factors (such as smoking, diet, exercise), genetic information, etc.
  - Goal: This application serves both prediction and inference. For healthcare providers, predicting patient outcomes helps in tailoring treatment plans. Moreover, understanding how different predictors affect recovery (inference) can guide research and policy decisions in healthcare.
- **Marketing - Customer Lifetime Value (CLV):**
  - Response: The lifetime value of a customer.
  - Predictors: Purchase history, engagement with marketing campaigns, demographic information, browsing behavior on the company website, customer service interactions, etc.
  - Goal: Primarily prediction, as businesses use this information to predict how valuable a customer might be in the long term, aiding in resource allocation for marketing and customer retention strategies. In some cases, there might be an interest in inference as well, particularly in understanding which factors most strongly influence CLV.

**(c) Describe three real-life applications in which cluster analysis might be useful.**

- **Market Segmentation in Marketing:**
  - Application: Businesses often use cluster analysis to identify distinct groups within their customer base.
  - Data used for clustering: Purchasing habits, customer preferences, demographic data, engagement with marketing channels.
  - Purpose: By understanding different market segments, businesses can tailor their products, services, and marketing strategies to meet the specific needs and preferences of each group. This targeted approach can lead to more effective marketing and higher customer satisfaction.
- **Genomic Data Analysis in Biology:**
  - Application: In biology, particularly in genomics, cluster analysis is crucial for understanding genetic similarities and differences.
  - Data used for clustering: Genetic sequences, expression levels of genes, protein-protein interaction data.
  - Purpose: Scientists use clustering to group genes or organisms based on genetic features, helping in identifying functional relationships, evolutionary patterns, and classifying new species or gene functions. This can guide research into disease mechanisms, drug discovery, and understanding evolutionary biology.
- **Document Clustering in Information Retrieval:**
  - Application: In information retrieval, such as search engines or digital libraries, cluster analysis helps in organizing large volumes of text data.
  - Data used for clustering: Keywords, text content, metadata, user interaction data with documents.
  - Purpose: By grouping similar documents, users can navigate information more efficiently, improving the search experience. It also aids in automatic topic extraction and summarization, which is useful in fields like academic research and legal document analysis.

**Question 5: What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred**

- **Very Flexible Approach**
  - *Advantages:* **Complex Pattern Recognition**: Better at detecting complex relationships in data. **High Predictive Accuracy**: More effective for datasets with high variance and intricate structures. **Adaptability**: Can adapt to a wide range of data shapes and structures.
  - *Disadvantages:* **Overfitting**: Prone to overfitting the training data. **Interpretability**: Often less interpretable. **Computational Intensity**: Requires more computational resources and time. **Data Requirements**: Needs more data to perform effectively.
- **Less Flexible Approach**
  - *Advantages:* **Interpretability**: Easier to understand and interpret. **Generalization**: Better at generalizing to unseen data. **Implicity**: Simpler to implement and requires less computational power. **Stability**: Less susceptible to noise in the training data.
  - *Disadvantages:* **Underfitting**: Can miss important complexities in data. **Predictive Accuracy**: May have lower predictive accuracy in complex data patterns. **Flexibility**: Not as adaptable to varying data structures and relationships.
- **When to Prefer Each Approach**
  - *Prefer More Flexible Approach:* When data has complex patterns that simpler models cannot capture. In scenarios where predictive accuracy is critical, especially in large datasets. When capturing as much information from the data is a priority.
  - *Prefer Less Flexible Approach:* When interpretability is crucial, such as in medicine or social sciences. In cases with smaller datasets to avoid overfitting. When computational resources are limited or simplicity and speed are desired. In situations with linear or nearly-linear data relationships or high noise level.

**Question 6: Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?**

**Definitions:**

- **Parametric Approach**: Parametric methods involve making assumptions about the form of the function that relates the predictors to the response. (i.e.: Linear Regression,Logistic Regression).
- **Non-Parametric Approach**: Non-parametric methods do not make explicit assumptions about the functional form of the relationship between predictors and response. Instead, they seek to estimate the relationship between predictors and response directly from the data. (i.e.: K-Nearest Neighbors (K-NN), Decision Trees, etc.).

- **Advantages of a Parametric Approach**
  - **Simplicity**: By assuming a specific form, parametric models are generally simpler and easier to understand.
  - **Efficiency**: They require fewer data points to estimate the parameters effectively.
  - **Interpretability**: The relationship between variables is often easier to interpret in parametric models.
  - **Predictive Performance**: In cases where the parametric form aligns well with the underlying data structure, these models can perform exceptionally well.
- **Disadvantages of a Parametric Approach**
  - **Bias**: If the chosen model form does not align well with the true underlying data structure, this can introduce bias.
  - **Flexibility**: Parametric models are less flexible in adapting to complex or non-linear relationships unless the correct form is known and specified.
  - **Overfitting to Model Assumptions**: There is a risk of overreliance on model assumptions, which might not hold in real-world scenarios.

**Question 7: The table below provides a training data set containing six observations, three predictors, and one qualitative response variable**

**(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0**

- Distance from Observation 1: 3.0
- Distance from Observation 2: 2.0
- Distance from Observation 3: 3.162
- Distance from Observation 4: 2.236
- Distance from Observation 5: 1.414
- Distance from Observation 6: 1.732

**(b) What is our prediction with K = 1? Why?**

With K=1 in K-nearest neighbors (K-NN), the prediction is made based on the single closest data point in the training set to the test point. The prediction is simply the response value of this nearest neighbor. From the previously calculated Euclidean distances, we see that the closest observation to the test point X1=X2=X3=0 is Observation 5, which has the smallest distance of 1.41. Observation 5 has the response value ”Green”. Therefore, with K=1, our prediction for the test point X1=X2=X3=0 using K-nearest neighbors is ”Green”.

**(c) What is our prediction with K = 3? Why?**

For K-nearest neighbors (K-NN) with $K = 3$, the prediction is based on the three closest data points in the training set to the test point. The prediction is typically the most common response value among these three nearest neighbors. From the Euclidean distances calculated earlier, the three closest observations to the test point $X1 = X2 = X3 = 0$ are:

- Observation 5 (Distance = 1.414) with response ”Green”
- Observation 6 (Distance = 1.732) with response ”Red”
- Observation 2 (Distance = 2.0) with response ”Red”

Among these three observations, there are two ”Red” responses and one ”Green” response. Since ”Red” is the most common response among the three nearest neighbors, the K-NN prediction with $K = 3$ for the test point $X1 = X2 = X3 = 0$ is ”Red”.

**(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?**

- **Small K Provides More Flexibility**: A smaller value of $K$ makes the K-NN algorithm more flexible and capable of adapting to more complex, non-linear decision boundaries. It allows the model to more closely follow the intricate patterns in the data.

- **Large K Smoothens the Boundary**: In contrast, a larger value of $K$ leads to a smoother, more linear decision boundary. This is because as $K$ increases, the model starts to consider more neighbors, effectively averaging out the local variations and thus, ironing out the non-linearities.

- **Risk of Overfitting vs. Underfitting**: While a smaller $K$ can capture complex decision boundaries, it also increases the risk of overfitting to the noise in the training data. On the other hand, a larger $K$ might underfit the data, especially when the true decision boundary is complex.

- **Balancing Bias and Variance**: The choice of $K$ involves balancing bias and variance. A small $K$ leads to low bias but high variance, while a large $K$ leads to high bias but low variance. In the case of a highly non-linear Bayes decision boundary, the risk of high bias (i.e., oversimplifying the model) is more concerning than high variance.

In summary, for a highly non-linear decision boundary, a smaller value of $K$ is generally preferred as it allows the model to capture the complex patterns in the data more effectively. However, care must be taken to avoid too small a value that might lead to overfitting.