that hypothesis tests, confidence intervals, and prediction intervals are sensitive to departures from independence and departures from equal variance. Hypothesis tests and confidence intervals for the slope and intercept are robust against departures from normality. Lastly, prediction intervals are very sensitive to departures from normality.

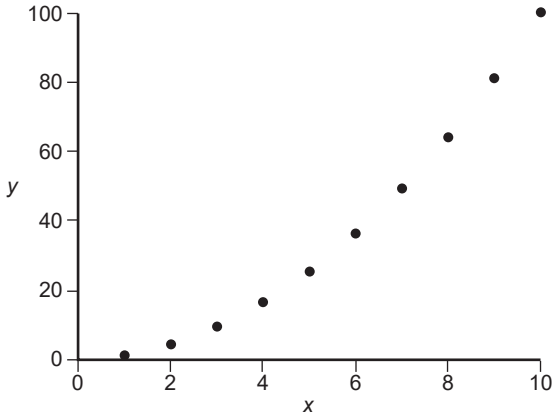
13.2 Exercises

Basic Concepts

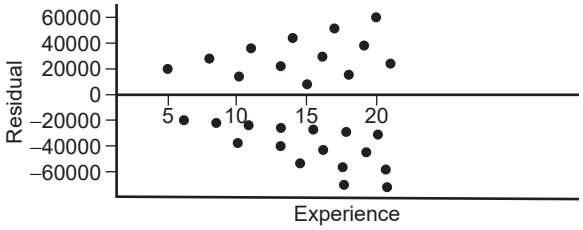
1. What are the assumptions of the simple linear model that need to be validated when doing a residual analysis?
2. What should a well-behaved residual plot look like?
3. List three ways to determine if the errors are normally distributed in a regression analysis.
4. How should a normal probability plot look to indicate normality?

Exercises

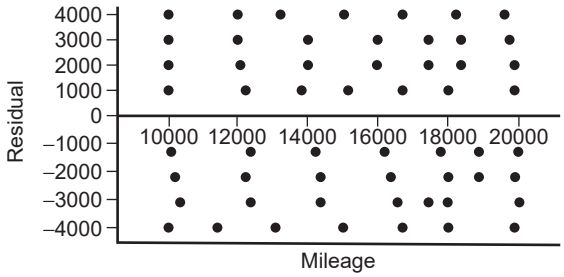
5. A scatterplot of y versus x for a dataset is given below. Which regression assumption is violated in this plot?



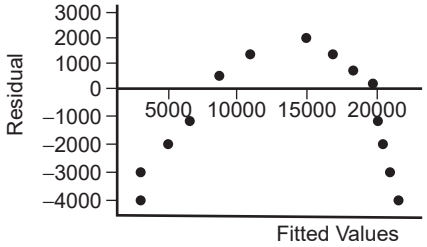
6. Based on the above plot, would you recommend fitting the regression model to predict the response given the predictor? Please explain.
7. A linear regression model was fitted to estimate the salary of an employee based on his/her experience. The plot of residuals of the regression model against the experience is given below. Which regression assumption is violated in this plot?



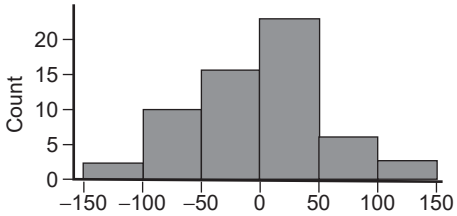
8. A regression model was fitted to predict the price of a used car using the mileage as the predictor. The plot of residuals of the regression model against the mileage is given below. State the regression assumption, if any, violated in this plot.



9. Observe the residuals vs. the fitted plot for the regression model of the price of a car against the age of the car. Is this model appropriate for predicting the price of the car using the age of the car? Explain.



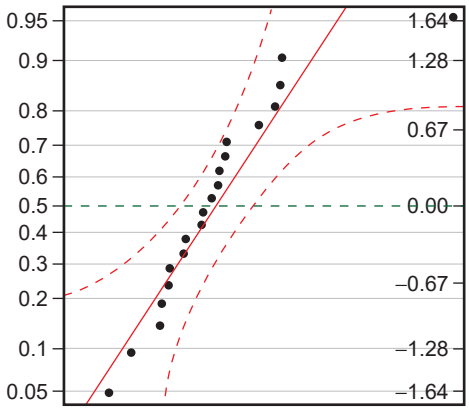
A simple regression model was fitted to estimate the credit score of customers based on their income. The histogram of residuals of the regression model is shown below. Use the histogram to answer the next two exercises.



10. Which assumption of the regression model can be checked using this plot?

11. Based on this plot, what can you say about the validity of the regression model?

A simple regression model was fitted to estimate the price of a used Honda Civic using the mileage as the predictor. The normal probability plot of regression residuals is shown below. Use this to answer the next two exercises.



12. Which assumption of the regression model can be checked using this plot?

13. Based on this plot, what can you say about the validity of the regression model?

14. Download the Pizza Delivery Data, which describes the relationship between Delivery Time (Minutes), the Number of Pizzas delivered, and the Distance (Miles). Use the data to answer the following questions.

a. Create a scatterplot of Delivery Time vs. Number of Pizzas. By examining the scatterplot, do you believe that the data follow a linear pattern?

b. Perform a residual analysis to check the assumptions of linearity, independence, normality, and equal variance. Are any of the assumptions violated? Justify your answer.

Data

The data can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Pizza Delivery Data**.

15. Download the Marathon Time Data, which has the finishing Marathon Times of 44 runners along with the total number of kilometers they run in training the 4 weeks prior to the race. Use the data to answer the following questions.

a. Create a scatterplot of Marathon Time vs. Km Run in 4 Weeks Prior. By examining the scatterplot, do you believe that the data follow a linear pattern?

b. Perform a residual analysis to check the assumptions of linearity, independence, normality, and equal variance. Are any of the assumptions violated? Justify your answer.

Data

The data can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Marathon Time**.

13.3 Evaluating the Fit of the Linear Regression Model

The goal in constructing most linear models is to use the independent variable, x, to explain or predict the dependent variable, y. The question we want to consider is, how much of the variation in y can be explained with the model? Before determining how much variation the model explains, it will be necessary to evaluate how much variability exists in the y-variable. This quantity is called the **total sum of squares (TSS)** and represents the total variation in the dependent variable, y.

Formula

Total Sum of Squares (TSS)

The total variation in y is given by the **total sum of squares (TSS)**.

$$\text{TSS} = \sum\left(y_i - \bar{y}\right)^2$$

If you think TSS looks a great deal like the numerator of the formula for the sample variance, you are right. TSS is the sum of the squared deviations about the mean of the dependent variable, y. If TSS were divided by $n - 1$ it would be the sample variance of y.

What is an Error?

An error $\left(y_i - \hat{y}_i\right)$ represents the model's inability to predict the variation in the dependent variable, y. If y didn't vary, for example if all y's were 6, its value would be easy to predict and the model's errors would all be zero. Adding all of the squared errors accumulates the total of all *unexplained* variation.

$$\text{SSE} = \sum\left(y_i - \hat{y}_i\right)^2$$