Chapter 9 - Multiple Regression

**Baby weights, Part I.** (9.1, p. 350) The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable *smoke* is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.

Estimate Std. Error t value Pr(*>|*t*|*) (Intercept) 123.05 0.65 189.60 0.0000

smoke -8.94 1.03 -8.65 0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don’t need to check linearity since the predictor has only two levels.)

1. Write the equation of the regression line.
2. Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.
3. Is there a statistically significant relationship between the average birth weight and smoking?

**ANSWERS**

1. In this context, the relationship between the variable X and the average birth weight (Y) of babies can be represented by the below equation:

Y = −8.94 X + 123.05 Y

where the slope (-8.94) indicates the rate at which average birth weight changes per unit increase in X, and the intercept (123.05) represents the initial estimated average birth weight when X is zero.

1. **For smokers:**

Y (smoker) = −8.94 × 1 + 123.05 = 114.

**For non-smokers:**

Y (non-smoker) = −8.94 × 0 + 123.05 = 123.05

This shows that smoking is associated with a decrease in average birth weight by 8.94 units compared to non-smokers.

1. When comparing the two groups, we observe an absolute difference of ∣114.11−123.05∣=8.94 ounces in the average birth weight of babies born to mothers who smoke versus those who don’t. Given that this study measures birth weight in ounces, a difference of 8.94 ounces is substantial.

**Absenteeism, Part I.** (9.4, p. 352) Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

eth sex lrn days

1 0 1 1 2

2 0 1 1 11

. . . . .

146 1 0 0 37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (eth: 0 - aboriginal, 1 - not aboriginal), sex (sex: 0 - female, 1 - male), and learner status (lrn: 0 - average learner, 1 - slow learner).

Estimate Std. Error t value Pr(*>|*t*|*) (Intercept) 18.93 2.57 7.37 0.0000

eth -9.11 2.60 -3.51 0.0000

sex 3.10 2.64 1.18 0.2411

lrn 2.15 2.65 0.81 0.4177

1. Write the equation of the regression line.
2. Interpret each one of the slopes in this context.
3. Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.
4. The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the *R*2 and the adjusted *R*2. Note that there are 146 observations in the data set.

**ANSWERS**

1. Y = 18.93 − 9.11 × ethnicity + 3.10 × gender + 2.15 × learning\_style.

In this equation, y is influenced by ethnicity, gender, and learning style, with coefficients indicating the respective impact of each variable on y.

1. **Ethnicity (eth):** If the individual is not Aboriginal, absenteeism decreases by 9.11 days.

**Gender (sex):** For male individuals, there is an increase of 3.10 days in absenteeism.

**Learning Style (lrn):** If the individual is classified as a slow learner, absenteeism rises by 2.15 days.

1. **(Given values)**

ethnicity = 0

gender = 1

learning\_style = 1

actual\_absences = 2

**(Calculating predicted absences)**

predicted\_absences = 18.93 - 9.11 \* ethnicity + 3.1 \* gender + 2.15 \* learning\_style

**(Calculating residual)**

residual = actual\_absences - predicted\_absences

residual = -22.18

The calculated residual is −22.18. This means the actual absences are 22.18 days lower than the predicted absences.

1. **(Given values)**

s = 146 - sample size

n = 3 - number of predictors

res\_var = 240.57 - residual variance

all\_var = 264.17 - total variance

**(Calculating R-squared)**

R2 = 1 - (res\_var / all\_var) = 0.08933641215883714

**(Calculating adjusted R-squared)**

adjusted\_R2 = 1 - (res\_var / all\_var) \* ((s - 1) / (s - n - 1)) = 0.0700970405847281

The calculated values are:

* R2=0.089 (indicating that approximately 8.9% of the variability in the data is explained by the model).
* Adjusted R2=0.070 (reflecting an adjusted proportion of variability explained, accounting for the predictors in the model). ​

​​

**Absenteeism, Part II.** (9.8, p. 357) Exercise above considers a model that predicts the number of days absent using three predictors: ethnic background (eth), gender (sex), and learner status (lrn). The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

Model Adjusted *R*2

* 1. Full model 0.0701
  2. No ethnicity -0.0033
  3. No sex 0.0676

4 No learner status 0.0723

Which, if any, variable should be removed from the model first?

**ANSWER**

The previous exercise yielded an adjusted R-squared of 0.0701. When "No learner status" is set to 0.0723, we see an improvement in R-squared, indicating that removing the learner status variable enhances the model's fit. Thus, the **learning style** (lrn) variable should be the first candidate for removal from the model.