

Prediction with Multiple Regression

For a specific set of values for the explanatory variables,

$$x_{1i}, x_{2i}, \dots, x_{Ki}$$

we can find the value of the response variable \hat{y} by plugging these values into the regression equation.

- If we are *estimating* the **mean of the response variable for the given values** $x_{1i}, x_{2i}, \dots, x_{Ki}$, we use the notation \hat{y}_m
- If we are *predicting* a **particular value of the response variable with explanatory variables** $x_{1i}, x_{2i}, \dots, x_{Ki}$, we use the notation \hat{y}_p

Confidence intervals for the mean and prediction intervals for the response variable can be obtained via SPSS as in Section 3.5

Example In the Meddicorp data, find a confidence interval for the conditional mean of the response variable given $x_1 = 500$ and $x_2 = 250$.

Example In the Meddicorp data, find a prediction interval for the response variable given $x_1 = 500$ and $x_2 = 250$.

Multicollinearity

Example Consider the data on Cal Ripkens RBIs, Homeruns, At Bats, and Hits during his career. This data is saved in the file

U:MT Student File Area/dgarth/stat378/calripken.sav

Test the overall fit of the regression. Use RBIs as the dependent variable.

Example (Continued) Now conduct hypotheses tests to determine whether each individual coefficient is zero.

Example (Continued) Now construct regression models for the RBIs with respect to each variable respectively.

Multicollinearity When linear relationships exist between the explanatory variables, a *potential* problem of multicollinearity exists, which may affect the regression analysis. This usually results in

1. Large t values for individual variables. Thus, we do not reject the hypotheses that these coefficients are zero. In other words, the model suggests that a particular variable is not important when in fact it is.
2. Unstable regression coefficients. Dropping one variable may result in large changes in the coefficients.

Detecting Multicollinearity There are many ways to do this. The book suggests several. We will focus on **Large F Statistics but Small t statistics**

- This is most effective if the F statistic is large, suggesting that the overall fit of the model is good, but *all* the individual t statistics are small, indicating that none of the individual variables are useful.

In Class Examples Do problem 10.

In Class Example Do problem 12