Multiple Regression (Hands-on)

Dr. Alex Marsella

2024-03-01

A few things to note about multiple regression

$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$ where k is the number of X variables.

- As you add additional X variables to a model, R-squared will always rise.
  - Adjusted R-squared will **only rise if the new variable you added has a statistically significant relationship to Y**.
  - Regular R-squared doesn't care about that, so it always increases as you add more variables.
- Interpreting individual variables in multiple regression is done "controlling for the other variables in the model".
  - We interpret the coefficient on X_1, **"controlling for" or "holding constant"** all other X_k in the model.
  - Calculus interpretation: Partial derivative of Y with respect to (whichever) X. Treats all other variables as constants.

Iris Data

I think I can try to predict Petal Length as a function of Sepal Length AND Sepal Width.

I theorize a relationship: $PetalLength = \beta_0 + \beta_1 SepalLength + \beta_2 SepalWidth + \epsilon$

```
# iris
```

Remember what our output tells us: $PredictedPetalLength = \hat{\beta}_0 + \hat{\beta}_1 SepalLength + \hat{\beta}_2 SepalWidth$

So then $PredictedPetalLength = -2.25 + 1.776 \times SepalLength - 1.339 \times SepalWidth$

Recall:

- We have a sample of 150 Irises, and from that, we can estimate β_1 with our estimate of the slope of the regression line: the sample estimator $\hat{\beta}_1$
- For those who took Calculus and not statistics, $\hat{\beta}_k$ is the derivative of Y with respect to X_k.
  - $PredictedPetalLength = \hat{\beta}_0 + \hat{\beta}_1 SepalLength + \hat{\beta}_2 SepalWidth$
  - How much does predicted petal length change as Sepal Length increases by 1, holding Sepal Width constant (controlling for Sepal Width)?
    - $\frac{\delta \hat{Y}}{\delta SepalLength} = \hat{\beta}_1$
  - How much does predicted petal length change as Sepal **Width** increases by 1, holding Sepal Length constant (controlling for Sepal Length)?
    - $\frac{\delta \hat{Y}}{\delta SepalWidth} = \hat{\beta}_2$

Motor Trend Car Road Tests

```
# mtcars regression
```

Medical Costs

Info contained here: https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download

```
# insurance
```

Exercise: Student Performance

- https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression

1. Estimate the the following multiple regression models.
    1. $Performance = \beta_0 + \beta_1 HoursStudied + \beta_2 PreviousScores + \epsilon$
    2. $Performance = \beta_0 + \beta_1 HoursStudied + \beta_2 PreviousScores + \beta_3 ExtraCurriculars + \epsilon$
    3. $Performance = \beta_0 + \beta_1 HoursStudied + \beta_2 PreviousScores + \beta_3 ExtraCurriculars + \beta_4 SleepHours + \beta_5 SampleQuestions + \epsilon$
2. For each of the three, interpret, in literal terms (referencing the variables)
    1. For the first model, interpret the regression coefficients.
    2. Compare the adjusted r-squared across models, does it increase or decrease as you add more variables? What does it mean?
    3. Interpret the F statistic for each model.
    4. For the third model, predict Performance score given:
        - Hours Studied = 7, Previous Score = 85, Extracurriculars = Yes, Sleephours = 8, Sample Questions = 2
        - Treat any insignificant coefficient as if it were equal to 0, since we failed to reject the null that $\beta_k = 0$ for that X_k.