Predictions I

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- The objective of inferential statistics is to draw conclusions about the population from the sample.
- So far we have used linear regression to draw conclusions about the nature of linear relationships in a population based on data.
- Next we are going to use linear regression to predict outcomes of combinations of explanatory variables.

Predictions

- A **prediction** (forecast) is a statement about a future event or unknown data.
- Can may use previous knowledge (statistical models) to make informed predictions.
- Sometimes referred to as *statistical learning*.

Review: Reading Models

- Understanding your results is extremely important.
- Visualizations can be useful to examine the relationships.
- Quantitative ways to read a model:
  1. **Read out the *model value*.**
  2. Characterize the relationship *described* by the model.

Review: Read out the Model Value

- Plug in specific values for the explanatory variables and read out the resulting model value.
- Essentially examining the fitted values for specific combinations of explanatory variables.
- Specific *point*, not a general description of the relationship.
- **Looking at these values for new values of X, sometimes denoted at X_***

Making Predictions

- Once the linear regression model is estimated predictions can be made.
- Predictions can be made on new combinations of values of explanatory variables.
- *Generally this should be done within the range of your sampled data.*
- Simple linear regression:

$$\widehat{E[Y]} = b_0 + b_i X_* \tag{1}$$

Making Predictions in R

- Create a data frame of different combinations of explanatory variables you are interested in predicting using your model (`model`).
- In R (predict m observations):
  - `new.data <- data.frame(XA = c(valueA1, valueA2, ..., valueAm), XB = c(valueB1, valueB2, ..., valueBm), ...)`
  - `predict(model,new.data)`

Example 1

- Run the given code to simulate some data.
- Estimate the linear regression model for $Y \sim 1 + X1$
- What do the coefficients and the Adjusted-R^2 say about the linear relationship?
- Use the `predict()` function to predict the Y values for the 20 simulated observations.
- Add these predicted observations to the scatterplot.
    - What do you notice?

Confidence Intervals for b_j

- Confidence intervals for our coefficient estimates can be obtained using R:
  - `confint(model, 'variable.name', level=0.95)`

$$b_j = \pm t_{\frac{\alpha}{2}, n-k} \cdot SE(b_j)$$

- The same information is used to calculate the p-value used to test the significance of the coefficient.

Example 2

- Use R to obtain a confidence interval for b_1 from Example 1.

Prediction Intervals

- Sometimes it may be important to assign confidence to your linear predictions.
- In general the prediction interval is larger than the confidence intervals.
- Now there are multiple sources of errors:
  1. Estimation Error
  2. Prediction Error
- Prediction intervals in R:
  - `predict(model,new.data, interval = ‘predict’)`

Example 3

- Use R to obtain a **prediction** interval for the predictions we made in Example 1.

Model Assumption Violations (If Violated:)

1. Linearity
   - **May lead to serious inaccuracies when making predictions.**
2. Normality (multivariate normal for multiple independent variables)
   - Causes problems in determining if model coefficients are significantly different from zero.
   - **Also causes problems in any confidence/prediction interval estimation.**
3. Homoscedasticity
   - As we are minimizing the residual sum of squares, extra *weight* may be given to observations with a higher variability during estimation.
   - **Also causes problems with prediction intervals.**
4. Independence
   - **May lead to bias (over/under estimate) the nature of the linear relationship.**

Example 4

- Load the *Football22.csv* data into R.
- Set your seed to 2020 and take a stratified sample of 75% with League as the stratifying variable.
- Using your stratified sample estimate the following linear regression model:

  Points $\sim$ 1 + Goals_For + Goals_Against
- Use the code provided to make predictions on the number of league points the clubs not in the stratified sample should earn.
  - Combine your results in a data frame.
  - How accurate are these predictions?
  - Examine the prediction intervals.

Exercise 1

- Take some time to make some predictions from other linear regression models we have estimated this term.
- Be sure to examine the prediction intervals.
  - Are they wide or narrow?

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/
2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- `predict()`
- `confint()`
- `anti_join()`