Multiple Linear Regression

UBCO MDS — DATA 570

The Multiple Linear Regression Model

- We assume that we can write the following model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Y is the random, quantitative response variable
- X_j is the j^{th} random predictor variable
- β_0 is the true intercept (unknown)
- β_j is the true j^{th} slope (unknown)
- ϵ is the true error

Estimates

- Notation for the estimates of Y, β_0, and β_j's follow the usual pattern (adding a "hat" ˆ).
- Calculations for $\hat{\beta}_j$ is messier than in the simple linear regression case. In this short module, we will focus on them...
- Common approach is still to use the least squares technique, minimizing $\sum(y_i - \hat{y}_i)^2$.

Interpretation

- How do we interpret, say $\hat{\beta}_2$?
- We interpret β_j as the average effect on Y of a one unit increase in X_j, *holding all other predictors fixed*
- This interpretation is not very useful in practice as predictors will often change together!

Example: Life Expectancy — SLR

- Data set contains Life Expectancy and Illiteracy rates for all 50 states from the 1970's.
- This dataset is available in R as `state.x77` in the `datasets` package
- We'll fit a simple linear regression with Life Expectancy as the response variable.

Example: Life Expectancy — SLR

```
Call:
lm(formula = stdata$Life.Exp ~ stdata$Illiteracy)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7169 -0.8063 -0.0349  0.7674  3.6675

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        72.3949     0.3383 213.973  < 2e-16 ***
stdata$Illiteracy  -1.2960     0.2570  -5.043 6.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.097 on 48 degrees of freedom
Multiple R-squared:  0.3463,Adjusted R-squared:  0.3327
F-statistic: 25.43 on 1 and 48 DF,  p-value: 6.969e-06
```

Example: Life Expectancy — MLR

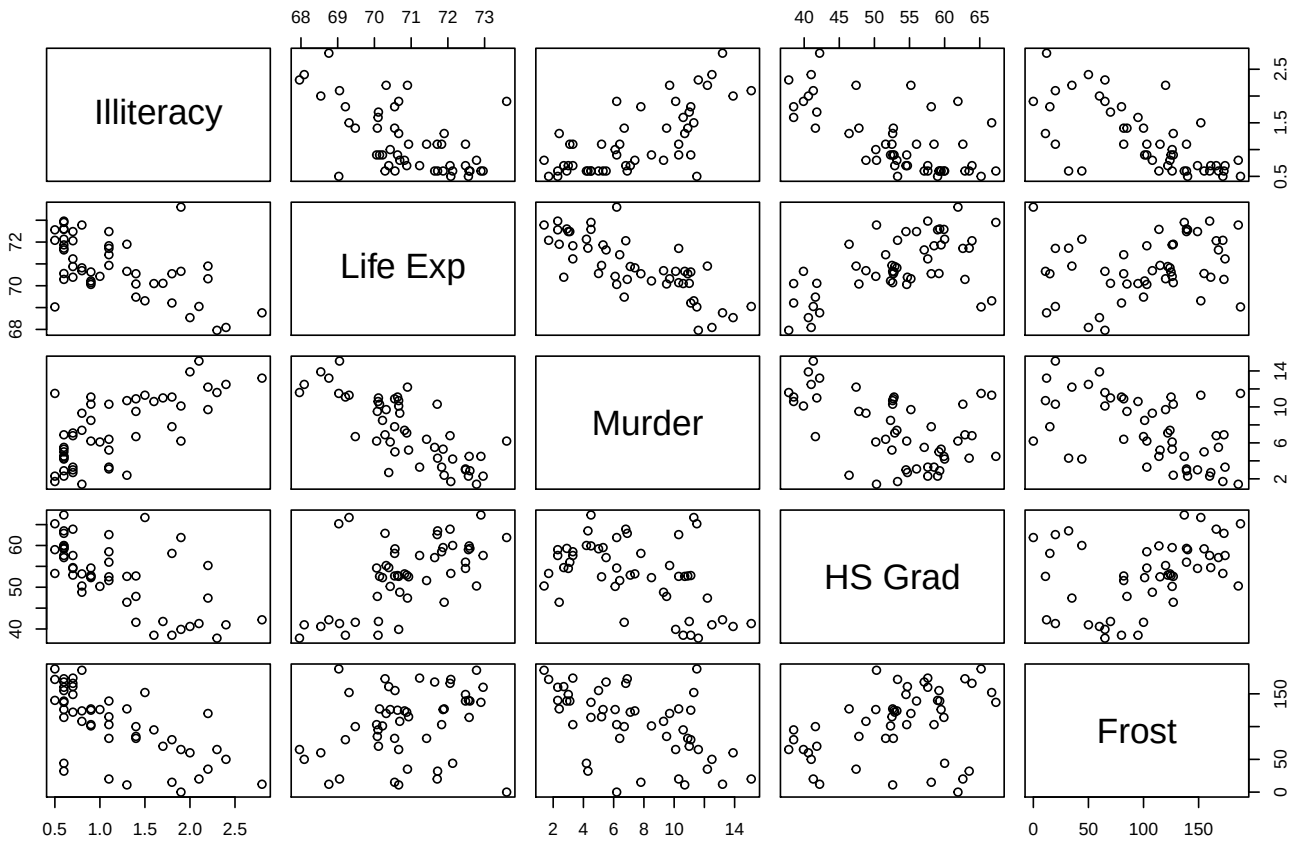
```
Call:
lm(formula = stdata$Life.Exp ~ stdata$Illiteracy + stdata$Murder +
    stdata$HS.Grad + stdata$Frost)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48906 -0.51040  0.09793  0.55193  1.33480

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.519958   1.320487  54.162  < 2e-16 ***
stdata$Illiteracy -0.181608   0.327846  -0.554  0.58236
stdata$Murder     -0.273118   0.041138  -6.639  3.5e-08 ***
stdata$HS.Grad     0.044970   0.017759   2.532  0.01490 *
stdata$Frost      -0.007678   0.002828  -2.715  0.00936 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7483 on 45 degrees of freedom
Multiple R-squared:  0.7146,Adjusted R-squared:  0.6892
F-statistic: 28.17 on 4 and 45 DF,  p-value: 9.547e-12
```

Example: Life Expectancy



All Predictors

- We want to know if our predictors are useful.
- Why should we avoid testing them all individually?
- To test all simultaneously amounts to assuming that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ and looking for evidence otherwise

Testing Model 'Usefulness'

- Hypotheses:
  H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
  H_a: not all β_j equal to 0

- Assumptions:
  Essentially the same as simple linear regression (linear relationships, independent and normally distributed errors with constant variance)

- Test statistic:

$$F = \frac{\frac{TSS - RSS}{p}}{\frac{RSS}{n-p-1}}$$

- Degrees of freedom (for comparing to $F^{\alpha}_{\nu_1,\nu_2}$):
  $\nu_1 = p$ and $\nu_2 = n - p - 1$

Output

- Each predictor has a t-statistic and p-value associated with it in the R output.
- These check whether each predictor is linearly related to the response, after adjusting for all the other predictors considered (hence the lack of stability when fitting different numbers of predictors).
- These are useful, but we have to be careful...especially for large numbers of predictors (because of the compounding error rates, we would expect to see variables meet our significance level threshold by chance).

Choosing Predictors

- There is a natural thought that probably arises when considering all these individualized hypothesis tests.
- If one (or more) of the variables does not appear significant, can't we just toss it out?
- This brings us into the topic of variable selection, which you will cover in more detail later in your program.
- As our first foray, we'll look at the most straightforward (and also oldest) techniques for doing so.
- First, we will need some way to measure the 'best model'...how? Assuming we have such a measure, what might be some options for finding variable subsets?

Measuring Model Fit

- We have discussed that RSE is a measure of the lack of fit and R^2 is a measure of goodness of fit.
- Note: $R^2 = Cor(Y, \hat{Y})^2$ (not quite the same as SLR). Why can we not use R^2 as a comparison of models with differing numbers of variables?
- R^2 will always increase by adding a predictor. In fact, once $n = p + 1$ then $R^2 = 1$.
- $RSE = \sqrt{RSS/(n - p - 1)}$ So RSE can increase when adding a useless predictor, but it more commonly will go down

Measuring Model Fit

- In summary, neither the RSE or R^2 are particularly useful when comparing different numbers of predictors!
- Where p is the number of variables in the model, some other options are...

Measuring Model Fit

- Adjusted $R^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}$, penalizes over-parameterization

  Larger=better

- $AIC = \frac{RSS + 2p\hat{\sigma}^2}{n\hat{\sigma}^2}$ is (approximately) 'Akaike's Information Criterion'.

  Smaller=better.

- $BIC = \frac{RSS + \log(n)p\hat{\sigma}^2}{n}$ is (approximately) the 'Bayesian Information Criterion'.

  Smaller=better.

Discussion

- You will learn alternative model fitting measures during other modules of your program

- Assuming we use one of these to measure how good a model is...what next? How do we start building models and measuring them? Discuss.

Choosing Predictors

- Forward Selection - Start with only the intercept (no predictors). Create p simple linear regressions, whichever predictor has the lowest RSS, add it into the model. Continue until, for instance, we see a decrease in our 'best model' measurement.

- Backward Selection - Start with all predictors. Remove the predictor with the largest p-value. Continue until, for instance, all variables are considered significant.

- Mixed Selection - Start with no predictors. Carry on with forward selection, but if any p-values for variables currently in the model reach above a threshold, remove that variable. Continue until, for instance, all variables in your model are significant and any additional variable would not be.

Further Note

- What happens if $p > n$?
- First: More coefficients β_j than observations n. In this case, our standard least squares estimates cannot be used!
- Second: F-statistic cannot be computed (for instance, $n - p - 1$ will be negative)!
- But this is a common problem with data nowadays! You will consider regularization approaches that get around this issue in DATA 571.

Types of Error

- Assuming the multiple linear model may not be true, we are introducing systematic bias by using it. We basically move forward with the assumption that our model is not overtly biased (since this is not measurable in realistic circumstances).
- Assuming the multiple linear model is (approximately) true, there will be error associated with building that model from a sample. Can think of it as differences between β_j and $\hat{\beta}_j$. We can use confidence intervals to estimate how far off we might be.
- If the multiple linear model was exactly f, there would still be irreducible error. That is, our predictions $\hat{y}_i \neq y_i$ for most (probabilistically all) i.
- Prediction intervals (where we expect to find y_i) must include both of these last two sources of error. Therefore, they must be larger than confidence intervals.


THE UNIVERSITY OF BRITISH COLUMBIA