IDS 702: Module 1.3

Model fitting and interpretation of coefficients

Dr. Olanrewaju Michael Akande

Back to our motivating example

Let's fit the following default MLR model to our Harris Trust and Savings Bank example using R.

$$\text{bsal}_i = \beta_0 + \beta_1\text{sex}_i + \beta_2\text{senior}_i + \beta_3\text{age}_i + \beta_4\text{educ}_i + \beta_5\text{exper}_i + \epsilon_i$$

We can estimate $\hat{\beta}$ in R directly as follows:

```
X <- model.matrix(~ sex + senior + age + educ + exper, data= wages)
y <- as.matrix(wages$bsal)
beta_hat <- solve(t(X)%*%X)%*%t(X)%*%y; beta_hat
```

```
##                     [,1]
## (Intercept) 6277.8933861
## sexFemale   -767.9126888
## senior       -22.5823029
## age            0.6309603
## educ          92.3060229
## exper          0.5006397
```

```
sigmasquared_hat <- t(y-X%*%beta_hat)%*%(y-X%*%beta_hat)/(nrow(X)-ncol(X))
SE_beta_hat <- sqrt(diag(c(sigmasquared_hat)*solve(t(X)%*%X))); SE_beta_hat
```

```
## (Intercept)   sexFemale      senior         age        educ       exper
## 652.2713190 128.9700022   5.2957316   0.7206541  24.8635404   1.0552624
```

IDS 702

Back to our motivating example

Let's fit the same MLR model using the lm command in R.

```
regwage <- lm(bsal~ sex + senior + age + educ + exper, data= wages)
summary(regwage)
```

```
##
## Call:
## lm(formula = bsal ~ sex + senior + age + educ + exper, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1217.36  -342.83   -55.61   297.10  1575.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6277.8934   652.2713   9.625 2.36e-15
## sexFemale   -767.9127   128.9700  -5.954 5.39e-08
## senior       -22.5823     5.2957  -4.264 5.08e-05
## age            0.6310     0.7207   0.876 0.383692
## educ          92.3060    24.8635   3.713 0.000361
## exper          0.5006     1.0553   0.474 0.636388
##
## Residual standard error: 508.1 on 87 degrees of freedom
## Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
## F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12
```

Interpretation of coefficients

- Each estimated slope is the amount y is expected to increase when the value of the corresponding predictor is increased by one unit, *holding the values of the other predictors constant*.

- For example, the estimated coefficient of educ is approximately 92.

  Interpretation: For each additional year of education for an employee, we expect baseline salary to increase by about $92, holding all other variables constant.

- That interpretation is a bit different when dealing with a binary variable (more generally, categorical/factor variables).

- For example, the estimated coefficient of sex (sexFemale) is approximately -768.

  Interpretation: For employees who started at the same time, had the same education and experience, and were the same age, women earned $768 less on average than men.

IDS 702

Which variable is the strongest predictor of the outcome?

- The coefficient that has the strongest linear association with the outcome variable is the one with the largest absolute value of T (referred to as t-value in the R output), the test statistic, which equals the coefficient over the corresponding SE.

- Note: T is NOT the size of the coefficient.

- The size of the coefficient is sensitive to scales of predictors, but T is not, since it is a standardized measure.

- Example: In our regression, seniority is a better predictor than education because it has a larger T.

IDS 702

Model fit

- How sure are we that this is actually a good model for this data?
- The easiest thing to do would be to look at the R-squared.
- R-squared has the same interpretation under both SLR and MLR, that is, the proportion of variation in the response variable, that is being explained by the regression fit.
- In this example, that proportion is approximately 52%. We will see if we can do better later.
- The adjusted R-squared is a modified version of R-squared that penalizes the original R-squared as extra variables are included in the model.
- In this example, we have approximately 48%, lower than the original 52%.
- We can do much better in assessing model fit, as we will see over the next few modules.

Centering

- How should we interpret the estimated intercept $\hat{\beta}_0 \approx 6278$?
- Generally speaking, we can say that the baseline salary for male employees, with zero age, zero seniority, zero education and zero experience is $6278.
- This is clearly not meaningful or realistic. Why?
- One way around this problem is centering. We can mean-center (can also scale if we want) continuous predictors to improve interpretation of the intercept.
- Centering does not really improve model fit, however it does help a lot with interpretability.

Centering

- So, for each continuous predictor,we will subtract its mean from every value, and use these mean centered predictors in our regression instead.
- The intercept can now be interpreted as the average value of Y at the average value of X, which is much more interpretable.
- Centering can be especially useful in models with interactions (which we are yet to explore).
- Centering can also help with multicollinearity (which we will also explore soon).
- Essentially, a transformed variable x_j^2 may be highly correlated with the untransformed counterpart x_j, which we want to avoid. Centering x_j before taking the square helps with that.
- Going forward, we will often mean center continuous predictors.

IDS 702

Centering

```
wages$agec <- c(scale(wages$age,scale=F))
wages$seniorc <- c(scale(wages$senior,scale=F))
wages$experc <- c(scale(wages$exper,scale=F))
wages$educc <- c(scale(wages$educ,scale=F))
regwagec <- lm(bsal~ sex + seniorc + agec + educc + experc, data= wages)
summary(regwagec)
```

```
##
## Call:
## lm(formula = bsal ~ sex + seniorc + agec + educc + experc, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1217.36  -342.83   -55.61   297.10  1575.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5924.0072    99.6588  59.443  < 2e-16
## sexFemale   -767.9127   128.9700  -5.954 5.39e-08
## seniorc      -22.5823     5.2957  -4.264 5.08e-05
## agec           0.6310     0.7207   0.876 0.383692
## educc         92.3060    24.8635   3.713 0.000361
## experc         0.5006     1.0553   0.474 0.636388
##
## Residual standard error: 508.1 on 87 degrees of freedom
## Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
## F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12
```

Centering

- Notice that the coefficients for the predictors have not changed but the intercept has changed.
- We interpret the intercept as the average baseline salary for male employees who are 474 months old, have 82 months of seniority, 12.5 years of education, and 101 months of experience.

```
colMeans(wages[,c("age","senior","educ","exper")])
```

```
##       age    senior      educ     exper
## 474.39785  82.27957  12.50538 100.92742
```

- Much more meaningful!

IDS 702

Some notes

- We can't say for sure that our model has not violated any of the assumptions. We must do model assessment just as with SLR.

- We will address these issues and more over the next few modules.

- Be very wary of extrapolation! Because there are several predictors, you can fall into the extrapolation trap in many ways.

  What do we mean by extrapolation?

- Finally, note that multiple regression shows association.

  It does NOT prove causality.

  Only a carefully designed observational study or randomized experiment or good causal inference methods can help show causality.

IDS 702

What's next?

Move on to the readings for the next module!

IDS 702