POL 345: Quantitative Analysis and Politics

Precept Handout 10

Week 12 (Verzani Chapter 10: 10.3)

Remember to complete the entire handout before precept. In this handout, we cover the following new materials:

- Using `summary()` to obtain the summary statistics of a linear regression model.
- Using `coef()` to obtain coefficients from a linear regression model.
- Using `predict()` to generate predicted values for the outcome variable based on an estimated linear regression model.
- Using `lm()` to estimate a linear multiple regression model, `summary()` to obtain the summary statistics of the model, and `coef()` to obtain the model coefficients.

1 Interpreting Regression Outputs

Just as we do with differences in means, we want to properly describe the uncertainty of the estimates produced via regression analysis. Recall that the simple regression model is given by:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

The estimated model is given by:

$$\widehat{Y_i} = \hat{\alpha} + \hat{\beta} X_i$$

where $\hat{\alpha}$ and $\hat{\beta}$ represent the estimated intercept and slope, respectively.

- `summary(model)` provides the summary statistics of the model. In particular, the following statistics are important
  - `Estimate`: point estimate of each coefficient
  - `Std. Error`: standard error of each estimate
  - `t value`: indicates the t-statistic of each coefficient under the null hypothesis that it equals zero
  - `Pr(>|t|)`: indicates the two-sided p-value corresponding to this t-statistic where asterisks indicate the level of statistical significance.
  - `Multiple R-squared`: The coefficient of determination
  - `Adjusted R-squared`: The coefficient of determination adjusting for the degrees of freedom
- **Resource Curse:** Beginning in the 1980's, a series of Political Science and Economics scholars began to explore what Richard Auty termed the *resource curse*: that countries with an abundance of natural resources experience poor economic conditions, as compared to countries with few natural resources. Oil, in particular, has been cited as a natural resource for which the resource curse is particularly pronounced. Alternatively, scholars have argued that it is not natural resources, but rather ethnic fractionalization that is associated with poor economic conditions. Many countries with an abundance of natural resources also have high levels of ethnic fractionalization. We explore these relationships using data from Ross's article "Does Oil Hinder Democracy?" (*World Politics*, 2001). The variables in the data are as follows:
  - `country`: country name
  - `year`: year
  - `gdppc`: Real GDP per capita, in international dollars
  - `ELF`: ethnic fractionalization index
  - `oil`: oil exports, as a percentage of GDP

  We begin by first exploring whether high levels of ethnic fractionalization are associated with low levels of log GDP per capita.

```
> load("curse.RData")
> ## Regress gdppc
> fit.1 <- lm(log(gdppc) ~ ELF, data = rescurse)
> summary(fit.1)
```

```
Call:
lm(formula = log(gdppc) ~ ELF, data = rescurse)

Residuals:
   Min     1Q Median     3Q    Max
-4.806 -1.636 -0.105  1.399  5.979

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.87209    0.07676   324.0   <2e-16 ***
ELF         -0.02336    0.00165   -14.1   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.08 on 1999 degrees of freedom
Multiple R-squared: 0.0909,        Adjusted R-squared: 0.0904
F-statistic:  200 on 1 and 1999 DF,  p-value: <2e-16
```

On average, a one unit increase ethnic fractionalization is associated with a 2.3% decrease in GDP per capita, all else equal. As shown by the small p-value, the estimated effect is statistically significant. This result supports the hypothesis that on average, higher levels of ethnic fractionalization are associated with lower levels of GDP per capita.

- We can construct confidence intervals of a estimated coefficient using the output. The point estimate can be obtained by `coef()` function. You can either base confidence intervals upon the normal approximation or the Student t distribution where the degrees of freedom are given in the output.

```
> est <- coef(fit.1)[2]
> ## normal approximation
> MoE.norm <- qnorm(0.975) * 0.001652
> ## t-distribution
> MoE.t <- qt(0.975, df = 1999)*0.001652
> low.norm <- est - MoE.norm
> high.norm <- est + MoE.norm
> low.norm; high.norm

       ELF
-0.0265958

       ELF
-0.0201201

> low.t <- est - MoE.t
> high.t <- est + MoE.t
> low.t; high.t

       ELF
-0.0265977
```

```
       ELF
-0.0201181
```

We observe that because of a large sample size the two confidence intervals are essentially identical. Over repeated sampling under the assumed model, 95% of time the true value of β falls in the constructed confidence interval. Again, this supports the hypothesis that on average, higher levels of ethnic fractionalization are associated with lower levels of GDP per capita.

2 Predictive Inference

Given our regression results, we often want to generate predicted values based on our inputs, which are given by,

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- To do so we'll need to use `predict(model, newdata, interval, level)`, where
  - `model` is an output of the `lm()` function
  - `newdata` is an optional dataframe of new values. If `newdata` is not provided, `predict()` will use the original data values supplied to the `lm()` function
  - Specifying `interval = "confidence"` will add confidence intervals to the output associated with the *expected value*, i.e., $\widehat{Y}_i$, while specifying `interval = "prediction"` will add confidence intervals to the output associated with the *predicted value*, i.e., $\widehat{Y}_i + \epsilon_i$;
  - `level` specifies the $1 - \alpha$ width for the confidence interval.

```
> new.obs <- data.frame(ELF = c(0, 31, 89)) ## min, median, max
> ## expected values
> predict(fit.1, new.obs, interval = "confidence", level = 0.95)

      fit     lwr     upr
1 24.8721 24.7216 25.0226
2 24.1480 24.0547 24.2413
3 22.7932 22.6014 22.9850

> ## Same point estimates but wider intervals for predicted values
> predict(fit.1, new.obs, interval = "prediction", level = 0.95)

      fit     lwr     upr
1 24.8721 20.7855 28.9587
2 24.1480 20.0631 28.2329
3 22.7932 18.7049 26.8815
```

3 Multiple Regression

Multiple regression is a generalized form of simple regression. Multiple linear regression allows for more than one explanatory variable to model the mean value of the dependent variable of interest. For example, when there are two explanatory variables, the model take the form of:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

This enables the assessment of the multivariate relationship between the outcome variable and multiple explanatory variables. The coefficient of determination also has the same interpretation as in simple regression; it represents the proportion of the original variance explained by all the explanatory variables in the model. The interpretation of additional explanatory variables is the same as in simple regression, with the assumption that all other variables are held constant.

- To fit a multiple regression model in **R**, we again use the `lm()` function. For example, with two explanatory variables `x1` and `x2`, we use the syntax `lm(y ~ x1 + x2 , data = mydata)`.
- **The Resource Curse Example:** We return to the resource curse example, now adding oil exports to the model. For the ELF, our null hypothesis remains $H_0 : \beta_1 = 0$, and our alternative hypothesis remains $H_1 : \beta_1 < 0$. Similarly, given the resource curse theory, for oil, our null hypothesis is $H_0 : \beta_2 = 0$, and our alternative hypothesis is $H_1 : \beta_2 < 0$.

```
> fit.2 <- lm(log(gdppc) ~ ELF + oil, data = rescurse)
> summary(fit.2)

Call:
lm(formula = log(gdppc) ~ ELF + oil, data = rescurse)

Residuals:
   Min     1Q Median     3Q    Max
-4.818 -1.646 -0.105  1.390  5.968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.88398    0.07875  315.99   <2e-16 ***
ELF         -0.02337    0.00165  -14.14   <2e-16 ***
oil         -0.00306    0.00453   -0.68      0.5
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.08 on 1998 degrees of freedom
Multiple R-squared: 0.0911,        Adjusted R-squared: 0.0902
F-statistic:  100 on 2 and 1998 DF,  p-value: <2e-16
```

The model suggests that, controlling for oil, on average, a one unit increase in the ELF index is on average associated with a 2.3% decrease in log GDP per capita, all else equal. Controlling for ethnic fractionalization, we find that oil exports do not have a statistically significant association

with a lower GDP per capita; we fail to reject the null hypothesis that oil has no association with the outcome variable on average after controlling for the ELF index. This result suggests that the negative association between oil exports and GDP per capita is weak once we control for ethnic fractionalization. Adding the oil variable to the model increases its explanatory power by only a modest amount, as indicated by the small difference in the coefficient of determination.

- As with simple regression, we can construct confidence intervals of a coefficient. The 95% confidence interval of the estimated coefficients for the ELF index may be computed as follows. Below, we show the confidence interval based on Student's t distribution, but it is also possible to use normal approximation.

```
> est <- coef(fit.2)[2]
> MoE <- qt(0.975, df = 1998)*0.001653
> ELF.low <- est - MoE; ELF.high <- est + MoE
> c(ELF.low, ELF.high) ## 95% CI

       ELF        ELF
-0.0266116 -0.0201280
```