

PLSC 503 – Spring 2017

Multivariate Regression II

February 16, 2017

Inference, In General

- Pick some $\mathbf{H}_A : \boldsymbol{\beta} = \boldsymbol{\beta}_A$
- Estimate $\hat{\boldsymbol{\beta}}$
- Determine distribution of $\hat{\boldsymbol{\beta}}$ under \mathbf{H}_A
- Form a *test statistic* $\hat{\mathbf{S}} = h(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$
- Assess $\Pr(\hat{\mathbf{S}}|\mathbf{H}_A)$

The Importance of $\mathbf{V}(\hat{\beta})$

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= E[\hat{\beta} - E(\hat{\beta})]^2 \\ &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\}\end{aligned}$$

Rewrite:

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

The Importance of $\mathbf{V}(\hat{\beta})$

Taking expectations:

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Estimating $\mathbf{V}(\hat{\beta})$

Empirical estimate:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{N - K}$$

Yields:

$$\widehat{\mathbf{V}(\hat{\beta})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Single Coefficient Hypothesis Tests

$$\hat{\beta} \sim \mathcal{N}[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$$

In practice, using $\hat{\sigma}^2$ means

$$\hat{\beta} - \beta \sim \mathbf{t}_{N-K}$$

Procedure:

- Choose a value of β_k that you want to test (say, $\beta_k = 0$),
- Calculate the t -statistic for the coefficient associated with X_k , which is:

$$\hat{t}_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\mathbf{v}(\hat{\beta}_k)}}$$

- Compare \hat{t}_k to a t distribution with $N - K$ degrees of freedom.

Multivariate Hypothesis Testing

E.g.: $H_0 : \beta_1 = \beta_2 = \dots = \beta_K = 0$

or: $H_0 : \beta_3 = \beta_6 = 0$

Generally: *Linear restrictions*:

$$\underset{q \times k}{\mathbf{R}} \underset{k \times 1}{\boldsymbol{\beta}} = \underset{q \times 1}{\mathbf{r}}$$

E.g.:

$$\beta_2 = -2 \iff (0 \ 1 \ 0) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = -2$$

Recall:

$$\mathbf{TSS} = \mathbf{MSS} + \mathbf{RSS}$$

Consider:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{Ui}$$

and the restriction:

$$H_a : \beta_2 = \beta_4 = 0.$$

Restricted model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + 0X_{2i} + \beta_3 X_{3i} + 0X_{4i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + u_{Ri} \end{aligned}$$

F-tests: Sums of Squared Residuals

“Unrestricted”:

$$\text{RSS}_U \equiv \hat{\mathbf{u}}_U' \hat{\mathbf{u}}_U = \sum_{i=1}^N \hat{u}_{Ui}^2$$

“Restricted”:

$$\text{RSS}_R \equiv \hat{\mathbf{u}}_R' \hat{\mathbf{u}}_R = \sum_{i=1}^N \hat{u}_{Ri}^2$$

F-statistic:

$$\begin{aligned}\mathbf{F} &= \frac{(\text{RSS}_R - \text{RSS}_U)/q}{\text{RSS}_U/(N - K)} \\ &= \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(N - K)}\end{aligned}$$

Testing:

$$\mathbf{F} \sim F_{q, N-K}$$

F-Test: Example

Consider:

$$\begin{aligned}H_b : \quad & \beta_1 + \beta_4 = 1 \\ & \beta_1 = 1 - \beta_4\end{aligned}$$

Implies:

$$\begin{aligned}Y_i &= \beta_0 + (1 - \beta_4)X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + u_{R'i} \\ &= \beta_0 + X_{1i} - \beta_4X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + u_{R'i} \\ &= \beta_0 + X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4(X_{4i} - X_{1i}) + u_{R'i}\end{aligned}$$

implying restricted model:

$$Y_i - X_{1i} = \beta_0 + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4(X_{4i} - X_{1i}) + u_{R'i}$$

Confidence Regions

$$F = \frac{(\hat{\beta}_q - \beta_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\beta}_q - \beta_q^H)}{q\hat{\sigma}^2}$$

Implies:

$$\Pr \left[\frac{(\hat{\beta}_q - \beta_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\beta}_q - \beta_q^H)}{q\hat{\sigma}^2} \leq F_{q, N-K} \right] = 1 - \alpha. \quad (1)$$

→ “confidence region” of all points satisfying:

$$(\hat{\beta}_q - \beta_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\beta}_q - \beta_q^H) \leq q\hat{\sigma}^2 F_{q, N-K}.$$

Multivariate Prediction

$$\hat{Y}_j = \mathbf{x}_j \hat{\beta}$$

Variance:

$$\widehat{\mathbf{V}}(\hat{Y}_j) = \hat{\sigma}^2 [1 + \mathbf{x}_j (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j']$$

Standard error:

$$\widehat{\text{s.e.}}(\hat{Y}_j) = \sqrt{\hat{\sigma}^2 [1 + \mathbf{x}_j (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j']}$$

Example: Africa Data

```
> library(RCurl)
> temp<-getURL("https://raw.githubusercontent.com/PrisonRodeo/PLSC503-Spring-2016-git/master/Data/africa2016.csv")
> Data<-read.csv(text=temp, header=TRUE)
> Data<-with(Data, data.frame(adrate, polity,
+                             subsaharan=as.numeric(subsaharan), muslperc, literacy))

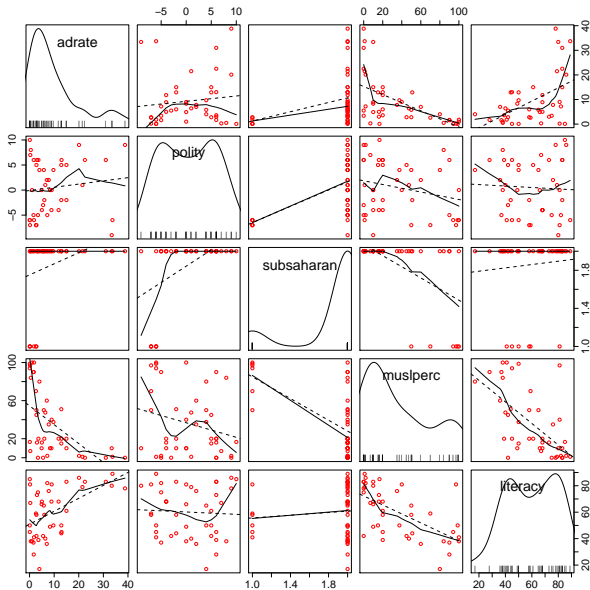
> summary(Data)
```

| adrate | polity | subsaharan | muslperc | literacy |
|-----------------|------------------|---------------|----------------|----------------|
| Min. : 0.100 | Min. : -9.0000 | Min. : 1.00 | Min. : 0.00 | Min. : 17.00 |
| 1st Qu.: 2.700 | 1st Qu.: -4.5000 | 1st Qu.: 2.00 | 1st Qu.: 10.00 | 1st Qu.: 43.00 |
| Median : 6.000 | Median : 0.0000 | Median : 2.00 | Median : 20.00 | Median : 61.00 |
| Mean : 9.365 | Mean : 0.5116 | Mean : 1.86 | Mean : 35.96 | Mean : 60.07 |
| 3rd Qu.: 12.900 | 3rd Qu.: 5.5000 | 3rd Qu.: 2.00 | 3rd Qu.: 55.50 | 3rd Qu.: 78.50 |
| Max. : 38.800 | Max. : 10.0000 | Max. : 2.00 | Max. : 100.00 | Max. : 89.00 |

```
> cor(Data)
```

| | adrate | polity | subsaharan | muslperc | literacy |
|------------|------------|-------------|-------------|------------|-------------|
| adrate | 1.0000000 | 0.11794182 | 0.33129420 | -0.5709233 | 0.51489444 |
| polity | 0.1179418 | 1.00000000 | 0.52819844 | -0.2391715 | -0.05079354 |
| subsaharan | 0.3312942 | 0.52819844 | 1.00000000 | -0.5772513 | 0.09472968 |
| muslperc | -0.5709233 | -0.23917151 | -0.57725134 | 1.0000000 | -0.61960385 |
| literacy | 0.5148944 | -0.05079354 | 0.09472968 | -0.6196039 | 1.00000000 |

Africa Data



A Regression

```
> model<-lm(adrate~polity+subsaharan+muslperc+literacy,data=Data)
> summary(model)
```

Call:

```
lm(formula = adrate ~ polity + subsaharan + muslperc + literacy,
    data = Data)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|---------|---------|--------|---------|
| -15.4681 | -4.3947 | -0.5251 | 3.4246 | 22.9358 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -4.39843 | 14.94744 | -0.294 | 0.7702 |
| polity | -0.01390 | 0.27969 | -0.050 | 0.9606 |
| subsaharan | 3.72969 | 5.43093 | 0.687 | 0.4964 |
| muslperc | -0.08689 | 0.06282 | -1.383 | 0.1747 |
| literacy | 0.16575 | 0.09433 | 1.757 | 0.0869 . |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.264 on 38 degrees of freedom

Multiple R-squared: 0.3771, Adjusted R-squared: 0.3115

F-statistic: 5.751 on 4 and 38 DF, p-value: 0.001013

Variance-Covariance Matrix of $\hat{\beta}$

```
> options(digits=4)
> vcov(model)
```

| | (Intercept) | polity | subsaharan | muslperc | literacy |
|-------------|-------------|-----------|------------|-----------|-----------|
| (Intercept) | 223.4259 | 1.088030 | -72.2628 | -0.771309 | -1.002421 |
| polity | 1.0880 | 0.078229 | -0.6642 | -0.000293 | 0.001968 |
| subsaharan | -72.2628 | -0.664212 | 29.4950 | 0.206067 | 0.171765 |
| muslperc | -0.7713 | -0.000293 | 0.2061 | 0.003946 | 0.004098 |
| literacy | -1.0024 | 0.001968 | 0.1718 | 0.004098 | 0.008898 |

Test $H_0 : \beta_{\text{polity}} = \beta_{\text{subsaharan}} = 0$:

```
> library(lmtest)
> modelsmall<-lm(adrate~muslperc+literacy,data=Data)
> waldtest(model,modelsmall)
```

Wald test

Model 1: adrate ~ polity + subsaharan + muslperc + literacy

Model 2: adrate ~ muslperc + literacy

| | Res.Df | Df | F | Pr(>F) |
|---|--------|----|------|--------|
| 1 | 38 | | | |
| 2 | 40 | -2 | 0.27 | 0.76 |

More tests...

Test $H_0 : \beta_{\text{muslperc}} = 0.1$:

```
> library(car)
> linearHypothesis(model,"muslperc=0.1")
```

Linear hypothesis test

Hypothesis:
muslperc = 0.1

Model 1: restricted model

Model 2: adrate ~ polity + subsaharan + muslperc + literacy

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|-----|--------|------|----|-----------|------|-----------|
| 1 | 39 | 3200 | | | | |
| 2 | 38 | 2595 | 1 | 605 | 8.85 | 0.0051 ** |
| --- | | | | | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

More tests...

Test $H_0 : \beta_{\text{literacy}} = \beta_{\text{muslperc}}$:

```
> linearHypothesis(model,"literacy=muslperc")
```

Linear hypothesis test

Hypothesis:

- muslperc + literacy = 0

Model 1: restricted model

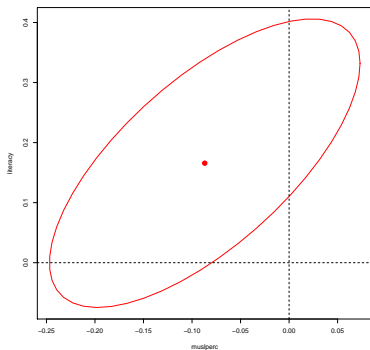
Model 2: adrate ~ polity + subsaharan + muslperc + literacy

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|------|----|-----------|------|-------------|
| 1 | 39 | 3534 | | | | |
| 2 | 38 | 2595 | 1 | 938 | 13.7 | 0.00067 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Confidence Regions / Ellipses

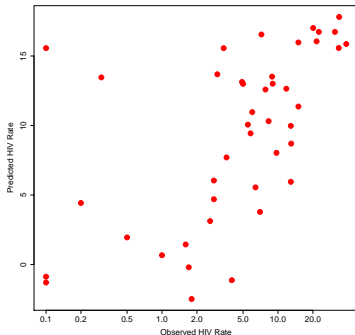
```
> confidenceEllipse(model=model,which.coef=c(4,5),  
                    xlab="Muslim Percentage",ylab="Literacy")  
> abline(h=0,v=0,lty=2)
```



Predicted Values

```
> hats<-fitted(model)
> # Or, alternatively:
> fitted<-predict(model,se.fit=TRUE, interval=c("confidence"))
> scatterplot(model$fitted~adrate,log="x",smooth=FALSE,boxplots=FALSE,
  reg.line=FALSE,xlab="Observed HIV Rate",ylab="Predicted HIV Rate",
  pch=16,cex=2)
```

Predicted and Actual HIV/AIDS Rates (X-Axis Logged)



An Even More Useful Plot

```
> library(plotrix)
> plotCI(Data$adrate,model$fitted,uiw=(1.96*(fitted$se.fit)),
         log="x",xlab="Observed HIV Rate",ylab="Predicted HIV Rate")
> lines(lowess(Data$adrate,Data$adrate),lwd=2)
```

Predicted and Actual HIV/AIDS Rates, with 95% C.I.s

