Soci708 – Statistics for Sociologists
Module 11 – Multiple Regression

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

1Adapted from slides for the course Quantitative Methods in Sociology (Sociology 6Z3) taught at McMaster University by Robert Andersen (now at University of Toronto)
Goals of This Module

▶ Review of *least-squares regression* analysis
  ▶ Simple and multiple regression
  ▶ Slope, intercept and $R^2$
  ▶ Standard error of the regression
▶ Inference for Regression
  ▶ Confidence intervals and hypothesis tests for the slope
  ▶ F-test for the entire regression model
▶ Assumptions of regression and how to check them
Inference in Multiple Regression (1)

- Recall that least-squares regression fits the equation:

\[
\hat{Y} = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k
\]

finding the values of \(A, B_1, B_2, \ldots, B_k\) that minimize the sum of the squared residuals:

\[
\sum \text{residual}^2 = \sum E_i^2 = \sum (y - \hat{y})^2
\]

- The slopes are now partial slopes (versus marginal slopes in simple regression)
  - The slope coefficient \(B_1\) represents the average change in \(y\) associated with a one-unit increase in \(x_1\), holding the other \(x\)'s constant
Inference in Multiple Regression (1)

- At this point explain again meaning of $B_k$s in context of an example
- Remember standardized coefficients
- Explain $R^2$ and how it relates to correlation between $\hat{Y}$ and $Y$
Inference in Multiple Regression (2)

- The standard deviation of $y$ around the regression surface (standard error of the regression) is again simply estimated:

$$SE = \sqrt{\sum E_i^2 / (n - k - 1)}$$

- There are $n - k - 1$ degrees of freedom, where $n$ is the sample size and $k$ is the number of slopes to estimate

- As in the simple regression case, $S_E$ is used to find the standard errors for each slope and CI and hypothesis are tested in exactly the same way
  - We will not go into details regarding this here because the formulas are complicated. A computer program will calculated the $SE$’s for us

- It is important to note, however, that we now use the $t$-distribution with $n - k - 1$ df
Multiple Regression in R

```
> # in R
> ed<-c(12,13,12,14,12,15,12)
> age<-c(45,35,27,60,23,30,49)
> income<-c(20000,22000,23000,25000,18000,30000,26000)
> reg.model<-lm(income~ed+age)
> summary(reg.model)

Coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8374.74   14195.13  -0.590   0.5869
   ed       2354.38    1103.60   2.133   0.0998 .
   age       39.88     100.33   0.398   0.7113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3248 on 4 degrees of freedom
Multiple R-Squared: 0.5591,   Adjusted R-squared: 0.3386
F-statistic: 2.536 on 2 and 4 DF,  p-value: 0.1944
```
Analysis of Variance and F Test (1)

- Recall that for the (simple or multiple) regression model
  \[(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\]

- It is an algebraic fact (that can be demonstrated) that the equality holds after one squares the deviations and sum them:
  \[\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2\]

- This can be written
  \[SST = SSM + SSE\]

where

\[SST = \sum (y_i - \bar{y})^2 \quad \text{(Sum of Squares Total)}\]
\[SSM = \sum (\hat{y}_i - \bar{y})^2 \quad \text{(Sum of Squares Model)}\]
\[SSE = \sum (y_i - \hat{y}_i)^2 \quad \text{(Sum of Squares Error)}\]
Analysis of Variance and F Test (2)

▸ Recall that to calculate $s_y^2$, the sample variance of $y$, we divide $SST$ by its degrees of freedom $(n - 1)$; likewise to calculate $s^2$, the error variance, we divide $SSE$ by its degrees of freedom $(n - k - 1)$.

▸ Each sum of squares has its degrees of freedom $DF$, and these add up to the total degrees of freedom $DFT = (n - 1)$:

\[
DFT = DFM + DFE \\
(n - 1) = k + (n - k - 1)
\]

▸ For each source of variation the mean square $MS$ is calculated as

\[
MS = \frac{\text{sum of squares}}{\text{degrees of freedom}}
\]
Analysis of Variance and F Test (3)

The three mean squares are thus:

\[ s^2_y = MST = \frac{SST}{DFT} = \frac{\sum (y_i - \bar{y})^2}{n - 1} \]

\[ MSM = \frac{SSM}{DFM} = \frac{\sum (\hat{y}_i - \bar{y})^2}{k} \]

\[ s^2 = MSE = \frac{SSE}{DFE} = \frac{\sum (y_i - \hat{y}_i)^2}{n - k - 1} \]

The null hypothesis that \( y \) is not linearly related to the \( x \) variables may be tested by comparing \( MSM \) and \( MSE \) using the \( F \) statistic

\[ F = \frac{MSM}{MSE} \]

Under the null hypothesis \( H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \) \( F \) is distributed as an \( F(k, n - k - 1) \) distribution
Analysis of Variance and F Test (4)

**ANALYSIS OF VARIANCE F TEST**

In the multiple regression model, the hypothesis

\[ H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \]

is tested by the analysis of variance F statistic

\[ F = \frac{\text{MSM}}{\text{MSE}} \]

The P-value is the probability that a random variable having the \( F(p, n-p-1) \) distribution is greater than or equal to the calculated value of the F statistic.

- A small P-value favors the alternative hypothesis\(^2\)

\[ H_a : \text{at least one of the } \beta_j \text{ is not 0} \]

\(^2\)From Moore & McCabe (2006, p.689)
### Analysis of Variance and F Test (5)

The Analysis of Variance (ANOVA) Table

The ANOVA table summarizes sums of squares, mean squares and the $F$ test

<table>
<thead>
<tr>
<th>Source of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$SSM = \sum (\hat{y}_i - \bar{y})^2$</td>
<td>$SSM/DFM$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = \sum (y_i - \hat{y}_i)^2$</td>
<td>$SSE/DFE$</td>
</tr>
<tr>
<td>Total</td>
<td>$SST = \sum (y_i - \bar{y})^2$</td>
<td>$SST/DFT$</td>
</tr>
</tbody>
</table>
Analysis of Variance and F Test (6)
Stata output with ANOVA table for regression of depression score

. * in Stata
  . reg total l10inc age female cath jewi none

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2686.95106</td>
<td>6</td>
<td>447.825177</td>
<td>F( 6, 249) = 6.35</td>
</tr>
<tr>
<td>Residual</td>
<td>17554.7989</td>
<td>249</td>
<td>70.5012006</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>20241.75</td>
<td>255</td>
<td>79.3794118</td>
<td>Adj R-squared = 0.1118</td>
</tr>
</tbody>
</table>

| total | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-------|-----|----------------------|
| l10inc  | -6.946661 | 1.72227 | -4.03 | 0.000 | -10.33874 -3.554586 |
| age     | -.074917  | .0334132 | -2.24 | 0.026 | -.1407254 -.0091085 |
| female  | 2.559048  | 1.095151 | 2.34  | 0.020 | .402108 4.715987 |
| cath    | .7527141  | 1.440845 | 0.52  | 0.602 | -2.085084 3.590512 |
| jewi    | 4.674959  | 1.917259 | 2.44  | 0.015 | .8988472 8.451071 |
| none    | 3.264667  | 1.400731 | 2.33  | 0.021 | .5058747 6.023459 |
| _cons   | 18.24958  | 2.971443 | 6.14  | 0.000 | 12.39721 24.10195 |
Analysis of Variance and F Test (7)

Stata output with ANOVA table for regression of depression score

. * in Stata
. reg total l10inc age female cath jewi none

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2686.95106</td>
<td>6</td>
<td>447.825177</td>
</tr>
<tr>
<td>Residual</td>
<td>17554.7989</td>
<td>249</td>
<td>70.5012006</td>
</tr>
<tr>
<td>Total</td>
<td>20241.75</td>
<td>255</td>
<td>79.3794118</td>
</tr>
</tbody>
</table>

Number of obs = 256
F( 6, 249) = 6.35
Prob > F = 0.0000
R-squared = 0.1327
Adj R-squared = 0.1118
Root MSE = 8.3965

Model = \[ SSM = \sum (\hat{y}_i - \bar{y})^2 \]
Residual = \[ SSE = \sum (y_i - \hat{y}_i)^2 \]
Total = \[ SST = \sum (y_i - \bar{y})^2 \]
MS = SS/Df...

What is SST/DFT = 79.379?

\[ F(6, 249) = \frac{MSM}{MSE} \]
\[ R\text{-squared} = \frac{SSM}{SST} \]
\[ Rooth \text{ MSE} = \sqrt{MSE} \]
Analysis of Variance and F Test (8)

Exercise: Can you recover the redacted figures?

. * in Stata
  . reg total l10inc age female cath jewi none

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2686.95106</td>
<td>xxx</td>
<td>447.825177</td>
<td>F( 6, 249) = xxxx</td>
</tr>
<tr>
<td>Residual</td>
<td>17554.7989</td>
<td>xxx</td>
<td>70.5012006</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>20241.75</td>
<td>255</td>
<td>79.3794118</td>
<td>R-squared = xxxxxx</td>
</tr>
</tbody>
</table>

| total      | Coef.       | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|------------|-------------|-----------|------|------|---------------------|
| l10inc     | -6.946661   | 1.72227   | xxxx|x.xxx | -10.33874 -3.554586 |
| age        | -.074917    | .0334132  | xxxx|x.xxx | -.1407254 -.0091085 |
| female     | 2.559048    | 1.095151  | 2.34 | 0.020 | .402108 4.715987   |
| cath       | .7527141    | 1.440845  | 0.52 | 0.602 | -2.085084 3.590512 |
| jewi       | 4.674959    | 1.917259  | 2.44 | 0.015 | .8988472 8.451071  |
| none       | 3.264667    | 1.400731  | 2.33 | 0.021 | .5058747 6.023459  |
| _cons      | 18.24958    | 2.971443  | 6.14 | 0.000 | 12.39721 24.10195  |
Extended Regression Analysis Example
Four regression models of CESD depression score (transformed) in Stata

.* in Stata
.use "D:\soci708\data\survey2b.dta", clear
.* CESD depression scale score is called ‘‘total’’
.* check distribution of total and transform (next 6 lines)
.histogram total, kdensity
.qnorm total
.ladder total
.generate sqrtdep=sqrt(total)
.histogram sqrtdep, kdensity
.qnorm sqrtdep
.* estimate 4 models and standardized coef. (next 5 lines)
.reg sqrtdep age female
.reg sqrtdep age female l10inc educatn
.reg sqrtdep age female l10inc educatn cath jewi none
.reg sqrtdep age female l10inc educatn cath jewi none drinks
.reg sqrtdep age female l10inc educatn cath jewi none drinks, beta
.* get descriptive stats and correlations for presentation
.su sqrtdep age female l10inc educatn cath jewi none drinks
.cor sqrtdep age female l10inc educatn cath jewi none drinks
Checking the Assumptions (1)

- Although regression can be an effective method to summarize the relationship between quantitative variables, some assumptions must be met.
- As with the other methods of inference we have discussed, these assumptions pertain to the population we want to make inferences about.
  - Of course, we do not have data on the whole population so cannot assess the assumptions directly.
  - We can, however, check whether the assumptions appear reasonable for the sample.
Checking the Assumptions (2)
Assumptions for linear regression are:

1. Linearity:
   - Is there a linear relationship between $y$ and $x$? We can assess this assumption in simple regression by looking at a scatterplot.

2. Constant Spread:
   - Is the spread of the $y$-values approximately the same regardless of the value of $x$? If the spread of $y$ changes with $x$, then we have a problem. A scatterplot of $y$ and $x$ or of the residuals against $x$ allows us to assess this.

3. Normality:
   - Are the residuals normally distributed (is there a skew or outliers)? If not normally distributed, we have a problem. We can check this assumption using a histogram or stemplot of the residuals.
Assessing Linearity

> # in R
> library(car)
> data(Prestige)
> attach(Prestige)
> par(pty="s") # to make square plot
> plot(education, prestige)
> abline(lm(prestige~education),
>       col="red", lwd=3)
Assessing Spread

> # in R
> data(Prestige)
> attach(Prestige)
> par(pty="s") # to make square plot
> mod1<-lm(prestige~education)
> plot(education, mod1$residuals,
>     xlab="Education",
>     ylab="Residuals")
> abline(h=0)
Assessing Normality

```r
> # in R
> mod1<-lm(prestige~education)
> hist(mod1$residuals,
    main="Histogram of residuals",
    xlab="Residual",
    col="red")
```

Choosing the Right Test

What is the research question?

- Test for means
  - single sample
    - z-test if \( \mu \) is known
      - t-test if \( \mu \) is unknown
        - \( df = n-1 \)
  - two samples
    - two sample t-test
      - \( df = \text{smallest of } n-1 \)
  - many samples
    - ANOVA
      - F-test
        - \( df = (l-1, N-l) \)

- Test for proportions
  - single sample
    - z-test
      - \( df = n-1 \)
  - two samples
    - two sample z-test
  - many samples
    - chi-square
      - \( df = (r-1)(c-1) \)

- Test for a regression slope
  - \( t \)-test
    - \( df = n-2 \)