Gov 2000 - 9. Multiple Regression: Interactions, F-tests, and Nonlinearities

Matthew Blackwell

November 10, 2015

1. Interactions
2. Nonlinear functional forms
3. Tests of multiple hypotheses

Where are we? Where are we going?

- Last few weeks: adding one variable to the bivariate regression
- This week: effects that vary between groups and other loose ends
- Next week: regression in full matrix form, formulas for all coefficients and SEs.

1/ Interactions

Two binary covariates

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Social pressure experiment:
  - $Y_i = 1$ for voted
  - $X_i = 1$ for neighbors treatment, $X_i = 0$ for civil duty mailer
  - $Z_i = 1$ for female, $Z_i = 0$ for male
- Parameters:
  - β_0: average turnout for males in the control group.
  - β_1: effect of neighbors treatment conditional on gender.
  - β_2: average difference in turnout between women and men conditional on treatment.
- β_1 averages across the effect for men and the effect for women.

Interactions

- How to allow to estimate the effect of neighbors for men and women separately?

1. Subset the data to men and women and run separate regressions.
   - No way to assess whether or not the effects are different from one another.
2. Include an interaction between the treatment and gender:
   - Add a third covariate that is $X_i \times Z_i$:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i$$

   - $X_i \times Z_i = 1$ for treated females ($X_i = 1$ and $Z_i = 1$), 0 otherwise

Binary interactions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i$$

- β_1 is the effect of treatment for men $(Z_i = 0)$:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \times 0 + \beta_3 X_i \times 0 + u_i \\ &= \beta_0 + \beta_1 X_i + u_i \end{aligned}$$

- $\beta_1 + \beta_3$ is the effect of treatment for women $(Z_i = 1)$:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \times X_i + \beta_2 \times 1 + \beta_3 X_i \times 1 + u_i \\ &= (\beta_0 + \beta_1) + (\beta_1 + \beta_3) \times X_i + u_i \end{aligned}$$

- β_3 is the difference in effects between women and men.

Hypothesis tests

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- Due to sampling variation, men and women will never have the exact same effect.
  - $\rightsquigarrow$ $\widehat{\beta}_3$ not exactly equal to 0 even if $\beta_3 = 0$.
- But how do we asses if the differences in the effects are "big enough" for us to say that the effect varies systematically by gender?
- We can test whether or not the effects for the two groups are different by testing the null hypothesis $H_0 : \beta_3 = 0$

$$\frac{\widehat{\beta}_3}{\widehat{SE}[\widehat{\beta}_3]}$$

Social pressure example

```r
summary(lm(voted ~ treat * female, data = social))
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.32274    0.00343   93.97  < 2e-16 ***
## treat          0.06180    0.00486   12.72  < 2e-16 ***
## female        -0.01640    0.00486   -3.38  0.00073 ***
## treat:female   0.00321    0.00687    0.47  0.63990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.475 on 76415 degrees of freedom
## Multiple R-squared:  0.00469,    Adjusted R-squared:  0.00465
## F-statistic:  120 on 3 and 76415 DF,  p-value: <2e-16
```

A note on linearity

- The linearity assumption says we can write Y_i as a linear function of the parameters:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i$$

- Linearity allows us to extrapolate to combinations of the covariates we don't observe.
- Linearity is usually violated when non-continuous outcomes (binary/categorical), but is satisfied in saturated models.
- A saturated model is one with discrete covariates and as many parameters as there are combinations of the covariates.
  - Same as estimating separate means for each combination of the covariates.
  - No extrapolation $\rightsquigarrow$ linearity holds by construction.

Saturated bivariate regression

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- If X_i is binary:

$$E[Y_i|X_i = 0] = \beta_0$$
$$E[Y_i|X_i = 1] = \beta_0 + \beta_1$$

- Model is saturated: β_1 is the difference in CEFs between $X_i = 1$ and $X_i = 0$.
  - No extrapolation, no linearity assumption.
- Compare this to when X_i is continuous:

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 \times x$$
$$E[Y_i|X_i = x + 1] = \beta_0 + \beta_1 \times (x + 1)$$

- Linearity assumes the effect (β_1) is constant across values of X_i.

Saturated model example

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i$$

- Four possible values of X_i and Z_i, four possible values of $\mathbb{E}[Y_i | X_i, Z_i]$:

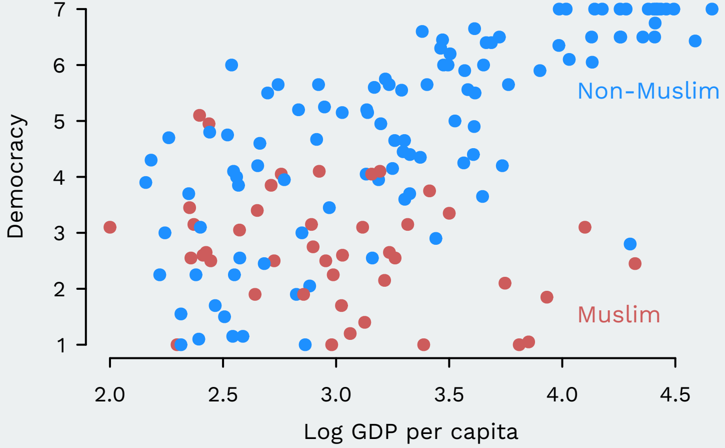
$$\begin{aligned}
E[Y_i | X_i = 0, Z_i = 0] &= \beta_0 \\
E[Y_i | X_i = 1, Z_i = 0] &= \beta_0 + \beta_1 \\
E[Y_i | X_i = 0, Z_i = 1] &= \beta_0 + \beta_2 \\
E[Y_i | X_i = 1, Z_i = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

- With binary covariates, including all interactions saturates the model.
- $\rightsquigarrow$ OK to use this model with a binary outcome.

One continuous, one binary covariate

- How do interactions work when a variable is continuous?
- Data comes from Fish (2002), "Islam and Authoritarianism."
- Basic relationship: does more economic development lead to more democracy?
- We measure economic development with log GDP per capita
- We measure democracy with a Freedom House score, 1 (less free) to 7 (more free)

Let's see the data



- Fish argues that Muslim countries are less likely to be democratic no matter their economic development

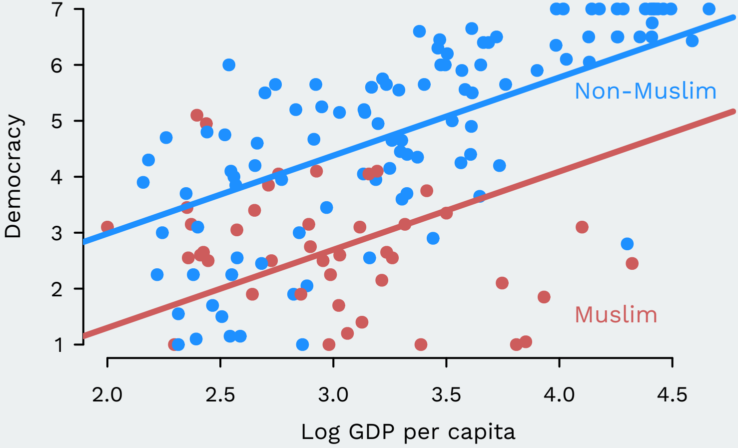
Controlling for religion

- muslim that is 1 when Islam is the largest religion in a country and 0 otherwise

```r
mod <- lm(fhrev ~ income + muslim, data = FishData)
summary(mod)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.189      0.556    0.34     0.73
## income         1.397      0.163    8.58  1.3e-14 ***
## muslim        -1.683      0.238   -7.07  5.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 146 degrees of freedom
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.515
## F-statistic: 79.6 on 2 and 146 DF,  p-value: <2e-16
```

Plotting the lines



- But the regression is a poor fit for Muslim countries
- Can we allow for different slopes for each group?

Interactions with a binary variable

- In this case, $Z_i = 1$ for the country being Muslim
- Interaction term is the product of the two marginal variables of interest:

$$income_i \times mulsim_i$$

- Here is the model with the interaction term:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Interactions with R (II)

- Easier/better to write the interaction term as `first*second`:

```r
mod.int <- lm(fhrev ~ income * muslim, data = FishData)
summary(mod.int)
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.349      0.540   -2.50    0.014 *
## income            1.859      0.159   11.70  < 2e-16 ***
## muslim            5.741      1.134    5.06  1.2e-06 ***
## income:muslim    -2.427      0.364   -6.66  5.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.13 on 145 degrees of freedom
## Multiple R-squared:  0.634,  Adjusted R-squared:  0.626
## F-statistic: 83.6 on 3 and 145 DF,  p-value: <2e-16
```

Data matrix with interactions

```r
head(model.matrix(mod.int))
```

```
##   (Intercept) income muslim income:muslim
## 1           1  2.925      1         2.925
## 2           1  3.214      1         3.214
## 3           1  2.824      0         0.000
## 4           1  3.762      0         0.000
## 5           1  3.188      0         0.000
## 6           1  4.436      0         0.000
```

Two lines in one regression

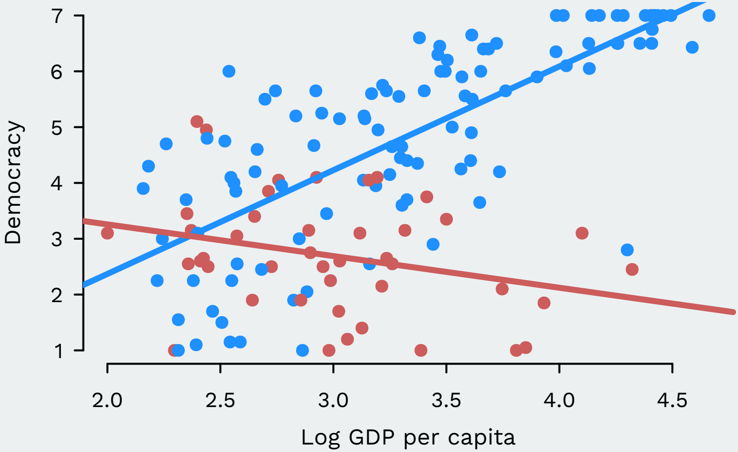
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- When $Z_i = 0$:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$
- When $Z_i = 1$:

$$\widehat{Y}_i = (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) X_i$$

Graphing interactions

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1 + \widehat{\beta}_3$



Interpretation of the coefficients

- β_0: average value of Y_i when both X_i and Z_i are equal to 0
- β_1: a one-unit change in X_i is associated with a β_1-unit change in Y_i when $Z_i = 0$
    - Model not saturated! Linearity in X_i!
- β_2: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group when $X_i = 0$
- β_3: change in the effect of X_i on Y_i between $Z_i = 1$ group and $Z_i = 0$

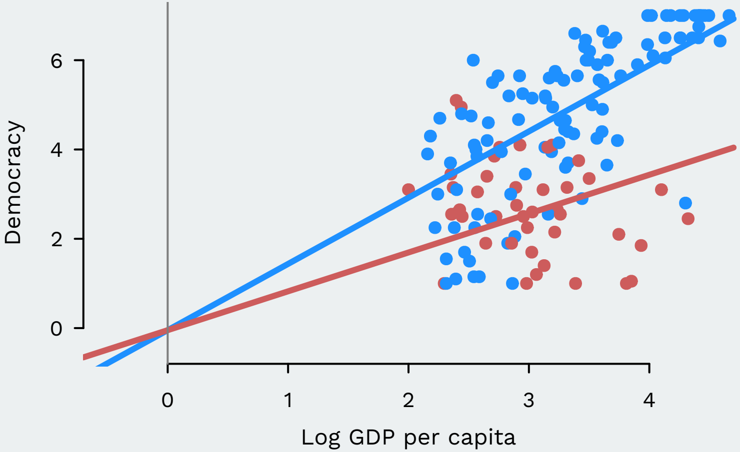
Lower order terms

- Always include the marginal effects (sometimes called the lower order terms)
- Imagine we omitted the lower order term for `muslim`:

```
wrong.mod <- lm(fhrev ~ income + income:muslim, data = FishData)
summary(wrong.mod)
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0465     0.5133   -0.09     0.93
## income          1.4837     0.1520    9.76  < 2e-16 ***
## income:muslim  -0.6137     0.0725   -8.46  2.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.22 on 146 degrees of freedom
## Multiple R-squared:  0.569,  Adjusted R-squared:  0.563
## F-statistic: 96.3 on 2 and 146 DF,  p-value: <2e-16
```

Omitting lower order terms



- What's the problem here?
- We've restricted the intercepts to be the same for both models:

Omitting lower order terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$

- Implication: no difference between Muslims and non-Muslims when income is 0
- Distorts slope estimates.
- Very rarely justified.

Interactions with two continuous variables

- Now let Z_i be continuous
- Z_i is the percent growth in GDP per capita from 1975 to 1998
- Is the effect of economic development for rapidly developing countries higher or lower than for stagnant economies?
- We can still define the interaction:

$$income_i \times growth_i$$

- And include it in the regression:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Example of continuous interaction

```r
mod.cont <- lm(fhrev ~ income * growth, data = FishData)
summary(mod.cont)
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.1066     0.6225   -0.17   0.8643
## income          1.2922     0.1941    6.66  5.3e-10 ***
## growth         -0.6172     0.2383   -2.59   0.0106 *
## income:growth   0.2395     0.0753    3.18   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 145 degrees of freedom
## Multiple R-squared:  0.433,  Adjusted R-squared:  0.422
## F-statistic: 36.9 on 3 and 145 DF,  p-value: <2e-16
```

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 0.5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 0.5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 0.5$
$Z_i = 1$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 1$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 1$
$Z_i = 5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 5$

General interpretation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i$$

- $\beta_1 \rightsquigarrow$ how the predicted outcome varies in X_i when $Z_i = 0$.
- $\beta_2 \rightsquigarrow$ how the predicted outcome varies in Z_i when $X_i = 0$
- $\beta_3 \rightsquigarrow$ the change in the effect of X_i given a one-unit change in Z_i:
$$\frac{\partial \mathbb{E}[Y_i|X_i, Z_i]}{\partial X_i} = \beta_1 + \beta_3 Z_i$$
- $\beta_3 \rightsquigarrow$ the change in the effect of Z_i given a one-unit change in X_i:
$$\frac{\partial \mathbb{E}[Y_i|X_i, Z_i]}{\partial Z_i} = \beta_2 + \beta_3 X_i$$

Standard errors for marginal effects

- What if we want to get a standard error for the effect of X_i at some level of Z_i?
- Marginal effect of X_i at some value Z_i:

$$\frac{\partial \widehat{\mathbb{E}[Y_i|X_i, Z_i]}}{\partial X_i} = \widehat{\beta}_1 + \widehat{\beta}_3 Z_i$$

- We already saw that $\widehat{\beta}_1$ is the effect when $Z_i = 0$. What about other values of Z_i?
- Use the properties of variances:

$$\begin{aligned}\text{Var}\left(\frac{\partial \widehat{\mathbb{E}[Y_i|X_i, Z_i]}}{\partial X_i}\right) &= \text{Var}(\widehat{\beta}_1 + Z_i \widehat{\beta}_3) \\ &= \text{Var}[\widehat{\beta}_1] + Z_i^2 \text{Var}[\widehat{\beta}_3] + 2 Z_i \text{Cov}[\widehat{\beta}_1, \widehat{\beta}_3]\end{aligned}$$

Standard errors for marginal effects

- Get the entries from the vcov() function (more on this next week):

```
## SE of effect of income at muslime = 1
var.inter <- vcov(mod.int)["income","income"] +
    1^2 * vcov(mod.int)["income:muslim","income:muslim"] +
    2 * 1 * vcov(mod.int)["income","income:muslim"]
sqrt(var.inter)
```

```
## [1] 0.3277
```

```
## SE when muslim = 0
sqrt(vcov(mod.cont)["income", "income"])
```

```
## [1] 0.1941
```

Recentering for interaction terms

- $\beta_1 \rightsquigarrow$ how the predicted outcome varies in X_i when $Z_i = 0$.
- A trick for getting R to calculate the standard errors for you is to recenter the variable so that 0 corresponds to the value you want to estimate.
- With binary Z_i, replace Z_i with $1 - Z_i$:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2(1 - Z_i) + \beta_3 X_i(1 - Z_i) + u_i$$

- Now, $\widehat{\beta}_1$ is the slope on X_i when $1 - Z_i = 0$, or, rearranging, when $Z_i = 1$.
- We “trick” R into calculating the standard errors for us

Recentering in R

- Use the `I()` syntax:

```
summary(lm(fhrev ~ income * I(1 - muslim), data = FishData))
```

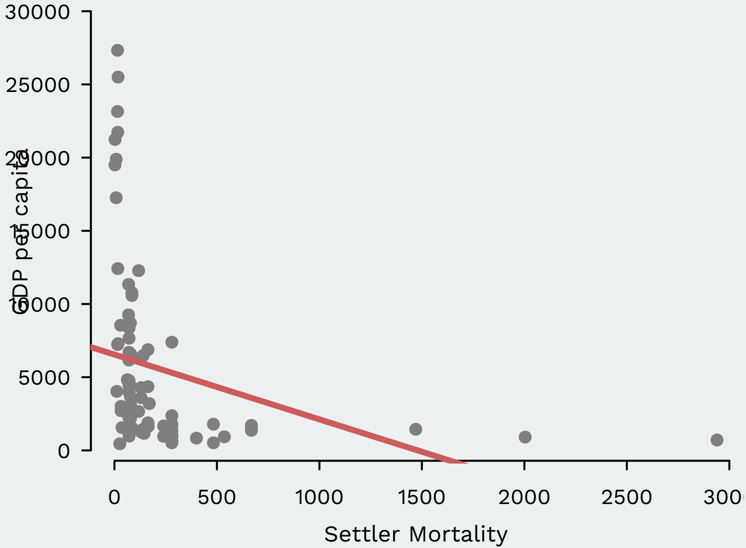
```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.392      0.997    4.41  2.0e-05 ***
## income                  -0.568      0.328   -1.73    0.085 .
## I(1 - muslim)           -5.741      1.134   -5.06  1.2e-06 ***
## income:I(1 - muslim)     2.427      0.364    6.66  5.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.13 on 145 degrees of freedom
## Multiple R-squared:  0.634,  Adjusted R-squared:  0.626
## F-statistic: 83.6 on 3 and 145 DF,  p-value: <2e-16
```

2/ Nonlinear functional forms

Logs of random variables

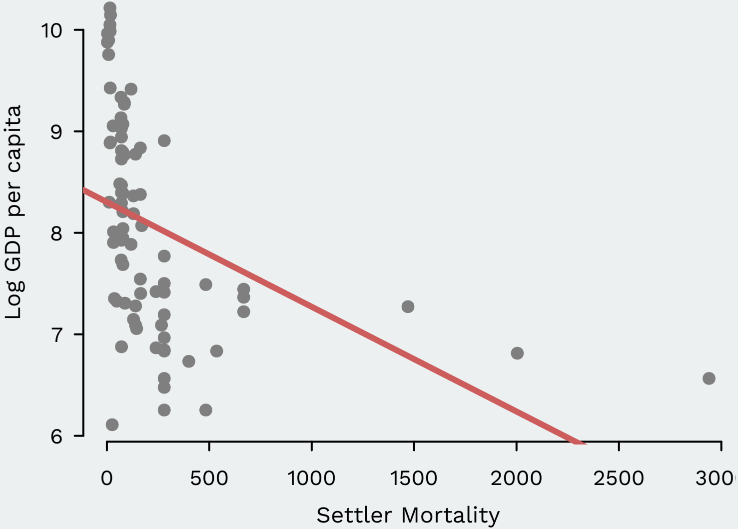
- We can account for non-linearity in X_i in a couple of ways
- One way: transform X_i or Y_i using the natural logarithm
- Useful when X_i or Y_i are positive and right-skewed
- Changes the interpretation of β_1:
  - Regress $\log(Y_i)$ on $X_i \rightarrow 100 \times \beta_1 \approx$ percentage increase in Y_i associated with a one-unit increase in X_i
  - Regress $\log(Y_i)$ on $\log(X_i) \rightarrow \beta_1 \approx$ percentage increase in Y_i associated with a one percent increase in X_i
  - Only useful for small increments, not for discrete r.v

Raw scales


30000
25000
20000
15000
10000
5000
0
GDP per capita
0
500
1000
1500
2000
2500
300
Settler Mortality

Log scale for Settler mortality

Log scale for GDP



Log scale for both

Logging variables

- Handy chart for interpreting logged variables:

Model	Equation	β_1 Interpretation
Level-Level	$Y = \beta_0 + \beta_1 X$	1-unit $\Delta X \rightsquigarrow \beta_1 \Delta Y$
Log-Level	$\log(Y) = \beta_0 + \beta_1 X$	1-unit $\Delta X \rightsquigarrow 100 \times \beta_1 \% \Delta Y$
Level-Log	$Y = \beta_0 + \beta_1 \log(X)$	1% $\Delta X \rightsquigarrow (\beta_1 / 100) \Delta Y$
Log-Log	$\log(Y) = \beta_0 + \beta_1 \log(X)$	1% $\Delta X \rightsquigarrow \beta_1 \% \Delta Y$

Adding a squared term

- Another approach: model relationship as a polynomial
- Add a polynomial of X_i to account for the non-linearity:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 X_i^2$$

- Similar to an "interaction" with itself: marginal effect of X_i varies as a function of X_i:

$$\frac{\partial \mathbb{E}[Y_i | X_i]}{\partial X_i} = \beta_1 + \beta_2 X_i$$

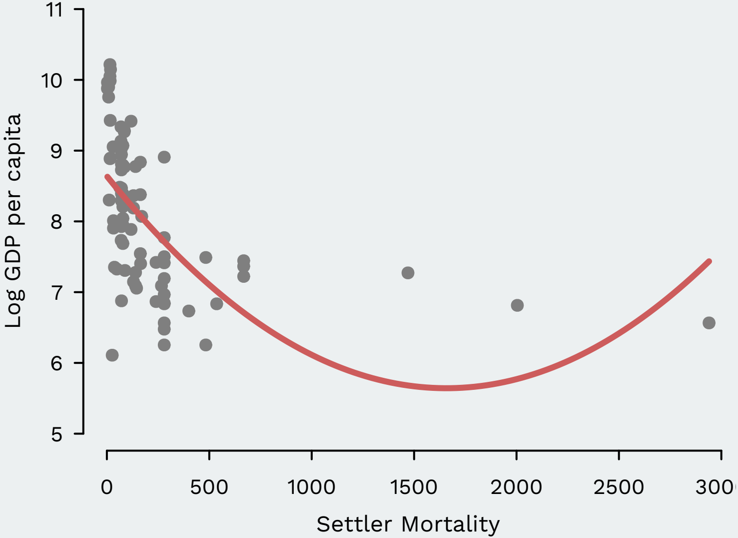
Adding a squared term in R

```
quad.mod <- lm(logpgp95 ~ raw.mort + I(raw.mort^2), data = ajr)
summary(quad.mod)
```

```
##
## Coefficients:
##                   Estimate   Std. Error t value   Pr(>|t|)
## (Intercept)    8.639495953  0.137819111   62.69    < 2e-16 ***
## raw.mort      -0.003615763  0.000663785   -5.45 0.00000058 ***
## I(raw.mort^2)  0.000001091  0.000000262    4.16 0.00008194 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.884 on 78 degrees of freedom
##   (82 observations deleted due to missingness)
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.304
## F-statistic: 18.4 on 2 and 78 DF,  p-value: 0.000000276
```

Non-linear functional form

- Plotting the results (see handout for R code):



3/ Tests of multiple hypotheses

Review of t-tests

- Null hypothesis:

$$H_0 : \beta_k = 0$$

- Alternative hypothesis:

$$H_A : \beta_k \neq 0$$

- Test statistic (t-statistic):

$$t = \frac{\widehat{\beta}_k}{\widehat{SE}[\widehat{\beta}_k]}$$

- $N(0, 1)$ distribution in large samples (under Assumptions 1-5)
- $t_{n-(k+1)}$ distribution under Assumptions 1-6 (when errors are conditionally Normal)

Joint null hypotheses

- What about more complicated null hypotheses?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$$

- Here we might want to test whether X_i belongs in the regression at all
- But that null hypothesis involves 2 parameters:

$$H_0 : \beta_1 = 0 \text{ and } \beta_3 = 0$$

- The alternative hypothesis:

$$H_A : \beta_1 \neq 0 \text{ or } \beta_3 \neq 0$$

- How can we test this null hypothesis?
- We will compare the predictive power of the model under the null and the model under the alternative

Unrestricted model

- Unrestricted model (alternative is true):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$$

- Estimates:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- SSR from unrestricted model:

$$SSR_u = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

Restricted model

- Restricted model (null is true):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i \\ &= \beta_0 + 0 \times X_i + \beta_2 Z_i + 0 \times X_i Z_i \\ Y_i &= \beta_0 + \beta_2 Z_i \end{aligned}$$

- Estimates:

$$\widetilde{Y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 Z_i$$

- SSR from restricted model model:

$$SSR_r = \sum_{i=1}^{n} (Y_i - \widetilde{Y}_i)^2$$

- If the null is true, then SSR_r and SSR_u should only be different due to sampling variation.
- The bigger the reduction in the prediction errors between SSR_r and SSR_u, the less plausible is the null hypothesis.

F statistic

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$$

- $(SSR_r - SSR_u)$: the increase in the variation in the residuals when we remove those βs
- q = number of restrictions (numerator degrees of freedom)
- $n - k - 1$: denominator/unrestricted degrees of freedom
- Intuition:

$$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

- Each of these is scaled by the degrees of freedom

F statistic in R

```r
ur.mod <- lm(fhrev ~ income * growth, data = FishData)
r.mod <- lm(fhrev ~ growth, data = FishData)
anova(r.mod, ur.mod)
```

```
## Analysis of Variance Table
##
## Model 1: fhrev ~ growth
## Model 2: fhrev ~ income * growth
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1    147 452
## 2    145 284  2       168 42.9 2.3e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

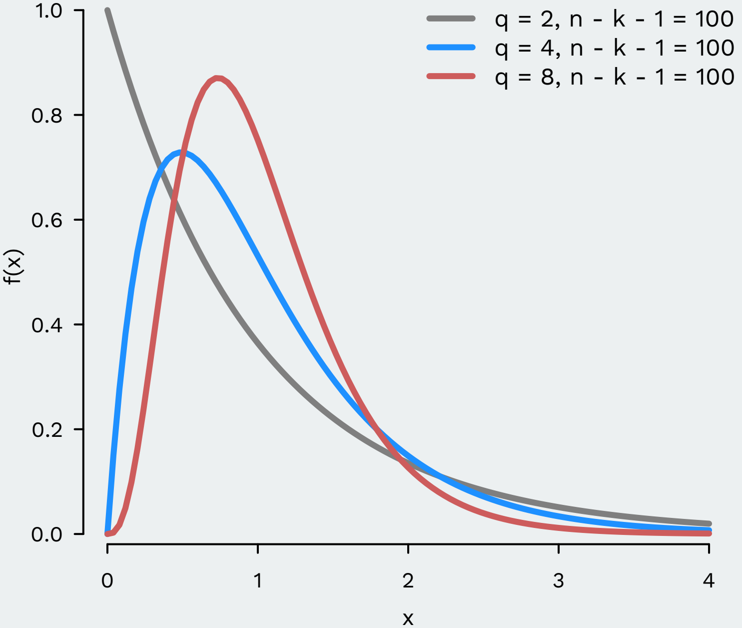
The F test

- What is the null distribution of this F statistic?
  - Assumptions 1-5 + large sample: F statistic has an approximately F distribution
  - Assumptions 1-6 (Normality): F statistic has an exact F distribution
  - Very similar to the t-test
- Either way, under the null:

$$\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-(k+1)}$$

- The F distribution tells us how much of a relative increase in the SSR we should expect if we were to add irrelevant variables to the model.
- Compare our observed F-statistic to the distribution under the null.

F distribution



F-test steps

1. Choose a Type I error rate, α.
   - Same interpretation as always: the proportion of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
   - Rejection region is the region $F > c$ such that $\mathbb{P}(F > c) = \alpha$
   - We can get this from R using the `qf()` function:

```
qf(0.05, 2, 100, lower.tail = FALSE)
```

```
## [1] 3.087
```

3. Reject if observed statistic is bigger than critical value

F-test p-values

- We might also want to calculate p-values.
- Probability of observing an F-statistic this large or larger given the null hypothesis is true.
- This is just the proportion of the distribution above the observed F-statistic.
- We can calculate this in R using the `pf()` function:

```
pf(5.2, 2, 100, lower.tail = FALSE)
```

```
## [1] 0.007105
```

F statistic for all variables

- "The" F-test: tests the null of all coefficients except the intercept being 0.
- In that case, the restricted model is just:

$$Y_i = \beta_0 + u_i$$

- And the estimate here would just be sample mean ($\widehat{\beta}_0 = \overline{Y}$)
- The SSR_r then would just be the sampling variation in Y:

$$SSR_f = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- Often reported with regression output.

Example of F-test for all variables

```
summary(ur.mod)
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.1066     0.6225   -0.17   0.8643
## income           1.2922     0.1941    6.66  5.3e-10 ***
## growth          -0.6172     0.2383   -2.59   0.0106 *
## income:growth    0.2395     0.0753    3.18   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 145 degrees of freedom
## Multiple R-squared:  0.433,  Adjusted R-squared:  0.422
## F-statistic: 36.9 on 3 and 145 DF,  p-value: <2e-16
```

Connection to t tests

- What about an F-test with just one coefficient equal to zero? $H_0 : \beta_1 = 0$
- We already can do this with an t-test. Is there a connection to the F-test?
- The F-statistic for a single restriction is just the square of the t-statistic:

$$F = t^2 = \left(\frac{\widehat{\beta}_1}{\widehat{SE}[\widehat{\beta}_1]} \right)^2$$

Multiple testing

- If we test all of the coefficients separately with a t-test, then we should expect that 5% of them will be significant just due to random chance.
- Illustration: randomly draw 21 variables, and run a regression of the first variable on the rest.
- By design, no effect of any variable on any other, but when we run the regression:

Multiple test example

```r
noise <- data.frame(matrix(rnorm(2100), nrow = 100, ncol = 21))
summary(lm(noise))
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.028039   0.113820   -0.25   0.8061
## X2          -0.150390   0.112181   -1.34   0.1839
## X3           0.079158   0.095028    0.83   0.4074
## X4          -0.071742   0.104579   -0.69   0.4947
## X5           0.172078   0.114002    1.51   0.1352
## X6           0.080852   0.108341    0.75   0.4577
## X7           0.102913   0.114156    0.90   0.3701
## X8          -0.321053   0.120673   -2.66   0.0094 **
## X9          -0.053122   0.107983   -0.49   0.6241
## X10          0.180105   0.126443    1.42   0.1583
## X11          0.166386   0.110947    1.50   0.1377
## X12          0.008011   0.103766    0.08   0.9387
## X13          0.000212   0.103785    0.00   0.9984
## X14         -0.065969   0.112214   -0.59   0.5583
## X15         -0.129654   0.111575   -1.16   0.2487
## X16         -0.054446   0.125140   -0.44   0.6647
## X17          0.004335   0.112012    0.04   0.9692
## X18         -0.080796   0.109853   -0.74   0.4642
## X19         -0.085806   0.118553   -0.72   0.4713
## X20         -0.186006   0.104560   -1.78   0.0791 .
## X21          0.002111   0.108118    0.02   0.9845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.999 on 79 degrees of freedom
## Multiple R-squared:  0.201,	Adjusted R-squared:  -0.00142
## F-statistic: 0.993 on 20 and 79 DF,  p-value: 0.48
```

Multiple testing gives false positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).
- But this is exactly what we expect: $1/20 = 0.05$ of the tests are false positives at the 0.05 level
- Also note that $2/20 = 0.1$ are significant at the 0.1 level. Totally expected!
- But notice the F-statistic: the variables are not jointly significant

Wrap up

- Interactions: allows us to see how the effect of one variable changes as a function of another
- F-tests: allows us to test the effect of multiple variables at the same time
- Non-linearity: logs and polynomials can make the linearity assumption more plausible
- Next time: regression in full matrix form