

Gov 2000 - 11. Interactions, Nonlinearities, and the F-test

Matthew Blackwell

Harvard University

mblackwell@gov.harvard.edu

Fall 2016

Where are we? Where are we going?

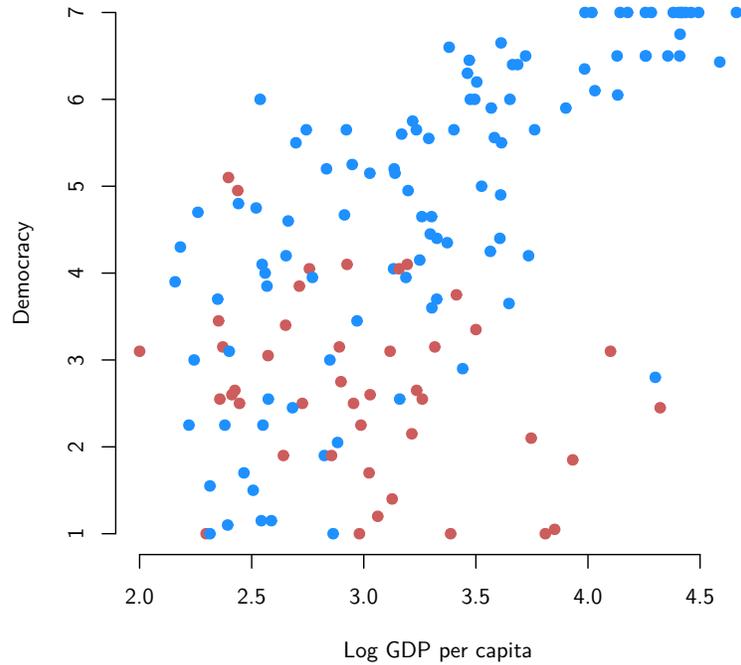
- Last few weeks: linear regression at its most general, matrix form
- This week: effects that vary between groups and other loose ends
- Next week: troubleshooting the linear model

INTERACTIONS

Data

Fish (2002) studies the relationship between predominantly Muslim societies and authoritarianism. The question he is exploring is whether or not countries with Islam as its largest religion are more likely to be authoritarian. To measure the level of democratization, we use the Freedom House score. But we also might think that its important to control for economic development since many Muslim countries have high GDP due to oil wealth. We can see the interplay between these variables in this graph (where red dots are Muslim countries, and blue dots are non-Muslim countries):

```
load("../data/FishData.RData")
plot(FishData$income, FishData$fhrev, ylab = "Democracy", xlab = "Log GDP per capita", pch = 19, bty = "n", col = "red")
```



Let's run a regression model with income being log GDP per capita and a binary variable `muslim` that is 1 when Islam is the largest religion in a country and 0 otherwise:

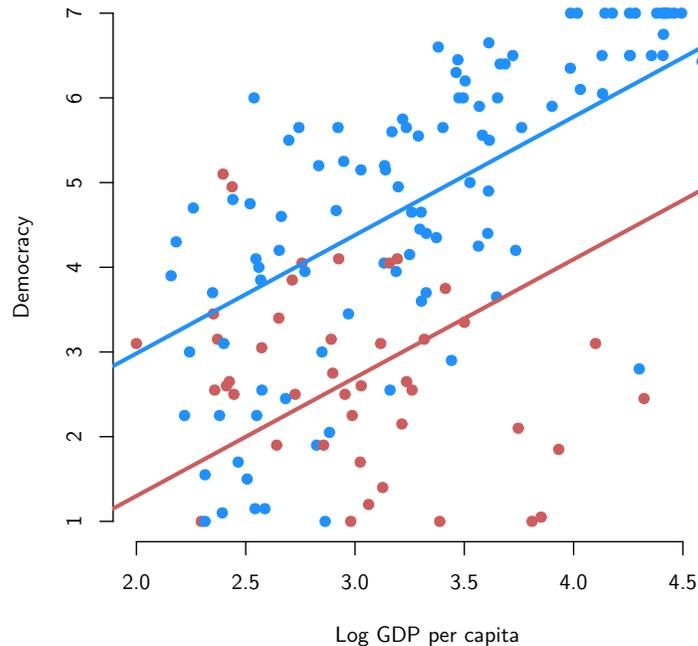
```
mod <- lm(fhrev ~ income + muslim, data = FishData)
summary(mod)
```

```
##
## Call:
## lm(formula = fhrev ~ income + muslim, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3961 -0.8276  0.2804  0.9425  3.2467
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1887     0.5560   0.339   0.735
## income        1.3970     0.1629   8.576 1.31e-14 ***
## muslim       -1.6827     0.2379  -7.074 5.82e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.282 on 146 degrees of freedom
## Multiple R-squared:  0.5216, Adjusted R-squared:  0.515
## F-statistic: 79.58 on 2 and 146 DF,  p-value: < 2.2e-16
```

Since `muslim` here is a binary variable, we can plot the two parallel regression lines implied by this model:

```
plot(FishData$income, FishData$fhrev, ylab = "Democracy", xlab = "Log GDP per capita", pch = 19, bty = "n", col = "blue",
      abline(a = coef(mod)[1], b = coef(mod)[2], col = "dodgerblue", lwd = 3)
      abline(a = coef(mod)[1] + coef(mod)[3], b = coef(mod)[2], col = "indianred", lwd = 3))
```



But looking at the data here, we might notice that the red line for Muslim countries does not fit the lines very well. Maybe there is a different relationship between income and democracy in Muslim and non-Muslim countries. An alternative way of thinking about this is that this model assumes that the “effect” of being a Muslim country is constant across levels of economic development. But this might be false! To get around this, we can add an **interaction term** to our model.

First, let’s focus on a situation with a continuous X_i (economic development) and binary Z_i (Muslim country). We can add another covariate to the baseline model

that allows the effect of income to vary by Muslim status. This covariate is called an interaction term and it is the product of the two **marginal** variables of interest:

$$income_i \times muslim_i$$

Here is the model with the interaction term:

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

Literally this last term is just a new covariate that is the X_i multiplied by Z_i . - Thus, the design matrix, \mathbf{X} looks like this:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 & Z_1 & X_1 \times Z_1 \\ 1 & X_2 & Z_2 & X_2 \times Z_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & Z_n & X_n \times Z_n \end{bmatrix}$$

In R, we simply add a new term to the regression which is `first:second` where `first` and `second` are the names of marginal variables:

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

```
##
## Call:
## lm(formula = fhrev ~ income + muslim + income:muslim, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.8460 -0.5705  0.0940  0.8517  2.6307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.3489    0.5400  -2.498  0.0136 *
## income         1.8592    0.1590  11.695 < 2e-16 ***
## muslim         5.7413    1.1338   5.064 1.23e-06 ***
## income:muslim -2.4267    0.3642 -6.662 5.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 145 degrees of freedom
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.6261
## F-statistic: 83.61 on 3 and 145 DF, p-value: < 2.2e-16
```

Let's look at the design matrix to see what this looks like:

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

```
## (Intercept)  income muslim income:muslim
## 1          1 2.925312      1      2.925312
## 2          1 3.214314      1      3.214314
## 3          1 2.824126      0      0.000000
## 4          1 3.762078      0      0.000000
## 5          1 3.187803      0      0.000000
## 6          1 4.435542      0      0.000000
```

Note that it is easier and better to write the interaction term as `first*second`, which adds each variable and its interaction to the model:

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

```
##
## Call:
## lm(formula = fhrev ~ income * muslim, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.8460 -0.5705  0.0940  0.8517  2.6307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.3489    0.5400  -2.498  0.0136 *
## income         1.8592    0.1590  11.695 < 2e-16 ***
## muslim         5.7413    1.1338   5.064 1.23e-06 ***
## income:muslim -2.4267    0.3642  -6.662 5.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 145 degrees of freedom
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.6261
## F-statistic: 83.61 on 3 and 145 DF, p-value: < 2.2e-16
```

How can we interpret this model? We'll repeat our exercise from a few weeks ago and plug in the two possible values of Z_i . First, when $Z_i = 0$:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\beta}_3 X_i Z_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 \times 0 + \hat{\beta}_3 X_i \times 0 \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_i\end{aligned}$$

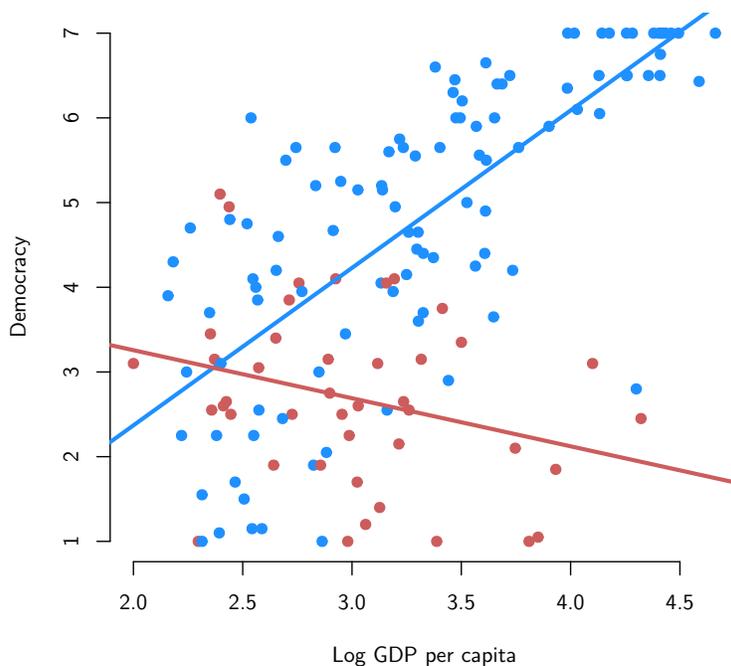
Next, when $Z_i = 1$:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\beta}_3 X_i Z_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 \times 1 + \hat{\beta}_3 X_i \times 1 \\ &= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) X_i\end{aligned}$$

Thus, we can think of this regression producing two regression lines:

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

```
plot(FishData$income, FishData$fhrev, ylab = "Democracy", xlab = "Log GDP per capita", pch = 19, bty = "n", col = "blue",
      abline(a = coef(mod.int)[1], b = coef(mod.int)[2], col = "dodgerblue", lwd = 3)
      abline(a = coef(mod.int)[1] + coef(mod.int)[3], b = coef(mod.int)[2] + coef(mod.int)[4], col = "indianred", lwd = 3))
```



In a more general context, we can interpret the coefficients from an interaction model as:

- $\hat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\hat{\beta}_1$: a one-unit change in X_i is associated with a $\hat{\beta}_1$ -unit change in Y_i when $Z_i = 0$
- $\hat{\beta}_2$: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group when $X_i = 0$
- $\hat{\beta}_3$: change in the effect of X_i on Y_i between $Z_i = 1$ group and $Z_i = 0$
- $\hat{\beta}_4$: change in the effect of Z_i on Y_i given a one-unit change in X_i

In a model with an interaction, we refer to the marginal variable, X_i and Z_i , as the **lower-order terms**. It is very important to always include these marginal effects. Why? Well, imagine we omitted the lower order term for `muslim`:

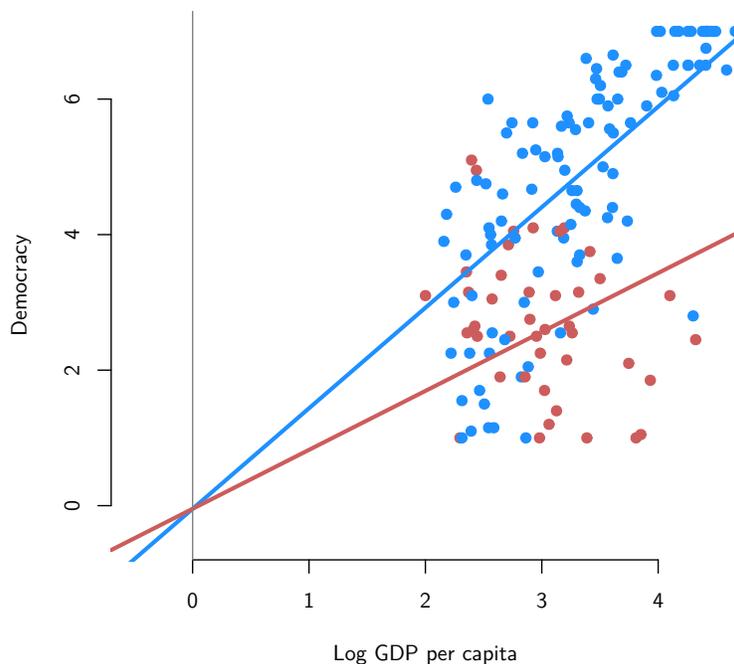
```
wrong.mod <- lm(fhrev ~ income + income:muslim, data = FishData)
summary(wrong.mod)
```

```
##
## Call:
## lm(formula = fhrev ~ income + income:muslim, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5338 -0.7332  0.2524  0.8582  3.0619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.04646   0.51333  -0.091   0.928
## income        1.48368   0.15202   9.760 < 2e-16 ***
## income:muslim -0.61372   0.07255  -8.460 2.56e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.217 on 146 degrees of freedom
## Multiple R-squared:  0.5689, Adjusted R-squared:  0.563
## F-statistic: 96.34 on 2 and 146 DF, p-value: < 2.2e-16
```

```

plot(FishData$income, FishData$fhrev, ylab = "Democracy", xlab = "Log GDP per capita", pch = 19, bty = "n", col = "blue",
      abline(a = coef(wrong.mod)[1], b = coef(wrong.mod)[2], col = "dodgerblue", lwd = 3)
      abline(a = coef(wrong.mod)[1], b = coef(wrong.mod)[2] + coef(wrong.mod)[3], col = "indianred", lwd = 3)
      abline(v = 0, col = "grey50")

```



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

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

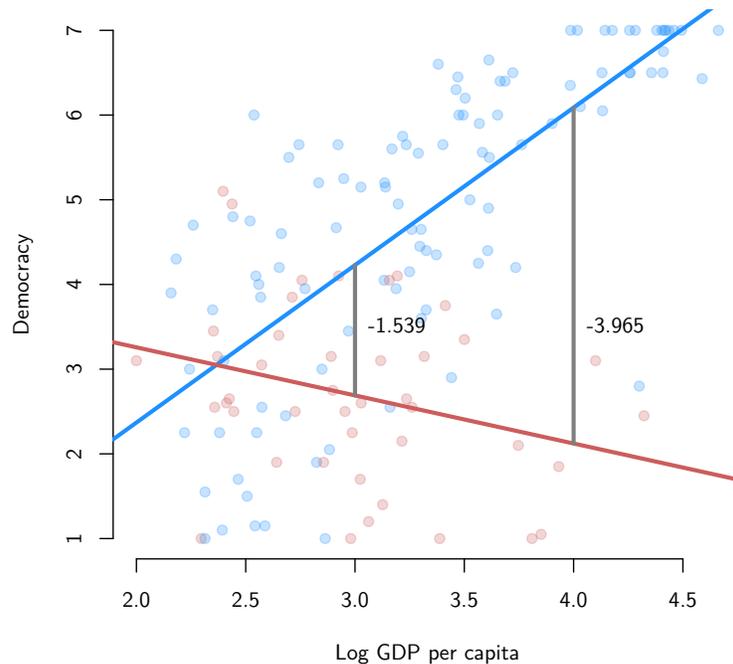
	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\hat{\beta}_0$	$\hat{\beta}_1$
Muslim country ($Z_i = 1$)	$\hat{\beta}_0 + 0$	$\hat{\beta}_1 + \hat{\beta}_3$

Basically, dropping the lower order term implies that there is no difference between Muslims and non-Muslims when income is 0. Or, practically, that the intercept is the same for the two groups, but the slopes differ. Distorts slope estimates. This is very rarely justified for a regression model.

We have interpreted the coefficient on the interaction term as the difference in the slopes between the $Z_i = 1$ and $Z_i = 0$. But we could also think of the interaction

term as showing how much the effect of Z_i changes given a one-unit change in X_i . That is, how much bigger or smaller does the effect of being a Muslim country change as we change log GDP per capita by one unit. We can think of the effect of `muslim` to be the difference between the two regression lines at any value of X_i . Here I've drawn the differences at $X_i = 3$ and $X_i = 4$. If you take the difference between these two values, you'll see that it is exactly equal to the coefficient on the interaction term.

```
plot(FishData$income, FishData$fhrev, ylab = "Democracy", xlab = "Log GDP per capita", pch = 19, bty = "n", col = "blue",
      abline(a = coef(mod.int)[1], b = coef(mod.int)[2], col = "dodgerblue", lwd = 3)
      abline(a = coef(mod.int)[1] + coef(mod.int)[3], b = coef(mod.int)[2] + coef(mod.int)[4], col = "indianred", lwd = 3)
p30 <- predict(mod.int, newdata = data.frame(income = 3, muslim = 0))
p31 <- predict(mod.int, newdata = data.frame(income = 3, muslim = 1))
p40 <- predict(mod.int, newdata = data.frame(income = 4, muslim = 0))
p41 <- predict(mod.int, newdata = data.frame(income = 4, muslim = 1))
segments(x0 = 3, y0 = p30, y1 = p31, lwd = 3, col = "grey50")
segments(x0 = 4, y0 = p40, y1 = p41, lwd = 3, col = "grey50")
text(x = 3, y = 3.5, paste(round(p31-p30, digits = 3)), pos = 4)
text(x = 4, y = 3.5, paste(round(p41-p40, digits = 3)), pos = 4)
```



More generally, we can think of the conditional effect of Z_i given a value of X_i as:

$$\mathbb{E}[Y_i | Z_i = 1, X_i] - \mathbb{E}[Y_i | Z_i = 0, X_i] = \beta_2 + \beta_3 X_i$$

From this it is clear that effect of Z_i is changing as X_i changes.

Interaction between two continuous variables

Now we move to a situation where we have two continuous variables interacted, so that Z_i is continuous. For example, Z_i could be the percent growth in GDP per capita from 1975 to 1998. With this variable, we might ask: 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:

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

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

##
## Call:
## lm(formula = fhrev ~ income * growth, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.0018 -0.9356  0.2241  0.9604  2.8338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1066     0.6225  -0.171   0.8643
## income        1.2922     0.1941   6.659 5.33e-10 ***
## growth       -0.6172     0.2383  -2.590  0.0106 *
## income:growth  0.2395     0.0753   3.180  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:  0.4332, Adjusted R-squared:  0.4215
## F-statistic: 36.95 on 3 and 145 DF, p-value: < 2.2e-16
```

```
head(model.matrix(mod.cont))
```

```
## (Intercept) income growth income:growth
## 1          1 2.925312  -0.8   -2.3402497
## 2          1 3.214314   0.2    0.6428628
## 3          1 2.824126  -1.6   -4.5186013
## 4          1 3.762078   0.6    2.2572469
## 5          1 3.187803  -6.6  -21.0394974
## 6          1 4.435542   2.2    9.7581919
```

With a continuous Z_i , we can have more than two values that it can take on. Thus, we have a set of regression lines for X_i based on the level of Z_i :

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

Thus, when we have our regression model like so:

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

We can interpret each of the coefficients as:

- The coefficient $\hat{\beta}_1$ measures how the predicted outcome varies in X_i when $Z_i = 0$.
- The coefficient $\hat{\beta}_2$ measures how the predicted outcome varies in Z_i when $X_i = 0$.
- The coefficient $\hat{\beta}_3$ is the change in the effect of X_i given a one-unit change in Z_i :

$$\frac{\partial \mathbb{E}[Y_i | X_i, Z_i]}{\partial X_i} = \beta_1 + \beta_3 Z_i$$

- The coefficient $\hat{\beta}_3$ is the change in the effect of Z_i given a one-unit change in X_i :

$$\frac{\partial \mathbb{E}[Y_i | X_i, Z_i]}{\partial Z_i} = \beta_2 + \beta_3 X_i$$

Hypothesis tests

Due to sampling variation, the two groups will never have the exact same slope. But how do we assess if the differences in the slopes are “big enough” for us to say that the effect varies by group? We can test whether or not the effects for the two groups are different by testing the null hypothesis $H_0 : \beta_3 = 0$

$$\frac{\hat{\beta}_3}{\widehat{\text{se}}[\hat{\beta}_3]}$$

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 ? We already saw that $\hat{\beta}_1$ is the effect when $Z_i = 0$. What about other values of Z_i ? To calculate the sampling variances (and thus the SEs), we need to use the properties of variances. Here is the expression:

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

The variances here are the usual variances and the $\text{Cov}[\hat{\beta}_1, \hat{\beta}_3]$ is the covariance between the estimator of the two coefficients. Each of these can come from the variance-covariance matrix we learned about that week:

$$\mathbb{V}[\hat{\beta}] = \hat{\sigma}_u^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Here, \mathbf{X} is the matrix with four columns: a column of 1's, a column for X_i , a column for Z_i , and a column for the interaction $X_i \times Z_i$.

Let's calculate the SE for the effect of income for a Muslim country. We can use the `vcov()` function to get the variances and covariances (more on this in the next few weeks):

```
## 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 * v
sqrt(var.inter)
```

```
## [1] 0.3277283
```

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

```
## [1] 0.1940696
```

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. So if we wanted to estimate the effect of being a Muslim country with the associated SEs, we could use $1 - Z_i$ in place of 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, $\hat{\beta}_1$ is the slope on X_i when $1 - Z_i = 0$, or, rearranging, when $Z_i = 1$. We “tricked” R into calculating the standard errors for us:

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

```
##
## Call:
## lm(formula = fhrev ~ income * I(1 - muslim), data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.8460 -0.5705  0.0940  0.8517  2.6307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.3924     0.9969   4.406 2.03e-05 ***
## income           -0.5675     0.3277  -1.732  0.0855 .
## I(1 - muslim)    -5.7413     1.1338  -5.064 1.23e-06 ***
## income:I(1 - muslim)  2.4267     0.3642   6.662 5.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 145 degrees of freedom
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.6261
## F-statistic: 83.61 on 3 and 145 DF, p-value: < 2.2e-16
```

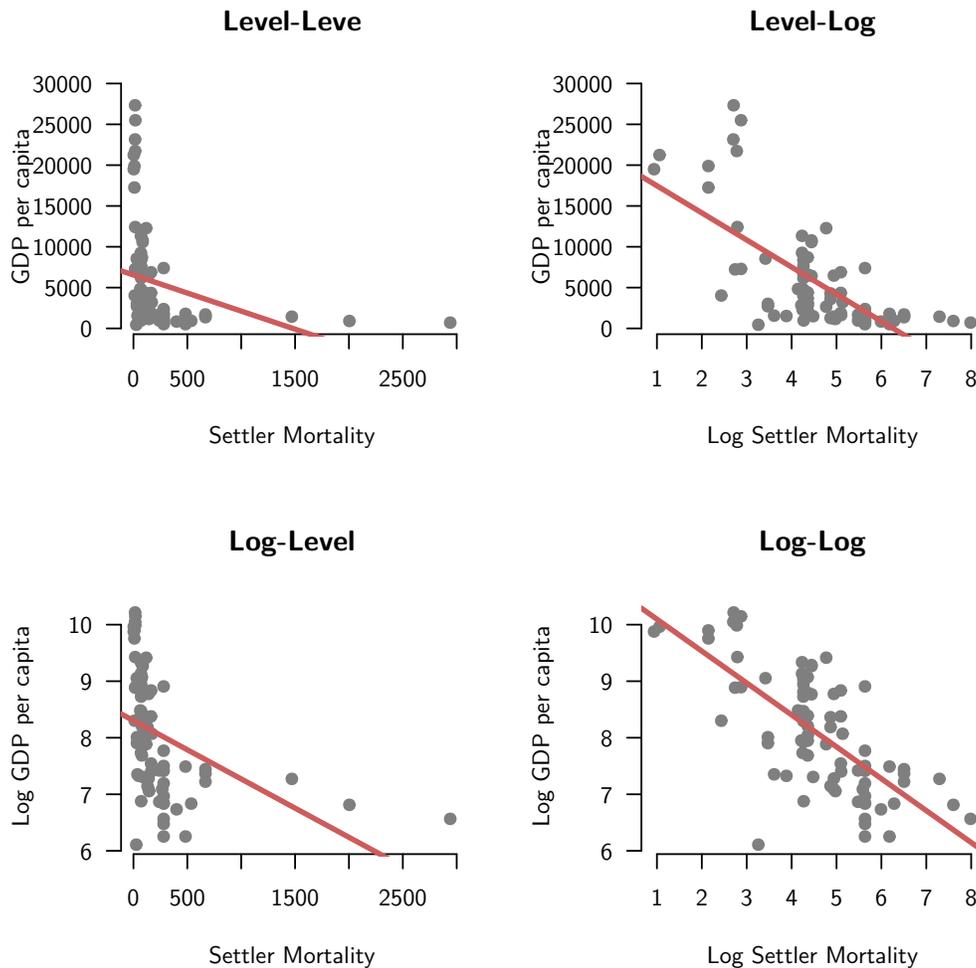
Notice that the SE is the same as we calculated before.

NONLINEAR FUNCTIONAL FORMS

What if we suspect or know that the relationship between X_i and Y_i is **non-linear**? Can we handle that in the linear model? Yes, it just means transforming these variables or including additional covariates.

Logs of random variables

One way to account for non-linearity is to transform X_i or Y_i using the natural logarithm. This approach is useful when X_i or Y_i are positive and right-skewed. We can see how transforming the variables from levels to logs changes the nature of the relationship and how the log-log relationship is clearly the one closest to linear:



Note that this changes the interpretation of β_1 . If we regress $\log(Y_i)$ on X_i , then $100 \times \beta_1$ can roughly be interpreted as the **percent increase** in Y_i associated with a one-unit increase in X_i . When we regress $\log(Y_i)$ on $\log(X_i)$, then β_1 is roughly the **percentage increase** in Y_i associated with a **one percent** increase in X_i . These interpretations are only useful for small increments, so you wouldn't want to use them with discrete r.v.s for instance. All of the different interpretations are summarized in this table:

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 to modeling nonlinearity is to model the relationship as a polynomial instead of a line. To do this, we can add a polynomial of X_i to account for the non-linearity:

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

This is similar to an “interaction” with itself: marginal effect of X_i varies as a function of X_i :

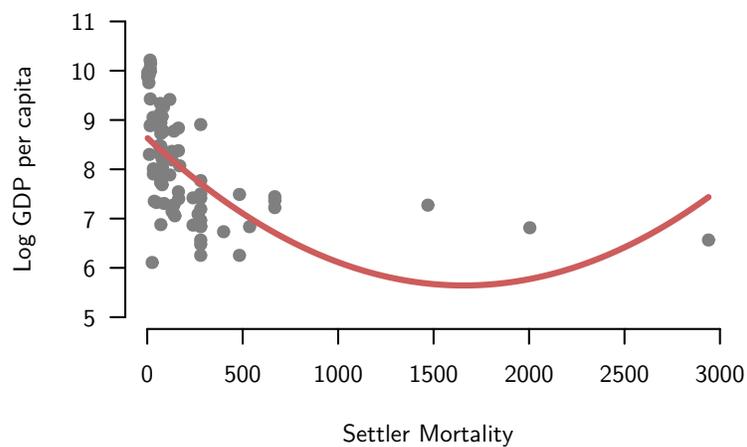
$$\frac{\partial \mathbb{E}[Y_i | X_i]}{\partial X_i} = \beta_1 + \beta_2 X_i$$

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

```
##
## Call:
## lm(formula = logpgp95 ~ raw.mort + I(raw.mort^2), data = ajr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.43698 -0.66321  0.00788  0.65436  1.63024
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.639e+00  1.378e-01  62.687 < 2e-16 ***
## raw.mort    -3.616e-03  6.638e-04  -5.447 5.77e-07 ***
## I(raw.mort^2) 1.091e-06  2.623e-07   4.157 8.19e-05 ***
```

```
## ---
## 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:  0.3211, Adjusted R-squared:  0.3037
## F-statistic: 18.45 on 2 and 78 DF,  p-value: 2.755e-07
```

We can plot the results using a function:



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):

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

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

Joint null hypotheses

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

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

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

- 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

Restricted versus unrestricted models

- Unrestricted model (alternative is true):

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

- Estimates:

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

- SSR from unrestricted model:

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

- 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:

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

- SSR from restricted model model:

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

- 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 β s
- 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

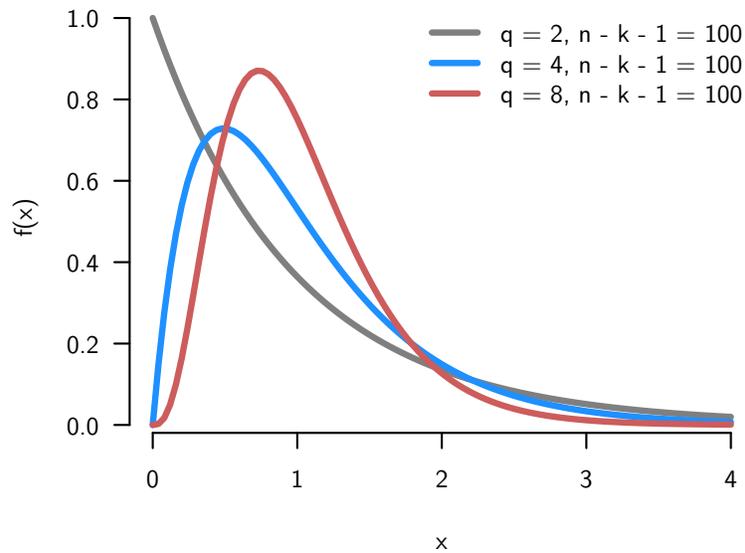
F statistic in R

```
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     147 452.13
## 2     145 284.09  2    168.04 42.885 2.337e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F distribution

```
curve(df(x, 2, 100), xlim = c(0,4), lwd = 3, col = "grey50", bty = "n", las = 1, ylab = "f(x)", xlab = "x")
curve(df(x, 4, 100), xlim = c(0,4), lwd = 3, col = "dodgerblue", add = TRUE)
curve(df(x, 8, 100), xlim = c(0,4), lwd = 3, col = "indianred", add = TRUE)
legend("topright", legend = c("q = 2, n - k - 1 = 100", "q = 4, n - k - 1 = 100", "q = 8, n - k - 1 = 100"), lwd
```



- Ratio of two χ^2 (Chi-squared) distributions

The F test

- The F test will test this null hypothesis, but what is the sampling distribution of this F statistic?
- Very similar to the t-test. We will assume either assumptions 1-5 and in large samples, or under 1-6 (including Normality).
- With these assumptions, when the null is true, then we have:

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

- Under the null hypothesis, we know that the F statistic follows an F distribution with degrees of freedom q and $n - (k + 1)$.
- Thus, we can perform a test of the null hypothesis by comparing our observed test statistic to the distribution of the statistic under the null.
- 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.
- If our calculated F statistic is large relative to the null distribution, then this means that there is more predictive power (bigger reductions in the SSR) than we would expect by random chance.
- To conduct the test, we simply choose an α , which has the same interpretation as always: the proportion of false positives you are willing to accept.

- Then we calculate the rejection region for the test. All F-tests are **one-sided tests**. Why? Because we only want to reject when the added covariates increase our predictive power (when the SSR goes up) and this is when the F statistic is big.
- So the rejection region is going to be 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] 3.087296
```

- We might also want to calculate p-values. These would be the 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] 0.00710471
```

F statistic for all variables

- Often, you'll an F-statistic reported along with the regression.
- This usually tests the null hypothesis of all the 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 ($\hat{\beta}_0 = \bar{Y}$)
- The SSR_r then would just be the sampling variation in Y :

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

Example of F-test for all variables

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

```
##
## Call:
## lm(formula = fhrev ~ income * growth, data = FishData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.0018 -0.9356  0.2241  0.9604  2.8338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1066     0.6225  -0.171   0.8643
## income        1.2922     0.1941   6.659 5.33e-10 ***
## growth       -0.6172     0.2383  -2.590  0.0106 *
## income:growth  0.2395     0.0753   3.180  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:  0.4332, Adjusted R-squared:  0.4215
## F-statistic: 36.95 on 3 and 145 DF,  p-value: < 2.2e-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?
- Yes, it turns out that 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:

```

set.seed(02138)
noise <- data.frame(matrix(rnorm(2100), nrow = 100, ncol = 21))
summary(lm(noise))

##
## Call:
## lm(formula = noise)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.1437 -0.5522  0.0697  0.6096  1.8470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0280393  0.1138198  -0.246  0.80605
## X2           -0.1503904  0.1121808  -1.341  0.18389
## X3             0.0791578  0.0950278   0.833  0.40736
## X4           -0.0717419  0.1045788  -0.686  0.49472
## X5             0.1720783  0.1140017   1.509  0.13518
## X6             0.0808522  0.1083414   0.746  0.45772
## X7             0.1029129  0.1141562   0.902  0.37006
## X8           -0.3210531  0.1206727  -2.661  0.00945 **
## X9           -0.0531223  0.1079834  -0.492  0.62412
## X10            0.1801045  0.1264427   1.424  0.15827
## X11            0.1663864  0.1109471   1.500  0.13768
## X12            0.0080111  0.1037663   0.077  0.93866
## X13            0.0002117  0.1037845   0.002  0.99838
## X14           -0.0659690  0.1122145  -0.588  0.55829
## X15           -0.1296539  0.1115753  -1.162  0.24872
## X16           -0.0544456  0.1251395  -0.435  0.66469
## X17            0.0043351  0.1120122   0.039  0.96923
## X18           -0.0807963  0.1098525  -0.735  0.46421
## X19           -0.0858057  0.1185529  -0.724  0.47134
## X20           -0.1860057  0.1045602  -1.779  0.07910 .
## X21            0.0021111  0.1081179   0.020  0.98447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9992 on 79 degrees of freedom

```

Multiple R-squared: 0.2009, Adjusted R-squared: -0.00142

F-statistic: 0.993 on 20 and 79 DF, p-value: 0.4797

- 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