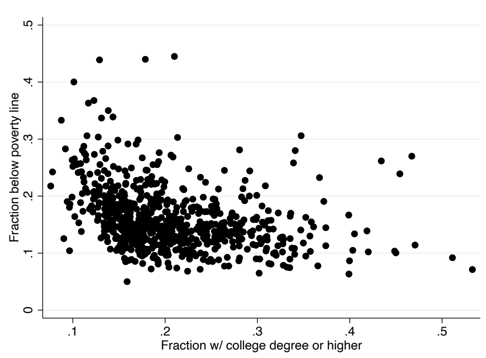
1 Polynomial Specifications

Not all relationships between variables are best captured by a linear model. For example, in the scatterplot below, the data look like a quadratic equation might fit the data better than a straight line. There might also be occasions when we have theoretical reasons to believe that a polynomial is the best model, like hours of studying and test scores or age and earnings. An important consequence of using a polynomial model is that the effect of a regressor will depend on the level of the regressor.



To estimate a **polynomial regression** of order r, we create new variables for different powers of the regressor and estimate a regression of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \cdots + \beta_r x_{1i}^r + u_i$$

```
. * Only Linear Term
. reg poor_share frac_coll, r

Linear regression                               Number of obs     =        741
                                                F(1, 739)         =      87.80
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1401
                                                Root MSE          =     5.0243

------------------------------------------------------------------------------
             |               Robust
  poor_share | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   frac_coll |  -.2917014   .0311316    -9.37   0.000    -.3528184   -.2305844
       _cons |   21.95317   .6855321    32.02   0.000     20.60735      23.299
------------------------------------------------------------------------------

.
. * Include Quadratic Term
. gen frac_coll2 = frac_coll^2

. reg poor_share frac_coll frac_coll2, r

Linear regression                               Number of obs     =        741
                                                F(2, 738)         =      84.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1909
                                                Root MSE          =      4.877

------------------------------------------------------------------------------
             |               Robust
  poor_share | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   frac_coll |  -1.156771   .1607295    -7.20   0.000    -1.472312   -.8412291
```

```
  frac_coll2 |   .0180537   .0034078     5.30   0.000     .0113634    .0247439
       _cons |   31.22327   1.786368    17.48   0.000      27.7163    34.73023
------------------------------------------------------------------------------
```

Above is the Stata output from a regression of the CZ-level percent of adults with a college degree or higher on the percent living below the poverty line, followed by a regression that also includes a quadratic term in the share college educated.

Q1: Suppose we are considering an increase in the share college educated in Boston, where the college-educated share in 2010 was about 40%. What would be a 1% increase in the college-educated share? A 1 percentage point increase?

Q2: Interpret the coefficient on `frac_coll` in the regression with only a linear term.

Q3: As a function of `frac_coll`, what is the predicted change in the poverty rate associated with a one-unit increase in `frac_coll` in the quadratic regression? (Hint: subtract your predicted value of `poor_share` when `frac_coll` $= x$ from the predicted value when `frac_coll` $= x+1$.)

Q4: What is the predicted change in the poverty rate associated with an increase in the college-educated share from 5% to 6%? Interpret.

Q5.a: What is the predicted change in the poverty rate associated with an increase in the college-educated share from 29% to 30%? Interpret.

Since we use multiple coefficients to calculate the effect of the college-educated share, we can't use the standard error on either of the coefficients by itself as the standard error for effects like those we calculated in questions 2-4. Instead, we'll apply a familiar formula from statistics to our answer from question 2:

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$
$$Var(\beta_1 + (1 + 2x)\beta_2) = Var(\beta_1) + (1 + 2x)^2Var(\beta_2) + 2(1 + 2x)Cov(\beta_1, \beta_2)$$

Note: Stata, R, and Python do not automatically output $Cov(\hat{\beta}_1, \hat{\beta}_2)$. You can use the command `matrix list e(V)` after your regression to display the variance-covariance matrix of the coefficients.

```
. matrix list e(V)

symmetric e(V)[3,3]
             frac_coll  frac_coll2       _cons
 frac_coll   .02583398
frac_coll2  -.00054008   .00001161
     _cons  -.28182377    .0057247   3.1911103
```

Q5.b: Use the above table to calculate the standard error on our estimate from question 4.

Rather than calculating it by hand, we can also get Stata to calculate the standard error for us by having it test if the expression is different from zero. Here is an example where $x = 5$ (e.g. a one-unit change from 5% to 6%). `lincom` stands for linear combinations of parameters.

```
. lincom frac_coll + frac_coll2*11

 ( 1)  frac_coll + 11*frac_coll2 = 0

------------------------------------------------------------------------------
  poor_share | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |  -.9581804   .1239248    -7.73   0.000    -1.201467   -.7148933
------------------------------------------------------------------------------
-
```

2 Log Specifications

2.1 Linear-Log Regressions

A **linear-log regression** has the following form:

$$y_i = \beta_0 + \beta_1 \log(x_{1i}) + \beta_2 x_{2i}$$

Because the logarithm is not linear, the effect of x_1 will depend on the level of x_1. In the case of logarithms, though, there is a useful approximation we can use. Subtracting as in question 2, we can find:

$$\begin{aligned}
\Delta y_i &= (\beta_0 + \beta_1 \log(x_{1i} + \Delta x_i) + \beta_2 x_{2i}) - (\beta_0 + \beta_1 \log(x_{1i}) + \beta_2 x_{2i}) \\
\Delta y_i &= \beta_1 \log(x_{1i} + \Delta x_i) - \beta_1 \log(x_{1i}) \\
\Delta y_i &= \beta_1 \log\left(\frac{x_{1i} + \Delta x_i}{x_{1i}}\right) \\
\Delta y_i &\approx \beta_1 \left(\frac{\Delta x_i}{x_{1i}}\right) \\
\Delta y_i &\approx \frac{\beta_1}{100} \left(\frac{100 \Delta x_i}{x_{1i}}\right)
\end{aligned}$$

In other words, a 1% change in x_1 is associated with a change of $0.01 \cdot \beta_1$ in y_i, holding x_2 constant. Make sure to take note that a 1% change is not the same as a 1 percentage point change.

Example 1 (Housing Prices) *Suppose we regress house price (in $1000s) on log square footage and number of bathrooms and estimate the following coefficients:*

$$\widehat{price} = -551 + 100.2 \cdot \log(SQFT) + 41 \cdot Baths$$

Q6: The median square footage of new homes built in the US in 1973 was 1,525. In 2010, it was 2,169. What is a 1% increase in square footage for a median home in 1973? In 2012?

Q7: Interpret the coefficient of log square footage.

2.2 Log-Linear Regressions

A **log-linear regression** has the following form:

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

By log subtraction rules, we find that the effect of a one-unit increase in x_1 with a corresponding change Δy_i will give us:

$$\begin{aligned} \log(y_i + \Delta y_i) - \log(y_i) &= (\beta_0 + \beta_1(x_{1i} + 1) + \beta_2 x_{2i}) - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \\ \log\left(\frac{y_i + \Delta y_i}{y_i}\right) &= \beta_1 \\ \frac{\Delta y_i}{y_i} &\approx \beta_1 \end{aligned}$$

We can multiply both sides of this equation by 100 to put this in percent changes: a one-unit change in x_1 is associated with a $(100 \cdot \beta_1)\%$ change in y, holding x_2 constant.

Example 2 (Returns to Education) *Suppose we regress hourly wage on years of education and years of experience and estimate the following coefficients:*

$$\widehat{\log(wage)} = 7.023 + 0.005 \cdot Edu + 0.024 \cdot Exper$$

Q8: Interpret the coefficient on years of education.

2.3 Log-Log Regressions

A **log-log regression** has the following form:

$$\log(y_i) = \beta_0 + \beta_1 \log(x_{1i}) + \beta_2 x_{2i}$$

Q9: Using the logarithm fact that $\log(u) - \log(v) = \log\left(\frac{u}{v}\right)$ and the approximation $\log(w+1) \approx w$ for small w, interpret a one-unit change in x_1. (Hint) Think about combining the interpretations of the linear-log and the log-linear regressions!

Example 3 (Cobb-Douglas Production) *Suppose we estimated a firm's production function and found the following coefficients:*

$$\widehat{\log(output)} = 4.461 + 0.227 \cdot \log(labor) + 0.76 \cdot \log(capital)$$

where labor, capital, and output are measured in dollars.

Q10: Interpret the coefficient on labor. What is the elasticity of output with respect to labor? What are the units of this elasticity?

3 F-statistics

We use F-test (allows heteroskedasticity) when there are multiple restrictions in the null hypothesis. For example, let's think about testing whether region has an effect on income in the following regression:

$$Income_i = \beta_0 + \beta_1 Education_i + \beta_2 TestScore_i + \beta_3 North_i + \beta_4 South_i + \beta_5 West_i + u_i$$

Q11: What is the omitted category for region?

Q12 a: What is the null and alternative hypothesis?

Q12 b: Why can't we use the p-value from our individual regressions and reject if one of them is significant?

Q13: What is the number of restrictions, q?

```
. reg income education iq  north south west, r
Linear regression                                      Number of obs =    6135
                                                       F(  5,  6129) =  141.91
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1681
                                                       Root MSE      =   31053
------------------------------------------------------------------------------
             |               Robust
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   3280.565   227.7613    14.40   0.000     2834.073    3727.058
          iq |   .2612454   .0165861    15.75   0.000     .2287308    .2937599
       north |  -4701.908   1382.968    -3.40   0.001    -7413.012   -1990.804
       south |  -4002.809   1278.274    -3.13   0.002    -6508.674   -1496.943
        west |  -2177.862   1450.285    -1.50   0.133    -5020.931    665.2069
       _cons |  -15685.47    2977.06    -5.27   0.000    -21521.55   -9849.384
------------------------------------------------------------------------------
. test  north south west
 ( 1)  north = 0
 ( 2)  south = 0
 ( 3)  west = 0
       F(  3,  6129) =    4.88
            Prob > F =    0.0022
```

Q14: How do we interpret the results of the test? Do we accept or reject the null hypothesis at the 5% level? What does that mean?

4 Calculating the P-Value and Critical Value

It's important to note, however, that Stata calculates the p-value incorrectly when using the `test` and `testparm` commands (as does R with `waldtest`).

The issue is the you want the F(q, ∞) distribution and not the F(q, N$-$K) distribution. q, the number of restrictions, are your numerator degrees of freedom, while N$-$K are your denominator degrees of freedom. Both Stata and R give p-values from the F(q, N$-$K) distribution. The test statistic is correct and does not change, but the p-value is incorrect.

You can calculate the p-value separately using the `chi2()` distribution function by using the F-statistic from `test` and the numerator degrees of freedom (the number of restrictions).

First, you can get the F-statistic from `test` or `testparm`:

```
reg yvar xvar1 xvar2 xvar3, robust
test xvar2 xvar3
```

In R, this is done with `waldtest` from the `lmtest` library.

```
mod1<-lm(calories ~ x2+ x3+ x4, data = snap)
mod1.fstatreg <-lm(calories ~ x3, data = snap)
ftest <- waldtest(mod1.fstatreg, mod1, vcov= vcovHC(mod1, type = "HC1"))

#display results
ftest
```

However, it is also possible to just increase the denominator degrees of freedom to a really big number 10^9 and add that as an option to `test` or `testparm`:

```
test xvar2 xvar3, df(1e+9)
```

Next, we can also calculate the critical values for a given numerator degrees of freedom and significance level for the F(q, ∞) distribution.

If we have two degrees of freedom and one degree of freedom and are using the $\alpha = 0.05$ significance level, we can calculate the critical values in Stata as such:

```
. display "Critical value = " invchi2(2,0.95)/2
Critical value = 2.9957323

. display "Critical value = " invchi2(1,0.95)
Critical value = 3.8414588
```

And in R:

```
> qchisq(0.95, 2)/2
[1] 2.995732
> qchisq(0.95, 1)
[1] 3.841459
```

And in Python:

```
> stats.f.ppf(0.95, 2, INF)
[1] 2.995732
> stats.f.ppf(0.95, 1, INF)
[2] 3.841459
```

Using this, we can then calculate more accurate p-values and critical values with the numerator dergrees of freedom and the F-statistic. Note that r(df) and r(F) in these examples are just placeholders for plugging in the corresponding degrees of freedom and the F-statistic found earlier.

```
reg yvar xvar1 xvar2 xvar3, robust
test xvar2 xvar3, df(1e+9)

display "p-value = " 1-chi2(r(df), r(df)*r(F))
display "5% critical value = " invchi2(r(df),0.95)/r(df)
```

And in R:

```
mod1<-lm(calories ~ x2+ x3+ x4, data = snap)
mod1.fstatreg <-lm(calories ~ x3, data = snap)
ftest <- waldtest(mod1.fstatreg, mod1, vcov=
vcovHC(mod1, type = "HC1"))

#display results
ftest

#get correct p-value
1-pchisq(ftest$F * ftest$Df,ftest$Df)

#p-value and critical value
"p-value" = 1 - pchisq(r(df)*r(F), r(df))
"Critical value" = qchisq(0.95, r(df))/r(df)
```

And in Python:

```
test_results = res.wald_test([
    \xvar1 = 0",
    \xvar2 = 0",
    \xvar3 = 0",
], scalar=True, use_f=True)
f_stat = test_results.fvalue
df = test_results.df_num
true_p = 1 { stats.chi2.cdf(fstat * df, df)
INF = 10 ** 10
crit_val = stats.f.ppf(.95, df, INF)
```

5 Interaction Terms

We can use interaction terms to calculate the relationship between two variables among different groups. Let's say we want to estimate the relationship between parent income and income, but we believe this may differ for those who have had a parent incarcerated. One way we can do this is to run separate regressions for the two groups:

For those without a parent incarcerated:

```
. regress wages parent_income if parent_incarcerated == 0 , r

Linear regression                               Number of obs     =      3,929
                                                F(1, 3927)        =     124.77
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0581
                                                Root MSE          =      51367

-------------------------------------------------------------------------------
              |               Robust
        wages | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
--------------+----------------------------------------------------------------
parent_income |   .2994191   .0268057    11.17   0.000     .2468647    .3519735
        _cons |   43025.92   1294.802    33.23   0.000     40487.37    45564.47
-------------------------------------------------------------------------------
```

For those with a parent who has been incarcerated:

```
.  regress wages parent_income if parent_incarcerated == 1 , r

Linear regression                               Number of obs     =         77
                                                F(1, 75)          =       0.10
                                                Prob > F          =     0.7514
                                                R-squared         =     0.0008
                                                Root MSE          =      30463

-------------------------------------------------------------------------------
              |               Robust
        wages | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
--------------+----------------------------------------------------------------
parent_income |   .0268471   .0844378     0.32   0.751    -.1413616    .1950558
        _cons |   38447.41   4290.696     8.96   0.000     29899.91    46994.92
-------------------------------------------------------------------------------
```

Regression with interaction term included:

```
. gen interaction_term = parent_income * parent_incarcerated

.  regress wages parent_incarcerated parent_income interaction_term  , r

Linear regression                               Number of obs     =      4,006
                                                F(3, 4002)        =      48.10
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0598
                                                Root MSE          =      51054

-------------------------------------------------------------------------------------
                    |               Robust
              wages | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
--------------------+----------------------------------------------------------------
parent_incarcerated |  -4578.508   4430.254    -1.03   0.301    -13264.27    4107.256
      parent_income |   .2994191   .0268122    11.17   0.000     .2468522    .3519861
   interaction_term |   -.272572   .0875808    -3.11   0.002    -.4442791   -.1008649
              _cons |   43025.92   1295.119    33.22   0.000     40486.76    45565.07
-------------------------------------------------------------------------------------
```

Q15: How do the coefficients in the first two regressions relate to the coefficient on the interaction term in the third regression?

Q16: Say a person's parent's income was X. Using only the third regression, what is their predicted income if their parents were incarcerated? If they were not?