Applied Statistics and Econometrics

Giuseppe Ragusa Luiss University

Lecture 9: Nonlinear Regression Functions

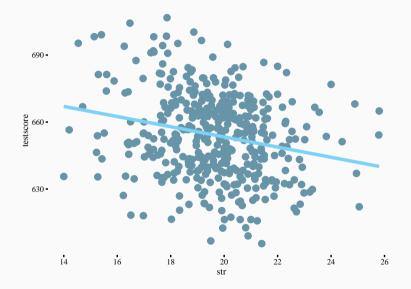
Outline

- Nonlinear regression functions general comments
- Nonlinear functions of one variable
- Nonlinear functions of two variables: interactions
- Application to the California Test Score data set

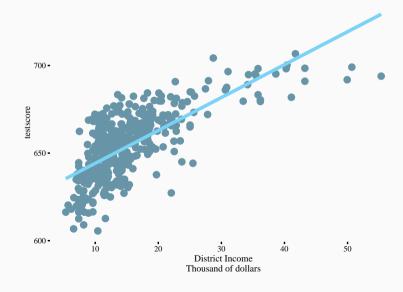
Nonlinear regression functions

- The regression functions so far have been linear in the X's
- But the linear approximation is not always a good one
- The multiple regression model can handle regression functions that are nonlinear in one or more X.

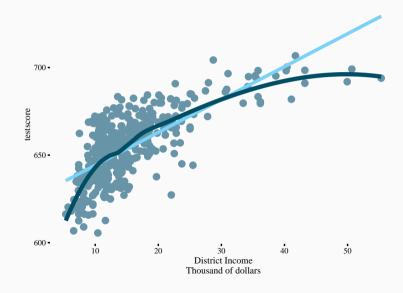
The testscore – str relation looks linear (maybe)...



But the testscore - income relation looks nonlinear...



But the testscore - income relation looks nonlinear...



Nonlinear Regression Population Regression Functions – General Ideas

- If a relation between Y and X is nonlinear: The effect on Y of a change in X depends on the value of X that is, the marginal effect of X is not constant
- A linear regression is mis-specified: the functional form is wrong
- The estimator of the effect on Y of X is biased: in general it isn't even right on average.
- ullet The solution is to estimate a regression function that is **nonlinear** in X

The general nonlinear population regression function

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i, \quad i = 1, \dots, n$$

Assumptions

- 1. $E(u_i|X_{1i},X_{2i},...,X_{ki})=0$ (same); implies that f is the **conditional expectation** of Y given the X's.
- 2. $(X_{1i}, X_{2i}, \dots, X_{ki}, Y_i)$ are i.i.d. (same).
- 3. Big outliers are rare (same idea; the precise mathematical condition depends on the specific f).
- 4. No perfect multicollinearity (same idea; the precise statement depends on the specific f).

The change in Y associated with a change in X_1 , holding X_2, \ldots, X_k constant is:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) - f(X_1, X_2, \dots, X_k)$$

Nonlinear Functions of a Single Independent Variable

We'll look at two complementary approaches:

Polynomials in X

The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial

Logarithmic transformations

Y and/or X is transformed by taking its logarithm this gives a "percentages" interpretation that makes sense in many applications

Polynomials in X

Approximate the population regression function by a polynomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \dots + \beta_r X_i^r + u_i$$

- This is just a linear multiple regression except that the regressors are powers of X!
- Estimation, hypothesis testing, etc. proceeds as in multiple regression estimated by OLS
- The coefficients are *difficult* to interpret, but the regression function itself is interpretable

Example: the testscore – income relation

*Income*_i= average district income in the ith district (thousands of dollars per capita)

• Quadratic specification:

$$testscore_i = \beta_0 + \beta_1 incomei + \beta_2 (incomei)^2 + u_i$$

Cubic specification:

$$testscore_i = \beta_0 + \beta_1 income_i + \beta_2 (income_i)^2 + \beta_3 (income_i)^3 + u_i$$

Estimation of the quadratic specification in $\ensuremath{\mathbb{R}}$

```
## In R, function of regressor can be used by using the I() construct
lm1 <- lm(testscore ~ income + I(income^2), data = CASchools)</pre>
summary_rob(lm1)
##
## Coefficients:
##
               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 607.30174 2.90175 209.29 <2e-16
## income 3.85099 0.26809 14.36 <2e-16
## I(income^2) -0.04231 0.00478 -8.85 <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 13 on 417 degrees of freedom
## Multiple R-squared: 0.556, Adjusted R-squared: 0.554
## F-statistic: 857 on 2 and Inf DF, p-value: <2e-16
```

Interpreting the estimated regression function:

$$testscore = \underset{(2.902)}{607.302} + \underset{(0.268)}{3.851} \\ income - \underset{(0.005)}{0.042} \\ income^2$$

(a) Plot predicted value:



Interpreting the estimated regression function, ctd:

(b) Compute "effects" for different X

$$testscore = \underset{(2.902)}{607.302} + \underset{(0.268)}{3.851} \\ income - \underset{(0.005)}{0.042} \\ income^2$$

Change in testscore when income increases \$1,000, from \$5,000

$$\Delta testscore = [\beta_0 + (5+1)\beta_1 + (5+1)^2\beta_2] - [\beta_0 + 5\beta_1 + 5^2\beta_2]$$
$$= 1 \times \beta_1 + [6^2 - 5^2]\beta_2$$

Which, using the estimated β 's, is

$$\Delta testscore = 3.85 - 0.47 = 3.39$$

Interpreting the estimated regression function, ctd:

$$\textit{testscore} = \underset{(2.902)}{607.302} + \underset{(0.268)}{3.851} \\ \textit{income} - \underset{(0.005)}{0.042} \\ \textit{income}^2$$

Estimated "effects" of an increase of \$1,000 at different values of X:

	$\widehat{\Delta testscore}$
from 5 to 6	3.39
from 25 to 26	1.69
from 45 to 46	0.00
from 53 to 54	-0.68
from 63 to 64	-1.52

Interpreting the estimated regression function, ctd:

$$\textit{testscore} = \underset{(2.902)}{607.302} + \underset{(0.268)}{3.851} \\ \textit{income} - \underset{(0.005)}{0.042} \\ \textit{income}^2$$

Estimated "effects" of an increase of \$2,000 at different values of X:

	$\widehat{\Delta testscore}$
from 5 to 7	6.69
from 25 to 27	3.30
from 45 to 47	-0.08
from 53 to 55	-1.44
from 63 to 65	-3.13

Estimation of a cubic specification in R

```
## We use T() twice
lm2 <- lm(testscore ~ income + I(income^2) + I(income^3), data = CASchools)</pre>
summary_rob(1m2)
##
## Coefficients:
##
               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.00e+02 5.10e+00 117.61 < 2e-16
## income 5.02e+00 7.07e-01 7.10 1.3e-12
## I(income^2) -9.58e-02 2.90e-02 -3.31 0.00094
## I(income^3) 6.85e-04 3.47e-04 1.98 0.04826
## ___
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 13 on 416 degrees of freedom
## Multiple R-squared: 0.558, Adjusted R-squared: 0.555
## F-statistic: 811 on 3 and Inf DF, p-value: <2e-16
```

Testing the null hypothesis of linearity

Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

- H_0 : population coefficients on $income^2$ and $income^3 = 0$
- H_1 : at least one of these coefficients is nonzero.

```
wald_test(lm2, testcoef = c("I(income^2)", "I(income^3)"))

## Wald test
##
## Null hypothesis:
## I(income^2) = 0
## I(income^3) = 0
##
## q W pvalue
## 2 75 4.3e-17
```

he hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

Testing the null hypothesis of linearity

Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

- H_0 : population coefficients on $income^2$ and $income^3 = 0$
- H_1 : at least one of these coefficients is nonzero.

```
wald_test(lm2, testcoef = c("I(income^2)", "I(income^3)"))

## Wald test
##
## Null hypothesis:
## I(income^2) = 0
## I(income^3) = 0
##
## q W pvalue
## 2 75 4.3e-17
```

The hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

Polynomial in multiple regression

• If we have a multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

and we think there is a nonlinear relationship between Y and one of the X's, say X_{ki} , we use a polynomial of that variable:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \beta_{k+1} X_{ki}^2 + \beta_{k+2} X_{ki}^3 + \ldots + u_i$$

Estimation proceed as usual

Polynomial in multiple regression

```
lm3 <- lm(testscore ~ str + expenditure + english + income + I(income^2), data = CASchools)</pre>
summary_rob(1m3)
##
## Coefficients:
##
               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.41e+02 1.16e+01 55.46 < 2e-16
        -4.39e-01 3.65e-01 -1.20 0.23
## str
## expenditure -9.33e-04 1.09e-03 -0.86 0.39
## english -4.51e-01 3.08e-02 -14.63 < 2e-16
## income 2.67e+00 2.35e-01 11.33 < 2e-16
## I(income^2) -2.44e-02 4.36e-03 -5.60 2.1e-08
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 10 on 414 degrees of freedom
## Multiple R-squared: 0.723, Adjusted R-squared: 0.72
## F-statistic: 1.39e+03 on 5 and Inf DF, p-value: <2e-16
```

Summary: polynomial regression functions

- Estimation: by OLS after defining new regressors
- Coefficients have complicated interpretations
- To interpret the estimated regression function:
 - plot predicted values as a function of X compute predicted $\Delta Y/\Delta X$ at different values of x
 - Hypotheses concerning degree r can be tested by t- and Wald-tests on the appropriate (blocks of) variable(s).
- Choice of degree r plot the data; t- and Wald-tests, check sensitivity of estimated effects; judgment.
- Or use model selection criteria (maybe later)

Logarithmic functions of Y and/or X

- ln(X) = the natural logarithm of X (we often uses log interchangeably)
- Logarithmic transforms permit modeling relations in "percentage" terms (like elasticities), rather than linearly.

Here is why:

$$\ln(x + \Delta x) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$$

In calculus,

$$\frac{d\ln(x)}{dx} = \frac{1}{x}$$

The three log regression specifications:

Case	Population regression function	
I. linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_{1i}) + \ldots + u_i$	
II. log-linear	$ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \ldots + u_i$	
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{1i}) + \ldots + u_i$	

 Table 1: Logarithmic transformation

- The interpretation of the slope coefficient differs in each case.
- The interpretation is found by applying the general "before and after" rule: "figure out the change in Y for a given change in X."
- Each case has a natural interpretation (for small changes in X)

Linear-log population regression function

1. Linear-log model

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

What is the interpretation of β_1 ? We calculate

$$\underbrace{E[Y_i|X_i = x + \Delta x] - E[Y_i|X_i = x]}_{\text{for the linear model this is } \beta_1 \Delta x}$$

For the linear-log

$$E[Y_i|X_i = x + \Delta x] - E[Y_i|X_i = x] = \beta_1 \left(\ln(x + \Delta x) - \ln(x) \right) \approx \beta_1 \frac{\Delta x}{x}$$

Linear-log population regression function

1. Linear-log model

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

What is the interpretation of β_1 ? We calculate

$$\underbrace{E[Y_i|X_i = x + \Delta x] - E[Y_i|X_i = x]}_{\text{for the linear model this is } \beta_1 \Delta x}$$

For the linear-log

$$E[Y_i|X_i = x + \Delta x] - E[Y_i|X_i = x] = \beta_1 \left(\ln(x + \Delta x) - \ln(x) \right) \approx \beta_1 \frac{\Delta x}{x}$$

- a 100% ($\Delta x/x = 1$) increase in X (multiplying X by 2) is associated with a β_1 change in Y.
- a 1% ($\Delta x/x = 0.01$) increase in X (multiplying X by 1.01) is associated with a .01 β_1 change in Y.

Example: *testscore* **vs.** ln(*Income*)

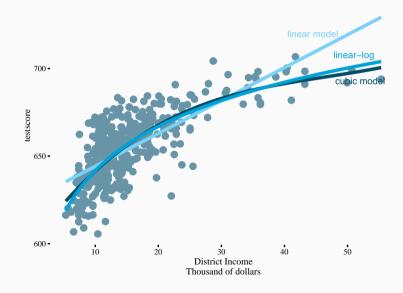
```
lm4 <- lm(testscore ~ I(log(income)), data = CASchools) ## Note: in R, natural logarithm is 'log'
summary_rob(lm4)
##
## Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
##
## (Intercept) 557.83 3.84 145.3 <2e-16
## I(log(income)) 36.42 1.40 26.1 <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 13 on 418 degrees of freedom
## Multiple R-squared: 0.563, Adjusted R-squared: 0.561
## F-statistic: 680 on 1 and Inf DF, p-value: <2e-16
```

Example: *testscore* **vs.** ln(*Income*)

```
lm4 <- lm(testscore ~ I(log(income)), data = CASchools) ## Note: in R, natural logarithm is 'log'
summary_rob(lm4)
##
## Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
##
## (Intercept) 557.83 3.84 145.3 <2e-16
## I(log(income)) 36.42 1.40 26.1 <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 13 on 418 degrees of freedom
## Multiple R-squared: 0.563, Adjusted R-squared: 0.561
## F-statistic: 680 on 1 and Inf DF, p-value: <2e-16
```

- a 1% increase in Income is associated with an increase in *testscore* of 0.36 points.
- Standard errors, confidence intervals, R^2 all the usual tools of regression apply here.

Example: *testscore* **vs.** ln(*income*)



Log-linear population regression function

2 Log-linear model

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

What is the interpretation of β_1 ? We calculate

$$E[\ln Y_i|X_i=x+\Delta x]-E[\ln Y_i|X_i=x]\approx E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right]$$

For the log-linear

$$E\left[\frac{\Delta y}{y}\middle|X\uparrow\Delta x\right] = \beta_1\left(x + \Delta x - x\right)$$
$$\approx \beta_1 \Delta x$$

Log-linear population regression function

2 Log-linear model

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

What is the interpretation of β_1 ? We calculate

$$E[\ln Y_i|X_i=x+\Delta x]-E[\ln Y_i|X_i=x]\approx E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right]$$

For the log-linear

$$E\left[\frac{\Delta y}{y} \middle| X \uparrow \Delta x\right] = \beta_1 \left(x + \Delta x - x\right)$$
$$\approx \beta_1 \Delta x$$

• a 1 unit increase ($\Delta x = 1$) increase in X is associated with a $100 \times \beta_1\%$ change in Y.

Example: California Data

```
lm4 <- lm(I(log(testscore)) ~ income, data = CASchools)</pre>
summary_rob(lm4)
##
## Coefficients:
##
              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.439362 0.002894 2225.2 <2e-16
## income 0.002844 0.000175 16.2 <2e-16
## ___
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.021 on 418 degrees of freedom
## Multiple R-squared: 0.498, Adjusted R-squared: 0.497
## F-statistic: 264 on 1 and Inf DF, p-value: <2e-16
```

Log-log population regression function

3 Log-log model

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

What is the interpretation of β_1 ? We calculate

$$E[\ln Y_i|X_i=x+\Delta x]-E[\ln Y_i|X_i=x]\approx E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right]$$

For the log-log

$$E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right] = \beta_1\left(\ln(x+\Delta x)-\ln x\right) \approx \beta_1\frac{\Delta x}{x}$$

Notice

$$\underbrace{\frac{100 \times \frac{\Delta y}{y}}{y}}_{\text{% change}} = \beta_1 \underbrace{\frac{100 \times \frac{\Delta x}{x}}{x}}_{\text{% change}}$$

Log-log population regression function

3 Log-log model

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

What is the interpretation of β_1 ? We calculate

$$E[\ln Y_i|X_i=x+\Delta x]-E[\ln Y_i|X_i=x]\approx E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right]$$

For the log-log

$$E\left[\left.\frac{\Delta y}{y}\right|X\uparrow\Delta x\right] = \beta_1\left(\ln(x+\Delta x)-\ln x\right) \approx \beta_1\frac{\Delta x}{x}$$

Notice

$$\underbrace{\frac{\Delta y}{y}}_{\text{% change}} = \beta_1 \underbrace{\frac{\Delta x}{x}}_{\text{% change}}$$

• a 1% increase ($\Delta x/x = 0.01$) in X is associated with a β_1 % change in Y.

Example: *In*(*testscore*) **vs.** *In*(*income*)

```
lm5 <- lm(I(log(testscore)) ~ I(log(income)), data = CASchools)</pre>
summary_rob(lm5)
##
## Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
##
## (Intercept) 6.33635 0.00592 1069.5 <2e-16
## I(log(income)) 0.05542 0.00214 25.8 <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.019 on 418 degrees of freedom
## Multiple R-squared: 0.558, Adjusted R-squared: 0.557
## F-statistic: 668 on 1 and Inf DF, p-value: <2e-16
```

Example: *In*(*testscore*) **vs.** *In*(*income*)

```
lm5 <- lm(I(log(testscore)) ~ I(log(income)), data = CASchools)</pre>
summary_rob(lm5)
##
## Coefficients:
##
                 Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.33635 0.00592 1069.5 <2e-16
## I(log(income)) 0.05542 0.00214 25.8 <2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.019 on 418 degrees of freedom
## Multiple R-squared: 0.558, Adjusted R-squared: 0.557
## F-statistic: 668 on 1 and Inf DF, p-value: <2e-16
```

An 1% increase in *income* is associated with an increase of .0554% in *testscore* (*income* up by a factor of 1.01, testscore up by a factor of 1.000554)

Graphics representation of the regression function

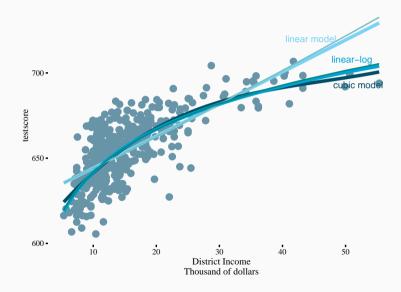
- Notice that the linear-log, the log-linear and the log-log model are linear in the parameters, but they imply a nonlinear relation between Y_i and X_i
- For instance,

$$\widehat{\ln Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i \implies \widehat{Y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

$$\widehat{\ln Y_i} = \hat{\beta}_0 + \hat{\beta}_1 \ln X_i \implies \widehat{Y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 \ln X_i)$$

Let's graph them all together

Graphical



Summary: Logarithmic transformations

- Three cases, differing in whether Y and/or X is transformed by taking logarithms.
- The regression is linear in the new variable(s) ln(Y) and/or ln(X), and the coefficients can be estimated by OLS.
- Hypothesis tests and confidence intervals are now implemented and interpreted "as usual."
- The interpretation of β_1 differs from case to case.
- The choice of specification (functional form) should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values

CPS Data set

For the purpose of this exercise, let

- earnings: average hourly earnings (sum of annual pretax wages, salaries, tips, and bonuses, divided by the number of hours worked annually).
- learnings: log average hourly earnings (sum of annual pretax wages, salaries, tips, and bonuses, divided by the number of hours worked annually).
- education: number of years of education.
- college: =1 if individual has at least a college degree
- female: =1 if "female".
- age: age in year.
- ullet neast: =1 if individual live in the Norteast region
- **mwest:** =1 if individual live in the Midwest region
- **south:** =1 if individual live in the South region
- west: =1 if the individual live in the West region

Regression: average hourly earnings

```
lm_CPS <- lm(earnings ~ college + female + age + mwest + south + west, data = CPS)</pre>
summary_rob(lm_CPS)
##
## Coefficients:
##
             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.70961 0.52926 20.23
                                        <2e-16
## college 8.43970 0.29790 28.33 <2e-16
## female -3.62800 0.25135 -14.43 <2e-16
## age 0.17317 0.01093 15.84
                                        <2e-16
## mwest -0.88502 0.35321 -2.51
                                        0.012
## south -0.00497 0.37790 -0.01
                                        0.989
## west -0.34601 0.37365 -0.93
                                        0.354
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 8.9 on 4993 degrees of freedom
## Multiple R-squared: 0.217, Adjusted R-squared: 0.216
## F-statistic: 1.18e+03 on 6 and Inf DF. p-value: <2e-16
```

Regression: average hourly earnings

```
lm_CPS <- lm(learnings ~ college + female + age + mwest + south + west, data = CPS)</pre>
summary_rob(lm_CPS)
##
## Coefficients:
##
              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.339565 0.031259 74.84 <2e-16
## college 0.434142 0.014596 29.74 <2e-16
## female -0.192023 0.013754 -13.96 <2e-16
## age 0.009910 0.000636 15.58
                                        <2e-16
## mwest -0.039427 0.018992 -2.08 0.038
## south -0.016779 0.020097 -0.83 0.404
## west
         -0.019039 0.020146 -0.95 0.345
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.49 on 4993 degrees of freedom
## Multiple R-squared: 0.205, Adjusted R-squared: 0.204
## F-statistic: 1.32e+03 on 6 and Inf DF. p-value: <2e-16
```

Regression: average hourly earnings

```
lm_CPS <- lm(learnings ~ college + female + age + I(age^2) + mwest + south + west,</pre>
   data = CPS)
summary rob(lm CPS)
##
## Coefficients:
##
              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.24e+00 9.53e-02 13.01
                                        <2e-16
## college 4.29e-01 1.43e-02 29.93
                                        <2e-16
## female -1.91e-01 1.36e-02 -14.09 <2e-16
## age 6.68e-02 4.72e-03 14.15
                                        <2e-16
## I(age^2) -6.81e-04 5.62e-05 -12.11 <2e-16
## mwest -3.66e-02 1.86e-02 -1.96
                                        0.05
## south -2.66e-02 1.98e-02 -1.34
                                        0.18
                                       0.34
## west
             -1.88e-02 1.98e-02 -0.95
## ____
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.48 on 4992 degrees of freedom
## Multiple R-squared: 0.229.Adjusted R-squared: 0.228
## F-statistic: 1.53e+03 on 7 and Inf DF, p-value: <2e-16
```

Regional differences

Do there appear to be important regional differences?

```
wald_test(lm_CPS, testcoef = c("mwest", "south", "west"), vcov = vcovHC)

## Wald test
##
## Null hypothesis:
## mwest = 0
## south = 0
## west = 0
##
## q W pvalue
## 3 4 0.26
```

Regression: average hourly earnings (years of education)

```
lm_CPS2 <- lm(learnings ~ education + female + age + I(age^2) + mwest + south + west,</pre>
   data = CPS)
summary rob(lm CPS2)
##
## Coefficients:
##
              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.07e-01 9.97e-02 1.07
                                        0.2848
## education 9.53e-02 2.75e-03 34.67
                                        <2e-16
## female -2.04e-01 1.32e-02 -15.44 <2e-16
## age 6.55e-02 4.62e-03 14.17 <2e-16
## I(age^2) -6.70e-04 5.49e-05 -12.20
                                        <2e-16
## mwest -4.76e-02 1.81e-02 -2.63
                                        0.0085
## south -2.93e-02 1.94e-02 -1.51
                                        0.1303
             -1.30e-02 1.92e-02 -0.68
## west
                                        0.4990
## ____
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.47 on 4992 degrees of freedom
## Multiple R-squared: 0.268.Adjusted R-squared: 0.267
## F-statistic: 1.86e+03 on 7 and Inf DF, p-value: <2e-16
```