MODEL SPECIFICATION

TODAY

- Dummy variables
  - Intercept
  - Slopes
- Polynomials - simple non-linear modeling
- Logarithms - elastic modeling

DUMMIES...

When things are not numeric

DUMMIES - DEFINITION

Qualitative attributes not easily measured as a number

Male/female

Region

Examples:

Likert scale

Industry type

No just 0/1 variables

DUMMIES - INTERCEPT

Easiest case:

Only two categories

Dummy moves intercept

BASIC SET-UP

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

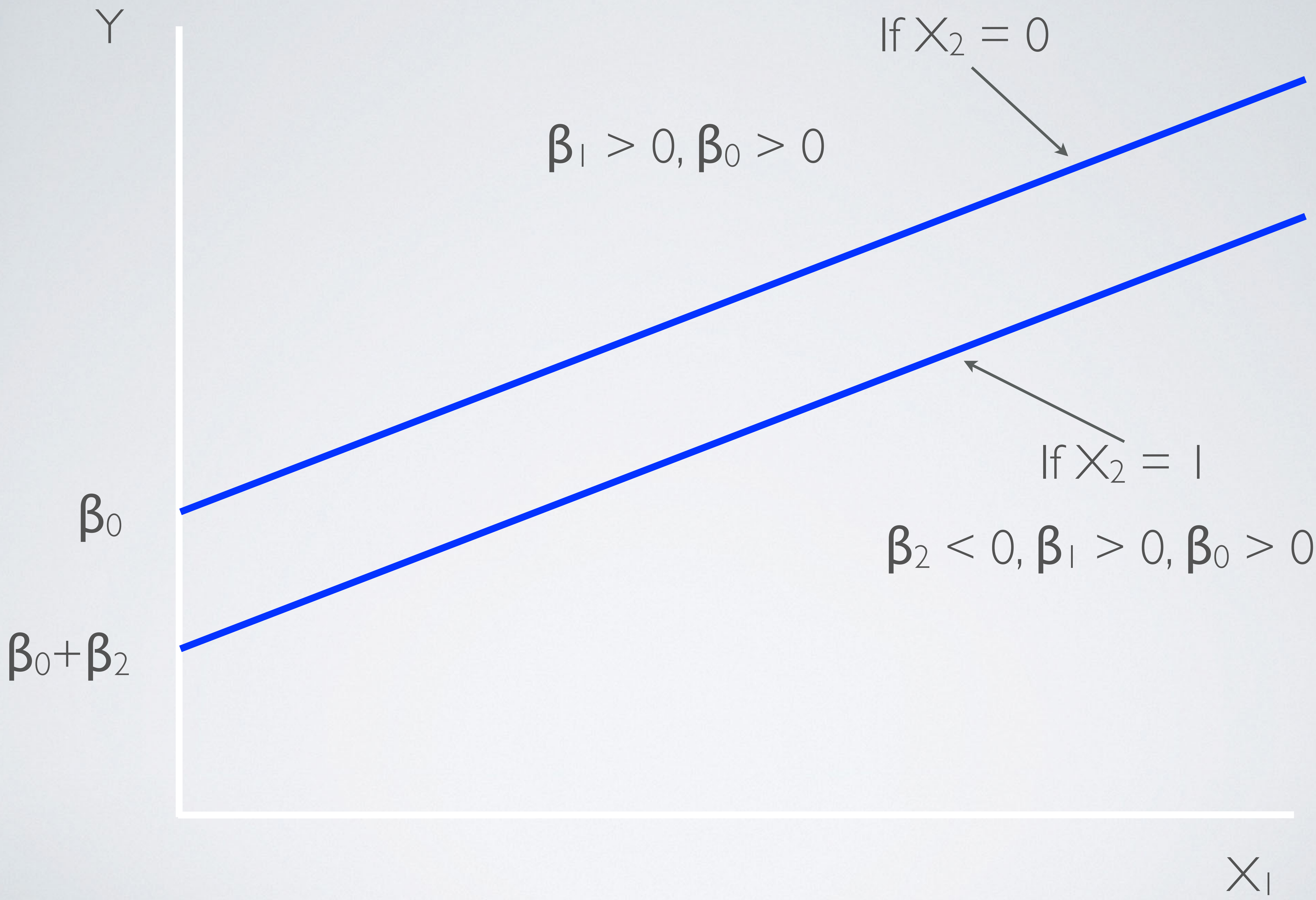
X_1: Continuous variable

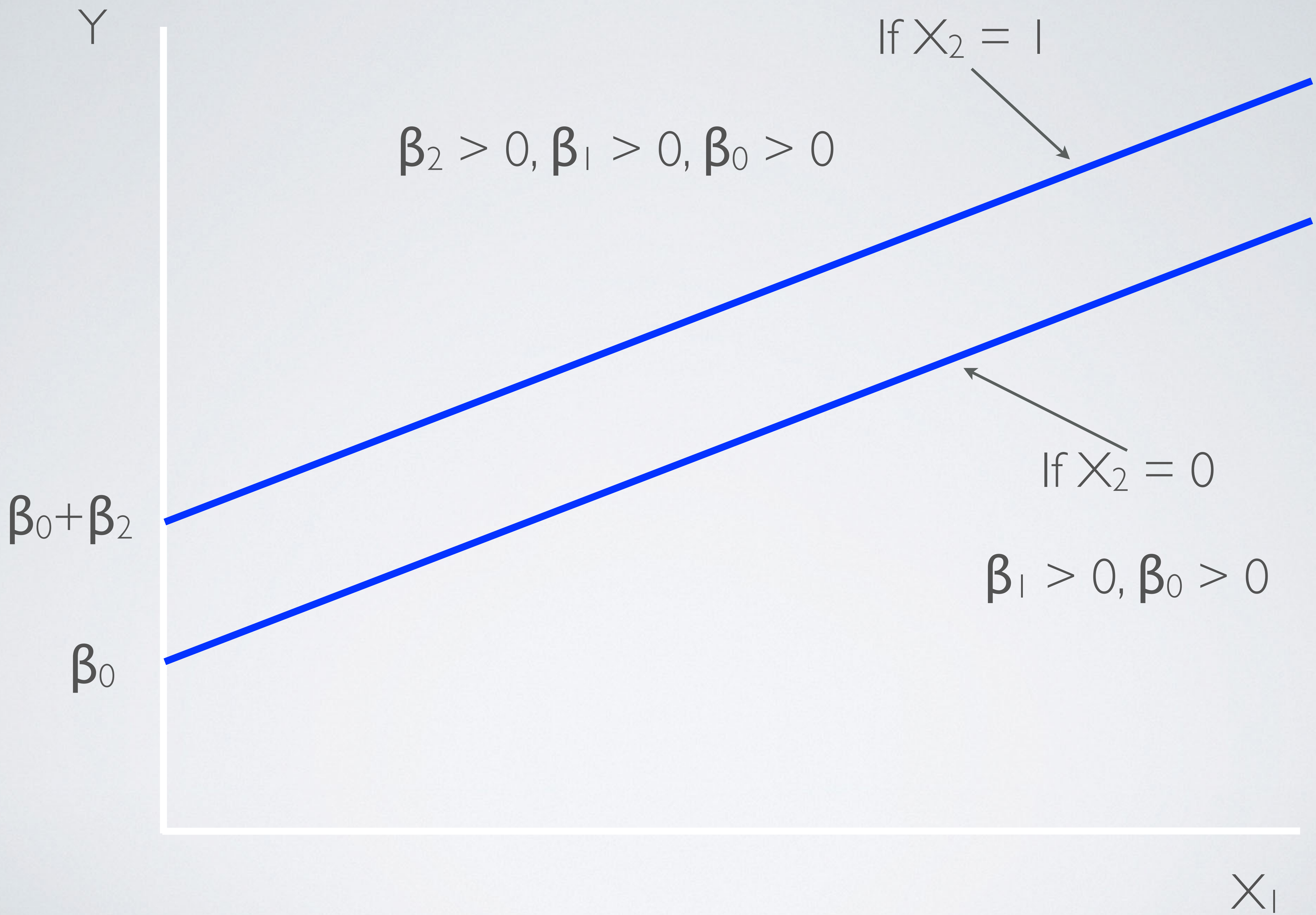
X_2: Dummy variable

$X_2 = 1$ if condition true

$X_2 = 0$ if condition false

Interpretation of β_2: The effect of condition true **RELATIVE to condition false/omitted category, holding X_1 constant**


Y
If $X_2 = 0$
$\beta_1 > 0, \beta_0 > 0$
If $X_2 = 1$
$\beta_2 < 0, \beta_1 > 0, \beta_0 > 0$
β_0
$\beta_0+\beta_2$
X_1



Y
If $X_2 = 1$
$\beta_2 > 0, \beta_1 > 0, \beta_0 > 0$
If $X_2 = 0$
$\beta_0+\beta_2$
$\beta_1 > 0, \beta_0 > 0$
β_0
X_1

EXAMPLE: INCOME BY SEX

Dummy moves intercept

$$\text{Income} = \beta_0 + \beta_1 \text{ female} + \beta_2 \text{ educ} + \beta_3 \text{ age} + \beta_4 \text{ hours} + \varepsilon$$

What is excluded category?

Men

WITHOUT DUMMIES

```
Call:
lm(formula = inc_1000 ~ educ + age + hrs, data = gss)

Residuals:
    Min      1Q  Median      3Q     Max 
-83.597 -19.122  -6.098   9.248 193.651 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -86.85764    7.51076 -11.564  < 2e-16 ***
educ          5.32744    0.39478  13.495  < 2e-16 ***
age           0.58032    0.09291   6.246 6.33e-10 ***
hrs           0.78317    0.07867   9.955  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.28 on 952 degrees of freedom
  (709 observations deleted due to missingness)
Multiple R-squared:  0.2536,	Adjusted R-squared:  0.2513 
F-statistic: 107.8 on 3 and 952 DF,  p-value: < 2.2e-16
```

DUMMIES - INTERCEPT

```
Call:
lm(formula = inc_1000 ~ female + educ + age + hrs, data = gss)

Residuals:
   Min     1Q Median     3Q    Max 
-89.56 -18.78  -5.61  10.21 184.07 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -75.24235    7.52224 -10.003  < 2e-16 ***
femaleTRUE  -15.60004    2.25500  -6.918 8.41e-12 ***
educ          5.47291    0.38598  14.179  < 2e-16 ***
age           0.57069    0.09071   6.291 4.80e-10 ***
hrs           0.64031    0.07953   8.051 2.44e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

DUMMIES - PREDICTIONS

For a male:		For a female:
0×-15.60	female	1×-15.60
$+\ 12 \times 5.47$	educ	$+\ 12 \times 5.47$
$+\ 45 \times 0.57$	age	$+\ 45 \times 0.57$
$+\ 40 \times 0.64$	hours	$+\ 40 \times 0.64$
$-\ 75.24$	constant	$-\ 75.24$
$= 41.65$		$= 26.05$

DUMMIES - INTERCEPT - AGAIN

More complicated case:

More than two categories

Dummy moves intercept

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 C_2 + \boldsymbol{\beta}_3 C_3 + \boldsymbol{\varepsilon}$$

C_2: Category 2

C_3: Category 3

$C_2 = 1$ if in category 2

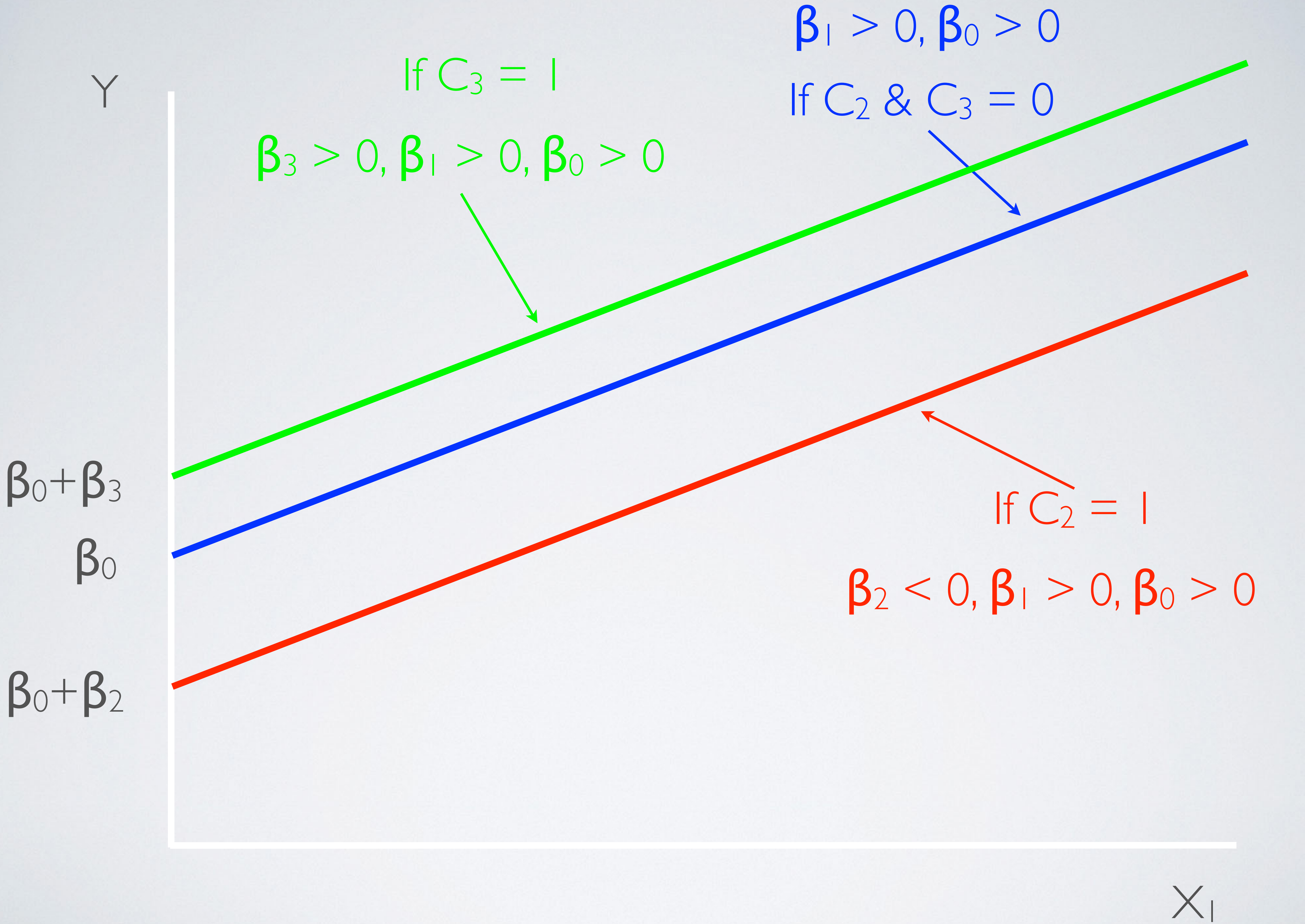
$C_2 = 0$ if not in category 2

$C_3 = 1$ if in category 3

$C_3 = 0$ if not in category 3

Interpretation of β_2: The effect of being in category 2 **RELATIVE to being in category 1, holding X_1 constant**

Interpretation of β_3: The effect of being in category 3 **RELATIVE to being in category 1, holding X_1 constant**



$\beta_1 > 0, \beta_0 > 0$
If $C_3 = 1$
If C_2 & $C_3 = 0$
Y
$\beta_3 > 0, \beta_1 > 0, \beta_0 > 0$
$\beta_0+\beta_3$
β_0
If $C_2 = 1$
$\beta_2 < 0, \beta_1 > 0, \beta_0 > 0$
$\beta_0+\beta_2$
X_1

DUMMIES - SLOPE

Dummy moves:

intercept

and

slope

BASIC SET-UP

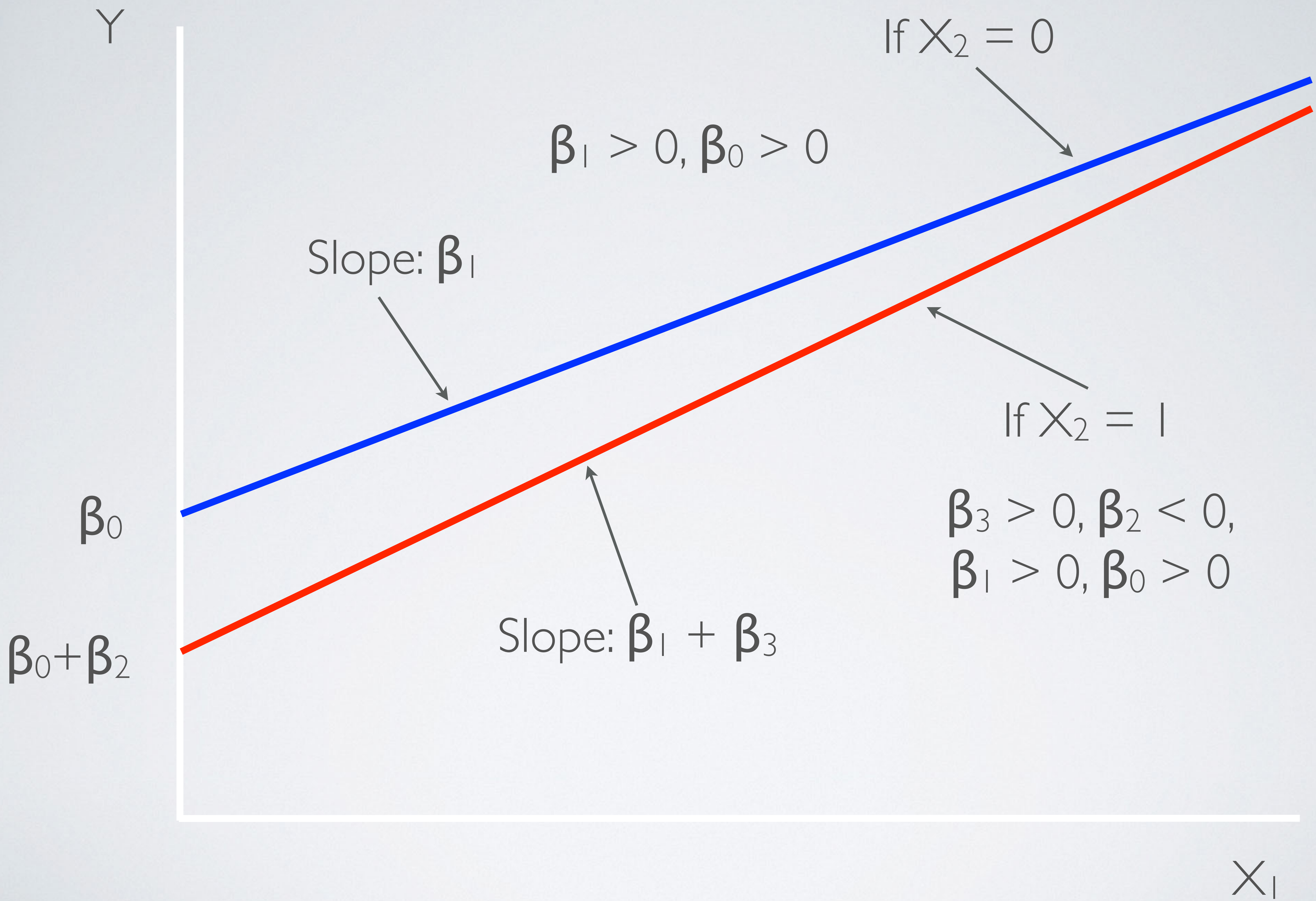
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2 + \varepsilon$$

X_1: Continuous variable

X_2: Dummy variable

$X_1 * X_2$: Interaction

Interpretation of β_3: The extra effect on slope
RELATIVE to condition false/omitted category


Y
If $X_2 = 0$
$\beta_1 > 0, \beta_0 > 0$
Slope: β_1
If $X_2 = 1$
β_0
$\beta_3 > 0, \beta_2 < 0,$
$\beta_1 > 0, \beta_0 > 0$
Slope: $\beta_1 + \beta_3$
$\beta_0+\beta_2$
X_1

EXAMPLE - SLOPE

Same example: Income of men and women

$$\text{Income} = \beta_0 + \beta_1 \text{ female} + \beta_2 \text{ educ} + \beta_3 \text{ female} \times \text{educ} + \beta_4 \text{ age} + \beta_5 \text{ hours} + \varepsilon$$

DUMMIES - SLOPE

```
Call:
lm(formula = inc_1000 ~ female * educ + age + hrs, data = gss)

Residuals:
    Min      1Q  Median      3Q     Max
-93.895 -18.859  -5.598  10.326 188.140

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -86.57879    8.76488  -9.878  < 2e-16 ***
femaleTRUE       11.46854   11.05527   1.037   0.2998
educ              6.30567    0.50897  12.389  < 2e-16 ***
age               0.56505    0.09049   6.244 6.42e-10 ***
hrs               0.64352    0.07932   8.113 1.52e-15 ***
femaleTRUE:educ  -1.94460    0.77760  -2.501   0.0126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

DUMMIES - PREDICTIONS

For a male:		For a female:
0×11.47	female	1×11.47
$+ 12 \times 6.31$	educ	$+ 12 \times 6.31$
$- 0 \times 1.94$	femaleXedu	-12×1.94
$+ 45 \times 0.57$	age	$+ 45 \times 0.57$
$+ 40 \times 0.64$	hours	$+ 40 \times 0.64$
$- 86.58$	constant	$- 86.58$
$= 40.51$		$= 28.7$

POLYNOMIALS

Simple non-linear modeling

PROBLEM

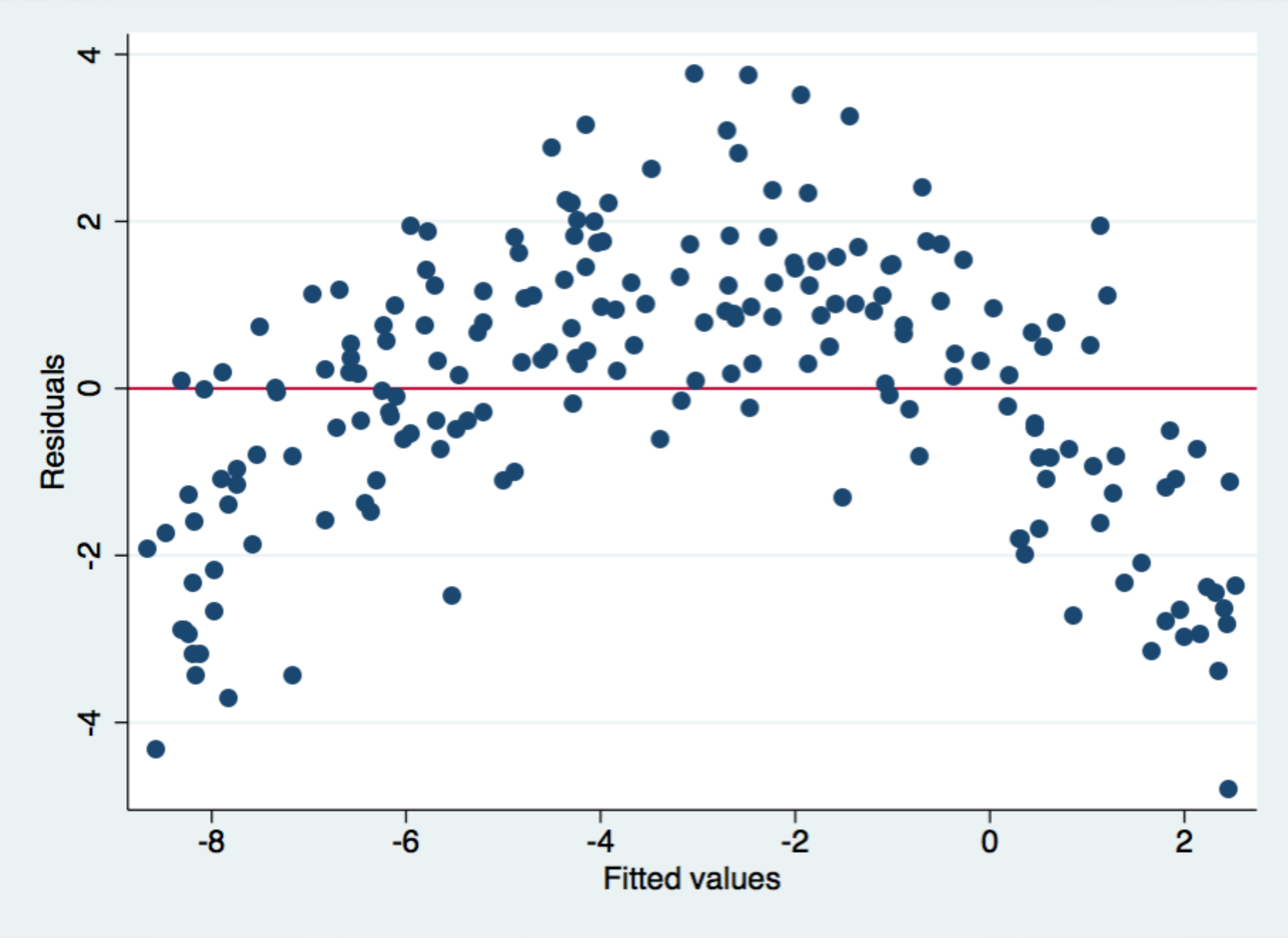
So far: Everything linear (in parameters)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Problem: a lot of effects are not linear

Which ones?
- Theory/common sense
- Model validation problems

MODEL VALIDATION



SOLUTIONS

Polynomial models

Double log and semi-log

POLYNOMIALS

Simplest model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

If effect of X_1 not linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \varepsilon$$

REAL MODELS HAVE CURVES

$\beta_1 > 0$, $\beta_2 < 0$

$\beta_1 < 0$, $\beta_2 > 0$

Remember: Holding X_2 constant

INCOME AND AGE

Question: how does age affect income?

Expected effect?

Data: General Social Survey

Sample: White women aged 20 to 60

SUMMARY STATISTICS

```
> summary(gss_women[c("age", "educ", "income", "hrs")])
      age             educ            income             hrs       
 Min.   :20.00   Min.   : 2.00   Min.   :   500   Min.   : 1.00  
 1st Qu.:32.00   1st Qu.:12.00   1st Qu.: 13750   1st Qu.:35.00  
 Median :42.00   Median :14.00   Median : 27500   Median :40.00  
 Mean   :41.14   Mean   :13.95   Mean   : 34623   Mean   :38.76  
 3rd Qu.:51.00   3rd Qu.:16.00   3rd Qu.: 45000   3rd Qu.:44.75  
 Max.   :60.00   Max.   :20.00   Max.   :200000   Max.   :80.00  
                                 NA's   :194      NA's   :207    
```

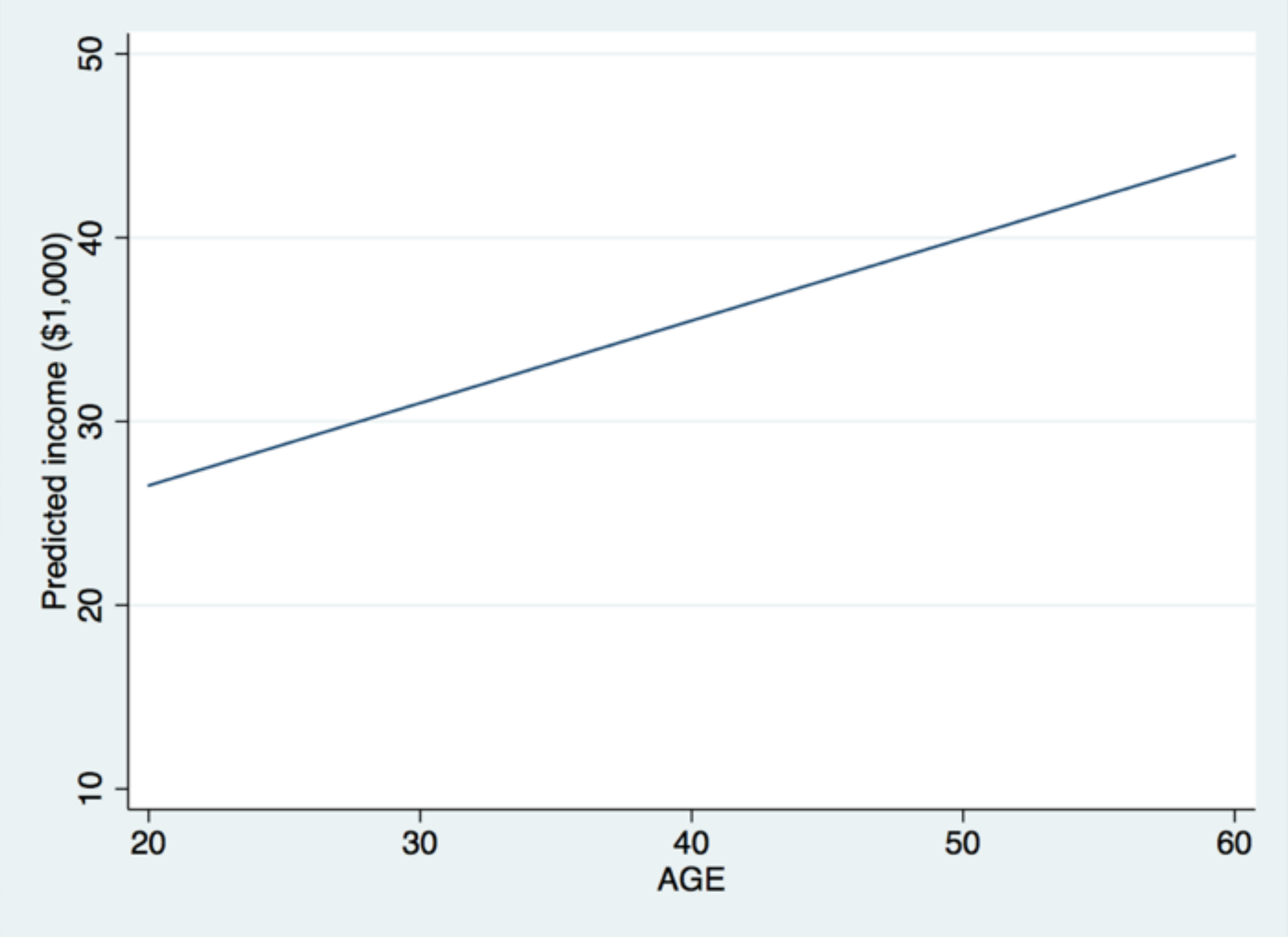
LINEAR MODEL

```
Call:
lm(formula = inc_1000 ~ educ + age + hrs, data = gss_women)

Residuals:
    Min      1Q  Median      3Q     Max
-53.718 -14.491  -3.522   8.534 174.162

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -65.3144    10.5630  -6.183 2.04e-09 ***
educ          4.1344     0.5271   7.843 7.70e-14 ***
age           0.4485     0.1274   3.521 0.000497 ***
hrs           0.6008     0.1237   4.859 1.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.54 on 301 degrees of freedom
  (272 observations deleted due to missingness)
Multiple R-squared:  0.2411,	Adjusted R-squared:  0.2336
F-statistic: 31.88 on 3 and 301 DF,  p-value: < 2.2e-16
```

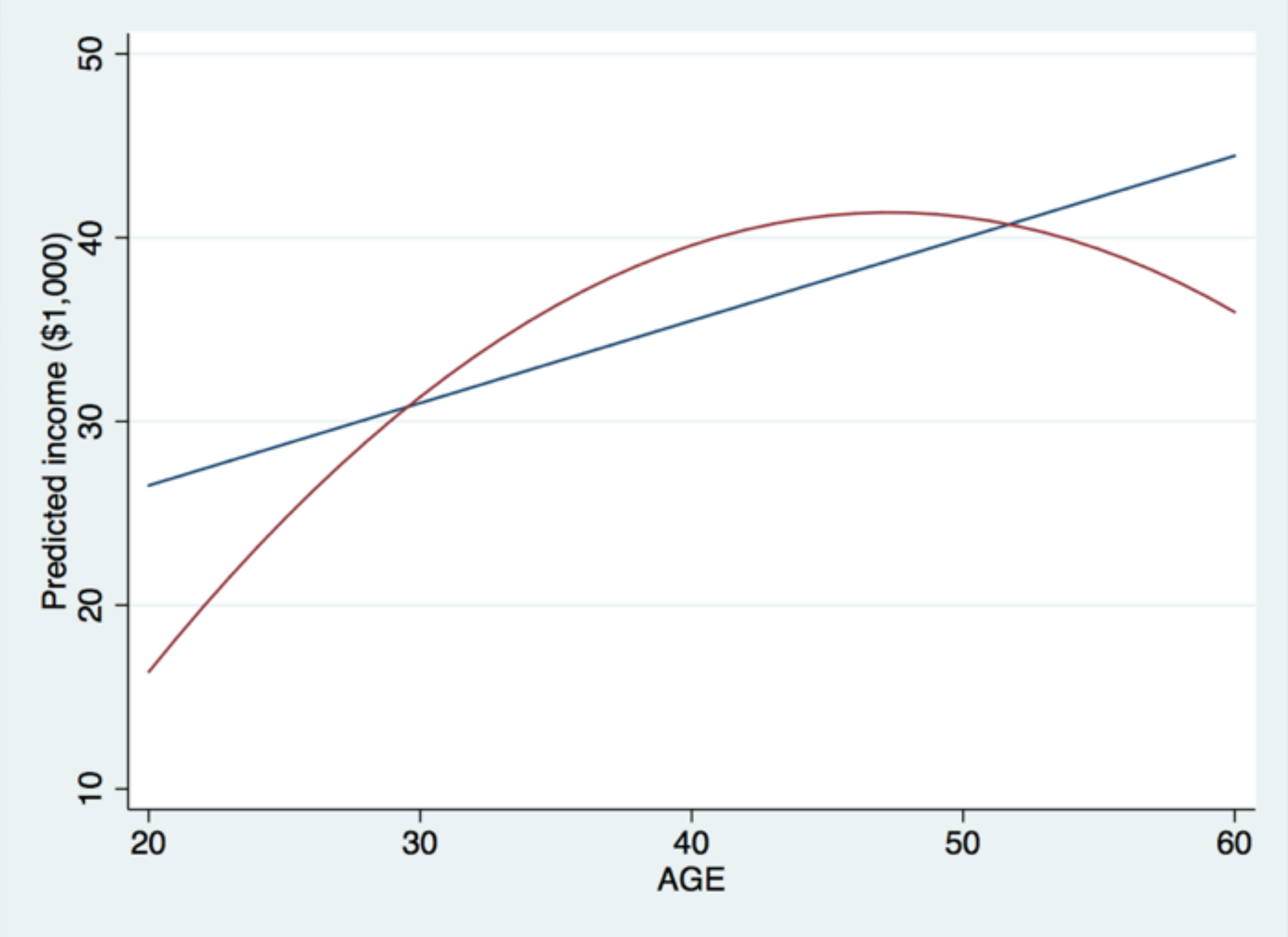
POLYNOMIAL MODEL

```
Call:
lm(formula = inc_1000 ~ educ + age + I(age^2) + hrs, data = gss_women)

Residuals:
    Min      1Q  Median      3Q     Max 
-57.081 -13.581  -3.895   8.195 171.262 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -113.67791   20.70894  -5.489 8.60e-08 ***
educ           3.99177    0.52434   7.613 3.51e-13 ***
age            3.17586    1.01598   3.126  0.00195 ** 
I(age^2)      -0.03358    0.01241  -2.705  0.00721 ** 
hrs            0.57890    0.12265   4.720 3.62e-06 ***
```

PREDICTED INCOME



PREDICTED INCOME



WHICH IS BETTER?

Polynomial more theoretically appealing

Adjusted R^2 higher for polynomial (24.93 vs 23.36)

Coefficient for Age^2 statistically significant

Linear model is easier to interpret

EDUCATION AND FERTILITY

Question: What is the effect of education on fertility?

Data: General Social Survey

Sample: women aged 46 to 55
(no more births)

DESCRIPTIVE STATISTICS

```
> summary(gss_kids[c("childs", "educ")])
     childs          educ      
 Min.   :0.000   Min.   : 2.00  
 1st Qu.:1.000   1st Qu.:12.00  
 Median :2.000   Median :14.00  
 Mean   :2.157   Mean   :13.78  
 3rd Qu.:3.000   3rd Qu.:16.00  
 Max.   :8.000   Max.   :20.00  
 NA's   :1                      
```

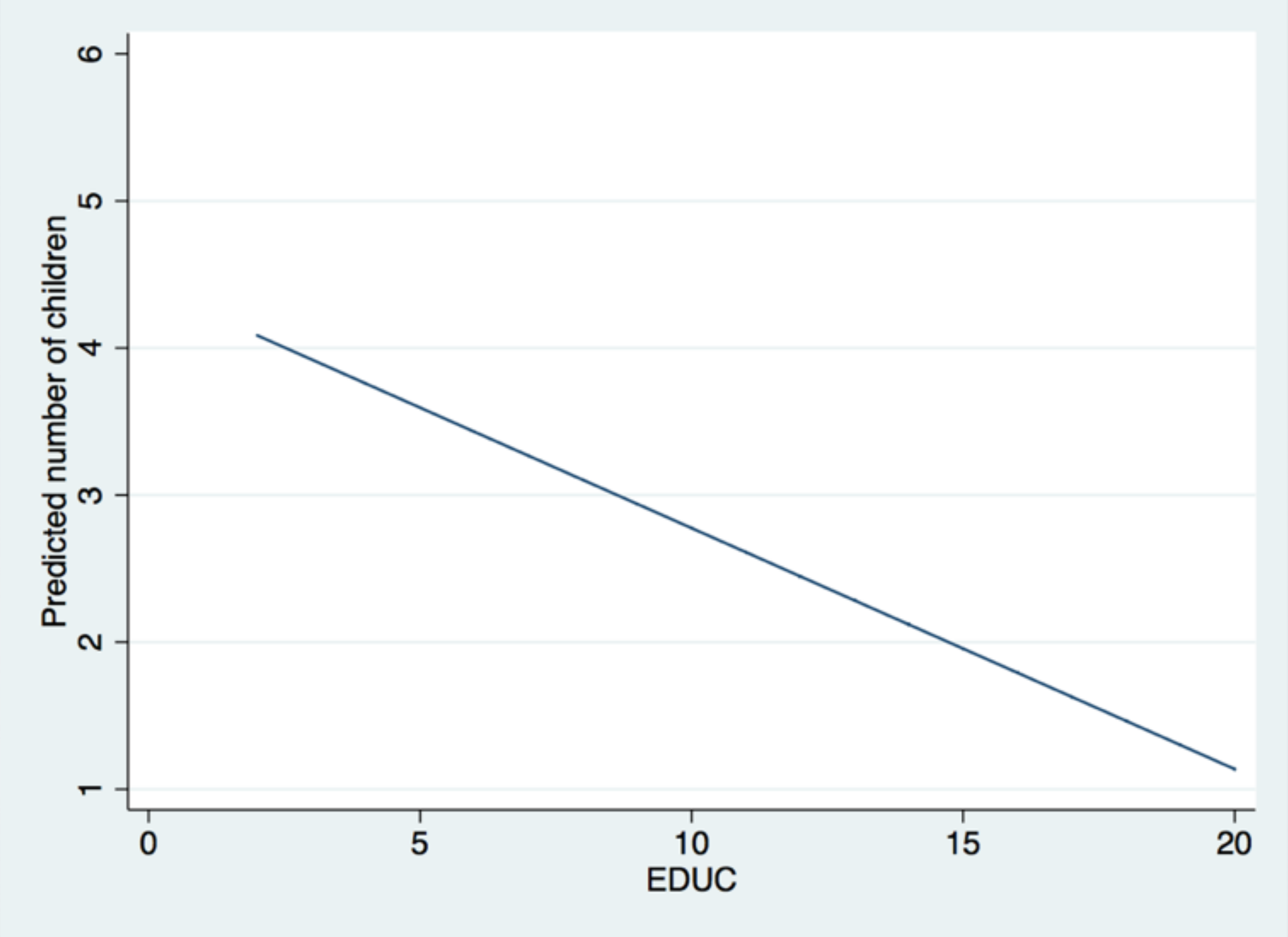
LINEAR MODEL

```
Call:
lm(formula = childs ~ educ, data = gss_kids)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1026 -1.1195 -0.1195  0.7167  5.5528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.41344    0.49364   8.941  < 2e-16 ***
educ        -0.16385    0.03504  -4.677 5.07e-06 ***
```

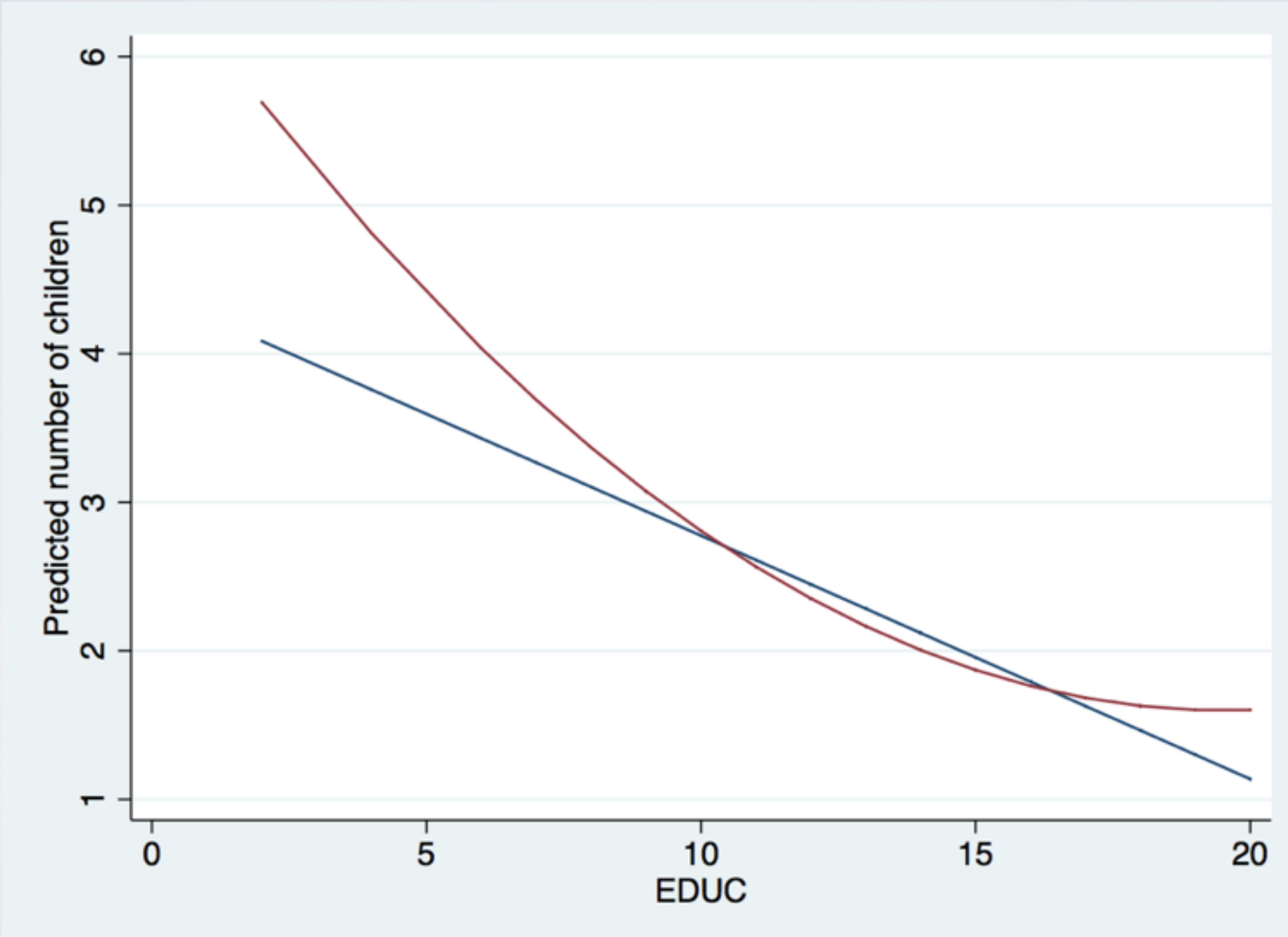
POLYNOMIAL MODEL

```
Call:
lm(formula = childs ~ educ + I(educ^2), data = gss_kids)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3680 -1.0202 -0.0740  0.7415  5.6479

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.680952   1.289424   5.181 4.97e-07 ***
educ        -0.520875   0.190945  -2.728  0.00689 **
I(educ^2)    0.013345   0.007017   1.902  0.05852 .
```

PREDICTED NUMBER OF KIDS



PREDICTED NUMBER OF KIDS



WHICH IS BETTER?

Polynomial more theoretically appealing

Adjusted R^2 higher for polynomial (9.66 vs 8.59)

Coefficient for $Educ^2$ statistically significant (at 10%)

Linear model is easier to interpret

If remove 5 lowest educated women:
Effect of $Educ^2$ very small

DROP 5 LEAST EDUCATED

```
Call:
lm(formula = childs ~ educ + I(educ^2), data = gss_kids_small)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9089 -1.0790 -0.0790  0.6597  5.6597

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.230589   2.501287   1.691   0.0922 .
educ        -0.180591   0.353676  -0.511   0.6101
I(educ^2)    0.001922   0.012215   0.157   0.8751
```

CONSIDERATIONS

Interpretation: Both X and X^2 change at the same time!

Transformation of X or X^2 to ease readability
(divide by, say, 100)

LOGARITHMS

And the economists who love them

SEMI-LOG - WHEN?

Ln of Y: Y changes in percent for a change in X

Wages and education

Ln of X: Y changes less and less for a change in X

Demand

Diminishing marginal returns

INTERPRETATION OF SEMI-LOGS

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

β_1*100 : The percent change in Y for a unit change in X_1 (approximately)

β_2*100 : The percent change in Y for a unit change in X_2 (approximately)

SEMI-LOG - RETURNS TO EDUCATION

```
Call:
lm(formula = log(income) ~ female + educ + age + hrs, data = gss)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3863 -0.3269  0.1118  0.4981  2.7369 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.957859   0.195073  35.668  < 2e-16 ***
femaleTRUE  -0.259098   0.058479  -4.431 1.05e-05 ***
educ         0.126939   0.010010  12.682  < 2e-16 ***
age          0.018117   0.002352   7.701 3.38e-14 ***
hrs          0.022245   0.002062  10.786  < 2e-16 ***
```

SEMI-LOG - INTERPRETATION

Women earn 25.9% less than a similar male,
holding everything else constant

A year of education increases your income by 12.7%,
holding everything else constant

SEMI-LOG - PREDICTIONS

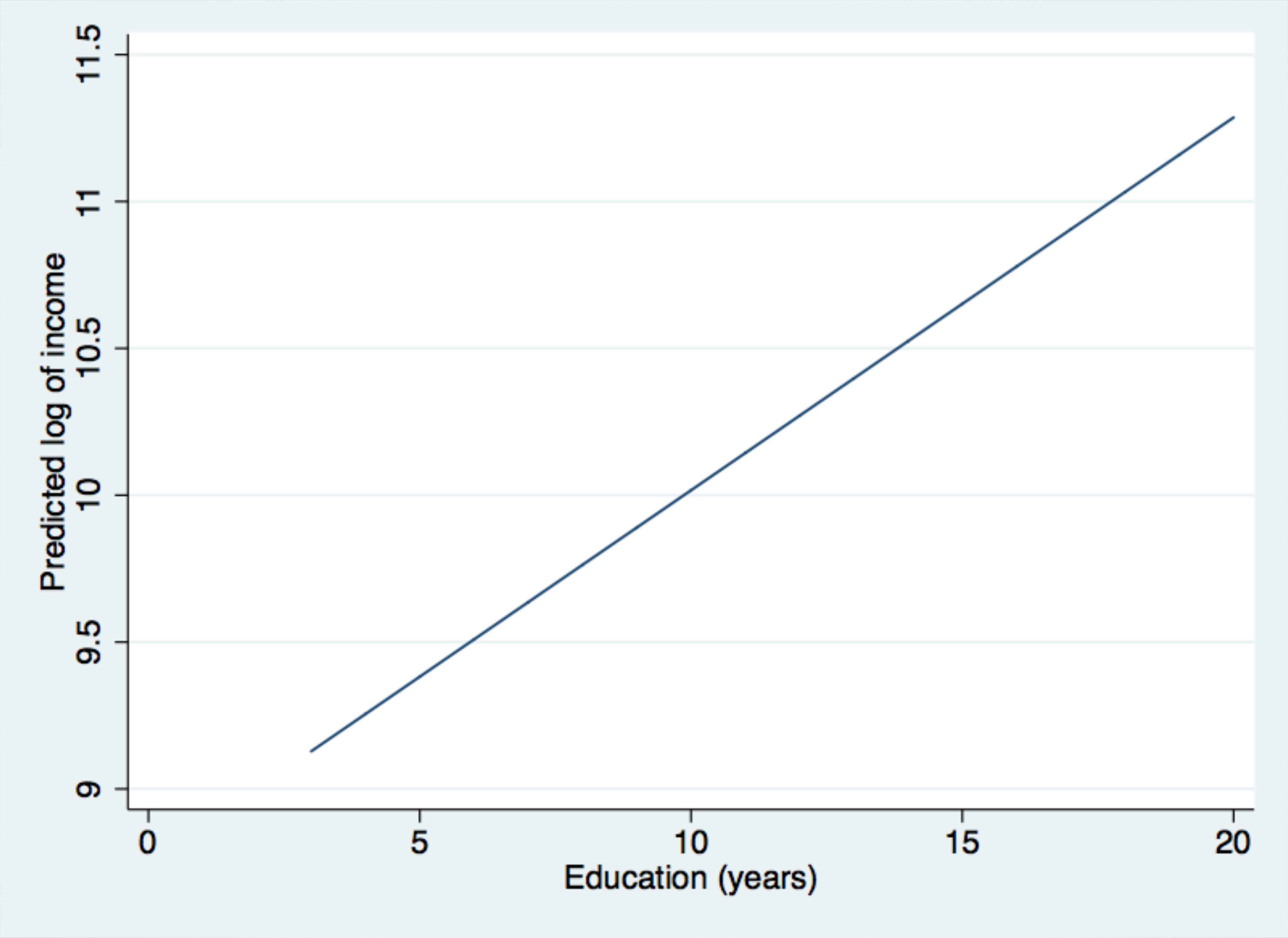
Same male as above

$$
\begin{aligned}
& 0 \times -0.259 \\
& + 12 \times 0.127 \\
& + 45 \times 0.018 \\
& + 40 \times 0.022 \\
& + 6.96
\end{aligned}
$$

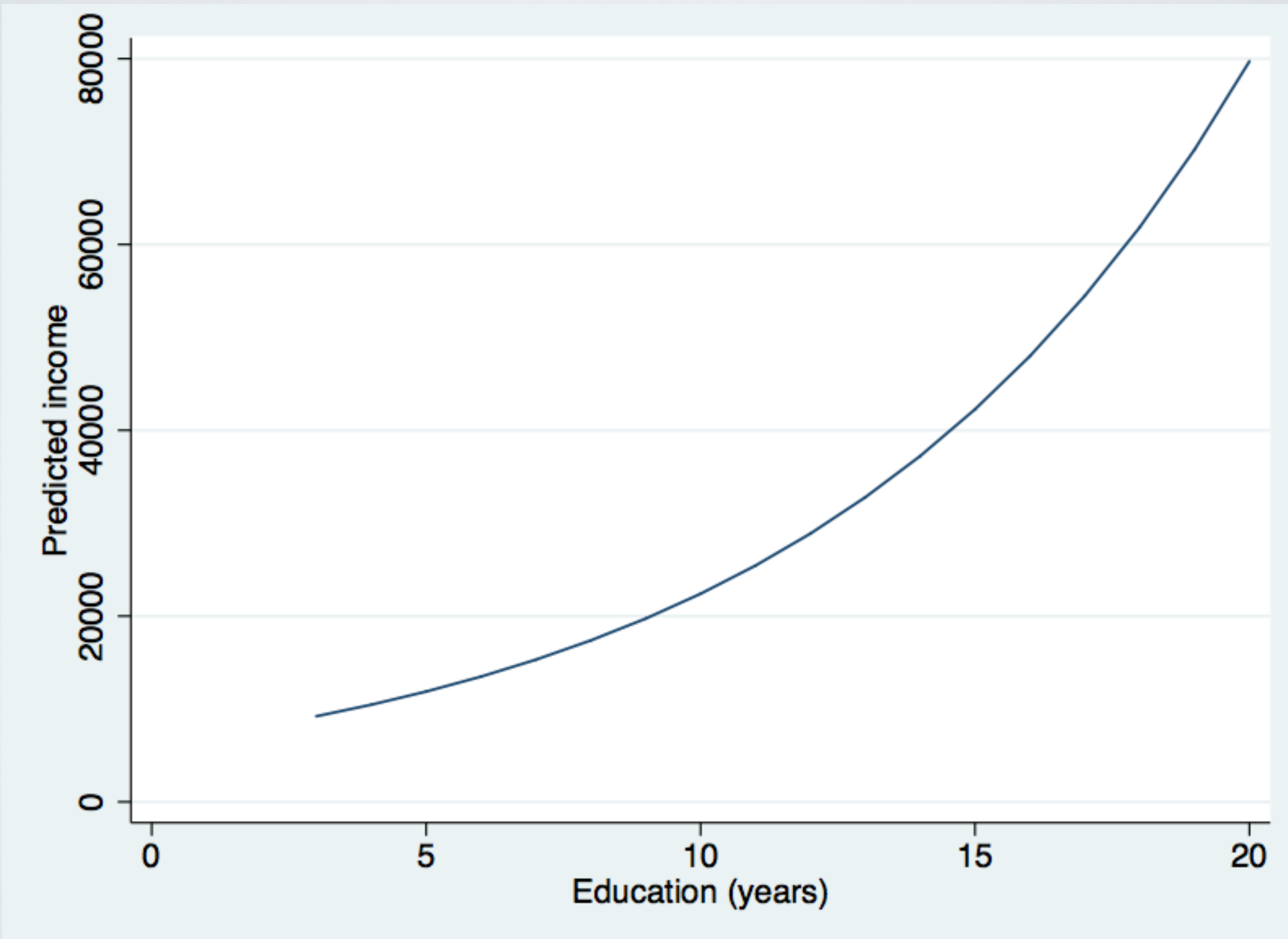
$= 10.174$

Income ($): e^10.174 = 26,212

GRAPHS OF PREDICTIONS



GRAPHS OF PREDICTIONS



INTERPRETATION OF SEMI-LOGS

$$Y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 + \varepsilon$$

β_1 : An increase of 1 in the log of X_1 increases Y by β_1 (not very interesting)

Better: A 1% increase in X_1 increases Y by $\beta_1/100$

ELASTICITIES - DOUBLE-LOG

Elasticity: Percent change in Y
for a 1 percent change in X

Elasticity constant for all X

Production functions

DOUBLE-LOG - EXAMPLE

Relation between income and education

$$\ln \text{income} = \beta_0 + \beta_1 \ln \text{educ} + \ldots + \varepsilon$$

LINEAR VERSION

```
Call:
lm(formula = log(income) ~ female + log(educ) + age + hrs, data = gss)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3774 -0.3397  0.1085  0.5070  2.7230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.842802   0.355121  13.637  < 2e-16 ***
femaleTRUE  -0.260374   0.059111  -4.405 1.18e-05 ***
log(educ)    1.485014   0.126480  11.741  < 2e-16 ***
age          0.018306   0.002377   7.702 3.36e-14 ***
hrs          0.022187   0.002084  10.646  < 2e-16 ***
```

DOUBLE-LOG - INTERPRETATION

A one percent increase in education is associated with a 1.5 percent increase in income

COMBINING EXAMPLES

Does the return to education vary by sex?

How would you estimate this?

COMBINING EXAMPLES

```
Call:
lm(formula = log(income) ~ female * educ + age + hrs, data = gss)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3992 -0.3410  0.1211  0.5034  2.7655

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.878030   0.227990  30.168  < 2e-16 ***
femaleTRUE      -0.068487   0.287568  -0.238    0.812
educ             0.132803   0.013239  10.031  < 2e-16 ***
age              0.018077   0.002354   7.680 3.96e-14 ***
hrs              0.022268   0.002063  10.792  < 2e-16 ***
femaleTRUE:educ -0.013693   0.020227  -0.677    0.499
```

INTERPRETATION - RETURN BY SEX

Men: Extra year of education ~ 13.3% increase in income

Women: Extra year of education ~ 11.9% increase in income

Note: not statistically significant different

Q&A ON CASE STUDY