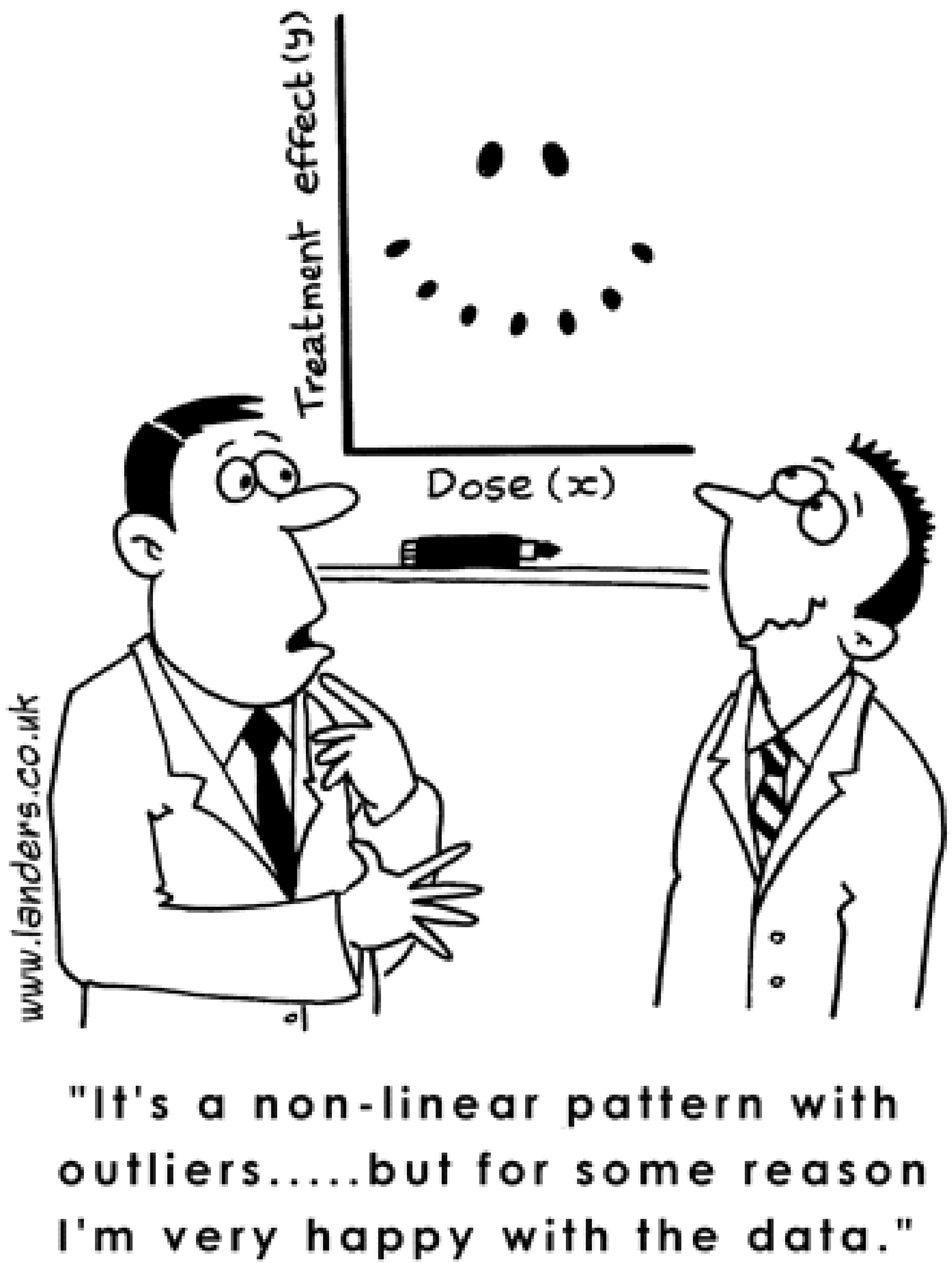


Imperial means
Intelligent Business

Econometrics for dummies

About qualitative and nonlinear relationships

by Ralf Martin (r.martin@imperial.ac.uk)

Objectives for this lecture

- Learn how to deal with modelling qualitative aspects of reality

- We can code those with dummy (binary) variables
  - As explanatory variables
  - As dependent variables

- Appreciate that dummy variables an important building block for constructing more sophisticated models
  - Allowing for non-linear relationships
  - Controlling for many potential confounding factors

- There are some other easy example of nonlinear models. Let’s look at those.

Looking at wage regressions as example again:

```
wage1 <-  read.csv("https://www.dropbox.com/s/9agc2vmamfztlel/WAGE1.csv?dl=1")


r1 <- lm(wage ~ female, wage1)
r1 %>% summary()
```

```
##
## Call:
## lm(formula = wage ~ female, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.114
## F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```

Change the x variable by 1 unit. What's the effect on the Y variable?

i.e. go from male to female, what's the effect on wages?

How can we interpret the coefficients?

$$Wage = \beta_0 + \beta_1 \times FEMALE + \epsilon$$

A closer look at those regression coefficients

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
```

```
wage1 %>% group_by(female) %>% summarize(mean(wage))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   female `mean(wage)`
##    <int>        <dbl>
## 1      0         7.10
## 2      1         4.59
```

Average wage

Notice, that the intercept (β_0) is the average wage for men. To get the average wage for women we need to add the coefficient:

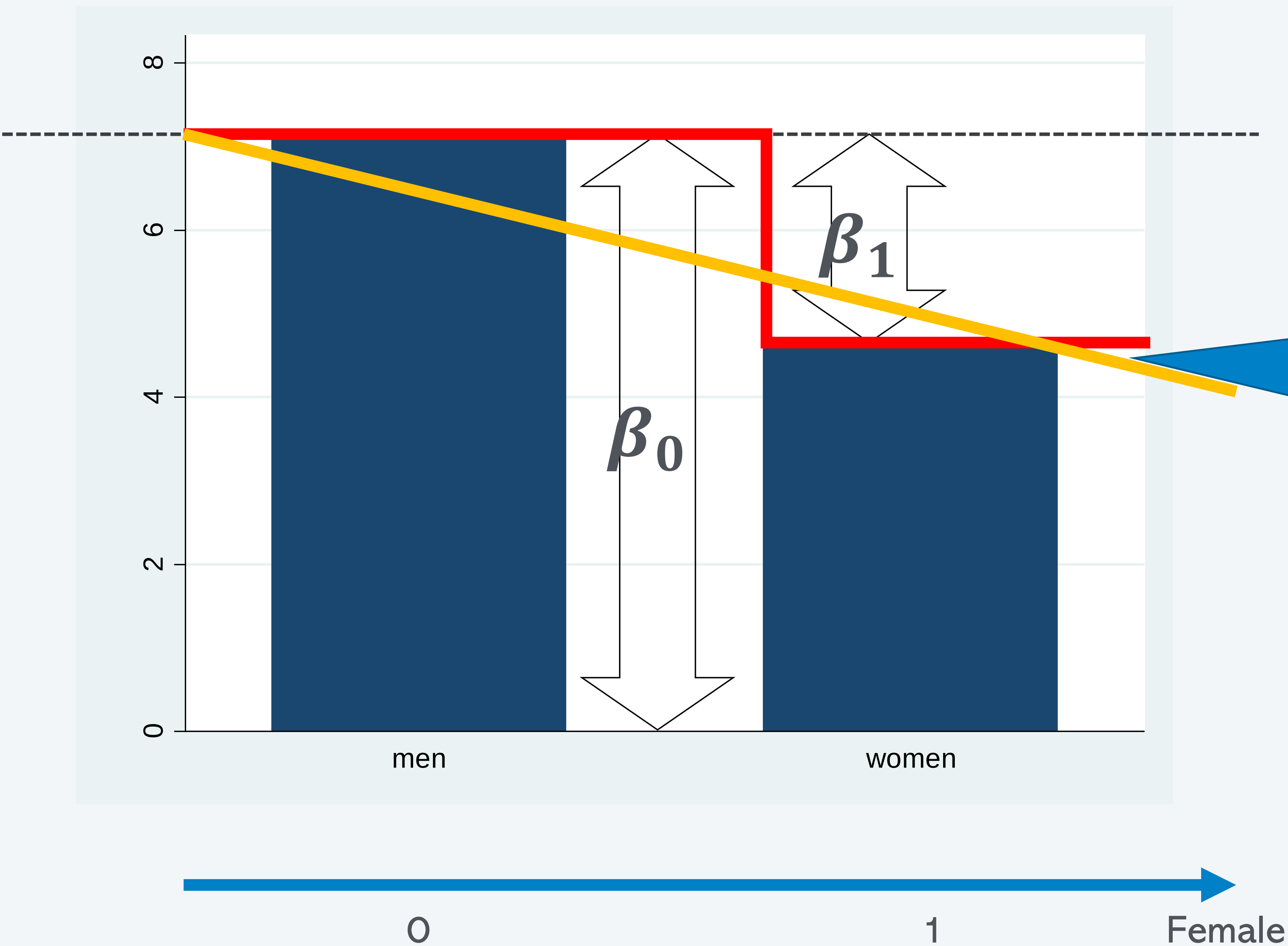
$$E\{wage|Women\} = \beta_0 + \beta_1 = 4.587...$$

where $E\{wage|Women\}$ is a mathematical way of saying the average (or expected) wage for women.

Conditional expectation

Dummies as bars

The underlying model: $Y = \beta_0 + \beta_1 FEMALE + \epsilon$



Instead of a linear relationship between the x and y variable we we get a nonlinear one

What about men?

```
> summary(lm(wage ~ male +female, wage1))

Call:
lm(formula = wage ~ male + female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5995 -1.8495 -0.9877  1.4260 17.8805 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.5877     0.2190  20.950  < 2e-16 ***
male          2.5118     0.3034   8.279 1.04e-15 ***
female            NA         NA      NA       NA    
```

- R dropped "female" because of multi-collinearity
- **We cannot see from our data what happens when female changes while male is kept constant.**
- Interpretation of coefficients: $E\{Y|Women\} = \beta_{Const}$

  $E\{Y|Men\} = \beta_{Cons}+\beta_{male}$

  i.e. $\beta_{male} = E\{Y|Men\} - E\{Y|Women\}$

2 cases in the data. We only need 2 parameters to represent those

Another option

```
> summary(lm(wage ~ 0+male +female, wage1))

Call:
lm(formula = wage ~ 0 + male + female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5995 -1.8495 -0.9877  1.4260 17.8805 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
male      7.099      0.210   33.81   <2e-16 ***
female    4.588      0.219   20.95   <2e-16 ***
---
```

- Instead of dropping "female" we can drop "male"
- However, we can also drop the constant as in the regression above
- Consequently $E\{Y|Women\} = \beta_{female}$
  $E\{Y|Men\} = \beta_{male}$

Main take-away

- Various ways to represent the same thing/model that men and women have a different average wage by including combinations of dummy variables from the following
  - "constant" : always equal to 1
  - "male" : equal to 1 for men
  - "female" : equal to 1 for women

- Which dummies we include exactly will affect the interpretation of the coefficients ($\beta' s$)

- If we include "constant" and "male" ("female") then "female" ("male") becomes the **reference category**

- The mean of the reference category is represented by the constant coefficient

Sets of dummies

- Dummies "male" and "female" classify the sample exhaustively
- We often have classifications with more than two categories
- e.g. rather than having the education in years we might have only a categorical variable capturing 3 levels of education:

```
wage1["educats"] = 0
wage1$educats[wage1$educ==12] = 1
wage1$educats[wage1$educ>12]  = 2


wage1 %>% group_by(educats) %>% summarize(n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   educats `n()`
##     <dbl> <int>
## 1       0   116
## 2       1   198
## 3       2   212
```

Your turn: How would you conduct regression analysis of the relationship between wage and education if all you had was the educats variable?

(a) e.g. lm(wage~educats)
(b) something else?

Regression on categorical variable

```
> summary(lm(wage ~ educats, wage1))

Call:
lm(formula = wage ~ educats, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2933 -2.2542 -0.9292  1.2301 17.6867 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.8751     0.2765  14.016   <2e-16 ***
educats       1.7091     0.1961   8.717   <2e-16 ***
```

Can you interpret this coefficient?

Creating sets of dummies from categories approach 1

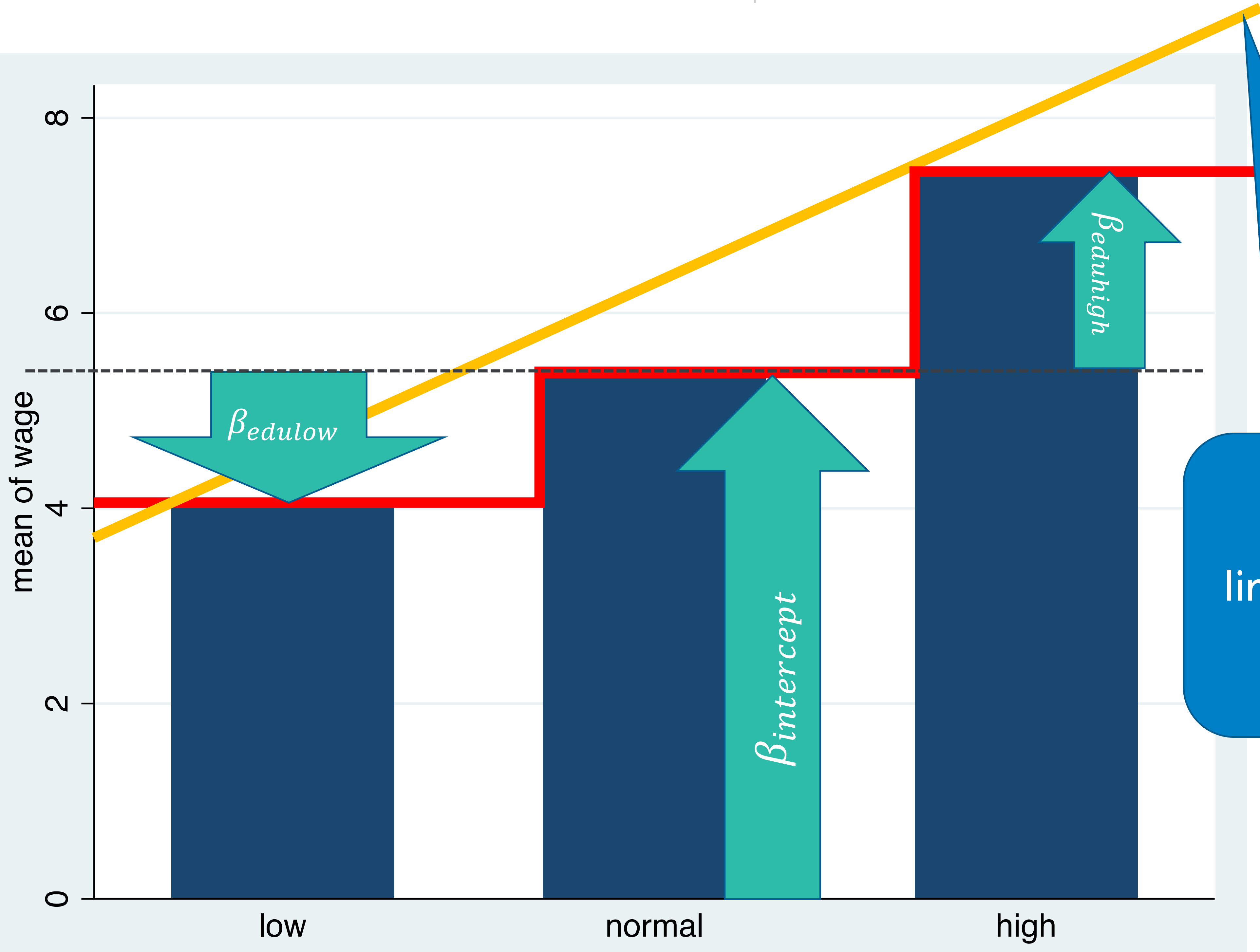
```
wage1=wage1 %>% mutate(edu_low=educats==0,
                       edu_normal=educats==1,
                       edu_high=educats==2)

reg=lm(wage~edu_low+edu_normal+edu_high,wage1)

reg %>% summary()
```

```
##
## Call:
## lm(formula = wage ~ edu_low + edu_normal + edu_high, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.393 -2.119 -1.033  1.245 17.587
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.3926     0.2372  31.165  < 2e-16 ***
## edu_lowTRUE     -3.3359     0.3989  -8.363 5.56e-16 ***
## edu_normalTRUE  -2.0213     0.3413  -5.922 5.78e-09 ***
## edu_highTRUE         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

- We cannot regress dummies for all categories and a constant (dummy variable trap)
- R makes sure we don't fall in the trap and drops one of the dummies (thank you R…that was a close call)

Interpretation

```
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.3714     0.2454  21.884  < 2e-16 ***
## edu_lowTRUE       -1.3146     0.4038  -3.255  0.00121 **
## edu_highTRUE       2.0213     0.3413   5.922 5.78e-09 ***
## edu_normalTRUE         NA         NA      NA       NA
```



Non-linear step function….however linear relationship looks like a fairly good approximation

Testing the validity of a linear model

Linear: A change of 1 in x variable always has the same effect on the y variable

This means in current context

Edu low to mid = Edu mid to high

How to say that in terms of model parameters?

$$Wage = \beta_0 + \beta_{edulow}\text{edu\_low} + \beta_{eduhigh}\text{edu\_high} + \epsilon$$

Test $\beta_{edulow} = -\beta_{eduhigh}$

Testing the validity of the linear model

```
reg=lm(wage ~ edu_low+edu_high, wage1)

linearHypothesis(reg,c("edu_lowTRUE=-edu_highTRUE"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## edu_lowTRUE  + edu_highTRUE = 0
##
## Model 1: restricted model
## Model 2: wage ~ edu_low + edu_high
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    524 6253.5
## 2    523 6238.6  1    14.888 1.2481 0.2644
```

What do you conclude?

High p-value, so we cannot reject. Hence, it would be valid to use the linear model here

Creating sets of dummies from categories approach 2

- Instead of creating a separate dummy variable for every category, we can tell R that we are dealing with a categorical/factor variable
- Good if we have a large number of categories

```
wage1 =wage1 %>% mutate(educatsf=factor(educats,label=c("low","normal","high")))
```

R creates dummy variables automatically

```
> summary(lm(wage ~ educatsf, wage1))

Call:
lm(formula = wage ~ educatsf, data = wage1)

Residuals:
   Min     1Q Median     3Q    Max 
-5.393 -2.119 -1.033  1.245 17.587 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.0567     0.3207  12.651  < 2e-16 ***
educatsfnormal   1.3146     0.4038   3.255  0.00121 ** 
educatsfhigh     3.3359     0.3989   8.363 5.56e-16 ***
---
```

```
> summary(lm(wage ~ 0+educatsf, wage1))

Call:
lm(formula = wage ~ 0 + educatsf, data = wage1)

Residuals:
   Min     1Q Median     3Q    Max 
-5.393 -2.119 -1.033  1.245 17.587 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
educatsflow      4.0567     0.3207   12.65   <2e-16 ***
educatsfnormal   5.3714     0.2454   21.88   <2e-16 ***
educatsfhigh     7.3926     0.2372   31.16   <2e-16 ***
---
```

Several Dummy Sets

```
> wage1["gender"]<-factor(wage1$female, label=c("male","female"))
> summary(lm(wage ~ educatsf+gender, wage1))

Call:
lm(formula = wage ~ educatsf + gender, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4467 -2.0765 -0.4759  0.9779 16.6133 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.1207     0.3280  15.611  < 2e-16 ***
educatsfnormal   1.6052     0.3817   4.206 3.06e-05 ***
educatsfhigh     3.2460     0.3755   8.644  < 2e-16 ***
genderfemale    -2.3735     0.2867  -8.278 1.06e-15 ***
---
```

- Implied model: $Wage = \beta_C$
  $$+\beta_{normal} normal + \beta_{high} high$$
  $$+\beta_{female} female + \epsilon$$
- Reference category for education: Low
- Reference category for gender: Male

Interpretation

$$Wage = \underline{\beta_C} + \beta_{normal} normal + \beta_{high} high$$
$$+ \beta_{female} female + \epsilon$$

$\beta_C = E\{WAGE|Educ\ low, Male\}$

Interpretation

$$Wage = \beta_C + \beta_{normal} normal + \beta_{high} high + \beta_{female} female + \epsilon$$

$$E\{WAGE|Educ\ Low, Female\} = \beta_C + \beta_{female}$$

$$E\{WAGE|Educ\ Normal, Female\} = \beta_C + \beta_{female} + \beta_{normal}$$

Dummies as dependent variables

- So far we discussed dummies as explanatory variables
- However, we might also have dummies as dependent variables
- E.g Bertrand Mullainathan Data: ../data/bm.dta
- We regress $CALL = \beta_0 + \beta_1 BLACK + \epsilon$

Share of non Black people that received a call back

- Hence, following the discussion in this lecture:
- $\beta_0 = E\{CALL|Non\ Black\} = \frac{\sum_{i \in NonBlack} CALL_i}{n} = P\{Call|NonBlack\}$

- $\beta_1 = E\{CALL|Black\} - E\{CALL|Non\ Black\}$
  $= P\{Call|Black\} - P\{Call|NonBlack\}$

- $\hat{\beta}_0$: share of non Black receiving call back
- $\hat{\beta}_1$: share of black receiving call - of share of non Black receiving call

- i.e. there is a natural interpretation of coefficients when regressing dummies on dummies

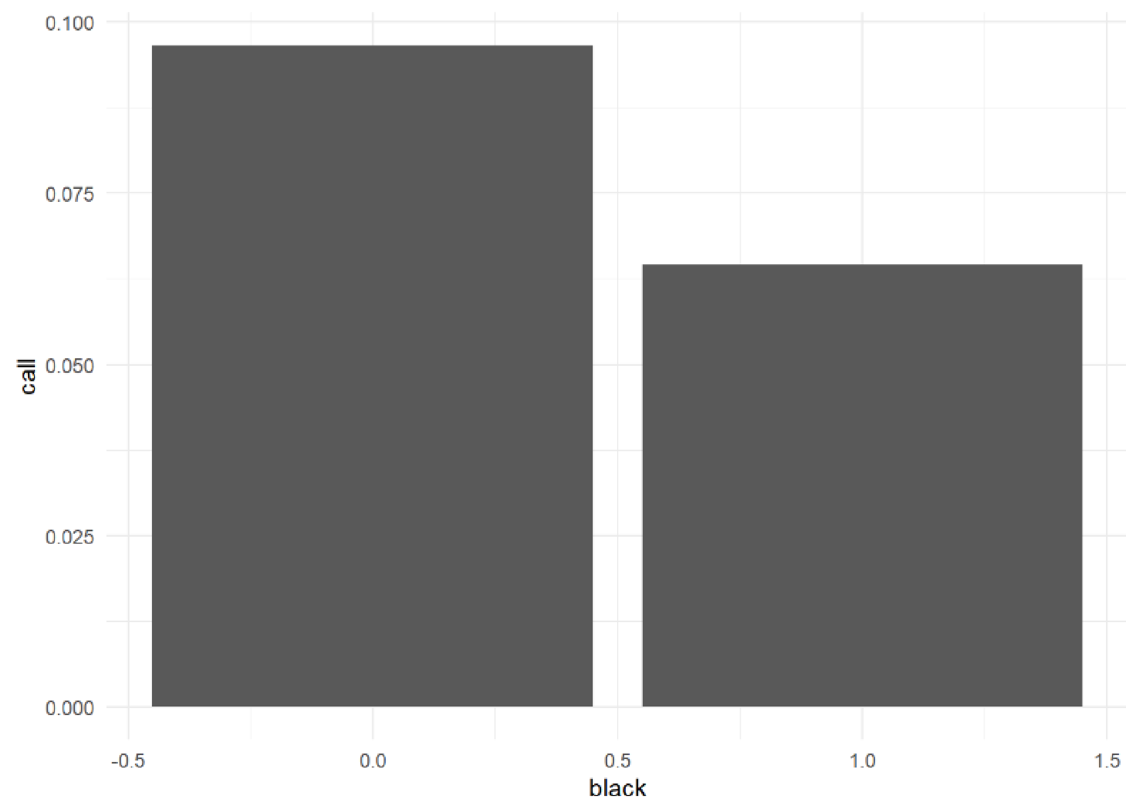
- Things are a bit less clear when regressing dummies on – say – a linear term

Dummies as dependent variables - In action

```r
summary(lm(call~black,bm))
```

```
##
## Call:
## lm(formula = call ~ black, data = bm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09651 -0.09651 -0.06448 -0.06448  0.93552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.096509   0.005505  17.532  < 2e-16 ***
## black       -0.032033   0.007785  -4.115 3.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 ' *' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.271      868 degrees of freedom
## Multiple R-squared:  0.003466,       ted R-squared:  0.003261
## F-statistic: 16.93 on 1 and 4       value: 3.941e-05
```

```r
library(ggplot2)
agg=bm %>% group_by(black) %>% summarise(call=mean(call))

ggplot(agg,aes(x=black,y=call))+geom_bar(stat="identity")+theme_m
inimal()
```



CVs with "black" sounding names have a 3.2% lower chance of receiving a call back

Dummies as dependent variables - linear model case

$$Call = \beta_1 + \beta_2 Experience + \epsilon$$

```
(linear0 <- lm(call ~ yearsexp         , data = bm)) %>% summary()
```

```
##
## Call:
## lm(formula = call ~ yearsexp, data = bm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20030 -0.08101 -0.07439 -0.06776  0.94218
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0545046  0.0071949   7.575 4.26e-14 ***
## yearsexp    0.0033136  0.0007716   4.295 1.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual      dard error: 0.2716 on 4868 degrees of freedom
## Multipl          d:  0.003774,   Adjusted R-squared:  0.00357
## F-stati             1 and 4868 DF,  p-value: 1.784e-05
```

Probability of call back increases by 0.3 percentage points with every additional year of experience

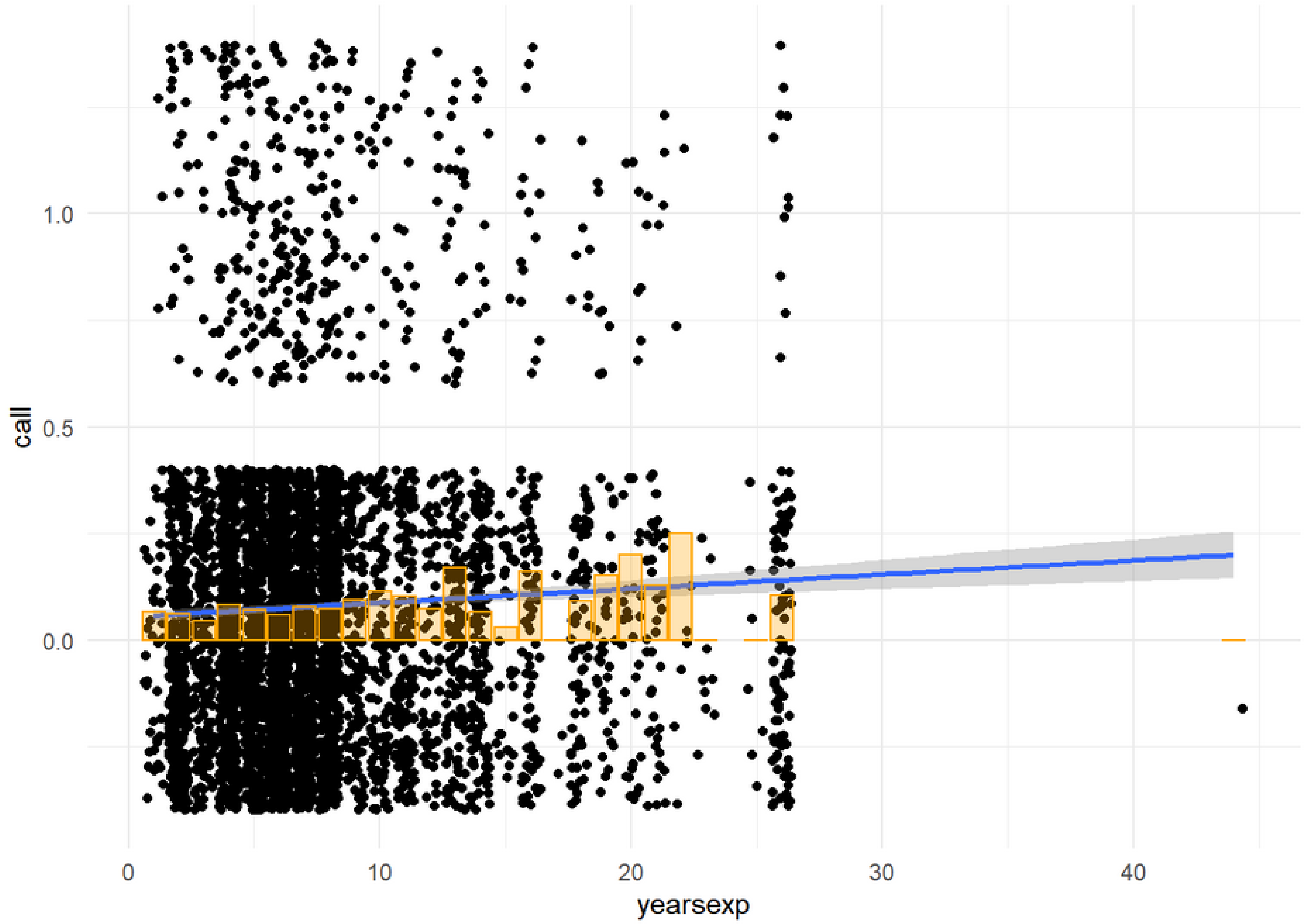
Alternatively use a model with a separate effect for every year of experience; e.g. with 13 years the probability goes up by 10 percentage points relative to reference group

```
(linear1 <- lm(call ~ factor(yearsexp), data = bm)) %>% summary()
```

```
##
## Call:
## lm(formula = call ~ factor(yearsexp), data = bm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25000 -0.08194 -0.07246 -0.05998  0.97059
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0666667  0.0404239   1.649   0.0992 .
## factor(yearsexp)2 -0.0041667  0.0429301  -0.097   0.9227
## factor(yearsexp)3 -0.0202749  0.0448679  -0.452   0.6514
## factor(yearsexp)4  0.0152700  0.0420835   0.363   0.7167
## factor(yearsexp)5  0.0043393  0.0421797   0.103   0.9181
## factor(yearsexp)6 -0.0066911  0.0415222  -0.161   0.8720
## factor(yearsexp)7  0.0109673  0.0420715   0.261   0.7943
## factor(yearsexp)8  0.0059977  0.0419680   0.143   0.8864
## factor(yearsexp)9  0.0276730  0.0457883   0.604   0.5456
## factor(yearsexp)10  0.0487179  0.0469013   1.039   0.2990
## factor(yearsexp)11  0.0373796  0.0453778   0.824   0.4101
## factor(yearsexp)12  0.0057971  0.0519596   0.112   0.9112
## factor(yearsexp)13  0.1021645  0.0459520   2.223   0.0262 *
## factor(yearsexp)14  0.0004474  0.0461260   0.010   0.9923
## factor(yearsexp)15 -0.0372549  0.0616186  -0.605   0.5455
## factor(yearsexp)16  0.0929078  0.0491566   1.890   0.0588 .
## factor(yearsexp)17 -0.0666667  0.1616955  -0.412   0.6801
## factor(yearsexp)18  0.0242424  0.0508830   0.476   0.6338
## factor(yearsexp)19  0.0855072  0.0568565   1.504   0.1327
## factor(yearsexp)20  0.1333333  0.0611152   2.182   0.0292 *
## factor(yearsexp)21  0.0609929  0.0565566   1.078   0.2809
## factor(yearsexp)22  0.1833333  0.1040473   1.762   0.0781 .
## factor(yearsexp)23 -0.0666667  0.0990179  -0.673   0.5008
## factor(yearsexp)25 -0.0666667  0.1101769  -0.605   0.5451
## factor(yearsexp)26  0.0391026  0.0483854   0.808   0.4190
## factor(yearsexp)44 -0.0666667  0.2741681  -0.243   0.8079
## ---
```

Effect of experience on call back

```
agg2=agg=bm %>% group_by(yearsexp) %>% summarise(call=mean(call))
ggplot(bm,aes(x=yearsexp,y=call) ) + geom_jitter()+
          geom_smooth(method="lm")+
          geom_bar(data=agg2,aes(x=yearsexp,y=call),
                  stat="identity",color="orange",
                  fill="orange",alpha=0.3)+
         theme_minimal()
```



Other non-linear relationships

- Relationship between explanatory and dependent variables may be non-linear

- There are general methods to deal with this
- However, in many cases we can avoid using different methods because many types of seemingly non-linear relationship can be represented in what boils down to a linear regression.
- e.g. suppose you suspect that the relationship between wage and education in wage1.dta is actually following a quadratic form:

$$Wage = \beta_0 + \beta_1 EDU + \beta_2 EDU^2 + \epsilon$$

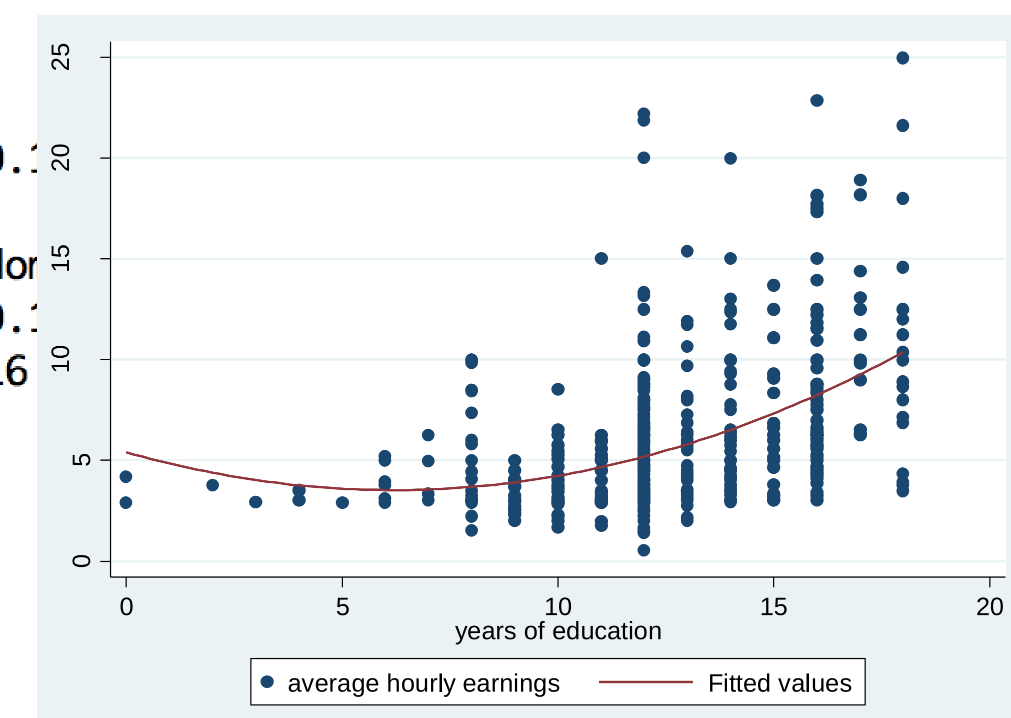
Your turn: Any ideas how to deal with this?

A square relationship

wage1["educ2"] =wage1$educ^2

```
>   summary(lm(wage ~ educ+educ2, wage1))

Call:
lm(formula = wage ~ educ + educ2, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8722 -2.0002 -0.7472  1.2642 17.0159 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.40769    1.45886   3.707 0.000232 ***
educ        -0.60750    0.24149  -2.516 0.012181 *  
educ2        0.04907    0.01007   4.872 1.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 **’ 0.01 ‘*’ 0.05 ‘.’ 0.

Residual standard error: 3.  on 523 degrees of freedor
Multiple R-squared:  0.201     Adjusted R-squared:  0.
F-statistic: 65.79 on 2 a     DF,  p-value: < 2.2e-16
```

Seems to be significant



How to interpret things?

Note what we do in the linear case $Y = \beta X + \epsilon$

$$\frac{\partial Y}{\partial X} = \beta$$

In the linear case β is the marginal effect of X on Y

We can work out the same thing in the nonlinear case

$$Y = \beta_1 X + \beta_2 X^2 + \epsilon \longrightarrow \frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

- *The marginal effect (how much Y changes in response to change in X) varies for different values of X*
- *We can also find the extreme point by looking at* $\frac{\partial Y}{\partial X} = 0$

log-linear relationships

- The most popular non-linear model is probably

$$Y = \exp(\beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon)$$

- To make it linear all that is required is to take the (natural) logarithm on both sides of the equation:

$$\ln Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

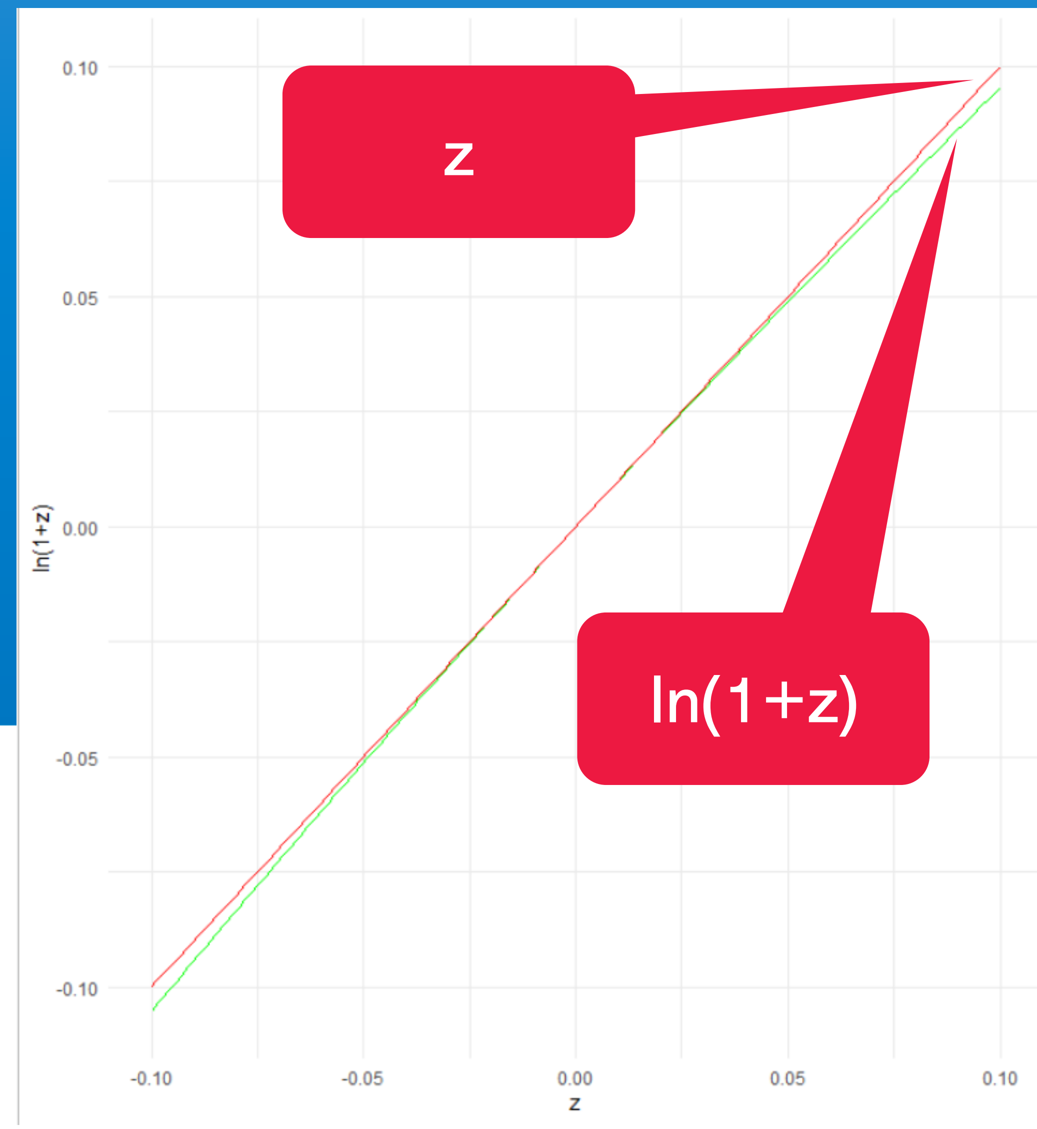
- One of the reasons why it's popular is the interpretation of the β coefficients it implies

Interpreting log-linear relationships

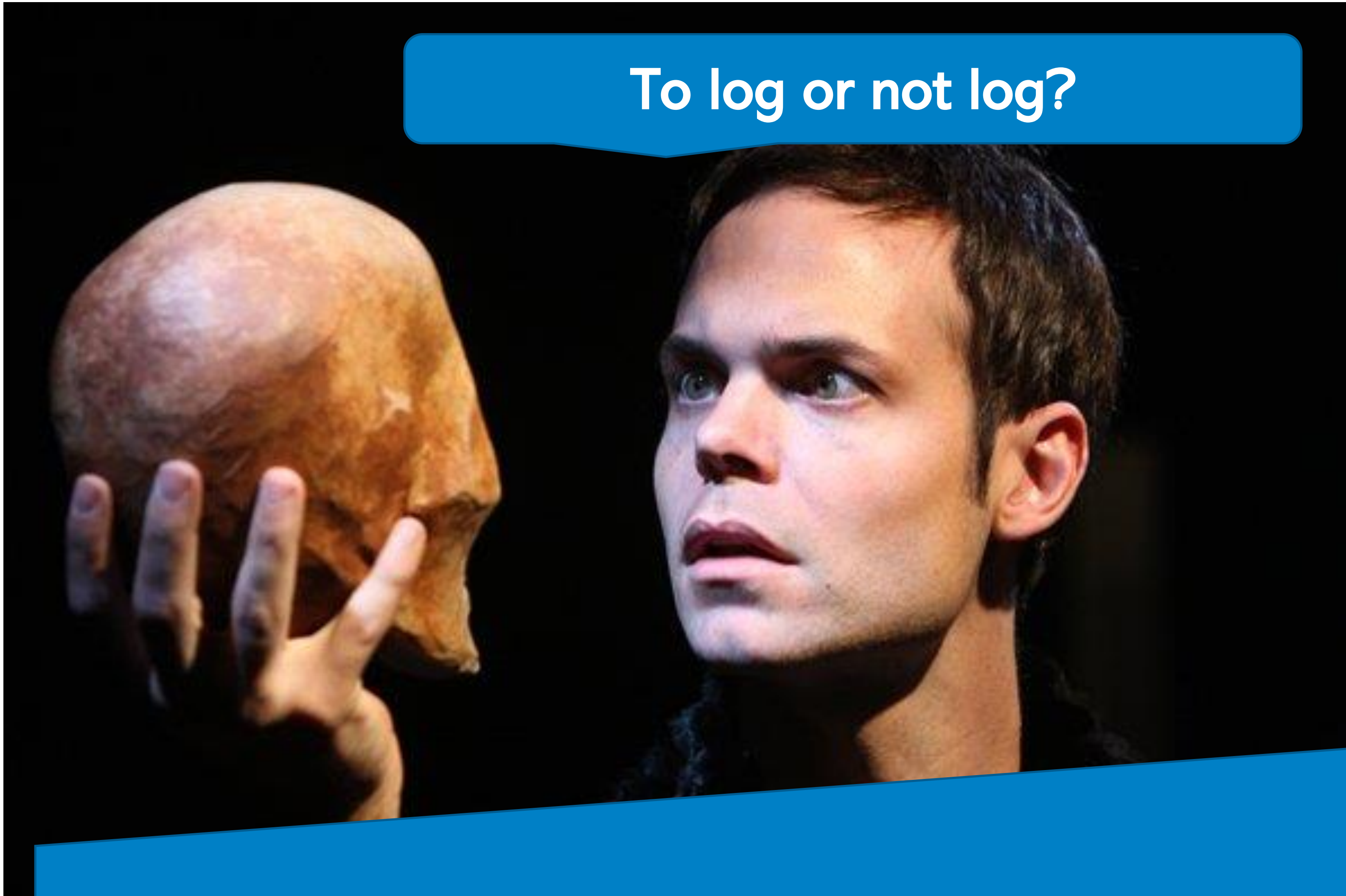
$$\ln Y = \cdots + \beta \quad X \quad + \epsilon$$

- As before: change in dependent variable for given change in X
- However now:
- $DepVar_a - DepVar_b$
- $= \ln Y_a - \ln Y_b$
- $= \ln\left(\frac{Y_a}{Y_b}\right)$
- $= \ln\left(1 + \frac{Y_a}{Y_b} - 1\right)$
- $\approx \frac{Y_a}{Y_b} - 1$
- $= \frac{Y_a - Y_b}{Y_b}$

Let $z = \frac{Y_a}{Y_b} - 1$



Hence, β captures (approximately) the Growth in dependent variable Y when we change X by 1 unit



Which is more plausible?

1. Change in X leads to fixed change in Y→ use Y
2. Change in X leads to fixed percentage change in Y→ use $\ln Y$

Going log crazy: log log

$$\ln Y = \cdots + \beta \ \ln X \ + \epsilon$$

- As before: change in dependent variable for given change in X
- However now:

$$\beta = \frac{DepVar_a - DepVar_b}{XVar_a - XVar_b} = \frac{\ln Y_a - \ln Y_b}{\ln X_a - \ln X_b}$$

$$= \frac{\frac{Y_a - Y_b}{Y_b}}{\frac{X_a - X_b}{X_b}}$$

Elasticity

Famous example: production functions

Cobb Douglas production function:

employment

output

$$Y = AL^{\alpha_L} K^{\alpha_K}$$

capital

Productivity shock: $A_0 \exp(\epsilon)$

Taking logs:

$$\ln Y = \alpha_0 + \alpha_L \ln L + \alpha_K \ln K + \epsilon$$

Elasticity of output with respect to a change in employment

Summary

- Don’t fall in the dummy variable trap
- The same model can be represented in several ways
- Be careful with interpretation of dummies
- A lot of stuff that looks non-linear at first glance is linear after all

Extra Slides

So many dummies

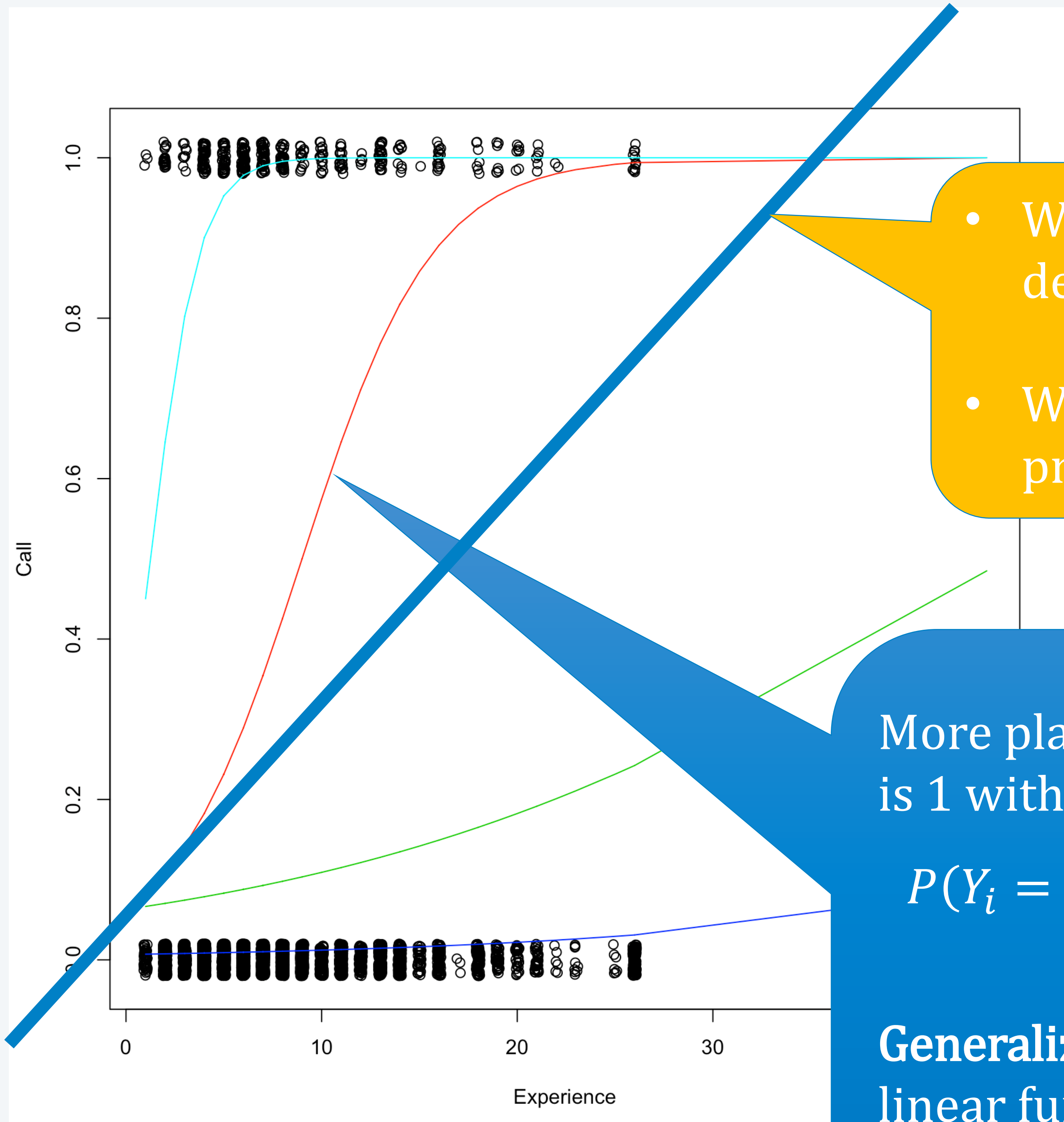
- Note that we often use large numbers of dummies in regression models
- Helps to control for a wide range of potential confounding factors

Accounts for sector level confounding factors

Examples:

- Sector dummies when using firm level data from several sectors
- Regional dummies
- Firm dummies when having panel data for firms
- Time dummies when having panel data

Revisiting dummies as dependent variables



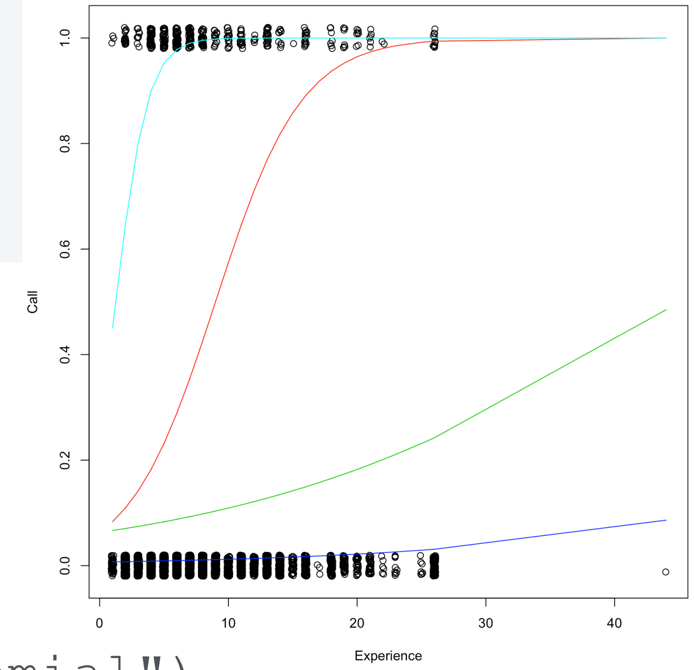
- We saw we could model the probability of the dependent variable equal to 1 as a line
$$P(Y_i = 1|X_i) = \beta_0 + \beta_1 X$$
- We could interpret β_1 as the change in probability in response to change in X

More plausible model probability that outcome dummy is 1 with the logistic function:

$$P(Y_i = 1|X_i) = \frac{1}{1 + \exp(-[\beta_0 + \beta_1 X])} = g(\beta_0 + \beta_1 X)$$

Generalized linear model: A linear model inside a non-linear function

Implement logit in R




```
> library(aod)
>   library(margins )
>   logit <- glm(call ~ yearsexp, data = bm, family = "binomial")
>   summary(logit)

Call:
glm(formula = call ~ yearsexp, family = "binomial", data = bm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7780  -0.4075  -0.3924  -0.3779   2.3598

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.75960    0.09620 -28.687  < 2e-16 ***
yearsexp     0.03908    0.00918   4.257 2.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2726.9  on 4869  degrees of freedom
Residual deviance: 2710.2  on 4868  degrees of freedom
```

How to interpret things?

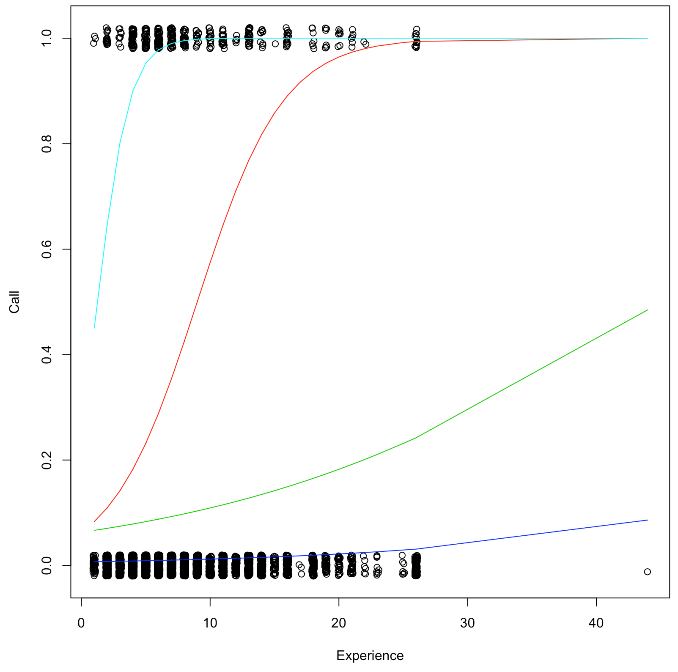
Note what we do in the linear case $Y = \beta X + \epsilon$

$$\frac{\partial Y}{\partial X} = \beta$$

In the linear case β is the marginal effect of X on Y

We can work out the same thing in the nonlinear case

How to interpret things?



Compute marginal effect on of variable on probability of Y=1

$$\frac{\partial P(Y_i = 1|X_i)}{\partial X_i} = \frac{exp(-\hat{\beta}X_i)\hat{\beta}}{[1 + exp(-\hat{\beta}X_i)]^2}$$

Not constant for different observations. So we have to compute for every observation separately. Or at specific observations we are interested in.
R margins() command will help

```
>   margins(logit)
Average marginal effects
glm(formula = call ~ yearsexp, fam

 yearsexp
 0.002881
```

Can be compared to linear probability model

```
## Coefficients:
##              Estimate  Std. Error
## (Intercept)  0.0545046  0.0071949
## yearsexp     0.0033136  0.0007716
## ---
```

Interactions?

In the context of the model regressing wages on skill and gender we might ask if income gap between man and women is the same irrespective of the educational group; e.g.

$$E\{Wage|Women, Low\} - E\{Wage|Men, Low\}$$

$=E\{Wage|Women, Normal\} - E\{Wage|Men, Normal\}$?

Notice that in terms of the model we used before it is impossible to have differences in the gap across education group.

$$Wage = \beta_C + \beta_{normal} normal + \beta_{high} high + \beta_{female} female + \epsilon$$

By construction the gap is always β_{female}

A more complex model

We can get a more complex model using interactions:

$$Wage = \beta_C + \beta_{normal} normal + \beta_{high} high$$

$$+\beta_{female} female$$

$$+\beta_{fem \times norm} normal \times female$$

$$+\beta_{fem \times high} high \times female \ + \epsilon$$

A more complex model

We can get a more complex model using interactions:

$$E\{Wage|Women, Low\} - E\{Wage|Men, Low\} = \beta_{female}$$

$$E\{Wage|Women, Normal\} - E\{Wage|Men, Normal\} = \beta_{female} + \beta_{fem \times norm}$$

How would you test if the wage gap is different for normally educated persons?

Interactions?

```
> summary(lm(wage ~ educatsf*gender, wage1))

Call:
lm(formula = wage ~ educatsf * gender, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5235 -1.8394 -0.4407  0.9131 16.5365 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   4.6780     0.4057  11.530  < 2e-16 ***
educatsfnormal                2.2683     0.5371   4.223 2.85e-05 ***
educatsfhigh                  3.7656     0.4989   7.548 1.99e-13 ***
genderfemale                 -1.3859     0.6059  -2.287   0.0226 *  
educatsfnormal:genderfemale  -1.3737     0.7644  -1.797   0.0729 .  
educatsfhigh:genderfemale    -1.1749     0.7567  -1.553   0.1211    
---
```

How does the wage gap differ for different educational groups?

Interactions and continuous variables

```
> summary(lm(wage ~ educatsf*exper, wage1))

Call:
lm(formula = wage ~ educatsf * exper, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3796 -1.9595 -0.6658  1.3310 16.3972 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.468072   0.501029   6.922 1.32e-11 ***
educatsfnormal       1.294190   0.650598   1.989  0.04720 *  
educatsfhigh         2.469483   0.617135   4.002 7.21e-05 ***
exper                0.028100   0.018721   1.501  0.13396    
educatsfnormal:exper 0.005466   0.026468   0.206  0.83648    
educatsfhigh:exper   0.077254   0.027425   2.817  0.00503 ** 
---
```

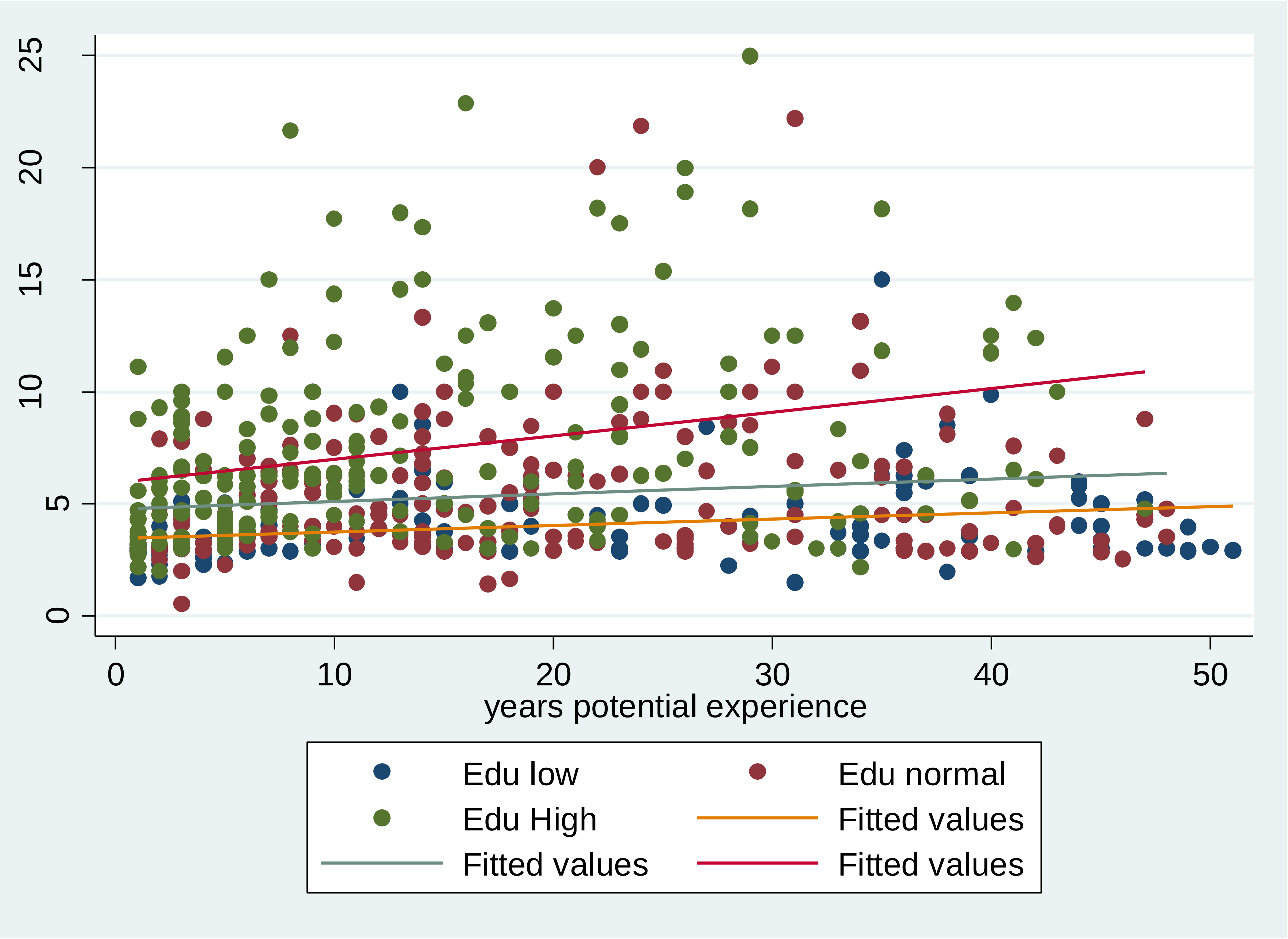
Your turn: Does experience affect the educational groups differently?

(a) No

(b) Yes: experience has a bigger impact for the highest educated

(c) Yes: experience has a smaller impact for the highest educated

Interactions and continuous variables



Relating the figure to the model

The overall model is:

$$Wage = \beta_C + \beta_{normal} normal + \beta_{high}\, high + \beta_{exper} exper$$

$$+\beta_{normal \times exper}\, normal \times exper \ + \beta_{high \times exper}\, high \times exper + \epsilon$$

Topline (EDU high): $\beta_C + \beta_{high} + \left(\beta_{exper} + \beta_{high \times exper}\right) \times exper$

Middle line (EDU normal): $\beta_C + \beta_{normal} + \left(\beta_{exper} + \beta_{normal \times exper}\right) \times exper$

Bottom line (EDU low): $\beta_C + \beta_{exper} \times exper$