Advanced Data Analysis I
Nonlinear Relations & Regression Assumptions

PA 541 Week 8

Michael D. Siciliano
Department of Public Administration
College of Urban Planning and Public Affairs

- Topics
  - Review (i) interactions and (ii) qualitative predictors
  - Non-linear relationships: quadratics (will discuss log models after the midterm)
  - Review our regression assumptions
    - Learn how to test for and correct heteroskedasticity and multicollinearity
  - Let’s put what we have learned into practice with a case study

Admin

- Next Week: **<u>Midterm</u>**
  - Two-sided sheet of hand-written notes allowed (honor system)
    - No other materials can be used during the exam.
    - You will submit your note sheet with your exam.
  - Midterm will be taken during class time. Designed to be finished in about 2 hours; but you have the full three hours to complete if needed.
  - Midterm will cover the first 8 weeks of class. Interpretation only; no coding.

- **On the Horizon**
  - **Week 9**: Midterm
  - **Week 10**: Model specification and data issues. Log Models. Data screening and cleaning. Methods for handling outliers and missing data.
  - **Week 11**: Spring Break
  - **Week 12**: Logistic Regression
  - **Week 13/14**: Panel Data

Science & technology

Feb 6th 2021 edition ›

Daughters and divorce

Parents of daughters are more likely to divorce than those with sons

But the difference only emerges when the children are teenagers

Starter Question

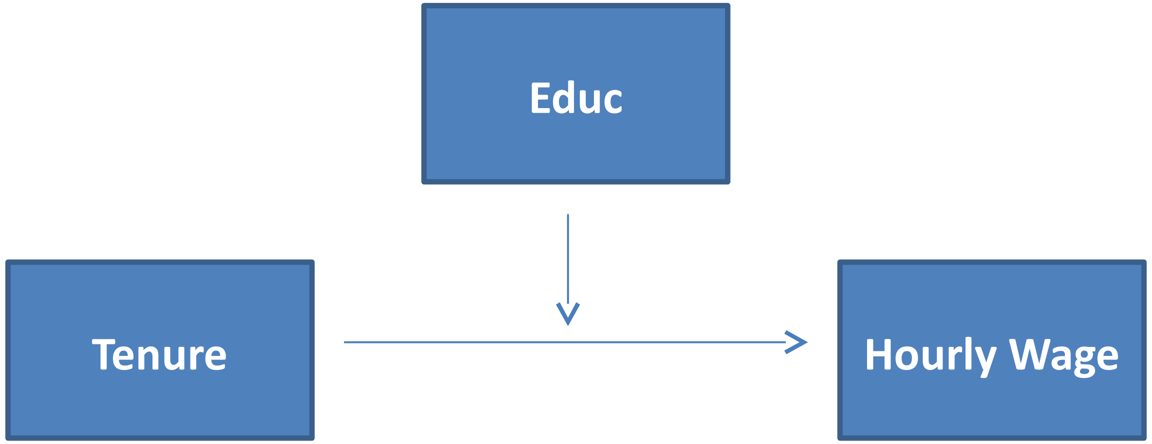
Someone commented on this article: “Correlation is not causation.” Does that old quip apply here?

Review Interactions

A model with an interaction

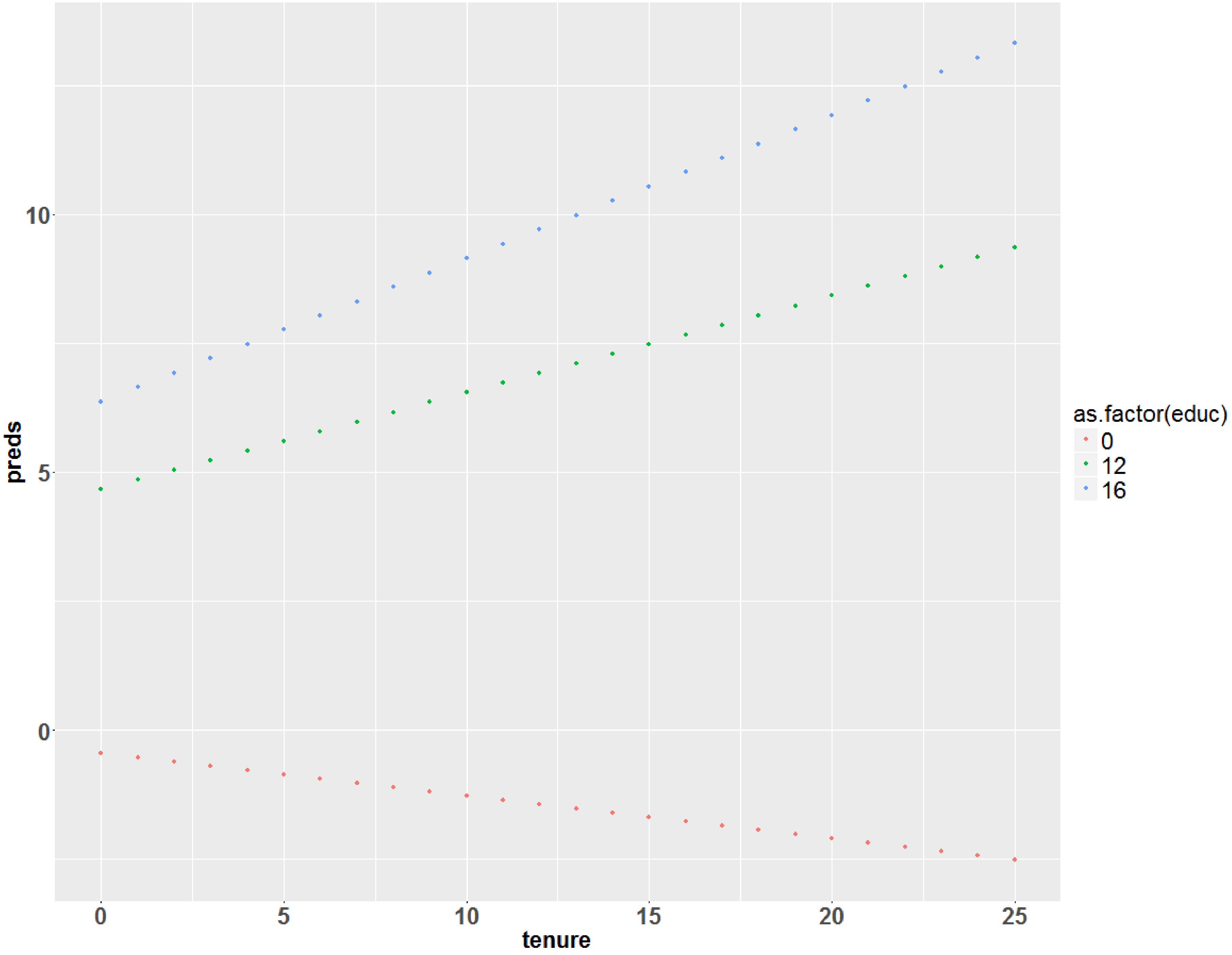
	Uncentered Model
(Intercept)	-0.449 (0.785)
educ	0.426 (0.061)***
tenure	-0.083 (0.074)
educ:tenure	0.023 (0.006)***
R^2	0.320
Adj. R^2	0.316
Num. obs.	525
RMSE	3.055

***p < 0.01, **p < 0.05, *p < 0.1



- How do we interpret each of the coefficients in the model? (assume education is the moderator)
- Does this mean that tenure is not important?

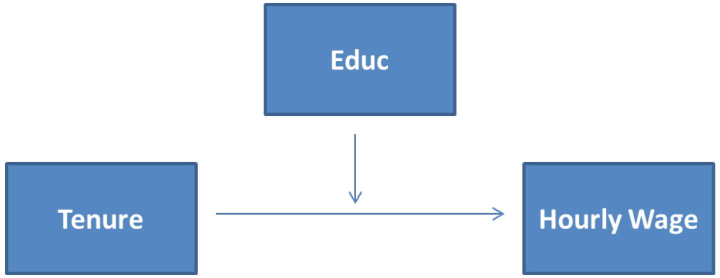
Uncentered model visually…



With education centered

	Uncentered Model	Centered Model
(Intercept)	-0.449 (0.785)	4.661 (0.167)***
educ	0.426 (0.061)***	
tenure	-0.083 (0.074)	0.188 (0.019)***
educ:tenure	0.023 (0.006)***	
educ12		0.426 (0.061)***
educ12:tenure		0.023 (0.006)***
R^2	0.320	0.320
Adj. R^2	0.316	0.316
Num. obs.	525	525
RMSE	3.055	3.055

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$



- What happened to the intercept?

With education centered cont.

	Uncentered Model	Centered Model
(Intercept)	-0.449 (0.785)	4.661 (0.167)***
educ	$\boxed{0.426\ (0.061)^{***}}$	
tenure	-0.083 (0.074)	0.188 (0.019)***
educ:tenure	0.023 (0.006)***	
educ12		$\boxed{0.426\ (0.061)^{***}}$
educ12:tenure		0.023 (0.006)***
R^2	0.320	0.320
Adj. R^2	0.316	0.316
Num. obs.	525	525
RMSE	3.055	3.055

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

- Why is the slope of *educ* and *educ12* the same?

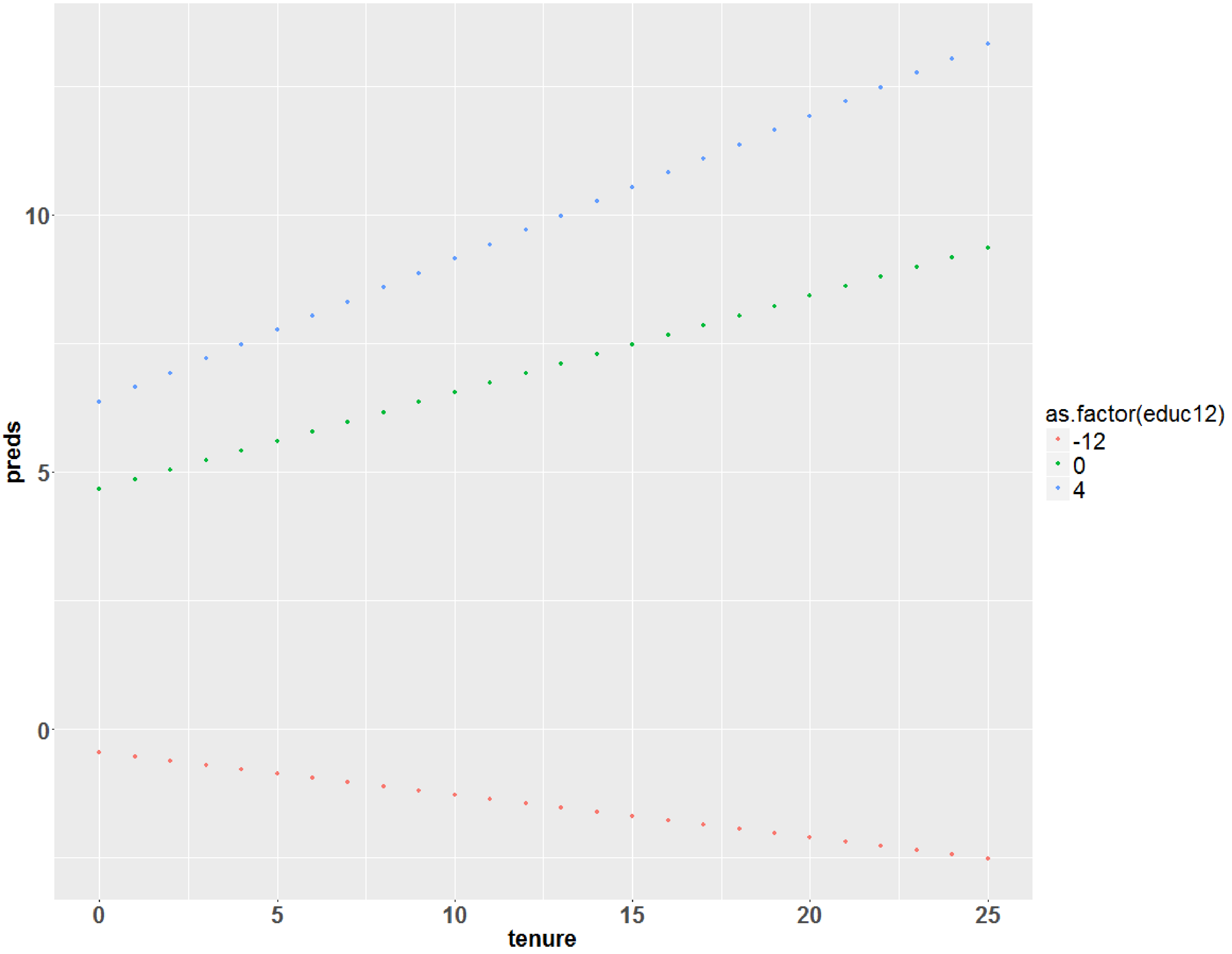
With education centered cont.

	Uncentered Model	Centered Model
(Intercept)	-0.449 (0.785)	4.661 (0.167)***
educ	0.426 (0.061)***	
tenure	-0.083 (0.074)	0.188 (0.019)***
educ:tenure	0.023 (0.006)***	
educ12		0.426 (0.061)***
educ12:tenure		0.023 (0.006)***
R^2	0.320	0.320
Adj. R^2	0.316	0.316
Num. obs.	525	525
RMSE	3.055	3.055

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

- What happened to the effect of tenure?

Centered model visually…



Is there any difference?

	Uncentered Model	**Centered Model**
(Intercept)	-0.449 (0.785)	4.661 (0.167)***
educ	0.426 (0.061)***	
tenure	-0.083 (0.074)	0.188 (0.019)***
educ:tenure	0.023 (0.006)***	
educ12		0.426 (0.061)***
educ12:tenure		0.023 (0.006)***
R^2	0.320	0.320
Adj. R^2	0.316	0.316
Num. obs.	525	525
RMSE	3.055	3.055

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

- What is the effect of an additional year of tenure for someone with a highschool education?
  - **Centered Model**: just the slope of tenure = **0.188**

- **Two takeaways:**
  - Be careful to avoid concluding that a variable involved in an interaction is not important based on its simple main effect.
  - Generally best to center predictor variables if there is an interest in interpreting simple main effects (especially if the zero point is outside the relevant range).

Review: Categorical predictors

Two Categorical Variable Model

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- Where
  - yi = hourly wage of a worker
  - D1i = 1 if the worker is a man, 0 otherwise
  - D2i = 1 if white, 0 otherwise
  - Xi = years of experience

- Based on the dummy variables there are four possible groups: white males, non-white males, white females, and

Model with no interactions

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- Mean earnings of non-white women
  - E(yi|xi, D1=0, D2=0) = α0 + βxi
- Mean earnings of non-white men
  - E(yi|xi, D1=1, D2=0) = (α0 + α1) + βxi
- Mean earnings of white women
  - E(yi|xi, D1=0, D2=1) = (α0 + α2) + βxi
- Mean earnings of white men

D1 is dummy for gender, where female = 0, male =1

D2 is dummy for race, where nonwhite = 0 white = 1

Model with interactions

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{1i} D_{2i} + \beta x_i + \varepsilon_i$$

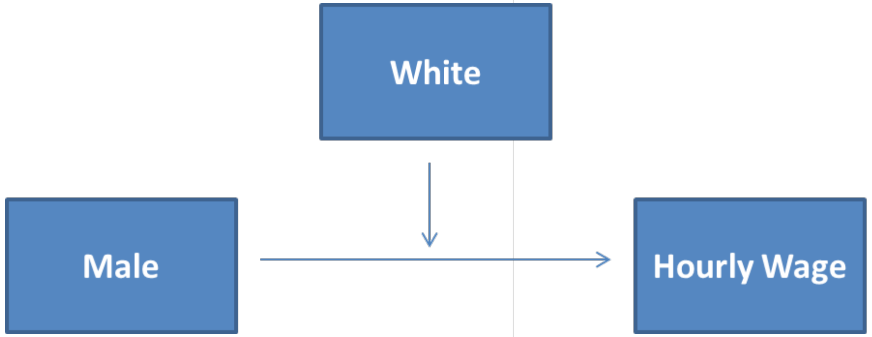
- Mean earnings of non-white women
  - E(yi|xi, D1=0, D2=0) = α0 + βxi
- Mean earnings of non-white men
  - E(yi|xi, D1=1, D2=0) = (α0 + α1) + βxi
- Mean earnings of white women
  - E(yi|xi, D1=0, D2=1) = (α0 + α2) + βxi
- Mean earnings of white men
  - E(yi|xi, D1=1, D2=1) = (α0 + α1 + α2 + α3) + βxi

D1 is dummy for gender, where female = 0, male =1

D2 is dummy for race, where nonwhite = 0 white = 1

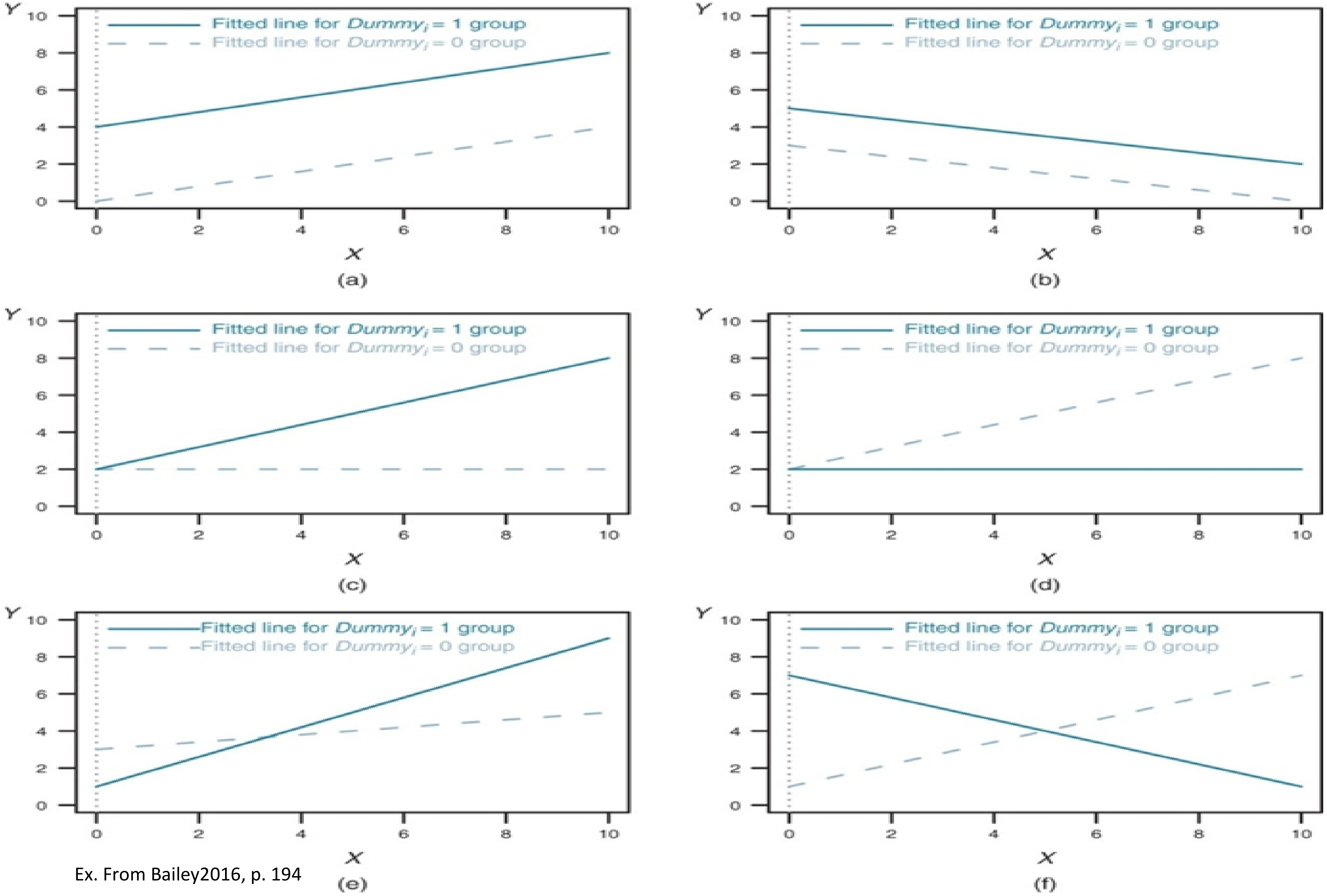
Output from Two Categorical Variable Model

Output from Categorical Interaction Model



Review Question

- For the following diagrams, write down whether each of the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ in the model is less than, equal to, or greater than zero.



Ex. From Bailey2016, p. 194

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Dummy_i + \beta_3 Dummy_i * X_i + e_i$$

Non-linear relationships

Can we still use linear models?

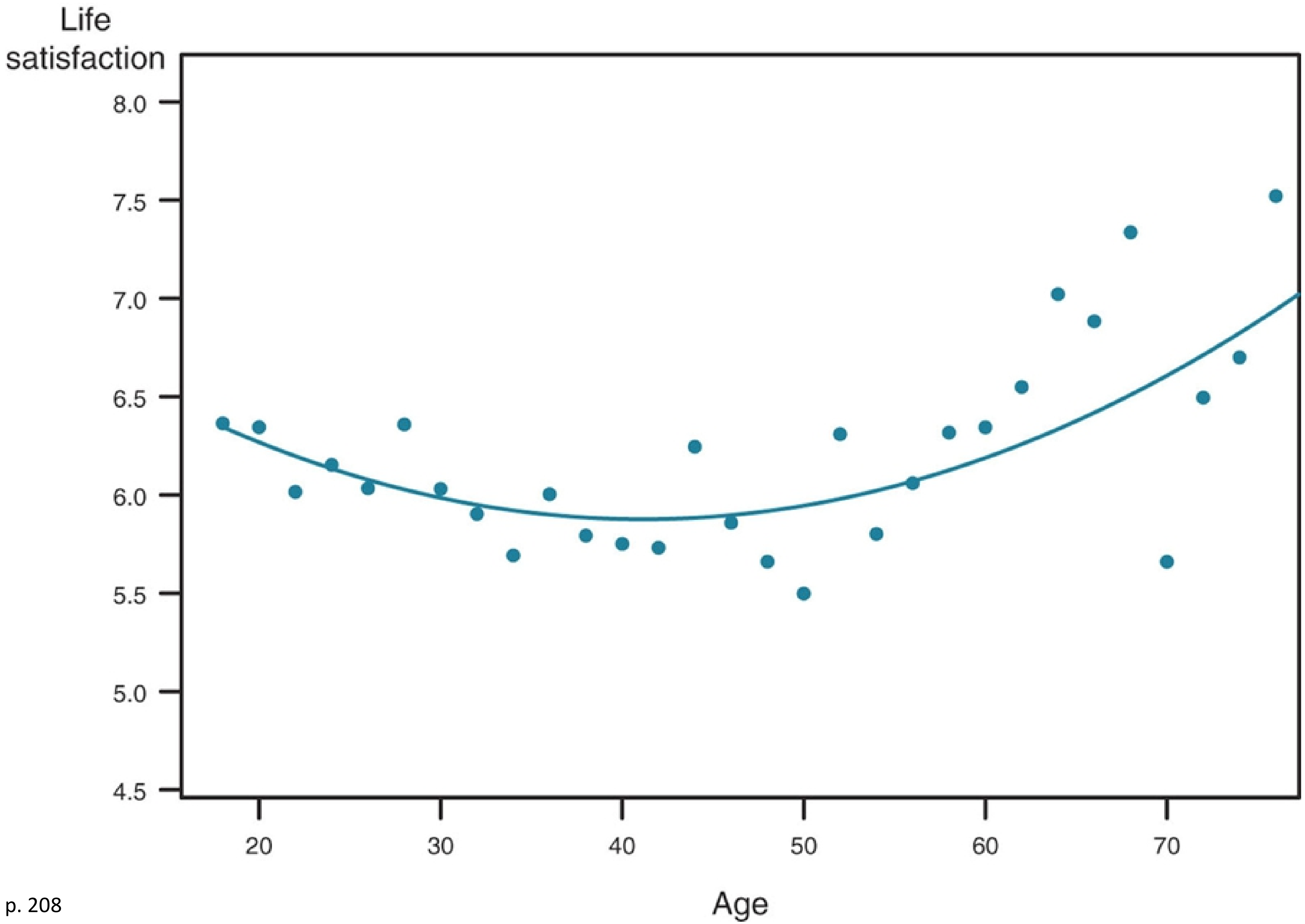
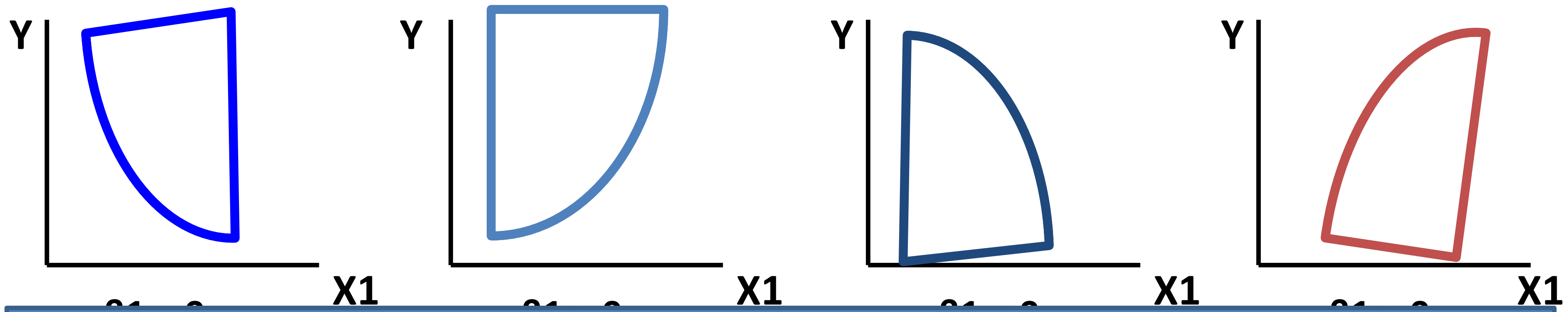


Image from Bailey 2016, p. 208

Quadratic Regression Model

Quadratic models may be considered when the scatter diagram takes on one of the following shapes:

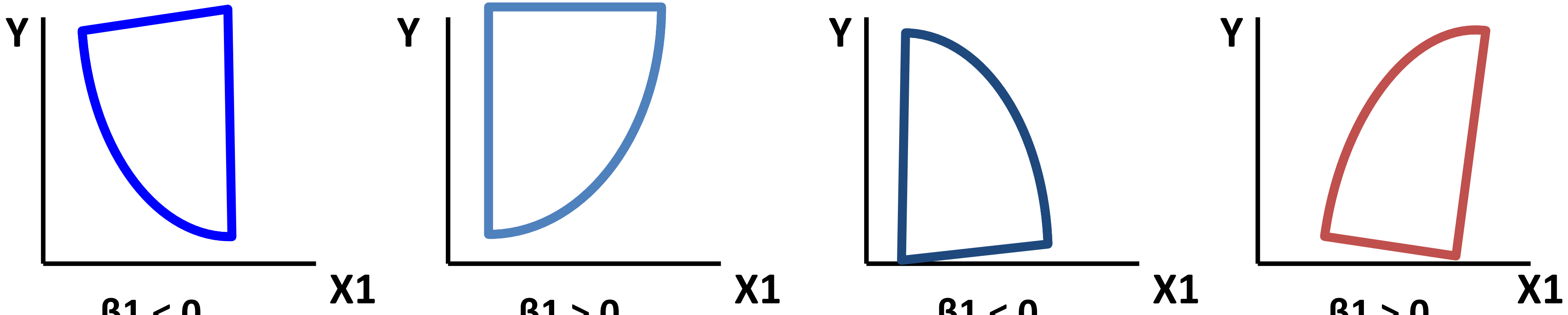


β1 = the coefficient of the linear term
β2 = the coefficient of the squared term

Image from Statistics for Business and Economics, 6e

Quadratic Regression Model

Quadratic models may be considered when the scatter diagram takes on one of the following shapes:



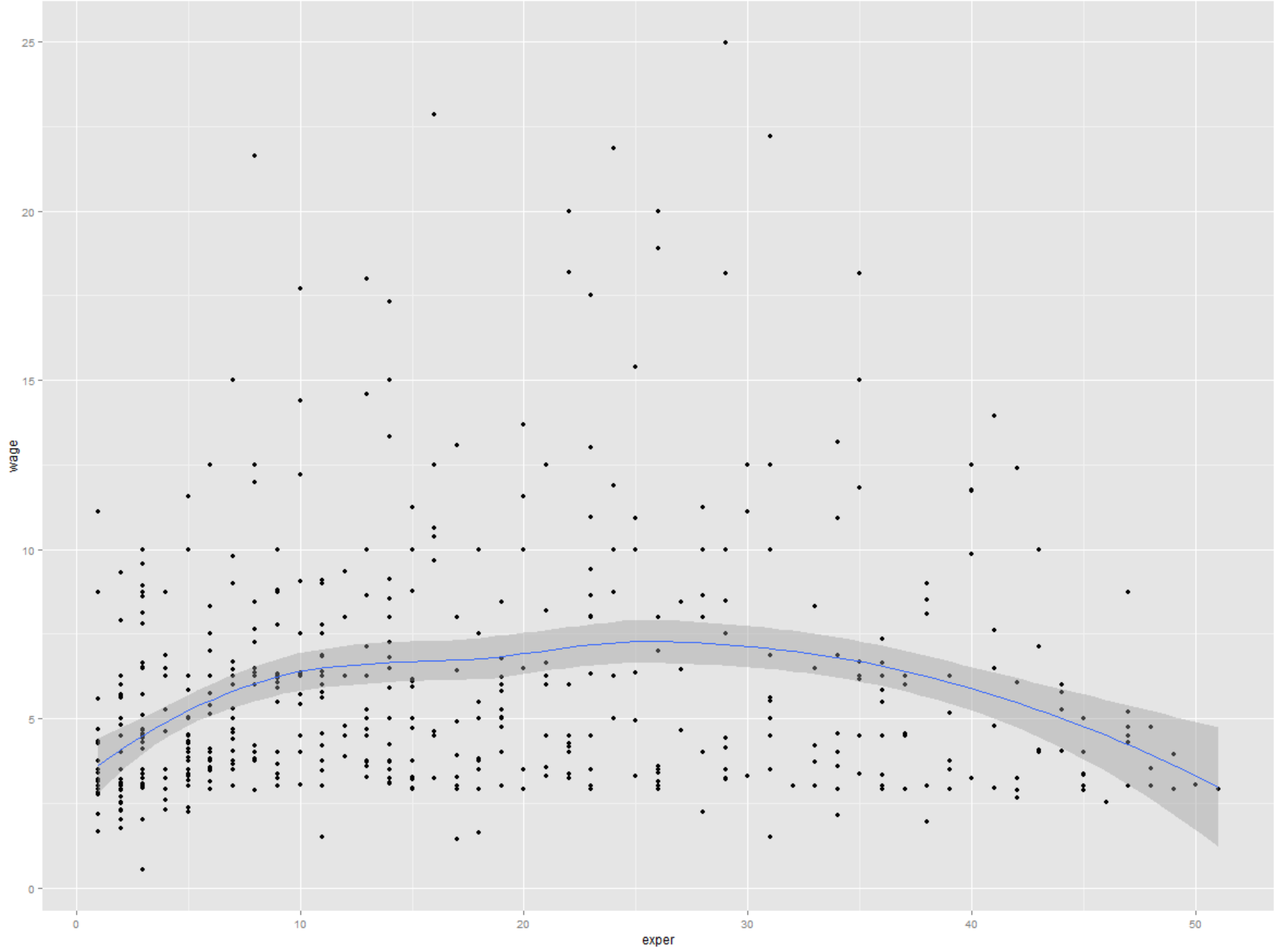
$\beta_1 < 0$	$\beta_1 > 0$	$\beta_1 < 0$	$\beta_1 > 0$
$\beta_2 > 0$	$\beta_2 > 0$	$\beta_2 < 0$	$\beta_2 < 0$

β_1 = the coefficient of the linear term
β_2 = the coefficient of the squared term

Image from Statistics for Business and Economics, 6e

Wage and experience

ggplot(data = wage, aes(x=exper, y=wage)) + geom_point() + stat_smooth()



- Here is a scatterplot of wage versus experience.
- Provides justification for considering a squared term in a model.

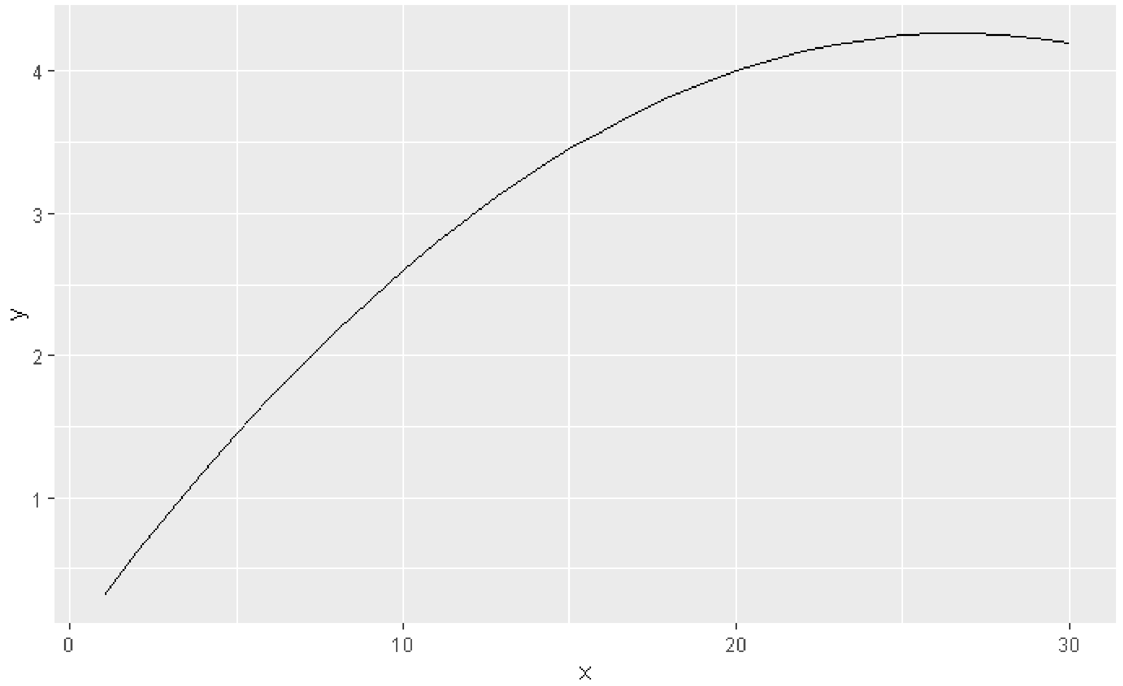
$$wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$$

```
> lm1 = lm(wage ~ exper + expersq, data=wage)
> summary(lm1)


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7254058  0.3459392  10.769  < 2e-16 ***
exper        0.2981001  0.0409655   7.277 1.26e-12 ***
expersq     -0.0061299  0.0009025  -6.792 3.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.524 on 523 degrees of freedom
Multiple R-squared: 0.09277,   Adjusted R-squared: 0.0893
F-statistic: 26.74 on 2 and 523 DF,  p-value: 8.774e-12
```

- How do we interpret the results?

Interpreting the Slope Coefficients

- The results on the previous slide indicate that the effect of experience on wage decreases as experience increases. Thus, there are diminishing marginal effects.
- Therefore, the relationship between experience and wage is not fixed. Specifically, the slope decreases as experience



Interpreting the Slope Coefficients

- Remember that our slope is simply the change in y divided by the change in x. We can calculate, at any value of experience, the slope of the relationship between experience and wage by taking the first derivative.

$$wage = 3.73 + 0.298exper + -0.0061exper^2 + \varepsilon$$

$$slope = 0.298 + 2(-0.0061exper) + \varepsilon$$

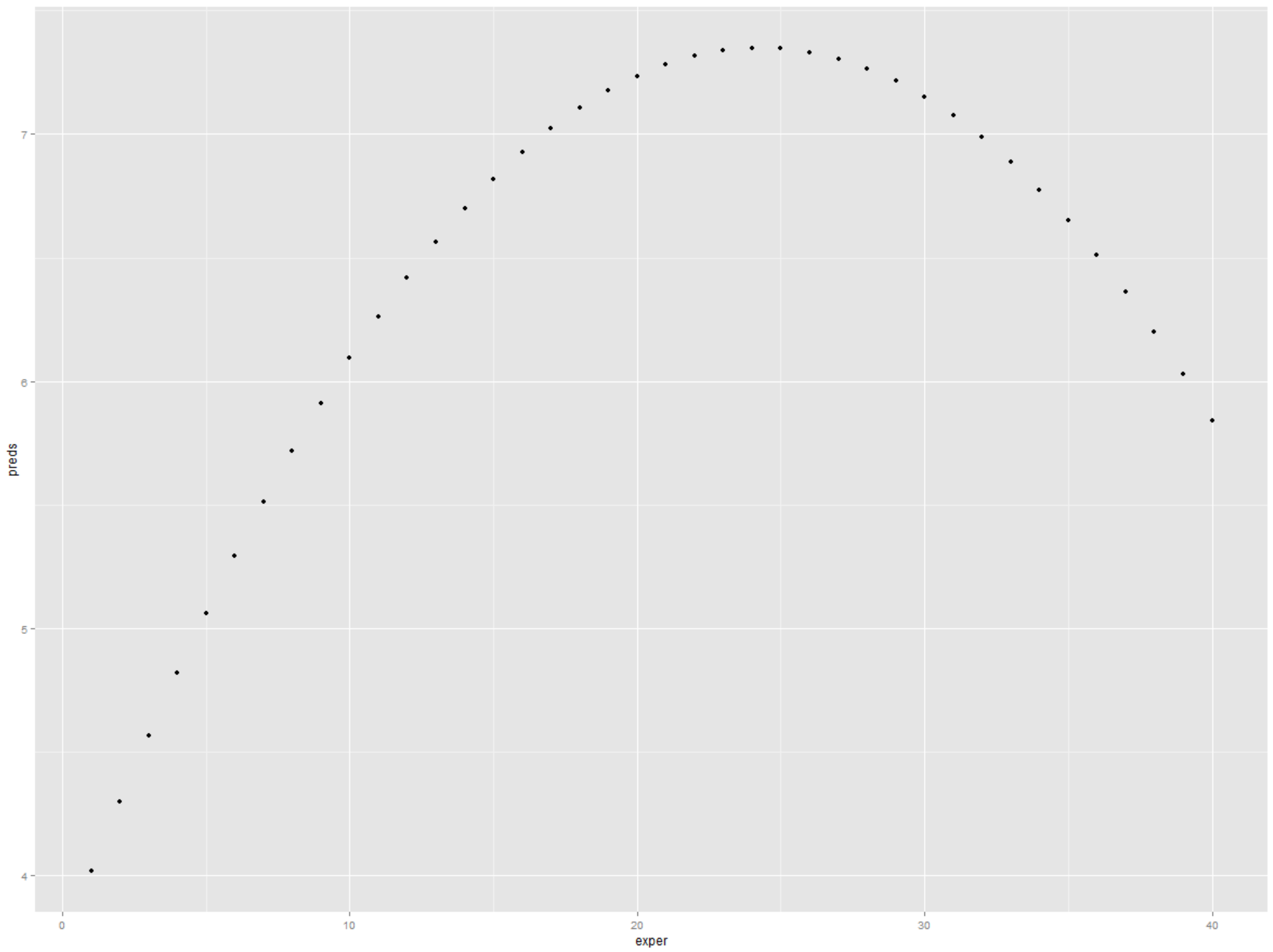
In General: $$\frac{\partial Y}{\partial X_1} = \beta_1 + 2\beta_2 X_1$$

Interpreting Slope Coefficients

- So, we see the first year of experience (going from 0 to 1) is worth
  - .298 + 2*-.0061*0 = 0.298
- The second year of experience (going from 1 to 2) is worth less
  - .298 + 2*-.0061*1 = 0.286
- Going from 10 to 11 years of experience is even less
  - .298 + 2*-.0061*10 = 0 .176

Plot the predicted relationship

```
> newdata = data.frame(exper = seq(from = 1, to = 40, by = 1))
> newdata$expersq = newdata$exper^2
> newdata$preds = predict(lm1, new = newdata)
> ggplot(data = newdata, aes(x=exper, y=preds)) + geom_point()
```



- In Week 10 we will talk about another type of non-linear relationship: logged dependent variables and logged independent variables.
- For now, remember:
  - A quadratic model includes an X variable raised to the power of 2. It has the following form:
    - $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon_i$
  - The effect of X in the model varies depending on the level of X. If we want to estimate the effect of a one-unit change in X we can take the derivative:
    - $\widehat{\beta_1} + 2\widehat{\beta_2}X$

- In Week 10 we will talk about another type of non-linear relationship: logged dependent variables and logged independent variables.
- For now, remember:
  - A quadratic model includes an X variable raised to the power of 2. It has the following form:
  - The effect of X in the model varies depending on the level of X. If we want to estimate the effect of a one-unit change in X we can take the derivative:

Testing Regression Assumptions

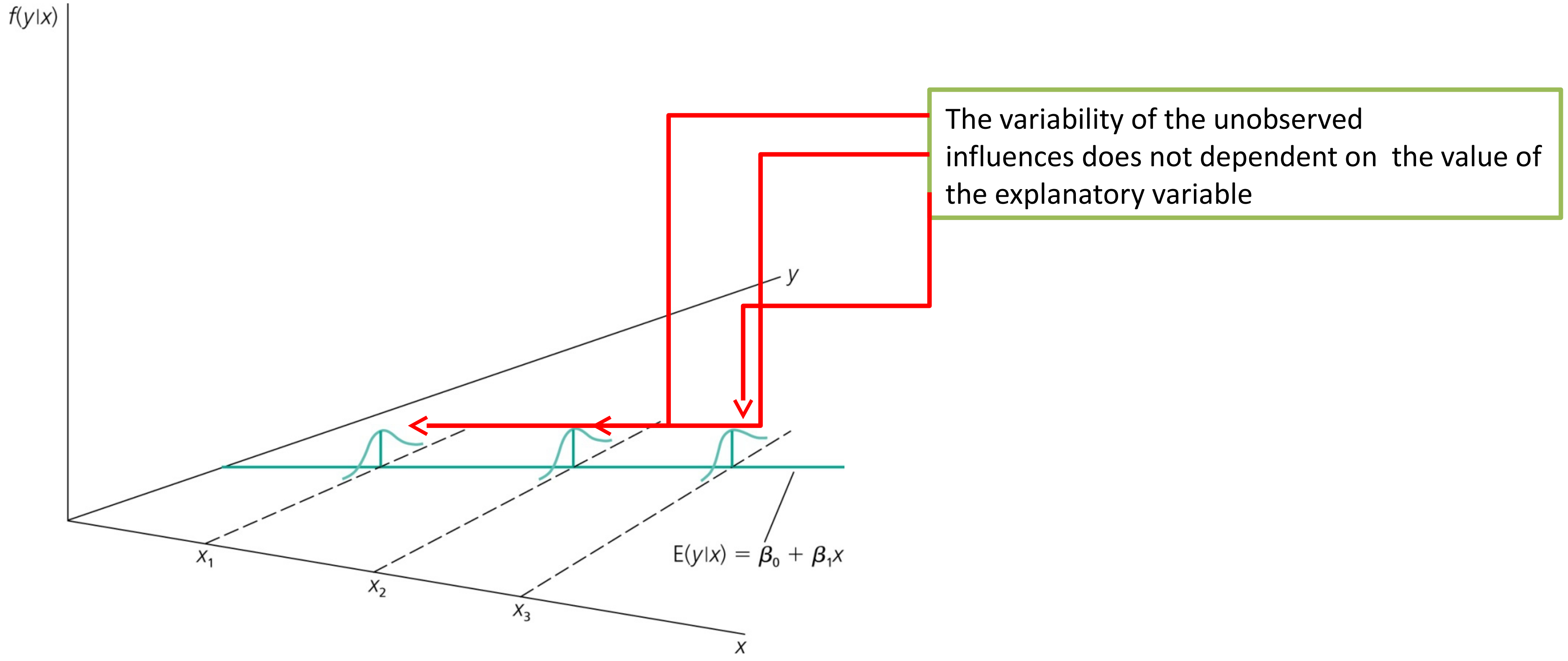
MLR Assumptions

- MLR. 1 - Linear in parameters
- MLR. 2 - Data drawn from a random sample (i.e., the errors are independent - no autocorrelation in the data)
- **MLR.3** - No perfect collinearity (no exact linear relationship among the independent variables)
- MLR.4 - Zero conditional mean for the error term
- **MLR.5** - Homoskedasticity (i.e., the errors have equal variance)
- MLR. 6 - Normality of the error term
- Validity (i.e. data maps to research question)

Assumption: Homoskedasticity

- One of the assumptions of the regression model is homoskedasticity; meaning that the variance in the disturbance term is equal regardless of the values of the predictor variables.
- However, as we have discussed, it is quite possible that the residual variance does in fact vary across observations. When this occurs, we state that our errors are heteroskedastic.

Graphical illustration of homoskedasticity

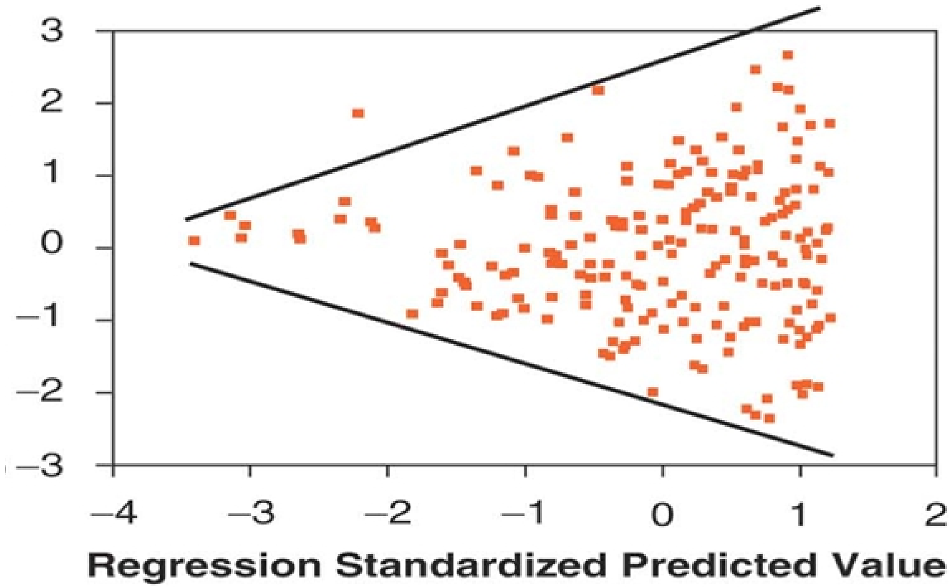


OLS Estimation when Homoskedasticity is violated

- What happens to our estimators when we have heteroskedastic errors:
  - **β estimates are unbiased**. This is true because the assumption necessary to show that our β values are unbiased estimates of the population parameters is that the four first assumptions are met. Heteroskedasticity only violates the assumption of the variance of the residuals, not the expected value. Hence, applying OLS will continue to yield unbiased estimates.
  - However, the **variances of our estimates will be incorrect**. Therefore, inferences drawn with the standard OLS variance formulas (t-tests, F-tests) will be incorrect.

Detecting Heteroskedasticity

- We have already gone through the informal procedures to test the assumption of homoskedasticity; namely checking the relationship between our IVs and the DV and checking the residual plots.



Detecting Heteroskedasticity cont..

- A formal test for heteroskedasticity is the **Breusch-Pagen** test.
- Assume that the other regression assumptions have been met. We can test to see whether any of the independent variables are significant predictors of our residuals (or more specifically, the square of our residuals).

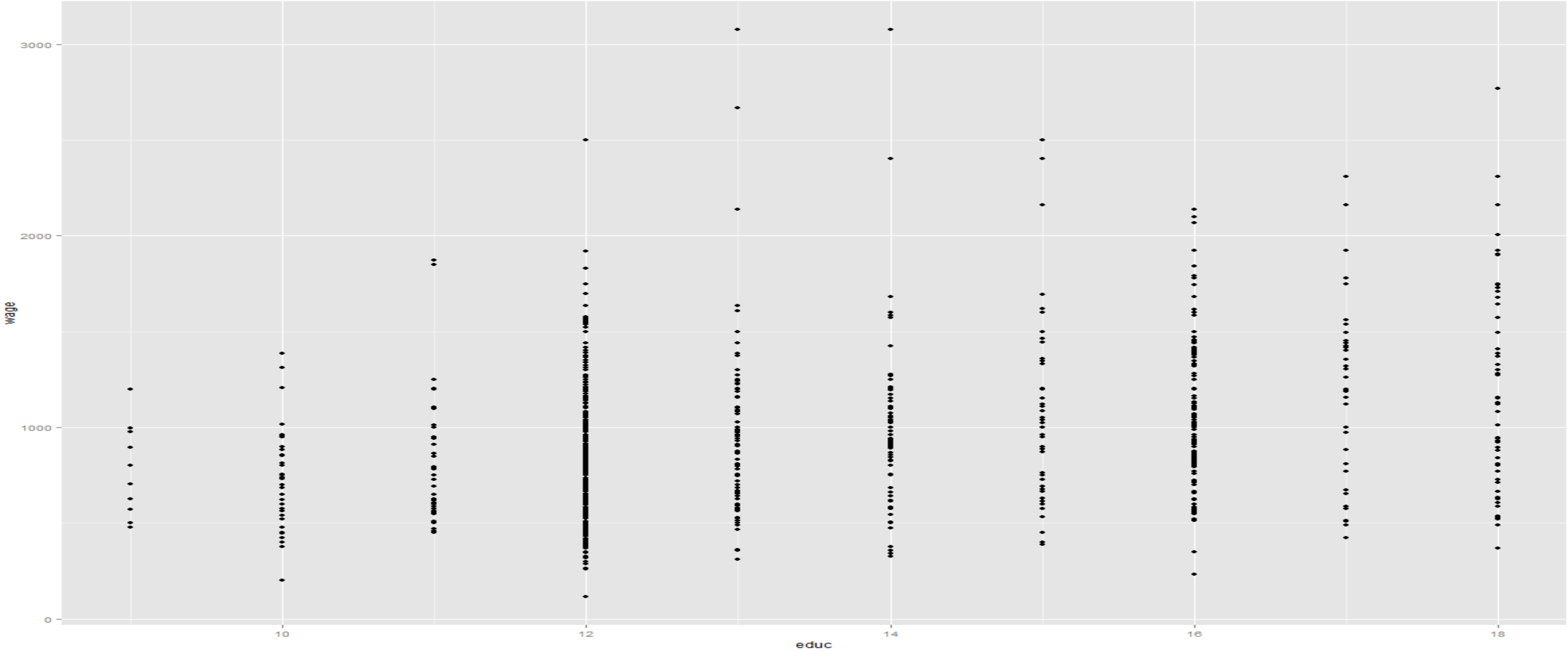
Detecting Heteroskedasticity cont..

- This should make intuitive sense; if some of our IVs are correlated with the squared-residuals, then the variance of the disturbance term is not independent of our observations (hence, in violation of our assumption).

- We can simply examine the F-statistic from the regression of our squared residuals on our predictors to test this assumption; if it is significant, we fail to meet the assumption of homogeneity.
  - **Significance here means that our predictor variables can explain**

Detecting Heteroskedasticity cont…

- Using the same dataset from above, we can run the model: wagei = β0 + β1educi + εi
- We know from previous discussions that education tends to hav



Testing for Heteroskedasticity

- Add a new variable to your dataset that is the square of the residuals from your regression model.
- Run a second regression of those squared residuals on education.
- $\varepsilon i2 = \beta 0 + \beta 1educi + ui$
- Check the overall F-test. If none of the independent variables (or in this case, just education) can explain the square of the residual then we have met the assumption of homoskedasticity.

Output from Breusch-Pagan Test

```
> m1 = lm(wage ~ educ, data=wage2)#run your intended model

> wage2$resid = resid(m1) #calculate the residuals for the model
> wage2$residsq = wage2$resid^2 #square the residuals
>
> m2 = lm(residsq ~ educ, data = wage2)#use those as the DV
> summary(m2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129206      65650  -1.968   0.0494 *
educ           20423       4811   4.245  2.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 323000 on 933 degrees of freedom
Multiple R-squared: 0.01895,   Adjusted R-squared: 0.0179
F-statistic: 18.02 on 1 and 933 DF,  p-value: 2.403e-05
```

- What do we conclude?

Dealing With Heteroskedasticity

- Once we have transformed variables and dealt with potential omitted variables, and still find heteroskedasticity in our data, how do we deal with it?

- Clearly, we need to be able to adjust our methods to allow us to draw correct inferences.

- Remember that the variances of our estimates are incorrect and not the point estimates themselves.

- **White's Heteroskedasticity-Consistent Variances**. Basically, since our estimates are unbiased when heteroskedasticity is present, we simply need to adjust our variance estimates to draw correct inferences.

- Using the OLS estimates along with "White Standard Errors" or more properly "heteroskedasticity consistent (HC) standard errors" has become standard practice among empirical researchers.

HC standard errors in R

- Correcting our standard errors in the presence of heteroskedasticity is straightforward in R.
- We simply run our intended regression and obtain the results.
- We know (i) that our coefficients are unbiased, but (ii) that our standard error estimates are not valid. We can correct for the standard errors after the fact.

- Here we see that the estimates of our standard errors are slightly larger, but not large enough to change the significance

- When we have heteroskedasticity we can:
  - Respecify the model
  - Use HC standard errors

Assumption: No Perfect Multicollinearity

- The regression assumption is actually that there is no perfect multicollinearity (sometimes referred to as singularity).
- In this section, we will deal with the issues that arise when multicollinearity exists.
  - As we have seen before R will not run models with perfect multicollinearity (it will just drop one of the violating variables).
  - Again, multicollinearity (i.e., something less than perfect multicollinearity) does not violate our regression assumptions, but its presence impacts our model.
- As Wooldridge states: "the problem of multicollinearity is not

Multicollinearity

- Though not well defined, all else being equal, it is better to estimate β_j when there is less correlation between x_j and the other independent variables.
- If high multicollinearity is present, often we can try to drop other independent variables from the model. Though as we saw with our discussion on omitted variable bias, dropping relevant variables leads to its own set of problems.

Why Multicollinearity is Problematic

- Mertler and Vannatta (2010) state:
  - Multicollinearity causes difficulty when attempting to determine the importance of individual IVs because the **individual effects are confounded due** to overlapping information.
  - Multicollinearity **increases the variances** of the regression coefficients; thus increasing the likelihood of a type II error (accepting the null when it should be rejected).

Multicollinearity, Tolerance, and the Variance Inflation Factor (VIF).

- We already know that one way to identify multicollinearity is to look at a simple correlation matrix. While effective, this method often misses some more subtle forms of multicollinearity.
  - Ex: If you used verbal GRE score, math GRE score, and total GRE score.
- An alternate method is to use tolerance and the VIF.
  - Tolerance is a measure of the collinearity among the predictor variables. Tolerance values range from 0 to 1, where values close to zero indicate multicollinearity.

Calculating VIF

- The calculation of VIF is straightforward. It is simply (1/(1-Rj2). Where Rj2 is the just r-squared from the regression model where predictor j is the dependent variable and the remaining predictors are the IVs.
- Hence, this model is estimating the amount of variance in one IV that can be predicted by the other IVs.
- No hard and fast rules for what value of VIF is cause for concern. However, Myers (1990) suggests that a value of 10 is the point at which to worry.
  - Menard (1995) suggests that values above 5 are worthy of concern

Obtaining Collinearity Diagnostics

- First of all, as we just saw, the diagnostic measures are intuitive and simple to calculate “by hand”.
- Second of all, R can produce these measures for us.
- Let's look at obtaining these values through both methods.

Collinearity Diagnostics cont…

- Assume we want to run the following model:

  wage = β0 + β1educi + β2experi + μi

- To obtain collinearity diagnostics we would run the following model and save the r-squared:

  educi = β0 + β2experi + μi

$$var(\widehat{\beta}_j) = \frac{\sigma^2}{N\, var(X_j)(1 - R_j^2)}$$

Collinearity Diagnostics cont…

- The output for collinearity diagnostics in R. Note, it is the same value that we calculated earlier.

```
> vif(lm2) #remember you want to use this with your original model
    educ    exper
1.261904 1.261904
```

As noted by Mertver and Vannata (2010) two ways to deal with multicollinearity are to drop the problem variable or to create a single new composite variable using the correlated IVs.

The square root of the VIF, is the factor by which the standard errors are inflated due to multicollinearity. Thus, a VIF of 4, means that the standard errors are double the size they would be if the predictors were all independent of one another.

If we had more than 2 predictors, each would have its own, different, VIF statistic.