Advanced Data Analysis I

Model Specification, Outliers, and Missing Data

PA 541 Week 10

Michael D. Siciliano

Department of Public Administration

College of Urban Planning and Public Affairs

Today’s Lecture

- Midterm
- Log models
- Outliers
- Missing Data (provide slides and code; but will not cover in class. Will not be on any exam or HW)
- Also going to hold off on model specification…will tackle in a later lecture.

Homework 2 Results

- Points possible: 72
- Average points: 55 (76%)
- Median points: 60 (83%)

Optional Lab

- Friday the 19th at 3pm
  - Will cover items from the midterm exam
  - Common interpretation issues
  - Recent lecture material

Remaining coursework

- **Week 11** – Spring Break
- **Week 12** – Logistic Regression
- **Week 13/14** – Panel Data
- **Week 15** – DAGs
- **Week 16** – Final Exam (similar format to midterm)

- **Homework 3 (April 12th):** Will distribute on March 29th. Will cover non-linear relationships, logistic regression, and first part of panel data (so weeks 10, 12, and 13.

- **Final Papers (May 5th)**

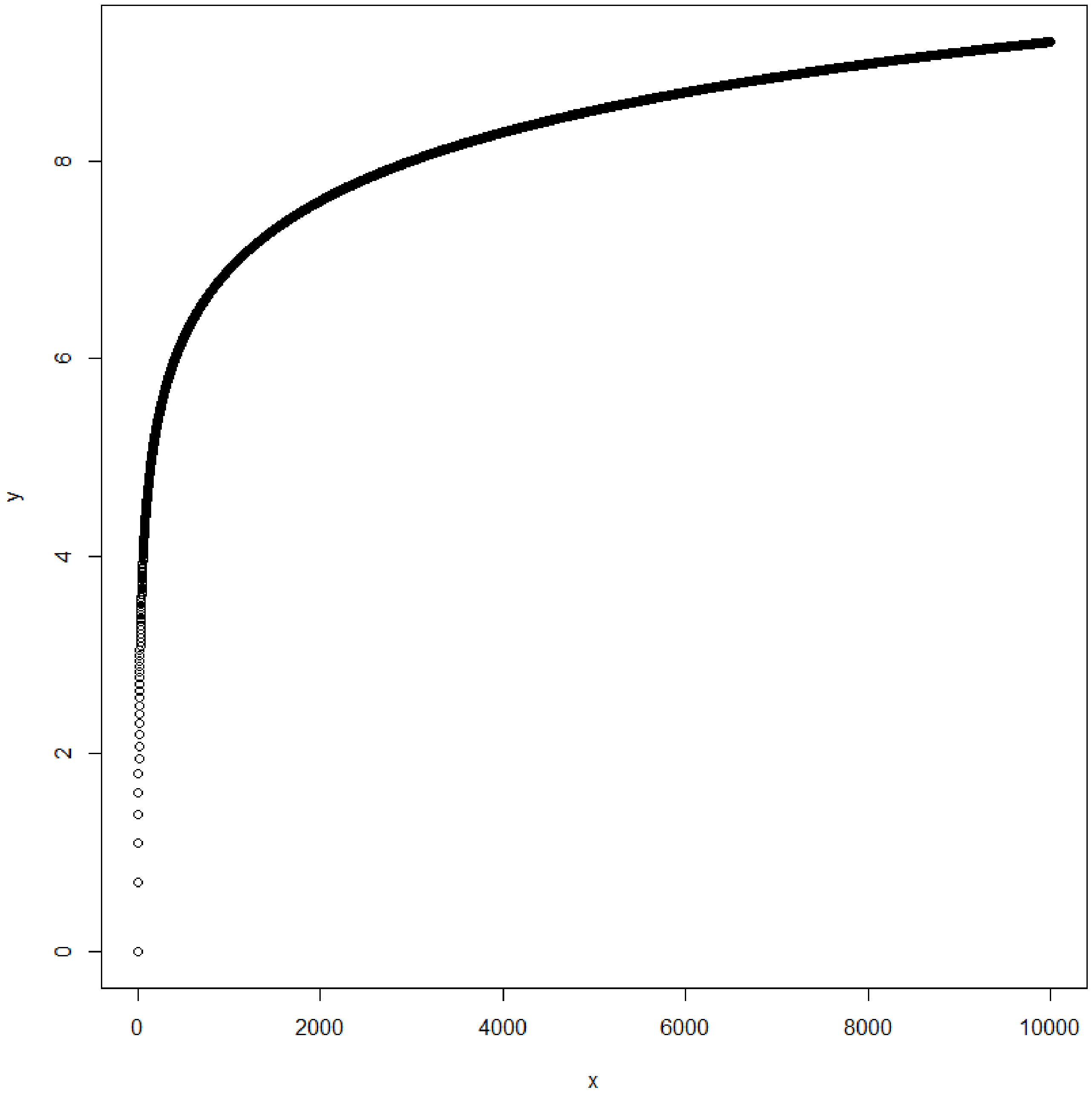
LOG MODELS

The Natural Logarithm

- The nonlinear function that plays the most important role in econometric analysis is the natural logarithm (Wooldridge, 2003 p. 685).
- In many instances econometrics texts will simply use Log and/or ln to denote natural log (I will also use them both to refer to the natural log).
- When y = log(x), the relationship between y and x indicates diminishing marginal returns.
    - However, this is different from the quadratic function we discussed earlier because when using a log function, the effect of x on y never becomes negative. In other words, the slope of the function asymptotically approaches zero as x gets larger, but it never quite reaches zero.

```
> x=seq(from =1, to = 10000, by=1)
> y=log(x)
> plot(x,y)
```

Note the log() function in R computes the natural logarithm by default.

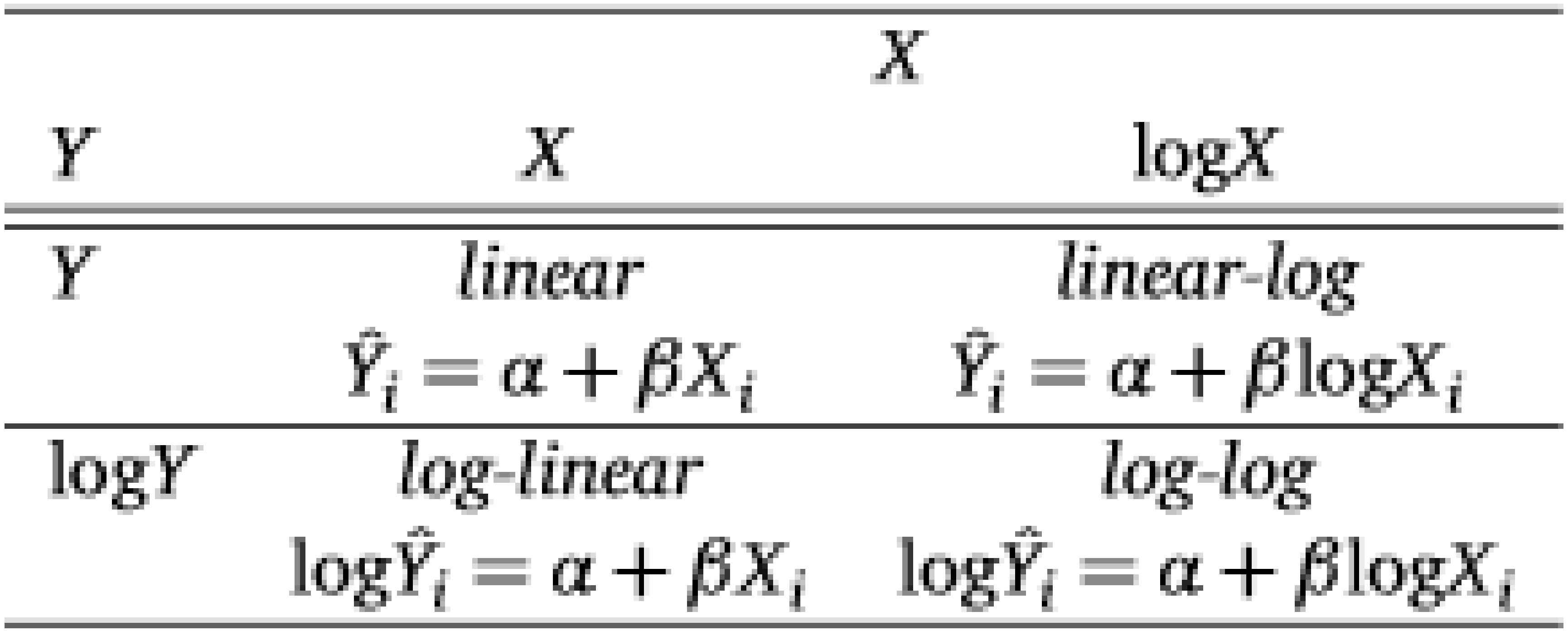


- We use logged variables not only for their ability to capture non-linear relationships, but because they allow us to interpret the effects in percentage terms.
  - I will demonstrate this below.

Two Types of Log Transformations

- As you have probably seen before, there are times when we want to conduct log transformations to make our data normally distributed. For example, we tend to log transform salaries as they tend to be significantly skewed to the right.
- Also, in many economic situations, theory may dictate a relationship between variables that is non-linear.
- Because of the frequency with which logs are used we will examine and interpret several types of log transformed models:
  - **log-log model** (both DV and IV are transformed)
  - **semi-log models** (log-lin and lin-log; either IV or DV are transformed but not both)

Log models: a typology



	X	
Y	X	$\log X$
Y	*linear* $\hat{Y}_i = \alpha + \beta X_i$	*linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$
$\log Y$	*log-linear* $\log \hat{Y}_i = \alpha + \beta X_i$	*log-log* $\log \hat{Y}_i = \alpha + \beta \log X_i$

Log-Log Model

- We can estimate the following generic linear regression model, known as the log-log model (remember that ln and log are used interchangeably to represent the natural log)
- $\log(y) = \beta_0 + \beta 1\log(x_{1i}) + \beta_2\log(x_{2i}) + \varepsilon_i$
- As with our quadratic model, it is linear in its parameters.
- One reason, for the use of log-log models especially among economists, is that **the regression coefficients can be interpreted as elasticities.**

Log-Log Model cont…

- From economics, recall that an elasticity is defined as:

$$\epsilon = \frac{Percentage\ Change\ in\ Y}{Percentage\ Change\ in\ X}$$

Log-Log Model cont…

- How do we define a percentage change?

- % change in y = {(New – Old) / Old } * 100

- For instance, if your favorite sports team had 20 wins last season and they have 25 wins this season you could easily calculate the percentage change in the number of wins as {(25-20)/20 }*100 = 25%.

Log-Log Models cont…

- If we allow the Δx to represent the change in x and define Δy as the change in y, then we can rewrite elasticity as:

$$\epsilon = \frac{Percentage\ Change\ in\ Y}{Percentage\ Change\ in\ X} \longrightarrow \epsilon = \frac{\left(\frac{\Delta y}{y}\right) * 100}{\left(\frac{\Delta x}{x}\right) * 100}$$

$$\varepsilon = \frac{\Delta y}{\Delta x} * \frac{x}{y}$$

Log-Log Models cont…

- How does all this relate to our log-log model?
- Using a bit of calculus, it can be shown that the Δln(x) = Δx/x. So, 100* Δln(x) = percentage change in x.
  - Ex. Log(51) – log (50) = 0.0198.
    - .0198*100 = 1.98%
  - Percentage change from 50 to 51 = (51-50)/50 = 0.02.
    - .02*100 = 2%
  - This is a very good approximation for small changes in X.
- If we run the following model

$$\text{Log}(y) = \beta_0 + \beta_1 \log(x_{1i}) + \mu_i$$

- Then the coefficient β_1 is defined as:

$$\beta_1 = \frac{Change\ in \ln y}{Change\ in \ln x} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x} * \frac{x}{y} = elasticity$$

Log-Log Models cont…

- Based on the previous slide, we can view the slope coefficients in log-log models as elasticities.
- Note, that the effect of a change in x on y is the same regardless of the size of x. Because of this, log-log models are often termed **constant elasticity models.**

- So, all of the equations on the previous slides simply show that when the DV is log transformed and the IV is log transformed we can view the relationship by saying a one percent change in x causes a β percentage change in y; as opposed to a one unit change in x causes a β unit change in y.

Log-Log Model Example

- Simple theory on the determinants of crime (as well as common sense) suggests that as enrollment in a university increases the number of crimes on campus also increases. We may estimate the simple regression model of: $\ln(\text{crime}_i) = \beta_0 + \beta_1 \ln(\text{enroll}_i) + \mu_i$
- Where crime is the number of criminal incidents on campus and enroll is the total number of students enrolled at the university.

Log-Log Model Example Output

```
> logm1 = lm(lcrime ~ lenroll, data=crime)
> summary(logm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -6.6279     1.0413  -6.365 7.08e-09 ***
lenroll       1.2693     0.1107  11.469  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8994 on 94 degrees of freedom
Multiple R-squared: 0.5832,	Adjusted R-squared: 0.5788 
F-statistic: 131.5 on 1 and 94 DF,  p-value: < 2.2e-16
```

- How do we interpret the coefficient on log(enroll)?
- We can then say that a 1% increase in enrollment leads to a 1.269% increase in crime.
- This is an interesting conclusion (more so than the commonsense fact that higher enrollment results in higher crime). Because the elasticity is greater than 1, thus in a relative sense, not just an absolute sense, crime is more of a problem on larger campuses.

Log-Log exact calculation of percentage change

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -6.6279     1.0413  -6.365 7.08e-09 ***
lenroll       1.2693     0.1107  11.469  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #Let's calculate the exact percentage change
> #what we need to do now is look at a 1% change in your level of
> #enrollment.  So let's say we are currently at 5000 students, a
> #1% increase would bring us to 5050 students
> a=-6.6279 + 1.2693*log(5000)
> b=-6.6279 + 1.2693*log(5000 + .01*5000)
> #Now both a and b are in log units, we need to exponentiate to get
> #them back to original units.
> e.a = exp(a) #= 65.56
> e.b = exp(b) # = 66.39
> (e.b-e.a)/e.a
[1] 0.01271005
> #If we multiply by 100 we get the percent increase in the outcome
> #variable (on the scale of the outcome variable).  So we say that a
> #1% change in enrollment leads to a 1.27% change in crime rate
```

Semi-log Models

- Semi-log models are models where either the dependent variable or the independent variable (but not both) have been log transformed.
- **Log-Lin model**: $\log(y_i) = \beta_0 + \beta_1 x_i + \mu_i$
- **Lin-Log model**: $y_i = \beta_0 + \beta_1 \log(x_i) + \mu_i$

Log-Lin Model

- Log-Lin model: $\log(y_i) = \beta_0 + \beta_1 x_i + \mu_i$
- How do we interpret the coefficient in the log-lin model?

$$\beta_1 = \frac{Change\ in \ln y}{Change\ in\ x} = \frac{\Delta y/y}{\Delta x} = \frac{\Delta y}{\Delta x} * \frac{1}{y}$$

- This interpretation is close to an elasticity, but not exactly. In the log-lin model the relationship between an elasticity and β_1 is:

$$\varepsilon = \frac{\Delta y}{\Delta x} * \frac{x}{y} = \left(\frac{\Delta y}{\Delta x} * \frac{1}{y}\right) x = \beta_1 * x$$

- Hence the elasticity depends on the value of x. In most instances, the elasticity is evaluated at the sample mean of x.

Log-Lin Model Example

- Let's return to our university crime dataset and run the following model (where enroll is used as the predictor as opposed to the log of enroll).
- $\mathbf{\ln(crime_i) = \beta_0 + \beta_1(enroll_i) + \mu_i}$
- From the regression output below we can see that a one person increase in enrollment increases the natural log of crime by .00008293. Or, equivalently, if university enrollment increases by 1,000 the increase in the natural log of crime is .08293. This relationship is difficult to understand. So, we turn to elasticities.

```
> logm2 = lm(lcrime ~ enroll, data=crime)
> summary(logm2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.940e+00  1.574e-01   25.04   <2e-16 ***
enroll      8.293e-05  7.797e-06   10.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9385 on 94 degrees of freedom
Multiple R-squared: 0.5462,    Adjusted R-squared: 0.5413 
F-statistic: 113.1 on 1 and 94 DF,  p-value: < 2.2e-16
```

Log-Lin Model Example cont…

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.940e+00  1.574e-01   25.04   <2e-16 ***
enroll      8.293e-05  7.797e-06   10.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We can determine the elasticity of crime with respect to enrollment based on:

$$\varepsilon = \frac{\Delta y}{\Delta x} * \frac{x}{y} = \left(\frac{\Delta y}{\Delta x} * \frac{1}{y}\right) x = \beta_1 * x$$

- We can calculate the elasticity at any value of X; again, we typically do this at the mean. If the sample mean of enroll is 16,016, then elasticity = β1 * x = .0000829*16,016 = 1.328. So, we can now say that, at the sample mean, a 1% increase in enrollment increases crime by 1.328%.

- Perhaps more usefully, when we have log-lin models we can say that the **$\%\Delta y = (100*\beta_1)\Delta x$.** Where we are measuring x in its actual units and not percentage change.
  - The origin of the above equation should be clear when we look at the original equation.
  - $\text{Log}(y) = \beta_0 + \beta_1 x$; so if we are looking at the effect of x, then: $\Delta\log(y) = \beta1\Delta x$.
  - So $100*\Delta\log(y) = (100*\beta1)\Delta x$.
  - Recall that $100*\ \Delta\ln(y)$ = percentage change in y. So, if we assume a 1 unit change in x, then $100*\beta1$ is the associated percentage change in Y.

- For our example, at any level of enrollment, if we increase enrollment by 1 student, we can expect a (100*.00008293) * 1 = .008293% increase in crime.

Log-Lin Model cont…

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.940e+00  1.574e-01   25.04   <2e-16 ***
enroll      8.293e-05  7.797e-06   10.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Generally, the quantity %Δy/Δx is called the semi-elasticity of y with respect to x. The semi-elasticity is the percentage change in y when x increases by one unit. Thus, the semi-elasticity is constant; the relationship between enroll and crime can be stated such that increasing enrollment by 1 person increases crime by .00829 percent.

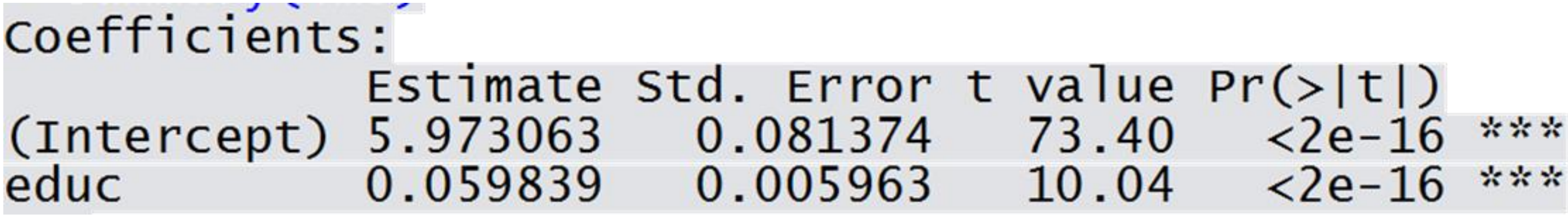
- Note that a .00829% change in crime causes a different absolute change in crime because it is based off the current level of crime. For example, keeping with the idea that a 1 person increase in enroll increases crime by .00829%, say we are at x = 1000 y = 20 and then we increase enroll by 1 person the crime rate only goes up by .0000829(20) or by .00165 crimes; but if we are at x = 20,000 and y = 123 and we increase enrollment by 1 person, crime goes up by .0000829(123) or .0102.

Another log-lin example

```
> lm3 = lm(lwage ~ educ, data = wage2) #NOTE DV is the Log of Wage
> summary(lm3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.973063   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4003 on 933 degrees of freedom
Multiple R-squared: 0.09742,   Adjusted R-squared: 0.09645
F-statistic: 100.7 on 1 and 933 DF,  p-value: < 2.2e-16
```

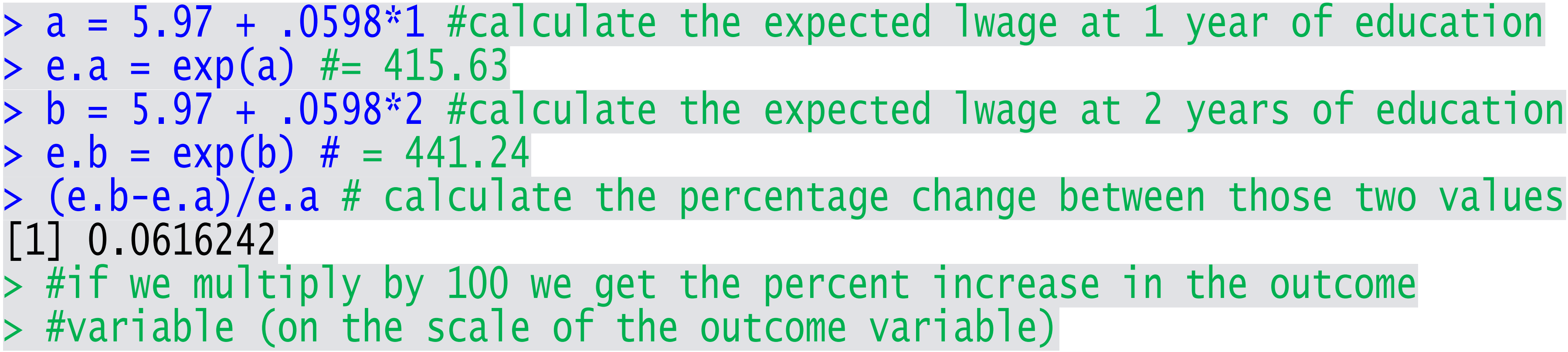
- How do we interpret the results?
  - Recall, the %Δy = (100*β1)Δx.
  - So, we can say that a one unit increase in the IV is associated with a (B1 * 100) percent increase in DV.
  - For our output then: we can say that for each additional year of education, wage goes up by 5.9%

Log-Linear model: Exact calculation of percentage change



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.973063   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
```

- Because this is only an approximation, if we want to calculate the actual change, we can always just use our regression equation and measure the difference in wages when education changes by 1 year. For example:



```
> a = 5.97 + .0598*1 #calculate the expected lwage at 1 year of education
> e.a = exp(a) #= 415.63
> b = 5.97 + .0598*2 #calculate the expected lwage at 2 years of education
> e.b = exp(b) # = 441.24
> (e.b-e.a)/e.a # calculate the percentage change between those two values
[1] 0.0616242
> #if we multiply by 100 we get the percent increase in the outcome
> #variable (on the scale of the outcome variable)
```

Lin-Log Model

- Lin-Log model: $y_i = \beta_0 + \beta_1 \log(x_i) + \mu_i$
- Let’s once again return to our university crime dataset and run the following model (where crime is used as the DV and not log(crime).
- **$\mathbf{crime_i = \beta_0 + \beta_1 \log(enroll_i) + \mu_i}$**
- From the regression output below we can see that a one unit increase in the log of enrollment increases crime by 400.382.

```
> logm3 = lm(crime ~ lenroll, data=crime)
> summary(logm3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3358.53     373.73  -8.987 2.62e-14 ***
lenroll       400.38      39.72  10.080  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 322.8 on 94 degrees of freedom
Multiple R-squared: 0.5194,    Adjusted R-squared: 0.5143
F-statistic: 101.6 on 1 and 94 DF,  p-value: < 2.2e-16
```

Lin-Log Model cont…

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3358.53     373.73  -8.987 2.62e-14 ***
lenroll       400.38      39.72  10.080  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- For lin-log models, we know that the $\Delta y = \beta_1 \Delta \log(x)$. This can be rewritten as $\Delta y = (\beta_1/100)[100*\Delta \log(x)]$.
  - Recall, 100* $\Delta \ln(x)$ = percentage change in x
- Thus, $\Delta y = (\beta1/100)(\%\Delta x)$. So, if we assume a 1 percent change in x, then $\beta1/100$ is the unit change in Y.

- In other words, $\beta_1/100$ is the unit change in y, when x increases by 1%. So, in our example, a one percent change in enrollment causes crime to increase by 4 units.

Lin-Log exact calculation of change

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3358.53     373.73  -8.987 2.62e-14 ***
lenroll       400.38      39.72  10.080  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #Let's calculate the exact percentage change
> #what we need to do now is look at a 1% change in your level of
> #enrollment.  So let's say we are currently at 5000 students, a
> #1% increase would bring us to 5050 students
> a = -3358.53 + 400.38*log(5000)
> b = -3358.53 + 400.38*log(5000 + .01*5000)
> #Now in this model, we interpret the percentage change in the predictor
> #on the unit change in crime.  Thus all we need to do is take the
following:
> b-a #and we get 3.98; approximately B1/100 unit change.
[1] 3.983913
```

Approximate Interpretations

- **Linear:** No transformations
  - DV = Intercept + B1 * IV + Error
  - "One unit increase in IV is associated with a (B1) unit increase in DV."
- **Log-Linear**: Outcome transformed
  - log(DV) = Intercept + B1 * IV + Error
  - "One unit increase in IV is associated with a (B1 * 100) percent increase in DV."
- **Linear-Log**: Predictor transformed
  - DV = Intercept + B1 * log(IV) + Error
  - "One percent increase in IV is associated with a (B1 / 100) unit increase in DV."
- **Log-Log**: Outcome transformed and Predictor transformed
  - log(DV) = Intercept + B1 * log(IV) + Error
  - "One percent increase in IV is associated with a (B1) percent increase in DV."

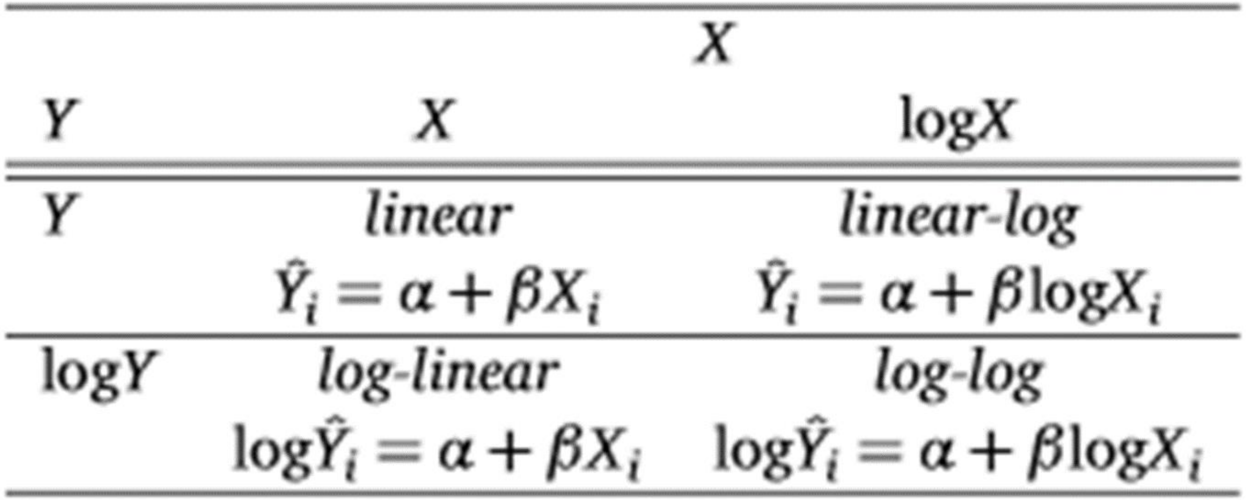
Again, you can get the exact interpretation by simply using the regression equation.

- In the end, I like Bailey’s (2019) advice:
  - While we can memorize the way units work in these various models, the safe course of action here is to simply accept that each time we use logged models, we’ll probably have to look up how units in logged models work…

But which model is best? Let’s Think Through an Example

- Consider the relationship between GDP (predictor) and Life Expectancy (outcome). What does each of the following models suggest about the relationship between the two variables and which ones seem to be reasonable relationships:

  - Lin – Lin model
  - Log – Lin model
  - Lin – Log model
  - Log – Log model

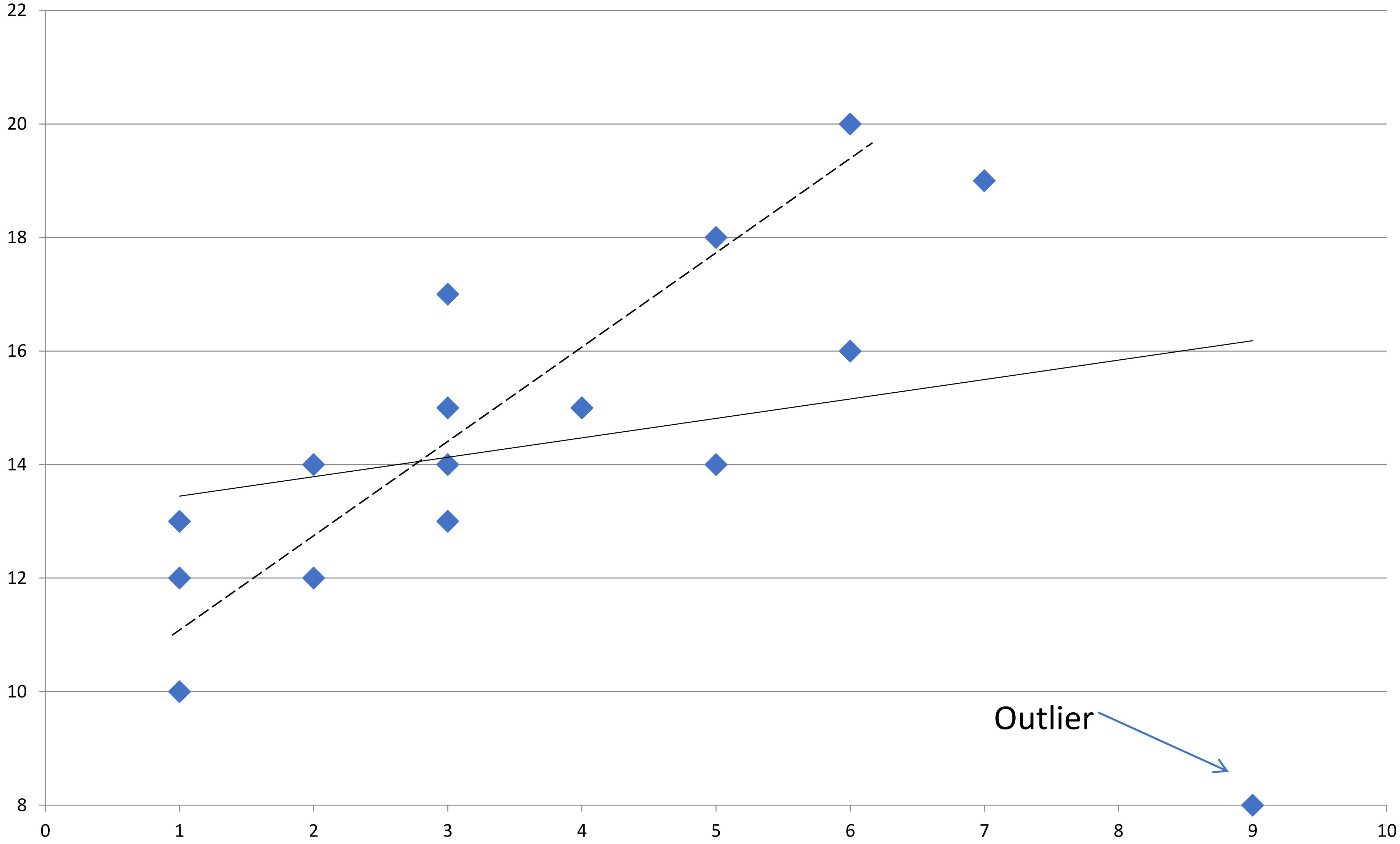


	X	
Y	X	$\log X$
Y	*linear* $\hat{Y}_i = \alpha + \beta X_i$	*linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$
$\log Y$	*log-linear* $\log \hat{Y}_i = \alpha + \beta X_i$	*log-log* $\log \hat{Y}_i = \alpha + \beta \log X_i$

OUTLIERS

Outliers

- A univariate outlier is an extreme value on one variable
- A multivariate outlier is a case with a strange combination of scores on two or more variables
  - For example, a 15-year-old boy is perfectly within bounds regarding age, and someone who earns $100,000 a year is in bounds regarding income, but a 15-year-old who earns $100,000 may be very unusual.
- Outliers can be found in IVs and DVs.
- Outliers lead to Type I and Type II errors.
- **The concern is that, especially in small samples, one weird observation can screw up the analysis.**


22
20
18
16
14
12
10
8
0
1
2
3
4
5
6
7
8
9
10
Outlier

Let’s look at an example

- Wooldridge discussing ‘RDCHEM’ data regarding R&D expenditures as a percentage of sales.

- Obs: 32

1. rd	R&D spending, millions $
2. sales	firm sales, millions $
3. profits	profits, millions $
4**. rdintens**	rd as percent of sales
5. profmarg	profits as percent of sales
6. salessq	sales^2
7. lsales	log(sales)
8. lrd	log(rd)

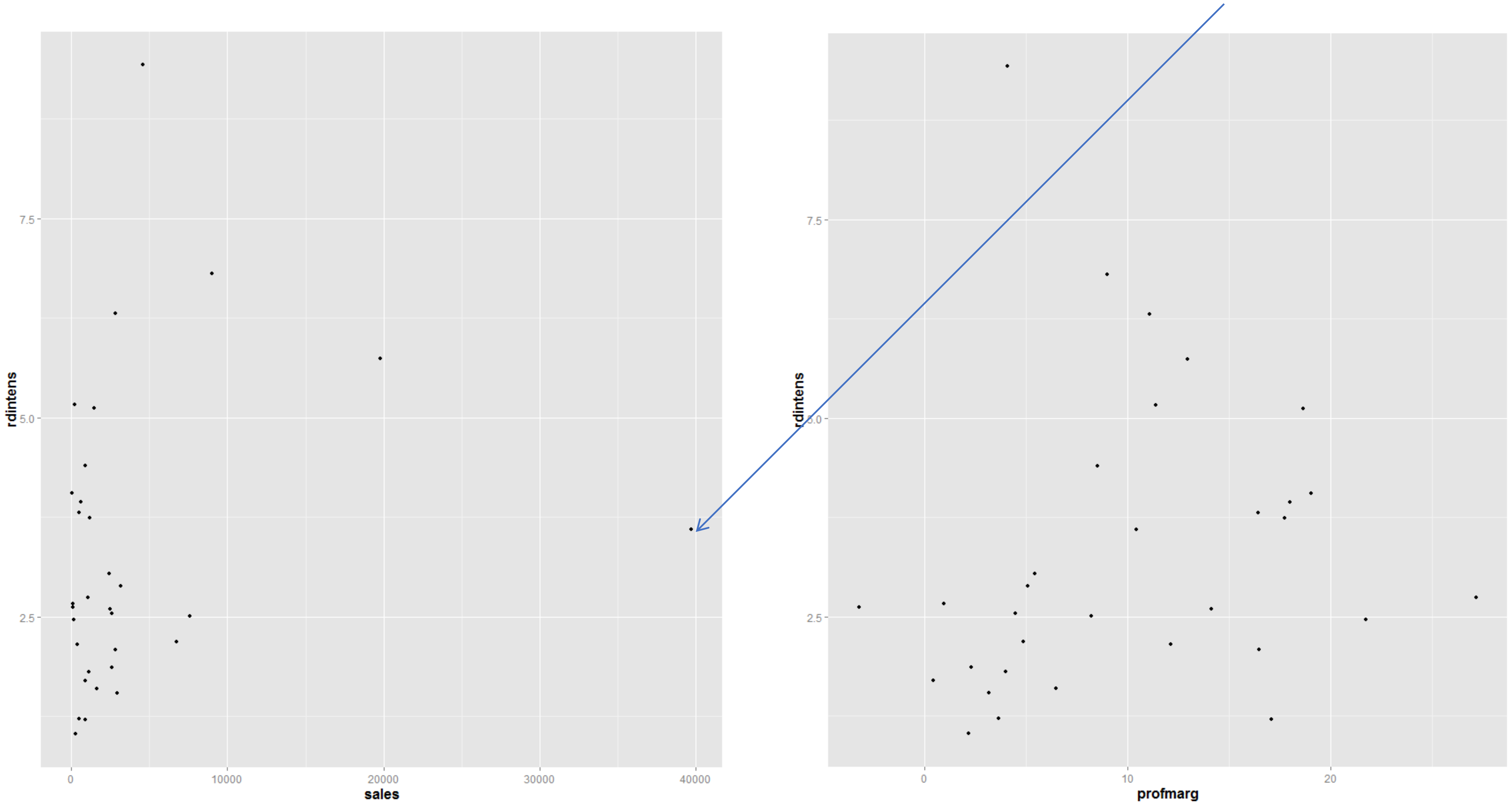
Wooldridge Example

- Let's assume we want to predict r&d intensity (defined as the total r&d expenditures as a percentage of sales) based on sales (in millions) and profits as a percentage of sales.

```
> out1 = lm(rdintens ~ sales + profmarg, data=chem)
> summary(out1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.625e+00  5.855e-01   4.484 0.000106 ***
sales       5.338e-05  4.407e-05   1.211 0.235638    
profmarg    4.462e-02  4.618e-02   0.966 0.341966    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.862 on 29 degrees of freedom
Multiple R-squared: 0.07612,	Adjusted R-squared: 0.0124 
F-statistic: 1.195 on 2 and 29 DF,  p-value: 0.3173
```

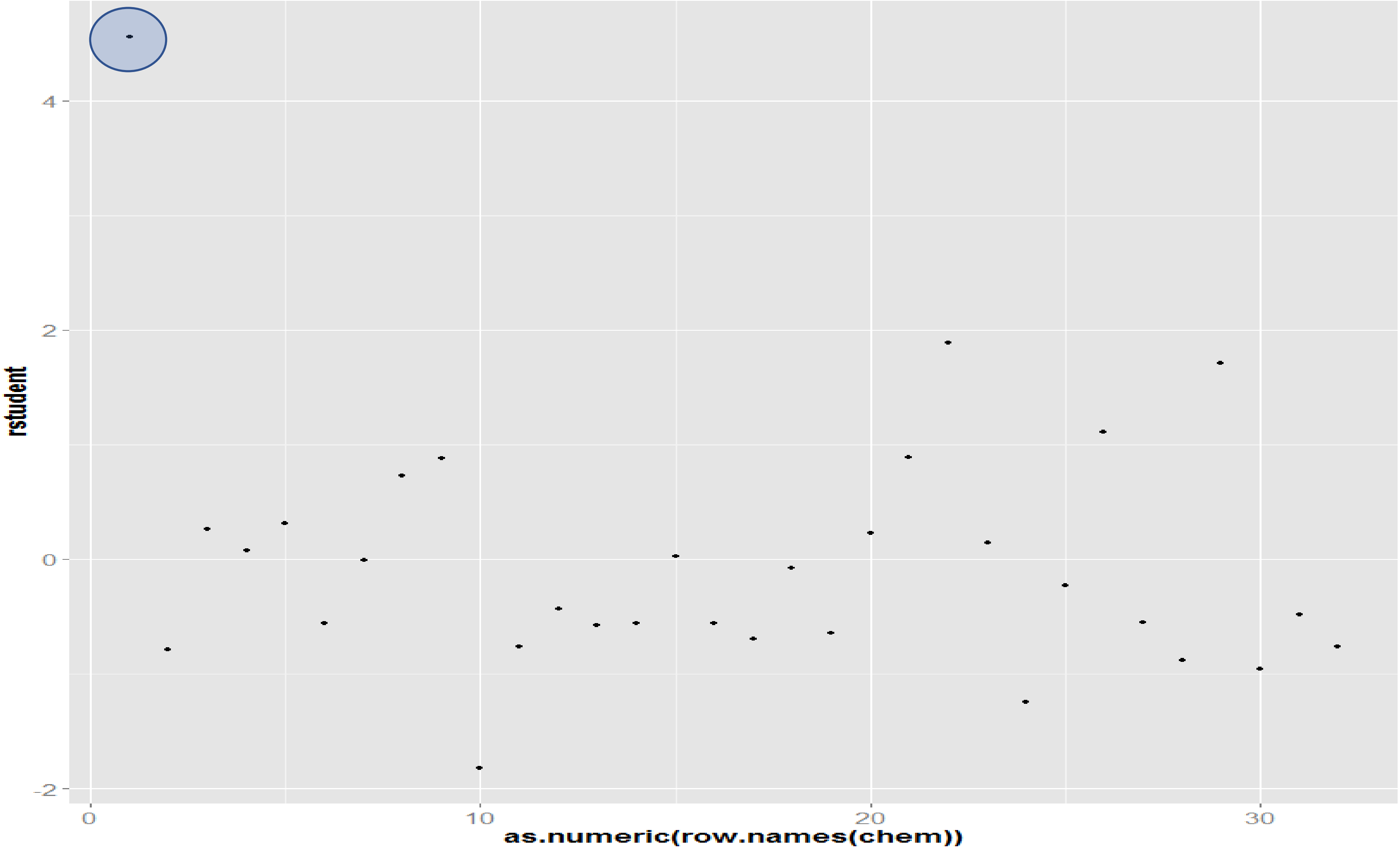
Note, our two predictors are not statistically significant even at the 10% level.

- First, look at some plots. Suspect a potential outlier based on sales.



Take look at the studentized residuals

```
chem$rstudent = rstudent(out1)
ggplot(data = chem, aes(x=as.numeric(row.names(chem)), y=rstudent)) + geom_point()
```



Rerun our model

```
> chem2 = chem
> chem2 = chem2[which(!chem2$rstudent>4),] #drop obs with large residual
> chem2 = chem2[which(!chem2$sales==max(chem2$sales)),] #drop obs with large 
sales value
>
> out3 = lm(rdintens ~ sales + profmarg, data=chem2)
> summary(out3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.939e+00  4.588e-01   4.226 0.000243 ***
sales       1.596e-04  6.457e-05   2.472 0.020029 *  
profmarg    7.007e-02  3.433e-02   2.041 0.051116 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.369 on 27 degrees of freedom
Multiple R-squared: 0.2711,	Adjusted R-squared: 0.2171 
F-statistic: 5.021 on 2 and 27 DF,  p-value: 0.014
```

- R-squared increases dramatically and we see sales and profmarg are now significant.
- Need to be careful though, we should not use outliers as a way to justify manipulating our data to get an intended result. Dealing with influential observations is a difficult endeavor.
- It should be noted that Wooldridge found the best fit, when a log transformation was taken on sales.   No observations were needed to be dropped.

Another outlier example…

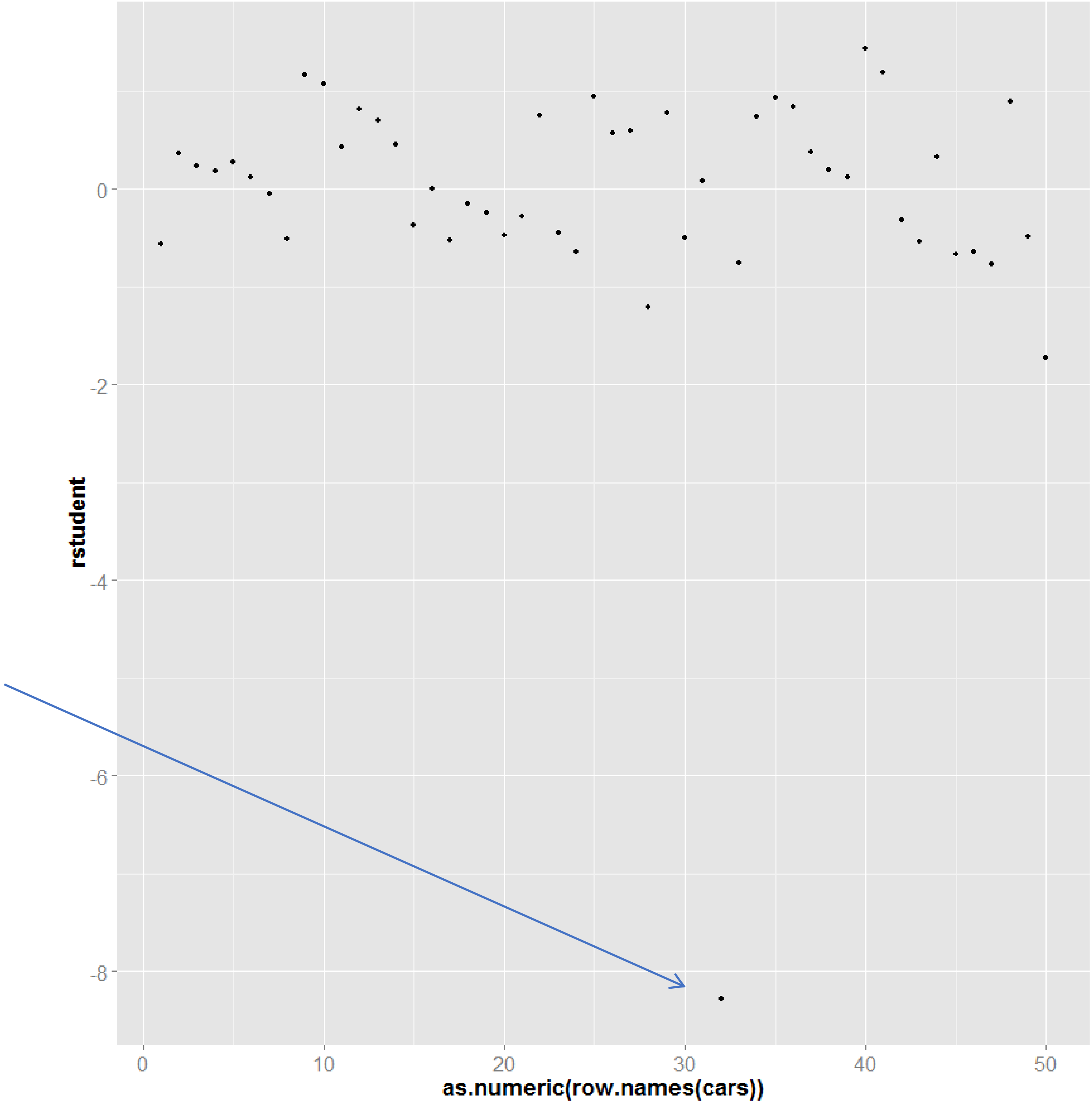
- We will use the ‘cars5’ dataset obtained from www.fueleconomy.gov.
- The dataset contains information for 50 new US passenger cars for the 2011 model year. Our goal is to explain Cgphm(City Gallons per 100 miles) using the following variables:
  - *Eng* – Engine size (liters)
  - *Cyl* – number of cylinders
  - *Vol* – interior passenger and cargo volume (hundreds of cubic feet)

Initial model results

```
> cm1 = lm(Cgphm  ~ Eng + Cyl + Vol, data = cars)
> summary(cm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.5420     0.3650   6.964 1.03e-08 ***
Eng           0.2350     0.1117   2.104 0.040886 *  
Cyl           0.2937     0.0762   3.855 0.000358 ***
Vol           0.4762     0.2786   1.709 0.094212 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3786 on 46 degrees of freedom
Multiple R-squared: 0.7584,	Adjusted R-squared: 0.7427 
F-statistic: 48.14 on 3 and 46 DF,  p-value: 3.114e-14
```

There is a clear outlier in the dataset. With an observed value that is much smaller than what our model would have predicted.

When we expect the dataset further, we find out that this vehicle is a hybrid. Thus, it is not surprising that is does not fit the pattern based on the other gasoline powered vehicles.



Updated results

```
> cars.b = cars #create a duplicate dataset
> cars.b = cars.b[which(!abs(cars.b$rstudent)==max(abs(cars.b$rstudent))),]
>
> cm2 = lm(Cgphm  ~ Eng + Cyl + Vol, data = cars.b)
> summary(cm2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.90476    0.23652  12.281 5.71e-16 ***
Eng          0.25943    0.07118   3.645 0.000691 ***
Cyl          0.22784    0.04917   4.634 3.08e-05 ***
Vol          0.49218    0.17743   2.774 0.008035 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2411 on 45 degrees of freedom
Multiple R-squared: 0.8566,    Adjusted R-squared: 0.847
F-statistic:  89.6 on 3 and 45 DF,  p-value: < 2.2e-16
```

Once we identify a potential outlier…

- Once we identify a potential outlier, we should investigate why the potential outlier has such an unusual response value on Y relative to their predicted values from the model.
- Some possible reasons:
  - Data have been incorrectly recorded; data entry error
  - A key predictor has been omitted from the model
  - One or more of the regression assumptions have been violated (try some transformations)
  - Misspecification of missing value codes
  - Outlier is not a member of intended population
  - Outlier is from intended population, but its distribution has more extreme values than the normal distribution

MISSING DATA

Missing Data

- One of the most pervasive problems in data analysis
- The seriousness of missing data depends on the:
  - Pattern of missing data
  - How much data is missing
  - Why the data is missing

Missing Data cont…

- Missing data are characterized by:
    - MCAR – missing completely at random
    - MAR – missing at random (though I don’t like this terminology)
    - MNAR – missing not at random

Missing completely at random

- Data are deemed missing completely at random when the probability of an observation, Xi, is missing is unrelated to the value of Xi or to the value of any other variable.
- Let’s take a question regarding family income:
  - Missing data on family income would not be considered MCAR if people with low incomes were less likely to report their family income than people with high incomes.
  - Similarly, if whites were more likely to omit reporting income than African Americans, the missing data would not be MCAR as missingness is correlated with race.

- MCAR means that any piece of data is just as likely to be missing as any other piece of data.
- When data are MCAR then analysis is unbiased.

Missing at random

- Data are missing at random if the probability of missing data on a variable is not a function of its own value after controlling for other variables in the design.
- Allison (2002) gives the following example:
  - Unmarried couples are less likely to report their income than married ones. Unmarried couples probably have lower incomes than married ones, and it would at first appear that missingness on income is related to income itself. But the data would still be MAR if the conditional probability of missingness were unrelated to the value of income within each marital category.
  - For MAR, the real question is whether the value of the variable determines the probability that it will be reported, or whether there is another variable, X, where the probability of missingness on Y is conditional on the levels of X.
    - To put it more formally, data are MAR if:
    - $p(Y \text{ missing} | Y,X) = p(Y \text{ missing} | x)$

Missing not at random

- If data are neither MCAR or MAR, then they can be classified as ‘missing not at random’ (MNAR).
  - Ex. People with low incomes are less likely to report their incomes.
- When data are MNAR we say that the mechanism controlling the missingness is nonignorable.
  - This means that we cannot sensibly solve whatever model we have unless we are also able to write a model that governs the missingness.
  - This is extremely difficult to do.

Identifying the mechanism (Baraldi and Enders 2010)

- Of the three missing data mechanisms, it is only possible to empirically test the MCAR mechanism.
  - Methodologists have proposed a number of tests, though they tend to have low power and do a poor job detecting deviations from a purely random process (Thoemmes and Enders 2007)
  - See Little's MCAR test.
- MAR and MNAR mechanisms are impossible to verify because they depend on unobserved data.
  - Demonstrating a relationship, or lack thereof, between the probability of missingness and the would be values of the incomplete variable requires knowledge of the missing values.
  - Thus, the MAR assumption that underlies the more advanced methods we will discuss is an untreatable assumption.
- Note, that the missing data mechanisms are not characteristics of the entire dataset, but they are assumptions that apply to specific analyses.

An intuitive MCAR test (Allison 2001)

- Create a dummy variable to indicate missingness on a particular variable. Regress the other variables in the data on the dummy indicator.
- If there are, in fact, no systematic differences on the fully observed variables between those with data present and those with missing data, then we may say that the data are *observed at random.*
- *On the other hand, just because the data pass this test* doesn't mean that the MCAR assumption is satisfied.
  - There must still be no relationship between missingness on a particular variable and the values of that variable.

Approaches to correct for missing data

- Listwise and pairwise deletion
- Mean/median substitution
- Regression substitution
- Expectation maximization
- Multiple imputation

- If only a few data points are missing (less and 5%) and the pattern is random then missing values do not pose a serious problem to your analysis.
- No firm guidelines for how much missing data can be tolerated for a particular sample size.

Deletion Methods

- In the presence of missing data, most statistical packages use **listwise** deletion which removes any row that contains a missing value from the analysis.
  - By default, R engages in Listwise deletion.
- **Pairwise** deletion, is another alternative that many have labeled as '*unwise deletion*'.
  - Allison (2001) states, the idea of pairwise deletion is to compute each of the necessary summary statistics using all the cases that are available. For example, to compute the covariance between two variables X and Z, we use all the cases that have data present for both X and Z. Once the summary measures have been computed, these can be used to calculate the parameters of interest, for example, regression coefficients.
  - The big problem with pairwise deletion is that the estimated standard errors and test statistics produced by conventional software are biased.

Example

- Using data from Cohen et al. (2003) we will look at predicted the salary of professors based on a number of variables. Only one variable in this dataset, CITM, has missing data.

```
➢cite = read.table("Cohen_ch11_data.txt", header=T, na.strings="NA")
➢head(cite)
```

```
  TIME  PUB  SEX  SALARY  CIT1  CITM
1    3   18    0   51876    13    50
2    6    3    0   54511     0    26
3    3    2    0   53425    10    50
4    8   17    1   61863    12    34
5    9   11    0   52926     9    41
6    6    6    1   47034    22    37
```

A Quick look at missing data

```
> countNAs <- function(x) {
+   sum(is.na(x))
+ }
>
> missing=list()
> for (i in 1:length(colnames(cite)))
+ missing[[i]] = countNAs(cite[,i])
> names(missing)=colnames(cite)
> missing
$TIME
[1] 0

$PUB
[1] 0

$SEX
[1] 0

$SALARY
[1] 0

$CIT1
[1] 0

$CITM
[1] 7
```

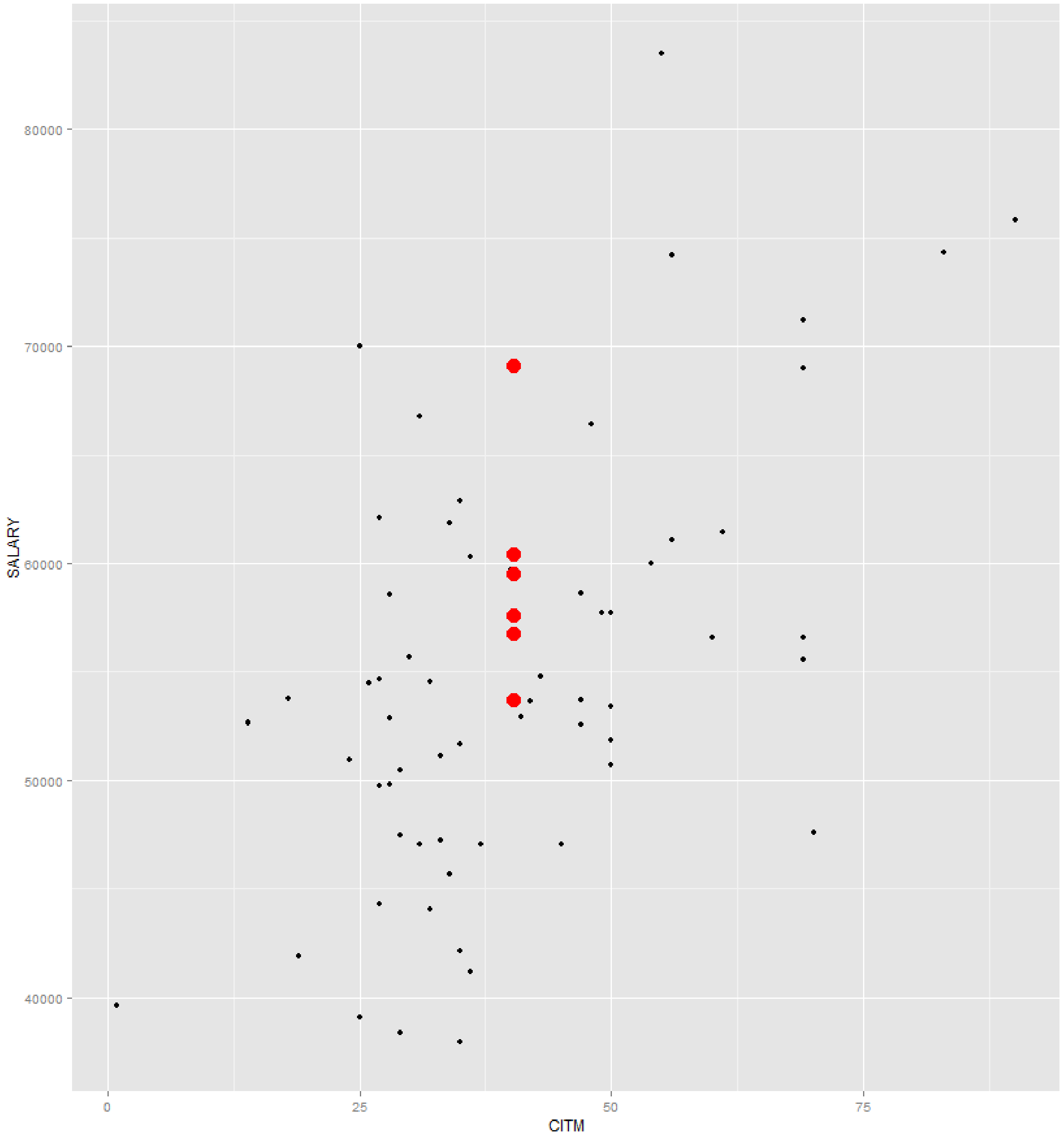
Regression results with listwise deletion

```
> lm1 = lm(SALARY ~ PUB + CITM, data=cite)
> summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40492.97    2505.39  16.162  < 2e-16 ***
PUB           251.75      72.92   3.452  0.00103 **
CITM          242.30      59.47   4.074  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7519 on 59 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared: 0.4195,   Adjusted R-squared: 0.3998
F-statistic: 21.32 on 2 and 59 DF,  p-value: 1.076e-07
```

Mean Substitution

- This is an old procedure that is still in use today despite its known biases and limitations.
- With this approach you simply replace any missing values for a variable with the observed mean (or median) for that variable.
- There are several problems with this approach:
  - It adds no new information as the overall mean with or without the missing data is identical.
  - However, the standard error of a regression coefficient will be smaller.

Scatterplot with updated data based on mean substitution



Regression results with mean substitution:

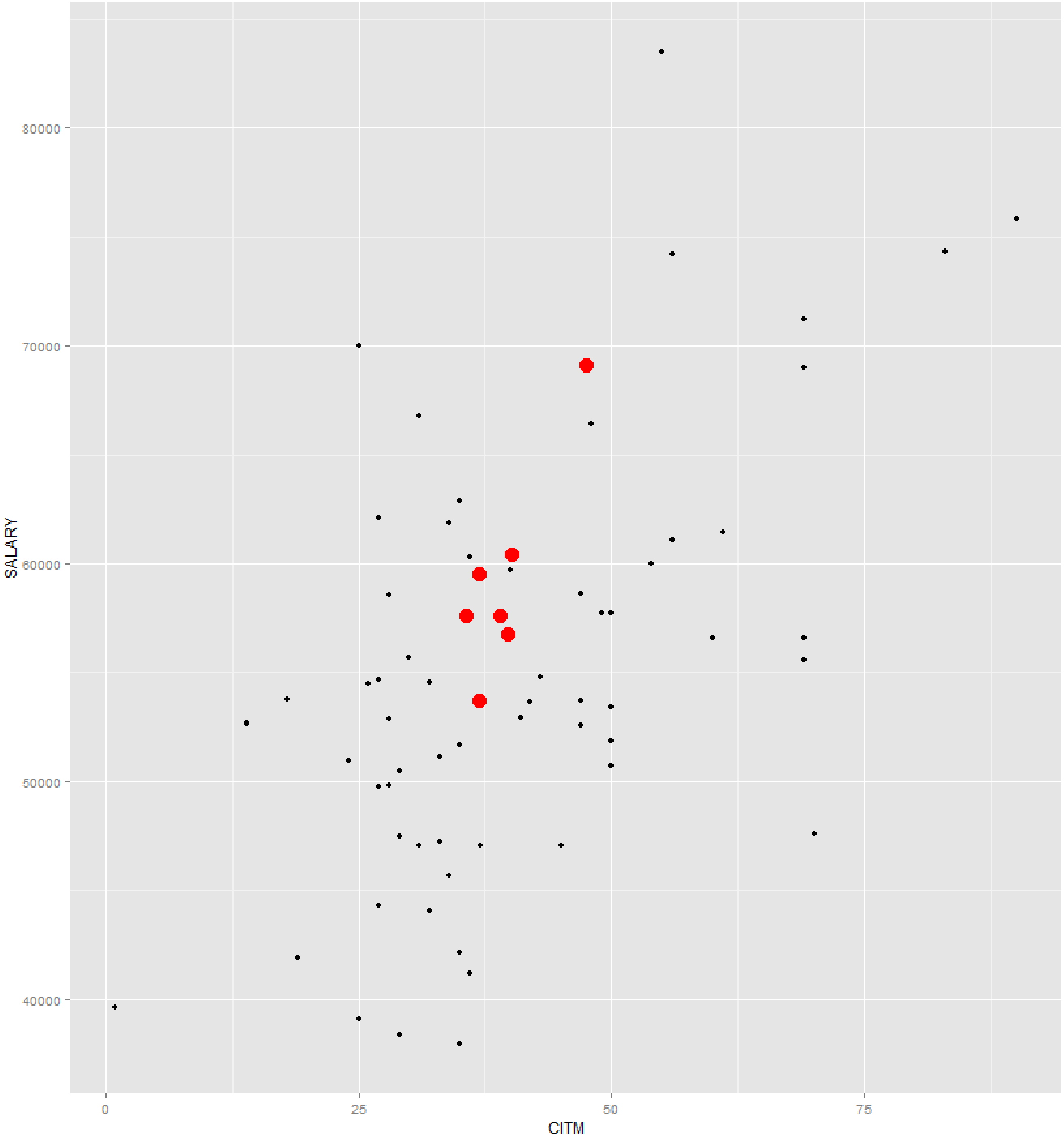
```
> lm2 = lm(SALARY ~ PUB + CITM, data=cite.mean)
> summary(lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40969.18    2417.34  16.948  < 2e-16 ***
PUB           255.46      69.17   3.693 0.000452 ***
CITM          241.29      57.78   4.176 8.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7326 on 66 degrees of freedom
Multiple R-squared: 0.4107,    Adjusted R-squared: 0.3928 
F-statistic:    23 on 2 and 66 DF,  p-value: 2.638e-08
```

Note: we see a drop in the standard error for both predictors.

Regression Substitution

- Another simple, though still problematic approach, is to use information on the other variables to predict the value of missing observations.
  - Suppose we are estimating a multiple regression model with several independent variables. One of those variables, X, has missing data for some of the cases.
  - For those cases with complete data, we regress X on all the other variables. Using the estimated equation, we generate predicted values for the cases with missing data on X.
  - These are substituted for the missing data, and the analysis proceeds as if there were no missing data.
- This method increases the correlation among the IVs (as some items have been calculated as a linear function of the others) and thus the regression coefficients are affected.
- Furthermore, this process likely underestimates the standard error as well by underestimating the variance of the imputed variable.

Scatterplot with updated data based on regression substitution



Results with regression substitution:

```
> #now we want to build a model to predict the variable with missing data
> lm3 = lm(CITM ~ PUB , data=cite)

> newdata=cite[is.na(cite$CITM),] #create a dataset that has just
observations with missing data

> pred.miss = predict(lm3, newdata=newdata)

> cite.reg = cite #create another duplicate dataset

> cite.reg$CITM = replace(cite.reg$CITM, is.na(cite.reg$CITM), pred.miss)

> lm4 = lm(SALARY ~ PUB + CITM , data=cite.reg)
> summary(lm4)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 41039.10    2401.36  17.090  < 2e-16 ***
PUB           250.45      69.54   3.602 0.000607 ***
CITM          242.30      57.92   4.183 8.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7324 on 66 degrees of freedom
Multiple R-squared: 0.4111,	Adjusted R-squared: 0.3932 
F-statistic: 23.04 on 2 and 66 DF,  p-value: 2.58e-08
```

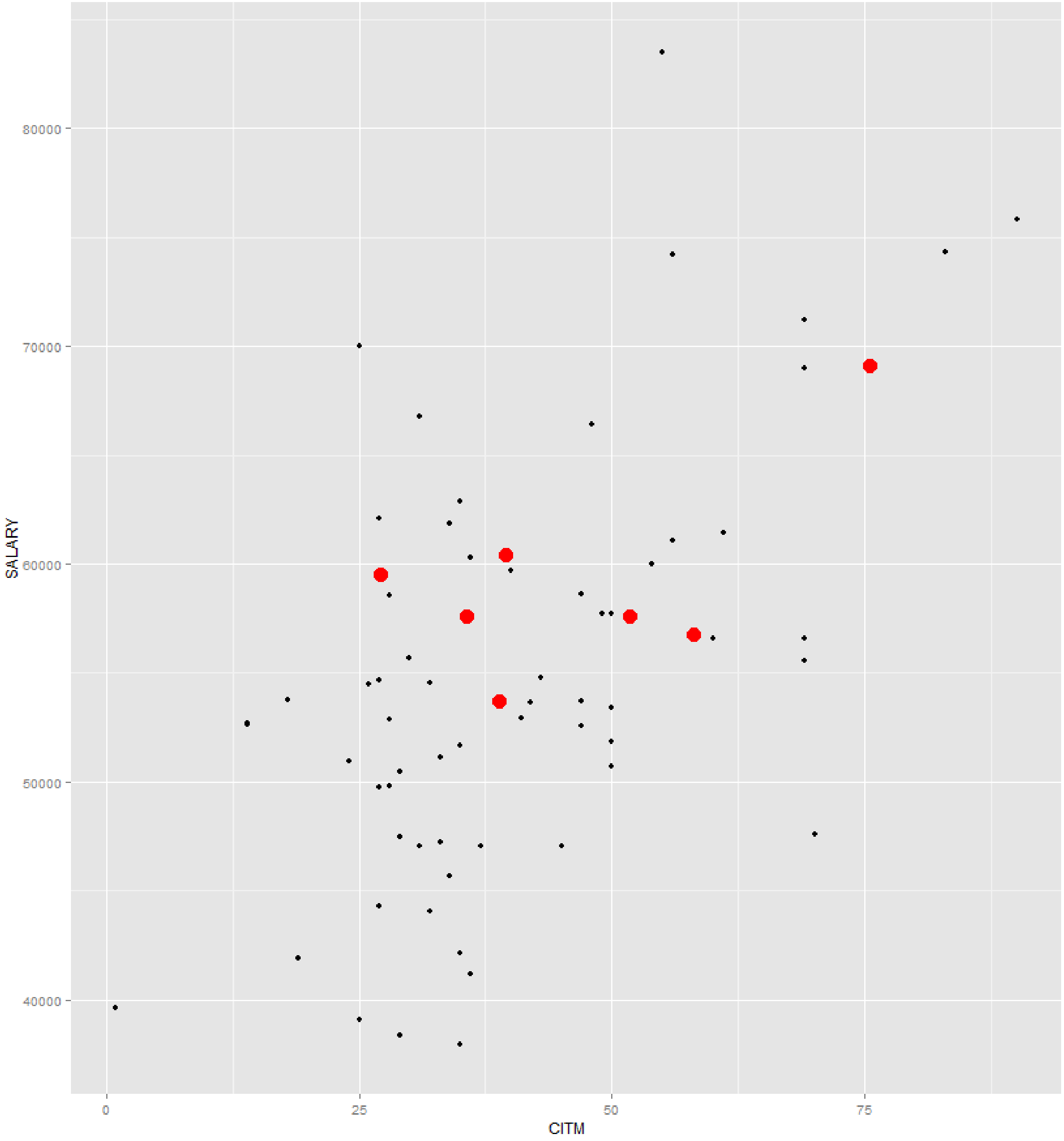
EM approach to missing data

- EM is an algorithm that obtains maximum likelihood estimators for parameters in statistical models.
    - We will discuss maximum likelihood when we cover logistic regression. For now I will simply cover the basic idea of this missing data method.
- Schafer (1999) phrased the problem well when he noted "If we knew the missing values, then estimating the model parameters would be straightforward. Similarly, if we knew the parameters of the data model, then it would be possible to obtained unbiased predictions for the missing values."
- Here we are going to do both.

EM cont…(Allison 2001)

- It’s called EM because it consists of two steps: an *Expectation step and a Maximization step. These two steps* are repeated multiple times in an iterative process that eventually converges to the ML estimates.
- The E-step essentially reduces to regression imputation of the missing values.
- Let’s suppose our data set contains four variables, *X1 through X4, and there is some missing data on each variable, in no particular pattern.*
  - EM begins by choosing staring values for the unknown parameters, that is the means and covariance matrix using standard formulas.
  - Based on the starting values of the parameters, we compute the coefficients for the regression of any one of the X’s on any subset of the other three. We then use the regression coefficients to generate the imputed values.
  - After the missing data have been imputed, the M-step consists of calculating new values for the means and covariance matrix, using the imputed data along with the nonmissing data.
    - Here, means are based on the usual formula, but the variances and covariances are modified based on the residuals of the regression equations used in the imputation process.
  - Once we have new estimates for the mean and covariances, we start over with the E-step. That is, we use the new estimates to produce new regression imputations for the missing values.
  - E-steps and M-steps are cycled through until the estimate converges, meaning there is very little change in the estimates from one iteration to the next.

Scatterplot with updated data based on EM



Results with EM approach using the ‘norm’ package

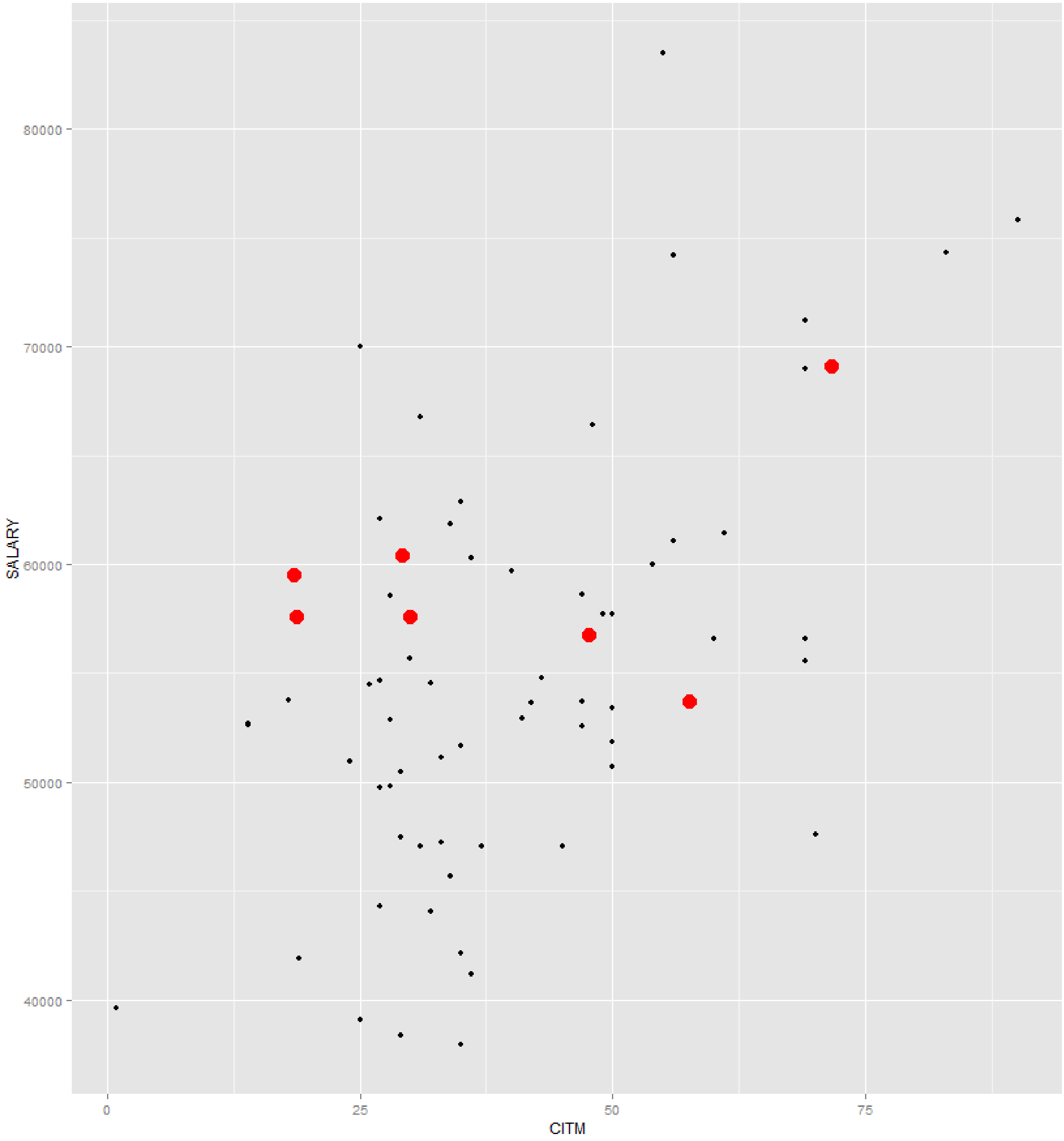
```
> mcite = as.matrix(cite) #need a data matrix not a dataframe for the
functions in norm
>
pcite = prelim.norm(mcite)
>
> emcite = em.norm(pcite)
Iterations of EM:
1...2...3...4...5...6...7...
>
> rngseed(1827) #need to set a random seed for this to work
> cite.em = imp.norm(pcite, emcite, mcite)
> #Performs maximum-likelihood estimation on the matrix of incomplete
> #data using the EM algorithm.
>
> cite.em = as.data.frame(cite.em)
>
> lm5 = lm(SALARY ~ PUB + CITM, data=cite.em)
> summary(lm5)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 41152.13    2331.99  17.647  < 2e-16 ***
PUB           245.17      69.34   3.536 0.000749 ***
CITM          237.51      55.14   4.307 5.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7278 on 66 degrees of freedom
Multiple R-squared: 0.4184,    Adjusted R-squared: 0.4008 
F-statistic: 23.74 on 2 and 66 DF,  p-value: 1.704e-08
```

Multiple Imputation

- Multiple refers to the fact that we impute multiple complete datasets and run our analysis on each one. We then combine the results across the multiple analyses (Rubin 1987).
- This approach was not used for many years due to a lack of good algorithms to carry it out and lack of software.
- Simulation methods known as Marcov Chain Monte Carlo (MCMC) have simplified the task considerable.
- The ‘Amelia’ package in R provides simple approach for conducting these methods.

Scatterplot with updated data based on multiple imputation

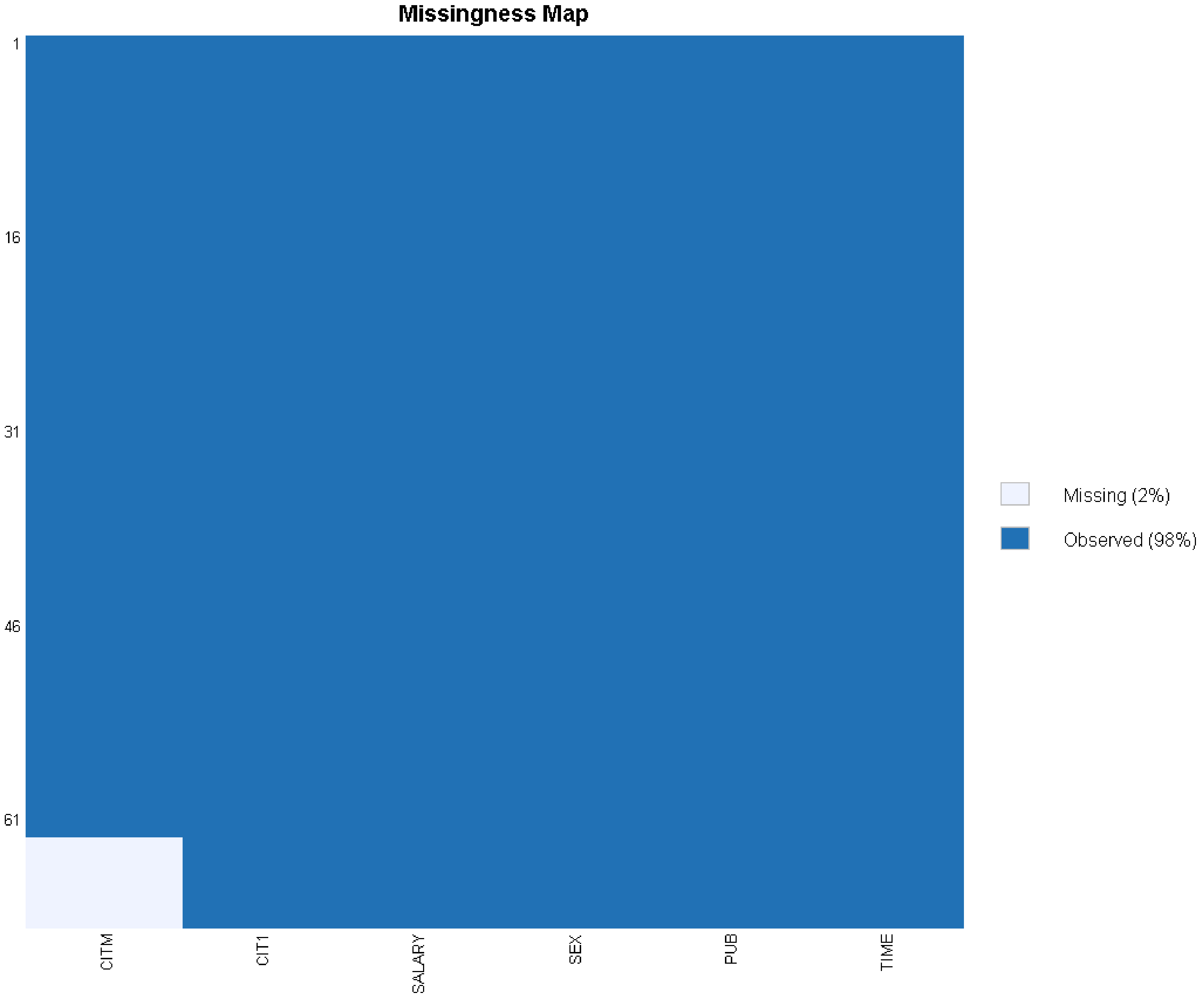
- Note, this is just one of several imputed datasets used to estimate the model.



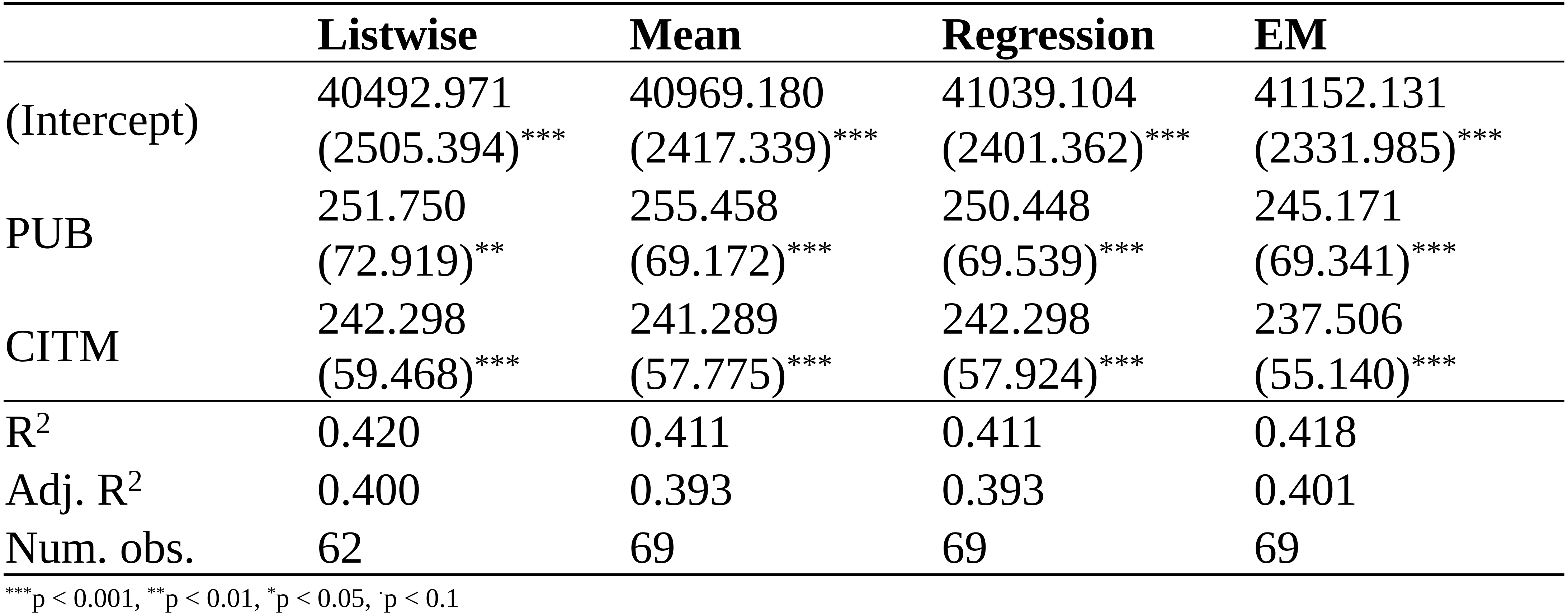
Results with multiple imputation using the ‘Amelia’ package

```
> a.out = amelia(cite, m=20)#here we impute 20 new datasets

> b.out = NULL
> se.out = NULL
> for (i in 1:a.out$m) {
+   ols.out = lm(SALARY ~ CITM + PUB, data=a.out$imputations[[i]])
+   b.out = rbind(b.out, ols.out$coef)
+   se.out = rbind(se.out, coef(summary(ols.out))[,2])
+ }
>
> combined.results = mi.meld(q=b.out, se = se.out)
> combined.results
$q.mi
     (Intercept)     CITM      PUB
[1,]    41588.79 228.5115 248.1051

$se.mi
     (Intercept)     CITM      PUB
[1,]     2378.24 56.95861 70.44416
```

- One cool function in Amelia: missmap.
- Here it is no that interesting since we only have one variable with missing data.
- This is a flexible function and you can also compare missing patterns by group and/or time periods.



Comparing Results



	Listwise	**Mean**	**Regression**	**EM**
(Intercept)	40492.971 (2505.394)***	40969.180 (2417.339)***	41039.104 (2401.362)***	41152.131 (2331.985)***
PUB	251.750 (72.919)**	255.458 (69.172)***	250.448 (69.539)***	245.171 (69.341)***
CITM	242.298 (59.468)***	241.289 (57.775)***	242.298 (57.924)***	237.506 (55.140)***
R^2	0.420	0.411	0.411	0.418
Adj. R^2	0.400	0.393	0.393	0.401
Num. obs.	62	69	69	69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

- Note, I didn’t list multiple imputation as it is based on 20 different datasets and hence, 20 different regression models.