PS0700
Basic Statistical Methods: Introduction to Regression Analysis

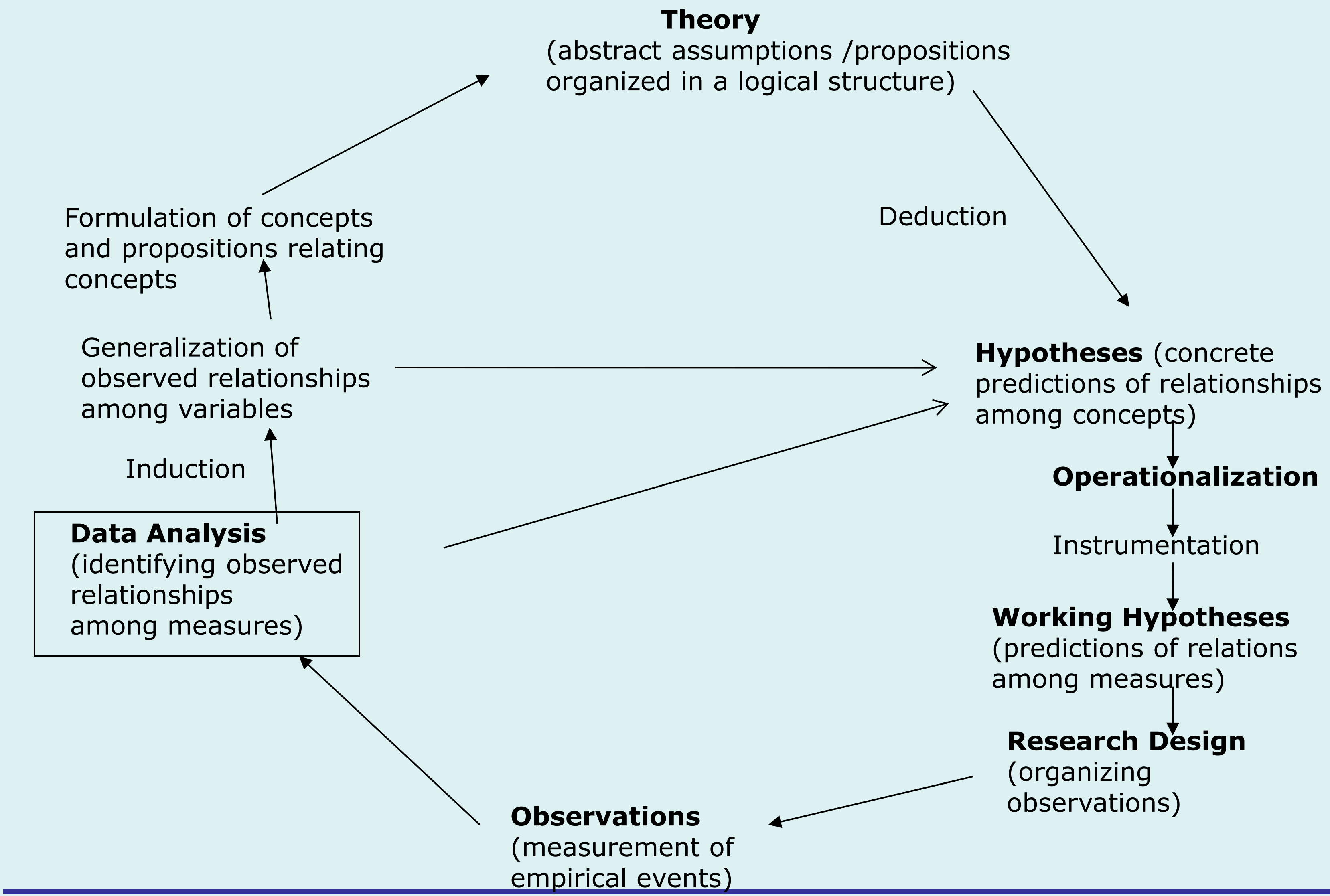
Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

Week 13

A Model of the Research Process



Theory
(abstract assumptions /propositions
organized in a logical structure)

Deduction

Hypotheses (concrete
predictions of relationships
among concepts)

Operationalization

Instrumentation

Working Hypotheses
(predictions of relations
among measures)

Research Design
(organizing
observations)

Observations
(measurement of
empirical events)

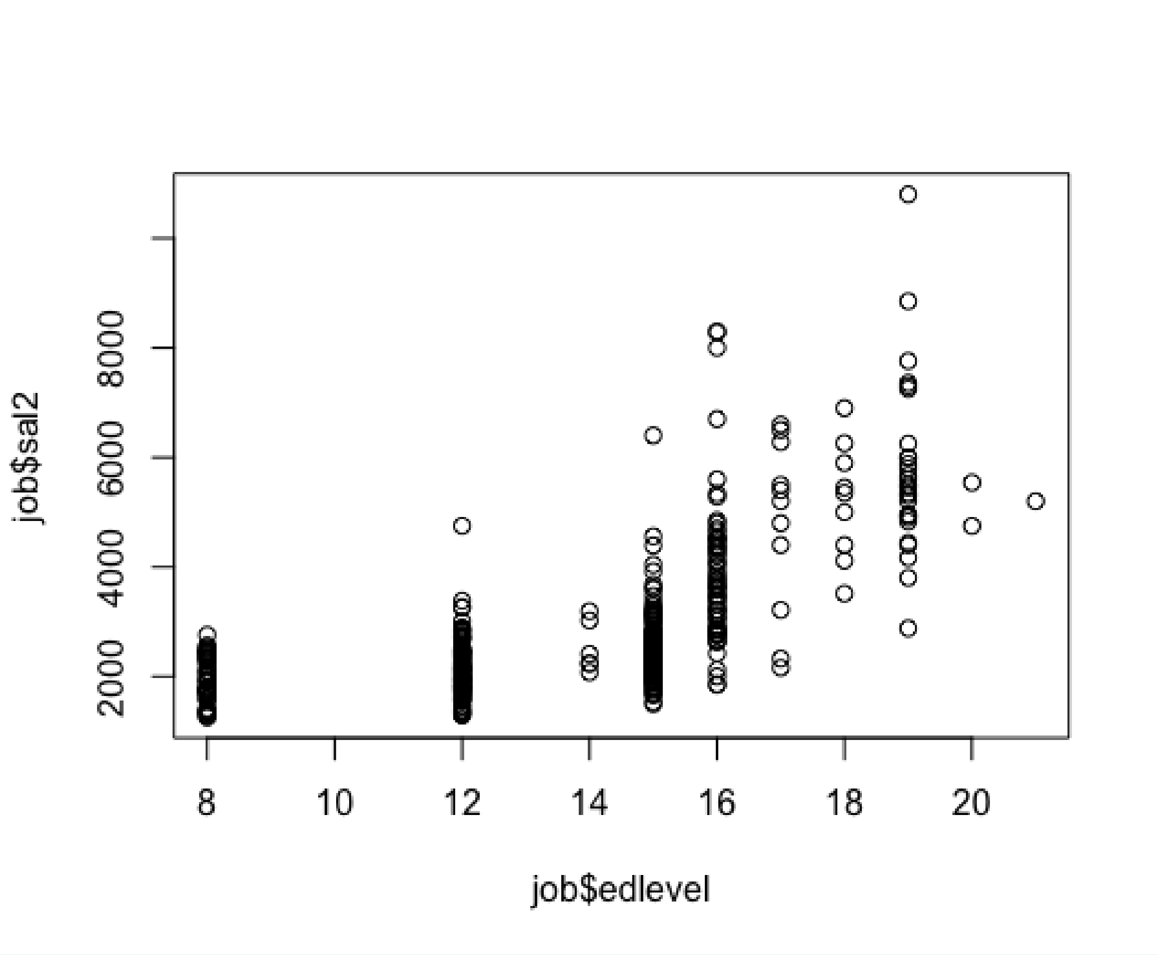
Data Analysis
(identifying observed
relationships
among measures)

Induction

Generalization of
observed relationships
among variables

Formulation of concepts
and propositions relating
concepts

Analyzing Relationships Between Interval-Ratio Independent and Dependent Variables: Correlation and Regression Analysis



This is the ***scatterplot*** of Y("sal2") against X("edlevel"), both interval-ratio variables

X (Edlevel) and Y(Sal2) are positively related – right?

There is a positive ***correlation*** between education and current salary among this sample of bank employees

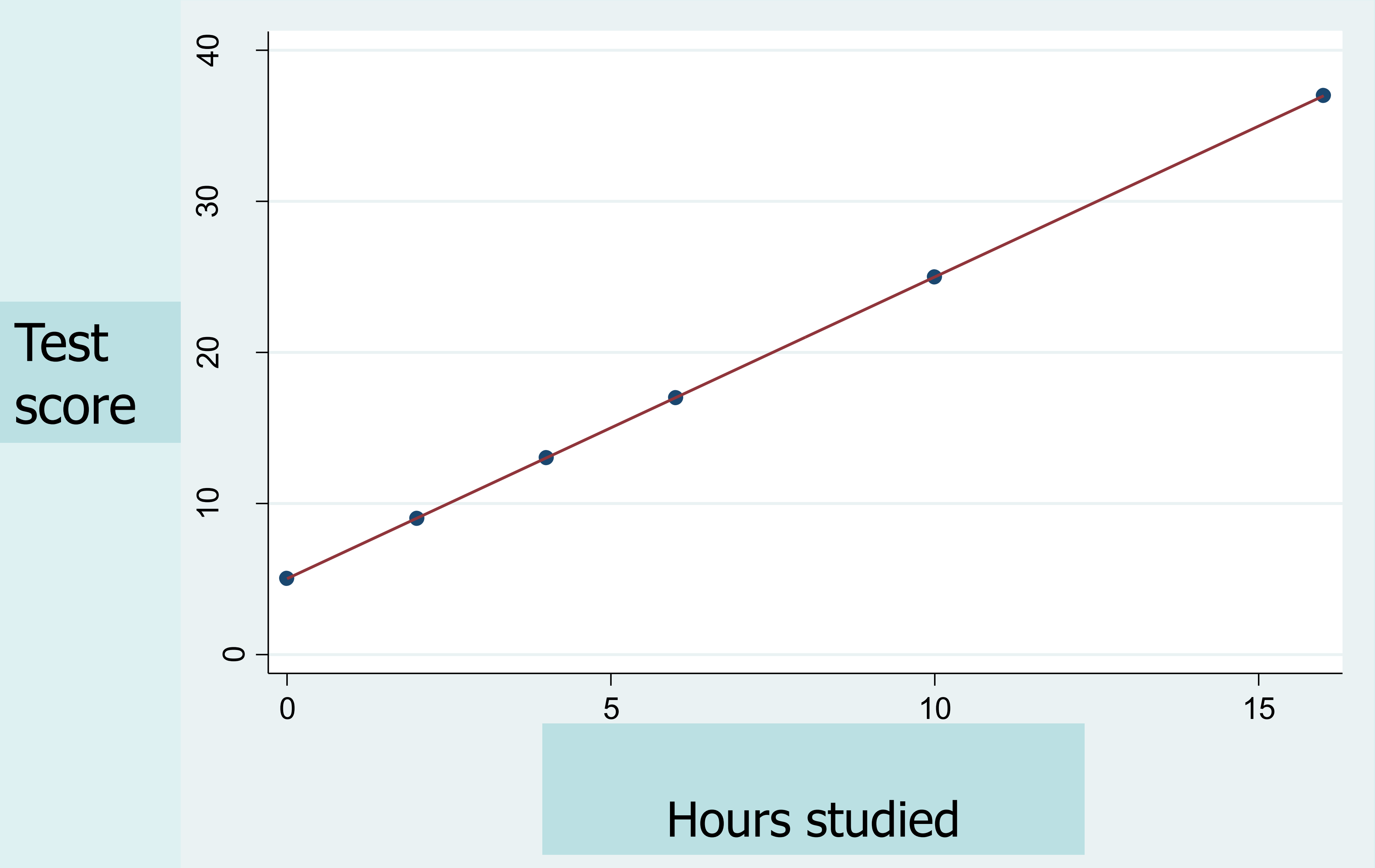
Can we be more precise about the nature of this relationship?

Regression Analysis

- Regression analysis is the procedure used for estimating the nature and exact form of the relationship between an *interval-ratio* Independent Variable (X) and an *interval-ratio* Dependent Variable (Y)
- Provides the "best fitting line" that characterizes the data, with the "exact form" of the relationship between X and Y being the **slope** of the line (represented by the symbol β)
- "As X changes by one unit, Y changes by β units"
- Also provides the foundation for *all* advanced statistical analyses – regression-type models are used for dichotomous and other categorical dependent variables, longitudinal and time-series models, "duration" models and many other examples from both pure and applied political science research

Regression Analysis: An Unrealistic Example

Set-up: Students in a statistics class take a 40 point test, and you plot the scores they received (Y) against the hours that they studied (X)


40
30
20
10
0
Test score
0
5
10
15
Hours studied

Regression Analysis: An Unrealistic Example

- The model is Y=5 + 2*X
- This is a "deterministic model," that is, the level of X *perfectly* explains the level of Y. There is no "error" in the relationship.
- The relationship can be graphed as a line, with the "starting point," or "intercept," being 5 when X = 0 and with Y increasing at a rate of change of 2 for every unit increase in X
- We say that the "slope" of this line is 2. The slope is defined as the amount that Y changes for each unit change in X. Changing X by 1 unit *always* produces a 2 (and *only* a 2) unit change in Y.
- So for every additional hour that a student studies, they get 2 more points on the test

- Is this realistic? That we would have a perfect prediction of test scores from study hours?

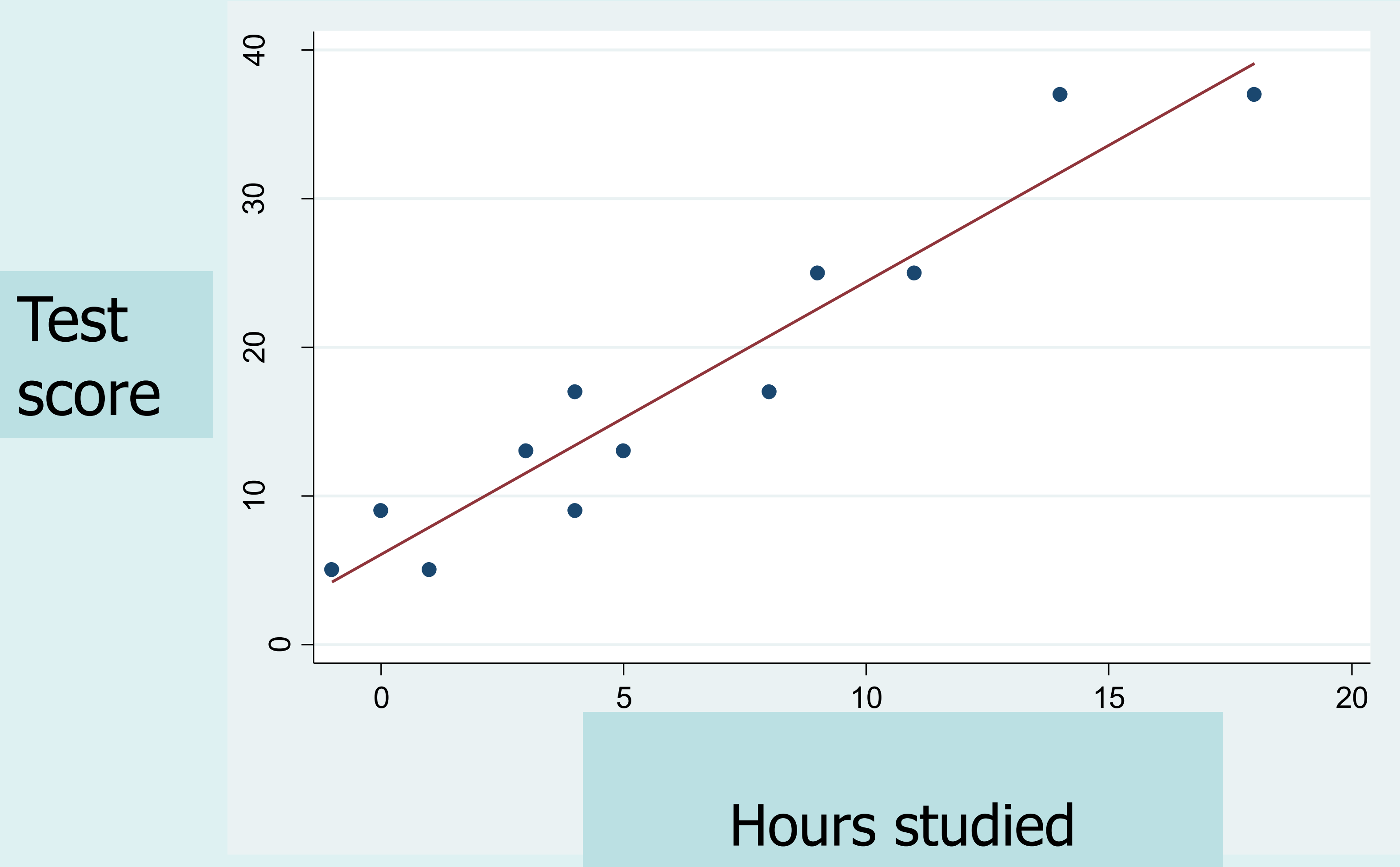
A More Realistic Example

- Problem: NO Social Science Relationships are Deterministic!!
- We *always* have a situation where

  $Y=\alpha+\beta*X+\varepsilon$

  Where ε represents some "error term," what we cannot explain about Y through knowledge of X
- In our example: Student test scores (Y) are a function of Hours Studied (X) **plus other factors**
- So: $Y=5+2*X + \varepsilon$

  and $Y'=5+2X$ (Y' means 'Predicted' Y)

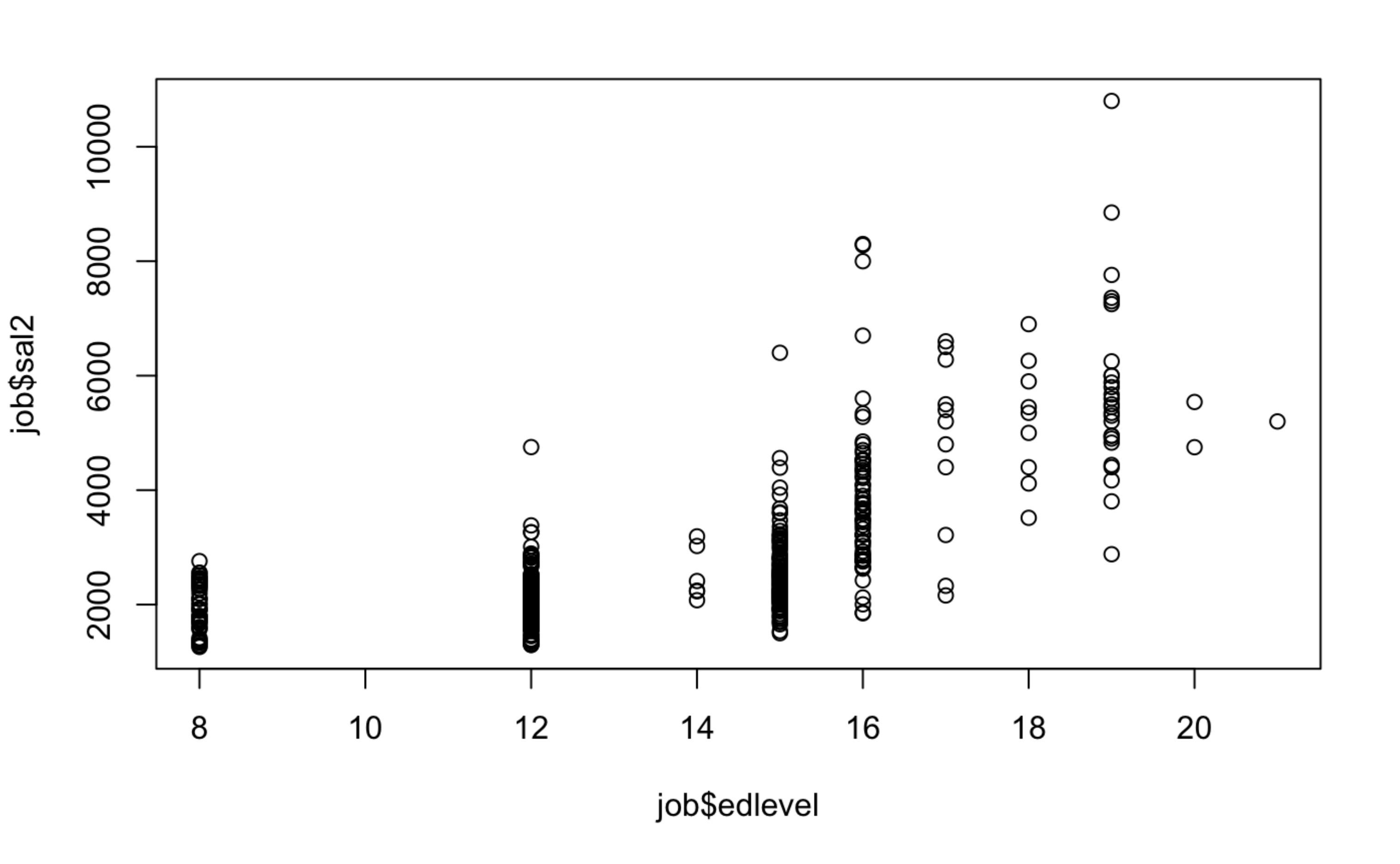

Test score
0
10
20
30
40
0
5
10
15
20
Hours studied

- So can see that the function 5+2x does not predict Y completely, there is some "residual" or unexplained portion of Y that is unrelated to X
- Some residuals are positive, meaning that the case is larger on Y than predicted from knowing X, and some residuals are negative, meaning that the case is smaller on Y than predicted from knowing X
- What is in the error term exactly?
  - Other variables that you have not included!
  - Measurement error in the variables
  - Intrinsic unpredictability of human behavior!

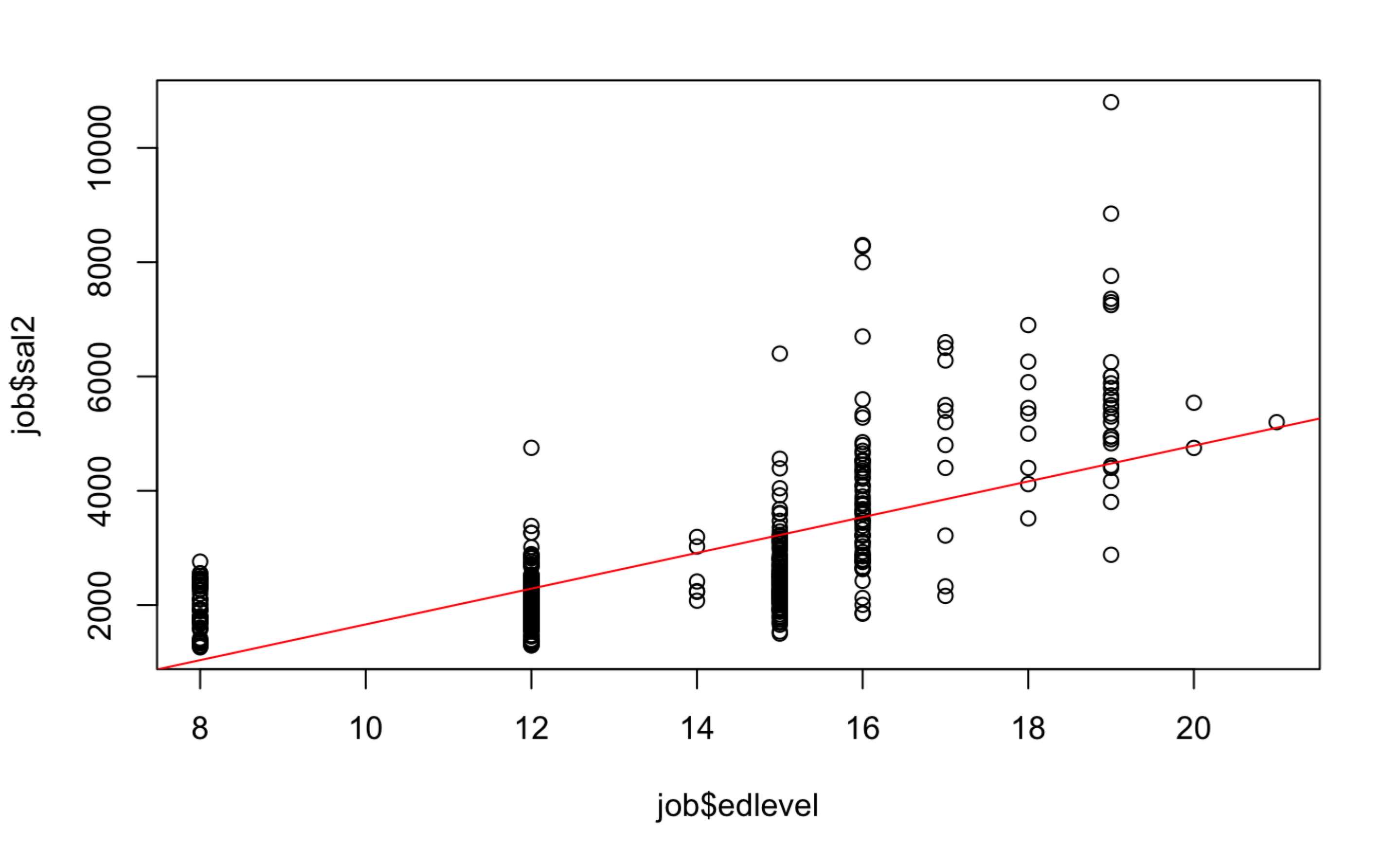
How Regression Analysis is Done in Practice

- We don't know the true form of the relationship, i.e. we don't know the values of α and β in the population. We need to estimate these parameters from a sample of data
- Produce a "scatterplot" graph of Y against X
- Figure out some method of generating the "best fitting line" to characterize the relationship between Y and X
- Estimate sample values of the intercept "a" (α) and slope "b" (β) of the line using that method
- Assess how well the line "fits" the data
- Test the hypothesis that the parameters (especially β) are 0 *in the population* (i.e., make inferences to the population value of β, given our sample value "b")
- Include additional variables in a **"multiple regression"** to better assess/explore the causal relationship between X and Y

An Example: Education and Current Salary (from our Job Training Dataset)



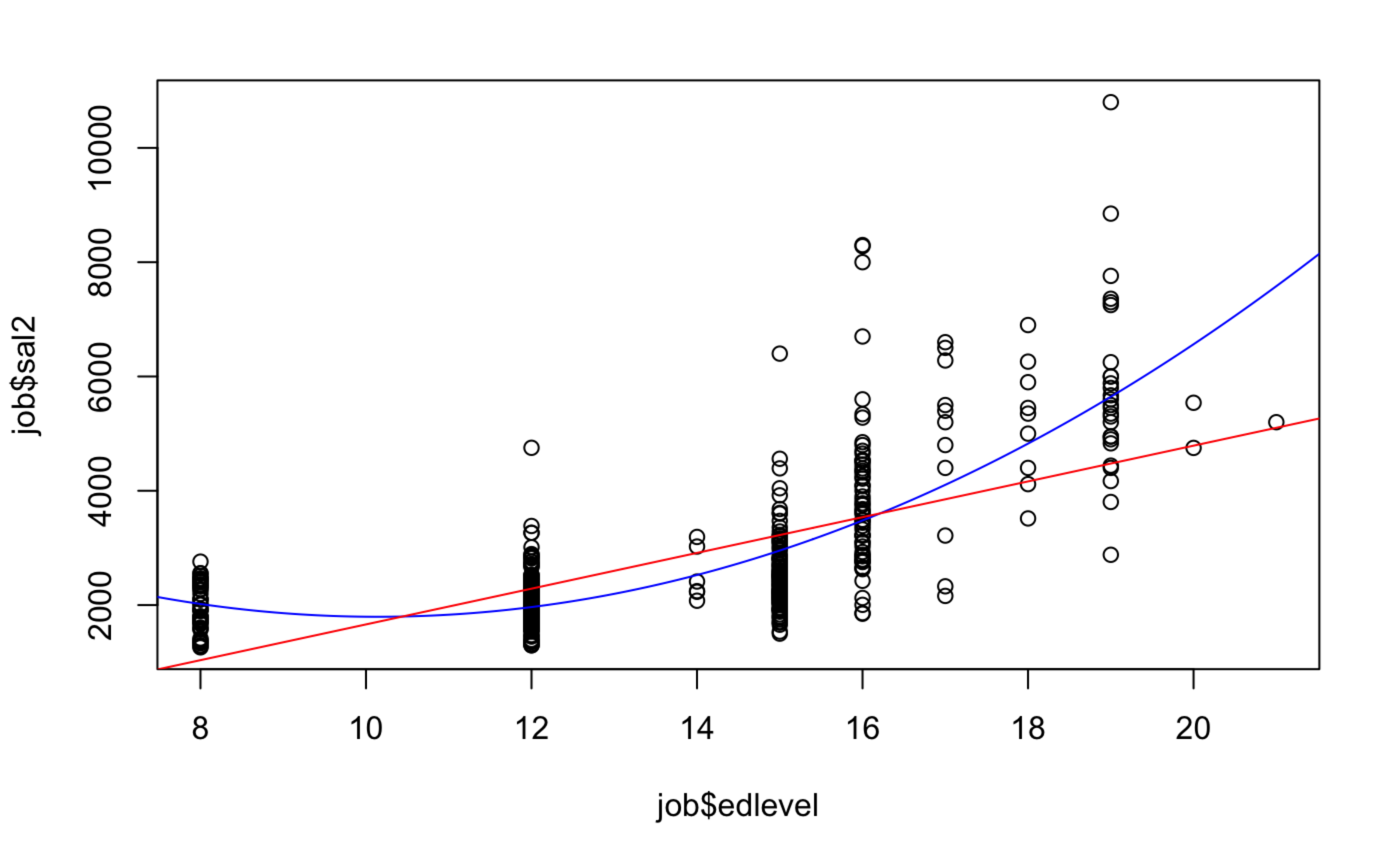
The Fitted Regression Line



```
job <- read_dta(file = "/Users/FINKEL/…..job_training.teaching.dta")
reg2 <- lm(sal2~edlevel,data=job)
            with(job,plot(edlevel, sal2), col="red")
            abline(reg2, col="red")
```

Notes:

- As education level increases, current salary increases (on average)
- Lots of errors in prediction from the "best fitting line"
- (Relationship may not be linear at all – see next slide!!)
- But assuming for now that it is a linear relationship, how do we get the best fitting line?


job$sal2
10000
8000
6000
4000
2000
8
10
12
14
16
18
20
job$edlevel

How to Fit a Regression Line

- Goal: Minimize errors of prediction, that is, find the line that leads to the smallest (in some sense) overall residuals from that line
- Residuals= $y_i – y_i$', where

  y_i = actual value of y and

  y_i'= predicted value of y from the line
- Several possible ways to minimize:
  - Minimize Σ ($y_i – y_i$')
    - Problem: Any line between X-bar and Y-bar will satisfy this criterion
  - Minimize Σ |$y_i – y_i$'|
    - Problem: Does not lend itself to application to statistical inference, i.e. not tied to normal sampling distributions, central limit theorem, etc.

The “Least Squares” Criterion

- Sample Equation: $y_i = a + b*X + e$
- Prediction of $y_i' = a + b*X$
- Minimize $\Sigma\ (y_i – y_i')^2$
- In English: find the line that produces the fewest *squared deviations* of the actual values on Y among cases in the sample from the predicted value of Y from the line
- Since $y_i'= a+b*X$, criteria of least squares yields:
- Minimize $\Sigma\ (y_i – a\text{-}b*X)^2$ or $\Sigma\ (a+b*X\text{-}\ y_i)^2$ with respect to “a” and “b”
- This is an (easy) calculus problem (but we don’t care about how this is done!!!)

The Least Squares Solution

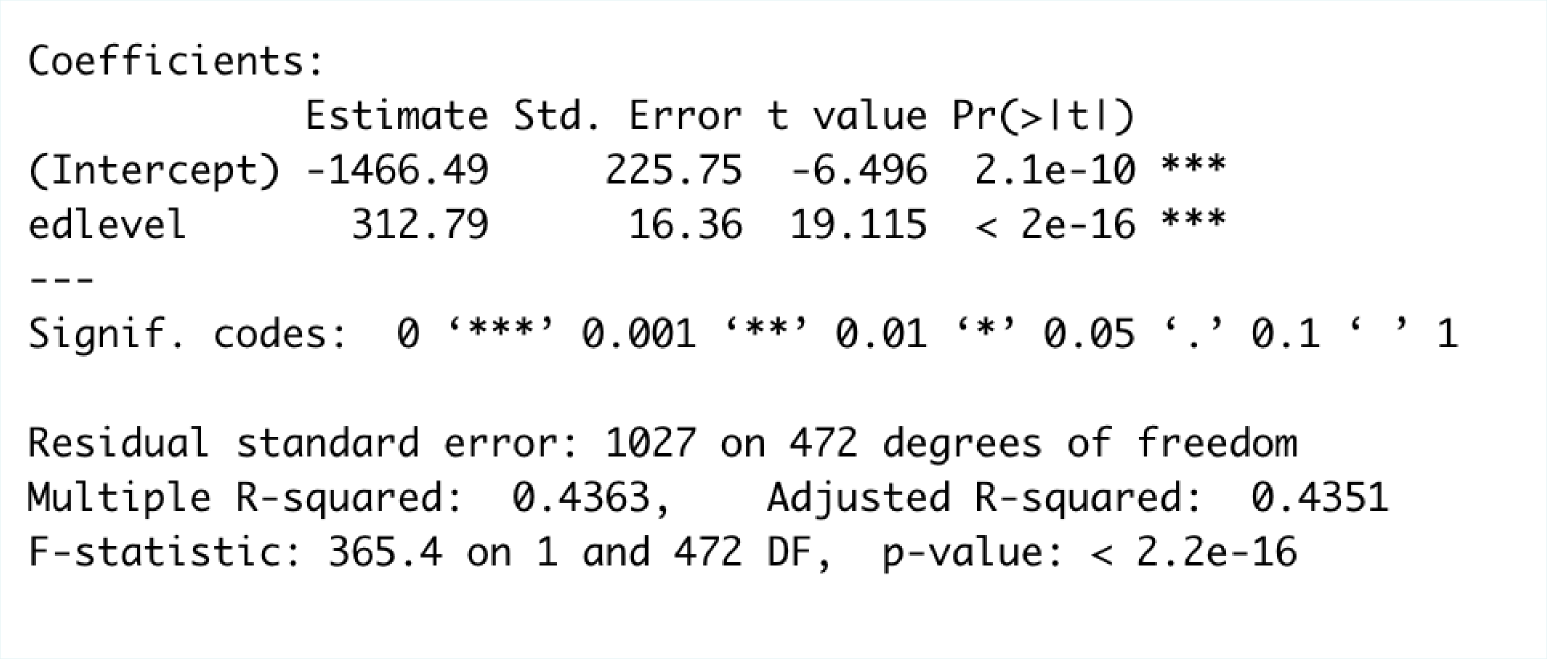
$$a = \bar{Y} - b * \bar{X}$$

$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Interpretations

- “a” Interpretation: the predicted place on the line for Y when X=0
  - May not be realistic to have X=0 so be cautious in this interpretation. It is a mathematical necessity but substantively not always meaningful
- “b” Interpretation: as X changes by 1 unit, Y changes on average by β units. This is the “estimated effect of X on Y”
- Note “b” formula: “Covariance Between X and Y, Divided by the Variance of X

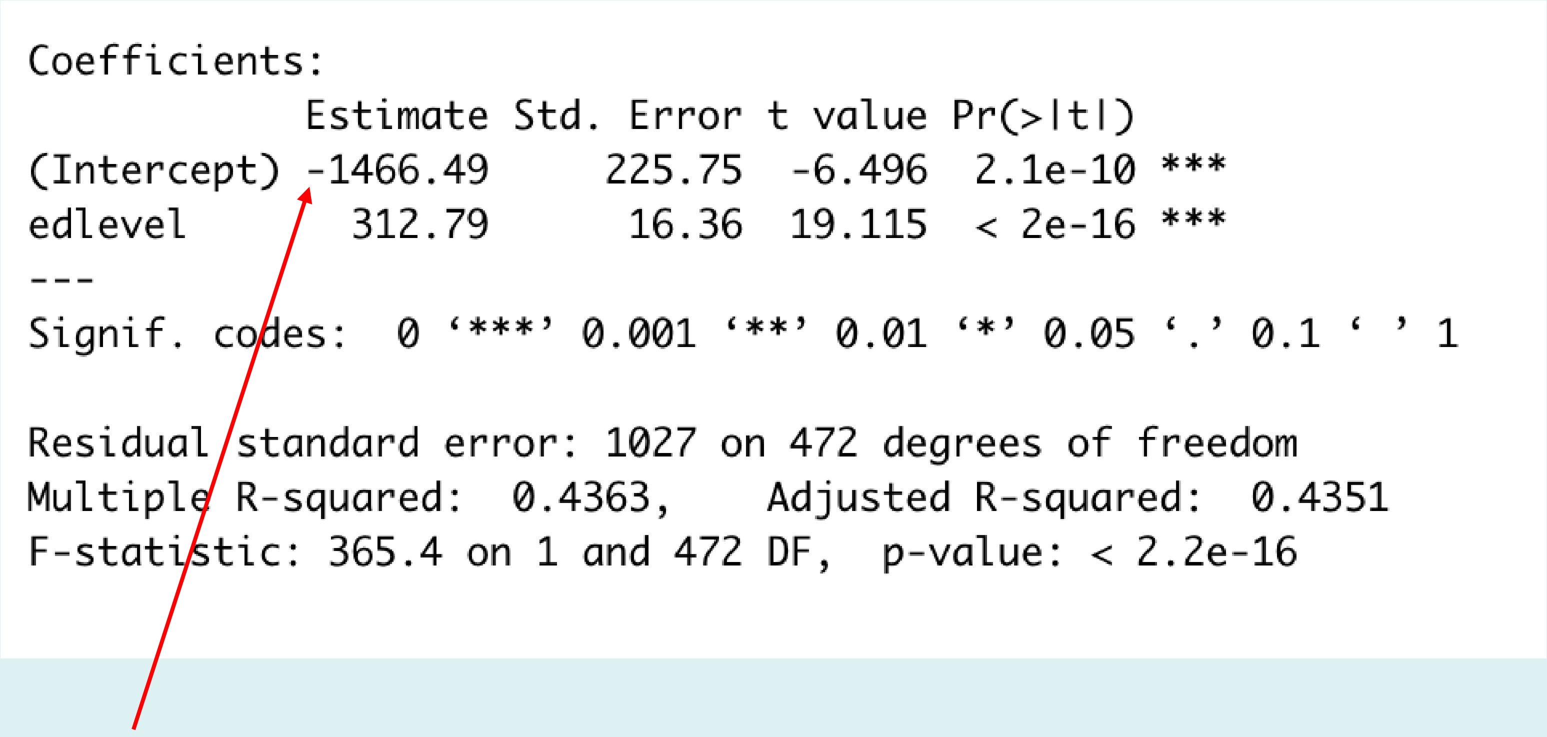
Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,    Adjusted R-squared:  0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

```
reg1 <- lm(sal1~edlevel, data=job)
summary(reg1)
```

Our R Results

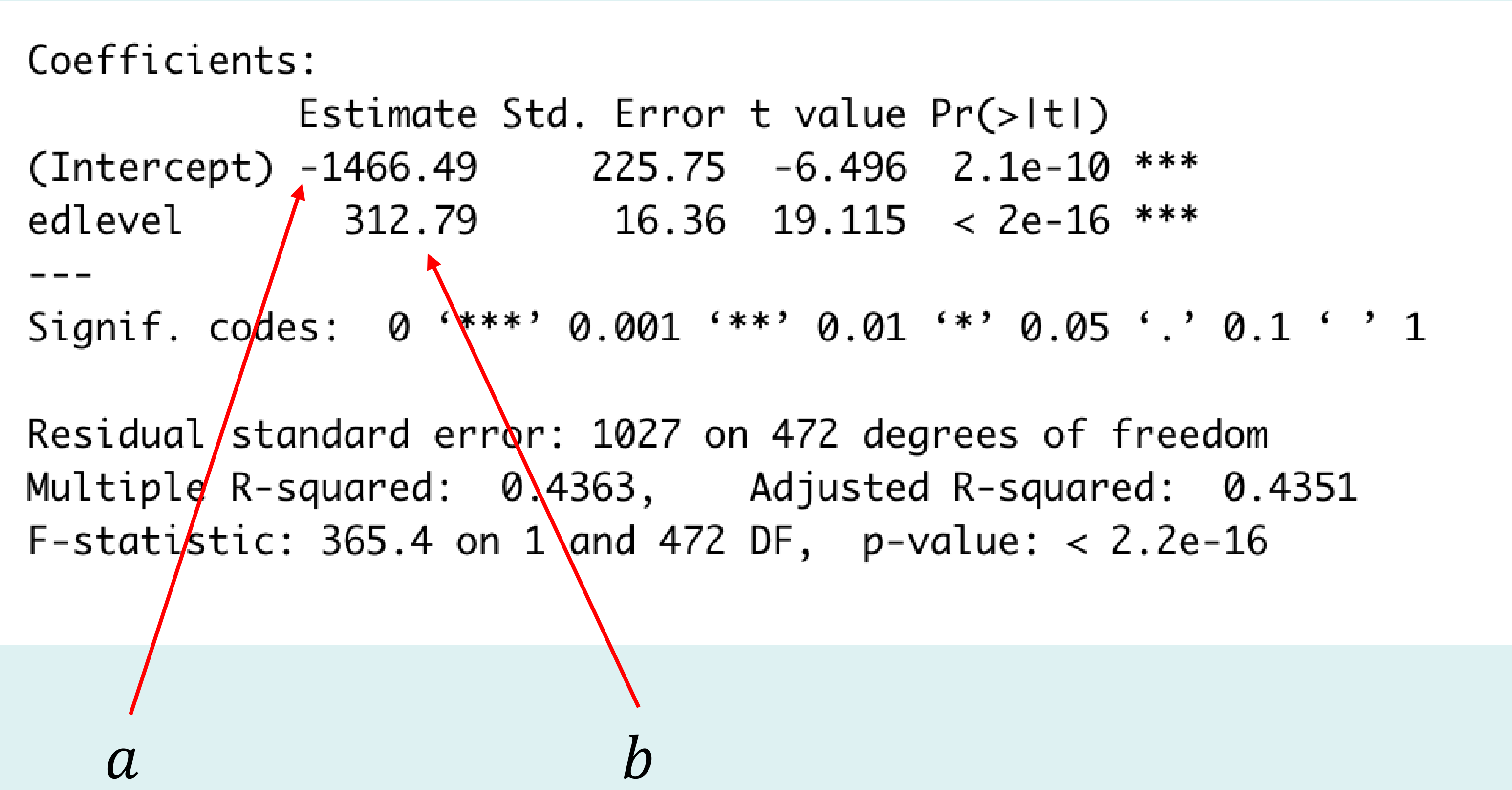


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,    Adjusted R-squared:  0.4351 
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

a

```
reg1 <- lm(sal1~edlevel, data=job)
summary(reg1)
```

Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,	Adjusted R-squared:  0.4351 
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

reg1 <- lm(sal1~edlevel, data=job)
summary(reg1)

- a = -1466.49 So when a person has NO education, we predict his/her salary to be -1466 dollars. Example of a nonsensical intercept because no one has NO education

- b = 312.79 So for every additional year of education a person has, we predict his/her salary to be 313 dollars more, on average

- Our best guess of any individual's current salary, conditioned on X (education) is

$y_i' = -1466.49 + 312.79 * \text{Education}$

This is the prediction from our regression line!!!

Notes On Bivariate Regression

- "b" represents *average* change in Y for a unit change in X. We will not necessarily be accurate in predicting individual values of Y, and one of the measures of the strength of the relationship will be how accurate in fact we are in predicting Y from X (see next slides)
- "b," nevertheless, represents our best estimate of the *substantive* relationship between X and Y, expressed in the given units of Y. That is, each year of education gives you 313 more dollars per month. ALWAYS INTERPRET THIS SUBSTANTIVELY – IS IT A LOT, A LITTLE, BIG, SMALL, OR WHAT?
- "b" is our estimate of the causal effect of X on Y, but what we have done so far *almost never* gives us the true picture of the causal relationship
  - Sampling Error – β in the overall population could still be 0
  - Reciprocal Causality in Non-Experimental Designs
  - Spuriousness caused by Omitted Z1, Z2, Z3, Z4, Etc.!!!!

Example of Regression Analysis in Political Science:
Lewis-Beck and Tien, “Election Forecasting for Turbulent Times”
PS: Political Science and Politics (October 2012)

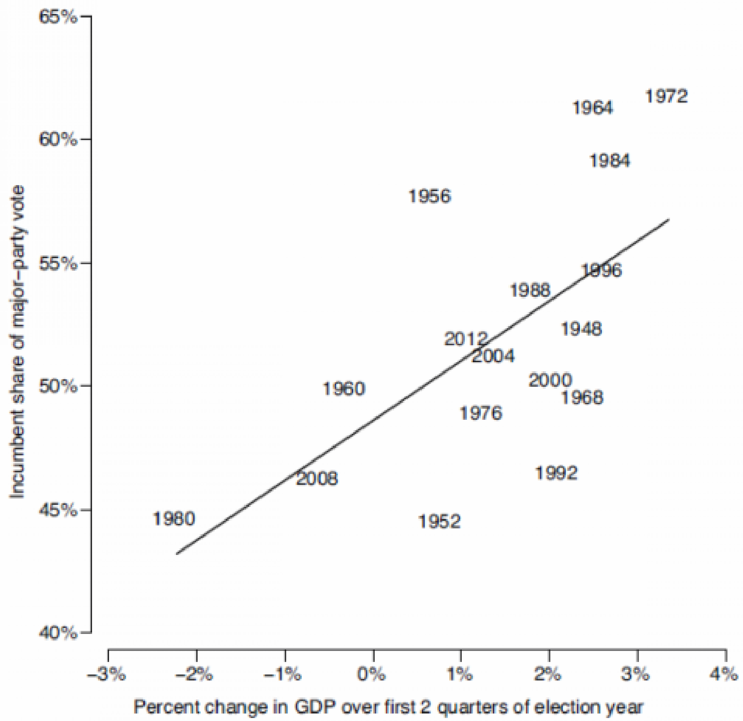
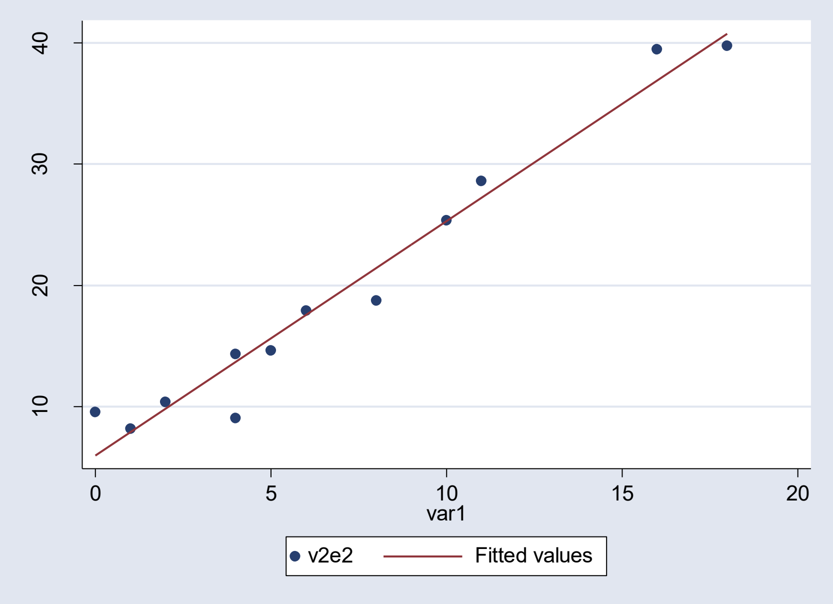


Figure 7.1.
Economic growth and presidential election outcomes, 1948–2012.
Note: The relationship between change in GDP and the vote—the diagonal line—is estimated without the 2012 election included. This shows how close the 2012 outcome was to what we would predict based on the historical relationship between GDP and the vote from 1948 to 2008.

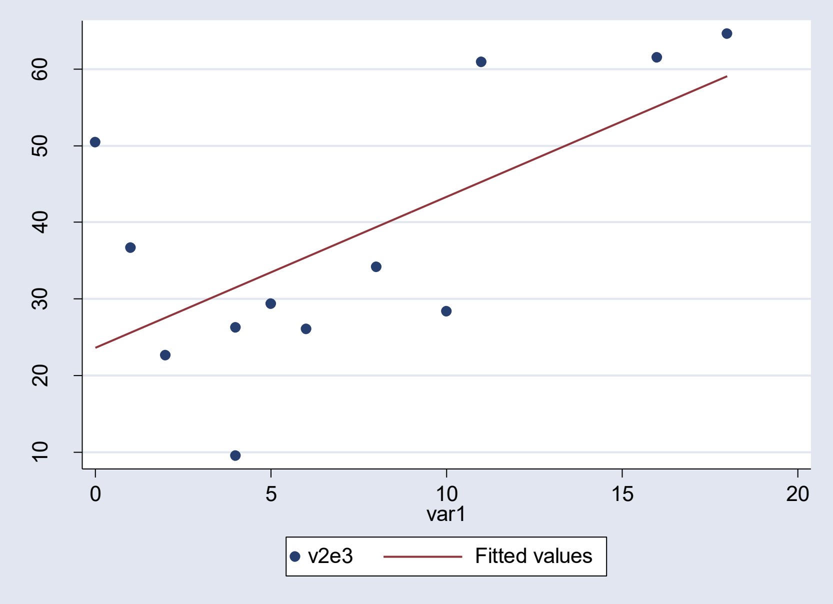
Assessing the Strength of a Regression Relationship

- Basic Question: How close are the points, on average, from the regression line? If very close, we have a "strong" relationship; if not close, we have a "weak" relationship
- Simplest measure of association in this regard is called the *correlation coefficient*, or "Pearson's correlation." Runs from -1, for a perfect negative relationship, to 0 for no relationship, to +1 for a perfect positive relationship
- Formula: $\frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{(x_i - \overline{X})^2 (y_i - \overline{Y})^2}}$


40
30
20
10
0
5
10
15
20
var1
● v2e2
Fitted values

← r=.98

r=.64 →


60
50
40
30
20
10
0
5
10
15
20
var1
● v2e3
Fitted values

Problems/Issues with Pearson‘s R

- Gives crude sense of correlation only, no exact “unit change” in Y interpretation like β
- Bidirectional statistic, in that r(XY)= r(YX), so says nothing about causality (reflecting the truism that “correlation does not equal causality”)
- We don‘t know what the meaning of the -1 to 1 scale is, aside from bigger numbers means stronger positive or negative relationship, and numbers closer to 0 mean a weak or no relationship.

More Useful Measure: R-squared

- Question: How much of total variation of Y around its mean $\bar{Y}$ is "explained" by X, and how much is residual or "unexplained" variation?
- Total Variation in Y = $(Y_i - \bar{Y})$
- Total Residual Variation in Y = $(Y_i - Y_i')$
- Total "Explained Variation", i.e. variation from the regression line to Ybar: $(Y_i' - \bar{Y})$
- So: $\Sigma\ (Y_i - \bar{Y}\ )^2 = \Sigma(Y_i' - \bar{Y}\ \ )^{2\ +}\ \ \Sigma\ (Y_i - Y_i')^2$

  Total Variation = Regression + Residual

  Variation Variation

Illustration of R-Squared

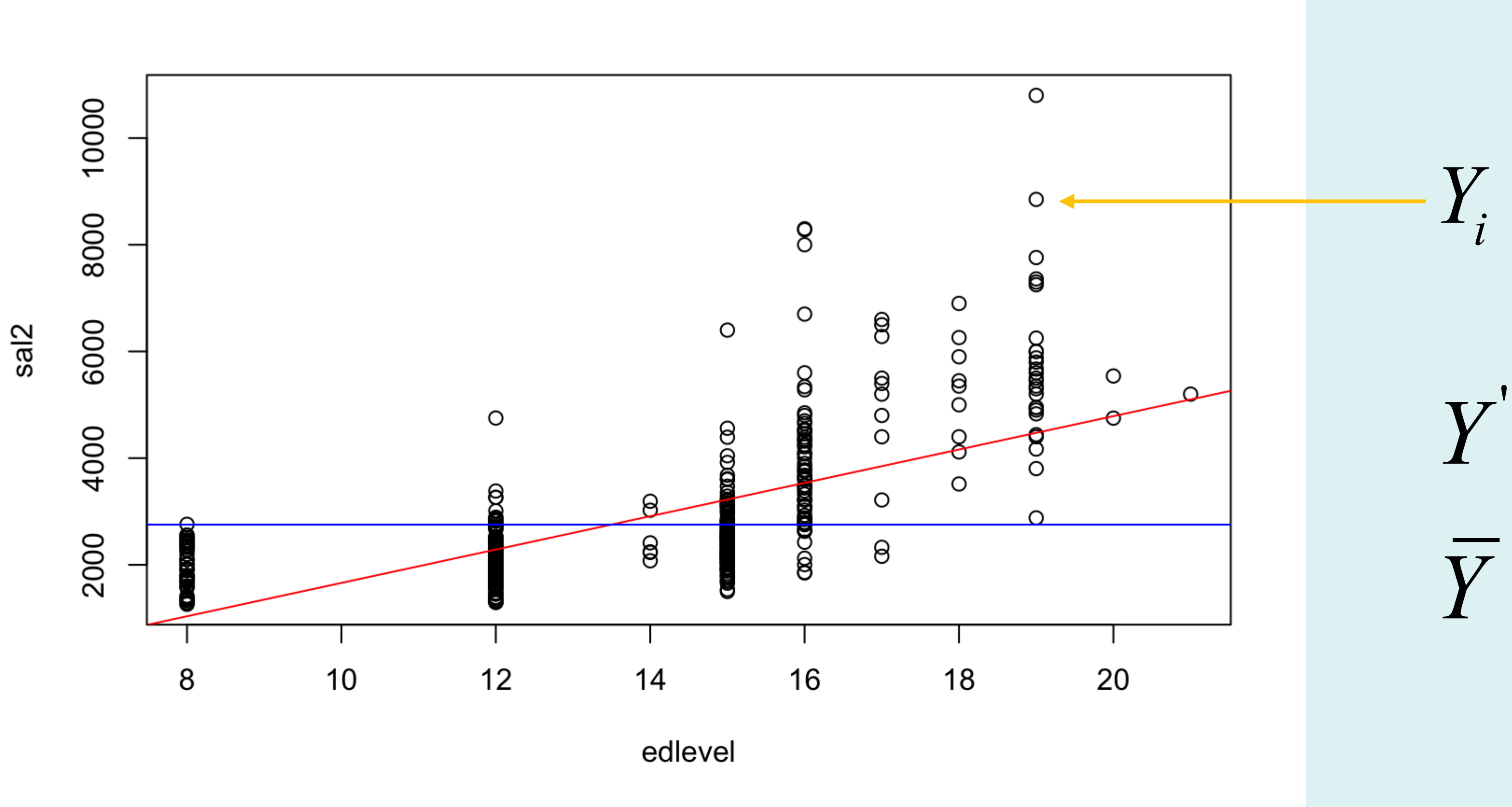


Illustration of R-Squared

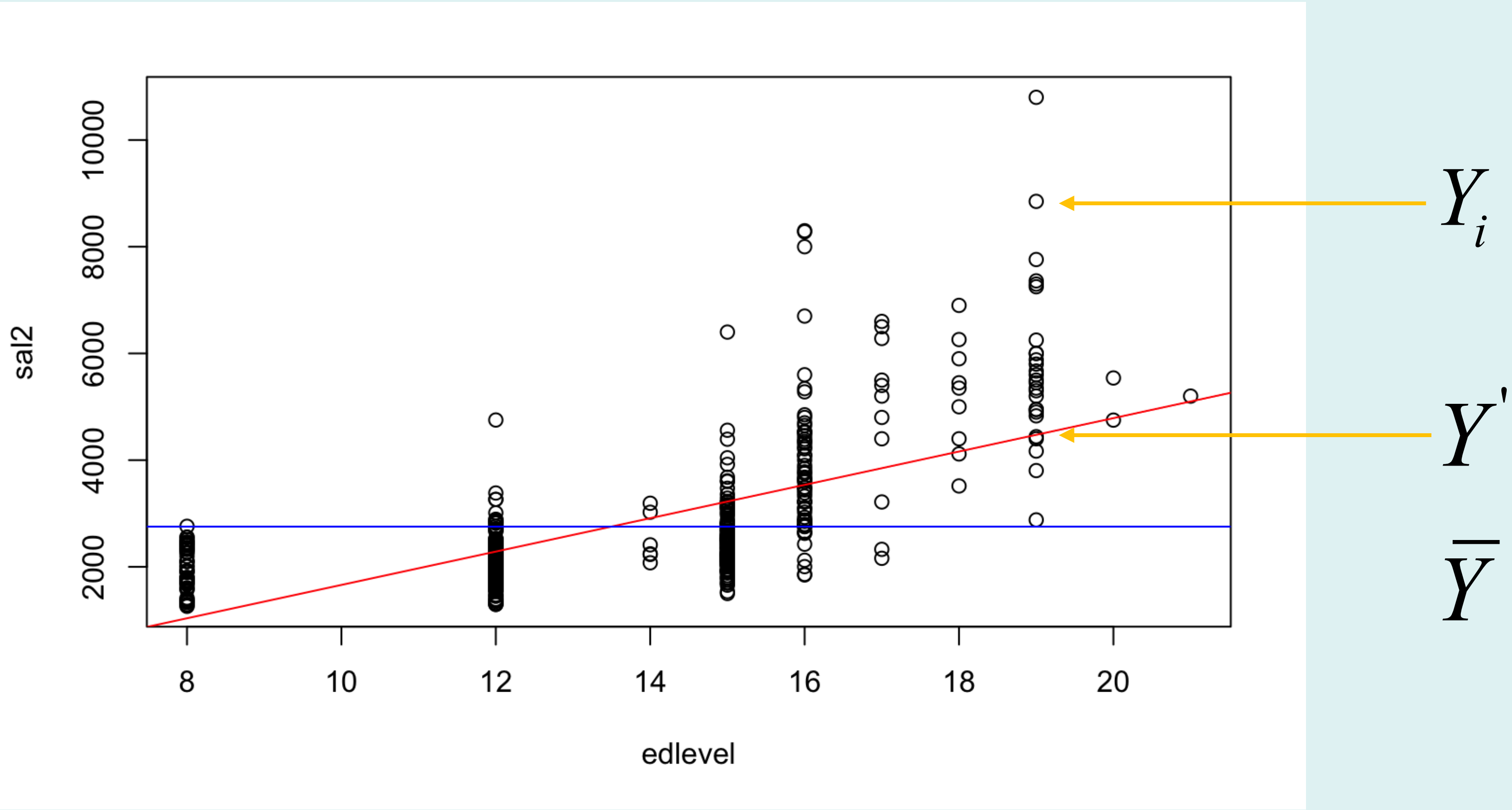
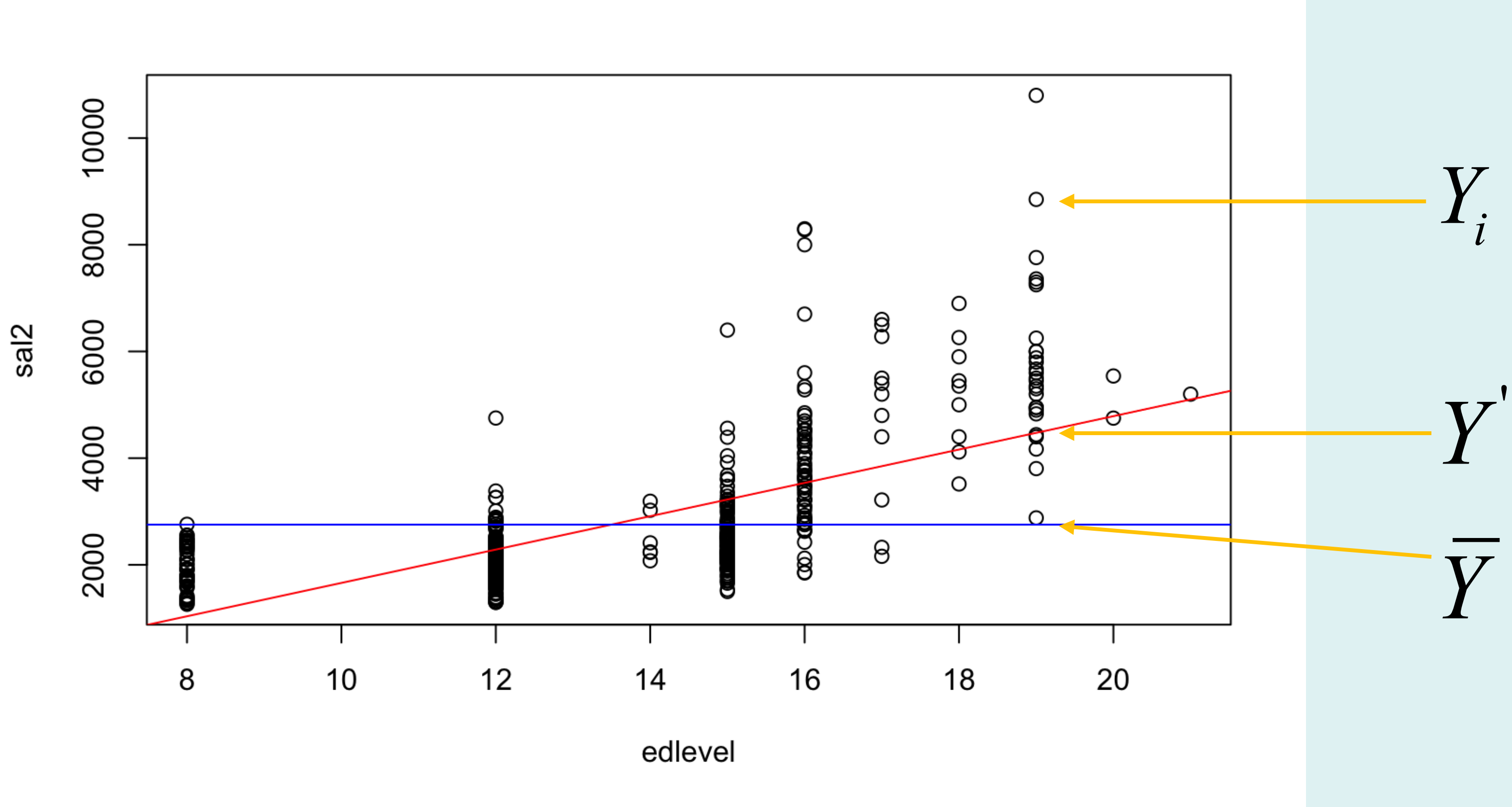


Illustration of R-Squared



$$Y_i - \bar{Y} \quad = \quad (Y_i - Y_i') \quad + \quad (Y_i' - \bar{Y})$$

Total Variation = Error Variation + Regression Variation

- R-squared= (Squared) Regression Variation Divided By (Squared) Total Variation

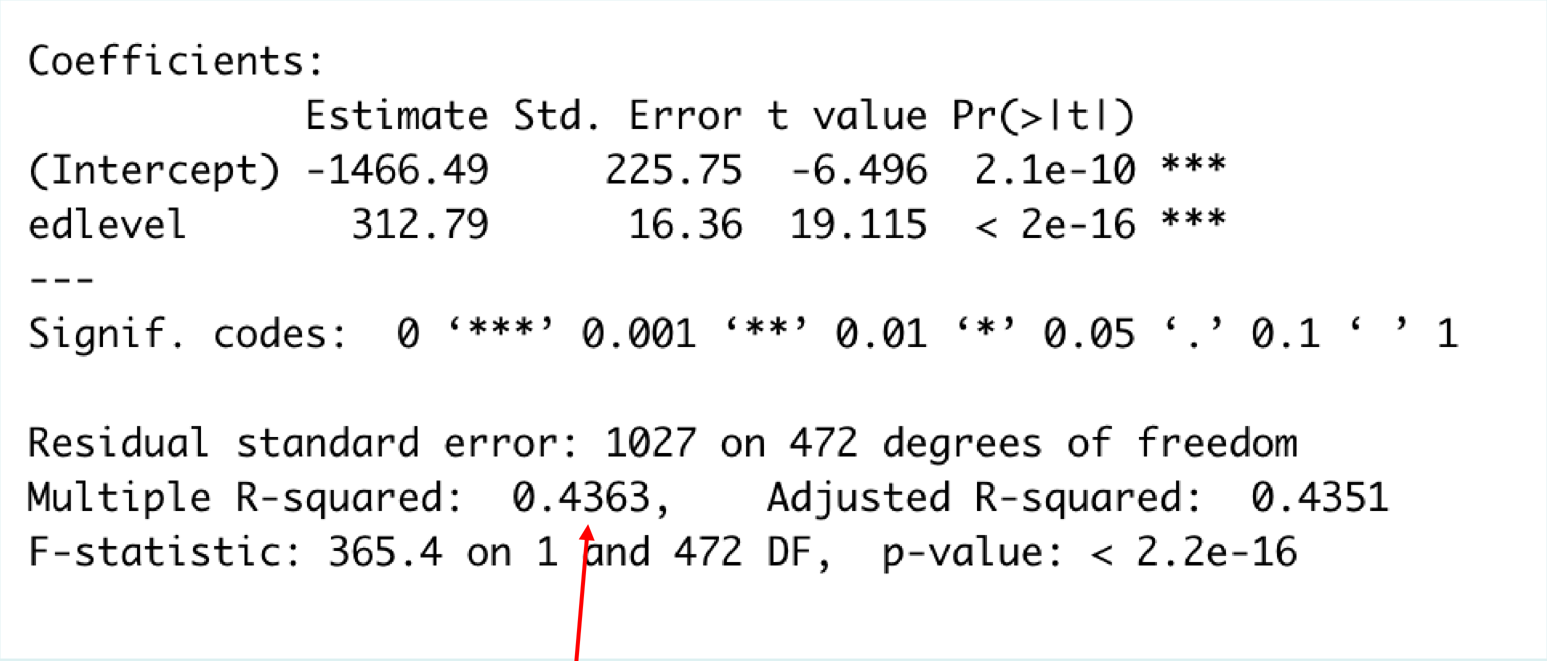
- It is the proportion of Total Variance “Explained” by the Regression Line

$$\frac{\sum_{i=1}^{n}(Y_i' - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- Goes from 0 to 1, so can interpret it as a proportion

- Our example: R-squared=.436

Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,	Adjusted R-squared:  0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

R-squared

43.6, or 44% of the variation in Sal2 is “explained” by variation in Edlevel. Education “explains” or “accounts for” nearly half of the total variation in current salary.

Notes on R-squared

- $R^2 = (r)^2$; that is, R^2 for a bivariate regression is equal to Pearson r, squared
- Low R^2 either means that you have a poor explanatory model, or the relationship is possibly non-linear, or you have little variation in your DV to explain in the first place
- High R^2 either means that you have a good model, or that you have a good statistical model but not necessarily a meaningful substantive model (e.g. Y at time t predicted by Y at time t-1)
- Therefore, treat R^2 as highly informative about model fit but don’t take it on face value automatically

Hypothesis Testing in Regression Analyses

- As with all statistical analyses, once we have established the nature of the regression relationship between X and Y in our sample, we need to see whether we can reject the idea that there is truly no relationship between X and Y in the overall population. In other words, we need now to conduct statistical inferences from our sample to the population
- The tool for this, as usual, is the *sampling distribution* – this time, the sampling distribution of *slopes*, which is what would result if we were to sample the population an infinite number of times, calculate the "Least Squares" slope and plot the distribution of these sample slopes.

- IF a series of assumptions are met (which we won‘t have time to discuss unfortunately), then such a sampling distribution would:
  - Be normally distributed
  - Be centered around the true population slope β
  - Have a standard deviation (“standard error”) of:

$$s.e._b = \frac{\sqrt{\frac{\sum(Y_i - Y_i')^2}{N-2}}}{\sqrt{(X_i - \bar{X})^2}}$$

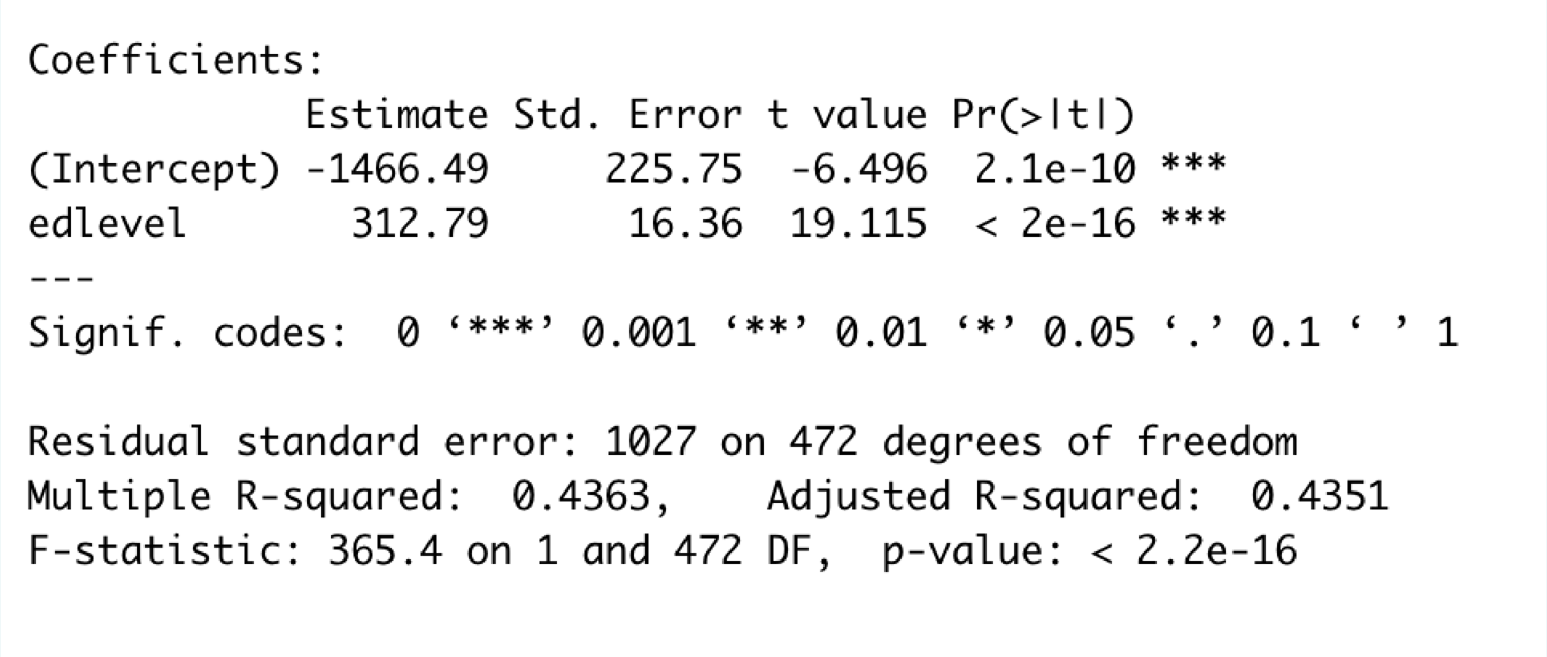
- So we now know how to test the null hypothesis that β= 0 in the population (right?)

- H_0: $\beta = 0$
- H_A: $\beta \neq 0$
- $\alpha = .05$ (as usual, the .05 significance level)
- Test statistic: t
- Calculation:

$t = (b(\text{sample}) – \beta_0)/ \text{s.e.}_b$

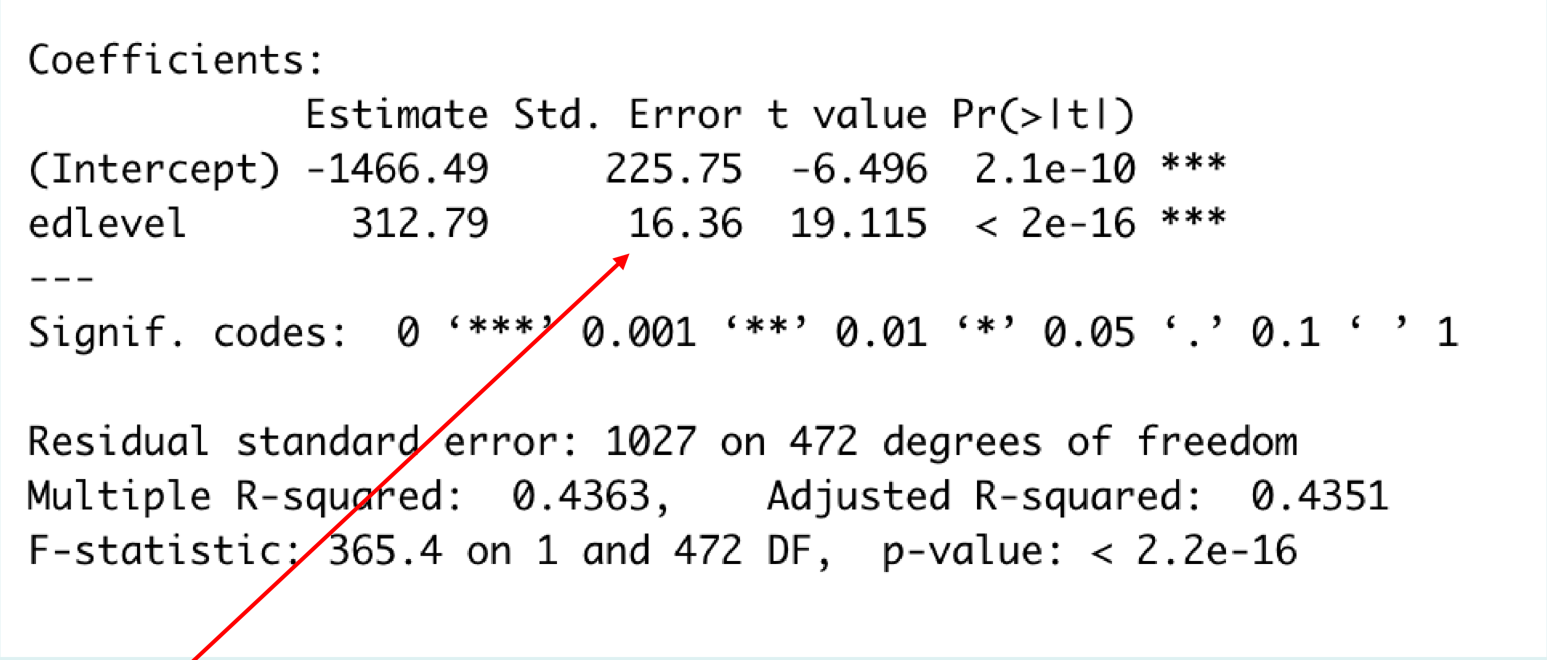
How many standard errors away from β_0 is our sample estimate b? If a lot (more than 1.96 or less than -1.96), we reject the Null at the .05 level. If not, we cannot reject the Null at the .05 level.

Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,	Adjusted R-squared:  0.4351 
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

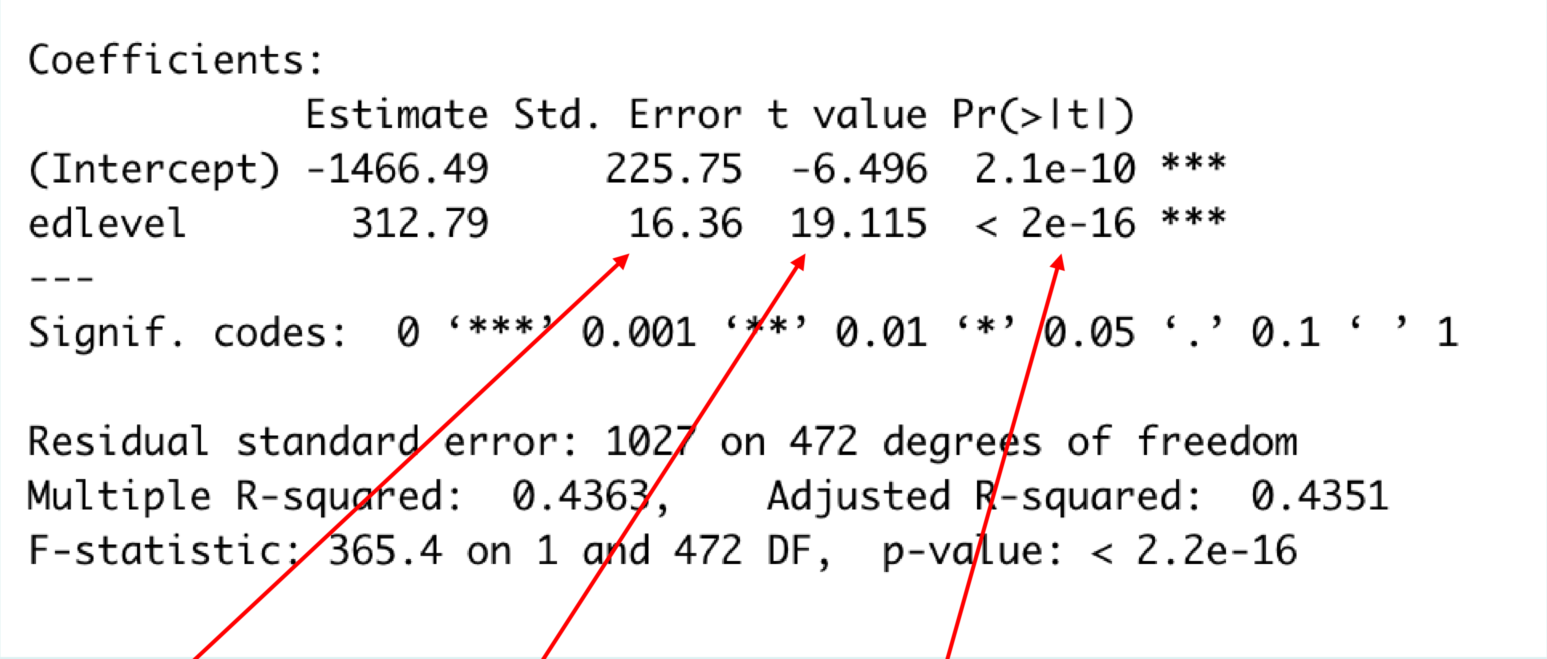
Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,    Adjusted R-squared:  0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

Standard error of b

Our R Results



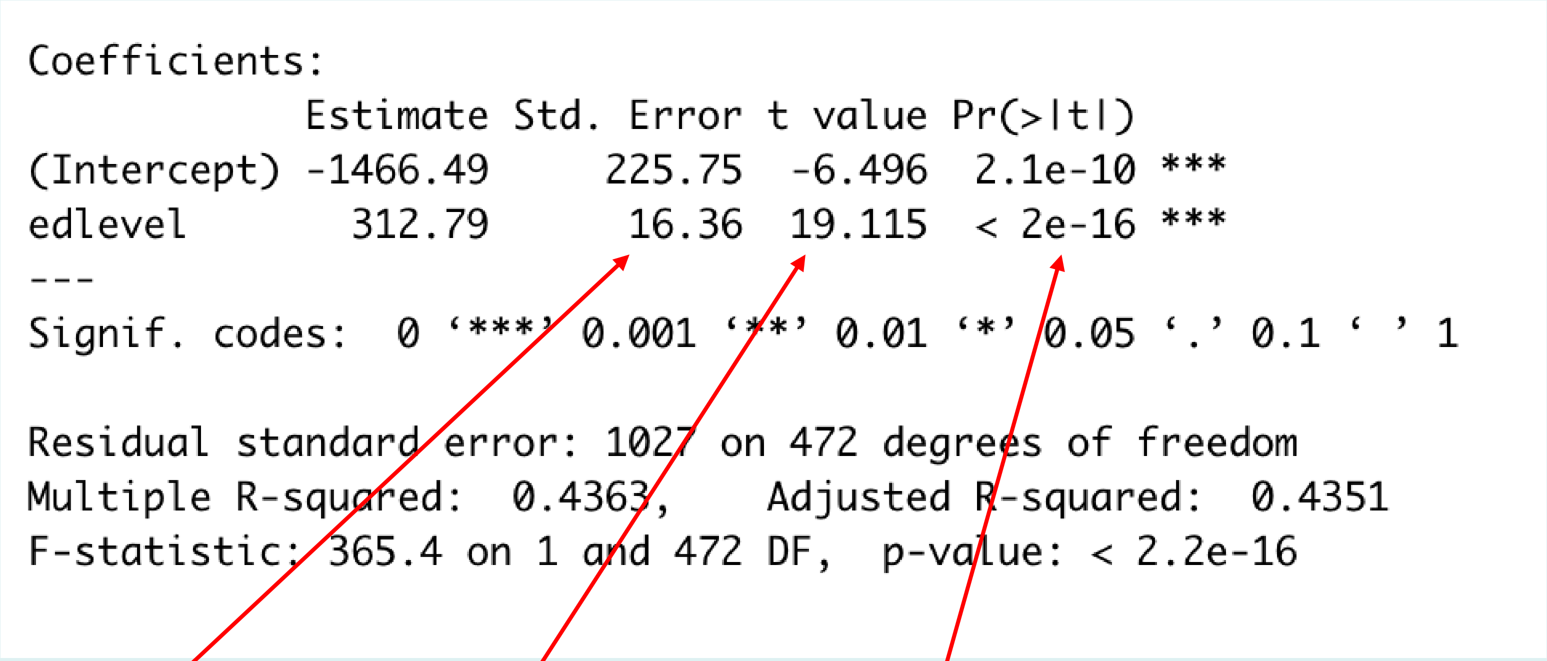
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,	Adjusted R-squared:  0.4351 
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

Standard error of b

t= (312.8 – 0)/16.36 = 19.12

Decision: Reject Null Hypotheses! The probability of observing a t value of 19.12 if the null hypothesis were true is infinitesimal!!!

Our R Results



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1466.49     225.75  -6.496  2.1e-10 ***
edlevel       312.79      16.36  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 472 degrees of freedom
Multiple R-squared:  0.4363,	Adjusted R-squared:  0.4351 
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```

Standard error of b

t= (312.8 – 0)/16.36 = 19.12

Decision: Reject Null Hypotheses! The probability of observing a t value of 19.12 if the null hypothesis were true is infinitesimal!!!

Remember statistical versus substantive significance!

Remember next steps: bringing in 3rd/4th/5th variables to control for possibly confounding Z factors!