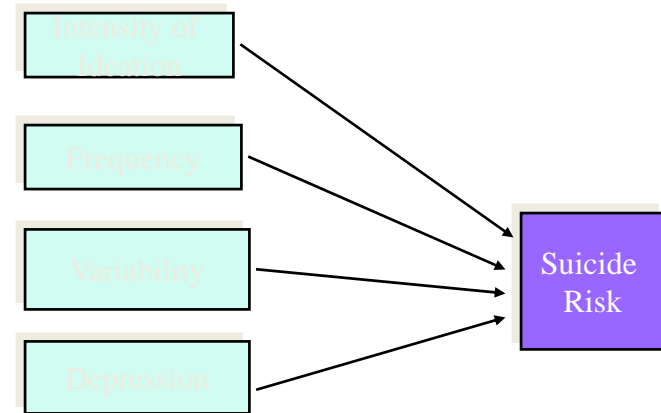


# Multiple Regression

# Multiple Regression: What Is It?

- Multiple regression is a collection of techniques in which there are *multiple* predictors of varying kinds and a single outcome
- We are interested in discovering the general and specific relations among the variables as well as the actual predictive ability of the model for future cases
- For much of the rest of the discussion we will focus on ordinary least squares regression with continuous outcomes



# Regression

- Using the covariances among a set of variables, regression is a technique that allows us to predict an outcome based on information provided by one or more other variables
- The regression equation, the linear model, simply extends the previous simple equation to include more variables

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

# The Best Fitting Surface

- Before we attempted to find the best fitting line to our 2d scatterplot of values
- With the addition of another predictor our cloud of values becomes 3 dimensional
- Now we are looking for what amounts to the best fitting plane
- With 3 or more predictors we get into hyperspace and are dealing with a regression surface
- Regression equation:

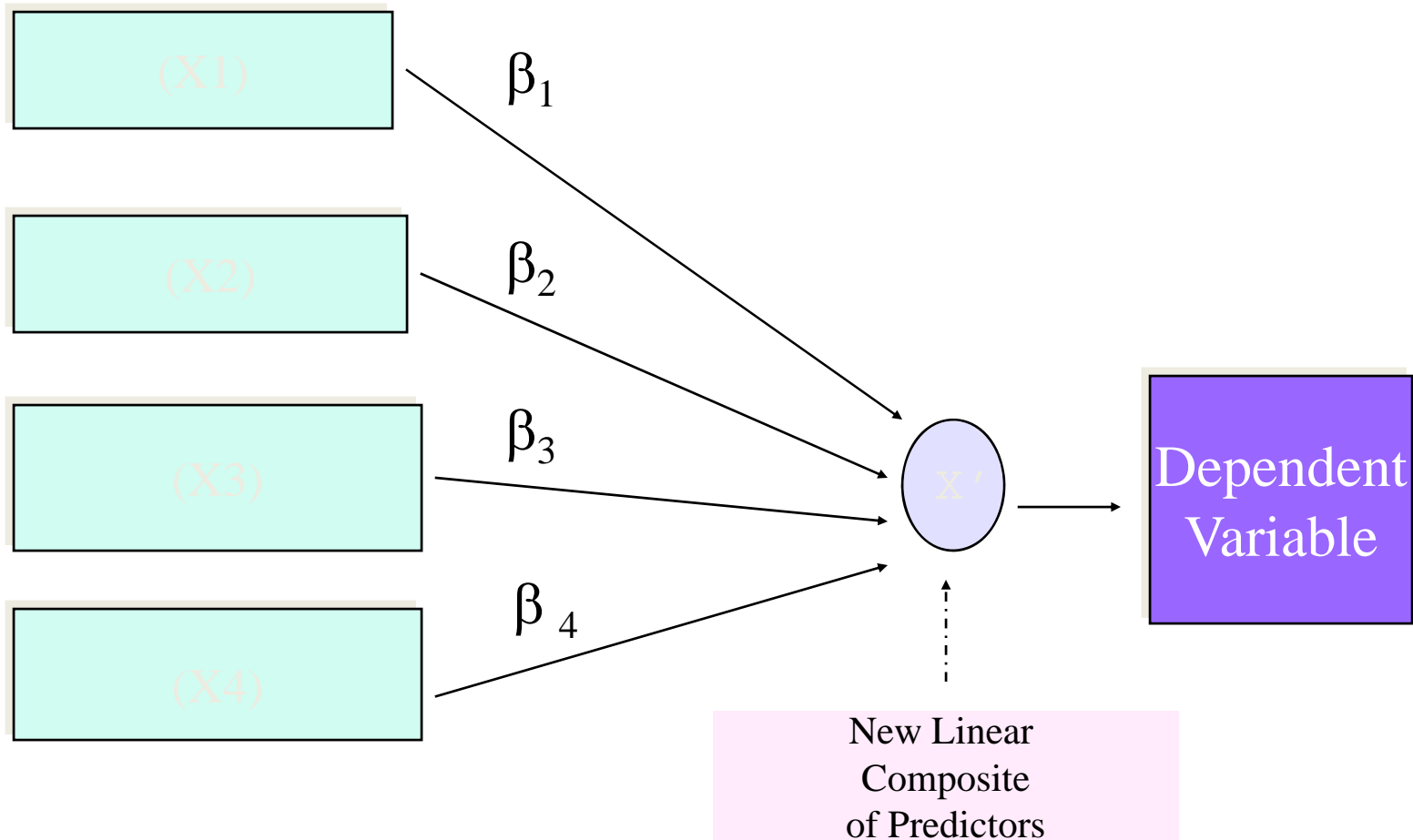
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$



# Linear combination

- The notion of a linear combination<sup>1</sup> is important for you to understand for MR and multivariate techniques in general
- Again, what MR analysis does is create a linear combination (weighted sum) of the predictors
- The weights are important to help us assess the nature of the predictor-DV relationships with consideration of the other variables in the model
- We then look to see how the linear combination in a sense matches up with the DV
- One way to think about it is that we extract relevant information from predictors to help us understand the DV

# Conceptual Understanding



# Understanding Regression

- To understand regression one must interpret it at both a model level (macro level) and at the level of individual predictors (micro level)
- Another interpretation distinction that can be made relates more to the goals of the study, which are not mutually exclusive in practice
  - Prediction vs. Explanation



# Steps in Regression

- Define your ideas/theories in scientific terms and specify the model(s) accordingly
  - At this point determine whether prediction or explanation is the primary goal but maintain a balance between the two
- Obtain good (not ‘adequate’<sup>1</sup>) measures of the variables of interest
- Inspect the data, especially visually. Know the basics of the variables involved (descriptive information).
  - However, for MR this is not where you’ll test model assumptions

# Steps in Regression

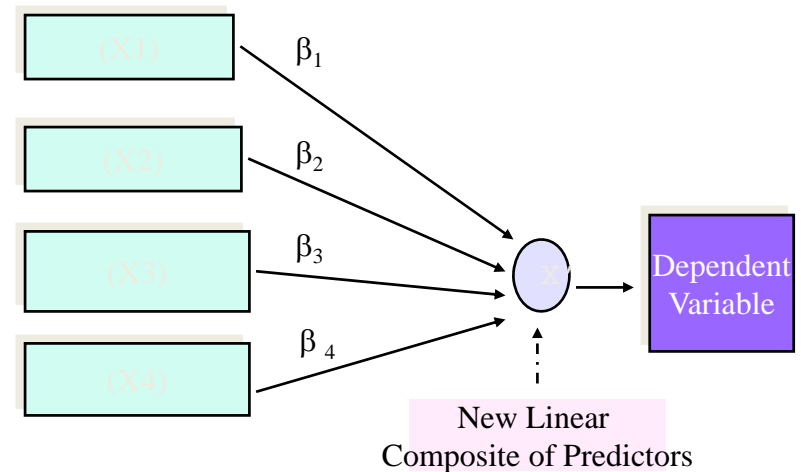
- Run the initial analysis to test assumptions graphically then statistically
  - Normality: bias, incorrect probability coverage
  - Homoscedasticity: model misspecification, inflated standard errors (inefficient estimates)
  - Independence: theoretical conclusions, inflated standard errors
  - Linearity: model misspecification
  - Lack of outliers: biased, inflated standard errors
  - Collinearity: inflated standard errors
- Do something about the problems
  - Compare initial output to robust regression
  - Use bootstrapped estimates
  - Create composites or drop variables with collinearity
  - Run more appropriate models
- Validate the model

# Considerations in multiple regression

- Type of regression
- Overall fit of model
- Assumptions
- Parameter estimates
- Variable importance
- Variable entry
- Relationships among the predictors
- Prediction

# Specifying the Models of Interest

- 'Simple' Multiple regression in a sense goes little beyond a correlation matrix (albeit a partial correlation matrix)
- It does not have interaction terms, indirect effects or multiple outcomes
- However within any collection of predictors are several models possible
- One must consider relations (potentially causal ones) among *all* the variables under consideration
  - Given 3 variables of which none are initially set as an outcome, there are possibly 20 or more models which might best represent the state of affairs
- Reliable measures must be used, or it will be all for naught
  - For some reason people try anyway



# Prediction vs. Explanation

- Doing a real regression involves much more than simply obtaining regression output from a statistical program
- First one must discern the goals of the regression, and there are two main kinds that are for the most part exclusive to one another in intent but do not have to be in practical application
- Prediction
  - With prediction there is much less if any concern over variable importance etc., and one is more concerned with how well (the model) works rather than how it works
    - Coefficients will be used to predict future data
  - Analytical example: Stepwise techniques
  - Applied example: prediction of graduate school success with incoming applicants
- Explanation
  - The goal here is to understand the individual relationships of variables with some outcome
    - Majority of psych usage falls almost entirely to this end
  - Analytical example: Sequential regression
  - Applied example: Personality factors involved in Aggression

# Prediction vs. Explanation

- There is a problem with going to far on either end of the spectrum
- While understandable in some, if not many research cases, putting all emphasis on prediction without understanding well the 'why and which' would be untenable for most psych research
- However, if one is entirely on the explanation side, one often finds a 'much ado about nothing' type of scenario
  - No point in saying which predictor is more important with a crappy model
- Modern approaches allow one to establish predictive validity much more easily, as well as provide better means to understand variable importance

Research Goals

---

Explanation

Prediction

# Reliability

- Collect reliable data
  - Reliability deals with measurement error (as opposed to sampling error or prediction error)
  - In a simple bivariate setting, less reliable measures underestimate the true relationship between the variables
  - In multiple regression, the relationships observed may vary wildly from the true relationship depending on the specifics
- In short, get good measures or suffer the consequences



# IED

- Initial examination of data includes:
- Getting basic univariate descriptive statistics for all variables of interest
- Graphical depiction of their distribution
- Assessment of simple bivariate correlations
- Inspection for missing data and determining how to deal with it<sup>1</sup>
- Start noting potential outliers, but realize that univariate outliers may not be model outliers
- The gist is you should know the data inside and out before the analysis is conducted<sup>2</sup>

# Running the Analysis

- You should know every aspect of an ANOVA table
- In the following the Model df is the number of predictors  $k$ , the Error df is  $N - k - 1$
- The  $F$  is a ratio of variance attributable to the model and that which is unaccounted for
- The square root of the  $MS_{\text{error}}$  is the standard error of estimate (aka residual standard error), a measure of predictive fit
- $R^2$  is the  $SS_{\text{Model}} / SS_{\text{Total}}$ , the amount of variance in the DV explained by the model, a measure of explanatory fit
- Often times you will simply see in-text reporting of  $F(\text{df}_1, \text{df}_2) = \text{value}$ ,  $MSE = \text{value}$ ; As you can see though, that is enough information to reconstruct the entire ANOVA table.

Source	SS	df	MS	F	$R^2$
Model	$\sum (\hat{Y} - \bar{Y})^2$	$k$	$SS_{\text{Model}} / df_{\text{Model}}$	$MS_{\text{Model}} / MS_{\text{Residual}}$	$SS_{\text{Model}} / SS_{\text{Total}}$
Residual	$\sum (\hat{Y} - Y)^2$	$N - k - 1$	$SS_{\text{Res}} / df_{\text{Residual}}$		
Total	$\sum (Y - \bar{Y})^2$	$N - 1$			



# Testing Assumptions

- When dealing with regression one must fit the model first to be able to test most of the assumptions, as they usually revolve around the residuals. These are the same assumptions as in simple regression though there are additional concerns.
- Assumptions and the problems violating them leads to
  - Normality: bias, incorrect probability coverage
  - Homoscedasticity: model misspecification, inflated standard errors (inefficient estimates)
  - Independence: theoretical conclusions, inflated standard errors
  - Linearity: model misspecification
  - Lack of outliers: biased, inflated standard errors
  - Collinearity: inflated standard errors
- The first approach must be graphical, glaring problems can easily be detected this way
  - This involves basic scatterplots, Density/QQ plots of residuals, residuals vs. fitted, influence plots, component-residual plots etc.
- Statistical inspection
  - Any normality test on the residuals (e.g. Shapiro-Wilks)
  - Breusch-Pagan test for heteroscedasticity
  - Durbin-Watson for autocorrelation<sup>1</sup>
  - RESET test for linearity
  - Many measures of outliers (Cook's distance, dfBetas, Mahalanobis' distance etc.)
  - Variance Inflation Factor for collinearity

# Dealing with Problems

- When violations of assumptions occur and/or outliers are present, steps must be taken to provide accurate model assessment
- This could include, transformations of variables, bootstrapped interval estimates, robust regression etc.
- Robust regression allows for comparisons to be made, and if notably different from classical approaches when there are data problems, typically to be preferred.

# Dealing with Bias: Validation

- Using the same data the model was fitted with to then predict leads to 'overfitting', i.e. model optimism
- Unless you have a very large sample relative to the number of predictors, assessments such as  $R^2$  are biased and often notably
- Modern approaches allow for easy validation when before sample sizes required were prohibitive

# Interpretation

- Model fit
  - $R^2$ 
    - Amount of the DV's variability is accounted for by the predictors
    - Statistical significance of the models are tested against a null of  $R^2 = 0$
    - More biased with more predictors and smaller N
    - Should always use a bias-adjusted  $R^2$  unless you have no interest in generalizing beyond your sample
  - Residual standard error (standard error of estimate)
    - Average error in prediction
    - Given the scale of the DV involved one should be able to get a sense of whether the RSE suggests a decent model
- Variable importance
  - Individually first
    - Raw coefficients tell you the same as they did in simple regression (except the effects of the other variables have been partialled out i.e. controlled for)
      - How much does the DV change with a one unit change in this predictor?
    - Statistical significance is a very poor measure for this and varies wildly with only slight changes in coefficients and their standard errors
  - Relative
    - How does the predictor perform relative to others?
- Interval estimates for  $R^2$  and coefficients give a sense of uncertainty in those estimates



# Example data

- Current salary predicted by educational level, previous experience, and minority status (N = 474)<sup>1</sup>
- As with any analysis, initial data analysis should be extensive prior to examination of the inferential analysis

# Initial examination of data

- We can use the graphics and descriptive statistics to give us a general feel for what's going on with the variables in question
- Start univariately, then proceed to bivariate relationships
- Univariately, we see:
  - An average salary of 34k but noticeable variability
  - Average education is beyond high school
  - Most have a few years of experience
  - The sample is overwhelmingly white (22% minority)
- Here we can also see that previous experience is not too well correlated with our dependent variable of current salary
  - Ack!

**Descriptive Statistics**

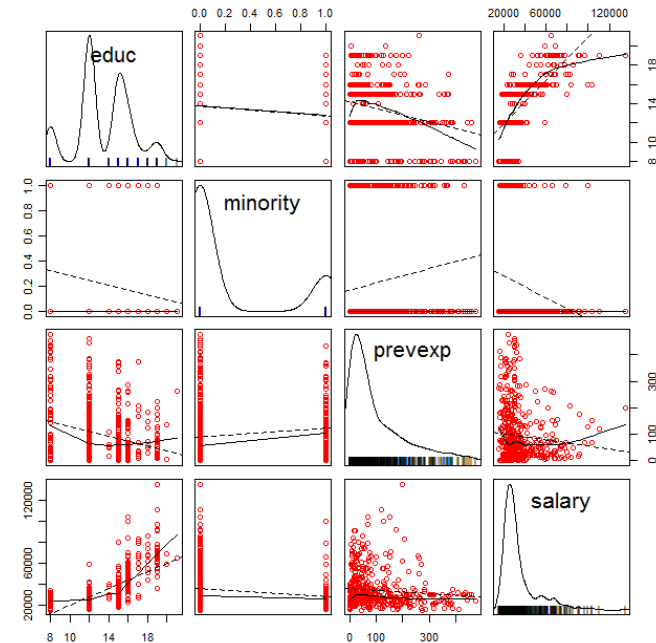
	Mean	Std. Deviation	N
Current Salary	\$34,419.57	\$17,075.661	474
Educational Level (years)	13.49	2.885	474
Previous Experience (months)	95.86	104.586	474
Minority Classification	.22	.414	474

**Correlations**

		Current Salary	Educational Level (years)	Previous Experience (months)	Minority Classification
Pearson Correlation	Current Salary	1.000	.661	-.097	-.177
	Educational Level (years)	.661	1.000	-.252	-.133
	Previous Experience (months)	-.097	-.252	1.000	.145
	Minority Classification	-.177	-.133	.145	1.000
Sig. (1-tailed)	Current Salary		.000	.017	.000
	Educational Level (years)	.000		.000	.002
	Previous Experience (months)	.017	.000		.001
	Minority Classification	.000	.002	.001	
N	Current Salary	474	474	474	474
	Educational Level (years)	474	474	474	474
	Previous Experience (months)	474	474	474	474
	Minority Classification	474	474	474	474

# Bivariate examination

- We'd also want to look at the scatterplots to further aid our assessment of the predictor-DV relationships
- There is noticeable skewness in the DV and the relationship between education and salary is clearly curvilinear
- Thus initial examination of the data already suggests misspecification of the model and likely violations of assumptions



# Starting point: Model Fit

- The multiple correlation coefficient is the correlation between the DV and the linear combination of predictors which minimizes the sum of the squared residuals
- More simply, it is the correlation between the observed values and the values that would be predicted by our model
- Its squared value ( $R^2$ ) is the amount of variance in the dependent variable accounted for by the independent variables, however it is biased
- Adjusted  $R^2$  would suggest initially decent model fit<sup>1</sup> with roughly 45% of the variance in the DV accounted for
- However the standard error of estimate suggests we are off on our predictions by over 12k typically
- The researcher must decide for themselves whether this would be acceptable
  - What does your gut tell you?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.672 <sup>a</sup>	.451	.448	\$12,688.728

a. Predictors: (Constant), Minority Classification, Educational Level (years), Previous Experience (months)

# Statistical significance of the model

- The ANOVA summary table tells us whether our model is statistically adequate
  - Is  $R^2$  different from zero
  - The regression equation is a better predictor than simply guessing the mean of the DV
- As with simple regression, the analysis involves the ratio of variance predicted to residual variance
- As we can see, it is reflective of the relationship of the predictors to the DV ( $R^2$ ), the number of predictors in the model, and sample size

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.224E10	3	2.075E10	128.868	.000 <sup>a</sup>
	Residual	7.567E10	470	1.610E8		
	Total	1.379E11	473			

a. Predictors: (Constant), Minority Classification, Educational Level (years), Previous Experience (months)

b. Dependent Variable: Current Salary

$$F = \frac{R^2(N - p - 1)}{(1 - R^2)p}$$

$$df = p, N - p - 1$$

# Variable Importance

- A first look at variable importance must entail looking at raw coefficients for each variable
- Below it's suggested that moving up one year of education would produce a \$4k increase in salary
- Previous experience suggests that a month more experience (and by extension a year) won't increase salary much, even though it's statistically significant
- Minorities suffer a drop of over \$4k in salary, but there is a lot of variability in that estimate

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-19424.234	3109.426		-6.247	.000					
	Educational Level (years)	3958.883	210.070	.669	18.846	.000	.661	.656	.644	.927	1.079
	Previous Experience (months)	14.028	5.804	.086	2.417	.016	-.097	.111	.083	.924	1.083
	Minority Classification	-4158.559	1430.535	-.101	-2.907	.004	-.177	-.133	-.099	.969	1.032

a. Dependent Variable: Current Salary

# Relative importance: Statistical significance

- Raw coefficients cannot clue us to variable importance with predictors of different scales
  - Moving 1 unit on education  $\neq$  moving 1 unit on race
- Statistical significance is largely useless also, in this case they're all statistically significant
  - And only slight changes in the coefficients and standard errors would produce large changes in p-values
- To begin with we can examine the output to determine which variables statistically significantly contribute to the model

# Relative Importance

- Standard metrics of relative importance include different themes on the predictor-DV correlation controlling for the effects of other predictors
- Standardized coefficients
- Partial correlation
- Semi-partial correlation

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-19424.234	3109.426		-6.247	.000					
	Educational Level (years)	3958.883	210.070	.669	18.846	.000	.661	.656	.644	.927	1.079
	Previous Experience (months)	14.028	5.804	.086	2.417	.016	-.097	.111	.083	.924	1.083
	Minority Classification	-4158.559	1430.535	-.101	-2.907	.004	-.177	-.133	-.099	.969	1.032

a. Dependent Variable: Current Salary

# Relative Importance

- Standardized regression coefficients get around that problem
- Now we can see how much the DV will change in standard deviation units with one standard deviation unit change in the predictor (*all others held constant*)
- Not very meaningful if you don't remember what the standard deviation of your variables are (hint, hint)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-19424.234	3109.426		-6.247	.000					
	Educational Level (years)	3958.883	210.070	.669	18.846	.000	.661	.656	.644	.927	1.079
	Previous Experience (months)	14.028	5.804	.086	2.417	.016	-.097	.111	.083	.924	1.083
	Minority Classification	-4158.559	1430.535	-.101	-2.907	.004	-.177	-.133	-.099	.969	1.032

a. Dependent Variable: Current Salary

# Relative Importance

- Standardized regression coefficients get around that problem
- Now we can see how much the DV will change in standard deviation units with one standard deviation unit change in the predictor (*all others held constant*)
- Not very meaningful if you don't remember what the standard deviation of your variables are (hint, hint)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-19424.234	3109.426		-6.247	.000					
	Educational Level (years)	3958.883	210.070	.669	18.846	.000	.661	.656	.644	.927	1.079
	Previous Experience (months)	14.028	5.804	.086	2.417	.016	-.097	.111	.083	.924	1.083
	Minority Classification	-4158.559	1430.535	-.101	-2.907	.004	-.177	-.133	-.099	.969	1.032

a. Dependent Variable: Current Salary

# Relative Importance

- However we still have other output to help us understand variable contribution
- Partial correlation is the contribution of a predictor after the contributions of the other predictors have been taken out of *both* the predictor and DV
  - When squared, it is a measure of variance *left over* that can be accounted for uniquely by the variable
- Semi-partial correlation is the unique contribution of an predictor after the contribution of other predictors have been taken *only* out of the predictor in question
  - When squared, it is a measure of that part of the *total* variance that can be accounted for uniquely by a specific variable
- Neither Previous Experience nor Minority status account for even an  $R^2$  of 1% after partialling out the effects of Education

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-19424.234	3109.426		-6.247	.000					
	Educational Level (years)	3958.883	210.070	.669	18.846	.000	.661	.656	.644	.927	1.079
	Previous Experience (months)	14.028	5.804	.086	2.417	.016	-.097	.111	.083	.924	1.083
	Minority Classification	-4158.559	1430.535	-.101	-2.907	.004	-.177	-.133	-.099	.969	1.032

a. Dependent Variable: Current Salary

# Relative Importance

- A better metric
- Averaging the semi-partial correlation over all possible models shows that Education contributes over 94% of the  $R^2$  value, with Minority taking up most of the rest
- Statistical testing shows education's contribution to be statistically greater than the other two, which are not distinct from one another

Relative importance metrics:

	lmg
educ	0.94053492
minority	0.04357798
prevexp	0.01588710

Differences between Relative Contributions:

	difference		Lower	Upper
	0.95			
educ-minority.lmg	0.8970	*	0.8214	0.9590
educ-prevexp.lmg	0.9246	*	0.8775	0.9593
minority-prevexp.lmg	0.0277		-0.0108	0.0662



# Relative Importance Summary

- There are multiple ways to estimate a variable's contribution to the model, and some may be better than others
- A general approach:
- Check simple bivariate relationships
  - If you don't see worthwhile correlations with the DV there you shouldn't expect much from your results regarding the model\*
    - Check for outliers and compare with robust measures also
  - You may detect that some variables are so highly correlated that one is redundant
- Statistical significance is not a useful means of assessing relative importance, nor is the raw coefficient typically
- Standardized coefficients and partial correlations are a first step
  - Compare standardized to simple correlations as a check on possible suppression
- Of typical output the semi-partial correlation is probably the more intuitive assessment
- The LMG is also intuitive, and is a natural decomposition of  $R^2$ , unlike the others

# Relative Importance Summary

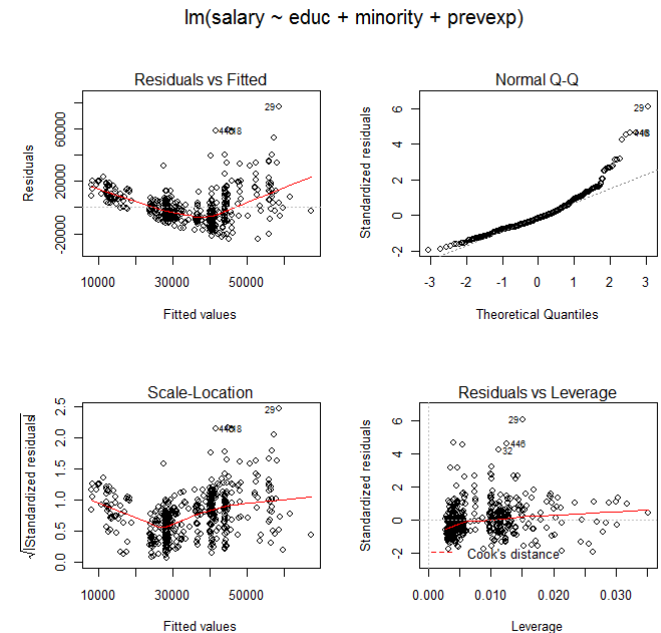
- One thing to keep in mind is that determining variable importance, while possible for a single sample, should not be overgeneralized
- Variable orderings *likely* will change upon repeated sampling
  - E.g. while one might think that war and bush are better than math (it certainly makes theoretical sense), saying that either would be better than the other would be quite a stretch with just one sample
- What you see in your sample is specific to it, and it would be wise to not make any bold claims without validation

# Hold the Phone

- All of the preceding doesn't matter
- Why?
- You didn't test the assumptions of the model

# Testing Model Adequacy

- Numerical tests tell us what is already made clear by the graphs
- The model violates normality, homoscedasticity and linearity assumptions along with having a couple outliers
- In short, drastic measures would need to take place to better capture the true nature of the relationships



Shapiro-Wilk normality test

```
data: RegModel.1$resid  
W = 0.8868, p-value < 2.2e-16
```

Breusch-Pagan test

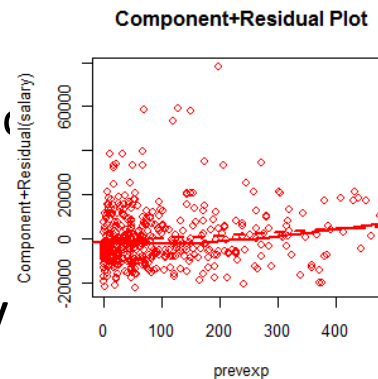
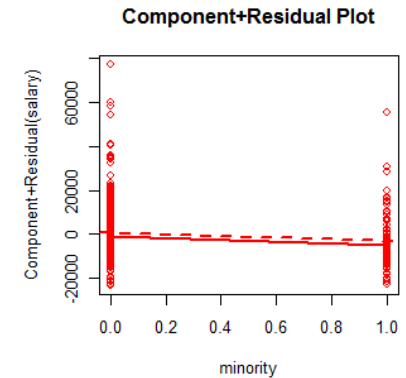
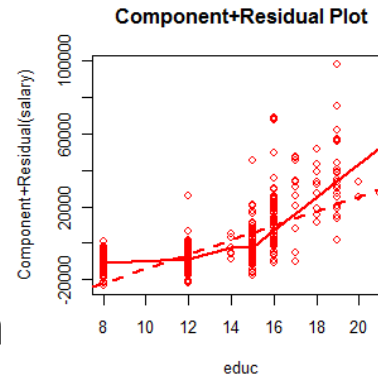
```
data: RegModel.1  
BP = 34.1522, df = 3, p-value = 1.84e-07
```

RESET test

```
data: salary ~ educ + minority + prevexp  
RESET = 27.357, df1 = 6, df2 = 464, p-value < 2.2e-16
```

# Testing Model Adequacy

- Plots examining each predictor relationship with the DV (still controlling for the others) suggest there is more to the story
- That the relationship between education and salary is curvilinear
  - In fact just a polynomial regression with educ (quadratic) would increase  $R^2$  to .59
- The relationship with Minority may be distorted by outliers



# Summary

- Even simple regression entails quite a bit to make sense of it all
- The first step is having a good model with good measures of the constructs of interest
- Divide examination into model and variable inspection, but only after testing model adequacy and even when ok, do not ignore the natural bias present in fitting the model to data that produced the model
  - Follow that up with validation
- Model fit entails measures such as adj  $R^2$  and the residual standard error
- Variable importance must first be measured in raw units, followed by appropriate order determination if one chooses to make inferences beyond the sample at hand

# Summary

- At this point there are a few key ideas regarding the regression analysis to note:
- What's being done?
  - Least squares approach
  - General linear model
  - Linear combination
- Model Fit
  - $R^2$
  - Error in prediction
- Variable Importance
  - Individual
  - Relative
- Validation of the model