

# Model Adequacy and Other Concerns

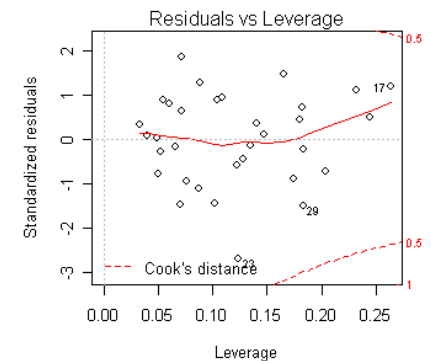
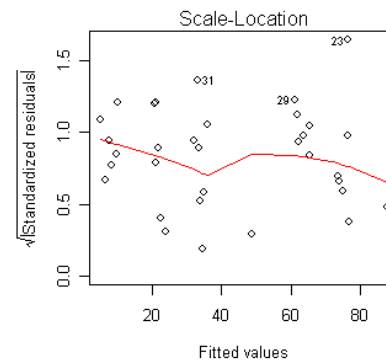
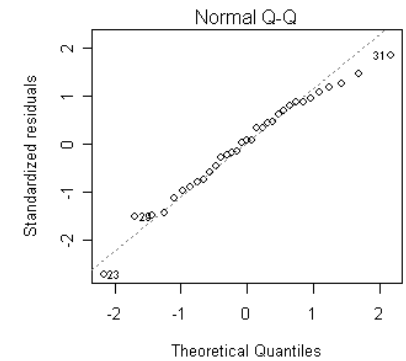
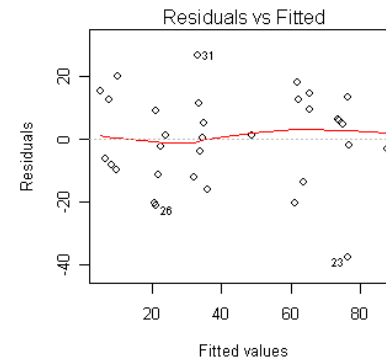
Identifying and Dealing with Problems

# Unmet Assumptions and Other Problems

- Normality
- Homoscedasticity
- Linearity
- Other issues
  - Collinearity
  - Outliers
  - Suppression
  - Prediction
    - Overfitting
    - Uncertainty: Interval estimates
  - Causality

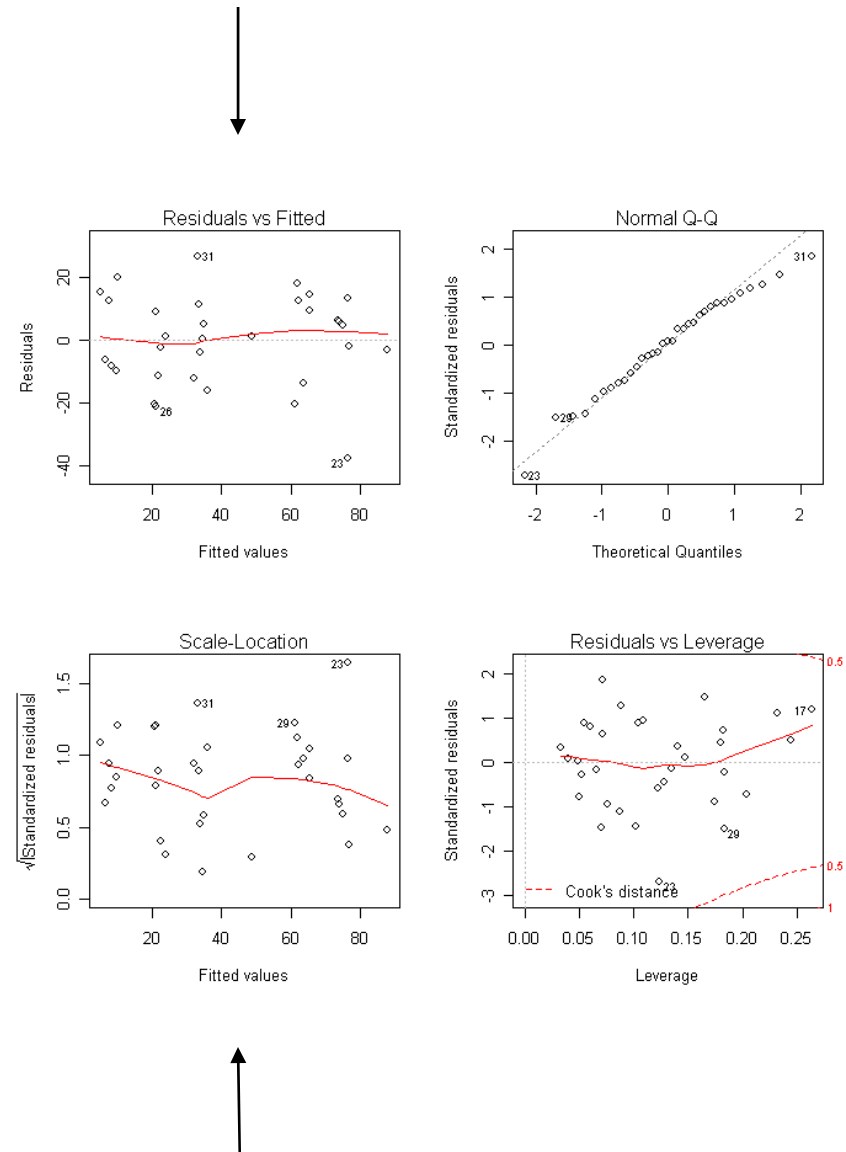
# Normal distribution of residuals

- Our normality assumption applies to the residuals
- One can simply save them and plot a density curve/histogram
- Often a quantile-quantile plot is readily available, and here we hope to find most of our data along a 45 degree line
- There are many statistical tests for normality, among the common ones are Shapiro-Wilks and Kolmogorov-Smirnov
  - Reject  $H_0 =$  unmet assumption



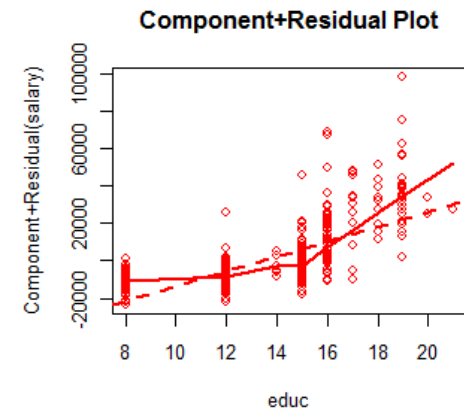
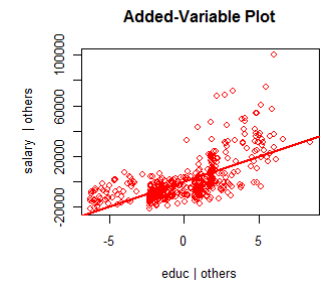
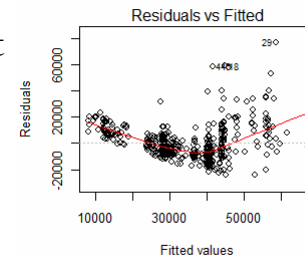
# Homoscedasticity

- We can check a plot of the residuals vs our predicted values to get a sense of the spread along the regression line
- We prefer to see kind of a blob about the zero line (our mean), with no readily discernable pattern
- This would mean that the residuals don't get overly large for certain areas of the regression line relative to others
- A common statistical test for this assumption is the Breusch-Pagan test
  - Again  $\text{Reject } H_0 = \text{unmet assumption}$



# Linearity

- A lack of linearity can be detected with the same plots sometimes, since those previous assumptions not being met may be due to a curvilinear relationship
- However there are plots to help us detect this better for specific variables
- Added-variable plots
  - Compute the residuals of regressing the response variable against all predictors except the one of interest
  - Compute the residuals from regressing the predictor in question against the remaining predictor variables
  - Plot the residuals from (1) against the residuals from (2)
  - Plot of unique variance in DV left vs. the predictor's unique variance
    - Good for detecting outliers
- Component + residual plots
  - An attempt to understand the relationship between the potentially nonlinear components in the model and predictor in question
    - Not so good for outliers, but typically better for detecting curvilinearity
- A statistical test would be the RESET test, which compares your model with those that add polynomial terms



# Collinearity

- Collinearity is simply correlation among predictors
- If it is too high it can lead to large standard errors for our coefficients
  - In terms of theory, this makes replication of variable importance findings/conclusions even less likely to hold up (remember that sampling variability already could produce notable order changes)
- However this is not restricted to simple bivariate relationships but regards more the  $R^2$  of this model for each predictor variable

$$\text{Pred}_1 = b_0 + b_1\text{Pred}_2 + b_2\text{Pred}_3\dots$$

- The larger the  $R^2$  the more that variable is redundant since its variance is already accounted for by the other predictors

# Collinearity diagnostics

- VIF
  - Variance inflation factor: how much will the standard error increase as a result of collinearity
  - Looking for VIF values that are large
    - E.g. individual VIF greater than 10 should be inspected
  - $VIF = 1 / \text{tolerance}$
- Tolerance
  - Proportion of a predictors' variance not accounted for by other variables
    - $1 - R^2$  for the model just indicated
  - Looking for tolerance values that are small, close to zero
    - Means they are not contributing anything new to the model
  - $\text{tolerance} = 1 / VIF$
- Other Indicators of Collinearity
  - Eigenvalues
    - Small values, close to zero
  - Condition index
    - Large values (15+)
- Solutions: create composite variables, drop from the model

# Outliers

- Initial results would be relatively useless if we are not meeting our assumptions and/or have overly influential data points
  - In fact, you shouldn't be really looking at the results unless you test assumptions and look for outliers, even though this requires running the analysis to begin with
- Various tools are available for the detection of outliers
- Ways to think about outliers
  - Leverage
  - Discrepancy
  - Influence
- Common methods
  - Standardized Residuals (ZRESID)
  - Studentized Residuals (SRESID)
  - Studentized Deleted Residuals (SDRESID)
- Thinking 'robustly'

# Outliers

- *Leverage* assesses outliers among the predictors (unusual profile)
  - Mahalanobis distance
    - Relatively high Mahalanobis suggests an outlier on one or more variables
  - Hat values
- *Discrepancy*
  - Measures the extent to which a case is in line with others
- *Influence*
  - A product of leverage and discrepancy
  - How much would the coefficients change if the case were deleted?
    - Cook's distance, dfBetas

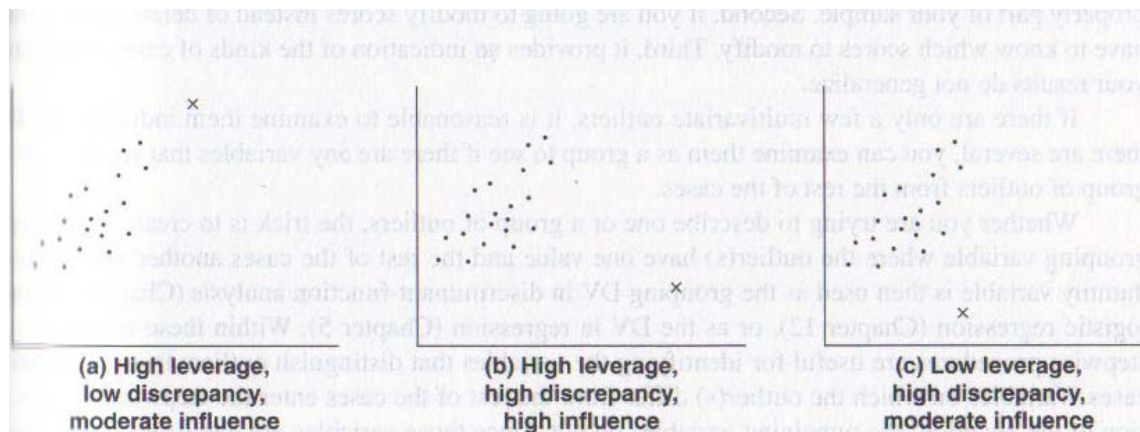


FIGURE 4.2 The relationships among leverage, discrepancy, and influence.

# Outliers

- Standardized Residuals (ZRESID)
  - Standardized errors in prediction
    - Mean 0, Sd = std. error of estimate
    - To standardize, divide each residual by its s.e.e.
  - At best an initial indicator (e.g. the  $\pm 2$  rule of thumb), but because the case itself determines what the variance would be, almost useless
- Studentized Residuals (SRESID)
  - Same thing but studentized residual recognizes that the error associated with predicting values far from the mean of X is larger than the error associated with predicting values closer to the mean of X
  - standard error is multiplied by a value that will allow the result to take this into account
- Studentized Deleted Residuals (SDRESID)
  - Studentized in which the standard error is calculated with the case in question removed from the others

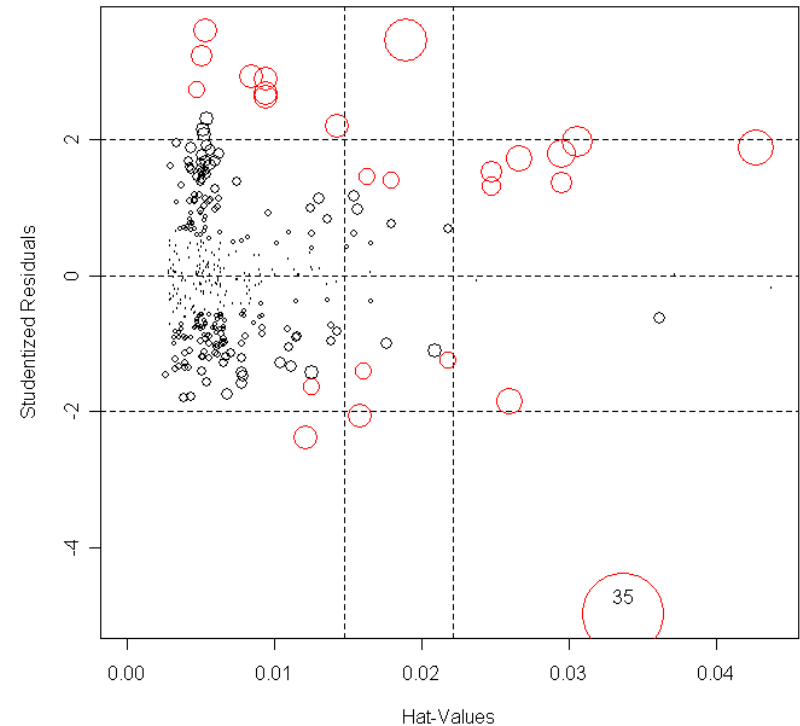
$$S_{e_i} = S_{Y.X} \sqrt{1 - \left[ \frac{1}{N} + \frac{(X - \bar{X})^2}{\sum x^2} \right]}$$

# Outliers

- Mahalanobis' Distance
  - Mahalanobis distance is the distance of a case from the centroid of the remaining points (point where the means meet in n-dimensional space)
- Cook's Distance
  - Identifies an influential data point whether in terms of predictor or DV
  - A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients.
  - With larger (relative) values, excluding a case would change the coefficients substantially.
- DfBeta
  - Change in the regression coefficient that results from the exclusion of a particular case
  - Note that you get DfBetas for each coefficient associated with the predictors
- Hat values
  - A measure of influence that each observed value has on each fitted value

# Outliers

- Influence plots
- With a couple measures of ‘outlierness’ we can construct a scatterplot to note especially problematic cases
  - After fitting a regression model in R-commander, i.e. running the analysis, this graph is available via point and click
  - Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the studentized-residual scale. Red are the cases with the most ‘noteworthy’ Cook’s distance.
- Here we have what is actually a 3-d plot, with 2 outlier measures on the x and y axes (studentized residuals and ‘hat’ values, a measure of leverage) and a third in terms of the size of the circle (Cook’s distance)
- For this example, case 35 appears to be a problem



# Outliers

- No matter the analysis, some cases will be the ‘most extreme’. However, none may really qualify as being overly influential.
- Whatever you do, always run some diagnostic analysis and do not ignore influential cases
- Assessing outliers, while there are rules of thumb (e.g. Cook’s  $D > 4/df_{res}$ ), determination is a *relative* notion specific to the data at hand
- It should be clear to interested readers whatever has been done to deal with outliers
- As noted before, the best approach to dealing with outliers when they do occur is to run a robust regression with capable software and compare results
  - However, outliers are your first indication the model is possibly misspecified
  - In short, your model is not adequate to capture those cases effectively

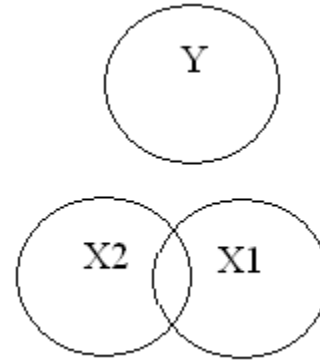
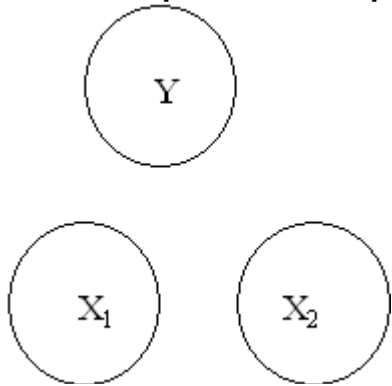
# Suppressor variables

- There are a couple of ways in which suppression can occur or be talked of, but the gist is that suppression masks the impact the predictor would have on the outcome if the third variable was not in the model
- In general suppression occurs when  $\beta_i$  falls outside the range of  $0 \rightarrow r_{yi}$
- Suppression in MR can entail some different relationships among predictors
  - For example one suppressor relationship would be where two variables,  $X_1$  and  $X_2$ , are positively related to  $Y$ , but when the equation comes out we get
    - $\hat{Y} = b_1X_1 - b_2X_2 + a$
- There is more than one kind of suppression
  - Classical
  - Net
  - Cooperative
- However to understand it, it is best to visualize the type of relationships that might occur in the simple setting

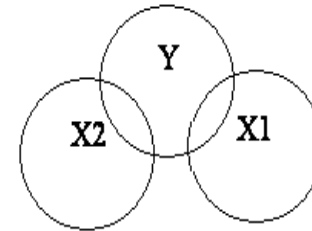
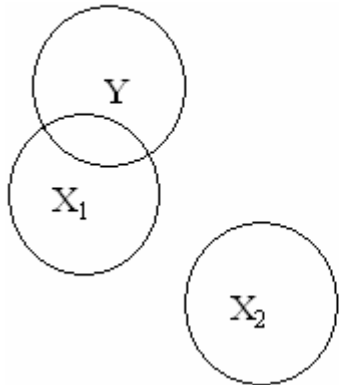
# Suppression

- Consider the following relationships

a. Complete independence:  $R^2_{Y.12} = 0$       b. Partial independence:  $R^2_{Y.12} = 0$  but  $r_{12} \neq 0$ ,



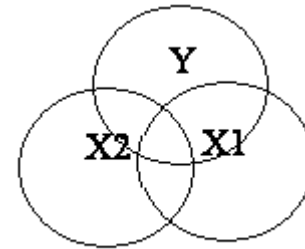
c. Partial independence:  $r_{12} = 0$ ,  $r_{Y2} = 0$ ,  $r_{Y1} \neq 0$       d. Partial independence again, both  $r_{Y1}$  and  $r_{Y2} \neq 0$ , but  $r_{12} = 0$



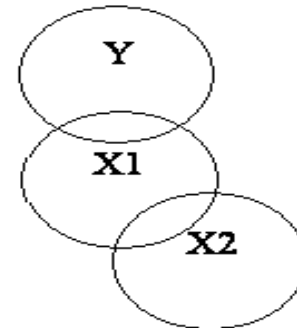
# Suppression

e. Normal situation, redundancy: no simple correlation = 0

- Each semi-partial correlation, and the corresponding beta, will be less than the simple correlation between  $X_i$  and  $Y$ . This is because the variables share variance and influence



f. Classical suppression:  $r_{Y2} = 0$



# Suppression: Technical side

- When dealing with standardized regression coefficients, note that

$$\beta_{Y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

$$\beta_{Y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}$$

*such that*

$$\hat{z}_y = \beta_{Y1.2}z_{X1} + \beta_{Y2.1}z_{X2}$$

# Suppression: Technical side

- Recall from previously →

$$\beta_{Y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

- If  $r_{y2} = 0$ , then →

$$\beta_{Y1.2} = \frac{r_{y1}}{1 - r_{12}^2}$$

- With increasingly shared variance between  $X_1$  and  $X_2$  we will have an inflated beta coefficient for  $X_1$

$$R^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}$$

- $X_2$  is *suppressing the error variance* in  $X_1$
- In other words, even though  $X_2$  is not correlated with  $Y$ , having it in the equation raises the  $R^2$  from what it would have been with just  $X_1$ .

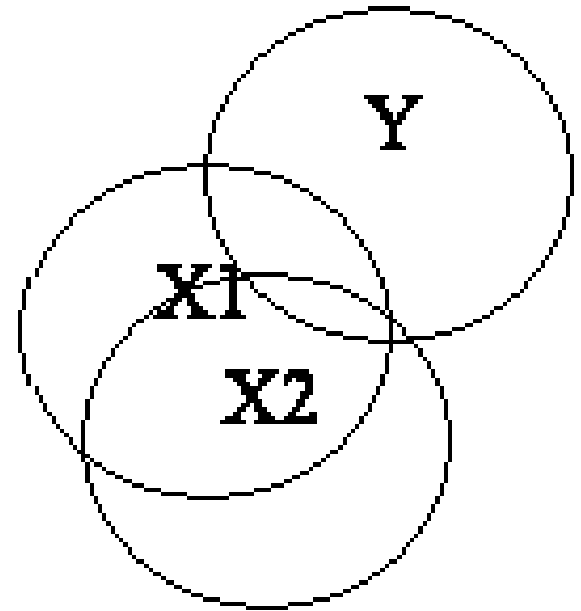
or

$$R_{y.12}^2 = \beta_{Y1.2}r_{y1} + \beta_{y2.1}r_{y2}$$

# Suppression

- Other suppression situations
- Net
  - All  $r$ s positive
  - $\beta_2$  ends up with a sign *opposite* that of its simple correlation with Y
  - It is always the X which has the smaller  $r_{yi}$  which ends up with a  $\beta$  of opposite sign
  - $\beta$  falls outside of the range  $0 \rightarrow r_{yi}$ , which is always true with any sort of suppression
- Cooperative
  - Predictors negatively correlated with one another, both positive with DV
    - Or positively with one another and negatively with Y
- Example of Classical<sup>1</sup>:
  - Cor X1-Y = .5
  - Cor X1-X2 = .6
  - Cor X2-Y = .05
  - Regression output (standardized coefficients)  
Coefficients:

( Intercept )	X1	X2
-9.506e-18	.7344	-.3906
  - Note the increase in both predictors relative to their simple correlation with Y
- Example of Cooperative:
  - Correlation between social aggressiveness ( $X_1$ ) and sales success (Y) = .29
  - Correlation between record keeping ( $X_2$ ) and sales success (Y) = .24
  - $r_{12} = -.30$
  - Regression coefficients for predictors = .398 and .359 respectively



# Suppression

- Gist: weird stuff can happen in MR, so take note of the relationship of the predictors and how it may affect your overall interpretation
- Compare the simple correlations of each predictor with the DV and compare to their respective beta coefficients<sup>1</sup>
  - If coefficient noticeably larger than simple correlation (absolute value) or of opposite sign one should suspect possible suppression
- This would probably indicate model misspecification (e.g. in the classical case you are retaining a predictor that has no influence on the outcome)

# Overfitting

- *Every* initial fit of a model is overfitted, as it is fit to the data that gave rise to the model itself
- Thus the issue with overfitting is one of external validity
- In some cases, some of the variation the parameters chosen are explaining is variation that is idiosyncratic to the sample
  - For example, the bias of an overestimated  $R^2$  leads to ‘shrinkage’ with new model
  - We would not see this variability in the population
- So the fit of the model is good, but it doesn’t generalize as well as one would think
- Capitalization on chance

# Overfitting Example

- Example from Lattin, Carroll, Green
- Randomly generated 30 variables to predict an outcome variable
- Using a best subsets approach, 3 variables were found that produce an  $R^2$  of .33 or 33% variance accounted for before bias-correction
- As one can see, even random data has the capability of appearing to be a decent fit

**TABLE 3.11** Results from best fitting regression using three out of 30 randomly generated variables

	Coefficient	Standard Error	<i>t</i>	<i>p</i>	
Intercept	0.1602	0.1274	1.26	0.215	
$X_1$	-0.3874	0.1324	-2.93	0.005	
$X_9$	0.5029	0.1319	3.81	0.000	
$X_{14}$	-0.2378	0.1177	-2.02	0.049	
$s = 0.8753$	$R^2 = 32.9\%$	$\bar{R}^2 = 28.6\%$			
	SS	<i>df</i>	MS=SS/ <i>df</i>	<i>F</i>	<i>p</i>
Regression	17.3160	3	5.7720	7.53	0.000
Error	35.2397	46	0.7661		
Total	52.5557	49			

# Validation

- The old way to deal with this issue is with a simple random split
- With large datasets one can randomly split the sample into two sets
  - Calibration sample: used to estimate the coefficients
  - Holdout sample: used to validate the model
- Typical suggestions required a 2:1 or 4:1 split and thus large total samples for the holdout sample to be viable
- How it works: Using the coefficients from the calibration set one can create predicted values for the holdout set
  - i.e. apply the model to the other data
- The squared correlation between the predicted values and observed values can then be compared to the  $R^2$  of the calibration set
- In previous example of randomly generated data the  $R^2$  for the holdout set was 0

# Other approaches

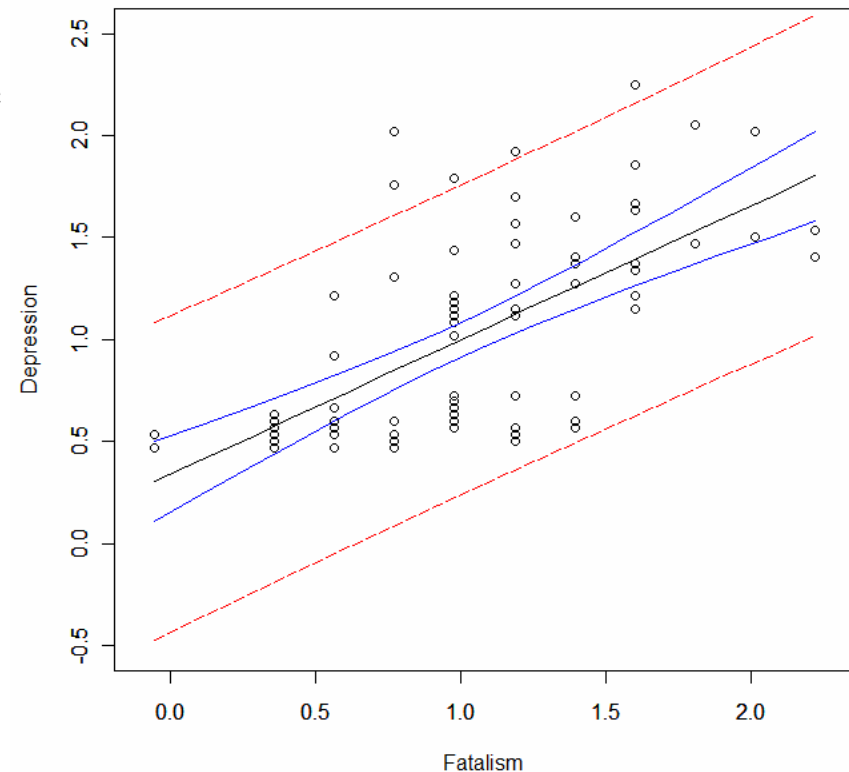
- K-fold cross validation
  - Create several (K) samples of the data of roughly equal size
  - Use the holdout approach with one sample, and obtain estimates (coefficients) from the others
  - Do this for each sample, obtain average estimates
- Jackknife Validation
  - Create estimates with a particular case removed
  - Use the coefficients obtained from analysis of the n-1 remaining cases to create a predicted value for the removed case
  - Do for all cases, and then compare the jackknifed  $R^2$  to the original
- Bootstrap Validation
  - With relatively smaller samples<sup>1</sup>, cross-validation may not be as feasible
  - One may instead resample (with replacement) from the original data to obtain estimates for the coefficients
  - Fit each bootstrap model to the original data and average

# Interval estimates

- Intervals of the coefficients, as with other statistics, give us a sense of uncertainty with our guess as to the true parameter in the population, and also provide a means to conduct a hypothesis test for that coefficient
- One may easily obtain bootstrap intervals for coefficients, which as before can serve as the hypothesis test of the coefficient vs. zero
  - Does the interval contain zero or not?

# Prediction

- They can also be represented graphically, but it is necessary to distinguish the confidence interval for our regression line from the *prediction interval* for the outcome variable given a particular predictor score
- Estimating the outcome given a specific value of the predictor
- In this manner we could
  - 1. Attempt to predict the mean of the outcome for that given value or...
  - 2. Predict the value in terms of the individual
- The wide intervals are interval estimates for individual scores
  - Given a score of 1 on Fatalism, what would we expect that person's score for Depression to be?
- The narrow ones are confidence intervals for the slope
  - What is the mean Depression score for people at 1 on the Fatalism scale?
- We are basically getting a sense of the regression at different values along the range of the predictor



# R<sup>2</sup>

- It is also important to note the boundaries of how good our model fit is
  - Interval estimates on effect sizes are something specifically noted by the APA taskforce as a step in the right direction, and they are easily obtained through formal methods or nonparametric (i.e. bootstrap)
  - Studies seem to be indicating better performance for the bootstrapped version
- Example from the MBESS package
  - Insert your own R<sup>2</sup>, sample size, number of predictors and desired confidence level
  - `ci.R2(R2=.42, N=82, K=1, conf.level=.95)`
  - 95% CI = (.25,.57)
- CIs for effect sizes are usually very wide except for very large effects with large samples, indicating the true effect is a hard thing to be sure about with just a single sample

# Causality

- Just because we have a significant F and decent  $R^2$ , we can't assume there's a causal relationship between the two variables
- Correlational relationships are not necessarily causal, but the regression model suggests the possibility of a causal relationship
  - i.e. the predictive arrow is flows from one variable to the outcome according to theory, previous research, logic etc.
  - If it seems odd to even remotely think causally given your specific research situation, it's probably a good bet your model is misspecified, and important factors are being left out.
- Many of the issues previously mentioned specifically deal with an implied structural model that suggests an *effect* of predictors on the outcome

# Summary of how to do a real regression analysis

- 1. Have an idea, grounded in reality/common sense/previous research
- 2. Propose a theoretical (possibly causal) model in which you have thought about other viable models (including how predictors might predict one another, moderating and mediating possibilities etc.)
  - Now choose *at least* one other plausible idea
- 3. Use reliable measures
- 4. Collect appropriate and enough data
- 5. Spend time with initial examination of data including obtaining a healthy understanding of the variables descriptively, missing values analysis if necessary, inspection of correlations etc.
- 6. Run the analysis. Might as well ignore for now.
- 7. With the model in place, test assumptions, look for collinearity, identify outliers. Take appropriate steps necessary to deal with any issues including bootstrapped regression or robust regression
- 8. Validate the model. Note any bias. Examine graphical displays of fit.
- 9. Interpret results. Focus on bias corrected estimates of  $R^2$ , interval estimates of coefficients and  $R^2$ , interpretable measures of variable importance (test for differences among them)