Outline

Many possible models

Table: Sixteen Possible Models

$E\, y = \beta_0$			1 model with no variables
$E\, y = \beta_0 + \beta_1 x_i$,	$i =$	1, 2, 3, 4	4 models with one variable
$E\, y = \beta_0 + \beta_1 x_i + \beta_2 x_j$,	$(i, j) =$	(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)	6 models with two variables
$E\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_j + \beta_3 x_k$,	$(i, j, k) =$	(1,2,3), (1,2,4), (1,3,4), (2,3,4)	4 models with three variables
$E\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$			1 model with all variables

- With k explanatory variables, there are 2^k possible linear models
- There are infinitely many nonlinear ones!!

Stepwise Regression Algorithm

Suppose that the analyst has identified one variable as the response, y, and k potential explanatory variables, $x_1, x_2, \ldots, x_k$.

- (i). Consider all possible regressions using one explanatory variable. For each of the k regressions, compute $t(b_1)$, the t-ratio for the slope. Choose that variable with the largest t-ratio. If the t-ratio does not exceed a pre-specified t-value (such as two), then do not choose any variables and halt the procedure.
- (ii). Add a variable to the model from the previous step. The variable to enter is the one that makes the largest significant contribution. To determine the size of contribution, use the absolute value of the variable's t-ratio. To enter, the t-ratio must exceed a specified t-value in absolute value.
- (iii). Delete a variable to the model from the previous step. The variable to be removed is the one that makes the smallest contribution. To determine the size of contribution, use the absolute value of the variable's t-ratio. To be removed, the t-ratio must be less than a specified t-value in absolute value.
- (iv). Repeat steps (ii) and (iii) until all possible additions and deletions are performed.

Drawbacks of Stepwise Regression

- The procedure "snoops" through a large number of models and may fit the data "too well."
- There is no guarantee that the selected model is the best.
  - The algorithm does not consider models that are based on nonlinear combinations of explanatory variables.
  - It also ignores the presence of outliers and high leverage points.
  - The algorithm does not even search all 2^k possible linear regressions.
- The algorithm uses one criterion, a t-ratio, and does not consider other criteria such as s, R^2, R, and so on.
- There is a sequence of significance tests involved. Thus, the significance level that determines the t-value is not meaningful.
- By considering each variable separately, the algorithm does not take into account the joint effect of independent variables.
- Purely automatic procedures may not take into account an investigator's special knowledge.

Variants of Stepwise Regression

- Forward selection. Add one variable at a time without trying to delete variables.
- Backwards selection. Start with the full model and delete one variable at a time without trying to add variables.
- Best regressions.

Residuals

- Standardized residual = residual divided by it's estimated standard error.
- Use standardized residuals because:
  - we can focus on relationships of interest
  - achieve carry-over of experience from one data set to another.
- Using $e_i = y_i - \hat{y}_i$ as the ith residual, here are three commonly used definitions:

$$\text{(a) } \frac{e_i}{s}, \quad \text{(b) } \frac{e_i}{s\sqrt{1-h_{ii}}}, \quad \text{(c) } \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}.$$

  - First choice is simple
  - Second choice, from theory, $\text{Var}(e_i) = \sigma^2(1-h_{ii})$. Here, h_{ii} is the ith leverage. It is calculated based on values of the explanatory variables and will be defined in Section 5.4.1.
  - Third choice is termed "studentized residuals". Idea: numerator is independent of the denominator.

Residual Analysis

- Role of residuals: If the model formulation is correct, then residuals $\approx$ random errors.
- Four types of Patterns:
  - Unusually large residuals
  - Residuals related to explanatory variables
  - Heteroscedastic residuals (Section 5.7)
  - Time patterns in residuals (start in Chapter 7)
- Method of attack: Look for patterns in the residuals. Use this information to improve the model specification.

Outliers

- An outlier is an observation that is not well fit by the model; these are observations where the residual is unusually large.
  - Unusual means what? Many packages mark a point if the |standardized residual| > 2.
- Options for handling outliers
  - Ignore them in the analysis but be sure to discuss their effects.
  - Delete them from the data set (but be sure to discuss their effects).
  - Create a binary variable to indicator their presence. (This will increase your R^2!)

Using Residuals to Select Explanatory Variables

- Residual analysis can help identify additional explanatory variables that may be used to improve the formulation of the model.
  - If the model is correct, then residuals should resemble random errors and contain no discernible patterns.
  - Thus, when comparing residuals to explanatory variables, we do not expect any relationships.
  - If we do detect a relationship, then this suggests the need to control for this additional variable.
- Ways of detecting relationships.
  - Calculate summary statistics and display the distribution of (standardized) residuals to identify outliers.
  - Calculate the correlation between the (standardized) residuals and additional explanatory variables to search for linear relationships.
  - Create scatter plots between the (standardized) residuals and additional explanatory variables to search for nonlinear relationships.

Example. Stock Liquidity

- Investors want high expected returns and low volatility. Sometimes, they also want to receive income from investments periodically (dividends).
- Investors also wish to be able to sell investments quickly, known as "liquidity."
- This study has $n = 126$ companies.
- For the trading activity variables, we examine
  - the three months total trading volume (VOLUME, in millions of shares),
  - the three months total number of transactions (NTRAN), and
  - the average time between transactions (AVGT, measured in minutes).
- For the firm size variables, we use the
  - opening stock price on January 2, 1985 (PRICE),
  - the number of outstanding shares on December 31, 1984 (SHARE, in millions of shares), and
  - the market equity value (VALUE, in billions of dollars) obtained by taking the product of PRICE and SHARE.

Example. Stock Liquidity

Table: Summary Statistics of the Stock Liquidity Variables

	Mean	Median	Standard deviation	Minimum	Maximum
VOLUME	13.423	11.556	10.632	0.658	64.572
AVGT	5.441	4.284	3.853	0.590	20.772
NTRAN	6436	5071	5310	999	36420
PRICE	38.80	34.37	21.37	9.12	122.37
SHARE	94.7	53.8	115.1	6.7	783.1
VALUE	4.116	2.065	8.157	0.115	75.437
DEB_EQ	2.697	1.105	6.509	0.185	53.628

Source: Francis Emory Fitch, Inc., Standard & Poor's Compustat, and University of Chicago's Center for Research on Security Prices.

Example. Stock Liquidity

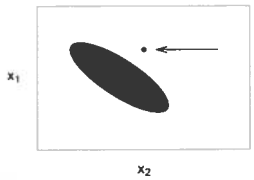
Table: Correlation Matrix of the Stock Liquidity

	AVGT	NTRAN	PRICE	SHARE	VALUE	DEB_EQ
NTRAN	−0.668					
PRICE	−0.128	0.190				
SHARE	−0.429	0.817	0.177			
VALUE	−0.318	0.760	0.457	0.829		
DEB_EQ	0.094	−0.092	−0.038	−0.077	−0.077	
VOLUME	−0.674	0.913	0.168	0.773	0.702	−0.052

- Liquidity variables, VOLUME, NTRAN and AVGT, are highly correlated

Influential Points

- Influential points are observations that potentially have a lot to say about a regression fit. They may have large residuals, high leverage, or both.
- We have seen that regression coefficients are weighted sums - not all observations are equal!
  - Intuitively, observations that are "far away" in the x-space have greater influence.
  - One can get a feel for influential observations by looking a summary statistics (mins, maxs) for each explanatory variable.
  - However, this is not sufficient - see the graph below.
- The ellipsoid represents most of the data. The arrow marks an unusual point that is not unusual for x_1 or x_2.



High Leverage Points

- An observation is a high leverage point if the value of h_{ii} is large.
  - By algebra, the average leverage is $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k+1}{n}$.
  - Some packages declare an observation to be a *high leverage point* if the leverage exceeds three times the average, that is, if $h_{ii} > 3(k+1)/n$.
- Options for handling high leverage points (similar to outliers)
  - Ignore them in the analysis but be sure to discuss their effects.
  - Delete them from the data set (but be sure to discuss their effects).
  - Choose another variable to represent the information.
  - Use a nonlinear transformation of an explanatory variable.

Cook's Distance

- Another measure of "influence." This measure considers both the explanatory and response variables.
- This distance, D_i, is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1)s^2}.$$

  $\hat{y}_{j(i)}$ is the prediction of the jth observation, computed leaving the ith observation out of the regression fit.
- Algebra shows that

$$D_i = \left(\frac{e_i}{se(e_i)}\right)^2 \frac{h_{ii}}{(k+1)(1-h_{ii})}.$$

  - $(e_i/se(e_i))^2$, is the square of the ith standardized residual.
  - $h_{ii}/((k+1)(1-h_{ii}))$, is attributable solely to the leverage.
  - The distance D_i is composed of a measure for outliers times a measure for leverage.

Example. Outliers and High Leverage Points

- In Section 2.6, we considered 19 "good," or base, points plus each of the three types of unusual points, labeled A, B and C.

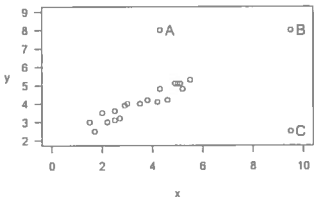


Table: Measures of Three Types of Unusual Points

Observation	Standardized residual $e/se(e)$	Leverage h	Cook's distance D
A	4.00	.067	.577
B	.77	.550	.363
C	-4.01	.550	9.832

What is Collinearity?

- *Collinearity*, or *multicollinearity*, occurs when one explanatory variable is, or nearly is, a linear combination of the other explanatory variables.
  - Useful to think of the explanatory variables as being highly correlated with one another.
- **Example** . Data

i	1	2	3	4
y_i	23	83	63	103
x_{i1}	2	8	6	10
x_{i2}	6	9	8	10

  - Joe Finance was asked to fit the model E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to a data set.
    - He got $\hat{y} = -87 + x_1 + 18x_2$.
    - Happy! $R^2 = 100\%$! For example, for $i = 3$, $\hat{y}_3 = -87 + 6 + 18(8) = 63 = y_3$.
  - Jane Actuary came along and fit the model $\hat{y} = -7 + 9x_1 + 2x_2$. Also got $R^2 = 100\%$! Who is right?
  - Answer: Both/neither. Infinite number of fits because of the perfect relation $x_2 = 5 + x_1/2$

Collinearity Facts

- Collinearity neither precludes us from getting good fits nor from making predictions of new observations. Note that in the above example we got perfect fits.
- Estimates of error variances and, therefore, tests of model adequacy, are still reliable.
- In cases of serious collinearity, standard errors of individual regression coefficients are larger than cases where, other things equal, serious collinearity does not exist.
  - With large standard errors, individual regression coefficients may not be meaningful.
  - Because a large standard error means that the corresponding t-ratio is small, it is difficult to detect the importance of a variable.

Variance Inflation Factors

- To detect collinearity, begin with a matrix of correlation coefficients of the explanatory variables.
  - This matrix is simple to create, easy to interpret and quickly captures linear relationships between pairs of variables.
  - A scatterplot matrix provides a visual reinforcement of the summary statistics in the correlation matrix.
- To capture more complex relationships among several variables, use a *variance inflation factor (VIF)*.
  - Suppose that the set of explanatory variables is labeled $x_1, x_2, ..., x_k$.
  - Run the regression using x_j as the "response" and the other x's $(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_k)$ as the explanatory variables.
  - Denote the coefficient of determination from this regression by R_j^2.
  - Define the variance inflation factor

$$VIF_j = \frac{1}{1 - R_j^2}, \quad \text{for } j = 1, 2, ..., k.$$

Variance Inflation Factors

- The variance inflation factor is $VIF_j = \frac{1}{1-R_j^2}$.
  - A larger R_j^2 results in a larger VIF_j; this means greater collinearity between x_j and the other x's.
- We use VIF because of the relation $se(b_j) = s\frac{\sqrt{VIF_j}}{s_{x_j}\sqrt{n-1}}$.
  - Here, $se(b_j)$ and s are from a *full* regression fit of y on $x_1, ..., x_k$.
  - Further, $s_{x_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ is the sample standard deviation of the jth variable x_j.
- Rule of thumb: When VIF_j exceeds 10 (which is equivalent to $R_j^2 > 90\%$), we say that severe collinearity exists. This may signal is a need for action.
  - Recall that $se(b_j) = s\sqrt{(j+1)st\text{diagonal element of } (\mathbf{X}'\mathbf{X})^{-1}}$.
  - When collinearity occurs, the matrix $\mathbf{X}'\mathbf{X}$ is close to zero.
  - The inverse of $\mathbf{X}'\mathbf{X}$ becomes large.

Example. Stock Liquidity - Continued

- Regression of VOLUME on PRICE, SHARE and VALUE.
- These 3 explanatory variables are not measures of trading activity.
- From a regression fit, we have $R^2 = 61\%$ and $s = 6.72$.

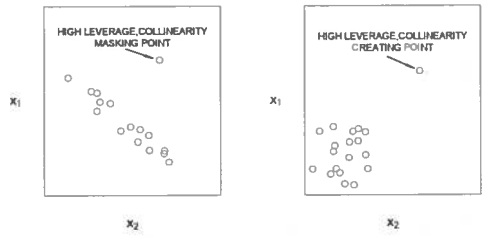
Table: Regression of VOLUME on PRICE, SHARE and VALUE

x_j	s_{x_j}	b_j	$se(b_j)$	$t(b_j)$	VIF_j
PRICE	21.37	-0.022	0.035	-0.63	1.5
SHARE	115.1	0.054	0.010	5.19	3.8
VALUE	8.157	0.313	0.162	1.94	4.7

- Because each *VIF* statistic is less than ten, there is little reason to suspect severe collinearity.
  - This is interesting because you may recall that there is a perfect relationship between PRICE, SHARE and VALUE in that we defined the market value to be VALUE = PRICE $\times$ SHARE.
  - The relationship is multiplicative, and hence is nonlinear. Because the variables are not linearly related, it is valid to enter all three into the regression model.

Collinearity and Leverage

- Both measures are based on the explanatory variables.
- Collinearity is about variables, leverage is about observations.
- They are related, as follows.
  - Left panel: With the exception of the marked point, x_1 and x_2 are highly linearly related.
  - Right panel: The highly linear relationship between x_1 and x_2 is primarily due to the marked point.



Options for Handling Collinearity

- Recode the variables by "centering" - that is, subtract the mean and divide by the standard deviation.
- Ignore the collinearity in the analysis but comment on it in the interpretation. Probably the most common approach.
- Replace one or more variables by auxiliary variables or transformed versions.
- Remove one or more variables. Easy. Which One? is hard.
  - Use interpretation. Which variable(s) do you feel most comfortable with?
  - Use automatic variable selection procedures to suggest a model.

Suppressor Variables

- Even if one explanatory variable is nearly a linear combination of the others, that does not mean that the information that it provides is redundant.
- A *suppressor variable* is an explanatory variable that increases the importance of other explanatory variables when included in the model.

Table: Correlation Matrix for the Suppressor Example

	x_1	x_2
x_2	0.972	
y	0.188	−0.022

- The regression of x_1 and x_2 on y yields $R^2 = 80.7\%$!!

Goodness of Fit

- How well does the model fit the data?
  - Criteria that measure the proximity of the fitted model and realized data are known as *goodness of fit* statistics.
  - Basic examples include the coefficient of determination (R^2), an adjusted version (R_a^2), the size of the typical error (s), and t-ratios for each regression coefficient.
- A general measure is *Akaike's Information Criterion*, defined as

$$AIC = -2 \times (\textit{fitted log likelihood}) + 2 \times (\textit{number of parameters})$$

  - For model comparison, the smaller the *AIC*, the better the fit.
  - This measures balances the fit (in the first part) with a penalty for complexity (in the second part)
  - It is a general measure - for linear regression, it reduces to

$$AIC = n\ln(s^2) + n\ln(2\pi) + n + k + 3.$$

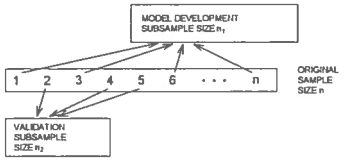
  - Stat packages often omit constants such as $n\ln(2\pi)$ when reporting *AIC* because they do not matter when comparing models.

Heteroscedasticity

- Heteroscedasticity means "different scatter."
- Our regression methods are based on the usual assumption of common variability, Assumption E3/F3 in Section 3.2, called *homoscedasticity* which stands for "same scatter."
- Strategies will depend on the extent of heteroscedasticity
  - For mild heteroscedasticity, use least squares estimators, perhaps with a "heteroscedasticity-corrected" standard error (Section 5.7.2)
  - If the amount of scatter depends on known variables or patterns, use "weighted least squares" (Chapter 13).
  - For severe heteroscedasticity, we will transform the dependent variable

Model Validation

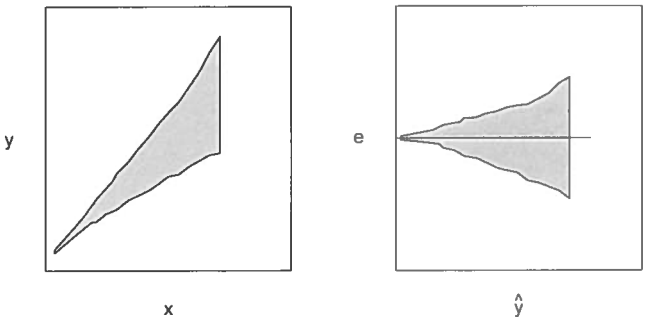
- Model validation is the process of confirming our proposed model.
- Concern: *data-snooping* - fitting many models to a single set of data.
- Response to concern: *out-of-sample validation*.
  - (i) Divide the data into *model development* and *validation subsamples*.



  - (ii) Using the model development subsample, fit a candidate model.
  - (iii) Using the Step (ii) model and the explanatory variables from the validation subsample, "predict" the dependent variables in the validation subsample, $\hat{y}_i$, where $i = n_1 + 1, ..., n_1 + n_2$.
  - (iv) Calc the *sum of squared prediction errors* $SSPE = \sum_{i=n_1+1}^{n_1+n_2} (y_i - \hat{y}_i)^2$. Repeat Steps (ii) through (iv) for each candidate model. Choose the model with the smallest *SSPE*.

Detecting Heteroscedasticity

- To detect heteroscedasticity graphically, plot the residuals versus the fitted values.
  - Left panel: The shaded area represents the data. The line is the true regression line.
  - Right panel: Residuals plotted versus the fitted values for data in the left panel.



Transformations

- As we saw in Section 1.3, transformations can serve to "shrink" spread out data and symmetrize a distribution.
  - This is both a strength and limitation of the transformation approach - a transformation simultaneously affects both the distribution and the heteroscedasticity.
- Power transformations, including the logarithmic transform, are most useful when the variability of the data grows with the mean.
  - Conversely, transformations will not help with patterns of variability that are non-monotonic.
- With transformations, we are implicitly optimizing the model in a new scale. This may not be helpful. For example, no one cares about the best predictions in terms of "log dollars."