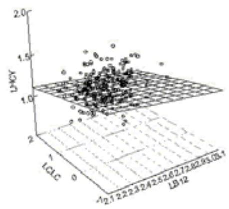
Multiple Regression



Motivations

- to make the predictions of the model more precise by adding other factors believed to affect the dependent variable to reduce the proportion of error variance associated with SSE
- to support a causal theory by eliminating potential sources of spuriousness

Multiple Regression

- method employed when the dependent variable is a function of two or more independent variables.
- necessary because few relations explained by bivariate models, other are determinants important.
- To expand methodology to include more than one independent variable

---

- First Question : which independent variables should be added?

- Answer :
  - Intuition
  - Theory
  - Empirical
  - Diagnostic Residuals
- Introduction of additional independent variables reduces *STOCHASTIC ERROR* - the error that arises because of inherent irreproducibility of physical or social phenomenon. i.e. Independent variables (that effect y) that are omitted.

- Expected $(y_i) = \hat{a} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$
  - Where : k = number of independent variables;
  - i = observation (y_i, x_i pairs) β = estimate (from this sample i = 1 .... n) of B, the population parameter

Example

	River	y_i(discharge)	x_{i1} (distance)	x_{i2} (basin)
i = 1	Nile	324	6690	3031.7
i = 2	Amazon	6630	12741	7050
i = 3	Chang Jiang	900	5797	1800
:	:	:	:	:
:	:	:	:	:

SPSS output

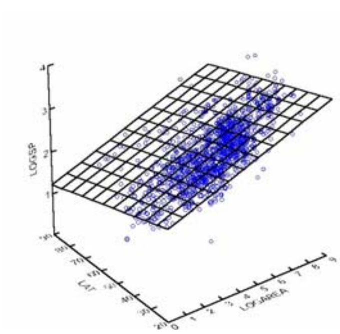
Coefficients						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-970.274	283.6289		-3.42093	0.001935
	basin/1000	0.413899	0.166816	0.481409	2.481173	0.019369
	length	0.21435	0.107501	0.386873	1.993934	0.055981
a	Dependent Variable: discharge cubic km					

$$Y(discharge) = \alpha + \beta_1(\mathbf{distance}) + \beta_2(\mathbf{basin})$$

For the Nile: β_1=6690, β_2=3031.7, α=-970.3

$$\hat{Y}_{Nile} = -970.3 + .214(6690) + .414(3031.7) = 1716.5$$

- y - $\hat{y}$= residual or an ERROR TERM
  error$_{(Nile)}$ =324-1716.5=-1392.5
- The slope of the plane is described by two parameters (up to k on a dimensional hypersurface), in this example :
- β_1 = slope in x_1 direction
- β_2 = slope in x_2 direction



- β_1 and β_2 are called *PARTIAL REGRESSION COEFFECIENTS* (because the coefficient only partially explains or predicts changes in Y)
- The plane is a least - squares plane that minimizes the sum of squared deviations in the y dimension.
- Ordinary least squares (OLS) - select the combination of (α, β_1, β_2, ... β_k) that minimizes the sum of squares deviations between y_is and x_is
- As with simple regression, the *y*-intercept disappears if all variables are standardized

$$\min s = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min s = \sum_{i=1}^{n} (y_i - [\alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}])^2 = \sum_{i=1}^{n} \varepsilon_i^{\,2}$$

How : Set partial derivatives of functions with respect to α, β_1, β_k (the unknowns) equal to zero and solve. End up with what are termed the "*NORMAL EQUATIONS*".

- They represent the additional effect of adding the variable if the other variables are controlled for

- Value of the parameter β expresses the relation between the dependent variables and the independent variables while holding the effects of all other variables in the regression constant.
- It is still the amount of change in y for each unit change in X, while holding contributions of other variables constant. Thus as independent variables are added to a regression model, change $\hat{\beta}s$

- Substantive significance versus statistical significance
  - Statistical significance is tested via F tests or t tests
  - Substantive can be evaluated several ways
    - Examine the unstandardized regression coefficient to see if its large enough to be concerned about
    - How much does the independent variable contribute to an increase in r^2 (as in stepwise regression)

$\hat{y}$

Multiple Correlation Coefficient

- SPSS for Windows outputs three coefficients :
  - (1) MULTIPLE r 0.88
  - (2) R - SQUARE 0.77 = 0.882
  - (3) ADJUSTED r^2 0.76
- same interpretation of 'r' of simple correlation coefficient
- the 'gross' correlation between y and $\hat{y}$, a measure of the scatter of y from the Least Square Surface.

MULTIPLE COEFFICIENT OF DETERMINATION

- r^2 = proportion of variance of the dependent variable accounted for by independent variables

$$r^2 = \frac{\text{variance accounted for by model}}{\text{total variance of } y}$$

Adjusted coefficient

- Is r^2 adjusted for the number of independent variables and sample size. Should report this in results.

$$r^2_{adjusted} = r^2 - \frac{k(1-r^2)}{N-k-1}$$

- If there is much *intercorrelation* (multicollinearity) between independent variables, adding other independent variables will not raise r^2 by much thus;
- Adding independent variables not related to each other will raise r^2 by a lot if these independent variables are, themselves, related to y.

Methods of regression

- All possible equations
- If there are 5 independent variables (n = 5), the number of ‘possible’ combinations of models = 31 plus the null model
  - for a total of 32
- If there are many independent variables we need a way to pick out the best equation

- Trade - off :
  - (a) Adding variables will *always* increase r^2, the percent of the variance explained, and predictions will be better.
  - (b) Verses explanation, clearer interpretation of the relationships between independent and dependent variables, parsimonious, clarity.
- Will MAXIMIZE r^2 while MINIMIZING the number of independent variables.

Forward Selection

- Picks the X variable with the highest r, puts in the model
- Then looks for the X variable which will increase r^2 by the highest amount
- Test for statistical significance performed (using the *F* test)
- If statistically significant, the new variable is included in the model, and the variable with the next highest r^2 is tested
- The selection stops when no variable can be added which significantly increases r^2

Backwards Elimination

- Starts with all variables in the model
- Removes the X variable which results in the smallest change in r^2
- Continues to remove variables from the model until removal produces a statistically significant drop in r^2

Stepwise regression

- Similar to forward selection, but after each new X added to the model, all X variables already in the model are re-checked to see if the addition of the new variable has effected their significance
- *Bizarre, but unfortunately true*: running forward selection, backward elimination, and stepwise regression on the same data often gives different answers

- The existence of suppressor variables may be a reason
  - A variable may appear statistically significant only when another a variable is controlled or held constant
  - This is a problem associated with the forward stepwise regression

- The RSQUARE method differs from the other selection methods in that RSQUARE always identifies the model with the largest r^2 for each number of variables considered. The other selection methods are not guaranteed to find the model with the largest r^2. The RSQUARE method requires much more computer time than the other selection methods, so a different selection method such as the STEPWISE method is a good choice when there are many independent variables to consider.

- Adjusted r^2 Selection (ADJRSQ)
  - This method is similar to the RSQUARE method, except that the adjusted r^2 statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted r^2 within the range of sizes.
- Mallows' C_p Selection
  - This method is similar to the ADJRSQ method, except that Mallows' C_p statistic is used as the criterion for model selection. Models are listed in ascending order of C_p.

Alternate approaches

- Mallow's C_p is available in SPSS using the command syntax but not as a selection method
- SAS does include it

$$C_p = \frac{SS(k\ variable\ model) - SS(p\ variable\ model)}{MS(k\ variable\ model)} + 2p - (k+1)$$

If the p variable model is as good as the k variable model The $C_p \leq p+1$

Types of errors

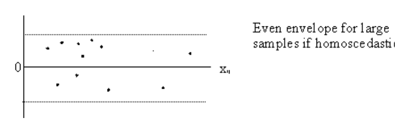
- <u>Specification Error</u>
- the wrong model was specified. There are 2 ways this kind of error can occur :
  - a) We may have the proper variables but the wrong *functional form*
  - model assumes the relationship are linear and additive. If violated, the least square estimates will be biased.
  - b) Wrong Independent Variables. When relevant variable is excluded, the remaining pick up some of the impact of that variable. The result is biased estimators, the direction of bias depends on the direction of the effect of the excluded variables.

- <u>Measurement Error</u>
  - 2 types of error - random and non - random
    - a) Random - results in lower r^2, partial slope coefficients are hard to achieve statistical significance.
    - b) Non - Random - brings up the question of the *validity* of the measurement.
- <u>Multicollinearity</u>

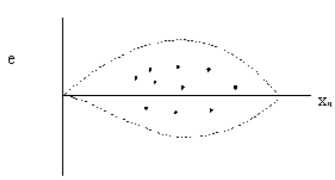
- <u>Heteroscedasticity</u>
  - the error term in a regression model does not have constant variance.
    - Situations where it can occur:
    - a) Where dependent variable is measured with error and the amount of error varies with the value of the independent variable.
    - b) When the unit of analysis is an aggregate and the dependent variable is an average of values for individual objects.
    - c) Interaction between independent variables in the model and another variable left out of the model.

- When present, the standard error of partial slope coefficients are no longer unbiased estimators of the true estimator.
- Standard Deviations - test of statistical significance based on these standard errors will be inaccurate.
- How to detect?
- Look at plot of residual against X.

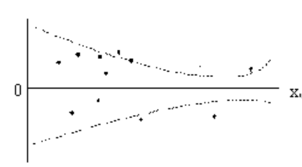
For large samples



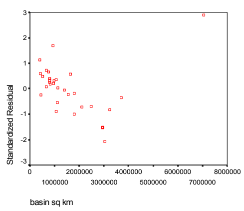
- For small samples

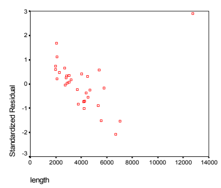


- Other forms probably indicate heteroscedasticity :



River data





- The impact of collinearity on the precision of estimation is captured by 1 / (1 - R^2) called the *Variance Inflation Factor*, VIF. The R^2 is the multiple regression of a particular x on the others.
- Probably better look at :
- The table below reveals the linear relationship between. Among the x's must be very strong before collinearity seriously degrades the precision of estimation.
- i.e. Not until r, approaches 0.9 that precision of estimation is halved.

Variance Inflation Factor

Example A: If R_j^2 =.00 then VIF_j = 1:

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{1-0} = 1}$$

Example B: If R_j^2 = .90 then VIF_j = 10:

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{1-.90} = 10}$$

Evidence of Multicollinearity

- Any VIF > 10
- Sum of VIFs > 10
- High correlation for pairs of predictors X_j and X_k
- Unstable estimates
  - *(i.e., the remaining coefficients change sharply when a suspect predictor is dropped from the model)*

Example: Estimating Body Fat

The regression equation is
Fat%1 = 18.6 + 0.0685 Age - 0.197 Height - 0.765 Neck - 0.051 Chest + 0.943 Abdomen - 0.731 Hip + 0.530 Thigh

Predictor	Coef	StDev	T	P	VIF
Constant	18.63	12.44	1.50	0.14	
Age	0.06845	0.09268	0.74	0.46	1.7
Height	-0.197	0.1087	-1.81	0.08	1.3
Neck	-0.765	0.3836	-1.99	0.05	4.4
Chest	-0.0514	0.1865	-0.28	0.78	10.9
Abdomen	0.9426	0.1731	5.45	0.00	17.6
Hip	-0.7309	0.2281	-3.20	0.00	15.9
Thigh	0.5299	0.2886	1.84	0.07	10.5

S = 4.188 R-Sq = 81.8% R-Sq(adj) = 78.7%

Problem: Several VIFs exceed 10.

Correlation Matrix of Predictors

	Age	Height	Neck	Chest	Abdomen	Hip
Height	-0.276					
Neck	0.176	0.201				
Chest	0.376	0.014	0.820			
Abdomen	0.442	-0.052	0.781	0.942		
Hip	0.314	-0.045	0.804	0.911	0.942	
Thigh	0.219	-0.037	0.823	0.859	0.890	0.938

Age and *Height* are relatively independent of other predictors.

Problem: *Neck*, *Chest*, *Abdomen*, and *Thigh* are highly correlated.

Solution: Eliminate Some Predictors

The regression equation is
Fat%1 = 0.8 + 0.0927 Age - 0.184 Height - 0.842 Neck + 0.637 Abdomen

Predictor	Coef	StDev	T	P	VIF
Constant	0.79	10.35	0.08	0.94	
Age	0.0927	0.09199	1.01	0.32	1.4
Height	-0.1837	0.1133	-1.62	0.11	1.2
Neck	-0.8418	0.3516	-2.39	0.02	3.2
Abdomen	0.63659	0.0846	7.52	0.00	3.6

S = 4.542   R-Sq = 77.0%   R-Sq(adj) = 75.0%

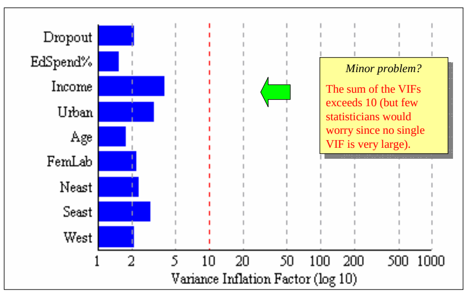
R^2 is reduced slightly, but all VIFs are below 10.

Stability Check for Coefficients

Variable	Run 1	Run 2	Run 3	Run 4	%Chg
Constant	18.63	17.67	19.89	0.79	-95.8%
Age	0.06845	0.0689	0.0200	0.0927	35.4%
Height	-0.1970	-0.1978	-0.2387	-0.1837	-6.8%
Neck	-0.7650	-0.8012	-0.5717	-0.8418	10.0%
Chest	-0.0514				
Abdomen	0.9426	0.9158	0.9554	0.6366	-32.5%
Hip	-0.7309	-0.7408	-0.5141		
Thigh	0.5299	0.5406			
Std Err	4.188	4.143	4.266	4.542	8.5%
R-Sq	81.8%	81.7%	80.2%	77.0%	-5.9%
R-Sq(adj)	78.7%	79.2%	77.9%	75.0%	-4.7%

There are large changes in estimated coefficients as high VIF predictors are eliminated, revealing that the original estimates were unstable. But the "fit" deteriorates when we eliminate predictors.

Example: College Graduation Rates

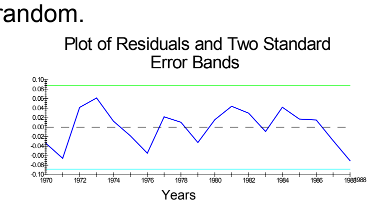


- Autocorrelation means that the error is not truly random, but depends upon its own past values, e.g.
  - $e_t = \rho\, e_{t-1} + v_t$
  - where ρ measures the correlation between successive errors and v is another error term, but a truly random one.

- Why does autocorrelation matter? If *e* is not truly random then it is, to some extent, predictable. If so, we ought to include that in our model. If our model exhibits autocorrelation, then it cannot be the best model for explaining *y*.
- If autocorrelation exists in the model, then the coefficient estimates are unbiased, but the standard errors are not. Hence inference is invalid. *t* and *F* statistics cannot be relied upon.

Detecting autocorrelation

- Graph the residuals – they should look random.



- Evidence here of **positive autocorrelation** ($\rho > 0$) – positive errors tend to follow positive errors, negative errors to follow negative errors.
- It looks likely the next error will be negative rather than zero.

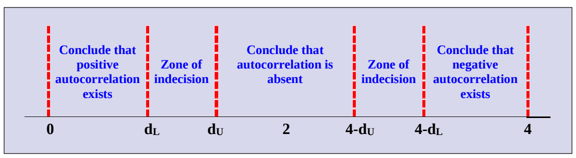
The Durham Watson Statistic

- => Test for Autocorrelation
- Small values indicate positive correlation and large values indicate negative correlation

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Durbin-Watson Test

Bounded by 0 and 4



lower bound d_L and upper bound d_U are dependent upon the data and must be calculated for each analysis

A table of values can be found at http://hadm.sph.sc.edu/courses/J716/Dw.html

		X variables, excluding the intercept									
Observations		1		2		3		4		5	
N	Prob.	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U
15	0.05	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
	0.01	0.81	1.07	0.7	1.25	0.59	1.46	0.49	1.70	0.39	1.96
20	0.05	1.20	1.71	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
	0.01	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
25	0.05	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
	0.01	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
30	0.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
	0.01	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
40	0.05	1.44	1.54	1.39	1.60	1.34	1.66	1.39	1.72	1.23	1.79
	0.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
50	0.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
	0.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
60	0.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
	0.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
80	0.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
	0.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
100	0.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
	0.01	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

- To formally test for serial correlation in your residuals:
  - Find the box corresponding to the number of X variables in your equation and the number of observations in your data. Choose the row within the box for the significance level ("Prob.") you consider appropriate. That gives you two numbers, a D-L and a D-U. If the Durbin-Watson statistic you got is less than D-L, you have serial correlation. If it is less than D-U, you probably have serial correlation, particularly if one of your X variables is a measure of time.

From: http://hadm.sph.sc.edu/courses/J716/Dw.html

River colinearity stats

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	0.833994	0.695546	0.673799	672.5264	1.545644
a	Predictors: (Constant), length, basin/1000				
b	Dependent Variable: discharge cubic km				

- Positive autocorrelation is present if a positive (negative) residual in one period is followed by another positive (negative) residual the next period.
- Negative autocorrelation is present if positive (negative) residuals are followed by negative (positive) residuals.

Multiple Regression: Caveats

- Try not to include predictor variables which are highly correlated with each other
- One X may force the other out, with strange results
- Overfitting: too many variables make for an unstable model
- Model assumes normal distribution for variables - widely skewed data may give misleading results

Spatial Autocorrelation

- First law of geography: "everything is related to everything else, but near things are more related than distant things" – Waldo Tobler
- Many geographers would say "I don't understand spatial autocorrelation" Actually, they don't understand the mechanics, they do understand the concept.

Spatial Autocorrelation

- Spatial Autocorrelation – correlation of a variable with itself through space.
  - If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated
  - If nearby or neighboring areas are more alike, this is *positive spatial autocorrelation*
  - *Negative autocorrelation* describes patterns in which neighboring areas are unlike
  - Random patterns exhibit *no spatial autocorrelation*

Why spatial autocorrelation is important

- Most statistics are based on the assumption that the values of observations in each sample are independent of one another
- Positive spatial autocorrelation may violate this, if the samples were taken from nearby areas
- Goals of spatial autocorrelation
  - Measure the strength of spatial autocorrelation in a map
  - test the assumption of independence or randomness

Spatial Autocorrelation

- It measures the extent to which the occurrence of an event in an areal unit constrains, or makes more probable, the occurrence of an event in a neighboring areal unit.

Spatial Autocorrelation

- Non-spatial independence suggests many statistical tools and inferences are inappropriate.
  - Correlation coefficients or ordinary least squares regressions (OLS) to predict a dependent variable assumes random samples
  - If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise.
  - They are biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed.

Indices of Spatial Autocorrelation

- Moran's I
- Geary's C
- Ripley's K
- Join Count Analysis

Spatial regression

- The existence of spatial autocorrelation can be used to improve regression analysis
- One can use spatial regression to allow the regression to make use of variables exhibiting like values of neighboring observations
- Use of this technique is often covered in GIS courses but is beyond the scope of this course

How Many Predictors?

Regression with an intercept can be performed as long as n exceeds p+1. However, for sound results desirable that n be substantially larger than p. Various guidelines have been proposed, but judgment is allowed to reflect the context of the problem.

Rule 1 (maybe a bit lax)

$n/p \geq 5$ (at least 5 cases per predictor)

Example: n = 50 would allow up to10 predictors

Rule 2 (somewhat conservative)

$n/p \geq 10$ (at least 10 cases per predictor

Example: n = 50 would allow up to 5 predictors

Binary Model Form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

X_2 is binary (0 or 1)

Explanation

If $X_2 = 0$ then $Y_i = \beta_0 + \beta_1 X_1 + \beta_2(0) + \varepsilon_i$

$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$

If $X_2 = 1$ then $Y_i = \beta_0 + \beta_1 X_1 + \beta_2(1) + \varepsilon_i$

$Y_i = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon_i$

The binary (also called dummy) variable shifts the intercept

Example: Binary Predictors

MPG = 27.52 – .00356 Weight + 2.51 Stick

Explanation

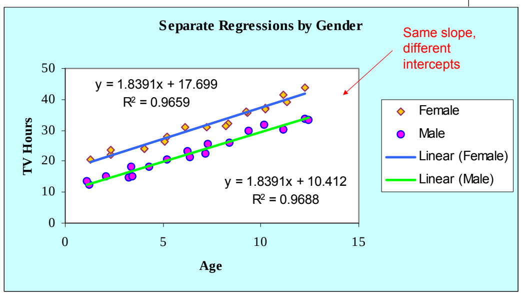
Define: **Stick = 1 if manual transmission**
Stick = 0 if automatic

If Stick = 0 then **MPG = 27.52 - .00356 Weight + 2.52(0)**
i.e., **MPG = 27.52 –.00356 Weight**

If Stick = 1 then **MPG = 27.52 - .00356 Weight + 2.51**
i.e., **MPG = 30.03 – .00356 Weight**

The binary variable shifts the *intercept*

Binaries Shift the Intercept



k-1 Binaries for k Groups?

That's right, for k groups, we only need k-1 binaries

Gender (male, female) requires only 1 binary (e.g., male) because male=0 would be female.

Season (fall, winter, spring, summer) requires only 3 binaries (e.g., fall, winter, spring) because fall=0, winter=0, spring=0 would be summer.

For provincial data, we might divide Canada into 4 regions, but in a regression, we omit *one* region.

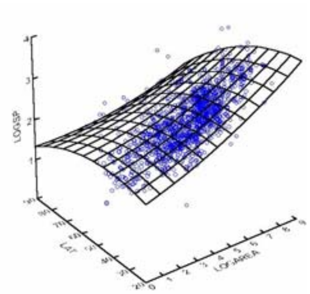
The omitted binary is the base reference point. No information is lost.

What about polynomials?

- Note that:
  $y = ax^3 + bx^2 + cx + d + e$
- can be expressed as:
  $y = ß0 + ß_1x_1 + ß_2x_2 + ß_3x_3 + e$
- if $x_1 = x^1$, $x_2 = x^2$, $x_3 = x3$

- So polynomial regression is considered a special case of linear regression.
- This is handy, because even if polynomials do not represent the *true* model, they take a variety of forms, and may be close enough for a variety of purposes.
- Fitting a response surface is often useful:
  $y = \alpha + ß_1x_1 + ß_2x_1^2 + ß_3x^2 + ß_4x_2^2 + ß_4x_1x_2 + \varepsilon$

This can fit simple ridges, peaks, valleys, pits, slopes, and saddles.



Interaction Terms

$$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \varepsilon_i$$

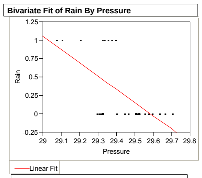
If we can reject $\beta_3 = 0$ there is a significant interaction effect

Pro
- ❖ **Detects interaction between any two predictors**
- ❖ **Multiple interactions are possible (e.g., $X_1X_2X_3$)**

Con
- ❖ **Becomes complex if many predictors**
- ❖ **Difficult to interpret the coefficient**

Logistic Regression

If the dependent variable Y is binary (0 or 1) and the X's are continuous, a linear regression model is inappropriate. The logistic regression is a non-linear model with the form $Y = 1/\{1+\exp[-(b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p)]\}$. Y interpreted as the *probability* of the binary event. Rain = f(Barometric Pressure). The binary Y is Rain = 0, 1.

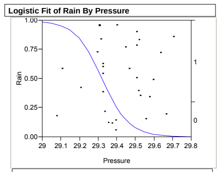


Linear Fit

Rain = 53.241865 - 1.7992749 Pressure

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	53.241865	13.80328	3.86	0.0006
Pressure	-1.799275	0.46911	-3.84	0.0007



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	405.362669	169.29517	5.73	0.0166
Pressure	-13.823388	5.7651316	5.75	0.0165

For log odds of 0/1

APPROPRIATE PROCEDURE FOR MULTIVARIABLE ANALYSIS: ANALYSIS OF ONE DEPENDENT VARIABLE AND MORE THAN ONE INDEPENDENT VARIABLE

Characterization of Variables to be Analyzed		
Dependent Variable	Independent Variables*	Appropriate Procedure(s)
Continuous	All categorical	ANOVA
Continuous	Some categorical, some continuous	ANOVA
Continuous	All continuous	Multiple linear regression
Ordinal	--	No formal multivariate procedure. Treat variables as if continuous (see above procedures) or perform log-linear analysis
Dichotomous	All categorical	Logistic regression; log-linear analysis
Dichotomous	Some categorical, some continuous	Logistic regression†
Dichotomous	All continuous	Logistic regression; discriminant function analysis
Nominal	All categorical	Log-linear analysis
Nominal	Some categorical, some continuous	Group the continuous variables and perform log-linear analysis
Nominal	All continuous	Discriminant function analysis; group the continuous variables and perform log-linear analysis

* Categorical variables include ordinal, dichotomous, and nominal variables.
† If the outcome is a time-related, dichotomous variable (such as live/die), then proportional-hazards (Cox) models are best

77

Results of Regression Assumption violations

The assumption of the absence of perfect multicollinearity

- if there is perfect multicollinearity then there are an infinite number of regressions that will fit the data
- 3 ways this can happen
  - a) you mistakenly put in independent variables that are linear combinations of each other
  - b) putting in as many dummy variables as the number of classes of the nominal variable you are trying to use
  - c) if the sample size is too small, ie the number of cases is less than the number of independent variables

- for example if you use 3 independent variables and 2 data points, the job is to find the plane of best fit but you only have 2 data points
- a line perfectly fits the 2 points, so any plane containing that line also fits

- The estimates of the partial slope coefficients will have high standard errors so that there will be high variability of the estimates between samples

Specification error: Leaving out a relevant independent variable

- Consequences: Biased partial slope coefficients

The assumption that the mean of the error term is zero

- can happen in 2 cases
  - 1) the error is a constant across all cases
  - 2) the error term varies - this is the more serious case
  - for case 1 - intercept is biased by an amount equal to the error term
    - it can happen with measurement error equal to a constant
- for case 2 - causes bias in the partial slope coefficients

The assumption of measurement without error

- a) random measurement error
  - if it affects the dependent variable the r^2 is attenuated and estimates are less efficient but unbiased
  - If it affects independent variable the parameter estimates are biased
- b) nonrandom measurement error
  - always leads to bias but amount and type depends on the error

The assumptions of linearity and additivity

- errors of this type are a kind of specificity error
- difficult to predict the effect

The assumptions of homoscedasticity and lack of autocorrelation

- assumption that the variance of the error term is constant
  - accuracy of data is constant across data
  - i.e. it doesn't get better or worse over time
- significance tests are invalid
  - likely a problem in time series models but also in cases of spatial autocorrelation

The assumption that the error term is normally distributed

- important for small samples to allow for significance testing

- for large samples you can test even if its not normal