

Multiple Regression

- Numeric Response variable (y)
- p Numeric predictor variables ($p < n$)
- Model:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Partial Regression Coefficients: $\beta_i \equiv$ effect (on the mean response) of increasing the i^{th} predictor variable by 1 unit, **holding all other predictors constant**
- Model Assumptions (Involving Error terms ε)
 - Normally distributed with mean 0
 - Constant Variance σ^2
 - Independent (Problematic when data are series in time/space)

Example - Effect of Birth weight on Body Size in Early Adolescence

- Response: Height at Early adolescence ($n = 250$ cases)
- Predictors ($p=6$ explanatory variables)
 - Adolescent Age (x_1 , in years -- 11-14)
 - Tanner stage (x_2 , units not given)
 - Gender ($x_3=1$ if male, 0 if female)
 - Gestational age (x_4 , in weeks at birth)
 - Birth length (x_5 , units not given)
 - Birthweight Group ($x_6=1, \dots, 6$ <1500g (1), 1500-1999g(2), 2000-2499g(3), 2500-2999g(4), 3000-3499g(5), >3500g(6))

Least Squares Estimation

- Population Model for mean response:

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Least Squares Fitted (predicted) equation, minimizing *SSE*:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \quad SSE = \sum \left(Y - \hat{Y} \right)^2$$

- All statistical software packages/spreadsheets can compute least squares estimates and their standard errors

Analysis of Variance

- Direct extension to ANOVA based on simple linear regression
- Only adjustments are to degrees of freedom:
 - $DF_R = p$ $DF_E = n - p^*$ ($p^* = p + 1 = \#Parameters$)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	<i>F</i>
Model	<i>SSR</i>	<i>p</i>	$MSR = SSR/p$	$F = MSR/MSE$
Error	<i>SSE</i>	$n - p^*$	$MSE = SSE/(n - p^*)$	
Total	<i>TSS</i>	$n - 1$		

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{SSR}{TSS}$$

Testing for the Overall Model - F -test

- Tests whether **any** of the explanatory variables are associated with the response
- $H_0: \beta_1 = \dots = \beta_p = 0$ (None of the x^s associated with y)
- H_A : Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2 / p}{(1 - R^2) / (n - p^*)}$$

$$R.R.: F_{obs} \geq F_{\alpha, p, n-p^*}$$

$$P - val : P(F \geq F_{obs})$$

Example - Effect of Birth weight on Body Size in Early Adolescence

- Authors did not print ANOVA, but did provide following:

- $n=250$ $p=6$ $R^2=0.26$

- $H_0: \beta_1=\dots=\beta_6=0$ $H_A: \text{Not all } \beta_i = 0$

$$\begin{aligned} T.S.: F_{obs} &= \frac{MSR}{MSE} = \frac{R^2 / p}{(1 - R^2) / (n - p^*)} = \\ &= \frac{0.26 / 6}{(1 - 0.26) / (250 - 7)} = \frac{.0433}{.0030} = 14.2 \end{aligned}$$

$$R.R.: F_{obs} \geq F_{\alpha, 6, 243} = 2.13$$

$$P - val : P(F \geq 14.2)$$

Testing Individual Partial Coefficients - t -tests

- Wish to determine whether the response is associated with a single explanatory variable, after controlling for the others
- $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (2-sided alternative)

$$T.S.: t_{obs} = \frac{\hat{\beta}_i}{s_{\hat{b}_i}}$$

$$R.R.: |t_{obs}| \geq t_{\alpha/2, n-p^*}$$

$$P\text{-val} : 2P(t \geq |t_{obs}|)$$

Example - Effect of Birth weight on Body Size in Early Adolescence

Variable	b	SE_b	t=b/SE_b	P-val (z)
Adolescent Age	2.86	0.99	2.89	.0038
Tanner Stage	3.41	0.89	3.83	<.001
Male	0.08	1.26	0.06	.9522
Gestational Age	-0.11	0.21	-0.52	.6030
Birth Length	0.44	0.19	2.32	.0204
Birth Wt Grp	-0.78	0.64	-1.22	.2224

Controlling for all other predictors, adolescent age, Tanner stage, and Birth length are associated with adolescent height measurement

Comparing Regression Models

- Conflicting Goals: Explaining variation in Y while keeping model as simple as possible (parsimony)
- We can test whether a subset of $p-g$ predictors (including possibly cross-product terms) can be dropped from a model that contains the remaining g predictors.

$$H_0: \beta_{g+1} = \dots = \beta_p = 0$$

- Complete Model: Contains all p predictors
- Reduced Model: Eliminates the predictors from H_0
- Fit both models, obtaining sums of squares for each (or R^2 from each):
 - Complete: $SSR_c, SSE_c (R_c^2)$
 - Reduced: $SSR_r, SSE_r (R_r^2)$

Comparing Regression Models

- $H_0: \beta_{g+1} = \dots = \beta_p = 0$ (After removing the effects of X_1, \dots, X_g , none of other predictors are associated with Y)
- $H_a: H_0$ is false

$$TS: F_{obs} = \frac{(SSR_c - SSR_r)/(p - g)}{SSE_c/[n - p^*]} = \frac{(R_c^2 - R_r^2)/(p - g)}{(1 - R_c^2)/[n - p^*]}$$

$$RR: F_{obs} \geq F_{\alpha, p-g, (n-p^*)}$$

$$P = P(F \geq F_{obs})$$

P -value based on F -distribution with $p-g$ and $n-p^*$ d.f.

Models with Dummy Variables

- Some models have both numeric and categorical explanatory variables (Recall **gender** in example)
- If a categorical variable has m levels, need to create $m-1$ dummy variables that take on the values 1 if the level of interest is present, 0 otherwise.
- The baseline level of the categorical variable is the one for which all $m-1$ dummy variables are set to 0
- The regression coefficient corresponding to a dummy variable is the difference between the mean for that level and the mean for baseline group, controlling for all numeric predictors

Example - Deep Cervical Infections

- Subjects - Patients with deep neck infections
- Response (Y) - Length of Stay in hospital
- Predictors: (One numeric, 11 Dichotomous)
 - Age (x_1)
 - Gender ($x_2=1$ if female, 0 if male)
 - Fever ($x_3=1$ if Body Temp $> 38C$, 0 if not)
 - Neck swelling ($x_4=1$ if Present, 0 if absent)
 - Neck Pain ($x_5=1$ if Present, 0 if absent)
 - Trismus ($x_6=1$ if Present, 0 if absent)
 - Underlying Disease ($x_7=1$ if Present, 0 if absent)
 - Respiration Difficulty ($x_8=1$ if Present, 0 if absent)
 - Complication ($x_9=1$ if Present, 0 if absent)
 - WBC $> 15000/mm^3$ ($x_{10}=1$ if Present, 0 if absent)
 - CRP $> 100\mu g/ml$ ($x_{11}=1$ if Present, 0 if absent)

Example - Weather and Spinal Patients

- Subjects - Visitors to National Spinal Network in 23 cities
Completing SF-36 Form
- Response - Physical Function subscale (1 of 10 reported)
- Predictors:
 - Patient's age (x_1)
 - Gender ($x_2=1$ if female, 0 if male)
 - High temperature on day of visit (x_3)
 - Low temperature on day of visit (x_4)
 - Dew point (x_5)
 - Wet bulb (x_6)
 - Total precipitation (x_7)
 - Barometric Pressure (x_7)
 - Length of sunlight (x_8)
 - Moon Phase (new, wax crescent, 1st Qtr, wax gibbous, full moon, wan gibbous, last Qtr, wan crescent, presumably had 8-1=7 dummy variables)

Modeling Interactions

- Statistical Interaction: When the effect of one predictor (on the response) depends on the level of other predictors.
- Can be modeled (and thus tested) with cross-product terms (case of 2 predictors):
 - $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
 - $X_2=0 \Rightarrow E(Y) = \alpha + \beta_1 X_1$
 - $X_2=10 \Rightarrow E(Y) = \alpha + \beta_1 X_1 + 10\beta_2 + 10\beta_3 X_1$
 $= (\alpha + 10\beta_2) + (\beta_1 + 10\beta_3)X_1$
- The effect of increasing X_1 by 1 on $E(Y)$ depends on level of X_2 , unless $\beta_3=0$ (t -test)

Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in Y with a small set of predictors
- Automated Procedures and all possible regressions:
 - Backward Elimination (Top down approach)
 - Forward Selection (Bottom up approach)
 - Stepwise Regression (Combines Forward/Backward)
 - C_p Statistic - Summarizes each possible model, where “best” model can be selected based on statistic

Backward Elimination

- Select a significance level to stay in the model (e.g. $SLS=0.20$, generally $.05$ is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest t -statistic (highest P -value).
 - If $P > SLS$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
 - If $P \leq SLS$, stop and keep current model
- Continue until all predictors have P -values below SLS

Forward Selection

- Choose a significance level to enter the model (e.g. $SLE=0.20$, generally $.05$ is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest t -statistic (lowest P -value)
 - If $P \leq SLE$, keep this variable and fit all two variable models that include this predictor
 - If $P > SLE$, stop and keep previous model
- Continue until no new predictors have $P \leq SLE$

Stepwise Regression

- Select SLS and SLE ($SLE < SLS$)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq SLE$ to enter
- Re-tests all “old variables” that have already been entered, must have $P \leq SLS$ to stay in model
- Continues until no new variables can be entered and no old variables need to be removed

All Possible Regressions - C_p

- Fits every possible model. If K potential predictor variables, there are $2^K - 1$ models.
- Label the Mean Square Error for the model containing all K predictors as MSE_K
- For each model, compute SSE and C_p where p^* is the number of parameters (including intercept) in model

$$C_p = \frac{SSE}{MSE_K} - (n - 2p^*)$$

- Select the model with the fewest predictors that has $C_p \approx p^*$

Regression Diagnostics

- Model Assumptions:
 - Regression function correctly specified (e.g. linear)
 - Conditional distribution of Y is normal distribution
 - Conditional distribution of Y has constant standard deviation
 - Observations on Y are statistically independent
- Residual plots can be used to check the assumptions
 - Histogram (stem-and-leaf plot) should be mound-shaped (normal)
 - Plot of Residuals versus each predictor should be random cloud
 - U-shaped (or inverted U) \Rightarrow Nonlinear relation
 - Funnel shaped \Rightarrow Non-constant Variance
 - Plot of Residuals versus Time order (Time series data) should be random cloud. If pattern appears, not independent.

Detecting Influential Observations

- ◆ **Studentized Residuals** – Residuals divided by their estimated standard errors (like t -statistics). Observations with values larger than 3 in absolute value are considered outliers.
- ◆ **Leverage Values (Hat Diag)** – Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than $2p^*/n$ are considered to be potentially highly influential, where p is the number of predictors and n is the sample size.
- ◆ **DFFITs** – Measure of how much an observation has effected its fitted value from the regression model. Values larger than $2\sqrt{p^*/n}$ in absolute value are considered highly influential. Use standardized DFFITS in SPSS.

Detecting Influential Observations

- ◆ **DFBETAS** – Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than $2/\sqrt{n}$ in absolute value are considered highly influential.
- ◆ **Cook's D** – Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than $4/n$ are considered highly influential.
- ◆ **COVRATIO** – Measure of the impact of each observation on the variances (and standard errors) of the regression coefficients and their covariances. Values outside the interval $1 \pm 3p^*/n$ are considered highly influential.

Variance Inflation Factors

- **Variance Inflation Factor (VIF)** – Measure of how highly correlated each **independent variable** is with the other predictors in the model. Used to identify **Multicollinearity**.
- Values larger than 10 for a predictor imply large inflation of standard errors of regression coefficients due to this variable being in model.
- Inflated standard errors lead to small t -statistics for partial regression coefficients and wider confidence intervals

Nonlinearity: Polynomial Regression

- When relation between Y and X is not linear, polynomial models can be fit that approximate the relationship within a particular range of X
- General form of model:

$$E(Y) = \alpha + \beta_1 X + \dots + \beta_p X^p$$

- Second order model (most widely used case, allows one “bend”):

$$E(Y) = \alpha + \beta_1 X + \beta_2 X^2$$

- Must be very careful not to extrapolate beyond observed X levels