

I. Review .....1  
 II. Fitting Equations to Data .....2  
 III. Simple Regression (one x).....3  
 IV. Multiple Regression .....5  
 V. Collinearity.....6  
 VI. Categorical Variables .....7  
 VII. Model Building.....8

I. Review

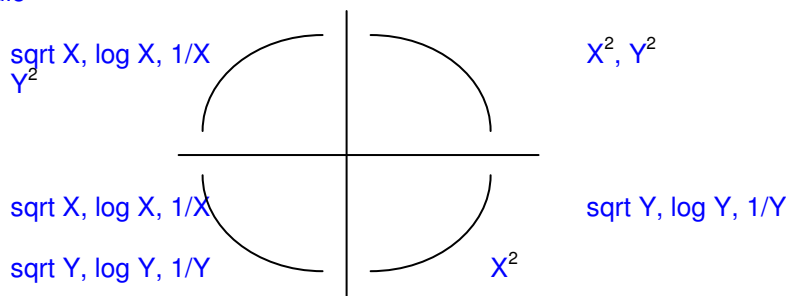
- **Expected Value** ~ avg, weighted by probabilities
- **variance** – (~ volatility / riskiness) – **also see VI below**
  - interpreted as – probability weighted avg squared deviation + LR avg squared deviation over infinitely many repetitions of X + risk of gamble + variance of population from which RV drawn + dispersion of probability distribution of X about  $\mu$
  - high variance ~ reduces  $E(X)$
- **t distributions** (*probability function t quantile on JMP calculator*) –
  - family of continuous probability distributions resembling normal distributions ~ bell shaped with fat tails (~ means more likely to get extreme observations)
  - defined by –  $\mu$ , var, and **df** (~ degrees of freedom ~ shape parameter that is always positive ~ as df rises → tails get lighter and distribution more closely approximates normal)
- **r (gamma) distributions** - family of continuous probability distributions ~ skewed (not symmetric) ~ used in queuing models for supply processes
- **Covariance**  $(Cov(X,Y)) = E [(X-\mu_X)(Y-\mu_Y)]$  (~ relation of return b/w random elements) (**interpretation = at process below**)
  - **problem1** – depends on units in which RV's are measured
  - **problem2** – difficult to calculate b/c  $X \cdot Y$  combos exist
- **Correlation**  $(\rho_{XY}) = Cov(X,Y) / \sigma_X \sigma_Y$  – (where  $-1 \leq \rho_{XY} \leq 1$ )
  - **measures** – degree of linear dependence b/w variables
  - **essentially** – represents a standardized covariance
  - **bene** – no dependence on units of RV (i.e., is a pure measure of linear association)
  - **interpretation** –
    - $\rho_{XY} = 1 \rightarrow$  line  $mx + b$        $\rho_{XY} = -1 \rightarrow$  line  $-mx + b$       (**interpretation = at process below**)
    - thickness of line = thickness of ellipse on scatter plot
    - thickness of ellipse =  $2(1-r^2)^{0.5}$
- **auto-correlation** – sequential dependence b/w adjacent errors about fitted relationship
  - relationship b/w rows
- **collinearity**
  - relationship b/w columns
- **heteroscedasticity** – lack of constant variance for noise terms
- **residual** – vertical deviation of a point from regression line
- **remember** -
  - rebalance portfolio b/c mizing investments reduces volatility
  - to increase Expected return → reduce volatility / variance
  - moments (mean, var, cov) may change in future (~ LTCM)
  - account for future inflation

- outcome of random variable w/ probability distribution  $p(x)$  can always be interpreted as a random selection from a population whose distribution has the shape  $p(x)$
- any RV w/ density  $f(x)$  always c/b interpreted as random selection from a population w/ distribution shape  $f(x)$
- interpreting covariance and correlation –
  - $Cov(X, Y) > 0$  &  $\rho_{XY} > 0$  → linear relationship (positive – where sign represents the tilt of association)
  - $Cov(X, Y) < 0$  &  $\rho_{XY} < 0$  → linear relationship (negative – where sign represents the tilt of association)
  - $Cov(X, Y) = 0$  &  $\rho_{XY} = 0$  → NO linear relationship (BUT does not imply independence b/c non-linear association may exist)
  - $Cov(X, Y) \neq 0$  &  $\rho_{XY} \neq 0$  → implies that X & Y dependent
  - **X & Y independent** → implies that  $Cov(X, Y) = 0$  &  $\rho_{XY} = 0$
- **Normal Distribution** -
  - 95% = w/i 2 SD
  - diagnostic = normal quantile plot
- **SE (x)** =  $s/(n^{0.5})$  [variability of statistic from one sample to the next]
- **Confidence Interval** – 95% ~ estimate +/- 2 SE (~ contains truth for 95% of samples)
- **Hypothesis test** –
  - **t-stat** – counts SEs from conjectured value
  - **p-value** –
    - measures plausibility of  $H_0$
    - if  $< 0.5$  → cannot reject  $H_0$  at .05 level of significance
    - tells how confident you are that true value of an estimate is zero

## II. Fitting Equations to Data

- **goal** = function fitting
- **sample correlation coefficient** –  $r = [(1/n-1) \text{ sum } (x_i - \bar{x})(y_i - \bar{y})] / (s_x)(s_y)$ 
  - estimates population correlation  $\rho_{xy}$
  - $-1 < \rho < 1$
  - problem – what if correlation is by chance?
  - solution –  $H_0: \rho_{xy} = 0$ 
    - JMP – provides p-value (“signif. prob”) to test  $H_0: \rho_{xy} = 0$
    - **ROT – if  $|r| > 2 / (n0.5)$  → reject  $H_0: \rho_{xy} = 0$  at the .05 level of significance**

- **Tulkey’s Rule**



- **remember**
  - log transformation ~ 1% change in log = change by 1% of slope  $b_0$
  - be suspicious of linear assumptions for big changes (may not hold beyond data observed)
  - do not fit through 0 b/c would worsen fit through observed data points
  - !!warning!! – intercept may be an extrapolation from the data set

### III. Simple Regression (one x)

- **assumptions** - re data characteristics –
  - generally
    - n independent pairs
    - noise = equal variance b/w data (homoscedasticity) + avg size ( $\mu$ ) = 0 (must be by way LS regression line is fitted) + independent (i.e., not auto-correlated) + normal distribution (Central Limit Theorem)
    - x explains / predicts y (where x = predictor; y = dependent variable)
  - less important = constant variation + normality
  - more important = independence + capturing non-linearities
- **Least Squares regression line** – summarizes relationship b/w x & y
  - $y = b_0 + b_1x$  (where minimizes sum of squared vertical distances b/w regression line and observed pts)
    - $b_0$  = intercept = y when x = 0
    - $b_1$  = slope = y rate of change given change in x (if 0  $\rightarrow$  y is independent of x)
    - **note** –
      - ◆ size of  $b_0$  &  $b_1$  is irrelevant (b/c influenced by units of measure)
      - ◆ regression  $\leftrightarrow$  cause = association
      - ◆ each pt on regression line = estimate of population mean of y given a value of x
      - ◆ regression line w/b better than an “avg” b/c regression line uses weights
  - **smoothing spline** – greater  $\lambda$  = line gets less wiggly
    - BUT - does not provide interpretable model
  - **transformations** –
    - captures non-linear patterns in linear form
    - goal - successful transformation yields scatterplot showing only linear association b/w  $y^*$  and  $x^*$  (where \* indicates transformed value)
    - **log transformation** –
      - ◆ coefficient of  $\log_e X / 100$  = expected change in Y for 1 percent change in X (~ 1% of slope)
      - ◆ intercept = expected Y when  $\log_x = 0$  (i.e., x = 1)
  - **cons** – influenced by outliers
- **signal noise decomposition**
  - **ideal situation** = y observations are treated as independent draws from normal distribution w/ mean  $\mu$  and SD  $\sigma$
  - **model** =  $\mu_y + \varepsilon_i = (B_0 + B_1x) + \varepsilon_i$ 
    - $\mu_y$  = signal =  $B_0 + B_1x$
    - $\varepsilon_i$  = error (~ iid,  $\mu = 0$ , normal distribution) = random error
      - ◆ **assumptions** = independent + equal variance ( $\sigma^2$ ) + normally distributed
    - **goal** = see signal thru noise (find guess at  $y^\wedge$ ) + use multiple estimates / returns to make better predictions
    - **how** –
      - ◆ i) find decomposition of model where  $y = y^\wedge + e_i$  ( $e_i = y_i - y^\wedge_i$ )
        - >  $y^\wedge$  = fitted / predicted values
        - >  $e_i$  = residual
      - ◆ ii) model is good when  $e_i$  manifest iid behavior + residuals are small
- **process**
  - i) observe data pairs
  - ii) !! plot data !! [**impossible to identify desired linearity from equation / parameters alone**]
  - iii) check SRM assumptions (or assume to hold) – look for gross deviations [see p. 2-12]
    - a) linear relationship b/w x & y (~ scatterplot)
      - ◆ **problem1** = non-linearity +
        - > i) is their at least dependence b/w assets – check T ratio and p value
        - > ii) check strength of linear signal – is weak if  $R^2$  is low (i.e., less is explained by model)

- ◆ **problem2** = auto-correlated residuals (~ “momentum” in a time series)
  - ◆ **problem3** = expanding residuals (violates assumption of equal variances)
  - ◆ **problem4** = outlier / influential point
    - > **i)** investigate pt. – identify / explain (~ use JMP pt labels)
      - **leverage** – increases the impact of an outlier + may affect fit even though not influential
      - **influence** – how much regression equation changes when single observation is deleted
    - > **ii)** remove outlier & re-run regression
    - > **iii)** consider – should observation be removed? **there is no test** BUT **Y** if **influential** + has special circumstances
      - **influential** – yes if slope changes significantly when observation removed
    - > **iv)** consider - is there a better model? goal = keep observation in
    - > **v)** if no better model → exclude observation and use this model
  - **b)** look for distortion from influential outliers (~ plot residuals or scatter plot of residuals vs. x → s/b no trends)
  - **c)** check that residuals manifest iid behavior (~ histogram & normal quantile plot of residuals)
    - ◆ **note** – having more residuals provides a better check of normality
  - **iv)** (if necessary) transform data to get (better) linear association
  - **v)** estimate **true regression line** ( $y = B_0 + B_1x$ ) w/ **LS regression line** ( $y^\wedge = b_0 + b_1x$ ) [ $b_0 = B^\wedge_0$ ;  $b_1 = B^\wedge_1 =$  LS estimates]
    - **true regression line** – never known
    - **LS regression line** – best estimate of true regression line →  $y = y^\wedge_i + e_i$ 
      - ◆  $y^\wedge =$  fitted value =  $B^\wedge_0 + B^\wedge_1x_i$
      - ◆  $e_i =$  residual =  $y_i - y^\wedge_i$
  - **vi)** consider causation – which way does it flow are there any other factors/
  - **vii)** can reject  $H_0$ ?
  - **viii)** interpret - retransform data to original units (but not to see fit)
- **RMSE** ( $\sigma_\epsilon$ ) – estimates SD of residuals (measures dispersion of residuals around LS regression line) ~ **variability around fit**
    - $\sigma_\epsilon = [(1 / n-2) (\text{sum } y_i - y^\wedge_i)^2]$
    - s/b better estimate than  $s_y$
    - $RMSE^2 =$  variance of residuals = avg squared deviation b/w data and the LS regression line
    - interpretation – if RMSE holds → 68% of data w/i 1 RMSE; 95% of data w/i 2 RMSE
    - measures fit (in units of the variable being measured) - lower RMSE ~ better
  - $R^2$  ~ goodness of fit ~ proportion of explained variation
    - measures – proportion of variation in response (Y) explained by regression line
    - higher number ~ more successful in explaining variability in response ~ closer points lie to line
    - BUT – is relative (so goodness depends on relativeness)
    - $= (1 - \text{Residual SS} / \text{Total SS}) = (\text{Total SS} - \text{Residual SS}) / \text{Total SS} = \text{Regression SS} / \text{Total SS}$
    - $= (\text{correlation b/w predictor and response})^2$
    - for a change in  $R^2$  → consider change in relative % of unexplained variation
  - **confidence intervals**
    - become smaller as sample size grows
    - slope estimate –
      - SE (estimated slope) gets smaller as N gets larger
      - as predictors are more dispersed → SE falls
      - as observations cluster around line → SD (error about line) and SE (estimated slope) get smaller
  - **remember**
    - line v spline v equation – line <> work for non-linear relation + equation can be optimized
    - more data → LS regression line gets closer to true regression line
    - regression line = signal IFF data is infinite (????)
    - **degrees of freedom**
      - n-1 – need 2 pts to find  $\sigma$  → n-1 eliminates when only have one pt
      - n-2 – need 3 pts to find variance from line → n-2 throws out first 2 pts
    - **leverage points** –
      - outlier in x direction – important b/c this is the independent variable / predictor
      - may not be influential BUT are the best candidates for being influential

- **statistical prediction intervals penalize for statistical error** (~ assume that fitted model holds even when extrapolated) – (model selection error / extrapolation penalty)

#### IV. Multiple Regression

- **key terms** -
  - **partial slope** – change in response for unit change in predictor, holding other predictors constant
  - **marginal slope** – change in response for unit change in predictor, not holding other predictors constant
  - **leverage plot** – shows association b/w each predictor / response one at a time
- **process** -
  - **i) regress data**
  - **ii) if too small increments, rescale**
  - **iii) consider fit** –  $R^2$ , RMSE, SE, t-stat, etc.
  - **iv) check diagnostics - save residuals**
    - **a) non-linearity**
      - ◆ **i) is there at least dependence b/w assets** – check T ratio and p value
      - ◆ **ii) check strength of linear signal** – is weak if  $R^2$  is low (i.e., less is explained by model)
    - **b) auto-correlated residuals** (~ “momentum” in a time series)
    - **c) heteroscedasticity** - expanding residuals (violates assumption of equal variances)
    - **d) outlier / influential point - re-consider parameter estimates (t-ratio, p-value) after excluding any points**
      - ◆ **i) Analyze + Multivariate + Analyze Outliers** – s/b only few dots above line
      - ◆ **ii) leverage plot** –
      - ◆ **iii) BUT** – leverage plots tell us the most about the slope so long as they are consistent with the data
      - ◆
      - ◆ **i) investigate pt.** – identify / explain (~ use JMP pt labels)
        - > **leverage** – increases the impact of an outlier + may affect fit even though not influential
        - > **influence** – how much regression equation changes when single observation is deleted
      - ◆ **ii) remove outlier & re-run regression**
      - ◆ **iii) consider** – should observation be removed? **there is no test** BUT **Y** if **influential** + has special circumstances
        - > **influential** – yes if slope changes significantly when observation removed
      - ◆ **iv) consider** - is there a better model? goal = keep observation in
      - ◆ **v) if no better model** → exclude observation and use this model
  - **v) consider relevance of predictors**
    - **a) correlation** –
      - ◆ ~ sensitive to outliers
      - ◆ as correlation rises → SE goes up as well
      - ◆ **i) Analyze + Multivariate** – see correlation matrix + ellipses in scatterplot matrix (higher correlation if narrow ellipse + tilted in 45 degree angle)
      - ◆ **ii) Spinning plot** – shows if data is in specific region (~ due to collinearity)
    - **b) SE of slope** –
      - ◆ determined by – error variation around fitted line + n + unique variation in predictor
      - ◆ correlation reduces the effective range of a single variable for estimating the coefficient of that variable (~ restricted to specific amount of other variable, less variation in the predictor is available); note – model may fit better but have a higher SE
    - **c) leverage plot** – shows contribution of each predictor to the multiple regression
      - ◆ distances pt to line = multiple regression residuals
      - ◆ shows sequence of simple regressions reflecting how each predictor enters the model
      - ◆ **slope of variable = significant if horizontal line ever lies outside the confidence bands**
      - ◆ help to identify outliers (especially if all outliers have same special circumstance)
    - **d) partial F statistic associated w/ added predictors** – measures how much residual explained by addition of extra predictors
      - ◆ = (change in  $R^2$  / number of extra predictors) / [(1 – new  $R^2$ ) / residual degrees of freedom] – p. 152
      - ◆ ROT –  $F > 4$  --> F ratio is significant

- vi) interpret
  - a) prediction –
  - b) prediction interval -
    - ◆ is narrower when more variation is explained
  - c) partial F statistic associated w/ added predictors – measure
- vii) consider practical issues (e.g., need more people to run more machines?)
- remember -
  - diagnostic key = residuals
  - leverage plots – allow diagnosing multi regression as sequence of bivariate plots
  - models = more easily interpreted w/ no correlation b/w predictors

## V. Collinearity

- **collinearity** ~ correlation among predictors
  - size of coefficient of one variable depends on the presence of other variables in the model
  - is a column based problem
  - makes regression coefficient values unstable
- **Variance Inflation Factor (VIF)** – index of collinearity on the variance of coefficient estimates in regression
  - tells how much variability inflated due to collinearity
  - VIF = 9 implies →
    - SE of coefficient estimate = 3 times larger than would be were the predictor unrelated to other predictors
    - CI for coefficient = 3 times longer than . . .
  - indicates the multiplicative factor by which each slope estimate is inflated by collinearity (p. 145)
- **F ratio** –
  - measures efficiency of adding variable to regression and converting it into a slope (~ by measuring the incremental amount of variance explained??)
  - analyzes one variable at time (~"honest")
- **Beta** –
  - ~ slope associated with the regression of price of individual stock on a market index
  - used to asses performance / risk of stock relative to market
- **diagnostics (134)**
  - i) scatterplot matrix – look for strong linear relationship b/w predictors
  - ii) counterintuitive signs on regression coefficients
  - iii) large SE on some regression coefficients – b/c little info to estimate them
  - iv) leverage plots – observations fall within narrow band in middle of plot ~ indicative of collinearity
  - v) high VIF – the actual harm dependson the resulting SE
  - vi) low value of t stats despite significant overall fit (~ F Stat)
- **interpretation**
  - if asking about
    - a) one predictor → **t stat**
    - b) some subset of predictors → **partial F test** (p. 127)
    - c) all predictors → **F stat** (from annova summary) – shows whether some combination of predictors is useful (but does not say which ones)
- **solutions to collinearity**
  - i) ignore - collinearity not necessarily bad if only predicting w/i data
  - ii) combine - collinear data into single index (BUT check that still meaningful)
  - iii) remove – one of the collinear predictors

- **multiplicative model** (???)
  - used to model production problems
  - $Y = \alpha X_1^{B_1} X_2^{B_2} \sigma$
  - expressed in logs → regression coefficients are interpreted as elasticities
  - note – multiple regression estimates the elasticities in the joint production function
    - marginal elasticity – accounts for all other changes which normally occur when one predictor changes
    - partial elasticity – holds all other predictors constant
  - to test the addition of predictors – check partial F stat (see above) (if  $>4$  → then significant)
- **remember** -
  - degrees of freedom (df) – we lose degree for (i) computing sample mean + (ii) estimating constant in regression + (iii) estimating the slope in regression
  - collinearity ~ less precise slope estimates b/c of loss of uncorrelated variation along x-axis

## VI. Categorical Variables

- **Categorical Variables** – takes attributes (<> numbers)
  - regression's production of parallel lines
    - ~ implies difference b/w categories does not depend on the relative amount of the predictor (i.e., impact of category <> depend on level of predictor)
    - JMP value (+/- difference from avg of both groups)
- **interaction** – impact of a predictor ( $X_2$ ) on the response (Y) depends on the level of another predictor ( $X_1$ )
  - relaxes assumption that categorical regression lines are parallel - allows each regression line to have separate slope and intercept
  - requires three variables
  - ~ identifying synergy
- **analysis of covariance** - used to check stat significance of difference b/w intercepts of parallel fitted lines
  - regression analysis combining categorical variable w/ other covariates
- **two sample t-test** – [Fit Y by X] + [Means / Anova / t-test]
- **interpretation**
  - **i)** regress response by predictor
    - **a)** distinguish observations (~ color)
    - **b)** is there a relationship?
  - **ii)** group by categorical variable
  - **iii)** fit lines (1 for each category)
  - **iv)** compare equations
    - **a)** are slopes same (i.e., lines parallel)?
      - ◆ check interaction b/w predictors – if coefficient <> stat significant → conclusion = slopes parallel
    - **b)** is difference in intercept stat significant?
      - ◆ a) add intercept adjustment factors to obtain total difference b/w intercepts
      - ◆ b) calculate t-ratio - if difference  $> 2$  SE → stat significant
  - **v)** check model significance – 2 levels = t-stat; 2+ levels = **partial F-stat**
    - **partial F-stat** = change in  $R^2$  per added variable / remaining variation per residual
  - **vi)** add interactions – to test significance of non-parallel slopes
  - **vii)** check linear fit & residual assumptions –
    - no auto-correlation,
    - equal variance (homoscedasticity),
    - normal distribution,
    - outliers – consider effect relative to size + consider relative proximity to other observations + exclude & compare results
- **remember**
  - interaction – continuous variable is centered in the interaction term (by subtracting the mean) to reduce collinearity

---

## VII. Model Building

- **step-wise regression** – “greedy” + adds var one at time (no removal) + make  $R^2$  grow as quickly as possible
- **response surface** – regression models that include all possible interactions and quadratic terms for each predictor
- **process**
  - **i)** regress observations
  - **ii)** check residual assumptions
  - **iii)** consider – what other factors explain residual variation
    - **a)** common sense
    - **b)** scatterplot matrix
  - **iv)** step-wise regression = **(a)** fit model platform + **(b)** pick additional relevant predictors – [**Effect Macros**] + [**Response Surface**]
    - **forward stepwise** – includes predictors one at a time (no removal)
    - **backward stepwise** – includes all and excludes non-significant
    - **partial forward** – lock in specific predictors + forward step wise
  - **v)** interpret results
- **validating regression models**
  - cross validation – sub-sample set aside for later application of derived model
  - **Bonferoni inequality** – utilizes stricter rule for assessing significance (and T/F including a variable)
    - = standard p-value / number of predictors
    - risk – overly conservative + JMP may not allow p-value so low (BUT OK if more efficient)
- **risks**
  - **data mining** – automated model fitting that derives algorithm for max increase in  $R^2$  (problem = misses relevant relationships in face of collinearity)
  - **over-fitting** – high  $R^2$  but solely due to chance (high F-ratio is misleading) (solution = low p-value)
- **remember**
  - comparing  $\log_{10}$  transformations = elasticity
  - statistical significance = p-value
  - substantive significance = RMSE