Ch15: Multiple Regression
12.1: One-way ANOVA

- *Please download: 10-MultRegr.xls*
- *HW7 this week*
- *Projects*

20 Mar 2012
Dr. Sean Ho

busi275.seanho.com

Trinity Western University

Outline for today

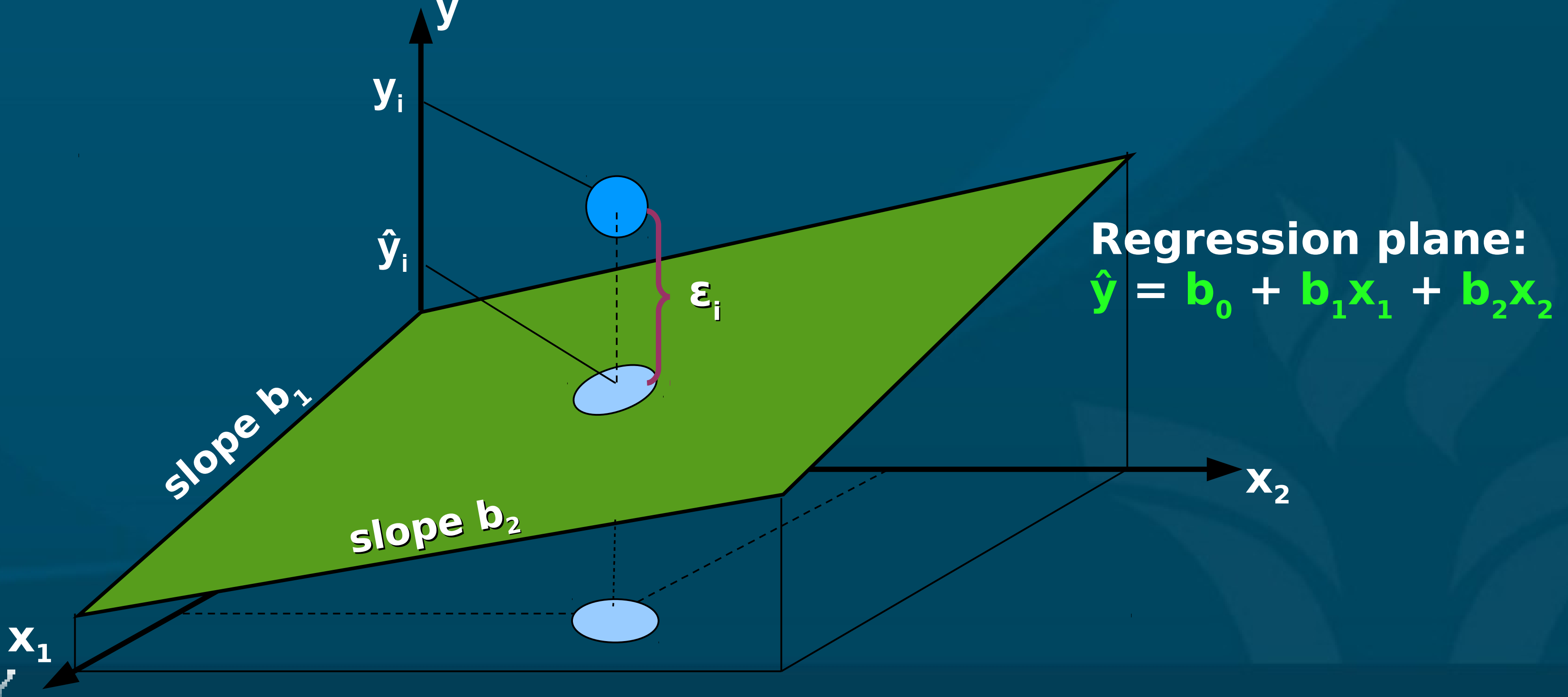
- Multiple regression
  - Interpreting Analysis ToolPak output
  - Ranking predictors
  - Moderation (interaction of predictors)
  - Regression diagnostics: check assumptions
  - Transforming variables
- One-way ANOVA
  - Assumptions & concepts: between vs. within
  - Global *F*-test
  - Follow-up analysis with Tukey-Kramer

Summary of hypothesis tests

- Distributions:
  - Sample mean with σ known → NORM
  - Sample mean with s known → T
  - Binomial proportion (dichot var) → NORM
- Data collection:
  - One sample (compare vs. threshold)
  - Paired data (compare vs. 0 diff)
  - Two samples (dichot. IV, compare means)
  - Multiple samples (nom. IV → 1-way ANOVA)
  - Multiple quant. vars (correl + regression)
- Conf. Int. (critical value) vs. *p*-value approach

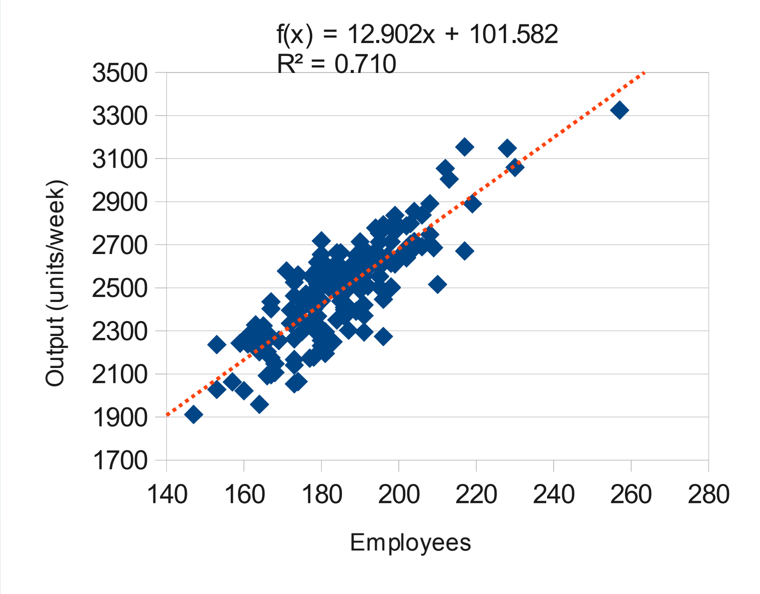
Multiple regression

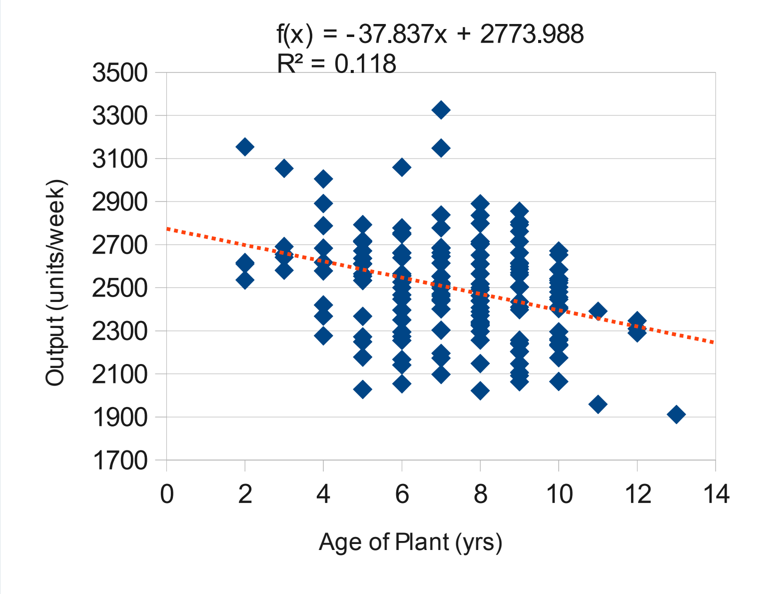
- 1 outcome (scale), k predictors (scale)
- Linear model: hyperplane
  - $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$
- Residuals still assumed normal, homoscedastic



Multiple regression in Excel

- Dataset: 10-MultRegr.xls
- DV (y): Output (units/wk)
- IV (x_1): Employees
- IV (x_2): Age of Plant (yrs)
- Pairwise scatters are helpful
  - Note R^2 for each predictor



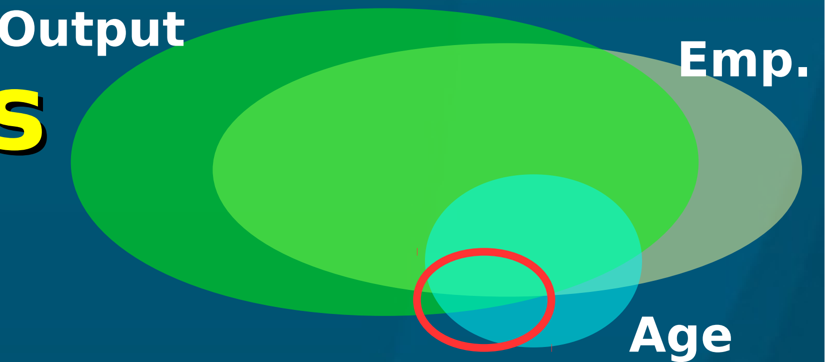


- Data → Analysis → Regression
  - Y Range: B1:B160
  - X Range: C1:D160
  - Check "Labels" and "Standardized Residuals"

Interpreting the output

- R Square (R^2): fraction of DV var explained
  - Adjusted R^2 compensates for adding more IVs
- ANOVA table: F, p, and dfs
  - "Number of employees and plant age significantly predicted output: $R^2 = .72$, $F(2, 156) = 200.7$, $p < .001$."
- Coefficient table:
  - For each predictor: slope b_i, t-score, and p
  - Both slopes are significantly nonzero
- Standardized residuals: z-scores
  - Can use to look for observations that don't fit the model (e.g., $|z| > 3$)

Unique contributions



- From the Employees scatter, it predicts Output pretty well (R^2 = 71%)
- Age? Not so well (R^2 = 12%)
- When use both together, why is R^2 only 72%?
  - Most of the 12% of variability in Output explained by Age is shared variability:
  - Age doesn't tell us much more about Output than we already knew from Employees
  - Age's unique contribution is only 1%
- Compare regression using all predictors against regression using all except Age

Drawing conclusions

- We see that Employees and Age do significantly predict Output (global *F* test), and
- Each predictor does contribute significantly (*t*-tests on slope), but
- The unique contribution of Age is very small, so
- Most of the predictive power is in the number of employees.

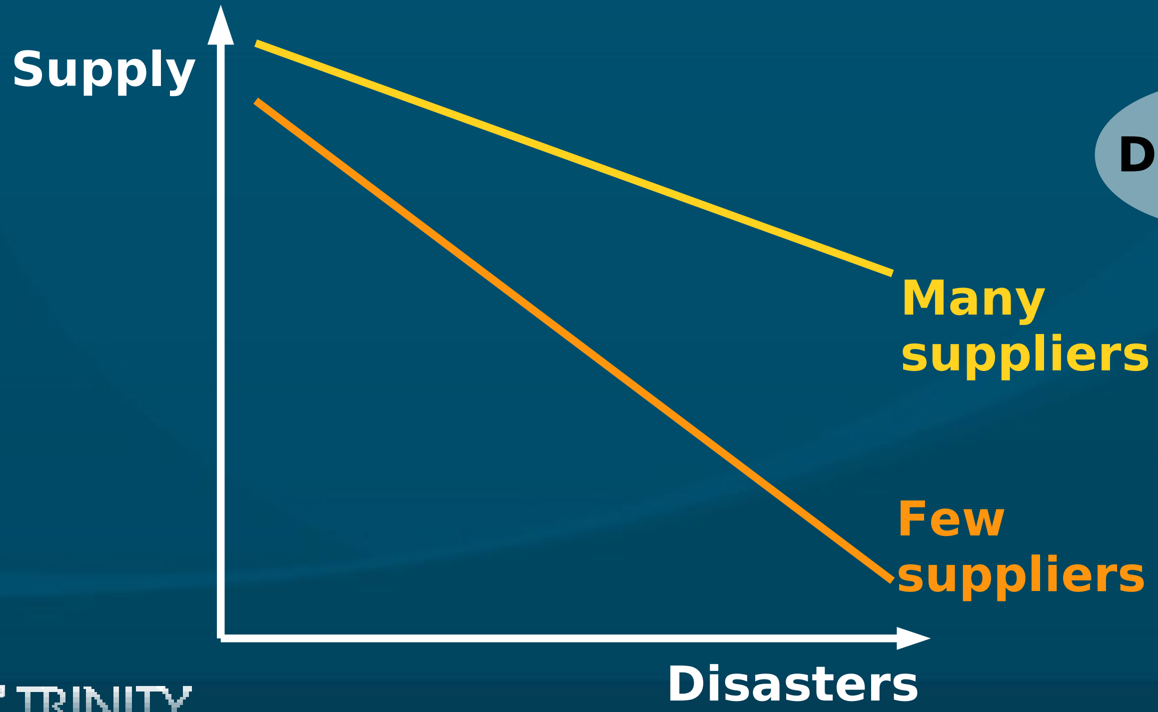
- In a formal write-up, you usually want to include details such as R^2, F, dfs, and p, for those who understand the stats and might want to replicate your results.

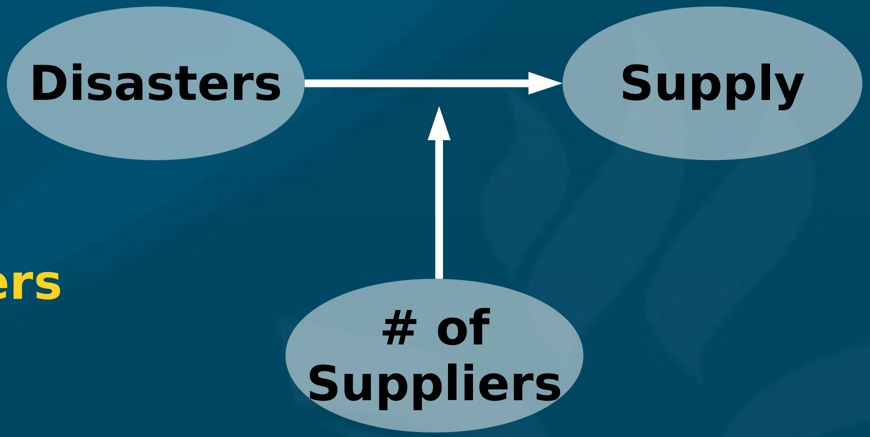
Outline for today

- Multiple regression
  - Interpreting Analysis ToolPak output
  - Ranking predictors
  - Moderation (interaction of predictors)
  - Regression diagnostics: check assumptions
  - Transforming variables
- One-way ANOVA
  - Assumptions & concepts: between vs. within
  - Global *F*-test
  - Follow-up analysis with Tukey-Kramer

Moderation

- Moderator: a predictor that affects the strength of another predictor's influence on the outcome
  - Interacts with the other predictor
- E.g., natural disasters may affect your supply, but having multiple suppliers buffers the effect





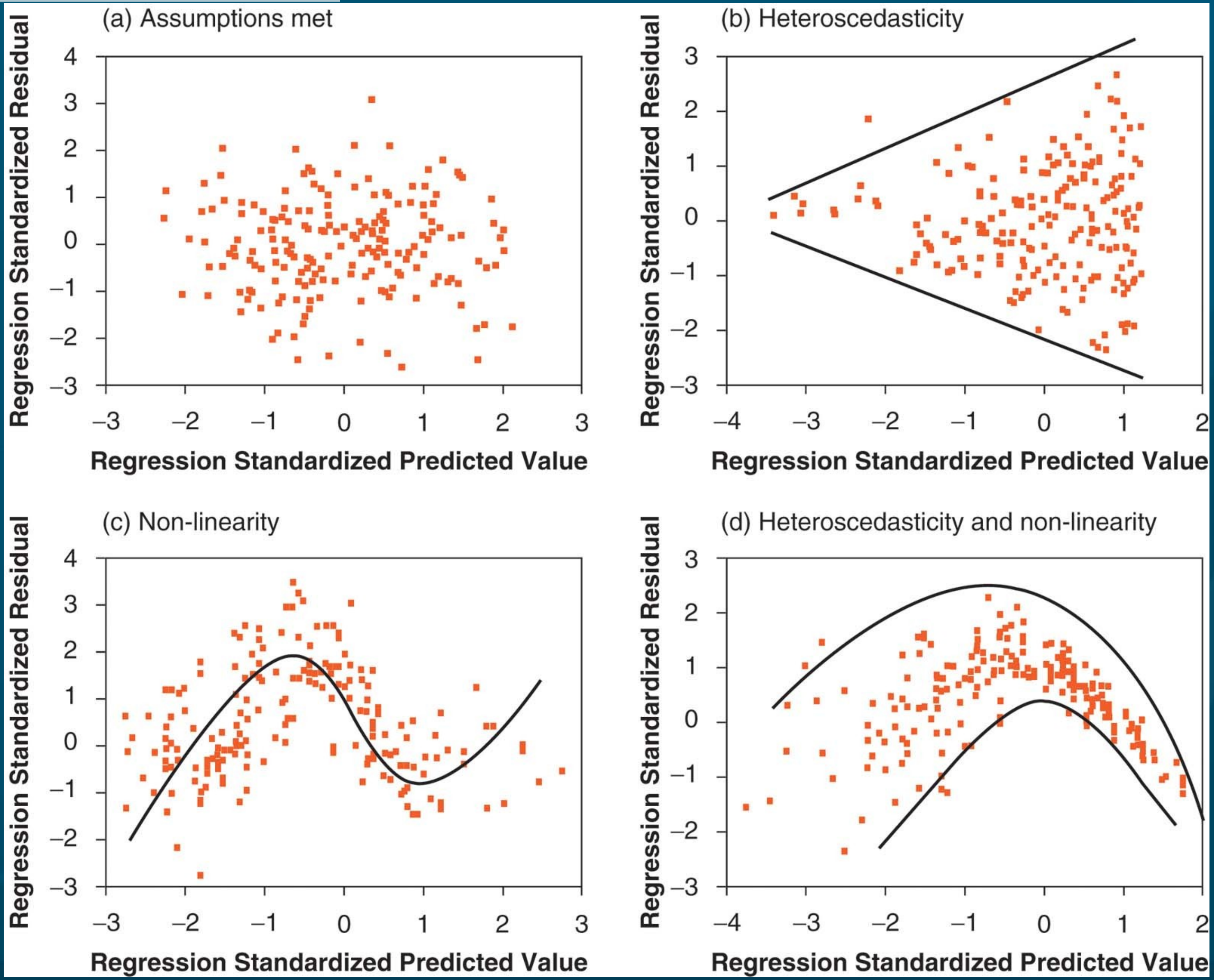
Testing for moderation

- How do we know if predictors are interacting?
- Add an interaction term to the regression:
  - $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$
- In Excel, centre both IVs (subtract their means), then make a 3rd column with the product
  - Include it in the regression as if it were an IV
- Check the *t*-test to see if the slope (b_{12}) of the interaction term is significantly nonzero
- If so, check R^2 both with and without the interaction term to see the size of its effect
- Also 3-way ($x_1 x_2 x_3$) and higher interactions!

Diagnostics: check assumptions

- Normality of residuals:
  - Check histogram of standardized residuals
- Homoscedasticity:
  - Residual plot: residuals vs. predicted values
  - Look for any odd or "fan shaped" patterns
- Linearity: curves on the residual plot
  - Try adding x_1^2 or x_2^2, etc. to the model
  - And/or apply transforms to variables
- Indep. of residuals (time series are usually bad)
- Collinearity of IVs: check correlations of IVs
- Outliers / influential points: see residual plot

Homoscedasticity & linearity

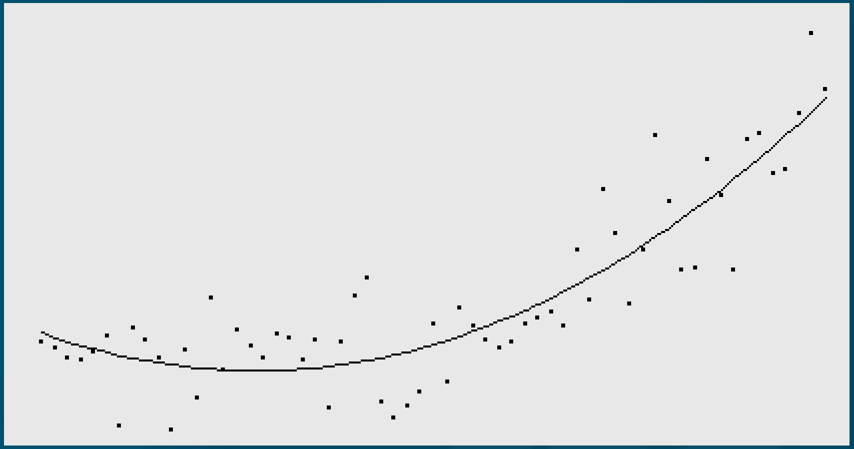


Transforms

- Some variables (either IVs or DV) may be so heavily skewed that they break assumptions (esp. heteroscedasticity and nonlinearity)
- You can try applying a transform to make them roughly more symmetric or normal
  - But strict normality is not required
  - E.g., log(income) is usually more normal
- The family of power transforms includes:
  - $\sqrt{x}$, x^2, $1/x$, $x^{-5.2}$, etc., as well as $\log(x)$
  - May need to shift ($x+c$) or reflect ($c-x$) first
  - The Box-Cox procedure "automatically" selects a power transform for your variable

Polynomial regression

- Add terms to the regression equation: e.g.,
  - $\hat{y} = b_0 + b_1 x + b_2 x^2$ (quadratic regression)
- Allows the regression line to curve
  - Quadratic ⇒ parabola



- Check whether quadratic term is needed:
  - t-test on its slope b_2 to check its significance
  - R^2 and adjusted-R^2 to see its contribution
  - partial *F*-test: model w/ vs. w/o the term

Automated predictor selection

- Adding predictors always increases R^2, but
  - Want to find the smallest set of predictors that still explains the outcome variable
  - Parsimony: simpler model to understand
- "Best subsets" automatic selection:
  - Several random combinations of predictors
- "Stepwise" regression adds/removes 1 predictor at a time to try to do the same
  - Backward: eliminate the least significant IV
  - Forward: add the next most significant IV
- Use PHStat add-on, or SPSS, Stata, R, etc.
- Not magic! What do the vars mean?

Outline for today

- Multiple regression
  - Interpreting Analysis ToolPak output
  - Ranking predictors
  - Moderation (interaction of predictors)
  - Regression diagnostics: check assumptions
  - Transforming variables
- One-way ANOVA
  - Assumptions & concepts: between vs. within
  - Global *F*-test
  - Follow-up analysis with Tukey-Kramer

ANOVA: Analysis of Variance

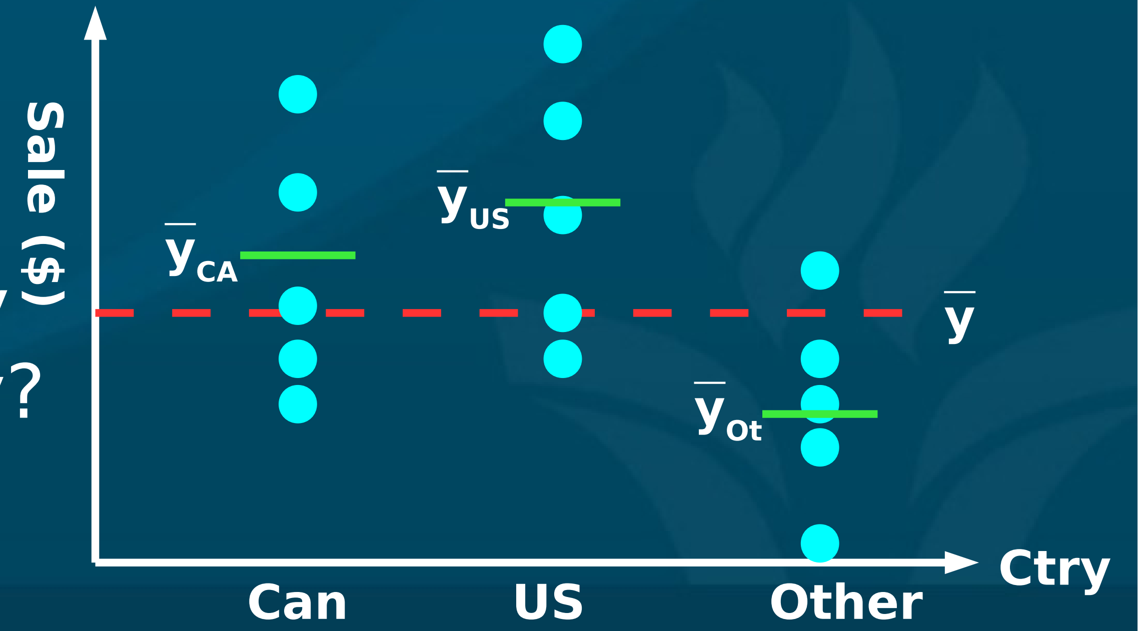
- 1 DV (scale) and one or more IVs (nominal)
  - One-way ANOVA: just one IV, with *k* levels
  - e.g., does country affect avg purchase amt?
    - Groups: Canada, US, China, UK, etc.
- The independent-groups t-test is a special case
  - One IV that is dichotomous
- ANOVA performs one global F-test to assess if the predictor has any effect on the outcome
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - Omnidirectional (generalization of 2-tailed)
- Follow-up tests then identify which groups differ

ANOVA: assumptions

- DV is continuous
  - If DV is dichotomous, try Logistic Regression
  - If all vars are nominal, try Log-Linear analysis
- Observations are independent, and Groups (levels of the IV) are independent
- DV is normally distributed within each group
  - If not, try transforming the DV
- Variance (SD) of DV in each group is roughly similar across all the groups (homoscedasticity)
  - Not crucial if n in each group is large and if balanced design: similar n in each group

ANOVA: concepts

- How much of variability in purchase amount is due to country of origin?
  - $SS_{tot} = SS_{Country} + SS_{residual}$
  - $SS_{Country}$ is "between-group" variation (SSB)
  - $SS_{residual}$ is "within-group" variation (SSW)
- Do the group means differ significantly?
  - F-test, p-value
- Fraction of variability explained by country?
  - η^2 (equiv. to R^2)



ANOVA table

- Model: Y = (offset due to group) + (residual ε)

-	**Group (Between)**	**Residual (Within)**
SS	$SSB = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$	$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$
df	**k – 1**	**n – k**
MS = SS/df	**MSB = SSB / (k – 1)**	**MSW = SSW / (n – k)**

- Test statistic is F = MSB / MSW
  - Model vs. residual (as in regression!)
  - Use FDIST() with two dfs to get p-value

Example: Delivery minivans

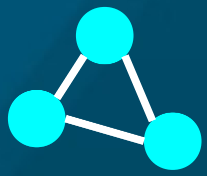
- Dataset: "Delivery" in 10-MultRegr.xls
  - (See p.496, #12-15)
- DV: operating cost per mile
  - IV: manufacturer (3 companies)
  - Unit of observation: one minivan (total $n=13$)
- ANOVA table: df = (2, 9)
  - SS = (6.07, 3.45) ⇒ MS = (3.04, 0.38)
  - ⇒ $F = 7.91$ ⇒ $p = 0.010$
- Reject H_0: operating costs per mile do differ significantly depending on manufacturer

Outline for today

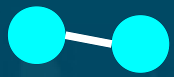
- Multiple regression
  - Interpreting Analysis ToolPak output
  - Ranking predictors
  - Moderation (interaction of predictors)
  - Regression diagnostics: check assumptions
  - Transforming variables
- One-way ANOVA
  - Assumptions & concepts: between vs. within
  - Global *F*-test
  - Follow-up analysis with Tukey-Kramer

Follow-up analysis

- ANOVA's global F test is an omnibus test:
  - Just says there is a difference somewhere
  - Doesn't tell us which groups differ!
- There may be sets of groups that don't differ significantly from each other



- Follow-up analysis tries to find these



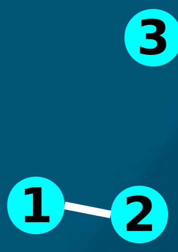
  - Post-hoc: try all pairs of groups
    - The multiple comparisons problem: a blind “shotgun” approach leads to inflated Type I error
  - Planned contrasts: if theory guides us to try certain comparisons of groups

Post-hoc: Tukey-Kramer

- Considers all possible pairings of groups
  - (Can vs. US), (Can vs. Other), (US vs. Other)
  - In general, k*(k-1) pairings!
- From table in Appendix J, find critical value for q
  - Test statistic for studentized range (like F)
  - Use α (.05 or .01) and both dfs to look up
- For each pairing (group i vs. group j):
  - Find standard error: $$SE = \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$
  - Compare difference of means: $|x_i - x_j|$ against critical range: (q)*(SE)
  - If larger, then these groups differ significantly

Delivery ex.: Tukey-Kramer

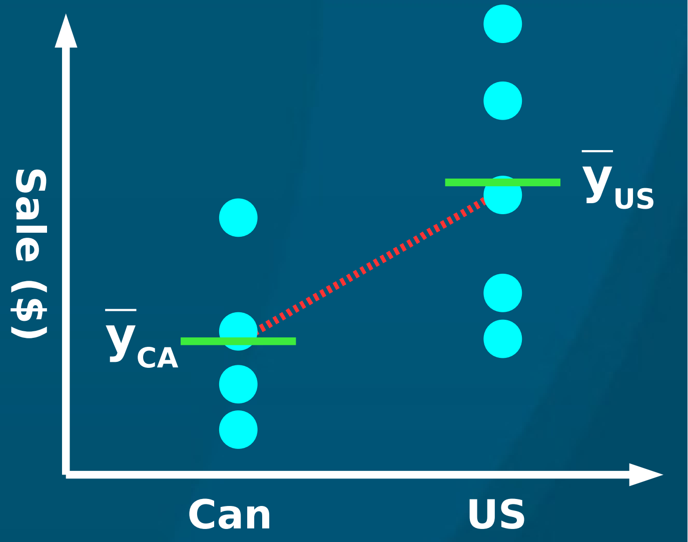
- Which manufacturers differ significantly?
- Appendix J (p.867): α=0.05 (95% conf)
  - df = (2, 9) ⇒ q = 3.20
- Calculate SE for each pairing
- Calculate critical range for each pair: q*SE
- Compare against mean differences:



q	3.20		
Pair:	1 vs 2	1 vs 3	2 vs 3
SE	0.303	0.279	0.317
Crit Range:	1.024	0.940	1.071
Mean diff:	0.633	1.175	1.808
Result	FALSE	TRUE	TRUE

- Conclusion: manufacturer 3 is the odd one out, with significantly higher operating costs

ANOVA vs. regression



- With only 1 dichotomous IV:
  - ANOVA = t-test = regression
  - Code the IV as 0/1
    - Intercept b_0 = mean of group 0 ($\bar{y}_0$)
    - Slope b_1 = difference of means
  - Effect size $\eta^2 = R^2$
- If the IV has multiple levels, use dummy coding:
  - Choose a reference level
  - Make k-1 dummy variables, for each of the other levels: each coded 0/1
  - Use multiple regression

Cty	US	Ot
Ca	0	0
US	1	0
Ot	0	1

TODO

- HW7 due Thu
- Projects: be pro-active and self-led
  - All groups have passed REB by now
  - Presentations on 10Apr (3 weeks from now!)
  - Remember your potential clients:
    what questions would they like answered?
  - Tell a story/narrative in your presentation