ch15: Multiple Regression

- **HW7 due Tues**
- **Please download: 17-Hawlins.xls**

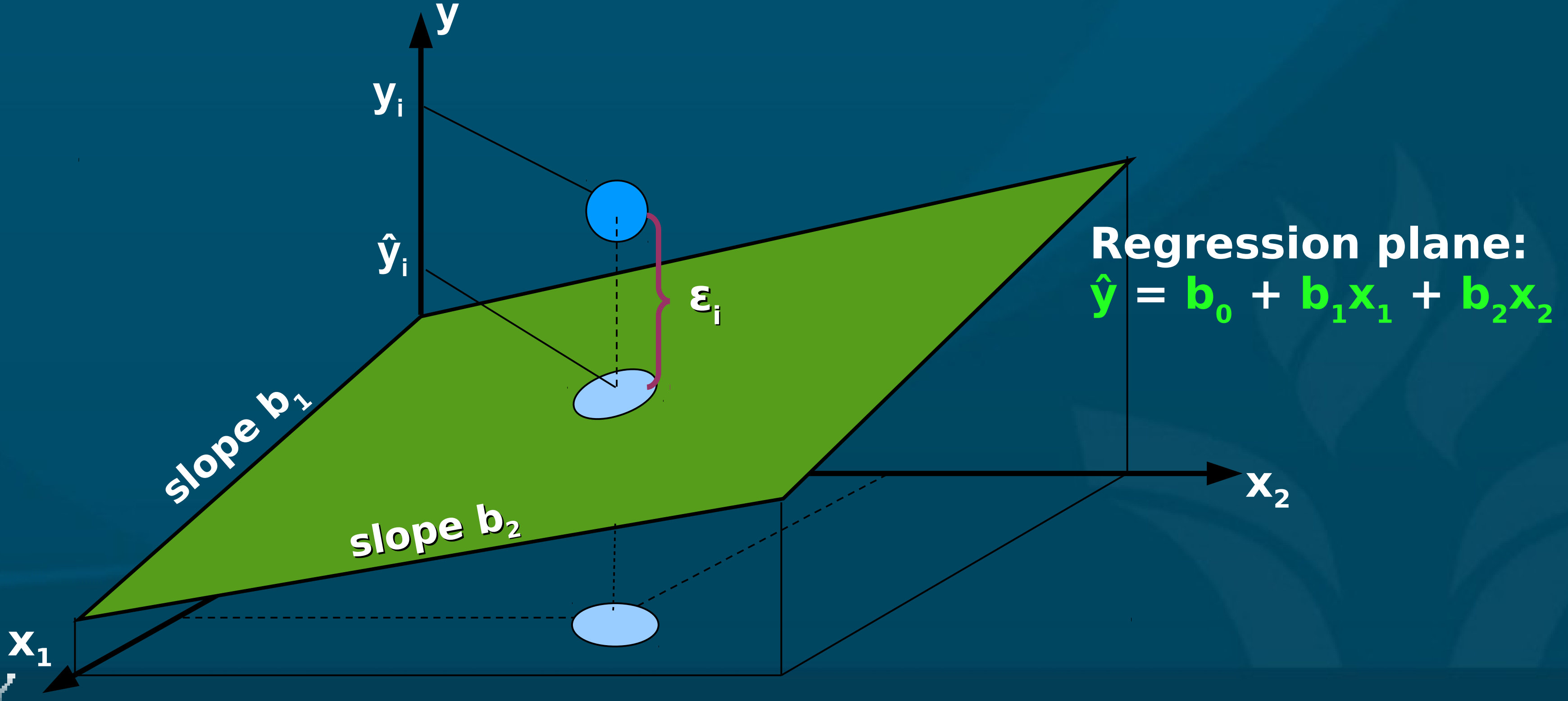
3 Nov 2011
BUSI275
Dr. Sean Ho

Outline for today

- Multiple regression model
  - Running it in Excel
  - Interpreting output
- Unique contributions of predictors
  - Automated predictor selection
- Moderation (interaction of predictors)
  - How to test for it
- Regression diagnostics: checking assumptions
  - Transforming variables

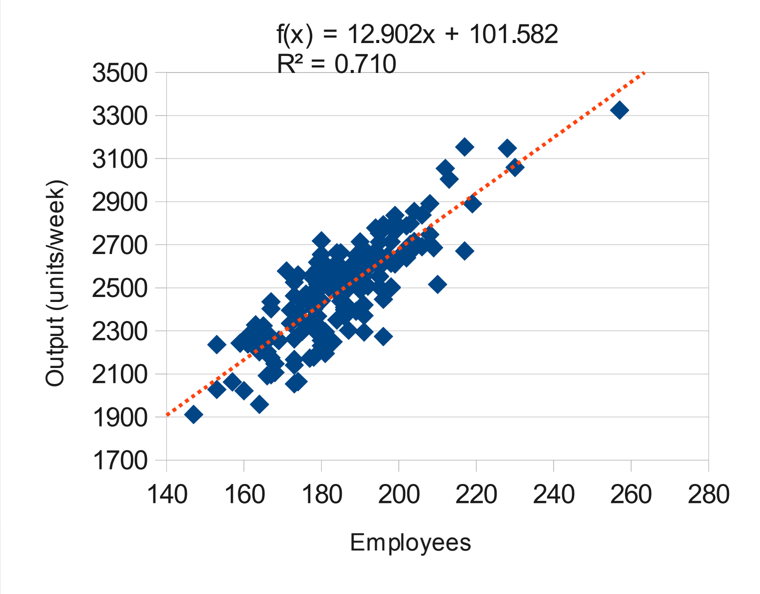
Multiple regression

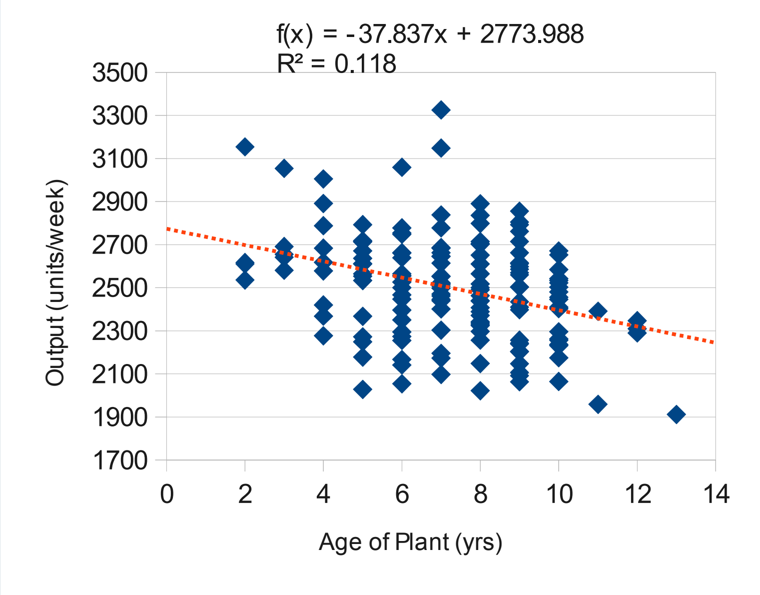
- 1 outcome (scale), k predictors (scale)
- Linear model: hyperplane
  - $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$
- Residuals still assumed normal, homoscedastic



Multiple regression in Excel

- Dataset: 17-Hawlins.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”

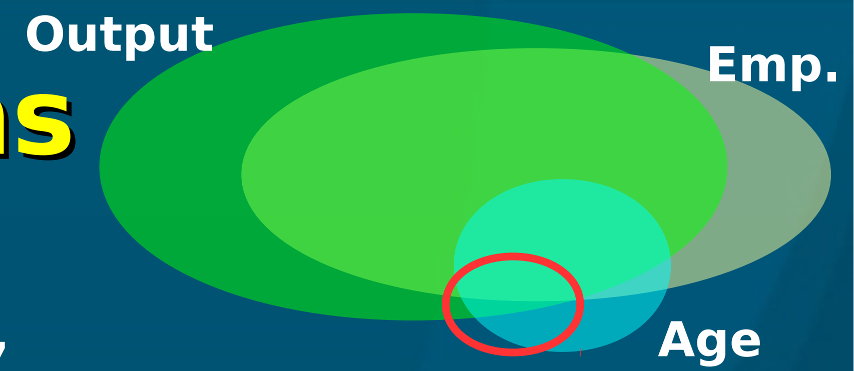




Interpreting the output

- 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.

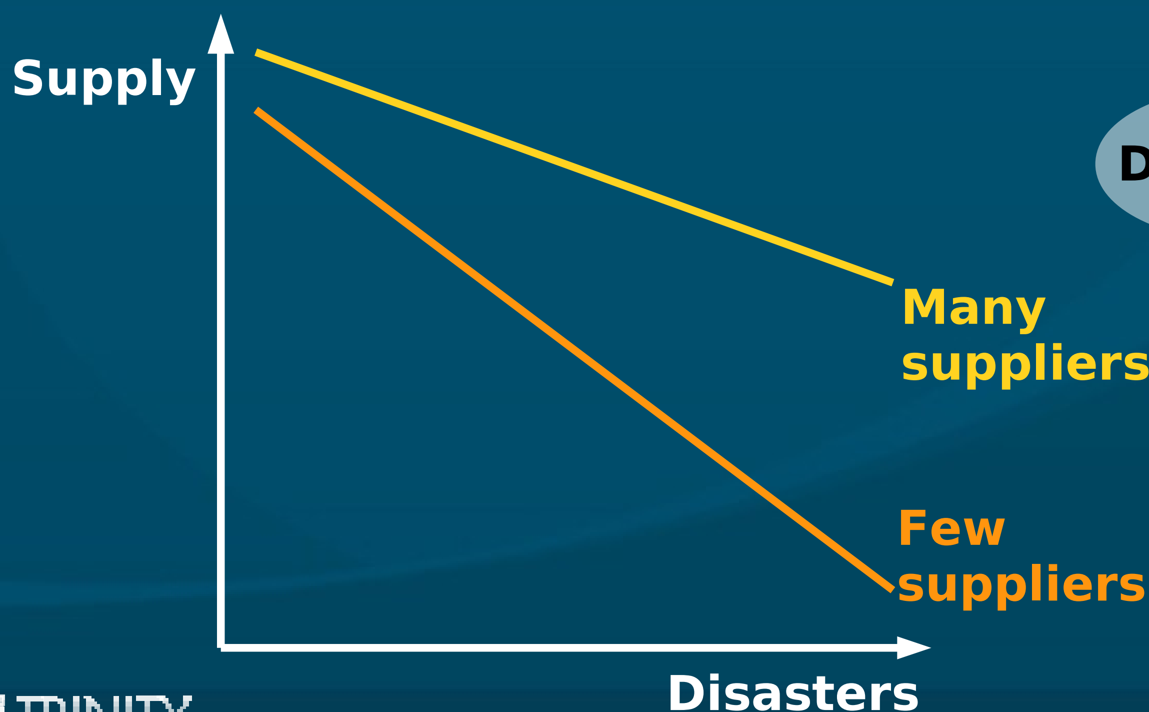
- In a formal write-up, you usually want to include details such as R^2, F, dfs, and p, for those who understand the statistics.

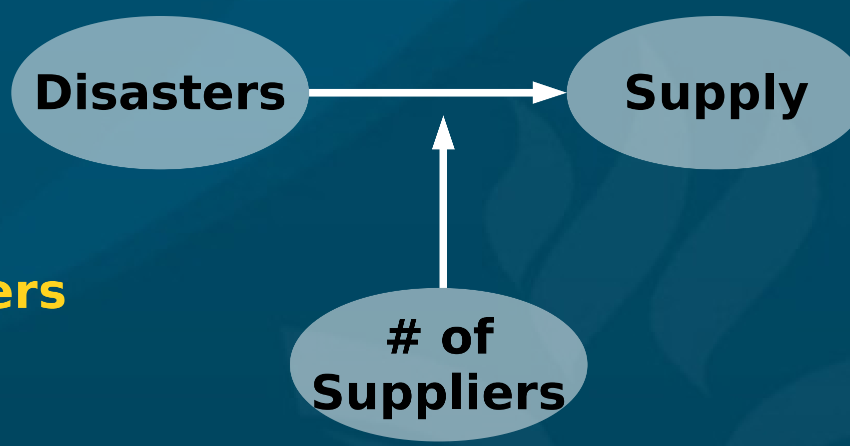
Automated predictor selection

- "Best subsets" regression uses several runs with different combinations of predictors to try to find the set that predicts best while using the fewest predictors
  - Parsimony: simpler model to understand
- "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
- Only in PHStat add-on, or SPSS, Stata, R, etc.

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





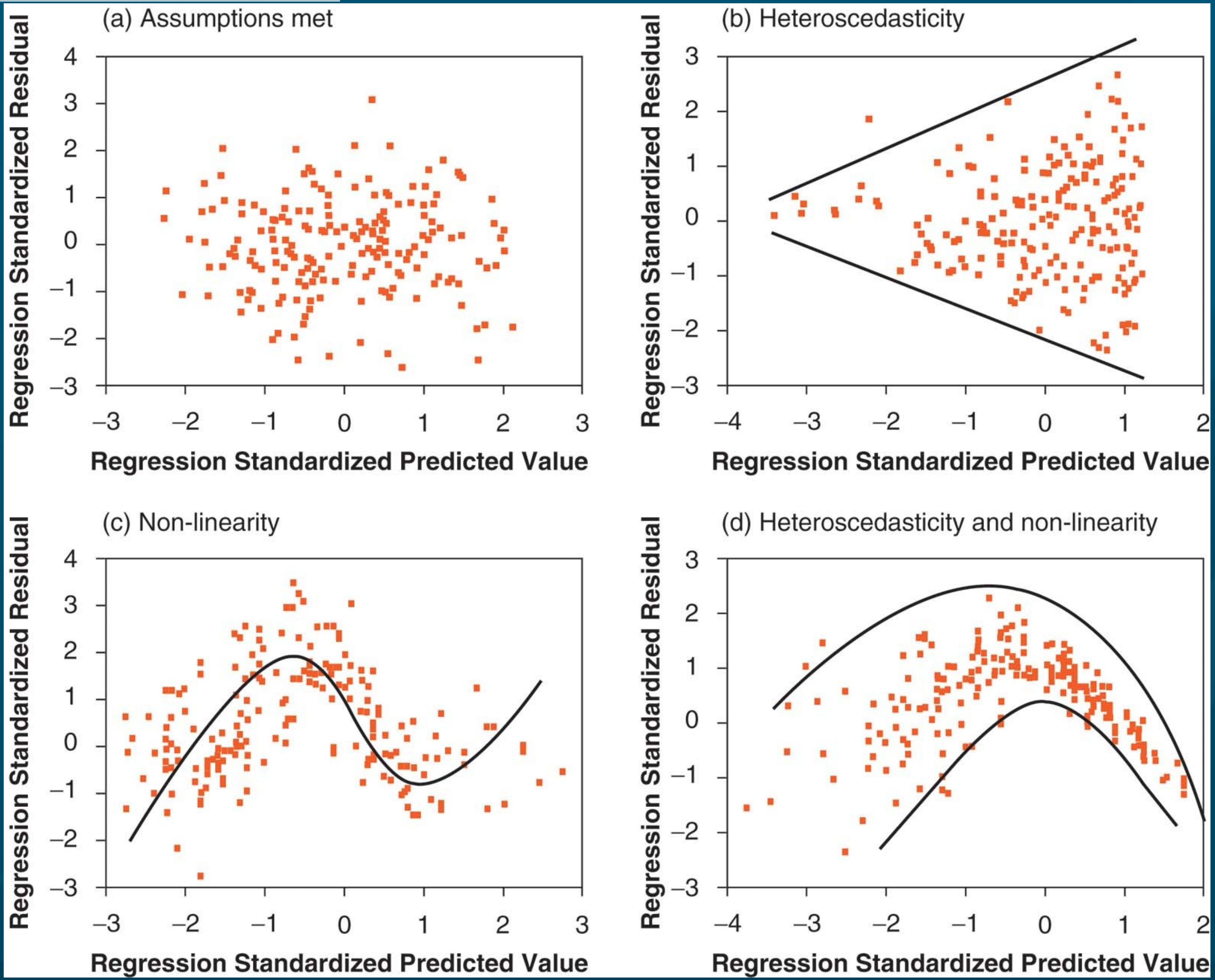
Testing for moderation

- 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

TODO

- HW7 (ch10,14): due Tue 8 Nov
- Projects:
  - Acquire data if you haven't already
    - If waiting for REB: try making up toy data so you can get started on analysis
  - Background research for likely predictors of your outcome variable
  - Read ahead on your chosen method of analysis (regression, time-series, logistic, etc.)