Stat 102

Introduction to Business Statistics
Class 16

Richard P. Waterman

Wharton

Table of contents I

1. Today's module
2. Last time
3. Leverage plots
    - Graphing the multiple regression coefficients
    - Constructing a leverage plot
4. Two-level categorical predictors
    - Categorical predictors: introduction
5. Interaction in regression analysis
6. Summary
7. Next time

Today's module

Topics to be covered in this module:

- Last time
- Leverage plots
- Introduction to categorical predictor variables (two-levels)
- Summary
- Next time

Last time

- Correlated predictors – collinearity
- The market model
- Diagnostics for collinearity
- Fix-ups for collinearity

Leverage plots

- Leverage plots display the partial regression coefficients
- They show the *direct effect* for each variable in the multiple regression
- There is one leverage plot for each variable in the regression
- Because they display partial association, they can only be constructed after the multiple regression has been run
- They play the role of a scatter-plot in simple regression, allowing you to diagnose the multiple regression, a variable at a time
- They are *clever* two-dimensional views of a high dimensional regression plan

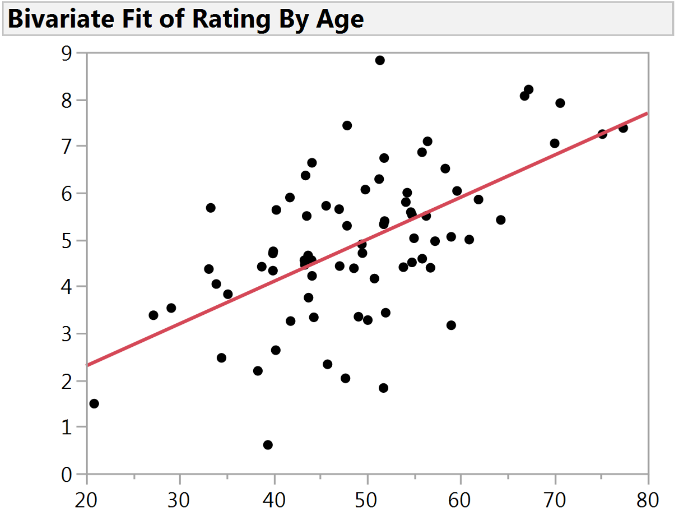
Using a leverage plot

Potential uses for a leverage plot include (everything you can do with a scatter-plot in simple regression):

- Assess linearity
- Spot leveraged data points
- Spot points with large residuals
- Assess heteroscedasticity
- Diagnose collinearity (a collapsed leverage plot)
- A useful graphical tool for explaining partial association to someone else.

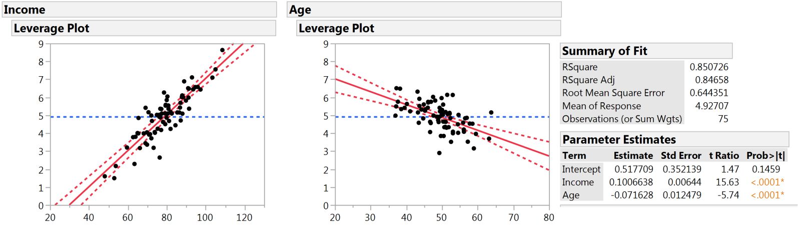
The smartphone data set

The simple regression of `Rating` against age with a *positive* marginal association.



Using a leverage plot

For the smartphone data set where we regressed `Rating` against both `income` and age, we have two leverage plots:



Summary of Fit

RSquare	0.850726
RSquare Adj	0.84658
Root Mean Square Error	0.644351
Mean of Response	4.92707
Observations (or Sum Wgts)	75

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.517709	0.352139	1.47	0.1459
Income	0.1006638	0.00644	15.63	<.0001*
Age	-0.071628	0.012479	-5.74	<.0001*

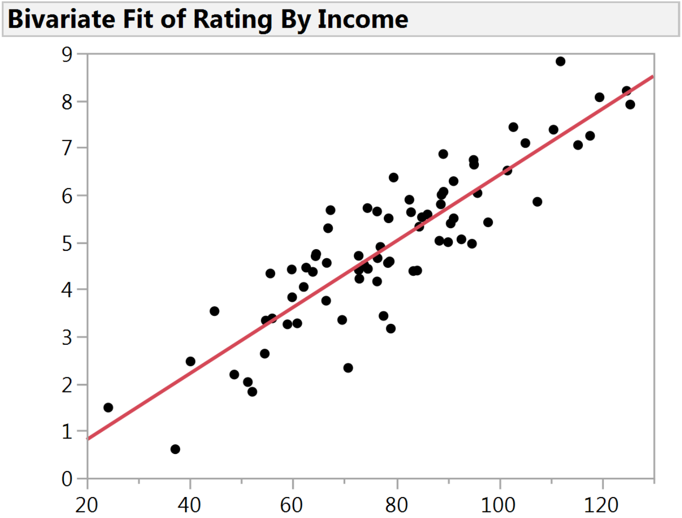
Note that the age leverage plot reveals a **negative** partial association. What do we hope to see in a leverage plot? Approximate linearity, just like in a simple regression.

Constructing a leverage plot

- Leverage plots are constructed to display partial association.
- It is not obvious how they achieve this goal, given that they look like simple regressions.
- Understanding their construction gives insight into what exactly it means to *control* for another explanatory variables, or the expression *accounting* for the other variables.
- The leverage plot has a three stage construction (we will demonstrate for x_1 in the presence of x_2):
  1. Regress x_2 out of y and save the residuals. Call these residuals e_1.
  2. Regress x_2 out of x_1 and save the residuals. Call these residuals e_2.
  3. Perform the simple regression of e_1 against e_2. The slope of this simple regression is exactly the partial regression slope in the multiple regression of y against x_1 and x_2.
- JMP adds on $\bar{y}$ to e_1 and $\bar{x_1}$ to e_2 so that the plot has the same location as the unadjusted data.

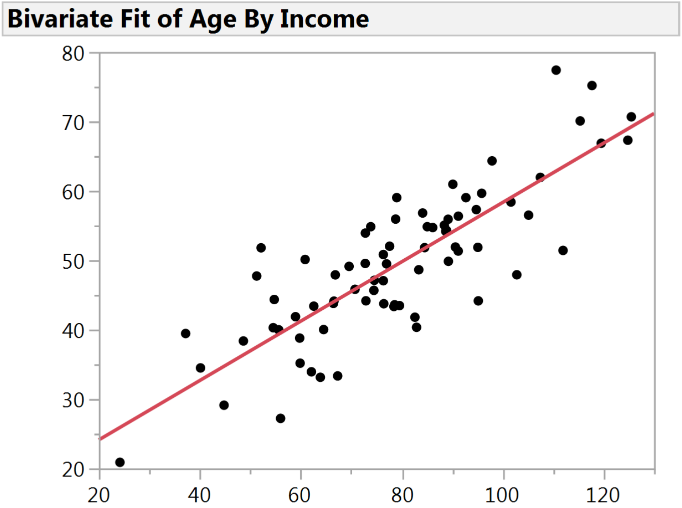
Step 1

Regress x_2 out of y and save the residuals. Call these residuals e_1:



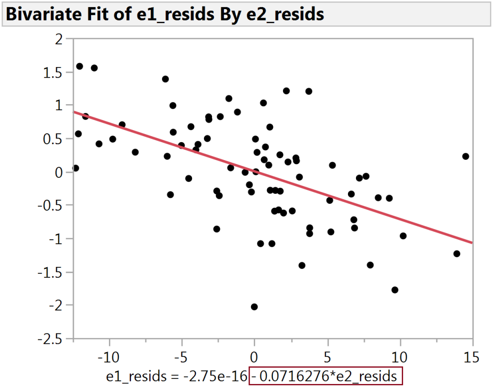
Step 2

Regress x_2 out of x_1 and save the residuals. Call these residuals e_2:



Step 3

Plot e_1 against e_2 and run a simple regression:



Bivariate Fit of e1_resids By e2_resids

e1_resids = -2.75e-16 - 0.0716276*e2_resids

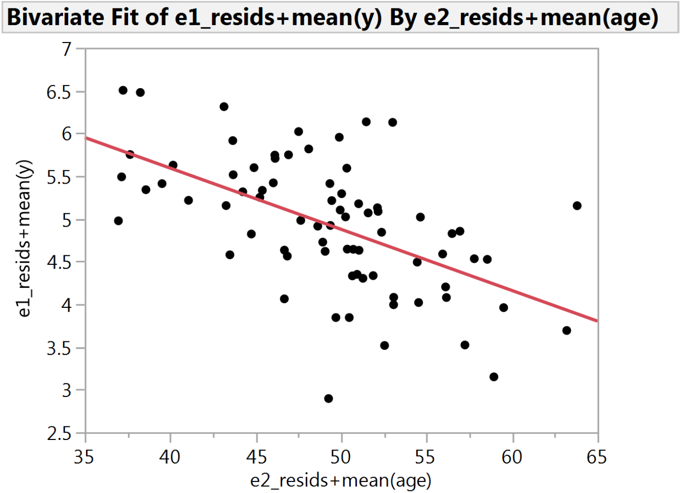
Summary of Fit

RSquare	0.31393
RSquare Adj	0.304532
Root Mean Square Error	0.639922
Mean of Response	-8.9e-17

- Note that the slope in this special simple regression is identical to the partial slope in the multiple regression. Note that the RMSE is essentially the same too. The slight difference is due to the division by (n -2) here, when it should really be divided by (n - 3).

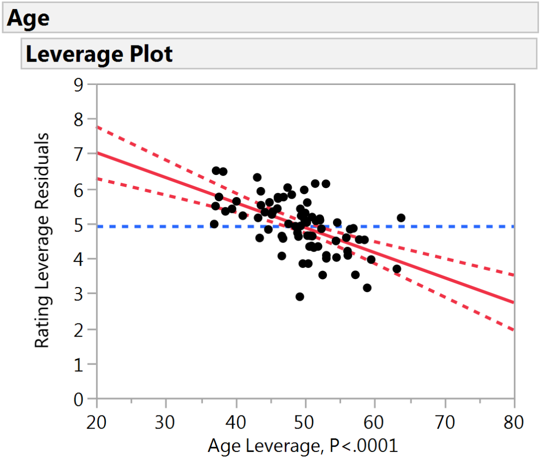
JMPs exact leverage plot

JMP adds on $\bar{y}$ to e_1 and $\bar{x_1}$ to e_2 so that the plot has the same location as the unadjusted data:



Additional information on the leverage plot

- The red dotted lines: 95% confidence bands for the partial slope
- The flat blue line: a visual null hypothesis of the partial slope equal to zero



Introduction to categorical predictor variables

- To date in the regression part of the course, all of the variables we have dealt with have been continuous.
- But there will be many instances where the are categorical variables we want to include too.
- Examples: Male/Female. East/West. India/China. Growth/Value
- We will start with two-level categoricals and then expand to multi-level categoricals.

Formulation

Three variables; Y, X_1 and X_2.
Y is the response variable, say Blood Pressure. X_1 is a two-level categorical, "Location" with levels Rural/Urban. X_2 is a continuous variable, say "age".
Recode X_1 where $X_1 = +1$ if Rural, and $X_1 = 0$ if Urban.
Write down the usual multiple regression equation:

$$Av(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Now condition on $X_1 = \mathit{Rural}$ to get (substitute in $X_1 = +1$),

$$Av(Y|X_1 = \mathit{Rural}, X_2) = \beta_0 + \beta_1 \times 1 + \beta_2 X_2.$$

Likewise for Urban:

$$Av(Y|X_1 = \mathit{Urban}, X_2) = \beta_0 + \beta_1 \times 0 + \beta_2 X_2.$$

Simplify to get

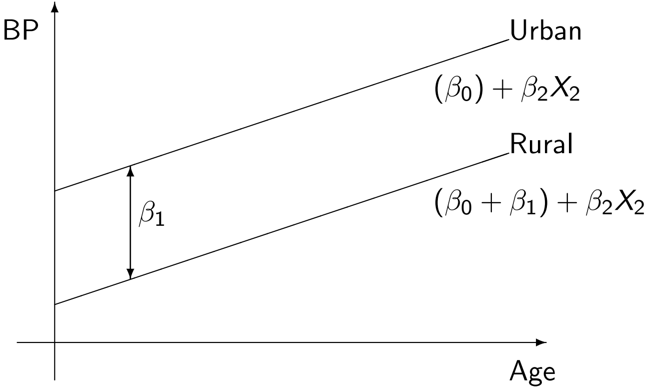
$$Av(Y|X_1 = Rural, X_2) = \overbrace{\beta_0 + \beta_1}^{Intercept} + \overbrace{\beta_2}^{Slope} X_2.$$

$$Av(Y|X_1 = Urban, X_2) = \overbrace{\beta_0}^{Intercept} + \overbrace{\beta_2}^{Slope} X_2.$$

Notice that these are the equations of parallel lines, with difference in height equal to:

$$Av(Y|X_1 = Rural, X_2) - Av(Y|X_1 = Urban, X_2) = \beta_1.$$

Graphically


BP
Urban
$(\beta_0) + \beta_2 X_2$
Rural
$(\beta_0 + \beta_1) + \beta_2 X_2$
β_1
Age

Parallel lines regression

Allows for different intercepts for the two groups.

- To implement in JMP:
- Make sure the variable is identified as **NOMINAL**.
- Add it just like any other X-variable.
  - Including the categorical variable allows you to fit a separate line to each group so that you can compare them.
  - Recognize that the comparison is between one group and the baseline (last in alphabet) group.
  - Recognize that the lines are forced to be **parallel**.

Parallel lines regression

- The regression estimate on the categorical variable is the difference between one group and the baseline group for the estimated Y-value.
- The height difference between the parallel lines is given by the estimated regression coefficient for the categorical variable.

Non-parallel lines formulation

Three variables; Y, X_1 and X_2.
Recode X_1 where $X_1 = +1$ if Rural, and $X_1 = 0$ if Urban.
Write down the usual multiple regression equation, but now with a **cross-product** term:

$$Av(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \, X_2.$$

Now condition on $X_1 = \mathit{Rural}$ to get (substitute in $X_1 = +1$),

$$Av(Y|X_1 = \mathit{Rural}, X_2) = \beta_0 + \beta_1 \times 1 + \beta_2 X_2 + \beta_3 \times +1 \times X_2.$$

Likewise for Urban:

$$Av(Y|X_1 = \mathit{Urban}, X_2) = \beta_0 + \beta_1 \times 0 + \beta_2 X_2 + \beta_3 \times 0 \times X_2.$$

Simplify to get

$$Av(Y|X_1 = Rural, X_2) = \overbrace{\beta_0 + \beta_1}^{Intercept} + \overbrace{(\beta_2 + \beta_3)}^{Slope} X_2.$$

$$Av(Y|X_1 = Urban, X_2) = \overbrace{\beta_0}^{Intercept} + \overbrace{\beta_2}^{Slope} X_2.$$

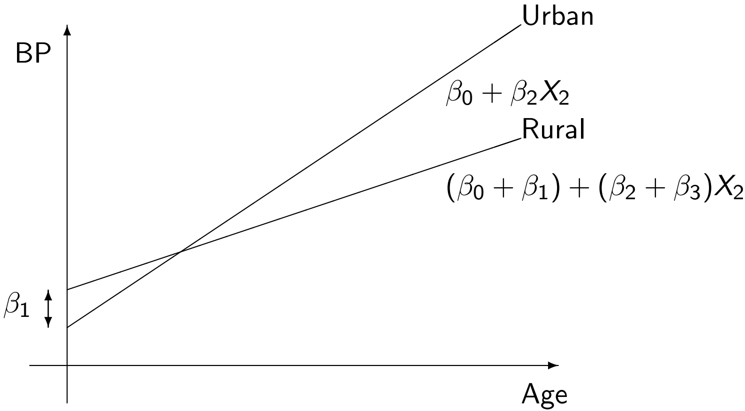
Notice that these are the equations of non-parallel lines, with difference in **intercepts** equal to:

$$Av(Y|X_1 = Rural, X_2 = 0) - Av(Y|X_1 = Urban, X_2 = 0) = \beta_1.$$

and difference in **slopes** equal to

$$(\beta_2 + \beta_3) - (\beta_2) = \beta_3.$$

Graphically



Non-parallel lines regression

Allows for different intercepts **and** different slopes for each group.

- Declare the categorical variable as **NOMINAL**.
- Add it just like any other X-variable but **also** add the cross-product term. Cross-product terms are sometimes known as **interaction** terms.
- Remember to uncheck "center polynomials" in the model dialog under the red triangle.
- The regression coefficient on the categorical variable tells you the difference between intercepts, comparing one group to the baseline group.
- The regression coefficient on the cross-product term tells you the difference between the slopes for the two groups, comparing the slope of one group to the baseline group.

Interaction

- Interaction. A three variable concept (Y, X_1, X_2). Generic description: the impact of X_1 on Y depends on the value of X_2.
- It allows the slope of one variable to depend on the **value** of another variable.
- Understand it as **synergy** between variables.
- A model with different intercepts and same slopes is OK. A model with different intercepts and different slopes is OK. A model with same intercepts but different slopes is **not** desirable.
- Good modeling practice: if you have an interaction term in the model (i.e. different slopes) then make sure you have the variables that make up the interaction in the model as well (even if they are not significant).

Summary

- Leverage plots
- Introduction to categorical predictor variables

Next time

- More on categorical variables
- More on *interaction*