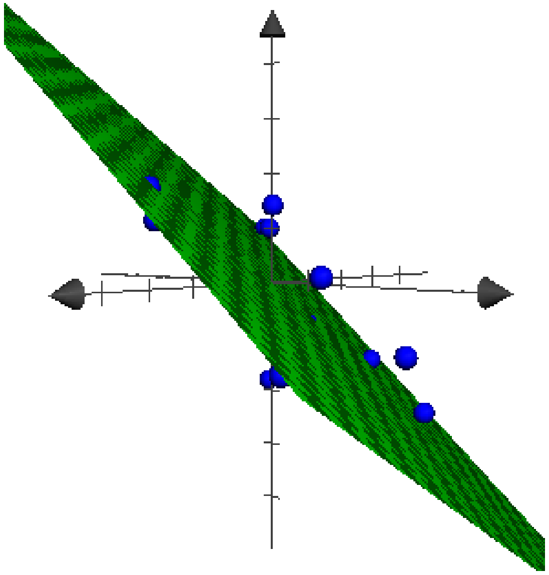
Stat 608 Chapter 5

Multiple Linear Regression

- Chapter 5:
  - Multiple predictor variables
  - ANCOVA
  - Polynomial Regression
  - Assumption that model is valid
- Chapter 6:
  - Leverage points
  - Transformations
  - Relationships between explanatory variables:
    - Multicollinearity
    - Interactions
- Chapter 7:
  - Variable Selection

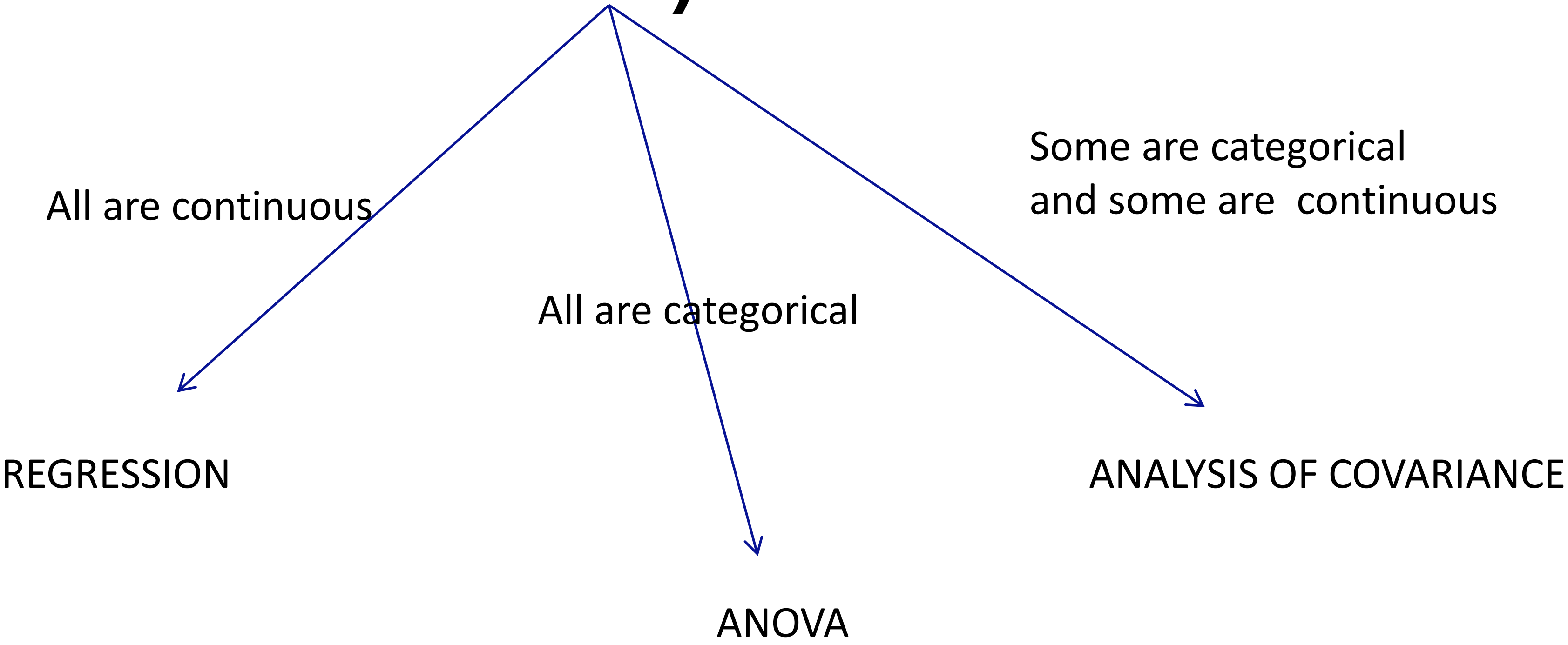
Types of Multiple Linear Regression

- Polynomial Regression (curves)
- Quantitative and Categorical explanatory variables: ANCOVA (separate lines)
- Many explanatory variables (multiple dimensions)



ANOVA & REGRESSION & ANALYSIS OF COVARIANCE

$$Y = X\beta + e$$



All are continuous

All are categorical

Some are categorical and some are continuous

REGRESSION

ANOVA

ANALYSIS OF COVARIANCE

ANOVA (ANalysis Of VAriance)

- One-way ANOVA, without an intercept:

$$y_i = \alpha_i + e_i, \, i = 1, 2, 3$$

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + e_i$$

Design Matrix:

Pro: (X’X) is diagonal: easy to invert!

Con: When we have two-way ANOVA, we have to cut out the last indicator variable…

ANOVA

- One-way ANOVA, with an intercept:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i$$

Design matrix:

Two-Way ANOVA

- Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

Design matrix:

Analysis of Covariance

- Suppose we have three groups, and want to compare the three means, holding the value of a quantitative variable x constant.

- It’s possible to create three separate regression lines, as shown below. We might also create three lines with separate intercepts, but the same slope, or three lines with separate slopes, but the same intercept.

Group 1: $y_{i1} = \beta_{01} + \beta_{11} x_i + e_i$

Group 2: $y_{i2} = \beta_{02} + \beta_{12} x_i + e_i$

Group 3: $y_{i3} = \beta_{03} + \beta_{13} x_i + e_i$

Interactions

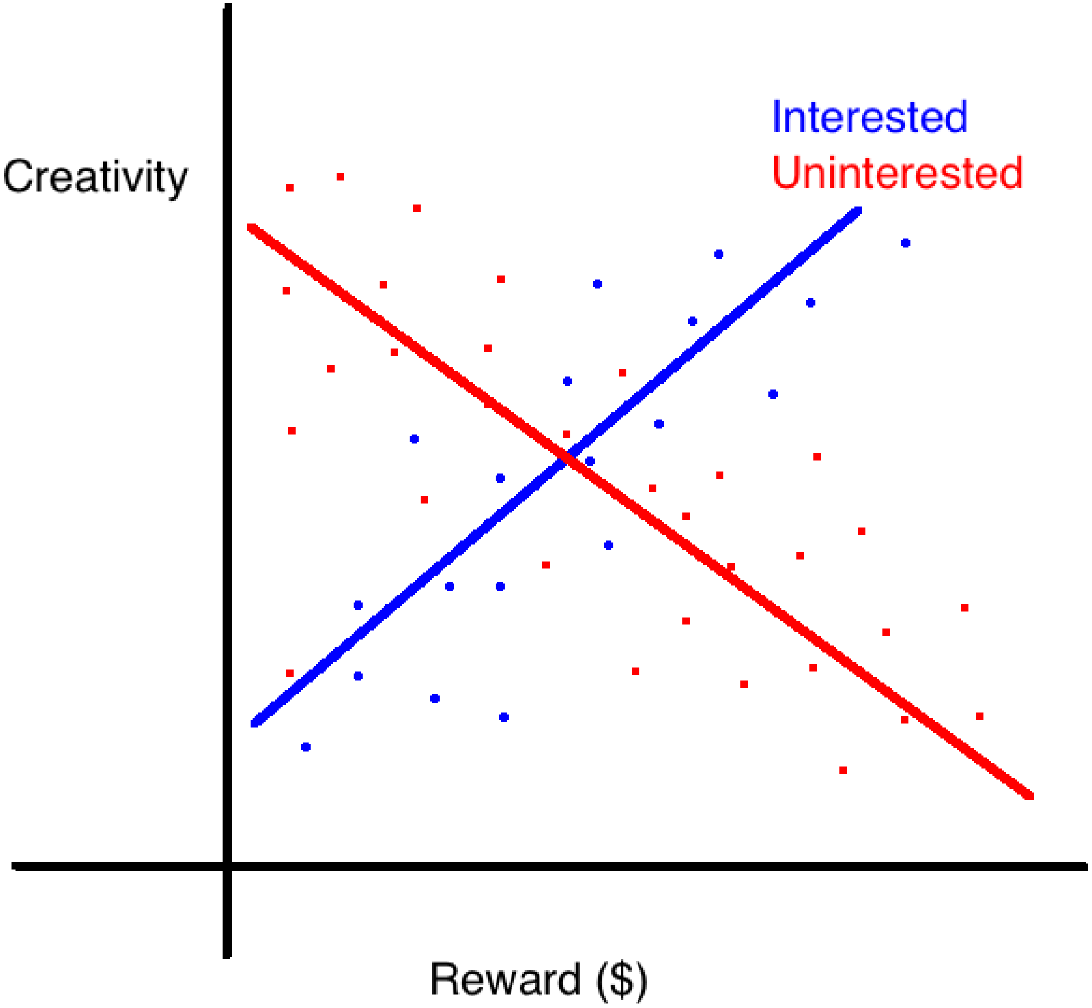
- An **interaction** between two input variables exists when the effect of one input (X_1) on the target variable (Y) is different for different values of the other input (X_2).

- Ex: Reward (X_1) has a positive effect on creativity (Y) for strongly interested (X_2) people but a detrimental effect for uninterested people.

- Ex: The amount of iron in food (Y) is higher when cooking in a cast iron pot (X_1). While tomatoes have a tiny amount of iron in them, the acidity in tomatoes means their presence in food (X_2) has a multiplicative effect on cast iron.

…

+ Interactions: ANCOVA



…

ANCOVA: Interactions

What would the model and design matrix look like in the case of the reward example? For the iron pot example?

…

ANCOVA Model

- Where x1 is a quantitative variable and x2 is an indicator (dummy) variable, write down the model for separate slopes, separate intercepts:

- Write down the model with separate slopes, but the same intercept:

- Write down the model with separate intercepts, but the same slope:

…

+

ANCOVA Example

Rats are randomly assigned to be fed 0, 2, 4, and 6 mg of one of two cancerous substances. The response variable y is the number of tumors recorded. What should the model look like? What should the design matrix look like?

…

Multiple Regression: Interactions

- If an interaction exists, there are many possible things that could be true:

- The relationship could change direction in the presence of the third variable: the relationship between X_1 and Y is positive before taking into account X_2, and negative afterward.

- The relationship might not change direction in the presence of a third variable, but merely have a dramatic multiplicative effect: Fast driving (X_1) is much more dangerous (Y) when drunk(X_2).

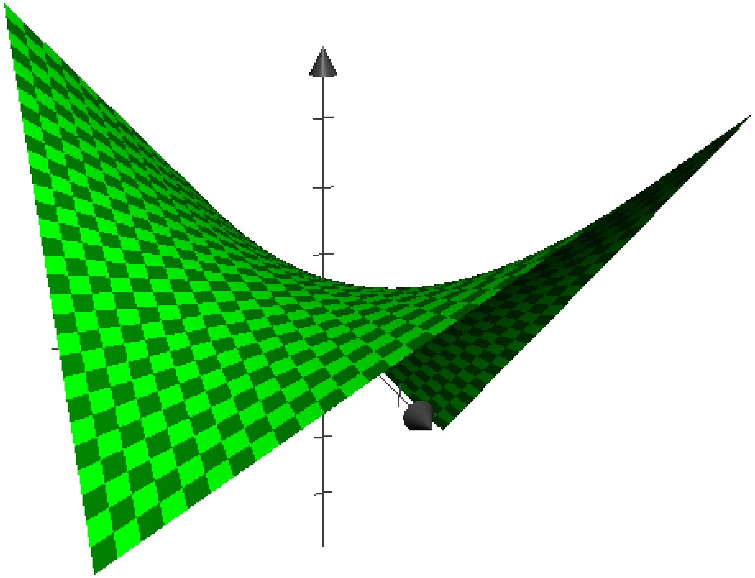
- Main effects might not be significant, while the interaction is significant.

- Cause and effect could run in many possible directions, but we can only scientifically establish cause and effect through direct experimentation.

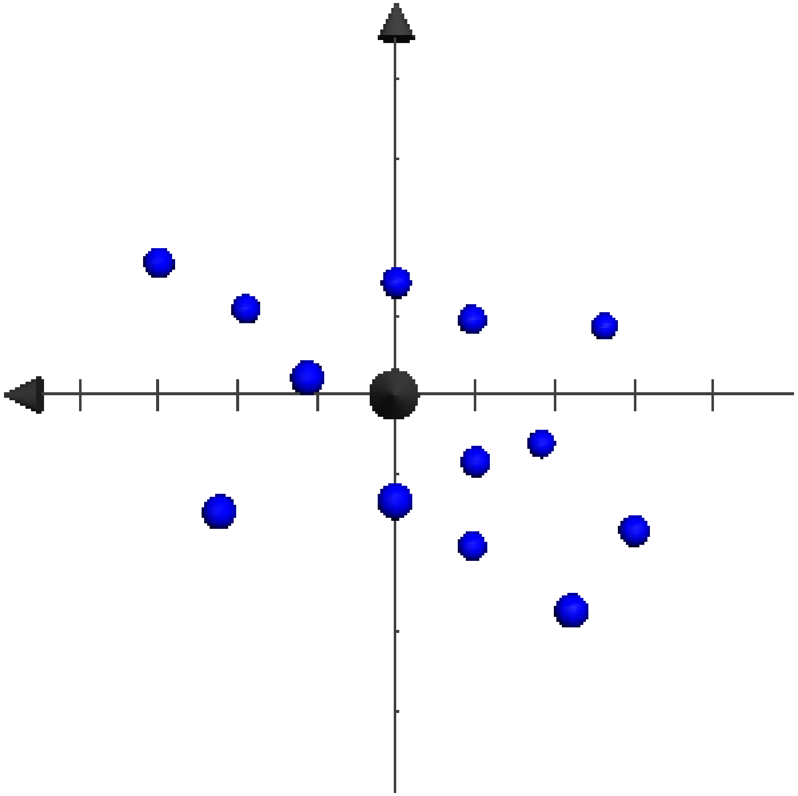
…

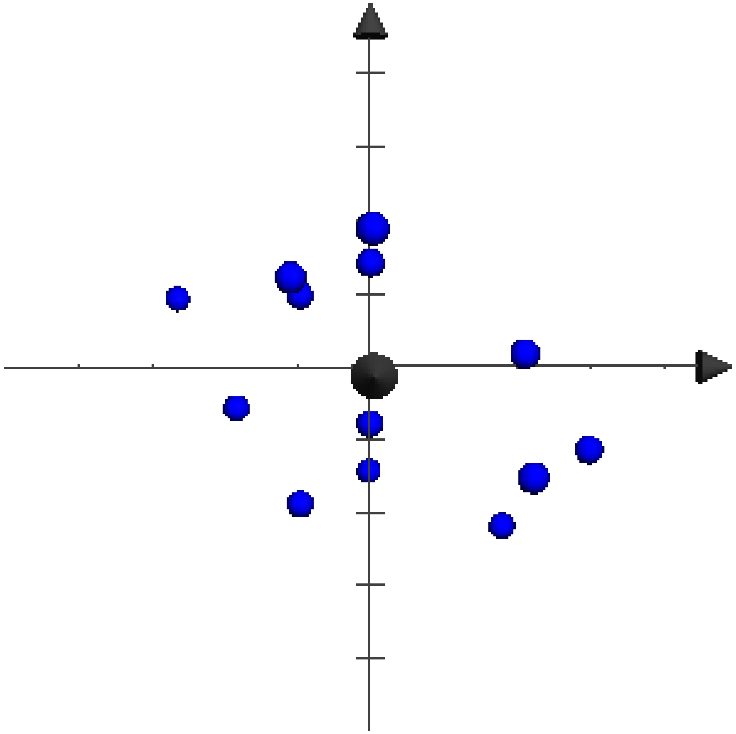
Multiple Regression: Interactions

- Interactions between quantitative variables: fit different kinds of surfaces.



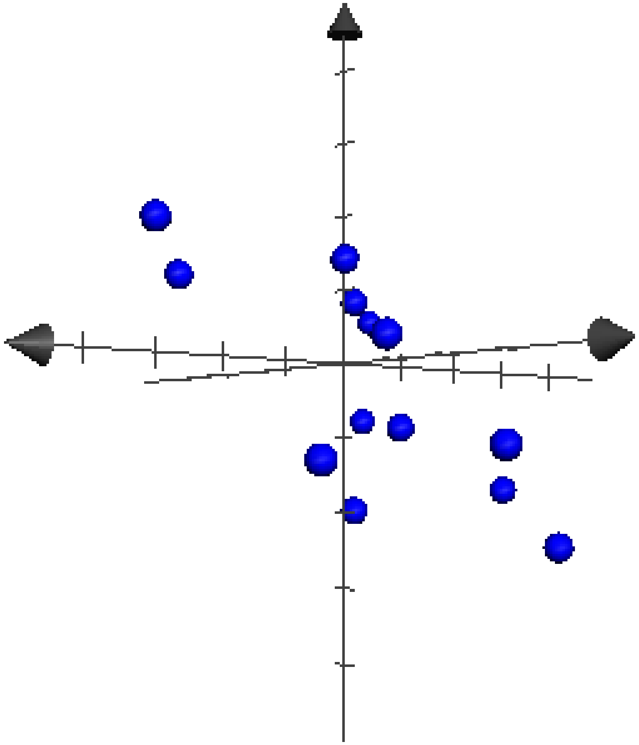
The Fallacy of Bivariate Thinking





The Fallacy of Bivariate Thinking

- Plots of x1 or x2 vs. y may seem to have no relationship with y.
- But the two variables working together may explain much more of the variation in y.
- Ex: Weight, Height, and Body fat percentage.



Rank of X

The rank of our design matrix X should be the number of columns of X. We say our design matrix is not full rank if it isn’t.

- If rank (X) < # columns = p + 1, that means there exist a linear combination of the other variables that adds up to one of the variables. Why do we need that extra variable??

- If rank(X) < p + 1, that means rank (X’ X) < p+1, so X’X is not invertible.

- If n < p + 1, rank(X) < p + 1 because the rank of X has to be less than or equal to both the number of columns and the number of rows of X. Get a bigger sample size or get rid of some variables.

Rank of X

- R error message:

```
Coefficients: (1 not defined because of
singularities)
```

- SAS error message:

```
Note:     Model is not full rank. Least-squares solutions for the
          parameters are not unique. Some statistics will be misleading.
          A reported DF of 0 or B means that the estimate is biased.
```

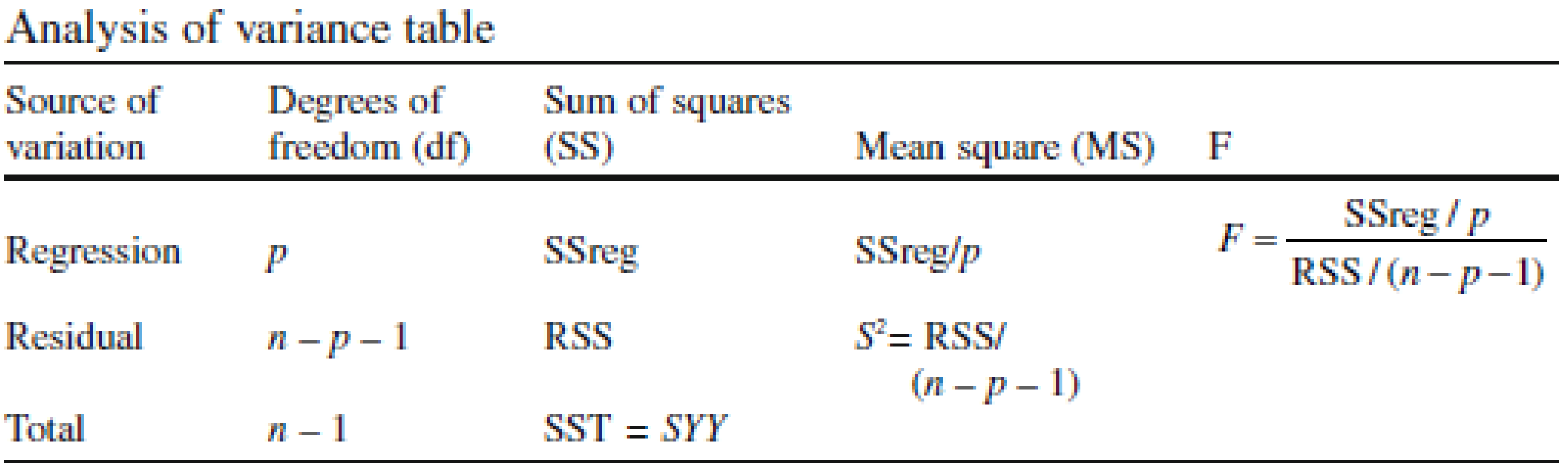
```
Note:      The following parameters have been set to 0, since the
           variables are a linear combination of other variables as
           shown.
```

ANOVA Table for Multiple Regression

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \text{ vs.}$$

$$H_a : \text{At least one } \beta_i \neq 0$$

$$F = \frac{SSReg/p}{RSS/(n-p-1)}$$



Analysis of variance table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	p	SSreg	SSreg/p	$F = \dfrac{\text{SSreg}/p}{\text{RSS}/(n-p-1)}$
Residual	$n-p-1$	RSS	$S^2 = \text{RSS}/(n-p-1)$	
Total	$n-1$	SST = SYY		

+

Tests & Confidence Intervals for Multiple Regression

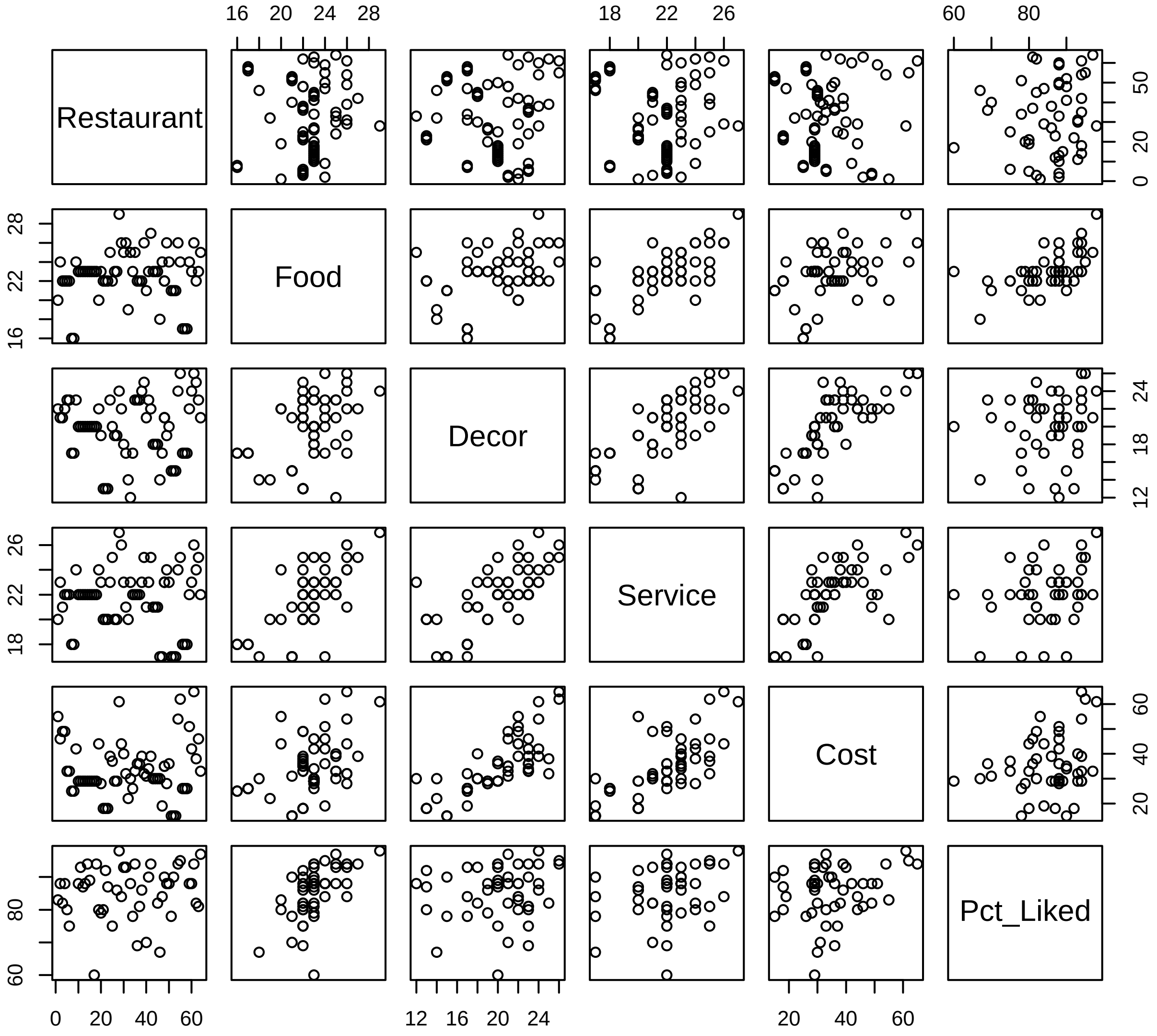
$$t_{n-p-1} = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

$$\hat{\beta}_i \pm t^*_{n-p-1}\, se(\hat{\beta}_i)$$

- Note: If we start conducting these tests for many of the variables, performing p separate t-tests, our overall Type I error increases.
- We also run into problems when the predictor variables are highly correlated with each other.

Italian Restaurants: Houston

plot(Italian)



Italian Restaurants: Houston

```
my.lm<- lm(Italian$food ~Italian$Service +
Italian$Pct_Liked + Italian$Cost)

anova(my.lm)
```

```
Response: Italian$Food
                  Df Sum Sq Mean Sq F value    Pr(>F)
Italian$Service    1 64.619  64.619  36.369 2.605e-07
Italian$Pct_Liked  1 36.510  36.510  20.548 4.133e-05
Italian$Cost       1  1.860   1.860   1.047    0.3116
Residuals         46 81.731   1.777

Residual standard error: 1.333 on 46 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared: 0.5575, Adjusted R-squared: 0.5287
F-statistic: 19.32 on 3 and 46 DF,  p-value: 2.989e-08
```

Italian Restaurants: Houston

```
my.lm<- lm(Italian$food ~Italian$Service +
Italian$Pct_Liked + Italian$Cost)

summary(my.lm)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.68923    2.66400   1.385  0.17278
Italian$Service    0.47849    0.11744   4.074  0.00018
Italian$Pct_Liked  0.11304    0.02461   4.594 3.38e-05
Italian$Cost      -0.02236    0.02185  -1.023  0.31156
```

Italian Restaurants: Houston

- The coefficient for cost is negative; does that make sense? Interpret the slope for Cost in context.

- The p-value for cost is large; does that make sense?

Italian Restaurants: Houston

- New model:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.77121    2.32996   4.623 2.97e-05
Italian$Pct_Liked  0.13047    0.02797   4.665 2.58e-05
Italian$Cost      0.03510    0.01927   1.821   0.0749
```

- The coefficient for %Liked is smaller than the coefficient for Service; does that mean Service is more important when it comes to predicting Food rating?

- The p-value for %Liked is smaller than the p-value for Cost; does that mean the association between %Liked and Food rating is stronger than the association between Cost and Food rating?

Italian Restaurants: Houston

- Calculate an approximate 95% confidence interval for the slope for Service.

Polynomial Regression

- Is the following model linear in the parameters?

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

- If a model is linear in the parameters, it means we can write it as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Polynomial Regression

- What does the design matrix look like for the following model?

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

+

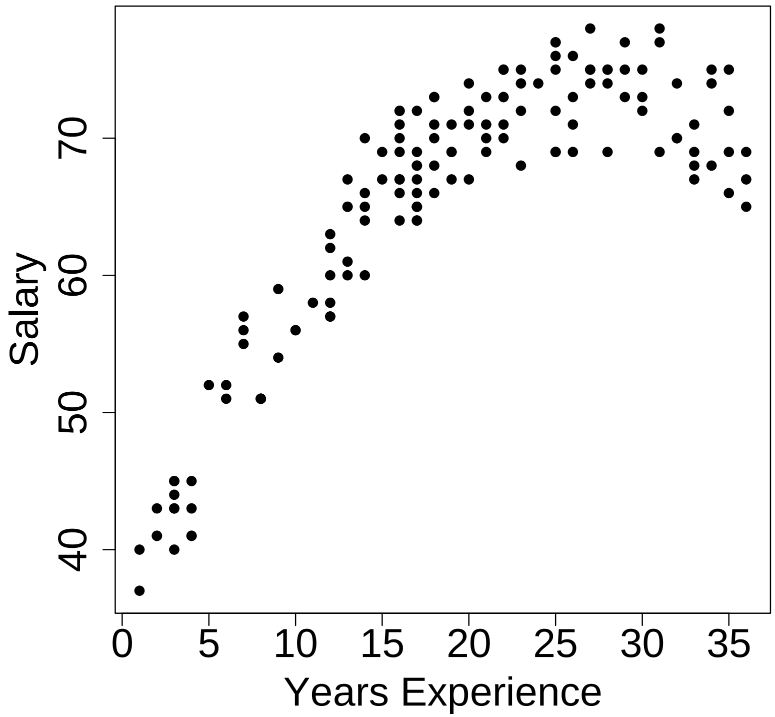
Salary Example

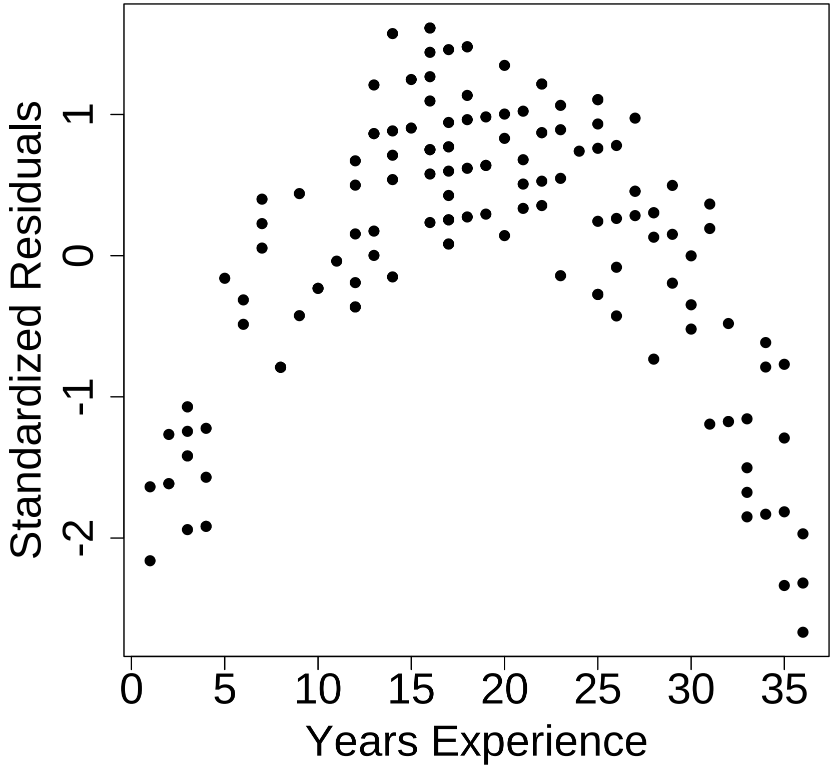
A salary curve relates salary to years of experience. Employees might use it find out where they stand among their peers. Personnel might use it to consider salary adjustments when hiring new professionals.

When we fit the simple linear regression model below, we get the residual plot on the next slide.

This image cannot currently be displayed.

Salary Example: SLR





Salary Example: Parabola

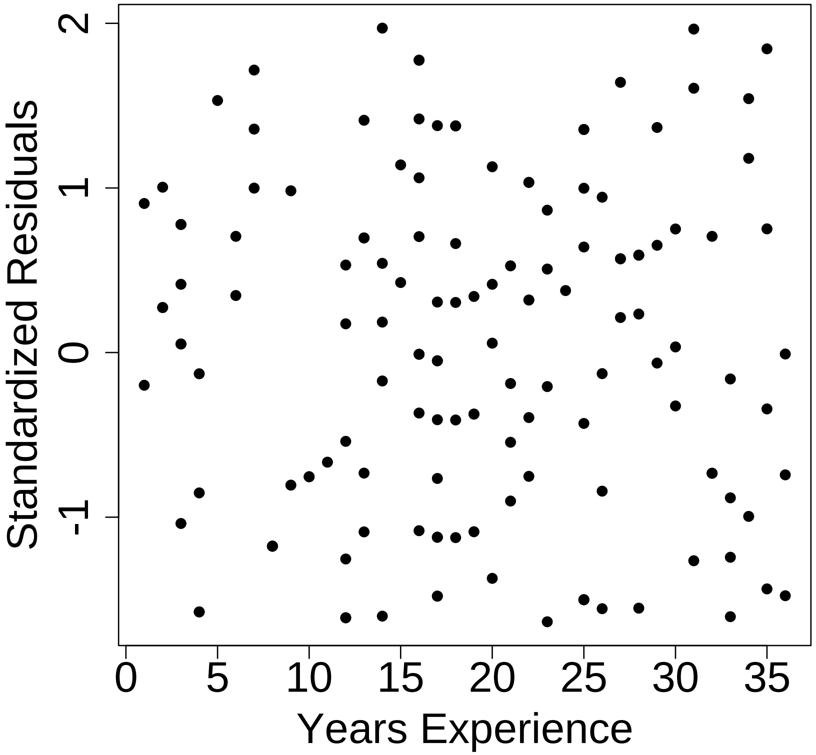
- Clearly there is a non-linear relationship between salary and years of experience.

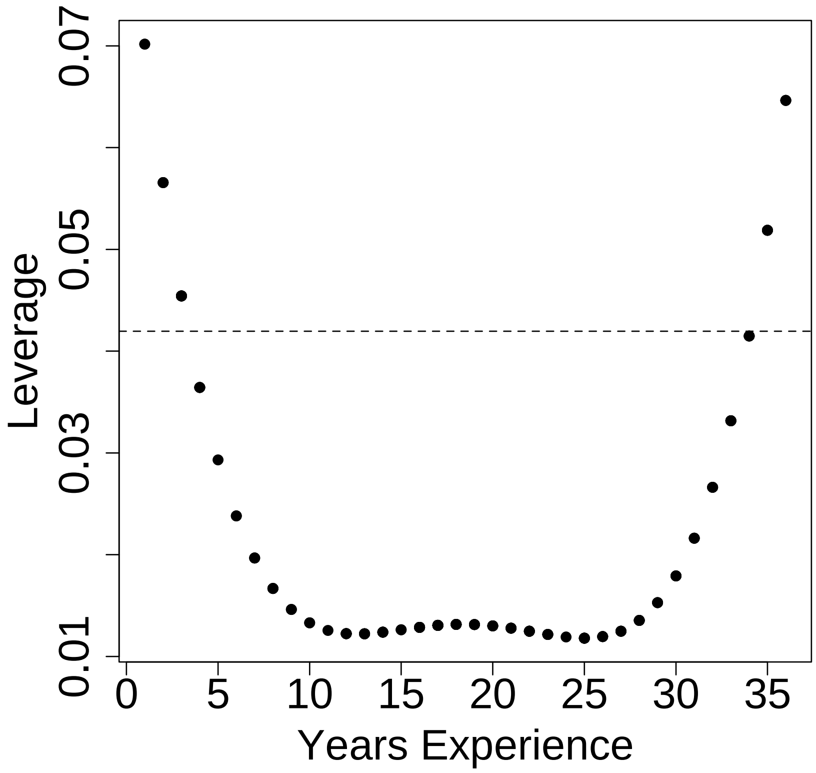
- Next we fit the model

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

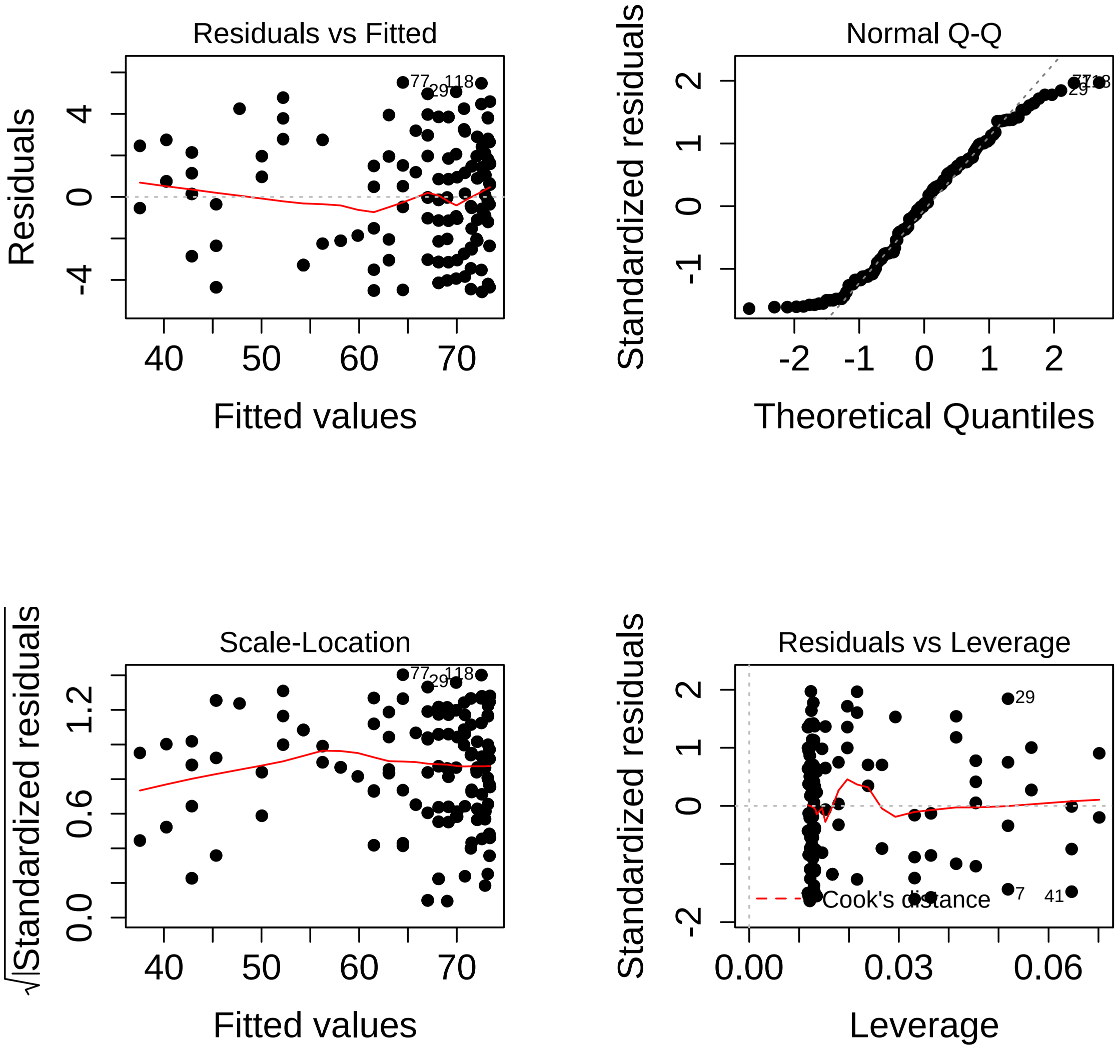
- The cutoff for high leverage is $h_{ii} > 2(p+1) / n$ when there are p predictors plus 1 intercept.

Salary Example: Parabola





Salary Example: Parabola



+ Model Reduction **Method 1 – Partial F-Test**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

Suppose we have the model above, and are Interested in testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0, k < p$$

against the alternative hypothesis

Ha: At least one of the parameters is not 0.

That is, the question is, “Can we drop these k variables from our model?” This can be tested using an F-test.

Model Reduction **Method 1 – Partial F-Test**

Let RSS(Full) be the residual sum of squares from the model with all the predictors 1, …, p. Let RSS(Reduced) be the residual sum of squares from the reduced model (with only the remaining predictors that we don't think are 0). Then the F-statistic for testing the above hypotheses is given by:

$$\begin{aligned} F &= \frac{(RSS(reduced) - RSS(full))/(df_{reduced} - df_{full})}{RSS(full)/df_{full}} \\ &= \frac{(RSS(reduced) - RSS(full))/k}{RSS(full)/(n-p-1)} \end{aligned}$$

Model Reduction **Method 2**

$$H_0 : \mathbf{A}_{r\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} = h_{r\times 1}$$

$$H_a : \text{At least one equality} \neq h_{r\times 1}$$

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - h)'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - h)/r}{\hat{\mathbf{e}}'\hat{\mathbf{e}}/(n-p-1)}$$

rank(A) = r

+ R^2

- R^2 is often defined as the proportion of the variability in the random variable Y explained by the regression model.

$$SSreg + RSS = SST$$

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$$

R^2: Adding Variables

- Adding irrelevant predictor variables to the regression equation often increases R^2.

- **Solution**: R^2 adjusted

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- The denominator is an unbiased estimate of the variance of Y with all slopes = 0, while the numerator is an unbiased estimate of the variance of the residuals.