Announcements

- Data analysis 8 due tonight at 11:59 pm

- Office hours today with me at 10 am (Weniger 273)

- Finals Week Review Sessions:
    - Monday, June 6 1-2:30 pm Weniger 275
    - Tuesday, June 7 2-3:30 pm Weniger 275
    - Come with prepared questions or just come study and ask questions as they come up

- If you have not already done so, please read through the Final Exam Info and FAQ page in the Final Exam module on Canvas

Week 10

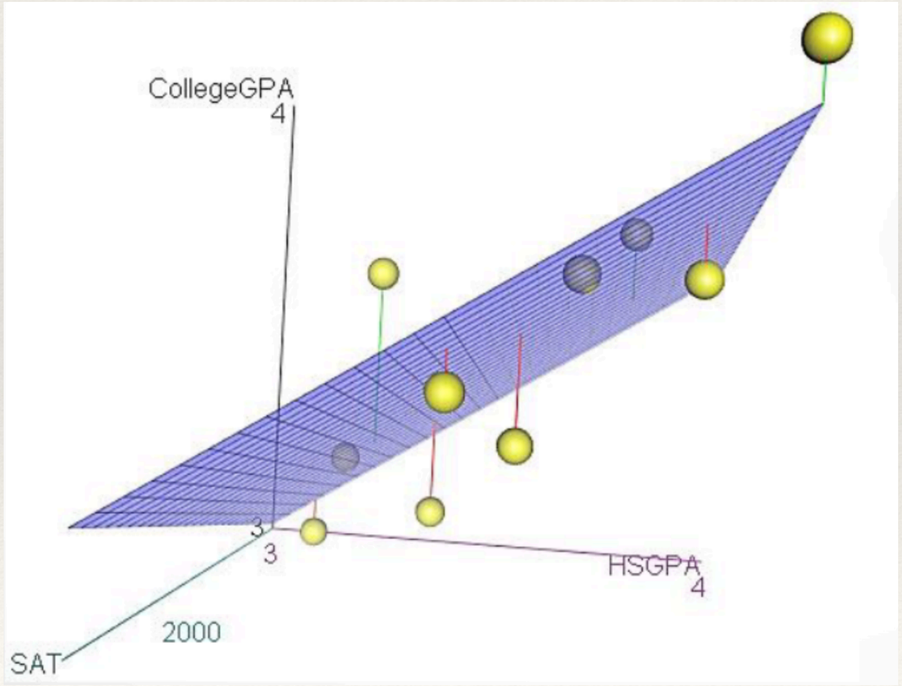
Introduction to Multiple Regression

ST 314
Introduction to Statistics for Engineers

Oregon State University

Multivariate Data

Along with HS GPA, suppose SAT scores may also be a good predictor of freshman year college GPA. Can we have more than one explanatory variable?



Yes!

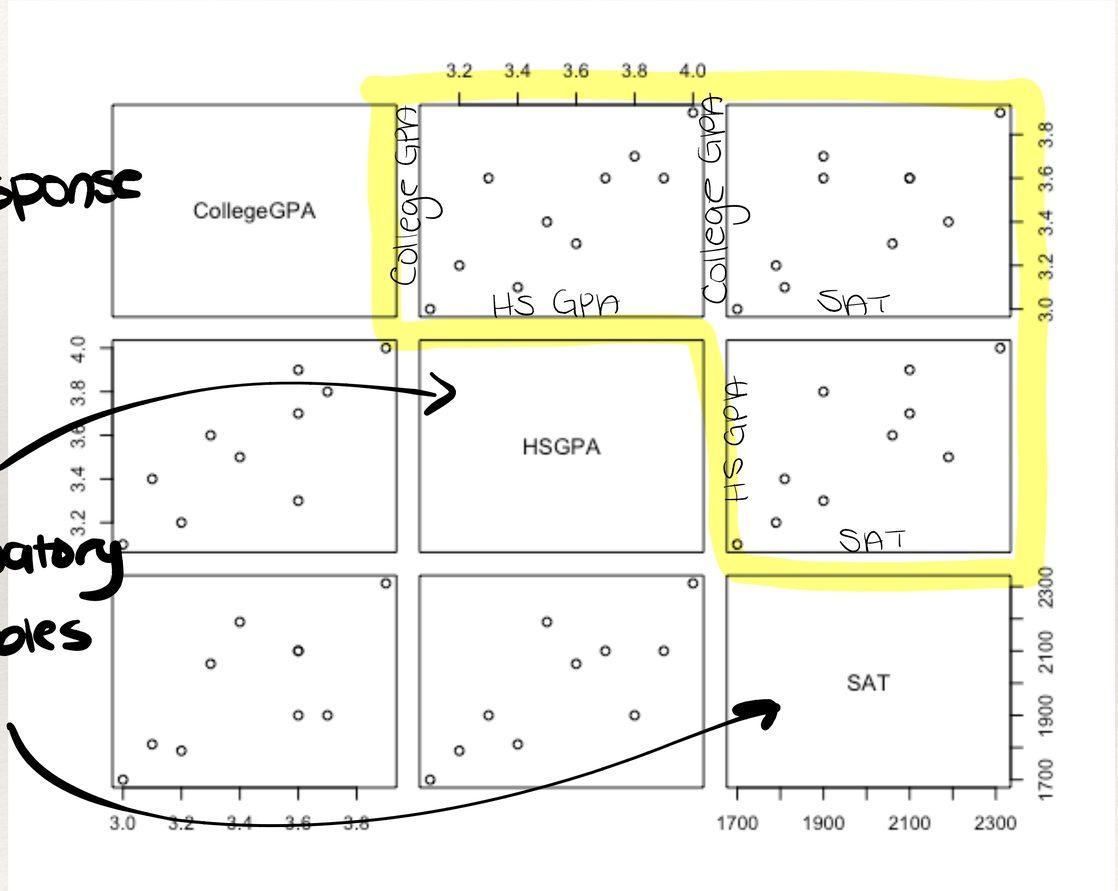
It's possible that modeling the response variable with multiple explanatory variables does a better job at explaining the variation in the response.

Explanatory and Response Variables

Explanatory Variable(s)	Response Variable
• may help to explain the response • can be quantitative or categorical • $x_1, x_2, \cdots, x_k$	• variable to be estimated or predicted • quantitative • y

Visualizing Multivariate Data

- As the number of explanatory variables increase, visualization can become tricky.
- We can visualization bivariate relationships between all involved variables using a scatter plot matrix.



Multiple Linear Regression Equation

- A multiple linear regression (MLR) models the relationship between one response variable and multiple ($k > 1$) explanatory variables
  ↑ # of explanatory variables
- General form of the estimated multiple linear regression model:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

$b_0, b_1, \ldots, b_k$ are estimates for the coefficients from the population regression equation $\beta_0, \beta_1, \ldots, \beta_k$

- The coefficient estimates are determined by finding the model that yields the smallest sum of squared error (just like in simple linear regression).

MLR Example

Consider a random sample of 141 eBay auction sales for the Nintendo Wii game Mario Kart. We are interested in modeling total sale price from starting price of the bid, number of wheels, and condition.

total.price	start.price	wheels	used
51.55	0.99	1	0
37.04	0.99	1	1
45.5	0.99	1	0
...			
54.51	1	2	0

$x_1 =$ starting price

$x_2 =$ number of wheels

$x_3 =$ 1 if used, 0 otherwise

$\hat{y} =$ predicted total price

Model output:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	41.79739	0.99545	41.988	< 2e-16	***
start.price	0.12376	0.02566	4.824	3.70e-06	***
wheels	6.83250	0.50964	13.406	< 2e-16	***
used	-5.69207	0.85817	-6.633	7.02e-10	***

Write the estimated MLR equation using the model output.

$\hat{y} = 41.797 + 0.124x_1 + 6.833x_2 - 5.692x_3$

used: $\hat{y} = 36.105 + 0.124x_1 + 6.833x_2$

new: $\hat{y} = 41.797 + 0.124x_1 + 6.833x_2$

Broken Used New

Used and New

$$\hat{y} = 41.797 + 0.124x_1 + 6.833x_2 - 5.692x_3$$

Broken, new, used

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3\,\text{Broken} + b_4\,\text{Used}$$

Broken: 1 if broken, 0 otherwise

Used: 1 if used, 0 otherwise

Coefficient of Determination, R^2

How do we know if the model is a good fit?

- The coefficient of Determination, R^2, defines the proportion of variation in the response variable that can be explained by the explanatory variables in the model.
- $R^2 = \frac{SSR}{SST}$ ← regression sum of squares
  ← total sum of squares

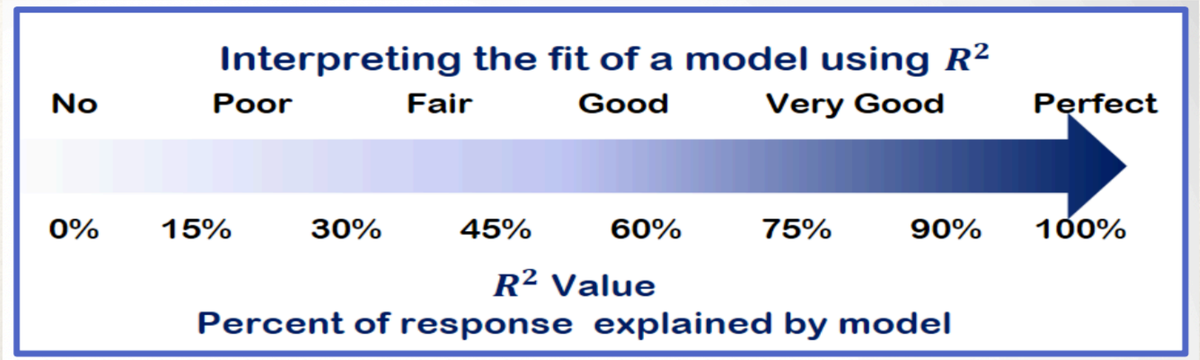
Coefficient of Determination, R^2

$$SST = SSR + SSE$$

Total Sum of Sqaures	Regression Sum of Sqaures	Error Sum of Squares
Total variability of the response variable $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ ↑ observed response	The explaincd variation in the response by the model $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ ↑ predicted response	Unexplained or "left over" variation → variation in the response that is not explained by model $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Interpreting R^2

The higher the R^2 value the better the explanatory variables are at explaining the variability in the response!



Model Output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.79739    0.99545  41.988  < 2e-16 ***
start.price  0.12376    0.02566   4.824 3.70e-06 ***
wheels       6.83250    0.50964  13.406  < 2e-16 ***
used        -5.69207    0.85817  -6.633 7.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.535 on 137 degrees of freedom
Multiple R-squared:  0.7577,     Adjusted R-squared:  0.7524
F-statistic: 142.8 on 3 and 137 DF,  p-value: < 2.2e-16
```

The explanatory variables in the model explain 75.77% of the variation in total selling price for Mario Kart in the eBay Auctions.

$R^2 =$ 0.7577

The model is a very good fit.

R^2 vs. Adjusted R^2

Cautions about R^2

- Different fields of study have different ideas of what R^2 values suggest a "good fit"
- Adding variables to the model will always increase R^2 even if they aren't helpful predictors of the response

Adjusted R^2 accounts for the number of variables and the sample size in the model.

Adjusted $R^2 = 1 - \frac{n-1}{n-(k+1)}\left(\frac{SSE}{SST}\right)$

R^2 vs. Adjusted R^2 Example

Compare the fits of the two models for predicting college GPA. What effect does adding SAT to the model have on the R^2 value? What about adjusted R^2? Which model is a "better" fit?

Model with HS GPA and SAT score as the explanatory variables:

```
Residual standard error: 0.1838 on 7 degrees of freedom
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.5912
F-statistic: 7.508 on 2 and 7 DF,  p-value: 0.01812
```

Model with only HS GPA as the explanatory variable:

```
Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

Adjusted R^2 Example

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.79739    0.99545  41.988  < 2e-16 ***
start.price  0.12376    0.02566   4.824 3.70e-06 ***
wheels       6.83250    0.50964  13.406  < 2e-16 ***
used        -5.69207    0.85817  -6.633 7.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.535 on 137 degrees of freedom
Multiple R-squared:  0.7577,    Adjusted R-squared:  0.7524
F-statistic: 142.8 on 3 and 137 DF,  p-value: < 2.2e-16
```

While accounting for sample size and the number of explanatory variables in the model, the explanatory variables in the model explain 75.24 % of the variation in total selling price for Mario Kart in the eBay Auctions.

The model is a very good fit.