# Multiple Regression 2

**Dr Tom Ilvento**
Department of Food and Resource Economics

## Overview

- We will continue our discussion of Multiple Regression with a special focus on looking at Residuals
- Residual Analysis looks at the error terms in the model
  - The error term, or residual, shows the part not explained by our model
  - Residuals can often point to problems in our model
- We will revisit dummy variables in the model when there are other independent variables in the model
- We will look at a short, but complete analysis to give you an idea of how multiple regression can improve our insights into models



## Residuals in Regression

- A Probabilistic Model has a deterministic component and a random error component, denoted as $e_i$ or $\varepsilon_i$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_{i1}$$

- The model is our "expectation" of the relationship. This is the deterministic part of the model

$$E(Y_i) = \beta_0 + \beta_1 X_{i1}$$

- The error component is very important
  - Observed in population/sample

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

  - Predicted from model

$$\hat{Y}_i = \beta_o + \beta_1 X_{i1}$$

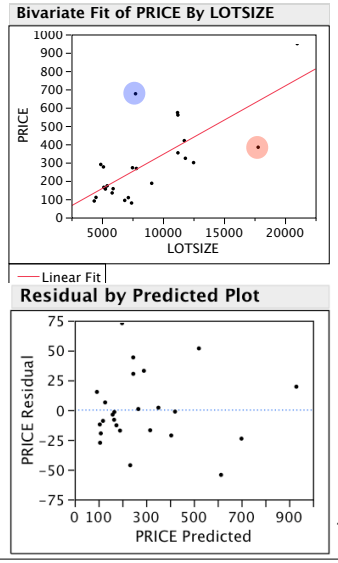
  - The difference between what we observe and what we predict

$$\varepsilon_i = Y - \hat{Y}$$



## Residuals or Error Terms

- The error terms will always have a mean of zero
  - Some will be **positive** and some **negative**
  - We want all the pattern in the data to be represented by the model
- Excel or other software will generate residuals and Predicted Y
  - We can plot residuals versus any X in the model
  - Or against **Predicted Y**
- It is usually best to work with **Standardized Residuals** (also called **Studentized**)
  - These are normed to have a mean of zero and standard deviation of 1
  - Standardized Residuals should be within ± 3 standard deviations



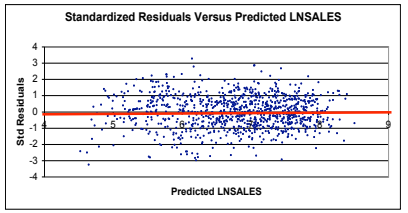
# Residual Analysis

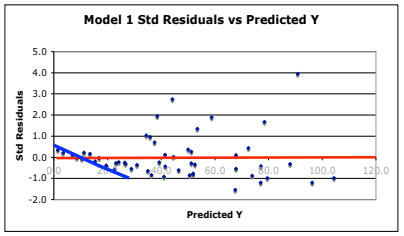
- The residuals are an important part of regression
  - Shows the fit of the model
  - Estimates the overall variance of the regression
  - Factors into the standard errors for coefficients
  - Residual Analysis shows the fit of the model and possible violations in assumptions
- We want a random pattern with standardized residuals that are no more than ±3 standard deviations
- Look for patterns in residuals to give clues
  - Bad fit
  - Outliers that don't fit the model well
  - Violation of assumptions
- **Residual Analysis can be thought of as a diagnostic tool to assess the aptness of your model**

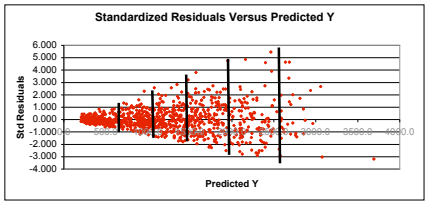


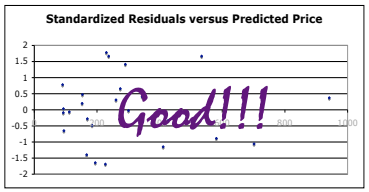
# Let's Look at Some Residuals from Models

- Remember, we want to see a random pattern



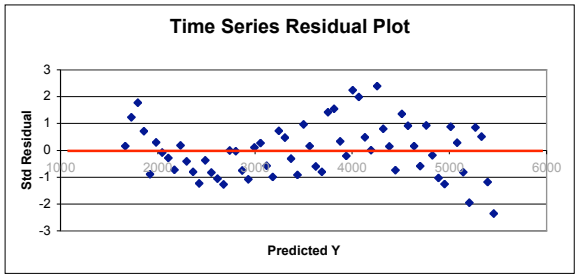






# Data with a time element often show a residual pattern

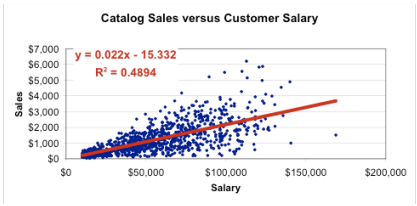
- Data with a time element - Time Series - often have error terms that are correlated with each other
- The residual in time t+1 is related to the error term in time t





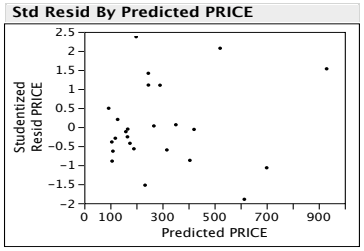

# Catalog Sales Data

- We want to look at the relationship between Sales and Salary
- We know there is a strong relationship, but the variance increases with increasing salary,
- We also know there are extreme values for salary and for sales
- As a result, residuals look bad and show heteroscadasticity
- Look what happens when we transform the data using the natural log of both variables
- Sometimes a transformation can solve the problem of a pattern in residuals



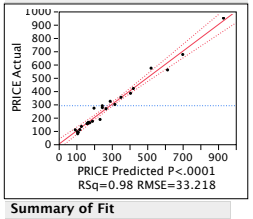
# Let's revisit the Apartment Data

- The model to the left shows the regression results that were discussed the previous lecture
- The residual plot below and the $R^2$ = .980 shows we have a very good fit to this model
- Next we will add dummy variables to reflect the condition of the apartment building




**Response PRICE**

**Actual by Predicted Plot**




**Summary of Fit**

| | |
|---|---|
| RSquare | 0.980 |
| RSquare Adj | 0.975 |
| Root Mean Square Error | 33.218 |
| Mean of Response | 290.574 |
| Observations (or Sum Wgts) | 25.000 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 5 | 1052904.6 | 210581 | 190.8408 |
| Error | 19 | 20965.3 | 1103 | Prob > F |
| C. Total | 24 | 1073869.9 | | <.0001* |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 92.789 | 28.688 | 3.23 | 0.0044* |
| #APTS | 4.140 | 1.490 | 2.78 | 0.0120* |
| AREA | 0.016 | 0.001 | 10.63 | <.0001* |
| LOTSIZE | 0.001 | 0.003 | 0.34 | 0.7411 |
| PARKING | 2.696 | 1.577 | 1.71 | 0.1036 |
| AGE | −0.853 | 0.299 | −2.86 | 0.0101* |



# Dummy variables in Multiple Regression

- We can also include dummy variables into the multiple regression model
- When other variables are in the model, the intercept now captures more than just the reference group
- But the interpretation of the coefficients for each dummy variable is still the same - how is that group different from the reference group?
  - The test test for the dummy variables is a test to see if there is a significant difference of that group with the reference group, ***holding all other factors constant***
  - Controlling for other variables makes this test more rigorous



# Apartment Data

- Let's see how our model changes when we add dummy variables of the condition of the apartment
- This is an expert judgement that the apartment condition is
  - **Excellent** (DUME)
  - **Good** (DUMG)
  - **Fair** we will let Fair be the reference category
- Now PRICE (per $1,000) is a function of the following attributes of the apartment building
  - #APTS
  - AREA
  - LOTSIZE
  - PARKING
  - AGE
  - DUMG
  - DUME

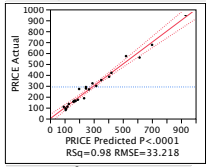


# Compare Models

- This model has a good fit and the coefficients are in the expected direction
- LOTSIZE and PARKING are insignificant and not important
- $R^2$ improved slightly to .988
- PARKING is now marginally significant
- DUME is positive and significant

**Regression of PRICE on Apartment Attributes**

**Actual by Predicted Plot**




**Summary of Fit**

| | |
|---|---|
| RSquare | 0.980 |
| RSquare Adj | 0.975 |
| Root Mean Square Error | 33.218 |
| Mean of Response | 290.574 |
| Observations (or Sum Wgts) | 25.000 |

**Analysis of Variance**

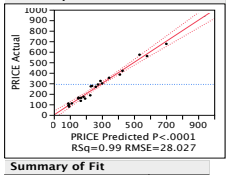
| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 5 | 1052904.6 | 210581 | 190.8408 |
| Error | 19 | 20965.3 | 1103 | Prob > F |
| C. Total | 24 | 1073869.9 | | <.0001* |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 92.7888 | 28.6883 | 3.23 | 0.0044* |
| #APTS | 4.1405 | 1.4898 | 2.78 | 0.0120* |
| AREA | 0.0155 | 0.0015 | 10.63 | <.0001* |
| LOTSIZE | 0.0010 | 0.0029 | 0.34 | 0.7411 |
| PARKING | 2.6958 | 1.5767 | 1.71 | 0.1036 |
| AGE | −0.8532 | 0.2988 | −2.86 | 0.0101* |

**PRICE Model including Condition**

**Actual by Predicted Plot**




**Summary of Fit**

| | |
|---|---|
| RSquare | 0.988 |
| RSquare Adj | 0.982 |
| Root Mean Square Error | 28.027 |
| Mean of Response | 290.574 |
| Observations (or Sum Wgts) | 25.000 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 7 | 1060515.8 | 151502 | 192.8648 |
| Error | 17 | 13354.1 | 786 | Prob > F |
| C. Total | 24 | 1073869.9 | | <.0001* |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 72.8287 | 31.9152 | 2.28 | 0.0356* |
| #APTS | 5.6445 | 1.3518 | 4.18 | 0.0006* |
| AREA | 0.0130 | 0.0015 | 8.75 | <.0001* |
| LOTSIZE | 0.0024 | 0.0025 | 0.97 | 0.3472 |
| PARKING | 2.9191 | 1.3670 | 2.14 | 0.0476* |
| AGE | −0.8385 | 0.2918 | −2.87 | 0.0105* |
| DumG | 7.1341 | 17.3156 | 0.41 | 0.6855 |
| DumE | 56.0228 | 20.8922 | 2.68 | 0.0158* |

# Compare the coefficients

**Table 1: Comparison of Two Models of PRICE on Apartment Attributes**

| Term | Model 1 | Model 2 |
|---|---|---|
| Intercept | 92.7888 | 72.8287 |
| #APTS | 4.1405** | 5.6445*** |
| AREA | 0.0155*** | 0.0130*** |
| LOTSIZE | 0.0010 | 0.0024 |
| PARKING | 2.6958 | 2.9191** |
| AGE | -0.8532** | -0.8385** |
| DUMG | | 7.1341 |
| DUME | | 56.0228** |
| | | |
| $R^2$ | 0.980 | 0.988 |

* p < .10; ** p<.05; ***p< .01

$$\hat{Y} = \$92.788 + \$4.140(\#\,\text{APTS}) + \$.0055(\text{AREA}) + \$.001(\text{LOTSIZE}) + \$2.696(\text{PARKING}) - \$.853(\text{AGE})$$

$$\hat{Y} = \$72.829 + \$5.645(\#\,\text{APTS}) + \$.0013(\text{AREA}) + \$.002(\text{LOTSIZE}) + \$2.919(\text{PARKING}) - \$.839(\text{AGE}) + \$56.023(\textit{Excellent}) + \$7.134(\textit{Good})$$

- **Model 1 is the first model**
  - There is a very good fit to this data
  - AREA, #APTS, and AGE are significant in the model
- **Model 2 adds in the Condition Dummy Variables**
  - $R^2$ is larger at .988
  - #APTS coefficient became larger and is more significant
  - AREA is reduced slightly
  - PARKING is larger and significant
  - DUME is positive and significant and adds about $56k in value



# The F-Test Hypothesis Test for a Regression Model

- **Ho:** Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$
- **Ha:** Ha: At least one $\beta_0 \neq 0$
- **Assumptions** Equal variances, normal distribution
- **Test Statistic** F* = 192.86 p < .001
- **Rejection Region** $F_{.05,\,7,\,17\ d.f.} = 2.10$
- **Conclusion:** F* > $F_{.05,\,7,\,17\ d.f.}$ or p < .001
  Reject Ho: Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

**We can see something is going on in the model**

**And we should explore the individual t-tests for the coefficients**



# The Hypothesis Test for a Regression Coefficient

- **Ho:** Ho: $\beta_{DUME} = 0$
- **Ha:** Ha: $\beta_{DUME} \neq 0$
- **Assumptions** Equal variances, normal distribution
- **Test Statistic** t* = 2.68 p = .0158
- **Rejection Region** $t_{.05/2,\,17} = 2.110$
- **Conclusion:** t* > $t_{.05,\,17}$ or p = .0158
  Reject Ho: $\beta_{DUME} = 0$

**Adding in Condition adds something to the model**

**The coefficient for DUME is significantly different from zero.**

**Apartments considered in Excellent condition get about $56,000 more in value compared with those in Fair condition, holding everything else constant.**



# An Analysis from start-to-finish

- A director of a training program for an insurance company wants to test to see if different methods of teaching produce different results.
- She randomly divides students into three groups - a traditional teaching group using lectures; a CD-ROM based learning approach; and a Web-based approach.
- Before taking the training the students took a proficiency test in basic math and computer skills.
- Following the training they took an end-of-training exam.
- For this research
  - The sample size is 30
  - The dependent variable is END-of-TRAINING
  - The independent variables are PROFICIENCY and training Type (represented by dummy variables)

# Here's the Data

- We need to create dummy variables for our Teaching Methods
- Decision: let the traditional method be the reference category so we can compare the alternative methods to the traditional
- The dummy variables look like this

| End-of-Training | Proficiency | Method |
|---|---|---|
| 14 | 94 | Traditional |
| 19 | 96 | Traditional |
| 17 | 98 | Traditional |
| 38 | 100 | Traditional |
| 40 | 102 | Traditional |
| 26 | 105 | Traditional |
| 41 | 109 | Traditional |
| 28 | 110 | Traditional |
| 36 | 111 | Traditional |
| 66 | 130 | Traditional |
| 38 | 80 | CD-ROM-based |
| 34 | 84 | CD-ROM-based |
| 43 | 90 | CD-ROM-based |
| 43 | 97 | CD-ROM-based |
| 61 | 97 | CD-ROM-based |
| 63 | 112 | CD-ROM-based |
| 93 | 115 | CD-ROM-based |
| 74 | 118 | CD-ROM-based |
| 76 | 120 | CD-ROM-based |
| 79 | 120 | CD-ROM-based |
| 55 | 92 | Web-based |
| 53 | 96 | Web-based |
| 55 | 99 | Web-based |
| 52 | 101 | Web-based |
| 35 | 102 | Web-based |
| 46 | 104 | Web-based |
| 57 | 107 | Web-based |
| 55 | 110 | Web-based |
| 42 | 111 | Web-based |
| 81 | 118 | Web-based |



# Let's look at some descriptive statistics

- END-of-TRAINING:
  - Mean is 48.667, larger than the median
  - The Stem-and-Leaf Plot does not show any major problems
- The Independent Variables
  - The mean PROFICIENCY is 104.267, less spread
  - The Training is equally spread across the three methods: the means are .333

**End-of-Training**

**Quantiles**

| | | |
|---|---|---|
| 100.0% | maximum | 93.000 |
| 99.5% | | 93.000 |
| 97.5% | | 93.000 |
| 90.0% | | 78.700 |
| 75.0% | quartile | 61.500 |
| 50.0% | median | 44.500 |
| 25.0% | quartile | 35.750 |
| 10.0% | | 19.700 |
| 2.5% | | 14.000 |
| 0.5% | | 14.000 |
| 0.0% | minimum | 14.000 |

**Moments**

| | |
|---|---|
| Mean | 48.667 |
| Std Dev | 19.707 |
| Std Err Mean | 3.598 |
| upper 95% Mean | 56.025 |
| lower 95% Mean | 41.308 |
| N | 30.000 |
| Sum Wgt | 30.000 |
| Sum | 1460.000 |
| Variance | 388.368 |
| Skewness | 0.314 |
| Kurtosis | -0.313 |
| CV | 40.494 |
| N Missing | 0.000 |

**Stem and Leaf**

| Stem | Leaf | Count |
|---|---|---|
| 9 | 3 | 1 |
| 8 | 1 | 1 |
| 7 | 469 | 3 |
| 6 | 136 | 3 |
| 5 | 235557 | 6 |
| 4 | 012336 | 6 |
| 3 | 45688 | 5 |
| 2 | 68 | 2 |
| 1 | 479 | 3 |

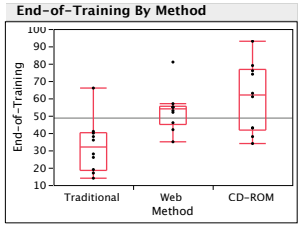
1|4 represents 14

| | End-of-Training | Proficiency | CDROM | Web |
|---|---|---|---|---|
| Mean | 48.667 | 104.267 | 0.333 | 0.333 |
| Standard Error | 3.598 | 2.063 | 0.088 | 0.088 |
| Median | 44.500 | 103.000 | 0.000 | 0.000 |
| Mode | 55.000 | 96.000 | 0.000 | 0.000 |
| Standard Deviation | 19.707 | 11.301 | 0.479 | 0.479 |
| Sample Variance | 388.368 | 127.720 | 0.230 | 0.230 |
| Coef of Variation | 40.494 | 10.839 | 143.839 | 143.839 |
| Kurtosis | -0.313 | -0.050 | -1.554 | -1.554 |
| Skewness | 0.314 | 0.050 | 0.745 | 0.745 |
| Range | 79.000 | 50.000 | 1.000 | 1.000 |
| Minimum | 14.000 | 80.000 | 0.000 | 0.000 |
| Maximum | 93.000 | 130.000 | 1.000 | 1.000 |
| Sum | 1460 | 3128 | 10 | 10 |
| Count | 30 | 30 | 30 | 30 |



# More background information

- If I look at End-of-Training by the Training Method I can see that the Traditional Method has the lowest average score, followed by Web and then CD-Rom
- The Simple ANOVA of End-of-Training by Training Method has an $R^2$ of .372; about 37.2% of the variability in the training test is explained by the training method
- The assumption of equal variances is not perfect, but not all together unreasonable
- In fact, an F-test for equal variances could not be rejected
- Note the means of each method




**Oneway Anova**

**Summary of Fit**

| | |
|---|---|
| Rsquare | 0.372 |
| Adj Rsquare | 0.325 |
| Root Mean Square Error | 16.188 |
| Mean of Response | 48.667 |
| Observations (or Sum Wgts) | 30.000 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Method | 2 | 4186.867 | 2093.43 | 7.9882 | 0.0019* |
| Error | 27 | 7075.800 | 262.07 | | |
| C. Total | 29 | 11262.667 | | | |

**Means for Oneway Anova**

| Level | Number | Mean | Std Error | Lower 95% | Upper 95% |
|---|---|---|---|---|---|
| Traditional | 10 | 32.500 | 5.119 | 21.996 | 43.004 |
| Web | 10 | 53.100 | 5.119 | 42.596 | 63.604 |
| CD-ROM | 10 | 60.400 | 5.119 | 49.896 | 70.904 |

Std Error uses a pooled estimate of error variance



# The Regression of the same results

- The Regression results are the same as ANOVA
- Except we get coefficients for each dummy variable
- Which estimate the difference in means from the reference group
  - The CDROM method has an average score that is 27.90 higher than the traditional method
  - The Web method has an average score that is 20.60 higher than the traditional method

**Whole Model**

**Summary of Fit**

| | |
|---|---|
| RSquare | 0.372 |
| RSquare Adj | 0.325 |
| Root Mean Square Error | 16.188 |
| Mean of Response | 48.667 |
| Observations (or Sum Wgts) | 30.000 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 2 | 4186.867 | 2093.43 | 7.9882 |
| Error | 27 | 7075.800 | 262.07 | Prob > F |
| C. Total | 29 | 11262.667 | | 0.0019* |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 32.500 | 5.119 | 6.35 | <.0001* |
| CDROM | 27.900 | 7.240 | 3.85 | 0.0007* |
| Web | 20.600 | 7.240 | 2.85 | 0.0084* |

**If I do a test of differences across all methods and adjust for Experiment-wise error, there is a significant difference between Traditional and the other two methods, but not between CDROM and Web**

## We should also look at the Correlations

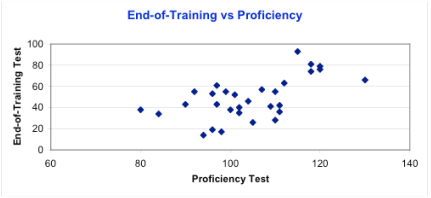
| | *End-of-Training* | *Proficiency* | *CDROM* | *Web* |
|---|---|---|---|---|
| **End-of-Training** | 1.000 | | | |
| **Proficiency** | 0.591 | 1.000 | | |
| **CDROM** | 0.428 | -0.062 | 1.000 | |
| **Web** | 0.162 | -0.017 | -0.500 | 1.000 |

- PROFICIENCY is moderately correlated with the End-of-Training score (.591)
- The CDROM method is positively correlated with End-of-Training (.428) indicating that those who received this training method had, on average, higher scores
- The correlation of End-of-Training with Web method was also positive, but much lower (.162)
- Since subject were assigned to each method randomly, there is little to no relationship between the teaching method dummy variables and the Proficiency score



## A closer look at End-of-Training and Proficiency

- We noted the positive correlation between End-of-Training score and the Proficiency score
- Those who tested as more proficient in math and computer skills tended to score higher at the end of the training
- Here's one way to think of Proficiency:
  - It is not the focus of the research, but
  - Proficiency is a good control variable to help assess the effectiveness of the training methods



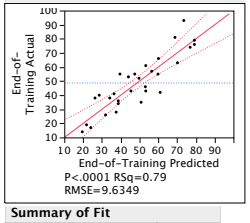



## Regression Results

- $R^2$ is moderately high, .786
- The F-test is significant at p < .001. We can safely conclude that at least of the coefficients are significantly different from zero
- When we examine the individual t-tests for the variables in the model
  - The coefficient for Proficiency is positive and significant at p< .001
  - The coefficient for Web is 22.289 and significant:
  - The CDROM coefficient is 30.377 and is significant and positive
- Looking at the standardized coefficients, CDROM was most important followed by Proficiency and Web.

**Whole Model**

**Actual by Predicted Plot**




**Summary of Fit**

| | |
|---|---|
| RSquare | 0.786 |
| RSquare Adj | 0.761 |
| Root Mean Square Error | 9.635 |
| Mean of Response | 48.667 |
| Observations (or Sum Wgts) | 30.000 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 3 | 8849.066 | 2949.69 | 31.7749 |
| Error | 26 | 2413.601 | 92.83 | Prob > F |
| C. Total | 29 | 11262.667 | | <.0001* |

**Lack Of Fit**

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Lack Of Fit | 24 | 2247.101 | 93.6292 | 1.1247 |
| Pure Error | 2 | 166.500 | 83.2500 | Prob > F |
| Total Error | 26 | 2413.601 | | 0.5759 |
| | | | | Max RSq |
| | | | | 0.9852 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| | Std Beta |
|---|---|---|---|---|---|
| Intercept | -86.270 | 17.034 | -5.06 | <.0001* | 0.000 |
| Proficiency | 1.126 | 0.159 | 7.09 | <.0001* | 0.646 |
| Web | 22.289 | 4.315 | 5.16 | <.0001* | 0.542 |
| CDROM | 30.377 | 4.323 | 7.03 | <.0001* | 0.739 |



## Compare the two models

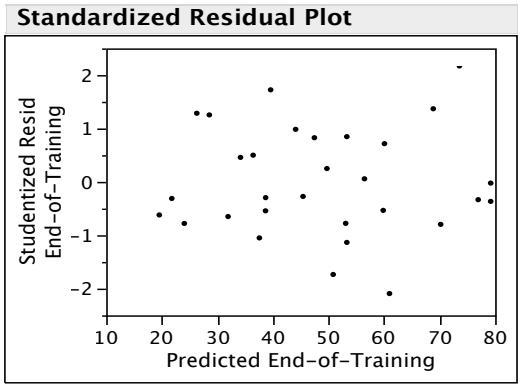
**Table 1: Comparison of Two Models of PRICE on Apartment Attributes**

| Term | Model 1 | Model 2 |
|---|---|---|
| **Intercept** | 32.500*** | -86.270*** |
| **Proficiency** | | 1.126*** |
| **Web** | 20.600*** | 22.289*** |
| **CDROM** | 27.900*** | 30.377*** |
| | | |
| **$R^2$** | 0.372 | 0.786 |

* p < .10; ** p<.05; ***p< .01

- $R^2$ increased from .372 to .786
- The coefficients for Web and CDROM increased after controlling for Proficiency
- The coefficient for the **CD-Rom** method is **30.377**, indicating this training method resulted in an average score that is 30 points higher than the traditional method, ***holding constant the student's proficiency***
- The coefficient for the **Web** method is **22.289**, indicating this training method resulted in an average score that is 22 points higher than the traditional method, ***holding constant the student's proficiency***

# Residual Results - RANDOM!!!






# Analysis Conclusions

- It was important to control for the students' proficiency in the analysis
- The CD-Rom teaching method produced a result that was on average 30 points higher than the traditional method
- The Web-based teaching method produced a result that was lower, but still on average 22 points higher than the traditional method
- Both alternative methods showed an improvement over the Traditional Method.
- The final decision on which method to use should factor in other things. The company may compare costs or other factors (students' perceptions of ease of use, convenience, how the instructor feels about the methods) to determine which alternative method is preferred.



# Summary

- Multiple regression is a powerful tool
- It allows us to make inferences about the effect of an independent variable on a dependent variable while controlling for other factors in the model – statistical control
- There is so much more to learn in regression
  - More complex tests
  - Non-linear relationships
  - Adding an element of time to the model
  - Fixes for data problems