UCLA STAT 251

Statistical Methods for the Life and Health Sciences

•**Instructor**: **Ivo Dinov**,

Asst. Prof. In Statistics and Neurology

University of California, Los Angeles, Winter 2003

http://www.stat.ucla.edu/~dinov/

Multiple Regression Analysis

Correlation Coefficient

Correlation coefficient ($-1 \le R \le 1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_Y) and the correlation coefficient, R. $R=1$, perfect positive correlation (straight line relationship), $R=0$, no correlation (random cloud scatter), $R=-1$, perfect negative correlation.

Computing $R(X,Y)$: (standardize, multiply, average)

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$X=\{x_1, x_2,\ldots, x_N\}$
$Y=\{y_1, y_2,\ldots, y_N\}$
(μ_X, σ_X), (μ_Y, σ_Y) sample mean / SD.

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

Student i	Height x_i	Weight y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	167	60	6	4.67	36	21.8089	28.02
2	170	64	9	8.67	81	75.1689	78.03
3	160	57	-1	1.67	1	2.7889	-1.67
4	152	46	-9	-9.33	81	87.0489	83.97
5	157	55	-4	-0.33	16	0.1089	1.32
6	160	50	-1	-5.33	1	28.4089	5.33
Total	966	332	0	≈ 0	216	215.3334	195.0

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x = \frac{966}{6} = 161\,\text{cm}, \quad \mu_y = \frac{332}{6} = 55\,\text{kg},$$

$$\sigma_x = \sqrt{\frac{216}{5}} = 6.573, \quad \sigma_y = \sqrt{\frac{215.3}{5}} = 6.563,$$

$$Corr(X,Y) = R(X,Y) = 0.904$$

Correlation Coefficient - Properties

Correlation is **quasi-invariant** w.r.t. linear transformations of X or Y

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) =$$

$$sign(a)\times sign(c)\times R(aX+b, cY+d), \quad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right) = \left(\frac{ax_k+b-(a\mu_x+b)}{|a|\times\sigma_x}\right) =$$

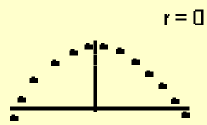
$$\left(\frac{a(x_k-\mu)+b-b}{|a|\times\sigma_x}\right) = sign(a)\times\left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y)=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=R(Y,X)$$

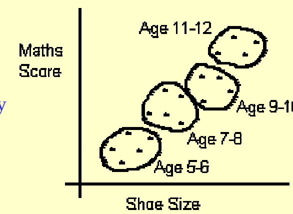
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



Correlation Coefficient - Properties

$$R(X,Y)=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=R(Y,X)$$

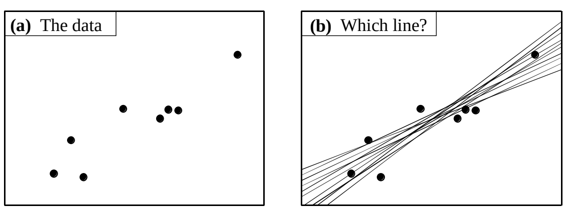
1. *R* measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.



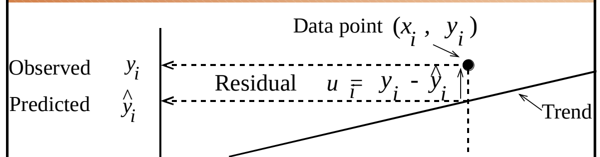
Fitting a line through the data

Show the Regression-Line Simulation Applet

RegressionApplet.html



The idea of a residual or prediction error



Least squares criterion

Least squares criterion: Choose the values of the parameters to *minimize the sum of squared prediction errors* (or sum of squared residuals),

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

The least squares line

Least-squares line

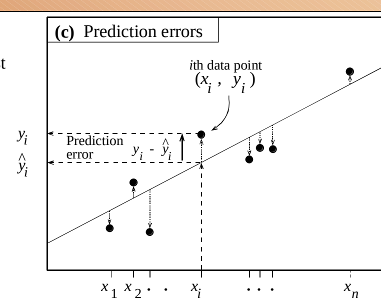
Choose line with smallest sum of squared prediction errors

Min $\Sigma\ (y_i-\hat{y}_i)^2$

Its parameters are denoted:

Intercept: $\hat{\beta}_0$

Slope: $\hat{\beta}_1$



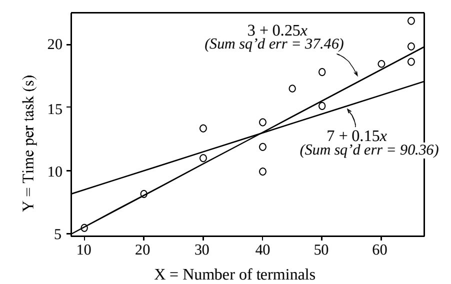
Least-squares line: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$

The least squares line

Least-squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

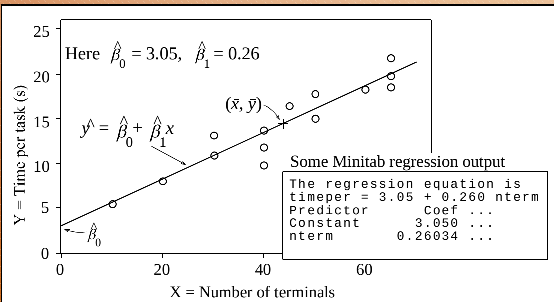
$$\hat{\boldsymbol{\beta}}_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \qquad \hat{\boldsymbol{\beta}}_0 = \bar{y} - \hat{\boldsymbol{\beta}}_1 \bar{x}$$

Computer timings data – linear fit



Two lines on the computer-timings data.

Adding the least squares line



Computer-timings data with least-squares line.

Example – Method/Hemi/Tissue/Value

1. C:\Ivo.dir\Research\Data.dir\WM_GM_CSF_tissueMaps.dir
2. SYSTAT: → regression Value = c_0+ c_1M+ c_2H+ c_3T
3. Results:

Effect	Coefficient	SE	t	P(2 Tail)
CONSTANT	1.02231E+05	9087	11.24911	0.00000
METHOD	-3703.77667	3635	-1.01887	0.31038 ← Insignif
TISSUE	-22623.47875	2226	-1.01E01	0.00000
HEMISPH	-2.13667	3635	-0.00059	0.99953

Effect	Coeff.	Lower < 95%>	Upper
CONSTANT	1.02231E+05	84231.33157	1.20231E+05
METHOD	-3703.77667	-10903.69304	3496.13971
TISSUE	-22623.47875	-27032.50908	-18214.44842
HEMISPH	-2.13667	-7202.05304	7197.77971

Introduction – Multiple Regression Analysis

- We extend the concept of **simple linear regression** as we investigate a response y which is affected by **several independent variables, $x_1, x_2, x_3, \ldots, x_k$.**
- Our objective is to use the information provided by the **x_i to predict the value of y.**

Example

- Let y be a student's college achievement, measured by his/her GPA. This might be a function of several variables:
  - x_1 = rank in high school class
  - x_2 = high school's overall rating
  - x_3 = high school GPA
  - x_4 = SAT scores
- We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Example

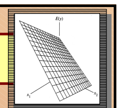
- Let y be the monthly sales revenue for a company. This might be a function of several variables:
  - x_1 = advertising expenditure
  - x_2 = time of year
  - x_3 = state of economy
  - x_4 = size of inventory
- We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Some Questions

- How well does the model fit?
- How strong is the relationship between y and the predictor variables?
- Have any assumptions been violated?
- How good are the estimates and predictions?

We collect information using n observations on the response y and the independent variables, x_1, x_2, x_3, $\ldots x_k$.

The General Linear Model



- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$
- where
- ✓ y is the response variable you want to predict.
- ✓ $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unknown constants
- ✓ $x_1, x_2, \ldots, x_k$ are independent predictor variables, measured without error.

The Random Error

- The deterministic part of the model,
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.
- describes average value of y for any fixed values of $x_1, x_2, \ldots, x_k$. The population of measurements is generated as y deviates from the **line of means** by an amount ε. We assume
  - ✓ ε are independent
  - ✓ Have a mean 0 and common variance σ^2 for any set $x_1, x_2, \ldots, x_k$.
  - ✓ Have a normal distribution.

Example

- Consider the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- This is a **first order model** (independent variables appear only to the first power).
- β_0 = ***y*-intercept** = value of $E(y)$ when $x_1 = x_2 = 0$.
- β_1 and β_2 are the **partial regression coefficients**—the change in y for a one-unit change in x_i **when the other independent variables are held constant**.
- Traces a **plane** in three dimensional space.

The Method of Least Squares

- The best-fitting prediction equation is calculated using a set of n measurements $(y, x_1, x_2, \ldots x_k)$ as

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

- We choose our estimates $b_0, b_1, \ldots, b_k$ to estimate $\beta_0, \beta_1, \ldots, \beta_k$ to minimize

$$\text{SSE} = \sum (y - \hat{y})^2$$

$$= \sum (y - b_0 - b_1 x_1 - \ldots - b_k x_k)^2$$

Example

- A computer database in a small community contains the listed selling price y (in thousands of dollars), the amount of living area x_1 (in hundreds of square feet), and the number of floors x_2, bedrooms x_3, and bathrooms x_4, for $n = 15$ randomly selected residences currently on the market.

Property	y	x_1	x_2	x_3	x_4
1	69.0	6	1	2	1
2	118.5	10	1	2	2
3	116.5	10	1	3	2
…	…	…	…	…	…
15	209.9	21	2	4	3

Fit a first order model to the data using the method of least squares.

Slide 26 STAT 251, UCLA, Ivo Dinov

Example

- **The first order model is**

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

fit using *Splus* with the values of y and the four independent variables entered into five columns of the output worksheet.

Regression equation

```
Regression Analysis: ListPrice versus SqFeet, NumFlrs, Bdrms, Baths
The regression equation is
ListPrice = 18.8 + 6.27 SqFeet - 16.2 NumFlrs - 2.67 Bdrms + 30.3 Baths

Predictor      Coef    SE Coef       T       P
Constant     18.763     9.207     2.04   0.069
SqFeet       6.2698    0.7252     8.65   0.000
NumFlrs     -16.203                      0.026
Bdrms        -2.673                      0.565
Baths        30.271                      0.001
```

Partial regression coefficients

Slide 27 STAT 251, UCLA, Ivo Dinov

The Analysis of Variance

- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = S_{yy} = \sum (y - \bar{y})^2$$

The **Total SS** is divided into two parts:

✓ **SSR** (sum of squares for regression): measures the variation explained by using the regression equation.

✓ **SSE** (sum of squares for error): measures the leftover variation not explained by the independent variables.

Slide 28 STAT 251, UCLA, Ivo Dinov

The ANOVA Table

Total *df* = **n -1**

Regression *df* = **k**

Error *df* = **n –1 – k = n – k -1**

Mean Squares

MSR = SSR/k

MSE = SSE/(n-k-1)

Source	df	SS	MS	F
Regression	k	SSR	SSR/k	MSR/MSE
Error	$n-k-1$	SSE	SSE/(n-k-1)	
Total	n -1	Total SS		

Slide 29 STAT 251, UCLA, Ivo Dinov

The Real Estate Problem

Another portion of the SYSTAT printout shows the ANOVA Table, with $n = 15$ and $k = 4$.

$\sqrt{MSE}$

```
S = 6.849      R-Sq = 97.1%      R-Sq(adj) = 96.0%

Analysis of Variance
Source            DF          SS          MS          F          P
Regression         4     15913.0
Residual Error    10       469.1
Total             14     16382.2

Source       DF      Seq SS
SqFeet        1     14829.3
NumFlrs       1         0.9
Bdrms         1       166.4
Baths         1       916.5
```

Sequential Sums of squares: conditional contribution of each independent variable to SSR given the variables already entered into the model.

Slide 30 STAT 251, UCLA, Ivo Dinov

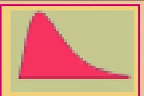
Testing the Usefulness of the Model

- The first question to ask is whether the regression model is of any use in predicting y.
- If it is not, then the value of y does not change, regardless of the value of the independent variables, $x_1, x_2, \ldots, x_k$. This implies that the partial regression coefficients, $\beta_1, \beta_2, \ldots, \beta_k$ are all zero.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ versus}$$
$$H_a : \text{at least one } \beta_i \text{ is not zero}$$

Slide 31 STAT 251, UCLA, Ivo Dinov

The F Test



- You can test the overall usefulness of the model using an F test. If the model is useful, MSR will be large compared to the unexplained variation, MSE.

To test H_0 : model is useful in predicting y is equivalent to

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

$$\text{Test Statistic: } F = \frac{MSR}{MSE}$$

Reject H_0 if $F > F_\alpha$ with k and $n-k-1$ df.

Measuring the Strength of the Relationship

- If the independent variables are useful in predicting y, you will want to know how well the model fits.
- The strength of the relationship between x and y can be measured using:

Multiple coefficient of determination :

$$R^2 = \frac{SSR}{\text{Total SS}}$$

Measuring the Strength of the Relationship

- Since Total SS = SSR + SSE, R^2 measures
  - ✓ the proportion of the total variation in the responses that can be explained by using the independent variables in the model.
  - ✓ the percent reduction the total variation by using the regression equation rather than just using the sample mean y-bar to estimate y.

$$R^2 = \frac{SSR}{\text{Total SS}} \quad \text{and } F = \frac{MSR}{MSE} = \frac{R^2 / k}{(1-R^2)/(n-k-1)}$$

Testing the Partial Regression Coefficients

- Is a particular independent variable useful in the model, *in the presence of all the other independent* variables? The test statistic is function of b_i, our best estimate of β_i.

The Real Estate Problem

Is the overall model useful in predicting list price? How much of the overall variation in the response is explained by the regression model?

```
S = 6.849        R-Sq = 97.1%      R-Sq(adj) = 96.0%

Analysis of Variance
Source             DF          SS          MS         F        P
Regression          4     15913.0      3978.3     84.80    0.000
                                9.1        46.9
Baths              1       916.5
```

R^2 = .971 indicates that 97.1% of the overall variation is explained by the regression model.

F = MSR/MSE = 84.80 with p-value = .000 is highly significant. The model is very useful in predicting the list price of homes.

...blem

...e independent ...ooms significant in ...s? Test using α = .05.

```
Regr                                   Bdrms, Baths
The
List                                   - 2.67 Bdrms + 30.3 Baths

Pred                                                   P
Constant       18.763       9.207        2.04      0.069
SqFeet         6.2698      0.7252        8.65      0.000
NumFlrs       -16.203       6.212       -2.61      0.026
Bdrms          -2.673       4.494       -0.59      0.565
Baths          30.271       6.849        4.42      0.001
```

To test H_0: $\beta_3 = 0$, the test statistic is t = -0.59 with p-value = .565.

The p-value is larger than .05 and H_0 is not rejected.

We cannot conclude that number of bedrooms is a valuable predictor in the presence of the other variables.

Perhaps the model could be refit without x_3.

Comparing Regression Models

The strength of a regression model is measured using R^2 = SSR/Total SS. This value will only increase as variables are added to the model.

- To fairly compare two models, it is better to use a measure that has been adjusted using *df*:

$$R^2(\text{adj}) = \left(1 - \frac{\text{MSE}}{\text{Total SS}/(n-1)}\right)100\%$$

Checking the Regression Assumptions

•Remember that the results of a regression analysis are only valid when the necessary assumptions have been satisfied.

- ✓ **ε** are independent
- ✓ Have a mean 0 and common variance σ^2 for any set $x_1, x_2, \ldots, x_k$.
- ✓ Have a normal distribution.

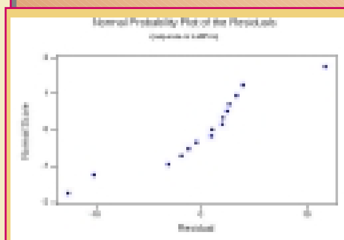
Diagnostic Tools

•We use the same diagnostic tools used in Chapter 11 and 12 to check the normality assumption and the assumption of equal variances.

1. **Normal probability plot** of residuals
2. Plot of **residuals versus fit** or **residuals versus variables**

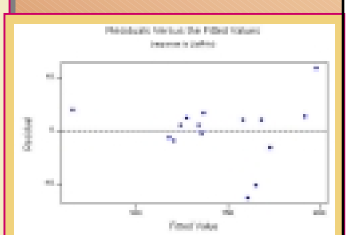
Normal Probability Plot

- ✓ If the normality assumption is valid, the plot should resemble a straight line, sloping upward to the right.
- ✓ If not, you will often see the pattern fail in the tails of the graph.



Residuals versus Fits

- ✓ If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line.
- ✓ If not, you will see a pattern in the residuals.



Estimation and Prediction

- **Once you have**
  - ✓ determined that the regression line is useful
  - ✓ used the diagnostic plots to check for violation of the regression assumptions.
- **You are ready to use the regression line to**
  - ✓ Estimate the average value of *y* for a given value of *x*
  - ✓ Predict a particular value of *y* for a given value of *x*.

Estimation and Prediction

- **Enter the appropriate values of $x_1, x_2, \ldots, x_k$ in SoftPackage to calculate**

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

- **and both the confidence interval and the prediction interval.**
- **Particular values of y are more difficult to predict, requiring a wider range of values in the prediction interval.**

The Real Estate Problem

- Estimate the average list price for a home with 1000 square feet of living space, one floor, 3 bedrooms and two baths with a 95% confidence interval.

We estimate that the average list price will be between $110,860 and $124,700 for a home like this.

```
Predicted Values for New Observations
New Obs     Fit     SE Fit        95.0% CI               95.0% PI
1         117.78     3.11     ( 110.86,  124.70)  (  101.02,  134.54)

Values of Predictors for New Observations
New Obs     SqFeet    NumFlrs     Bdrms      Baths
1             10.0       1.00      3.00       2.00
```


Using Regression Models

When you perform multiple regression analysis, use a step-by step approach:

1. Obtain the fitted prediction model.
2. Use the analysis of variance F test and R^2 to determine how well the model fits the data.
3. Check the t tests for the partial regression coefficients to see which ones are contributing significant information in the presence of the others.
4. If you choose to compare several different models, use R^2(adj) to compare their effectiveness.
5. Use diagnostic plots to check for violation of the regression assumptions.

A Polynomial Model

- A response y is related to a single independent variable x, but not in a linear manner. The polynomial model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$$

- When $k = 2$, the model is **quadratic**:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

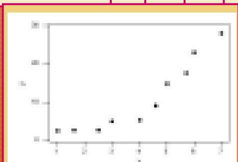
- When $k = 3$, the model is **cubic**:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

Example

- A market research firm has observed the sales (y) as a function of mass media advertising expenses (x) for 10 different companies selling a similar product.

Company	1	2	3	4	5	6	7	8	9	10
Expenditure, x	1.0	1.6	2.5	3.0	4.0	4.6	5.0	5.7	6.0	7.0
Sales, y	2.5	2.6	2.7	5.0	5.3	9.1	14.8	17.5	23.0	28.0



Since there is only one independent variable, you could fit a linear, quadratic, or cubic polynomial model. Which would you pick?

Two Possible Choices

A straight line model: $y = \beta_0 + \beta_1 x + \varepsilon$

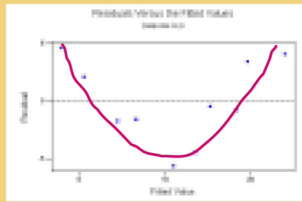
A quadratic model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

Printout for the straight line:

Overall F test is highly significant, as is the t-test of the slope. $R^2 = .856$ suggests a good fit. Let's check the residual plots…

```
Regression Analysis: y versus x
The regression equation is
y = - 6.47 + 4.34 x
Predictor        Coef     SE Coef        T       P
Constant       -6.465       2.795    -2.31   0.049
x              4.3355      0.6274     6.91   0.000
S = 3.725      R-Sq = 85.6%     R-Sq(adj) = 83.9%
Analysis of Variance
Source            DF          SS          MS        F        P
Regression         1      662.46      662.46    47.74    0.000
Residual Error     8      111.00       13.88
Total              9      773.46
```

Example



There is a strong pattern of a "curve" leftover in the residual plot.

This indicates that there is a curvilinear relationship unaccounted for by your straight line model. You should have used the quadratic model!

Use SYSTAT to fit the quadratic model:

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

The Quadratic Model

```
Regression Analysis: y versus x, x-sq
The regression equation is
y = 4.66 - 3.03 x + 0.939 x-sq

Predictor        Coef     SE Coef          T        P
Constant        4.657       2.443       1.91    0.098
x              -3.030       1.395      -2.17    0.067
x-sq           0.9389      0.1739       5.40    0.001
S = 1.752      R-Sq = 97.2%        R-Sq(adj) = 96.4%

                                   MS         F        P
                               375.99    122.49    0.000
                                 3.07
```

Overall F test is highly significant, as is the *t*-test of the quadratic term β_2. $R^2 = .972$ suggests a very good fit.

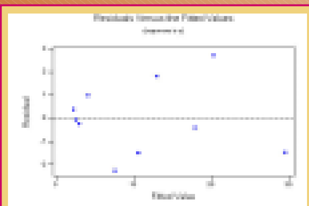
Let's compare the two models, and check the residual plots.

Which Model to Use?

Use R^2(adj) to compare the models:

The straight line model: $y = \beta_0 + \beta_1 x + \varepsilon$ $\quad R^2(\text{adj}) = 83.9\%$

The quadratic model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ $\quad R^2(\text{adj}) = 96.4\%$



The quadratic model is better.

There are no patterns in the residual plot, indicating that this is the correct model for the data.

Using Qualitative Variables

- Multiple regression requires that the response *y* be a quantitative variable.
- Independent variables can be either quantitative or qualitative.
- **Qualitative variables** involving *k* categories are entered into the model by using *k*-1 **dummy variables**.
- **Example:** To enter **gender** as a variable, use
  - $x_i = 1$ if male; 0 if female

Example

- Data was collected on 6 male and 6 female assistant professors. The researchers recorded their salaries (y) along with years of experience (x_1). The professor's gender enters into the model as a dummy variable: $x_2 = 1$ if male; 0 if not.

Professor	Salary, y	Experience, x_1	Gender, x_2	Interaction, x_1x_2
1	$50,710	1	1	1
2	49,510	1	0	0
…	…	…	…	…
11	55,590	5	1	5
12	53,200	5	0	0

Example

- We want to predict a professor's salary based on years of experience and gender. We think that there may be a difference in salary depending on whether you are male or female.
- The model we choose includes experience (x_1), gender (x_2), and an interaction term (x_1x_2) to allow salary's for males and females to behave differently.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Computer Out

What is the regression

Is the overall model useful in predicting y?

The overall F test is F = 346.24 with p-value = .000. The value of R^2 = .992 indicates that the model fits very well.

models.

```
Regression Analysis: y versus x1, x2, x1x2
The regression equation is
y = 48593 + 969 x1 + 867 x2 + 260

Predictor      Coef      SE Coef
Constant     48593.0      207.9
x1            969.00      63.67
x2             866.7      305.3
x1x2          260.13      87.06
```

Is there a difference in the relationship between salary and years of experience, depending on the gender of the professor?

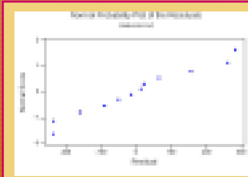
Yes. The individual t-test for the interaction term is t = 2.99 with p-value = .017. This indicates a significant interaction between gender and years of experience.

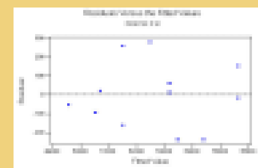
Exam

- Have any of the regres violated, or have we f

It does not appear from the diagnostic plots that there are any violations of assumptions.

The model is ready to be used for prediction or estimation.





Testing Sets of Parameters

- Suppose the demand y may be related to five independent variables, but that the cost of measuring three of them is very high.
- If it could be shown that these three contribute little or no information, they can be eliminated.
- You want to test the null hypothesis
- $\boldsymbol{H_0 : \beta_3 = \beta_4 = \beta_5 = 0}$ — that is, the independent variables x_3, x_4, and x_5 contribute no information for the prediction of y—versus the alternative hypothesis:
- $\boldsymbol{H_a :}$ **At least one of β_3, β_4, or β_5 differs from 0** — that is, at least one of the variables x_3, x_4, or x_5 contributes information for the prediction of y.

Testing Sets of Parameters

- To explain how to test a hypothesis concerning a set of model parameters, we define two models:
- **Model One (reduced model)**

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

- **Model Two (complete model)**

$$E(y) = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r}_{\text{terms in model 1}} + \underbrace{\beta_{r+1} x_{r+1} + \beta_{r+2} x_{r+2} + \cdots + \beta_k x_k}_{\text{additional terms in model 2}}$$

Testing Sets of Parameters

- The test of the hypothesis
- $\boldsymbol{H_0 : \beta_3 = \beta_4 = \beta_5 = 0}$
- $\boldsymbol{H_a :}$ **At least one of the β_i differs from 0**
- uses the test statistic $F = \dfrac{(\text{SSE}_1 - \text{SSE}_2)/(k-r)}{\text{MSE}_2}$

where F is based on $df_1 = (k - r)$ and $df_2 = n - (k + 1)$.
The rejection region for the test is identical to other analysis of variance F tests, namely $F > F_\alpha$.

Stepwise Regression

- A stepwise regression analysis fits a variety of models to the data, adding and deleting variables as their significance in the presence of the other variables is either **significant** or **nonsignificant**, respectively.
- Once the program has performed a sufficient number of iterations and no more variables are significant when added to the model, and none of the variables are nonsignificant when removed, the procedure stops.
- These programs **always fit first-order models** and are not helpful in detecting curvature or interaction in the data.

Important Points

- **Causality:** Be careful not to deduce a causal relationship between a response *y* and a variable *x*.
- **Multi-collinearity:** Neither the size of a regression coefficient nor its *t*-value indicates the importance of the variable as a contributor of information. This may be because two or more of the predictor variables are highly correlated with one another; this is called **multi-collinearity**.

Slide 62 STAT 251, UCLA, Ivo Dinov

Multicollinearity

- **Multicollinearity** can have these effects on the analysis:
  - The estimated regression coefficients will have **large standard errors**, causing imprecision in confidence and prediction intervals.
  - Adding or deleting a predictor variable may cause significant changes in the values of the other regression coefficients.

Slide 63 STAT 251, UCLA, Ivo Dinov

Multicollinearity

- How can you tell whether a regression analysis exhibits **multicollinearity**?
  - The value of R^2 **is large**, indicating a good fit, but the **individual *t*-tests are nonsignificant**.
  - The signs of the regression coefficients are contrary to what you would intuitively expect the contributions of those variables to be.
  - A matrix of correlations, generated by the computer, shows you which **predictor variables are highly correlated with each other** and with the response *y*.

Slide 64 STAT 251, UCLA, Ivo Dinov

Basic Concepts

I. The General Linear Model

1. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
2. The random error ε has a normal distribution with mean 0 and variance σ^2.

II. Method of Least Squares

1. Estimates $b_0, b_1, \ldots, b_k$ for $\beta_0, \beta_1, \ldots, \beta_k$, are chosen to minimize SSE, the sum of squared deviations about the regression line $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$.
2. Least-squares estimates are produced by computer.

Slide 65 STAT 251, UCLA, Ivo Dinov

Basic Concepts

III. Analysis of Variance

1. Total SS = SSR + SSE, where Total SS = S_{yy}. The ANOVA table is produced by computer.
2. Best estimate of σ^2 is $$\text{MSE} = \frac{\text{SSE}}{n-k-1}$$

IV. Testing, Estimation, and Prediction

1. A test for the significance of the regression, $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, can be implemented using the analysis of variance F test: $$F = \frac{\text{MSR}}{\text{MSE}}$$

Slide 66 STAT 251, UCLA, Ivo Dinov

Basic Concepts

2. The strength of the relationship between *x* and *y* can be measured using $$R^2 = \frac{\text{SSR}}{\text{Total SS}}$$ which gets closer to 1 as the relationship gets stronger.
3. Use residual plots to check for nonnormality, inequality of variances, and an incorrectly fit model.
4. Significance tests for the partial regression coefficients can be performed using the Student's t test with error $df = n - k - 1$: $$t = \frac{b_i - \beta_i}{\text{SE}(b_i)}$$

Slide 67 STAT 251, UCLA, Ivo Dinov

Basic Concepts

5. Confidence intervals can be generated by computer to estimate the average value of y, $E(y)$, for given values of $x_1, x_2, \ldots, x_k$. Computer-generated prediction intervals can be used to predict a particular observation y for given value of $x_1, x_2, \ldots, x_k$. For given $x_1, x_2, \ldots, x_k$, prediction intervals are always wider than confidence intervals.

Basic Concepts – Model Building

1. The number of terms in a regression model cannot exceed the number of observations in the data set and should be considerably less!
2. To account for a curvilinear effect in a **quantitative** variable, use a second-order polynomial model. For a cubic effect, use a third-order polynomial model.
3. To add a **qualitative** variable with k categories, use $(k-1)$ dummy or indicator variables.
4. There may be interactions between two qualitative variables or between a quantitative and a qualitative variable. Interaction terms are entered as $\beta x_i x_j$.
5. Compare models using R^2(adj).