H. Diagnostic plots of residuals

1. Plot residuals versus fitted values almost always
   a. For simple reg. this is about the same as residuals vs. x
   b. Look for outliers, curvature, increasing spread (funnel or horn shape); then take appropriate action

2. If data were collected over time, plot residuals versus time (to check for time trend and for serial correlation)

3. If normality is important, use normal probability plot (QQ plot in S-PLUS is same thing, but with axes reversed)
   a. Plot the residuals on the x-axis and the expected values of the ordered observations from a normal distribution on y-axis
   b. A straight line is expected if distribution is normal
II. Multiple Regression Models

A. Data

1. | Y   | X_1 | X_2 | X_3 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>15</td>
<td>-37</td>
<td>3.331</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>59</td>
<td>1.111</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

B. Linear regression models (Sect. 9.2.1)

1. Model with 2 X’s: \( \mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)

2. Ex: Y: 1st year GPA, X_1: Math SAT, X_2: Verbal SAT

3. Ex: Y = log(tree volume), X_1: log(height), X_2: log(diameter)
C. Important notes about interpretation of $\beta$’s

1. Causal conclusions can be made from randomized experiments but not from observational studies!!!!

2. Geometrically, $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ describes a plane: For a fixed value of $X_1$ the mean of $Y$ changes by $\beta_2$ for each one-unit increase in $X_2$

3. The meaning of a coefficient depends on what explanatory variables are included!!!! $\beta_1$ in $\mu(Y|X_1) = \beta_0 + \beta_1 X_1$ is not the same as $\beta_1$ in $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
D. Special constructed explanatory variables

1. *Polynomial terms*, eg $X^2$, for curvature (see Display 9.6)

2. *Indicator variables* to model effects of *categorical* variables
   
   a. One indicator variable ($X=0,1$) to distinguish 2 groups; Ex: $X = 1$ for females, 0 for males
   
   b. $(K-1)$ indicator variables to distinguish $K$ groups; Ex: $X_2 = 1$ if fertilizer B was used, 0 if A or C was used $X_3 = 1$ if fertilizer C was used, 0 if A or B was used

3. *Product terms* for interaction

\[ \mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) \]

So $\mu(Y|X_1,X_2=7) = (\beta_0 + \beta_2 7) + (\beta_1 + \beta_3 7)X_1$

$\mu(Y|X_1,X_2=9) = (\beta_0 + \beta_2 9) + (\beta_1 + \beta_3 9)X_1$

“The effect of $X_1$ on $Y$ depends on the value of $X_2$”
E. Hypothetical sex discrimination example

1. Data: \( Y_i = \text{salary for teacher i} \), \( X_{1i} = \text{their years of experience} \),
   \( X_{2i} = 1 \) if they were a male and 0 if they were a female

<table>
<thead>
<tr>
<th>i</th>
<th>( Y )</th>
<th>( X_1 )</th>
<th>Gender</th>
<th>( X_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23000</td>
<td>4</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>39000</td>
<td>30</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>29000</td>
<td>17</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>25000</td>
<td>7</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

categorical
factor

Indicator variable
2. Parallel lines model: \( \mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)

So for all females: \( \mu(Y|X_1,X_2=0) = \beta_0 + \beta_1 X_1 \)

and for all males: \( \mu(Y|X_1,X_2=1) = \beta_0 + \beta_1 X_1 + \beta_2 \)

For the subpopulation of teachers at any particular years of experience, the mean for males is \( \beta_2 \) more than that for females.
3. Model with interaction \( \mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 X_2) \)

So for females: \( \mu(Y|X_1, X_2=0) = \beta_0 + \beta_1 X_1 \)

and for males: \( \mu(Y|X_1, X_2=1) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 \)

The mean salary for unexperienced males is \( \beta_2 \) (dollars) more than for unexperienced females. The rate of increase in salary with increasing experience is \( \beta_3 \) (dollars/year) more for males than for females.
4. Modeling curvature (parallel quadratic curves):
\[ \mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 \]

5. Further multiple regression modeling, for example:
\[ \mu(\text{salary}|...) = \beta_0 + \beta_1 \text{exper} + \beta_2 \text{educ} + \beta_3 \text{exper}^2 + \beta_4 \text{male.ind.} \]
F. Notes about indicator variables

1. A t-test for $H_0: \beta_1 = 0$ in the regression of $Y$ on a single indicator variable $I_B$, $\mu(Y|I_B) = \beta_0 + \beta_2 I_B$ is the 2-sample t-test

2. Regression when all explanatory variables are categorical is “analysis of variance”

3. Regression with categorical variables and one numerical $X$ is often called “analysis of covariance”

4. Short-hand notation: $\mu(\text{salary}|\text{service, SEX}) = \text{service} + \text{SEX}$

   **Lower case for a numerical explanatory variable**

   **Upper case for a categorical factor (this represents a set of indicator variables)**

   Interaction: service + SEX + service:SEX (S-PLUS notation)
G. Strategies and Graphical Tools

1. Strategy for data analysis with models (Section 9.4)

2. Graphical tools for exploration and communication
   a. Matrix of scatterplots (9.5.1)
   b. Coded scatterplot (9.5.2)  
      (different plotting codes for different categories)
   c. Jittered scatterplot (9.5.3)
   d. Point identification
The Plots2D and Annotation tool bars will be useful to us. These may be repositioned.
Left click on a column; ctrl left click on other columns (response var. last); then click Scatter Matrix button.
Move the cursor arrow to a point on the graph and it will identify the point by row label and variable values.
1. Insert legend; then edit, move it.
2. Data to Plot
   - Data Set: case0901
   - y Columns: flowers
   - z Columns: time
3. Vary Symbols
   - Vary Size By: None
   - Vary Color By: z Column
   - Vary Style By: z Column
4. Coded Scatterplot
   - Time = 1
   - Time = 2
Jittered Scatterplot (to separate points visually)
To label points
H. EXAMPLE

1. Rainfall and Corn Yield (example 9.15)

   a. Initial scatterplot of yield vs rainfall, and residual plot from simple linear regression fit
b. Quadratic fit

That looks better. Since data were collected over time we should check for time trend and serial correlation, by plotting residuals vs. time
d. Include year in the model: yield ~ rain + (rain^2) + year

There does appear to be a trend. There’s no obvious serial correlation (more on that in Ch. 15).
e. Partly because of the outliers and partly because we suspect that the effect of rain might be changing over 1890 to 1928 (because of improvements in agricultural techniques, including irrigation), it seems appropriate to further investigate the interactive effect of year and rainfall on yield.

This conditional scatterplot shows the effect of rainfall on yield to be smaller in later time periods (trellis plot in S-PLUS)
**f. Final model: yield ~ rain + (rain^2) + year + rain:year**

Coefficients:

|           | Value     | Std. Error | t value | Pr(>|t|) |
|-----------|-----------|------------|---------|----------|
| (Intercept)| -1909.4647| 486.2435   | -3.9270 | 0.0004   |
| RAINFALL  | 158.8411  | 44.5681    | 3.5640  | 0.0011   |
| YEAR      | 1.0012    | 0.2554     | 3.9193  | 0.0004   |
| I(RAINFALL^2) | -0.1862  | 0.0720     | -2.5876 | 0.0143   |
| RAINFALL:YEAR | -0.0806 | 0.0234     | -3.4391 | 0.0016   |

Residual standard error: 3.028 on 33 degrees of freedom
Multiple R-Squared: 0.5706
F-statistic: 10.96 on 4 and 33 degrees of freedom, the p-value is 9.127e-006

**g. Summary of findings**

As evident in the scatterplot below, the mean yearly yield of corn in six midwestern states from 1890 to 1927 increased with increasing rainfall up to a certain optimum rainfall, and then leveled off or decreased with rain in excess of that amount (the p-value from a t-test for the quadratic effect of rainfall on mean corn yield is .014). There is strong evidence, however, that the effect of rainfall changed over this period of observation (p-value from a t-test for the interactive effect of year and rainfall is .002). Representative quadratic fits to the regression of corn yield on rainfall are shown in the plot—for 1890, 1910, and 1927. It is apparent that less rainfall was needed to produce the same mean yield as time progressed. The “optimum” rainfall amounts for these three years were 17, 18, and 9 inches respectively.
Yearly Corn Yield Versus Rainfall for Six Midwestern States between 1890 and 1927
(Plotting Code is Year)
2. S-PLUS: Conditional plots (Trellis graphs) as on p. 40 above

   a. It is desired to make a 2D plot for each of several levels or intervals of a third variable

   b. (1) make the ordinary 2D plot (of y vs x, say)
      (2) On the standard tool bar, click “conditioning mode” icon
      (3) Highlight the column of the data set corresponding to the third variable; move the cursor over a data entry so that the pointer is displayed (i.e. not the down arrow).
      (4) Left click and drag the column into the graph, between the top of the plot and the green area; a dashed-lined box will appear; move the cursor in the box and release.
3. S-PLUS Adding lines (or points) to plots (as on p. 42 above)

   a. (1) Make a plot; make sure it is highlighted. (2) From the “INSERT” submenu, select “Plot”. (3) Add the plot you want to include on the existing one.

   b. Note: for p. 42 above I created a new column called `x`, with 100 rainfall values ranging from 6 to 17 (using DATA, Fill). Then I created 3 new columns of predicted values using the regression output (p. 41 above) with `year=1890, 1910, 1927`.