

## Chapter 6

Consider a population regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

(The following discussion also applies to multiple regression equations)

- For a given value of the predictor variable(s)  $x_i$ , the distribution of the response variable  $y_i$  is the **Conditional Distribution** of the variable  $y_i$ .
- For a given value of the predictor variable(s)  $x_i$ , the mean of the variable  $y_i$  is the **Conditional Mean** of the variable  $y_i$ .

### Assumptions for Regression Inferences

1. **Population Regression Line** There are constants  $\beta_0$  and  $\beta_1$  such that, for each value  $x_i$  of the predictor variable, the conditional mean of the response variable is  $\beta_0 + \beta_1 x_i$ .
2. **Equal Standard Deviations:** The conditional standard deviations of the response variable have the same value  $\sigma_e$  for each value of the predictor variable.
3. **Normal Populations:** For each value of the predictor variable, the conditional distribution of the response variable is normal.
4. **Independent Observations:** The observations of the response variable are independent of one another.

### Standard Error of the Estimate

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

Recall that

$$SSE = \sum (y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

**Interpretation:**  $s_e$  is how much, on average, the predicted values of the response variable differ from the observed values of the response variable.

$s_e$  is used to estimate the common conditional standard deviation  $\sigma_e$  of assumption 2.

### Example

The Price of a used car depends on its age. 11 Orions are sampled, and their prices and ages are as follows:

Car	Age $x$	Price $y$ (100s)
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

- Estimate the mean age of a 5 year old Orion.
- Estimate the common standard deviation of the price of an Orion.
- Illustrate the conditional distribution of the price of a 5 year old Orion. (See Handout)

**Q:** How do we find the population regression equation?

**A:** We estimate it with a regression equation for sample data.

- The regression equation for the Orion data is  $\hat{y} = 195.47 - 20.26x$ .
- Notice the difference between the sample regression equation and the population regression equation.
- A different sample will yield a different regression line.

### Analysis of Residuals

**The Problem** To use sample data to determine whether assumptions 1-3 are met.

## Residuals

$$e_i = y_i - \hat{y}_i$$

**Residual Plot** A scatterplot of the residual values vs the predictor variable.

**Example** Use SPSS to make a Residual Plot of the Orion Data

- Run **Analyze/Regression/Linear** to find the regression line
- In the box where you define the dependent and independent variables, select the **Save** option.
- Check the **Unstandardized Residuals** box. (This will compute the residuals for each observed value of  $x$  and save them on the data sheet).
- Click OK to find the regression line.
- The residuals will be saved in the column, **Res\_\_1**
- Now make a scatter plot with Res1 on the vertical axis and AGE on the horizontal axis.

**Using Residual Plots to test whether assumptions 1 through 3 of the regression model are met**

- Assumption 1 implies the residuals should be scattered about the  $x$  axis.
- Assumption 2 implies the variation of the observed values of the response variable is constant from one variable to the next, so *the residuals fall roughly in a horizontal band*.
- Assumption 3 implies that the horizontal band should be *centered* and *symmetric* about the  $x$  axis.

## Standardized Residuals

- These are residuals divided by their standard deviation

$$\hat{e}_{is} = \frac{\hat{e}_i}{stdev(\hat{e}_i)}$$

- The standardized residuals have a mean of 0 and a standard deviation of 1
- We often plot the standardized residuals rather than the residuals.
- The residual plot and standardized residual plot will look the same,
- The standardized residuals typically range from  $-3$  to  $3$ , by the empirical rule
- The standardized residual plot will be more useful for determining whether the normality assumption is violated.

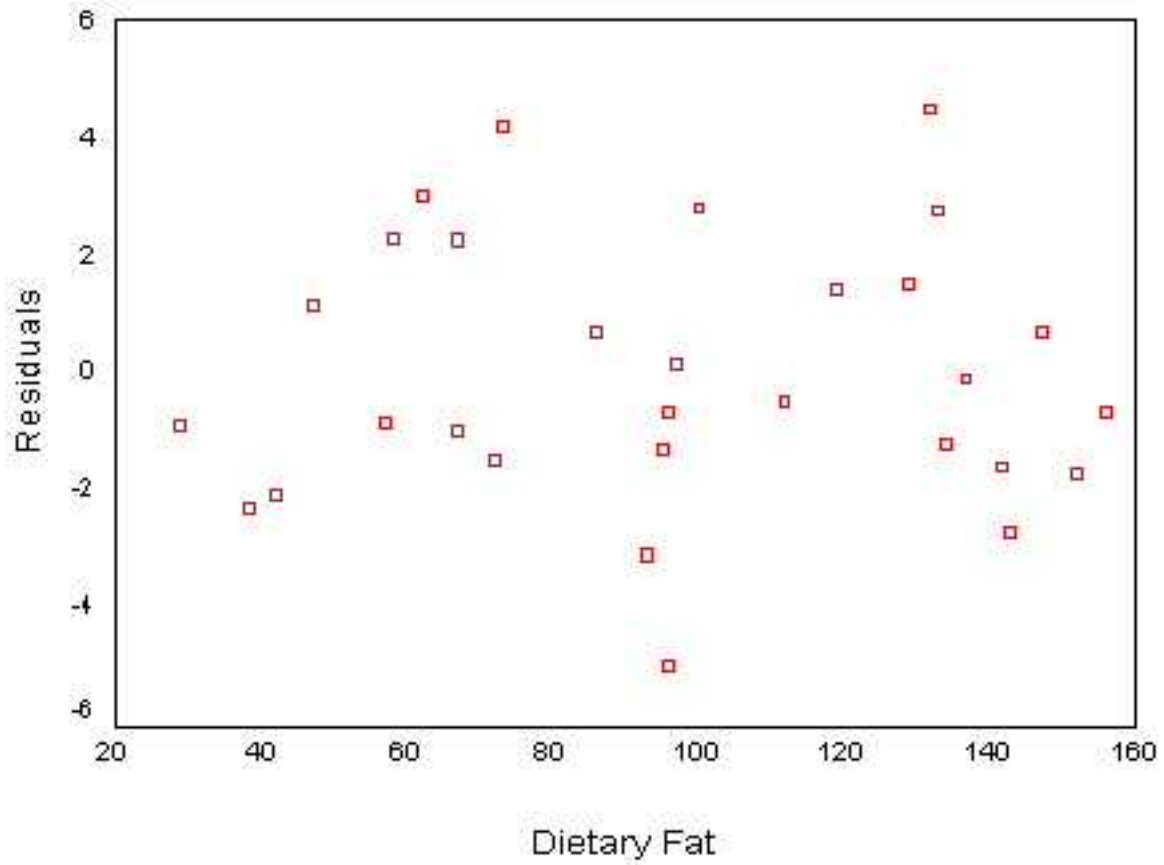
**Example** The data in the file

**U:MT Student File Area/dgarth/stat378/fat.sav**

gives the average dietary fat intake for citizens in 30 countries, as well as the death rate from prostate cancer in those countries.

- Obtain a Scatter Diagram on the data. Does obtaining a regression equation for the data seem reasonable?
- Construct a standardized residual plot to determine whether the assumptions for regression inferences are satisfied.

## Residual Plot of Prostate Data



## Using Residual Plots to Assess the Assumptions and Choose a Model

In general, given a data set, follow these steps

1. Run the full regression, test the overall fit. Save residuals and predicted values.
2. Make a scatter plot of the response versus each explanatory variable to see what type of curve might be the best choice. (Linear, polynomial, reciprocal, logarithmic, or log of dependent variable)
3. Make standardized residuals plots, plotting the standardized residuals versus the predicted values and each explanatory variables.
4. If the scatter plot suggests a transformation be made, and/or if the residual plots suggest an assumption is violated, choose an appropriate transformation.
5. Come up with the new model, and check the residual plots again to make sure assumptions are met
6. If applicable, compare the  $R^2$ ,  $R_{adj}^2$ , and  $s_e$  values to make sure the transformed model is an improvement. (This doesn't work if the dependent variable is transformed).
7. Eliminate any unnecessary variables.

**Example** Consider the telemarketing data stored in the **telemark6.sav** file in the Chapter 6 folder on the U: drive. Go through the steps to come up with a model using CALLS as the dependent variable and MONTHS as the explanatory variable.

- Run the full regression. Save the predicted values and the residuals.
- Make a scatter plot of CALLS vs MONTHS. What transformation does this suggest?
- Make a standardized residual plot of the data, plotting residuals vs MONTHS. Which assumptions, if any, appear to be violated?
- Try a quadratic model. Save the predicted values and residuals.

- Check to see that the transformed model is an improvement.

- Make residual plots to check that the assumptions are met for the quadratic model.

**Example (Page 219, problem 2)** The relationship between PROFIT and expenditure on research and development (RD) and the calculated risk (RISK) for a company is contained in the file **rd6.sav** in the chapter six folder. Find a model using PROFIT as the explanatory variable.

- Run the full regression. Save the predicted values and the residuals.
- Make a scatter plot of the dependent variable vs the explanatory variables. What transformation does this suggest?
- Make a standardized residual plot of the data, plotting residuals vs predicted values and MONTHS. Which assumptions, if any, appear to be violated?
- Try a logarithmic transformation first. Save the predicted values and residuals.
- Check to see that the transformed model is an improvement.
- Make residual plots to check that the assumptions are met for the logarithmic model.
- Try a quadratic model. Save the predicted values and residuals.

- Check to see that the transformed model is an improvement over the linear model.
- Make residual plots to check that the assumptions are met for the quadratic model.

**Assessing the constant variance assumption** A “cone shaped” residual plot indicates a violation of the constant variance assumption. Often this is remedied by taking a log transformation of the dependent variable.

**Example** A physician collected data on the age  $x$  and peak heart rate  $y$  of 10 patients. The data is stored in the file

U:MT Student File Area/dgarth/stat378/heart.sav

- Use residual plots to test the assumption of constant variance.
- Make a log transformation of both the dependent and independent variables.

### Assessing the Normality Assumption

- Make a histogram of the residuals vs the explanatory variables. The histogram should be bell shaped.
  - Use the program **Graphs/Interactive/Histogram** or **Graphs/Legacy Dialogs/Histograms**
- If this assumption is violated, then the residuals will not be centered around the horizontal axis.
- The spacing of the residuals should also follow the empirical rule.
  - 68% of the *standardized* residuals should lie between  $-1$  and  $1$
  - 95% should lie between  $-2$  and  $2$
  - 99% should lie between  $-3$  and  $3$
- In addition, since if the residuals are normally distributed, the **Normal Probability Plot** of the residuals ought to be linear

**Normal Probability Plot** This is a scatterplot of the residuals versus the *expected z scores* for each residual.

- In other words, if a residual came from a normal distribution, the expected  $z$  score for that residual is the  $z$  score we would expect that residual to have, *had it come from a normal distribution*
- Usually the horizontal axis plots the residuals, while the vertical axis plots either the expected  $z$  score or the cumulative probability for that expected  $z$  score
- In SPSS, the horizontal axis plots the cumulative probability of each observed residual.

**Example** The file **comnodes6.sav** in the chapter 6 folder on the U: drive contains the data for the COST of communications nodes in terms of the number of ports NUMPORTS and the bandwidth BANDWIDTH. Test the normality assumption.

- Make residual plots versus the predicted values as well as each explanatory variable.
- Make a normal probability plot.

### Remarks

- The linearity of the normal probability plot has nothing to say about whether the response variable is linearly related to the explanatory variables. It only says whether the residuals are normally distributed.
- In correcting for a normality violation, check first for linearity or constant variance violations. These violations can create apparent normality violations as well.

### Corrections for Violations in the Normality Assumptions

- Obtain a larger sample size
- Take the log of the dependent variable.

**Example (Page 266, Problem 17)** The file **realest6.sav** contains data on the properties in Tarrant County. The variables are as follows.

VALUE	The appraised value of the property
SIZE	Size in square feet
CONDITION	physical condition index
DEPRECIATION	depreciation factor

Develop an equation that might be useful for assessing values. Keep the following in mind.

- Check for violations in assumptions
- Check for transformations that give a better fit or fix violated assumptions

Once you have this equation, use it to assess the value of a 1400 square foot house with a physical condition index of 0.7 and a depreciation factor of 0.02.