# Remedial Measures, Brown-Forsythe test,F test

Frank Wood

February 18, 2010

# Remedial Measures

- How do we know that the regression function is a good explainer of the observed data?
  - Plotting
  - Tests
- What if it is not? What can we do about it?
  - Transformation of variables(next lecture)

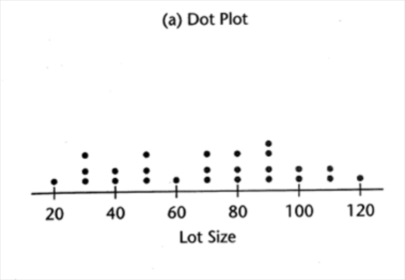
# Graphical Diagnostics for the Predictor Variable

- Dot Plot
  - Useful for visualizing distribution of inputs
- Sequence Plot
  - Useful for visualizing dependencies between error terms
- Box Plot - Useful for visualizing distribution of inputs

Toluca manufacturing example again: production time vs. lot size

# Dot Plot

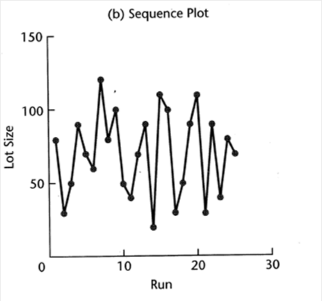
Figure:




- How many observations per input value?
- Range of inputs?

# Sequence Plot

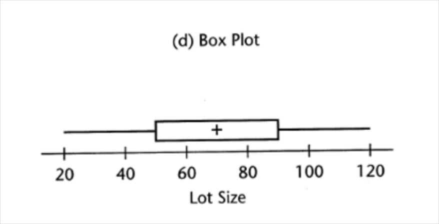
Figure:




- If observations are made over time, is there a correlation between input and position in observation sequence?

# Box Plot

Figure:




- Shows
  - Median
  - 1st and 3rd quartiles
  - Maximum and minimum

# Residuals

- Remember, the definition of residuals:

$$e_i = Y_i - \hat{Y}_i$$

- And the difference between that and the unknown true error

$$\epsilon = Y_i - E(Y_i)$$

- In a normal regression model the $\epsilon_i$'s are assumed to be iid $N(0, \sigma^2)$ random variables. The observed residuals $e_i$ should reflect these properties.

# Remember: residual properties

- Mean

$$\bar{e}_i = \frac{\sum e_i}{n} = 0$$

- Variance

$$s^2 = \frac{(e_i - \bar{e})^2}{n-2} = \frac{SSE}{n-2} = MSE$$

# Nonindependence of Residuals

- The residuals $e_i$ are not independent random variables - The fitted values $\hat{Y}_i$ are based on the same fitted regression line.
  - The residuals are subject to two constraints
  - Sum of the $e_i$'s equals 0 - Sum of the products $X_i e_i$'s equals 0
- When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals $e_i$ can reasonably safely be ignored.

# Definition: semistudentized residuals

- Like usual, sine the standard deviation of $\epsilon_i$ is $\sigma$(itself estimated by square root of MSE) a natural form of standardization to consider is
$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$
- This is called a semistudentized residual.

# Departures from Model...

To be studied by residuals

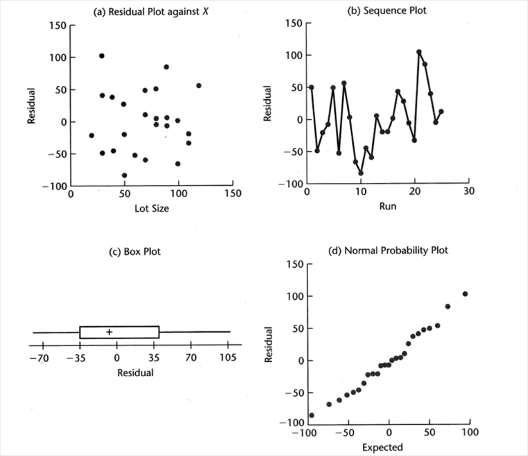
- Regression function not linear
- Error terms do not have constant variance
- Error terms are not independent
- Model fits all but one or a few outlier observations
- Error terms are not normally distributed
- One or more predictor variables have been omitted from the model

# Diagnostics for Residuals

- Plot of residuals against predictor variable
- Plot of absolute or squared residuals against predictor variable
- Plot of residuals against fitted values
- Plot of residuals against time or other sequence
- Plot of residuals against omitted predictor variables
- Box plot of residuals
- Normal probability plot of residuals

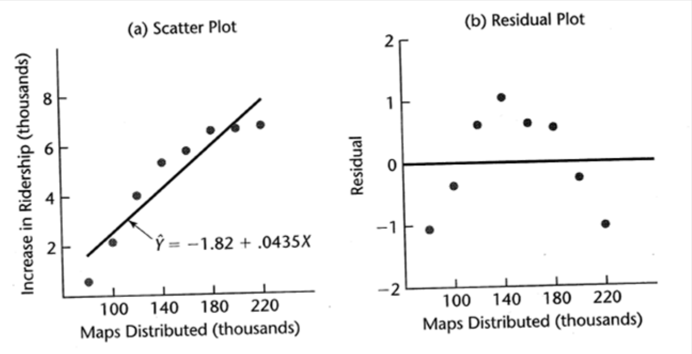
# Diagnostic Residual Plots

Figure:



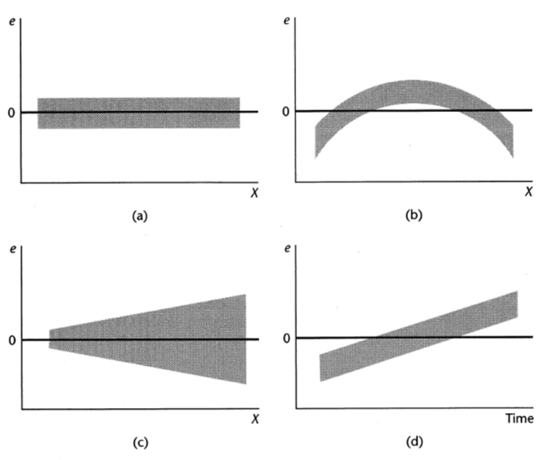
# Scatter and Residual Plot

Figure:



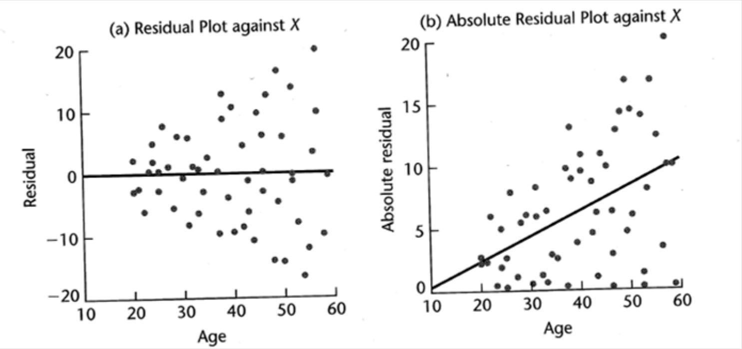
# Prototype Residual Plots

Figure:



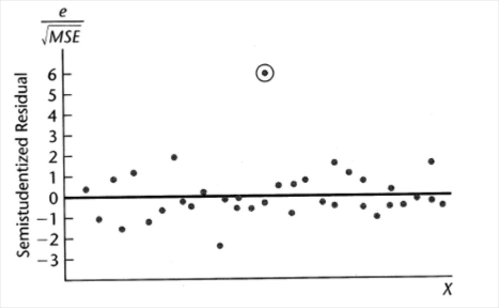
# Nonconstancy of Error Variance

Figure:



# Presence of Outliers

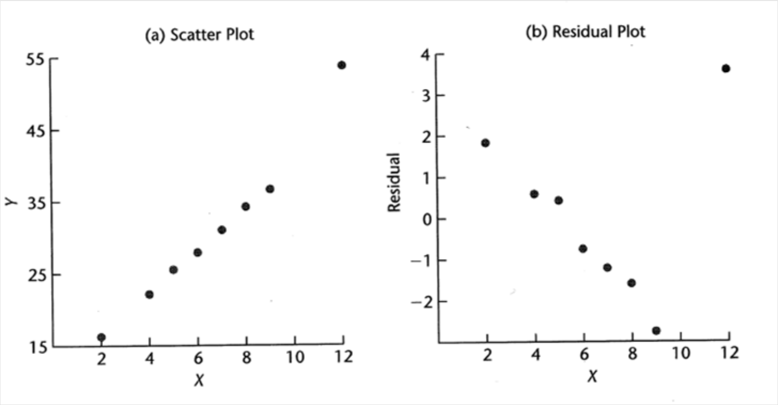
Figure:




Outliers can strongly effect the fitted values of the regression line.

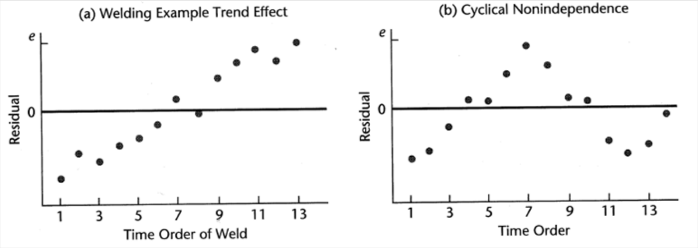
# Outlier effect on residuals

Figure:



# Nonindependence of Error Terms

Figure:

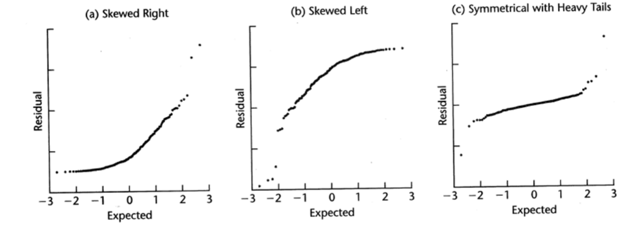



Sequential observations

# Non-normality of Error Terms

- Distribution plots
- Comparison of Frequencies
- Normal probability plot

Figure:



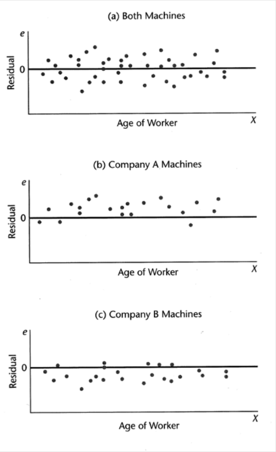
# Normal probability plot

- For a $N(0, MSE^{1/2})$ random variable, a good approximation of the expected value of the k-th smallest observation in a random sample of size n is
$$\sqrt{MSE}[z(\frac{k-.375}{n+.25})]$$
- A normal probability plot consists of plotting the expected value of the k-th smallest observation against the observed k-th smallest observation

# Omission of Important Predictor Variables

- Example
  - Qualitative variable
  - Type of machine
- Partitioning data can reveal dependence on omitted variable(s)
- Works for quantitative variables as well
- Can suggest that inclusion of other inputs is important

Figure:



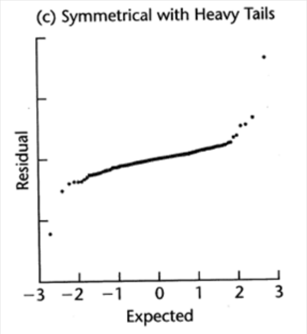
# Tests Involving Residuals

- Tests for randomness
- Tests for constancy of variance
- Tests for outliers
- Tests for normality

# Correction Test for Normality

- Calculated the coefficient of correlation between residuals $e_i$ and their expected values under normality
$$r = \sqrt{(R^2)}$$
- Tables(B.6 in the book) given critical values for the null hypothesis (normally distributed errors) holding.

Figure:



# Tests for Constancy of Error Variance

- Brown-Forsythe test does not depend normality of error terms.
  - The Brown-Forsythe test is applicable to simple linear regression when
  - The variance of the error terms either increases or decreases with X
  - Sample size is large enough to ignore dependencies between the residuals
- Basically a t-test for testing whether the means of two normally distributed populations are the same

# Brown-Forsythe Test

- Divide X into $X_1$(the low values of X) and $X_2$(the high values of X)
- Let $e_{i1}$ be the error terms for $X_1$ and vice versa
- let $n = n_1 + n_2$
- The Brown-Forsythe test uses the absolute deviations of the residuals around their group median

$$d_{1i} = |e_{1i} - \tilde{e}_1|$$

# Brown-Forsythe Test

- The test statistic for comparing the means of the absolute deviations of the residuals around the group medians is
$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
where
$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i1} - \bar{d}_1)^2}{n-2}$$

## Brown-Forsythe Test

- If $n_1$ and $n_2$ are not extremely small
$$t^*_{BF} \sim t(n-2)$$
approximately
- From this confidence intervals and test can be constructed.

# F test for lack of fit

- Formal test for determining whether a specific type of regression function adequately fits the data.
- Assumptions(usual):
  - $Y|X$
  - iid - normally distributed - same variance $\sigma^2$
- Requires: repeat observations at one or more X levels(called replicates)

# Example

- 12 similar branches of a bank offered gifts for setting up money market accounts
- Minimum initial deposits were specific to qualify for the gift
- Value of gift was proportional to the specified minimum deposit
- Interested in: relationship between specified minimum deposit and number of new accounts opened

# F Test Example Data and ANOVA Table

Figure:

(a) Data

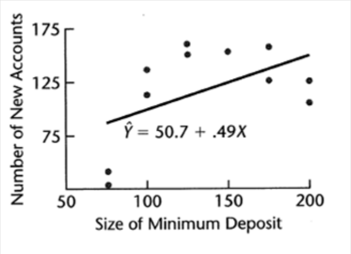
| Branch $i$ | Size of Minimum Deposit (dollars) $X_i$ | Number of New Accounts $Y_i$ | Branch $i$ | Size of Minimum Deposit (dollars) $X_i$ | Number of New Accounts $Y_i$ |
|---|---|---|---|---|---|
| 1 | 125 | 160 | 7 | 75 | 42 |
| 2 | 100 | 112 | 8 | 175 | 124 |
| 3 | 200 | 124 | 9 | 125 | 150 |
| 4 | 75 | 28 | 10 | 200 | 104 |
| 5 | 150 | 152 | 11 | 100 | 136 |
| 6 | 175 | 156 | | | |

(b) ANOVA Table

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 5,141.3 | 1 | 5,141.3 |
| Error | 14,741.6 | 9 | 1,638.0 |
| Total | 19,882.9 | 10 | |

# Fit

Figure:



Number of New Accounts
175
125
75
$\hat{Y} = 50.7 + .49X$
50
100
150
200
Size of Minimum Deposit

# Data Arranged To Highlight Replicates

Figure:

| Replicate | Size of Minimum Deposit (dollars) $j=1$ $X_1=75$ | $j=2$ $X_2=100$ | $j=3$ $X_3=125$ | $j=4$ $X_4=150$ | $j=5$ $X_5=175$ | $j=6$ $X_6=200$ |
|---|---|---|---|---|---|---|
| $i=1$ | 28 | 112 | 160 | 152 | 156 | 124 |
| $i=2$ | 42 | 136 | 150 | | 124 | 104 |
| Mean $\bar{Y}_j$ | 35 | 124 | 155 | 152 | 140 | 114 |

- The observed value of the response variable for the i-th replicate for the j-th level of X is $Y_{ij}$
- The mean of the Y observations at the level $X = X_j$ is $\bar{Y}_j$

# Full Model vs. Regression Model

- The full model is

$$Y_{ij} = \mu_j + \epsilon_{ij} \qquad \text{Full model}$$

  where
  - $\mu_j$ are parameters $j = 1, ..., c$
  - $\epsilon_{ij}$ are iid $N(0, \sigma^2)$

- Since the error terms have expectation zero

$$E(Y_{ij}) = \mu_{ij}$$

# Full Model

- In the full model there is a different mean(a free parameter) for each $X_i$
- In the regression model the mean responses are constrained to lie on a line

$$E(Y) = \beta_0 + \beta_1 X$$

# Fitting the Full Model

- The esstimators of $\mu_j$ are simply
$$\hat{\mu}_j = \bar{Y}_j$$
- The error sum of squares of the full model therefore is
$$SSE(F) = \sum\sum(Y_{ij} - \bar{Y}_j)^2 == SSPE$$

# Degrees of Freedom

- Ordinary total sum of squares had n-1 degrees of freedom.
- Each of the j terms is a ordinary total sum of squares
  - Each then has $n_j - 1$ degrees of freedom
- The number of degrees of freedom of SSPE is the sum of the component degrees of freedom

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

# General Linear Test

- Remember: the general linear test proposes a reduced model null hypotheses
  - this will be our normal regression model
- The full model will be as described (one independent mean for each level of X)

$$H_0 : E(Y) = \beta_0 + \beta_1 X$$
$$H_a : E(Y) \neq \beta_0 + \beta_1 X$$

# SSE For Reduced Model

The SSE for the reduced model is as before
- remember

$$
\begin{aligned}
SSE(R) &= \sum \sum [Y_{ij} - (b_0 + b_1 X_j)]^2 \\
&= \sum \sum (Y_{ij} - \hat{Y}_{ij})^2
\end{aligned}
$$

- and has n-2 degrees of freedom $df_R = n - 2$

# SSE(R)

Figure:

(a) Data

| Branch $i$ | Size of Minimum Deposit (dollars) $X_i$ | Number of New Accounts $Y_i$ | Branch $i$ | Size of Minimum Deposit (dollars) $X_i$ | Number of New Accounts $Y_i$ |
|---|---|---|---|---|---|
| 1 | 125 | 160 | 7 | 75 | 42 |
| 2 | 100 | 112 | 8 | 175 | 124 |
| 3 | 200 | 124 | 9 | 125 | 150 |
| 4 | 75 | 28 | 10 | 200 | 104 |
| 5 | 150 | 152 | 11 | 100 | 136 |
| 6 | 175 | 156 | | | |

(b) ANOVA Table

| Source of Variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | 5,141.3 | 1 | 5,141.3 |
| Error | 14,741.6 | 9 | 1,638.0 |
| Total | 19,882.9 | 10 | |

# F Test Statistic

From the general linear test approach

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$
$$F^* = \frac{SSE - SSPE}{(n-2)-(n-c)} \div \frac{SSPE}{n-c}$$

where a little algebra takes us to the next slide

# F Test Rule

- From the F test we know that large values of $F^*$ lead us to reject the null hypothesis:
  If $F^* \leq F(1-\alpha; c-2, n-c)$, conclude $H_0$
  If $F^* > F(1-\alpha; c-2, n-c)$, conclude $H_a$
- For this example we have

$$SSPE = 1{,}148.0 \qquad n - c = 11 - 6 = 5$$

$$SSE = 14{,}741.6$$

$$SSLF = 14{,}741.6 - 1{,}148.0 = 13{,}593.6 \qquad c - 2 = 6 - 2 = 4$$

$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5}$$

$$= \frac{3{,}398.4}{229.6} = 14.80$$

# Example Conclusion

- If we set the significance level to $\alpha = .01$
- And look up the value of the F inv-cdf $F(.99, 4, 5) = 11.4$
- We can conclude that the null hypothesis should be rejected.