Chapter 13.1: Assessing Model Adequacy

Instructor: Dr. Arnab Maity
So far we have learned

- What is simple linear regression model \( y = \beta_0 + \beta_1 x + \epsilon \)
- How to interpret model parameters
- How to estimate model parameters (least squares)
- Inference for slope \( \beta_1 \) (t-test and CI, ANOVA)
- Prediction of a new \( y \) observation based on a new \( x \) value

There are two crucial assumptions we make that allows us to do inference on \( \beta_1 \) and the prediction. They are

- The response \( y \) has a linear relationship with \( x \)
- The errors \( \epsilon \) have a normal distribution
- The errors have constant variance \( \sigma^2 \) (that does not depend on the value of \( x \)’s)

In this chapter, we will learn how to check these conditions to ensure that the model fits the data well.

---

**Fitting and diagnostic procedure**

**Observed data:** \((y_1, x_1), \ldots, (y_n, x_n)\)

1. **Make a scatterplot of the data:** plot of the \((x_i, y_i)\) pairs to see whether the relationship seems linear or not. Calculating correlation between \( x \) and \( y \) might help here since large (positive or negative) correlation also means strong linear relationship. If the relationship looks linear, go to next step. Otherwise, look for other regression methods, or consider transforming the variables suitably (we will learn this later).

2. **Fit a linear regression model** to the data \( y = \beta_0 + \beta_1 x + \epsilon \) where the intercept and slope are estimated using the least squares.

3. **Calculate the coefficient of determination,** \( R^2 \). This provides the proportion of variability in the response that is explained by the model. We would like this to be high.

4. **Use inferential methods** (confidence interval or hypothesis testing) to make formal claims about \( \beta_1 \), the slope parameter. Here you can formally check the extent of association between \( x \) and \( y \) using the model utility test.

5. **Verify the model assumptions:** The simple linear regression model makes two important assumptions: (1) The errors \( \epsilon \) have a normal distribution and (2) The errors have constant variance \( \sigma^2 \) (that does not depend on the value of \( x \)’s).
The last step (**Verify the model assumptions:**) in the previous page is of great importance. Without these assumptions, all our inference methods (interval and testing) may be completely misleading. We can check these assumptions using residuals obtained from the model fit. Recall, residuals are
\[
\hat{e}_i = y_i - (\hat{\beta}_0 + x_i\hat{\beta}_1) = y_i - \hat{y}_i,
\]
where \(\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1\) is the fitted values of \(y_i\). We will use these residuals in further steps. Alternatively, we can also use the **standardized residuals**
\[
\hat{e}_i^* = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{XX}}}}.
\]
The denominator is simply an estimator of \(sd(\hat{e}_i)\).

**Checking for constant variance:** There are some basic plots (called bf residual plots) that we use to check whether the error variance \(\sigma^2\) is constant or not.

- Plot the residuals (or standardized residuals) on the vertical on the vertical axis and the \(x\)'s on the horizontal axis.

- Plot the residuals (or standardized residuals) on the vertical on the vertical axis and the \(\hat{y}\)'s on the horizontal axis.

What do we expect to see from these plots?
Example 1: Corrosion of steel reinforcing bars is the most important durability problem for reinforced concrete structures. Carbonation of concrete results from a chemical reaction that lowers the pH value by enough to initiate corrosion of the rebar. Representative data on $x =$ carbonation depth (mm) and $y =$ strength (M Pa) for a sample of core specimens taken from a particular building follows (read from a plot in the article “The Carbonation of Concrete Structures in the Tropical Environment of Singapore,” Magazine of Concrete Res., 1996: 293-300). Data are provided in Example 12.13.

Simple linear regression results:
Dependent Variable: strength
Independent Variable: carbonation_depth
strength = 27.182936 - 0.29756123 carbonation_depth
Sample size: 18
R (correlation coefficient) = -0.87497382
R-sq = 0.76557918
Estimate of error standard deviation: 2.864026

Parameter estimates:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>DF</th>
<th>95% L. Limit</th>
<th>95% U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>27.183</td>
<td>1.651</td>
<td>16</td>
<td>23.682</td>
<td>30.684</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.298</td>
<td>0.041</td>
<td>16</td>
<td>-0.385</td>
<td>-0.210</td>
</tr>
</tbody>
</table>

Residuals vs. depth  
Standardized residuals vs depth  
Standardized residuals vs predicted depth

Based on these plots, do you think the assumption of constant variance is reasonable?
Example 2: The following results and plots are from another data set where the true relation is \( y_i = 1 + 2x_i + e_i \), and \( \text{var}(e_i) = x_i^2 \).

![Scatter plot of \( y \) vs \( x \)](attachment)

**Simple linear regression results:**
- **Dependent Variable:** \( y \)
- **Independent Variable:** \( x \)
- \( y = 1.188185 + 1.845873 \times x \)
- **Sample size:** 50
- **R (correlation coefficient):** 0.81491048
- **R-sq:** 0.66407909
- **Estimate of error standard deviation:** 1.391739

Based on these plots, do you think the assumption of constant variance is reasonable?
Example 3: The following results and plots are from another data set where the true relation is $y_i = 1 + e_i$, and $\text{var}(e_i) = x_i^2$. (Note: The true value of $\beta_1 = 0$ in this example)

Simple linear regression results:
Dependent Variable: yy
Independent Variable: x
$yy = 1.2636969 - 0.33028641 \times$  
Sample size: 50
$R$ (correlation coefficient) = $-0.2985273$
$R$-sq = $0.08911855$
Estimate of error standard deviation: 1.119399

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>Alternative</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.2636969</td>
<td>0.2921148</td>
<td>$\neq 0$</td>
<td>48</td>
<td>4.326028</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.33028641</td>
<td>0.15241123</td>
<td>$\neq 0$</td>
<td>48</td>
<td>-2.167074</td>
<td>0.0352</td>
</tr>
</tbody>
</table>

Residuals vs. $y$  Standardized residuals vs $y$  Standardized residuals vs predicted $y$

Based on these plots, do you think the assumption of constant variance is reasonable?

What effect this might have on our inference?
**Verifying normality assumption:** We use the quantile quantile plot, (in short Q-Q plot) of the residual for this purpose. Here are the Q-Q plots of the standardized residuals for all the three examples we saw before.

![Example 1](image1.png) ![Example 2](image2.png) ![Example 3](image3.png)

Based on these plots, do you think the assumption of normality is reasonable?

What effect this might have on our inference?
**Difficulties and Remedies**

Although we hope that our analysis will yield plots like those of Figure 13.1, quite frequently the plots will suggest one or more of the following difficulties:

1. A nonlinear probabilistic relationship between $x$ and $y$ is appropriate.
2. The variance of $\epsilon$ (and of $Y$) is not a constant $\sigma^2$, but instead depends somehow on $x$.
3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.
4. The error variable $\epsilon$ does not have a normal distribution.
5. When the subscript $i$ indicates the time order of the observations, the $\epsilon_i$'s exhibit dependence over time.
6. One or more relevant independent variables have been omitted from the model.

Figure 13.2 presents residual plots corresponding to items 1–3, 5, and 6. In Chapter 4, we discussed patterns in normal probability plots that cast doubt on the assumption of an underlying normal distribution. Notice that the residuals from the data in Figure 13.2(d) with the circled point included would not by themselves necessarily suggest further analysis, yet when a new line is fit with that point deleted, the new line differs considerably from the original line. This type of behavior is more difficult to identify in multiple regression. It is most likely to arise when there is a single (or very few) data point(s) with independent variable value(s) far removed from the remainder of the data.

![](image)

**Figure 13.2** Plots that indicate abnormality in data: (a) nonlinear relationship; (b) nonconstant variance; (c) discrepant observation; (d) observation with large influence; (e) dependence in errors; (f) variable omitted