Chapter 3. Diagnostics and Remedial Measures

So far, we took data \((X_i, Y_i)\) and we assumed

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, 2, \ldots, n, \]

where

- \(\epsilon_i \sim iid \sim N(0, \sigma^2)\),
- \(\beta_0, \beta_1\) and \(\sigma^2\) are unknown parameters,
- \(X_i\)'s are fixed constants.

**Question:**

What are the possible mistakes or violations of these assumptions?
1. Regression function is not linear \( E(Y) \neq \beta_0 + \beta_1 X \)

2. Error terms do not have a constant variance

3. Error terms are not independent

We will use **Residual Plots** to diagnose the problems

**Residuals:** \( e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) \)

**Sample Mean:** \( \bar{e} = \frac{1}{n} \sum_i e_i = 0 \)

**Sample Var:**

\[
\frac{1}{n-1} \sum_i (e_i - \bar{e})^2 = \frac{1}{n-1} \sum_i e_i^2 \approx MSE
\]

We will sometimes use standardized (semistudentized) residuals
Nonlinearity of Regression Function (1.)

Residual plot against the predictor variable, $X$. Or use a residual plot against the fitted values, $\hat{Y}$.

Look for systematic tendencies!

Example:
\[ y = 1 + 2x + \varepsilon \quad \varepsilon \sim \mathcal{N}(0,3) \]

\[ y = 1 + 2x + 3 \sin(x/2) + \varepsilon \quad \varepsilon \sim \mathcal{N}(0,3) \]

\[ y = 1 + 0.1x + \varepsilon \quad \varepsilon \sim \chi_1 \]
Nonconstancy of Error Variance (2.)

We diagnose nonconstant error variance by observing a residual plot against $X$ and looking for structure.

Example:
**Modified Levene Test**

1. Divide residuals into two groups. For this example, low and high salary groups, because the variance is suspected to depend on salary.

2. Calculate $d_{i1} = |e_{i1} - \tilde{e}_1|$ and $d_{i2} = |e_{i2} - \tilde{e}_2|$, where $e_{ij}$ is the $i^{th}$ residual in group $j$ and $\tilde{e}_j$ is the median of residuals in group $j$.

3. Conduct two-sample t-test with $d_{ij}$.
Nonindependence of Error Terms (3.)

We diagnose nonindependence of errors **over time** or **in some sequence** by observing a residual plot against time (or the sequence) and looking for a trend (see textbook, p. 101, for typical plots).

**Example:**

![Diagram of residual plots](image-url)
But, if the data is like
day 1: \((X_1, Y_1)\)
day 2: \((X_2, Y_2)\)
::
day \(n\): \((X_n, Y_n)\)

then we can see the **effect of learning**.
Model fits all but a few observations (4.)

**Example:** LS Estimates with 2 outlying points (solid) and without them (dashed).

**Rule of Thumb:**

Outliers are detected by observing a plot of $e_i^*$ vs. $X_i$. 
Errors not normally dist’d (5.)

We assumed $\epsilon_1, \ldots, \epsilon_n$ iid $N(0, \sigma^2)$ but we can’t observe these error terms!

We will be convinced that this assumption is reasonable, if $e_1, \ldots, e_n$ appear to be iid $N(0, MSE)$.

Fact: If $e_1, \ldots, e_n$ iid $N(0, MSE)$, then one can show that the expected value of the $i$th smallest is

$$\sqrt{MSE} \left[ z \left( \frac{i - 3/8}{n + 1/4} \right) \right], \quad i = 1, 2, \ldots, n$$

Then we have pairs

<table>
<thead>
<tr>
<th>residual</th>
<th>expected residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{\text{min}}$</td>
<td>$\sqrt{MSE} \left[ z \left( \frac{1-0.375}{n+0.25} \right) \right]$</td>
</tr>
<tr>
<td>$e_{\text{2nd smallest}}$</td>
<td>$\sqrt{MSE} \left[ z \left( \frac{2-0.375}{n+0.25} \right) \right]$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$e_{\text{max}}$</td>
<td>$\sqrt{MSE} \left[ z \left( \frac{n-0.375}{n+0.25} \right) \right]$</td>
</tr>
</tbody>
</table>
Notice: If $Y_1, \ldots, Y_4$ iid $N(0, \sigma^2)$, then

$$E(Y_1) = \cdots = E(Y_4) = 0,$$

and

$$E(\bar{Y}) = 0,$$

but

$$E(Y_{\text{min}}) = \sigma \left[ z \left( \frac{1 - 0.375}{4 + 0.25} \right) \right] = \sigma z(0.147) = -1.05\sigma,$$

$$E(Y_{2\text{nd}}) = \sigma \left[ z \left( \frac{2 - 0.375}{4 + 0.25} \right) \right] = \sigma z(0.382) = -0.30\sigma,$$

$$E(Y_{3\text{rd}}) = \sigma \left[ z \left( \frac{3 - 0.375}{4 + 0.25} \right) \right] = \sigma z(0.618) = +0.30\sigma,$$

$$E(Y_{\text{max}}) = \sigma \left[ z \left( \frac{4 - 0.375}{4 + 0.25} \right) \right] = \sigma z(0.853) = +1.05\sigma,$$
Points on a straight line: Errors are normal (left)
Points on a curve: Errors are not normal (right)
Omission of important predictors (6.)

Example: $X_i = \#\text{years of education}, Y_i = \text{salary}$

Means, that a better model would be (Multiple Regression Model)
Lack of Fit Test

Suppose we want to test whether the relationship between \(X\) and \(Y\) is linear vs the possibility that it is NOT linear.

Test for: \[ H_0 : E(Y) = \beta_0 + \beta_1 X \]

versus \(H_a : \) Not \(H_0\)

Here, \(H_0\) includes the cases when either or both \(\beta_0\) and \(\beta_1\) are zero.

We can’t use this test unless there are multiple \(Y\)’s observed at at least 1 value of \(X\).
Can we use this test when $X=$day and $Y=$stock price?

Can we use this test when $X=$weight and $Y=$height and those are measured with a super accurate measure?
**New Notation:** \( Y \) values are observed at \( c \) different levels of \( X \), say \( X_1, X_2, \ldots, X_c \).

\( n_j \) such \( Y \) values, say \( Y_{1j}, Y_{2j}, \ldots, Y_{nj} \), are observed at level \( X_j, j = 1, 2, \ldots, c, n_j \geq 1 \).

Let \( \bar{Y}_j = \frac{1}{n_j} \sum_i Y_{ij} \) be the average of the \( Y \)'s at \( X_j \) and \( \hat{Y}_j = b_0 + b_1 X_j \) the fitted mean under the SLR.

The data now look like

at \( X_1 \): (\( X_1, Y_{11} \)), (\( X_1, Y_{21} \)), \ldots, (\( X_1, Y_{n_1} \)) \( \Rightarrow \bar{Y}_1 \)

at \( X_2 \): (\( X_2, Y_{12} \)), (\( X_2, Y_{22} \)), \ldots, (\( X_2, Y_{n_2} \)) \( \Rightarrow \bar{Y}_2 \)

\vdots

at \( X_c \): (\( X_c, Y_{1c} \)), (\( X_c, Y_{2c} \)), \ldots, (\( X_c, Y_{nc} \)) \( \Rightarrow \bar{Y}_c \)
The less restricting model puts **no structure** on the means at each level of $X$ (Full model).

**Full model:** $Y_{ij} = \mu_j + \epsilon_{ij}$, where $\hat{\mu}_j = \bar{Y}_j$

**Reduced model:** $Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$

**F-test !!!!**
Note that

\[ Y_{ij} - \hat{Y}_j = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_j) \]

Let’s partition the \( SSE \) into 2 pieces

\[ SSE = SSPE + SSLF \]

where

\[
\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{c} \sum_{i=1}^{n_j} (\bar{Y}_j - \hat{Y}_j)^2
\]

- If \( SSPE \approx SSE \), it says that the means (\( \triangle \)) are close to the fitted values (\( \Box \)). That is, even if we fit a less restrictive model, we can’t reduce the amount of unexplained variability.

- If \( SSLF \approx SSE \), the means (\( \triangle \)) are far away from the fitted values (\( \Box \)) and the (linear) restriction seems unreasonable.

Thus,
Formal Test for: $H_0 : E(Y) = \beta_0 + \beta_1 X$  
$H_A : E(Y) \neq \beta_0 + \beta_1 X$

Let

$$M_{SLF} = \frac{SSLF}{c-2} \quad \text{and} \quad M_{SPE} = \frac{SSPE}{n-c}$$

$$F = \frac{\frac{SSE(R)-SSE(F)}{df(R)-df(F)}}{\frac{SSE(F)}{df(F)}}$$

$$= \frac{SSPE}{\frac{SSE-SSPE}{n-2-(n-c)}}$$

$$= \frac{\frac{SSLF}{c-2}}{\frac{SSPE}{n-c}} \sim F_{c-2,n-c}$$

$F \uparrow \iff SSE \gg SSPE \iff$ The model is bad.

Test Statistic:

Rejection Rule:
This fits nicely into our ANOVA Table:

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SSR$</td>
<td>1</td>
<td>$MSR$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE$</td>
<td>$n - 2$</td>
<td>$MSE$</td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>$SSLF$</td>
<td>$c - 2$</td>
<td>$MSLF$</td>
</tr>
<tr>
<td>Pure Error</td>
<td>$SSPE$</td>
<td>$n - c$</td>
<td>$MSPE$</td>
</tr>
<tr>
<td>Total</td>
<td>$SSTO$</td>
<td>$n - 1$</td>
<td></td>
</tr>
</tbody>
</table>

**Example:** Suppose that the house prices follow a SLR in $#$bedrooms. The estimated regression function is

$$\hat{E}(\text{price}/1,000) = -37.2 + 43.0(\text{#bedrooms})$$
Because $F^* = MSLF/MSPE = 1.432 / 1.281 = 1.12 < F(0.95; 3, 88) = 2.71$ we do not reject $H_0$. 
Motivation: Consider the function \( y = x^2 \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

If you have \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) and you know \( y = f(x) \), then \((f(x_1), y_1), (f(x_2), y_2), \ldots, (f(x_n), y_n)\) will be on a **straight line**.

What follows are two situations in which transformations may help:
**Situation 1:** nonlinear regression function with constant error variances (1.)

Note that $E(Y)$ doesn’t appear to be a linear function of $X$, that is, the points do not seem to lie on a line. The spread of the $Y$’s at each level of $X$ appears to be constant, however.

Typical remedy –
Transform $X$

We consider
Do not transform $Y$ because this will disturb the spread of the $Y$’s at each level $X$. 
**Situation 2:** nonlinear regression function with non-constant error variances (1. with 2.)

Note that $E(Y)$ isn’t a linear function of $X$.
The variance of the $Y$’s at each level of $X$ is increasing with $X$.

Typical remedy –
Transform $Y$
(or maybe $X$ and $Y$)
We consider
And hope that both problems are fixed.
Prototypes for Transforming $Y$

Try $\sqrt{Y}$, $\log_{10} Y$, or $1/Y$

Prototypes for Transforming $X$

Use $\sqrt{X}$ or $\log_{10} X$ (left); $X^2$ or $\exp(X)$ (middle); $1/X$ or $\exp(-X)$ (right).
Model: \( Y = \beta_0 + \beta_1 h(X) + \epsilon \) where \( X' = h(X) \)
Model: \( g(Y) = \beta_0 + \beta_1 h(X) + \epsilon \) where \( Y' = g(Y) \) and \( X' = h(X) \)