LECTURE 6

BOGDAN ICHIM

Residual Plots

The residual $\widehat{\epsilon}_i$ is a measure of how closely a model agrees with the observation y_i. One can simply use the residuals $\widehat{\epsilon}_i = y_i - \widehat{y}_i$, or the **standardized residuals**, which have approximately a variance of 1. One should check for:

- isolated lack-of-fit (a few unusual observations, which may be potential outliers);
- systematic lack-of-fit (the general behavior of the data differs from what the model predicts, which means that the model is bad and it should be adjusted).

The following plots may be useful to check the model assumptions.

Zero Mean and Constant Variance

Plot the standardized residuals against the fitted values $\widehat{y}_i$.

Remark 1. The points should be evenly scattered around zero, with no systematic pattern.

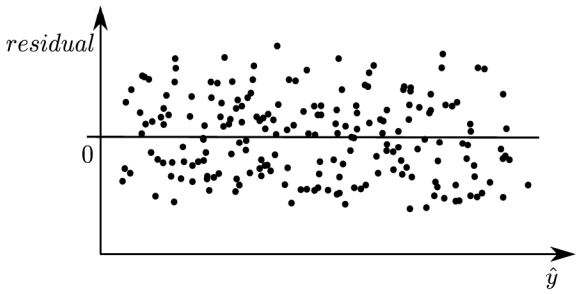

FIGURE 1. Random scatter (satisfactory)

In the case of triangle or diamond shapes one can try to transform y in order to obtain approximate constancy of the error variance.

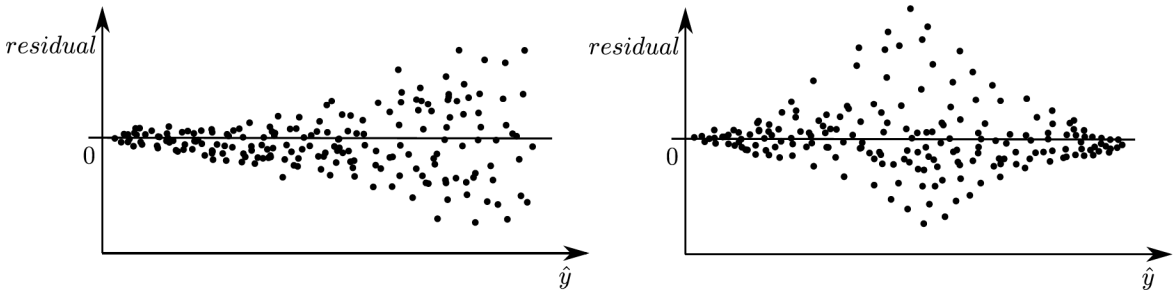

FIGURE 2. Left figure: increasing spread. Right figure: non-constant spread

Linearity

The standardized residuals can be plotted against the individual explanatory variables (predictors). All such plots should indicate random scatter of equal width around zero.

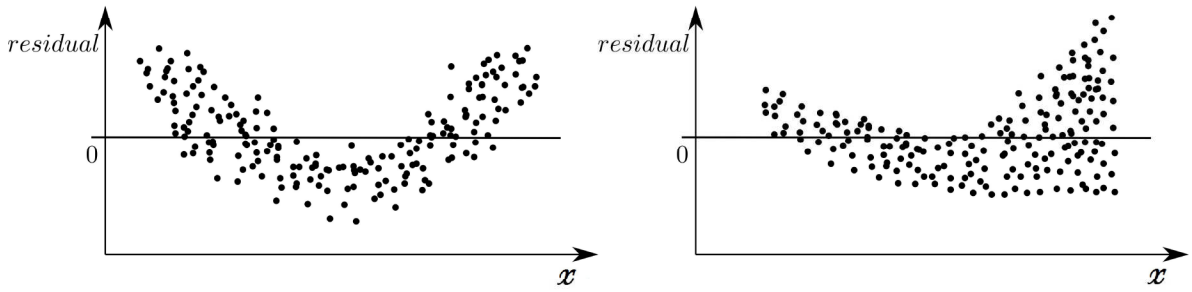

FIGURE 3. Left figure: quadratic, constant spread. Right figure: non-linear, increased spread.

Some particular situations to consider:

(1) For predictors that are already in the model non-linearity suggests that higher order terms involving those predictors should be added to the model. Transform X and/or Y in order to achieve approximate linearity.

(2) Two separate straight lines: fit two separate models. For example one for males and one for females.

(3) For predictors which are not included in the model, they should be perceived as noise. Any systematic residual patterns suggests that those predictors should be added to the model.

Normality

One can look at the histogram of the standardized residuals. Alternatively, one can check the normal **QQ plot**, in which departures from normality are indicated by deviations from a straight line.

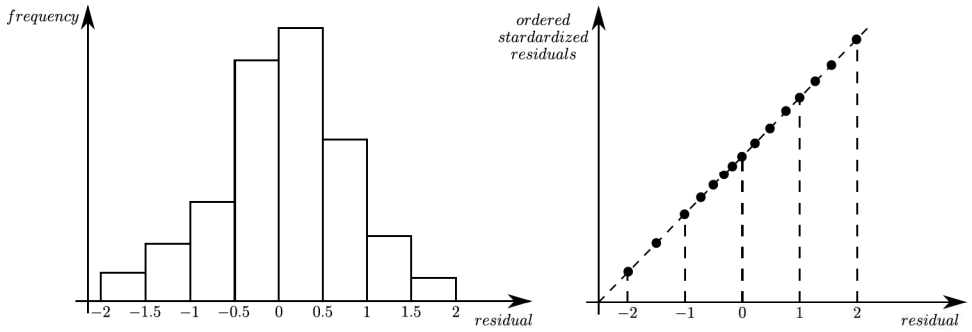

FIGURE 4. Left figure: histogram. Right figure: QQ plot.

Remark 2. There may be good reasons for such departures! If those happen, one can try to use models with different probability distributions for Y. For example, if Y represents counts, one can try to use a Poisson distribution. Note that in R the GLM library implements such models.

Independence

One can plot the residuals against the **serial order** in which the observations were taken. If the model assumptions are correct, the plot should come in the form of a random scatter plot with no visible pattern or trend.

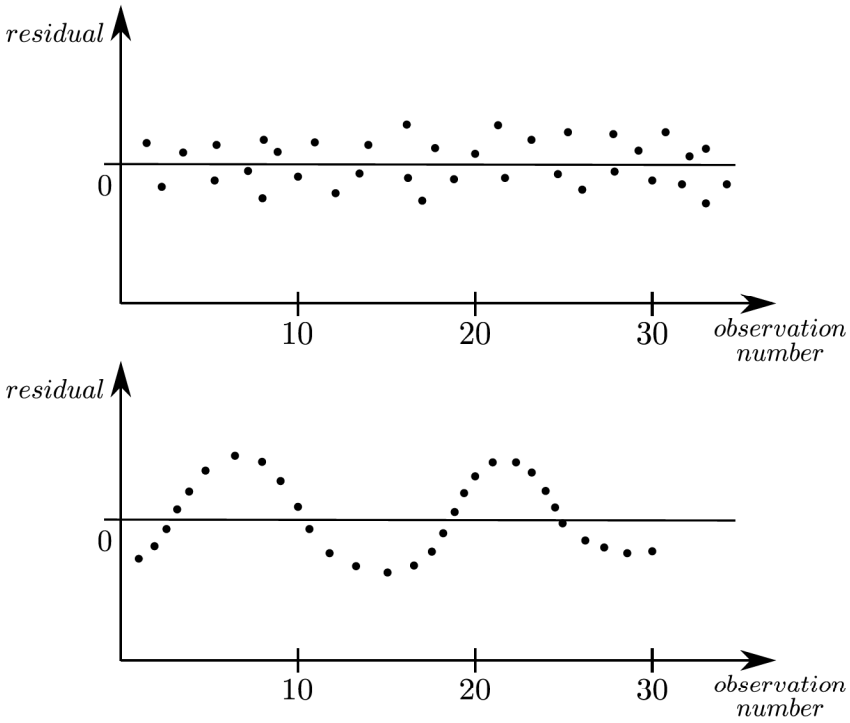

FIGURE 5. Top figure: uncorrelated residuals (satisfactory).
Bottom figure: tracking in residuals

Remark 3. Correlations appear frequently in the context of **time-series** data, which consist of observations for which measurements are obtained at discrete points in time. In many cases observations that are obtained at the adjacent points in time will have **positively correlated errors**. We may see **tracking** in residuals - that is, adjacent residuals may have similar values. In the context of time-series data, many specialized methods have been developed to properly take account of correlations of the residuals. See for example ARMA, ARIMA, ARCH, GARCH. Their presentation is beyond the scope of this course.

Outliers Plot

There is an ongoing debate on outliers. What is or is not an outlier depends heavily on the nature of the data and on the task at hand.

Possible outliers may be indicated by points with large standardized residuals. Under normality assumptions, approximately 95% of the observations should have standardized residuals in the range $(-2.0, 2.0)$.

Rule of Thumb: An observation with an absolute value of the standard residual ≥ 2.5 may be a potential outlier and the accuracy of such an observation should be investigated.

The Logistic Model

(The Logistic Regression Model)

The linear model assumes that the response variable Y is **quantitative**. However, in many situations, the response variable is **qualitative** (categorical). For example, eye colour: blue, green, brown.

Predicting qualitative responses is a process that is known as **classification**.

Assume that we have a binary (two levels) qualitative response. One approach is to encode the response using a 0/1 **dummy variable** and predict 0 if $\widehat{y} \leq 0.5$ and 1 otherwise.

Remark 4. The classification that we get if we use a linear model to predict a binary response encoded as above is the same as for **Linear Discriminant Analysis** (LDA). The estimates that we get might be outside the $[0, 1]$ interval, making them hard to interpret as probabilities.

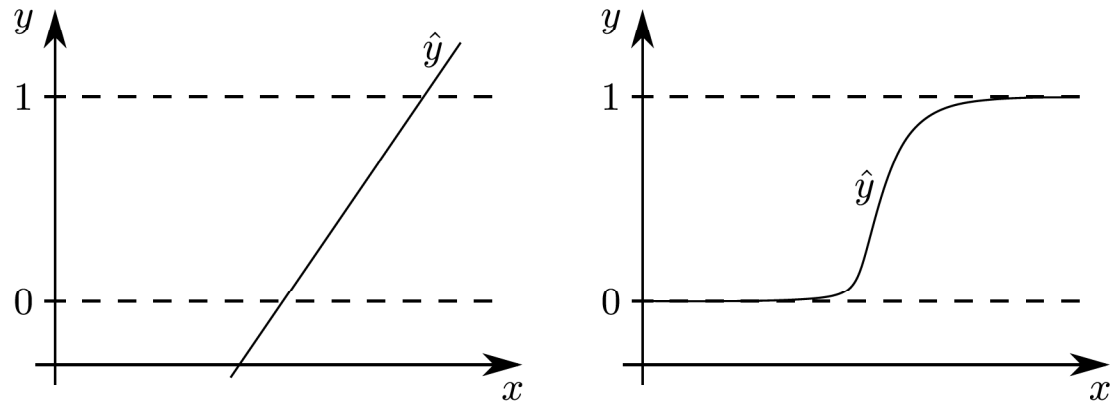

FIGURE 6. Left figure: estimated Y using the linear model. Right figure: predicted probabilities for $Y = 1$ using the logistic model

The **logistic model** uses the **logistic function**

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

in order to model the relationship between the probability of $Y = 1$ and X.

We compute

$$\iff P(Y = 1|X)(1 + e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$

$$\iff P(Y = 1|X) = e^{\beta_0+\beta_1 X}(1 - P(Y = 1|X))$$

$$\iff \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = e^{\beta_0+\beta_1 X} \quad \Big| \ln()$$

$$\iff \ln\Big(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\Big) = \beta_0 + \beta_1.X$$

The fraction

$$\frac{P(Y=1|X)}{1-P(Y=1|X)}$$

is called the **odds** and can take any value in $[0,\infty)$. Odds are traditionally used in horse racing instead of probabilities, since they relate more naturally to the correct betting strategy.

Similarly,

$$\ln \frac{P(Y=1|X)}{1-P(Y=1|X)}$$

is called the **log-odds** or **logit**.

Remark 5. The logistic model has a logit that is linear in X!

Multiple Logistic Model

Let $X=(X_1,...,X_p)$. This model uses the **multivariate logistic function**

$$P(Y=1|X)=\frac{e^{\beta_0+\sum_{i=1}^{p}\beta_i X_i}}{1+e^{\beta_0+\sum_{i=1}^{p}\beta_i X_i}}.$$

The model is linearized by similar computations. It follows that

$$\ln \frac{P(Y=1|X)}{1-P(Y=1|X)}=\beta_0+\sum_{i=1}^{p}\beta_i X_i.$$