Chapter 5, Model Building and Residual Analysis

1 Comparing Regression Models

There are several ways to compare regression models. In our class we will use Adjusted r^2

2 Residual Analysis

Recall that the model is

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

and we assumed that

- ϵ_i is normally distributed with mean zero and variance σ^2
- ϵ_is are independent

If these assumptions do not hold, they invalidate our analysis Diagnostics:

- Examine appropriateness of model & and detect violations of model assumptions
- Typical violations
  - Regression function is not linear
  - Error terms do not have constant variance
  - Error terms are not independent
  - One or more observations are outliers
  - Error terms are not normally distributed
  - One or more important predictors have been omitted from the model

We use the residuals to examine important departures from the simple linear regression model

$y = \beta_0 + \beta_1 x + \epsilon$ with independent and identically normally distributed errors

- The ith residual $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$.

- The residuals are used to estimate the errors
- $\sum_{i=1}^{n} e_i = 0.$
- $var(e_i) = \sigma^2(1 - h_{ii})$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The residuals are not independent

We use plots of the residuals to answer these questions. The plots that are commonly used are

1. plot the residuals against the predictor variable
2. plot the residuals against the fitted values
3. plot the residuals against the time (important if data collected over time)
4. plot the residuals against omitted variables
5. box plot for the residuals
6. normal plot for the residuals

We should check for:

1. Nonlinearity of the the regression function: this can be studied by a plot of the residuals against the predictor variable or equivalently by a plot of the residuals against the fitted values. The plot should not show any particular pattern.

2. Nonconstancy of the error variance: plot residuals versus the predictor variable or the fitted values

3. Presence of outliers: outliers are extreme observations. An outlier may dramatically change the regression line (when this is the case the outliers in an influential case). Outlier residuals can be identified from residual plots of residuals versus x or $\hat{y}$ as well as box plots.

Figure 1: Linear in top not linear in the bottom

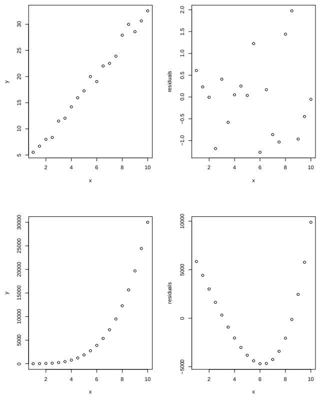
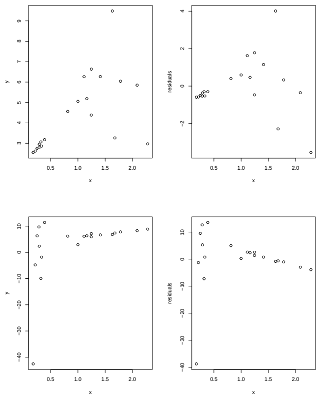


Figure 2: Noconstant Variance



4. Nonindependent errors: plot residuals versus time to see if there is any cyclical pattern. The residuals are always dependent but this dependency decreases with the sample size.

5. Nonnormality of error terms: make a box plot of the residuals

- An outlier is a data point whose response y does not follow the general trend of the rest of the data.

- A data point has high leverage if it has ”extreme” predictor x values.

Figure 3: Presence of outliers

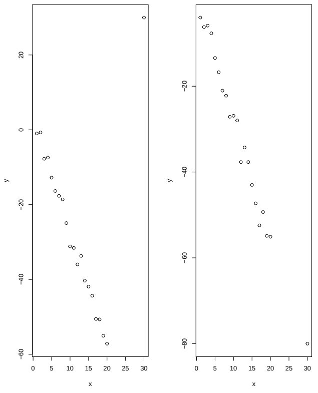
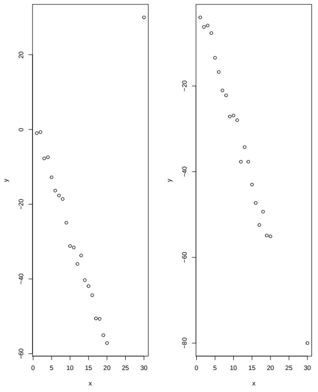


Figure 4: Presence of outliers



With a single predictor, an extreme x value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor).

- A data point is influential if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

- Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential

- One advantage of the case in which we have only one predictor is that we can look at simple scatter plots in order to identify any outliers and influential data points.

- It turns out that

$$\begin{aligned}
\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \ldots + h_{in}y_n \\
\hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \ldots + h_{2n}y_n \\
\vdots \quad &\vdots \quad \vdots \\
\hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \ldots + h_{nn}y_n
\end{aligned}$$

  for some numbers $h_{ij}, i, j = 1, 2, \ldots, n$

- h_{ii} i measures the distance of the x values of the ith case from the center of the experimental region (ignores the response)

- The leverage h_{ii}, quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$. That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$.

- On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$. It's for this reason that the h_{ii} are called the leverages.

- In simple linear regression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Some important properties of the leverages:

1. The leverage h_{ii} is a measure of the distance between the x value for the ith data point and the mean of the x values for all n data points.

2. The leverage h_{ii} is a number between 0 and 1, inclusive.

3. The sum of the h_{ii} equals k+1, the number of parameters (regression coefficients including the intercept).

The first bullet indicates that the leverage h_{ii} quantifies how far away the ith x value is from the rest of the x values. If the ith x value is far away, the leverage h_{ii} will be large; and otherwise not.

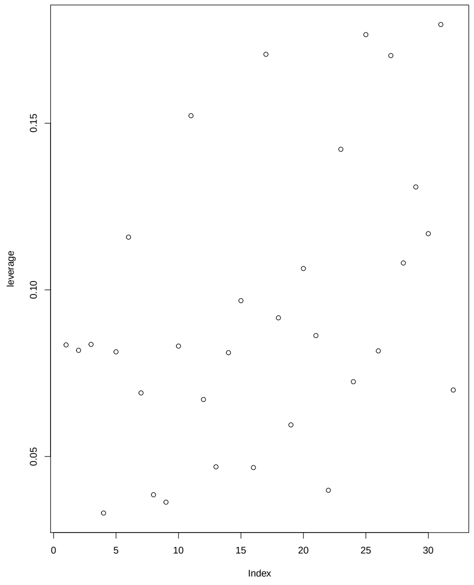
Identifying data points whose x values are extreme

- The great thing about leverages is that they can help us identify x values that are extreme and therefore potentially influential on our regression analysis.

- How? All we need to do is determine when a leverage value should be considered large. A common rule is to flag any observation whose leverage value, h_{ii}, satisfies

$$h_{ii} > \frac{2(k+1)}{n}$$

Example (Clocks continued)



```
fit<-lm(Price~Age+Bidders)
>leverage<-hat(model.matrix(fit))
> leverage<-hat(model.matrix(fit))
> leverage
 [1] 0.08348378 0.08185951 0.08363981 0.03302920
 [5] 0.08140369 0.11582144 0.06907473 0.03851475
 [9] 0.03629957 0.08313134 0.15225853 0.06710040
[13] 0.04690377 0.08115071 0.09675641 0.04666806
[17] 0.17068377 0.09161438 0.05948491 0.10640592
[21] 0.08627957 0.03986110 0.14218854 0.07243952
[25] 0.17658508 0.08168793 0.17029360 0.10806706
[29] 0.13087893 0.11686406 0.17962005 0.06994990

> data[leverage> 2*3/32,]
[1] Age     Bidders Price
<0 rows> (or 0-length row.names)
> plot(leverage)
```

Identifying Outliers (Unusual y Values)

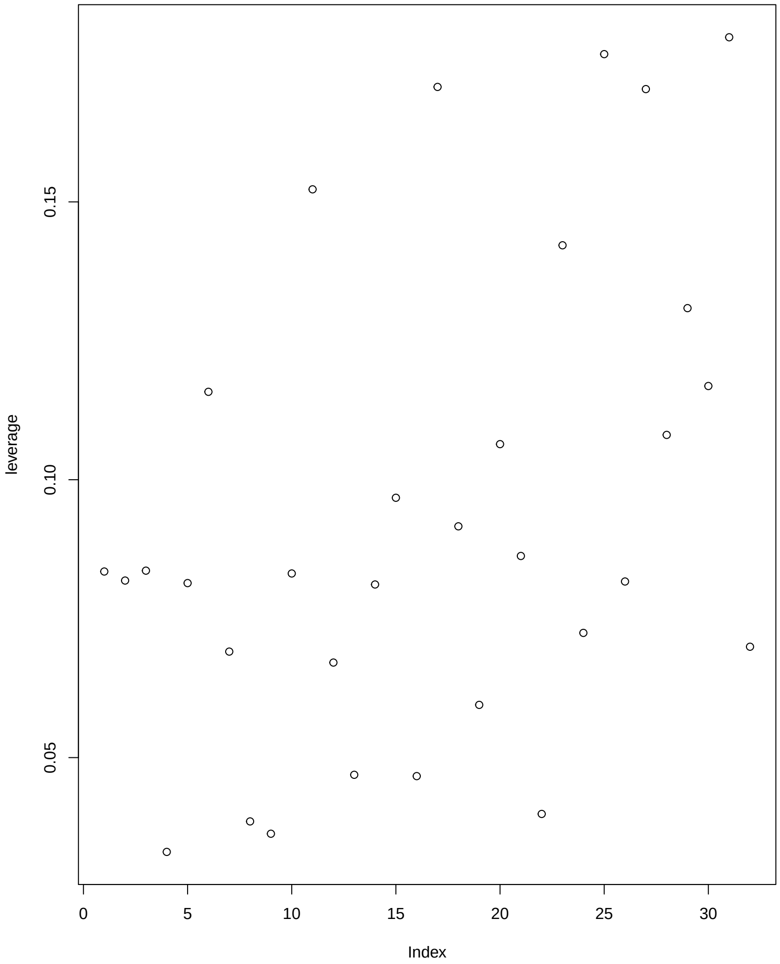
- Residuals: The ith residual is defined as $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$.
- Studentized residuals (or internally studentized residuals) are defined for each observation, $i = 1, ..., n$ as an ordinary residual divided by an estimate of its standard deviation:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

- An observation with an internally studentized residual that is larger than 3 (in absolute value) is generally deemed an outlier. [Sometimes, the term "outlier" is reserved for an observation with an externally studentized residual that is larger than 2 in absolute value, we consider externally studentized residuals in the next section.]

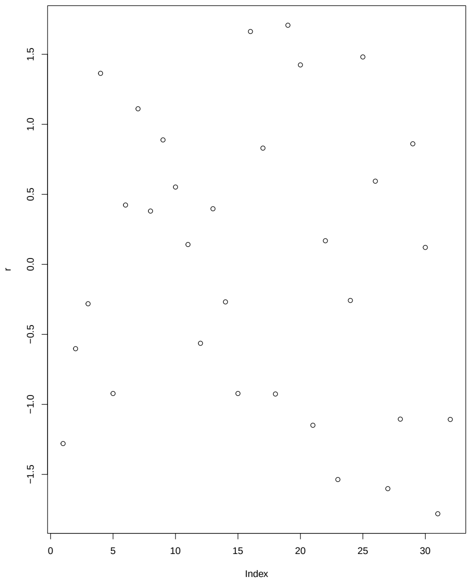
```
> r = rstudent(fit)
> data[abs(r)>2,]
[1] Age     Bidders Price
<0 rows> (or 0-length row.names)
> plot(r)
```

Example (Clocks continued)



Influential points

- An influential point is one if removed from the data would significantly change the fit.
- An influential point may either be an outlier or have large leverage, or both, but it will tend to have at least one of those properties.
- Cook's distance is a commonly used influence measure that combines



these two properties. It can be expressed as

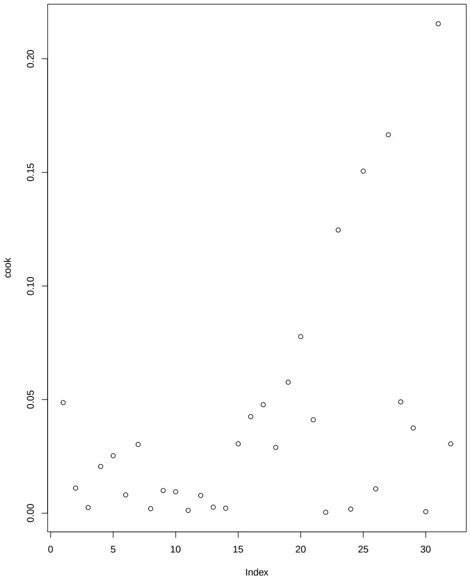
- Typically, points with cook's distance greater than 1 are classified as being influential.
- We can compute the Cook's distance using the following commands cook = cooks.distance(fit)

Example (Clocks continued) To determine whether the ith case in an influential case or no, we its cooks distance D_i and the rule is

- If $D_i < F_{[0.80]}^{(k+1,n-k-1)}$, the ith case is not influential
- If $D_i > F_{[0.50]}^{(k+1,n-k-1)}$, the ith case is influential
- if $F_{[0.80]}^{(k+1,n-k-1)} < D_i < F_{[0.50]}^{(k+1,n-k-1)}$, inconclusive

Normal Assumption

- A Normal probability plot of the residuals can be used to check the normality assumption.



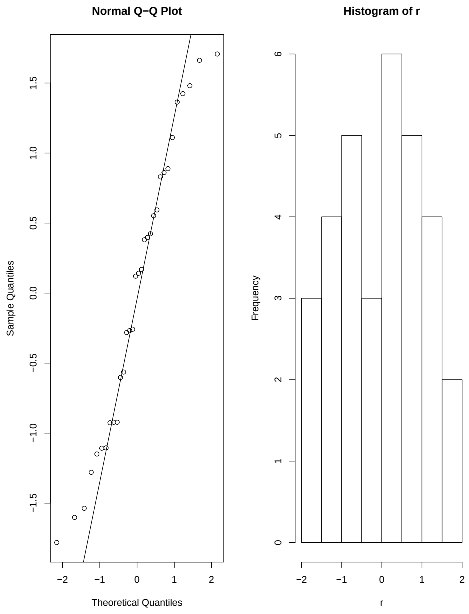
- Here each residual is plotted against its expected value under normality. To make normal probability plots , as well as a histogram, type:

  ```
  > qqnorm(fit$res)
  > qqline(fit$res)
  > hist(fit$res)
  ```

Example (Clocks continued)

3 Multicollinearity

- What is the problem when predictor variables are strongly correlated? (or when one predictor variable is a linear combination of the others).
- Multicollinearity is the situation where one or more predictor variables are ?nearly? linearly related to the others.
- If one of the predictors is almost perfectly predicted from the set of other variables, then you have multicollinearity.



- How does this affect an analysis?

- Effects of multicollinearity:

  1. the estimated slopes have high variability; the b_is s have large standard errors
  2. High std. err. of b_i may mean few β_is are significantly different from 0 (testing $H_0 : \beta_i = 0$), even when a relationship truly exists.
  3. The estimated regression coefficients can vary widely from one sample to the next.

- Diagnosis of multicollinearity

  – examine pairwise correlations

  – look for b_is with unusual slopes (+/-)

  – use the Variance Inflation Factor (VIF) to make the call:

$$VIF_j = \frac{1}{1 - r_j^2}$$

where R_j^2 is the % of variability in X_j explained by all the other predictors (Compute r_j^2 by regressing X_j on the other variables).

- When r_j^2 is close to 1 (i.e. you can predict X_j very well from the other predictors), VIF_j will be large.
- Generally the multicollinearity between independent variables in considered severe if
  1. The largest $VIF > 10$ (which means $r_j^2 > 0.90$ for some j)
  2. The mean $V\bar{I}F$ of the variance inflation factors is substantially greater than 1.
- To comput VIF in R you need to install the package fmsb (do the following install.packages(fmsb) hit enter and then type library(fmsb) and hit enter)
- Then use VIF(fit) (fit is the object were you stored your regression results)
- Example (Clocks)

```
> VIF(fit)
[1] 9.32076
```

4 Stepwise Regression