Spring 2014
Stat 8004: Statistical Methods II
Lecture 5

Instructor: Jichun Xie

In this lecture, we discuss the diagosis of linear models.

1 Outliers and Leverage Points

1.1 Definition and Diagnosis

- Outliers: the points $(\mathbf{x}_i, y_i)$ with y_i "extrodinarily" large or small.
- Leverage points: the points $(\mathbf{x}_i, y_i)$ with some x_{ik} "extrodinarily" large or small.

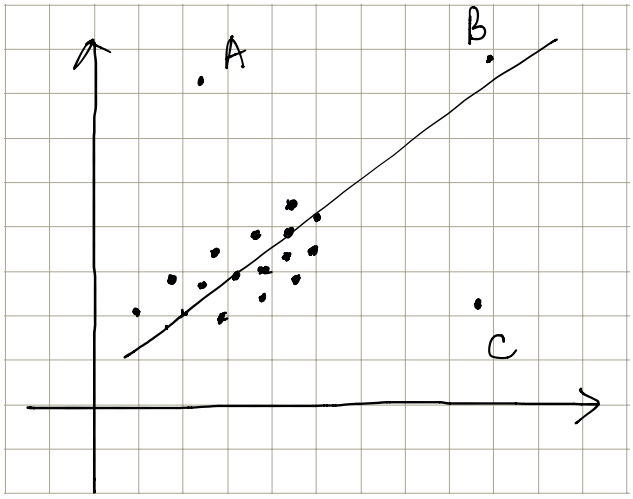

Figure 1: Outlier and Leverage Points

In Figure 1,

- A is an outlier

- B is an outlier and a leverage point
- C is also an outlier and a leverage point

How to detect outliers or leverage points? Here is an *ad-hoc* method.

Take the interquartile rage (IQR): $q_{3/4} - q_{1/4}$ of your data and multiply it by 1.5. Substract that number from $q_{1/4}$ and add that number to $q_{3/4}$. Any point lying outside these points can be considered as an outlier/leverage point.

Example: $\mathbf{y} = (12, 18, 19, 21, 25)$. Is there any outlier?

$$IQR = 3, \quad 1.5IQR = 4.5, \quad 18 - 4.5 = 13.5, \quad 21 + 4.5 = 25.5.$$

Therefore, we can consider 12 as an outlier.

Remark: In practice, the detection of outliers will be very flexible. As long as it affect the regression too much, it can be considered as an outlier.

From Figure 1, it is easy to see that some outliers/leverage points will affect the regression a lot, while some won't. How to decide whether we should delete a certain outlier/leverage point?

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \text{ where } \mathbf{e} \sim N(0, \sigma^2\mathbf{I}).$$

From prevoius lectures, we have

$$\hat{\mathbf{Y}} = \mathbf{P}_X\mathbf{Y}, \quad \hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y},$$

where $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}$ is a projection matrix.

$$\mathbb{E}(\hat{\mathbf{e}}) = 0, \quad \mathrm{Var}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{P}_X).$$

Suppose $\mathbf{P}_X = (h_{ij})_{n\times n}$. It is the projection of the column space of $\mathbf{X}$; sometimes, we call it "hat matrix".

Define *Internally Studentized Residual*:

$$r_i = \frac{e_i}{\hat{\sigma}(1 - h_{ii})^{1/2}},$$

where $\hat{\sigma}^2 = \frac{1}{n-p}\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$. It can be shown that $\frac{r_i^2}{n-p} \sim Beta(\frac{1}{2}, \frac{1}{2}(n-p-1))$.

The residuals and the estimator of σ^2 can be affected by outliers. To eliminate the effect, we define *Externally Studentized Residual*:

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}(1 - h_i)^{1/2}}.$$

The estimator $\hat{\sigma}$ is replaced by the estimator $\hat{\sigma}_{(i)}$, which is calculated in the usual way from the $n-1$ data points that remain after deleting the ith observation. In this way, even if the ith observation is an outlier, $\hat{\sigma}_{(i)}$ will no be affected.

If $|t_i|$ is too large, it might be an indicator that the ith observation deviates from the regression line too much. Usually, we consider deleting the ith point if $|t_i| > 2$. While fitting a linear regression, we need to check the data first and see if we need to delete any points before we finalize our results.

1.2 Example

Example: Olympic Records for High Jump, Discus and Long Jump. The data recorded winning heights or distances (inches) for the High Jump, Discus and Long Jump events at the Olympics up to 1996.

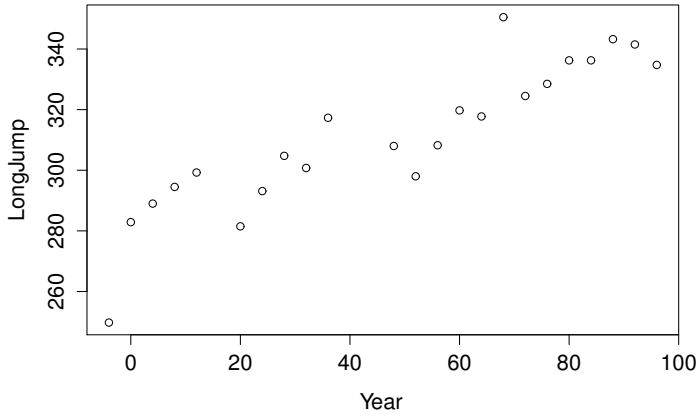

Figure 2: Long Jump vs. Year

Use IQR to check outliers:

```
> lj.sum <-summary(olympic$LongJump)
> IQR <- lj.sum[5] - lj.sum[2]
> names(IQR) <- "IQR"
> lower <- lj.sum[2] - 1.5*IQR
> upper <- lj.sum[5] + 1.5*IQR
> idx <- which( (olympic$LongJump < lower)|(olympic$LongJump >
   upper))
> idx
integer(0)
```

And then we fit the regression model, get the summary.

```
> fit <- lm(LongJump ~ Year, data = olympic)
> summary(fit)
```

```
Call:
lm(formula = LongJump ~ Year, data = olympic)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2984  -4.1304   0.0141   5.6293  25.3032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 278.77891    4.18428  66.625  < 2e-16 ***
Year          0.68262    0.07346   9.292 6.89e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.99 on 21 degrees of freedom
Multiple R-squared: 0.8044, Adjusted R-squared: 0.7951
F-statistic: 86.35 on 1 and 21 DF,  p-value: 6.892e-09
```

Next, we obtain the externally studentized residuals and plot.

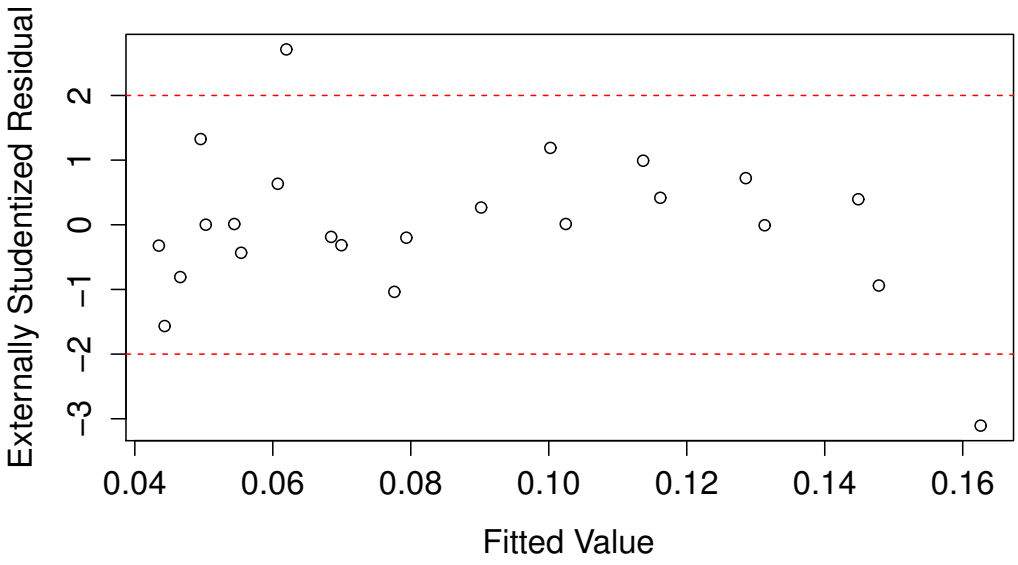

Figure 3: Externally Studentized Residual vs. Fitted Value

It is easy to see that there are two points might affect the regression line too much. They are the Point 1 and Point 16. After removing the points, we redo the regression. Here is the result:

```
Call:
lm(formula = LongJump ~ Year, data = olympic1)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0444  -3.7031   0.4079   4.7534  12.6333

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 283.60762    3.05274   92.90   <2e-16 ***
Year          0.58532    0.05288   11.07    1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.34 on 19 degrees of freedom
Multiple R-squared: 0.8657, Adjusted R-squared: 0.8587
F-statistic: 122.5 on 1 and 19 DF,  p-value: 1.001e-09
```

Compared with the results with all the points, the new regression has a larger R^2 value, indicating it might fit the data better.

2 Heteroscedasticity

2.1 Definition

Usual Model:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i, \tag{1}$$

where $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$.

We now have, instead,

$$\mathrm{Var}(e_i) = \sigma_i^2,$$

where σ_i^2 may depend either on the mean $\mathbb{E}(y_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, and possibly other parameters, or on a vector of (possibly additional) explanatory variables z_i. In this case, the ordinary least square estimator $\hat{\boldsymbol{\beta}}$ is not the "best" unbiased estimator of $\boldsymbol{\beta}$ if the variance σ_i^2 are not equal. We need to check the variances for equality and, if necessary, use more efficient estimation methods.

Now suppose the true model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

We know that $\hat{\mathbf{Y}} = \mathbf{P}_X\mathbf{Y}$ and $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$.

$$\text{Var}(\hat{\mathbf{e}}) = \text{Var}\{(\mathbf{I} - \mathbf{P}_X)\mathbf{y}\} = (\mathbf{I} - \mathbf{P}_X)\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{P}_X).$$

It leads to

$$\text{Var}(\hat{e}_i) = (1 - h_{ii})^2\sigma_i^2 + \sum_{k \neq i} h_{ik}^2\sigma_k^2.$$

Usually, $h_{ik} \leq h_{ii}$ for $k \neq i$, so very often large σ_i^2 are indicated by large residuals, though this will not be the case for high-leverage points.

$$\mathbb{E}(\hat{e}_i) = 0, \quad \text{Var}(\hat{e}_i) = \mathbb{E}(\hat{e}_i^2).$$

Define $b_i = \frac{\hat{e}_i^2}{1-h_{ii}}$. If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, then

$$\mathbb{E}(b_i) = (1 - h_{ii})\sigma^2 + \sum_{k:k\neq i} \frac{h_{ik}^2}{1 - h_{ii}}\sigma^2 = \sigma^2.$$

This is because $\mathbf{I} - \mathbf{P}_X$ is idempotent,

$$(1 - h_{ii})^2 + \sum_{k \neq i} h_{ik}^2 = 1 - h_{ii}.$$

Thus, when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, b_i has constant expectation.

2.2 How to check heteroscedasticity?

We can plot b_i *vs.* the fitted value. If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we should observe a random scatter plot. Note that the internally standardized residual

$$r_i = \frac{b_i}{\hat{\sigma}},$$

where $\hat{\sigma}$ is a constant across all i. Therefore, we can also plot r_i *vs.* the fitted value and see if there is any pattern.

2.3 Example

Example 1: Selling Price of Antique Grandfather Clocks. The data give the selling price at auction of 32 antique grandfather clocks. Also recorded is the age of the clock and the number of people who made a bid.

Variable	Description
Age	Age of the clock (years)
Bidders	Number of individuals participating in the bidding
Price	Selling price (pounds sterling)

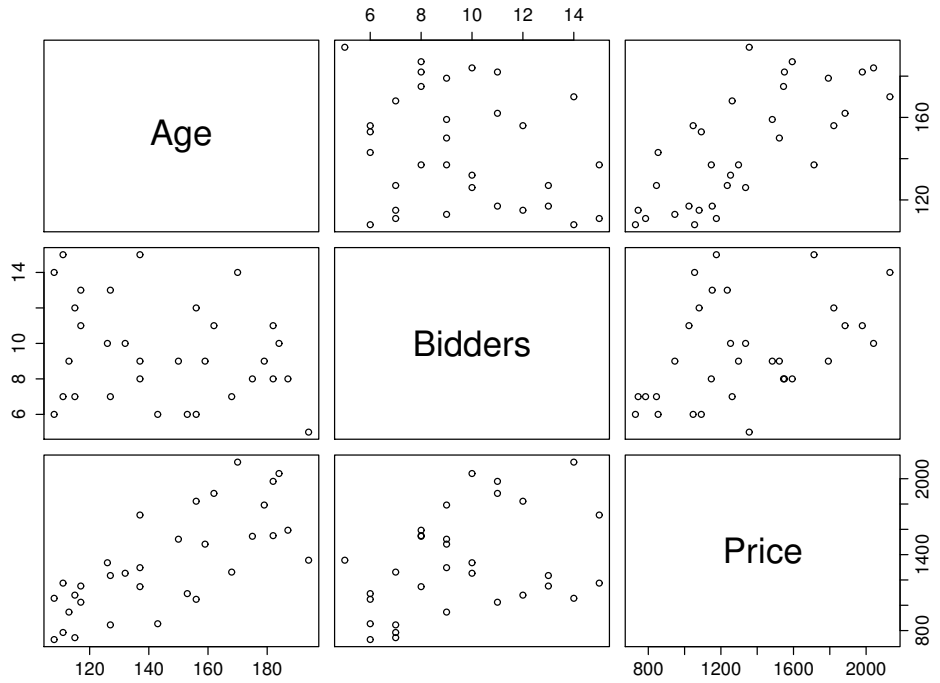

Figure 4: The scatterplot matrix for the clock auction price

```
> fit <- lm(Price~Age + Bidders, data = auction)
> summary(fit)

Call:
lm(formula = Price ~ Age + Bidders, data = auction)

Residuals:
   Min     1Q Median     3Q    Max
-207.2 -117.8   16.5  102.7  213.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
    .    0.1         1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared: 0.8927, Adjusted R-squared: 0.8853
```

```
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15
```

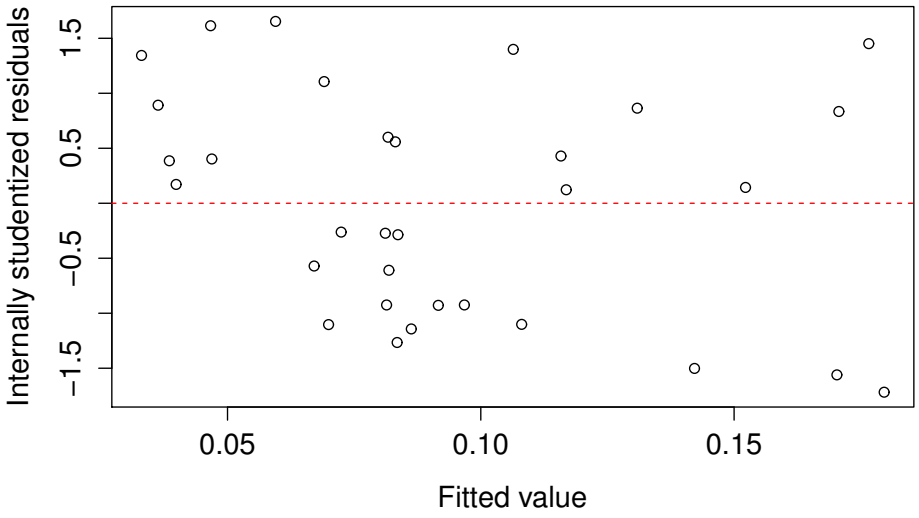

Figure 5: Internally studentized residual *vs.* fitted value

Both "Bidder" and "Age" have significant effect on the auction price of the clocks. R^2 is 0.89, indicating a large proportion of the variance of the auction price can be explained by these two variables. From Figure 5, we can see there is no obvious pattern of the internally studentized residuals.

What if we leave the covariate "Bidders' out, and only include the covariate "Age"?

```
> fit2 <- lm(Price ~ Age, data = auction)
> summary(fit2)

Call:
lm(formula = Price ~ Age, data = auction)

Residuals:
    Min      1Q  Median      3Q     Max
-485.29 -192.66   30.75  157.21  541.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -191.66     263.89  -0.726    0.473
Age            10.48       1.79   5.854  2.1e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
    .    0.1         1
```

```
Residual standard error: 273 on 30 degrees of freedom
Multiple R-squared: 0.5332, Adjusted R-squared: 0.5177
F-statistic: 34.27 on 1 and 30 DF,  p-value: 2.096e-06
```

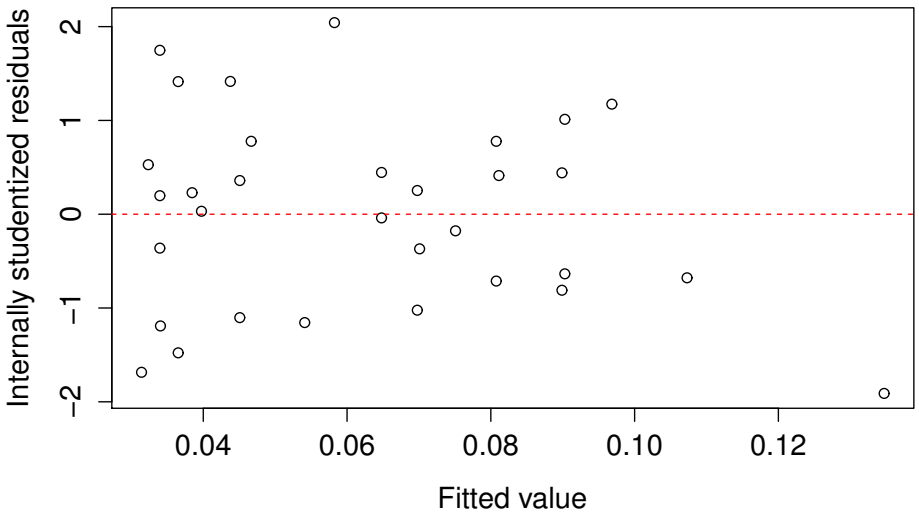

Figure 6: Internally studentized residual *vs.* fitted value

From Figure 6, we can observe some pattens. It seems that the variance of r_i is larger when the fitted values are small.

In reality, there might be multiple reasons if the internally studentized residual plot shows some pattern. The model could leave some important variables out, or the true underlying model might be higher order terms of the existing covariates, or we should do some transformation for the outcome variable, *etc.*.

3 Normality Assumption and Q-Q Plot

3.1 Q-Q Plot

3.1.1 Quantile Function

Consider a continous and strictly monotonic distribution function:

$$F : R \to (0, 1), \quad F(x) = \mathbb{P}(X \leq x) = p.$$

Tge quantile function is defined as

$$Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}.$$

3.1.2 Definition of Q-Q Plot

A Q-Q plot is a plot of the quantiles of two distributions against each other, or a plot on estimates of the quantiles. The pattern of points in the plot is used to compare two distributions.

Types of Q-Q plots:

- One known distribution *vs.* another known distribution
- One data set (empirical distribution) *vs.* one known distribution
- One data set (empirical distribution) *vs.* another data set (empirical distribution)

3.1.3 Normal Q-Q Plots

Suppose $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. We can rank $Z_1, \ldots, Z_n$ from the smallest to the largest:

$$\text{Order Statistics: } Z_{(1)} \leq \ldots \leq Z_{(n)}.$$

Since $Z_i \overset{iid}{\sim} N(\mu, \sigma^2)$,

$$\frac{Z_i - \mu}{\sigma} \overset{i.i.d.}{\sim} N(0, 1).$$

$$\Phi\left(\frac{Z_{(i)} - \mu}{\sigma}\right) \approx \frac{i}{n},$$

wehre Φ is the CDF of $N(0, 1)$. Thus

$$\frac{Z_{(i)} - \mu}{\sigma} \approx \Phi^{-1}\left(\frac{i}{n}\right)$$
$$Z_{(i)} \approx \sigma\Phi^{-1}\left(\frac{i}{n}\right) + \mu$$

This means that if Z_i are *i.i.d.* normal, we should expect to observe a linear trend between the order statistics of Z_i and the quantile of normal $\Phi^{-1}(i/n)$. Some researchers argue that we might do some heuristic corrections. The idea is to change i/n to $(i + a)/(n + b)$ to smooth the curve, with a and b equal to certain small numbers.

3.2 Linear Model and Normal Assumption

When we consider linear models, we need the normality assuption to do confidence interval construction and hypothesis testing.

Consider the linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Recall the internally studentized residual:

$$r_i = \frac{e_i}{\hat{\sigma}(1-h_{ii})^{1/2}},$$

where $\hat{\sigma} = \frac{1}{n-p}\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_X)\mathbf{Y}$.

It is easy to see that $r_i \xrightarrow{d} N(0,1)$ if $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Therefore, to check the normality assumption, we can plot Q-Q plot of r_i *vs.* $N(0,1)$ and check whether it is a line. In practice, we can also plot $\hat{e}_i$ *vs.* $N(0,1)$. Under most circumstances, they look similar.

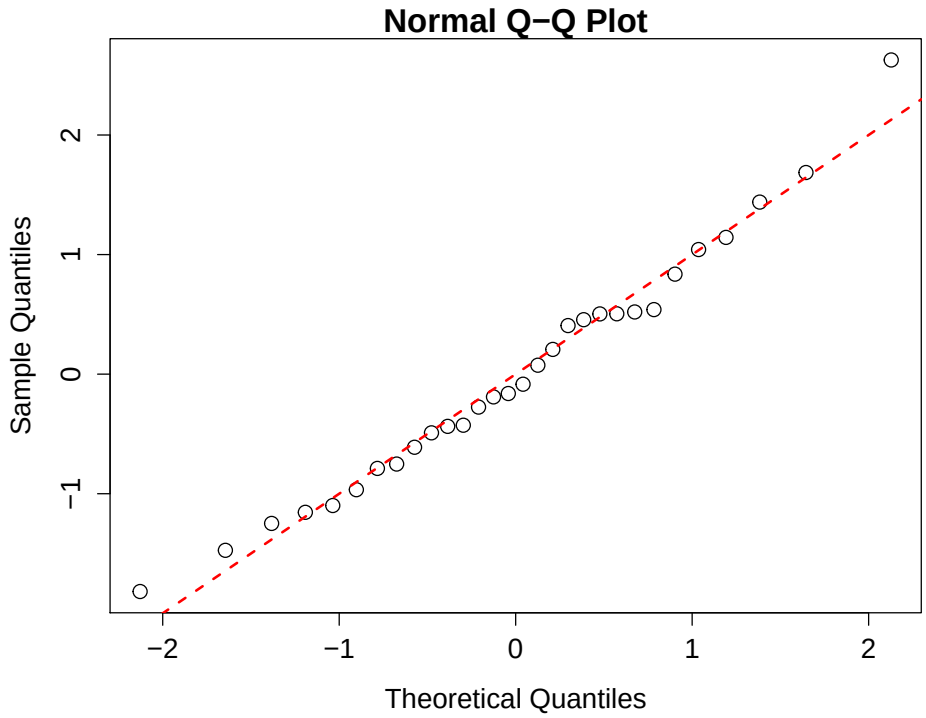

Figure 7: Q-Q Plot of Internally Studentized Residuals

Figure 7 shows the $Q-Q$ plots of internally studentized residuals of the regression model in the cheese taste example. We can clearly observe the linear pattern. Some pattern can be observed in Figure 8.

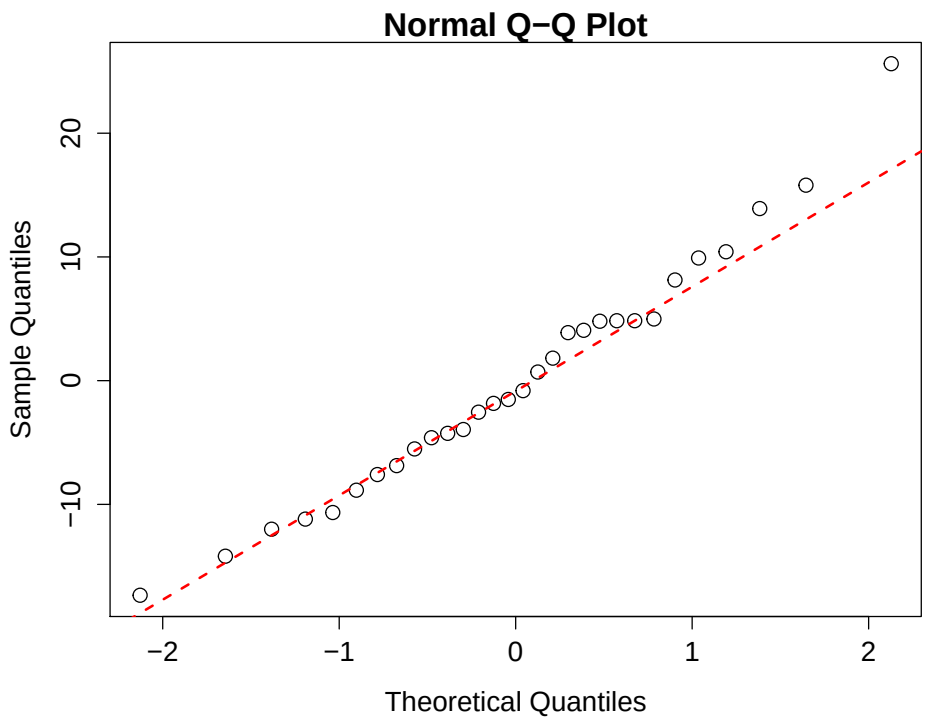

Figure 8: Q-Q Plot of Residuals

4 Collinearity and Variable Selection

Consider the linear model (1). Usually, there are more observations than the number of covariates, *i.e.* $n > p$.

Suppose $\mathbf{X}$ is a full-rank matrix, *i.e.* $\text{Rank}(X) = n$. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}, \quad \text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

4.1 Definition of Collinearity

If there is some variable $\mathbf{X}_i \approx \sum_{j \neq i} \alpha_j \mathbf{X}_j$. What will happen?

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is close to a singular matrix, and therefore can hardly be inverted.
- Even if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, it will be highly "unstable".

One way to show the problem of collinearity. Without loss of generality, suppose $X_p \approx \sum_{j=1}^{p-1} \alpha_j X_j$. Let $\mathcal{S} = \{1, \ldots, p-1\}$.

$$
\begin{aligned}
(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} &= \left[\begin{pmatrix} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} \\ \mathbf{X}_p^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} & \mathbf{X}_p^{\mathrm{T}} \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} (\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1} + \frac{1}{k}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}\mathbf{X}_p\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1} & -\frac{1}{k}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_p \\ -\frac{1}{k}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X})^{-1} & \frac{1}{k} \end{pmatrix},
\end{aligned}
$$

where $k = \mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{\mathcal{S}}})\mathbf{X}_p$.

If $\mathbf{X}_p \approx \sum_{j=1}^{p-1} \alpha_j \mathbf{X}_j$. Then $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{\mathcal{S}}})\mathbf{X}_p \approx 0$. $\Rightarrow k \approx 0$. $\Rightarrow 1/k \approx \infty$. Note that $\mathrm{Var}(\hat{\beta}_p) = 1/k$. This means that , if X_p can be almost linearly represented by other covariates, $\mathrm{Var}(\hat{\beta}_p)$ will be very large. In other words, $\hat{\beta}_p$ is very unstable.

Another way to show the problem of collinearity:

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \ldots + \mathbf{X}_{p-1}\beta_{p-1} + \mathbf{X}_p\beta_p + \mathbf{e},$$

where $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$. If $\mathbf{X}_p \approx \sum_{j=1}^{p-1} \alpha_j X_j$, then

$$
\begin{aligned}
\mathbf{Y} &\approx \mathbf{X}_1\beta_1 + \ldots + \mathbf{X}_{p-1}\beta_{p-1} + (\sum_{j=1}^{p-1} \alpha_j \mathbf{X}_j)\beta_p + \mathbf{e} \\
&= \mathbf{X}_1(\beta_1 + \alpha_1\beta_p) + \ldots + X_{p-1}(\beta_{p-1} + \alpha_{p-1}\beta_p) + \mathbf{e}
\end{aligned}
$$

It will reduce to an almost non-full rank model. The solution of $\hat{\boldsymbol{\beta}}$ is close to non-unique (unstable).

Therefore, when there are lots of covariates, it increases the probability of collinearity. It is better to reduce the number of covariates. In practice, we can use scatter plot matrix to examine whether there is any collinearity.

4.2 Example

Example: Chedder Cheese Tasting. As cheese ages, various chemical processes take place that determine the taste of the final product. This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The variables "Acetic" and "H2S" are the natural logarithm of the concentration of acetic asid and hydrogen sulfide respectively. The variable "Lactic" has not been transformed. There are 30 observations.

Variable	Description
Taste	Subjective taste test score, obtained by combining the scores of several tasters
Acetic	Natural log of concentration of acetic acid
H2S	Natural log of concentration of hydrogen sulfide
Lactic	Concentration of lactic acid

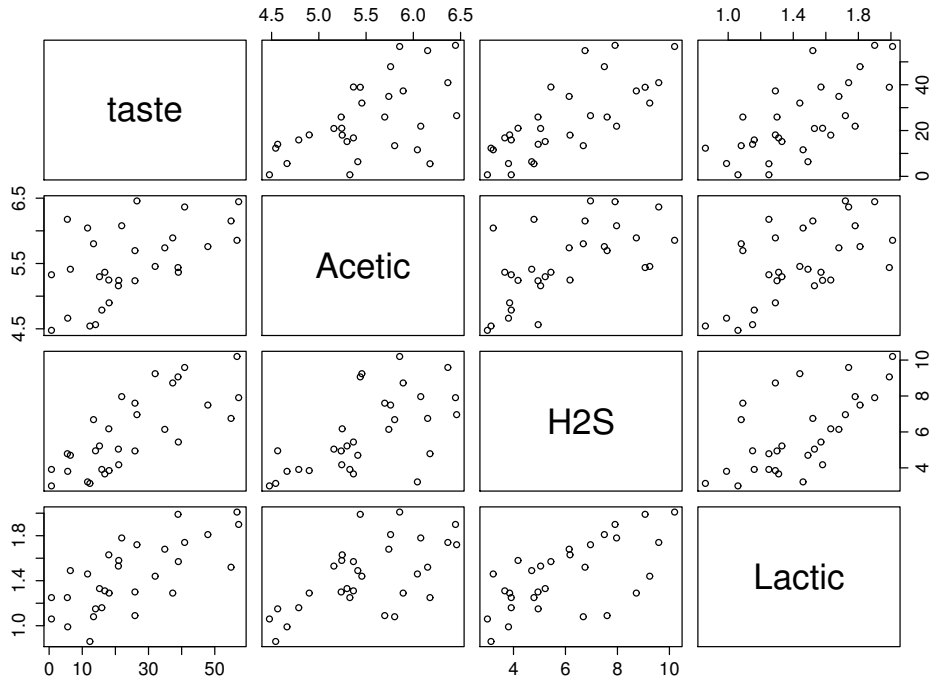

Figure 9: Cheese Taste

It seems that there is some collinearity between Acetic and Lactic, and also Acetic and H2S.

```
> fit1 <- lm(taste ~ Acetic + H2S + Lactic, data = cheese)
> summary(fit1)

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518, Adjusted R-squared: 0.6116
```

```
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

4.3 Choosing Subset

How to choose the subset? There are many methods for choosing a best subset.

4.3.1 Goodness of Fit Criteria

One of the commonly used criteria is R^2. Recall that when we discussed testing hypothesis for linear models, we introduced SSR, which is

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

The R^2 is just the ratio of RSS and SST_m.

Recall that

$$SST_m = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

and also

$$SST_m = SSE + SSR.$$

Therefore,

$$R^2 = \frac{SSR}{SST_m} = 1 - \frac{SSE}{SST_m}.$$

In other words, R^2 measures how much proportion of the variance of Y could be estimated by the linear combination of $X_1, \ldots, X_p$. In the traditional setting (when the number of covariates are not too large), the larger the R^2 is, the better the fitting is.

In the chedder cheese tasting example, if we include all the variables in the model, $R^2 = 0.65$.

Removing any one of the covariate will simply reduce R^2, as long as this covariate cannot be linearly represented by other covariates. Why?

For example, in the cheese example, if we only include "H2S" and "Lactic", then R^2 will reduce to 0.65.

```
> fit2 <- lm(taste ~ H2S + Lactic, data = cheese)
> summary(fit2)

Call:
lm(formula = taste ~ H2S + Lactic, data = cheese)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared: 0.6517, Adjusted R-squared: 0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

Appearently, R^2 is not a criteria for whether we should remove a variable or not, since it will always prefore the full model. However, we can look at the adjusted R^2, which is

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{SSE}{SST_m}\frac{n-1}{n-p-1} = 1 - \frac{MSE}{MST_m}.$$

Compare to R^2, $\bar{R}^2$ penalizes on the complexity of the model. Therefore, when removing one covariate in the model, also R^2 will always decrease, $\bar{R}^2$ might still increase.

In the cheese example, when we include all the variables, $\bar{R}^2 = 0.61$. After we remove "Acetic", $\bar{R}^2 = 0.63$. The reduced model seems to work better than the full model.

When we have a small number of covariates, we can try all the subsets of covariates and compare the $\bar{R}^2$. Suppose in the full model, there are p covariates, the number of all the subsets is 2^p. When p is large, it is impossible to try all the reduced-models. Is there any method that can choose the subset in a more efficient way?

4.3.2 Forward, Backward and Stepwise Selection

Forward Selection.

Step 1 Start from the model with only the intercept term.

Step 2 Add one moe variable in the model. Check the results for all the models at this stage. Choose the model with the p-value of the newly added variable significant an the smallest.

Step 3 Repeat Step 2 util the p-value of all the newly added variables are not significant.

Let's discuss the cheese tast example.

Add one variable.

Added variable	p-value
Acetic	1.66E-3
H2S	1.37E-6
Lactic	1.41E-5

The most significant one is "H2S". Add it to the model. Fit two models, the one with "H2S" and "Acetic", and the one with "H2S" and "Lactic". Check the significance level of "Acetic" and "Lactic".

Added variable	p-value
Acetic	0.42
Lactic	0.02

We should therefore include "Lactic" in the model. Now add the only left "Acetic" in the model. It turns out that it's p-value is 0.94. It is not significant. We do not include the variable in the model. Therefore, the final model is the one with "H2S" and "Lactic".

In R, the *step* function can do the forward selection automatically. But instead of using p-value as an criterion, it uses Akaike information criterion (AIC).

$$AIC = 2k - 2\log(L).$$

The *step* chooses the model with the smallest AIC.

```
> step(lm(taste ~1, data = cheese),
  scope=list(lower=~1,upper=~ H2S + Lactic + Acetic),
  direction="forward")
Start:  AIC=168.29
taste ~ 1

          Df Sum of Sq    RSS    AIC
+ H2S      1    4376.7 3286.1 144.89
+ Lactic   1    3800.4 3862.5 149.74
+ Acetic   1    2314.1 5348.7 159.50
<none>                 7662.9 168.29

Step:  AIC=144.89
taste ~ H2S

          Df Sum of Sq    RSS    AIC
+ Lactic   1    617.18 2669.0 140.65
<none>                 3286.1 144.89
+ Acetic   1     84.41 3201.7 146.11

Step:  AIC=140.65
taste ~ H2S + Lactic
```

```
          Df  Sum of Sq     RSS     AIC
<none>                    2669.0  140.65
+ Acetic   1    0.55427 2668.4  142.64

Call:
lm(formula = taste ~ H2S + Lactic, data = cheese)

Coefficients:
(Intercept)          H2S       Lactic
    -27.592        3.946       19.887
```

It also ends up with the model with "H2S" and "Lactic".

Backward Selection.

Step 1 Start from the full model with all the vairables in the model.

Step 2 Check the p-value of all the variables in the model. Remove the one with the largest non-significant p-value.

Step 3. Repeat Step 2 until all the p-values are significant.

Similar to the forward selection algorithm. We can also change the p-value criterion to AIC. Then, we can use the *step* function in R to do the selection.

```
> step(lm(taste ~ Acetic + H2S + Lactic, data = cheese),
+      direction="backward")
Start:  AIC=142.64
taste ~ Acetic + H2S + Lactic

          Df  Sum of Sq     RSS     AIC
- Acetic   1       0.55 2669.0  140.65
<none>                  2668.4  142.64
- Lactic   1     533.32 3201.7  146.11
- H2S      1    1007.66 3676.1  150.25

Step:  AIC=140.65
taste ~ H2S + Lactic

          Df  Sum of Sq     RSS     AIC
<none>                  2669.0  140.65
- Lactic   1     617.18 3286.1  144.89
- H2S      1    1193.52 3862.5  149.74

Call:
lm(formula = taste ~ H2S + Lactic, data = cheese)
```

```
Coefficients:
(Intercept)          H2S       Lactic
    -27.592        3.946       19.887
```

It also ends up with the model with "H2S" and "Lactic".

Stepwise Selection.

For forward selection, once a covariate is added to the model, there is no chance that it could be removed. For backward selection, once a covariate is removed from the model, there is no chance that it could be added back. Under some cases, it might not be a good idea. For example, one covariate might be significant because it is highly correlated to another significant variable. Once the other one is added, this variable is no longer significant and you might want to remove it from the model.

Step 1 Do forward step.

Step 2 If some variable is added, do backward step and check if all the variables are significant. If not, remove the one with the largest p-value.

Step 3 Repeat Step 1 and Step 2 until no more variables can be added in Step 1.

The p-value criterion can also be changed to AIC.

```
> step(lm(taste ~ 1, data = cheese),
+      scope=list(lower=~1,upper=~ H2S + Lactic + Acetic),
+      direction="both")
Start:  AIC=168.29
taste ~ 1

          Df Sum of Sq    RSS    AIC
+ H2S      1    4376.7 3286.1 144.89
+ Lactic   1    3800.4 3862.5 149.74
+ Acetic   1    2314.1 5348.7 159.50
<none>                7662.9 168.29

Step:  AIC=144.89
taste ~ H2S

          Df Sum of Sq    RSS    AIC
+ Lactic   1     617.2 2669.0 140.65
<none>                3286.1 144.89
+ Acetic   1      84.4 3201.7 146.11
- H2S      1    4376.7 7662.9 168.29

Step:  AIC=140.65
taste ~ H2S + Lactic
```

```
          Df Sum of Sq    RSS    AIC
<none>                  2669.0 140.65
+ Acetic   1      0.55 2668.4 142.64
- Lactic   1    617.18 3286.1 144.89
- H2S      1   1193.52 3862.5 149.74

Call:
lm(formula = taste ~ H2S + Lactic, data = cheese)

Coefficients:
(Intercept)          H2S       Lactic
    -27.592        3.946       19.887
```