Linear Regression (part 2)

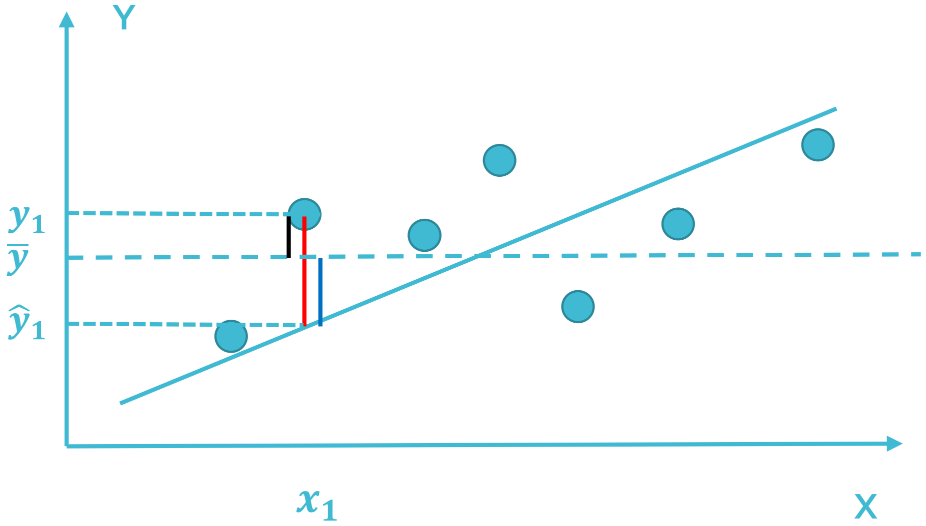
DSA 6000: Data Science and Analytics, Fall 2019

Wayne State University

Assessing the Accuracy of the Model

- Residual Sum of Squares (RSS): $\sum_{i=1}^{n}\left(y_i - \hat{f}(X_i)\right)^2$, measures the amount of variability in Y that is left unexplained after performing the regression.
- Residual Standard Error (RSE) is an estimate of the standard deviation of ϵ.
  - RSE = $\sqrt{\left(\frac{1}{n-p-1}\right) RSS}$
  - Note that RSE depends on p, so adding a useless predictor to the model increases $\left(\frac{1}{n-p-1}\right)$, overall RSE might also increase
- RSE represents the average amount that the response will deviate from the true regression line. It is a measure of the *lack of fit* of the model to the data, in the units of Y.
- R^2 statistic: the proportion of variance in Y that is explained by the model.
  - $R^2 = \frac{TSS-RSS}{TSS}$, where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares.
  - Adding an extra predictor will always increase R^2
  - Adjusted R^2 accounts for the model complexity

Decompose the TSS



- Suppose the linear model is fit by the OLS method, then
- Total variation in Y is decomposed into two parts:
  - Variation explained by the model
  - Variation left unexplained by the model
- $Total\ SS = Explained\ SS + Residual\ SS$
- $\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

About the F-statistic

- F-statistic is used for testing whether at least one of the predictors has a significant effect on the response variable.
- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_1: at least one β_j is non-zero
- A large F-statistic value will lead to rejection of H_0. The rejection threshold depends on both n and p.
- Why use F test since we already have the t test?
- For a model with many predictors (i.e., large p), it can happen that the p-value for some individual predictor(s) is small (e.g., < 0.05), but the model as a whole fails the F test (i.e., fail to reject H_0).
  - For instance, if there are 100 variables, all unrelated to Y, the p-values for about 5% of the variables will be below 0.05 **by chance**. We would expect to see about 5 small p-values even in the absence of any true association between the predictors and the response.
  - F-statistic is immune to this type of fallacy.

Interpreting LR outputs

```
Call:
lm(formula = sales ~ TV + radio, data = ad, subset = trainIndex)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8720 -0.8629  0.2989  1.1603  2.9526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.839160   0.373493   7.602 4.17e-12 ***
TV          0.045219   0.001721  26.275  < 2e-16 ***
radio       0.191949   0.010121  18.965  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.767 on 137 degrees of freedom
Multiple R-squared:  0.8881,    Adjusted R-squared:  0.8865 
F-statistic: 543.8 on 2 and 137 DF,  p-value: < 2.2e-16
```

Everything else held equal, one more unit of ad expense on TV will on average increase sales by 0.045 unit.

Model with Interaction Effects

```
> summary(lm(sales~ TV + radio, data= ad))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
```

- **Model**: Sales = 2.921 + 0.046*TV + 0.188*radio + ϵ
- **Interpretation**: e.g., regardless of the spending on TV advertisement, holding it at the same level, a unit change in radio advertisement will cause 0.188 unit of change in sales in the same direction

```
> summary(lm(sales~ TV + radio + TV:radio, data= ad))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
```

- **Model**: Sales = 6.75 + 0.019*TV + 0.029*radio + 0.001*TV*radio + ϵ
- **Interpretation**: The effect of radio advertising on sales now depends on the amount of the TV advertising.
- Positive interaction effect is called **synergy**, negative called **friction**.

Deciding on Important Variables

- Variable Selection is studied extensively in Chapter 6.
- Three classical approaches:
  - **Forward selection**: start with null model, add a single variable at a time whose addition results in the lowest RSS (or highest R^2) among all possible single-variable additions, until Adj. R^2 (or AIC, or BIC) starts the deteriorate.
  - **Backward selection**: start with full model, remove the variable with the largest p-value, refit and repeat, until all p-values are small enough.
  - **Mixed selection**: start with null model, add a single variable at a time as in Forward selection, if some variable in the updated model has a large p-value, remove it; repeat until all variables in the model have a small p-value and all remaining variables would have a high p-value if included in the model.

Potential Problems in a Linear Model

- Linear regression makes strong assumptions about the relationship between Y and X
  - Recall: What are the assumptions?
- Real data rarely conform to all those assumptions
- Potential problems:
  - Nonlinearity of the response-predictor relationship
  - Correlation of error terms
  - Non-constant variance of the error terms
  - Outliers
  - High-leverage points
  - Collinearity

Nonlinearity

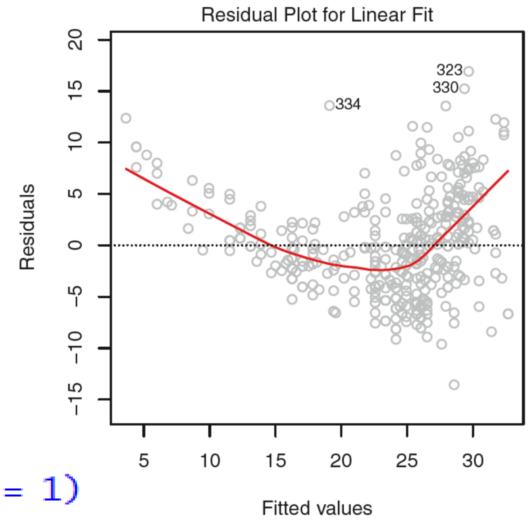
- The relationship between Y and X is nonlinear, but we fit a linear model
- Can be revealed by **diagnostic plots** (e.g., the residual plot)
  - plot() a linear fit will generate a series of diagnostic plots
  - You can select a specific plot by using the "which = " option
- Solutions:
  - Apply nonlinear transformation on predictors, or include higher-order terms of predictors, and fit a linear model using the transformed variables (how the coefficients are interpreted will change)
  - Use nonlinear models (Chapter 7)

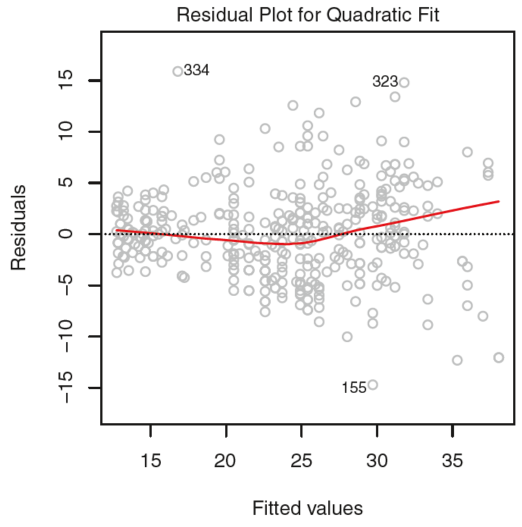
Nonlinear relationships

```
> summary(lm(mpg ~ poly(horsepower, 2), data = Auto))

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            23.4459     0.2209  106.13   <2e-16 ***
poly(horsepower, 2)1 -120.1377     4.3739  -27.47   <2e-16 ***
poly(horsepower, 2)2   44.0895     4.3739   10.08   <2e-16 ***
```

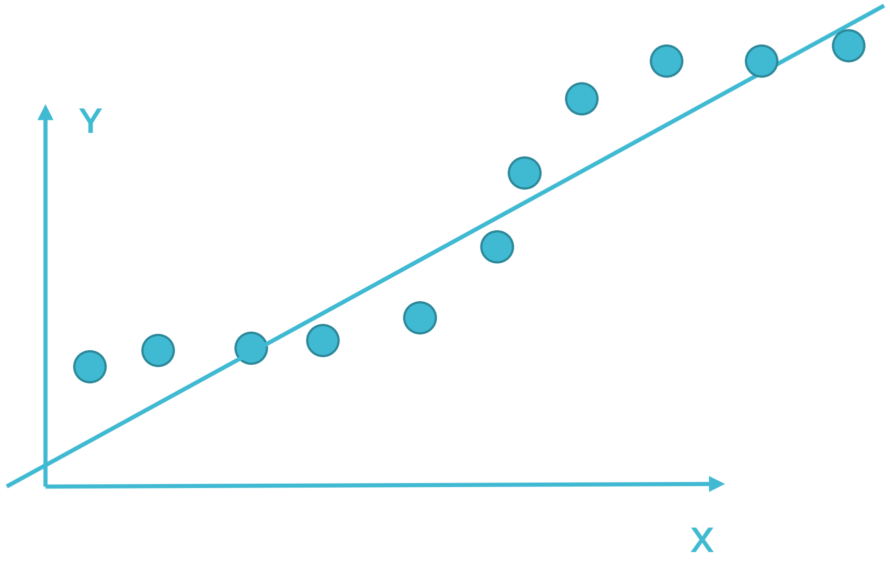
- **Model**: mpg = 23.5 -120*horsepower + 44*horsepower^2 + ϵ
- It models the nonlinear relationship between mpg and horsepower
- Coefficients are estimated by OLS, so it is still a linear fit.
- What makes you try including a higher order term of horsepower in the first place? The diagnostic plots.

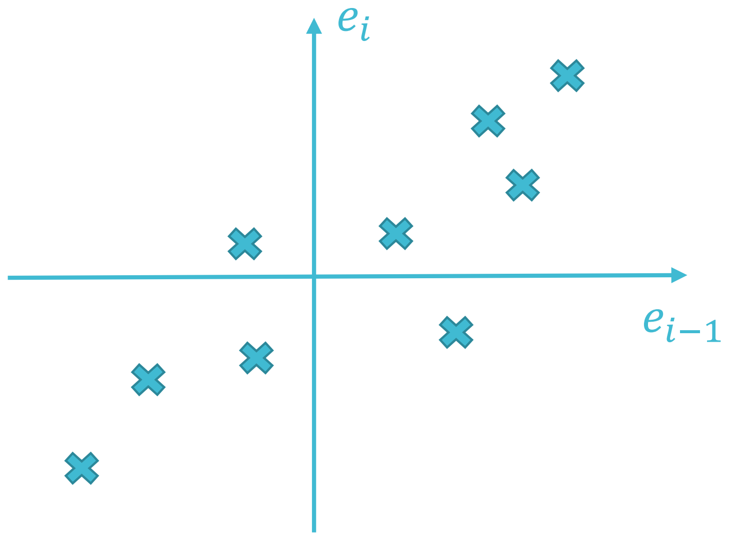





```
plot(lm(mpg ~ horsepower, data = Auto), which = 1)
plot(lm(mpg ~ poly(horsepower,2), data = Auto), which = 1)
```

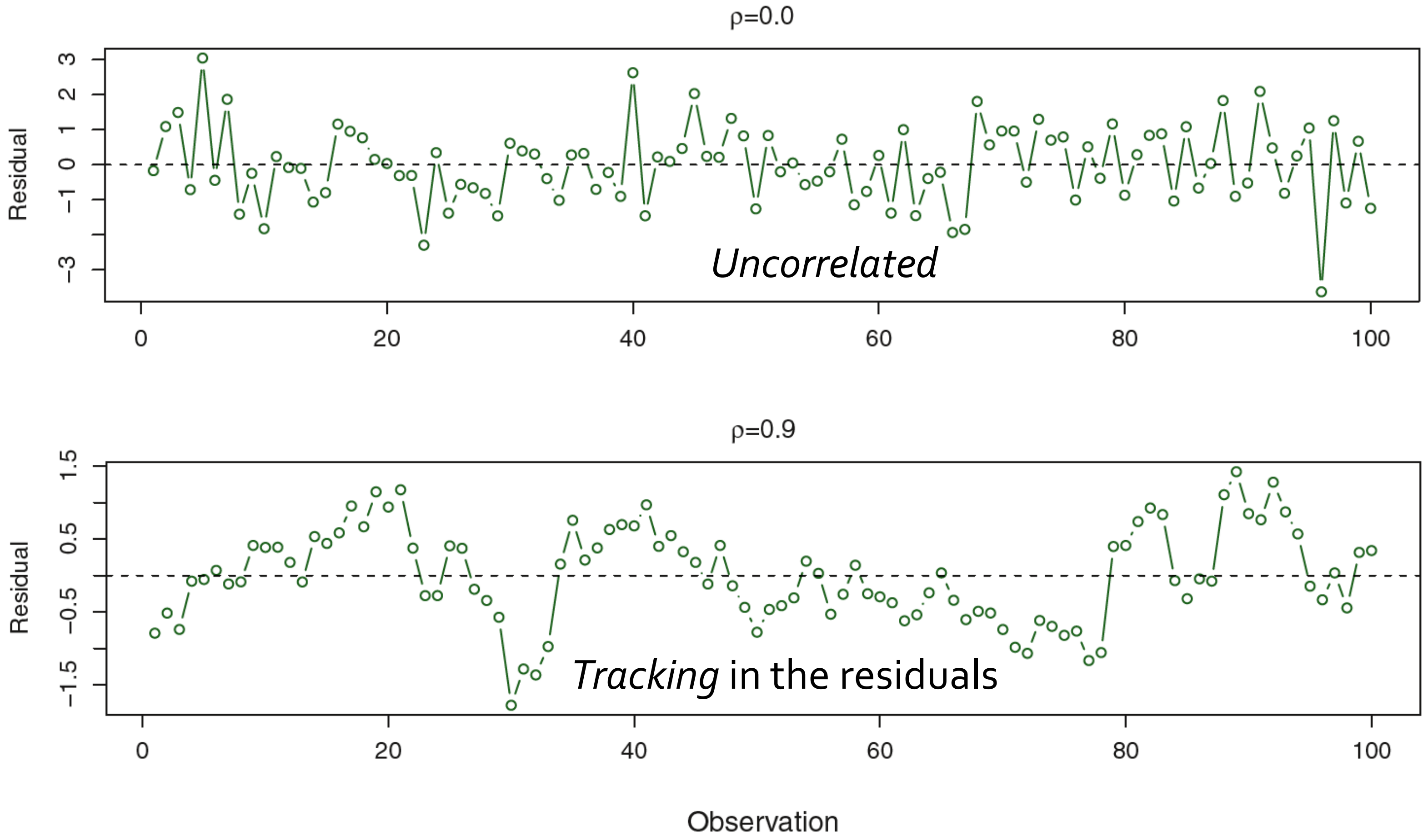
Correlation of error terms





- Autocorrelation: statistical dependence of errors on preceding errors
- Can be tested using Durbin-Watson test
- Positive autocorrelation can make $SE(\hat{\beta}_j)$ an underestimate of σ_{β_j}, and can over-estimate R^2

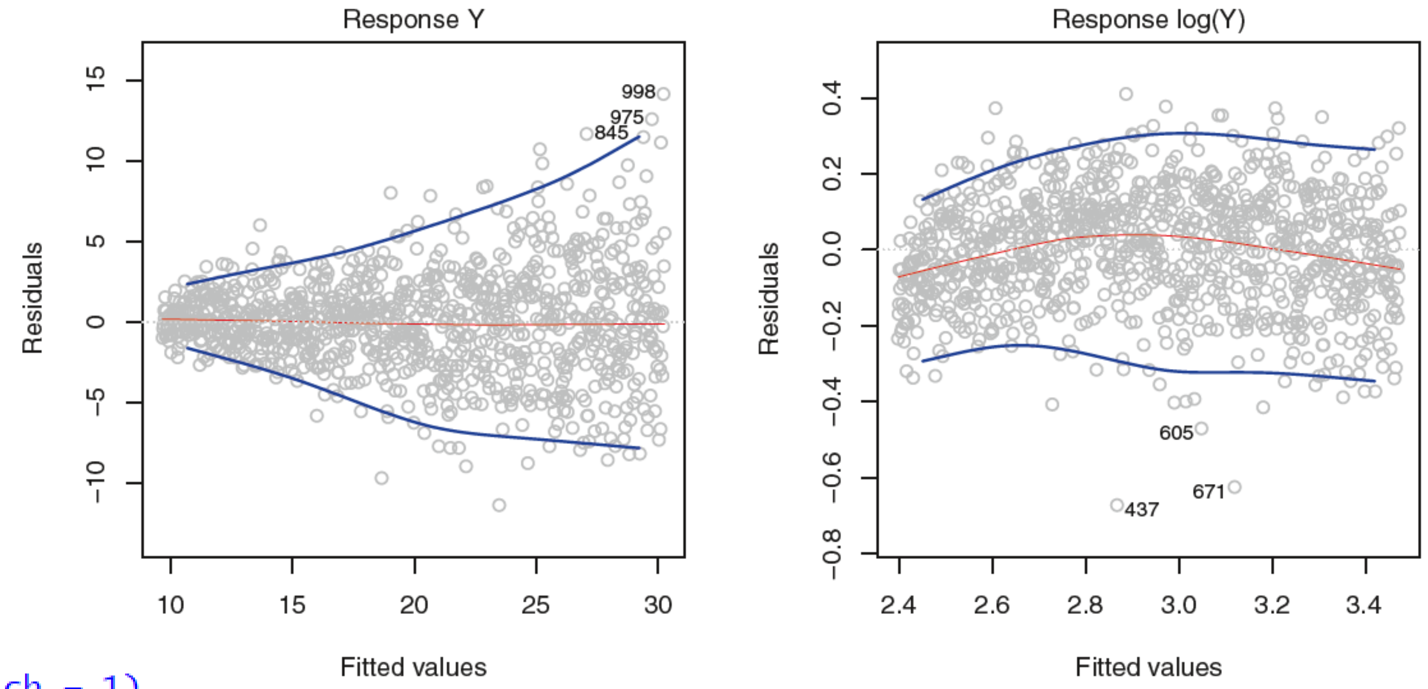
Correlation of error terms



- Occurs most often in data taken from observations over time
- Can occur outside the time series data.
- May be due to undiscovered nonlinearity or to missing variables
- Good experimental design can mitigate the risk of such correlations

Non-constant variance of the error terms (Heteroscedasticity)

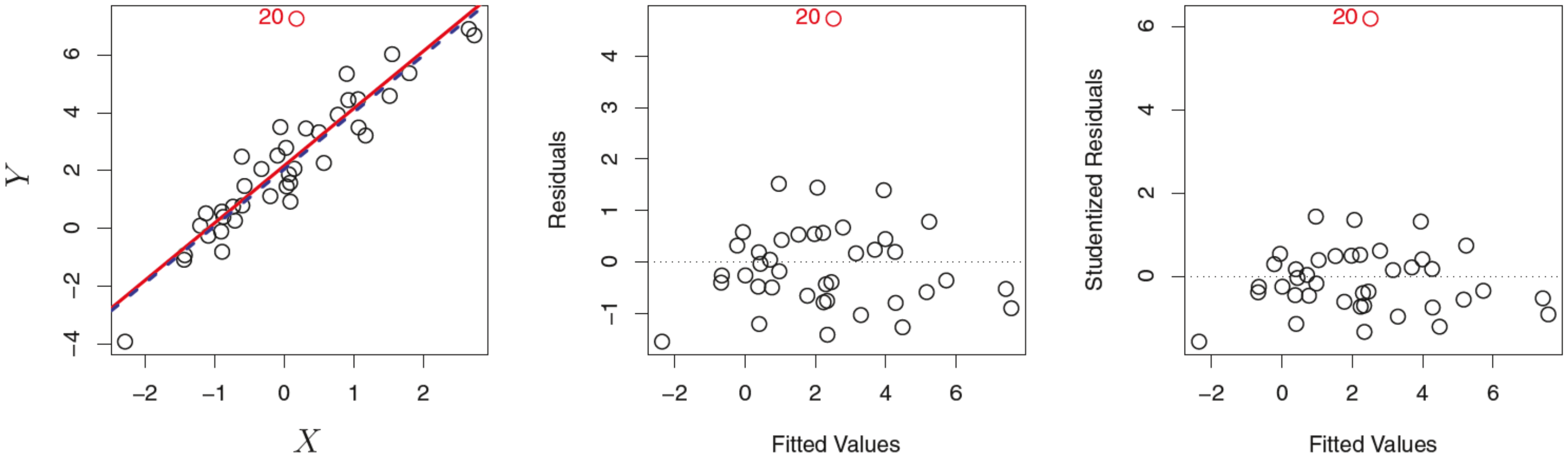
- Violates the assumption: $Var(\epsilon_i) = \sigma^2$
- Funnel shape in the residual plot
- If each X_i is an average of n_i raw data points, can use Weighted Least Squares (using n_i as the weight for obs i) to mitigate
- Nonlinear transform, e.g., $\log(Y)$, $\sqrt{Y}$ can also mitigate, but introduce nonlinearity




```
> plot(lm(sales ~ TV, data = ad), which = 1)
```

Outliers

- An outlier is a point for which y_i is far from the value predicted by the model.
- An outlier may not drastically affect the coefficient estimate, but it can hurt the model's explanatory power, e.g., the R-squares.

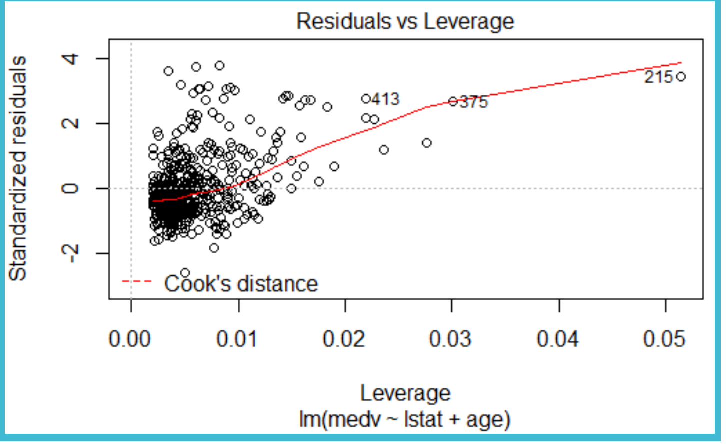


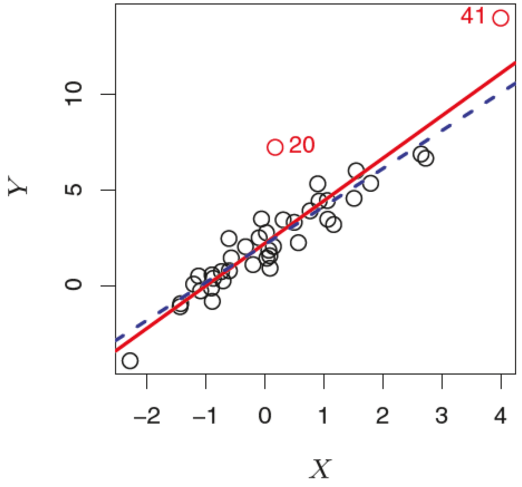
Left: Exclusion of the outlier obs 20 caused little change in the fit (red solid line v.s. black dashed line)
Middle: Residual plot can reveal an outlier
Right: We would expect all studentized residuals to fall between -3 to 3. Any residual outside the normal range can be regarded as an outlier

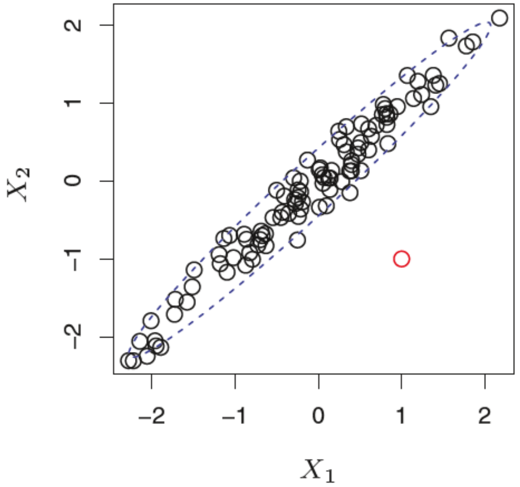
```
plot(lm(mpg ~ horsepower, data = Auto), which = 1)
```

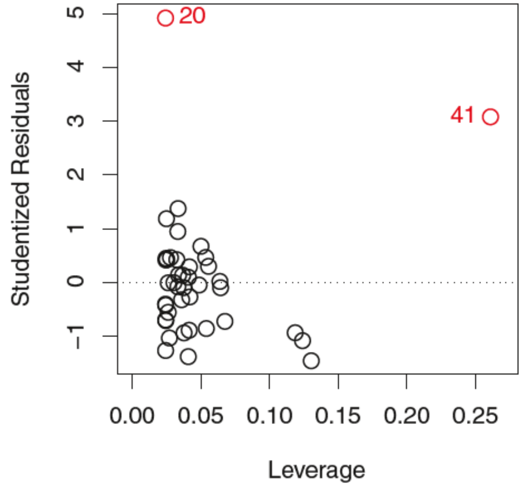
High-leverage points

- A high-leverage point has an unusual independent variable value
- Easy to identify in simple linear regression, hard to identify graphically in multiple linear regression
- We can use the *leverage statistic* to quantify the leverage of each point
- The average leverage for all the observations is always equal to $(p+1)/n$
- So if an observation's leverage greatly exceeds the average, it is a high-leverage point









Multicollinearity

- **Collinearity** refers to the situation in which two or more independent variables are closely related to one another.
- It reduces the accuracy of the coefficient estimate
  - A variable may have a significant effect on the response, but we may failed to recognize it (i.e., fail to reject $H_0: \beta_j = 0$) due to its large standard error caused by collinearity
- Pairwise collinearity can be detected by pairwise scatter plots (pairs() function in R), but scatter plots cannot detect **multicollinearity**.
- Multicollinearity can be assessed by variance inflation factor (VIF).
  - VIF of a variable j is $1/(1 - R_j^2)$, where R_j^2 is the R-square statistic obtained by fitting a linear model using X_j as the response and using all other predictors as independent variables.
- VIF > 5 indicates problematic multicollinearity due to presence of the variable. The variable should be removed.