STA 5820
Chapter 3
Linear Regression

Kazuhiko Shinki

Wayne State University

Overview:

Most of the essential features of linear model have already been discussed in (i) Chapters 6 and 11 in STA 5030 and (ii) Chapter 2 of this course.
In this chapter, we will discuss

(a) '`lm`' function's outputs which have not been discussed.

(b) how to check assumption of a linear model, and how to deal with violation on assumptions.

(c) how to handle qualitative (categorical) predictors.

- testing categorical predictors
- handling ordered predictors

3.a Outputs of lm function

```
> Dad <- read.table("Advertising.csv",sep=",",header=T)
> attach(Dad)
> LMB <- lm(sales ~ TV + I(TV^2) + radio)
> summary(LMB)
Residuals:
    Min      1Q  Median      3Q     Max
-7.3860 -0.8822 -0.0498  0.9613  3.5725

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.288e+00  3.588e-01   3.588 0.000421 ***
TV           7.844e-02  4.985e-03  15.736  < 2e-16 ***
I(TV^2)     -1.136e-04  1.677e-05  -6.775 1.42e-10 ***
radio        1.930e-01  7.293e-03  26.465  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.517 on 196 degrees of freedom
Multiple R-squared:  0.9167,        Adjusted R-squared:  0.9154
F-statistic:   719 on 3 and 196 DF,  p-value: < 2.2e-16
```

3.a Outputs of lm function

Recall that the above output says that the estimated model is:

$$\textbf{\textit{sales}} = 1.288 + 0.07844\textit{TV} - 0.0001136\textit{TV}^2 + 0.193\textit{radio} + \textit{error}$$

Below we use the notation $\boldsymbol{y = f(x) + \epsilon = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p + \epsilon}$.

In our Advertising example, $\boldsymbol{p = 3}$, $\boldsymbol{x_1 = TV}$, $\boldsymbol{x_2 = TV^2}$ and $\boldsymbol{x_3 = radio}$. The sample size $\boldsymbol{n = 200}$.

3.a Outputs of lm function

Residuals

Recall that the residual is the estimated error.
When the model is $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \epsilon$ and $\hat{\boldsymbol{f}}$ is the estimated $\boldsymbol{f}$, the estimated error $\hat{\epsilon}$ is defined by $\boldsymbol{y} = \hat{\boldsymbol{f}}(\boldsymbol{x}) + \hat{\epsilon}$.

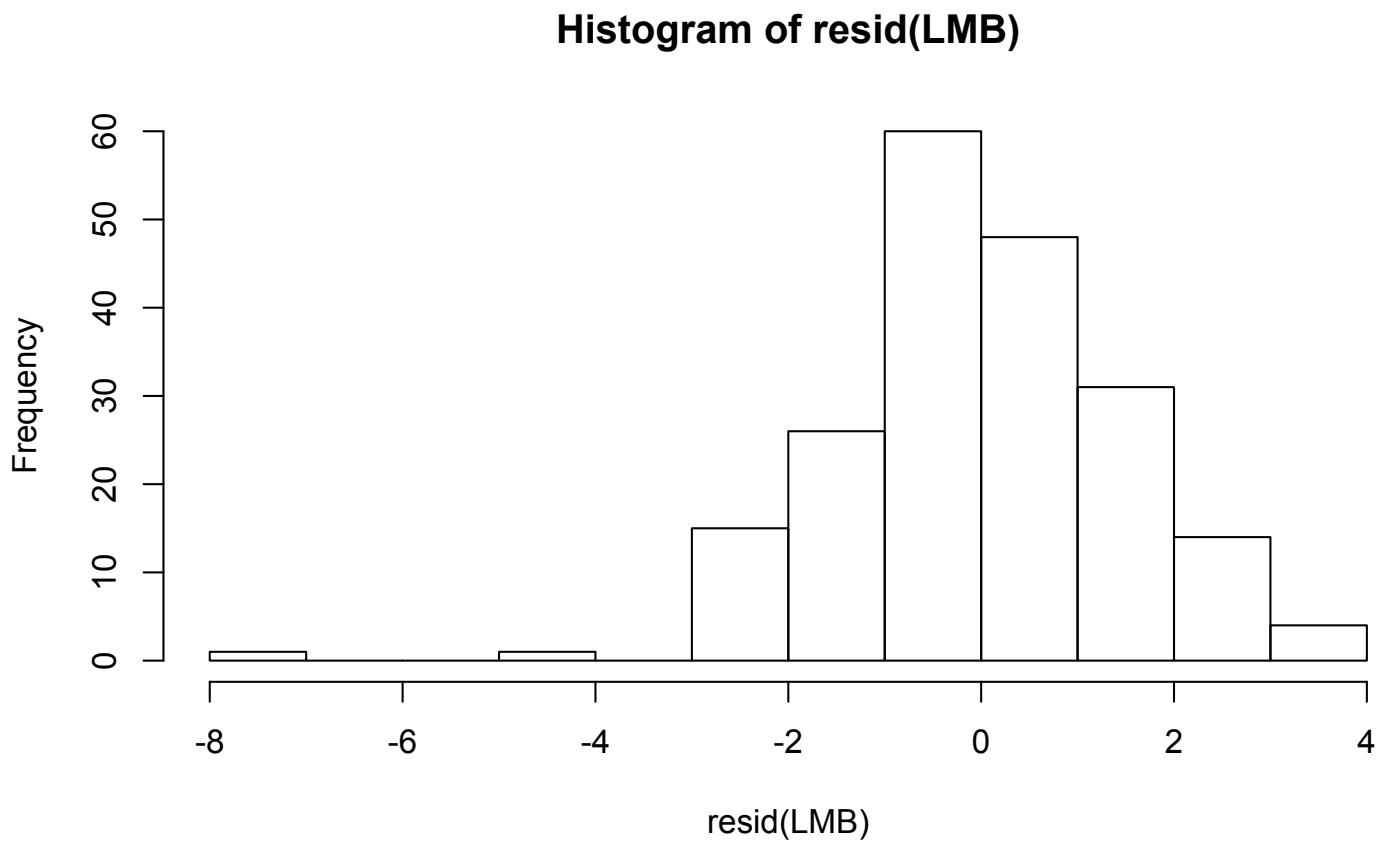
```
Residuals:
    Min      1Q  Median      3Q     Max
-7.3860 -0.8822 -0.0498  0.9613  3.5725
hist(resid(LMB))
```

The R otuput includes some statistics on the distribution of the residuals.

- The minimum of the residuals is -7.3860.
- The 1st quartile of the residuals is -0.8822.
- The median of the residuals is -0.0498.
- The 3rd quartile of the residuals is 09613.
- The maximum of the residuals is 3.5725.

3.a Outputs of lm function

Figure: The histogram of the resudials.



3.a Outputs of lm function

Residual standard error

A linear regression estimated by least square assumes that ϵ follows a distribution with mean 0 and a standard deviation σ.

```
Residual standard error: 1.517 on 196 degrees of freedom
```

R output indicates that estimated standard deviation $\hat{\sigma}$ is 1.517.
This is calculated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1 \cdots - \beta_p x_p)^2 \tag{1}$$

When we assume that ϵ follows $N(0, \sigma^2)$, $\hat{\epsilon}/\hat{\sigma}$ follows a t-distribution with 196 degrees of freedom.

3.a Outputs of lm function

Remark:

ϵ and $\hat{\epsilon}$ have different distributions.

For example, imagine a linear regression

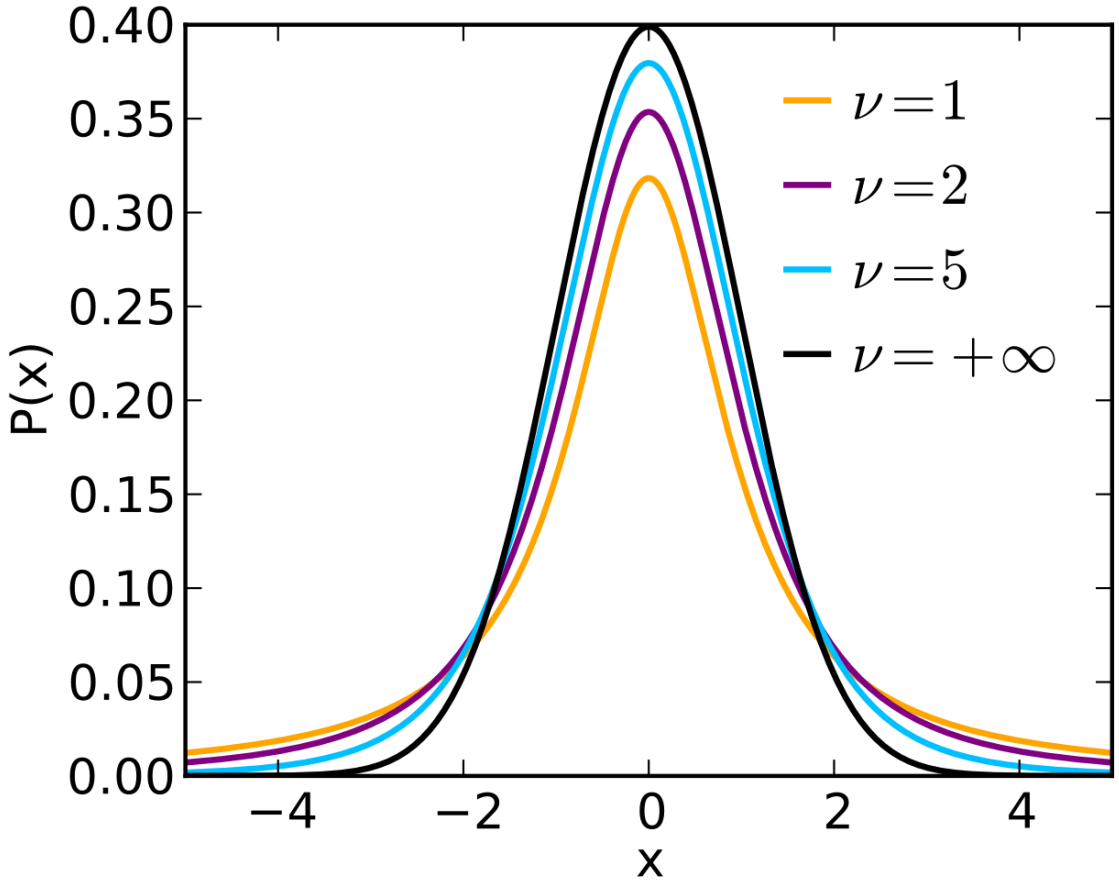
$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

with two observations $(x_1, y_1), (x_2, y_2)$.

Since we can always determine the line which goes through any two points, the residuals are always zero but errors are almost always non-zero.

3.a Outputs of lm function

The t-distribution is a symmetric distribution whose tails are fatter than a normal distribution. The figure shows the density of t-distribution for several different degrees of freedom. ***t***-distribution with ∞ degrees of freedom is a standard normal distribution.



3.a Outputs of lm function

R^2: Coefficient of determination

```
Multiple R-squared:  0.9167,      Adjusted R-squared:  0.9154
```

The R^2 is called coefficient of determinanation, and represents the ratio of sum of squared errors explained by predictors.
There are two versions. These are calculated as:

$$\textbf{Multiple } R^2 = 1 - \frac{SS_{Err}}{SS_{Tot}}$$

$$\textbf{Adjusted } R^2 = 1 - \frac{SS_{Err}/(n-p-1)}{SS_{Tot}/(n-1)}$$

where p is the number of predictors (except intercept) and

$$SS_{Err} := (Y - X\hat{\beta})'(Y - X\hat{\beta}) = \sum (y_i - x_i\hat{\beta})^2$$

$$SS_{Tot} := (Y - \bar{Y})'(Y - \bar{Y}) = \sum (y_i - \bar{y})^2,$$

3.a Outputs of lm function

where $\bar{\boldsymbol{Y}}$ is an n by 1 vector whose components are all $\bar{\boldsymbol{y}}$.

In short, the multiple $\boldsymbol{R^2}$ is just a ratio, while the adjusted $\boldsymbol{R^2}$ penalizes a large number of predictors for a fair comparison of models with a different number of predictors.

~~To be concrete, the expectation of the adjusted $\boldsymbol{R^2}$ is always zero when all predictors do not have any true relationship with $\boldsymbol{y}$.~~

$\boldsymbol{R^2}$ is always between 0 and 1, and tends to be larger when the number of predictors becomes large (even if there are no true relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$).

3.a Outputs of lm function

Coefficient of determination and correlation coefficient

When there is only one predictor $\boldsymbol{x}$, the coefficient of determination $\boldsymbol{R^2}$ is equal to the square of correlation coefficient $\boldsymbol{r}$ between $\boldsymbol{x}$ and $\boldsymbol{y}$.

```
> summary(lm(sales~radio))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
radio        0.20250    0.02041   9.921   <2e-16 ***
---
Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,        Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16

> cor(sales,radio)
[1] 0.5762226
> cor(sales,radio)^2
[1] 0.3320325
```

3.a Outputs of lm function

F-test

```
F-statistic:   719 on 3 and 196 DF,  p-value: < 2.2e-16
```

This is the result for a hypothesis testing:

- A null hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (i.e., the regression model is not useful at all), against
- an alternative hypothesis H_1: $\beta_j \neq 0$ for some j.

3.a Outputs of lm function

The test statistic F is given by

$$F = \frac{(SS_{Tot} - SS_{Err})/p}{SS_{Err}/(n-p-1)}. \tag{2}$$

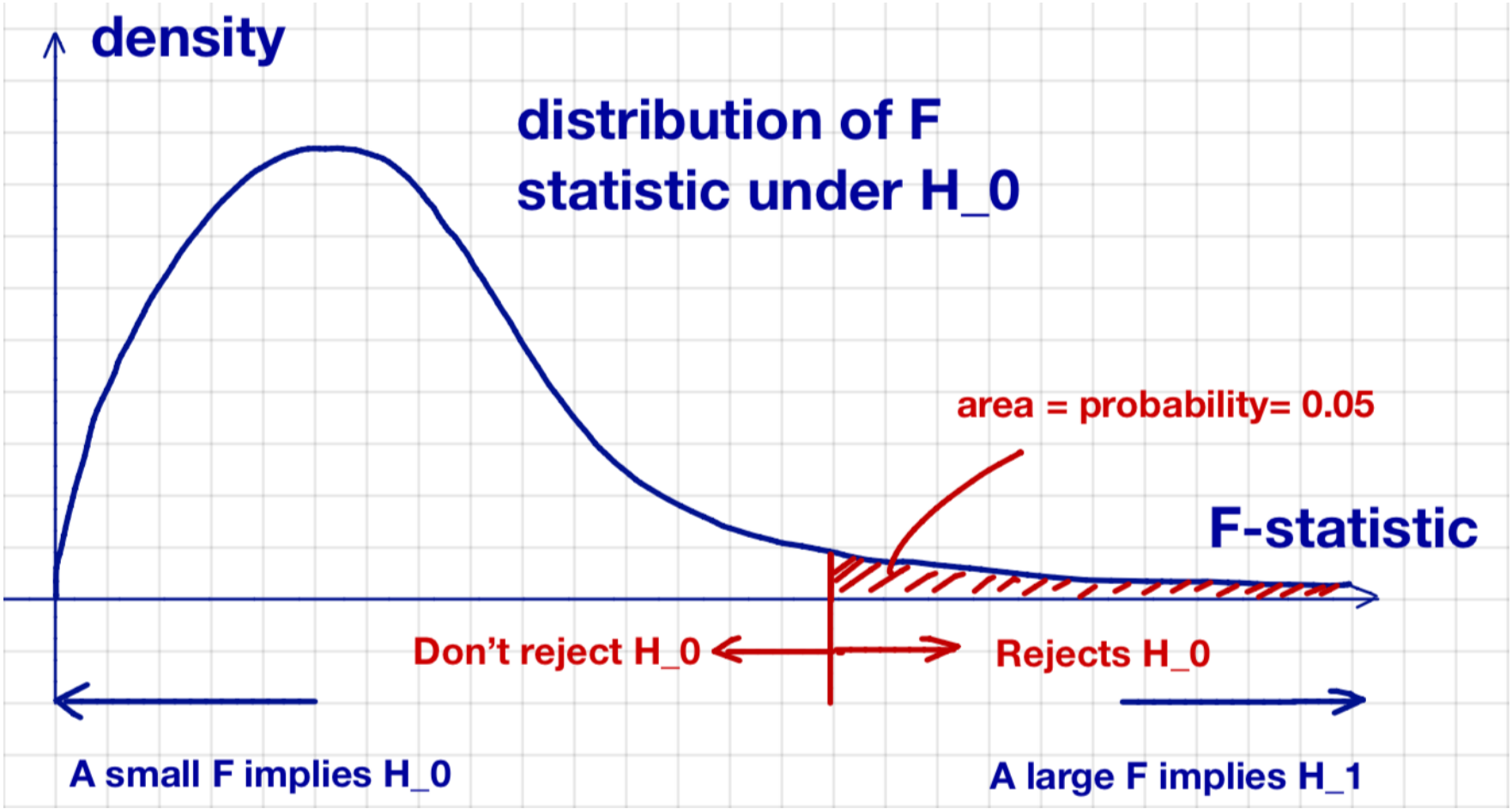
The value of F is larger when SS_{Err} gets smaller compared to SS_{Tot}.

Since SS_{Err} is the error size of regression model and SS_{Tot} is the total variability of y, this happens only if the regression model works good.

Therefore, we reject the null hypothesis when the value of F is larger than some threshold value.

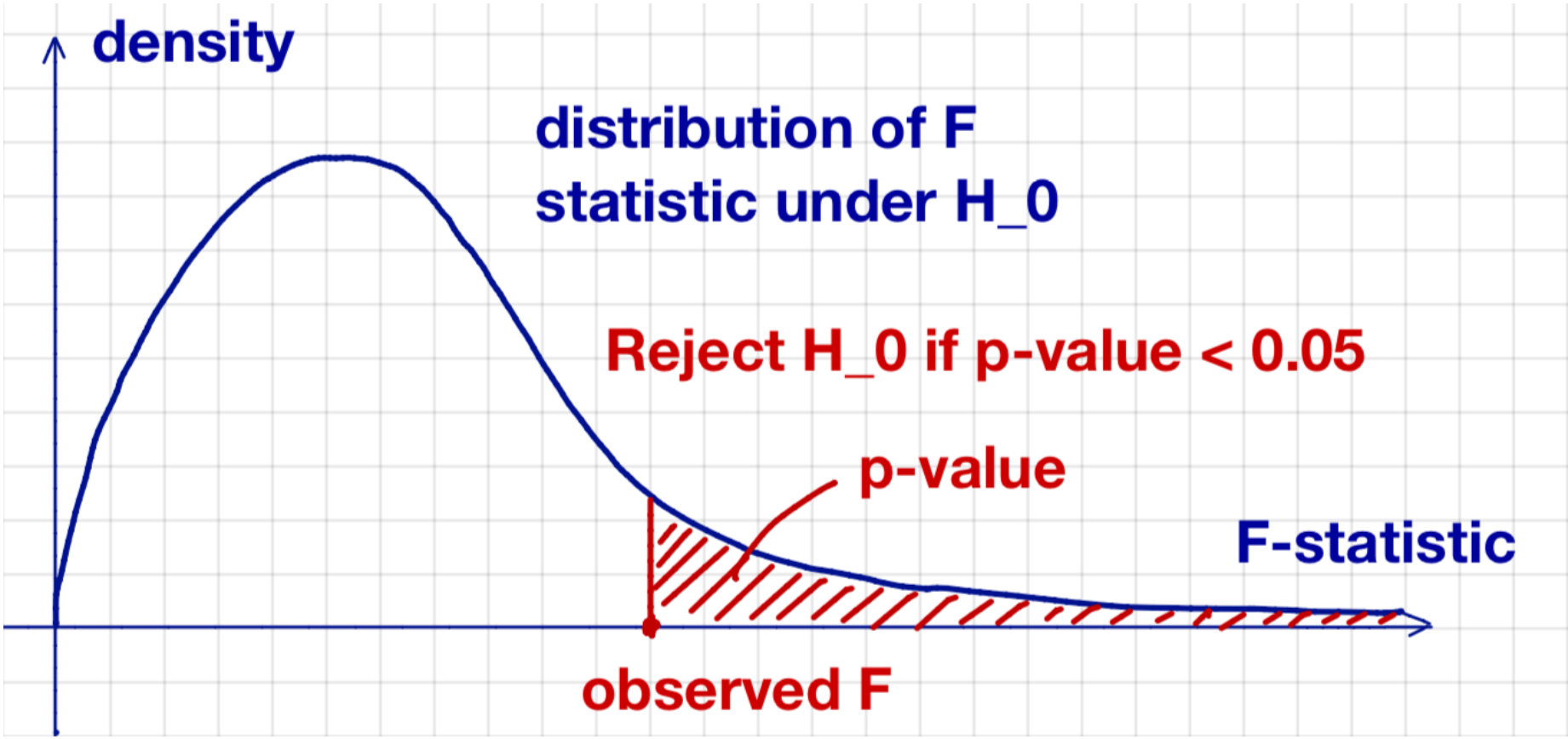
3.a Outputs of lm function

An illustration of ***F***-test:



3.a Outputs of lm function

An illustration for ***p***-value in ***F***-test:



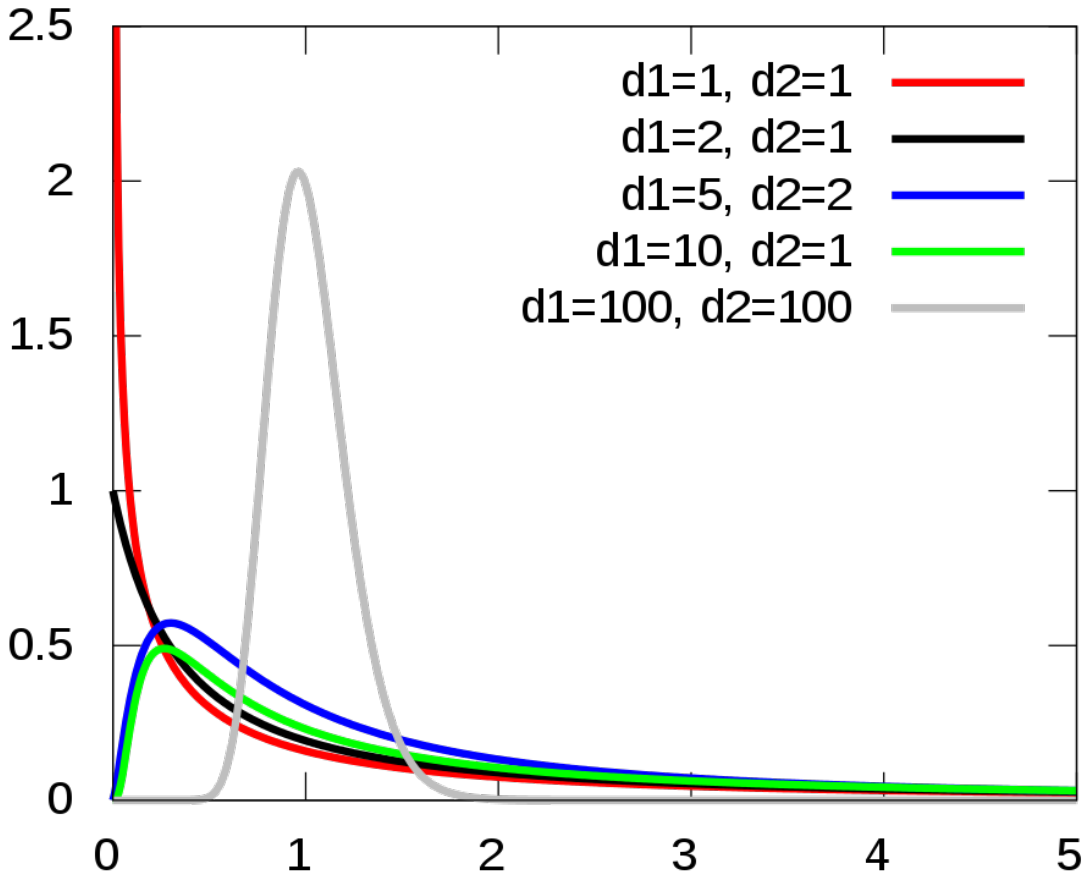
3.a Outputs of lm function

It is known that $\boldsymbol{F}$ follows an F-distribution with degrees of freedom $\boldsymbol{(p, n-p-1)}$ under $\boldsymbol{H_0}$.

An $\boldsymbol{F}$-distribution has two parameters ("numerator degrees of freedom" and and "denominator degrees of freedom"), and its shape depends on the two parameters.

3.a Outputs of lm function

Figure: A graph of F-distributions.



3.a Outputs of lm function

```
F-statistic:   719 on 3 and 196 DF,  p-value: < 2.2e-16
```

The R output indicates that $\boldsymbol{F}$-statistic is 196, and it follows an F distribution with $\boldsymbol{df = (3, 196)}$ under $\boldsymbol{H_0}$.

Since p-value $= \boldsymbol{2 \cdot 10^{-16} < 0.05}$, $\boldsymbol{H_0}$ is rejected.

The regression model is useful to explain $\boldsymbol{y}$-value.

3.b How to check assumptions of linear model

A linear model with least square estimation assumes

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \ IID \tag{3}$$

where IID means “independent and identically-distributed”.

There are seveal possible violation of assumptions:

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms (heteroskedasticity).
- Outliers.
- High-leverage points.
- Collinearity.

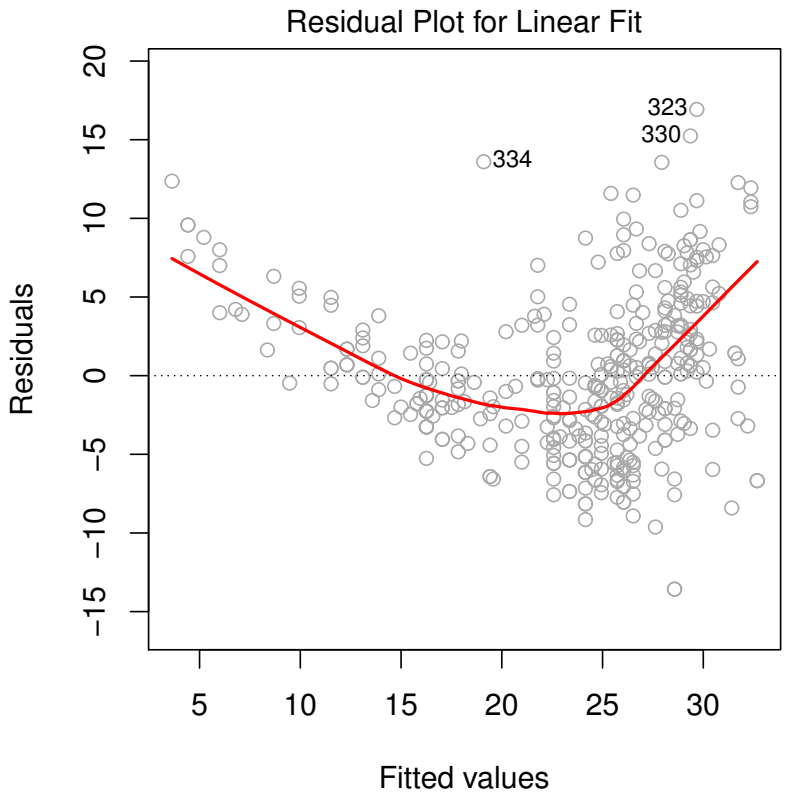
3.b How to check assumptions of linear model

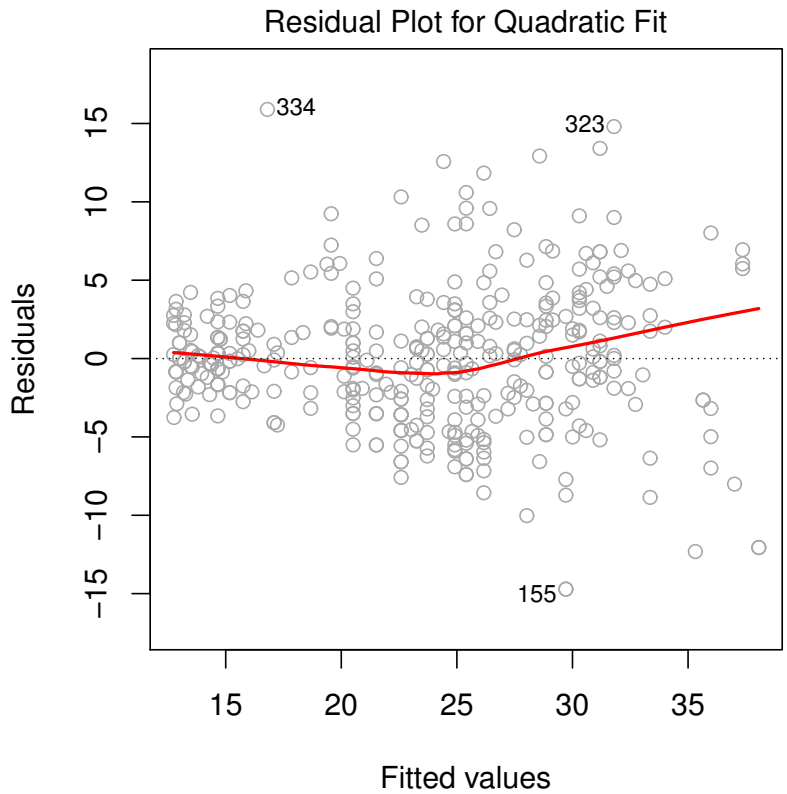
Non-linearity

When residual plots have some (non-linear) patterns in trend, it implies that more information can be extracted from $\boldsymbol{x}$ to explain $\boldsymbol{y}$. That is, there is non-linear relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.b How to check assumptions of linear model

Figure 3-9: Left: residual plots of $\boldsymbol{mpg} = \beta_0 + \beta_1 \boldsymbol{HP} + \epsilon$.
Right: residual plots of $\boldsymbol{mpg} = \beta_0 + \beta_1 \boldsymbol{HP} + \beta_2 \boldsymbol{HP}^2 + \epsilon$.





3.b How to check assumptions of linear model

Remedies for non-linearity

- Transform $\boldsymbol{x}$, $\boldsymbol{y}$, or both. Common functions used are $\boldsymbol{\log(x)}$, $\boldsymbol{x^{\alpha}}$ with some α.
- Add more predictors in the linear model. E.g., $\boldsymbol{y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon}$.

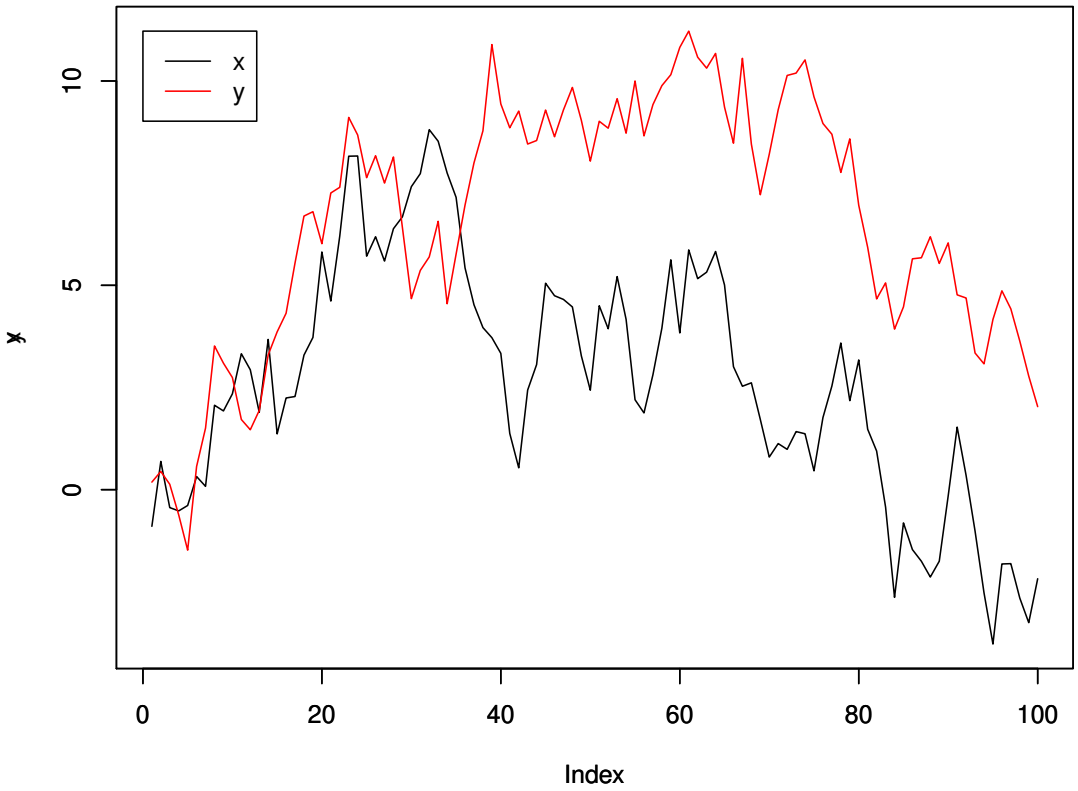
3.b How to check assumptions of linear model

Correlation of error terms

When ϵ is correlated each other, standard errors of parameter estimates are underestimated.

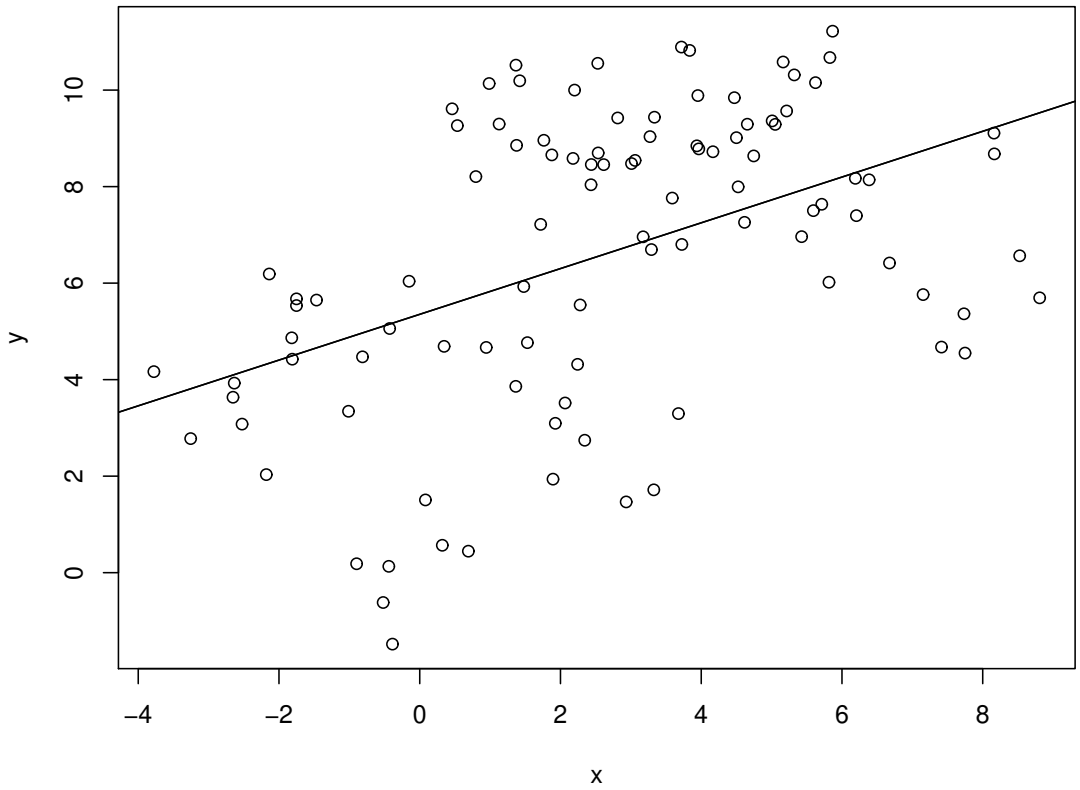
3.b How to check assumptions of linear model

Figure: Two series $\boldsymbol{x}$ and $\boldsymbol{y}$ are time series data with high auto-correlation (that is, adjacent observations such as $\boldsymbol{x_t}$ and $\boldsymbol{x_{t+1}}$ are highly correlated). Two series are similar by coincidence (actually two paths are independent random walks).



3.b How to check assumptions of linear model

Figure: However, (x_t, y_t) looks highly related when we ignore time structure.



3.b How to check assumptions of linear model

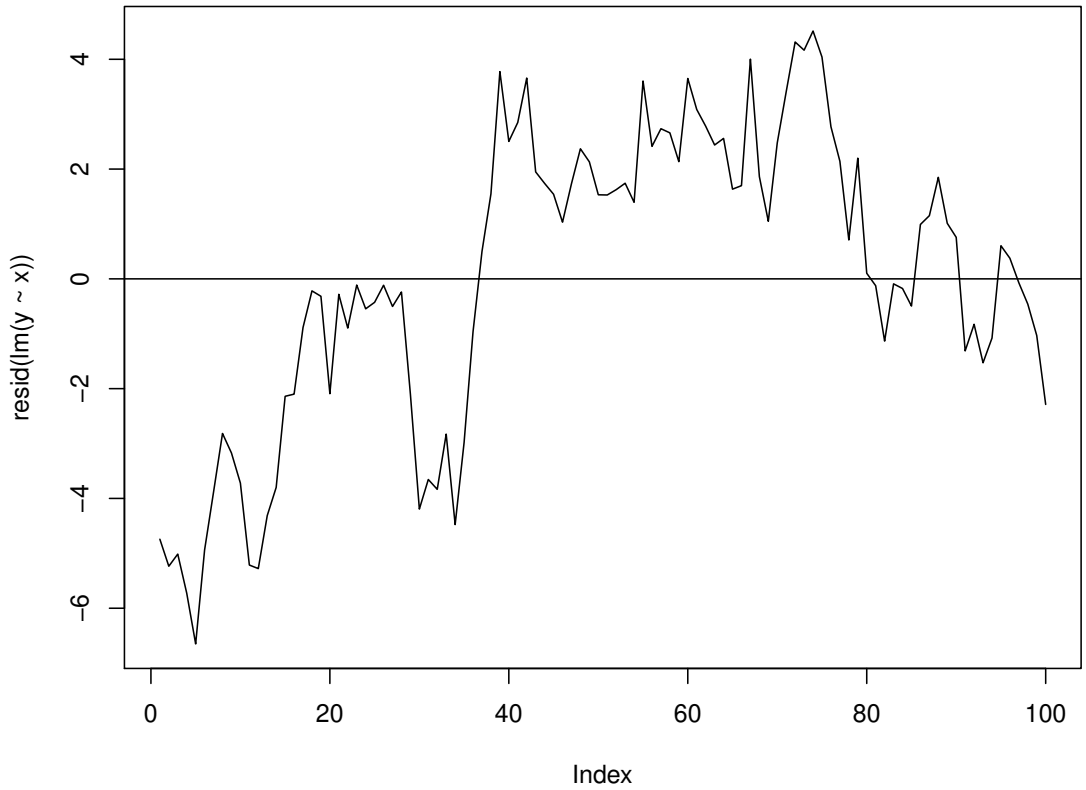
As a result, lm function returns a highly significant coefficient.

```
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.35331    0.36719  14.579  < 2e-16 ***
x            0.47407    0.09334   5.079 1.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.734 on 98 degrees of freedom
Multiple R-squared:  0.2084,        Adjusted R-squared:  0.2003
F-statistic: 25.79 on 1 and 98 DF,  p-value: 1.813e-06
```

3.b How to check assumptions of linear model

The problem: Independent assumption of ϵ is clearly violated.

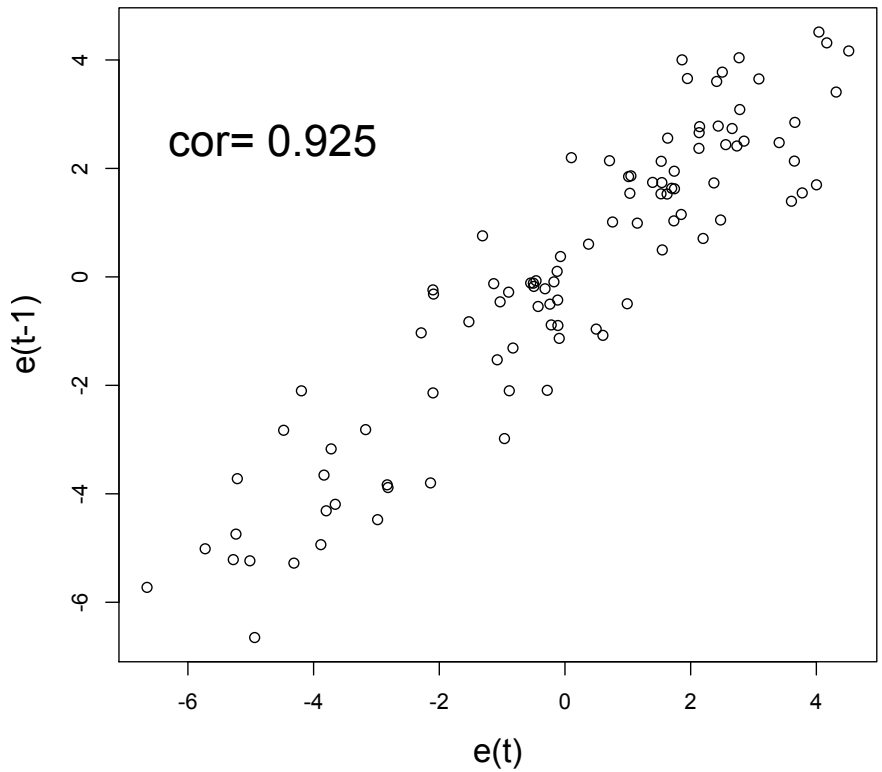
Figure: Residuals by time order



3.b How to check assumptions of linear model

Correlation between $\hat{\epsilon}_t$ and $\hat{\epsilon}_{t-1}$ is 0.925, very high.

```
> e <- resid(lm(y~x))
> plot(e[2:n], e[1:(n-1)],xlab="e(t)",ylab="e(t-1)",cex.lab=1.5) # (e(t),e(t-1))
> text(min(e)+2,max(e)-2, cex=2,
+         labels=paste("cor=",round(cor(e[2:n], e[1:(n-1)]),3))) # correlation
```



3.b How to check assumptions of linear model

Remedies for correlation of error terms

- Take the first difference $\Delta x_t := x_t - x_{t-1}$ of the time series (for both x_t and y_t). It often removes time correlation.
- Estimate parameters with accounting for correlation between errors (generalized least square, GLS).

3.b How to check assumptions of linear model

GLS gives more reasonable standard errors for parameters:

```
> library(nlme) # for gls
> GLS1 <- gls(y~x, correlation=corAR1())
> summary(GLS1)
Generalized least squares fit by REML
  Data: NULL
       AIC     BIC    logLik
  285.2752 295.615 -138.6376
 Parameter estimate(s):
      Phi
0.9808276
Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 3.752359  3.549629 1.057113  0.2931
x           0.133462  0.084324 1.582719  0.1167

 Correlation:
  (Intr)
x -0.012
Residual standard error: 4.966842
Degrees of freedom: 100 total; 98 residual
```

3.b How to check assumptions of linear model

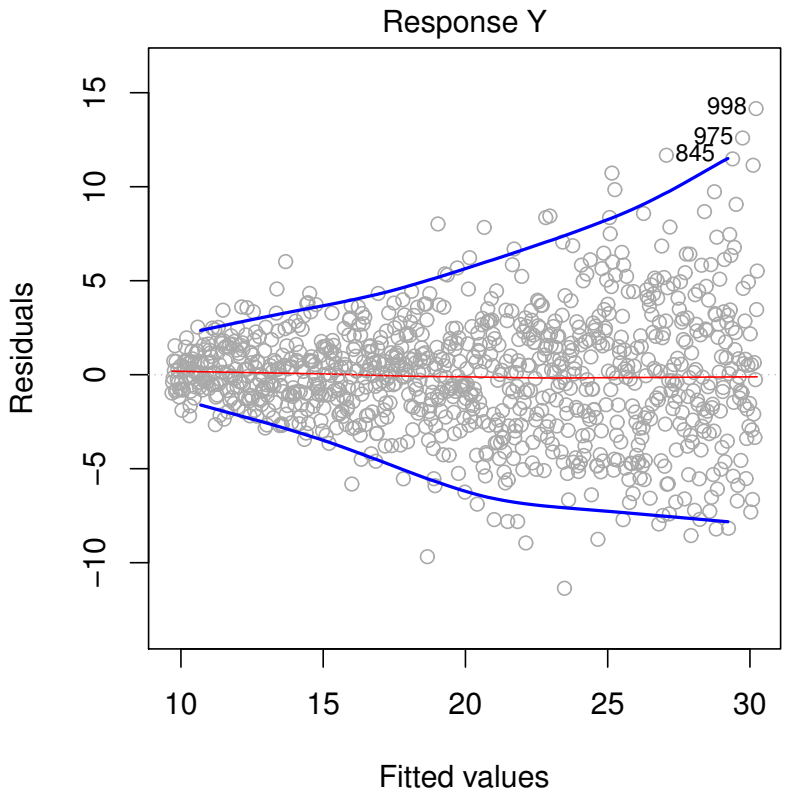
Non-constant variance of error terms

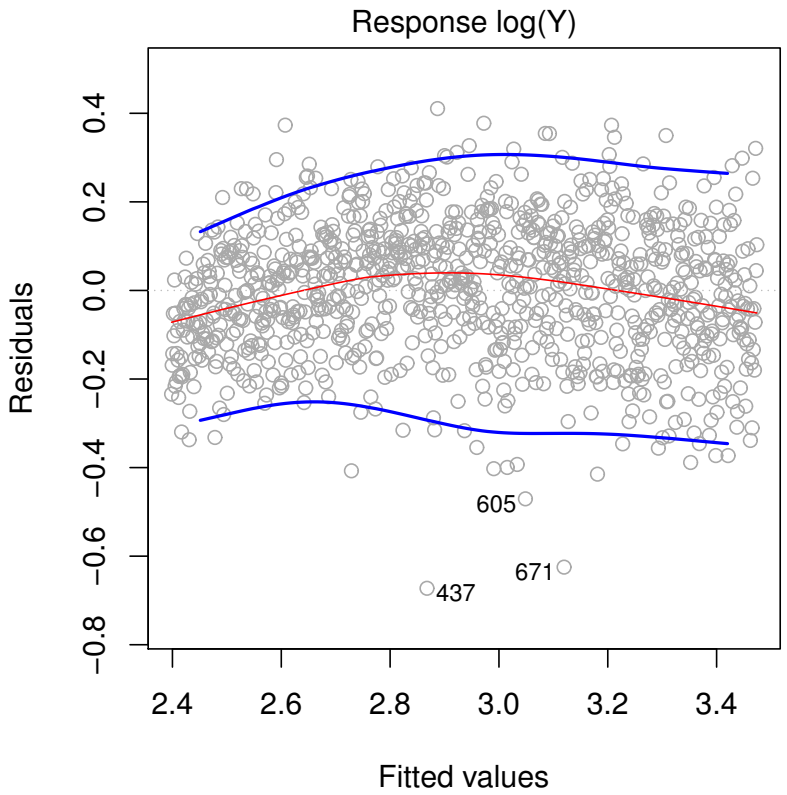
$\epsilon_1, \cdots, \epsilon_n$ may have different variances.

Namely, $\epsilon_i \sim N(0, \sigma_i^2)$ and σ_i depends on i.

3.b How to check assumptions of linear model

Example 1: (Left:) Residual plot indicates error size is larger for larger fitted values. (Right:) After transforming $\boldsymbol{Y}$ by logarithm, residual plot has equal variance.





3.b How to check assumptions of linear model

Example 2:

Suppose that you want to investigate relationship between height and weight of college students, and collected mean height and mean weight data from each of 100 colleges. There are 100 data points, but errors are not expected identially distributed. A small error in height-weight relationship is expected for colleges with a larger enrollment.

You may use a different estimation procedure such as weighted least square, WLS.

3.b How to check assumptions of linear model

Remedies for non-constant variance of error terms

- Transform the response variable $\boldsymbol{y}$. Common functions are $\mathbf{log}\ \boldsymbol{y}$, $\boldsymbol{y}^{\alpha}$ with some α.
- Use an appropriate estimation method such as weighted least square.
- For time series data with different error size by time period, use models such as GARCH.

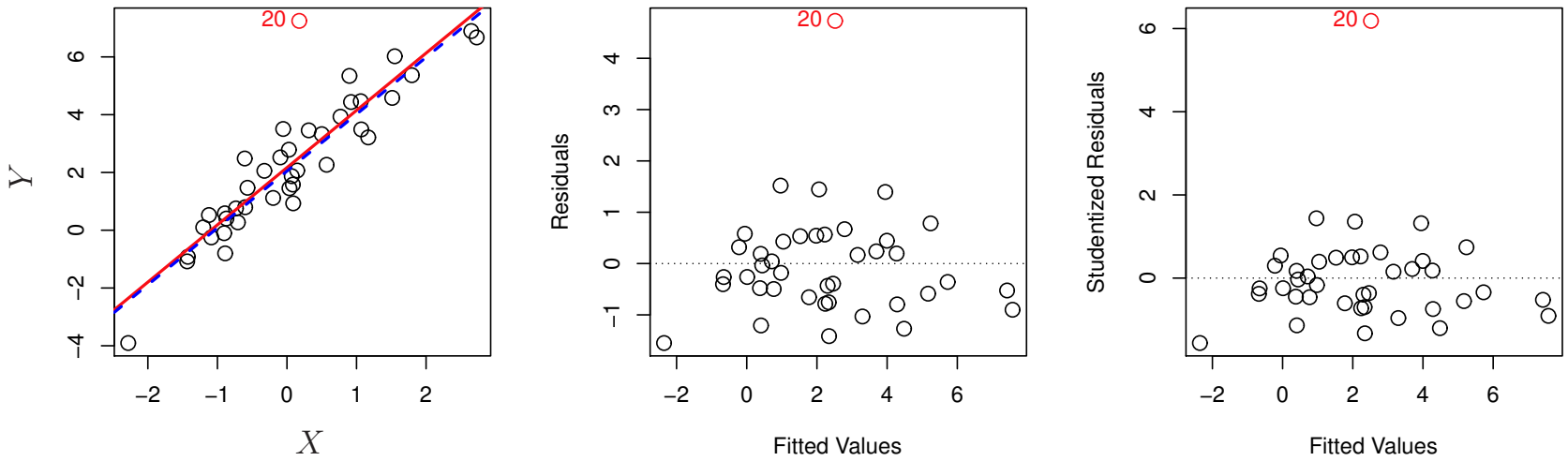
3.b How to check assumptions of linear model

Outliers

There are sometimes extremely large errors, called outliers. This violates normality assumtpion of errors. Also this makes estimates more unreliable.

3.b How to check assumptions of linear model

Example: The 20th observations is an outlier.



Note: Standardized residual adjusts the size of errors by its expected size of errors. A residual tends to be larger when it is closer to the center in predictor value.

3.b How to check assumptions of linear model

Remedies for outliers

- Check if any recording errors in data collection.
- Check if the estimation is robust. Does the result change if the outlier(s) are dropped?
- Use a more robust estimation method (e.g., quantile regression, M-estimation).

3.b How to check assumptions of linear model

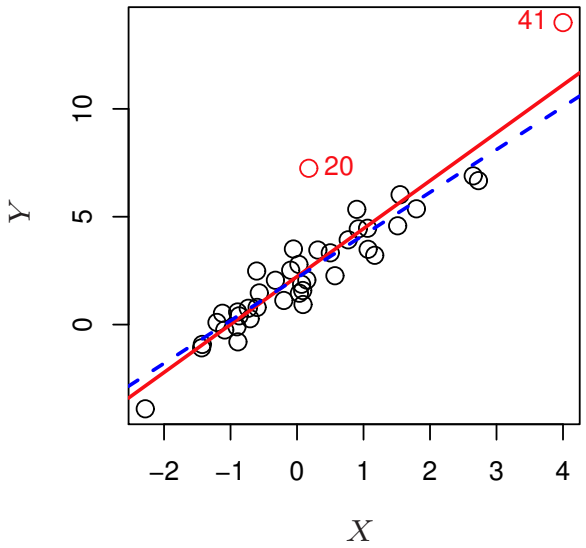
High Leverage points

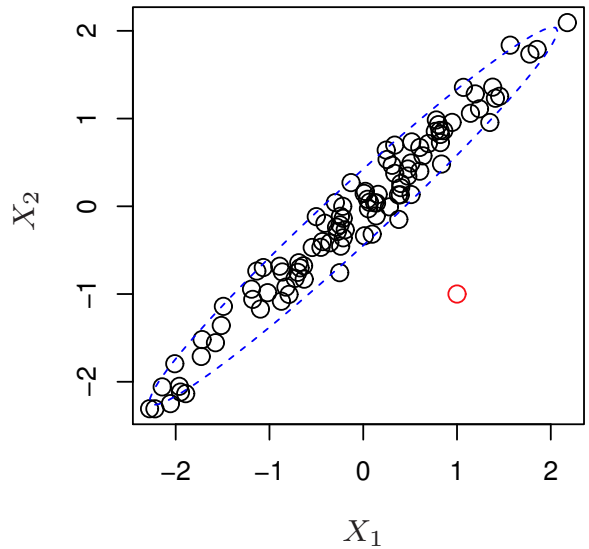
When the $\boldsymbol{x}$ value of an observation is far away from the center, it may have too much influence on the regression line.

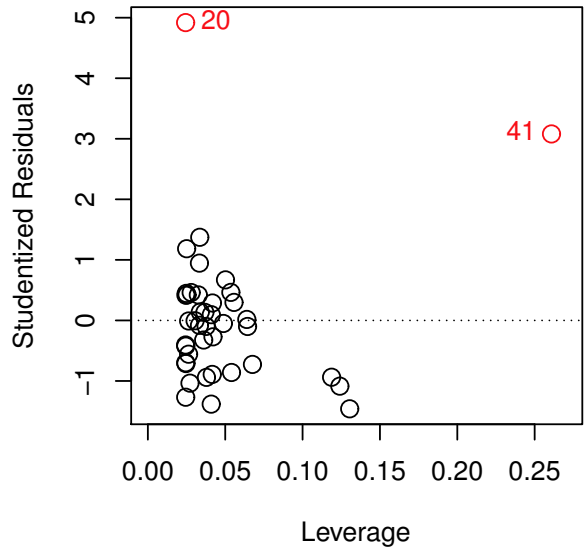
This is not good since a few observations almost determines the regression result. Often such observations may have larger errors than the other observations.

3.b How to check assumptions of linear model

Figure: The 41st observation is influential point (left and right). An influential point may not be extreme in any specific predictor (center).







3.b How to check assumptions of linear model

The leverage statistic of an observation (x_i, y_i) is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

for a simple linear regression. (The formula for multiple regressions is different.)

h_i is between $1/n$ and 1, and the mean leverage is $(p+1)/n$ where p is the number of predictors.

3.b How to check assumptions of linear model

Remedies for high leverage

- Transformation of $\boldsymbol{x}$ (or both $\boldsymbol{x}$ and $\boldsymbol{y}$) may make the point less influential.
- Check if the estimation is robust. Does the result change if the outlier(s) are dropped?
- Separate influential points from the others, and use another model for them.

3.b How to check assumptions of linear model

Collinearity

When some predictors $x_1, \cdots, x_q$ are nearly linearly dependent, we say there is collinearity in predictors.

This fact inflates the standard errors of estimated coefficients, making more predictors look less significant.

3.b How to check assumptions of linear model

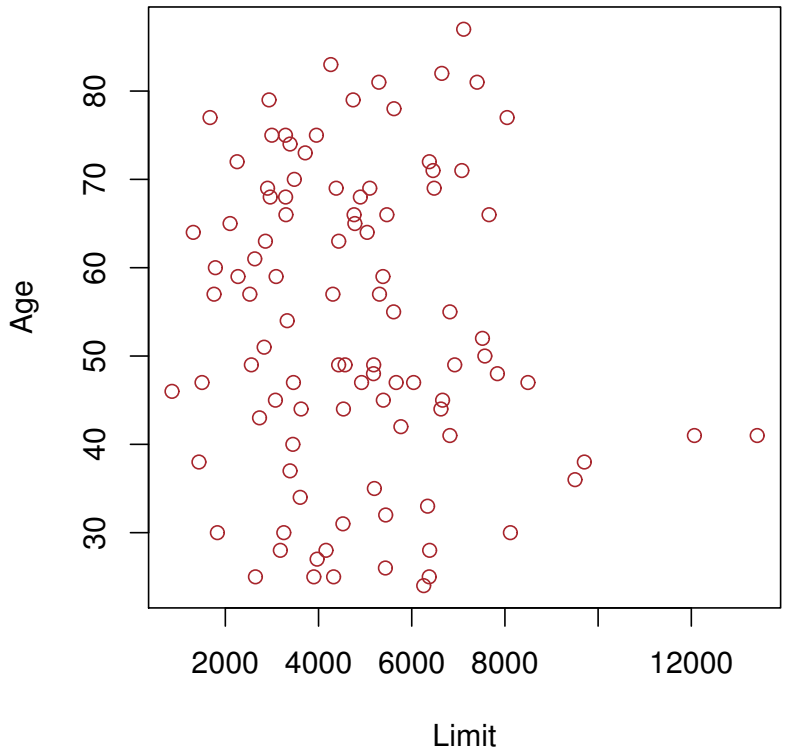
Example: Credit Data

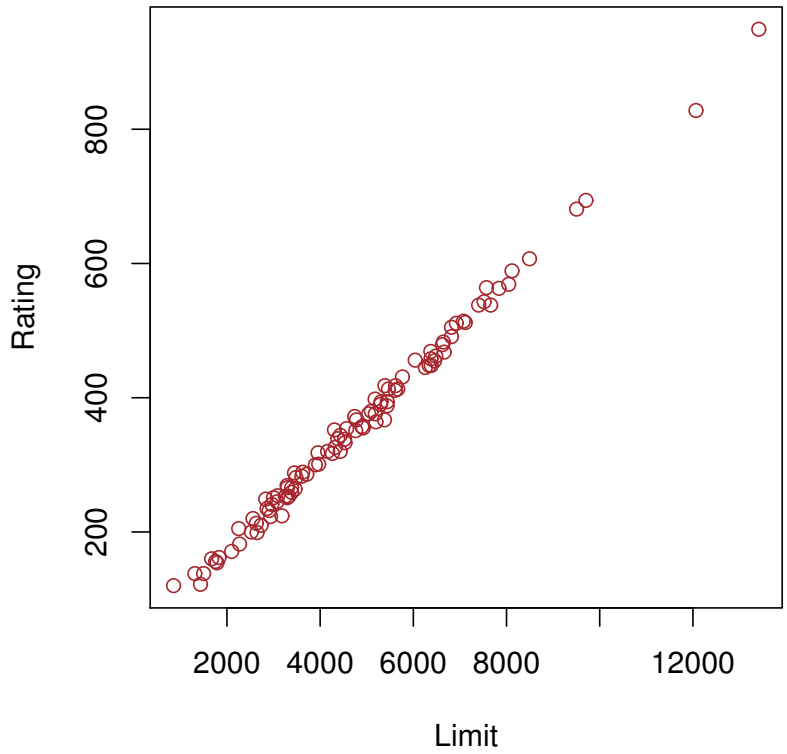
Want to regress `'balance'` on

- `'age'` and `'limit'`, where the two variable are almost independent.
- `'rating'` and `'limit'`, where the two variable are highly dependent.

3.b How to check assumptions of linear model

Figure: Credit Data





3.b How to check assumptions of linear model

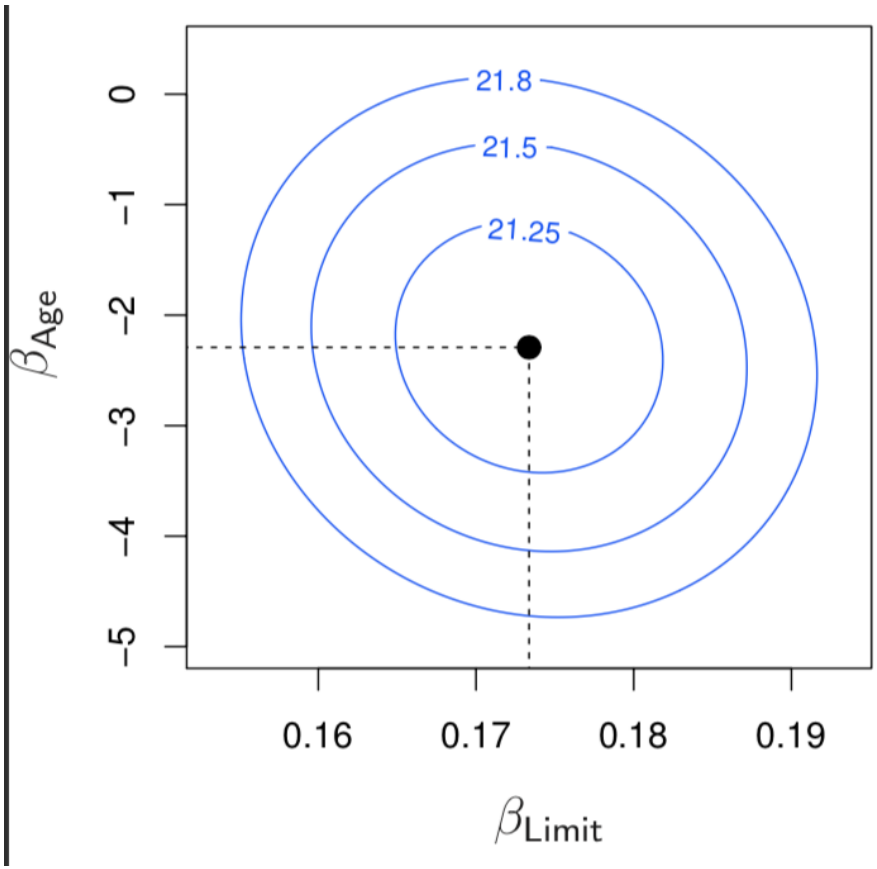
Table: Estimation results. Since ‘`rating`’ and ‘`limit`’ are highly correlated, it is hard to distinguish the effects from the two variables. Consequently, standard errors of the coefficients are larger in Model 2.

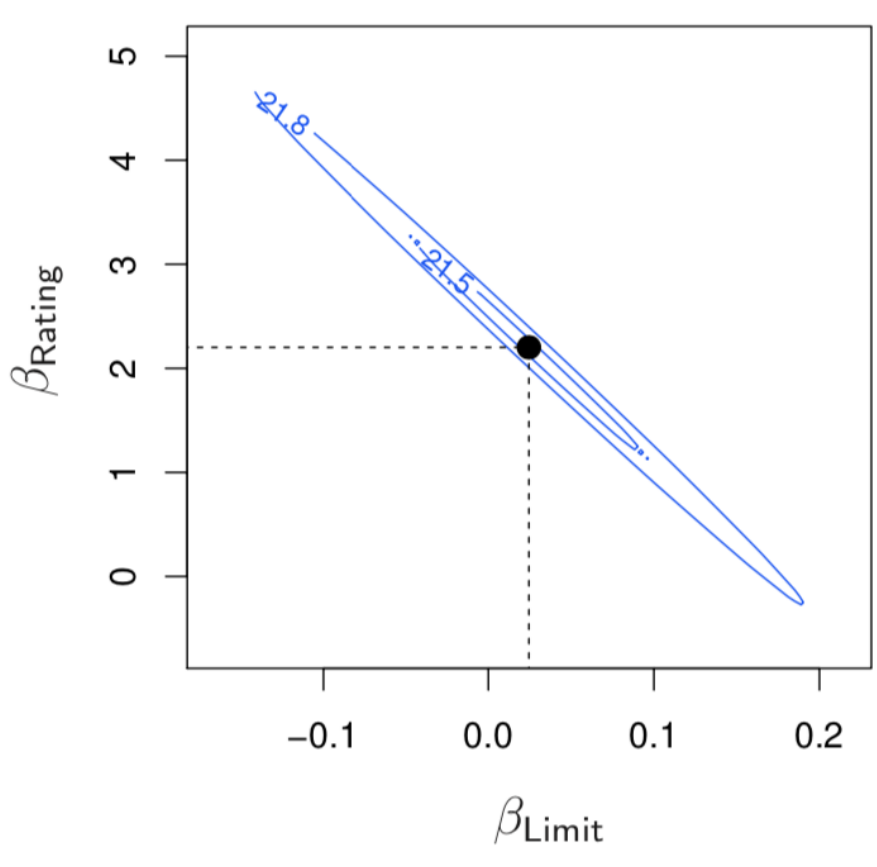
		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

3.b How to check assumptions of linear model

Figure: Confidence regions (2-dimensional version of confidence interval) of the two regression models above. (Left:) Model 1. Confidence interval for β_{Limit} is very narrow (consider project the contour on to x-axis). (Right:) Model 2. Confidence intervals for β_{Limit} is very wide, while the confidence region is still small in area.





3.b How to check assumptions of linear model

The Variance Inflation Factor (VIF) represents how many times $\mathbf{Var}\hat{\beta}_j$ gets larger (relative to squared residual standard error) when the model changes from $\boldsymbol{y} = \beta_0 + \beta_j \boldsymbol{X}_j + \epsilon$ to $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \cdots + \beta_p \boldsymbol{X}_p + \epsilon$.

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - \boldsymbol{R}^2_{\boldsymbol{X}_j | \boldsymbol{X}_{-j}}}$$

where $\boldsymbol{R}^2_{\boldsymbol{X}_j | \boldsymbol{X}_{-j}}$ is the multiple $\boldsymbol{R}^2$ when $\boldsymbol{X}_j$ is regressed on the other predictors.

When VIF is more than 10, multicollinearity is regarded as severe.

3.b How to check assumptions of linear model

Example of VIF:

Regress `balance` on `Limit` and two other predictors, and evaluate the standard error of the coefficient of `Limit`.

```
> summary(lm(Balance ~ Limit))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.928e+02  2.668e+01  -10.97   <2e-16 ***
Limit        1.716e-01  5.066e-03   33.88   <2e-16 ***
---
Residual standard error: 233.6 on 398 degrees of freedom

> summary(lm(Balance ~ Age + Rating + Limit))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -259.51752   55.88219  -4.644 4.66e-06 ***
Age           -2.34575    0.66861  -3.508 0.000503 ***
Rating         2.31046    0.93953   2.459 0.014352 *
Limit          0.01901    0.06296   0.302 0.762830
---
Residual standard error: 229.1 on 396 degrees of freedom
```

3.b How to check assumptions of linear model

The standard error for `Limit` inflated from 0.005066 (simple regression) to 0.06296 (multiple regression). Relative to residual standard error, the ratio is

$$\frac{0.06296/229.1}{0.005066/233.6} = 12.67.$$

If we consider the ratio of variances, $12.67^2 = 160.58$. This is obtained by

```
> library(car) # for vif
> vif(lm(Balance ~ Age + Rating + Limit)) # Rounding error made 160.59 be 160.58.
      Age     Rating      Limit
 1.011385 160.668301 160.592880
```

3.b How to check assumptions of linear model

The VIF can also be calculated directly by definition.

```
> R2 <- summary(lm(Limit ~ Age + Rating))$r.squared  #$
> 1/(1-R2) # VIF
[1] 160.5929
```

3.b How to check assumptions of linear model

Remedies for collinearity

- Ignore the issue if you only focus on prediction.
- Exclude similar predictors beforehand.
- Combine similar predictors beforehand (e.g., principal component analysis, state space models).
- Use a variable selection method which drops more variables (e.g., Lasso, ridge regression).

3.c Qualitative (categorical) predictors

Example: Property Sales and the number of bedrooms

Consider a forecasting problem for sales in the Property Sales Data in Troy, MI. Sales may depend on the number of bedrooms. The number of bedrooms is either 0,2,3,4,5 or 6, so it can be regarded as either quantitative or qualitative variables.

```
> str(DPS)
'data.frame':        490 obs. of  7 variables:
 $ Date : int  40535 40486 40452 40263 40459 40399 40479 40451 40283 40420 ...
 $ Price: int  140000 90000 113000 96900 230000 232000 200000 189000 235000 415000 .
 $ Style: Factor w/ 5 levels "BI-LEVEL","BUNGALOW",..: 3 4 4 4 3 3 3 3 3 3 ...
 $ Sqft : int  2117 1064 960 960 2494 2321 2235 2269 2631 3718 ...
 $ Year : int  1952 1950 1972 1976 1987 1988 1984 1974 1994 2004 ...
 $ Bed  : int  4 3 3 3 4 3 4 4 4 4 ...
 $ Bath : int  2 1 1 1 2 2 2 2 2 3 ...
> table(DPS$Bed) #
  0   2   3   4   5   6
  5   8 235 235   6   1
```

3.c Qualitative (categorical) predictors

Consider the following models:

$$\textbf{Price} = \beta_0 + \beta_1 \textbf{Sqft} + \beta_2 \textbf{Year} + \beta_3 \textbf{Bath} + \beta_4 \textbf{Bed} + \epsilon \quad (4)$$

$$\textbf{Price} = \beta_0 + \beta_1 \textbf{Sqft} + \beta_2 \textbf{Year} + \beta_3 \textbf{Bath} + \beta_4 \textbf{Bed} + \beta_{4,i} + \epsilon_{ij} \quad (5)$$

where

- ***Bed*** is a quantitative variable, and
- $\beta_{4,i}$ represents a factor (dummy variable) for $\boldsymbol{i}$ bedrooms ($\boldsymbol{i = 2, 3, 4, 5, 6}$) as 0 bedroom as a baseline. (Note: One of the 5 factors is redundant because there are 6 parameters ($\beta_4, \beta_{4,2}, \cdots, \beta_{4,6}$) to represent 5 levels other than the baseline.)
- The first model assumes that the price change for one additional bedroom is constant, while the second model assumes a non-linear relationship between price and the number of bedrooms.

3.c Qualitative (categorical) predictors

Sample R code:

```
> LM.l <- lm(Price ~ Sqft+Year+Bath+Bed) # model (3)
> Bed.f <- as.factor(Bed) # Number of Bedroom as factor
> LM.f <- lm(Price ~ Sqft+Year+Bath+Bed+Bed.f) # model (4)
```

3.c Qualitative (categorical) predictors

Results for the model (4):

```
> summary(LM.l)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.546e+06  2.475e+05  -6.246 9.20e-10 ***
Sqft         1.094e+02  4.071e+00  26.865  < 2e-16 ***
Year         7.780e+02  1.268e+02   6.138 1.74e-09 ***
Bath         1.043e+04  3.949e+03   2.641  0.00854 **
Bed         -8.837e+03  3.168e+03  -2.790  0.00548 **
...
Residual standard error: 40640 on 485 degrees of freedom
Multiple R-squared: 0.8259,        Adjusted R-squared: 0.8245
F-statistic: 575.2 on 4 and 485 DF,  p-value: < 2.2e-16
```

The number of bedrooms is a significant factor (p-value= **0.00548**).

3.c Qualitative (categorical) predictors

Results for the model (5):

```
> summary(LM.f)
...
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1699196.6   251265.9  -6.763 3.94e-11 ***
Sqft              111.8        4.2  26.613  < 2e-16 ***
Year              833.7      126.7   6.581 1.22e-10 ***
Bath             9902.4     3907.2   2.534  0.01158 *
Bed             -31086.1     7275.5  -4.273 2.33e-05 ***
Bed.f2           83608.0    23314.3   3.586  0.00037 ***
Bed.f3          109130.7    22260.2   4.903 1.30e-06 ***
Bed.f4          125143.1    27276.6   4.588 5.72e-06 ***
Bed.f5          168758.0    37235.0   4.532 7.37e-06 ***
Bed.f6                NA         NA      NA       NA
...
Residual standard error: 39730 on 481 degrees of freedom
Multiple R-squared: 0.835,        Adjusted R-squared: 0.8322
F-statistic: 304.2 on 8 and 481 DF,  p-value: < 2.2e-16
```

3.c Qualitative (categorical) predictors

The above result indicates:
when we introduce a paremeter for each level of ***Bed***, many of them are significant predictors of ***Price***.

3.c Qualitative (categorical) predictors

The following ANOVA table can test the overall significance of the factor:

```
> anova(LM.f)
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq   F value    Pr(>F)
Sqft        1 3.7140e+12 3.7140e+12 2352.7231 < 2.2e-16 ***
Year        1 6.0205e+10 6.0205e+10   38.1382 1.404e-09 ***
Bath        1 1.2500e+10 1.2500e+10    7.9184  0.005094 **
Bed         1 1.2853e+10 1.2853e+10    8.1424  0.004511 **
Bed.f       4 4.1659e+10 1.0415e+10    6.5976 3.562e-05 ***
Residuals 481 7.5930e+11 1.5786e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table reads: after accounting for ***Sqft***, ***Year***, ***Bath*** and ***Bed*** as continuous variables, ***Bed.f*** is a sigfinicant factor in the model (p-value $= \mathbf{3.562 \times 10^{-5}}$).

3.c Qualitative (categorical) predictors

AIC also indicates the model (5) is a better prediction model.

```
> AIC(LM.l) # model (3)
[1] 11797.75

> AIC(LM.f) # model (4)
[1] 11779.58
```

Note:
The same analysis can be done for the number of bathrooms. In fact, the number of bathrooms as a factor is also a significant predictor of price. (The result is not shown in slides.)

3.d Removing additive assumption

Removing additive assumption

Suppose we want to predict Y by X_1 and X_2. A basic multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

This model assumes that X_1 and X_2 do not have interaction. A model with an interaction term is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

3.d Removing additive assumption

Example 1: Sales and advertising

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

3.d Removing additive assumption

Other examples:

- Crop yield may be affected by the amount of fertilizer only when sunlight is not enough.
- Southpaws pitch well only when the batter is left-handed.

3.e KNN and linear regression

Remember K-Nearest neighborhood (KNN) method, which use K closest observations to determine the predicted value at x.

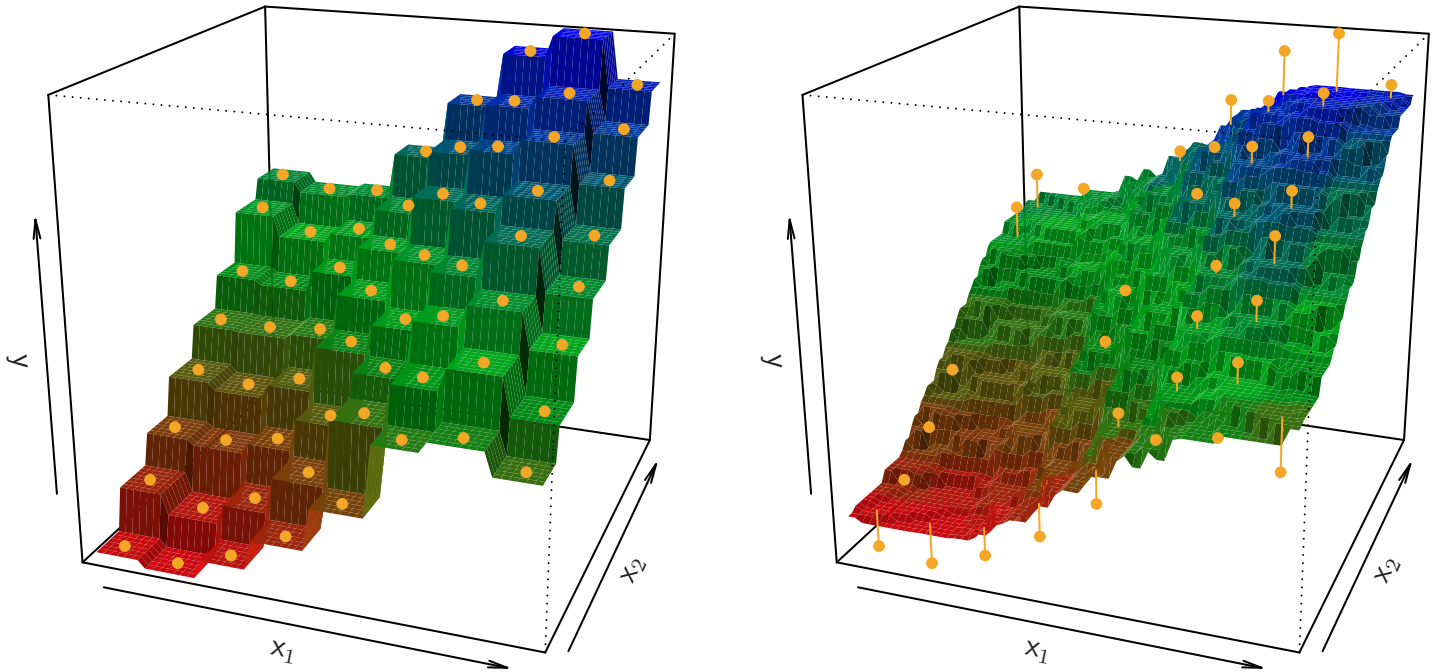
While it is used for categorization problem in Chapter 2, it can also be used for regression problem. To be exact, define the fitted value at x_0 by

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_0 \in N_0} y_i$$

where N_0 is a set of the K closest points to x_0 within the data set.

3.e KNN and linear regression

Figure illustrates KNN in regression problems with two predictors x_1 and x_2 (left: $K = 1$; right: $K = 9$).



KNN is more flexible in the shape of $\hat{f}$ than linear regression, while it lacks features such as significant of regression coefficients.

Memo