Statistical Modeling
BANA 7042

Lecture 1: A review of linear regression models

Topics to be reviewed

- Continuous distributions
- The modeling idea of linear regression
- Estimation (Least squares and maximum likelihood approaches)
- Hypothesis testing ($\beta = 0$)
- Goodness-of-fit (how well the model fits the data)

The following two will be reviewed along with our study of logistic Reg.

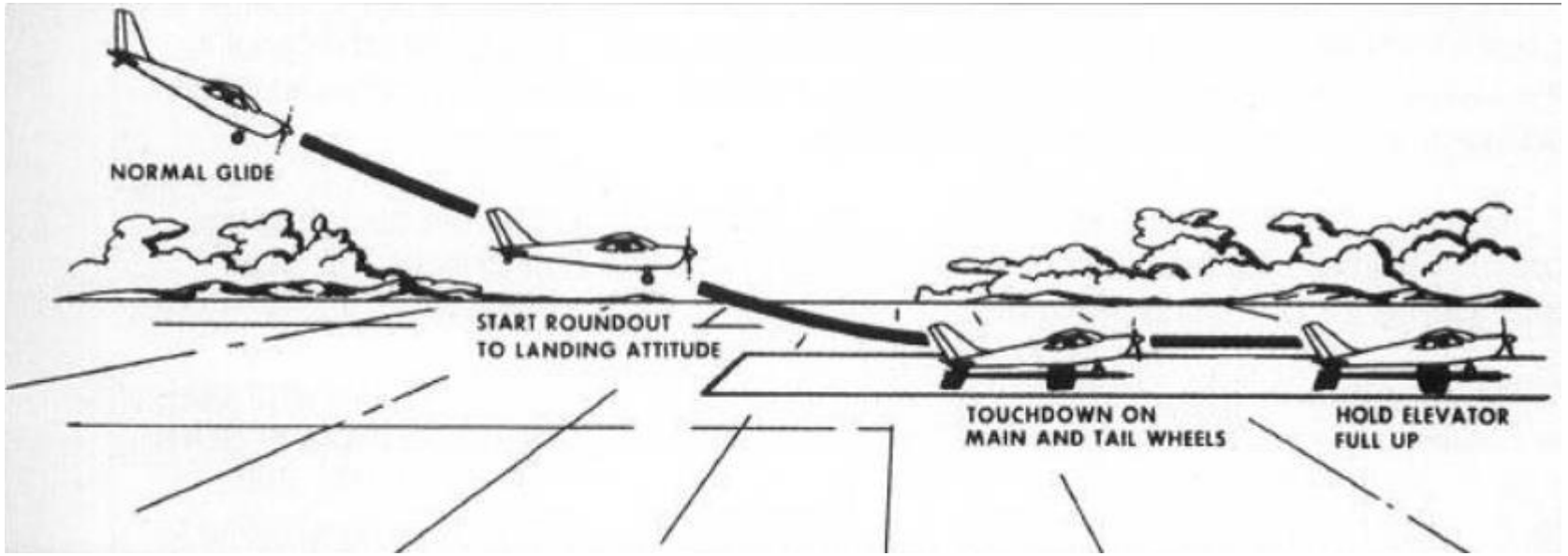
- Diagnostics (check model assumptions)
- Variable selection

Let us work on a real FAA project

Project description

- **Background**: Flight landing.
- **Motivation**: To reduce the risk of landing overrun.
- **Goal**: To study what factors and how they would impact the landing distance of a commercial flight.
- **Data**: Landing data (landing distance and other parameters) from 950 commercial flights. See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights). ttt

Landing procedure



Landing overrun

Variable dictionary

- **Aircraft**: The make of an aircraft (Boeing or Airbus).
- **Duration** (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.
- **No_pasg**: The number of passengers in a flight.
- **Speed_ground** (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
- **Speed_air** (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
- **Height** (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.
- **Pitch** (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.
- **Distance** (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Take a look at the data

```
> FAA<-read.csv("E:/FAA1.csv",header=T)
> head(FAA)
  aircraft  duration no_pasg speed_ground speed_air   height    pitch distance
1   boeing  98.47909      53    107.91568  109.3284 27.41892 4.043515 3369.836
2   boeing 125.73330      69    101.65559  102.8514 27.80472 4.117432 2987.804
3   boeing 112.01700      61     71.05196        NA 18.58939 4.434043 1144.922
4   boeing 196.82569      56     85.81333        NA 30.74460 3.884236 1664.218
5   boeing  90.09538      70     59.88853        NA 32.39769 4.026096 1050.264
6   boeing 137.59582      55     75.01434        NA 41.21496 4.203853 1627.068
```

Characteristics of the variables

```
> str(FAA)
'data.frame':   800 obs. of  8 variables:
 $ aircraft    : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
 $ no_pasg     : int  53 69 61 56 70 55 54 57 61 56 ...
 $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
 $ speed_air   : num  109 103 NA NA NA ...
 $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
 $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
 $ distance    : num  3370 2988 1145 1664 1050 ...
```

Summary of each variable

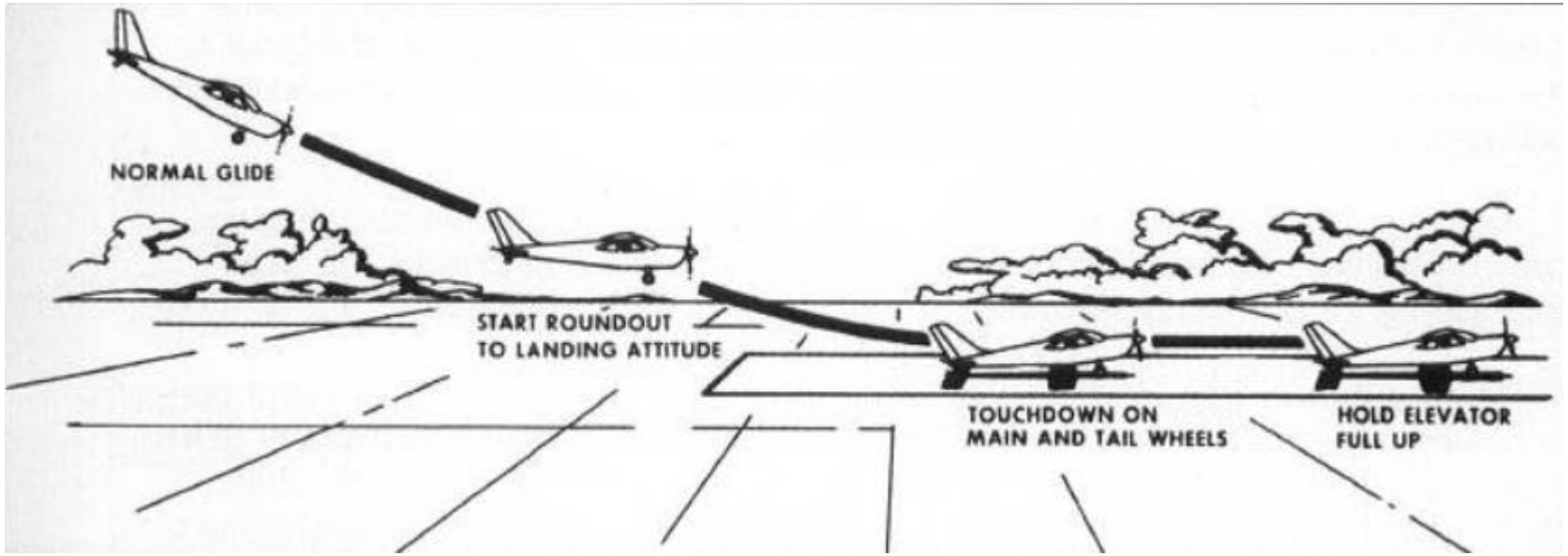
```
> summary(FAA)
   aircraft       duration         no_pasg       speed_ground
 airbus:400   Min.   : 14.76   Min.   :29.00   Min.   : 27.74
 boeing:400   1st Qu.:119.49   1st Qu.:55.00   1st Qu.: 65.87
              Median :153.95   Median :60.00   Median : 79.64
              Mean   :154.01   Mean   :60.13   Mean   : 79.54
              3rd Qu.:188.91   3rd Qu.:65.00   3rd Qu.: 92.33
              Max.   :305.62   Max.   :87.00   Max.   :141.22

   speed_air          height           pitch          distance
 Min.   : 90.00   Min.   :-3.546   Min.   :2.284   Min.   :  34.08
 1st Qu.: 96.16   1st Qu.:23.338   1st Qu.:3.658   1st Qu.: 900.95
 Median :100.99   Median :30.147   Median :4.020   Median :1267.44
 Mean   :103.83   Mean   :30.122   Mean   :4.018   Mean   :1544.52
 3rd Qu.:109.48   3rd Qu.:36.981   3rd Qu.:4.388   3rd Qu.:1960.44
 Max.   :141.72   Max.   :59.946   Max.   :5.927   Max.   :6533.05
 NA's   :600
```

Keep these numbers in mind

```
> summary(FAA)
   aircraft       duration         no_pasg       speed_ground
 airbus:400   Min.   : 14.76   Min.   :29.00   Min.   : 27.74
 boeing:400   1st Qu.:119.49   1st Qu.:55.00   1st Qu.: 65.87
              Median :153.95   Median :60.00   Median : 79.64
              Mean   :154.01   Mean   :60.13   Mean   : 79.54
              3rd Qu.:188.91   3rd Qu.:65.00   3rd Qu.: 92.33
              Max.   :305.62   Max.   :87.00   Max.   :141.22

   speed_air         height           pitch          distance
 Min.   : 90.00   Min.   :-3.546   Min.   :2.284   Min.   :  34.08
 1st Qu.: 96.16   1st Qu.:23.338   1st Qu.:3.658   1st Qu.: 900.95
 Median :100.99   Median :30.147   Median :4.020   Median :1267.44
 Mean   :103.83   Mean   :30.122   Mean   :4.018   Mean   :1544.52
 3rd Qu.:109.48   3rd Qu.:36.981   3rd Qu.:4.388   3rd Qu.:1960.44
 Max.   :141.72   Max.   :59.946   Max.   :5.927   Max.   :6533.05
 NA's   :600
```

Landing distance (LD)

- The variable of our central interest



How to study the LD?

- Question: Is it useful to test LD=4000 feet?
- Answer: No. Because LD is not a constant, but a **random** variable.

- We are interested in answering questions such as
  - What is the chance that LD>4000 feet? (higher risk of landing overrun)
  - What is the chance that LD<500 feet? (higher risk of undershoot)
  - What is the chance that 1000<LD<2000 feet? (safer landing)

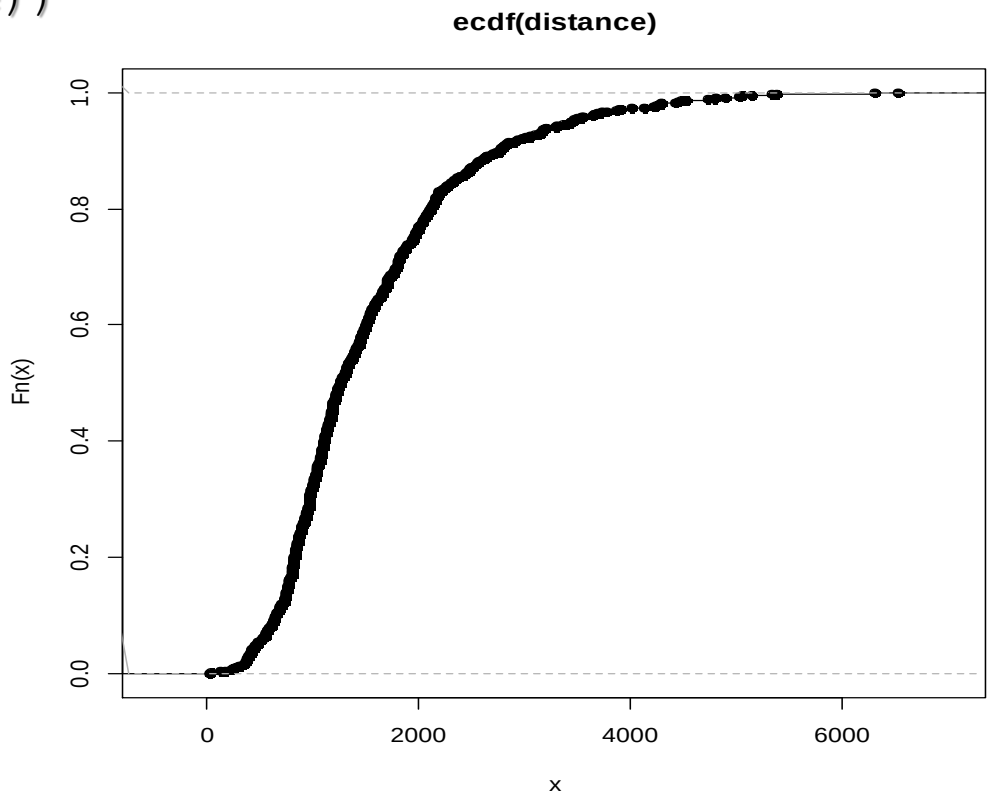
- In other words, we need to study the **distribution** of LD.

How to study the distribution of a continuous variable?

- A nonparametric approach
  - Calculate $F(x) = \Pr(LD \leq x)$ for **every** x in the range of LD.
  - Such a function $F(x)$ is called the **cumulative distribution function** (CDF).

- Questions:
  1. How to calculate $F(4000) = \Pr(LD < 4000)$ using R?
  2. How many values we need to report in order to fully describe the distribution of LD?

```
> CDF<-ecdf(FAA$distance)
> summary(ecdf(distance))
Empirical CDF:     800 unique values with summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.08  900.95 1267.44 1544.52 1960.44 6533.05
> plot(ecdf(distance))
```



How to study the distribution of a continuous variable?

- A parametric approach
  - Make a parametric assumption: The LD follows a normal distribution.

$$LD \sim N(\mu, \sigma^2)$$

  - The distribution involves two parameters only:
    - The mean μ and the variance σ^2

- Questions:
  1. How to estimate μ and σ^2 using R?
  2. How many values we need to report in order to fully describe the distribution of LD?

Principle 1: Dimension reduction

- How to study the **distribution** of LD?
  - A nonparametric approach
    - Calculate $F(x) = \Pr(LD \leq x)$ for **every** x in the range of LD.
    - Dimension = sample size.
  - A parametric approach
    - Make an assumption. For example: $LD \sim N(\mu, \sigma^2)$
    - Estimate two parameters only
    - Dimension = 2.

- A **compromise** between simplicity and precision.

No free lunch!

- How to check the assumption $LD \sim N(\mu, \sigma^2)$?

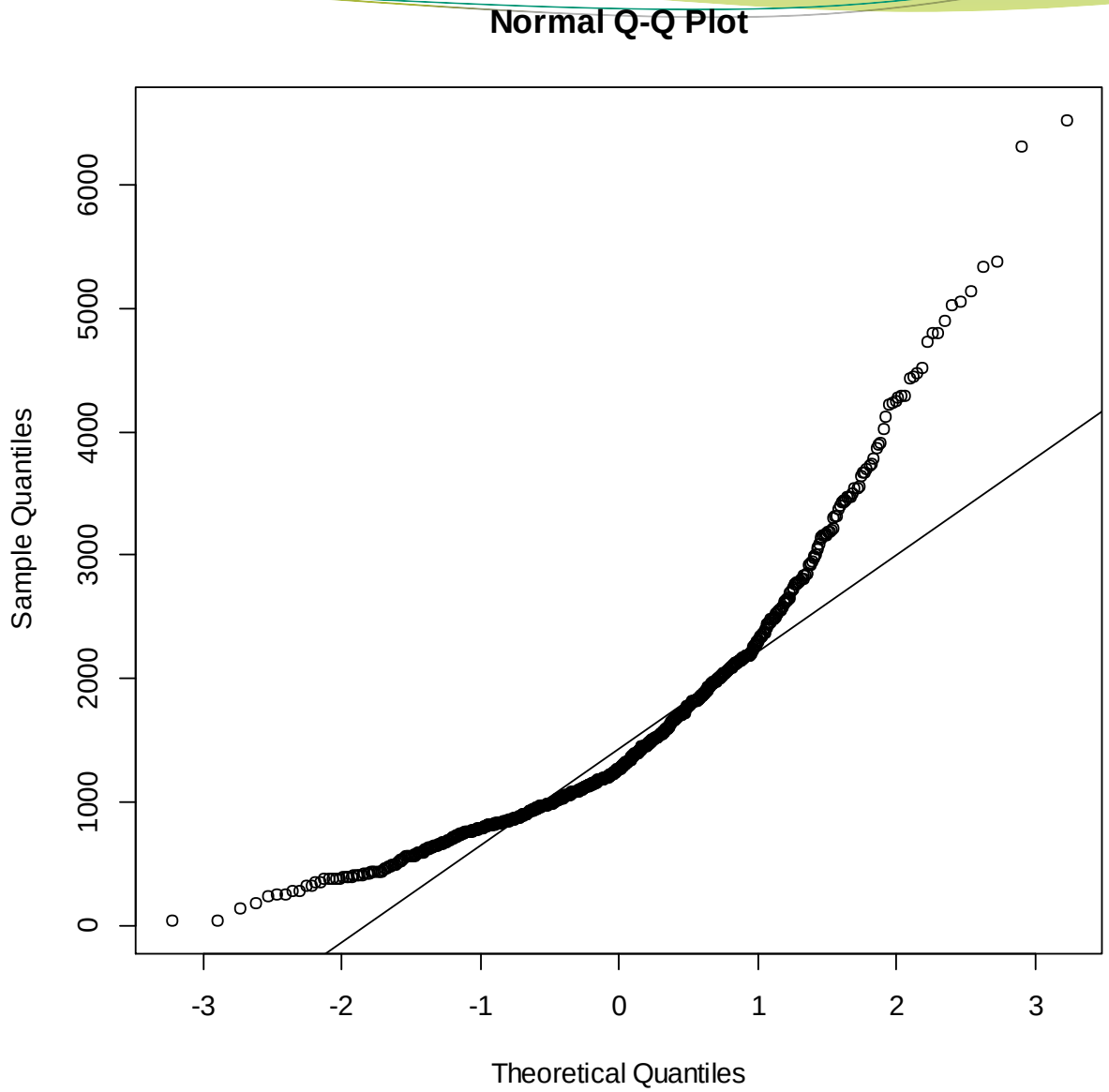
- If the assumption does not hold, what would you do?

No free lunch!

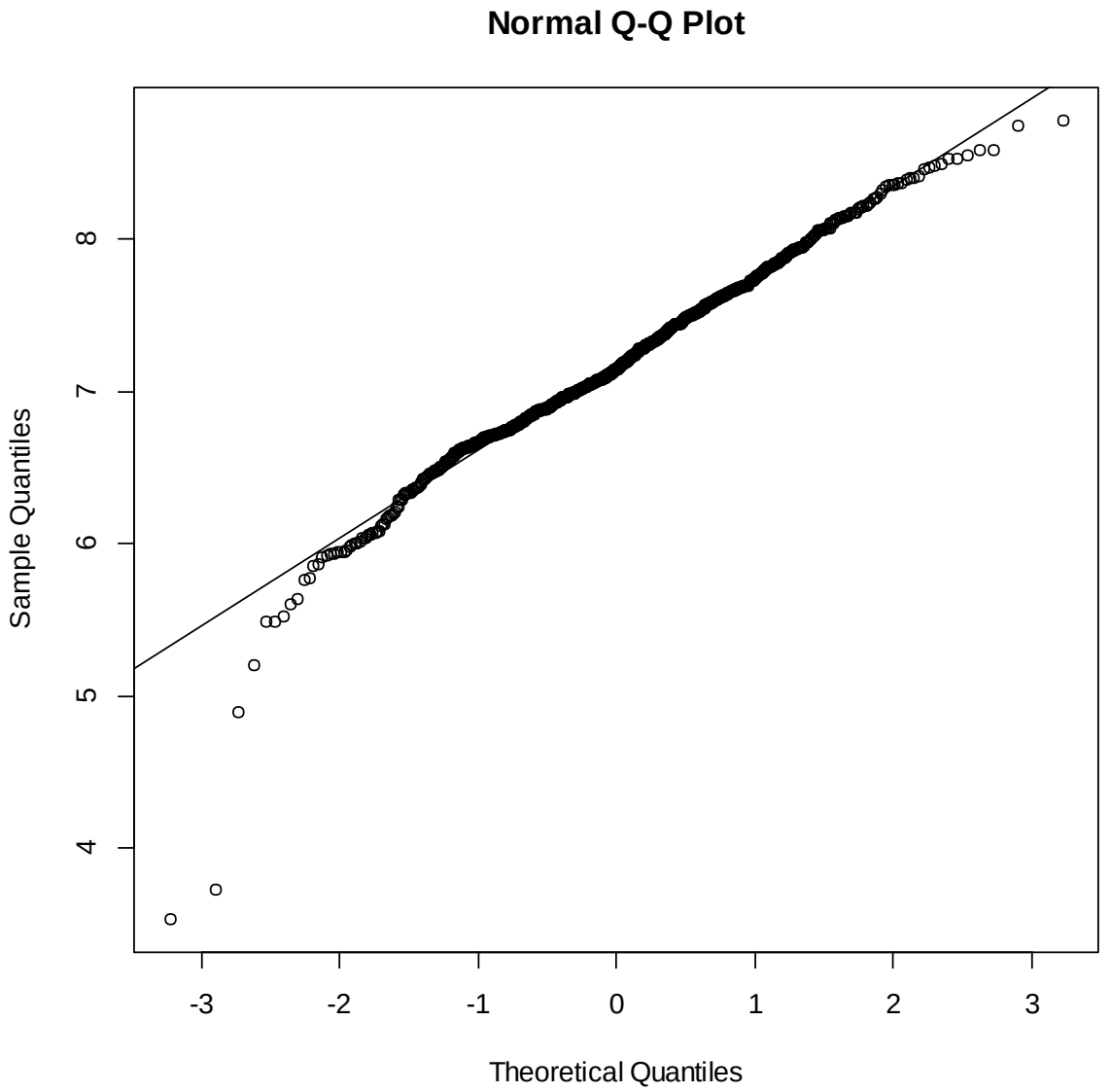
- How to check the assumption $LD \sim N(\mu, \sigma^2)$?
  - Visualization: Quantile-Quantile (QQ) plot.
  - Testing: Kolmogorov-Smirnov test.

- If the assumption does not hold, what would you do?
  - Transformation (e.g., logarithm)
  - Consider more flexible models (e.g., t / Gamma distribution)

```
> qqnorm(FAA$distance)
> qqline(FAA$distance)
```



```
> qqnorm(log(FAA$distance))
> qqline(log(FAA$distance))
```



Associate the response to covariates

Suppose the normality assumption holds

$$LD \sim N(\mu, \sigma^2)$$

How can we associate an explanatory variable (or predictor, covariate..) X to the response variable $Y = LD$?

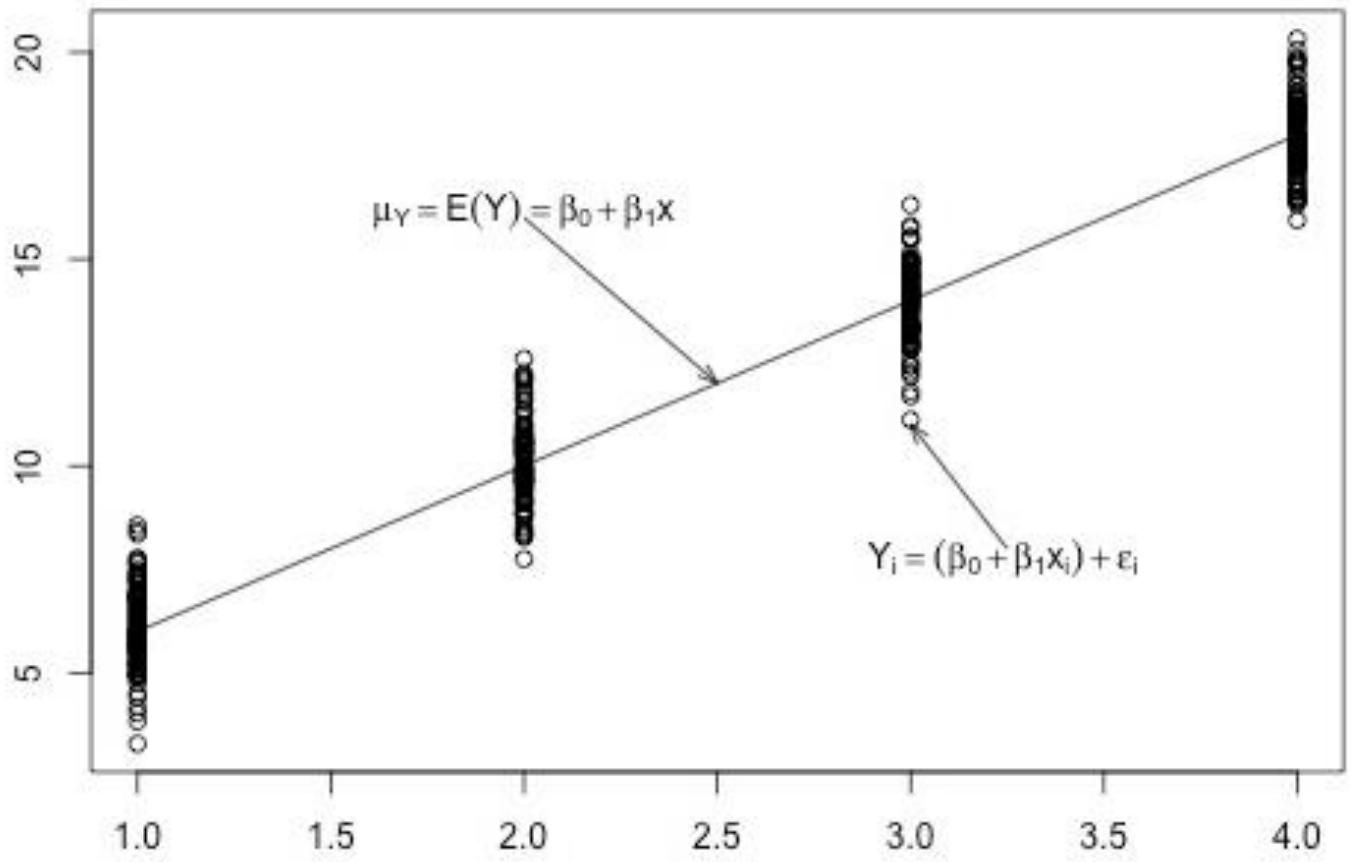
Establish association through the mean

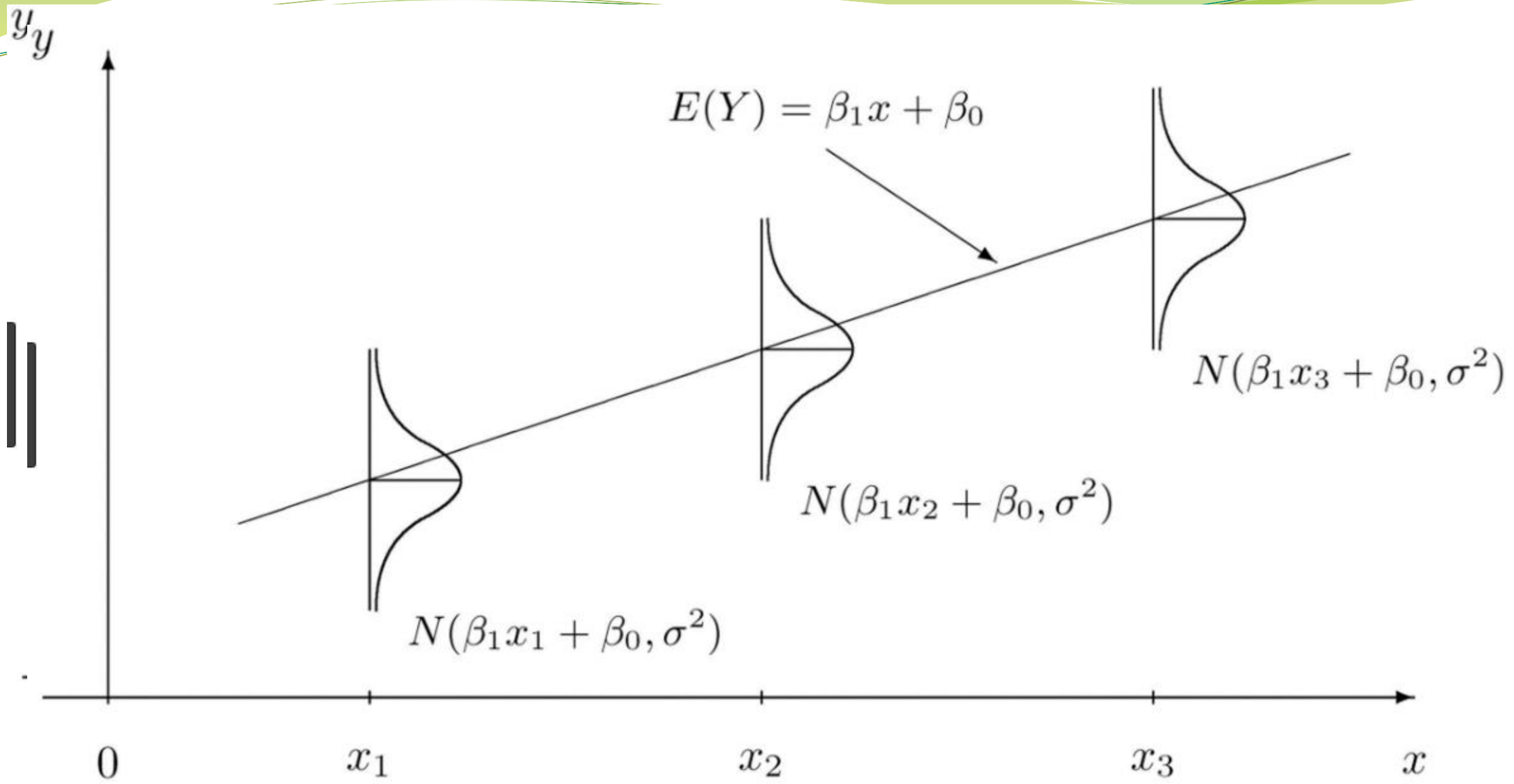
- We will make two additional assumptions on top of

$$LD \sim N(\mu, \sigma^2)$$

1. LD relies on X only through its mean, e.g, $LD \sim N(\mu(X), \sigma^2)$.
2. Linearity: $\mu(X) = \beta_0 + X\beta_1$

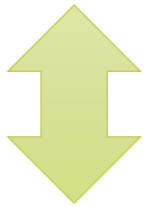
- In other words, $LD \sim N(\beta_0 + X\beta_1, \sigma^2)$


20
15
10
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
$\mu_Y = E(Y) = \beta_0 + \beta_1 x$
$Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$


y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

It is a linear regression!

$$LD \sim N(\beta_0 + X\beta_1, \sigma^2)$$

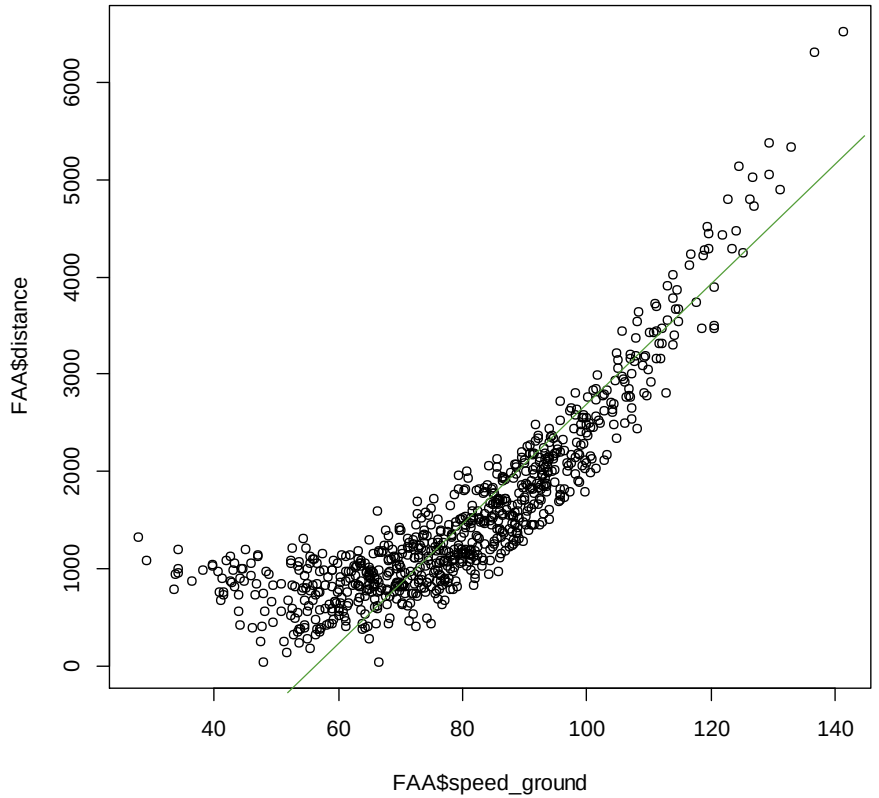


$$LD = \beta_0 + X\beta_1 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

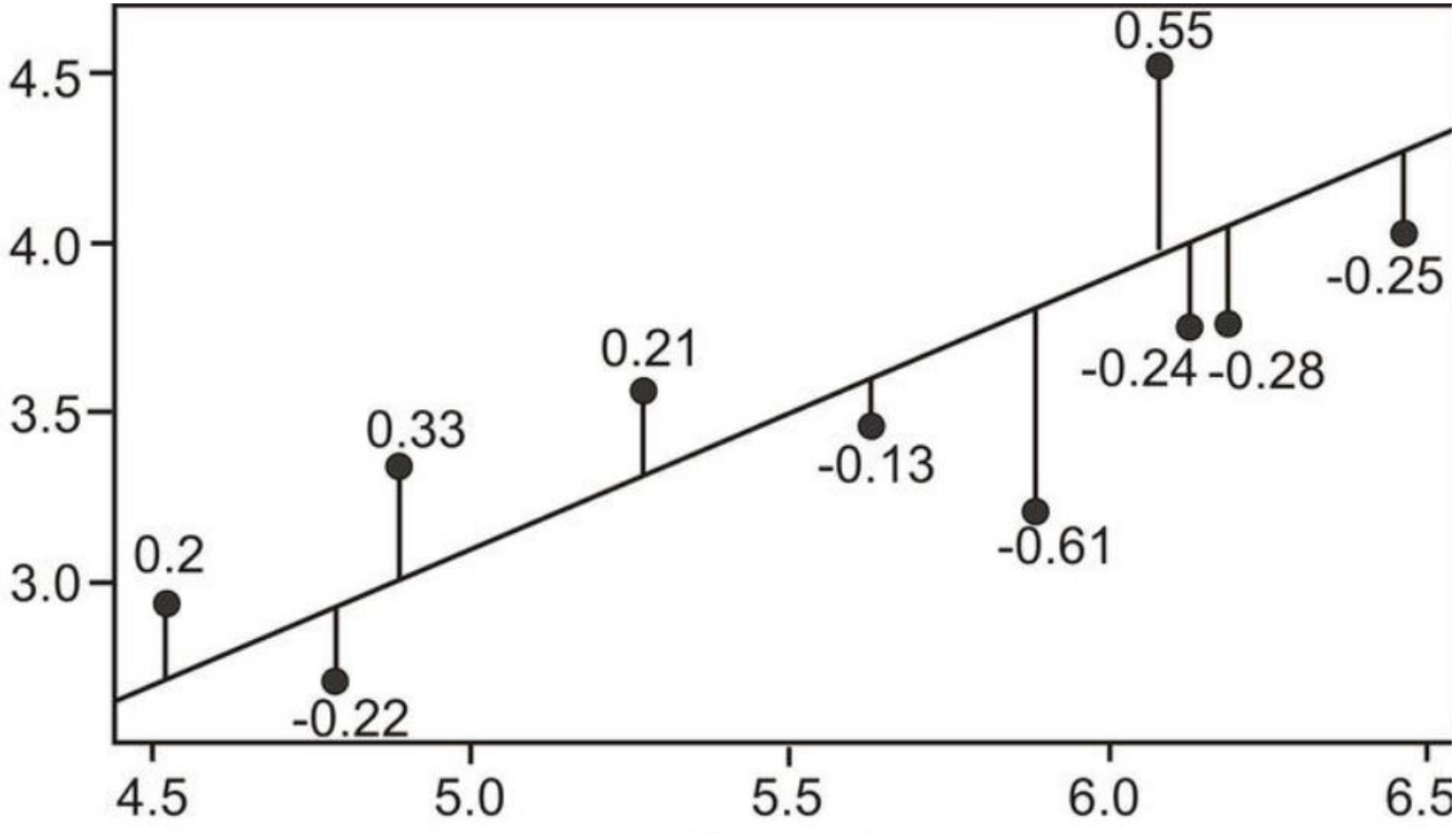
Fit a linear regression model

```
> model<-lm(distance~speed_ground,data=FAA)
> summary(model)

Call:
lm(formula = distance ~ speed_ground)

Residuals:
    Min      1Q  Median      3Q     Max
-957.96 -323.26  -82.72  207.20 2391.37

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1804.8738    71.3171  -25.31   <2e-16 ***
speed_ground    42.1088     0.8715   48.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 473.8 on 798 degrees of freedom
Multiple R-squared:  0.7453,    Adjusted R-squared:  0.7449
F-statistic:  2335 on 1 and 798 DF,  p-value: < 2.2e-16
```

How to find the “best” line?



Estimation: least square approach



Least square estimates

- Find the parameter values to minimize

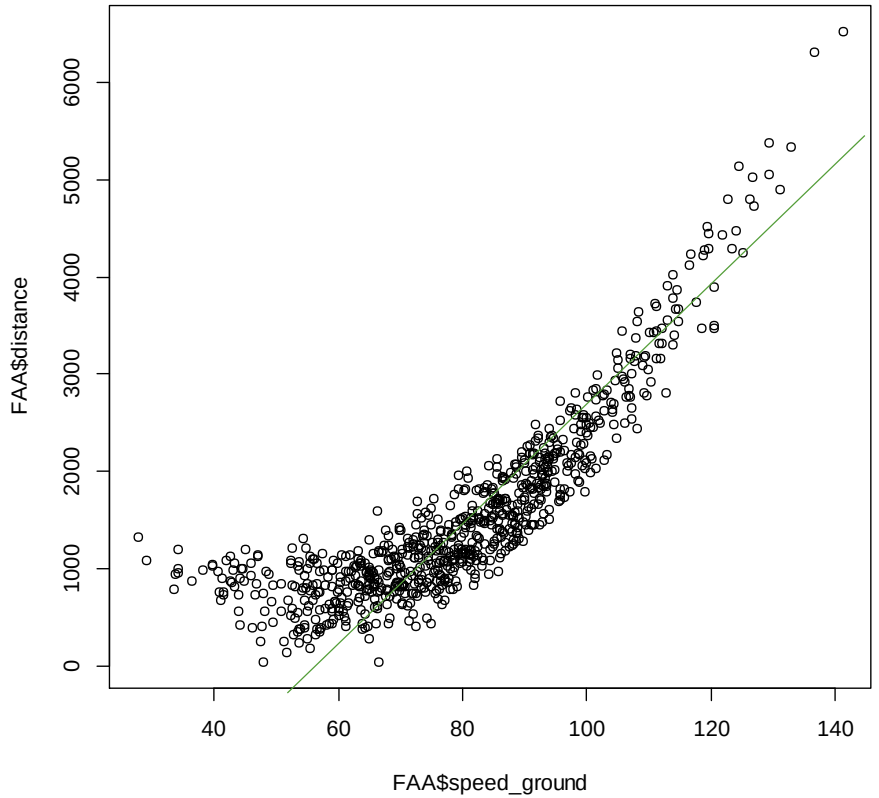
$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_i\beta_1))^2$$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1804.8738    71.3171  -25.31   <2e-16 ***
speed_ground    42.1088     0.8715   48.32   <2e-16 ***
```

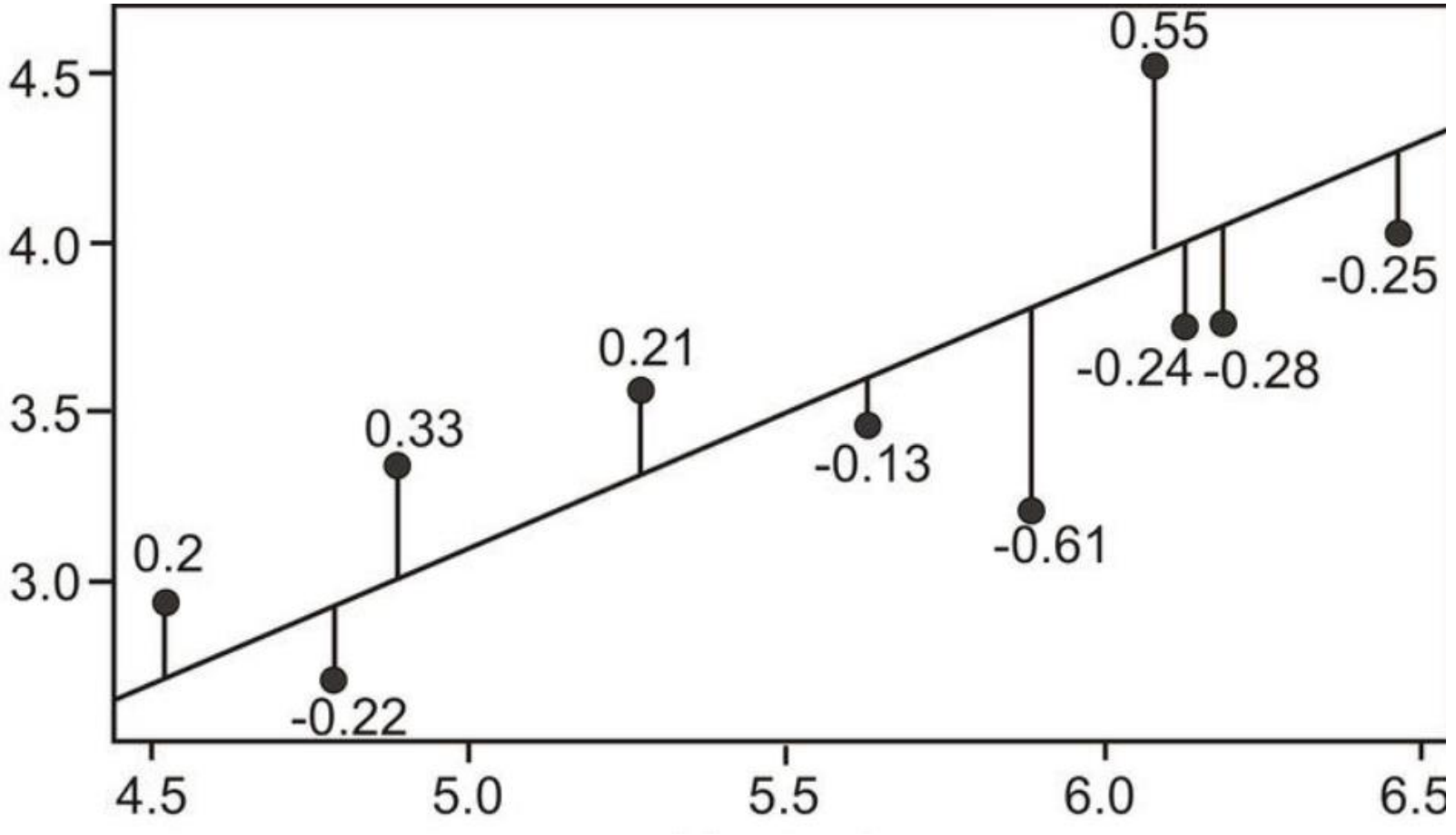
Prediction and residuals

```
> predict(model)
     1      2      3      4      5      6      7      8      9     10
2739.3 2475.7 1187.0 1808.6  717.0 1353.9  487.1  599.6 1793.1  797.3
    11     12     13     14     15     16     17     18     19     20
 459.7 4141.7 2127.7 2154.9  870.7  249.0 1713.6 1850.5 2608.5 3222.1

> residuals(model)
      1       2       3       4       5       6       7       8       9
 630.51  512.08  -42.12 -144.41  333.30  273.18  318.20  -25.99  -94.06
     10      11      12      13      14      15      16      17      18
 340.43  615.71 2391.37    0.97  149.97  219.23  694.10   79.97   60.37
```

What are predicted values?



What are residuals?



Definitions

For a linear regression model

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

- Predicted value: $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1$
- Residual: $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + x_i\hat{\beta}_1)$ (1)

Please use R to double check (1).

```
> head(residuals(model))
   1    2    3    4    5    6
 631  512  -42 -144  333  273

> head(FAA$distance-predict(model))
   1    2    3    4    5    6
 631  512  -42 -144  333  273


> all.equal(residuals(model),FAA$distance-predict(model))
[1] TRUE
```

Given a response (LD) and a predictor, what is the first question you would ask?

Association? Yes or No?

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

- Testing

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

  - The null hypothesis $H_0 : \beta_1 = 0$
  - The alternative hypothesis $H_1 : \beta_1 \neq 0$

- Question: why are we testing the specific value of zero?

Principle 2: innocent until proven guilty

- Under the presumption of innocence, the legal burden of proof is thus on the prosecution, which must collect and present compelling evidence to the trier of fact. The trier of fact (a judge or a jury) is thus restrained and ordered by law to consider only actual evidence and testimony presented in court. The prosecution must, in most cases prove that the accused is guilty beyond reasonable doubt. If reasonable doubt remains, the accused must be acquitted.

P-value: a summary of the evidence against the null hypothesis

```
> summary(model)

Call:
lm(formula = distance ~ speed_ground)

Residuals:
    Min      1Q  Median      3Q     Max
-957.96 -323.26  -82.72  207.20 2391.37

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1804.8738    71.3171  -25.31   <2e-16 ***
speed_ground    42.1088     0.8715   48.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 473.8 on 798 degrees of freedom
Multiple R-squared:  0.7453,    Adjusted R-squared:  0.7449
F-statistic:  2335 on 1 and 798 DF,  p-value: < 2.2e-16
```

P-value

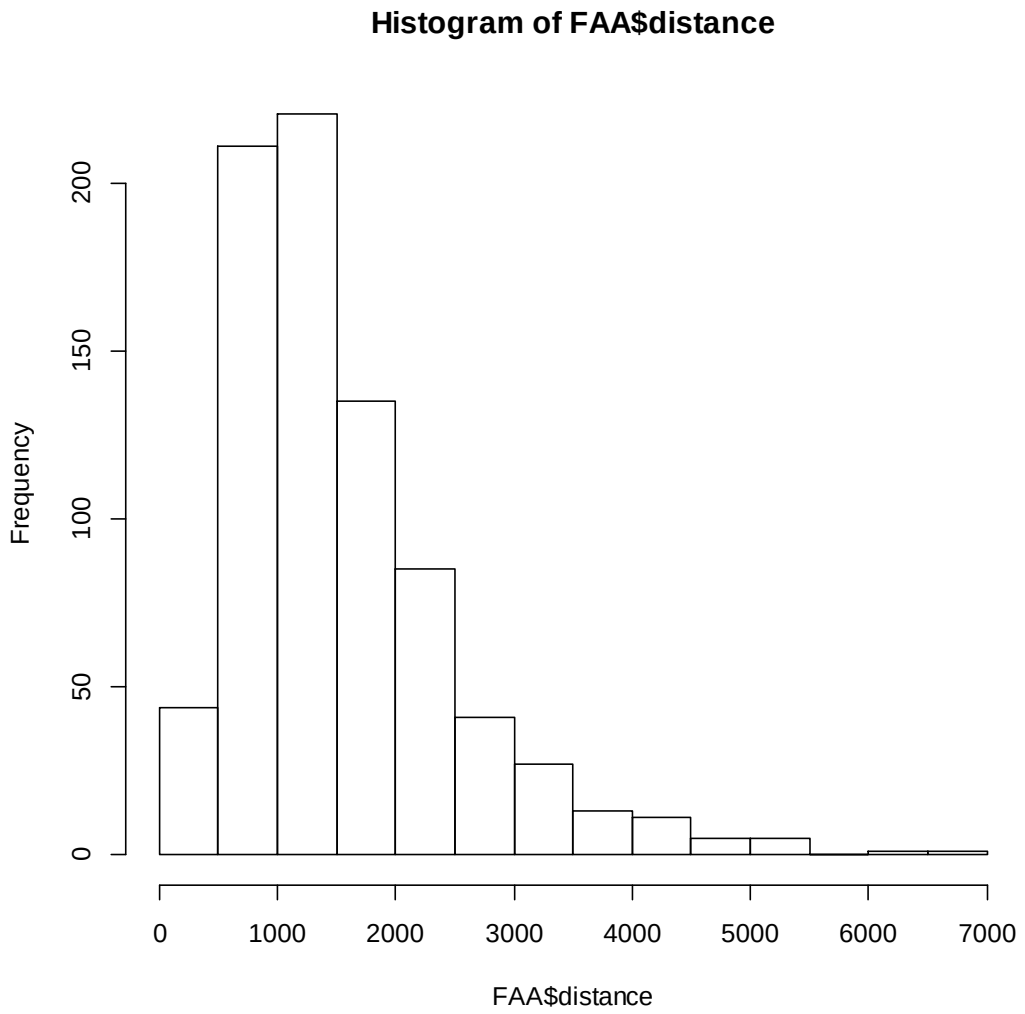
- It is a numeric value (probability) between 0 and 1.
- It measures the likelihood of having observed the given outcome, assuming that the X is innocent (e.g., X has nothing to do with Y).
- The smaller the P-value, the stronger the contradiction between the outcome (facts) and the innocent assumption (the null hypothesis).
- The threshold for judging “beyond reasonable doubt”: 0.05 (0.01 or 0.1).

P-value

$$\Pr_{H_0}(\text{Having observed the given (or more extreme) outcome})$$

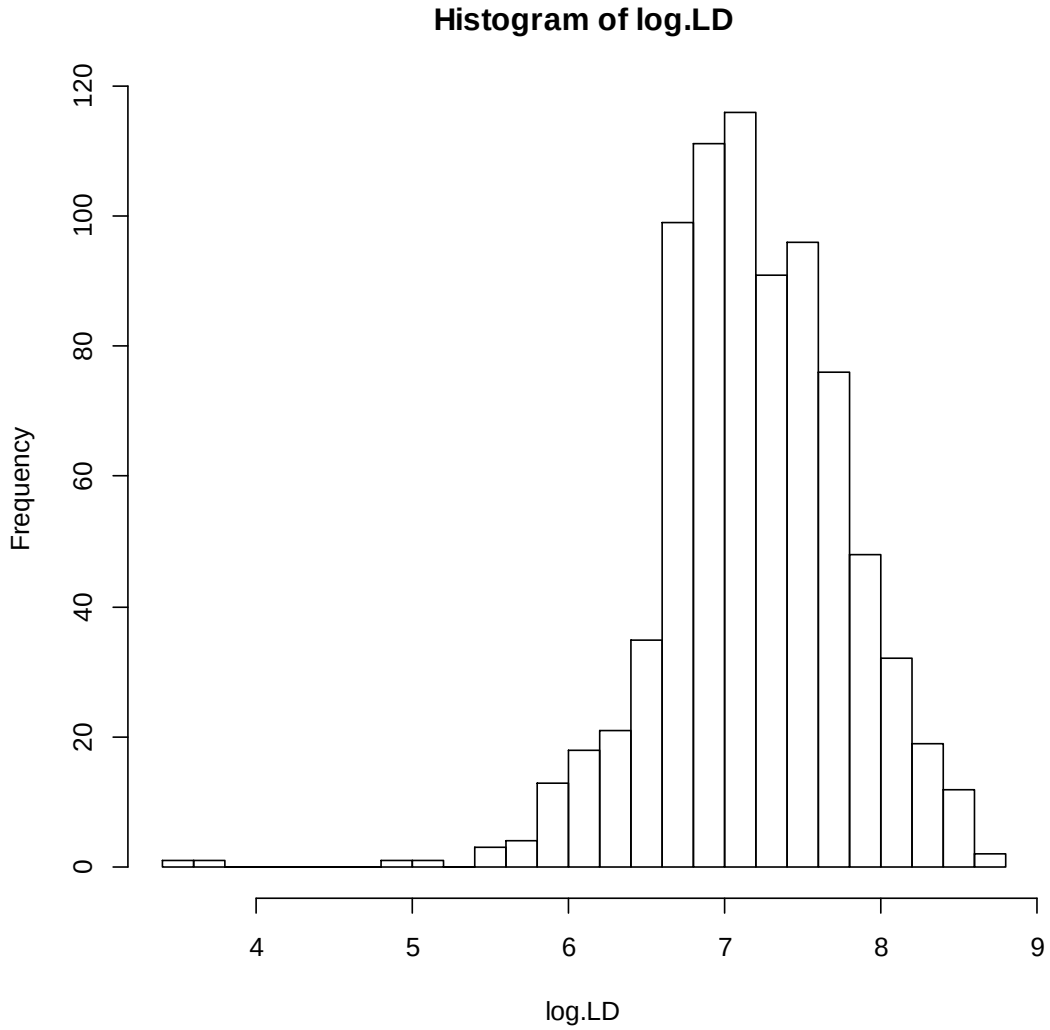
Exercise

```
> hist(FAA$distance)
```



```
> log.LD<-log(FAA$distance)
> hist(log.LD,breaks=20)
```

Assuming log.LD is normally distributed, we want to test if its mean is 6. How can we derive the P-value?



```
> t<-(mean(log.LD)-
6)/(sd(log.LD)/sqrt(length(log.LD)))
> t
[1] 55.16771
> p.value<-2*(1-pnorm(t))
> p.value
[1] 0
```

T test versus F test

- Consider the case with more than one predictors

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- T-test: $H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$
- F-test: $H_0 : \beta_1 = \beta_2 = 0 \quad \text{versus} \quad H_1 : \text{otherwise}$

- Question: why do we want to test $H_0 : \beta_1 = \beta_2 = 0$?

```
> model2<-lm(distance~speed_ground+height,data=FAA)
> summary(model2)

Call:
lm(formula = distance ~ speed_ground + height, data = FAA)

Residuals:
   Min     1Q Median     3Q    Max
-770.1 -338.1  -61.4  173.0 2477.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2221.652     83.068  -26.74   <2e-16 ***
speed_ground    42.169      0.833   50.64   <2e-16 ***
height          13.677      1.559    8.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 453 on 797 degrees of freedom
Multiple R-squared:  0.768,     Adjusted R-squared:  0.767
F-statistic: 1.32e+03 on 2 and 797 DF,  p-value: <2e-16
```

Given that we know a response (LD) is associated with some predictors, what is the second question you would ask?

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- The LD is a random variable. Calculate its variance using R.
- Calculate the variance of predicted values $\hat{y}_i$ using R.
- Calculate the variance of residuals $\hat{\epsilon}_i$ using R.
- Question: What is the relationship among the three variances?

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

```
> var(FAA$distance)
[1] 880281.35
> var(predict(model2))
[1] 675784.83
> var(residuals(model2))
[1] 204496.52

> var(predict(model2))+var(residuals(model2))
[1] 880281.35

> var(predict(model2))/var(FAA$distance) ## What is this ratio?
[1] 0.76769186
```

R squared

$$var(Y) = var(\beta_0 + X_1\beta_1 + X_2\beta_2) + var(\epsilon)$$

- `var(FAA$distance)`: Total Sum of Squares (TSS)

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- `var(residuals(model2))`: Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- ANOVA: TSS = (TSS-RSS) + RSS

R squared

- R squared is the proportion of var(Y) that can be explained by X.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

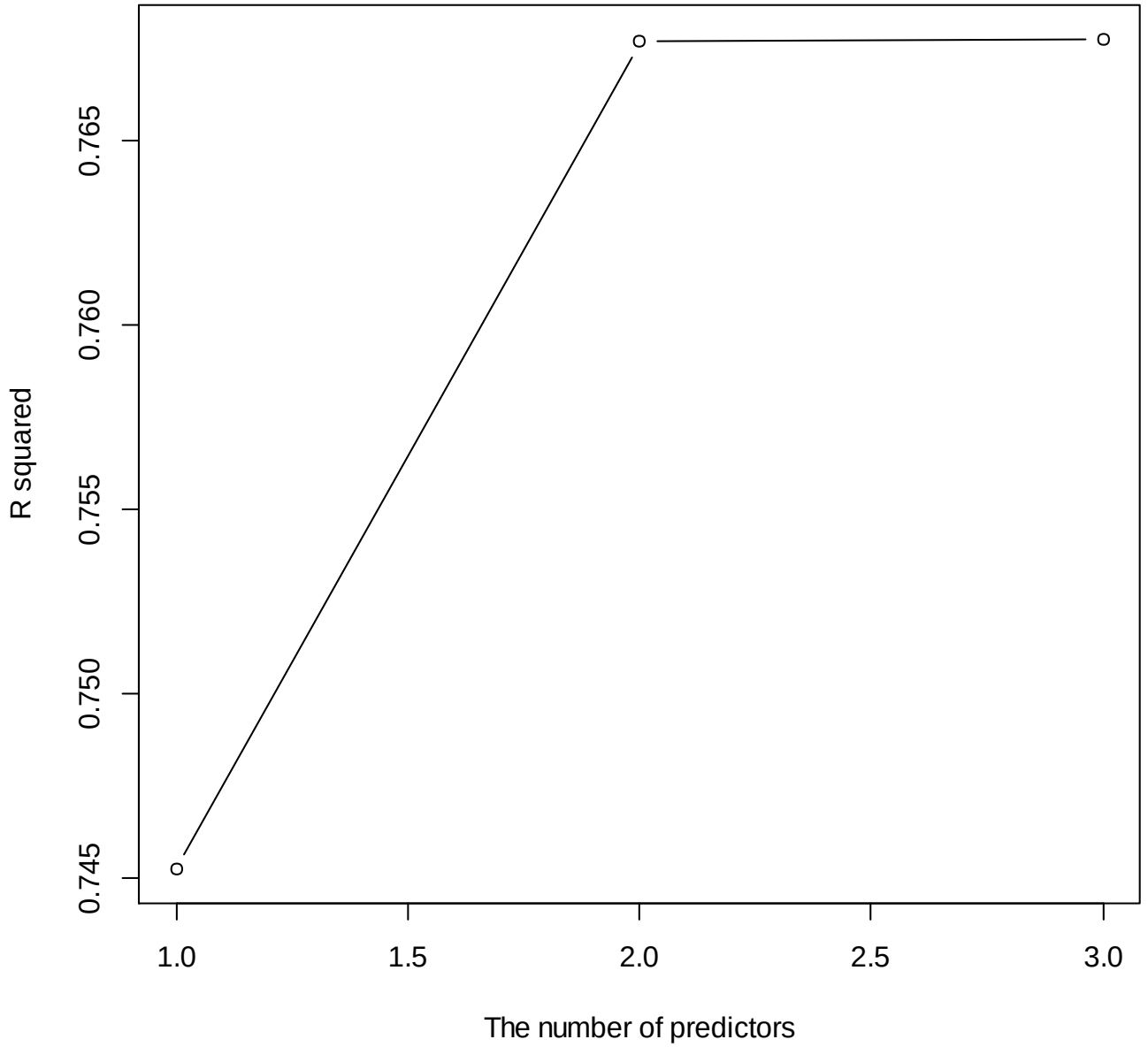
- R squared is a value between 0 and 1.

Exercise

```
> model3<-lm(distance~speed_ground+height+duration,data=FAA)
```

- How many models you have in R so far?
- Compare the R squared of those models.
- Draw a plot to show how R squared changes as the number of X increases.

```
> r.squared.1<-summary(model)$r.squared; print(r.squared.1)
[1] 0.745253
> r.squared.2<-summary(model2)$r.squared; print(r.squared.2)
[1] 0.7676919
> r.squared.3<-summary(model3)$r.squared; print(r.squared.3)
[1] 0.767767

> plot(c(1,2,3),c(r.squared.1,r.squared.2,r.squared.3),type="b",ylab="R
squared",xlab="The number of predictors")
```


R squared
0.745
0.750
0.755
0.760
0.765
1.0
1.5
2.0
2.5
3.0
The number of predictors

The monotonicity of R squared

- As the number of X increases,
  - The R squared always increases, and
  - The explanatory power of the linear model always increases.
- However …

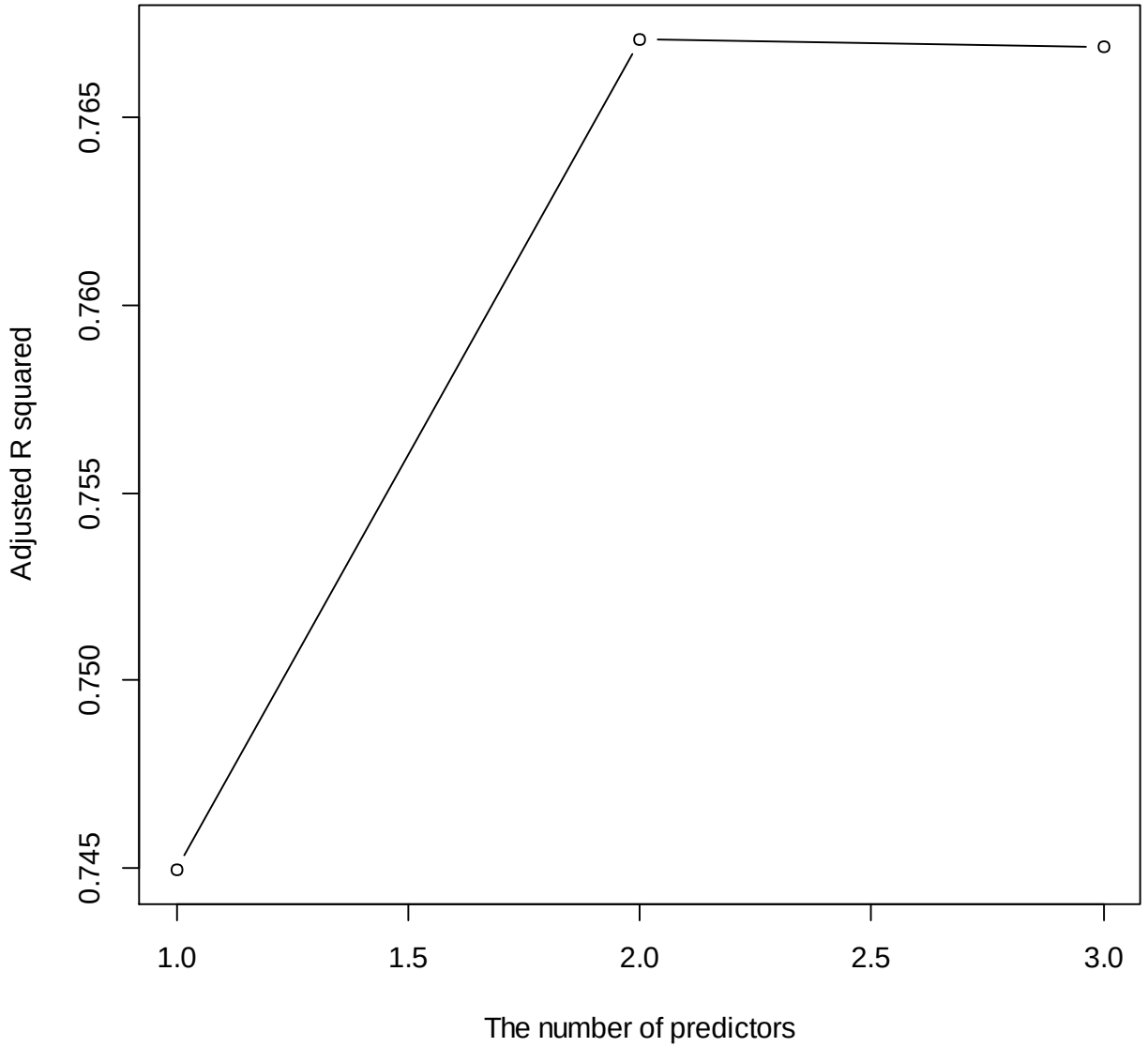
However ….

The predictive power of the linear model may decrease.

Exercise

```
adj.r.squared.1<-summary(model)$adj.r.squared;
print(adj.r.squared.1)
adj.r.squared.2<-summary(model2)$adj.r.squared;
print(adj.r.squared.2)
adj.r.squared.3<-summary(model3)$adj.r.squared;
print(adj.r.squared.3)


plot(c(1,2,3),c(adj.r.squared.1,adj.r.squared.2,adj.r.
squared.3),type="b",ylab="Adjusted R
squared",xlab="The number of predictors")
```

```
> print(adj.r.squared.1)
[1] 0.7449338
> print(adj.r.squared.2)
[1] 0.7671089
> print(adj.r.squared.3)
[1] 0.7668918
```



Adjusted R squared

$$R_a^2 = 1 - \frac{RSS}{TSS} \times \frac{n-1}{n-p}$$

- The adjusted R squared applies a penalty on the number of X.
- Does this make sense?

Principle 3: No need of Yao and O'Neals in the same team!

Principle 3: No need of Yao and O'Neals in the same team!

- To decide if we want to include another person X in our team, we will access the **additional value** X will bring to the team, rather than the **stand-alone value** of X.

- This principle is used in the adjusted R squared, AIC, BIC, and other variable selection approaches.