AUA CS 108, Statistics, Fall 2019

Lecture 44 (The Last One)

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

06 Dec 2019

Contents

Last Lecture ReCap

- We will skip this, sorry!

Intro to Linear Regression

Let us recall that we have defined the **Regression Function** by:

$$\mathit{RegFun}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}).$$

Intro to Linear Regression

Let us recall that we have defined the **Regression Function** by:

$$\mathit{RegFun}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}).$$

and we model the dependency of Y on $\mathbf{X}$ by

$$Y = \mathit{RegFun}(\mathbf{X}) + \varepsilon.$$

Intro to Linear Regression

Let us recall that we have defined the **Regression Function** by:

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}).$$

and we model the dependency of Y on $\mathbf{X}$ by

$$Y = RegFun(\mathbf{X}) + \varepsilon.$$

Here $\mathbb{E}(\varepsilon) = 0$, for any $\mathbf{x}$.

Intro to Linear Regression

Let us recall that we have defined the **Regression Function** by:

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}).$$

and we model the dependency of Y on $\mathbf{X}$ by

$$Y = RegFun(\mathbf{X}) + \varepsilon.$$

Here $\mathbb{E}(\varepsilon) = 0$, for any $\mathbf{x}$.

Now, unfortunately, we cannot calculate $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, since we do not have the Distribution of $Y|\mathbf{X} = \mathbf{x}$. We just have Data.

Intro to Linear Regression

Let us recall that we have defined the **Regression Function** by:

$$\mathit{RegFun}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}).$$

and we model the dependency of Y on $\mathbf{X}$ by

$$Y = \mathit{RegFun}(\mathbf{X}) + \varepsilon.$$

Here $\mathbb{E}(\varepsilon) = 0$, for any $\mathbf{x}$.

Now, unfortunately, we cannot calculate $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, since we do not have the Distribution of $Y|\mathbf{X} = \mathbf{x}$. We just have Data.

And also we will consider the simplest case, when we seek the Regression Function among the Linear Functions.

Intro to Linear Regression

The simplest Regression Model is the Linear Model: we will assume that

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta^T \cdot \mathbf{x}.$$

Intro to Linear Regression

The simplest Regression Model is the Linear Model: we will assume that

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta^T \cdot \mathbf{x}.$$

Hence, in the Linear Regression Problem, we assume

$$Y = \beta_0 + \beta^T \cdot \mathbf{X} + \varepsilon,$$

where ε is a r.v., for each value of $\mathbf{X}$, with $\mathbb{E}(\varepsilon) = 0$.

Note: For each value of X, ε can be a different r.v.!

Intro to Linear Regression

The simplest Regression Model is the Linear Model: we will assume that

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta^T \cdot \mathbf{x}.$$

Hence, in the Linear Regression Problem, we assume

$$Y = \beta_0 + \beta^T \cdot \mathbf{X} + \varepsilon,$$

where ε is a r.v., for each value of $\mathbf{X}$, with $\mathbb{E}(\varepsilon) = 0$.

Note: For each value of X, ε can be a different r.v.!

Here, if $\mathbf{X} = (X^1, ..., X^d)$ is d-Dim, then

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix},$$

Intro to Linear Regression

The simplest Regression Model is the Linear Model: we will assume that

$$RegFun(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta^T \cdot \mathbf{x}.$$

Hence, in the Linear Regression Problem, we assume

$$Y = \beta_0 + \beta^T \cdot \mathbf{X} + \varepsilon,$$

where ε is a r.v., for each value of $\mathbf{X}$, with $\mathbb{E}(\varepsilon) = 0$.

Note: For each value of X, ε can be a different r.v.!

Here, if $\mathbf{X} = (X^1, ..., X^d)$ is d-Dim, then

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix},$$

so our Model, in the expanded way, is

$$Y = \beta_0 + \beta_1 \cdot X^1 + \beta_2 \cdot X^2 + ... + \beta_d \cdot X^d + \varepsilon.$$

Simple Linear Regression

Now we simplify further the story: we consider the 1D Explanatory Variable case, i.e., $\mathbf{X} = X$ is 1D.

Simple Linear Regression

Now we simplify further the story: we consider the 1D Explanatory Variable case, i.e., $\mathbf{X} = X$ is 1D. In that case, we will have our LR Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

Simple Linear Regression

Now we simplify further the story: we consider the 1D Explanatory Variable case, i.e., $\mathbf{X} = X$ is 1D. In that case, we will have our LR Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

This is called the **Simple Linear Regression Model**.

Example

Example:

Example

Example: (of Prob type, not of a Stat type, a Generative Model)

Example

Example: (of Prob type, not of a Stat type, a Generative Model)
Assume we have the actual Relationship between X and Y:

$$Y = -1 + 3 \cdot X + \varepsilon$$

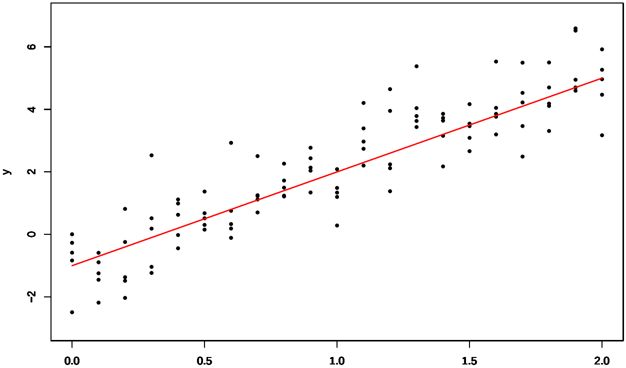
where $X \in [0, 2]$ and for each value of X, $\varepsilon \sim \mathcal{N}(0, 1)$, and for different values of X, ε-s are Independent.

Example

Example: (of Prob type, not of a Stat type, a Generative Model)
Assume we have the actual Relationship between X and Y:

$$Y = -1 + 3 \cdot X + \varepsilon$$

where $X \in [0, 2]$ and for each value of X, $\varepsilon \sim \mathcal{N}(0, 1)$, and for different values of X, ε-s are Independent.



Simple Linear Regression

Again, let us assume we have a Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

Simple Linear Regression

Again, let us assume we have a Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

The Problem is the Following: we have a Dataset of pairs:

$$(x_1, y_1), ..., (x_n, y_n),$$

and we want to Estimate β_0 and β_1.

Simple Linear Regression

Again, let us assume we have a Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

The Problem is the Following: we have a Dataset of pairs:

$$(x_1, y_1), ..., (x_n, y_n),$$

and we want to Estimate β_0 and β_1.

Of course, as we have done it many-many times, to generalize, we Model this by a Random Sample (Y_k-s are Random, but not X_k-s)

$$(X_1, Y_1), ..., (X_n, Y_n),$$

coming from a Model

$$Y_k = \beta_0 + \beta_1 \cdot X_k + \varepsilon_k, \qquad \mathbb{E}(\varepsilon_k) = 0$$

Simple Linear Regression

Again, let us assume we have a Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0.$$

The Problem is the Following: we have a Dataset of pairs:

$$(x_1, y_1), ..., (x_n, y_n),$$

and we want to Estimate β_0 and β_1.

Of course, as we have done it many-many times, to generalize, we Model this by a Random Sample (Y_k-s are Random, but not X_k-s)

$$(X_1, Y_1), ..., (X_n, Y_n),$$

coming from a Model

$$Y_k = \beta_0 + \beta_1 \cdot X_k + \varepsilon_k, \qquad \mathbb{E}(\varepsilon_k) = 0$$

where ε_k-s are Independent, and our aim is to find good Estimators for β_0 and β_1.

Simple Linear Regression

Now, the idea of the Ordinary Least Squares Method for Estimating the Parameters β_0, β_1 is the following: Find

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{argmin} \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k \Big)^2 .$$

Simple Linear Regression

Now, the idea of the Ordinary Least Squares Method for Estimating the Parameters β_0, β_1 is the following: Find

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{argmin} \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k\Big)^2.$$

Here, the solution $(\hat{\beta}_0, \hat{\beta}_1)$ will be a pair of r.v.s, since Y_k-s are r.v.s. So we will obtain *Estimators* for β_0 and β_1.

Simple Linear Regression

Now, the idea of the Ordinary Least Squares Method for Estimating the Parameters β_0, β_1 is the following: Find

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{argmin} \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k\Big)^2.$$

Here, the solution $(\hat{\beta}_0, \hat{\beta}_1)$ will be a pair of r.v.s, since Y_k-s are r.v.s. So we will obtain *Estimators* for β_0 and β_1. If we will have an Observation (y_k, x_k), $k = 1, ..., n$, then we will solve

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{argmin} \sum_{k=1}^{n} \Big(y_k - \beta_0 - \beta_1 \cdot x_k\Big)^2,$$

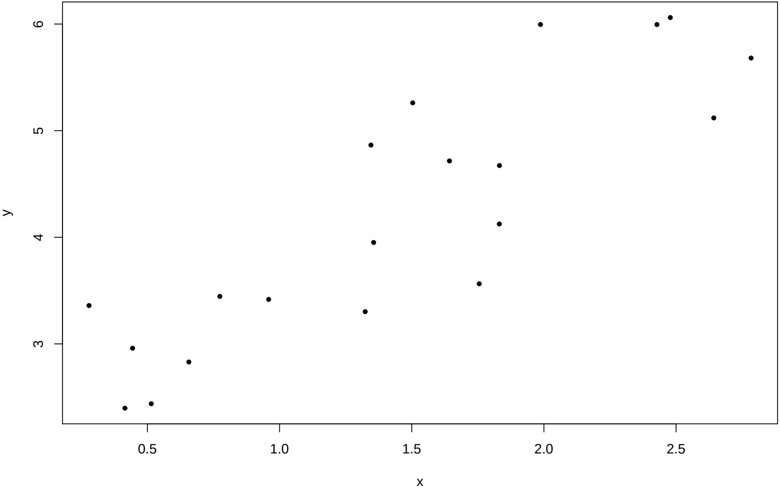
and we will find *Estimates* for β_0 and β_1.

Simple Linear Regression

Geometrically, the Problem is to find the "best fit line":

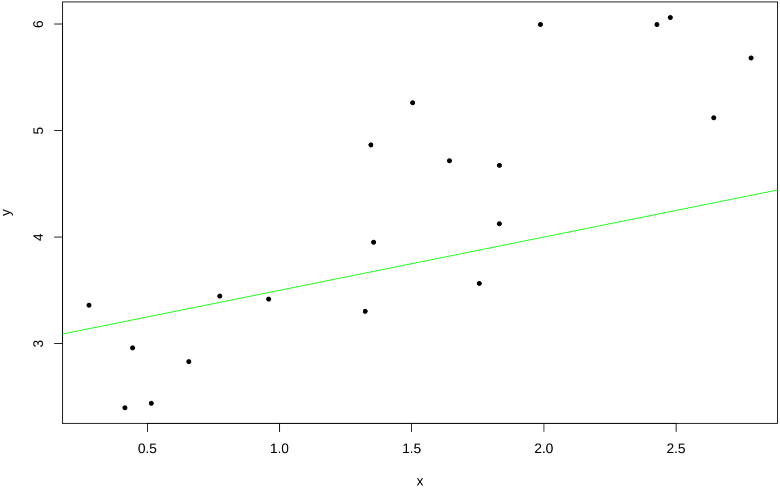
Simple Linear Regression

Geometrically, the Problem is to find the "best fit line":



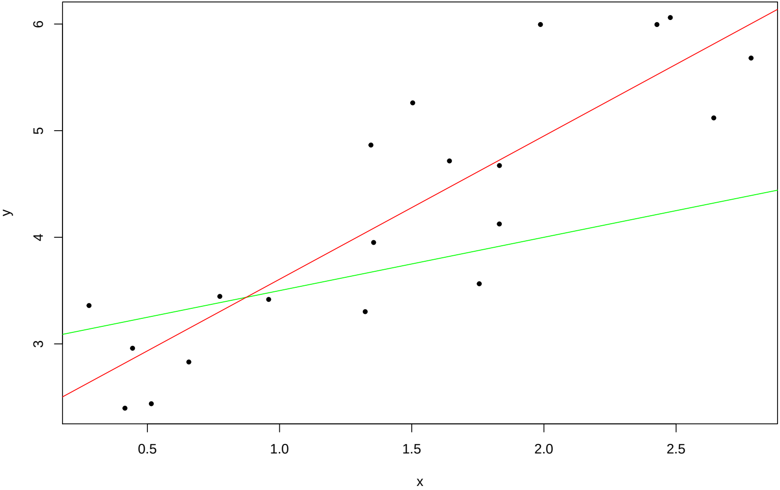
Simple Linear Regression

Geometrically, the Problem is to find the “best fit line”:



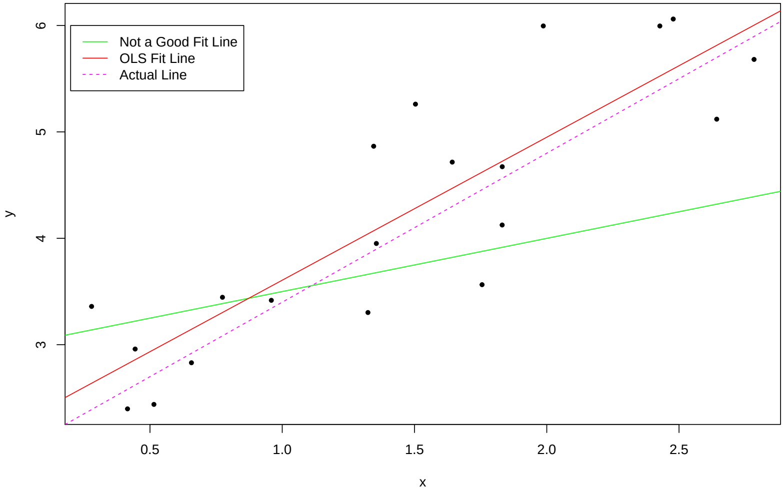
Simple Linear Regression

Geometrically, the Problem is to find the "best fit line":



Simple Linear Regression

Geometrically, the Problem is to find the "best fit line":



Simple Linear Regression

Calculation of the best fit line is easy: we just define

$$\varphi(\beta_0, \beta_1) = \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k \Big)^2, \qquad (\beta_0, \beta_1) \in \mathbb{R}^2,$$

and using our Calc 3 knowledge, find the Minimum Point of φ by solving the System

$$\begin{cases} \dfrac{\partial \varphi}{\partial \beta_0} = 0 \\ \dfrac{\partial \varphi}{\partial \beta_1} = 0 \end{cases}$$

Simple Linear Regression

Calculation of the best fit line is easy: we just define

$$\varphi(\beta_0, \beta_1) = \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k \Big)^2, \qquad (\beta_0, \beta_1) \in \mathbb{R}^2,$$

and using our Calc 3 knowledge, find the Minimum Point of φ by solving the System

$$\begin{cases} \dfrac{\partial \varphi}{\partial \beta_0} = 0 \\ \dfrac{\partial \varphi}{\partial \beta_1} = 0 \end{cases}$$

The Solution is:

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})(Y_k - \overline{Y})}{\sum_{k=1}^{n} (X_k - \overline{X})^2}$$

Simple Linear Regression

Calculation of the best fit line is easy: we just define

$$\varphi(\beta_0, \beta_1) = \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k \Big)^2, \qquad (\beta_0, \beta_1) \in \mathbb{R}^2,$$

and using our Calc 3 knowledge, find the Minimum Point of φ by solving the System

$$\begin{cases} \dfrac{\partial \varphi}{\partial \beta_0} = 0 \\ \dfrac{\partial \varphi}{\partial \beta_1} = 0 \end{cases}$$

The Solution is:

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})(Y_k - \overline{Y})}{\sum_{k=1}^{n} (X_k - \overline{X})^2} = cor(X, Y) \cdot \frac{sd(Y)}{sd(X)} = \rho_{XY} \cdot \frac{S_Y}{S_X}$$

Simple Linear Regression

Calculation of the best fit line is easy: we just define

$$\varphi(\beta_0, \beta_1) = \sum_{k=1}^{n} \Big(Y_k - \beta_0 - \beta_1 \cdot X_k\Big)^2, \qquad (\beta_0, \beta_1) \in \mathbb{R}^2,$$

and using our Calc 3 knowledge, find the Minimum Point of φ by solving the System

$$\begin{cases} \dfrac{\partial \varphi}{\partial \beta_0} = 0 \\ \dfrac{\partial \varphi}{\partial \beta_1} = 0 \end{cases}$$

The Solution is:

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sum_{k=1}^{n}(X_k - \overline{X})^2} = cor(X, Y) \cdot \frac{sd(Y)}{sd(X)} = \rho_{XY} \cdot \frac{S_Y}{S_X}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \cdot \overline{X}.$$

Simple Linear Regression

The obtained Line

$$y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

is called the **Regression Line**.

Simple Linear Regression

The obtained Line

$$y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

is called the **Regression Line**.

Note: So the $cor(X, Y)$ is not the Slope of the Regression Line, but

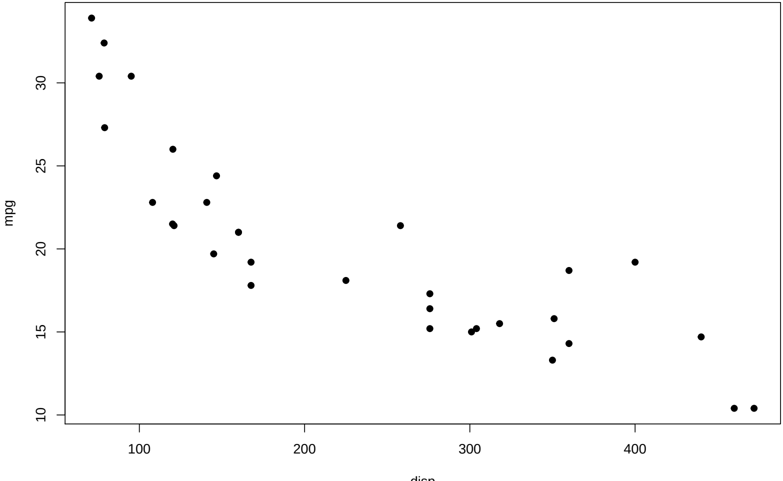
$$cor(X, Y) \cdot \frac{sd(Y)}{sd(X)}$$

is. Recall our Descriptive Statistics part!

Example

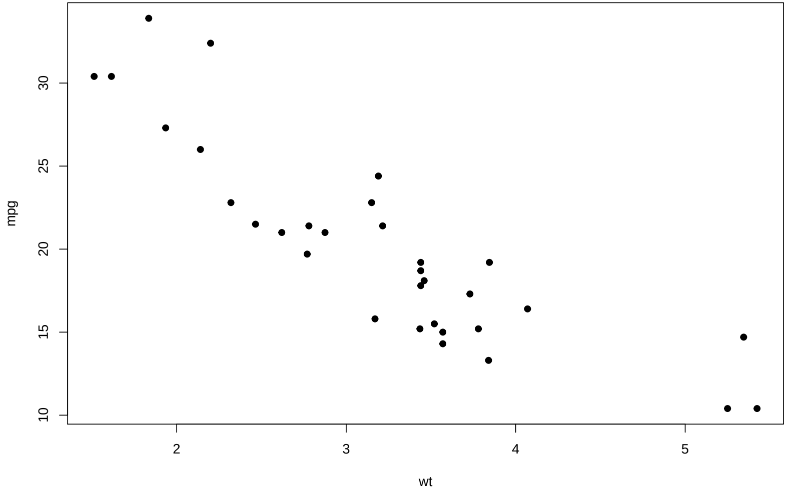
Example: We will use the `mtcars` Dataset from **R**:

```r
plot(mpg ~ disp, data = mtcars, pch = 19)
```



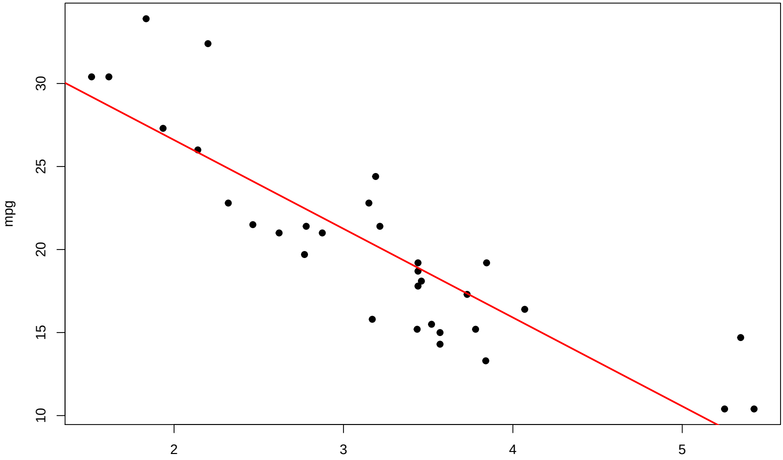
Example, Cont'd

```
plot(mpg ~ wt, data = mtcars, pch = 19)
```



Example, Cont'd

```
plot(mpg ~ wt, data = mtcars, pch = 19)
model <- lm(mpg ~ wt, data = mtcars)
abline(model, col = "red", lwd = 2)
```



Example, Cont'd

```r
model <- lm(mpg ~ wt, data = mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Example, Cont'd

Now, we predict the value of mpg for a new values of a wt Variable:

```r
pred <- predict(model, data.frame(wt=4.7))
pred
```

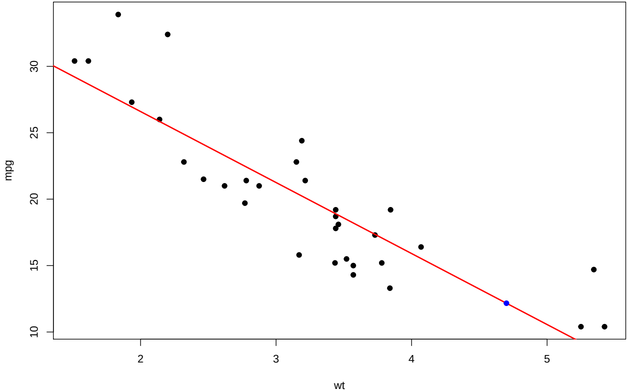
```
##        1
## 12.16611
```

Example, Cont'd

Now, we predict the value of mpg for a new values of a wt Variable:

```
pred <- predict(model, data.frame(wt=4.7))
pred
```

```
##        1
## 12.16611
```



Example

```r
x <- rnorm(100, mean = -1, sd = 1)
y <- runif(100, 2, 10)
z <- 2.7 - 1.7*x + 13.5*y + rnorm(100)
head(x)
```

```
## [1]  0.5271618 -2.0473435  0.4363159 -0.1708086 -0.7298618 -1
```

```r
head(y)
```

```
## [1] 3.977900 9.480366 2.867116 6.922705 7.797675 7.675757
```

```r
head(z)
```

```
## [1]  55.39792 133.41440  42.64337  95.71013 109.41316 108.627
```

Example

```r
mod1 <- lm(z ~ x); summary(mod1)
```

```
##
## Call:
## lm(formula = z ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.755 -27.454   2.906  28.465  50.215
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   84.135      4.202  20.022   <2e-16 ***
## x             -3.130      2.917  -1.073    0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 31.23 on 98 degrees of freedom
## Multiple R-squared:  0.01161,    Adjusted R-squared:  0.00152
## F-statistic: 1.152 on 1 and 98 DF,  p-value: 0.2859
```

Example, Cont'd

```r
mod2 <- lm(z ~ x + y); summary(mod2)
```

```
##
## Call:
## lm(formula = z ~ x + y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25835 -0.69422 -0.04329  0.72539  2.12851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92180    0.29018   10.07   <2e-16 ***
## x           -1.50675    0.09194  -16.39   <2e-16 ***
## y           13.46759    0.04283  314.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.9827 on 97 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 5.001e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

Properties of the Estimators: $\hat{\beta}_0$ and $\hat{\beta}_1$

Here we assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Properties of the Estimators: $\hat{\beta}_0$ and $\hat{\beta}_1$

Here we assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Fact 1: Estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are UnBiased:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad Var(\hat{\beta}_0) = \frac{\sigma^2}{n} \cdot \frac{\sum_{k=1}^n X_k^2}{\sum_{k=1}^n (X_k - \overline{X})^2}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{k=1}^n (X_k - \overline{X})^2}$$

Properties of the Estimators: $\widehat{\sigma^2}$

Fact 2: Assume σ^2 is unknown.

Properties of the Estimators: $\widehat{\sigma^2}$

Fact 2: Assume σ^2 is unknown. Then

$$s^2 = \frac{1}{n-2} \cdot \sum_{k=1}^{n} (\hat{\varepsilon}_k)^2$$

is an UnBiased Estimator for σ^2, and

$$\widehat{\sigma^2} = \frac{1}{n} \cdot \sum_{k=1}^{n} (\hat{\varepsilon}_k)^2$$

is the MLE for σ^2.

Properties of the Estimators: $\widehat{\sigma^2}$

Fact 2: Assume σ^2 is unknown. Then

$$s^2 = \frac{1}{n-2} \cdot \sum_{k=1}^{n} (\hat{\varepsilon}_k)^2$$

is an UnBiased Estimator for σ^2, and

$$\widehat{\sigma^2} = \frac{1}{n} \cdot \sum_{k=1}^{n} (\hat{\varepsilon}_k)^2$$

is the MLE for σ^2. Here

$$\hat{\varepsilon}_k = Y_k - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_k$$

is the k-th **residual**.

Goodness-of-Fit Tests

Intro to GoF Tests

Here, we have a DataSet $x_1, x_2, ..., x_n$, and a Statistical Model, and we want to see how good our Model is fitting the Data.

Intro to GoF Tests

Here, we have a DataSet $x_1, x_2, ..., x_n$, and a Statistical Model, and we want to see how good our Model is fitting the Data.

First we consider Pearson's χ^2-Test: a famous GoF Test for the Multinomial Distribution.

Goodness-of-Fit Tests: Pearson's χ^2 Test

Model: Here we assume that the result of an Experiment can be one of the $A_1, ..., A_m$ (different classes), with Probabilities

$$p_1 = \mathbb{P}(A_1), ..., p_m = \mathbb{P}(A_m).$$

Goodness-of-Fit Tests: Pearson's χ^2 Test

Model: Here we assume that the result of an Experiment can be one of the $A_1, ..., A_m$ (different classes), with Probabilities

$$p_1 = \mathbb{P}(A_1), ..., p_m = \mathbb{P}(A_m).$$

Data: We have the results of a repetition of the previous Experiment: The results are: the number of A_1 shown is X_1, the number of A_2 shown is X_2, . . . , the number of A_m shown is X_m;

Goodness-of-Fit Tests: Pearson's χ^2 Test

Model: Here we assume that the result of an Experiment can be one of the $A_1, ..., A_m$ (different classes), with Probabilities

$$p_1 = \mathbb{P}(A_1), ..., p_m = \mathbb{P}(A_m).$$

Data: We have the results of a repetition of the previous Experiment: The results are: the number of A_1 shown is X_1, the number of A_2 shown is X_2, . . . , the number of A_m shown is X_m;

Null Hypothesis:

$$\mathcal{H}_0: \text{ The Actual Probabilities are } p_1, p_2, ..., p_m.$$

vs

$$\mathcal{H}_1: \mathcal{H}_0 \text{ is not correct.}$$

Goodness-of-Fit Tests: Pearson's χ^2 Test

Model: Here we assume that the result of an Experiment can be one of the $A_1, ..., A_m$ (different classes), with Probabilities

$$p_1 = \mathbb{P}(A_1), ..., p_m = \mathbb{P}(A_m).$$

Data: We have the results of a repetition of the previous Experiment: The results are: the number of A_1 shown is X_1, the number of A_2 shown is X_2, . . . , the number of A_m shown is X_m;

Null Hypothesis:

$$\mathcal{H}_0 : \text{ The Actual Probabilities are } p_1, p_2, ..., p_m.$$

vs

$$\mathcal{H}_1 : \mathcal{H}_0 \text{ is not correct.}$$

Significance Level: $\alpha \in (0, 1)$;

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics:

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k}$

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$

Here usually one constructs the following χ^2-Table:

	A_1	A_2	...	A_m
Observed Freq., O_k	X_1	X_2	...	X_m
Expected Freq., E_k	$n \cdot p_1$	$n \cdot p_2$	...	$n \cdot p_m$

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$

Here usually one constructs the following χ^2-Table:

	A_1	A_2	...	A_m
Observed Freq., O_k	X_1	X_2	...	X_m
Expected Freq., E_k	$n \cdot p_1$	$n \cdot p_2$	...	$n \cdot p_m$

Assumption: We assume that $n \cdot p_k \geq 5$ for any k;

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$

Here usually one constructs the following χ^2-Table:

	A_1	A_2	...	A_m
Observed Freq., O_k	X_1	X_2	...	X_m
Expected Freq., E_k	$n \cdot p_1$	$n \cdot p_2$	...	$n \cdot p_m$

Assumption: We assume that $n \cdot p_k \geq 5$ for any k;

Distrib of the Test-Statistics Under $\mathcal{H}_0$: $\chi^2 \approx \chi^2(m-1)$;

Goodness-of-Fit Tests: Pearson's χ^2 Test

Test Statistics: $\chi^2 = \sum_{k=1}^{m} \frac{(X_k - n \cdot p_k)^2}{n \cdot p_k} = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$

Here usually one constructs the following χ^2-Table:

	A_1	A_2	...	A_m
Observed Freq., O_k	X_1	X_2	...	X_m
Expected Freq., E_k	$n \cdot p_1$	$n \cdot p_2$	...	$n \cdot p_m$

Assumption: We assume that $n \cdot p_k \geq 5$ for any k;

Distrib of the Test-Statistics Under $\mathcal{H}_0$: $\chi^2 \approx \chi^2(m-1)$;

Rejection Region: $\chi^2 > \chi^2_{m-1,1-\alpha}$

Example

Example: I am claiming that, for my Stat courses, the percentage of A-grade students is 15%, of B-grade students is 25%, of C-grades are 20%, for D I have 15%, and all others are Failing the course.

Example

Example: I am claiming that, for my Stat courses, the percentage of A-grade students is 15%, of B-grade students is 25%, of C-grades are 20%, for D I have 15%, and all others are Failing the course. Now, one of my courses finished with the following result:

$$\#A = 27,\ \#B = 22,\ \#C = 10,\ \#D = 10,\ \#F = 12.$$

Is this Data supporting my claim?

Solution:

Example

Example: I am claiming that, for my Stat courses, the percentage of A-grade students is 15%, of B-grade students is 25%, of C-grades are 20%, for D I have 15%, and all others are Failing the course. Now, one of my courses finished with the following result:

$$\#A = 27,\ \#B = 22,\ \#C = 10,\ \#D = 10,\ \#F = 12.$$

Is this Data supporting my claim?

Solution: We have $n = 27 + 22 + 10 + 10 + 12 = 81$. Next, we make the Table:

	A	B	C	D	E
Obs. Frq., O_k	27	22	10	10	12
Exp. Frq., E_k	$81 \cdot 0.15$	$81 \cdot 0.25$	$81 \cdot 0.2$	$81 \cdot 0.15$	$81 \cdot 0.25$

Example

Example: I am claiming that, for my Stat courses, the percentage of A-grade students is 15%, of B-grade students is 25%, of C-grades are 20%, for D I have 15%, and all others are Failing the course. Now, one of my courses finished with the following result:

$$\#A = 27, \ \#B = 22, \ \#C = 10, \ \#D = 10, \ \#F = 12.$$

Is this Data supporting my claim?

Solution: We have $n = 27 + 22 + 10 + 10 + 12 = 81$. Next, we make the Table:

	A	B	C	D	E
Obs. Frq., O_k	27	22	10	10	12
Exp. Frq., E_k	$81 \cdot 0.15$	$81 \cdot 0.25$	$81 \cdot 0.2$	$81 \cdot 0.15$	$81 \cdot 0.25$

Now, we can calculate the TS:

$$\chi^2 = \sum_{k=1}^{5} \frac{(O_k - E_k)^2}{E_k} = \frac{(27 - 81 \cdot 0.15)^2}{81 \cdot 0.15} + ... + \frac{(12 - 81 \cdot 0.15)^2}{81 \cdot 0.15}$$

Example, Cont'd

The rest is in **R**:

```
obsd <-  c(27, 22, 10, 10, 12)
expd <- 81* c(0.15, 0.25, 0.2, 0.15, 0.25)
xi2 <- sum((obsd-expd)^2/expd)
xi2
```

```
## [1] 24.41564
```

```
q <- qchisq(1-0.05, df = length(obsd)-1)
q
```

```
## [1] 9.487729
```

```
xi2 > q
```

```
## [1] TRUE
```

Example, Cont'd

```r
obsd <-  c(27, 22, 10, 10, 12)
p <-  c(0.15, 0.25, 0.2, 0.15, 0.25)
chisq.test(obsd, p = p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obsd
## X-squared = 24.416, df = 4, p-value = 6.592e-05
```

Kolmogorov-Smirnov Test

```r
x <- rnorm(50, mean = 3, sd = 1)
ks.test(x, y = "pnorm", mean = 0, sd = 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.87385, p-value = 8.882e-16
## alternative hypothesis: two-sided
```

Kolmogorov-Smirnov Test

```r
x <- rnorm(50, mean = 3, sd = 1)
ks.test(x, y = "pnorm", mean = 0, sd = 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.87385, p-value = 8.882e-16
## alternative hypothesis: two-sided
```

```r
x <- rexp(50, rate = 3.1)
ks.test(x, y = "pnorm", mean = 0, sd = 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.50157, p-value = 3.672e-12
## alternative hypothesis: two-sided
```

Example

```r
x <- runif(40)
y <- rexp(30)
ks.test(x,y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.54167, p-value = 4.101e-05
## alternative hypothesis: two-sided
```

Shapiro-Wilk test

```r
x <- rnorm(25, mean = -2, sd = 10)
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.94969, p-value = 0.2467
```

Fitting a Distribution Family, **R**

```r
library(MASS)
x <- rexp(45, rate = 3.4254)
fitdistr(x, densfun = "exponential")
```

```
##      rate
##   4.8935081
##  (0.7294811)
```

I FSYO!