Example 4.1

Kevin Cummiskey

October 15, 2019

Example 4.1 Recovering Polyphenols (pg 272)

Lesson 16 - Simple Linear Regression

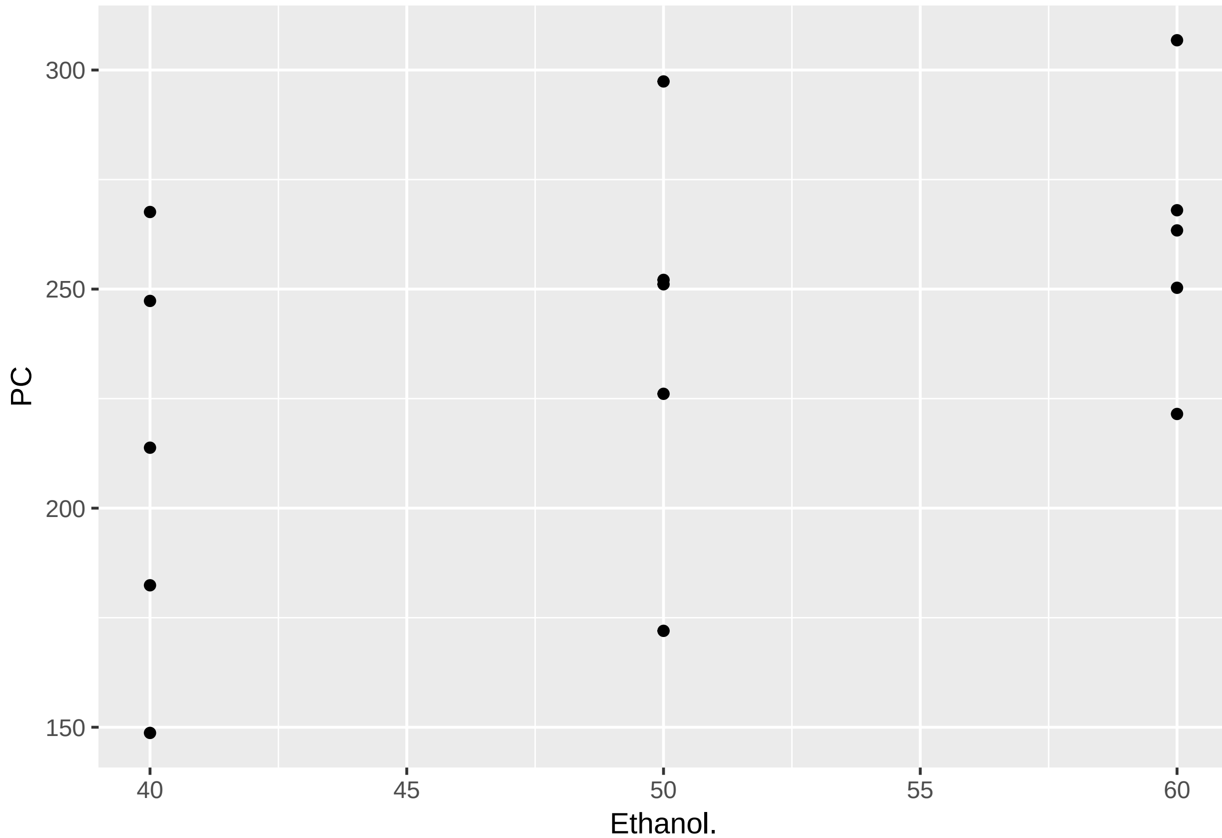
Objectives:

1. Describe association between two quantivative variables.
2. Interpret least-squares regression models.
3. Compare and contrast separate versus linear regression models.

What is the observational unit? explanatory variable? response variable?

What type of study is this?

```
grapes = read.table(file = "http://www.isi-stats.com/isi2/data/Polyphenols.txt", header = T)
grapes %>% ggplot(aes(x = Ethanol., y = PC)) + geom_point()
```



Write a statistical model for a separate mean for each ethanol level.

Fit the model.

```r
grapes$Ethanol_cat = factor(grapes$Ethanol.)
contrasts(grapes$Ethanol_cat) = contr.sum
model_anova = lm(PC ~ Ethanol_cat, data = grapes)
summary(model_anova)
```

```
##
## Call:
## lm(formula = PC ~ Ethanol_cat, data = grapes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.74 -21.60   1.84  23.85  57.66
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    237.90      10.91  21.805 5.08e-11 ***
## Ethanol_cat1   -25.94      15.43  -1.681    0.119
## Ethanol_cat2     1.84      15.43   0.119    0.907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.26 on 12 degrees of freedom
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.09794
```

```
## F-statistic:  1.76 on 2 and 12 DF,  p-value: 0.2137
```

```r
anova(model_anova)
```

```
## Analysis of Variance Table
##
## Response: PC
##             Df  Sum Sq Mean Sq F value Pr(>F)
## Ethanol_cat  2  6285.4  3142.7    1.76 0.2137
## Residuals   12 21427.1  1785.6
```

What would you conclude from this model?

Instead, let's say we fit the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$$

where $i = 1, \ldots, 15$ is grape, x_i is the ethanol concentration used on the ith grape, and y_i is the PC of the ith grape.

How is this model different that the ANOVA model?

Let's fit a regression model.

```r
model_regression = lm(PC ~ Ethanol., data = grapes)
summary(model_regression)
```

```
##
## Call:
## lm(formula = PC ~ Ethanol., data = grapes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65.90 -21.55   0.92  24.31  59.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  112.800     65.081   1.733   0.1067
## Ethanol.       2.502      1.285   1.948   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.62 on 13 degrees of freedom
## Multiple R-squared:  0.2259, Adjusted R-squared:  0.1663
```

```
## F-statistic: 3.793 on 1 and 13 DF,  p-value: 0.07338
```

```r
anova(model_regression)
```

```
## Analysis of Variance Table
##
## Response: PC
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Ethanol.   1   6260  6260.0  3.7935 0.07338 .
## Residuals 13  21453  1650.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What would you conclude from this output? What has changed from the ANOVA model?

Lesson 17 - Inference for Simple Linear Regression

Objectives:

1. Simulation-based inference for relationship between quantiative variables.
2. Theory-based approach for relationship between quantiative variables.
3. Evaluate validity conditions for theory-based tests.

Why do we want to conduct inference?

Write the null and alternative hypothesis to test whether there is an association between ethanol concentration used and PC.

Simulation-based approach

Describe how we would conduct a simulation to conduct this test.

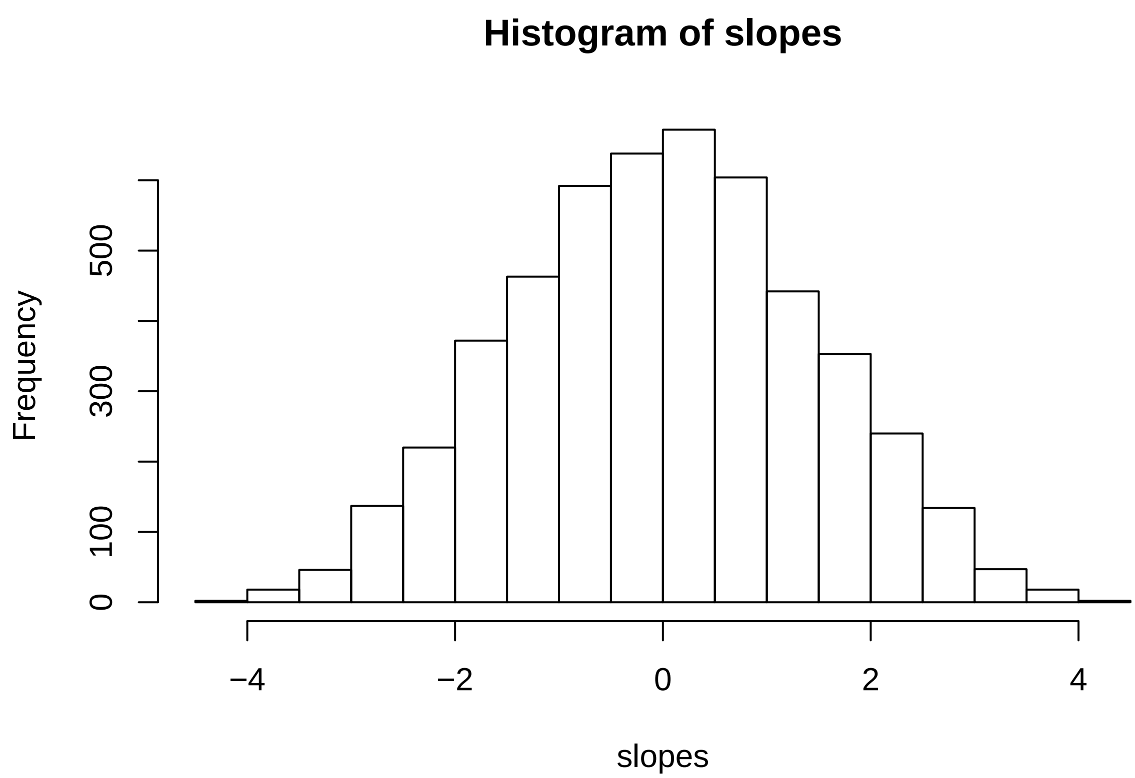
Let's conduct a simulation-based test (pg 291).

```r
m = 5000 #number of iterations
slopes = c() # empty vector
grapes.sim = grapes #copy of data

for(i in 1:m){
  grapes.sim$PC.sim = sample(grapes.sim$PC) #shuffle the response
  model.sim = lm(PC.sim ~ Ethanol., data = grapes.sim) # fit model to shuffled data
  slopes[i] = coef(model.sim)[2] # extract the slope from the model
}
```

Here is a plot of the distribution of the simulated slopes:

```r
hist(slopes)
```



Here is the *p*-value:

```r
sum(slopes > coef(model_regression)[2])/m
```

```
## [1] 0.0402
```

What would we conclude?

Theory-based test

The p-value is in the linear model object:

```
summary(model_regression)
```

```
##
## Call:
## lm(formula = PC ~ Ethanol., data = grapes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65.90 -21.55   0.92  24.31  59.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  112.800     65.081   1.733   0.1067
## Ethanol.       2.502      1.285   1.948   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.62 on 13 degrees of freedom
## Multiple R-squared:  0.2259, Adjusted R-squared:  0.1663
## F-statistic: 3.793 on 1 and 13 DF,  p-value: 0.07338
```

A t confidence interval for the population slope β_1 is (pg295):

```
b1 = summary(model_regression)$coefficients[2,1]
se_b1 = summary(model_regression)$coefficients[2,2]
tstar = qt(0.975,13)

upper = b1 + tstar * se_b1
lower = b1 - tstar * se_b1

lower
```

```
## [1] -0.2732065
```

```
upper
```

```
## [1] 5.277207
```

```
# or you can just do
confint(model_regression)
```

```
##                   2.5 %     97.5 %
## (Intercept) -27.7982917 253.398292
## Ethanol.     -0.2732065   5.277207
```