

Modeling in R and Tidying Results

linear models and broom

2019-08-21

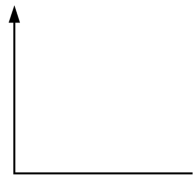
 **This is not a course in a
regression**

Modeling in R

```
lm(y ~ x + z, data = df)
```

Modeling in R

```
lm(y ~ x + z, data = df)
```



model
function

Modeling in R

```
lm(y ~ x + z, data = df)
```

variables
in your
data



your
data



Modeling in R

```
lm(y ~ x + z, data = df)
```

outcome

predictors
or
exposure

Modeling in R

lm() = **Linear Regression (OLS)**

Modeling in R

`lm()` = Linear Regression (OLS)

`glm()` = Generalized Linear Model
(default family = Gaussian)

Modeling in R

```
lm(price ~ carat, data = diamonds)
```

Modeling in R

```
lm(price ~ carat, data = diamonds)
```

```
##
```

```
## Call:
```

```
## lm(formula = price ~ carat, data = diamonds)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          carat
```

```
##          -2256          7756
```

Modeling in R

```
lm(price ~ carat, data = diamonds) %>%  
summary()
```

Modeling in R

```
##  
## Call:  
## lm(formula = price ~ carat, data = diamonds)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -18585.3   -804.8    -18.9    537.4   12731.7   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  -2256.36      13.06  -172.8  <2e-16 ***   
## carat         7756.43      14.07   551.4  <2e-16 ***   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1549 on 53938 degrees of freedom  
## Multiple R-squared:  0.8493,    Adjusted R-squared:  0.8493   
## F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

Modeling in R

```
##  
## Call:  
## lm(formula = price ~ carat, data = diamonds)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -18585.3   -804.8    -18.9    537.4   12731.7   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  -2256.36      13.06  -172.8  <2e-16 ***   
## carat         7756.43      14.07   551.4  <2e-16 ***   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1549 on 53938 degrees of freedom  
## Multiple R-squared:  0.8493,    Adjusted R-squared:  0.8493   
## F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

broom: tidy models

`tidy()`

`glance()`

`augment()`



broom: tidy models

tidy() = **model coefficients**

glance()

augment()



broom: tidy models

tidy()

glance() = **model fit**

augment()



broom: tidy models

tidy()

glance()

augment() = model predictions



broom: tidy models

tidy()

glance()

augment()

NOT a core member of the tidyverse. Need to load
with library(broom)



Modeling in R

```
library(broom)  
lm(price ~ carat, data = diamonds) %>%  
  tidy()
```

Modeling in R

```
library(broom)
lm(price ~ carat, data = diamonds) %>%
  tidy()
```

```
## # A tibble: 2 x 5
```

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-2256.	13.1	-173.	0
2	carat	7756.	14.1	551.	0

Modeling in R

```
lm(price ~ carat, data = diamonds) %>%  
  glance()
```

Modeling in R

```
lm(price ~ carat, data = diamonds) %>%  
  glance()
```

```
## # A tibble: 1 x 11  
##   r.squared adj.r.squared sigma statistic p.value    df  
##   <dbl>      <dbl> <dbl>      <dbl>   <dbl> <int>  
## 1    0.849        0.849 1549.    304051.     0     2  
## # ... with 5 more variables: logLik <dbl>, AIC <dbl>,  
## #   BIC <dbl>, deviance <dbl>, df.residual <int>
```

Modeling in R

```
lm(price ~ carat, data = diamonds) %>%  
  augment()
```

Modeling in R

```
lm(price ~ carat, data = diamonds) %>%  
  augment()
```

```
## # A tibble: 53,940 x 9  
##   price carat .fitted .se.fit .resid   .hat .sigma .cooksd  
##   <int> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>   <dbl>  
## 1    326 0.23   -472.    10.4   798.  4.52e-5 1549.  6.00e-6  
## 2    326 0.21   -628.    10.6   954.  4.71e-5 1549.  8.92e-6  
## 3    327 0.23   -472.    10.4   799.  4.52e-5 1549.  6.02e-6  
## 4    334 0.290   -7.00     9.77   341.  3.98e-5 1549.  9.66e-7  
## 5    335 0.31    148.     9.57   187.  3.82e-5 1549.  2.78e-7  
## 6    336 0.24   -395.    10.3   731.  4.42e-5 1549.  4.93e-6  
## 7    336 0.24   -395.    10.3   731.  4.42e-5 1549.  4.93e-6  
## 8    337 0.26   -240.    10.1   577.  4.24e-5 1549.  2.94e-6  
## 9    337 0.22   -550.    10.5   887.  4.61e-5 1549.  7.56e-6  
## 10   338 0.23   -472.    10.4   810.  4.52e-5 1549.  6.18e-6  
## # ... with 53,930 more rows, and 1 more variable:  
## #   .std.resid <dbl>
```


Try it yourself

Work your way through the exercises. If anything in particular is giving you trouble, we'll work through it together.

Resources

R for Data Science: A comprehensive but friendly introduction to the tidyverse.
Free online.

UCLA IDRE: Useful resources on modeling in R and other languages