Anova and Diagnostics

Emmenta Janneh

2024-02-17

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.3.2
```

```
library(car) # companion of applied regression
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
hibbs <- read_csv("https://dcgerard.github.io/stat_415_615/data/hibbs.csv")
```

```
## Rows: 16 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): inc_party_candidate, other_candidate
## dbl (3): year, growth, vote
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

To do an ANOVA approach, first fit the linear model, then use the `Anova()` function from the `car` package.

```
lm_hibbs <- lm(vote ~ growth, data = hibbs)
Anova(lm_hibbs)
```

```
## Anova Table (Type II tests)
##
## Response: vote
##           Sum Sq Df F value  Pr(>F)
## growth    273.63  1  19.321 0.00061 ***
## Residuals 198.27 14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Value	Meaning	More
273.63	SSE(R) - SSE(F)	SSR also SSTO
198.27	SSE(F)	SSE
1	df(R) - df(F)	
14	df(F)	
19.321	F-Statics	
0.00061	P-Value	

Var = SSE(F) / df(F)

R has an `anova()` function, and we will use it, nut it is less useful in multiple linear regression.

```
anova(lm_hibbs)
```

```
## Analysis of Variance Table
##
## Response: vote
##           Df Sum Sq Mean Sq F value  Pr(>F)
## growth     1 273.63 273.632  19.321 0.00061 ***
## Residuals 14 198.27  14.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use `glance()` to get R^2 in R

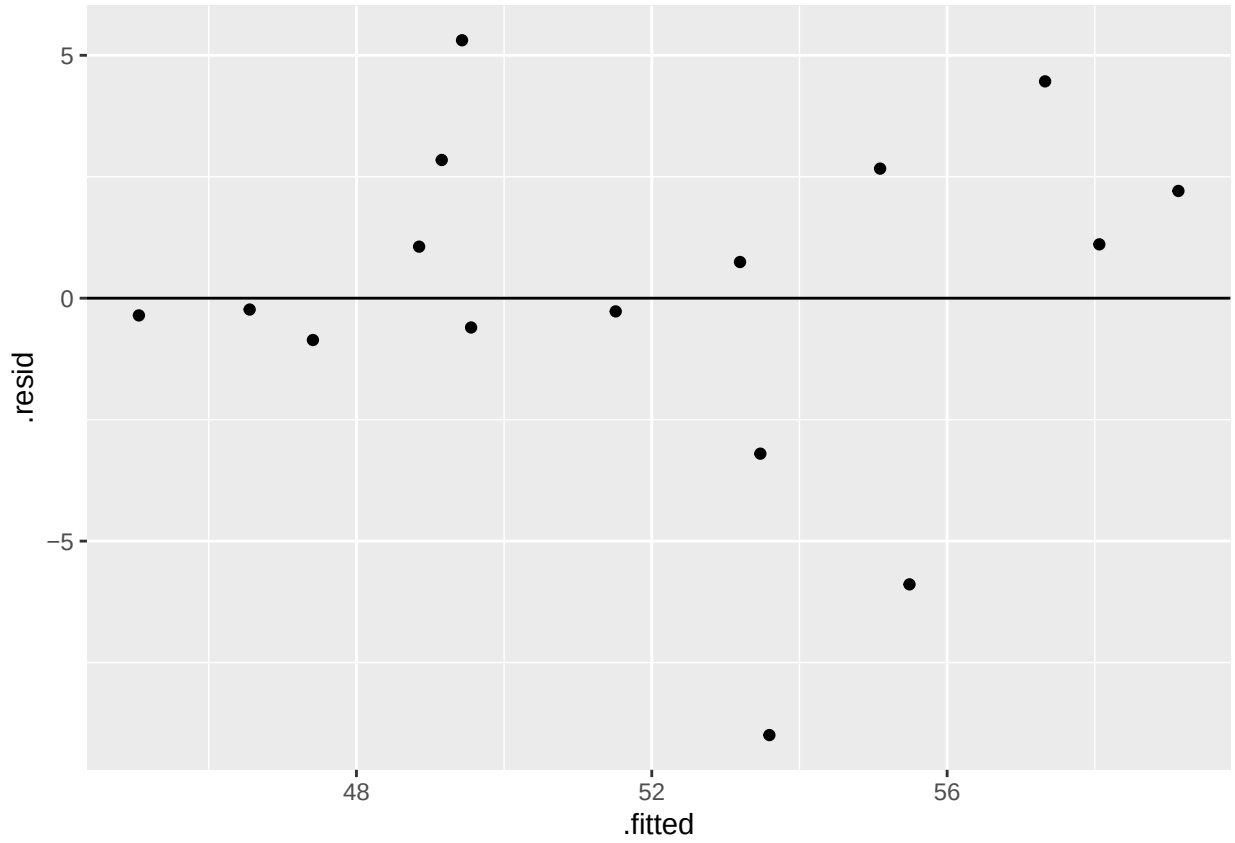
```
glance(lm_hibbs)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.580         0.550  3.76      19.3 0.000610     1  -42.8  91.7  94.0
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

57.98% of the variability in incumbent vote-share was explained by economic growth.

To make a fits vers residuals plot, get the fitted values and the residuals with `augment()`. Make a scatterplot of `.fitted` versus the `.resid` variables

```
a_hibbs <- augment(lm_hibbs)
ggplot(a_hibbs, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



For sandwich estimate, use the `lmtest` and `sandwich` packages

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.3.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.2
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.3.2
```

```
## vcon. tells the software how to estimate the standard errors
## vconHC is a function to estimate sandwich standard errors (Heteroskedastic Consistent)
cout <- coeftest(lm_hibbs, vcon. = vcovHC)
tidy(cout)
```

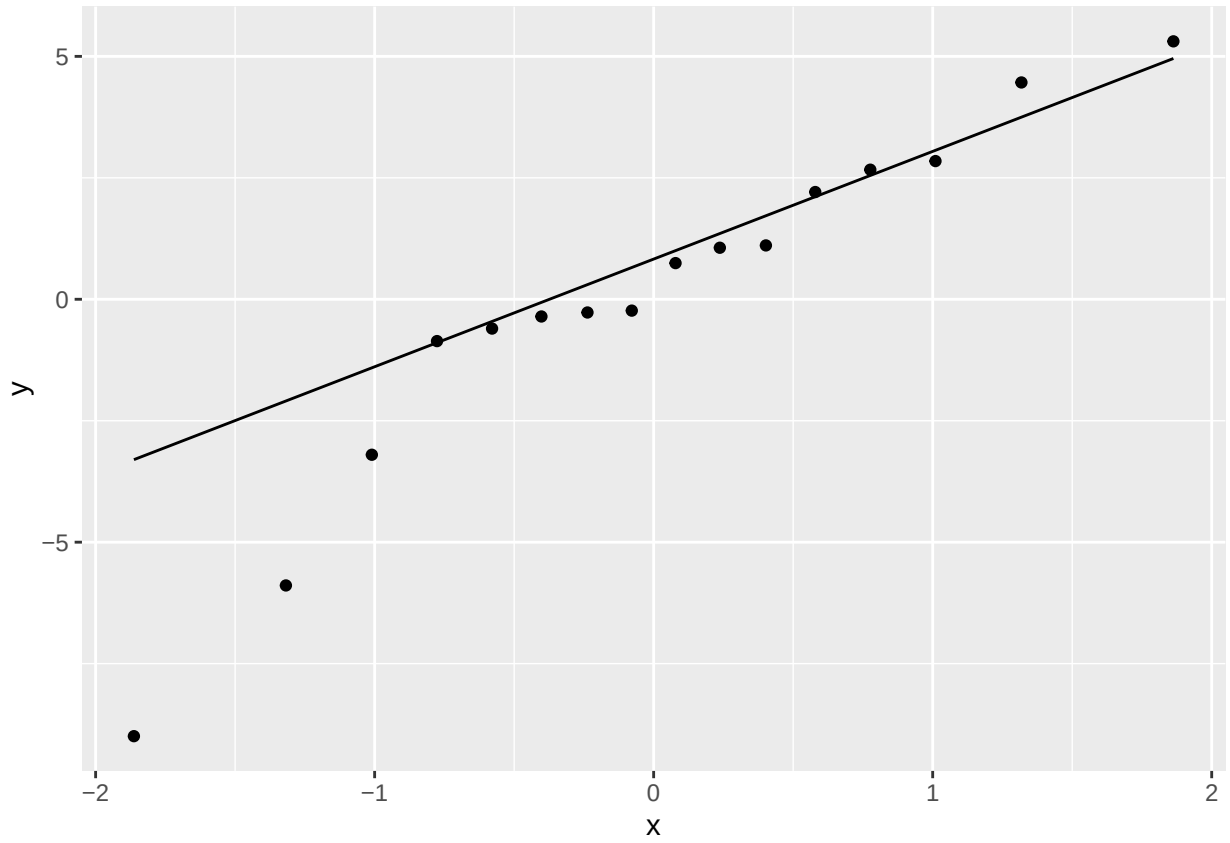
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.2      1.62      28.5  8.41e-14
## 2 growth          3.06     0.696      4.40 6.10e- 4
```

```
tidy(lm_hibbs)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.2      1.62      28.5  8.41e-14
## 2 growth          3.06     0.696      4.40 6.10e- 4
```

Only make qq-plots of residuals. To do so, use the `sample` aesthetic in `ggplot()`, and use `geom_qq()`.

```
ggplot(a_hibbs, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```



Only check normality last, because other violations will result in non-normal looking QQ-plots.