QCB 508 – Week 9

John D. Storey

Spring 2020

Contents

OLS in R 3
- Example: Davis Data .......................................................... 3
- Weight Regressed on Height + Sex ........................................ 3
- Residual Distribution .......................................................... 4
- Normal Residuals Check ...................................................... 4
- One Variable, Two Scales ................................................... 5
- Interactions ..................................................................... 6
- More on Interactions .......................................................... 6
- Visualizing Three Different Models ....................................... 7

Categorical Explanatory Variables 7
- Example: Chicken Weights .................................................. 7
- Factor Variables in \texttt{lm()} .............................................. 8
- Plot the Fit .................................................................. 8
- ANOVA (Version 1) .......................................................... 9
- \texttt{anova()} ................................................................ 9
- How It Works ................................................................ 10
- Top of Design Matrix ....................................................... 10
- Bottom of Design Matrix .................................................. 11
- Model Fits ................................................................... 11
- Another ANOVA Function .................................................. 11

Variable Transformations 12
- Rationale ................................................................... 12
- Power and Log Transformations ........................................... 12
- Diamonds Data ................................................................. 12
- Nonlinear Relationship ...................................................... 12
- Regression with Nonlinear Relationship .............................. 13
- Residual Distribution ........................................................ 13
- Normal Residuals Check .................................................... 14
- Log-Transformation ........................................................... 15
- OLS on Log-Transformed Data ............................................ 16
- Residual Distribution ........................................................ 16
- Normal Residuals Check .................................................... 17
- Tree Pollen Study ............................................................... 18
- Tree Pollen Count by Week ............................................... 19
- A Clever Transformation .................................................... 19
- \texttt{week} Transformed ..................................................... 19

OLS Goodness of Fit: Theory 20
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pythagorean Theorem</td>
<td>20</td>
</tr>
<tr>
<td>OLS Normal Model</td>
<td>21</td>
</tr>
<tr>
<td>Projection Matrices</td>
<td>21</td>
</tr>
<tr>
<td>Decomposition</td>
<td>21</td>
</tr>
<tr>
<td>Distribution of Projection</td>
<td>22</td>
</tr>
<tr>
<td>Distribution of Residuals</td>
<td>22</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>22</td>
</tr>
<tr>
<td>Submodels</td>
<td>22</td>
</tr>
<tr>
<td>Hypothesis Testing</td>
<td>22</td>
</tr>
<tr>
<td>Generalized LRT</td>
<td>23</td>
</tr>
<tr>
<td>Nested Projections</td>
<td>23</td>
</tr>
<tr>
<td>$F$ Statistic</td>
<td>23</td>
</tr>
<tr>
<td>$F$ Distribution</td>
<td>24</td>
</tr>
<tr>
<td>$F$ Test</td>
<td>24</td>
</tr>
<tr>
<td>OLS Goodness of Fit: R</td>
<td>24</td>
</tr>
<tr>
<td>Example: Davis Data</td>
<td>24</td>
</tr>
<tr>
<td>Comparing Linear Models in R</td>
<td>24</td>
</tr>
<tr>
<td>ANOVA (Version 2)</td>
<td>25</td>
</tr>
<tr>
<td>Comparing Two Models with <code>anova()</code></td>
<td>25</td>
</tr>
<tr>
<td>When There’s a Single Variable Difference</td>
<td>25</td>
</tr>
<tr>
<td>Calculating the F-statistic</td>
<td>25</td>
</tr>
<tr>
<td>Calculating the Generalized LRT</td>
<td>26</td>
</tr>
<tr>
<td>ANOVA on More Distant Models</td>
<td>26</td>
</tr>
<tr>
<td>Compare Multiple Models at Once</td>
<td>27</td>
</tr>
<tr>
<td>Monogenic Trait Model</td>
<td>27</td>
</tr>
<tr>
<td>Genotypes Under HWE</td>
<td>27</td>
</tr>
<tr>
<td>Inbreeding</td>
<td>27</td>
</tr>
<tr>
<td>Kinship</td>
<td>27</td>
</tr>
<tr>
<td>Kinship Examples</td>
<td>28</td>
</tr>
<tr>
<td>Covariance of Genotypes</td>
<td>28</td>
</tr>
<tr>
<td>Additive Trait Model</td>
<td>28</td>
</tr>
<tr>
<td>General Trait Model</td>
<td>28</td>
</tr>
<tr>
<td>OLS and Dependence</td>
<td>28</td>
</tr>
<tr>
<td>Variance of Trait</td>
<td>29</td>
</tr>
<tr>
<td>Variance Decomposition</td>
<td>29</td>
</tr>
<tr>
<td>Covariance of Trait</td>
<td>29</td>
</tr>
<tr>
<td>Multivariate Distribution of Trait</td>
<td>29</td>
</tr>
<tr>
<td>Heritability</td>
<td>30</td>
</tr>
<tr>
<td>Polygenic Trait Model</td>
<td>30</td>
</tr>
<tr>
<td>Fisher (1918)</td>
<td>30</td>
</tr>
<tr>
<td>Assumptions</td>
<td>30</td>
</tr>
<tr>
<td>Variance of Trait</td>
<td>30</td>
</tr>
<tr>
<td>Heritability</td>
<td>31</td>
</tr>
<tr>
<td>Covariance of Trait</td>
<td>31</td>
</tr>
<tr>
<td>Normal Approximation</td>
<td>31</td>
</tr>
<tr>
<td>Variance Components</td>
<td>31</td>
</tr>
<tr>
<td>Association Testing</td>
<td>32</td>
</tr>
<tr>
<td>Extras</td>
<td>32</td>
</tr>
<tr>
<td>Source</td>
<td>32</td>
</tr>
<tr>
<td>Session Information</td>
<td>32</td>
</tr>
</tbody>
</table>
OLS in R

Example: Davis Data

```r
> data("Davis", package="carData")
> htwt <- tbl_df(Davis)
> htwt[12,c(2,3)] <- htwt[12,c(3,2)]
> head(htwt)

# A tibble: 6 x 5
  sex weight height repwt repht
   <fct> <int> <int> <int> <int>
1   M     77   182    77    180
2   F     58   161    51    159
3   F     53   161    54    158
4   M     68   177    70    175
5   F     59   157    59    155
6   M     76   170    76    165
```

R implements OLS of multiple explanatory variables exactly the same as with a single explanatory variable, except we need to show the sum of all explanatory variables that we want to use.

```r
> lm(weight ~ height + sex, data=htwt)

Call:
  lm(formula = weight ~ height + sex, data = htwt)

Coefficients:
(Intercept)    height         sexM
     -76.6167     0.81058     8.22689

Weight Regressed on Height + Sex

```r
> summary(lm(weight ~ height + sex, data=htwt))

Call:
  lm(formula = weight ~ height + sex, data = htwt)

Residuals:
     Min      1Q  Median      3Q     Max
-20.1310  -4.8840  -0.6399  5.1600  41.4900

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -76.6167    15.7150  -4.875  2.23e-06 ***
height                        0.8106     0.0953   8.506  4.50e-15 ***
sexM                           8.2269    1.7105   4.810  3.00e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.066 on 197 degrees of freedom
Multiple R-squared: 0.6372,  Adjusted R-squared: 0.6335
F-statistic: 173 on 2 and 197 DF,  p-value: < 2.2e-16
```
Residual Distribution

```r
myfit <- lm(weight ~ height + sex, data=htwt)
plot(myfit, which=1)
```

Normal Residuals Check

```r
plot(myfit, which=2)
```
One Variable, Two Scales

We can include a single variable but on two different scales:

```r
> htwt <- htwt %>% mutate(height2 = height^2)
> summary(lm(weight ~ height + height2, data=htwt))
```

Call:
```
  lm(formula = weight ~ height + height2, data = htwt)
```

Residuals:
```
    Min     1Q  Median     3Q    Max
-24.265 -5.159  -0.499  4.549  42.965
```

Coefficients:
```
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.117140 175.246872 0.611 0.542
height     -1.632719   2.045524 -0.798 0.426
height2    0.008111   0.005959 1.361 0.175
```
Interactions

It is possible to include products of explanatory variables, which is called an interaction.

```r
> summary(lm(weight ~ height + sex + height:sex, data=htwt))
```

```
Call:
lm(formula = weight ~ height + sex + height:sex, data = htwt)

Residuals:
     Min      1Q  Median      3Q     Max
-20.869  -4.835  -0.897   4.429  41.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.6730    22.1342  -2.063  0.0404 *
height      0.6227     0.1343   4.637  6.46e-06 ***
sexM        -55.6571    32.4597  -1.715  0.0880 .
height:sexM  0.3729     0.1892   1.971  0.0502 .
---
Signif. codes:  * 0.05  ** 0.01  *** 0.001

Residual standard error: 8.007 on 196 degrees of freedom
Multiple R-squared: 0.6442, Adjusted R-squared: 0.6388
F-statistic: 118.3 on 3 and 196 DF, p-value: < 2.2e-16
```

More on Interactions

What happens when there is an interaction between a quantitative explanatory variable and a factor explanatory variable? In the next plot, we show three models:

- Grey solid: `lm(weight ~ height, data=htwt)`
- Color dashed: `lm(weight ~ height + sex, data=htwt)`
- Color solid: `lm(weight ~ height + sex + height:sex, data=htwt)`
Categorical Explanatory Variables

Example: Chicken Weights

```r
> data("chickwts", package="datasets")
> head(chickwts)
weight   feed
1   179 horsebean
2   160 horsebean
3   136 horsebean
4   227 horsebean
5   217 horsebean
6   168 horsebean
> summary(chickwts$feed)
 casein horsebean linseed meatmeal soybean sunflower
    12     10     12     11     14     12
```
Factor Variables in \texttt{lm()}

\begin{verbatim}
> chick_fit <- \texttt{lm(weight ~ feed, data=chickwts)}
> summary(chick_fit)

Call:
\texttt{lm(formula = weight ~ feed, data = chickwts)}

Residuals:
\begin{tabular}{rrrrr}
Min & 1Q & Median & 3Q & Max \\
-123.909 & -34.413 & 1.571 & 38.170 & 103.091 \\
\end{tabular}

Coefficients:
\begin{tabular}{rrrrr}
 & Estimate & Std. Error & \texttt{t value} & Pr(\texttt{>|t|}) \\
(Intercept) & 323.583 & 15.834 & 20.436 & < 2e-16 *** \\
feedhorsebean & -163.383 & 23.485 & -6.957 & 2.07e-09 *** \\
feedlinseed & -104.833 & 22.393 & -4.682 & 1.49e-05 *** \\
feedmeatmeal & -46.674 & 22.896 & -2.039 & 0.045567 * \\
feedsoybean & -77.155 & 21.578 & -3.576 & 0.000665 *** \\
feedsunflower & 5.333 & 22.393 & 0.238 & 0.812495 \\
\end{tabular}

---

Signif. codes: \texttt{0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1}

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared: 0.5417, Adjusted R-squared: 0.5064
F-statistic: 15.36 on 5 and 65 DF, \texttt{p-value: 5.936e-10}
\end{verbatim}

Plot the Fit

\begin{verbatim}
> plot(chickwts$feed, chickwts$weight, xlab="Feed", ylab="Weight", las=2)
> points(chickwts$feed, chick_fit$fitted.values, col="blue", pch=20, cex=2)
\end{verbatim}
ANOVA (Version 1)

ANOVA (analysis of variance) was originally developed as a statistical model and method for comparing differences in mean values between various groups.

ANOVA quantifies and tests for differences in response variables with respect to factor variables.

In doing so, it also partitions the total variance to that due to within and between groups, where groups are defined by the factor variables.

anova()

The classic ANOVA table:

```
> anova(chick_fit)
Analysis of Variance Table

Response: weight
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
feed        5 231129 46226.0 15.36544 5.936e-10 ***
Residuals  65 195556  3009.0
```
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> n <- length(chick_fit$residuals) # n <- 71
> (n-1)*var(chick_fit$fitted.values)
[1] 231129.2
> (n-1)*var(chick_fit$residuals)
[1] 195556
> (n-1)*var(chickwts$weight) # sum of above two quantities
[1] 426685.2
> (231129/5)/(195556/65) # F-statistic
[1] 15.36479

How It Works

> levels(chickwts$feed)
[1] "casein"  "horsebean" "linseed"  "meatmeal" "soybean"  "sunflower"
> head(chickwts, n=3)
  weight feed
1  179  horsebean
2  160  horsebean
3  136  horsebean
> tail(chickwts, n=3)
  weight feed
69 222  casein
70 283  casein
71 332  casein
> x <- model.matrix(weight ~ feed, data=chickwts)
> dim(x)
[1] 71  6

Top of Design Matrix

> head(x)
   (Intercept) feedhorsebean feedlinseed feedmeatmeal feedsoybean feedsunflower
1       1          1          1            0            0            0
2       1          1          1            0            0            0
3       1          1          1            0            0            0
4       1          1          1            0            0            0
5       1          1          1            0            0            0
6       1          1          1            0            0            0

---
Bottom of Design Matrix

```r
> tail(x)

  (Intercept) feedhorsebean feedlinseed feedmeatmeal feedsoybean feedsunflower
66       1        0           0           0          0         0
67       1        0           0           0          0         0
68       1        0           0           0          0         0
69       1        0           0           0          0         0
70       1        0           0           0          0         0
71       1        0           0           0          0         0
```

Model Fits

```r
> chick_fit$fitted.values %>% round(digits=4) %>% unique()

```

```r
> chickwts %>% group_by(feed) %>% summarize(mean(weight))

# A tibble: 6 x 2
  feed     mean(weight)
  <fct>           <dbl>
1 casein         324.
2 horsebean      160.
3 linseed        219.
4 meatmeal       277.
5 soybean        246.
6 sunflower      329.
```

Another ANOVA Function

```r
> aov(weight ~ feed, data=chickwts)
Call:
aov(formula = weight ~ feed, data = chickwts)

Terms:
  feed Residuals
Sum of Squares 231129.2 195556.0
Deg. of Freedom 5 65

Residual standard error: 54.85029
Estimated effects may be unbalanced
```

```r
> summary(aov(weight ~ feed, data=chickwts))

Df  Sum Sq Mean Sq F value Pr(>F)
feed     5 231129  46226 15.37 5.94e-10 ***
Residuals 65 195556  3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Variable Transformations

Rationale

In order to obtain reliable model fits and inference on linear models, the model assumptions described earlier must be satisfied.

Sometimes it is necessary to transform the response variable and/or some of the explanatory variables.

This process should involve data visualization and exploration.

Power and Log Transformations

It is often useful to explore power and log transforms of the variables, e.g., \( \log(y) \) or \( y^\lambda \) for some \( \lambda \) (and likewise \( \log(x) \) or \( x^\lambda \)).

You can read more about the Box-Cox family of power transformations.

Diamonds Data

```r
> data("diamonds", package="ggplot2")
> head(diamonds)
```

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>carat</td>
<td>cut</td>
<td>color</td>
<td>clarity</td>
<td>depth</td>
<td>table</td>
<td>price</td>
<td>x</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>0.23</td>
<td>Ideal</td>
<td>E</td>
<td>SI2</td>
<td>61.5</td>
<td>55</td>
<td>326</td>
<td>3.95</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>Prem-</td>
<td>E</td>
<td>SI1</td>
<td>59.8</td>
<td>61</td>
<td>326</td>
<td>3.89</td>
</tr>
<tr>
<td>3</td>
<td>0.23</td>
<td>Good</td>
<td>E</td>
<td>VS1</td>
<td>56.9</td>
<td>65</td>
<td>327</td>
<td>4.05</td>
</tr>
<tr>
<td>4</td>
<td>0.290</td>
<td>Prem-</td>
<td>I</td>
<td>VS2</td>
<td>62.4</td>
<td>58</td>
<td>334</td>
<td>4.2</td>
</tr>
<tr>
<td>5</td>
<td>0.31</td>
<td>Good</td>
<td>J</td>
<td>SI2</td>
<td>63.3</td>
<td>58</td>
<td>335</td>
<td>4.34</td>
</tr>
<tr>
<td>6</td>
<td>0.24</td>
<td>Very-</td>
<td>J</td>
<td>VVS2</td>
<td>62.8</td>
<td>57</td>
<td>336</td>
<td>3.94</td>
</tr>
</tbody>
</table>

Nonlinear Relationship

```r
> ggplot(data = diamonds) +
  geom_point(mapping=aes(x=carat, y=price, color=clarity), alpha=0.3)
```
Regression with Nonlinear Relationship

```r
> diam_fit <- lm(price ~ carat + clarity, data=diamonds)
> anova(diam_fit)
Analysis of Variance Table
Response: price
        Df Sum Sq Mean Sq F value  Pr(>F)
carat    1 7.291e+11 7.291e+11 435639.9 < 2.2e-16 ***
clarity  7 3.908e+10  5.583e+09  3335.8 < 2.2e-16 ***
Residuals 53931 9.026e+10  1.673e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Distribution

```r
> plot(diam_fit, which=1)
```
Normal Residuals Check

\texttt{> plot(diam_fit, which=2)}
Log-Transformation

```r
> ggplot(data = diamonds) +
+   geom_point(aes(x=carat, y=price, color=clarity), alpha=0.3) +
+   scale_y_log10(breaks=c(1000, 5000, 10000)) +
+   scale_x_log10(breaks=1:5)
```
OLS on Log-Transformed Data

```r
> diamonds <- mutate(diamonds, log_price = log(price, base=10),
+                      log_carat = log(carat, base=10))
> ldiam_fit <- lm(log_price ~ log_carat + clarity, data=diamonds)
> anova(ldiam_fit)
Analysis of Variance Table

Response: log_price

                      Df Sum Sq Mean Sq F value    Pr(>F)
log_carat             1 9771.9  9771.9 1452922.6 < 2.2e-16 ***
clarity              7  339.1   48.4    7203.3 < 2.2e-16 ***
Residuals           53931  362.7   0.0
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' ' 1

Residual Distribution

> plot(ldiam_fit, which=1)
```
Normal Residuals Check

> plot(ldiam_fit, which=2)
Tree Pollen Study

Suppose that we have a study where tree pollen measurements are averaged every week, and these data are recorded for 10 years. These data are simulated:

```r
> pollen_study
# A tibble: 520 x 3
  week year pollen
  <int> <int>  <dbl>
1     1   2001  1842.
3     3   2001  2381.
4     4   2001  2141.
5     5   2001  2210.
6     6   2001  2585.
7     7   2001  2392.
8     8   2001  2105.
9     9   2001  2278.
10    10   2001  2384.
# ... with 510 more rows
```
Tree Pollen Count by Week

```
> ggplot(pollen_study) + geom_point(aes(x=week, y=pollen))
```

A Clever Transformation

We can see there is a linear relationship between pollen and week if we transform week to be number of weeks from the peak week.

```
> pollen_study <- pollen_study %>%
  + mutate(week_new = abs(week-20))
```

Note that this is a very different transformation from taking a log or power transformation.

**week Transformed**

```
> ggplot(pollen_study) + geom_point(aes(x=week_new, y=pollen))
```
OLS Goodness of Fit: Theory

Pythagorean Theorem

![Pythagorean Theorem Diagram](image)

Figure 1: PythMod

Least squares model fitting can be understood through the Pythagorean theorem: \( a^2 + b^2 = c^2 \). However, here we have:
\[ \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]

where the \( \hat{Y}_i \) are the result of a linear projection of the \( Y_i \).

**OLS Normal Model**

In this section, let’s assume that \( (X_1, Y_1), \ldots, (X_n, Y_n) \) are distributed so that

\[ Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + E_i \]

\[ = X_i \beta + E_i \]

where \( E|X \sim \text{MVN}_n(0, \sigma^2 I) \). Note that we haven’t specified the distribution of the \( X_i \) rv’s.

**Projection Matrices**

In the OLS framework we have:

\[ \hat{Y} = X(X^T X)^{-1}X^T Y. \]

The matrix \( P_{n \times n} = X(X^T X)^{-1}X^T \) is a projection matrix. The vector \( Y \) is projected into the space spanned by the column space of \( X \).

Project matrices have the following properties:

- \( P \) is symmetric
- \( P \) is idempotent so that \( PP = P \)
- If \( X \) has column rank \( p \), then \( P \) has rank \( p \)
- The eigenvalues of \( P \) are \( p \) 1’s and \( n - p \) 0’s
- The trace (sum of diagonal entries) is \( \text{tr}(P) = p \)
- \( I - P \) is also a projection matrix with rank \( n - p \)

**Decomposition**

Note that \( P(I - P) = P - PP = P - P = 0 \).

We have

\[ \|Y\|_2^2 = Y^T Y = (PY + (I - P)Y)^T (PY + (I - P)Y) \]

\[ = (PY)^T PY + ((I - P)Y)^T ((I - P)Y) \]

\[ = \|PY\|_2^2 + \|(I - P)Y\|_2^2 \]

where the cross terms disappear because \( P(I - P) = 0 \).

Note: The \( \ell_p \) norm of an \( n \)-vector \( w \) is defined as

\[ \|w\|_p = \left( \sum_{i=1}^{n} |w_i|^p \right)^{1/p} \]

Above we calculated

\[ \|w\|_2^2 = \sum_{i=1}^{n} w_i^2. \]
Distribution of Projection

Suppose that $Y_1, Y_2, \ldots, Y_n \sim \text{Normal}(0, \sigma^2)$. This can also be written as $Y \sim \text{MVN}_n(0, \sigma^2 I)$. It follows that

$$PY \sim \text{MVN}_n(0, \sigma^2 PIP^T).$$

where $PIP^T = PP^T = PP = P$.

Also, $(PY)^T (PY) = Y^T P^T PY = Y^T PY$, a quadratic form. Given the eigenvalues of $P$, $Y^T PY$ is equivalent in distribution to $p$ squared iid Normal(0,1) rv’s, so

$$\frac{Y^T PY}{\sigma^2} \sim \chi^2_p.$$

Distribution of Residuals

If $PY = \hat{Y}$ are the fitted OLS values, then $(I - P)Y = Y - \hat{Y}$ are the residuals.

It follows by the same argument as above that

$$\frac{Y^T (I - P)Y}{\sigma^2} \sim \chi^2_{n-p}.$$

It’s also straightforward to show that $(I - P)Y \sim \text{MVN}_n(0, \sigma^2(I - P))$ and $\text{Cov}(PY, (I - P)Y) = 0$.

Degrees of Freedom

The degrees of freedom, $p$, of a linear projection model fit is equal to

- The number of linearly dependent columns of $X$
- The number of nonzero eigenvalues of $P$ (where nonzero eigenvalues are equal to 1)
- The trace of the projection matrix, $\text{tr}(P)$.

The reason why we divide estimates of variance by $n - p$ is because this is the number of effective independent sources of variation remaining after the model is fit by projecting the $n$ observations into a $p$ dimensional linear space.

Submodels

Consider the OLS model $Y = X\beta + E$ where there are $p$ columns of $X$ and $\beta$ is a $p$-vector.

Let $X_0$ be a subset of $p_0$ columns of $X$ and let $X_1$ be a subset of $p_1$ columns, where $1 \leq p_0 < p_1 \leq p$. Also, assume that the columns of $X_0$ are a subset of $X_1$.

We can form $\hat{Y}_0 = P_0Y$ where $P_0$ is the projection matrix built from $X_0$. We can analogously form $\hat{Y}_1 = P_1Y$.

Hypothesis Testing

Without loss of generality, suppose that $\beta_0 = (\beta_1, \beta_2, \ldots, \beta_{p_0})^T$ and $\beta_1 = (\beta_{p_0+1}, \beta_{p_0+2}, \ldots, \beta_{p_1})^T$.

How do we compare these models, specifically to test $H_0 : (\beta_{p_0+1}, \beta_{p_0+2}, \ldots, \beta_{p_1}) = 0$ vs $H_1 : (\beta_{p_0+1}, \beta_{p_0+2}, \ldots, \beta_{p_1}) \neq 0$?

The basic idea to perform this test is to compare the goodness of fits of each model via a pivotal statistic. We will discuss the generalized LRT and ANOVA approaches.
Generalized LRT
Under the OLS Normal model, it follows that \( \hat{\beta}_0 = (X_0^T X_0)^{-1} X_0^T Y \) is the MLE under the null hypothesis and \( \hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y \) is the unconstrained MLE. Also, the respective MLEs of \( \sigma^2 \) are

\[
\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_{0,i})^2}{n}
\]

\[
\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_{1,i})^2}{n}
\]

where \( \hat{Y}_0 = X_0\hat{\beta}_0 \) and \( \hat{Y}_1 = X_1\hat{\beta}_1 \).

The generalized LRT statistic is

\[
\lambda(X,Y) = \frac{L(\beta_1, \hat{\sigma}_1^2; X,Y)}{L(\hat{\beta}_0, \hat{\sigma}_0^2; X,Y)}
\]

where \( 2 \log \lambda(X,Y) \) has a \( \chi^2_{p_1-p_0} \) null distribution.

Nested Projections
We can apply the Pythagorean theorem we saw earlier to linear subspaces to get:

\[
\|Y\|_2^2 = \|(I - P_1)Y\|_2^2 + \|P_1Y\|_2^2
\]

\[
= \|(I - P_1)Y\|_2^2 + \|P_1 - P_0\|_2^2 + \|P_0Y\|_2^2
\]

We can also use the Pythagorean theorem to decompose the residuals from the smaller projection \( P_0 \):

\[
\|(I - P_0)Y\|_2^2 = \|(I - P_1)Y\|_2^2 + \|(P_1 - P_0)Y\|_2^2
\]

\( F \) Statistic
The \( F \) statistic compares the improvement of goodness in fit of the larger model to that of the smaller model in terms of sums of squared residuals, and it scales this improvement by an estimate of \( \sigma^2 \):

\[
F = \frac{\|P_1 - P_0\|_2^2/(p_1 - p_0)}{\|(I - P_1)Y\|_2^2/(n - p_1)}
\]

\[
= \frac{\sum_{i=1}^{n}(\hat{Y}_{1,i} - \hat{Y}_{0,i})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_{1,i})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_{0,i})^2/(n - p_1)}
\]

Since \( \|P_1 - P_0\|_2^2 - \|(I - P_1)Y\|_2^2 = \|(P_1 - P_0)Y\|_2^2 \), we can equivalently write the \( F \) statistic as:

\[
F = \frac{\|(P_1 - P_0)Y\|_2^2/(p_1 - p_0)}{\|(I - P_1)Y\|_2^2/(n - p_1)}
\]

\[
= \frac{\sum_{i=1}^{n}(\hat{Y}_{1,i} - \hat{Y}_{0,i})^2/(p_1 - p_0)}{\sum_{i=1}^{n}(Y_i - \hat{Y}_{1,i})^2/(n - p_1)}
\]
**F Distribution**

Suppose we have independent random variables $V \sim \chi^2_a$ and $W \sim \chi^2_b$. It follows that

$$\frac{V/a}{W/b} \sim F_{a,b}$$

where $F_{a,b}$ is the $F$ distribution with $(a, b)$ degrees of freedom.

By arguments similar to those given above, we have

$$\frac{\| (P_1 - P_0)Y \|^2}{\sigma^2} \sim \chi^2_{p_1 - p_0}$$

$$\frac{\| (I - P_1)Y \|^2}{\sigma^2} \sim \chi^2_{n - p_1}$$

and these two rv’s are independent.

**F Test**

Suppose that the OLS model holds where $E|X \sim MVN_n(0, \sigma^2 I)$.

In order to test $H_0 : (\beta_{p_0+1}, \beta_{p_0+2}, \ldots, \beta_{p_1}) = 0$ vs $H_1 : (\beta_{p_0+1}, \beta_{p_0+2}, \ldots, \beta_{p_1}) \neq 0$, we can form the $F$ statistic as given above, which has null distribution $F_{p_1 - p_0, n - p_1}$. The p-value is calculated as $Pr(F^* \geq F)$ where $F$ is the observed $F$ statistic and $F^* \sim F_{p_1 - p_0, n - p_1}$.

If the above assumption on the distribution of $E|X$ only approximately holds, then the $F$ test p-value is also an approximation.

**OLS Goodness of Fit: R**

Example: Davis Data

```r
> data("Davis", package="carData")
> htwt <- tbl_df(Davis)
> htwt[12,c(2,3)] <- htwt[12,c(3,2)]
> head(htwt)
# A tibble: 6 x 5
  sex weight height repwt repht
   <fct> <int> <int> <int> <int>
1 M      77     182    77   180
2 F      58     161    51   159
3 F      53     161    54   158
4 M      68     177    70   175
5 F      59     157    59   155
6 M      76     170    76   165
```

Comparing Linear Models in R

Example: Davis Data

Suppose we are considering the three following models:
f1 <- lm(weight ~ height, data=htwt)
f2 <- lm(weight ~ height + sex, data=htwt)
f3 <- lm(weight ~ height + sex + height:sex, data=htwt)

How do we determine if the additional terms in models f2 and f3 are needed?

ANOVA (Version 2)

A generalization of ANOVA exists that allows us to compare two nested models, quantifying their differences in terms of goodness of fit and performing a hypothesis test of whether this difference is statistically significant.

A model is nested within another model if their difference is simply the absence of certain terms in the smaller model.

The null hypothesis is that the additional terms have coefficients equal to zero, and the alternative hypothesis is that at least one coefficient is nonzero.

Both versions of ANOVA can be described in a single, elegant mathematical framework.

Comparing Two Models with anova()

This provides a comparison of the improvement in fit from model f2 compared to model f1:

```r
> anova(f1, f2)
Analysis of Variance Table
Model 1: weight ~ height
Model 2: weight ~ height + sex
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 14321
2 197 12816 1 1504.9 23.133 2.999e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When There’s a Single Variable Difference

Compare above anova(f1, f2) p-value to that for the sex term from the f2 model:

```r
> library(broom)
> tidy(f2)
# A tibble: 3 x 5
term estimate std.error statistic p.value
<chr>   <dbl>    <dbl>     <dbl>    <dbl>
1 (Intercept) -76.6   15.7      -4.88    2.23e-6
2 height     0.811   0.0953    8.51     4.50e-15
3 sexM       8.23    1.71      4.81     3.00e-15
```

Calculating the F-statistic

```r
> anova(f1, f2)
Analysis of Variance Table
Model 1: weight ~ height
Model 2: weight ~ height + sex
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 14321
2 197 12816 1 1504.9 23.133 2.999e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

How the F-statistic is calculated:

```r
n <- nrow(htwt)
ss1 <- (n-1)*var(f1$residuals)
ss1
## [1] 14321.11
ss2 <- (n-1)*var(f2$residuals)
ss2
## [1] 12816.18
((ss1 - ss2)/anova(f1, f2)$Df[2])/(ss2/f2$df.residual)
## [1] 23.13253
```

Calculating the Generalized LRT

```r
> anova(f1, f2, test="LRT")
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height + sex
    Res.Df RSS Df Sum of Sq F Pr(>F)
1   198 14321
2  197 12816 1  1504.9 23.13 0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(lmtest)
> lrtest(f1, f2)
Likelihood ratio test

Model 1: weight ~ height
Model 2: weight ~ height + sex
    #Df LogLik Df Chisq Pr(>Chisq)
1   3 -710.9
2  4 -699.8 1  22.05 2.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

These tests produce slightly different answers because `anova()` adjusts for degrees of freedom when estimating the variance, whereas `lrtest()` is the strict generalized LRT. See here.

**ANOVA on More Distant Models**

We can compare models with multiple differences in terms:

```r
> anova(f1, f3)
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height + sex + height:sex
    Res.Df RSS Df Sum of Sq   F Pr(>F)
1   198 14321
```

26
Compare Multiple Models at Once

We can compare multiple models at once:

```r
> anova(f1, f2, f3)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Model</th>
<th>formula</th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>weight ~ height</td>
<td>198</td>
<td>14321</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>weight ~ height + sex</td>
<td>197</td>
<td>12816</td>
<td>1</td>
<td>1504.93</td>
<td>23.4712</td>
<td>2.571e-06 ***</td>
</tr>
<tr>
<td>3</td>
<td>weight ~ height + sex + height:sex</td>
<td>196</td>
<td>12567</td>
<td>1</td>
<td>249.04</td>
<td>3.8841</td>
<td>0.05015 .</td>
</tr>
</tbody>
</table>

Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1

Monogenic Trait Model

Genotypes Under HWE

Let $X$ be a rv representing a SNP genotype, coded as a reference allele count: $X \in \{0, 1, 2\}$.

Under Hardy-Weinberg equilibrium (HWE), we have shown that

$$X \sim \text{Binomial}(2, p)$$

where $p$ is the allele frequency of the reference allele. Recall that $E[X] = 2p$, $\text{Var}[X] = 2p(1 - p)$.

Inbreeding

We also considered a population-level inbreeding model, where $f$ is the probability that the alleles are identical-by-descent (IBD) and $p$ is the ancestral population allele frequency. In this case:

$$
\begin{align*}
\Pr(X = 0) &= (1 - p)^2 + p(1 - p)f \\
\Pr(X = 1) &= 2p(1 - p)(1 - f) \\
\Pr(X = 2) &= p^2 + p(1 - p)f
\end{align*}
$$

Recall that $E[X] = 2p$ and $\text{Var}[X] = 2p(1 - p)(1 + f)$.

Kinship

Define the kinship between two individuals to be the probability that random alleles (at a random locus), one chosen from each of two individuals, are IBD.

Denote the kinship probability by the parameter $\phi$.

The kinship of an individual with itself is $\phi = \frac{1}{2}(1 + f)$.  

---

Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
Kinship Examples
Assume that the founders of a pedigree are unrelated and no one is inbred.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>1/2</td>
</tr>
<tr>
<td>Parent / child</td>
<td>1/4</td>
</tr>
<tr>
<td>Siblings</td>
<td>1/4</td>
</tr>
<tr>
<td>Half-siblings</td>
<td>1/8</td>
</tr>
<tr>
<td>First cousins</td>
<td>1/16</td>
</tr>
</tbody>
</table>

Covariance of Genotypes
Consider individuals \( j \) and \( k \), with genotypes \( X_j \) and \( X_k \) and kinship \( \phi_{jk} \). It can be shown that:

\[
\text{Cov}[X_j, X_k] = 4p(1 - p)\phi_{jk}
\]

Prove the above as an exercise. Note that

\[
\text{Var}[X_j] = \text{Cov}[X_j, X_j] = 4p(1 - p)\phi_{jj} = 4p(1 - p)\frac{1}{2}(1 + f_j) = 2p(1 - p)(1 + f_j)
\]

Additive Trait Model
Let \( X_1, X_2, \ldots, X_n \) come from the above genotype model. Assume that \( E_1, E_2, \ldots, E_n \) \( \overset{iid}{\sim} \) Normal(0, \( \sigma_e^2 \)), and \( X \) and \( E \) are independent.

Generate trait values \( Y_1, Y_2, \ldots, Y_n \) by:

\[
Y_j = \alpha + \beta X_j + E_j.
\]

This assumes additive effects only from the genetic locus.

General Trait Model
If we allow for additive and dominance effects, then we can write this as:

\[
Y_j = \beta_0 + \beta_1 1(X_j = 1) + \beta_2 1(X_j = 2) + E_j.
\]

However, we will only consider the additive model.

OLS and Dependence
Note that if \( \phi_{jk} > 0 \), then \( Y_j \) and \( Y_k \) are dependent random variables because \( X_j \) and \( X_k \) are dependent random variables.

However, by our assumptions, \( E_j | X_j = E_j \) and \( E_1, E_2, \ldots, E_n \) \( \overset{iid}{\sim} \) Normal(0, \( \sigma_e^2 \)). Thus, the assumptions of OLS are met.

In what sense is OLS useful and not useful in this setting?
Variance of Trait

Even the alleles within an individual can be dependent, so

\[ \text{Var}[Y_j] = \text{Var}[\alpha + \beta X_j + E_j] \]
\[ = \text{Var}[\beta X_j] + \text{Var}[E_j] \]
\[ = \beta^2 \text{Var}[X_j] + \sigma_e^2 \]
\[ = \beta^2 4p(1-p)\phi_{jj} + \sigma_e^2 \]
\[ = 2 \left[ \beta^2 2p(1-p) \right] \phi_{jj} + \sigma_e^2 \]

Variance Decomposition

Let \( \sigma_a^2 = \beta^2 2p(1-p) \) be the additive genetic variance and \( \sigma_e^2 \) be the non-genetic variance. We then have that

\[ \text{Var}[Y_j] = 2\sigma_a^2 \phi_{jj} + \sigma_e^2 \]

and when individual \( j \) is not inbred, then \( \phi_{jj} = 1/2 \) and

\[ \text{Var}[Y_j] = \sigma_a^2 + \sigma_e^2. \]

Covariance of Trait

\[ \text{Cov}[Y_j, Y_k] = \text{Cov}[\alpha + \beta X_j + E_j, \alpha + \beta X_k + E_k] \]
\[ = \text{Cov}[\beta X_j, \beta X_k] + \text{Cov}[E_j, E_k] \]
\[ = \beta^2 \text{Cov}[X_j, X_k] \]
\[ = \beta^2 4p(1-p)\phi_{jk} \]
\[ = 2 \left[ \beta^2 2p(1-p) \right] \phi_{jk} \]
\[ = 2\sigma_a^2 \phi_{jk} \]

Multivariate Distribution of Trait

Putting this all together, we have that:

\[ \mathbf{Y} | \mathbf{X} \sim \text{MVN}_n(\alpha \mathbf{1} + \beta \mathbf{X}, \sigma_e^2 \mathbf{I}) \]

\[ \mathbf{E}[\mathbf{Y}] = \alpha \mathbf{1} + 2p\beta \mathbf{1} \]

\[ \text{Var}[\mathbf{Y}] = 2\sigma_a^2 \Phi + \sigma_e^2 \mathbf{I} \]

where \( \mathbf{E}[Y_j] = \alpha + \beta \mathbf{E}[X_j] = \alpha + 2p\beta \) and \( \Phi \) is the \( n \times n \) kinship matrix with \( (j,k) \) entry equal to \( \phi_{jk} \).
Heritability

Conceptually, heritability is a measure of the proportion of variation in a trait attributable to genetics. Broad sense heritability involves terms from the trait model that includes dominance effects. Narrow sense heritability is defined as

\[ \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \]

since \( \text{Var}[Y] = \sigma_a^2 + \sigma_e^2 \) when the individual is outbred.

Polygenic Trait Model

Fisher (1918)

In the Fisher (1918) paper, RA Fisher developed a polygenic linear model of the genetic basis of a quantitative trait. In doing so, he resolved the dispute between “blending inheritance” from the biometric school and discrete inheritance from the Mendelian school. This paved the way for modern genetics as it is practiced today.

Fisher (1918) was the first genomics paper.

Fisher’s polygenic trait model is the primary model used in GWAS analyses and estimates of genome-wide inheritance today – over 100 years later!

Fisher’s polygenic model lead him to formulating a precise mathematical description of Darwinian selection in terms of genetic inheritance as we understand it today.

Assumptions

We now consider the additive trait model where there are \( m \) independent SNPs contributing to the trait. For each SNP \( i = 1, 2, \ldots, m \), there are genotypes \( X_{i1}, X_{i2}, \ldots, X_{in} \) corresponding to the \( n \) individuals. The ancestral allele frequency of SNP \( i \) is \( p_i \), and the dependence among the \( X_{i1}, X_{i2}, \ldots, X_{in} \) is parameterized by the \( n \times n \) kinship matrix \( \Phi \), as in the single locus model.

Again, let \( E_1, E_2, \ldots, E_n \overset{iid}{\sim} \text{Normal}(0, \sigma_e^2) \) and generate trait values \( Y_1, Y_2, \ldots, Y_n \) by:

\[ Y_j = \alpha + \sum_{i=1}^{m} \beta_i X_{ij} + E_j. \]

This again assumes additive effects only.

Variance of Trait

\[
\begin{align*}
\text{Var}[Y_j] &= \text{Var}[\alpha + \sum_{i=1}^{m} \beta_i X_{ij} + E_j] \\
&= \sum_{i=1}^{m} \text{Var}[\beta_i X_{ij}] + \text{Var}[E_j] \\
&= \sum_{i=1}^{m} \beta_i^2 \text{Var}[X_{ij}] + \sigma_e^2 \\
&= \sum_{i=1}^{m} \beta_i^2 4 p_i(1 - p_i) \phi_{jj} + \sigma_e^2 \\
&= 2 \left[ \sum_{i=1}^{m} \beta_i^2 2 p_i(1 - p_i) \right] \phi_{jj} + \sigma_e^2
\end{align*}
\]
Heritability

Noting that we now have

\[ \sigma^2_a = \sum_{i=1}^{m} \beta_i^2 2p_i(1 - p_i), \]

narrow sense heritability in the polygenic model is still defined as

\[ \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e}, \]

since \( \text{Var}[Y] = \sigma^2_a + \sigma^2_e \) when the individual is outbred.

Covariance of Trait

\[
\text{Cov}[Y_j, Y_k] = \text{Cov}\left[\alpha + \sum_{i=1}^{m} \beta_i X_{ij} + E_j, \alpha + \sum_{i=1}^{m} \beta_i X_{ik} + E_k\right]
\]

\[= \text{Cov}\left[\sum_{i=1}^{m} \beta_i X_{ij}, \sum_{i=1}^{m} \beta_i X_{ik}\right] + \text{Cov}[E_j, E_k]\]

\[= \sum_{i=1}^{m} \text{Cov}[\beta_i X_{ij}, \beta_i X_{ik}] = \sum_{i=1}^{m} \beta_i^2 \text{Cov}[X_{ij}, X_{ik}]\]

\[= \sum_{i=1}^{m} \beta_i^2 4p_i(1 - p_i)\phi_{jk} = 2 \sum_{i=1}^{m} \beta_i^2 2p_i(1 - p_i) \phi_{jk}\]

Normal Approximation

In the model \( Y_j = \alpha + \sum_{i=1}^{m} \beta_i X_{ij} + E_j \), Fisher noted that each term \( \beta_i X_{ij} \) is an instance of Mendelian inheritance.

However, taken as a whole and applying the CLT, \( \sum_{i=1}^{m} \beta_i X_{ij} \) can be treated as an instance of approximately continuous inheritance.

This yielded the MVN approximation

\[ Y \sim \text{MVN}_n(\alpha 1 + 2p\beta 1, 2\sigma_a^2 \Phi + \sigma_e^2 I), \]

and allowed Fisher to unify the biometric and Mendelian frameworks of genetic inheritance.

Lange (1978) rigorously proved the CLT under this model.

Variance Components

If we have a good estimate of \( \Phi \) (and that’s a big IF), then we can perform variance components analysis to write out the Normal log-likelihood function

\[ \ell(\sigma_a^2, \sigma_e^2; Y, \hat{\Phi}) \]

and use numerical methods to form estimates of \( \sigma_a^2 \) and \( \sigma_e^2 \).

There are a variety of approaches for doing this. See, for example, Chapter 8 of MSMGA by Lange, the lme4 package in R, or the GCTA genomics software.
Association Testing

Let’s suppose we are interested in testing the hypothesis, \( H_0 : \beta_k = 0 \) vs \( H_1 : \beta_k \neq 0 \) for some SNP \( k \). Assuming that \( \sum_{i=1}^{m} \beta_i X_{ij} \approx \sum_{i \neq k} \beta_i X_{ij} \), we can approximate:

\[
Y|X_k \sim \text{MVN}_n(\overline{\alpha} 1 + \beta_k X_k, 2\sigma_a^2 \Phi + \sigma_e^2 I),
\]

where \( \overline{\alpha} = \alpha + 2p\beta \).

Suppose we have estimates \( \hat{\sigma}_a^2, \hat{\Phi}, \) and \( \hat{\sigma}_e^2 \) available. A GLS regression model can then be fit to test the hypothesis \( H_0 : \beta_k = 0 \) vs \( H_1 : \beta_k \neq 0 \) for each SNP \( k = 1, 2, \ldots, m \).

This **linear mixed effects model** implementation of the polygenic trait model is utilized to test for associations between SNPs and a quantitative trait.

Extras

Source
License
Source Code

Session Information

```r
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS 10.15.3

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:

attached base packages:
[1] stats graphics grDevices utils datasets methods
[7] base

other attached packages:
[1] lmtest_0.9-37 zoo_1.8-7 broom_0.5.2
[4] carData_3.0-3 forcats_0.5.0 stringr_1.4.0
[7] dplyr_0.8.4 purrr_0.3.3 readr_1.3.1
[10] tidyr_1.0.2 tibble_2.1.3 ggplot2_3.2.1
[13] tidyverse_1.3.0 knitr_1.28

loaded via a namespace (and not attached):
[1] tidyselect_1.0.0 xfun_0.12 haven_2.2.0
[4] lattice_0.20-40 colorspace_1.4-1 vctrs_0.2.3
[7] generics_0.0.2 htmltools_0.4.0 yaml_2.2.1
[10] utf8_1.1.4 rlang_0.4.5 pillar_1.4.3
[13] withr_2.1.2 glue_1.3.1 DBI_1.1.0
[16] dbplyr_1.4.2 modelr_0.1.6 readxl_1.3.1
```

32