Analysis of Quantitative data
Linear regression

Anne Segonds-Pichon
v2020-12
Association between 2 continuous variables
One variable X and One variable Y
One predictor
Correlation
Signal-to-noise ratio

\[
\frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}
\]

\[
\frac{\text{Signal}}{\text{Noise}} = \text{statistical significance}
\]

\[
\frac{\text{Signal}}{\text{Noise}} = \text{no statistical significance}
\]
Signal-to-noise ratio and Correlation

\[ \frac{\text{Signal}}{\text{Noise}} = \frac{\text{Similarity}}{\text{Variability}} \]

- Signal is \textit{similarity} of behaviour between variable \( x \) and variable \( y \).

- \textbf{Coefficient of correlation:} \( r = \frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}} \)

\[
r = \frac{\text{COV}_{xy}}{\text{SD}_x \text{SD}_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1) \text{SD}_x \text{SD}_y}
\]
Correlation

• Most widely-used correlation coefficient:
  • Pearson product-moment correlation coefficient “r”
    • The magnitude and the direction of the relation between 2 variables
    • It is designed to range in value between -1 and +1
    • -0.6 < r > +0.6 : exciting

• Coefficient of determination “r²”
  • It gives the proportion of variance in Y that can be explained by X (in percentage).
    • It helps with the interpretation of r
    • It’s basically the effect size
Correlation

Variable 1 vs Variable 2

$r = -0.34$, $p = 0.0002$, $r^2 = 0.12$

Variable 1 vs Variable 2

$r = -0.83$, $p = 0.04$, $r^2 = 0.68$

Power!!
Correlation
Assumptions

- **Assumptions for correlation**
  - Regression and linear Model (lm)

- **Linearity**: The relationship between $X$ and the mean of $Y$ is linear.

- **Homoscedasticity**: The variance of residual is the same for any value of $X$.

- **Independence**: Observations are independent of each other.

- **Normality**: For any fixed value of $X$, $Y$ is normally distributed.
Correlation
Outliers and High leverage points

- **Outliers**: the observed value for the point is very different from that predicted by the regression model.
Correlation
Outliers and High leverage points

• **Leverage points**: A leverage point is defined as an observation that has a value of $x$ that is far away from the mean of $x$.

• Outliers and leverage points have the potential to be **Influential observations**:
  – Change the slope of the line. Thus, have a large influence on the fit of the model.

• One method to find influential points is to compare the fit of the model **with** and **without** the dodgy observation.
Correlation
Outliers and High leverage points

All good
Correlation
Outliers and High leverage points

Outlier but not influential value

With outlier
Without outlier

Outlier but not influential value
Correlation
Outliers and High leverage points

High leverage but not influential value
Correlation

Outliers and High leverage points

Outlier and High leverage: Influential value
Correlation: Two more things

Thing 1: Pearson correlation is a parametric test

First assumption for parametric test: Normality

Correlation: bivariate Gaussian distribution

Symmetry-ish of the values on either side of the line of best fit.
Correlation: Two more things

Thing 2: Line of best fit comes from a regression

**Correlation:** nature and strength of the association

**Regression:** nature and strength of the association and prediction

\[ Y = A + B \times X \]
Correlation: correlation.csv

• Questions:
  • What is the nature and the strength of the relationship between X and Y?
  • Are there any dodgy points?
Correlation: correlation.csv

**Question:** are there any dodgy points?

```r
read_csv("correlation.csv") -> correlation

correlation %>%
  ggplot(aes(variable.x, variable.y, colour=Gender)) +
  geom_point(size=3, colour="sienna2")
```
Correlation: correlation.csv

• For the lines of best-fit: 3 new functions:

\begin{verbatim}
lm(y~x, data=) -> fit
coefficients(fit) -> cf.fit (vector of 2 values)
geom_abline(intercept=cf.fit[1], slope=cf.fit[2])
\end{verbatim}

\begin{verbatim}
lm(variable.y ~ variable.x, data=correlation) -> fit.correlation
coefficients(fit.correlation) -> coef.correlation
coef.correlation
\end{verbatim}

(Intercept) variable.x
8.379803 3.588814
intercept slope
Correlation: correlation.csv

correlation %>%
ggplot(aes(variable.x, variable.y, label = ID)) +
geom_point(size=3, colour="sienna2") +
geom_abline(intercept = coef.correlation[1], slope = coef.correlation[2]) +
geom_text(hjust = 0, nudge_x = 0.15)
par(mfrow=c(2,2))
plot(fit.correlation)

**Linearity, homoscedasticity and outlier**

**Normality and outlier**

The **Cook’s distance** is a combination of each observation’s leverage and residual values; the higher the leverage and residuals, the higher the Cook’s distance (influential observation).

- It summarizes how much all the values in the regression model change when the ith observation is removed.
- Consensus: cut-off point = 1 (0.5).
**Correlation: correlation.csv**

---

**Line of best fit:** \( Y = 8.38 + 3.59X \)

**Summary:**

```
summary(fit.correlation)
call: 
  lm(formula = variable.y ~ variable.x, data = correlation)

Residuals: 
     Min      1Q  Median      3Q     Max 
-40.034  -3.414   0.867   5.723  17.265

Coefficients: 
             Estimate Std. Error t value Pr(>|t|) 
(Intercept)  8.3798     4.1195   2.034  0.0548 . 
variable.x  3.5888     0.6225   5.765 1.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.93 on 21 degrees of freedom
Multiple R-squared:  0.6128,   Adjusted R-squared:  0.5943
F-statistic: 33.23 on 1 and 21 DF,  p-value: 1.01e-05
```

---

**Correlation %>% cor_test(variable.x, variable.y)**

<table>
<thead>
<tr>
<th>var1</th>
<th>var2</th>
<th>cor</th>
<th>statistic</th>
<th>p</th>
<th>conf.low</th>
<th>conf.high</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>variable.x</td>
<td>variable.y</td>
<td>0.78</td>
<td>5.764871</td>
<td>1.01e-05</td>
<td>0.5471597</td>
<td>0.9034793</td>
<td>Pearson</td>
</tr>
</tbody>
</table>
Correlation: correlation.csv

Have a go: Remove ID 23, then re-run the model and plot the graph again. Then decide what you want to do with ID 21 and 22.

correlation %>%
   filter(ID != 23) -> correlation.23
correlation %>%
  filter(ID != 23) -> correlation.23

lm(variable.y ~ variable.x, correlation.23) -> fit.correlation.23
summary(fit.correlation.23)

Call:
  lm(formula = variable.y ~ variable.x, data = correlation.23)

Residuals:
  Min 1Q Median 3Q Max
-5.049 -2.784 -1.446 1.679 16.915

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7103     1.8338   2.023  0.0566 .
variable.x   4.8436     0.2971  16.303 5.13e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.695 on 20 degrees of freedom
Multiple R-squared: 0.933,  Adjusted R-squared: 0.9265
F-statistic: 265.8 on 1 and 20 DF,  p-value: 5.13e-13
Correlation: correlation.csv

correlation.23 %>%
  filter(ID != 21) -> correlation.23.21

lm(variable.y ~ variable.x, correlation.23.21) -> fit.correlation.23.21
summary(fit.correlation.23.21)

cor_test(variable.x, variable.y)

Call:
lm(formula = variable.y ~ variable.x, data = correlation.23.21)

Residuals:
  Min     1Q  Median     3Q    Max
-4.3636 -1.8607 -0.5376  2.2987  5.0434

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.4679    1.0757  2.2944   0.0333 *
variable.x      4.9272    0.1719 28.6614  < 2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 19 degrees of freedom
Multiple R-squared:  0.9774, Adjusted R-squared:  0.9762
F-statistic: 821.4 on 1 and 19 DF,  p-value: < 2.2e-16
Extra exercise

Correlation: exam.anxiety.csv

• **Question:** Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

• Build a fit for the boys and a fit for the girls
  • `data %>% filter() lm(y~x, data=)`

• Plot the 2 lines of best fit on the same graph
  • `coefficients() geom_abline()`

• Check the assumptions visually from the data and with the output for models
  • `par(mfrow=c(2,2)) plot(fit.male)`

• Filter out misbehaving values based on the standardised residuals
  • `rstandard() add_column()`

• Plot the final (improved!) model
  • `bind_rows()`
Correlation: exam.anxiety.csv

**Question:** Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

```r
read_csv("exam.anxiety.csv") -> exam.anxiety

exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender)) + geom_point(size=3)
```
• Is there a relationship between time spent revising and exam anxiety?

```
exam.anxiety %>%
  filter(Gender=="Female") -> exam.anxiety.female
lm(Anxiety~Revise, data=exam.anxiety.female) -> fit.female
coefficients(fit.female) -> cf.fit.female
```

```
exam.anxiety %>%
  filter(Gender=="Male") -> exam.anxiety.male
lm(Anxiety~Revise, data=exam.anxiety.male) -> fit.male
coefficients(fit.male) -> cf.fit.male
```

Fit for the females

Fit for the males
• Is there a relationship between time spent revising and exam anxiety?

```r
exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender)) +
  geom_point(size=3) +
  geom_abline(intercept=cf.fit.male[1], slope=cf.fit.male[2]) +
  geom_abline(intercept=cf.fit.female[1], slope=cf.fit.female[2])
```
Correlation: exam anxiety.csv
Assumptions, outliers and influential cases

```
par(mfrow=c(2,2))
plot(fit.male)
```

![Residuals vs Fitted plot](image1)
![Normal Q-Q plot](image2)
![Scale-Location plot](image3)
![Residuals vs Leverage plot](image4)
Correlation: exam anxiety.csv
Assumptions, outliers and influential cases

```r
plot(fit.female)
```
Correlation: exam anxiety.csv

```r
exam.anxiety %>%
  group_by(Gender) %>%
  cor_test(Revise, Anxiety) %>%
  ungroup()
```

```
summary(fit.male)
```

Anxiety = 84.19 - 0.53*Revise

```
summary(fit.female)
```

Anxiety = 91.94 - 0.82*Revise
Correlation: exam.anxiety.csv

Influential outliers: Boys

\[
\text{rstandard}(\text{fit.male}) \rightarrow \text{st.resid.m}
\]

\[
\text{exam.anxiety.male} \%>\% \text{add_column}(\text{st.resid.m}) \%>\% \text{filter(abs(st.resid.m)<3)} \rightarrow \text{exam.anxiety.male.clean}
\]

\[
\text{lm(Anxiety~Revise, data=exam.anxiety.male.clean)} \rightarrow \text{fit.male2}
\]

\[
\text{summary(fit.male2)}
\]

Call:
\[
\text{lm(formula = Anxiety ~ Revise, data = exam.anxiety.male.clean)}
\]

Residuals:
\[
\begin{array}{ccccc}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-22.0296 & -3.8704 & 0.5626 & 6.0786 & 14.2525 \\
\end{array}
\]

Coefficients:
\[
\begin{array}{ccccc}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
\text{(Intercept)} & 86.97461 & 1.64755 & 52.790 & < 2e-16 *** \\
\text{Revise} & -0.60752 & 0.06326 & -9.603 & 7.59e-13 *** \\
\end{array}
\]

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.213 on 49 degrees of freedom
Multiple R-squared: 0.653, Adjusted R-squared: 0.6459
F-statistic: 92.22 on 1 and 49 DF, p-value: 7.59e-13

\[
\text{exam.anxiety.male.clean} \%>\% \text{cor_test(Revise, Anxiety)}
\]

\[
\begin{array}{cccccc}
\text{var1} & \text{var2} & \text{cor} & \text{statistic} & \text{p} & \text{conf.low} & \text{conf.high} \\
\text{Revise} & \text{Anxiety} & -0.81 & -9.602995 & 7.59e-13 & -0.8863013 & -0.6850763 \\
\end{array}
\]
**Correlation: exam.anxiety.csv**

Influential outliers: Girls

\[ \text{rstandard}(\text{fit.female}) \rightarrow \text{st.resid.f} \]

\[ \text{exam.anxiety.female} \%\% \]
\[ \text{add_column}(\text{st.resid.f}) \%\% \]
\[ \text{filter(abs(st.resid.f) < 3)} \rightarrow \text{exam.anxiety.female.clean} \]

\[ \text{lm}(\text{Anxiety}\sim\text{Revise}, \text{data=}\text{exam.anxiety.female.clean}) \rightarrow \text{fit.female2} \]

\[ \text{summary(fit.female2)} \]

```
Call:
  lm(formula = Anxiety ~ Revise, data = exam.anxiety.female.clean)

Residuals:
   Min     1Q  Median     3Q    Max
-18.7518 -5.7069 -0.7782  3.2117  18.5538

Coefficients:  
            Estimate Std. Error t value Pr(>|t|) 
(Intercept)   92.24536   1.93591  47.656  < 2e-16 *** 
Revise       -0.87504   0.07033  -12.444   < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.849 on 48 degrees of freedom 
Multiple R-squared: 0.7633,Adjusted R-squared: 0.7584
F-statistic: 154.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

\[ \text{exam.anxiety.female.clean} \%\% \]
\[ \text{cor_test}(\text{Revise}, \text{Anxiety}) \]
**Question:** Is there a relationship between time spent revising and exam anxiety? Yes!

```r
bind_rows(exam.anxiety.female.clean, exam.anxiety.male.clean) -> exam.anxiety.clean
coefficients(fit.male2) -> cf.fit.male2
coefficients(fit.female2) -> cf.fit.female2
exam.anxiety.clean %>%
  ggplot(aes(Revise, Anxiety, colour=Gender))+geom_point(size=3)+
  geom_abline(aes(intercept=cf.fit.male2[1], slope=cf.fit.male2[2]), colour="orange")+
  geom_abline(aes(intercept=cf.fit.female2[1], slope=cf.fit.female2[2]), colour="purple")+
  scale_colour_manual(values = c("purple", "orange"))
```
Correlation: exam.anxiety

Influential outliers: Another check

exam.anxiety.male %>% shapiro_test(st.resid.m)

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>st.resid.m</td>
<td>0.6992772</td>
<td>5.05199e-09</td>
</tr>
</tbody>
</table>

eexam.anxiety.male.clean %>% shapiro_test(st.resid.m)

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>st.resid.m</td>
<td>0.9539309</td>
<td>0.04607996</td>
</tr>
</tbody>
</table>

eexam.anxiety.female %>% shapiro_test(st.resid.f)

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>st.resid.f</td>
<td>0.9442729</td>
<td>0.01828732</td>
</tr>
</tbody>
</table>

eexam.anxiety.female.clean %>% shapiro_test(st.resid.f)

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>st.resid.f</td>
<td>0.9757888</td>
<td>0.4258592</td>
</tr>
</tbody>
</table>
• Difference between boys and girls?

```r
lm(Anxiety~Revise*Gender, data=exam.anxiety.clean) -> fit.genders
summary(fit.genders)
```

```
Call:
  lm(formula = Anxiety ~ Revise * Gender, data = exam.anxiety.clean)

Residuals:
    Min     1Q  Median     3Q    Max
-22.0296 -5.6022 -0.3294  5.6091  18.5538

Coefficients:  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      92.24536    1.86694   49.410  < 2e-16 ***
Revise          -0.87504    0.06783  -12.901  < 2e-16 ***
GenderMale      -5.27075    2.53296  -2.081    0.04008 *
Revise:GenderMale  0.26752    0.09445   2.832   0.00562 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.534 on 97 degrees of freedom
Multiple R-squared:  0.7228,   Adjusted R-squared:  0.7142
F-statistic:  84.32 on 3 and 97 DF,  p-value: < 2.2e-16
```