Quantitative Factors

STAT3022 Applied Linear Models Lecture 17

2020/02/17

THE UNIVERSITY OF SYDNEY

Today

1. Quantitative factor
2. Polynomial regression and ANOVA
3. Nesting of linear effects
4. Bartlett test to assess homoscedasticity assumption

Factor or numerical variables?

Sometimes it is unclear as to whether a particular explanatory variable should be regarded as a factor (categorical explanatory variable) or a numerical covariate.

We will see that polynomial regression is the key to understand the problem.

Example - Vitamin C and tooth growth

- This example concerns an (old) experiment into the effects of vitamin C on tooth growth.
- 30 guinea pigs were divided (at random) into three groups of ten and dosed with vitamin C (administered in orange juice).
  - Group 1 dose was low (0.5mg vit. C),
  - Group 2 dose was medium (1mg vit. C) and
  - Group 3 dose was high (2mg vit. C).
- Length of odontoblasts (teeth) measured as response variable.

Reference: Bliss (1952) The Statistics of Bioassay. Academic Press.

Replicate	0.5	1	2
1	4.2	16.5	23.6
2	11.5	16.5	18.5
3	7.3	15.2	33.9
4	5.8	17.3	25.5
5	6.4	22.5	26.4
6	10.0	17.3	32.5
7	11.2	13.6	26.7
8	11.2	14.5	21.5
9	5.2	18.8	23.3
10	7.0	15.5	29.5

Getting the data in the right form

- You will get the data in all sorts of forms especially when they are grouped in some way.
- Strive to get your data in the form ready for analysis.
- Remember the tidy data principle: each row is an observation, each column is a variable.
- We have 30 observations so the data frame should contain 30 rows.

```
dat <- datwide %>%
  select(-Replicate) %>%
  gather(key=Dose, value=Length) %>%
  mutate(Dose=as.numeric(Dose),
         Dose.fac=case_when(
           Dose=="0.5" ~ "Low",
           Dose=="1" ~ "Medium",
           Dose=="2" ~ "High"
         )) %>%
  # relevel first so order is meaningful
  mutate(Dose.fac=factor(Dose.fac,
                         level=c("Low", "Medium", "High")))
```

```
head(datwide) #short format
```

```
     Replicate  0.5    1    2
[1,]         1  4.2 16.5 23.6
[2,]         2 11.5 16.5 18.5
[3,]         3  7.3 15.2 33.9
[4,]         4  5.8 17.3 25.5
[5,]         5  6.4 22.5 26.4
[6,]         6 10.0 17.3 32.5
```

```
head(dat,5) #long format
```

```
  Length Dose.fac Dose
1    4.2      Low  0.5
2   11.5      Low  0.5
3    7.3      Low  0.5
4    5.8      Low  0.5
5    6.4      Low  0.5
```

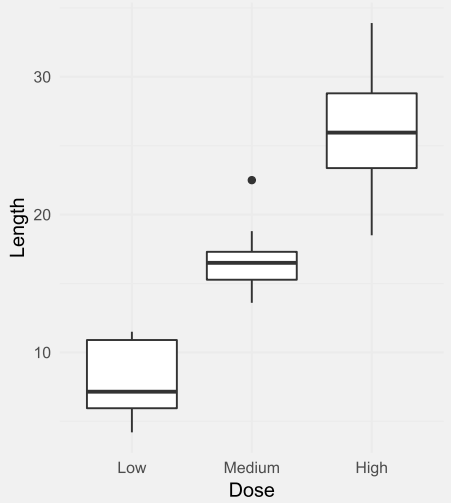
Example - Vitamin C and tooth growth

```
skimr::skim(dat)
```

```
  Skim summary statistics
   n obs: 30
   n variables: 3

  ── Variable type:factor ─────────────────────────────────────────────────────
   variable missing  n n_unique ordered
   Dose.fac       0 30        3   FALSE

  ── Variable type:numeric ────────────────────────────────────────────────────
   variable missing  mean   sd  p0  p25  p50  p75 p100     hist
       Dose       0  1.17 0.63 0.5  0.5  1    2    2
     Length       0 16.96 8.27 4.2 11.2 16.5 23.1 33.9
```

- For the plot it seems highly plausible that mean tooth growth depends on vitamin C levels.
- Could the dependence of tooth growth on vitamin C be adequately modelled by a linear regression of `Length` on `Dose`, or is an ANOVA type model with `Length` and `Dose.fac` to be preferred?

```
ggplot(dat,
  aes(Dose.fac, Length)) +
  geom_boxplot() +
  labs(x="Dose")
```



Simple linear regression or one-way ANOVA

```
mdat <- dat %>% group_by(Dose) %>%
  summarise(Length=mean(Length))
ggplot(dat, aes(Dose, Length)) + geom_point() +
  geom_smooth(method="lm", se=F) +
  geom_point(data=mdat, aes(Dose, Length),
             color="red", size=8, alpha=1/3) +
  theme_bw(base_size=18)
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
M.SLR <- lm(Length ~ Dose, dat) #simple linear reg
anova(M.SLR)
```

```
Analysis of Variance Table

Response: Length
          Df  Sum Sq Mean Sq F value    Pr(>F)
Dose       1 1601.34 1601.34  117.95 1.509e-11 ***
Residuals 28  380.15   13.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
```

```
M.ANOVA <- lm(Length ~ Dose.fac, dat) #1-way ANOVA
anova(M.ANOVA)
```

```
Analysis of Variance Table

Response: Length
          Df Sum Sq Mean Sq F value    Pr(>F)
Dose.fac   2 1649.5  824.74  67.072 3.357e-11 ***
Residuals 27  332.0   12.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
```

Simple linear regression or one-way ANOVA

Take Dose as continuous in **simple linear regression**?

```
print(broom::tidy(M.SLR), digits = 6)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     3.29      1.43      2.31 2.85e- 2
2 Dose           11.7       1.08     10.9  1.51e-11
```

Take Dose as categorical in **one-way ANOVA model**?

```
print(broom::tidy(M.ANOVA), digits = 6)
```

```
# A tibble: 3 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         7.98      1.11      7.20 9.71e- 8
2 Dose.facMedium      8.79      1.57      5.61 6.02e- 6
3 Dose.facHigh       18.2       1.57     11.6  5.58e-12
```

Note that although RSS is lower, it also has 1 df less. Both models are highly significant.

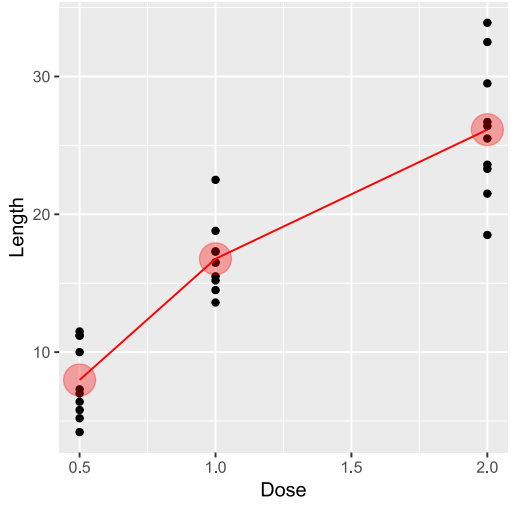
Recall: simple linear model passes through $(\bar{x}, \bar{Y}_{..})$

- The blue dot here is the $(\bar{x}, \bar{Y}_{..})$.
- Remember that the one-way ANOVA model can predict the red dots but cannot predict for factor levels not included in the model.

```
`geom_smooth()` using formula 'y ~ x'
```

Can we consider another linear model that goes through the mean of each treatment group?

That is a linear model that goes through $(x_i, \bar{Y}_{i.})$?

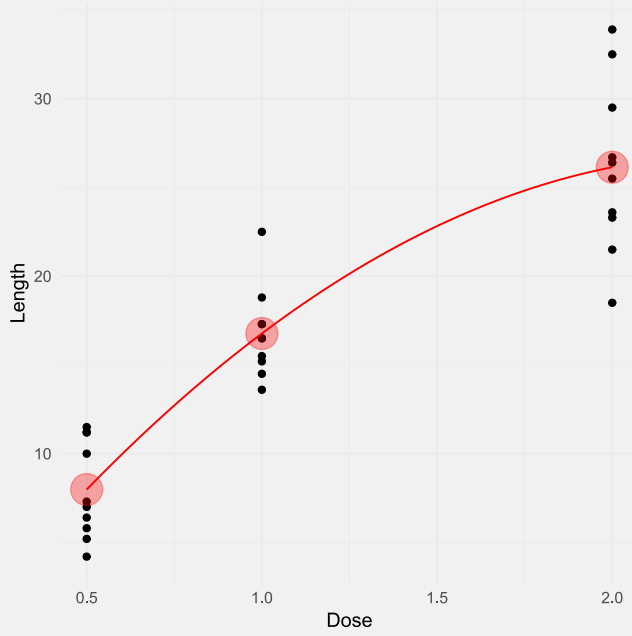


What about polynomial regression?

- Suppose the ith treatment corresponds to a measurement $x_i \in \mathrm{R}$.
- We can plot the sample mean response at each x_i value.
- If we have t treatments then we have t points on the plot.
- By looking at the way the mean values vary with x_i we can put forward a model for $E(Y|x)$.
- Through any t points we can fit a polynomial of degree $(t-1)$.

Thus the most general polynomial regression model for this situation is

$$Y_{ij} = \beta_0 + \beta_1 x_i + \ldots + \beta_{t-1} x_i^{t-1} + \epsilon_{ij}, \quad \epsilon_{ij} \sim NID(0, \sigma^2).$$



Polynomial regression equivalent to ANOVA

- The polynomial model has the same number of freely determined parameters (t) as the one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

- Both models give exactly the same fit and have the same RSS!
- The advantage of the polynomial model is that it suggests a range of models if

$$H_0 : \alpha_1 = \ldots = \alpha_t$$

or equivalently $\beta_1 = \ldots = \beta_{t-1} = 0$ does not hold because there are different possibility for the order of the polynomial model.

```
M.poly <- lm(Length ~ poly(Dose,2),data=dat) #ortho
data.frame(M.poly$residuals, M.ANOVA$residuals)
```

```
   M.poly.residuals M.ANOVA.residuals
1             -3.78             -3.78
2              3.52              3.52
3             -0.68             -0.68
4             -2.18             -2.18
5             -1.58             -1.58
6              2.02              2.02
7              3.22              3.22
8              3.22              3.22
9             -2.78             -2.78
10            -0.98             -0.98
11            -0.27             -0.27
12            -0.27             -0.27
13            -1.57             -1.57
14             0.53              0.53
15             5.73              5.73
16             0.53              0.53
```

Nesting of linear effects

- The estimator for the linear term is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_i - \bar{x})Y_{ij}}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{t} n_i(x_i - \bar{x})\bar{Y}_{i\cdot}}{\sum_{i=1}^{t} n_i(x_i - \bar{x})^2}$$

- Recall: Within SS = Total SS - Treatment SS = TotSS - TSS.
- To test adequacy of fit of the linear regression we use

$$f = \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS}_{H_1})/(t-2)}{\mathrm{RSS}_{H_1}/(n-t)}$$

- The difference between the two RSS in H_0 with linear regression and H_1 with one-way ANOVA is

$$\mathrm{RSS}_{H_0} - \mathrm{RSS}_{H_1} = \mathrm{TotSS} - \hat{\beta}_1^2 S_{xx} - (\mathrm{TotSS} - \mathrm{TSS})$$

```
anova(M.ANOVA)
```

```
Analysis of Variance Table

Response: Length
          Df Sum Sq Mean Sq F value    Pr(
Dose.fac   2 1649.5  824.74  67.072 3.357e
Residuals 27  332.0   12.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*
```

```
anova(M.poly) #RSS same orthogonal or not
```

```
Analysis of Variance Table

Response: Length
              Df Sum Sq Mean Sq F value
poly(Dose, 2)  2 1649.5  824.74  67.072 3.
Residuals     27  332.0   12.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*
```

The two RSS are the same.

LinRegSS is a 1 df component of TSS

```
anova(M.SLR)
```

```
Analysis of Variance Table

Response: Length
          Df  Sum Sq Mean Sq F value    Pr(>F)
Dose       1 1601.34 1601.34  117.95 1.509e-11 ***
Residuals 28  380.15   13.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' (
```

```
anova(M.poly)
```

```
Analysis of Variance Table

Response: Length
              Df Sum Sq Mean Sq F value    Pr(>F)
poly(Dose, 2)  2 1649.5  824.74  67.072 3.357e-11 ***
Residuals     27  332.0   12.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' (
```

- The simple linear regression model is nested within the order 2 polynomial regression model.
- Then there must be a contrast of LinRegSS within TSS in one-way ANOVA model which has the same RSS as the order 2 polynomial regression.
- What would the coefficients of the treatment means in the contrast of the one-way ANOVA be?

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{t} n_i(x_i - \bar{x})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})}{\sum_{i=1}^{t} n_i(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{t} n_i(x_i - \bar{x})\bar{Y}_{i\cdot}}{\sum_{i=1}^{t} n_i(x_i - \bar{x})^2}.$$

Linear Regression Sum of Squares

- Any contrast coefficients which are proportional to

$$c_i = n_i(x_i - \bar{x})$$

satisfy $\sum_{i=1}^{t} c_i = 0$ as $\sum_{i=1}^{t} c_i = \sum_{i=1}^{t} n_i(x_i - \bar{x}) = 0$.

- So a 1 df component corresponding to the TSS is

$$\frac{\left(\sum_{i=1}^{t} c_i \bar{Y}_{i\cdot}\right)^2}{\sum_{i=1}^{t} \frac{c_i^2}{n_i}} = \frac{\left[\sum_{i=1}^{t} n_i(x_i - \bar{x})\bar{Y}_{i\cdot}\right]^2}{\sum_{i=1}^{t} \frac{[n_i(x_i - \bar{x})]^2}{n_i}} = \frac{\left[\sum_{i=1}^{t} n_i(x_i - \bar{x})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})\right]^2}{\sum_{i=1}^{t} n_i(x_i - \bar{x})^2} = \frac{S_{xy}^2}{S_{xx}} = \hat{\beta}_1^2 S_{xx} = \text{LinRegSS}$$

- You can extend this idea to test adequacy of fit for higher order polynomials.

Nesting of linear effects

- To test if the sample means vary linearly with x is equivalent to testing
  - $H_0 : \beta_2 = \beta_3 = \ldots = \beta_{t-1} = 0$
  - or a test for adequacy of fit of a linear regression.
- If H_0 is true, $Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$.
- The linear model under H_0 is nested not only in polynomial regression but also in ANOVA model.
- Recall that for a simple linear regression model under H_0 has

$$S_{yy} = \text{TotSS} = \text{RSS}_{H_0} + \hat{\beta}_1^2 S_{xx}.$$

- $\hat{\beta}_1^2 S_{xx}$ for LinRegSS is 1601.34 in `Length`.

- Thus, we calculate the RSS as

$$\text{RSS}_{H_0} = S_{yy} - \hat{\beta}_1^2 \sum_{i=1}^{t} \sum_{j=1}^{n_i} (x_i - \bar{x})^2,$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{t} n_i x_i$.

```
anova(M.SLR)
```

```
Analysis of Variance Table

Response: Length
          Df  Sum Sq Mean Sq F value    Pr(>F)
Dose       1 1601.34 1601.34  117.95 1.509e-11 ***
Residuals 28  380.15   13.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
```

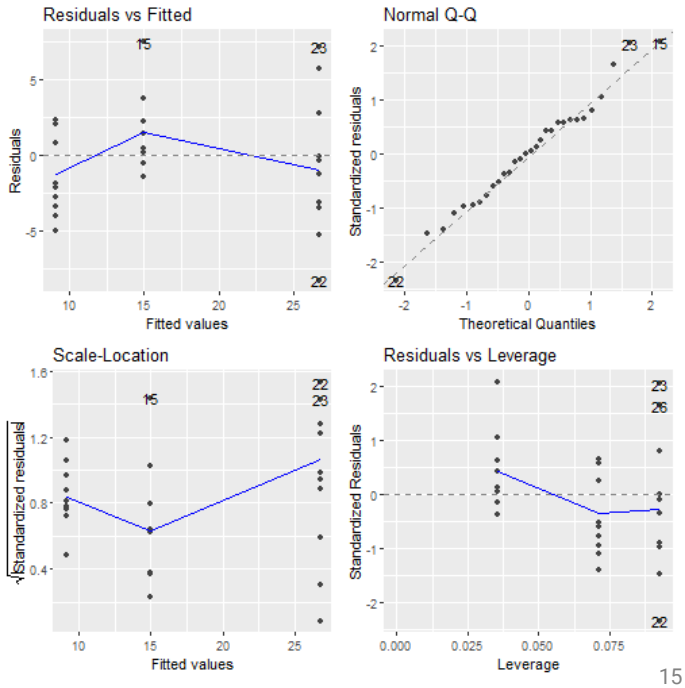
Example - Tooth: Linear regression vs ANOVA

```
anova(M.SLR, M.ANOVA)
```

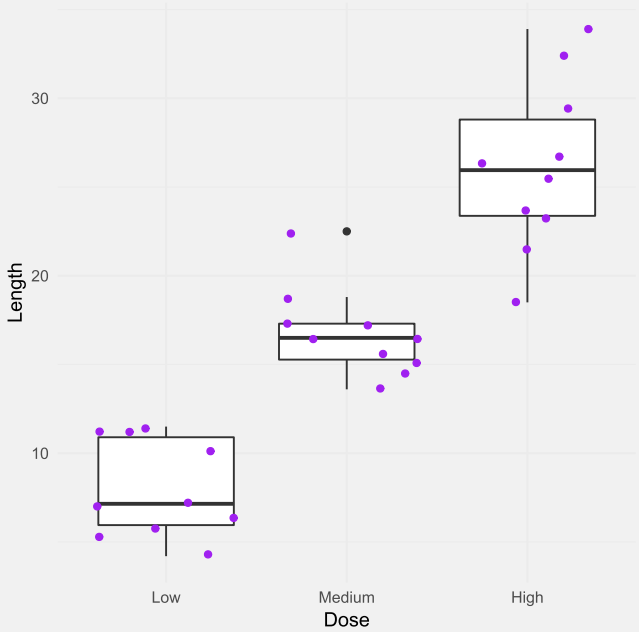
```
Analysis of Variance Table

Model 1: Length ~ Dose
Model 2: Length ~ Dose.fac
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     28 380.15
2     27 332.00  1    48.146 3.9155 0.05813 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
```

- The corresponding p-value is $p = 0.058$. We would retain H_0 at the 5% significance level.
- So we should use the simple linear regression.

```
autoplot(M.SLR)
```



Homogeneity of Variances



- The spread of `Length` in the `Medium` group seems to be smaller than other groups.
- We have reasons to doubt if the variances are equal across groups.
- We can conduct a *statistical test* to test this model assumption.

Bartlett test

- Let s_i^2 denote the sample variance in treatment group $i = 1, \ldots, t$. The pooled variance is defined as

$$s_p^2 = \sum_{i=1}^{t} (n_i - 1) \frac{s_i^2}{(n-t)}.$$

- The Bartlett test compares $H_0 : \sigma_1^2 = \ldots = \sigma_t^2$ vs $H_1 : \sigma_i^2 \neq \sigma_j^2, \exists i \neq j$.
- The Bartlett test can be calculated with `bartlett.test()` and is defined as

$$B = \frac{(n-t)\log s_p^2 - \sum_{i=1}^{t}(n_i - 1)\log s_i^2}{1 + [\frac{1}{3(t-1)}](\sum_{i=1}^{t} \frac{1}{n_i - 1} - \frac{1}{n-t})} \approx \chi_{t-1}^2$$

Tooth: Bartlett test in R

```
bartlett.test(Length ~ Dose.fac, data=dat)
```

```
    Bartlett test of homogeneity of variances

data:  Length by Dose.fac
Bartlett's K-squared = 4.5119, df = 2, p-value = 0.104
```

- Thus, we conclude from the Bartlett test that there is not sufficient evidence in the data against the homoscedasticity assumption in the one-way ANOVA model even though the spread of `Length` in the `medium` group is smaller.
- But the spread is not small enough to be significant.