STATISTICAL MODELLING Part IIC, Michaelmas 2019
Practical 5: ANOVA and ANCOVA By courtesy of Dr. S. A. Bacallado and Dr. R. Shah

ANOVA solutions

Why are there no `QualityGood` or `PhotoControl` coefficients in `EssayMarksLM2`?

This is due to the corner point constraints, which select Good and Control as base categories for the factors Quality and Photo, respectively.

Write out the models for `EssayMarksLM4` and `EssayMarksLM5`

If Y_{ijk} are the marks of the kth essay with quality i and photo j. Then both models have

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N(0, \sigma^2)$$

In `EssayMarksLM4`, $\alpha_{\text{good}} = 0$, $\beta_{\text{control}} = \beta_{\text{attractive}} = 0$, and $\gamma_{ij} \neq 0$ only when $i =$ good and $j =$ unattractive.

In `EssayMarksLM5`, $\alpha_{\text{good}} = 0$, $\beta_{\text{control}} = 0$, and $\gamma_{ij} \neq 0$ only when $j =$ unattractive. As the output of `summary(EssayMarksLM5)` shows, the resulting design matrix is not full rank; in particular, it has rank 5, so one parameter cannot be estimated.

Of course, there are many other ways of choosing the corner-point constraints which would generate the same fitted values $\hat{Y}$ (and neither of which would avoid the rank 5 of the resulting matrix in `EssayMarksLM5`).

What is the most appropriate model according to AIC?

```
AIC(EssayMarksLM1, EssayMarksLM2, EssayMarksLM3, EssayMarksLM4, EssayMarksLM5)
```

```
##                df      AIC
## EssayMarksLM1  3 361.5300
## EssayMarksLM2  5 355.0405
## EssayMarksLM3  7 357.0200
## EssayMarksLM4  5 353.6479
## EssayMarksLM5  6 355.3348
```

The most appropriate model according to the AIC is `EssayMarksLM4`. Recall that the AIC approximates the KL divergence between the estimated and true models, so the lower the AIC, the better the fit.

Would it be possible to compare `EssayMarksLM2` and `EssayMarksLM5` through an F-test?

Yes, as the predictors in the first model are a subset of the second's. Indeed,

```
anova(EssayMarksLM2,EssayMarksLM5)
```

```
## Analysis of Variance Table
##
## Model 1: Mark ~ Quality + Photo
## Model 2: Mark ~ Quality + Photo + Quality:Photo_grp
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     56 1104.5
## 2     55 1073.5  1    30.957 1.586 0.2132
```

How about `EssayMarksLM2` and `EssayMarksLM4`?

In this case it isn't possible, because the first model has a coefficient for `PhotoAttractive` which is missing in the second, while the second has interaction terms which are missing in the first. We can check that the `anova` function does not perform a test:

```
anova(EssayMarksLM2,EssayMarksLM4)
```

```
## Analysis of Variance Table
##
## Model 1: Mark ~ Quality + Photo
## Model 2: Mark ~ Quality * Photo_grp
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     56 1104.5
## 2     56 1079.2  0     25.34
```

ANCOVA solutions

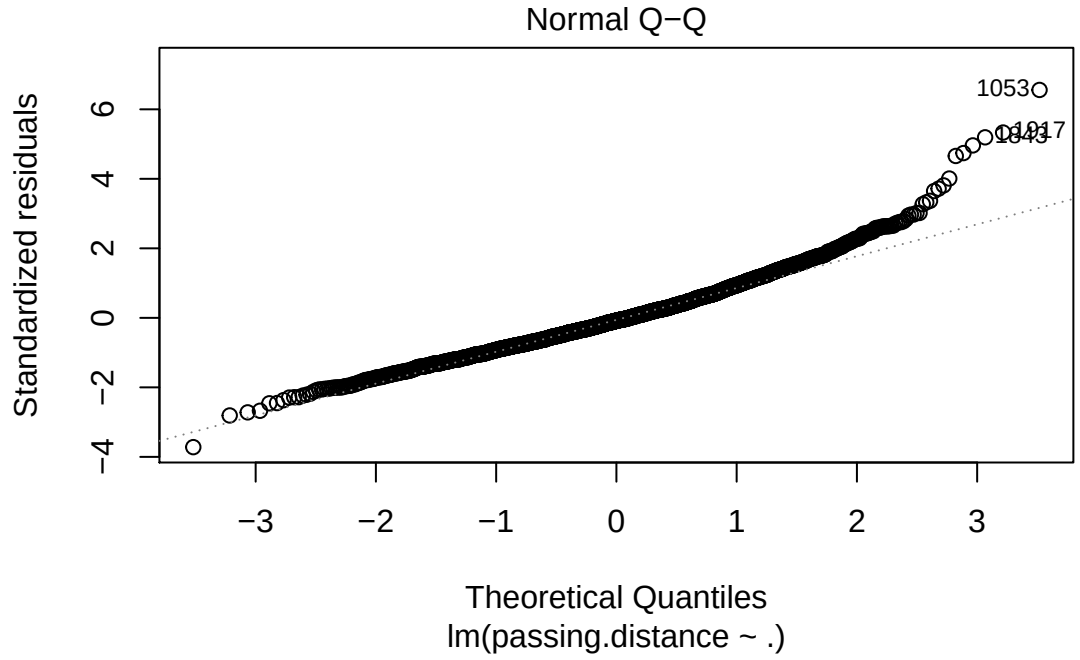
1) We fit a linear model with passing distance as response.

```
LM1 <- lm(passing.distance~.,data=Cycling)
summary(LM1)
```

```
##
## Call:
## lm(formula = passing.distance ~ ., data = Cycling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35650 -0.24557 -0.02978  0.20395  2.38417
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.542068   0.070501  21.873  < 2e-16 ***
## vehicleHGV                  0.120386   0.069337   1.736 0.082654 .
## vehicleLGV or minibus       0.222897   0.060571   3.680 0.000239 ***
## vehicleordinary car         0.249142   0.057225   4.354 1.40e-05 ***
## vehiclePTW                  0.286114   0.091877   3.114 0.001868 **
## vehicleSUV or pickup        0.253435   0.064498   3.929 8.77e-05 ***
## vehicletaxi                 0.163443   0.077215   2.117 0.034393 *
## colourblue                  0.002712   0.027392   0.099 0.921147
## colourgreen                -0.035421   0.038246  -0.926 0.354471
## colourother                 0.008794   0.057026   0.154 0.877455
## colourred                   0.006758   0.030186   0.224 0.822873
## coloursilver or grey       -0.002018   0.028110  -0.072 0.942783
## colourunknown              -0.223170   0.114354  -1.952 0.051112 .
## colourwhite                -0.034458   0.034875  -0.988 0.323236
## streetone-way (one lane)   -0.237461   0.122866  -1.933 0.053398 .
## streetone-way (two lanes)   0.172107   0.102252   1.683 0.092478 .
## streetresidential street    0.098060   0.059605   1.645 0.100075
## streetrural                 0.030756   0.259313   0.119 0.905598
## streeturban street         -0.059849   0.026713  -2.240 0.025160 *
## timeBefore 9:00            -0.001454   0.033372  -0.044 0.965264
## timeMiddle                  0.008013   0.018040   0.444 0.656946
## helmetYes                  -0.062755   0.015737  -3.988 6.88e-05 ***
## kerb                       -0.283093   0.024719 -11.453  < 2e-16 ***
## citySalisbury               0.027352   0.031190   0.877 0.380608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3653 on 2281 degrees of freedom
## Multiple R-squared:  0.102,  Adjusted R-squared:  0.09299
## F-statistic: 11.27 on 23 and 2281 DF,  p-value: < 2.2e-16
```

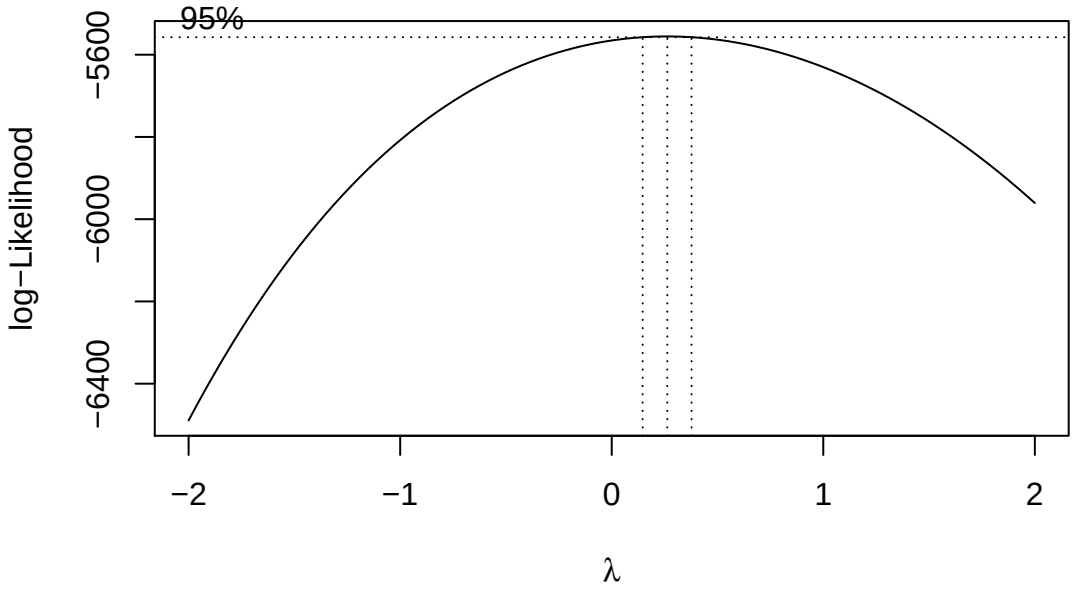
Interestingly, most of the vehicle types are significantly different from the base category `bus`. The type of street, the use of a helmet, and the presence of a kerb are also significant. However, the colour of the vehicle is unimportant. The QQ–plot looks bad:

```
plot(LM1,which=2)
```



The function `boxcox` fits a Box–Cox transform of the response at various values of λ and plots the maximum (or profile) log-likelihood as a function of λ. The plot shows a confidence region for the optimal value of the parameter which is assymptotically correct.

```
library(MASS)
boxcox(LM1)
```



Since the optimal value of λ is around 1/3, we repeat the fit after a cubic root transform of the response.

```r
LM2 <- lm(I(passing.distance)^(1/3)~.,data=Cycling)
summary(LM2)
```

```
##
## Call:
## lm(formula = I(passing.distance)^(1/3) ~ ., data = Cycling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46643 -0.05808 -0.00094  0.05447  0.44701
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.1463520  0.0172405  66.492  < 2e-16 ***
## vehicleHGV                  0.0278091  0.0169557   1.640 0.101121
## vehicleLGV or minibus       0.0568815  0.0148122   3.840 0.000126 ***
## vehicleordinary car         0.0634070  0.0139939   4.531 6.17e-06 ***
## vehiclePTW                  0.0705333  0.0224677   3.139 0.001715 **
## vehicleSUV or pickup        0.0642441  0.0157725   4.073 4.80e-05 ***
## vehicletaxi                 0.0386228  0.0188824   2.045 0.040926 *
## colourblue                 -0.0007789  0.0066986  -0.116 0.907441
## colourgreen                -0.0127362  0.0093528  -1.362 0.173412
## colourother                 0.0026996  0.0139453   0.194 0.846517
## colourred                   0.0004741  0.0073816   0.064 0.948799
## coloursilver or grey       -0.0025214  0.0068741  -0.367 0.713800
## colourunknown              -0.0523189  0.0279644  -1.871 0.061485 .
## colourwhite                -0.0090830  0.0085285  -1.065 0.286978
## streetone-way (one lane)   -0.0615784  0.0300457  -2.049 0.040529 *
## streetone-way (two lanes)   0.0392349  0.0250048   1.569 0.116764
## streetresidential street    0.0244551  0.0145760   1.678 0.093530 .
## streetrural                 0.0137793  0.0634128   0.217 0.827998
## streeturban street         -0.0153955  0.0065326  -2.357 0.018520 *
## timeBefore 9:00            -0.0007510  0.0081609  -0.092 0.926684
## timeMiddle                  0.0019798  0.0044114   0.449 0.653627
## helmetYes                  -0.0141585  0.0038484  -3.679 0.000239 ***
## kerb                       -0.0706982  0.0060448 -11.696  < 2e-16 ***
## citySalisbury               0.0082328  0.0076273   1.079 0.280526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08933 on 2281 degrees of freedom
## Multiple R-squared:  0.1066, Adjusted R-squared:  0.09754
## F-statistic: 11.83 on 23 and 2281 DF,  p-value: < 2.2e-16
```

2) The effect of wearing a helmet is highly significant, although probably not practically important—a helmet reduces the passing distance by 1.4 cm on average, while our estimate of the irreducible error σ is 8.9 cm.

3) Since the effect of wearing a helmet is linear in our model, a two-sample test would be able to detect it. However, controlling for all the other variables is very important because it reduces the variance of the response — we need fewer samples to detect a small effect.

   We assume that the data were collected in a randomized trial, where the decision to wear a helmet or not in each setting was random and independent from all other variables. If the data were observational, it would become even more important to control for any relevant covariates.

4) Doing backward selection with `stepAIC`, we obtain the following sequence of models.

```
stepAIC(LM2,direction="backward")
```

```
## Start:  AIC=-11111.33
## I(passing.distance)^(1/3) ~ vehicle + colour + street + time +
##     helmet + kerb + city
##
##           Df Sum of Sq    RSS    AIC
## - colour   7   0.05929 18.260 -11118
## - time     2   0.00249 18.204 -11115
## - city     1   0.00930 18.210 -11112
## <none>                 18.201 -11111
## - street   5   0.12088 18.322 -11106
## - helmet   1   0.10801 18.309 -11100
## - vehicle  6   0.26640 18.468 -11090
## - kerb     1   1.09152 19.293 -10979
##
## Step:  AIC=-11117.83
## I(passing.distance)^(1/3) ~ vehicle + street + time + helmet +
##     kerb + city
##
##           Df Sum of Sq    RSS    AIC
## - time     2   0.00322 18.264 -11121
## - city     1   0.00618 18.267 -11119
## <none>                 18.260 -11118
## - street   5   0.12953 18.390 -11112
## - helmet   1   0.10460 18.365 -11107
## - vehicle  6   0.33755 18.598 -11088
## - kerb     1   1.07508 19.335 -10988
##
## Step:  AIC=-11121.42
## I(passing.distance)^(1/3) ~ vehicle + street + helmet + kerb +
##     city
##
##           Df Sum of Sq    RSS    AIC
## - city     1   0.00511 18.269 -11123
## <none>                 18.264 -11121
## - street   5   0.12975 18.393 -11115
## - helmet   1   0.10864 18.372 -11110
## - vehicle  6   0.33730 18.601 -11091
## - kerb     1   1.35644 19.620 -10958
##
## Step:  AIC=-11122.78
## I(passing.distance)^(1/3) ~ vehicle + street + helmet + kerb
##
##           Df Sum of Sq    RSS    AIC
## <none>                 18.269 -11123
## - helmet   1   0.11628 18.385 -11110
## - street   5   0.26467 18.533 -11100
## - vehicle  6   0.34229 18.611 -11092
## - kerb     1   1.35135 19.620 -10960
```

```
##
## Call:
```

```
## lm(formula = I(passing.distance)^(1/3) ~ vehicle + street + helmet +
##     kerb, data = Cycling)
##
## Coefficients:
##               (Intercept)                 vehicleHGV
##                   1.15048                    0.02536
##     vehicleLGV or minibus        vehicleordinary car
##                   0.05553                    0.06613
##                vehiclePTW       vehicleSUV or pickup
##                   0.05947                    0.06650
##               vehicletaxi   streetone-way (one lane)
##                   0.04077                   -0.06020
## streetone-way (two lanes)   streetresidential street
##                   0.04034                    0.02512
##               streetrural         streeturban street
##                   0.01594                   -0.02056
##                 helmetYes                       kerb
##                  -0.01437                   -0.07007
```

The best model has the predictors `vehicle`, `helmet`, `street`, and `kerb`. It might be possible to improve the model by grouping categories of vehicle and street.