S631 HW10

Erik Parker

November 4th, 2017

1. Using the data *Robey.txt*, obtain type I and II anovas for the two linear models

$$tfr \sim region + contraceptors + region : contraceptors$$

and

$$tfr \sim contraceptors + region + region : contraceptors$$

Observe the models are the same except for the order of the main effects.

```r
rm(list = ls())

library(alr4)

robey <- Robey

m1 <- lm(tfr ~ region + contraceptors + region:contraceptors, data = robey)

m2 <- lm(tfr ~ contraceptors + region + region:contraceptors, data = robey)


# type I
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: tfr
##                      Df Sum Sq Mean Sq  F value Pr(>F)
## contraceptors         1 87.672  87.672 266.8706 <2e-16 ***
## region                3  1.677   0.559   1.7018 0.1812
## contraceptors:region  3  0.365   0.122   0.3706 0.7746
## Residuals            42 13.798   0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# type II
Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                      Sum Sq Df  F value    Pr(>F)
## contraceptors        45.045  1 137.1158 8.226e-15 ***
## region                1.677  3   1.7018    0.1812
## contraceptors:region  0.365  3   0.3706    0.7746
## Residuals            13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Type I
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: tfr
##                       Df Sum Sq Mean Sq  F value    Pr(>F)
## region                 3 44.304  14.768  44.9534 3.576e-13 ***
## contraceptors          1 45.045  45.045 137.1158 8.226e-15 ***
## region:contraceptors  3  0.365   0.122   0.3706    0.7746
## Residuals             42 13.798   0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Type II
Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                       Sum Sq Df  F value    Pr(>F)
## region                 1.677  3   1.7018    0.1812
## contraceptors         45.045  1 137.1158 8.226e-15 ***
## region:contraceptors   0.365  3   0.3706    0.7746
## Residuals             13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Interpret each line of the output for type I and II anova tests for the first model.

Interpretation of the Type I anova

First, the type I anova is referred to as a "sequential" analysis of variance, where models are fit according to the order that regressors are entered into the mean function. So, in the first line, the null hypothesis being tested is that the mean function is fully described just by the intercept, while the alternative hypothesis is that the first regressor, *region*, significantly influences the response and so its addition improves the explanatory power of the model. The resulting p_value for this test is close to zero, meaning that we can reject the null hypothesis, here that the mean function is fully described by the intercept alone, and so we can say that *region* has a meaningful impact when added to an otherwise empty model. Because this test is sequential, the next line we address is the second one, where the null hypothesis being tested is that the previous regressor, *region*, alone is sufficient to describe the mean function. The alternative here is that the addition of *contraceptors* to the model significantly improves the variance seen in the response. Again we see that the resulting p-value for this test is low, so we are able to reject the null hypothesis that the first regressor alone adaquately explains the mean function, and conclude that *contraceptors* should be incuded in the model. Finally, the last line of the type I anova is testing the null hypothesis that the two main effects addressed so far fully describe the mean function. The alternative is that the interaction term adds explanatory power to the model and should be included. We see from the large p-value that we are not able to reject the null hypothesis in this case, and so would not include the interaction term. Overall, if we trust the results from the type I anova, our final model would include both main effects.

The second test performed is with a type II anova which follows the marginality principle. This means that we first test the influence of higher order terms (interactions), and only move to testing the lower order terms (lower interactions and main effects) if the higher order terms are found to be insignificant. So, here we read the test output from bottom to top, and start with the third line, representing the test of *region:contraceptors*. The null hypothesis here is that the mean function is fully described by the preceeding main effects, and the alternative then is that the interaction term in non-zero and adds to the model. With a large p-value resulting from this test we can conclude, like above, that the interaction is not significant and not necessary to include in the model. Because this higher order term was not significant, we then move to test the two main effects, first the effect of *contraceptors*. Unlike above, type II tests are not sequential and test for the effects of adding each main effect to a model already containing the others. So here the null hypothesis for the *contraceptors* line is that the mean function is fully described just by the presence of *region*, and the alternative is that *contraceptors*, when added, further explains the variability seen in the response. We get a p-value near zero for this test, allowing us to reject the null and conclude that *contraceptors* should be included in the model. Finally, the line above tests the null hypothesis that the mean function is fully described by *contraceptors*, and the influence of *region* is zero, while the alternative hypothesis states that the influence of *region* is non-zero. This test results in a large p-value, leading to the conclusion that the null hypothesis can't be rejected and the effect of *region* on the mean function is not different from zero. Overall, the use of the type II anova leads us to a different conclusion than the type I, and we see here that the mean function is best described just by the main effect *contraceptors*.

b) Why does the type II anova provide the same output for both models, but the type I anova doesn't?

As discussed in part a, the type I anova is sequential, and so the order in which the the regressors are entered into the model are important as that is the order in which they will be tested. In this model, *contraceptors* is really the only significant regressor, and so when it is tested second in the first model it allows *region* to turn up as significant even though it really isn't when the influence of *contraceptors* is already accounted for.

c) Using F-tests to compare full and reduced models, choose the most appropriate model. Do you obtain the same conclusions you obtained previously (in HW08)? Explain why or why not.

```r
m3 <- lm(tfr ~ region + contraceptors, data = robey)

m4 <- lm(tfr ~ contraceptors, data = robey)

anova(m3, m2)
```

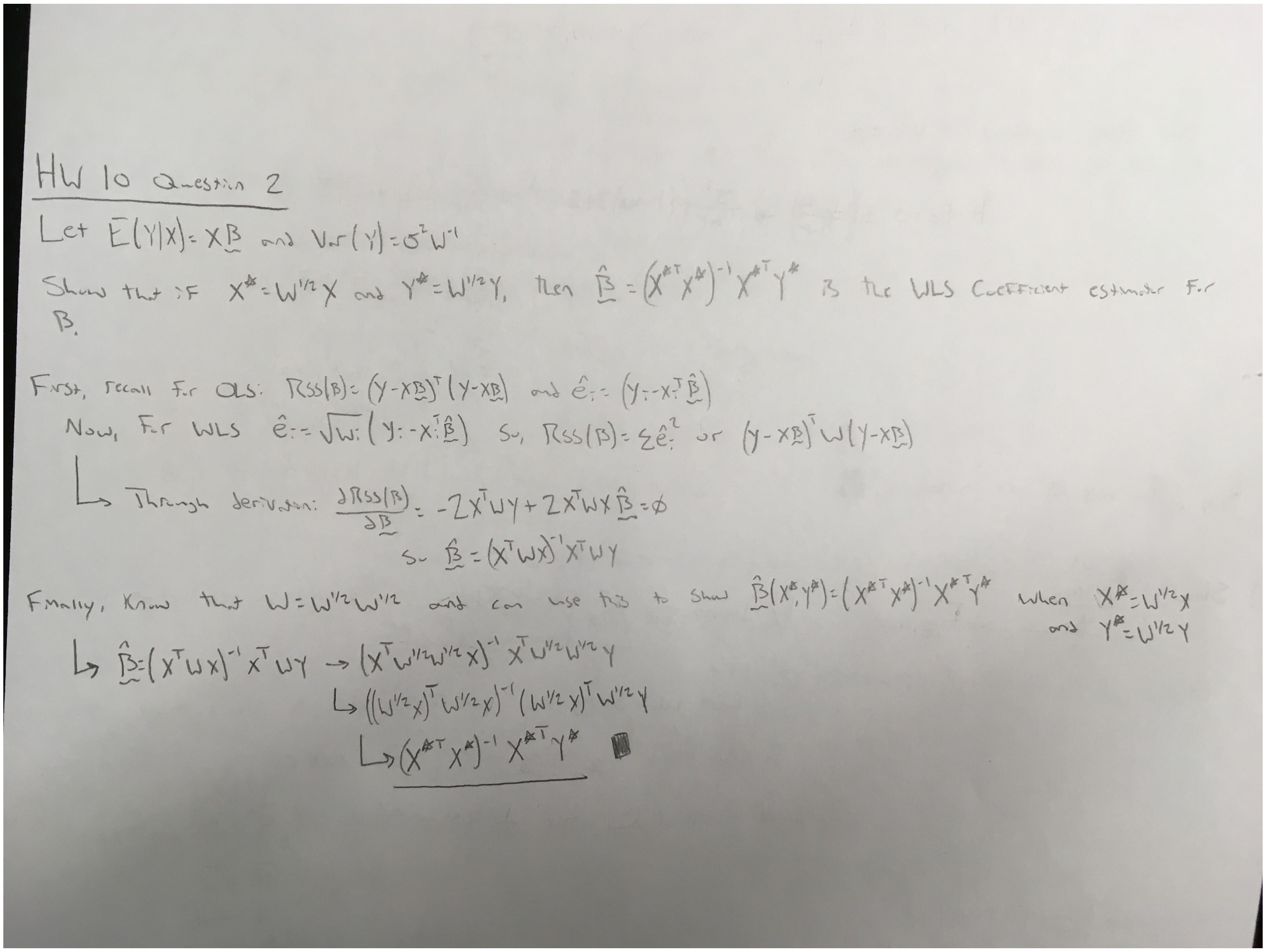
```
## Analysis of Variance Table
##
## Model 1: tfr ~ region + contraceptors
## Model 2: tfr ~ contraceptors + region + region:contraceptors
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 14.163
## 2     42 13.798  3   0.36524 0.3706 0.7746
```

```r
anova(m4, m3)
```

```
## Analysis of Variance Table
##
## Model 1: tfr ~ contraceptors
## Model 2: tfr ~ region + contraceptors
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 15.840
```

```
## 2     45 14.163  3    1.6772 1.7764 0.1652
```

So, with these F-tests, it seems like the most appropriate model is m4, where the only regressor is the main effect of *contraceptors*. This is the same conclusion I reached in HW 8, because there I saw graphically and through the significance tests on individual coefficients for different models, that the only significant regressor here is the continuous one.


HW 10 Question 2

Let $E(Y|X) = X\beta$ and $Var(Y) = \sigma^2 W^{-1}$

Show that if $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$, then $\hat{\beta} = (X^{*T}X^*)^{-1}X^{*T}Y^*$ is the WLS coefficient estimator for β.

First, recall for OLS: $RSS(\beta) = (Y - X\beta)^T(Y - X\beta)$ and $\hat{e}_i = (y_i - x_i^T\hat{\beta})$

Now, for WLS $\hat{e}_i = \sqrt{w_i}(y_i - x_i^T\hat{\beta})$ so, $RSS(\beta) = \sum \hat{e}_i^2$ or $(Y - X\beta)^T W (Y - X\beta)$

↳ Through derivation: $\frac{\partial RSS(\beta)}{\partial \beta} = -2X^TWY + 2X^TWX\hat{\beta} = \emptyset$

so $\hat{\beta} = (X^TWX)^{-1}X^TWY$

Finally, know that $W = W^{1/2}W^{1/2}$ and can use this to show $\hat{\beta}(X^*, Y^*) = (X^{*T}X^*)^{-1}X^{*T}Y^*$ when $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$

↳ $\hat{\beta} = (X^TWX)^{-1}X^TWY \to (X^TW^{1/2}W^{1/2}X)^{-1}X^TW^{1/2}W^{1/2}Y$

↳ $((W^{1/2}X)^TW^{1/2}X)^{-1}(W^{1/2}X)^TW^{1/2}Y$

↳ $(X^{*T}X^*)^{-1}X^{*T}Y^*$ ∎

Figure 1: Question 2 proof

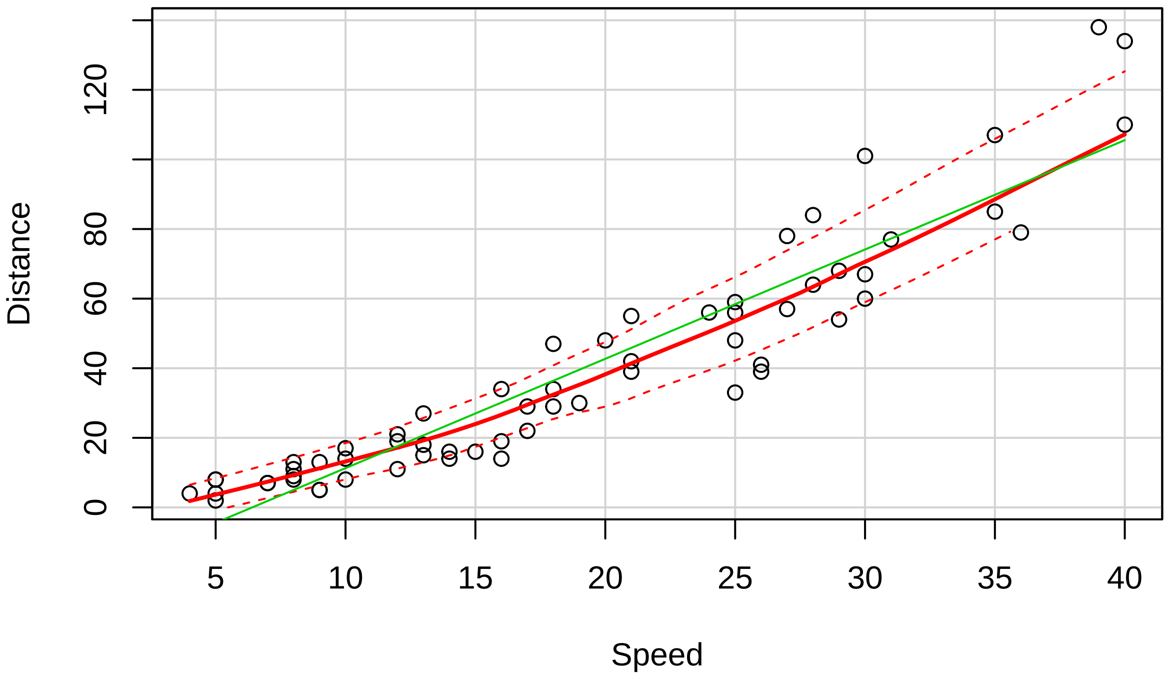
3. ALR 7.6.1 - 7.6.4

ALR 7.6: The (hypothetical) data in the file give automobile stopping *Distance* in feet and *Speed* in mph for n = 62 trials of various automobiles

7.6.1: Draw a scatterplot of *Distance* versus *Speed*. Explain why this supports fitting a quadratic regression model.

```
auto <- stopping

scatterplot(Distance ~ Speed, data = auto, boxplots = FALSE)
```



This plot supports fitting a quadratic regression model because the datapoints are not distributed along a line, there is a clear curve to the data which is well illustrated by the better fit of the red loess curve compared to the green OLS line.

7.6.2: Fit the quadratic model with constant variance. Compute the score test for nonconstant variance for the alternatives that a) variance depends on the mean, b) variance depends on $Speed$, c)variance depends on $Speed$ and $Speed^2$. Is adding $Speed^2$ helpful?

```
mauto <- lm(Distance ~ Speed + I(Speed^2), data = auto)

Z1 = with(auto, ncvTest(mauto))
Z2 = with(auto, ncvTest(mauto, ~Speed))
Z3 = with(auto, ncvTest(mauto, ~Speed + I(Speed^2)))
table1 = rbind(with(Z1, c(Df, ChiSquare, p)), with(Z2, c(Df, ChiSquare, p)),
    with(Z3, c(Df, ChiSquare, p)))
row.names(table1) = c("Fitted Values (mean)", "Speed", "Speed and Speed^2")
colnames(table1) = c("df", "Test statistic", "p-Value")
table1
```

```
##                      df Test statistic      p-Value
## Fitted Values (mean)  1       22.97013 1.645386e-06
## Speed                 1       23.39216 1.321162e-06
## Speed and Speed^2     2       23.46559 8.026245e-06
```

Based on the results of the test for nonconstant variance, it seems like the addition of the quadratic term $Speed^2$ does ever so slightly help. The p-value for the test containing both $Speed$ and $Speed^2$ is very slightly larger than that for the test just with $Speed$, meaning we are just a very slight amount closer to accepting the null hypothesis that the variance is constant.

7.6.3: Refit the quadratic assuming $Var(Distance|Speed) = Speed * \sigma^2$. Compare the estimates and their standard errors with the unweighted case.

```r
mauto2 <- lm(Distance ~ Speed + I(Speed^2), data = auto, weights = 1/Speed)

summary(mauto)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + I(Speed^2), data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5192  -5.4527  -0.5519   3.8442  27.9373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.58036    5.10266   0.310    0.758
## Speed        0.41607    0.55641   0.748    0.458
## I(Speed^2)   0.06556    0.01303   5.033 4.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.927 on 59 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9115
## F-statistic: 315.3 on 2 and 59 DF,  p-value: < 2.2e-16
```

```r
summary(mauto2)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + I(Speed^2), data = auto, weights = 1/Speed)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0037 -1.4120 -0.1054  1.2586  5.0984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.32590    3.09898   0.428    0.670
## Speed        0.44801    0.42065   1.065    0.291
## I(Speed^2)   0.06479    0.01122   5.777 3.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.011 on 59 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9204
## F-statistic: 353.8 on 2 and 59 DF,  p-value: < 2.2e-16
```

When we compare the weighted and unweighted models, we can see that in general the standard errors decreased when compared to the unweighted model. The estimated coefficients themselves are a little more mixed, with the intercept and quadratic term decreasing slightly and the coefficient for *Speed* increasing slightly. There is also some movement in the significance values for these estimated coefficients, and a slight increase in the R^2. But overall, there seems to be nothing that would change our overall interpretation.

7.6.4: Based on the unweighted model, use a sandwich estimator to correct for nonconstant variance. Compare results to 7.6.3.

```
library(lmtest)

coeftest(mauto, vcov = hccm)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1.580363   4.295827  0.3679 0.7142767
## Speed       0.416068   0.630317  0.6601 0.5117625
## I(Speed^2)  0.065556   0.017248  3.8008 0.0003439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, when we use the sandwich estimator of variance to correct for nonconstant variance in our original, unweighted *auto* model, we end up with quite similar results. The standard errors of the estimated coefficients for the intercept and *Speed* regressor both increase slightly compared to the weighted model, while the error of the quadratic term decreases. Compared to the unweighted model, only the standard error of the intercept coefficient decreases, while it increases for both regressors. Most importantly though, none of the p-values obtained change much at all when we use the sandwich estimators - the most change is seen in the decrease in significance of the quadratic term, from a p-value of $4.8e^{-6}$ to $3.4e^{-4}$. This isn't enough of a change to really alter the interpretation of this model though.