ASM - Second delivery

Laura Cebollero, Alexandre Rodriguez

12th of October, 2018

Example 2 ANCOVA questions

Let us assume the correct model to work with on these questions is the last one where a log-log scale has been applied:

```
model_init <- lm(pcBfat~ssf+sex)
std_res <- stdres(model_init)
ais2 <- ais[-which(std_res>3),]

model <- lm(log(pcBfat)~log(ssf)+sex, data=ais2)
```

1) Deduce which is the change in the pcBfat by changing one unit the ssf?

Using the log-log model, we can check its betas:

```
model$coefficients
```

```
## (Intercept)    log(ssf)        sexm
##  -0.8028811   0.8278677  -0.2358411
```

In this case, we can see that it follows the formula:

pcBFat = 0.827x (ssf) - 0.235x (sexm) - 0.802 + e

So by changing **one unit the ssf**, **pcBFat changes 0.235** each time, in logarithmic scale.

To obtain the change without a logarithmic scale we can use an exponential.

```
exp(model$coefficients[2])
```

```
## log(ssf)
## 2.288434
```

In this case, it means that of one unit in ssf means a change of 2.28 units in the real scale of pcBFat.

2) Which is the pcBfat predicted for a male with a ssf of 54?

Using the linear model we have created on question 1 we can predict this value easily:

```
(pred_log <- predict(model, newdata = data.frame(sex = "m", ssf=54)))
```

```
##        1
## 2.263629
```

```
exp(pred_log)
```

```
##        1
## 9.617927
```

In this case, the predicted pcBfat is 9.6 in the real scale, and 2.26 in the logarithmic one.

3) Which is the pcBfat predicted for a female with a sff of 48?

Applying the same methodology:

```
(pred_log <- predict(model, newdata = data.frame(sex = "f", ssf=48)))
```

```
##        1
## 2.401961
```

```
exp(pred_log)
```

```
##        1
## 11.04482
```

Now the estimated pcBfat is 11.04 in the real scale and 2.40 in the logarithmic scale.

4) Which has to be the ssf of a male in order to have a pcBfat of 14.7? and for a female?

For this case we have to change our linear model so that it predicts the ssf instead of pcBfat. We are not applying a logarithmic scale in this example.

```
model_2 <- lm(ssf~pcBfat+sex)
model_2$coefficients
```

```
## (Intercept)      pcBfat        sexm
##  -19.108742    5.943254   15.550944
```

Once we have created the model we can proceed onto the predictions:

```
predict(model_2, newdata = data.frame(sex = "m", pcBfat=14.7))
```

```
##        1
## 83.80804
```

```
predict(model_2, newdata = data.frame(sex = "f", pcBfat=14.7))
```

```
##       1
## 68.2571
```

For the male to have a pcBfat of 14.7 units it has to have an ssf of 83.8080.

In case of the female to have the same pcBfat, it has to have a lower value than the male one: 68.2571.

5) Interpret the standard error value.

In order to see the error and interpret it we will first compute the anova table associated with the linear model that we used to predict the values.

```
Anova(model, type="II")
```

```
## Anova Table (Type II tests)
##
## Response: log(pcBfat)
##            Sum Sq  Df F value    Pr(>F)
## log(ssf)  17.9469   1 3027.90 < 2.2e-16 ***
## sex        1.8776   1  316.77 < 2.2e-16 ***
## Residuals  1.1736 198
```

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

Since we know the degrees of freedom and the Sum of Squares of the Residuals we can use this data to calculate the standard error with the following formula:

$$\sigma^2 = \frac{SSE}{d.f.}$$

Therefore:

$$\sigma = \sqrt{\frac{SSE}{d.f.}}$$

In our case, the estimated standard error is $\sqrt{\frac{1.1736}{199}} = 0.0767$ To know if this value is too big or otherwise we can compare it to the mean of the response variable. Considering that the mean of ssf is 69.02 this error represents only a 0.11%, which is very little.

Birthweights analysis

First we load the data and take a quick peek to it to check its nature:

```
load("birthw.RData")
summary(data)
```

```
##      sex       age.weeks.    birthweight.g.
##  Female:12   Min.   :35.00   Min.   :2412
##  Male  :12   1st Qu.:37.00   1st Qu.:2785
##              Median :38.50   Median :2952
##              Mean   :38.54   Mean   :2968
##              3rd Qu.:40.00   3rd Qu.:3184
##              Max.   :42.00   Max.   :3473
```

We can see that it only has 3 columns:

- **Sex of the baby**. Factor. Either female or male.
- **Gestational age in weeks**. Numeric. Ranges from 35 to 42.
- **Birthweight in grams**. Numeric. Ranges from 2412 to 3473.

If we take a look at how the data is loaded we can see that it's interpreting their classes correctly:

```
sapply(data, class)
```

```
##            sex     age.weeks. birthweight.g.
##       "factor"      "integer"      "integer"
```

And from the summary we can also observe that there is no missing data in any of these fields.

So we are ready to proceed analyzing the data.

Outliers detection

Now, we are going to try to detect outliers.

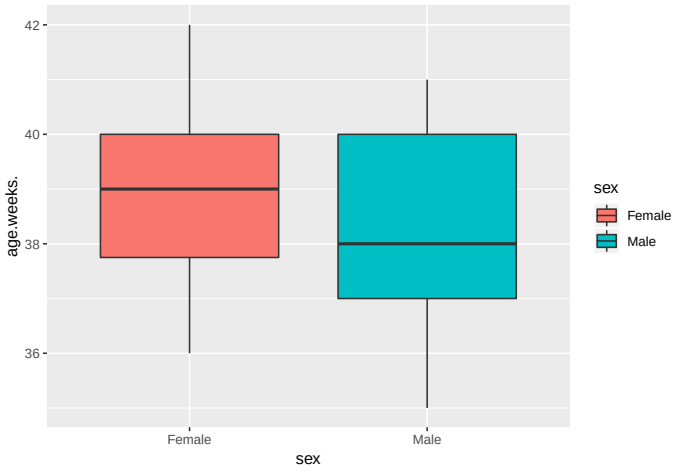


Figure 1: Boxplot of gestation age variable by sex

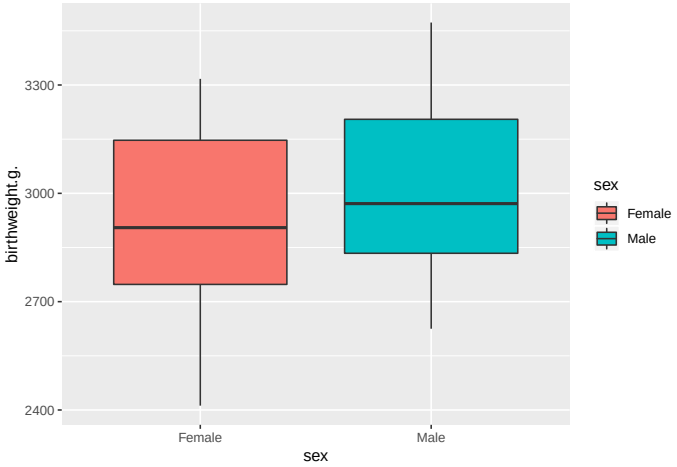


Figure 2: Boxplot of Birthweight variable by sex

Since the range for age is way shorter and lower than the birtweight one, we are going to plot the data in different boxplots' graphics. Also, we will plot them separated by their gender. This will, in some way, show us if there is some relevant difference between both genders.

```
ggplot(data, aes(x=sex, y=age.weeks., fill=sex)) +
geom_boxplot()
```

```
ggplot(data, aes(x=sex, y=birthweight.g., fill=sex)) +
geom_boxplot()
```

From these plots we cannot infere that there exist outliers.

However we can see some differences between the two genders (Female and Male respectively):

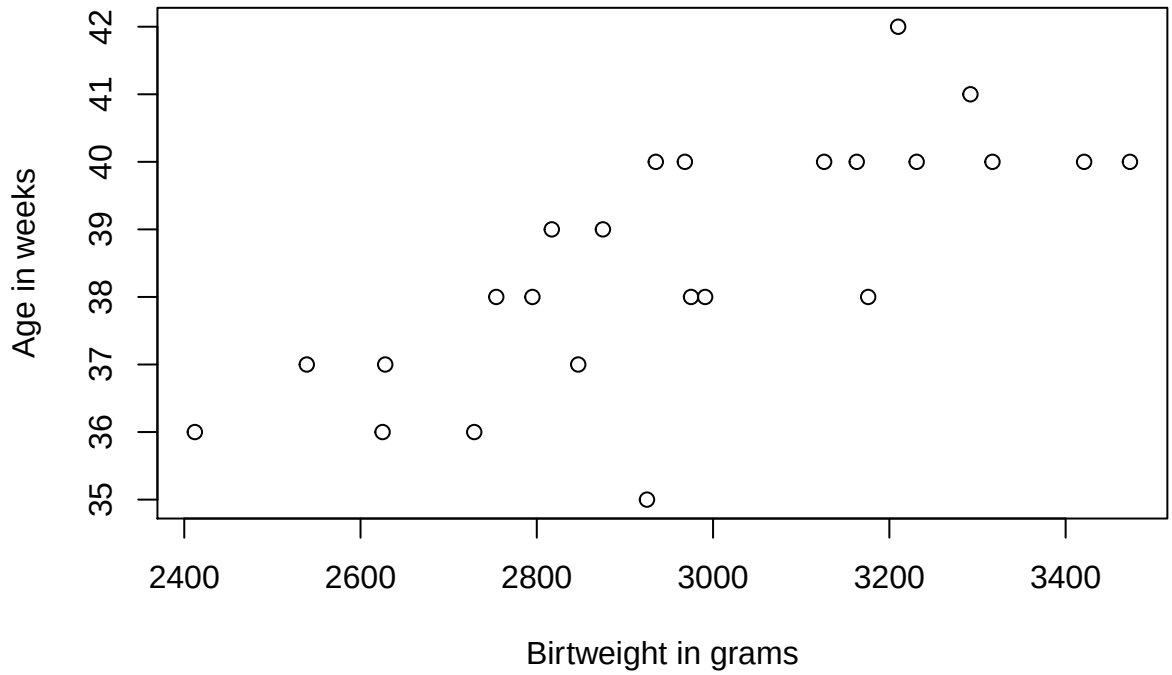
- We can see that the dispersion of the 'Male' gender is slightly greater than the 'Females' and while the median for Females lies on 39 it is 38 for Males.
- The whiskers cover the same range for both genders but it starts on 36 for 'Female' and on 35 for 'Male'. None of the boxplots are symmetric.
- For Birthweigh there is a clear difference between 'Female' and 'Male'. Clearly Males weigh more than

Females on birth: Female's median is roughly 2900 while Male's is almost 3000.

Let's check now if these two variables are correlated. Our initial Hypothesis (H_0) is that they are.

First we will check it with scatterplots:

```
plot(data$age.weeks. ~ data$birthweight.g., ylab = "Age in weeks", xlab = "Birtweight in grams")
```



It seems that there may be a slight positive linear relation between the variables. Let's check the correlation:

```
cor(data$age.weeks., data$birthweight.g.)
```

```
## [1] 0.7443298
```

Their correlation value is quite high so we cannot reject our H_0 stating that they are related.

Variables influence on each other

To check the possible influence of the baby's gender and gestational age onto the birthweight we are going to create a linear model.

```
birth_lm <- lm( data$birthweight.g. ~ data$age.weeks. + data$sex, data)
summary(birth_lm)
```

```
##
## Call:
## lm(formula = data$birthweight.g. ~ data$age.weeks. + data$sex,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257.49 -125.28  -58.44  169.00  303.98
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -1773.32     794.59  -2.232   0.0367 *
## data$age.weeks.   120.89      20.46   5.908 7.28e-06 ***
## data$sexMale      163.04      72.81   2.239   0.0361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 177.1 on 21 degrees of freedom
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6057
## F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05
```

The baseline of this linear model is that the baby's gender is female.

We can see how this model gives a lot of importance to the gestational age, but not so much to the sex of the baby, with a low value on the adjusted R^2 of 0.605. We can see that all variables in this model are important because the p-value is in all cases lower than 0.05, which tells us that we can reject the null hipothesis that the coefficient for each variable is 0. This means that both `weeks` and `sex` are important variables that should be taken into account when building a linear model to describe the variable `birhtweight`.

Let's try removing the gender of the baby and see the model results:

```
birth_lm2 <- lm( data$birthweight.g. ~ data$age.weeks., data)
summary(birth_lm2)
```

```
##
## Call:
## lm(formula = data$birthweight.g. ~ data$age.weeks., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.03 -158.29    8.35   88.15  366.50
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1485.0      852.6  -1.742   0.0955 .
## data$age.weeks.    115.5       22.1   5.228 3.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 192.6 on 22 degrees of freedom
## Multiple R-squared:  0.554,  Adjusted R-squared:  0.5338
## F-statistic: 27.33 on 1 and 22 DF,  p-value: 3.04e-05
```

Now the adjusted R^2 has an even lower value of 0.533. We should take into account that the adjusted R^2 penalizes models that have many parameters. However, this is not the case since we have lost one parameter in respect to the previous model. Therefore this model can be considered worse than the previous one because it only captures 53% of the variability of the variable `weeks`, which is lower than the 60% from the previous model.

Model selection

If we were to choose a model we would choose the **first one** since it has the highest Adjusted R^2 value. This model looks like

$$Y = 120.89w + 163.04g - 1773.32 + e$$

(where w is the variable and g is the variable gender, which will account for 163.04 if the baby is a male). This model makes more sense because it takes into account the gender of the baby which is recommended because, as we saw before, male babies weigh more than female babies.

This model passes the **omnibus test**. If we look at the F-statistic the p-value = 2.194e-05, meaning that the model, globally, is explanatory, for it is below 0.05, and so it explains an important part of the variability.

If we compute the ANOVA table related to this selected model:

```
Anova(birth_lm, type = "II")
```

```
## Anova Table (Type II tests)
##
## Response: data$birthweight.g.
##                  Sum Sq Df F value    Pr(>F)
## data$age.weeks. 1094940  1 34.9040 7.284e-06 ***
## data$sex         157304  1  5.0145   0.03609 *
## Residuals        658771 21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can confirm that both variables are significant for their p-value is also below 0.05.

By looking at the sum of squares and seeing that both variables have the same degrees of freedom, we can also state and confirm that the gestation age in weeks has a way larger influence than sex.

And 1094940 variability respect to the total is explained by the Age gestation variable, whereas a way smaller number of 157304 is explained by the baby's gender variable.

If we check the std. error estimation:

```
a = Anova(birth_lm, type = "II")
attributes(a)
```

```
## $names
## [1] "Sum Sq"  "Df"      "F value" "Pr(>F)"
##
## $class
## [1] "anova"      "data.frame"
##
## $row.names
## [1] "data$age.weeks." "data$sex"        "Residuals"
##
## $heading
## [1] "Anova Table (Type II tests)\n" "Response: data$birthweight.g."
```

```
(res_e <- sqrt(a$`Sum Sq`[3]/a$Df[3]))
```

```
## [1] 177.1159
```

```
(mean_birthw <- mean(data$birthweight.g.))
```

```
## [1] 2967.667
```

```
(100 * res_e) / mean_birthw
```

```
## [1] 5.968187
```

We can see that it is quite large, but if we compare it to the mean of the birthweight this value is not that large in comparison, for it amounts to almost 6% of the mean.

Hypothesis checking

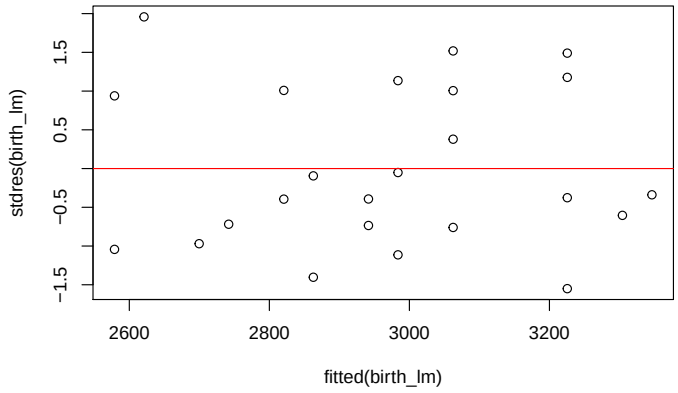
Now we are going to perform a residual error analysis to check the hypothesis of:

- Normality
- Homocedasticity
- Independence

First we standarize the residuals:

```
hist(stdres(birth_lm), col="darkred")
```


```
plot(fitted(birth_lm), stdres(birth_lm))
abline(h = 0, col='red')
```



In order to see how our residuals are distributed we used two plots: A histogram and a plot of the standardized residuals and the bodyweight.

From the histogram we can see that the distribution of the residuals is somewhat centered close to 0 —most of them between 0 and -0.5— with some values close to the 1.5 values, both positive and negative. We can see that there are no values above a value of 2 in absolute value, which is a really good sign. Also, judging from the histogram we can expect a fairly symmetrical plot, since even though it is not perfectly balanced, 9 values have standardized residuals above 0 —which represents a 37.5% of the predictions— and the 15 remaining predictions have a residual lower than 0.

Looking at the scatterplot we can confirm that the distribution of the points looks quite symmetrical. Also we can't really see any particular pattern so we can say that the residuals follow a **random** pattern. This suggests that there might be some linear relationship between the independent and dependant variable. Hence a linear model is appropiate for the data and it makes sense to use linear regression for the analysis fo the data.

Now we are going to check whether the slope of our model changes depending on the gender of the baby.

```
birth_lm_gender <- lm(data$birthweight.g. ~ data$age.weeks. + data$sex + data$age.weeks.:data$sex, data
summary(birth_lm_gender)
```

```
##
## Call:
## lm(formula = data$birthweight.g. ~ data$age.weeks. + data$sex +
##     data$age.weeks.:data$sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -246.69 -138.11  -39.13  176.57  274.28
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2141.67    1163.60  -1.841 0.080574 .
## data$age.weeks.                 130.40      30.00   4.347 0.000313 ***
## data$sexMale                    872.99    1611.33   0.542 0.593952
## data$age.weeks.:data$sexMale    -18.42      41.76  -0.441 0.663893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 180.6 on 20 degrees of freedom
## Multiple R-squared:  0.6435, Adjusted R-squared:   0.59
## F-statistic: 12.03 on 3 and 20 DF,  p-value: 0.000101
```

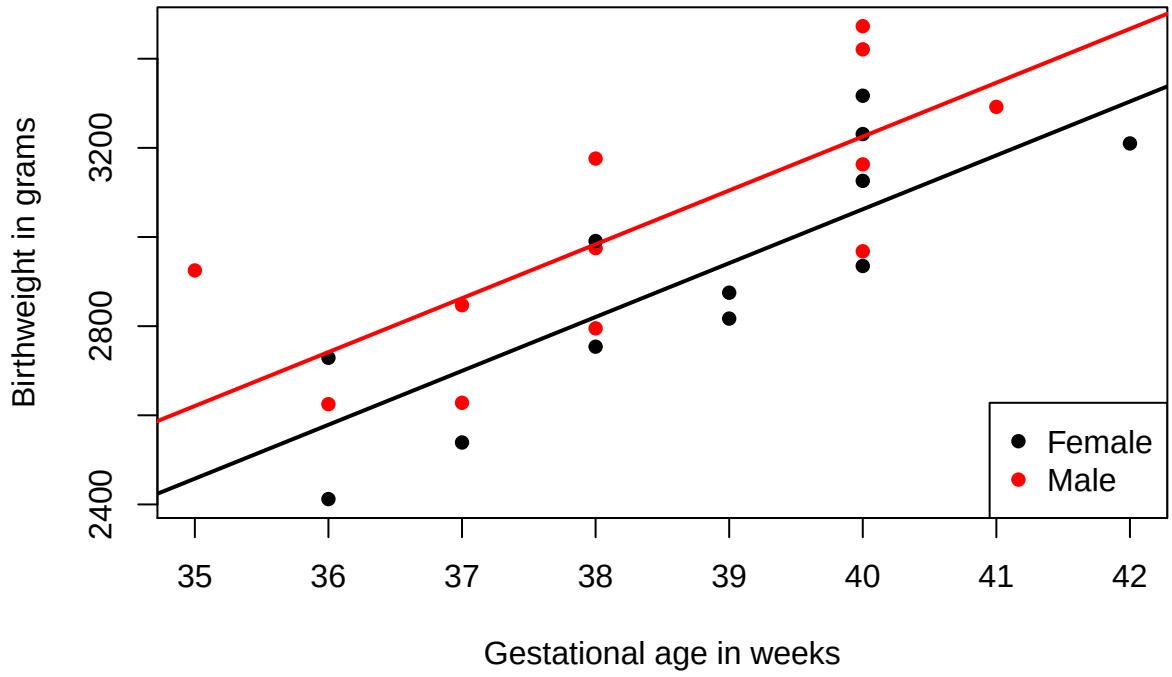
We can see that the interactive **term age.weeks.:~sexMale** is not significative, so we can conclude that the rate of change of the birthweight is the same for each one of the genders: female and male.

The adjusted R^2 has slightly increased, as it always happens when you add more parameters to your model, thus making it more complex, so we determine that this increase is not important for our study.

Finally, we are going to compare both slopes graphically with the existing data and check that they have, in fact, the same slope but a different intercept.

```
plot(x=data$age.weeks.,
    y= data$birthweight.g.,
    col = data$sex,
    pch = 16,
    xlab = "Gestational age in weeks",
    ylab = "Birthweight in grams")

legend('bottomright',
    legend = levels(data$sex),
    col = 1:2,
    cex = 1,
    pch = 16)


I = -1773.32
age_weeks = 120.89

line_gend_female <- I + 0
line_gend_male <- I + 163.04

abline(line_gend_female, age_weeks,lty=1, lwd=2, col = 1)
```

```r
abline(line_gend_male, age_weeks,lty=1, lwd=2, col = 2)
```



Conclusions

- The model appropiate when the baby's gender is female is:

$$BirthWeight = -1773.32 + 120.89Age$$

and for when the baby is a male:

$$BirthWeight = -1610.28 + 120.89Age$$

- Without taking into account the baby's gender, **an increment of one unit of the Baby's age will account for 120.89 grams in the baby's birthweight**.
- If we imagine the case of a baby at the 20th week of gestation (Age = 20), then when the baby's gender is female, the estimated weight of that baby is:

$$BirthWeight = -1773.32 + 120.89 * 20$$

which translates to:

$$BirthWeight = -1773.32 + 2417,8$$

$$BirthWeight = 644,48$$

and for when the baby is a male:

$$BirthWeight = -1610.28 + 120.89 * 20$$

$$BirthWeight = -1610.28 + 2417,8$$

$$BirthWeight = 807,52$$