Introduction into Linear Regression

Vitaly Belik

Institute for Veterinary Epidemiology and Biostatistics

March 11, 2019

(Lineare) Regression

$$Y = f(X_1, ..., X_n)$$

Regression ist ein weit gefasster Begriff für eine Reihe von Methoden, die zur Vorhersage einer *Antwort*-Variable (oder einer *abhängigen, resultierenden*) aus einer oder mehreren *Prädiktor*-Variablen (*unabhängigen, erklärenden*) verwendet werden.

Ziele der Regression

- *Bestimmung* der erklärenden Variablen, die sich auf die Antwortvariable beziehen
- *Beschreibung* der Form der Beziehung
- *Bereitstellen* einer Gleichung für die *Vorhersage* der Antwortvariablen aus den erklärenden Variablen

Aus der Regression konnte keine *Kausalität* direkt abgeleitet werden!

Regression example (1)

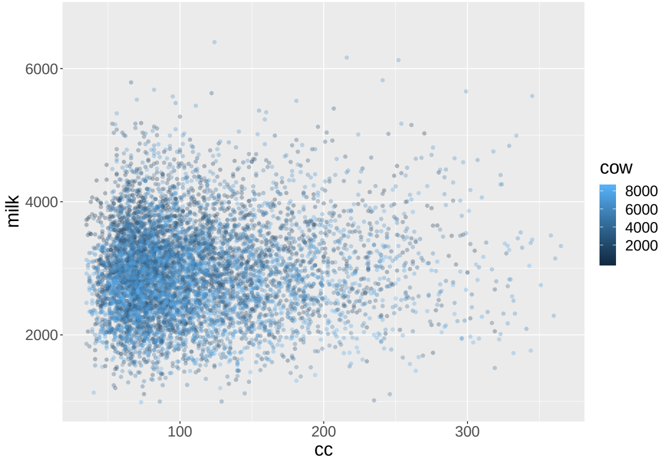

Figure 1: Impact of CC interval (interval from calving to conception) on Milk volume. 9383 lactation records in 42 year-round calving herds. http://projects.upei.ca/ver/data-and-samples/.

Regression example (2)

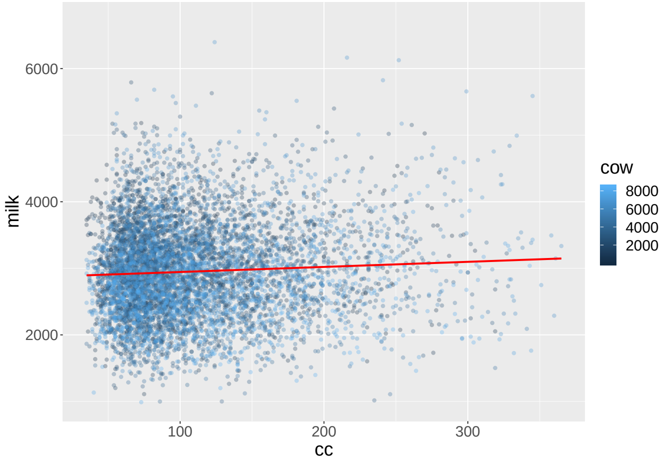

Figure 2: Impact of CC interval (interval from calving to conception) on Milk volume. 9383 lactation records in 42 year-round calving herds. http://projects.upei.ca/ver/data-and-samples/.

Regression example (3)

```r
library(ggplot2)
load("~/Lehre/WS1819/Spring_School/ver-master/data/ver2_data_R/daisy2.rdata")
ggplot(daisy2, aes(x = cc, y = milk120, colour = cow)) +
theme(text = element_text(size=18)) +
geom_point(alpha = 0.3) +
labs(x = "cc", y = "milk") +
geom_smooth(fill=NA,color="red",size=1,method="lm")
```

Regression example: model output

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(milk120~cc, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = milk120 ~ cc, data = daisy2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2031.7  -490.9   -35.4   432.1  3434.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2867.6913    18.9499 151.330  < 2e-16 ***
## cc             0.7652     0.1526   5.014 5.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 705.1 on 7032 degrees of freedom
##   (2349 observations deleted due to missingness)
## Multiple R-squared:  0.003562,   Adjusted R-squared:  0.00342
## F-statistic: 25.14 on 1 and 7032 DF,  p-value: 5.471e-07
```

Regression example: ANOVA table

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: milk120
##             Df     Sum Sq  Mean Sq F value    Pr(>F)
## cc           1   12495528 12495528  25.137 5.471e-07 ***
## Residuals 7032 3495627193   497103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA table

$$\hat{Y} = \beta_0 + \beta_1 X$$

To assess how much information the variable X contains about the variable Y, we consider how much of the *sum of squares* (SS) of the variable Y could be explained by the knowledge of the variable X.

Table 1: Decomposition of sums of squares in regression model with k predictor variables [after Dohoo, p. 327]

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Model	$SSM = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$dfM = k$	$MSM = \frac{SSM}{dfM}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$dE = n - (k+1)$	$MSE = \frac{SSE}{dfE}$	
Total	$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$	$dfT = n - 1$	$MST = \frac{SST}{dfT}$	

$MSE = \sigma^2$ and σ is called *standard error of prediction*.

F-test to assess the model significance

H_0: $\beta_1 = \beta_2 ... \beta_k = 0$, $\beta_0 \neq 0$

H_1: at least some of the coefficients $\beta_i \neq 0$, $i \neq 0$

- Variables are chosen in a way maximizing F-statistic
- F-test has a straightforward meaning if independent variables are manipulated treatments in a controlled experiment
- Caution if applied to observational variables (influenced by the number of variables, their correlations and sample size)

t-test (with $n-(k+1)$) degrees of freedom (1)

H_0: $\beta_i = \beta^* \; e.g. \beta^* = 0$

H_1: $\beta_i \neq \beta^* \; e.g. \beta^* = 0$

$$t = \frac{\beta_i - \beta^*}{SE(\beta_i)}$$

$SE(\beta_i)$ is a *standard error of the esimated coefficient*.

In the case of single predictor

$$SE(\beta_1) = \sqrt{\frac{MSE}{SSX(\beta_1)}},$$

where $SSX_1 = \sum_{i=1}^{n} \left(X_{1,i} - \bar{X_1}\right)^2$ is *sum of square of the variable X*.

t-test (with $n - (k + 1)$) degrees of freedom (2)

- Variables are chosen in a way maximazing t-statistic ensuring its significance
- t-test has a straightforward meaning only if independent variables are manipulated treatments in a controlled experiment
- Caution is required if applied to observational variables (influenced by the number of variables, their correlations and sample size)

Intervals for prediction (1)

Two sources of variation of prediction estimates:

- from the estimation of regression parameters (SE)
- from the variation associated with a new observation x^*(variation about the regression equation for the mean)

Error of the mean

For simple regression (single predictor) *error of the mean* of a large number of new observations

$$SE_{\text{mean}}(Y|x^*) = \sigma\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

Intervals for prediction (1)

Standard error for a new observation

For simple regression (single predictor)

$$SE_{\mathrm{obs}}(Y|x^*) = \sigma\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

95% Confidence interval

$$95\% CI = Y \pm t_{0.05}(SE)$$

Coefficient of determination, R^2

- R^2 describes the amount of variance in the outcome variable expained by the predictor variables.
- R^2 is a squared *correlation coefficient* between the predicted and observed Y-values

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R^2

Because R^2 alwasy increases with the number of variables, *adjusted* R^2 is considered

$$R^2_{\text{adjusted}} = 1 - \frac{MSE}{MST}$$

Akaike Information Criterion (AIC)

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

- One should prefer a model with smaller AIC

```
model1 <- lm(milk120~cc, data = daisy2)
AIC(model1)
```

```
## [1] 112227.5
```

```
model2 <- lm(milk120~parity, data = daisy2)
AIC(model2)
```

```
## [1] 104745.6
```

Another simple regression

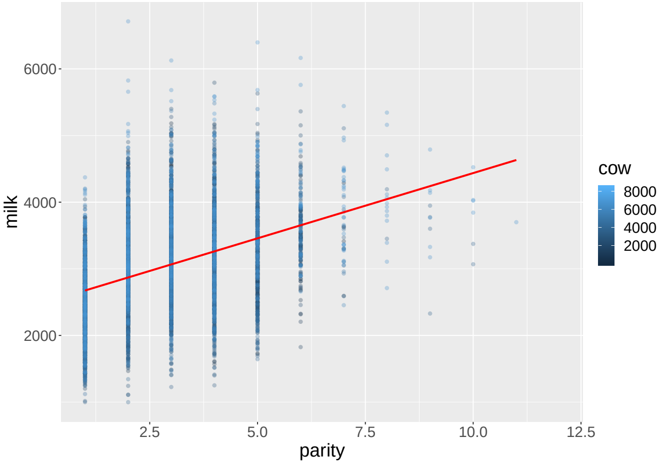

Figure 3: Impact of parity on Milk volume. 9383 lactation records in 42 year-round calving herds. http://projects.upei.ca/ver/data-and-samples/.

Another simple regression: model output

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(milk120~parity, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = milk120 ~ parity, data = daisy2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2010.3  -454.6   -36.3   413.5  3842.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2478.286     17.176  144.28   <2e-16 ***
## parity       195.807      5.385   36.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.8 on 6608 degrees of freedom
##   (2773 observations deleted due to missingness)
## Multiple R-squared:  0.1667, Adjusted R-squared:  0.1666
## F-statistic:  1322 on 1 and 6608 DF,  p-value: < 2.2e-16
```

Another simple regression: ANOVA table

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: milk120
##             Df     Sum Sq   Mean Sq F value    Pr(>F)
## parity       1  589631568 589631568  1322.1 < 2.2e-16 ***
## Residuals 6608 2947090043    445988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Another simple regression: prediction interval (1)

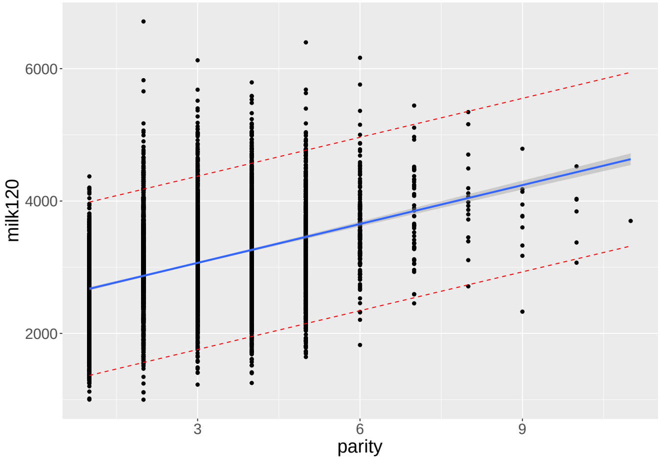

Figure 4: Prediction interval for the regression milk volume on parity.

Another simple regression: prediction interval (2)

```
temp_var <- predict(model, interval="prediction")
new_df0 <- na.omit(data.frame("milk120" = daisy2$milk120, "parity" = daisy2$parity))
new_df <- cbind(new_df0, temp_var)
ggplot(new_df, aes(x=parity, y=milk120))+
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)
```

[Implementation idea from https://rpubs.com/Bio-Geek/71339]

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \varepsilon$$

$$Y = \beta_0 + \sum_{i}^{n} \beta_i X_i + \varepsilon$$

Multiple regression 1: model output

```r
model <- lm(milk120 ~ parity + dyst + twin + rp + vag_disch, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = milk120 ~ parity + dyst + twin + rp + vag_disch,
##     data = daisy2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2014.4  -451.0   -32.9   412.0  3902.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2482.496     17.497 141.883   <2e-16 ***
## parity       195.788      5.406  36.220   <2e-16 ***
## dyst         -80.862     43.001  -1.880   0.0601 .
## twin           9.374     57.866   0.162   0.8713
## rp           -63.599     33.118  -1.920   0.0549 .
## vag_disch    123.049     53.271   2.310   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.5 on 6604 degrees of freedom
##   (2773 observations deleted due to missingness)
## Multiple R-squared:  0.1681, Adjusted R-squared:  0.1675
## F-statistic: 266.9 on 5 and 6604 DF,  p-value: < 2.2e-16
```

Multiple regression 1: ANOVA table

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: milk120
##             Df     Sum Sq   Mean Sq   F value  Pr(>F)
## parity       1  589631568 589631568 1323.4967 < 2e-16 ***
## dyst         1    1414549   1414549    3.1751 0.07481 .
## twin         1       6718      6718    0.0151 0.90227
## rp           1    1140835   1140835    2.5607 0.10960
## vag_disch    1    2377030   2377030    5.3355 0.02093 *
## Residuals 6604 2942150910    445510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple regression 2: model output

```r
library(lubridate)
daisy2$date <- as.character(daisy2$calv_dt)
daisy2$date <- ymd(daisy2$date)
daisy2$mth <- month(daisy2$date)
daisy2$aut_calv <- with(daisy2, ifelse(mth %in% c(9:12), "fall", "other"))
daisy2$hs100 <- daisy2$herd_size / 100  # herd size scaled by dividing by 100
daisy2$hs100_ct <- daisy2$hs100 - mean(daisy2$hs100)  # centered

model1 <- lm(wpc ~ aut_calv + hs100_ct  + twin + rp + vag_disch + dyst + rp*vag_disch, data = daisy2)
summary(model1)
```

```
##
## Call:
## lm(formula = wpc ~ aut_calv + hs100_ct + twin + rp + vag_disch +
##     dyst + rp * vag_disch, data = daisy2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78.30 -39.98 -16.77  26.03 254.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.5593     0.9815  74.948  < 2e-16 ***
## aut_calvother -2.4442     1.2819  -1.907  0.05659 .
## hs100_ct       4.6369     0.5674   8.172 3.53e-16 ***
## twin          15.5310     4.9822   3.117  0.00183 **
## rp            12.1403     2.9645   4.095 4.26e-05 ***
## vag_disch      6.6528     5.7964   1.148  0.25111
## dyst           5.9930     3.3819   1.772  0.07642 .
## rp:vag_disch  -2.6061    10.5217  -0.248  0.80439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.99 on 7462 degrees of freedom
##   (1913 observations deleted due to missingness)
## Multiple R-squared:  0.01417,    Adjusted R-squared:  0.01324
## F-statistic: 15.32 on 7 and 7462 DF,  p-value: < 2.2e-16
```

Multiple regression 2: ANOVA table

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: wpc
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## aut_calv         1    14341   14341  4.7421 0.0294644 *
## hs100_ct         1   193855  193855 64.1018 1.361e-15 ***
## twin             1    38759   38759 12.8165 0.0003458 ***
## rp               1    62035   62035 20.5129 6.015e-06 ***
## vag_disch        1     5460    5460  1.8054 0.1791075
## dyst             1     9662    9662  3.1950 0.0739055 .
## rp:vag_disch     1      186     186  0.0613 0.8043854
## Residuals     7462 22566410    3024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions (1)

Previously we considered only the *main effects* of the predictor variables. This assumes the association between Y and X_i is the same at all levels (values) of X_j ($i \neq j$) and the association of X_j to Y is the same at all levels (values) of X_i.

If this assumption is violated and the effect of one predictor variable depends on the values of another predictor variable, the addition of *Interaction term* could significantly improve the model.

For continuous variables interaction could be included in the following way

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Interactions, example 1 (1)

For interval from wait period to conception (wpc)

```
model <- lm(wpc~milk120*parity, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = wpc ~ milk120 * parity, data = daisy2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73.17 -39.16 -16.11  25.41 247.09
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.3914728  6.0499361  10.809   <2e-16 ***
## milk120         0.0019932  0.0020838   0.957    0.339
## parity         -1.6645556  2.0620816  -0.807    0.420
## milk120:parity  0.0004894  0.0006480   0.755    0.450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.34 on 5753 degrees of freedom
##   (3626 observations deleted due to missingness)
## Multiple R-squared:  0.002014,   Adjusted R-squared:  0.001494
## F-statistic:  3.87 on 3 and 5753 DF,  p-value: 0.008884
```

Interactions, example 1 (2)

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: wpc
##                  Df   Sum Sq Mean Sq F value   Pr(>F)
## milk120           1    31148 31147.8 10.9496 0.000942 ***
## parity            1      260   260.0  0.0914 0.762404
## milk120:parity    1     1623  1622.6  0.5704 0.450124
## Residuals      5753 16365322  2844.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(model)
```

```
## [1] 62130.22
```

Interactions, example 1 (3)

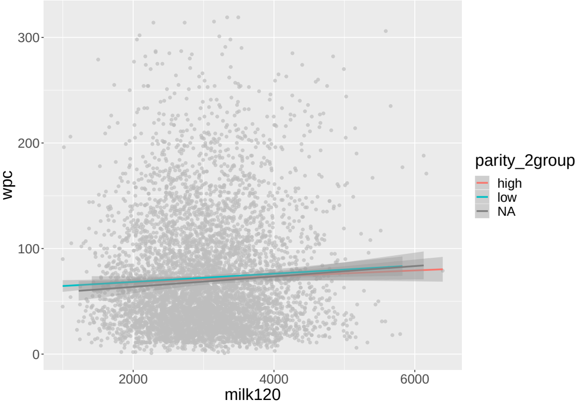

Figure 5: Impact of interaction between milk volume (milk120) and parity on the interval from waiting period to conception (wpc).

Interactions, example 1 (4)

```r
library(dplyr)
d <- data.frame("milk120" = daisy2$milk120, "wpc"= daisy2$wpc , "parity" = daisy2$parity)
d = na.omit(d)
x <- d$parity
d$parity_2group <-
  case_when(x > median(x) ~ "high",
            x < median(x) ~ "low")
count(d, parity_2group)
d %>%
  ggplot() +
  aes(x = milk120, y = wpc, group = parity_2group, color = parity_2group) +
  geom_point(color = "grey", alpha = .7) +
  geom_smooth(method = "lm")
```

[Visualization idea from: https://sebastiansauer.github.io/vis_interaction_effects/]

Interactions, example 2 (1)

```r
daisy2$milk120k <- daisy2$milk120/1000 # scaling of the variables
model <- lm(wpc~as.factor(dyst)*milk120k, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = wpc ~ as.factor(dyst) * milk120k, data = daisy2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72.84 -40.60 -16.48  26.00 254.17
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               67.7666     2.8723  23.594   <2e-16 ***
## as.factor(dyst)1          30.3670    15.4199   1.969   0.0490 *
## milk120k                   1.8437     0.9442   1.953   0.0509 .
## as.factor(dyst)1:milk120k  -8.1006     5.3236  -1.522   0.1281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55 on 7030 degrees of freedom
##   (2349 observations deleted due to missingness)
## Multiple R-squared:  0.0014, Adjusted R-squared:  0.0009739
## F-statistic: 3.285 on 3 and 7030 DF,  p-value: 0.01989
```

Interactions, example 2 (2)

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: wpc
##                          Df   Sum Sq Mean Sq F value  Pr(>F)
## as.factor(dyst)             1    13967 13966.8  4.6169 0.03169 *
## milk120k                    1     8844  8844.1  2.9235 0.08734 .
## as.factor(dyst):milk120k    1     7004  7004.3  2.3154 0.12815
## Residuals                7030 21266817  3025.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model)
```

```
## [1] 76343.14
```

Interactions, example 2 (3)

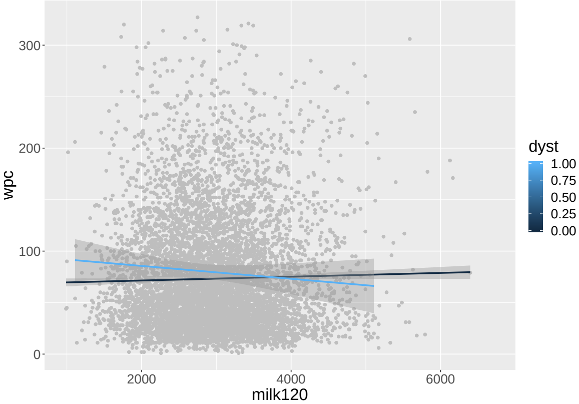

Figure 6: Impact of interaction between milk volume (milk120) and dystocia (dyst) on the interval from waiting period to conception (wpc).

Interactions, example 2 (4)

```r
dyst_0 <- filter(daisy2, dyst == "0")
dyst_1 <- filter(daisy2, dyst == "1")
ggplot(daisy2) +
  aes(x = milk120, y = wpc, color = dyst) +
  geom_point(color = "grey") +
  geom_smooth(method = "lm", data = dyst_0) +
  theme(text = element_text(size = 20)) +
  geom_smooth(method = "lm", data = dyst_1)
```

Model diagnostics

Independence

Homoscedasticity

Normal distribution of residuals

Linearity

Independence

- usually is clear from the nature of the data
- lack of independence e.g. in case of multiple observations of a single animal or the same herd
- lack of independence if serial correlations are present like measurements during particular season

Homoscedasticity (1)

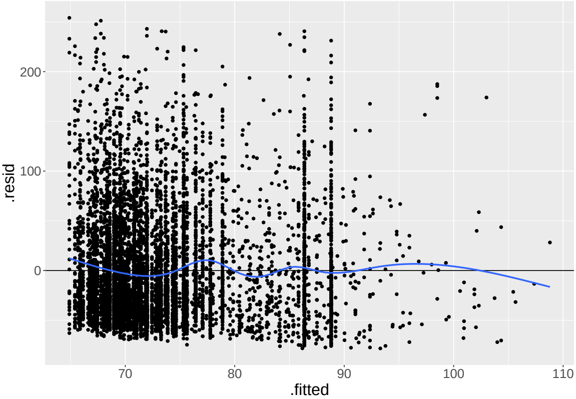

Figure 7: Evaluation of the homoscedasticity assumption for interval from waiting period to conception (wpc) regression.

Homoscedasticity (2)

```
library(car)
ncvTest(model1)
library(lmtest)
bptest(model1)
ggplot(model1, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0) +
  geom_point() +
  geom_smooth(se = FALSE)

library(car)
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 18.21011, Df = 1, p = 1.9783e-05
```

```
library(lmtest)
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 19.601, df = 7, p-value = 0.006499
```

Normal distribution of residuals (1)

Q-Q plot

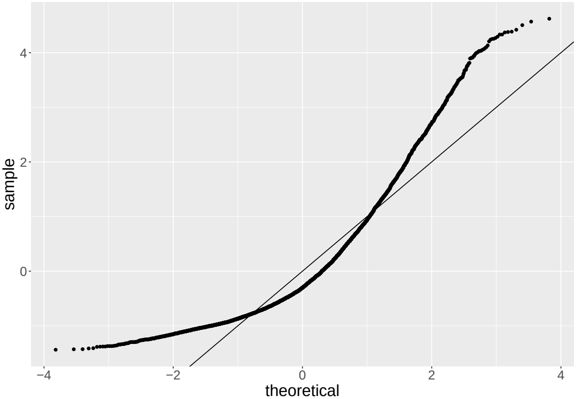

Figure 8: Evaluation of the normality of residuals assumption for interval from waiting period to conception (wpc) regression.

Normal distribution of residuals (2)

```r
ggplot(model1, aes(sample = .stdresid)) +
  stat_qq() +
  theme(text = element_text(size = 20)) +
  geom_abline()
```

Shapiro-Wilk's test

$p < 0.05$ indicates *non-normality*

```r
shapiro.test(resid(model1)[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model1)[0:5000]
## W = 0.87175, p-value < 2.2e-16
```

Leverage

Given a residual $r_i = Y_i - \hat{Y}_i$ its variance is given as

$$\mathrm{var}(r_i) = \sigma^2(1 - h_i),$$

with h_i called *leverage* of observation i. It indicates the potential of this observation to have a major impact on this model.

For a *simple regression*:

$$h_i = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{SSX}.$$

Leverage depends only on the predictor. Also for leverage holds

$$\frac{1}{n} < h_i < 1.$$

Residuals vs. leverage

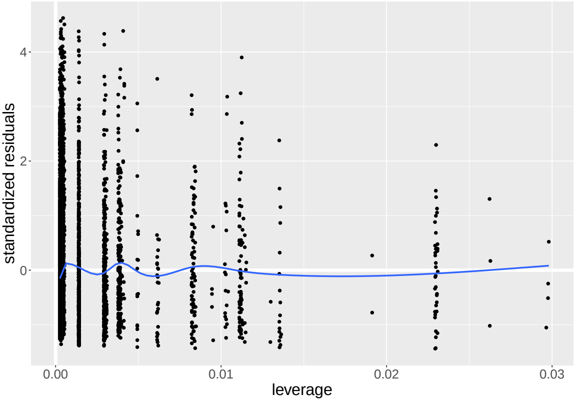

Figure 9: Residuals vs. leverage for interval from waiting period to conception (wpc) regression.

Residuals vs. leverage (2)

```r
ggplot(model1, aes(.hat, .stdresid)) +
  geom_vline(size = 2, colour = "white", xintercept = 0) +
  geom_hline(size = 2, colour = "white", yintercept = 0) +
  geom_point() +
  geom_smooth(se = FALSE)
```

References

- Dohoo I, Martin W, Stryhn H, Veterinary Epidemiologic Research, 2d edition (2009)
- Kabacoff R, R in action, 2d edition (2015)
- Crawley M J, The R book (2007)