Dunn book reading group Chapter 02

Ed Harris

20/02/2020

Dunn Ch02 notes

2.1 Linear regression (special case of GLM)

- Notation and assumptions
- Least squares
- Multiple regression
- Examples in R
- Variations and model selection

2.2 Definitions

- random component y
- systematic component: p explanatory vars $x_1, x_2, \ldots, x_p$
- Assumption of constant variance (across) **or**
- known *prior weights* exist

$u_i = \beta_0 + \beta_1 x_{1i}, \beta_2 x_{2i}, \ldots, \beta_p x_{pi}$ (2.1)
(aka multiple linear regression)

$E[y_i] = u_i$

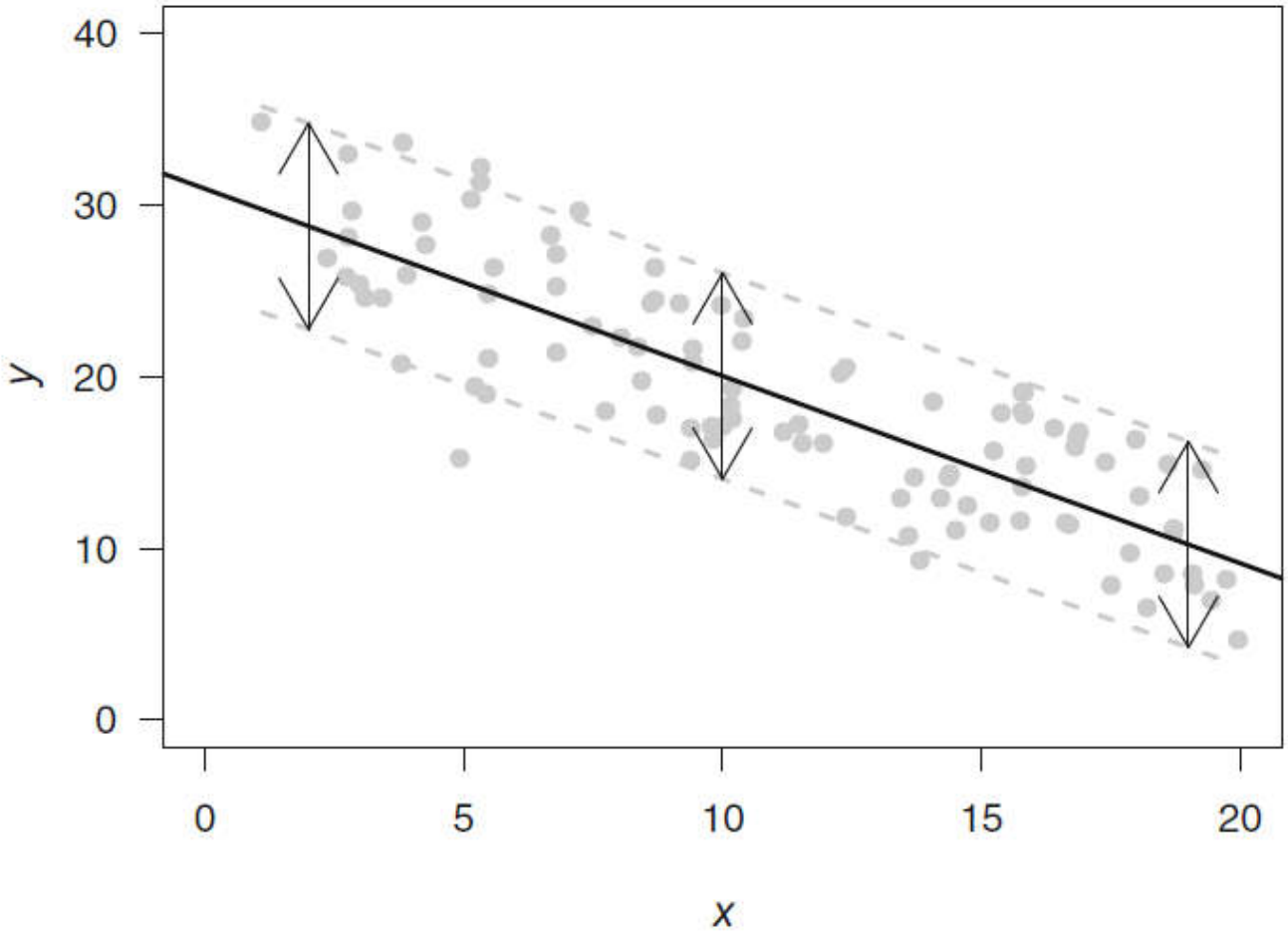
$var[y_i] = \sigma^2 / w_i$ specify prior weights

β_0 the *intercept* (y value when all explanatory vars are zero)

$\mu = \beta_0 + \beta_1 x_1$ *simple, plain old* linear regression (i.e., all prior weights == 1)

Assumptions
* **SUITABILITY** same regression model valid for all obs.
* **LINEARITY** model is linear…

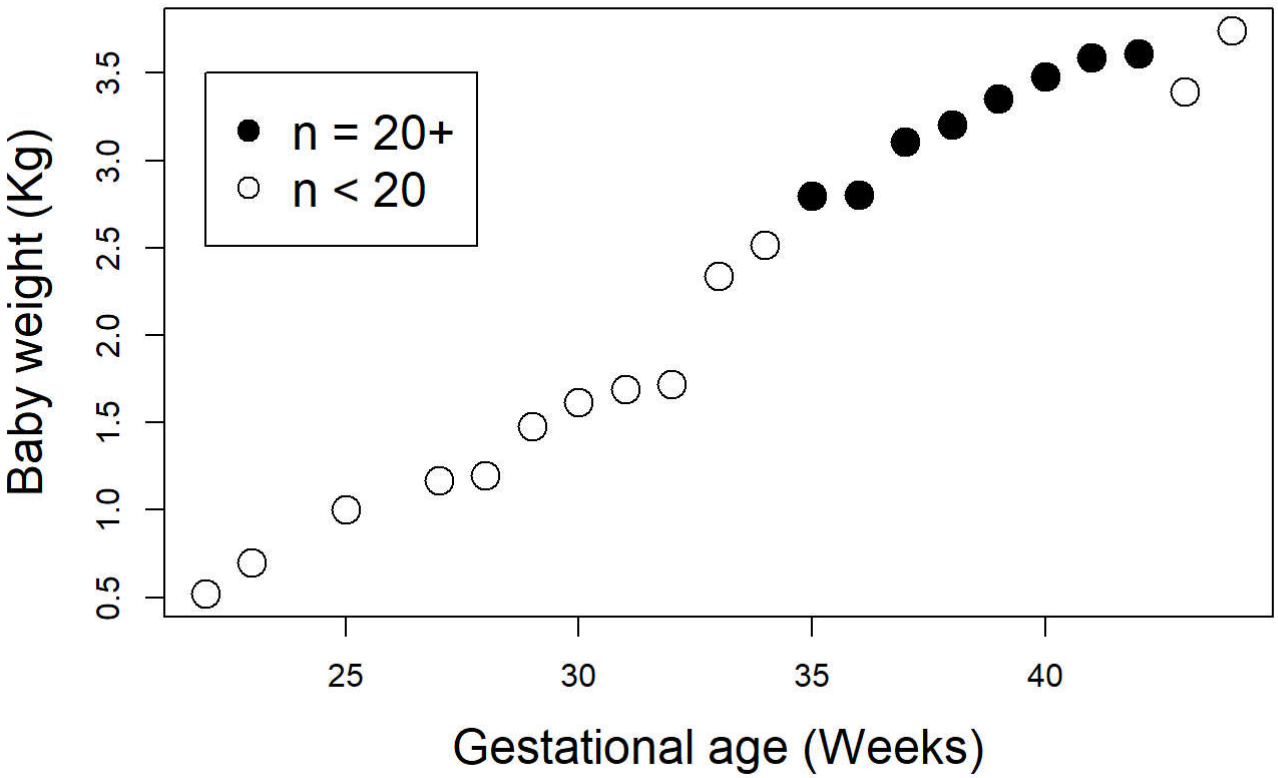
* **INDEPENDENCE** of observations
* **CONSTANT VARIANCE**



Example 2.1

library(GLMsData); data(gestation); str(gestation)

Weight = human birthweight (Kg); Age = gestational age (weeks)



NB, imbalance of observations

```
gestation
```

```
##    Age Births Weight    SD
## 1   22      1  0.520    NA
## 2   23      1  0.700    NA
## 3   25      1  1.000    NA
## 4   27      1  1.170    NA
## 5   28      6  1.198 0.121
## 6   29      1  1.480    NA
## 7   30      3  1.617 0.589
## 8   31      6  1.693 0.319
## 9   32      7  1.720 0.438
## 10  33      7  2.340 0.313
## 11  34      7  2.516 0.572
## 12  35     29  2.796 0.448
## 13  36     43  2.804 0.444
## 14  37    114  3.108 0.344
## 15  38    222  3.204 0.444
## 16  39    353  3.353 0.427
## 17  40    401  3.478 0.408
## 18  41    247  3.587 0.440
## 19  42     53  3.612 0.371
## 20  43      9  3.390 0.408
## 21  44      1  3.740    NA
```

```
sum(gestation$Births)
```

```
## [1] 1513
```

##2.2 Simple regression and Least-Squares

For our model, birth **Weight ~ Age**
$\mu_i = y_i - \beta_0 - \beta_1 x_i$

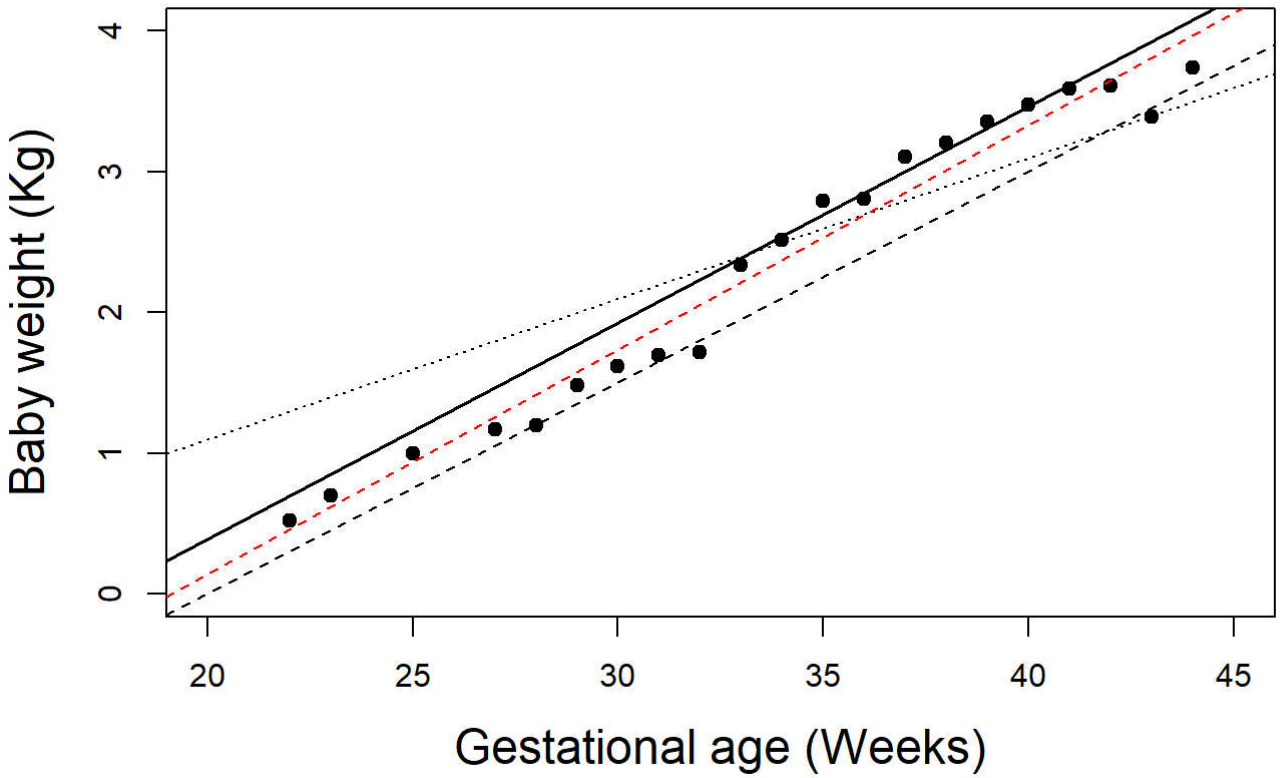
residual error in our model $e_i = y_i - \mu_i = y_i - \beta_0 - \beta_1 x_i$ (2.3)

Sum of Squares equation in "general form" explicitly including prior weights
$S(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (y_i - u_i)^2$ (SSE)

Example 2.2 - look at the effect of different slopw and intercept estimates

```
y <- gestation$Weight
x <- gestation$Age
wts <- gestation$Births
# Try these values for beta0 and beta1
beta0.A <- -0.9 #intercept
beta1.A <- 0.1 #slope
mu.A <- beta0.A + beta1.A * x
SA <- sum( wts*(y - mu.A)^2 ) #same as (SSE)
SA # goals is to minimise this!
```

```
## [1] 186.1106
```




```
## Analysis of Variance Table
##
## Response: Weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 22.1235  22.123   553.7 1.627e-15 ***
## Residuals 19  0.7592   0.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## (Intercept)         Age
##   -3.049879    0.159483
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Age        1 157.68 157.675  262.34 1.416e-12 ***
## Residuals 19  11.42   0.601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

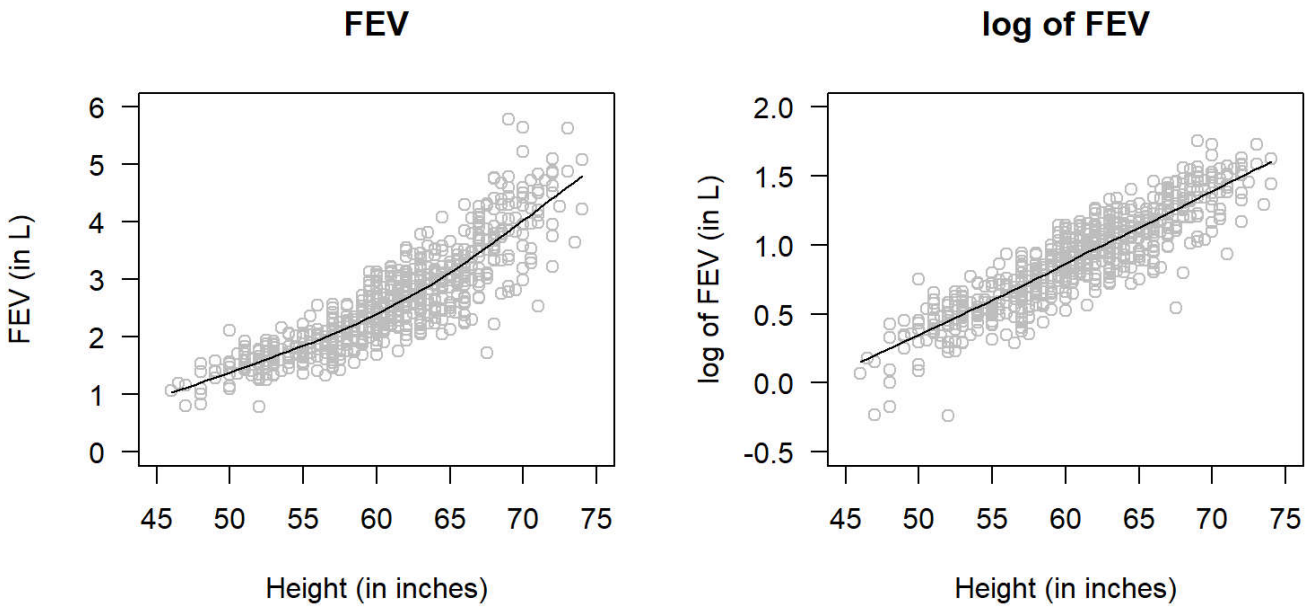
```
##
## Call:
## lm(formula = Weight ~ Age, data = gestation, weights = Births)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62979 -0.60893 -0.30063 -0.08845  1.03880
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.678389   0.371172  -7.216 7.49e-07 ***
## Age          0.153759   0.009493  16.197 1.42e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7753 on 19 degrees of freedom
## Multiple R-squared:  0.9325, Adjusted R-squared:  0.9289
## F-statistic: 262.3 on 1 and 19 DF,  p-value: 1.416e-12
```

```
## (Intercept)         Age
##  -2.6783891   0.1537594
```

NB derivation of SSE for B_0 and B_1:
*

2.3.3 Variance estimation **NB the estimate of variance s^2 is given as Mean Sq. error for residuals using anova()**
NB the estimate of standard deviation s is given as residual standard error for residuals using summary()

##2.4 Multiple Regression *Estimation of coefficients analogous to simple linear reg.
Lung data example



Discuss appropriateness of logged y var to regression assumptions

##2.5 Matrix Notation ""

**

##2.6 Linear regression in R

```
gest.wtd <- lm( Weight ~ Age, data=gestation,
weights=Births) # The prior weights
summary(gest.wtd)
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = gestation, weights = Births)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62979 -0.60893 -0.30063 -0.08845  1.03880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.678389   0.371172  -7.216 7.49e-07 ***
## Age          0.153759   0.009493  16.197 1.42e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7753 on 19 degrees of freedom
## Multiple R-squared:  0.9325, Adjusted R-squared:  0.9289
## F-statistic: 262.3 on 1 and 19 DF,  p-value: 1.416e-12
```

"~" = "as a function of"

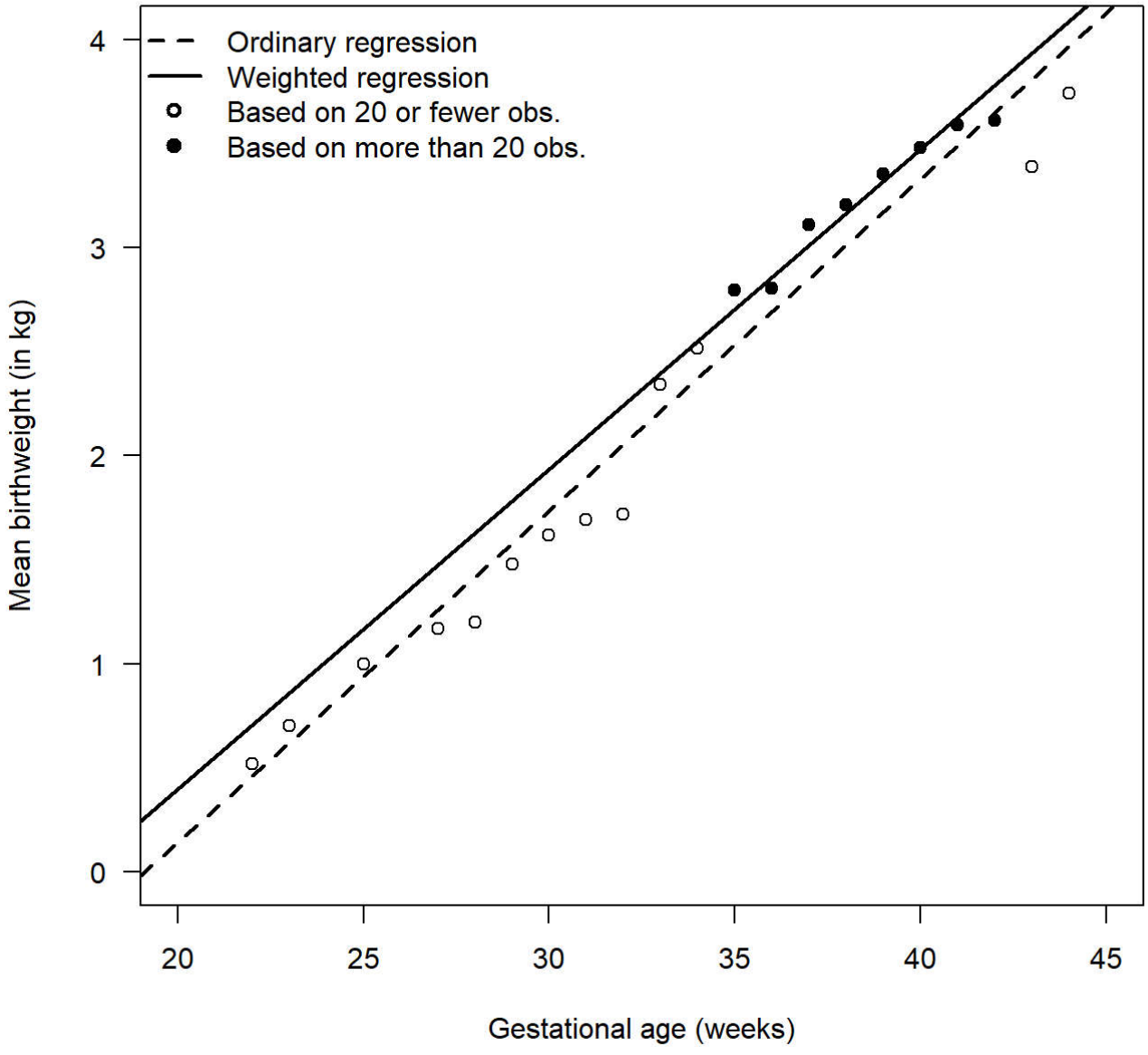
note weights

```
gest.ord <- lm( Weight ~ Age, data=gestation)
coef(gest.ord)
```

```
## (Intercept)         Age
##   -3.049879    0.159483
```

Better fit with weights?

```
plot( Weight ~ Age, data=gestation, type="n",
las=1, xlim=c(20, 45), ylim=c(0, 4),
xlab="Gestational age (weeks)", ylab="Mean birthweight (in kg)" )
points( Weight[Births< 20] ~ Age[Births< 20], pch=1, data=gestation )
points( Weight[Births>=20] ~ Age[Births>=20], pch=19, data=gestation )
abline( coef(gest.ord), lty=2, lwd=2)
abline( coef(gest.wtd), lty=1, lwd=2)
legend("topleft", lwd=c(2, 2), bty="n",
lty=c(2, 1, NA, NA), pch=c(NA, NA, 1, 19), # NA shows nothing
legend=c("Ordinary regression", "Weighted regression",
"Based on 20 or fewer obs.","Based on more than 20 obs."))
```



##Smoking example 2.15 Consider fitting the Model (2.14) to the lung capacity data (lungcap), using age, height, gender and smoking status as explanatory variables, and log(fev) as the response.

```
# Recall, Smoke has been declared previously as a factor
lm( log(FEV) ~ Age + Ht + Gender + factor(Smoke), data=lungcap )
```

```
##
## Call:
## lm(formula = log(FEV) ~ Age + Ht + Gender + factor(Smoke), data = lungcap)
##
## Coefficients:
##    (Intercept)             Age              Ht         GenderM
##       -1.94400         0.02339         0.04280         0.02932
## factor(Smoke)1
##       -0.04607
```

```
mylm <- lm( log(FEV) ~ Age + Ht + Gender + factor(Smoke), data=lungcap )
summary(mylm)
```

```
##
## Call:
## lm(formula = log(FEV) ~ Age + Ht + Gender + factor(Smoke), data = lungcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63278 -0.08657  0.01146  0.09540  0.40701
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.943998   0.078639 -24.721  < 2e-16 ***
## Age             0.023387   0.003348   6.984  7.1e-12 ***
## Ht              0.042796   0.001679  25.489  < 2e-16 ***
## GenderM         0.029319   0.011719   2.502   0.0126 *
## factor(Smoke)1 -0.046068   0.020910  -2.203   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 649 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.8095
## F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

##2.7 Interpreting regression coefficients

Challenge yourself with your own data: take responsibility for the analysis and the NUMERICAL OUTPUTS to make sense. I consider this to be the principle distinction between someone who is merely good at performing statistics, and someone who is a good scientist…

- Implications of logging our Y var in previous example
- coefficient values and predictions
- explicit interpretation of Smoke variable

```
#Coefficient est. of Smoke on log FEV
# -0.04607
#Now, relate back to actual FEV
exp(-0.04607) #yawn not that large
```

```
## [1] 0.9549751
```

```
#help(lungcap)
# FEV
# the forced expiratory volume in LITRES, a measure of lung capacity; a numeric vector

#omfg this is in Litres...
```

##2.8 Inference of linear models

Author pretends to now that we have only bee interested in the parameter estimates, without any actual interest in making a statistical inference.

Here is where We need to start being explicit about the distribution of residuals.

$y_i \sim N(\mu_i, \sigma^2/w_i)$

In practice, we formally relate this to the estimation of the model parameters, also assuming an explicit distribution.

$\hat{\beta}_j \sim N(\beta_j, var[\hat{\beta}_j])$

We use this estimate for hypothesis testing - this is where we get our p-value from

Is the estimate of a particular coefficient different to zero…?

```
mylm1 <- lm( log(FEV) ~ Age + Ht + Gender + factor(Smoke), data=lungcap )
confint(mylm1)
```

```
##                       2.5 %       97.5 %
## (Intercept)    -2.098414941 -1.789581413
## Age             0.016812109  0.029962319
## Ht              0.039498923  0.046092655
## GenderM         0.006308481  0.052330236
## factor(Smoke)1 -0.087127344 -0.005007728
```