STAT 425

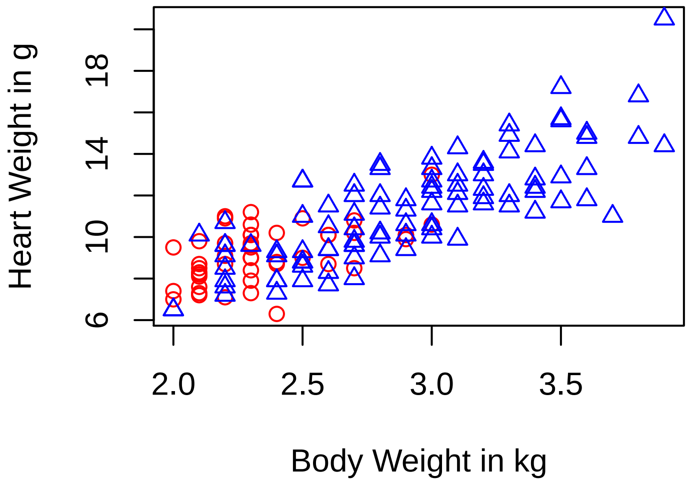
Simple Linear Regression. Part 1

An example

The cats data set from the MASS library

```
library(MASS)
help(cats)
summary(cats)
```

```
##  Sex         Bwt             Hwt       
##  F:47   Min.   :2.000   Min.   : 6.30  
##  M:97   1st Qu.:2.300   1st Qu.: 8.95  
##         Median :2.700   Median :10.10  
##         Mean   :2.724   Mean   :10.63  
##         3rd Qu.:3.025   3rd Qu.:12.12  
##         Max.   :3.900   Max.   :20.50  
```



Heart Weight in g
6
10
14
18
2.0
2.5
3.0
3.5
Body Weight in kg

- The goal is to describe the relationship between Hwt (heart weight) and Bwt (body weight). As a starting point, we assume the relationship is linear.
- Data of the form: $(y_i, x_i), i = 1, \ldots, n$ where $y_i, x_i \in \mathbb{R}$.
- Apparently the data won't be able to fit on a straight line.
  Assume $y_i = \beta_0 + \beta_1 x_i + e_i$.
  (β_0, β_1) : unknown regression coefficients
  e_i's: random errors often assumed to have zero mean and variance σ^2

Simple Linear Regression Overview (I)

- How to use Least Squares (LS) to estimate (β_0, β_1)? We can obtain an explicit expression $(\hat{\beta}_0, \hat{\beta_1})$. There is a nice connection between the LS estimate of the slope, β_1, and sample correlation/variance of X and Y, which will help you to remember the expression.
- Throughout the class we'll use some jargon: fitted value, residual, Residual Sum of Squares (RSS), R-squared (used to assess the overall model fit).
- How would the LS fitting/inference be affected if the data, X and/or Y, are shifted and/or scaled (i.e., linear transformed)?
- SLR without the intercept: fit a regression line passing through the origin.
- How to use **R** to carry out all the analysis and produce relevant graphs.

Parameter estimation by Least squares

We would like to choose a line which is close to the data points. We measure the closeness by squared errors [1].

Least Squares Estimation: find (β_0, β_1) that minimize the residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To find the solution, we can take the derivatives w.r.t. β_0 and β_1 and equate to zero.

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

[1] Why squared error? Why not absolute error?

Re-arrange the equations,

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i, & (1) \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i. & (2) \end{aligned}$$

From (1), we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Plug it back to (2),

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

$$\beta_1 \left(\sum x_i^2 - \sum x_i \bar{x} \right) = \sum x_i y_i - \sum x_i \bar{y}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}.$$

Some equalities (basically centering one side is the same as centering both sides for cross-products):

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i.$$

So the LS estimates of (β_0, β_1) can be expressed as

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\mathsf{Sxy}}{\mathsf{Sxx}} = r_{\mathsf{XY}} \left(\frac{\mathsf{Syy}}{\mathsf{Sxx}} \right)^{1/2}, \end{aligned}$$

where

$$\begin{aligned} \mathsf{Sxy} &= \sum (x_i - \bar{x})(y_i - \bar{y}), \\ \mathsf{Sxx} &= \sum (x_i - \bar{x})^2, \quad \mathsf{Syy} = \sum (y_i - \bar{y})^2, \\ r_{\mathsf{XY}} &= \frac{\mathsf{Sxy}}{\sqrt{(\mathsf{Sxx})(\mathsf{Syy})}} \quad \text{(the sample correlation).} \end{aligned}$$

- Recall that SLR assumes the dependence between X and Y is linear.
- It is not surprising that the LS estimates are related to the sample correlation between X and Y.
- Correlation is exactly the measure used to quantify the linear dependence between two variables [2].

[2] We can build an example in where variables X and Y have a non-linear relationship and their correlation is zero

Suppose that the mean and variance of X and Y, and the correlation between X and Y r_{xy} are known. Given a value of x, what is the best guess of y?
It seems reasonable to use the *unit-free location/scale invariant* value of x multiplied by r_{xy} to get a *unit-free location/scale invariant* value of y as follows:

$$\frac{y-\mu_y}{\sigma_y} \approx r_{xy}\frac{x-\mu_x}{\sigma_x}$$

3

By using the sample estimates of the means, variances and correlation coefficient we get the corresponding sample expression:

$$\frac{y-\bar{y}}{\sqrt{S_{yy}}} \approx r_{xy}\frac{x-\bar{x}}{\sqrt{S_{xx}}} \rightarrow y-\bar{y} \approx r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}(x-\bar{x})$$

[3] If you want to get x as a function of y, you need to multiply by r_{xy} on the y side of the equation.

A final equation of y as a function of x is given by:

$$y \approx \left(\bar{y} - r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\bar{x}\right) + \left(r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\right)x$$

Some jargon:

- Fitted value or prediction at x_i: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual at x_i: $r_i = y_i - \hat{y}_i$. If you plug-in the equations from page 7 for $\hat{\beta}_i$, you can show that:

$$\sum_i r_i = 0, \quad \sum_i x_i r_i = 0$$

[4]

- Residual Sum of Squares (RSS): $\sum_i r_i^2$
- The error variance is estimated as:

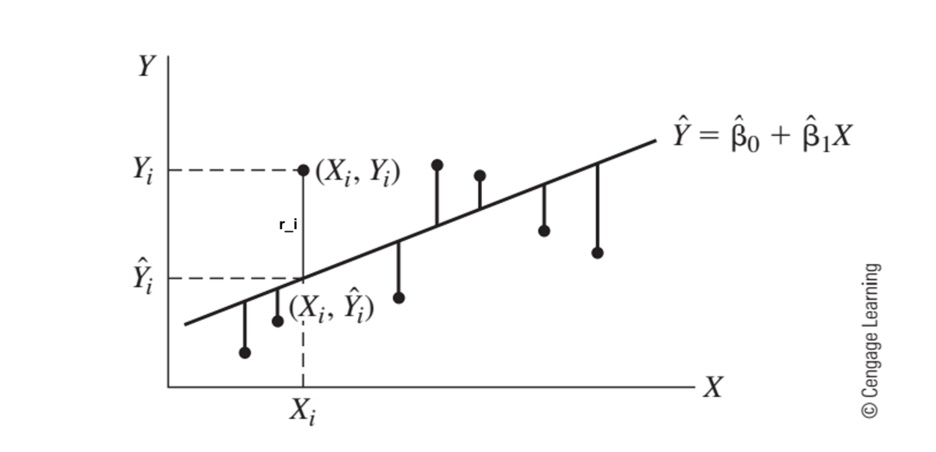
$$\hat{\sigma}^2 = \frac{1}{n-2} RSS = \frac{1}{n-2} \sum_i r_i^2$$

- residual degrees of freedom (df): $n - 2$. Normally $df =$ sample size $-$ number of parameters

---

[4] $\sum_i r_i = 0$ implies that the mean of $\hat{y}_i = \bar{y}$

LS fitted linear regression



Goodness of fit: R-square

The total variation of y (Total Sum of Squares (TSS)) can be decomposed into the sum of the total variation of the fitted values $\hat{y}$ (FSS) and the Residual Sum of Squares (RSS):

$$\begin{aligned} TSS = \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\ &= RSS + FSS \end{aligned} \tag{3}$$

[5]

Note: The average of the $\hat{y}_i$ ($\bar{\hat{y}}$) is the same as the average of the y_i. This is true because the intercept is included in the model.

[5] The cross product $\sum_i r_i(\hat{y}_i - \bar{y}) = \hat{\beta}_0 \sum_i r_i + \hat{\beta}_1 \sum_i r_i x_i - \bar{y} \sum_i r_i = 0$

A common measure on how well the model fits the data is the so-called coefficient of determination or simply R-square:
For a given data set where TSS is fixed, the smaller the RSS, the larger the R-squared.

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{FSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

We can also show that $R^2 = r_{XY}^2$.
Note also that $R^2 = \frac{Var(\hat{y})}{Var(y)}$. This ratio measures how much variation in the original data y_i's is explained or reduced by the LS fitting. If Y and X are strongly linear dependent, a linear function of X can help to reduce the uncertainty (i.e., variation) of Y.

Fitting a Linear Model in R

```
out = lm(Hwt~Bwt, data = cats)
summary(out)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## Bwt           4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466,Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Model output is stored in the out object. out is a list in **R**.

Extract Information and make some calculations

```
names(out)

##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"

out$coef

## (Intercept)         Bwt
##  -0.3566624   4.0340627

cor(Hwt,Bwt)^2

## [1] 0.6466209

var(out$fitted.values)/var(Hwt)

## [1] 0.6466209

1 - sum(out$res^2)/sum((Hwt-mean(Hwt))^2)

## [1] 0.6466209

summary(out)$r.sq

## [1] 0.6466209
```

Different ways to calculate the R-square

How affine transformations on the data affect the Regression?

Affine transformation: $\tilde{y} = ay + b$ where a and b ar known constants.
Changes of scale in X or Y are also affine transformations.
Suppose we have a SLR model of Y on X.

- If we rescale the data $\tilde{y} = ay + b$, and then regress $\tilde{y}$ on x. How would the LS estimates and R^2 be affected?
- If we re-scale the covariates x $\tilde{x} = ax + b$, and then regress y on $\tilde{x}$. How would the LS estimates and R^2 be affected?
- If we regress X on Y instead, will the LS line be the same? How about R^2?

In **R**:

```
out1<-lm(Hwt ~ I(Bwt*1000), data=cats)
out2<-lm(I(Hwt+1) ~ Bwt, data=cats)
out3<-lm(Bwt ~ Hwt,data=cats)
cbind(out$coef, out1$coef, out2$coef, out3$coef)

##                   [,1]         [,2]      [,3]      [,4]
## (Intercept) -0.3566624 -0.356662433 0.6433376 1.0196367
## Bwt          4.0340627  0.004034063 4.0340627 0.1602902

cbind(summary(out)$r.square,summary(out1)$r.square,
summary(out2)$r.square,summary(out3)$r.square)

##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.6466209 0.6466209 0.6466209 0.6466209
```

Regression through the Origin

Sometimes we want to fit a line with no intercept (regression through the origin): $y_i \approx \beta_1 x_i$. For example, x_i denotes the intensity level of various exercises and y_i denotes the additional calories you burn with those exercises.

We can estimate β_1 using the LS principle

$$\min_{\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \Longrightarrow \hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

The ordinary definition of R-square is no longer meaningful; you could have RSS bigger than TSS, and therefore have a negative R-square, if you use formula $R^2 = 1 - \text{RSS}/\text{TSS}$.

The ordinary R-square measures the effect of X after removing the effect of the intercept by centering both y_i's and $\hat{y}_i$'s. For regression models with no intercept, we shouldn't do the centering when computing R-square.

Let's look at the following decomposition (slightly different from (3))

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i^2.$$

Then define R-square for regression with no intercept as

$$\tilde{R}^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\text{RSS}}{\sum_i y_i^2}.$$

Some remarks

- We will use the hat symbol for the estimators/estimates of the true model parameters: $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$ are the LS estimators of the population parameters: $(\beta_0, \beta_1, \sigma^2)$
- These estimators are a function of the sample data. If we have a different sample, we will have a different set of estimators. These implies the estimators are random variables.
- As a next step we will check the properties of these estimators and we will determine their probability distributions.