Linear Regression

Felipe José Bravo Márquez

June 4, 2021

Introduction

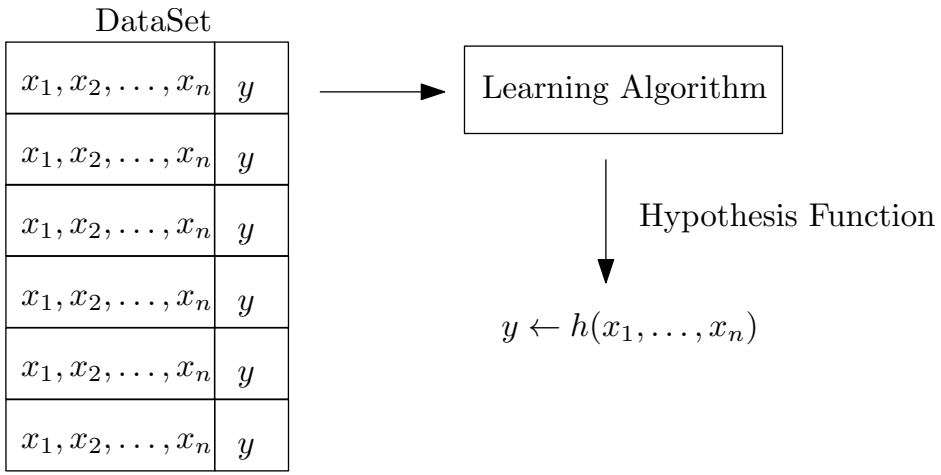
- A regression model is used to model the relationship of a numerical dependent variable **y** with n independent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ [Wasserman, 2013].
- The dependent variable **y** is also called **target**, **outcome**, or **response** variable.
- The independent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are also called **covariates**, **attributes**, **features**, por **predictor variables**.
- Roughly speaking we want to know the expected value of **y** from the values of **x**:

$$\mathbb{E}(y|x_1, x_2, \ldots, x_n)$$

- We use these models when we believe that the response variable **y** can be modeled by other independent variables.
- To perform this type of analysis we need a dataset consisting of m observations that include both the response variable and each of the attributes.
- We refer to the process of **fitting** a regression function as the process in which from the data we infer a hypothesis function h that allows us to **predict** unknown **y** values using the values of the attributes.

Introduction (2)

- This process of fitting a function from data is referred to in the areas of data mining and machine learning as **training**.
- In those disciplines, functions are said to **learn** from data.
- Since we need observations where the value of **y** is known to learn the function, such techniques are referred to as **supervised learning** techniques.
- When **y** is a categorical variable we have a **classification** problem.



Simple Linear Regression

- In simple linear regression, we have a single independent variable x to model the dependent variable **y**.
- The following linear relationship between the variables is assumed:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall i$$

- The parameter β_0 represents the intercept of the line (the value of y when x is zero).
- The parameter β_1 is the slope and represents the change of **y** when we vary the value of **x**. The greater the magnitude of this parameter the greater the linear relationship between the variables.
- The ϵ_i values correspond to the errors or **residuals** associated with the model.
- We have to find a linear function or straight line h_β that allows us to find an estimate of y, $\hat{y}$ for any value of x with the minimum expected error.

$$h(x) = \beta_0 + \beta_1 x$$

Least Squares

- The ordinary least squares method is used to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the sum of squared errors (SSE) of the observed data.
- Suppose we have *m* observations of **y** and **x**, we compute the sum of squared errors (SSE) as follows:

$$SSE = \sum_{i=1}^{m}(y_i - h(x_i))^2 = \sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_i)^2 \quad (1)$$

- To find the parameters that minimize the error we calculate the partial derivatives of SSE with respect to β_0 and β_1. Then we equal the derivatives to zero and solve the equation to find the parameter values.

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

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x)x_i = 0 \quad (3)$$

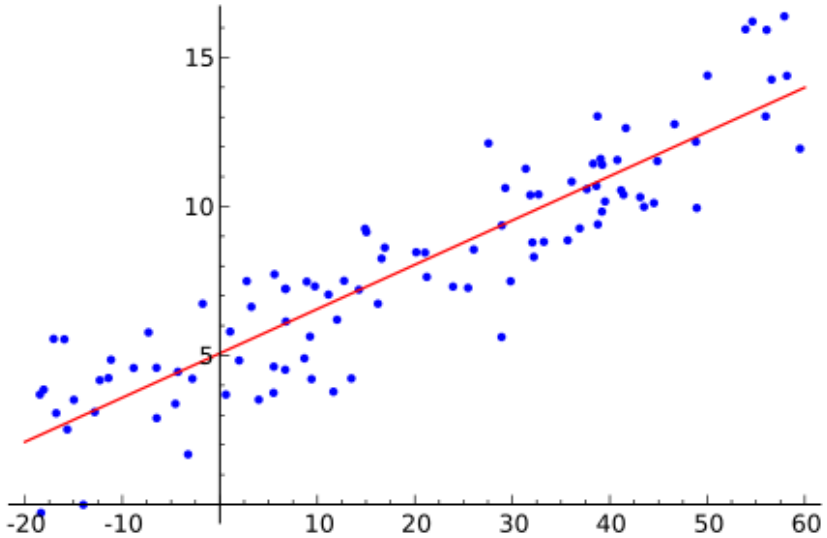
Least Squares (2)

- From the above system of equations the normal solutions are obtained:

$$\hat{\beta}_1 = \frac{\sum_i^m (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^m (x_i - \overline{x})^2} \tag{4}$$

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \tag{5}$$

- The fitted model represents the line of least squared error.



Coefficient of Determination R^2

- Once we have fitted our linear model we must evaluate the quality of the model.
- A very common metric is the coefficient of determination R^2.
- It is calculated from errors that are different than the SSE squared errors.
- The total sum of squares (SST) is defined as the predictive error when we use the mean $\overline{y}$ to predict the response variable y (it is very similar to the variance of the variable):

$$\text{SST} = \sum_{i}^{m} (y_i - \overline{y})^2$$

- The regression sum of squares (SSM), on the other hand, represents the amount of error in the regression model:

$$\text{SSM} = \sum_{i}^{m} (\hat{y}_i - \overline{y})^2$$

- SSM indicates the variability of the values predicted by the model with respect to the mean.

Coefficient of Determination R^2 (2)

- It can be proved that all the above errors are related as follows: $SST = SSE + SSM$
- The coefficient of determination for a linear model R^2 is defined as:

$$R^2 = \frac{\text{SSM}}{\text{SST}} = \frac{\sum_i^m (\hat{y}_i - \overline{y})^2}{\sum_i^m (y_i - \overline{y})^2} \tag{6}$$

- An alternative but equivalent definition:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_i^m (y_i - \hat{y}_i)^2}{\sum_i^m (y_i - \overline{y})^2} \tag{7}$$

- The coefficient takes values between 0 to 1 and the closer its value is to 1 the higher the quality of the model.
- The value of R^2 is equivalent to the linear correlation (Pearsons) between y and $\hat{y}$ squared.

$$R^2 = \text{cor}(y, \hat{y})^2$$

Assumptions of the Linear Model

Whenever we fit a linear model we are implicitly making certain assumptions about the data.

Assumptions

1. Linearity: the response variable is linearly related to the attributes.
2. Normality: errors have zero mean normal distribution: $\epsilon_i \sim N(0, \sigma^2)$.
3. Homoscedasticity: errors have constant variance (same value σ^2).
4. Independence: errors are independent of each other.

Probabilistic Interpretation

- Considering the above assumptions, we are saying that the probability density (PDF) of the errors ϵ is defined by a normal of zero mean and constant variance:

$$\text{PDF}(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

- Moreover all ϵ_i are IID (independent and identically distributed).
- This implies that:

$$\text{PDF}(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right)$$

- Which implies that the distribution of **y** given the values of **x** and parameterized by β follows a normal distribution.
- Let's estimate the β using maximum likelihood estimation.

Probabilistic Interpretation

- The likelihood function of β can be written as:

$$\mathcal{L}(\beta) = \prod_{i=i}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right)$$

- and the the log likelihood $l_n(\beta)$:

$$\begin{aligned} l_n(\beta) &= \log \mathcal{L}(\beta) && (8)\\ &= \log \prod_{i=i}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right) && (9)\\ &= \sum_{i=1}^{m} \log \prod_{i=i}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right) && (10)\\ &= m \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{\sigma^2} \times \frac{1}{2} \sum_{i=1}^{m} (y_i - h_\beta(x_i))^2 && (11) \end{aligned}$$

Probabilistic Interpretation

- Hence, maximizing $l_n(\beta)$ gives the same answer as minimizing

$$\frac{1}{2}\sum_{i=1}^{m}(y_i - h_\beta(x_i)^2$$

- which is equivalent as minimizing SSE.
- Then, if one estimates the parameters of β using maximum likelihood estimation one arrives at the same results as doing least squares estimation.
- This tells us that when we estimate the model parameters using least squares we are making the same probabilistic assumptions mentioned above.

Standard errors for regression models

- If we want to make inferences about the regression parameter estimates, then we also need an estimate of their variability.
- To compute this, we first need to compute the residual variance or error variance for the model.
- That is, how much variability in the dependent variable is not explained by the model.
- We need to compute the man squared error:

$$MS_{error} = \frac{SSE}{df} = \frac{\sum_{i=1}^{m}(y_i - h(x_i))^2}{m - p}$$

- The degrees of freedom (df) are determined by subtracting the number of estimated parameters (2 in this case: β_0 and β_1) from the number of observations (m)
- We can compute the standard error for the model as: $SE_{model} = \sqrt{MS_{error}}$

A significance test for β

- In order to get the standard error for a specific regression parameter estimate, SE_{β_x}, we need to rescale the standard error of the model by the square root of the sum of squares of the X variable:

$$SE_{\hat{\beta}_x} = \frac{SE_{model}}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2}}$$

- Now we can compute a t statistic to tell us the likelihood of the observed parameter estimates compared to some expected value under the null hypothesis.
- In this case we will test against the null hypothesis of no effect (i.e. $\beta_{H_0} = 0$):

$$t_{m-p} = \frac{\hat{\beta} - \beta_{H_0}}{SE_{\hat{\beta}_x}} = \frac{\hat{\beta}}{SE_{\hat{\beta}_x}}$$

- Later we will see that R automatically reports the t-statistics and p-values of all coefficients of a linear model.
- This allows us to determine whether the linear relationship between the two variables (y and x) is significant.

Example: a model of height

- We are going to work with the dataset `Howell1` that has demographic data from Kalahari !Kung San people collected by Nancy Howell in the late 1960s.
- The !Kung San are the most famous foraging population of the twentieth century, largely because of detailed quantitative studies by people like Howell. [McElreath, 2020]

Figure: By Staehler - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=45076017`

Example: a model of height

- Each observation corresponds to an individual.
- The variables of the dataset are:
  1. height: Height in cm
  2. weight: Weight in kg
  3. age: Age in years
  4. male: Gender indicator
- Let's explore the linear correlations between the variables

```
> library(rethinking)
> data(Howell1)
> d <- Howell1
> cor(d)
           height    weight          age         male
height 1.0000000 0.9408222 0.683688567 0.139229021
weight 0.9408222 1.0000000 0.678335313 0.155442866
age    0.6836886 0.6783353 1.000000000 0.005887126
male   0.1392290 0.1554429 0.005887126 1.000000000
```

Example: a model of height

- We can see that there is a positive correlation between `height` and `age`.
- Let's filter out the non-adult examples because we know that height is strongly correlated with age before adulthood.

```
d2 <- d[ d$age >= 18 , ]
```

- Now age doesn't correlate with height:

```
> cor(d2)
            height     weight          age       male
height  1.0000000  0.7547479 -0.10183776 0.69999340
weight  0.7547479  1.0000000 -0.17290430 0.52445271
age    -0.1018378 -0.1729043  1.00000000 0.02845498
male    0.6999934  0.5244527  0.02845498 1.00000000
```

- Let's model height as a function of weight using a simple linear regression:

$$\text{height}(\text{weight}) = \beta_0 + \beta_1 * \text{weight}$$

- In R the linear models are created with the command `lm` that receives as parameter a formula of the form `y~x` ($y = f(x)$).

Example: a model of height

```
> reg1<-lm(height~weight,d2)
> reg1

Call:
lm(formula = height ~ weight, data = d2)

Coefficients:
(Intercept)       weight
    113.879        0.905
```

- We can see that the coefficients of the model are $\beta_0 = 113.879$ and $\beta_1 = 0.905$.
- We can directly access the coefficients and store them in a variable:

```
> reg1.coef<-reg1$coefficients
> reg1.coef
(Intercept)      weight
113.8793936   0.9050291
```

Example: a model of height

- We can view various indicators about the linear model with the command **summary**:

```
> summary(reg1)

Call:
lm(formula = height ~ weight, data = d2)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

- We can see that β_0 and β_1 are both statistically significantly different from zero.

Example: a model of height

- We see that the coefficient of determination R^2 has a value of 0.57 which is not so good but acceptable.
- We can conclude that the weight while providing useful information to model a part of the variability of the height of the !Kung people, is not enough to build a highly reliable model.
- We can store the results of the command `summary` in a variable then access the coefficient of determination:

```
> sum.reg1<-summary(reg1)
> sum.reg1$r.squared
[1] 0.5696444
```

- We can also access the fitted values which are the values predicted by my model for the data used:

```
> reg1$fitted.values
       1        2        3        4        5        6
157.1630 146.9001 142.7180 161.8839 151.2362 170.8895
```

Example: a model of height

- We can check that the squared linear correlation between my fitted and observed values for the response variable is equivalent to the coefficient of determination:

```
> cor(d2$height,reg1$fitted.values)^2
[1] 0.5696444
```

- Suppose now that we know the weight for two !Kung people but I don't know the height.
- We could use my linear model to predict the height of these two people.
- To do this in R we must use the command `predict.lm` which receives the linear model and a data.frame with the new data:

```
> new.weights<-data.frame(weight=c(50,62))
> predict.lm(object=reg1,newdata=new.weights)
       1        2
159.1308 169.9912
> # this is equivalent to:
> reg1.coef[1]+reg1.coef[2]*new.weights[1:2,]
[1] 159.1308 169.99122
```

Multivariate Linear Regression

- Suppose we have n independent variables: $x_1, x_2, \ldots, x_n$.
- In many cases more variables can better explain the variability of the response variable **y** than a single one.
- A multivariate linear model is defined as follows:

$$y_i = \beta_0 + \beta_1 x_{i,1} + +\beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} + \epsilon_i \quad \forall i \in \{1, m\}$$

- In essence we are adding a parameter for each independent variable.
- Then, we multiply the parameter by the variable and add that term to the linear model.
- In the multivariate model all the properties of the simple linear model are extended.
- The problem can be represented in a matrix form:

$$Y = X\beta + \epsilon$$

Multivariate Linear Regression (2)

- Where Y is a vector $m \times 1$ response variables:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

- X is a $m \times (n+1)$ matrix with the explanatory variables. We have m observations of the n variables. The first column is constant equal to 1 ($x_{i,0} = 1 \quad \forall i$) to model the intercept variables β_0.

$$X = \begin{pmatrix} x_{1,0} & x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,0} & x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,0} & x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix}$$

Multivariate Linear Regression (2)

- Then, β is a $(n+1) \times 1$ vector of parameters.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

- Finally, ϵ is a $m \times 1$ vector with the model errors.

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

- Using matrix notation, we can see that the sum of squared errors (SSE) can be expressed as:

$$\text{SSE} = (Y - X\beta)^T (Y - X\beta)$$

- Minimizing this expression by deriving the error as a function of β and setting it equal to zero leads to the normal equations:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multivariate Linear Regression in R

- Now we will study a multiple linear regression for the `Howell1` data.
- Let's add the variable **age** as an additional predictor for **height**.
- We know that age is good at predicting height for non-adults so we will work with original dataset.
- Let's fit the following linear multi-variate model:

$$\text{height} = \beta_0 + \beta_1 * \text{weight} + \beta_2 * \text{age}$$

- In R to add more variables to the linear model we add them with the operator **+** :

```
reg2<-lm(height~weight+age,d)
```

Multivariate Linear Regression in R (7)

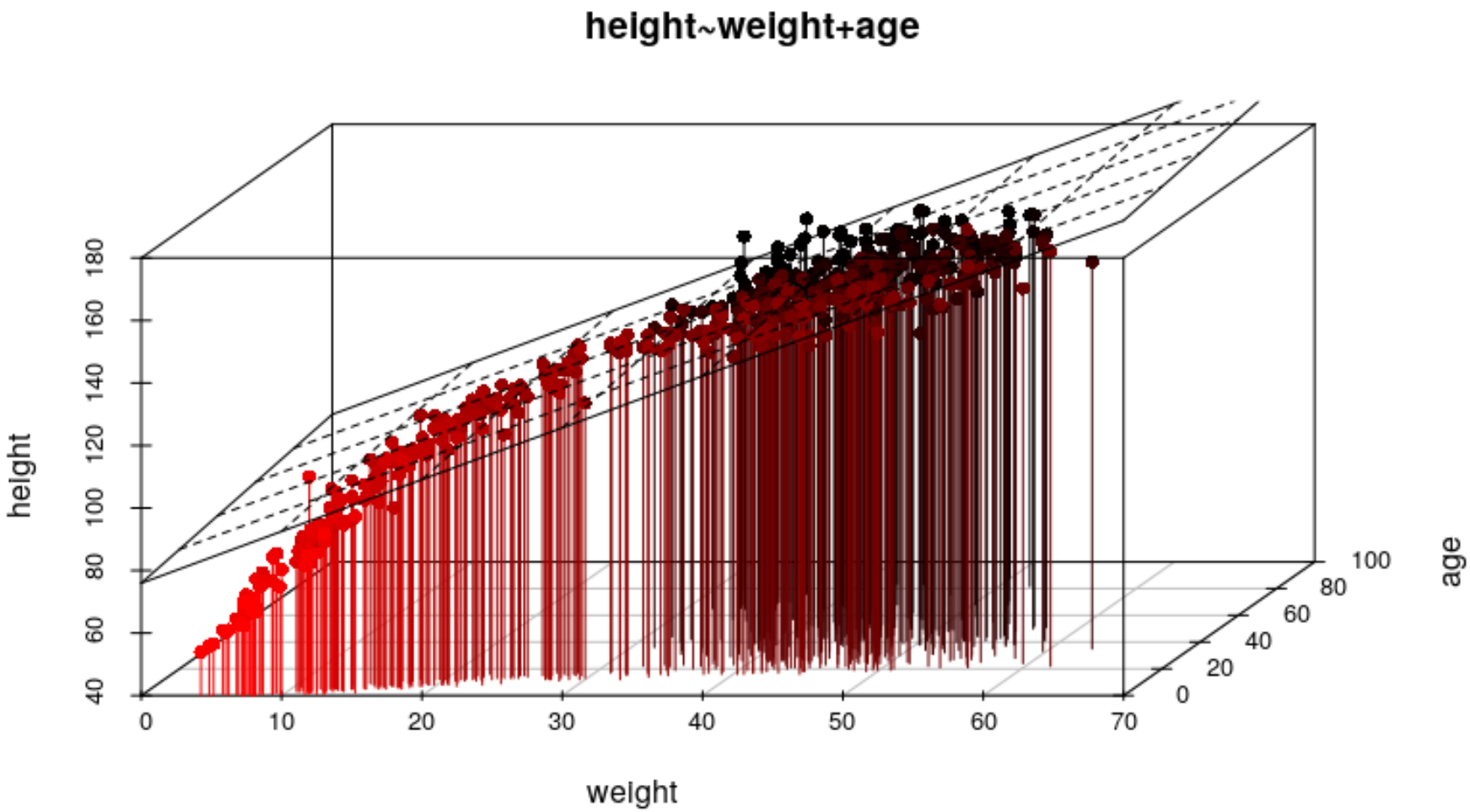
```
> summary(reg2)

Call:
lm(formula = height ~ weight + age, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-29.0350  -5.4228   0.7333   6.4919  19.6964

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.96329    1.04216  72.890  < 2e-16 ***
weight       1.65710    0.03656  45.324  < 2e-16 ***
age          0.11212    0.02594   4.322 1.84e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.214 on 541 degrees of freedom
Multiple R-squared:  0.889,Adjusted R-squared:  0.8886
F-statistic:  2166 on 2 and 541 DF,  p-value: < 2.2e-16
```

Multivariate Linear Regression in R (8)

- When we had a simple regression we could see the fitted model as a line.
- Now that we have two independent variables we can see the fitted model as a plane.
- If we had more independent variables our model would be a hyper-plane.
- We can plot the plane of our linear model of two independent variables and one dependent variable in R as follows:

```
library("scatterplot3d")
s3d <- scatterplot3d(d[,c("weight","age","height")],
                     type="h", highlight.3d=TRUE,
                     angle=55, scale.y=0.7, pch=16,
                     main="height~weight+age")
s3d$plane3d(reg2, lty.box = "solid")
```

Multivariate Linear Regression in R (9)



R^2 adjusted

- The coefficient of determination R^2 tends increases when extra explanatory variables are added to the model.
- R^2 adjusted or $\overline{R}^2$ penalizes for the number of variables p.

$$\overline{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

where m is the number of examples.

Polynomial Regression

- Polynomial regression uses powers of a variable—squares and cubes—as extra attributes.
- The most common polynomial regression is a parabolic model of the mean:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad \forall i$$

- Let's fit a polynomial regression for the height variable using a parabolic model for the weight.
- Because the square or cube of a large number can be truly massive we are going to standardize the weight by substracting the mean and dividing by the standard deviation.

```
d$weight_s <-( d$weight - mean(d$weight) )/sd(d$weight)
```

- Then we fit the model as follows:

```
> reg4 <- lm(height~weight_s+I(weight_s^2),d)
> reg4

Call:
lm(formula = height ~ weight_s + I(weight_s^2), data = d)

Coefficients:
  (Intercept)       weight_s  I(weight_s^2)
      146.660         21.415         -8.412
```

Conclusion

Bonus: Four Cardinal Rules of Statistics by Daniela Witten

Now that we have concluded the chapter on Frequentist inference, it is good to discuss the points raised by Daniela Witten in a tweet.

One: Correlation does not imply causation

- Yes, I know you know this, but it's so easy to forget! Yeah, YOU OVER THERE, you with the p-value of 0.0000001 — yes, YOU!! That's not causation.
- No matter how small the p-value for a regression of IQ onto shoe size is, that doesn't mean that big feet cause smarts!! It just means that grown-ups tend to have bigger feet and higher IQs than kids.
- So, unless you can design your study to uncover causation (very hard to do in most practical settings — the field of causal inference is devoted to understanding the settings in which it is possible), the best you can do is to discover correlations. Sad but true.

Bonus: Four Cardinal Rules of Statistics by Daniela Witten

Two: a p-value is just a test of sample size

- Read that again — I mean what I said! If your null hypothesis doesn't hold (and null hypotheses never hold IRL) then the larger your sample size, the smaller your p-value will tend to be.
- If you're testing whether mean=0 and actually the truth is that mean=0.000000001, and if you have a large enough sample size, then YOU WILL GET A TINY P-VALUE.
- Why does this matter? In many contemporary settings (think: the internet), sample sizes are so huge that we can get TINY p-values even when the deviation from the null hypothesis is negligible. In other words, we can have STATISTICAL significance w/o PRACTICAL significance.
- Often, people focus on that tiny p-value, and the fact that the effect is of **literally no practical relevance** is totally lost.
- This also means that with a large enough sample size we can reject basically ANY null hypothesis (since the null hypothesis never exactly holds IRL, but it might be "close enough" that the violation of the null hypothesis is not important).
- Want to write a paper saying Lucky Charms consumption is correlated w/blood type? W/a large enough sample size, you can get a small p-value. (Provided there's some super convoluted mechanism with some teeny effect size. . . which there probably is, b/c IRL null never holds)

Bonus: Four Cardinal Rules of Statistics by Daniela Witten

Three: seek and you shall find

- If you look at your data for long enough, you will find something interesting, even if only by chance!
- In principle, we know that we need to perform a correction for multiple testing if we conduct a bunch of tests.
- But in practice, what if we decide what test(s) to conduct AFTER we look at data? Our p-value will be misleadingly small because we peeked at the data. Pre-specifying our analysis plan in advance keeps us honest. . . but in reality, it's hard to do!!!

- Everyone is asking me about the mysterious and much-anticipated fourth rule of statistics. The answer is simple: we haven't figured it out yet.... that's the reason we need to do research in statistics

References I

McElreath, R. (2020).
Statistical rethinking: A Bayesian course with examples in R and Stan.
CRC press.

Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.