Class 5a: Multiple Linear Regression

Business Forecasting

Roadmap

This set of classes

- What is a multiple linear regression

Motivation

- Suppose that you are administering a hospital
- You need to know how many doctors, nurses and beds you need
- So you want to predict how long a patient will stay at the urgent care

- You collect the data on
    - The Duration of the visit
    - The type of patient
    - How many other people there are currently at urgent care
    - What kind of problem they came with
    - What type of bed they got

- If we know these factors, can we predict how long patient will stay?

Data

Show 10 entries

ID	Duration	Occupancy	SEXO	EDAD	TIPOCAMA	MOTATE
2693326	22	3	FEMENINO	19	SIN CAMA	MÉDICA
3687260	113	8	FEMENINO	50	CAMA DE OBSERVACION	MÉDICA
8332891	11	1	FEMENINO	20	SIN CAMA	GINECO-OBSTÉTRICA
2719030	15	1	FEMENINO	22	SIN CAMA	MÉDICA
2671304	15	1	FEMENINO	4	SIN CAMA	MÉDICA
5450507	67	4	FEMENINO	48	SIN CAMA	GINECO-OBSTÉTRICA
2782600	320	22	FEMENINO	78	NO ESPECIFICADO	MÉDICA
2247738	380	12	MASCULINO	42	SIN CAMA	MÉDICA
4385048	7	2	MASCULINO	26	SIN CAMA	MÉDICA
2984341	29	3	FEMENINO	55	CAMA DE OBSERVACION	MÉDICA

Showing 1 to 10 of 4,998 entries

Previous 1 2 3 4 5 … 500 Next

Multiple linear regression

Suppose that the outcome y_i (duration) is a linear function of x_1 (occupancy) and x_2 (age)

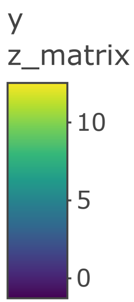
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

- β_0 represents the value of y_i when x_1 and x_2 are 0.
- β_1 represents the change in y_i while changing x_1 by one unit and keeping x_2 constant
- β_2 represents the change in y_i while changing x_2 by one unit and keeping x_1 constant

Multiple linear regression

100 observations simulated from an a regression line:

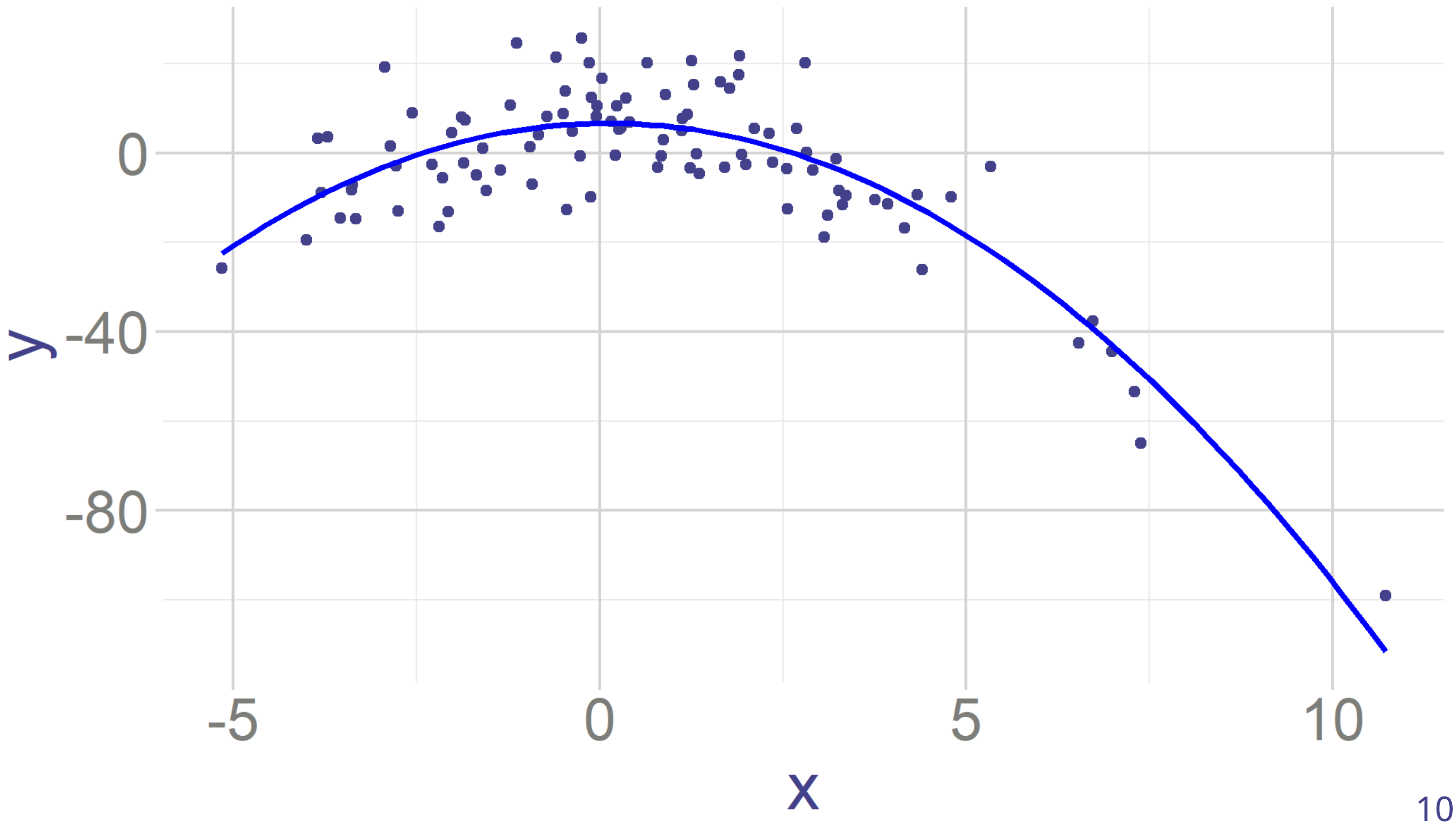
$$y_i = 5 + 2x_{i1} + 1x_{i2} + u_i$$



Multiple linear regression

100 observations simulated from an a regression line:

$$y_i = 5 + 2x_i - 1x_i^2 + u_i$$



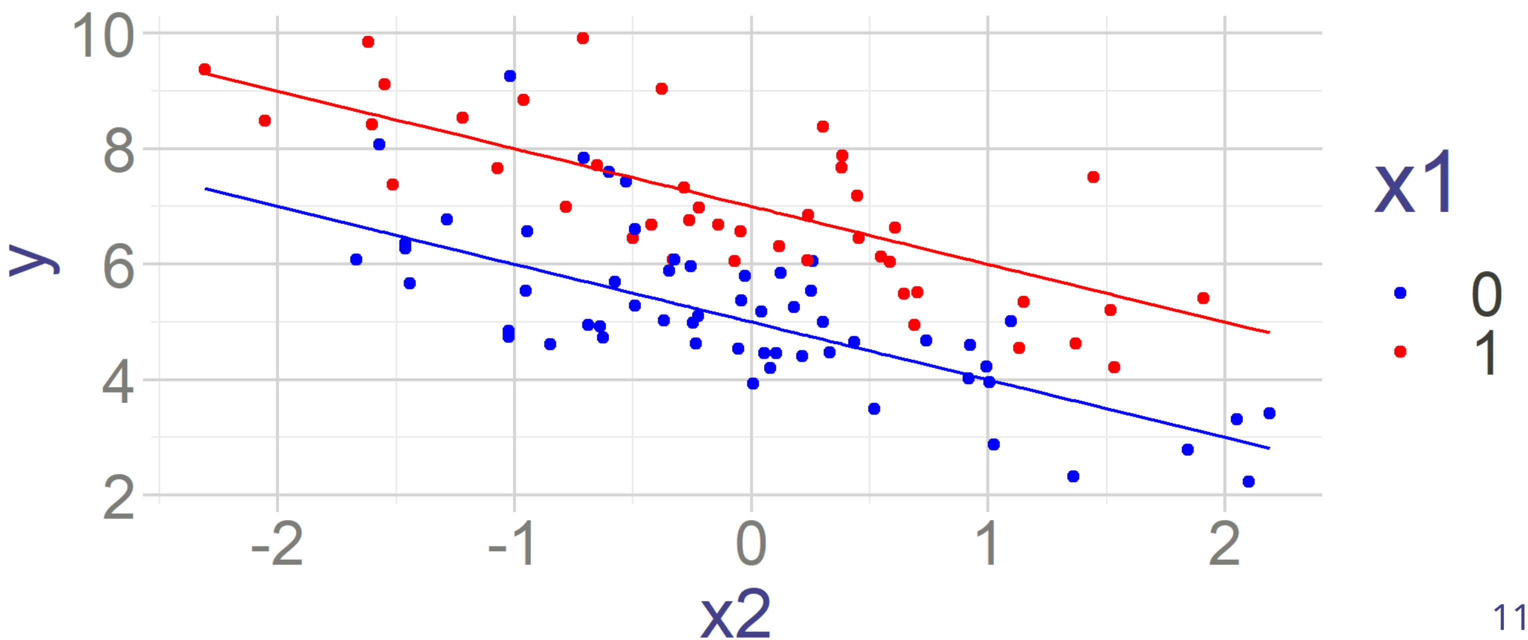
Multiple linear regression

Suppose that:

$$x_1 = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

100 observations simulated from an a regression line:

$$y_i = 5 + 2x_{i1} - 1x_{i2} + u_i$$



Multiple linear regression

Now imagine a regression with k variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

- Maybe you are trying to predict customer spending based on what they looked at and x_{ij} represent how long customer i looked at item j
- Maybe you are trying to predict sales in a store i, and x_{ij} represent prices of the products, their competitors' products, how many people live around and how rich are they etc...
- We can no longer visualize it (because we can't visualize more than 3 dimensions)

Multiple linear regression

We can also write it in the vector form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k, x_{ik} + u_i$$

In vector form is:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\substack{\mathbf{y} \\ n\times 1}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ldots. & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}}_{\substack{\mathbf{X} \\ n\times(k+1)}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\substack{\beta \\ (k+1)\times 1}} + \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}}_{\substack{\mathbf{u} \\ n\times 1}}$$

Full Rank

Important Assumption: **X is full rank**

- Has same rank as the number of parameters: $p = k + 1$
- Also known as: no perfect multicolinearity
- Technically: columns of X should be linearly independent

- Intuitively: none of the variables are perfectly correlated. If they are perfectly correlated, then we don't need one of the columns because we can perfectly predict one column with information from another column.

- Suppose that one column is income in USD, and the second one is income measured in Pesos. They are perfectly correlated. Once we know income in USD, income in Pesos does not bring any additional information. We would not be able to estimate the effect of both income in USD and income in Pesos at the same time.

$$\text{Full Rank Matrix:} \quad \text{Matrix Not of Full Rank:}$$
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 4 \\ 4 & 5 & 10 \\ 7 & 8 & 16 \end{bmatrix}$$

Multiple linear regression

Goal:

- Estimate the vector of parameters β

Procedure

- Find

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

- Which minimizes the squared errors in the problem:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + e_i$$

- That is minimize

$$SSE = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

Multiple linear regression

- We can do it with scalars

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\right) = 0$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_{i1}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\right) = 0$$

$$\vdots$$

$$\frac{\partial SSE}{\partial \hat{\beta}_k} = -2\sum_{i=1}^{n} x_{ik}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\right) = 0$$

- We have $k+1$ equations with $k+1$ unknowns.

Multiple linear regression

- Or we can do it with vectors
- First rewrite the sum of squares:

$$SSE(b) = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \mathbf{y'y} - \mathbf{2b'X'y} + \mathbf{b'X'Xb}$$

- Then minimize it with respect to $\mathbf{b}$

$$\frac{\partial}{\partial \mathbf{b}}(\mathbf{y'y} - \mathbf{2b'X'y} + \mathbf{b'X'Xb}) = -\mathbf{2X'y} + \mathbf{2X'Xb}$$

- $\hat{\beta}$ is the solution of such minimization (our OLS estimator)

$$\begin{aligned} -\mathbf{2X'y} + \mathbf{2X'X}\hat{\beta} &= 0 \\ \mathbf{X'X}\hat{\beta} &= \mathbf{X'y} \\ \hat{\beta} &= (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'y} \end{aligned}$$

Multiple linear regression

Looking more closely at the **first order condition**:

$$\underbrace{\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}}_{\mathbf{X}'\mathbf{X}} \underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}}_{\hat{\beta}} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}}_{\mathbf{X}'\mathbf{y}}$$

Looking more closely and it's **solution**:

$$\underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}}_{\hat{\beta}} = \underbrace{\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}^{-1}}_{(\mathbf{X}'\mathbf{X})^{-1}} \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}}_{\mathbf{X}'\mathbf{y}}$$

Special Case: k=1

What if we have just one x?

$$\underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}}_{\hat{\beta}} = \underbrace{\begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}^{-1}}_{(\mathbf{X}'\mathbf{X})^{-1}} \underbrace{\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \end{bmatrix}}_{\mathbf{X}'\mathbf{y}}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^n x_{i1}^2}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} & \frac{-\sum_{i=1}^n x_{i1}}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \\ \frac{-\sum_{i=1}^n x_{i1}}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} & \frac{n}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \end{bmatrix}$$

which gives:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x}_1 \frac{\sum (x_{1i} y_i - n\bar{y}\bar{x}_1)}{\sum_{i=1}^n x_{i1}^2 - n\bar{x}_1^2} \\ \frac{\sum_i x_{1i} y_i - n\bar{x}_1\bar{y}}{\sum_{i=1}^n x_{i1}^2 - n\bar{x}_1^2} \end{bmatrix}$$

Predictions

To make predictions based on the estimated regressors we use:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

Or in the vector form:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ is called a hat matrix.

Residuals

To get residuals, we calculate:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

Or in the vector form:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = y - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Example with numbers

Dataset:

Student	Hours Studied (x_1)	Hours Slept (x_2)	Exam Score (y)
1	3	8	80
2	4	7	85
3	6	6	92
4	5	7	88

X matrix:

$$X = \begin{bmatrix} 1 & 3 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 6 \\ 1 & 5 & 7 \end{bmatrix}$$

Response Vector (y) :

$$y = \begin{bmatrix} 80 \\ 85 \\ 92 \\ 88 \end{bmatrix}$$

We are trying to find:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Example with numbers

Multiply X' by X:

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 4 & 6 & 5 \\ 8 & 7 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 3 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 6 \\ 1 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 18 & 28 \\ 18 & 86 & 123 \\ 28 & 123 & 198 \end{bmatrix}$$

Find the inverse $(X'X)^{-1}$

$$(X'X)^{-1} = \begin{bmatrix} 474.75 & -30 & -48.5 \\ -30 & 2 & 3 \\ -48.5 & 3 & 5 \end{bmatrix}$$

Example with numbers

Next let's find $X'y$

$$X'y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 4 & 6 & 5 \\ 8 & 7 & 6 & 7 \end{bmatrix} \begin{bmatrix} 80 \\ 85 \\ 92 \\ 88 \end{bmatrix} = \begin{bmatrix} 345 \\ 1572 \\ 2403 \end{bmatrix}$$

So, our coefficients are:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 474.75 & -30 & -48.5 \\ -30 & 2 & 3 \\ -48.5 & 3 & 5 \end{bmatrix}}_{(X'X)^{-1}} \underbrace{\begin{bmatrix} 345 \\ 1572 \\ 2403 \end{bmatrix}}_{X'y} = \begin{bmatrix} 83.25 \\ 3 \\ -1.5 \end{bmatrix}$$

Interpretation

- Score with 0 hours of sleep and 0 of studying is 83.25
- 1 more hour of studying (without changing sleep hours) increases score by 3
- 1 more hour of sleep (without changing study hours) decreases score by 1.5

Example with numbers

We can find predicted values:

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 3 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 6 \\ 1 & 5 & 7 \end{bmatrix} \begin{bmatrix} 83.25 \\ 3 \\ -1.5 \end{bmatrix} = \begin{bmatrix} 80.25 \\ 84.75 \\ 92.25 \\ 87.75 \end{bmatrix}$$

And the residuals:

$$e = y - \hat{y} = y - X\hat{\beta} = \begin{bmatrix} 80 \\ 85 \\ 92 \\ 88 \end{bmatrix} - \begin{bmatrix} 80.25 \\ 84.75 \\ 92.25 \\ 87.75 \end{bmatrix} = \begin{bmatrix} -0.25 \\ 0.25 \\ -0.25 \\ 0.25 \end{bmatrix}$$

Example from data:

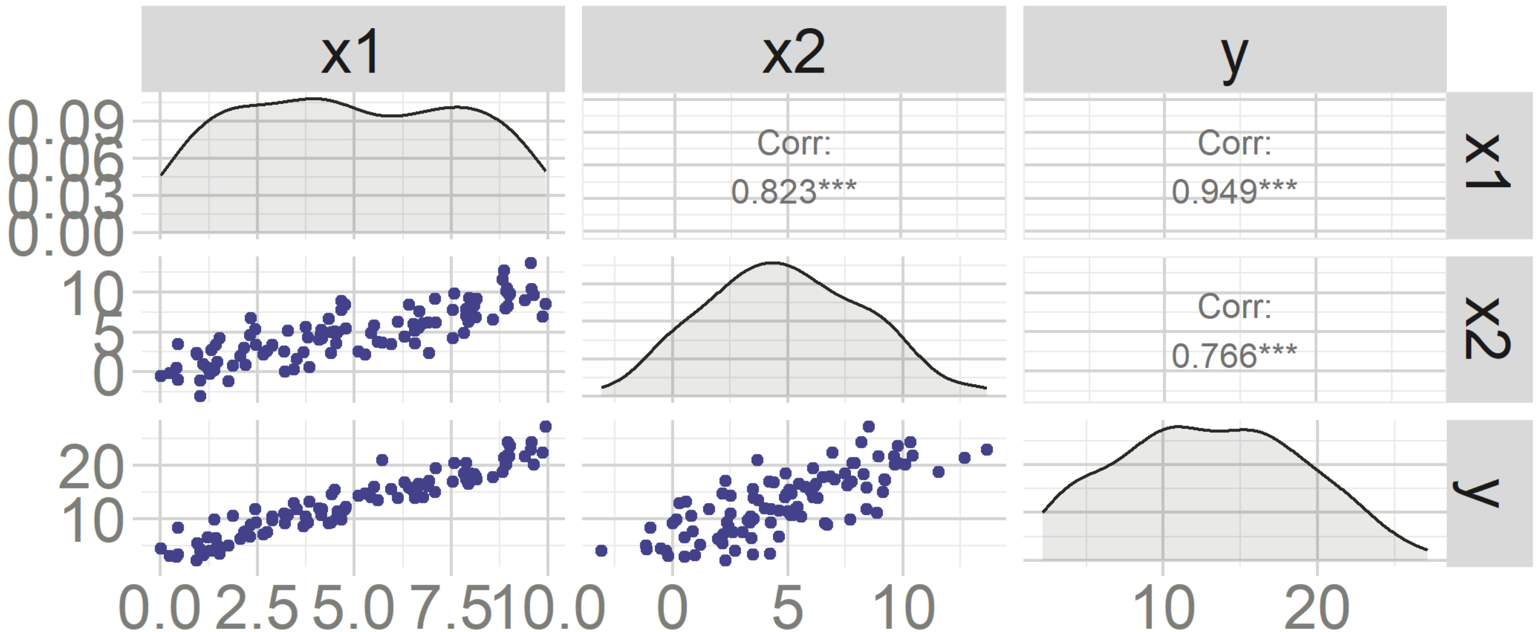
```r
# Fit a linear regression model
lm_model <- lm(Duration ~ Occupancy+EDAD, data = Sample_urg)
# Display the summary of the linear regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD, data = Sample_urg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.65  -26.61  -17.27   -0.57 1252.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.23422    2.48416   9.353  < 2e-16 ***
## Occupancy    3.70354    0.10090  36.705  < 2e-16 ***
## EDAD         0.20626    0.06747   3.057  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.99 on 4995 degrees of freedom
## Multiple R-squared:  0.2169,    Adjusted R-squared:  0.2166
## F-statistic: 691.8 on 2 and 4995 DF,  p-value: < 2.2e-16
```

Correlations vs Coefficients

Note, that x_1 and x_2 can both have positive correlation with y_i, but different coefficients!

- Suppose x_1 is study hours, x_2 is coffee cups drunk by a student, and y is student's score on the exam.



Correlations vs Coefficients

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.779 -1.422 -0.418  1.096  6.305
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13966    0.38033   8.255 7.68e-13 ***
## x1           2.06132    0.11686  17.640  < 2e-16 ***
## x2          -0.08510    0.09798  -0.868    0.387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.88 on 97 degrees of freedom
## Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8997
## F-statistic: 445.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

- Why coffee has 0 impact?
- Because it only helps to study longer, but comparing students who study the same amount, drinking more coffee is not better.

OLS Properties

- As usual, we asked whether it's unbiased and what is its variance.
- **Unbiased**:

$$
\begin{aligned}
E(\hat{\beta}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}))) \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}))) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta) + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\
&= \beta + 0 = \beta
\end{aligned}
$$

Where $E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})$ if $E(u|X) = 0$ (our usual assumption).

- **Variance**

$$
Var(\hat{\beta}) = Cov(\hat{\beta}) = \underbrace{\begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & cov(\hat{\beta}_0, \hat{\beta}_k) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \dots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_0) & cov(\hat{\beta}_k, \hat{\beta}_1) & \dots & var(\hat{\beta}_k) \end{bmatrix}}_{(k+1)\times(k+1)}
$$

- So it's a matrix with variance of single parameters on the diagonal and covariances off the diagonal.

Variance

First, note that:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + \mathbf{u} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Let's use this

$$\begin{aligned} var(\hat{\beta}) &= \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])'] \\ &= \mathbb{E}[(X'X)^{-1}X'\mathbf{u}((X'X)^{-1}X'\mathbf{u})'] = (X'X)^{-1}X'\mathbb{E}[\mathbf{u}\mathbf{u}']X(X'X)^{-1} \\ &= (X'X)^{-1}X'(I\sigma^2)X(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

So

$$var(\hat{\beta}_k) = \sigma^2(X'X)^{-1}_{k+1,k+1}$$

where $(X'X)^{-1}_{k+1,k+1}$ is element in k row and k column of $(X'X)^{-1}$ matrix.

- Because first coefficient is β_0
- And standard deviation is just square root of this!

Variance

- Where the hell do we get the σ^2 from?!
- Same as before:

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n - p}$$

- Where e_i is fitted residual and p is number of parameters $p = k + 1$
- This is called mean squared error as well

The easiest way to compute this sum is:

$$\sum_i e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$$

Gauss Markov Theorem (Again)

Assumptions

- $E(u_i|X) = 0$
- $var(u_i) = \sigma^2$
- $cov(u_i, u_j) = 0$
- X is full rank

NO NEED FOR NORMALITY

Theorem: OLS is BLUE: Best, Linear, Unbiased Estimator

- It has the lowest variance among linear and unbiased estimators
- What's a linear estimator?
  - It's an estimator where β coefficients are linear functions of outcomes
  - Anything of the form $b = Cy$ where C is p x n matrix.
  - So $b_1 = c_{11}y_1 + c_{12}y_2 + \ldots + c_{13}y_3$
  - Example $b_1 = \frac{1}{n}y_1 + \ldots + \frac{1}{n}y_n$
- How is OLS linear? $\hat{\beta} = Cy = \underbrace{(X'X)^{-1}X'}_{C}y$

Categorical Variables in a Regression

- Suppose we want to learn whether mode of work affects workers productivity.
- Each worker can be in one of these 3 categories:
  - Fully at the office
  - Fully remote
  - Hybrid

Show 8 entries

WorkerID	Productivity	WorkMode
1	93	Fully remote
2	75	Fully remote
3	108	Hybrid
4	115	Hybrid
5	75	Hybrid
6	88	Hybrid
7	107	Fully remote
8	115	Fully at the office

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 … 13 Next

- How do we estimate the impact of categorical variable?
- We turn it into a series of binary variables (or indicator variables)!

$$D_{i,Remote} = \begin{cases} 1 & WorkMode_i = FullyRemote \\ 0 & otherwise \end{cases}$$

$$D_{i,Hybrid} = \begin{cases} 1 & WorkMode_i = Hybrid \\ 0 & otherwise \end{cases}$$

Show 6 entries

WorkerID	Productivity	WorkMode	WorkModeFully.at.the.office	WorkModeFully.remote	WorkModeHybrid
1	112	Fully at the office	1	0	0
2	124	Hybrid	0	0	1
3	108	Hybrid	0	0	1
4	76	Fully at the office	1	0	0
5	125	Fully remote	0	1	0
6	111	Fully at the office	1	0	0

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 … 17 Next

- For each person, only one of these dummies is equal to 1!

- We will add these dummies into a regression, but not all of them!
- If we have m categories, we will add m-1 dummies. Why?

$$y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \ldots + \beta_{m-1} D_{im-1} + u_i$$

- In our Example:

$$y_i = \beta_0 + \beta_1 D_{i,Hybrid} + \beta_2 D_{i,Remote} + u_i$$

- Because otherwise X would not be full rank!

$$\text{Full Rank Matrix:} \quad \text{Matrix Not of Full Rank:}$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

- Intuitively, if I know that the values of $D_{i,Hybrid}$ and $D_{i,Remote}$, I know the value of $D_{i,Office}$
- Ex: if they don't work hybrid and don't work remote, I know they work at the office
- So including it does not bring any new information

- R automatically transform categorical variable to dummies and excludes one of them

```
# Fit a linear regression model
lm_model <- lm(Productivity ~ WorkMode, data = d)
# Display the summary of the linear regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = Productivity ~ WorkMode, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.774 -12.636   0.946  14.410  34.667
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           101.590      2.695  37.697   <2e-16 ***
## WorkModeFully remote   -7.256      4.087  -1.775    0.079 .
## WorkModeHybrid          6.184      4.050   1.527    0.130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.83 on 97 degrees of freedom
## Multiple R-squared:  0.09125,    Adjusted R-squared:  0.07251
## F-statistic:  4.87 on 2 and 97 DF,  p-value: 0.009652
```

Interpretation of Coefficients

- Coefficient on a dummy D_1 tells us by how much y changes when we change category from the excluded one to the category 1.
- In our example
  - Excluded category is: work fully at the office - this is our comparision group
  - $\beta_{hybrid} = 6.184$: employees working in hybrid mode have on average 6.184 higher productivity score compared to the ones working at the office
  - $\beta_{remote} = -7.256$: employees working in fully remotely have on average 7.256 lower productivity score compared to the ones working at the office
  - The t-test on these coefficients tells us whether these differences in means across categories are significant!
- Bottom line: the coefficients on the dummies show the average difference between y in that category compared to the excluded category (holding everything else unchanged)

Example

Suppose we have a categorical variable representing education level. We run a regression of income on the education level. Interpret the coefficients.

Show 6 entries

WorkerID	Income	Education
1	71497	Master
2	80993	Bachelor
3	87772	Master
4	79617	Bachelor
5	68597	PhD
6	75982	Bachelor

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 … 17 Next

```r
# Fit a linear regression model
lm_model <- lm(Income ~ Education, data = d)
# Display the summary of the linear regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = Income ~ Education, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25868 -10865  -1413  10204  28280
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          70342       3125  22.509  < 2e-16 ***
## EducationPhD         14639       4008   3.652 0.000424 ***
## EducationMaster      22303       4157   5.365 5.59e-07 ***
## EducationBachelor    16993       4273   3.977 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13980 on 96 degrees of freedom
## Multiple R-squared:  0.2401,    Adjusted R-squared:  0.2164
## F-statistic: 10.11 on 3 and 96 DF,  p-value: 7.517e-06
```

Interactions

Consider a regression:

$$\text{Duration}_i = \beta_0 + \beta_1 \text{Occupancy}_i + \beta_2 \text{Male}_i + u_i$$

- Where Male is a for patient i being male
- We assumed that occupancy has always the same effect, independent of your gender
- But what if occupancy matters more for men?
- In other words: one additional patient on urgent care increases duration by more if you are a men?
- Why? Maybe because when there is a lot of patients, doctors prioritize women (or men)
- We want allow the coefficient on occupancy to differ by gender. How?

Interactions

- Run the regression:

$$\text{Duration}_i = \beta_0 + \beta_1 \text{Occupancy}_i + \beta_2 \text{Male}_i + \beta_3 \text{Occupancy}_i * \text{Male}_i + u_i$$

- What's the coefficient on Occupancy when you are a woman $Male_i = 0$?

$$\text{Duration}_i = \beta_0 + \beta_1 \text{Occupancy}_i + \beta_2 \text{Male}_i + \beta_3 \text{Occupancy}_i * 0 + u_i$$
$$\text{Duration}_i = \beta_0 + \beta_1 \text{Occupancy}_i + \beta_2 \text{Male}_i + u_i$$

- What's the coefficient on Occupancy when you are a man $Male_i = 1$?

$$\text{Duration}_i = \beta_0 + \beta_1 \text{Occupancy}_i + \beta_2 \text{Male}_i + \beta_3 \text{Occupancy}_i * 1 + u_i$$
$$\text{Duration}_i = \beta_0 + (\beta_1 + \beta_3) \text{Occupancy}_i + \beta_2 \text{Male}_i + u_i$$

We can estimate β_3 and it will tell us by how much bigger is the coefficient on occupancy for men compared to the coefficient on occupancy for women.

- β_1 is the coefficient for women
- $\beta_1 + \beta_3$ is the coefficient for women
- β_3 is the difference in slopes, which we can test like other coefficients

```
##
## Call:
## lm(formula = Duration ~ Occupancy * SEXO, data = Sample_urg[Sample_urg$SEXO
##     "NO ESPECIFICADO", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1030.01   -26.49   -17.87    -1.11  1297.28
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               30.8015     1.8861  16.331   <2e-16 ***
## Occupancy                  2.6903     0.1264  21.278   <2e-16 ***
## SEXOMASCULINO             -4.8637     3.2324  -1.505    0.132
## Occupancy:SEXOMASCULINO    2.6174     0.2031  12.889   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.27 on 4994 degrees of freedom
## Multiple R-squared:  0.2441,    Adjusted R-squared:  0.2437
## F-statistic: 537.6 on 3 and 4994 DF,  p-value: < 2.2e-16
```

- One additional patients increases duration for women by 2.69 minutes
- One additional patients increases duration for men by 2.69+2.61=5.2 minutes
- What could be reasons for this?

Interactions

- More generally, we can rewrite a regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} * x_{i2} + u_i$$

- As

$$y_i = \beta_0 + (\beta_1 + \beta_3 x_{i2})x_{i1} + \beta_2 x_{i2} + +u_i$$

- β_3 answers the following question:

- If I increase x_{i2} by one, by how much the coefficient on x_{i1} changes?

Interactions

- Suppose you want to know who benefits the most from working from home. You collect survey data for each employee on the job satisfaction, whether they work in the office or from home, and the distance between the office and home

- Who do you think benefits most from working from home?

- How would you test this?

$$\text{Satisfaction}_i = \beta_0 + \beta_1 \text{WFH}_i + \beta_2 \text{Distance}_i + \beta_3 \text{WFH}_i * \text{Distance}_i + u_i$$

- What's the interpretation of β_3?

- By how much the effect of working from home on satisfaction changes when we increase distance by one unit (km)

- Which sign do you expect β_3 to have?

Goodness of fit

- We can use again the R square to measure the goodness of fit.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

- However, there is one problem with it.
    - Even if we add variables unrelated to y, the R^2 would typically still increase by a bit
    - Even if in population there is 0 relationship with this variable, our sample is small so we will never get exactly 0 relationship
    - Sampling noise will make coefficient slightly positive or negative
    - So the increase in R^2 will reflect that noise in our sample
    - The more coefficients we include, the higher R^2
    - We can adjust it, by accounting for the number of parameters used

$$R^2_{Adj} = 1 - \frac{\sum(y_i - \hat{y}_i)^2/(n-p)}{\sum(y_i - \bar{y}_i)^2/(n-1)}$$

- More parameters -> $\downarrow (n-p) \rightarrow \uparrow \sum(y_i - \hat{y}_i)^2/(n-p) \rightarrow \downarrow R^2_{Adj}$
- So it balances off the mechanical effect of higher R^2 due to more regressors

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD, data = Sample_urg[Sample_urg$SEXO
##     "NO ESPECIFICADO", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.65  -26.61  -17.27   -0.57 1252.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.23422    2.48416   9.353  < 2e-16 ***
## Occupancy    3.70354    0.10090  36.705  < 2e-16 ***
## EDAD         0.20626    0.06747   3.057  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.99 on 4995 degrees of freedom
## Multiple R-squared:  0.2169,    Adjusted R-squared:  0.2166
## F-statistic: 691.8 on 2 and 4995 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD + Random_var, data = Sample_urg[Sa
##     "NO ESPECIFICADO", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.70  -26.88  -17.31   -0.41 1253.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.76414    3.85025   5.393 7.25e-08 ***
## Occupancy    3.70326    0.10090  36.701  < 2e-16 ***
## EDAD         0.20566    0.06747   3.048  0.00231 **
## Random_var   0.45292    0.53938   0.840  0.40111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.99 on 4994 degrees of freedom
## Multiple R-squared:  0.217,    Adjusted R-squared:  0.2165
## F-statistic: 461.4 on 3 and 4994 DF,  p-value: < 2.2e-16
```

- Adding random variable increased R^2 but not R^2_{Adj}

Statistical Properties of OLS

Inference

- Let's add the assumption that errors are normally distributed:

$$\mathbf{u} \sim N(0, \sigma I)$$

Which means that:

$$y \sim N(X\beta, \sigma I)$$

- With inference we can:
    - Do hypothesis testing on single coefficients, ex: $H_0 : \beta_2 = 0$
    - Find confidence intervals for a single coefficients
    - Do hypothesis testing on multiple coefficients: ex: $H_0 : \beta_1 = \beta_2$

Test for a Single Coefficient

Under the above assumptions:

$$\hat{\beta} \sim N(\beta, \sigma\sqrt{(X'X)^{-1}})$$

And

$$\hat{\beta}_j \sim N(\beta, \sigma\sqrt{(X'X)^{-1}_{j+1,j+1}})$$

Normalizing we get that:

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{(X'X)^{-1}_{j+1,j+1}}} \sim t_{n-p}$$

- This test statistic has student t distribution with n-p degrees of freedom
  - Because the $\frac{s^2(n-p)}{\sigma^2} \sim \chi_{n-p}$
- Where p is the number of parameters (coefficients)
- $p = k + 1$: k regressors and 1 intercept

Test for a single coefficient

Suppose:

- $H_0 : \beta_j = \beta_{j0}$
- $H_A : \beta_j \neq \beta_{j0}$

Then, we use test statistic:

$$t_{test} = \frac{\hat{\beta}_j - \beta_{j0}}{s\sqrt{(X'X)^{-1}_{j+1,j+1}}}$$

And we reject if $t_{test} > t_{\alpha/2,n-p}$ or $t_{test} < -t_{\alpha/2,n-p}$

Where $t_{\alpha/2,n-p}$ is $1 - \alpha/2$ quantile of student t with n-p degrees of freedom
NOTE: This is a test for β_j given all other regressors. It's not the same as the test statistic with only one regressor!

Example

Suppose:

- $H_0 : \beta_{Age} = 0$
- $H_A : \beta_{Age} \neq 0$

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD, data = Sample_urg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.65  -26.61  -17.27   -0.57 1252.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.23422    2.48416   9.353  < 2e-16 ***
## Occupancy    3.70354    0.10090  36.705  < 2e-16 ***
## EDAD         0.20626    0.06747   3.057  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.99 on 4995 degrees of freedom
## Multiple R-squared:  0.2169,    Adjusted R-squared:  0.2166
## F-statistic: 691.8 on 2 and 4995 DF,  p-value: < 2.2e-16
```

Confidence Interval for a Single Coefficient

We can also use this distribution to construct confidence intervals:

An interval for β_j with confidence level $1-\alpha$ is:

$$\begin{aligned} CI_{1-\alpha} &= \{\hat{\beta}_j - t_{\alpha/2,n-p} SE(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2,n-p} SE(\hat{\beta}_j)\} \\ &= \{\hat{\beta}_j - t_{\alpha/2,n-p} s \sqrt{(X'X)^{-1}_{j+1,j+1}}, \hat{\beta}_j + t_{\alpha/2,n-p} s \sqrt{(X'X)^{-1}_{j+1,j+1}}\} \end{aligned}$$

Intepretation:

- We are $1-\alpha$ % confident that the true parameter is within this CI
- If we take repeated samples, $1-\alpha$ % of such constructed confidence intervals would contain true β

Example:

For our age coefficient we had:

- $\hat{\beta}_{Age} = 0.206$
- $SE(\hat{\beta}) = 0.067$
- Our $n = 5000$ so we can use normal approximation

So 95% CI for β_{Age} is:

$$
\begin{aligned}
CI_{1-\alpha} &= \{\hat{\beta}_j - t_{\alpha/2,n-p}SE(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2,n-p}SE(\hat{\beta}_j)\} \\
&= \{0.206 - 1.96 * 0.067, 0.206 + 1.96 * 0.067\} \\
&= \{0.075, 0.337\}
\end{aligned}
$$

- Note that the CI does not contain 0
- What does it imply for hypothesis testing with $H_0 : \beta_{age} = 0$?

CI for mean response

Suppose that we want an average prediction for individuals with these characteristics:

$$\mathbf{x_0} = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

Ex: What's average income (y), for people who whave 12 years of education $x_{01} = 12$ (2 other people are there) and are age 50 $x_{02} = 50$

How accurate is our prediction?

$$\hat{y}_0 = \mathbf{x_0}'\hat{\beta}$$

The prediction is unbiased:

$E(\hat{y}_0) = \mathbf{x_0}'\beta$

and it's variance is:

$$\begin{aligned} var(\hat{y}_0) &= var(\mathbf{x_0}'\hat{\beta}) \\ &= \mathbf{x_0}'var(\hat{\beta})\mathbf{x_0} \\ &= \sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0} \end{aligned}$$

So it's distribution is:

$\hat{y}_0 \sim N(\mathbf{x_0}'\beta, \sqrt{\sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}})$

Hence:

$CI_{1-\alpha} = \{\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}}\sqrt{\sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}\}$

Exanmple

What's the 95% CI for average wait time when there is 10 people at the Urgent Care $x_{occupancy} = 10$ for a person who is of age 52 $x_{age} = 52$?

- What do we need to answer this question?
- $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_{occupancy}, \hat{\beta}_{age}\} = \{23.236, 3.7, 0.2\}$
- $\sqrt{\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}} = \sqrt{[1, 10, 52](\mathbf{X}'\mathbf{X})^{-1}[1, 10, 52]'} = 0.021$
- $\sigma = 98.97$
- Prediction: $\hat{y_0} = 23.236 * 1 + 3.7 * 10 + 0.2 * 52 = 70.636$
- Standard Deviation: $SE(\hat{y_0}) = \sqrt{\sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}} = 2.07837$

$$CI_{95} = \{70.636 \pm 1.96 * 2.07837\} \approx \{67, 75\}$$

Exanmple

Or simply in R:

```
lm_model <- lm(Duration ~ Occupancy+EDAD, data = Sample_urg)
new_data<- data.frame(Occupancy= c(10), EDAD=52)
predict(lm_model, newdata = new_data, interval = "confidence", level = (
```

```
## $fit
##        fit      lwr      upr
## 1 70.9952 66.93326 75.05714
##
## $se.fit
## [1] 2.071955
##
## $df
## [1] 4995
##
## $residual.scale
## [1] 98.99182
```

CI for new observation

Reminder :

- When we look at average response, u_i doesn't play a role (because on average errors are 0)
- When we look at a single observation, u_i matters, so it increases the variance of prediction error

So variance is now the previous variance plus the variance of u_i

$$\begin{aligned} var(y_0 - \hat{y}_0) &= var(x_0\beta + u_i - x_0\hat{\beta}) \\ &= var(u_0) + var(x_0\hat{\beta}) \\ &= \sigma^2 + \sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0} \end{aligned}$$

So the confidence interval for a single observation is slightly wider:

$$CI_{1-\alpha} = \{\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}}\sqrt{\sigma^2(1 + \mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0})}\}$$

We are less certain about predicting outcome for a single person, compared to average outcome among namy people.

Testing for the significance of the regression

- Does our model helps to explain any variation in y_i?
- $H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$
- $H_A : \beta_j \neq 0$ for at least one j
- It's the same procedure as before!
  - **Explained variation** should be large compared to **unexplained variation** if the model works
- We can again do the decomposition in SST, SSR, and SSE:
  - SS_T is total sum of squares $\sum_i (y_i - \bar{y})^2$, n-1 DoF
  - SS_R is regression sum of squares $\sum_i (\hat{y}_i - \bar{y})^2$, k DoF
  - SS_E is residual error sum of squares $\sum_i (y_i - \hat{y}_i)^2$, n-k-1 DoF

Source	Sum of Squares	Degrees of Freedom	DoF
Regression	13557462	2	k
Residual Error	48947909	4995	n-k-1
Total	62505371	4997	n-1

Testing for the significance of the regression

F-stat and its distribution under the null

$$F_{stat} = \frac{SSR/(k)}{SSE/(n-k-1)} \sim \underbrace{F_{k,n-k-1}}_{\text{Dist under } H_0}$$

Alternative way to think about it:

- $H_0 : y = \beta_0 + u$ restricted model
- $H_A : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

If H_A is true, restricted model should explain more of y

$$F_{stat} = \frac{SSR/(k)}{SSE/(n-k-1)} = \frac{\frac{\overbrace{SSR_{H_A} - SSR_{H_0}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}} == \frac{\frac{\overbrace{SSE_{H_0} - SSE_{H_A}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}}$$

- SSR_{H_A} is the regression sum of square from unrestricted model with k degrees of freedom (2)
- SSR_{H_0} is the regression sum of squres from the restricted model wiht k_0 degrees of freedom (0) - it's the number of regressors in restricted model

Testing for the significance of the regression

```r
linearHypothesis(lm_model, c("Occupancy=0", "EDAD=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Occupancy = 0
## EDAD = 0
##
## Model 1: restricted model
## Model 2: Duration ~ Occupancy + EDAD
##
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   4997 62505371
## 2   4995 48947909  2  13557462 691.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a coefficient

- We can use the above logic to test how much more we can explain by including one more coeffcient
- Suppose we want to compare a regression model with only occupancy vs both occupancy and age
- $H_0 : y = \beta_0 + \beta_1 Occupancy + u$ restricted model
- $H_A : y = \beta_0 + \beta_1 Occupancy + \beta_2 Age + u$ unrestricted model

$$F_2 = \frac{\frac{\overbrace{SSR_{H_A} - SSR_{H_0}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}} = \frac{\frac{\overbrace{SSE_{H_0} - SSE_{H_A}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}} \sim \underbrace{F_{k-k_0, n-k-1}}_{\text{Dist under } H_0}$$

- In our case $k = 2$ and $k_0 = 1$, so the null distribution is $F_{1, n-3}

Adding a coefficient

```
## Analysis of Variance Table
##
## Response: Duration
##             Df   Sum Sq  Mean Sq   F value    Pr(>F)
## Occupancy    1 13465872 13465872 1374.1553 < 2.2e-16 ***
## EDAD         1    91590    91590    9.3465  0.002246 **
## Residuals 4995 48947909     9799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Sequential testing:**
- Occupancy - F_1 is the additional effect of including Occupancy to a model without any regressors
  - $H_0 : y = \beta_0 + u$ restricted model
  - $H_A : y = \beta_0 + \beta_1 Occupancy + u$ unrestricted model

- EDAD - F_2 is the additional effect of including Age once we already have Occupancy in the model
  - $H_0 : y = \beta_0 + \beta_1 Occupancy + u$ restricted model
  - $H_A : y = \beta_0 + \beta_1 Occupancy + \beta_2 Age + u$ unrestricted model

Adding a coefficient

- F_k (last coef) is equivalent to t_k^2 in our full model
- But F_1 is not equivalent to t_1^2 in our full model

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 23.2342216 2.48416121  9.352944  1.255500e-20
## Occupancy    3.7035443 0.10090081 36.704802 2.396768e-261
## EDAD         0.2062604 0.06746688  3.057209  2.245903e-03
```

```
## Analysis of Variance Table
##
## Response: Duration
##             Df   Sum Sq  Mean Sq   F value    Pr(>F)
## Occupancy    1 13465872 13465872 1374.1553 < 2.2e-16 ***
## EDAD         1    91590    91590    9.3465  0.002246 **
## Residuals 4995 48947909     9799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a coefficient

- Why reordering variables changes F_{stats}?

```
## Analysis of Variance Table
##
## Response: Duration
##             Df   Sum Sq  Mean Sq   F value    Pr(>F)
## Occupancy    1 13465872 13465872 1374.1553 < 2.2e-16 ***
## EDAD         1    91590    91590    9.3465  0.002246 **
## Residuals 4995 48947909     9799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: Duration
##             Df   Sum Sq  Mean Sq  F value    Pr(>F)
## EDAD         1   355320   355320   36.259 1.851e-09 ***
## Occupancy    1 13202142 13202142 1347.243 < 2.2e-16 ***
## Residuals 4995 48947909     9799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Because it changes which regressors we already have in the model
- Do squares always add up to the same thing?

Testing multiple coefficients

Suppose we have a model with three predictors

$$y = \beta_0 + \beta_1 Occupancy + \beta_2 Age + \beta_3 Male + u$$

We can test for a subset of predictors, for example if Age and Sex matter

- $H_0 : \beta_2 = \beta_3 = 0 \rightarrow y = \beta_0 + \beta_1 Occupancy + u$
- $H_A : \beta_2 \neq 0$ or $\beta_3 \neq 0 \rightarrow y = \beta_0 + \beta_1 Occupancy + \beta_2 Age + \beta_3 Male + u$

$$F_{test} = \frac{\frac{\overbrace{SSR_{H_A} - SSR_{H_0}}^{\text{Extra Sum of Squares}}}{3-1}}{\frac{SSE_{H_A}}{n-3-1}} = \frac{\frac{\overbrace{SSE_{H_0} - SSE_{H_A}}^{\text{Extra Sum of Squares}}}{3-1}}{\frac{SSE_{H_A}}{n-3-1}} \sim F_{2,n-4}$$

```
## Linear hypothesis test
##
## Hypothesis:
## EDAD = 0
## SEXOMASCULINO = 0
##
## Model 1: restricted model
## Model 2: Duration ~ Occupancy + EDAD + SEXO
##
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1   4996 49039499
## 2   4994 48728235  2    311264 15.95 1.244e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing for multiple coefficients

A cool thing about the regression is that we can test relationships between the coefficients:

For example:

- Is the impact of additional year of experience the same as impact of additional year of work experience in a regression:

$$income_i = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + u_i$$

- That corresponds to null hypothesis $H_0 : \beta_1 = \beta_2$ or $H_0 : \beta_1 - \beta_2 = 0$

Another Example:

- Suppose that employees can go through a sales training, and/or get a better office (these are binary variables). We want to evaluate impact of these measures on their sales:

$$Sales_i = \beta_0 + \beta_1 \text{training}_i + \beta_2 \text{office}_i + u_i$$

- We wonder if giving an employee all three would increase sales by more than 100: $H_A : \beta_1 + \beta_2 > 100$

Relationships between coefficients

Suppose we have a model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

- We want to test if the difference between impact of x_1 and x_2 is equal to c

Hypothesis

- $H_0 : \beta_1 - \beta_2 = c$
- $H_A : \beta_1 - \beta_2 \neq c$
  - Special case: $c = 0$ => testing equality $\beta_1 = \beta_2$ **Test statistic and its distribution under the null**

$$T_{test} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - c}{SE(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - c}{\sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2)}} \sim t_{n-k-1}$$

- Calculate p-value as $2P(t_{n-k-1} > |T_{test}|)$

Relationships between coefficients

- In the same way we can test whether one coefficient is larger than another by some amount

Hypothesis

- $H_0 : \beta_1 - \beta_2 = c$
- $H_A : \beta_1 - \beta_2 > c$
  - Special case: $c = 0$ => testing inequality $\beta_1 > \beta_2$

Test statistic and its distribution under the null

$$T_{test} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - c}{SE(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - c}{\sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2)}} \sim t_{n-k-1}$$

- Calculate p-value as $P(t_{n-k-1} > T_{test})$
- If alternative is $H_A : \beta_1 - \beta_2 < c$, then $P(t_{n-k-1} < T_{test})$

Example

- Test if one more person at the hospital has larger effect than being one year older

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD + SEXO, data = Sample_urg)
##
## Coefficients:
##   (Intercept)      Occupancy           EDAD  SEXOMASCULINO
##       18.5463         3.6803         0.2047        13.8988

##                (Intercept)     Occupancy          EDAD SEXOMASCULINO
## (Intercept)     7.12075807 -0.0481948400 -0.1296097547 -2.8941142167
## Occupancy      -0.04819484  0.0101612131 -0.0005422531 -0.0143206543
## EDAD           -0.12960975 -0.0005422531  0.0045323658 -0.0009536548
## SEXOMASCULINO -2.89411422 -0.0143206543 -0.0009536548  8.5804013484
```

Example

- Hypotheses:
  - $H_0 : \beta_O = \beta_A$
  - $H_A : \beta_O > \beta_A$
- Calculate the test statistic

$$T_{test} = \frac{\beta_O - \beta_A}{\sqrt{var(\hat{\beta_O}) + var(\hat{\beta_A}) - 2cov(\hat{\beta_O}, \hat{\beta_A})}} = \frac{3.6803 - 0.2047}{\sqrt{0.01 + 0.0045 - 2 * (-0.00054)}} = 27.84$$

- Calculate p-value

$$P - value = P(t_{n-k-1} > T_{test}) = P(t_{4994} > 27.84) \approx 0$$

Conclusion

- we reject that impact of one more year is smaller or equal to the impact of one more person

Sum of coefficients

Suppose we have a model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

- We want to test if the sum of impact of x_1 and x_2 is equal to c

Hypothesis

- $H_0 : \beta_1 + \beta_2 = c$
- $H_A : \beta_1 + \beta_2 \neq c$

Test statistic and its distribution under the null

$$T_{test} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - c}{SE(\hat{\beta}_1 + \hat{\beta}_2)} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - c}{\sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) + 2cov(\hat{\beta}_1, \hat{\beta}_2)}} \sim t_{n-k-1}$$

- Calculate p-value as $P(t_{n-k-1} > T_{test})$
- If $H_A : \beta_1 + \beta_2 < c$, then $P(t_{n-k-1} < T_{test})$
- If $H_A : \beta_1 + \beta_2 > c$, then $P(t_{n-k-1} > T_{test})$

Example

- Test if the total impact of increasing occupancy by one person and being male is larger than 17

```
##
## Call:
## lm(formula = Duration ~ Occupancy + EDAD + SEXO, data = Sample_urg)
##
## Coefficients:
##   (Intercept)      Occupancy           EDAD  SEXOMASCULINO
##       18.5463         3.6803         0.2047        13.8988

##                (Intercept)     Occupancy          EDAD SEXOMASCULINO
## (Intercept)     7.12075807 -0.0481948400 -0.1296097547 -2.8941142167
## Occupancy      -0.04819484  0.0101612131 -0.0005422531 -0.0143206543
## EDAD           -0.12960975 -0.0005422531  0.0045323658 -0.0009536548
## SEXOMASCULINO -2.89411422 -0.0143206543 -0.0009536548  8.5804013484
```

Standarized Coefficients

- Coefficients depend on the units of measurement of the x
- Since x can have different units or magnitudes, we can't directly compare them

Example:

$$\text{ecobici trips}_i = \beta_0 + \beta_1 \text{temperature}_i + \beta_2 \text{polution}_i + u_i$$

- It doesn't make sense to compare β_1 to β_2 to see what has bigger effect
- These variables have very different magnitudes
  - Increasing temperature by one unit (1 degree celcius) is different than increasing polution by one unit (1 µg/m3)
- To make them directly comparable, we want to make them unitless (standarized)
- Does increasing temperature by one standard deviation has the same effect as inreasing polution by one standard deviation?

Standarized coeffcients

Basically, we standardize all the variables and run the regression:

$$\frac{y_i - \bar{y}}{s_y} = \gamma_1 \frac{x_{i1} - \bar{x}_1}{s_{x_1}} + \gamma_2 \frac{x_{i2} - \bar{x}_2}{s_{x_2}} + \ldots + \gamma_k \frac{x_{ik} - \bar{x}_k}{s_{x_k}} + u_i$$

So then γ_k measures the impact of one standard deviation increase of x_k on standard deviation in y

But there is a short cut to calculate these standard coefficients

$$\gamma_k = \beta_k \frac{s_{x_k}}{s_y}$$

Example

Urgent Care duration example:

- $s_y = 111.82$
- $s_{Age} = 20.82$
- $s_{Occupancy} = 13.921$

We calculated that $\hat{\beta}_{Age} = 0.206$ and $\hat{\beta}_{Occupancy} = 3.703$

Standardized coefficients

$$\hat{\gamma}_{Age} = \hat{\beta}_{Age} \frac{s_{Age}}{s_y} = 0.206 \frac{20.82}{111.82} = 0.0383$$

$$\hat{\gamma}_{Occupancy} = \hat{\beta}_{Occupancy} \frac{s_{Occupancy}}{s_y} = 3.703 \frac{13.921}{111.82} = 0.461$$

- Changing age by one standard deviation increases duration by 3.8% of a standard deviation
- Changing occupancy one standard deviation increases duration by 46% of a standard deviation