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)

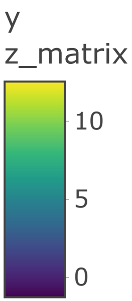
$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+u_i$$

- 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

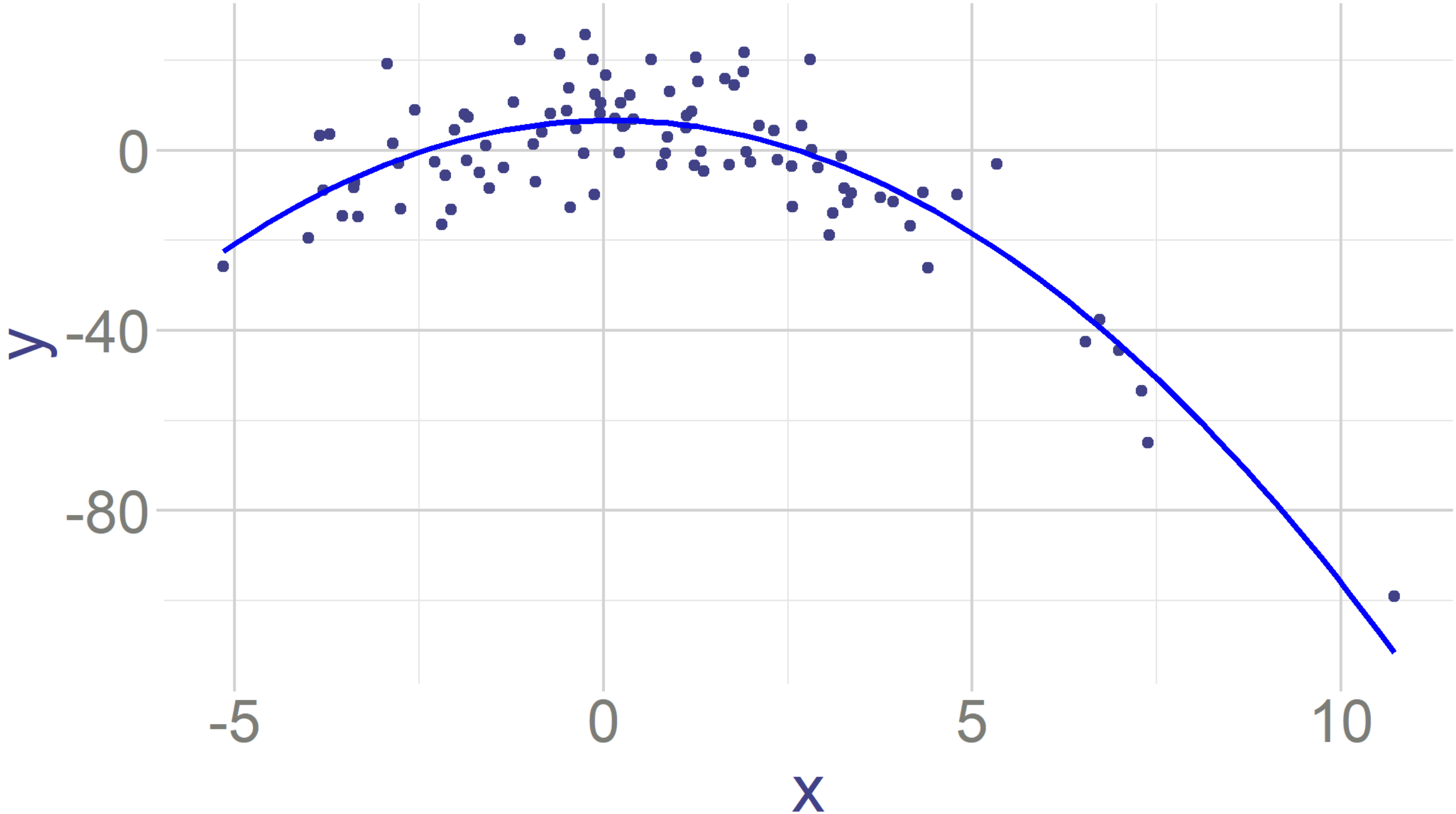
100 observations simulated from an a regression line:

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



Multiple linear regression

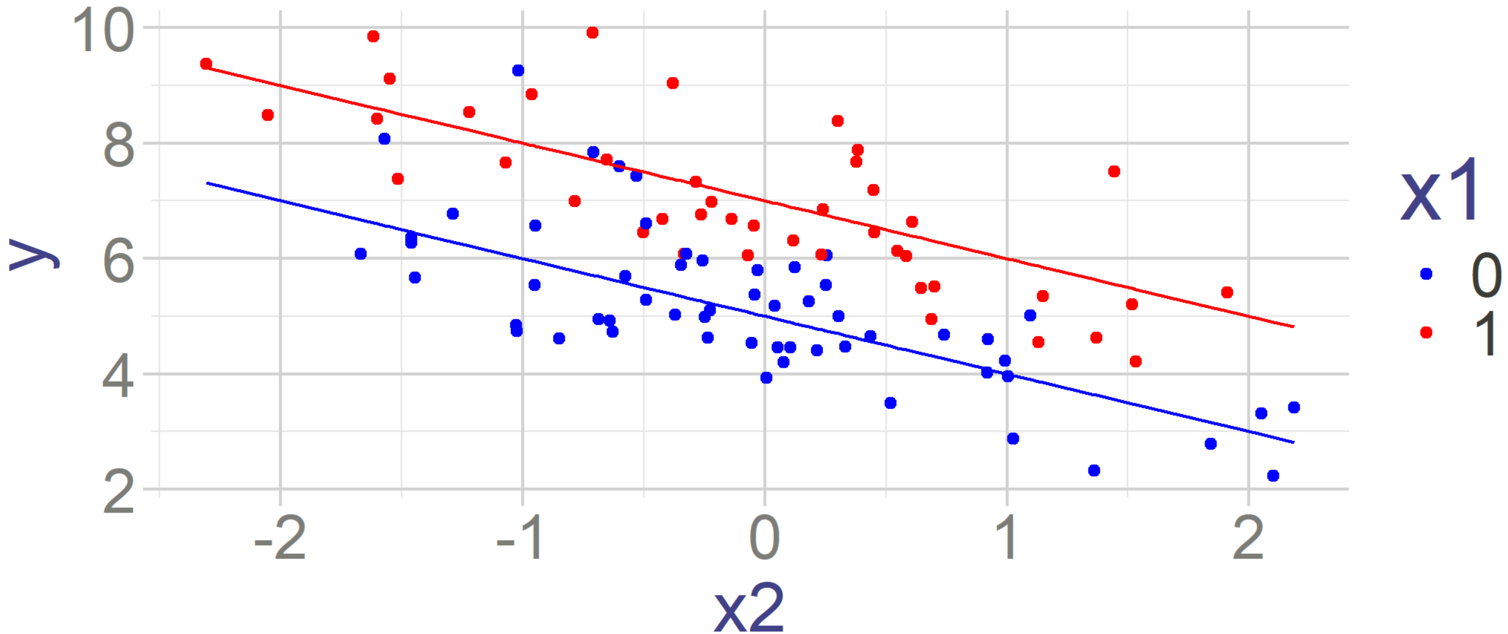
100 observations simulated from an a regression line: $$y_i=5+2x_{i}-1x_i^2+u_i$$



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:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_kx_{ik}+u_i$$

- 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:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_k,x_{ik}+u_i$$ In vector form is:

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

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

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.

$$ \begin{array}{cc} \text{Full Rank Matrix:} & \text{Matrix Not of Full Rank:} \\ \left[\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right] & \left[\begin{array}{ccc} 1 & 2 & 4 \\ 4 & 5 & 10 \\ 7 & 8 & 16 \end{array}\right] \end{array} $$

Multiple linear regression

Goal:

- Estimate the vector of parameters $\mathbf{\beta}$

Procedure

- Find

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

- Which minimizes the squared errors in the problem:

$$y_i=b_0+b_1x_{i1}+b_2x_{i2}+...+b_kx_{ik}+e_i$$

- That is minimize

$$SSE=\sum_ie_i^2=\sum_i(y_i-\hat{y}_i)^2=\mathbf{e}'\mathbf{e}=(\mathbf{y-\hat{y}})'(\mathbf{y-\hat{y}})=(\mathbf{y-Xb})'(\mathbf{y-Xb})$$

Multiple linear regression

- We can do it with scalars

$$\begin{align*} \frac{\partial SSE}{\partial \hat{\beta}_0} & = -2\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1}+...+\hat{\beta}_k x_{ik})\right)=0 \\ \frac{\partial SSE}{\partial \hat{\beta}_1} & = -2\sum_{i=1}^{n} x_{i1} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1}+...+\hat{\beta}_k x_{ik})\right)=0 \\ \vdots \\ \frac{\partial SSE}{\partial \hat{\beta}_k} & = -2\sum_{i=1}^{n} x_{ik} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1}+...+\hat{\beta}_k x_{ik})\right)=0 \\ \end{align*}$$

- 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: $$SSE(b)=(\mathbf{y-Xb})'(\mathbf{y-Xb})=\mathbf{y'}\mathbf{y-2b'X'y}+\mathbf{b'X'Xb}$$
- Then minimize it with respect to \(\mathbf{b}\)

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

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

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

Multiple linear regression

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

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

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

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

Practice theory

Lista 5.1 Q9

Special Case: k=1

What if we have just one x?

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

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

which gives:

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

Predictions

To make predictions based on the estimated regressors we use:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+...+\hat{\beta}_kx_{ik}$$

Or in the vector form:

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

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

Residuals

To get residuals, we calculate:

$$e_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+...+\hat{\beta}_kx_{ik}$$

Or in the vector form: $$\mathbf{e}=\mathbf{y-\hat{y}}=y-\mathbf{X\hat{\beta}}=\mathbf{y}-\mathbf{X\mathbf{(X'X)}^{-1}\mathbf{X'y}}=\mathbf{(I-H)y}$$

Summary

- We are trying to find \(\beta\)s which minimize the prediction error
- It turns out that we get minimal errors when we set \(\beta\) to be: \(\hat{\beta}=(X'X)^{-1}X'y\)

Practice Theory

Lista 5.1 Q7

Example with numbers

Similar to Lista 5.1 Q8. What's the impact of hours studied and hours slept on the exam score?

$$\small \begin{align*} \text{Dataset:} \\ &\begin{array}{|c|c|c|c|} \hline \text{Student} & \text{Hours Studied (}x_1\text{)} & \text{Hours Slept (}x_2\text{)} & \text{Exam Score (}y\text{)} \\ \hline 1 & 3 & 8 & 80 \\ 2 & 4 & 7 & 85 \\ 3 & 6 & 6 & 92 \\ 4 & 5 & 7 & 88 \\ \hline \end{array} \\ \text{X matrix}: \\ & X = \begin{bmatrix} 1 & 3 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 6 \\ 1 & 5 & 7 \\ \end{bmatrix} \\ \text{Response Vector } (y): \\ & y = \begin{bmatrix} 80 \\ 85 \\ 92 \\ 88 \\ \end{bmatrix} \end{align*}$$

What these matrices mean and do?!

- Now we have more, and we want to know SEPARATE impact of each.
- What would happen if we change just that one thing (hours of sleep) and keep everything else constant (hours studied)?
- But how do we keep other things constant? Hours slept and hours studied could be correlated
- So when we compare people with different hours of sleep, there is risk we also compare people with different hours studied
- How do we know that the change we see is coming from hours of sleep, not hours of studied?
- Cross terms in the matrix discount this correlation
- Intuitively, this makes variables "uncorrelated" and allows to get their own impacts.

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:

$$\beta=\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \end{bmatrix} = \underbrace{\begin{bmatrix} 474.75 & -30 & -48.5 \\ -30 & 2 & 3 \\ -48.5 & 3 & 5 \\ \end{bmatrix}}_{(X'X)^{-1}} \underbrace{\begin{bmatrix} 345 \\ 1572 \\ 2403 \\ \end{bmatrix}}_{X'y}= \begin{bmatrix} 83.25 \\ 3 \\ -1.5 \\ \end{bmatrix}$$

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
```

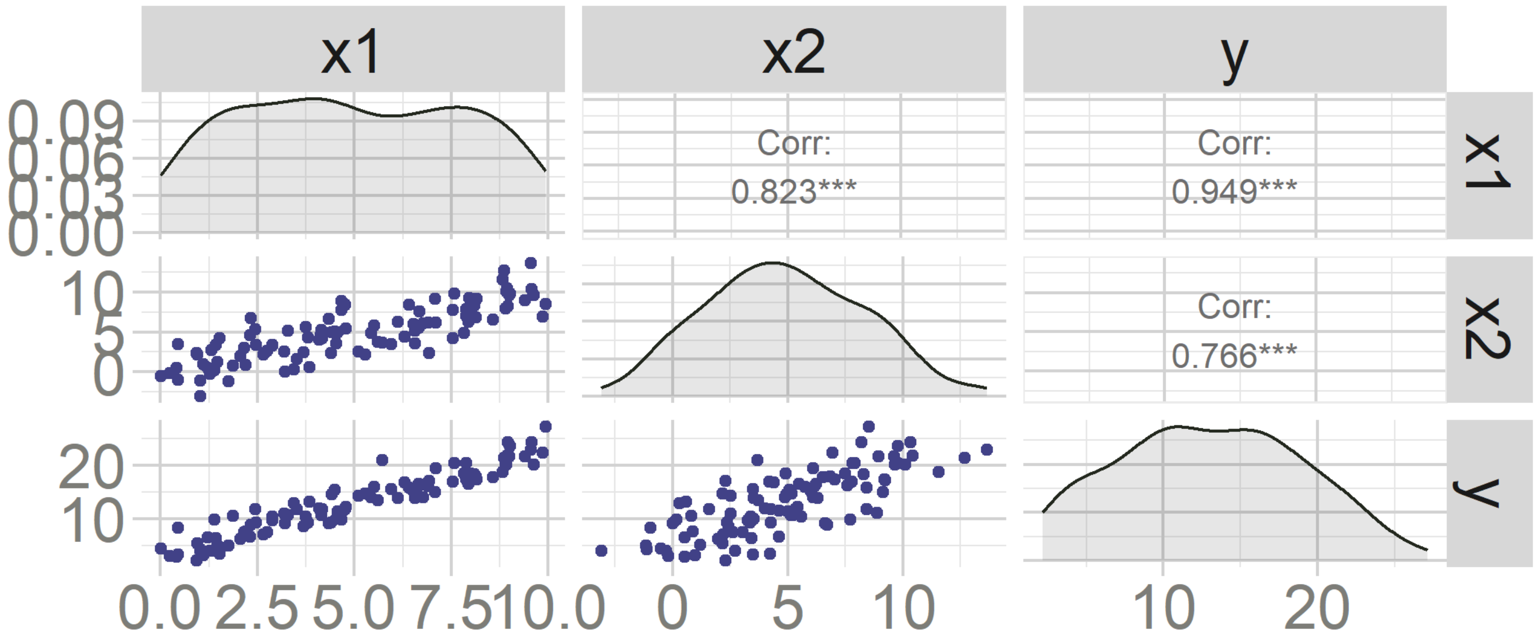
Practice

Predict how long a patient will stay if there 10 other patients in the hospital and the patient is 50 years old.

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**:

\[\small \begin{align*} E(\hat{\beta}) & =E(\mathbf{(X'X)^{-1}X'y})=E(\mathbf{(X'X)^{-1}X'(X\beta+u)})) \\ & = E(\mathbf{(X'X)^{-1}X'(X\beta+u)}))=E(\mathbf{(X'X)^{-1}X'X\beta})+E(\mathbf{(X'X)^{-1}X'u}) \\ & = \beta+\mathbf{(X'X)^{-1}X'E(u)}=\beta+0+\beta \end{align*} \]

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

- **Variance**

$$\small Var(\hat{\beta})=Cov(\hat{\beta})=\underbrace{\begin{bmatrix} var(\hat{\beta_0}) & cov(\hat{\beta_0}, \hat{\beta_1}) & ... & cov(\hat{\beta_0}, \hat{\beta_k}) \\ cov(\hat{\beta_1}, \hat{\beta_0}) & var(\hat{\beta_1}) & ... & cov(\hat{\beta_1}, \hat{\beta_k}) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\hat{\beta_k}, \hat{\beta_0}) & cov(\hat{\beta_k}, \hat{\beta_1}) & ... & var(\hat{\beta_k}) & \end{bmatrix}}_{\substack{ \\ (k+1) \times (k+1)}}$$

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

Example

10. [5 **puntos**] An estimation of a linear regression model of the form:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_1 x_{i2}$$

with 30 observations reported the following results:

$$\hat{\beta} = (0.4510, 0.34703, 0.41490)^T \quad ; \quad \operatorname{cov}(\hat{\beta}) = \begin{pmatrix} 0.07784 & -0.00808 & -0.00738 \\ & 0.0175 & -0.00014 \\ & & 0.00164 \end{pmatrix}$$

with $s^2 = 0.02389$ and $R^2 = 0.8752$. If you intent to test the hypothesis

$$\mathrm{H}_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad \mathrm{H}_0 : \beta_1 \neq \beta_2$$

then, the corresponding observed test statistic is approximately:

d) none of the above

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

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

- Where two last inequalies come from?
  - From \(\sigma^2\) being constant and errors having zero covariance.

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 \(k+1\) row and \(k+1\) column of \((X'X)^{-1}\) matrix. First one is intercept!

Example:

$$\text{cov}(\hat{\beta})= \begin{bmatrix} 0.07784 & -0.00808 & -0.00738 \\ -0.00808 & 0.0175 & -0.00014 \\ -0.00738 & -0.00014 & 0.00164 \end{bmatrix}=$$ And the residuals: $$\text{cov}(\hat{\beta})=\underbrace{0.02389}_{\sigma^2} \times \underbrace{\begin{bmatrix} 3.2583 & -0.3382 & -0.3089 \\ -0.3382 & 0.7325 & -0.0059 \\ -0.3089 & -0.0059 & 0.0686 \end{bmatrix}}_{(X'X)^{-1}}$$

\(var(\hat{\beta}_1)=0.02389 \times 0.7325=0.0175\) \
(cov(\hat{\beta}_1,\hat{\beta}_0)=0.02389 \times -0.3382=-0.00808\)

Variance

- Where the hell do we get the \(\sigma^2\) 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 \(p=k+1\)
- k is number of variables
- +1 because of intercept
- This is called mean squared error as well

The easiest way to compute this sum is:

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

Gauss Markov Theorem (Again)

Assumptions

- $E(u_i)=0$
- $var(u_i)=\sigma^2$
- $cov(u_i,u_j)=0$
- 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 $b_1=c_{11}y_1+c_{12}y_2+...+c_{13}y_3$
  - Example $b_1=\frac{1}{n}y_1+...+\frac{1}{n}y_n$

- 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)!

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

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

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?

$$y_i=\beta_0+\beta_1D_{i1}+\beta_2D_{i2}+...+\beta_{m-1}D_{im-1}+u_i$$

- In our Example:

$$y_i=\beta_0+\beta_1D_{i,Hybrid}+\beta_2D_{i,Remote}+u_i$$

- Because otherwise X would not be full rank!

$$\begin{array}{cc} \text{Full Rank Matrix:} & \text{Matrix Not of Full Rank:} \\ \left[\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{array}\right] & \left[\begin{array}{cccc} 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 1 & 0 \end{array}\right] \end{array}$$

- 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

```r
# 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)
```

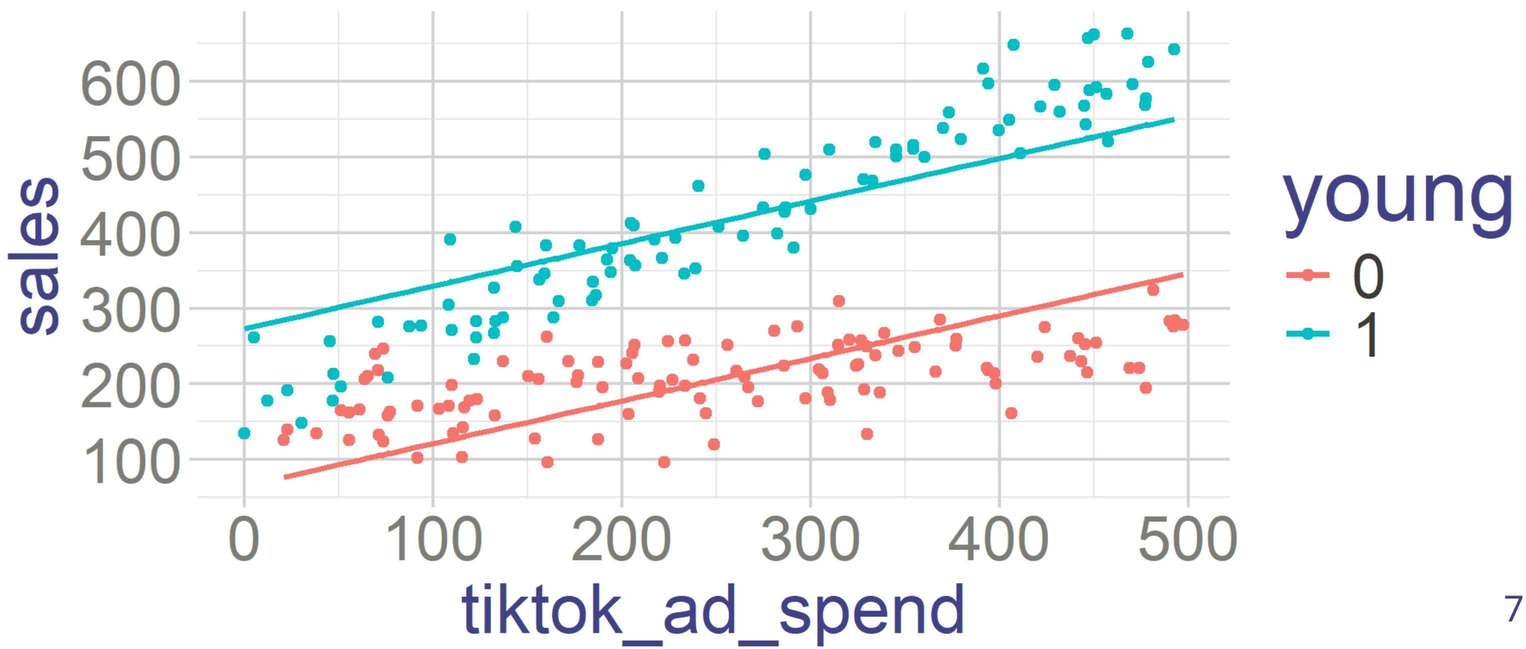
```
##
## Call:
## lm(formula = Income ~ Education, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25868 -10865  -1413  10204  28280
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          70342       3125  22.509  < 2e-16 ***
## EducationPhD         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:

$$\text{Purchases}_i=\beta_0+\beta_1\text{TikTok ads}_i+\beta_2\text{Age<25}_i+u_i $$

- Where tiktok ads is how much the company spends on tiktok ads
- Age<25 is a dummy variable that is 1 if the person is younger than 25
- Purchases is how much person i spent on clothes from a company.

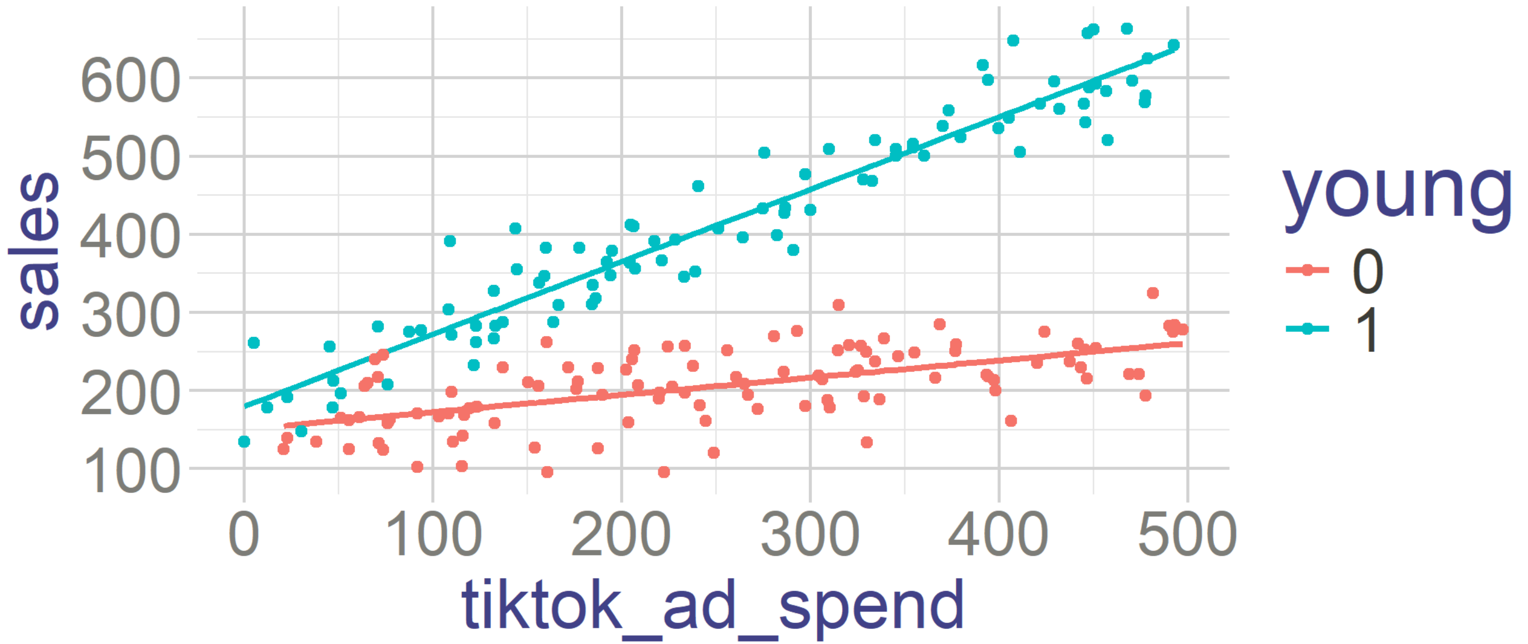


Interactions

```
##
## Call:
## lm(formula = sales ~ tiktok_ad_spend + young, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -141.883  -50.858    0.821   47.398  145.724
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      64.73544   10.12217   6.395 1.14e-09 ***
## tiktok_ad_spend   0.56359    0.03244  17.373  < 2e-16 ***
## young1          208.29233    8.89550  23.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.69 on 197 degrees of freedom
## Multiple R-squared:  0.8149,    Adjusted R-squared:  0.813
## F-statistic: 433.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

Interactions

- Do ads on tik tok affect in the same way older and younger people?
- In other words: one additional dollar on tik-tok ads increases purchases more if you are young?
- We want allow the coefficient on ads to differ by age group.



Interactions

- Run the regression:

$$\text{Purchases}_i=\beta_0+\beta_1\text{TikTok ads}_i+\beta_2\text{Age<25}_i+\beta_3\text{TikTok ads}_i*\text{Age<25}_i+u_i $$

- What's the coefficient on ads when you are older $Age<25_i=0$?

$$\begin{align*} & \text{Purchases}_i=\beta_0+\beta_1\text{TikTok ads}_i+\beta_2*0+\beta_3\text{TikTok ads}_i*0 +u_i \\ & \text{Purchases}_i=\beta_0+\beta_1\text{TikTok ads}_i+u_i \end{align*}$$

- What's the coefficient on ads when you are younger $Age<25_i=1$?

$$\begin{align*} & \text{Purchases}_i=\beta_0+\beta_1\text{TikTok ads}_i+\beta_2*1+\beta_3\text{TikTok ads}_i*1 +u_i \\ & \text{Purchases}_i=\beta_0+\beta_2+(\beta_1+\beta_3)\text{TikTok ads}_i+u_i \end{align*}$$

We can estimate β_3 and it will tell us by how much bigger is the coefficient on ads for young compared to the coefficient on ads for old.

- β_1 is the slope for old

```
##
## Call:
## lm(formula = sales ~ tiktok_ad_spend * young, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.867  -25.348   -1.574   24.928  110.078
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            150.44290    8.13287  18.498  < 2e-16 ***
## tiktok_ad_spend          0.22153    0.02865   7.733 5.38e-13 ***
## young1                  29.35544   11.85879   2.475   0.0142 *
## tiktok_ad_spend:young1   0.70583    0.04115  17.152  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.74 on 196 degrees of freedom
## Multiple R-squared:  0.926,	Adjusted R-squared:  0.9249
## F-statistic: 817.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

- One additional dollar on ads increases purchases among old by 0.222 dollars
- One additional dollar on ads increases purchases among young by 0.222+0.706=0.928 dollars

Interactions

- More generally, we can rewrite a regression:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i1}*x_{i2}+u_i$$

- As

$$y_i=\beta_0+(\beta_1+\beta_3x_{i2})x_{i1}+\beta_2x_{i2}+u_i$$- β_3 answers the following question:

- If I increase x_{i2} by one, by how much the coefficient on x_{i1} changes?
- If x_{i2} is categorical: by how much the effect of x_{i1} changes when $x_{i2}=1$ compared to when $x_{i2}=0$?
- Suppose y is mortality, x_1 is temperature, x_2 is age
  - What would you expect β_3 to be?

Interactions

Exam example

2. [**5 puntos**] You are conducting a study to investigate the effects of a new diet and an exercise programme on weight loss. Consider 100 subjects who recorded *weight loss* (w) measured in kg at the end of three months in the program. Some of the subjects participating in the study were asked to implement the new diet. All individuals exercised regularly during the study period and recorded the number of weekly exercise hours (h). A following linear model is fitted based on the collected data:

$$\hat{w} = -1.4 + 7.2d + 1.5h + 3.4dh$$

where d is a dummy variable indicating whether the individual followed the new diet. Then, without making any inference, it can be said that:

a) on average, people who neither exercise nor follow the new diet have a negative weight, which does not make sense.

b) on average, people who follow the new diet but do not exercise have an increase in weight greater than 5 kg.

c) on average, for people who do not follow the new diet, exercising an hour per week decreases their weight by aproximately 4.5 kg.

d) on average, for people who follow the new diet, exercising an hour per week decreases their weight by 10.7 kg.

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?

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

- 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?
- Exercise: Lista 5.1, 10 a) and b)

Goodness of fit

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

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

- 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

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

- More parameters -> $\downarrow(n-p)\rightarrow\uparrow\sum(y_i-\hat{y}_i)^2/(n-p)\rightarrow\downarrow R_{Adj}^2$
- 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
## -773.18  -26.60  -17.26   -0.47 1253.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.09257    2.49595   9.252  < 2e-16 ***
## Occupancy    3.70288    0.10091  36.693  < 2e-16 ***
## EDAD         0.20588    0.06747   3.051  0.00229 **
## Random_var   0.02755    0.04680   0.589  0.55616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99 on 4994 degrees of freedom
## Multiple R-squared:  0.217,    Adjusted R-squared:  0.2165
## F-statistic: 461.2 on 3 and 4994 DF,  p-value: < 2.2e-16
```

- Adding random variable increased R^2 but decreased 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
```

Example

10. [5 **puntos**] An estimation of a linear regression model of the form:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_1 x_{i2}$$

with 30 observations reported the following results:

$$\hat{\beta} = (0.4510, 0.34703, 0.41490)^T \quad ; \quad \text{cov}(\hat{\beta}) = \begin{pmatrix} 0.07784 & -0.00808 & -0.00738 \\ & 0.0175 & -0.00014 \\ & & 0.00164 \end{pmatrix}$$

with $s^2 = 0.02389$ and $R^2 = 0.8752$. If you intent to test the hypothesis

$$\text{H}_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad \text{H}_0 : \beta_1 \neq \beta_2$$

then, the corresponding observed test statistic is approximately:

d) none of the above

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:

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

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:

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

- Note that the CI does not contain 0
- What does it imply for hypothesis testing with $H_0: \beta_{age}=0$?
- Exercise: lista 5.1 4a) and b)

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:

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

So it's distribution is:

\(\hat{y}_0 \sim N(\mathbf{x_0}'\beta, \sqrt{\sigma^2\mathbf{x_0}'(\mathbf{X}'\math

Hence:

\(CI_{1-\alpha}=\{\hat{y}_0\pm t_{n-2,\frac{\alpha}

Example

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?
- $\hat{\beta}=\{\hat{\beta_0}, \hat{\beta}_{occupancy}, \hat{\beta}_{age}\}=\{23.236, 3.7, 0.2\}$
- $\hat{\sigma}=98.97$
- $(X'X)^{-1}=$

```
##               (Intercept)     Occupancy          EDAD
## (Intercept)  6.297395e-04 -5.434372e-06 -1.331625e-05
## Occupancy   -5.434372e-06  1.038940e-06 -5.573690e-08
## EDAD        -1.331625e-05 -5.573690e-08  4.644967e-07
```

Example

- Prediction: $\hat{y_0}=23.236*1+3.7*10+0.2*52=70.636$
- $\sqrt{\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}=\sqrt{[1, 10, 52](\mathbf{X}'\mathbf{X})^{-1}[1, 10, 52]'}=0.021$
- Standard Deviation: $SE(\hat{y_0})=\sqrt{\sigma^2\mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}=$
$$CI_{95}=\{70.636 \pm 1.96*2.07837\} \approx \{67, 75\}$$

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

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

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 \(y_i\)?
- \(\small H_0: \beta_1=\beta_2=...\beta_k=0\)
- \(\small H_A: \beta_j \neq 0\) 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

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

Testing for the significance of the regression

Alternative way to think about it:

- \(\small H_0: y=\beta_0+u\) restricted model (\(x_1\) and \(x_2\) don't matter)
- \(\small H_A: y=\beta_0+\beta_1x_{1}+\beta_2x_{2}+u\)

If \(H_A\) is true, unrestricted model should explain more of \(y\)

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

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

Testing for the significance of the regression

```
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
```

Example

7. [**5 puntos**] Suppose your local gym is offering three types of subscriptions: a limited subscription (you can only attend 4 classes a week), a full subscription (you can attend unlimited number of classes), and a luxury subscription (unlimited classes, plus genetic tests, nutritional and sleep coaching). Your gym measures how much muscles a member gained after 3 months since the start of the subscription (m_i measured in kilograms). To evaluate these programs, you regress m_i on the category of membership represented using dummies or indicator variables and a term for the intercept. You perform an ANOVA to assess the significance of the regression and obtained the following ANOVA table. Unfortunately, your dog ate parts of the printout of the table and only this information has been preserved:

Tabla 2: Analysis of Variance Table

Source	Df	Sum Squares
Regression		500
Residuals		
Total	256	1200

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
- $\small H_0: y=\beta_0+\beta_1Occupancy+u$ restricted model
- $\small H_A: y=\beta_0+\beta_1Occupancy+\beta_2Age+u$ unrestricted model

$$\small F_{2}=\frac{\frac{\overbrace{SSR_{H_A}-SSR_{H_0}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}}=\frac{\frac{\overbrace{SSE_{H_0}-SSE_{H_A}}^{\text{Extra Sum of Squares}}}{k-(k_0)}}{\frac{SSE_{H_A}}{n-k-1}} \sim \underbrace{F_{k-k_0, n-k-1}}_{\text{Dist under } H_0}$$

- 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
  - \(\small H_0: y=\beta_0+u\) restricted model
  - \(\small H_A: y=\beta_0+\beta_1Occupancy+u\) unrestricted model

- EDAD - \(F_2\) is the additional effect of including Age once we already have Occupancy in the model
  - \(\small H_0: y=\beta_0+\beta_1Occupancy+u\) restricted model
  - \(\small H_A: y=\beta_0+\beta_1Occupancy+\beta_2Age+u\) 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

$$\small y=\beta_0+\beta_1Occupancy+\beta_2Age+\beta_3Male+u$$ We can test for a subset of predictors, for example if Age and Sex matter

- \(H_0:\beta_2=\beta_3=0 \rightarrow y=\beta_0+\beta_1Occupancy+u\)
- \(H_A:\beta_2\neq 0 \text{ or } \beta_3\neq0 \\ \rightarrow y=\beta_0+\beta_1Occupancy+\beta_2Age+\beta_3Male+u\)

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

```
## 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 education 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 both of them would increase sales by more than 100: $H_A: \beta_1+\beta_2>100$

Relationships between coefficients

Suppose we have a model:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+...\beta_kx_k+u$$

- 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

$$\small T_{test}=\frac{\beta_{O}-\beta_{A}}
{\sqrt{var(\hat{\beta_O})+var(\hat{\beta_A})-2cov(\hat{\beta_O},\hat{\beta_A})}}=\frac{3.68
0.2047}{\sqrt{0.01+0.0045-2*(-0.00054)}}=27.84$$

- Calculate p-value

$P-value=P(t_{n-k-1}>T_{test})=P(t_{4994}>27.84) \approx0$

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:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+...\beta_kx_k+u$$

- 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:

$$\frac{y_i-\bar{y}}{s_y}=\gamma_1\frac{x_{i1}-\bar{x}_{1}}{s_{x_1}}+\gamma_2\frac{x_{i2}-\bar{x}_2}{s_{x_2}}+...+\gamma_k\frac{x_{ik}-\bar{x}_k}{s_{x_k}}+u_i$$ So then \(\gamma_k\) 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

$$\begin{align*} & \hat{\gamma}_{Age} =\hat{\beta}_{Age}\frac{s_{Age}}{s_{y}}=0.206\frac{20.82}{111.82}=0.0383 \\ &\hat{\gamma}_{Occupancy} =\hat{\beta}_{Occupancy}\frac{s_{Occupancy}}{s_y}=3.703\frac{13.921}{111.82}=0.461 \end{align*}$$

8. **[20 puntos]** A publisher wants to predict book sales during the first month after its release (y), measured in dollars. For this purpose, a linear model is fitted using historical information corresponding to 65 previous launches and the sales they scored during the first month. The information includes data on the number of promotional events (x_1), expenditure on digital *marketing* in dollars (x_2), an indicator variable whether this is the first book of the author (x_3) or if the author is a debutant, where $x_3 = 1$ if affirmative and zero otherwise, and lastly the number of pages in the book (x_4).

From the estimation, the regression equation was recovered as

$$\hat{y} = 2,500 + 300x_1 + 7x_2 + 1500x_3 - 10x_4$$

as well as the following statistics

$$s_Y = 500.00 \ ; \ s_{X_1} = 300.00 \ ; \ s_{X_2} = 1,000.00 \ ; \ s_{X_3} = 0.50 \ ; \ s_{X_4} = 150.00$$

where $s_{.}$ denotes the standard deviations for the variable of interest as well as for the regressors, and

$$s^2 = 500 \quad ; \qquad \text{cov}(\hat{\beta}) = \begin{pmatrix} 5000.00 & -10.00 & 0.00 & 5.00 & 0.00 \\ & 2000.00 & -5.00 & -3.00 & 0.20 \\ & & 50.00 & 0.00 & -0.05 \\ & & & 1000.00 & 0.00 \\ & & & & 25 \end{pmatrix}$$

and where s^2 denotes the estimated variance of the error term in the model. Based on the information provided, answer the following questions justifying your answer:

a) [2 puntos] Determine the sales that the model would predict for a book launch that will have 5 promotional events with the author having had other publications before, where the new book will have 300 pages, and a budget of 4,000 dollars in digital *marketing* expenses.

b) [3 puntos] Determine which of the variables, the number of promotional events or the number of pages, is the most important.

c) [7 puntos] Through an appropriate hypothesis test, determine if the variability of book sales during the first month can be explained by this linear model. Justify your answer by using the p—value. Indicate the assumptions in which your basing your work.

d) [8 puntos] Determine if the effect of spending an additional dollar on digital *marketing* counteracts or nullifies the effect of having an additional page. Argue using a significance level of $\alpha = 0.05$, making your calculations explicit and clearly defining the rejection region associated with the test. Indicate the assumptions in which your basing your work.

Practice

Lista 05.1

- 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