Grado en Estadística

TRABAJO FIN DE GRADO

Modelización estadística de ventas en el sector retail

Marta Venegas Pardo

Sevilla, Noviembre de 2021

Índice general

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen...

Aquí comienza mi resumeeenennenenenenenenneennenemn

Abstract

Abstract...

Introduccion

Esta es la introducción de mi trabajo

- Definición

El sector retail, o también conocido como comercio minorista, es un sector económico que se encarga de proveer de bienes y servicios al consumidor final, ya que integra al grueso de comerciantes de un determinado lugar.

Este sector aglutina a comerciantes y empresas encargadas de la comercialización, ofreciendo de una gran variedad de productos y servicios a los consumidores.

(Una tienda, un supermercado, una librería, son claros ejemplos de lo que es el sector retail.)

- Objetivos
- Estimación de la demanda
- Asociación entre los dos productos

Índice de figuras

Índice de tablas

Capítulo 1

La ciencia de datos en el sector retail

Modelización estadística de ventas en el sector retail

- Objetivo: predicción de ventas, es decir, predicción de la demanda de productos

1.1. Big Data en el sector retail (COMO TITULAR)

Con R...

Capítulo 2

Análisis de cesta de la compra (Market basket análisis)

Este capítulo está por ver

2.1. Introducción

2.2. Definición

2.3. Desarrollo

Capítulo 3

Modelos estadísticos

A continuación se exponen los modelos estadísticos que estudiaremos de cara a predecir las ventas de los productos.

3.1. Análisis de series temporales

Aplicaremos este modelo de predicción para tratar de identificar los patrones de la demanda anterior a lo largo del tiempo y luego proyectar (predecir) los patrones en el futuro.

Se define una serie temporal como una sucesión de datos ordenados en el tiempo que corresponden a una misma variable. Los datos son suelen ser tomados en intervalos regulares de tiempo.

Nuestro objetivo dentro del análisis de series temporales será identificar el proceso estocástico que ha sido capaz de generar la serie de estudio.

(No se si añadirlo) Se dice proceso estocástico a una colección o familia de variables aleatorias $\{X_t, \text{ con } t \in T\}$ que siguen la misma ley de distribución y están relacionadas entre sí, pudiendo por este motivo, describir la información de estas variables en términos de medias, variaciones y covarianzas.

A continuación encontramos las cuatro etapas en un análisis descriptivo de series temporales para elegir un modelo que se adecue a nuestros datos:

- **Representación gráfica de la serie**. Para tener así una primera aproximación del comportamiento de la serie y la existencia de posibles tendencias.
- **Modelización**: Se trata de encontrar el modelo que mejor se ajuste a los datos.
- **Validación de los modelos**: Es necesario saber si el modelo ajustado es adecuado o no, por lo que es muy importante el estudio de los residuos.
- **Predicciones**: Una vez construido y validado un modelo, realizaremos estimaciones del futuro con nuevas observaciones.

En un enfoque clásico de series temporales, asumiremos que el comportamiento de la variable con respecto al tiempo se compone de cuatro componentes:

1. **Tendencia**: Se trata del movimiento suave y regular de la serie a largo plazo. La tendencia existe cuando hay un aumento o disminución a largo plazo de los datos.

Puede ser lineal (ajuste mediante una recta) o no lineal (aproximación mediante una curva, como por ejemplo logarítmica o exponencial)

2. **Ciclo**: Componente de tipo oscilante caracterizada por movimientos recurrentes en torno a la tendencia de la serie y que se repiten cada año pero sin una frecuencia fija.
3. **Componente estacional**: Se trata de movimientos regulares dentro de la serie con una periodicidad menor a un año, es decir, aquello que ocurre generalmente y con la misma intensidad año tras año en los mismos períodos, por ejemplo, en la misma época del año o día de la semana. Vamos a denotar por L al número de estaciones.
4. **Componente irregular**: Se trata de las variaciones de la serie sin un comportamiento sistemático y que no son explicadas por las otras tres componentes

Existen diferentes modelos de combinación de las componentes. Para describir los modelos necesitamos primero una nomenclatura básica. Denotando por X_t al valor de la variable en el instante t, se tiene:

$$X_t = f(T_t, E_t, I_t)$$

donde:

- T_t: Valor de la tendencia en el instante t
- E_t: Valor de la componente estacional en el instante t
- I_t: Valor de la componente irregular en el instante t (ruido).

Por tanto, los modelos que puede adoptar la función f son los siguientes:

- **Modelo multiplicativo**: La composición de la serie se realiza mediante el producto de sus componentes.
$$X_t = T_x \times E_t \times I_t$$
- **Modelo aditivo**: Las componentes se agregan para formar la serie temporal.
$$X_t = T_x + E_t + I_t$$
- **Modelo mixto**: La composición de la serie de la parte irregular viene de forma aditiva y la parte regular de forma multiplicativa.
$$Xt = Tx \times Et + It$$

http://www5.uva.es/estadmed/datos/series/series2.htm

Tras haber detectado el modelo mas adecuado, podremos conocer el comportamiento de la serie a largo plazo.

El siguiente paso realizar una estimación de la tendencia, T_t, habiendo eliminado previamente la componente estacional para impedir que estas oscilaciones perturben la identificación de la tendencia.

Para estimar T_t, debemos hacer una hipótesis sobre su forma:

- **Tendencia determinista**: Se supone que la tendencia es una función determinística del tiempo:
$$T_t = a + bt \quad a, b \in \mathbb{R}$$

Siendo a y b constantes, que se estimarán mediante un modelo de regresión lineal.

Sin embargo, el método que aplicaremos será el que exponemos a continuación:

- **Tendencia evolutiva (método de medias móviles)**: Este método consiste en definir la tendencia como una serie suavizada. Sunpondremos que la tendencia de la serie es una función que evoluciona lentamente y que podremos aproximar función simple del tiempo, suponiendo así una recta.

Una vez identificada la tendencia, procedemos a hacer un análisis de la estacionalidad de la serie, con el objetivo de:

- **Desestacionalizar la serie**, es decir, eliminar las oscilaciones periodicas que se repiten a lo largo de los años, haciendo así que los datos de distintas estaciones sean comparables. La serie desestacionalizada la conseguimos diferenciando la serie.
- **Realizar predicciones**, ya que si nuestros datos están afectados por una componente estacional, necesitaremos una estimación de esta de cara a realizar una predicción

Para desestacionalizar la serie, emplearemos los índices de variación estacional asociados a cada estación, ya que se suponen constantes año a año. Con esta técnica, se evidencian las diferencias en cada período, por ejemplo, podemos ver la diferencia del volumen de ventas en función de la época del año (mes, día de la semana, estación,...) Estos índices reflejan la cantidad fija o proporción en la que se modifica la tendencia en cada estación.

Una vez calculados estos índices, se desestacionaliza la serie, eliminando así el efecto de cada estación.

Por último, procedemos a realizar las predicciones. Para ello, necesitamos que se cumpla la condición de estacionariedad, es decir, la media y la varianza permanecen constantes en el tiempo (no tiene raíces unitarias). En el caso de no imponer esta condición de estacionariedad, predeciríamos carasterísticas que no serán las mismas en el futuro que en el pasado.

Se tiene que todo proceso lineal es estacionario, por tanto, obtendremos trabajaremos con series estacionarias, y de lo contrario, podremos aplicar los mismos métodos a series no estacionarias realizando las transformaciones pertinentes para conseguir la estacionariedad.

En nuestro caso, aplicaremos la metodología Box-Jenkis como método predictivo.

3.1.1. Metodología Box-Jenkis

Esta metodología tiene en cuenta la dependencia existente entre los datos, es decir, cada observación en el instante t será modelada a partir de los valores pasados. Los modelos se conocen con el nombre de ARIMA (modelos integrados autorregresivos de medias móviles), que deriva de las siguientes componentes: AR (Autorregresivo) , I (integrado), MA(Medias móviles)

El siguiente paso es identificar el modelo más adecuado a través del estudio de la función de autocorrelación (FAC) y la función de autocorrelación parcial (FAP).

Nota: el método recomienda como mínimo 50 observaciones en la serie temporal.

Fases de la metodología Box-Jenkis:

1. Identificar el la estructura ARIMA que sigue la serie a través del estudio de la función de autocorrelación simple (FAS) y la función de autocorrelación parcial (FAP). Determinar el modelo arima consiste en identidicar los órdenes p y q de su estructura autoregresiva y de medias móviles
2. Estiamción de parámetros: Una vez tenemos identificado el modelo, estimamos los parámetros AR y MA del modelo por el método de máxima verosimilitud, obteniendo el error estándar y los resíduos del modelo

   Nota: Es muy importante comprobar que las estimaciones son significativamente no nulas.
3. Diagnosis del modelo: Comprobamos que los resíduos sigan un proceso de ruido blanco mediante el Test de Ljung-Box.

   Si hemos identificado varios modelos y todos ellos pasan la diagnosis, nos quedaremos con uno de ellos según el criterio del menor AIC
4. Predicción: una vez identificado y validado el mejor modelo, se realizan las predicciones con éste.

3.2. Modelo 2 (decidir)

Capítulo 4

Proceso de ciencia de datos (Data science process)

4.1. Introducción

La ciencia de datos es la combinación de múltiples campos, como la estadística, la inteligencia artificial (IA), el análisis de datos,... con el objetivo de extraer información de valor de esos datos. La ciencia de datos, abarca las siguientes etapas: recolección de los datos, limpieza, análisis exploratorio, construcción y validación de modelos y predicciones.

Una parte importante de la ciencia de datos es el Aprendizaje Automático o Machine Learning (ML). Se trata de un subcampo dentro de la ciencia de datos, concretamente, una subcategoría de la inteligencia artificial. Está basada en algoritmos, y consiste en que éstos descubran de manera autónoma patrones recurrentes del conjunto de datos. Los algoritmos de ML al detectar patrones en los datos, aprenden y mejoran el rendimiento en la ejecución de una tarea o al hacer predicciones. Una vez entrenado y validado el modelo, el algoritmo podrá encontrar patrones en nuevos datos (predicciones)

Para la correcta explicación de las técnicas que se van a describir, es necesario la definición del aprendizaje estadístico supervisado. **

El aprendizaje estadístico supervisado es una de las principales herramientas del aprendizaje automático y consiste en una serie de técnicas para deducir una función a partir de una serie de datos de entrenamiento. El objetivo es crear o estimar una función capaz de predecir el valor deseado después de haber visto una serie de ejemplos. Para ello, tiene que generalizar a partir de los datos presentados anteriormente a las nuevas situaciones no vistas previamente. La salida de la función puede ser un valor numérico (como en problemas de regresión) o una etiqueta de clase (como en los de clasificación)

4.2. Etapas del proceso de ciencia de datos

4.2.1. Conocimiento del negocio (Knowledge of Bussiness)

En esta primera etapa, es fundamental la definición del problema que nos ocupa, la definición de unos objetivos claros y la metodología para cumplirlos.

Esto implica la comprensión de los requisitos del proyecto desde el punto de vista de negocio, utilizando las perspectivas de negocio para determinar a que problemas podemos dar respuesta mediante el uso de la minería de datos.

4.2.2. Adquisición de los datos (Collect the data)

Explicación de los datos, fuente, explicación de las variables,...

Consiste en explicar como se ha llevado a cabo la adquisición de los datos, la identificación de las distintas fuentes y la explicación de los mismos.

4.2.3. Preparación de los datos (data preparation)

Raramente encontraremos los datos preparados para su análisis, ya que normalmente es necesario la limpieza y la transformación de los mismos. Para ello, es necesario llevar a cabo un paso previo llamado pre-procesamiento de los datos.

Fases del data cleaning:

- Eliminación de duplicados (filas y columnas)
- Datos erróneos (ej: precios negativos)
- Detección de valores faltantes: decidir si eliminar esos registros o imputarlos
- Detección de outliers (decidir si mantener, quitar o tratar a parte)
- Unificación de variables (unificación de unidades,...)
- Creación de variables a partir de otras ya existentes si fuera necesario

Preparación de los datos

- Reformateo de variables, por ejemplo, formatos horarios.
- Categorización,...
- Selección de variables (Feature selection): elegir las mejores variables que alimenten nuestros algoritmos dictarán la máxima calidad que podemos conseguir, ya que no todas las variables explican el problema que queremos modelar. Podemos resumir esto con la siguiente frase: “Garbage in, garbage out”, es decir, si entra basura saldrá basura. Refiriendonos con basura a ruido en los datos o información pobre.

4.2.4. Análisis exploratorio de datos (EDA)

El análisis exploratorio se utiliza para ver lo que nos pueden ofrecer los datos antes de la etapa del modelado y se lleva a cabo para resumir las principales características del conjunto de datos a través de diferentes tareas:

- Estudio descriptivo de los datos: La estadística descriptiva es la parte de la estadística dedicada a la ordenación y tratamiento de la información por medio de gráficas y tablas, además de la obtención de parámetros útiles para explicar la información
- Visualizaciones de los datos:
    - Análisis univariante: Empleado para observar diferentes características de interés, tratar de identificar patrones en los datos o ver la distribución de las variables . Algunos ejemplos serían los gráficos de caja y bigote o histogramas

- Análisis multivariante: Donde tratamos de ver la asociación o relación que pueden tener las distintas variables de interés. Encontramos los gráficos de barras o gráficos de dispersión entre los ejemplos de representaciones multivariantes.

- Relación entre las variables

Este tipo de análisis permite obtener medidas descriptivas de un conjunto de datos para poder extraer conclusiones referentes a una muestra o población.

4.2.5. Modelado

En la etapa de modelado aplicaremos algoritmos de aprendizaje automático. Para llevar a cabo esta fase y con el objetivo de obtener mayor robustez en los modelos, aplicaremos técnicas de validación cruzada.

La validación cruzada es una técnica de remuestreo de datos utilizada para evaluar la capacidad de generalización de los modelos de predicción proporcionando una estimación del rendimiento del modelo final en datos nuevos y evitar el sobreajuste. El sobreajuste aparece cuando un modelo que se adapta perfectamente a los datos de entrenamiento obteniendo unas métricas muy buenas pero que luego es incapaz de generalizar con datos nuevos.

El primer paso consiste en la división de los datos en el conjunto de datos de entrenamiento, validación y testeo. Para ello, generamos un conjunto de entrenamiento y otro de testeo a partir del conjunto de datos inicial. A continuación, volvemos a dividir los datos de entrenamiento en datos de entrenamiento y validación, obteniendo así tres conjuntos de datos independientes: entrenamiento, validación y testeo.

El conjunto de datos de entrenamiento es aquel que utilizamos para probar diferentes hiperparametrizaciones de cada modelo para ver cual es la más óptima. La hiperparametrización variará en función de los parámetros aplicables a cada algoritmo utilizado.

Una vez hayamos entrenado los modelos, pasamos a la fase de validación, donde aplicaremos a los datos de validación los diferentes algoritmos con la configuración de parámetros que mejor haya funcionado en el conjunto de datos de entrenamiento.

El modelo con el que obtengamos las mejores métricas será el que posteriormente apliquemos a los datos de testeo.

A continuación vamos a exponer tres algoritmos de aprendizaje automático que posteriormente aplicaremos a nuestros datos

4.2.5.1. Máquinas de vector soporte (Support Vector Machines SVMs)

Las máquinas de vector soporte son un conjunto de algoritmos de aprendizaje estadístico supervisado pertenecientes a la familia de los clasificadores lineales.

Suponiendo que tenemos ejemplos de sólo dos categorías y sin pérdida de generalidad, una SVM construye un hiperplano en un espacio de dimensionalidad muy alta. Este hiperplano separa de forma óptima los puntos de una clase de la otra. La característica fundamental de estos algoritmos es el concepto de "separación óptima", ya que se busca el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él

mismo al tiempo que clasifica correctamente tantos puntos de entrenamiento como sea posible. Los algoritmos SVM representan el hiperplano óptimo con vectores de soporte.

En nuestro caso al ser la variable volumen de ventas una variable numérica, vamos a centrarnos en la variante SVM para regresión, tambien conocida como SVR (support vector regressor). El caso del problema de regresión es una generalización del problema de clasificación, en la que el modelo devuelve un valor continuo, es decir, un modelo de regresión estima una función multivariante de valor continuo.

Dado un conjunto de ejemplos de entrenamiento $S = \Big\{(x_1, y_1), \ldots, (x_n, y_n)\Big\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, en el que se asume que todos los valores y_i de todos los ejemplos de S pueden ser ajustados mediante un hiperplano, nuestro objetivo será encontrar los parámetros $w = (w_1, \ldots, w_d)$ que permitan definir el hiperplano de regresión

$$y = f(x) = (w_1 x_1 + \cdots + w_d x_d) + b = \langle w, x \rangle + b \ , \ b \in \mathbb{R}$$

La generalización de SVM a SVR se logra introduciendo una región insensible a ϵ alrededor de la función. Esta región se conoce como tubo épsilon. Este tubo reformula el problema de optimización para encontrar el tubo que mejor se aproxime a la función al tiempo que equilibra el error de predicción, es decir, se formula un problema de optimización definiendo una función de pérdida a minimizar insensible a ϵ y encontrando el tubo más plano que contiene a la mayoría de instancias de entrenamiento.

Se dice **rudio, perturbación aleatoria o tubo épsilon** y se representa por $\epsilon \sim N(0, \sigma^2)$, al error en la medición del valor y, por tanto, $y = f(x) + \epsilon$

El valor de ϵ determina el ancho del tubo, y un valor más pequeño indica menor tolerancia al error, cuando más pequeño sea el valor de ϵ, el límite del tubo se desplaza hacia dentro, habiendo más puntos de datos alrededor del límite, lo que indica más vectores de soporte.

Se define la **función de pérdida lineal** $\epsilon-$ insensible, y se representa como L_ϵ a una función lineal en el que la función de pérdida toma valor nulo y viene definida de la siguiente forma:

$$L_\epsilon = \begin{cases} 0 & \text{si} |y - f(x)| \leq \epsilon, \\ |y - f(x)| - \epsilon & \text{en caso contrario} \end{cases}$$

4.2.5.2. KNN

4.2.5.3. Árboles de regresión (XGBoost Model)

4.2.5.4. Evaluación y presentación de resultados (+análisis del error)

- Predicciones con el mejor modelo
- Final de la historia de una forma ordenada y resumida
- Señalar posibles mejoras y recomendaciones para proyectos futuros

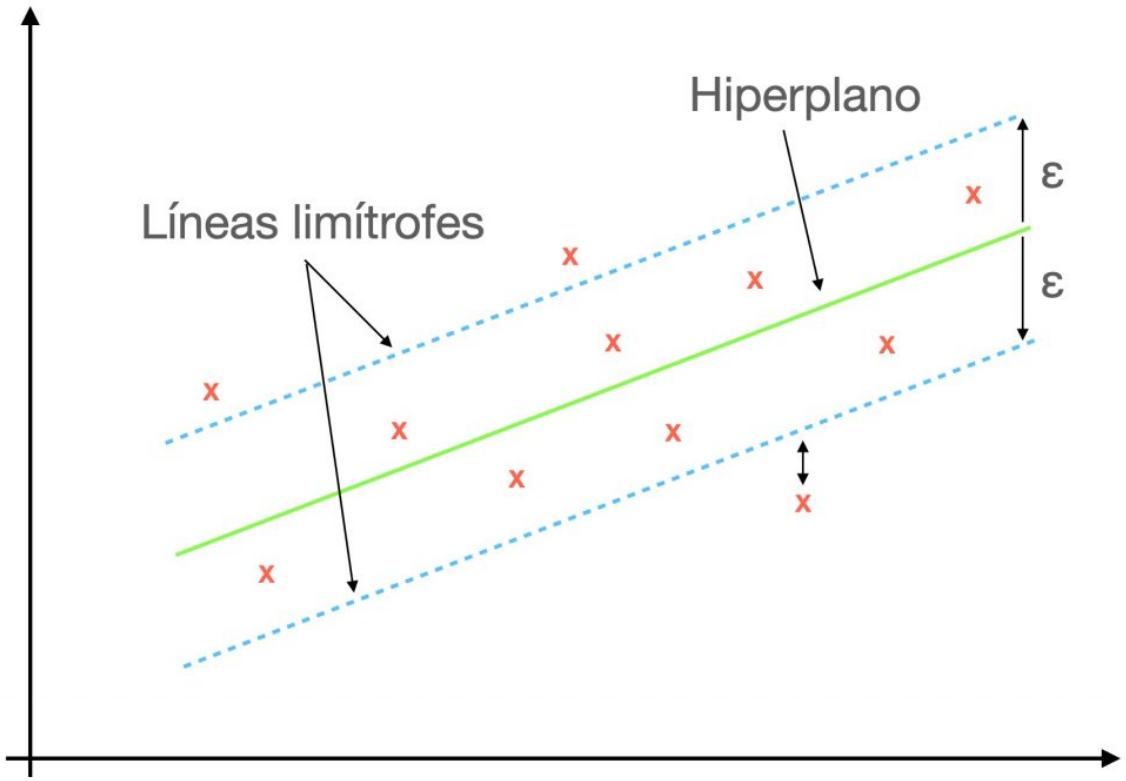

Figura 4.1: Vectores de Soporte de Regresión

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2021. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.11.

Mariette Awad and Rahul Khanna. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. URL https://doi.org/10.1007/978-1-4302-5990-9_4.

Carles Morales Boada. Introducción al feature selection (o cómo escoger las variables adecuadas). Disponible en https://www.linkedin.com/pulse/introducción-al-feature-selection-o-cómo-escoger-las-morales-boada/?originalSubdomain=es.

Andrea Mesta Carmona. *Técnicas estadísticas en predicción de la demanda en el sector comercio*. Facultad de Matemáticas, Universidad de Sevilla, 2021.

Guillermo Corres, Alejandra Esteban, Juan García, and Claudia Zárate. Análisis de series temporales. Disponible en https://economipedia.com/definiciones/sector-retail.html#:~:text=El%20sector%20retail%2C%20o%20comercio,comerciantes%20de%20un%20determinado%20lugar., 01 2009.

Carlos Alberto Cárdenas Ojeda, Sandra Patricia Ramos Soler. Análisis exploratorio de datos en r, 2015. URL https://repositorio.uptc.edu.co/handle/001/8178.

M. Dolores Jiménez Gamero. *Apuntes de la asignatura Series Temporales*, 2021.

Elena Campo León. *Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervisado*. Facultad de Ciencias, Universidad de Zaragoza.

Pedro L. Luque-Calvo. *Escribir un Trabajo Fin de Estudios con R Markdown*, 2017.

Pedro L. Luque-Calvo. *Cómo crear Tablas de información en R Markdown*, 2019.

Francisco Coll Morales. Sector retail. Disponible en https://economipedia.com/definiciones/sector-retail.html#:~:text=El%20sector%20retail%2C%20o%20comercio,comerciantes%20de%20un%20determinado%20lugar.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2021a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.3.5.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2021. URL https://yihui.org/knitr/. R package version 1.36.