Analítica de datos

Extensiones del modelo de regresión lineal

Pontificia Universidad JAVERIANA Bogotá

Profesor: Nicolás Velásquez

Extensiones del Modelo de Regresión Lineal

- Hasta ahora hemos asumido linealidad de nuestro modelo.

- Estimación de modelos No-Lineales:
  - Interacción de variables
  - Transformación de relacionales no lineales

Interacción de variables

Hasta ahora hemos asumido que el efecto de una variable es independiente de otra:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

El efecto del incremento en una unidad de X_1 en Y, es siempre β_1 e independiente de X_2.

Si pensamos que el efecto de X_1 depende del valor de X_2, entonces debemos agregar una tercera variable de interacción al modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

El termino de interacción es $X_1 X_2$

Interacción de variables

Ejemplo Advertising:

$$\widehat{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times radio$$

Suponga que el **gasto en publicidad radial incrementa la efectividad de la publicidad en TV**, esto es, la pendiente del efecto de la TV aumenta a medida que aumenta el gasto en radio:

$$\widehat{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (TV \times radio)$$

El efecto del gasto publicitario en TV sobre las ventas es:

$$\frac{dy\ (sales)}{dx\ (TV)} = \beta_1 + \beta_3 \times radio$$

Interacción de variables

En Marketing, este efecto se conoce como sinergia(HALO/HORN) y en estadísticas como **efecto de interacción.**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

El termino de interacción es $X_1 X_2$

Interacción de variables

Ejemplo Advertising:

$$\widehat{Sales} = 6.75 + 0.019 \times TV + 0.029 \times radio + 0.0011 \times (TV \times radio)$$

Interpretación:

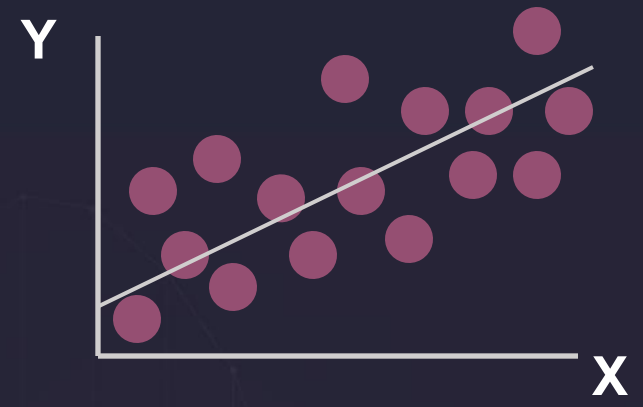
Los resultados sugieren que un incremento de $1000 en el gasto publicitario en TV esta asociado con un incremento en las ventas de: $(0.019 + 0.0011 \times unidades\ de\ radio) \times 1000$

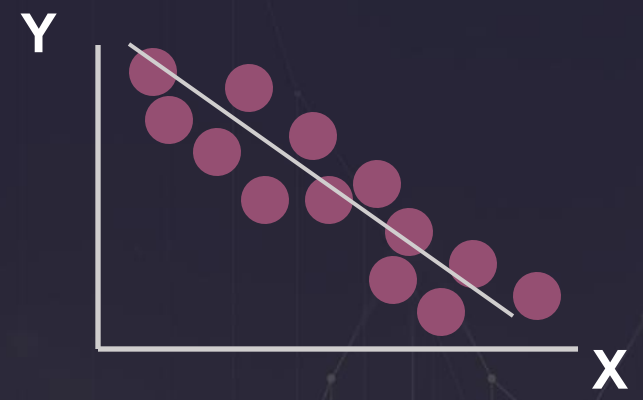
Los resultados sugieren que un incremento de $1000 en el gasto publicitario en radio esta asociado con un incremento en las ventas de: $(0.029 + 0.0011 \times unidades\ de\ TV) \times 1000$

TRANSFORMACIÓN DE VARIABLES NO LINEALES

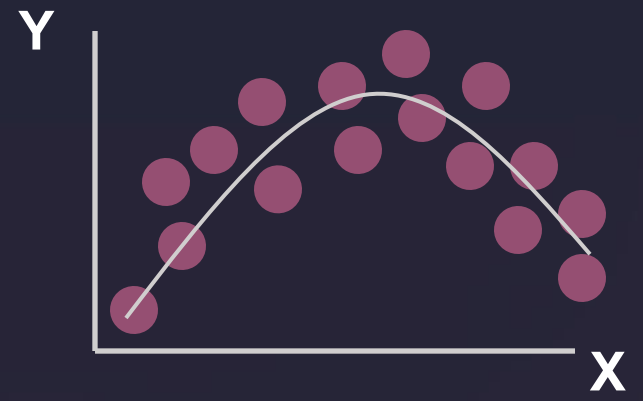
Tipos de relaciones entre dos variables

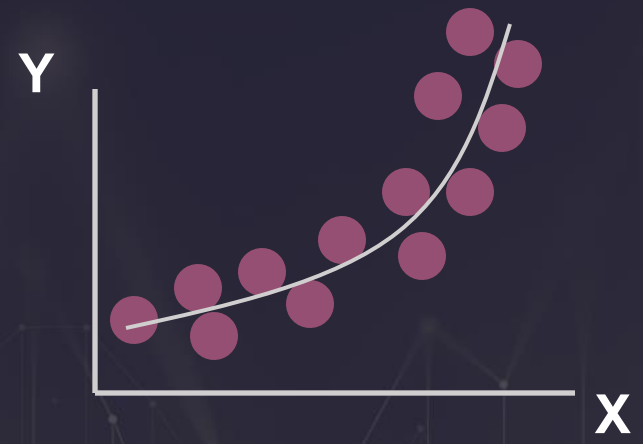
Relaciones lineales





Relaciones NO lineales





RELACIONES NO LINEALES ¿COMO LINEALIZAR UNA RELACION ENTRE DOS VARIABLES?

Relaciones no lineales: Relación Cuadrática

- Relación cuadrática ¿Cómo convertirla en una relación lineal?

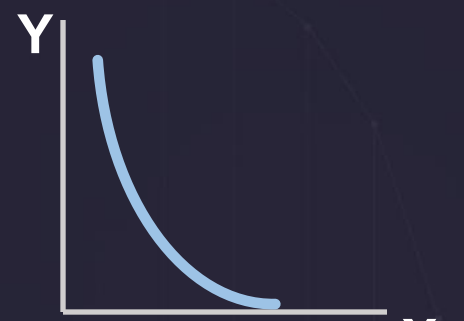
  - Dado que una de las propiedades de la regresión lineal es la **linealidad**:
  - La relación entre X y Y debe ser lineal.
  - Si esta relación no es lineal, **podemos hacer una transformación al modelo y agregar un termino cuadrático**, como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Modelo de regresión cuadrática

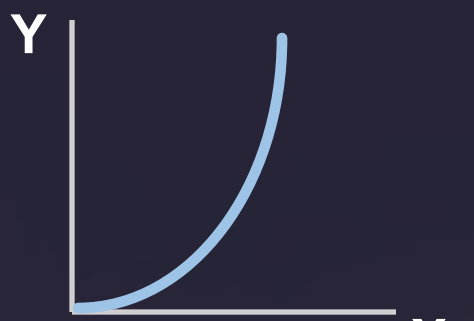
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Podemos considerar un modelo cuadrático cuando el diagrama de dispersión muestral los siguientes patrones:



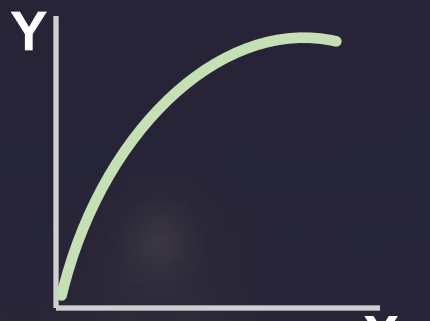
$\beta_1 < 0$

$\beta_2 > 0$



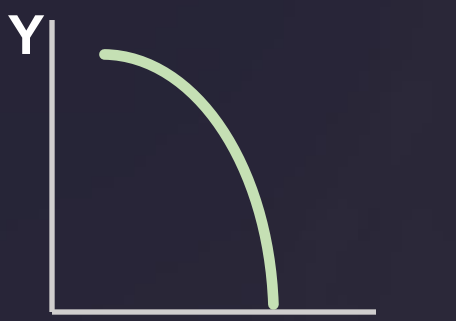
$\beta_1 > 0$

$\beta_2 > 0$



$\beta_1 > 0$

$\beta_2 < 0$



$\beta_1 < 0$

$\beta_2 < 0$

β_1 = coeficiente del término lineal
β_2 = coeficiente del término cuadrático

¿Hay evidencia de que mi término cuadrático tiene un efecto sobre Y?

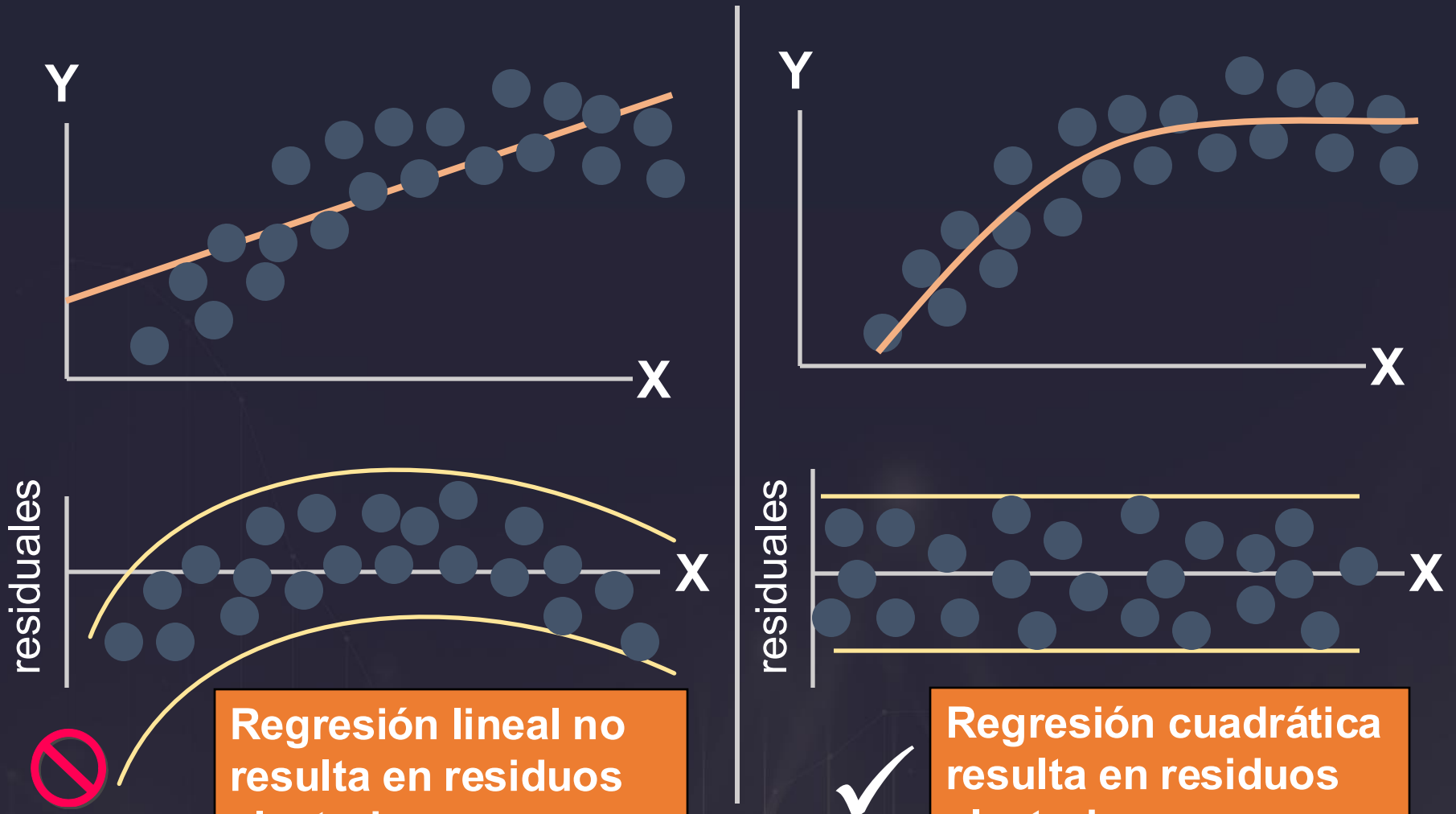
- Hipótesis

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

- Como siempre, si p-value < nivel de significancia (α), rechazamos la hipótesis nula. De lo contrario, no la rechazamos.

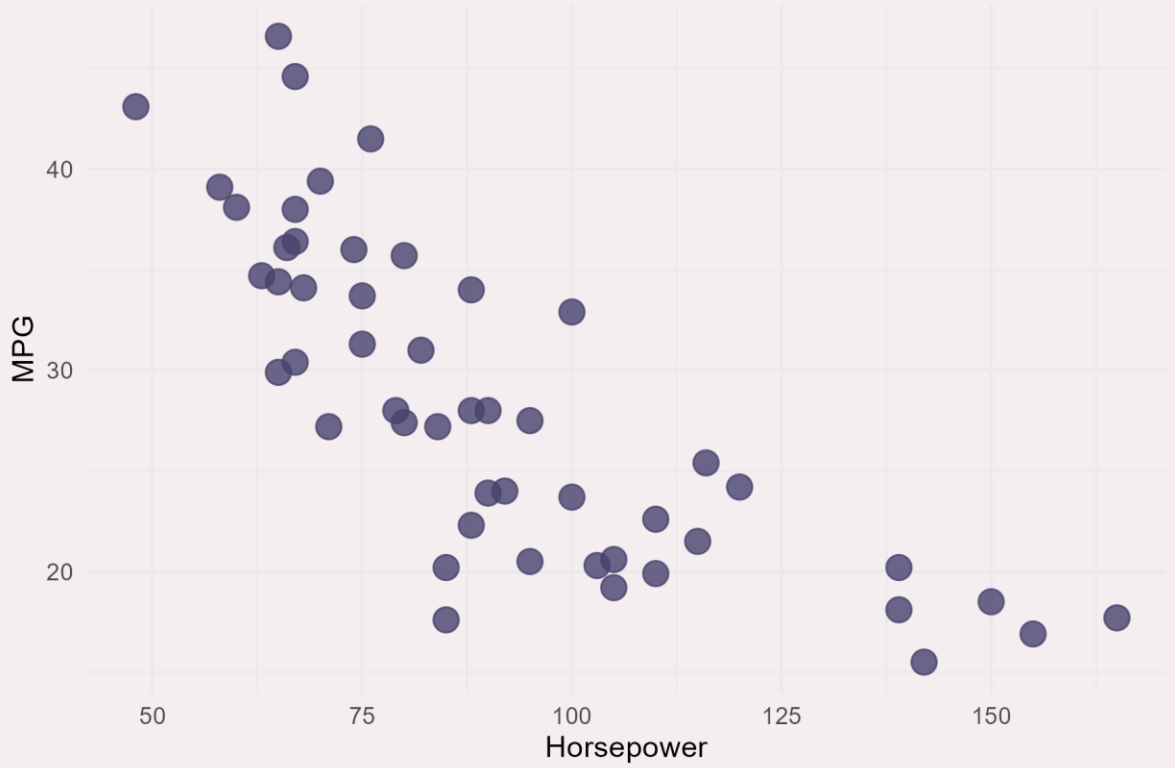
¿Hemos solucionado el problema de no linealidad?



Ejemplo: Relación Cuadrática

En la base de datos Auto, podemos estudiar la relación entre rendimiento (MPG – Millas por Galón) y Potencia (HorsePower):

Creen este gráfico en R.



Ejemplo: Relación Cuadrática

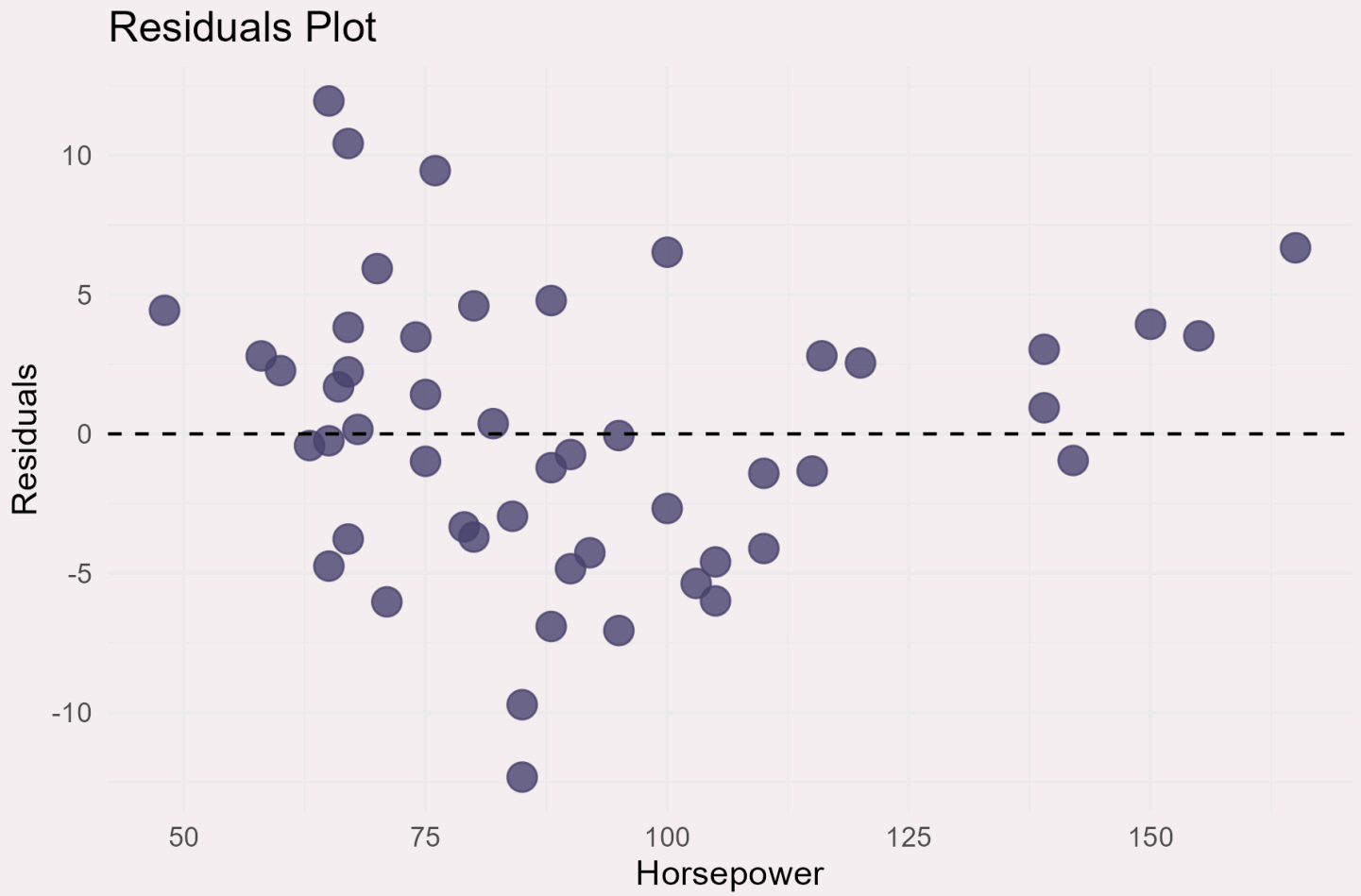
- Resultados regresión simple: $\hat{Y}$ **= 50.01 -0.24HP P-value:0.56**

```
# el gráfico Autos

Auto |> ggplot(aes(Horsepower, MPG)) +
geom_point(color = "#48446e", size = 4,alpha = 0.8)+
theme_minimal()

ggsave("grafico_autos.png", dpi = 300, width = 6, height = 4, units = "in")


# Y la regresión autos
lm_model_autos <- lm(MPG ~ Horsepower, data = Auto)

# Muestra los resultados del modelo
summ(lm_model_autos)
```

P-value de b1 es bajo, y R2 es muy alto. Sin embargo, los residuos no son aleatorios.

Ejemplo: Relación Cuadrática

- Resultados regresión simple: $\hat{Y} = 50.01 - 0.24HP$



Ejemplo: Relación Cuadrática

```
# create a data frame of the residuals
residuals_df_1 <- data.frame(
Horsepower = Auto$Horsepower,
Residuals = lm_model_autos$residuals
)

# create the plot
ggplot(residuals_df_1, aes(x = Horsepower, y = Residuals)) +
geom_point(color = "#48446e", size = 4,alpha = 0.8) +
geom_hline(yintercept = 0, linetype = "dashed") +
labs(x = "Horsepower", y = "Residuals", title = "Residuals Plot")+
theme_minimal()

ggsave("residuals_hp.png", dpi = 300, width = 6, height = 4, units = "in")
```

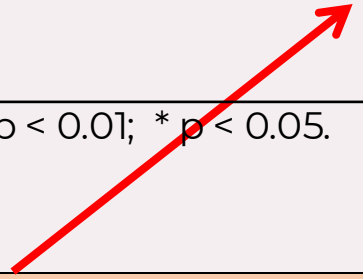
Ejemplo: Relación Cuadrática

- **Resultados regresión cuadrática:**

$$\hat{Y} = 77.1 - 0.8035HP + 0.0027\,(HP)^2$$

	Cuadrática
(Intercept)	77.1071 ***
	(7.8711)
Horsepower	-0.8035 ***
	(0.1596)
I(Horsepower^2)	0.0027 ***
	(0.0008)
N	50
R2	0.7029

*** p < 0.001; ** p < 0.01; * p < 0.05.



El término cuadrático es estadísticamente significativo a niveles convencionales (p-value muy pequeño)

Ejemplo: Relación Cuadrática

- Resultados regresión cuadrática:

$$\hat{Y} = 56.9 - 0.4662HP + 0.0012\,(HP)2$$

Regression Statistics	
R Square	0.70
Adjusted R Square	0.69
Standard Error	4.548

El modelo cuadrático explica 69% de la variación en Y.

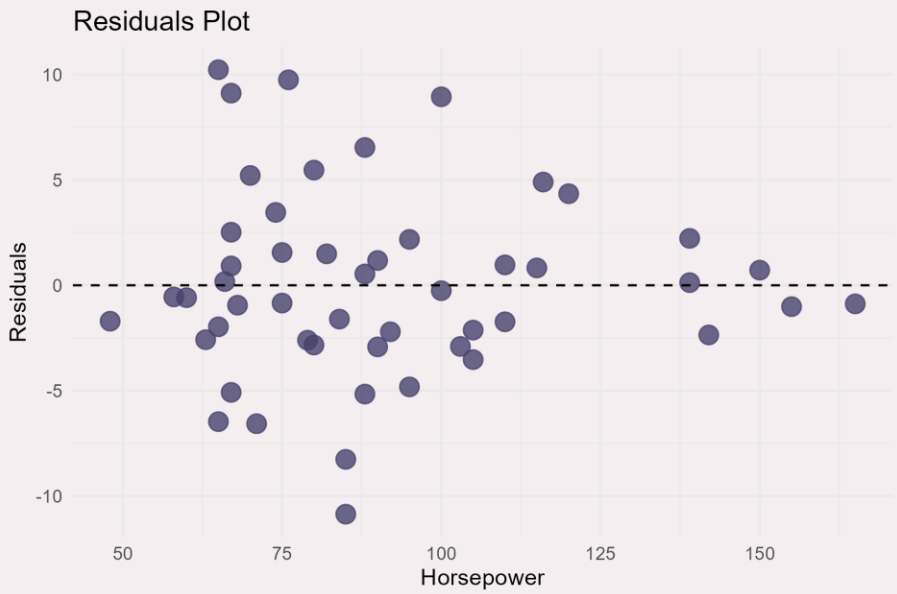
Ejemplo: Relación Cuadrática

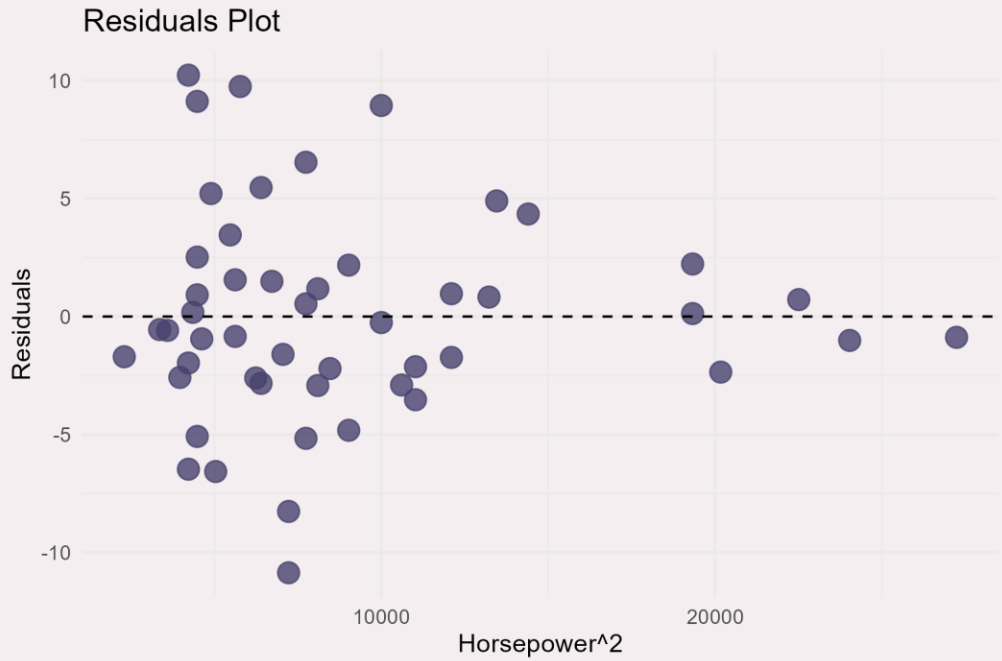
```
# fit the linear model with a quadratic term for Horsepower
lm_model_autos_quad <- lm(MPG ~ Horsepower + I(Horsepower^2), data = Auto)

summ(lm_model_autos_quad)
summary(lm_model_autos_quad)



export_summs(lm_model_autos_quad,digits = 4,
to.file = "docx", file.name = "quadratic.docx")
```

Ejemplo: Relación Cuadrática

- Resultados regresión cuadrática:

$$\hat{Y} = 56.9 - 0.4662HP + 0.0012\,(HP)2$$





Los dos diagramas de dispersión muestran residuos aleatorios. No hay patrón (i.i.d.)

Ejemplo: Relación Cuadrática

```r
# create a data frame of the residuals
residuals_df <- data.frame(
Horsepower = Auto$Horsepower,
Residuals = lm_model_autos_quad$residuals,
HP_squared = Auto$Horsepower^2
)
# create the plot HorseP
ggplot(residuals_df, aes(x = Horsepower, y = Residuals)) +
geom_point(color = "#48446e", size = 4,alpha = 0.8) +
geom_hline(yintercept = 0, linetype = "dashed") +
labs(x = "Horsepower", y = "Residuals", title = "Residuals Plot")+
theme_minimal()

ggsave("residuals_hpquad.png", dpi = 300, width = 6, height = 4, units = "in")



# create the plot HorseP^2
ggplot(residuals_df, aes(x = HP_squared  , y = Residuals)) +
geom_point(color = "#48446e", size = 4,alpha = 0.8) +
geom_hline(yintercept = 0, linetype = "dashed") +
labs(x = "Horsepower^2", y = "Residuals", title = "Residuals Plot")+
theme_minimal()

ggsave("residuals_hp2quad.png", dpi = 300, width = 6, height = 4, units = "in")
```

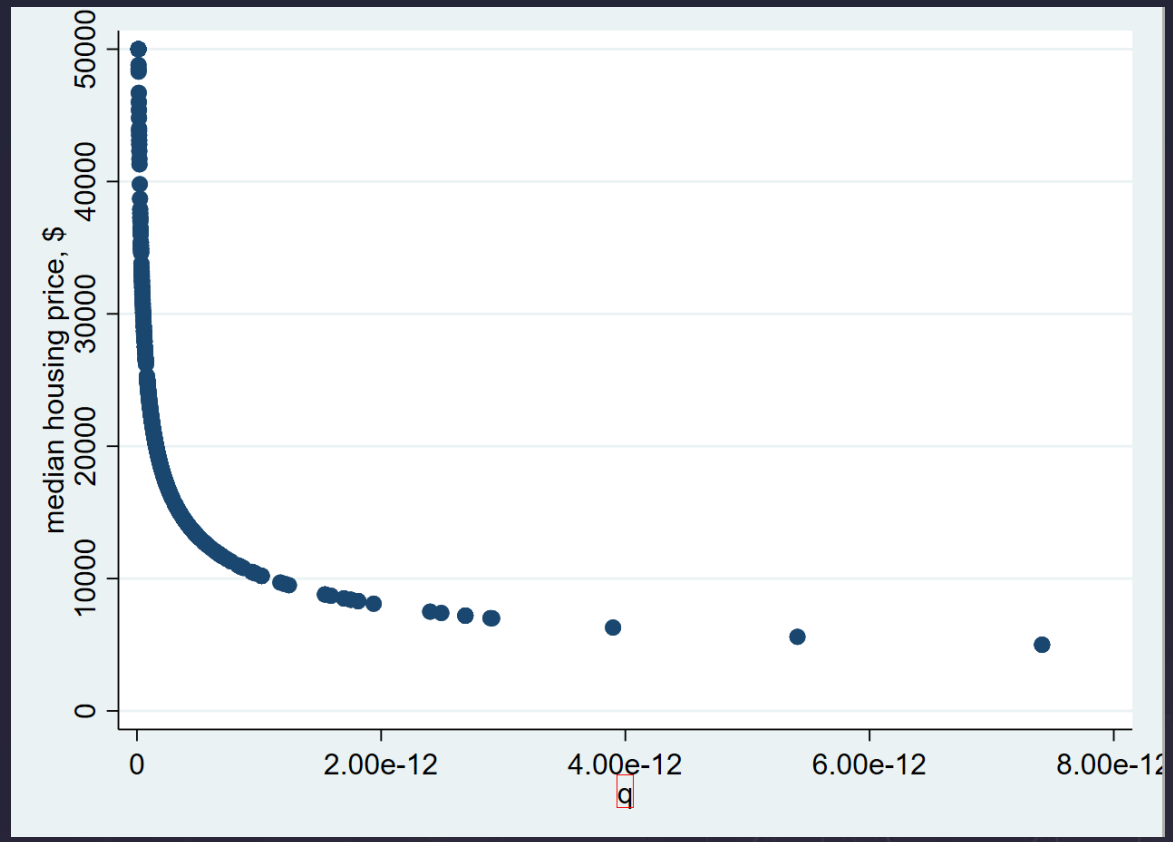
Transformación logarítmica

- Podemos considerar una transformación logarítmica de las variables de nuestro modelo (tanto variable dependiente e independientes) cuando:
  - Exista una relación no-lineal (exponencial) entre la variable dependiente y la explicativa.
  - Alguna de las variables tenga una distribución sesgada (muy distinta a la normal).

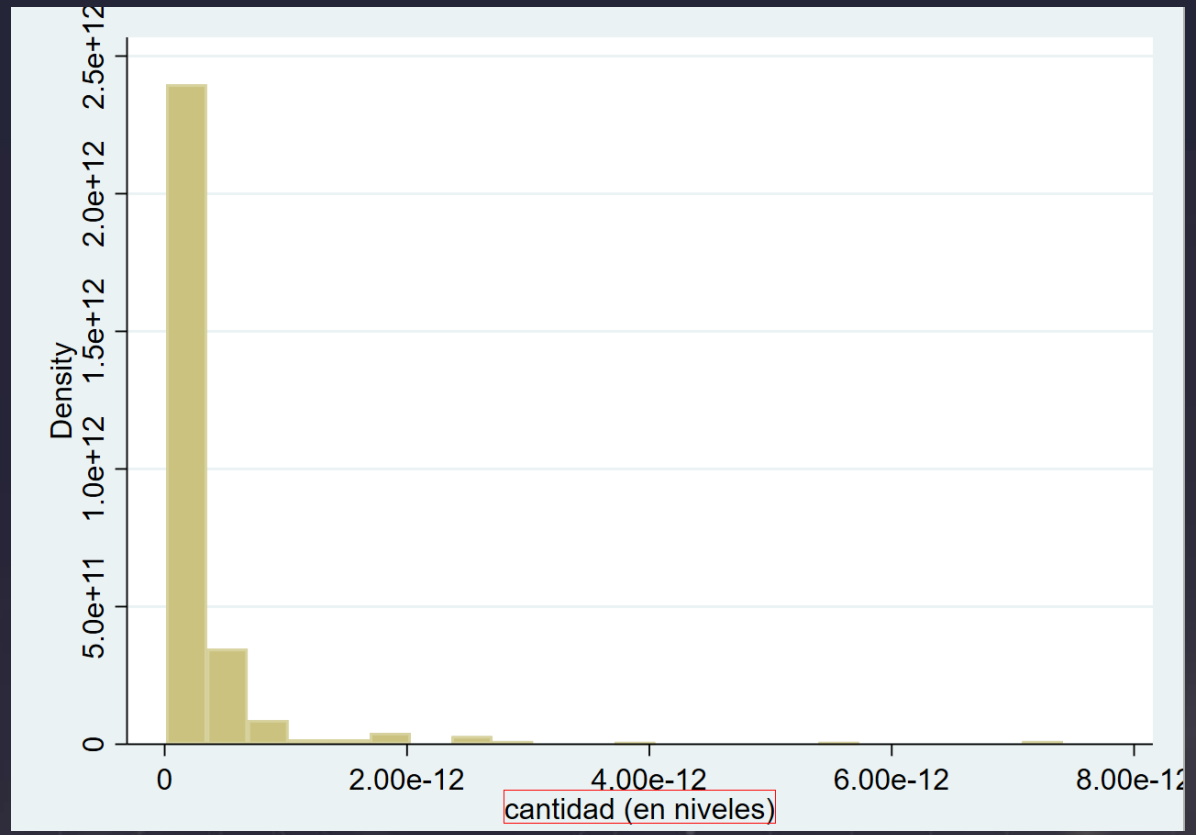
Transformación logarítmica

Relacion exponencial entre precio y cantidad:

$$Q = a \times P^b$$

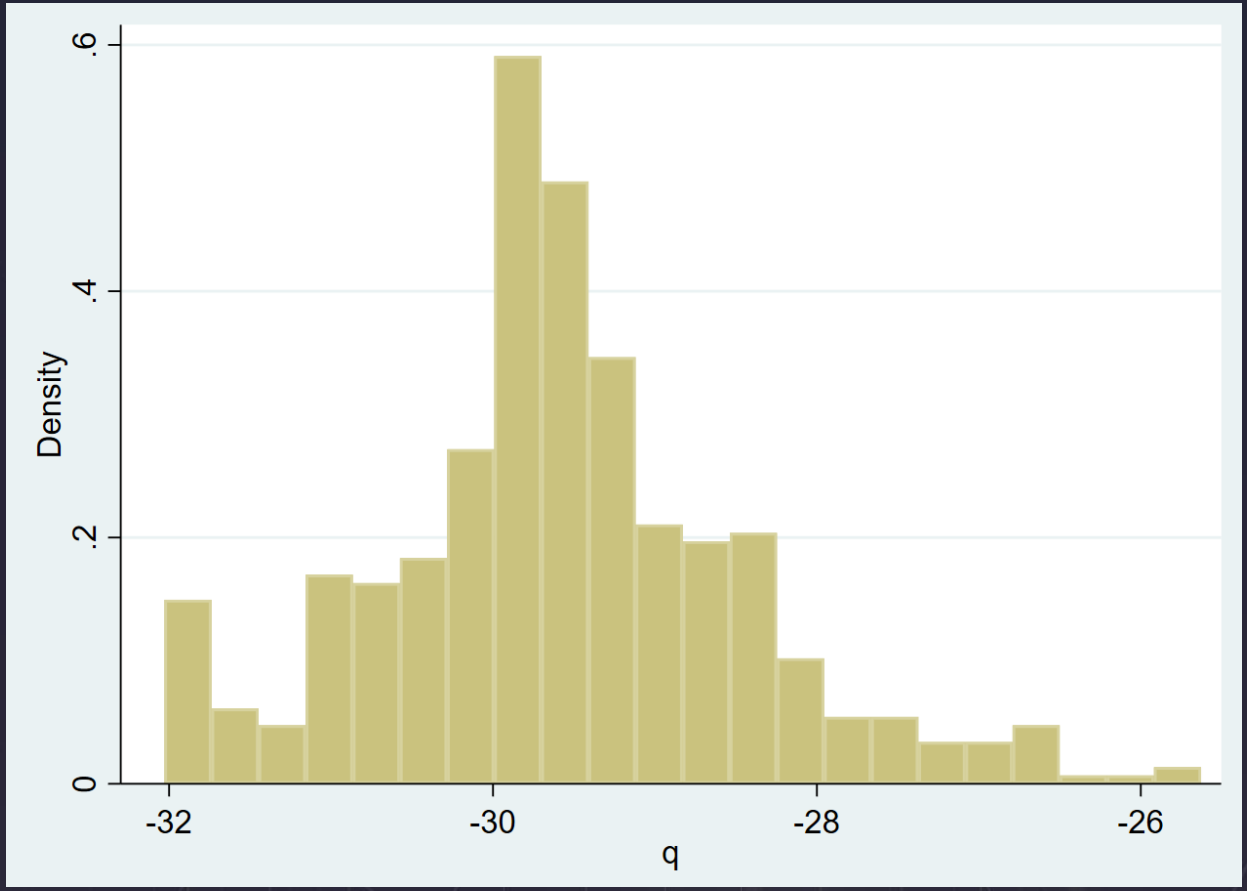


Distribución sesgada: ejemplo distribución sesgada a la derecha



Transformación logarítmica

Distribución cantidad en logaritmos



Transformación logarítmica

Ejemplo: Demanda de casas

```
. reg lsales lprice

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(1, 504)       >  99999.00
       Model |  653.567376         1  653.567376   Prob > F        =    0.0000
    Residual |  .001079196       504  2.1413e-06   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  653.568456       505  1.29419496   Root MSE        =    .00146

------------------------------------------------------------------------------
      lsales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lprice |  -2.779749   .0001591 -1.7e+04   0.000    -2.780062   -2.779437
       _cons |   34.88897   .0015831   2.2e+04   0.000     34.88586    34.89208
------------------------------------------------------------------------------
```

Interpretación de B1: si el precio sube en un 1% las ventas disminuirán en un 2,78% (elasticidad)

Transformación logarítmica

Interpretación después de la transformación logarítmica

Modelo	Variable dependiente	Variable explicativa	Interpretacion de β_1
Nivel - Nivel	Y	X	$\Delta Y = \beta_1 \Delta X$
Log – Nivel	Log(Y)	X	$\%\Delta Y = 100\beta_1 \Delta X$
Log - log	Log(Y)	Log(X)	$\%\Delta Y = \beta_1 \%\Delta X$

REGRESIÓN LINEAL CON VARIABLE DEPENDIENTE DUMMY

- Hasta el momento hemos hecho regresiones en los que la variable explicada es una variable numérica.

- Sin embargo, la variable explicada también puede ser una variable categórica!

- Veremos cómo estimar e interpretar un modelo de regresión cuando la variable explicada es una variable ***binaria.***

Modelos de variable dependiente binaria

El modelo de regresión lineal cuando la variable explicada es binaria se denomina:

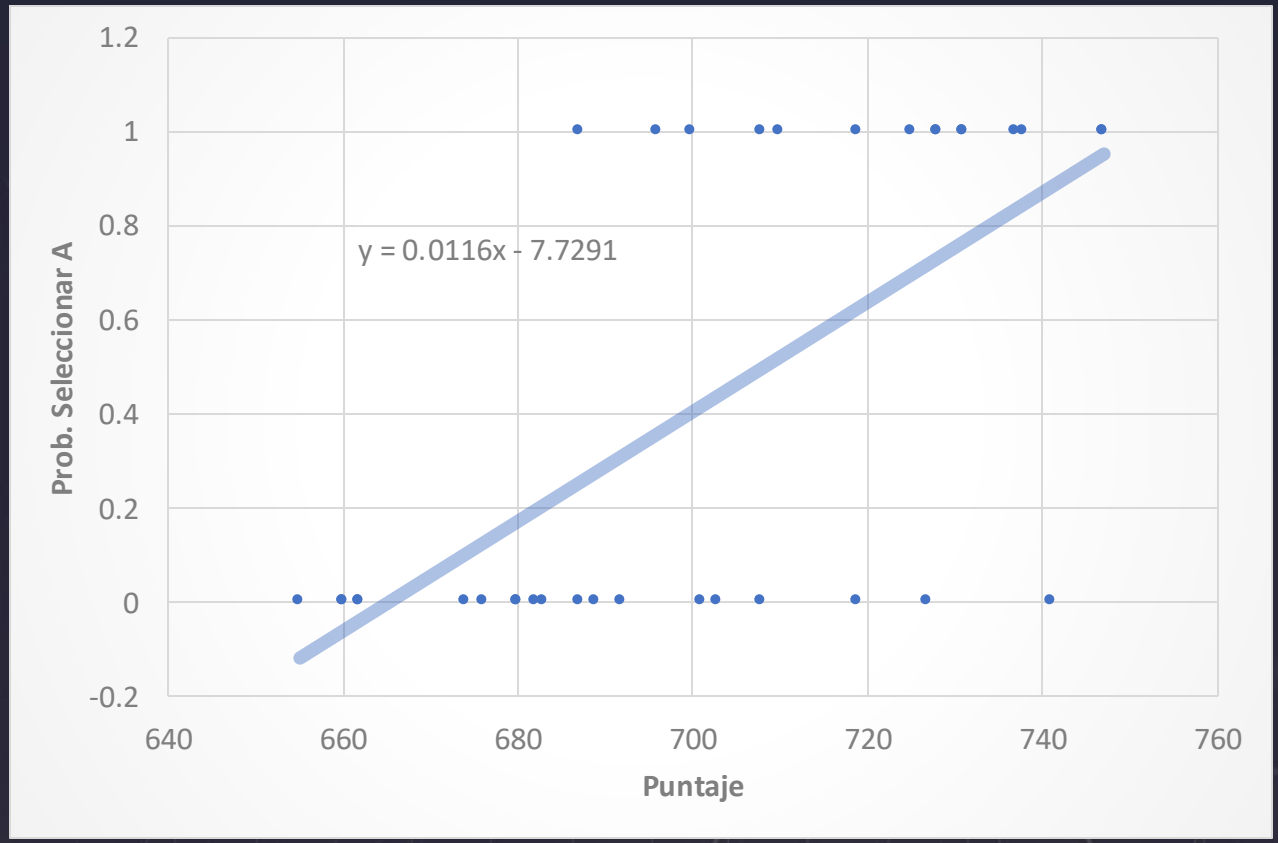
Modelo de Probabilidad Lineal

- En este caso, los coeficientes miden el efecto de las variables explicativas sobre la **PROBABILIDAD** de un evento:
  - ¿Cuál es el efecto de perder el empleo sobre la probabilidad que un cliente entre en mora?
  - ¿Cuál es el efecto del salario ofrecido sobre probabilidad que un ejecutivo acepte una oferta laboral?
  - ¿Cuál es el efecto de la nota en un quizz sobre la probabilidad de aprobar la materia?

Modelos de variable dependiente binaria

El modelo de regresión lineal cuando la variable explicada es binaria se denomina:

Modelo de Probabilidad Lineal



Modelos de variable dependiente binaria

El modelo de probabilidad lineal pronostica valores por fuera de [0,1]

Necesitamos un modelo **NO LINEAL** para que los pronósticos estén en [0,1] y por lo tanto puedan interpretarse como probabilidades

El modelo de regresión (posiblemente) mas usado en la practica es:

Regresión Logística

- Nos permite modelar distintas situaciones como:
  - ¿Cuál es la probabilidad que un cliente pague una deuda?
  - ¿Cuál es la probabilidad de que un cliente compre mi producto?
  - ¿Cuál es la probabilidad que un colaborador acepte una oferta laboral?

Modelo de regresión Logística

Modelo de regresión logística, se basa en una función logística que garantiza la predicción de probabilidad este entre 0 y 1.

$$p = P(y = 1|x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Para calcular los coeficientes usamos la siguiente transformación:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Note que el coeficiente β_1 **NO** representa el aumento en la probabilidad de que *Y=1* cuando *X* aumenta en una unidad.

Modelo de regresión Logística

Consideremos el ejemplo de selección de universidad:

Y= Selección=$\begin{cases} 1\ si\ selecciona\ A \\ 0\ si\ selecciona\ B \end{cases}$

X= puntaje del examen (GMAT)

Aplicando el modelo logístico obtenemos que: $log\left(\frac{p}{1-p}\right) = -48.47 + 0.0683X$

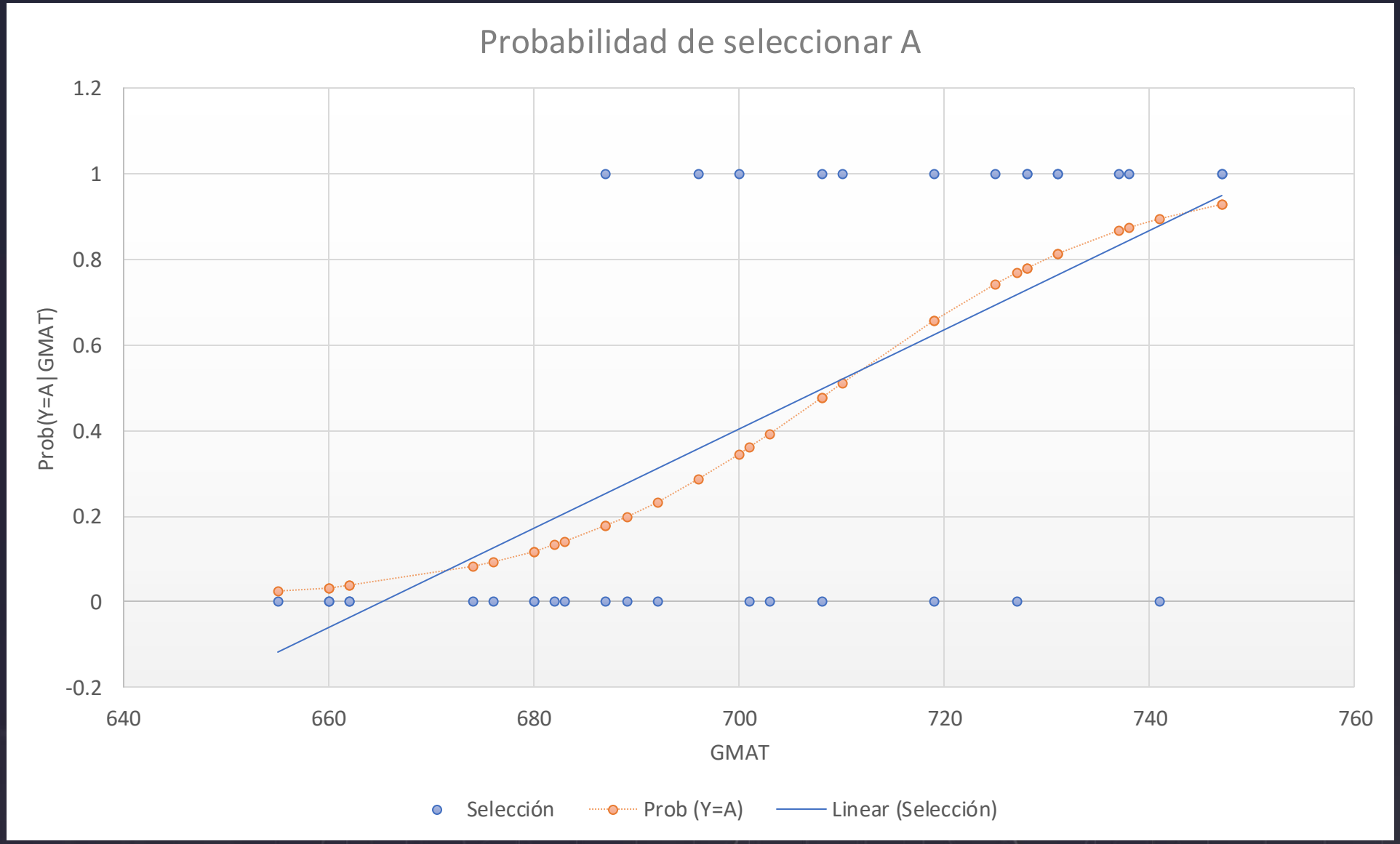
Modelo de regresión Logística

Consideremos el ejemplo de selección de universidad:

$$Prob(y = 1|GMAT) = \frac{e^{-48.47+0.0683\times GMAT}}{1+e^{-48.47+0.0683\times GMAT}}$$

Predicción de Probabilidad de seleccionar "A" dado un nivel de GMAT		
GMAT	**Prob (y=A|GMAT)**	**Prob (y=A|GMAT)**
700	$\frac{e^{-48.47+0.0683\times 700}}{1+e^{-48.47+0.0683\times 700}}$	34.46%
650	$\frac{e^{-48.47+0.0683\times 650}}{1+e^{-48.47+0.0683\times 650}}$	1.7%

Modelo de regresión Logística



Modelo de regresión Logística

Imagina que una tienda online quiere estimar la **probabilidad de que un cliente compre un producto (Compra = 1)** en función de su **nivel de ingreso (en millones de pesos)**.

Cliente	Ingreso (millones)	Compra (1 = sí, 0 = no)
1	1.0	0
2	1.5	0
3	2.0	0
4	3.0	1
5	3.5	1
6	4.0	1

Modelo de regresión Logística

Modelo

Proponemos un modelo lineal simple:

$$P(\text{Compra}_i = 1) = \beta_0 + \beta_1 \times \text{Ingreso}_i + \varepsilon_i$$

Supongamos que al estimarlo obtenemos:

$$\hat{P}(\text{Compra}_i = 1) = -0.8 + 0.4 \times \text{Ingreso}_i$$

Interpretación

- **β_0 = -0.8**: si el ingreso fuera 0, la probabilidad predicha de compra sería -0.8 (no tiene sentido práctico, pero sirve para ilustrar la limitación del modelo).
- **β_1 = 0.4**: cada aumento de **1 millón en ingreso** aumenta la **probabilidad de compra en 0.4 (40 puntos porcentuales)**.

Modelo de regresión Logística

Modelo

Proponemos un modelo lineal simple:

$$P(\text{Compra}_i = 1) = \beta_0 + \beta_1 \times \text{Ingreso}_i + \varepsilon_i$$

Supongamos que al estimarlo obtenemos:

$$\hat{P}(\text{Compra}_i = 1) = -0.8 + 0.4 \times \text{Ingreso}_i$$

Ingreso	Probabilidad estimada	Resultado
1.0	-0.8 + 0.4×1 = **-0.4**	❌ (valor inválido)
2.0	-0.8 + 0.4×2 = **0.0**	0%
3.0	-0.8 + 0.4×3 = **0.4**	40%
4.0	-0.8 + 0.4×4 = **0.8**	80%

Modelo de regresión Logística

Conclusión

✅ **Ventajas:**

•Fácil de entender e interpretar (una recta).

•Muestra cómo cambia la probabilidad esperada según una variable explicativa.

⚠️ **Limitaciones:**

•Puede predecir probabilidades **menores que 0 o mayores que 1**, lo que **no tiene sentido**.

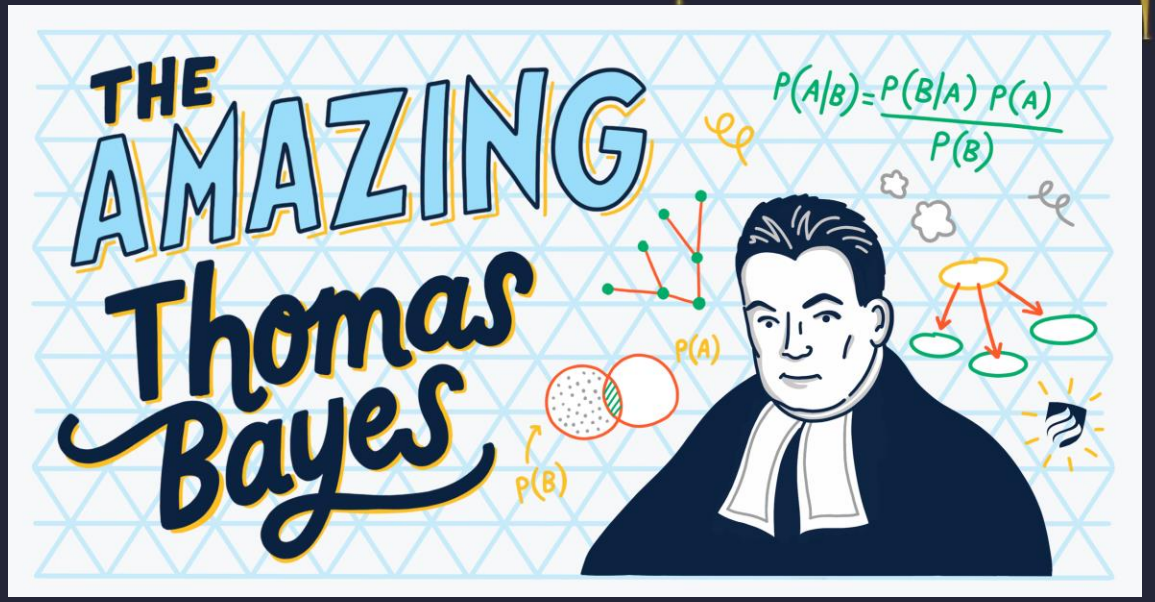
•No asegura que los errores tengan varianza constante (problema de heterocedasticidad).

👉 Por eso luego se introduce el **modelo logit o probit**, que corrigen estos problemas manteniendo la interpretación probabilística.

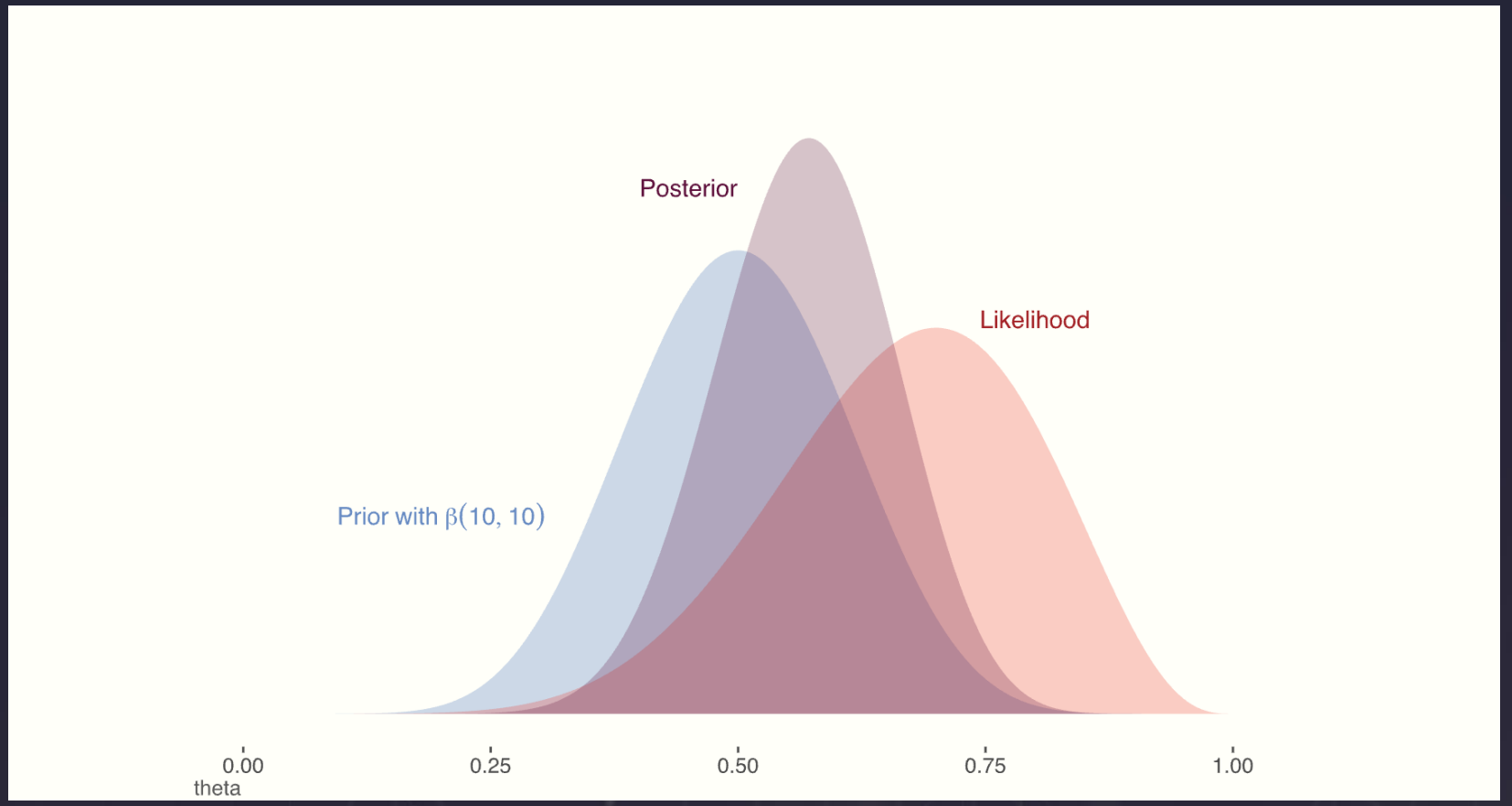
Teras: ULTIMO CASO

Predicting Customer Churn at QWE INC

BAYESIAN ESTIMATION



BAYES


Posterior
Likelihood
Prior with β(10, 10)
0.00
0.25
0.50
0.75
1.00
theta

Bayes theorem tells us how to use data to make inference on parameters of interest (i.e, θ which is often a vector of parameters)

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

The likelihood defines how data depend on parameters of interest

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

The prior is the strength of our belief in each possible value of θ before the data has been taken into account

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

The posterior is the strength of our belief in each possible value of θ after the data has been taken into account

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

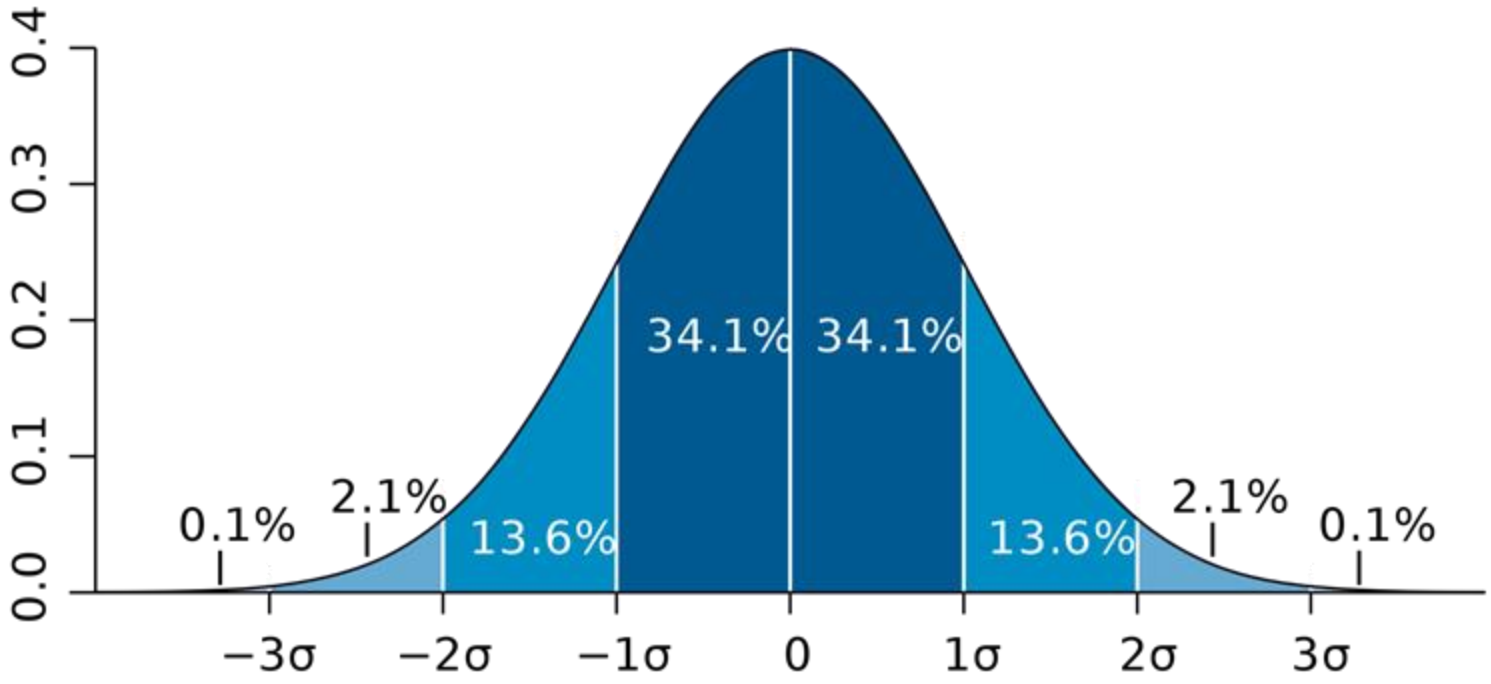
Key characteristics about priors

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

- Influence inference
- Are subjective: there's no "right" prior
- Are a probability distribution

A probability distribution defines probabilities of different possible outcomes of a random variable

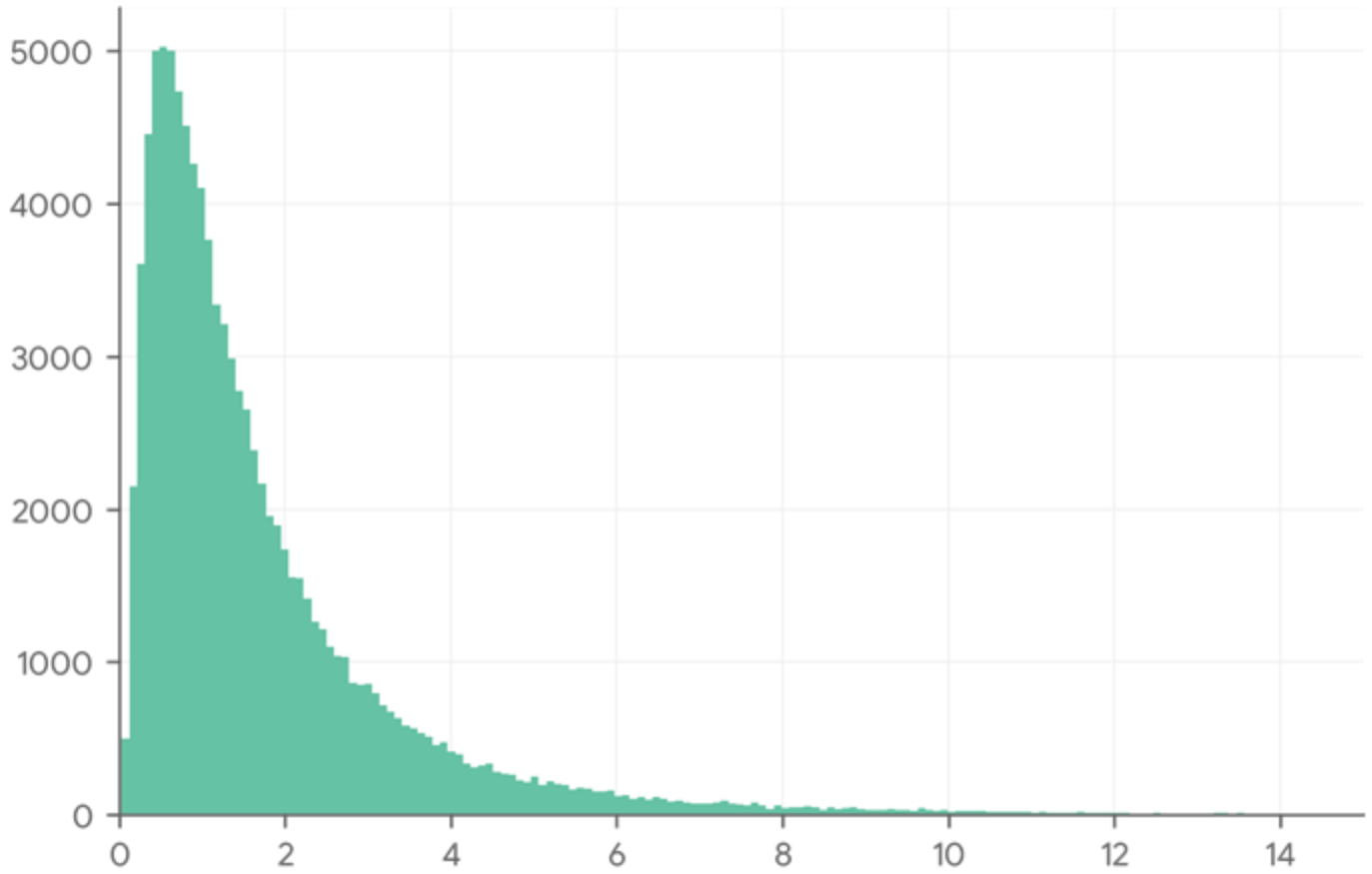
- For discrete distributions, the sum of the probabilities of all possible outcomes equals 1
- For continuous distributions, the integral of the probabilities of all possible outcomes equals 1
- The area under the curve for any range of numbers is the probability of an outcome in that range



Our default media priors are “moderately informative”

- Doesn’t allow negative ROIs, which are hard to believe and to explain. Negative ROIs may make other channels have unbelievably high ROIs.
- States our *a priori assumptions* that
  - Median of advertising ROIs is 1.22
  - 80% of ROIs are between 0.5 and 6.0
  - 99% of ROIs are below 10
- **Summary:** our default prior regularizes results towards something we know to be reasonable, while also allowing the data to have influence

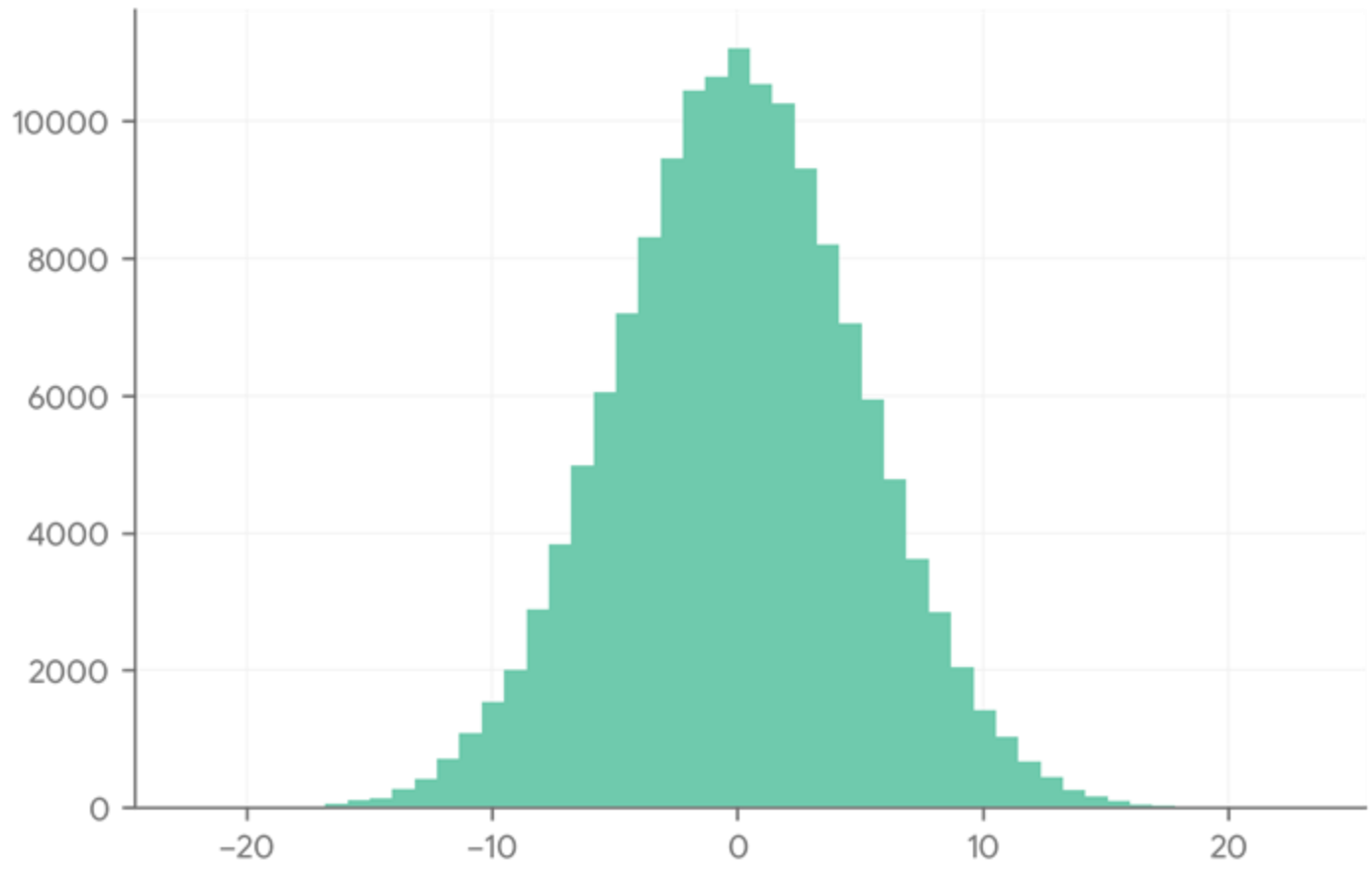
Default ROI prior: LogNormal(0.2, 0.9)



Our default control priors are “uninformative”

- Control variables are non-media variables used to debias the causal estimates of media
- The kinds of controls a user may put in the model is wide-ranging
- The impact a control can have on the KPI can be highly dependent on the control variable and on the advertiser

Default prior on the effect of a “control” variable



Our philosophy around priors

- Short justifications for each of our default priors can be found here. We tend to like moderately informative default priors when possible.
- Priors on media effects can be calibrated using the results of past experiments.

JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D., Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D., Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D., Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

ABSTRACT

BACKGROUND

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and the resulting coronavirus disease 2019 (Covid-19) have afflicted tens of millions of people in a worldwide pandemic. Safe and effective vaccines are needed urgently.

METHODS

In an ongoing multinational, placebo-controlled, observer-blinded, pivotal efficacy trial, we randomly assigned persons 16 years of age or older in a 1:1 ratio to receive two doses, 21 days apart, of either placebo or the BNT162b2 vaccine candidate (30 μg per dose). BNT162b2 is a lipid nanoparticle–formulated, nucleoside-modified RNA vaccine that encodes a prefusion stabilized, membrane-anchored SARS-CoV-2 full-length spike protein. The primary end points were efficacy of the vaccine against laboratory-confirmed Covid-19 and safety.

RESULTS

A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo; BNT162b2 was 95% effective in preventing Covid-19 (95% credible interval, 90.3 to 97.6). Similar vaccine efficacy (generally 90 to 100%) was observed across subgroups defined by age, sex, race, ethnicity, baseline body-mass index, and the presence of coexisting conditions. Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient. The safety profile of BNT162b2 was characterized by short-term, mild-to-moderate pain at the injection site, fatigue, and headache. The incidence of serious adverse events was low and was similar in the vaccine and placebo groups.

CONCLUSIONS

A two-dose regimen of BNT162b2 conferred 95% protection against Covid-19 in persons 16 years of age or older. Safety over a median of 2 months was similar to that of other viral vaccines. (Funded by BioNTech and Pfizer; ClinicalTrials.gov number, NCT04368728.)

The authors' affiliations are listed in the Appendix. Address reprint requests to Dr. Absalon at Pfizer, 401 N. Middletown Rd., Pearl River, NY 10965, or at judith.absalon@pfizer.com.

*A complete list of investigators in the C4591001 Clinical Trial Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Polack and Thomas contributed equally to this article.

This article was published on December 10, 2020, and updated on December 16, 2020, at NEJM.org.

N Engl J Med 2020;383:2603-15.
DOI: 10.1056/NEJMoa2034577

STATISTICAL ANALYSIS

The safety analyses included all participants who received at least one dose of BNT162b2 or placebo. The findings are descriptive in nature and not based on formal statistical hypothesis testing. Safety analyses are presented as counts, percentages, and associated Clopper–Pearson 95% confidence intervals for local reactions, systemic events, and any adverse events after vaccination, according to terms in the *Medical Dictionary for Regulatory Activities* (MedDRA), version 23.1, for each vaccine group.

Analysis of the first primary efficacy end point included participants who received the vaccine or placebo as randomly assigned, had no evidence of infection within 7 days after the second dose, and had no major protocol deviations (the population that could be evaluated). Vaccine efficacy was estimated by $100 \times (1 - \mathrm{IRR})$, where IRR is the calculated ratio of confirmed cases of Covid-19 illness per 1000 person-years of follow-up in the active vaccine group to the corresponding illness rate in the placebo group. The 95.0% credible interval for vaccine efficacy and the probability of vaccine efficacy greater than 30% were calculated with the use of a Bayesian beta-binomial model. The final analysis uses a success boundary of 98.6% for probability of vaccine efficacy greater than 30% to compensate for the interim analysis and to control the overall type 1 error rate at 2.5%. Moreover, primary and secondary efficacy end points are evaluated sequentially to control the

Let’s talk about the likelihood portion of Bayes Theorem

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

Our model (simplified)

KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Our model (unsimplified)

$$y_{g,t} = \mu_t + \tau_g + \sum_{c=1}^{C} Y_{g,c} z_{g,t,c}$$

$$+ \sum_{m=1}^{M} \beta_{g,m} HillAdstock(\{x_{g,t-s,m}\}_{s=0}^{L}; \alpha_m, ec_m, slope_m)$$

$$+ \sum_{n=1}^{N} \beta_{g,n}^{(rf)} Adstock(\{r_{g,t-s,n} \cdot Hill(f_{g,t-s,n}; ec_n^{(rf)}, slope_n^{(rf)})\}_{s=0}^{L}; \alpha_n^{(rf)})$$

$$+ \epsilon_{g,t}$$

Our model (simplified)

KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Modeling geo effects in Meridian

- Geo effects: dummy indicators $\boldsymbol{\tau} = [\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_{G-1}]$ for all geos (other than a baseline geo)

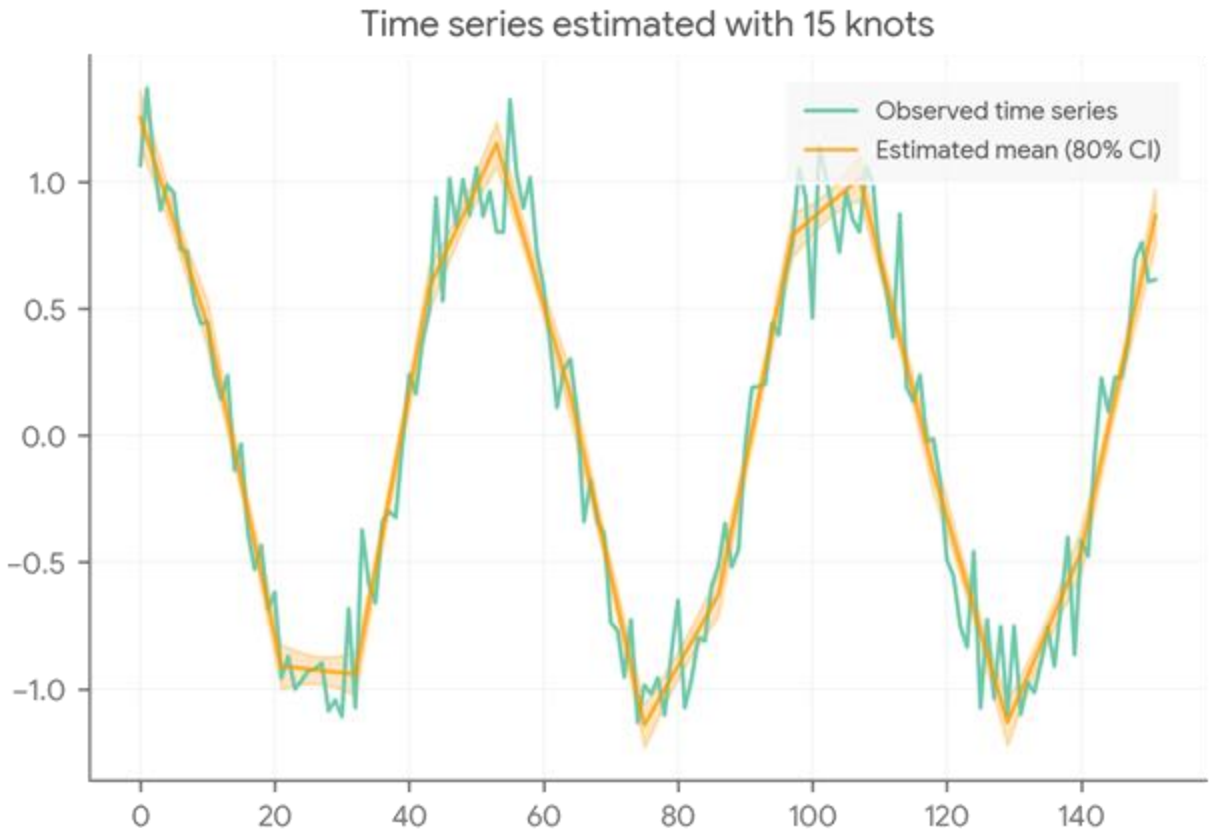
Our model (simplified)

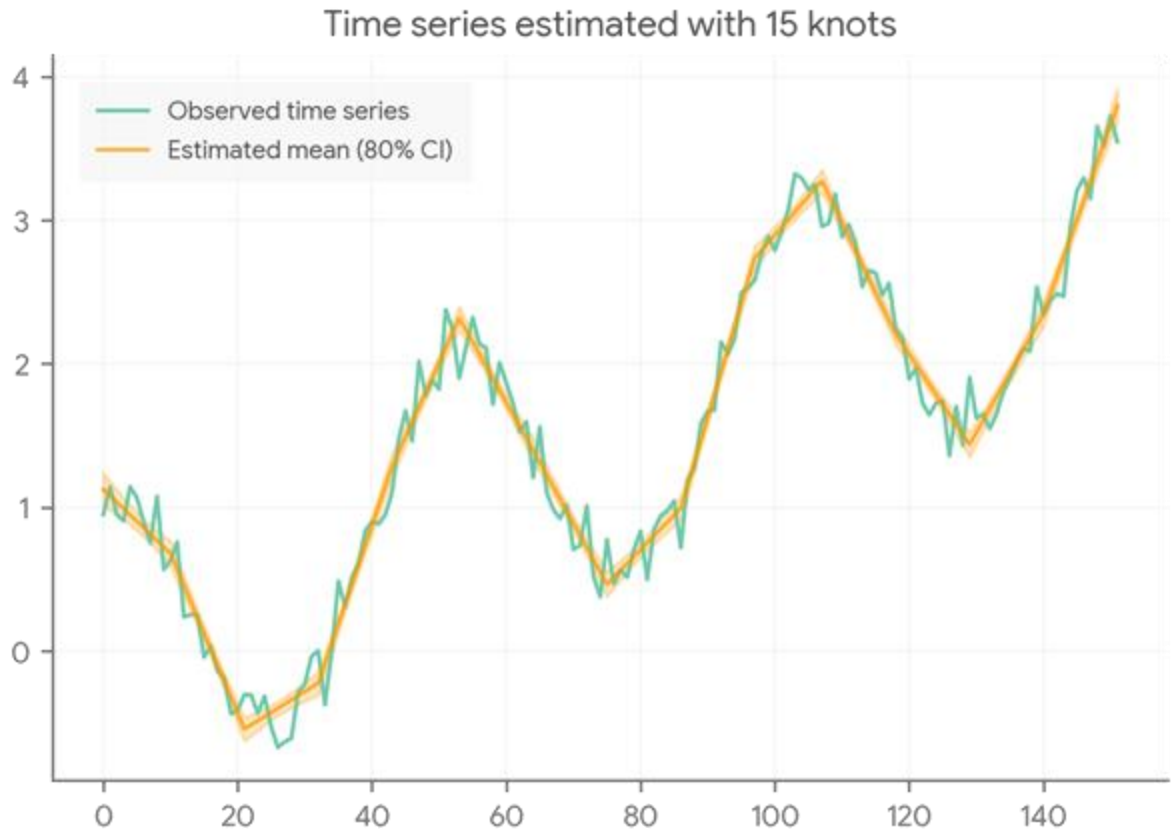
KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Modeling time effects in Meridian

- Time effects: $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_T]$ for each of the *T* time periods
  - We assume time effects are the same for all geos
- K ≤ T many knots are set at times within the range 1, 2, …, T
  - Knots are a hyperparameter: users can set either the knot locations or the number of evenly spaced knots
- Each τ_t is a weighted average of its two neighboring knots; with closer knots getting more weight (L1 distance)
  - E.g., say there are knots at time 9 and time 18. The estimate for τ_{16} will be influenced by both the knot at time 9 and the knot at time 18, with the knot at time 18 getting more influence.
- Bayesian inference is done on K ≤ T knot parameters instead of the T many τ parameters

Knot-based approach can capture trend and seasonality; even with relatively few knots





Pros and cons of the knot-based approach

Pros

- Very flexible
- Captures arbitrarily complex trend and seasonality patterns
- Captures multiple trend and seasonality patterns simultaneously

Cons

- Tends to be over-parameterized relative to models built to capture specific kinds of seasonality
- The number of knots is a hyperparameter

Our model (simplified)

KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Hierarchical modeling of media effects in Meridian

- Media effects: $\beta_{m,g}$ for each media channel *m* and each geo *g*
  - Media is assumed to have a different effect for each geo
- Geo-effects within a media channel are hierarchical, meaning that although each geo has its own media effect, the geos are informing each other and sharing data
- Partial pooling is a compromise between two extremes:
  - Complete pooling: assumes media effect is identical across geos
  - No pooling: assumes all geos are completely different from each other
- The hierarchical model learns how similar the geos are
  - If geos tend to have very similar media effects, the model understands that it can use data from geo 1 to influence the estimate for geo 2 (b/c the effects are similar)
  - If geos tend to have similar media effects, except for perhaps an outlier geo, that outlier geo will be regularized towards the rest of the geos
  - This is all done according to probability rules and Bayes theorem!

Pros and cons of a hierarchical model

Pros:

- Granular geo-level results
- Better inference for the causal effect of media
- Narrower credible intervals

Cons

- Requires geo-granular data for media execution, controls, and KPI
- Fitting the model is slower than a non-hierarchical model would be

Saturation and carryover effects of media in Meridian

$$\beta_{m,g}\, Hill(Adstock(x^{*}_{t,m,g}, \alpha_m, L), ec_m, slope_m)$$

- Data: $x_{t,\,m,\,g}$ is media execution (clicks, impressions, etc) at time period *t*, for media channel *m*, in geography *g*
  - $x^{*}_{t,\,m,\,g} = \{x_{s,\,m,\,g}\,,\ t - L \le s \le t\}$ is time series
- Adstock function aggregates the media execution time series so that the KPI at time *t* can be impacted by media execution before time *t* (i.e., carryover effects)
- Hill function adjusts the aggregated media execution so that relationship between the amount of media execution and the KPI is non-linear (i.e., media execution has diminishing returns and eventually is saturated)

Model can support two-dimensional reach and frequency or one-dimensional media execution

$$\beta_{m,g} Adstock(r^*_{t,m,g} Hill^*(f^*_{t,m,g}, K_m, S_m), \alpha_m, L)$$

- Data: $r^*_{t,m,g}$ and $f^*_{t,m,g}$ are reach and average frequency time series
- Model assumes the relationship between the KPI and frequency can become saturated
- Model assumes a linear relationship between KPI and reach

Our model (simplified)

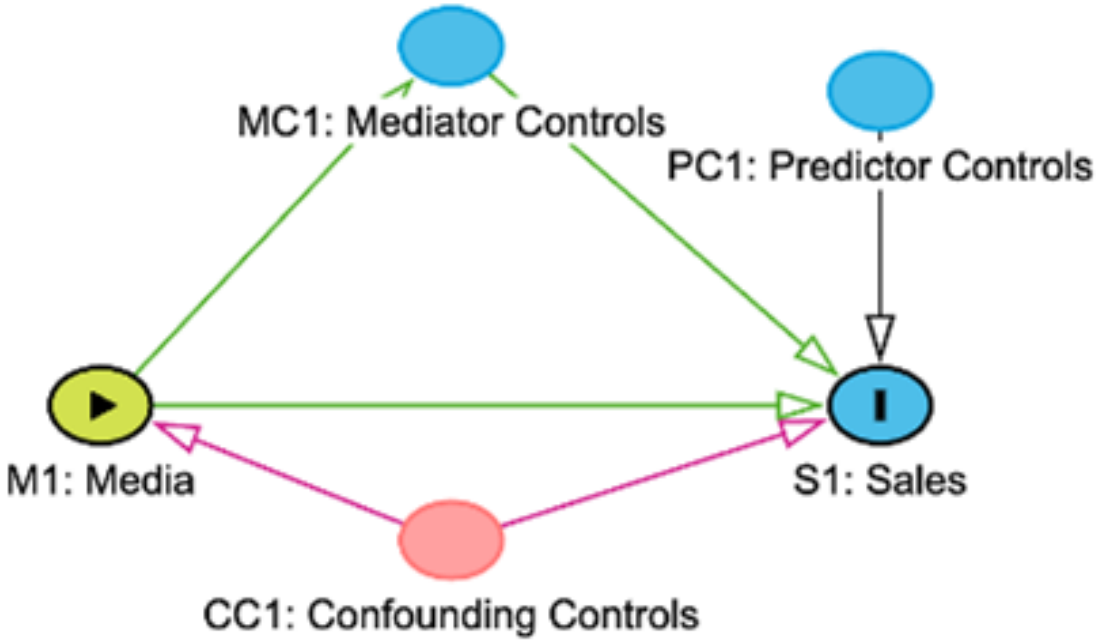
KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Hierarchical modeling of control effects in Meridian

- Control effects: $\gamma_{c,g}$ for each control variable *c* and each geo *g*
  - Control is assumed to have a different effect for each geo
- Geo-effects within a control are hierarchical
- No Hill/Adstock transformation is done on the control variables

How to select control variables: assume a causal diagram

- DAG tells us how we should select control variables in order to make **causal claims** (even with observational data):
  - We *must* include CC
  - We *must not* include MC
  - We c*an* include PC



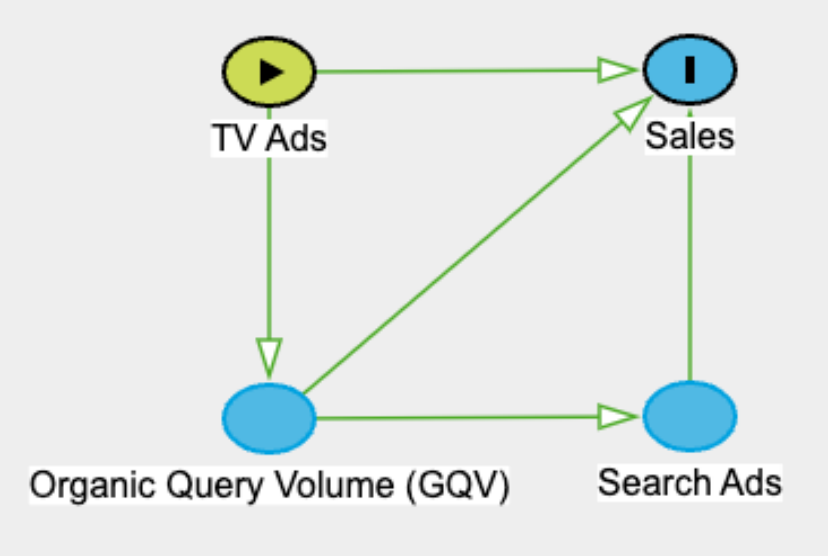
Correlation is not causation (or, prediction is not causation)

- For prediction problems, include every variable that points towards sales!
- Naive thinking: confounder variables are a *type* of variable that points towards sales, so it's tempting to think that if you solve a prediction problem you get a causal inference solution for free.
- The naive thinking is wrong:
  - You can get good prediction while missing an important confounder
  - You can get good prediction while including a mediator variable

Just because we know how to *predict* sales,
doesn't mean we know how to *increase* sales

Correlation is not causation (or, prediction is not causation)

- Simulate TV, GQV, and SEARCH according to the DAG on the right
- SALES is generated according to a linear model: SALES = TV + GQV + SEARCH
- The true causal effect of TV on SALES is 1.9 when considering it's direct effect on SALES (1.0) and it's indirect effect on SALES via GQV and SEARCH (0.9)
- Consider every possible combination of TV, GQV, and Search and fit a linear model on a large train set; assess predictive accuracy on a disjoint test set and assess causal inference estimate of TV on SALES



Our model (simplified)

KPI = Geo Effects +
Time Effects +
{Media x Geo} Effects +
{Control x Geo} Effects +
Noise

Gaussian noise with option to make it geo-dependent

$$y_{g,t} = \text{geo, time, media, and control effects} + \varepsilon_{g,t}$$

- Choice between
  - All KPI across geos and time are assumed to have same noise ($\varepsilon_{g,t} \sim N(0,\boldsymbol{\sigma})$)
  - All KPI across time, within a geo, are assumed to have same noise ($\varepsilon_{g,t} \sim N(0,\boldsymbol{\sigma}_g)$)

Let’s talk about the output of the model

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)d\theta}$$

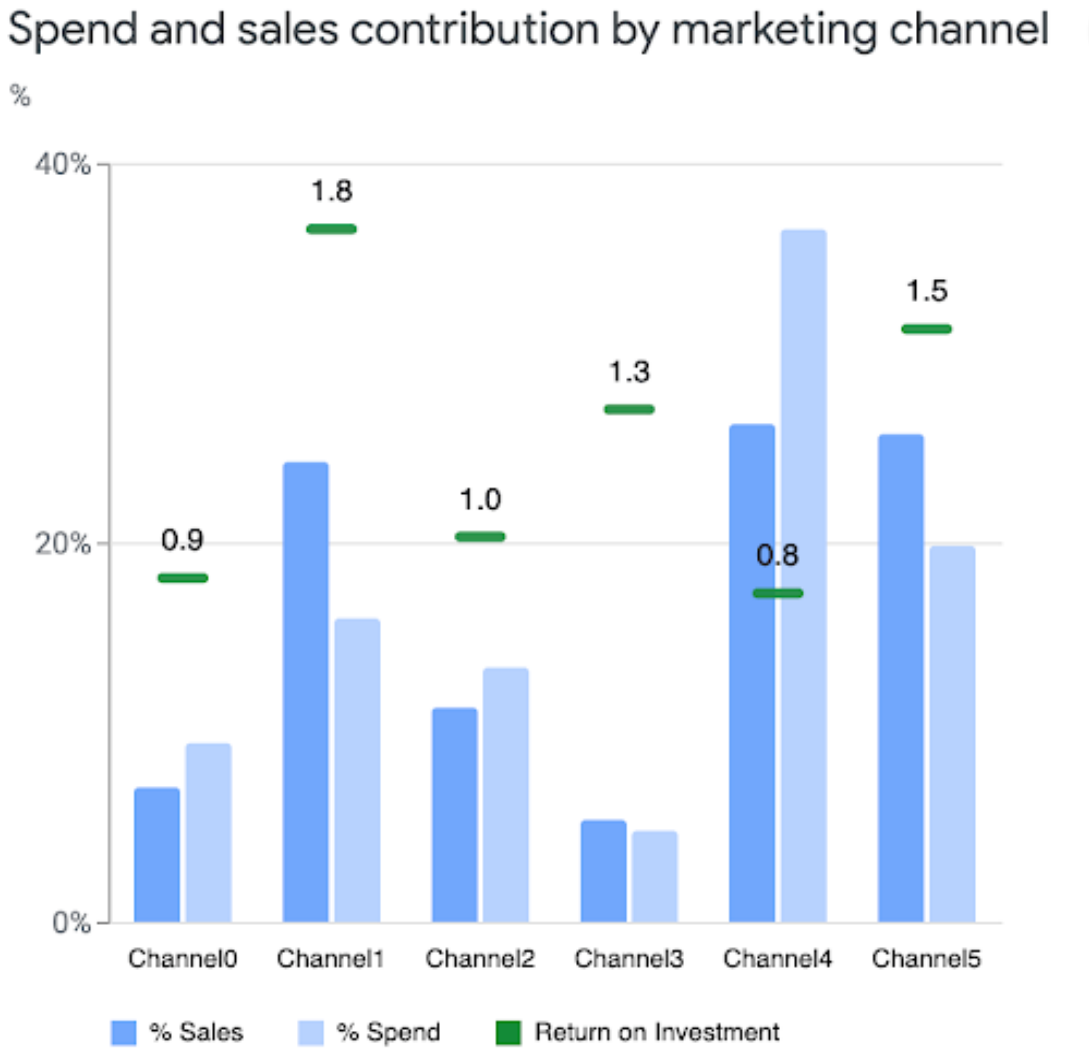
Bayesian models typically return posterior samples

- The output of Bayes theorem, the **posterior**, is a probability distribution
  - Common probability distributions: Normal, Binomial, Poisson
- In most practical situations, the integral at the bottom of Bayes theorem is extremely complex and intractable
  - Consequence: the posterior is not a cleanly stated distribution like Normal, Binomial, etc.
- Instead, **Markov Chain Monte Carlo (MCMC)** theory tells us how to get **posterior samples** θ_1 , …, θ_S for model parameters θ and any arbitrarily large value of S. Sampling from the posterior is powerful:
  - The mean of θ can be determined by taking the mean of θ_1 , …, θ_S
  - The median of θ can be determined by taking the median of θ_1 , …, θ_S

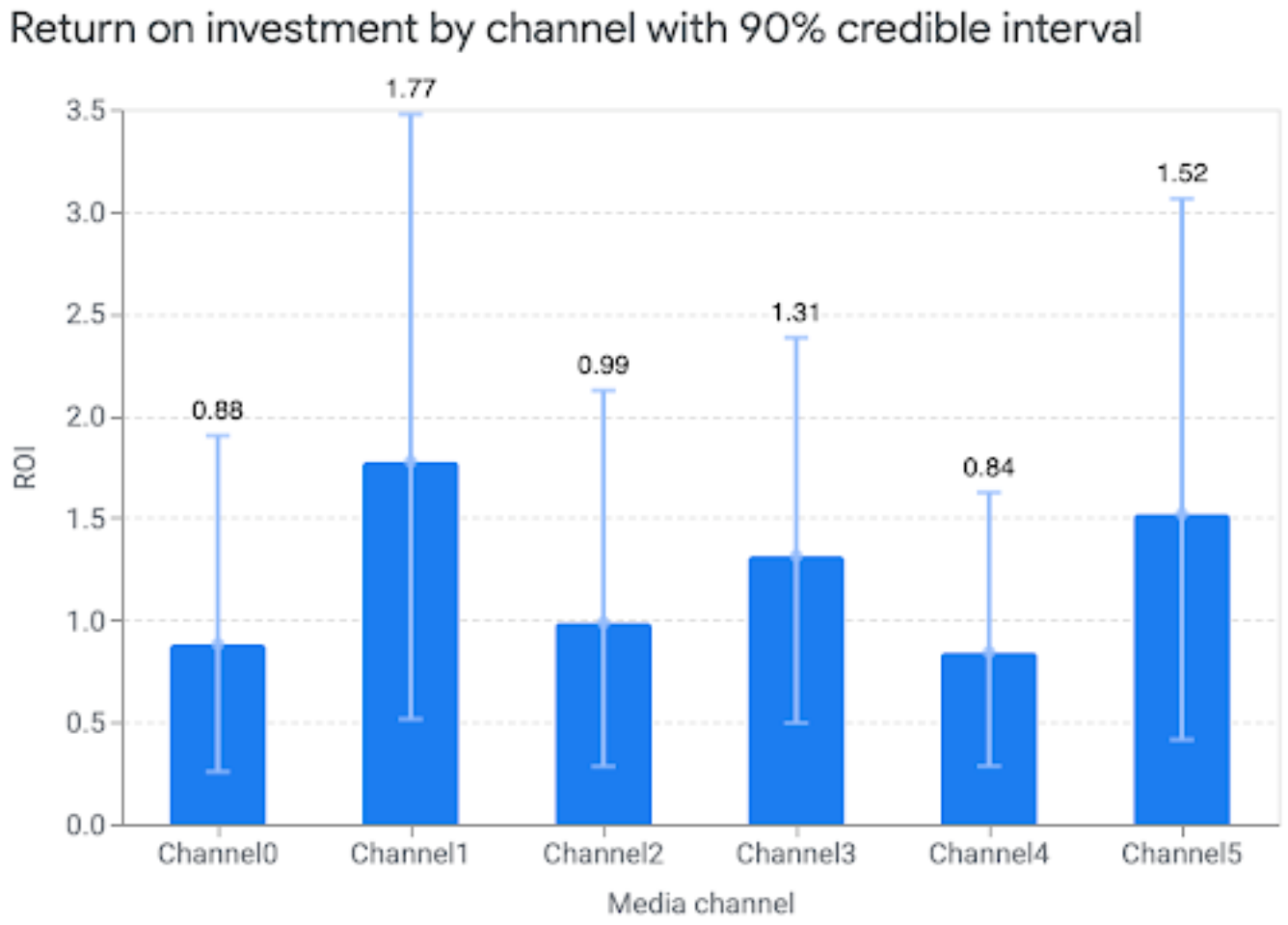
“Post processing” on posterior samples helps us interpret results

- Given posterior samples θ_1 , …, θ_S of some parameter θ, you can get a posterior distribution for $f(\theta)$ by calculating $f(\theta_1)$, …, $f(\theta_S)$
- Meridian does a lot of “post processing” of these posterior samples to help users interpret the model results:
  - `Analyzer.ROI`: full posterior distribution for ROI
  - `Analyzer.expected_impact`: full posterior distribution for the expected KPI (or expected revenue if `revenue_per_kpi` provided) given any set of media or control variables

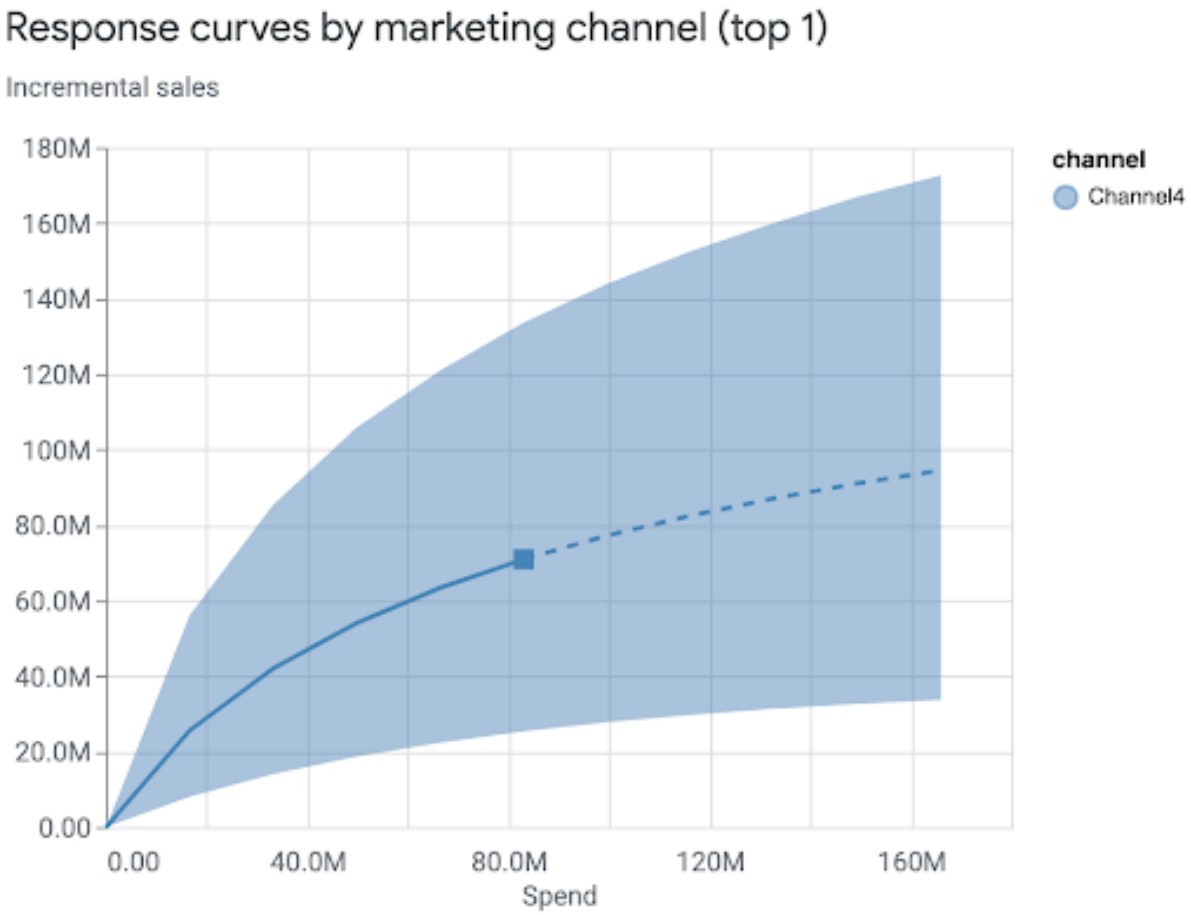
Visualization of posterior inference in Meridian

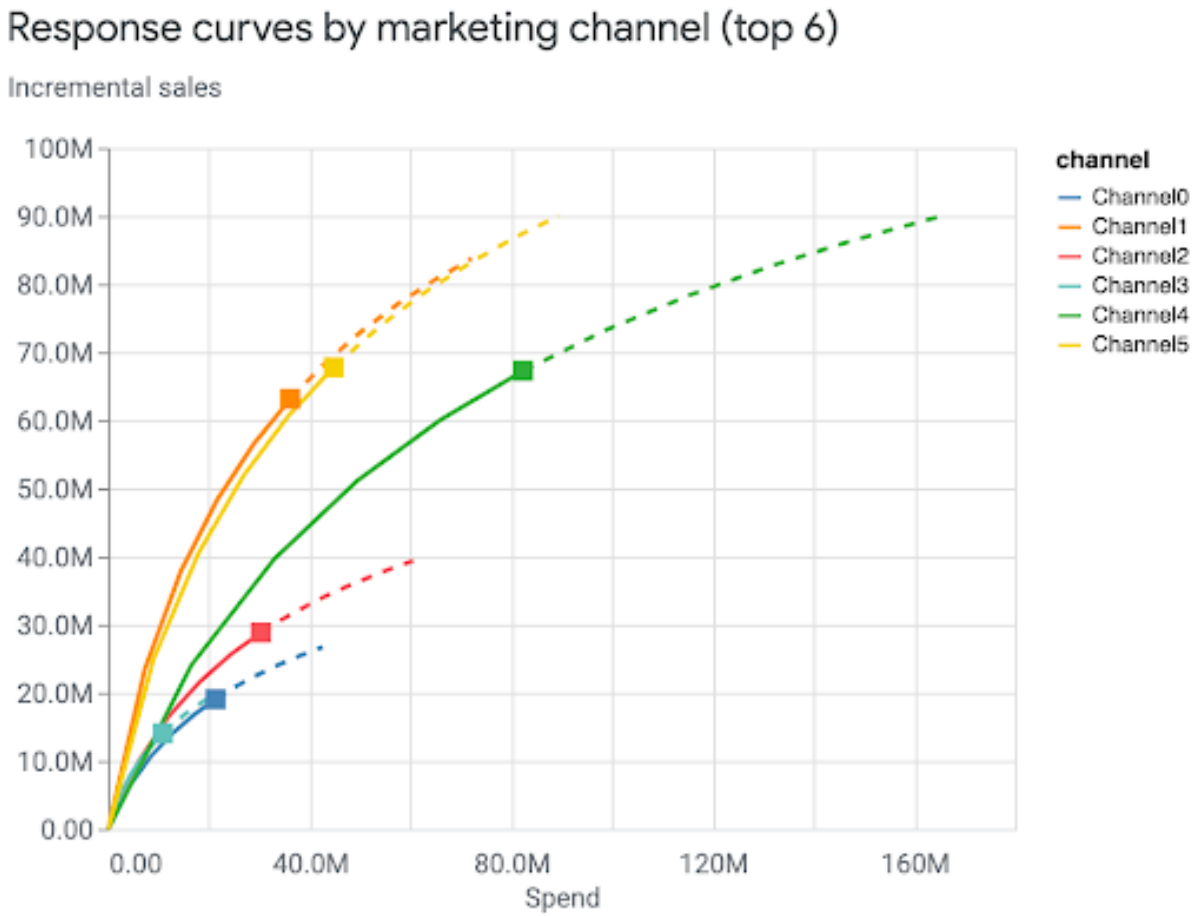


Visualization of posterior inference in Meridian



Visualization of posterior inference in Meridian





What are the pain points of using a Bayesian MMM for causal inference?

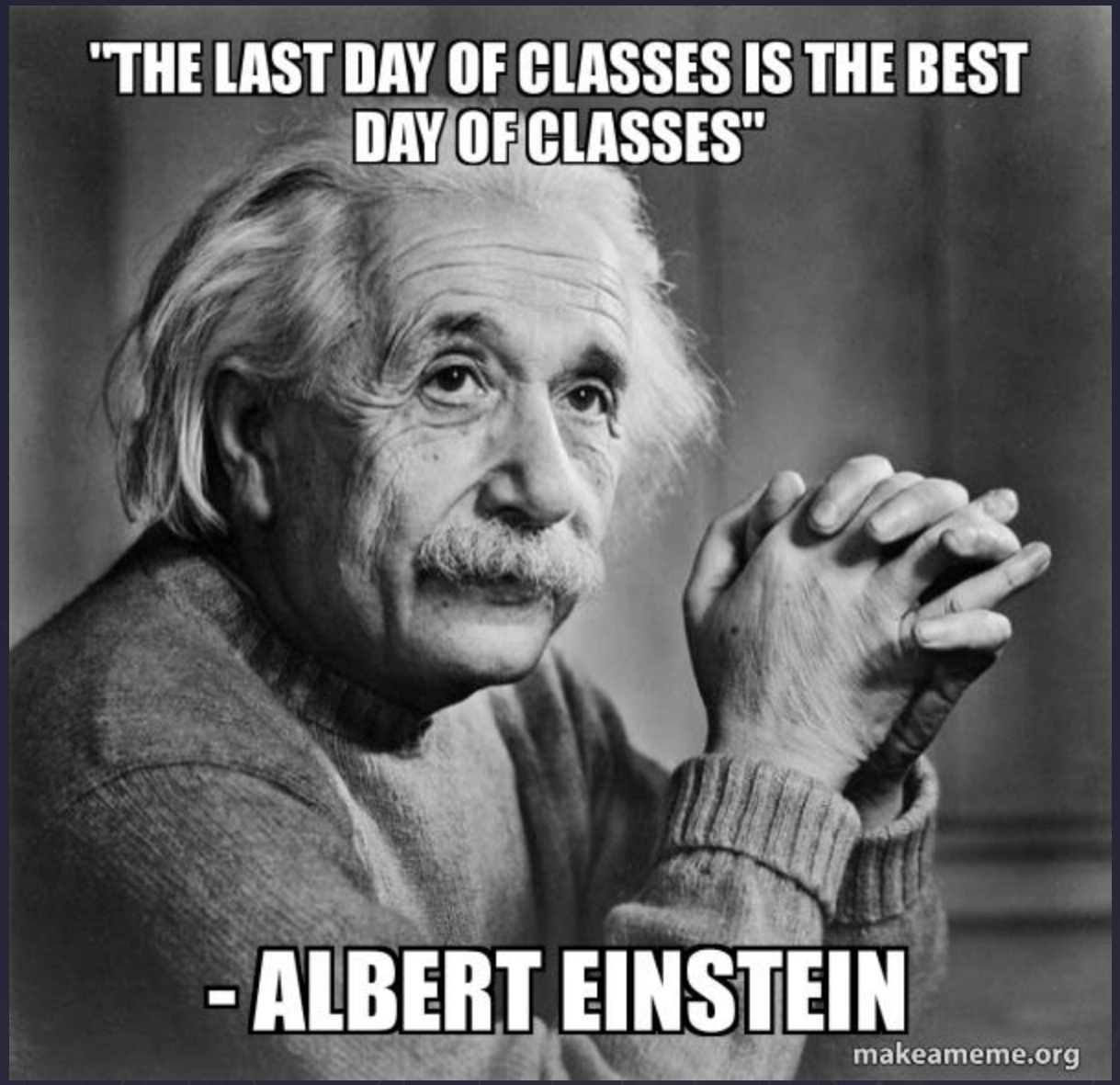
- MCMC convergence: MCMC theory tells us that we are sampling from the posterior as the number of samples *S* goes to infinity. How do we know *S* is large enough?
    - Research effort put into Meridian to make convergence easier.
    - Tools to check convergence.
- Our model has some hyperparameters that *can* be tuned: number of knots, default priors, model settings, etc.
    - Research effort put into meaningful defaults.
- Using observational data to make causal claims requires strong assumptions.

Resources for more details

- Our priors
- Our model specification
- Time-effects
- Geo-hierarchy
- Adstock/Hill (or, Saturation and Carryover)
- Reach and frequency
- ROI priors
- Getting started with using the code

ULTIMO CASO

Data Analytics Simulation: Strategic Decision Making


"THE LAST DAY OF CLASSES IS THE BEST DAY OF CLASSES"
- ALBERT EINSTEIN
makeameme.org