



Semana 2

Especificaciones de Regresión Lineal

DCDPP - Datos para la evaluación de Políticas Públicas | PUC | 26 de septiembre, 2022

 **Pablo A. Celhay** |  pacelhay@uc.cl

Outline

1. Sesgo por variable omitida
2. Modelo general de Regresión Múltiple
 - Bondad de ajuste
 - Supuestos
 - Variables dummy
 - Test de hipótesis
3. Regresión no lineal
 - Interacciones
 - Logaritmos
 - Especificaciones cuadráticas
4. Modelo de probabilidad lineal
5. Modelo probit
6. Modelo logit

1. Sesgo por variable omitida

Sesgo por variable omitida

Supuesto principal para estimación insesgada:

$$E[\mu_i | X_i] = 0$$

- μ_i, X_i son independientes
- Si hay otras variables en μ_i que se relacionan con X_i entonces → sesgo

Si (S1) no se cumple entonces este sesgo le llamamos "sesgo por variable omitida"

Sesgo por variable omitida

- La FRP de interés es:

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

- Pero estimamos:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$$

- ¿Cómo es la relación de $\hat{\alpha}_1$, nuestro estimador, con β_1 , el valor verdadero?

Sesgo por variable omitida

- Recordemos que para que $\hat{\alpha}_1$ recupere el valor verdadero del "efecto" de X_1 sobre Y entonces $\hat{\alpha}_1$ debe ser un estimador insesgado de β_1 .
 $\rightarrow E[\hat{\alpha}_1] = \beta_1$
- $\hat{\alpha}_1$ estará sesgado siempre cuando:
 1. X_{2i} es relevante para Y_i
 2. X_{2i} esta correlacionado(a) con X_{1i}
- Si (1) **y** (2) se cumplen entonces $\hat{\alpha}_1$ es un estimador sesgado de β_1 .
 - Esto es crucial al mirar un reporte o estudio con regresión lineal que afirma estimar un efecto causal
 - Típicamente no se controla por este potencial sesgo porque a veces no tenemos datos sobre X_{2i} (no observable)

Sesgo por variable omitida

- Magnitud del sesgo: $Sesgo = \alpha_1 - \beta_1$
- Corramos (imaginariamente) la siguiente regresión

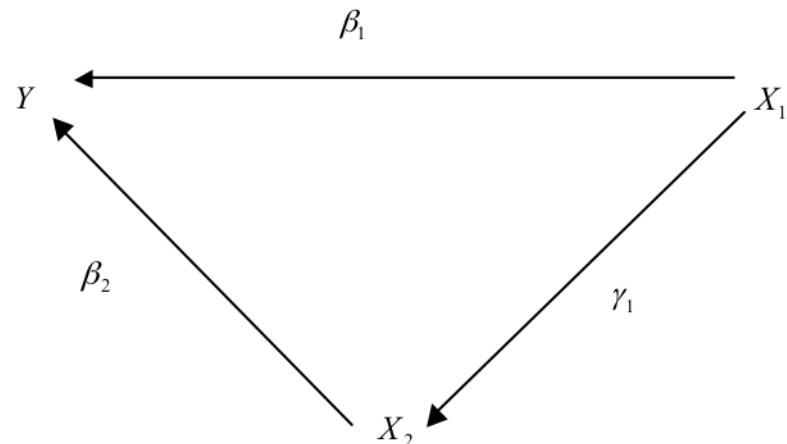
$$X_{2i} = \gamma_0 + \gamma_1 X_{1i} + w_i$$

- ¿Qué es γ_1 ?
- Sustituir esto en la FRP y obtenemos: $\alpha_1 = \beta_1 + \gamma_1 \cdot \beta_2$
- Sustituir esto en la fórmula de sesgo y obtenemos: $Sesgo = \gamma_1 \cdot \beta_2$

Discusión en clase: ¿De qué depende el sesgo?

Sesgo por variable omitida

- Magnitud del sesgo: $Sesgo = \alpha_1 - \beta_1 = \gamma_1 \cdot \beta_2$
- Si $\gamma_1 = 0$ o $\beta_2 = 0$ entonces no hay sesgo por variable omitida:
- Intuición
 1. $\beta_2 = 0$: Entonces X_2 no debería ser parte del modelo inicialmente
 2. $\gamma_1 = 0$: Entonces no hay correlación entre X_1 y X_2
- En términos de magnitudes, ¿un sesgo grande de qué depende?



Sesgo por variable omitida

- Signo del sesgo:
 - El signo o la dirección del sesgo depende de los signos de γ_1 y β_2
 - Lo primero a notar es que el signo de γ_1 es el signo de la correlación entre X_1 y X_2 , entonces el signo del sesgo depende de $\text{corr}(X_1, X_2)$ y de β_2 .
 - Dado que β_2 es un parámetro poblacional no podemos saber exactamente si es positivo o negativo. Y si X_2 no se observa no podemos saber exactamente cual es la $\text{corr}(X_1, X_2)$.
- Sin embargo podemos hacer algunas conjeturas basándonos en la intuición sobre el problema bajo estudio

Cuadro 1: Signo del sesgo por variable omitida

$\text{corr}(X_1, X_2) > 0$		
$\beta_2 > 0$		
$\beta_2 < 0$		

Sesgo por variable omitida

- Para tener una idea y afirmar si es que una variable omitida estaría generando una subestimación o sobre estimación del efecto es necesario conocer el signo del sesgo y además el signo de β_1
- Si es que el sesgo hace que nuestro estimador sea mayor (en valor absoluto) que lo que debería ser estamos sobre estimando el efecto de X sobre Y . Esto pasa cuando β_1 y el sesgo tienen el mismo signo

Sesgo por variable omitida: Ejemplo 1

Estudiaremos la efectividad de un programa de control prenatal antes de los 2 meses de embarazo sobre el peso del niño o niña al nacer.

Variables:

- `bwght`: peso al nacer (en gramos)
- `monpre_bin`: = 1 si recibe tratamiento; = 0 si no recibe tratamiento.
- `cigs`: Promedio de consumo diario de cigarrillos

Variable	Obs	Media	Std	Min	Max
bwght	1827	3401.600	576.919	360	5204
monpre_bin	1827	0.767	0.423	0	1
cigs	1827	1.092	4.227	0	40

Sesgo por variable omitida: Ejemplo 1

- Signo del sesgo:
 - El signo o la dirección del sesgo depende de los signos de γ_1 y β_2
 - Lo primero a notar es que el signo de γ_1 es el signo de la correlación entre X_1 y X_2 , entonces el signo del sesgo depende de $\text{corr}(X_1, X_2)$ y de β_2 .
 - Dado que β_2 es un parámetro poblacional no podemos saber exactamente si es positivo o negativo. Y si X_2 no se observa no podemos saber exactamente cual es la $\text{corr}(X_1, X_2)$.

Discusión en clase: ¿Qué podríamos saber con estos datos?

Correlación entre edad (omitida) y tratamiento (X_1)

Sesgo por variable omitida: Ejemplo 1

Correlación entre variable omitida (X_2) y tratamiento (X_1)

```
cor.test(df$monpre_bin, df$cigs, method = "pearson")
```

Pearson's product-moment correlation

```
data: df$monpre_bin and df$cigs
t = -4, df = 1716, p-value = 0.0003
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1340 -0.0401
sample estimates:
cor
-0.0873
```

Sesgo por variable omitida: Ejemplo 1

```
df %>%
  group_by(monpre_bin) %>%
  dplyr::summarise(Obs=n(),
    Media=mean(cigs, na.rm=TRUE),
    Std=sd(cigs, na.rm=TRUE),
    Min=min(cigs, na.rm=TRUE),
    Max=max(cigs, na.rm=TRUE))
```

```
# A tibble: 2 × 6
  monpre_bin    Obs   Media     Std     Min     Max
  <dbl>     <int>   <dbl>   <dbl>   <int>   <int>
1       0        426    1.75    5.46     0      40
2       1      1401   0.884   3.73     0      40
```

Sesgo por variable omitida: Ejemplo 1

```
summary(lm_robust(bwght ~ monpre_bin, data = df, se_type = "stata"), digits=4)
```

Call:

```
lm_robust(formula = bwght ~ monpre_bin, data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	3378.8	26.7	126.59	0.000	3326	3431.2	1825
monpre_bin	29.7	30.9	0.96	0.337	-31	90.3	1825

Multiple R-squared: 0.000473 , Adjusted R-squared: -7.42e-05

F-statistic: 0.921 on 1 and 1825 DF, p-value: 0.337

Discusión en clase ¿Cuál es la interpretación del coeficiente sobre `monpre_bin`?

Sesgo por variable omitida: Ejemplo 1

```
m1 ← lm_robust(bwght ~ monpre_bin + cigs, data = df, se_type = "stata")
summary(m1)
```

Call:

```
lm_robust(formula = bwght ~ monpre_bin + cigs, data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	3397.8	26.97	126.00	0.000000	3344.9	3450.69	1715
monpre_bin	32.3	31.09	1.04	0.299036	-28.7	93.27	1715
cigs	-11.2	3.14	-3.58	0.000352	-17.4	-5.09	1715

Multiple R-squared: 0.00788 , Adjusted R-squared: 0.00672

F-statistic: 6.86 on 2 and 1715 DF, p-value: 0.00108

Discusión en clase: ¿Cuál es la interpretación del coeficiente sobre `monpre_bin`?

Discusión en clase: ¿Cambia respecto a la regresión anterior? ¿Cuál es la dirección del cambio y la razón del cambio en esta dirección?

Sesgo por variable omitida: Ejemplo 2

- Salarios, capacitación y educación
- Quiero estimar una regresión de salarios sobre horas de capacitación laboral y educación

Discusión en clase: ¿Cómo es la regresión?

- Suponga que en una ciudad particular, un subsidio para talleres de capacitación grande fue ofrecido a trabajadores con bajos niveles de educación así que educación y horas de capacitación están (**¿negativa o positiva?**) mente correlacionados.
- Tenemos datos de salario y horas de capacitación solamente y estimamos la siguiente regresión:

$$\text{salario}_i = \alpha_0 + \alpha_1 \text{horas cap}_i + \varepsilon_i$$

Discusión en clase: ¿Qué tan buena es la estimación de α_1 por MCO?

2. Modelo general de Regresión Múltiple

El modelo general de regresión múltiple

- Hasta ahora aprendimos:
 - ¿Qué es una regresión?
 - Interpretación de coeficientes en una regresión
 - Cómo estimamos estos coeficientes (de variables continuas o binarias)
 - Test de hipótesis para un coeficiente
 - Condiciones bajo las cuáles estos son insesgados
 - Cómo conjeturar sobre la dirección y magnitud del sesgo por variable omitida
- Ahora mantendremos todo esto pero lo generalizaremos para el caso de regresión múltiple con más de 2 variables
 - Específicamente... con k variables, donde $k = 1, \dots, K$

El modelo general de regresión múltiple

- La FRP es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \mu_i$$

- Nuevamente queremos minimizar:

$$\sum_i^n (Y_i - \hat{Y}_i)^2$$

¿Qué es eso?

- Con:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

El modelo general de regresión múltiple

El coeficiente asociado a X_{1i} mide el cambio en Y_i dado un aumento en X_{1i} , dejando constante todo el resto de las variables dle modelo $(X_{2i}, X_{3i}, \dots, X_{ki})$

Ejemplo: Peso del niño/a al nacer post tratamiento

- `bwght`: birth weight, grams
- `monpre_bin`: treatment
- `cigs`: avg cigarettes per day
- `mage`: mother's age, years
- `meduc`: mother's educ, years
- `npvis`: total number of prenatal visits
- `fage`: father's age, years
- `feduc`: father's educ, years
- `omaps`: one minute apgar score
- `fmaps`: five minute apgar score
- `drink`: avg drinks per week
- `male`: =1 if baby male
- `mblk`: =1 if mother black
- `fblk`: =1 if father black
- `magesq`: mage²
- `npvissq`: npvis²

El modelo general de regresión múltiple

Call:

```
lm_robust(formula = bwght ~ ., data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	1638.2495	502.564	3.260	0.001	652.495	2624.004	1595
monpre_bin	13.3727	33.086	0.404	0.686	-51.525	78.270	1595
cigs	-7.4611	3.129	-2.384	0.017	-13.599	-1.323	1595
mage	49.0195	26.753	1.832	0.067	-3.455	101.494	1595
meduc	-5.2332	8.515	-0.615	0.539	-21.936	11.470	1595
npvis	3.5344	10.204	0.346	0.729	-16.479	23.548	1595
fage	5.6667	3.355	1.689	0.091	-0.914	12.247	1595
feduc	6.7703	7.908	0.856	0.392	-8.741	22.282	1595
omaps	28.3770	17.492	1.622	0.105	-5.932	62.686	1595
fmaps	63.8662	36.262	1.761	0.078	-7.261	134.993	1595
drink	-18.5829	26.430	-0.703	0.482	-70.424	33.259	1595
vlbw	-1950.7856	150.810	-12.935	0.000	-2246.591	-1654.980	1595
male	84.0183	26.800	3.135	0.002	31.451	136.585	1595
mblk	-190.6614	133.310	-1.430	0.153	-452.143	70.820	1595
fblk	203.0290	128.227	1.583	0.114	-48.482	454.540	1595
magesq	-0.8710	0.440	-1.981	0.048	-1.734	-0.009	1595
npvissq	0.0932	0.322	0.289	0.773	-0.539	0.726	1595

Multiple R-squared: 0.109 , Adjusted R-squared: 0.0999

El modelo general de regresión múltiple

¿Qué representa el coeficiente de cigs?

Call:

```
lm_robust(formula = bwght ~ ., data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	1638.2495	502.564	3.260	0.001	652.495	2624.004	1595
monpre_bin	13.3727	33.086	0.404	0.686	-51.525	78.270	1595
cigs	-7.4611	3.129	-2.384	0.017	-13.599	-1.323	1595
mage	49.0195	26.753	1.832	0.067	-3.455	101.494	1595
meduc	-5.2332	8.515	-0.615	0.539	-21.936	11.470	1595
npvis	3.5344	10.204	0.346	0.729	-16.479	23.548	1595
fage	5.6667	3.355	1.689	0.091	-0.914	12.247	1595
feduc	6.7703	7.908	0.856	0.392	-8.741	22.282	1595
omaps	28.3770	17.492	1.622	0.105	-5.932	62.686	1595
fmaps	63.8662	36.262	1.761	0.078	-7.261	134.993	1595
drink	-18.5829	26.430	-0.703	0.482	-70.424	33.259	1595
vlbw	-1950.7856	150.810	-12.935	0.000	-2246.591	-1654.980	1595
male	84.0183	26.800	3.135	0.002	31.451	136.585	1595
mblk	-190.6614	133.310	-1.430	0.153	-452.143	70.820	1595
fblk	203.0290	128.227	1.583	0.114	-48.482	454.540	1595
magesq	-0.8710	0.440	-1.981	0.048	-1.734	-0.009	1595
npvisca	0.0032	0.322	0.289	0.772	0.520	0.726	1595

El modelo general de regresión múltiple

¿Cómo interpretamos el coeficiente de mage?

Call:

```
lm_robust(formula = bwght ~ ., data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	1638.2495	502.564	3.260	0.001	652.495	2624.004	1595
monpre_bin	13.3727	33.086	0.404	0.686	-51.525	78.270	1595
cigs	-7.4611	3.129	-2.384	0.017	-13.599	-1.323	1595
mage	49.0195	26.753	1.832	0.067	-3.455	101.494	1595
meduc	-5.2332	8.515	-0.615	0.539	-21.936	11.470	1595
npvis	3.5344	10.204	0.346	0.729	-16.479	23.548	1595
fage	5.6667	3.355	1.689	0.091	-0.914	12.247	1595
feduc	6.7703	7.908	0.856	0.392	-8.741	22.282	1595
omaps	28.3770	17.492	1.622	0.105	-5.932	62.686	1595
fmaps	63.8662	36.262	1.761	0.078	-7.261	134.993	1595
drink	-18.5829	26.430	-0.703	0.482	-70.424	33.259	1595
vlbw	-1950.7856	150.810	-12.935	0.000	-2246.591	-1654.980	1595
male	84.0183	26.800	3.135	0.002	31.451	136.585	1595
mblk	-190.6614	133.310	-1.430	0.153	-452.143	70.820	1595
fblk	203.0290	128.227	1.583	0.114	-48.482	454.540	1595
magesq	-0.8710	0.440	-1.981	0.048	-1.734	-0.009	1595
npvisca	0.0032	0.322	0.289	0.772	0.520	0.726	1595

El modelo general de regresión múltiple

¿Cómo interpretamos el coeficiente de male?

Call:

```
lm_robust(formula = bwght ~ ., data = df, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	1638.2495	502.564	3.260	0.001	652.495	2624.004	1595
monpre_bin	13.3727	33.086	0.404	0.686	-51.525	78.270	1595
cigs	-7.4611	3.129	-2.384	0.017	-13.599	-1.323	1595
mage	49.0195	26.753	1.832	0.067	-3.455	101.494	1595
meduc	-5.2332	8.515	-0.615	0.539	-21.936	11.470	1595
npvis	3.5344	10.204	0.346	0.729	-16.479	23.548	1595
fage	5.6667	3.355	1.689	0.091	-0.914	12.247	1595
feduc	6.7703	7.908	0.856	0.392	-8.741	22.282	1595
omaps	28.3770	17.492	1.622	0.105	-5.932	62.686	1595
fmaps	63.8662	36.262	1.761	0.078	-7.261	134.993	1595
drink	-18.5829	26.430	-0.703	0.482	-70.424	33.259	1595
vlbw	-1950.7856	150.810	-12.935	0.000	-2246.591	-1654.980	1595
male	84.0183	26.800	3.135	0.002	31.451	136.585	1595
mblk	-190.6614	133.310	-1.430	0.153	-452.143	70.820	1595
fblk	203.0290	128.227	1.583	0.114	-48.482	454.540	1595
magesq	-0.8710	0.440	-1.981	0.048	-1.734	-0.009	1595
npvisca	0.0032	0.322	0.289	0.772	0.520	0.726	1595

Bondad de ajuste: ¿Qué tan bueno es el modelo para explicar Y?

¿Qué tanto se ajusta la línea estimada a los datos?

$$R^2 = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{(Y_i - \bar{Y})^2}$$
$$= \frac{\text{suma de lo explicado al cuadrado}}{\text{suma total al cuadrado}}$$
$$= 1 - \frac{\text{suma de lo NO explicado al cuadrado}}{\text{suma total al cuadrado}}$$
$$= 1 - \frac{\sum_i^n (\hat{\mu}_i)^2}{(Y_i - \bar{Y})^2}$$

Bondad de ajuste: ¿Qué tan bueno es el modelo para explicar Y?

- R^2 oscila entre 0 y 1
- Añadir variables al modelo no reduce el R^2 y en general este aumenta
- El R^2 es la proporción de la variación muestral de la variable dependiente que viene explicada por las variables independientes

Discusión en clase: ¿Cómo interpretamos el R^2 en el caso anterior?

Discusión en clase: ¿Un R^2 bajo (e.g., 0.01) es "malo"? ¿Es poco confiable el estimador del efecto de X sobre Y en este caso?

- Un R^2 más grande significa mejor predicción del modelo (ajuste de la línea a los datos)
- Un R^2 más grande NO significa mayor validez interna. No está relacionado con el sesgo del parámetro estimado

Supuestos

S1 Esperanza condicional del error es cero

- $E[\mu_i | X_{1i}, X_{2i}, \dots, X_{ki}] = 0$
- **Discusión en clases:** ¿Qué pasa si hay otras variables $k + 1$ en μ ?

S2 Muestra i.i.d

S3 *Outliers* son poco probables

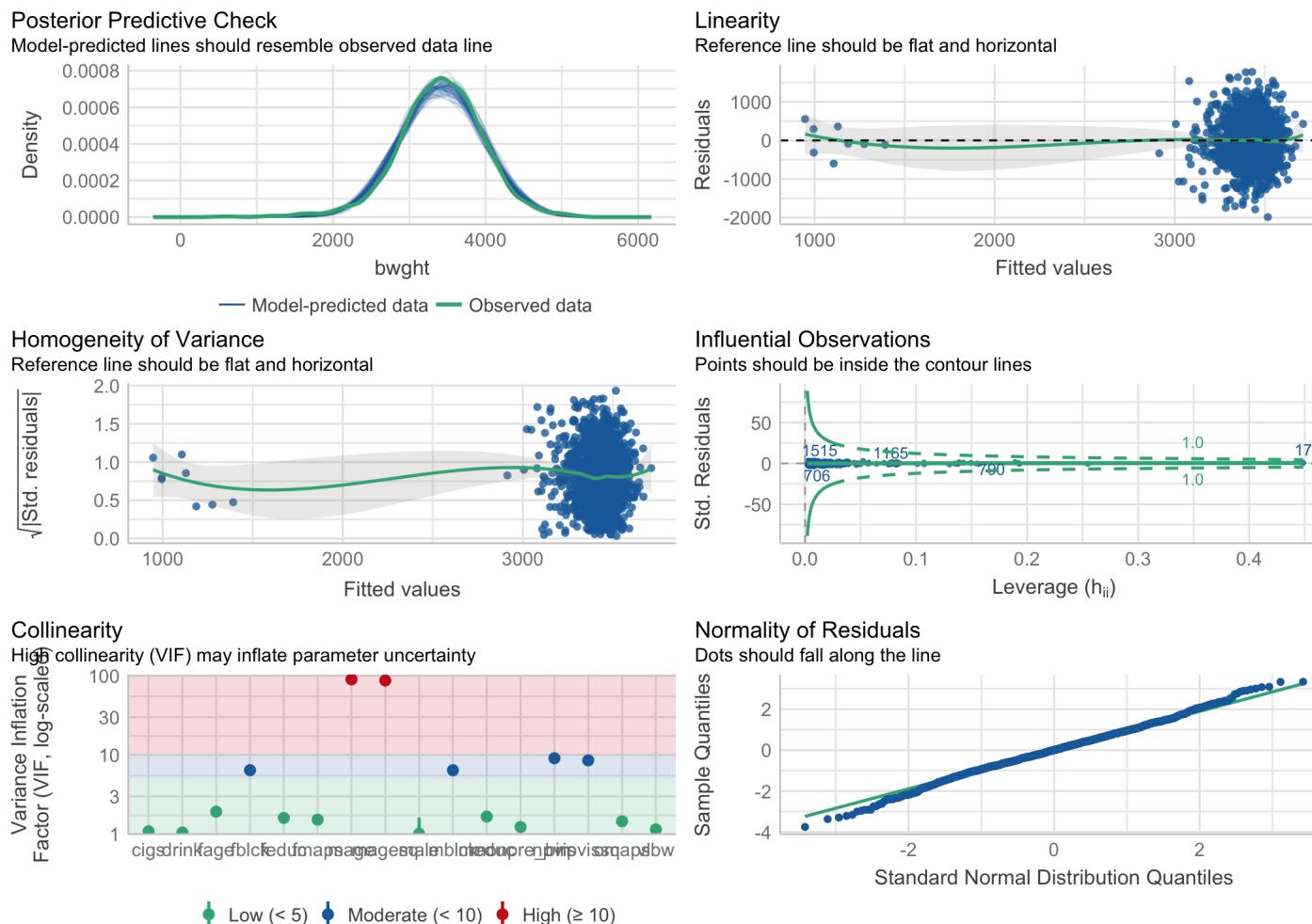
S4 Homocedasticidad

S5 No hay multicolinealidad perfecta

- Multicolinealidad perfecta: una variable X puede ser escrita como una función lineal de la otra

Si (S1) - (S5) se cumplen entonces $E[\hat{\beta}_j] = \beta_j \forall j = 0, 1, 2, \dots, k$

Supuestos: Peso del niño/a al nacer post tratamiento**



Variables dummy en Regresión Múltiple

Ejemplo de brechas salariales entre hombres y mujeres

```
dw ← wage1  
dw ← dw %>% mutate(male=1-female) %>% relocate(wage, female, male)  
glimpse(dw)
```

Rows: 526

Columns: 25

```
$ wage      <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.18,...  
$ female    <int> 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1,...  
$ male      <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0,...  
$ educ      <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, 16,...  
$ exper     <int> 2, 22, 2, 44, 7, 9, 15, 5, 26, 22, 8, 3, 15, 18, 31, 14, 10, ...  
$ tenure    <int> 0, 2, 0, 28, 2, 8, 7, 3, 4, 21, 2, 0, 0, 3, 15, 0, 0, 10, 0, ...  
$ nonwhite   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ married    <int> 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0,...  
$ numdep    <int> 2, 3, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 1, 1, 0, 0, 3, 0, 0,...  
$ smsa       <int> 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
$ northcen   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ south      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ west       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
$ construc   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ ndurman   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ trcomppu  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ trade      <int> 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

Variables dummy en Regresión Múltiple

Ejemplo de brechas salariales entre hombres y mujeres

```
m3 ← lm_robust(wage ~ female, data = dw, se_type = "stata")
```

```
summary(m3)
```

Call:

```
lm_robust(formula = wage ~ female, data = dw, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	7.10	0.251	28.24	0	6.61	7.59	524
female	-2.51	0.298	-8.44	0	-3.10	-1.93	524

Multiple R-squared: 0.116 , Adjusted R-squared: 0.114

F-statistic: 71.2 on 1 and 524 DF, p-value: 0.000000000000000313

Variables dummy en Regresión Múltiple

¿Qué pasa si agregamos la binaria para hombres?

```
m3 ← lm_robust(wage ~ female + male, data = dw, se_type = "stata")
```

```
summary(m3)
```

Call:

```
lm_robust(formula = wage ~ female + male, data = dw, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	60798767908212	NaN	NaN	NaN	NaN	NaN	523
female	-60798767908208	NaN	NaN	NaN	NaN	NaN	523
male	-60798767908205	NaN	NaN	NaN	NaN	NaN	523

Multiple R-squared: 0.115 , Adjusted R-squared: 0.112

F-statistic: NA on 2 and 523 DF, p-value: NA



Variables dummy en Regresión Múltiple

- Otras variables independientes:

$$\hat{salario}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot mujer_i + \hat{\beta}_2 educ_i$$

Discusión en clase: ¿Cuál es la interpretación de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$?

Discusión en clase: Graficar la relación entre educación y salarios según esta ecuación

Variables dummy en Regresión Múltiple

- Variable con más de dos categorías
- Ejemplo: Escolaridad separada en 3 grupos
 - Escolaridad < 12
 - $12 \geq$ Escolaridad < 16
 - $16 \geq$ Escolaridad

Discusión en clase: ¿Por qué hacer esto y no dejar escolaridad continua?

Cuadro 2: Categorías

	Esc<12	12>=Esc<16	16>=Esc
Lesshs	1	0	0
highsch	0	1	0
College	0	0	1

Variables dummy en Regresión Múltiple

Discusión en clase: ¿Cómo se ve la base de datos?

```
dw ← dw %>% mutate(  
  lesscholl = if_else(educ < 12, 1, 0),  
  highscholl = if_else(educ ≥ 12 & educ < 16, 1, 0),  
  college = if_else(educ ≥ 16, 1, 0)  
)  
  
dw %>% select(wage, female, male, educ, lesscholl, highscholl, college) %>% glimpse()
```

Rows: 526

Columns: 7

```
$ wage      <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.1...  
$ female    <int> 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1,...  
$ male      <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0,...  
$ educ      <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, ...  
$ lesscholl  <dbl> 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
$ highscholl <dbl> 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1,...  
$ college   <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0,...
```

Variables dummy en Regresión Múltiple

- Variable con más de dos categorías

```
m3 ← lm_robust(wage ~ highscholl + college, data = dw, se_type = "stata")
```

Call:

```
lm_robust(formula = wage ~ highscholl + college, data = dw, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	4.06	0.184	22.00	0	3.69	4.42	523
highscholl	1.55	0.258	6.01	0	1.05	2.06	523
college	4.89	0.511	9.57	0	3.89	5.90	523

Multiple R-squared: 0.187 , Adjusted R-squared: 0.184

F-statistic: 52.5 on 2 and 523 DF, p-value: <0.0000000000000002

Discusión en clase: ¿Cuál es el grupo base?

Discusión en clase: ¿Cómo interpretamos el coeficiente de `highschool`? ¿Es significativo?

Discusión en clase: ¿Por qué elegir tres categorías y no dos?

Variables dummy en Regresión Múltiple

Interpretando una tabla típica de regresión

```
m3a <- lm_robust(wage ~ female, data = dw, se_type = "stata")
m3b <- lm_robust(wage ~ male, data = dw, se_type = "stata")
m3c <- lm_robust(wage ~ educ, data = dw, se_type = "stata")
m3d <- lm_robust(wage ~ exper, data = dw, se_type = "stata")
m3e <- lm_robust(wage ~ female + educ + exper + tenure, data = dw, se_type = "stata")
```

Variables dummy en Regresión Múltiple

Variable dependiente: wage

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	7.099 (0.251)***	4.588 (0.159)***	-0.905 (0.725)	5.373 (0.204)***	-1.568 (0.826)
female	-2.512 (0.298)***				-1.811 (0.254)***
male		2.512 (0.298)***			
educ			0.541 (0.061)***		0.572 (0.061)***
exper				0.031 (0.011)**	0.025 (0.010)**
tenure					0.141 (0.028)***
R ²	0.116	0.116	0.165	0.013	0.364
Adj. R ²	0.114	0.114	0.163	0.011	0.359
Num. obs.	526	526	526	526	526
RMSE	3.476	3.476	3.378	3.673	2.958

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

Test de hipótesis en Regresión Múltiple

- En regresión múltiple podríamos estar interesadas(os) en *testear* hipótesis de manera simultánea
- Modelo con cuatro variables independientes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \mu_i$$

Discusión en clase: ¿Por qué hacer *test* de hipótesis?

- La meta en regresión lineal es usar un modelo estadístico y una muestra
- Siempre que tenemos una muestra estaremos condicionando la validez del modelo a variación muestral
- Test de hipótesis nos entregan nociones de la significancia estadística
- Capacidad estadística de nuestra estimación para explicar el modelo poblacional/verdadero

Test de hipótesis en Regresión Múltiple

- Hasta ahora esto se veía así:
 - $H_0 : \beta_2 = 0; H_A : \beta_2 \neq 0$
 - Usamos el $t - test$
- Pero ahora queremos saber sobre la significancia simultánea
 - Usamos el $F - test$, test de Fischer
 - Típicamente se utiliza para revisar:
 1. $H_0 : \beta_1 = 0 \text{ y } \beta_2 = 0; H_A : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0$
 2. $H_0 : \beta_1 = \beta_2; H_A : \beta_1 \neq \beta_2$
 - En esta clase nos enfocamos en el primer uso
- Intuición para el $F - test$:
 - ¿Testear la significancia de una variable singular a la del modelo?
 - ¿Testear no linealidades en un modelo de regresión? (ej., $edad$ y $edad^2$)
 - Testear por separado dos variables no sirve: necesitamos ajustar por la correlación entre t_1 y t_2

Test de hipótesis en Regresión Múltiple

Ejemplo: Resultados al nacer (Apgar al minuto) y variables explicativas

- `omaps`: one minute apgar score
- `cigs`: avg cigarettes per day
- `drink`: avg drinks per week
- `meduc`: mother's educ, years
- `feduc`: father's educ, years
- `male`: =1 if baby male

Test de hipótesis en Regresión Múltiple

Ejemplo: Resultados al nacer (Apgar al minuto) y variables explicativas

```
m4 ← lm_robust(omaps ~ cigs + feduc + meduc + male, data = df, se_type = "stata")
```

Test de hipótesis en Regresión Múltiple

Call:

```
lm_robust(formula = omaps ~ cigs + feduc + meduc + male, data = df,  
          se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	7.8507	0.20762	37.81	0.000	7.444	8.258	1664
cigs	-0.0129	0.00846	-1.52	0.128	-0.029	0.004	1664
feduc	0.0240	0.01476	1.63	0.104	-0.005	0.053	1664
meduc	0.0165	0.01618	1.02	0.307	-0.015	0.048	1664
male	-0.0190	0.05423	-0.35	0.726	-0.125	0.087	1664

Multiple R-squared: 0.00885 , Adjusted R-squared: 0.00647

F-statistic: 2.8 on 4 and 1664 DF, p-value: 0.0246

Discusión en clase: ¿Cómo interpretamos el coeficiente de β_{feduc} ? ¿Es significativo?

Discusión en clase: ¿Cómo interpretamos el coeficiente de β_{meduc} ? ¿Es significativo?

Discusión en clase: ¿Concluimos entonces que la educación de **padres en conjunto** no importa para apgar?

Test de hipótesis en Regresión Múltiple

Ejemplo: Resultados al nacer (Apgar al minuto) y variables explicativas

Discusión en clase: ¿Cómo podemos diseñar un test que nos permita revisar la hipótesis de que educación de padres importa?

- $H_0 : \beta_{meduc} = \beta_{feduc} = 0; H_A : \beta_{meduc} \neq 0 \text{ y/o } \beta_{feduc} \neq 0$
- Así como acabamos de calcular test t para cada coeficiente por separado, ahora calculamos el test F para la significancia conjunta.

```
linearHypothesis(m4, c("feduc = 0", "meduc = 0"))
```

Linear hypothesis test

Hypothesis:
feduc = 0
meduc = 0

Model 1: restricted model

Model 2: omaps ~ cigs + feduc + meduc + male

	Res.Df	Df	Chisq	Pr(>Chisq)
1	1666			
2	1664	2	8.12	0.017 *

Discusión en clase: ¿Cómo interpretamos?
¿Qué concluimos?

Test de hipótesis en Regresión Múltiple

Discusión en clase: ¿Por qué los t no eran significativos y el F si?

- Clave: correlación entre ambas variables aumenta los errores estándar en los coeficientes de estas variables

```
cor.test(df$feduc, df$meduc, method = "pearson")
```

Pearson's product-moment correlation

```
data: df$feduc and df$meduc
t = 30, df = 1777, p-value <0.0000000000000002
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.554 0.615
sample estimates:
 cor
0.585
```

Test de hipótesis en Regresión Múltiple

- Intuición
- El test F se puede pensar tomando dos regresiones como ejemplo
 1. Regresión restringida que impone H_0 :

$$omaps_i = \alpha_0 + \alpha_1 cigs_i + \alpha_2 male_i + v_i$$

1. Regresión no-restringida que no impone H_0 :

$$omaps_i = \beta_0 + \beta_1 cigs_i + \beta_2 male_i + \beta_3 meduc_i + \beta_4 feduc_i + \mu_i$$

Discusión en clase: ¿Qué pasa con el R^2 cuando pasamos de (1) a (2)?

Test de hipótesis en Regresión Múltiple

- F – *test* mira el aumento en el R^2 cuando agregamos variables de interés. En otras palabras, medirá si estas variables adicionales agregan algo al modelo.
 1. Si son variables relevantes: R^2 debería aumentar notoriamente
 2. Si no son variables relevantes: R^2 debería \pause{...} disminuir/aumentar poco/mucho?
- Fórmula simple:

$$F_{q,N-k-1} = \frac{(R_{NR}^2 - R_R^2)/q}{(1 - R_{NR}^2)/(N - k - 1)}$$

- **q** Número de coeficientes *testeados*
- **NR** No restringido; incluye las variables a *testear*
- **R** Restringido; no incluye las variables a *testear*
- **N** Número de observaciones
- **k** Número de variables en la versión no restringida

Test de hipótesis en Regresión Múltiple

- Fórmula simple:

$$F_{q,N-k-1} = \frac{(R_{NR}^2 - R_R^2)/q}{(1 - R_{NR}^2)/(N - k - 1)}$$

- **q** Número de coeficientes *testeados*
- **NR** No restringido; incluye las variables a *testear*
- **R** Restringido; no incluye las variables a *testear*
- **N** Número de observaciones
- **k** Número de variables en la versión no restringida

El F crítico cambia con q

$q \rightarrow$	2	3	4	5	6	7	8	9	10
5% critical value	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
1% critical value	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Discusión en clase: Caso anterior: ¿q, k, N, F crítico?

Test de hipótesis en Regresión Múltiple

```
m4a <- lm_robust(omaps ~ cigs + male, data = na.omit(df), se_type = "stata")
m4b <- lm_robust(omaps ~ cigs + male + feduc + meduc, data = na.omit(df), se_type = "stata")
```

Test de hipótesis en Regresión Múltiple

Variable dependiente: wage

	Modelo Restringido	Modelo NO Restringido
(Intercept)	8.419 (0.035)***	7.883 (0.206)***
cigs	-0.017 (0.009)	-0.014 (0.009)
male	-0.022 (0.054)	-0.023 (0.055)
feduc		0.025 (0.015)
meduc		0.014 (0.017)
R ²	0.004	0.009
Adj. R ²	0.003	0.006
Num. obs.	1612	1612
RMSE	1.096	1.094

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

Test de hipótesis en Regresión Múltiple

Discusión en clase: Caso anterior: R^2_{NT} , R^2_R ¿?

Discusión en clase: ¿Cómo calculamos el F?

```
waldtest(m4a, m4b)
```

Wald test

Model 1: omaps ~ cigs + male

Model 2: omaps ~ cigs + male + feduc + meduc

 Res.Df Df Chisq Pr(>Chisq)

1 1609

2 1607 2 7.45 0.024 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3. Regresión no lineal

Regresión no lineal

- Por ahora hemos visto sólo ejemplos de relaciones **lineales** entre Y y X :

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

$$\text{salario}_i = \beta_0 + \beta_1 \text{escol}_i + \mu_i$$

$$\text{ptjeacad}_i = \beta_0 + \beta_1 \text{tamañoclase}_i + \mu_i$$

- Lineal \rightarrow efecto es el mismo para todo i . \rightarrow Poco realista

Discusión en clase: ¿Cómo estimamos un efecto del tamaño de la clase que sea distinto para niñas y niños?

- En regresiones no lineales el cambio en Y asociado a un cambio en X_1 depende del valor de X_1 o de otra variable X_2
- Hoy: interacciones

Interacciones

- Al escribir nuestra FRP, inmediatamente estamos haciendo supuestos de cómo se relacionan la variable dependiente con las independientes:

$$\text{salario}_i = \beta_0 + \beta_1 \text{escol}_i + \beta_2 \text{mujer}_i + \mu_i$$

- Esta FRP permite que el salario sea distinto para hombres y mujeres
- Pero no permite que la asociación entre educación y salarios sea distinta entre hombres y mujeres

Discusión en clase: Gráficamente, ¿cómo es la relación entre educación y salarios según esta regresión?

Interacciones

Discusión en clase: ¿Cuál es el cambio en salarios asociado a un año adicional de educación para hombres? ¿Para mujeres?

```
m5 ← lm_robust(wage ~ female + educ, data = dw, se_type = "stata")
```

```
summary(m5)
```

Call:

```
lm_robust(formula = wage ~ female + educ, data = dw, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.623	0.7287	0.855	0.393	-0.809	2.054	523
female	-2.273	0.2702	-8.414	0.000	-2.804	-1.743	523
educ	0.506	0.0599	8.456	0.000	0.389	0.624	523

Multiple R-squared: 0.259 , Adjusted R-squared: 0.256

F-statistic: 69.1 on 2 and 523 DF, p-value: <0.0000000000000002

Interacciones

Discusión en clase: ¿Es esto realista?

- Lo que nos gustaría capturar es una asociación entre salarios y educación distinta para hombres y mujeres
- Para estos efectos hacemos una interacción:

$$\text{salario}_i = \beta_0 + \beta_1 \text{escol}_i + \beta_2 \text{mujer}_i + \beta_3 \text{escol}_i \times \text{mujer}_i + \mu_i$$

- En R: `mujer_esc=escol*mujer`
- La FRM entonces es:

$$\hat{\text{salario}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{escol}_i + \hat{\beta}_2 \text{mujer}_i + \hat{\beta}_3 \text{mujer}_{esc}_i$$

Interacciones

$$\hat{salario}_i = \hat{\beta}_0 + \hat{\beta}_1 escol_i + \hat{\beta}_2 mujer_i + \hat{\beta}_3 mujer_esc_i$$

Hombre: $mujer_i = 0$

Mujer: $mujer_i = 1$

$$\hat{salario}_i = \hat{\beta}_0 + \hat{\beta}_1 escol_i \quad \hat{salario}_i = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) escol_i$$

Discusión en clase: ¿Cuál es el cambio en salarios asociado a un cambio en escolaridad para hombres? ¿Para mujeres?

Discusión en clase: ¿Cuál es la interpretación de $\hat{\beta}_1$? ¿de $\hat{\beta}_3$?

Interacciones

```
dw ← dw %>% mutate(female_educ=educ*female)
m5 ← lm_robust(wage ~ female + educ + female_educ, data = dw, se_type = "stata")
```

```
summary(m5)
```

Call:

```
lm_robust(formula = wage ~ female + educ + female_educ, data = dw,
          se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.200	0.8717	0.230	0.818	-1.512	1.913	522
female	-1.199	1.4608	-0.820	0.412	-4.068	1.671	522
educ	0.539	0.0734	7.350	0.000	0.395	0.684	522
female_educ	-0.086	0.1238	-0.695	0.488	-0.329	0.157	522

Multiple R-squared: 0.26 , Adjusted R-squared: 0.256

F-statistic: 49.1 on 3 and 522 DF, p-value: <0.0000000000000002

Discusión en clases: ¿Cuál es el cambio en salario con el cambio en un año de escolaridad para hombres? ¿Para mujeres?

Interacciones

- Mas ejemplos: Interacción entre dos variables binarias

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{1i} \times D_{2i} + \mu_i$$

$$\text{salario}_i = \beta_0 + \beta_1 \text{mujer}_i + \beta_2 \text{ecivil}_i + \beta_3 \text{ecivil}_i \times \text{mujer}_i + \mu_i$$

→ La diferencia en salarios entre hombres y mujeres además depende del estado civil
→ La diferencia en salarios entre personas casadas y no casadas depende de su sexo

- Mas ejemplos: Interacción entre dos variables continuas

$$\text{salario}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{educ}_i \times \text{exper}_i + \mu_i$$

→ Los retornos a la educación dependen de la experiencia → Los retornos a la experiencia dependen de la educación

Interacciones

- Marco general para interacción entre variables binarias

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{1i} \times D_{2i} + \mu_i$$

D _{1i} = 0	D _{1i} = 1
D _{2i} = 0	
D _{2i} = 1	

- Matemáticamente:

- $\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$
- $\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$

Interacciones: Estudio de caso

Diferencias en accidentes fatales de tránsito entre lugares con distinta regulación para el consumo de cerveza

```
dfat ← Ecdat::Fatality %>% filter(year=1988) %>%
  mutate(jaild = if_else(jaild == "yes", 1, 0),
        comserd = if_else(comserd == "yes", 1, 0))
```

- `state`: state ID code
- `year`: year
- `mrall`: traffic fatality rate (deaths per 10000)
- `beertax`: tax on case of beer
- `jaild`: mandatory jail sentence ?
- `comserd`: mandatory community service ?
- `vmiles`: average miles per driver
- `unrate`: unemployment rate
- `perinc`: per capita personal income

Interacciones: Estudio de caso

Diferencias en accidentes fatales de tránsito entre lugares con distinta regulación para el consumo de cerveza

Variable	Obs	Media	Std	Min	Max
mrall	48	2.070	0.521	1.2311	3.24
beertax	48	0.480	0.435	0.0433	2.19
jaild	48	0.292	0.459	0.0000	1.00
comserd	48	0.208	0.410	0.0000	1.00
vmiles	48	8.616	1.115	5.7899	11.81
unrate	48	5.456	1.838	2.4000	10.90
perinc	48	14893.527	2628.106	10698.7490	22193.46

Interacciones: Estudio de caso

Análisis de regresión

```
m5a <- lm_robust(mrall ~ jaild, data = dfat, se_type = "stata")  
  
m5b <- lm_robust(mrall ~ jaild + beertax + comserd + vmiles + unrate + perinc,  
                  data = dfat, se_type = "stata")  
m5c <- lm_robust(mrall ~ jaild + beertax + comserd + vmiles + unrate + perinc + jaild:beertax,  
                  data = dfat, se_type = "stata")
```

Interacciones: Estudio de caso

Variable dependiente mral: traffic fatality

	Model 1	Model 2	Model 3
(Intercept)	1.961 (0.085)***	0.577 (0.762)	0.648 (0.816)
jaild	0.371 (0.158)*	0.121 (0.164)	0.156 (0.227)
beertax		0.145 (0.108)	0.163 (0.113)
comserd		0.097 (0.183)	0.103 (0.183)
vmiles		0.223 (0.049)***	0.218 (0.054)***
unrate		0.051 (0.035)	0.049 (0.036)
perinc		-0.000 (0.000)*	-0.000 (0.000)*
jaild:beertax			-0.077 (0.340)
R ²	0.107	0.637	0.638
Adj. R ²	0.088	0.584	0.574
Num. obs.	48	48	48

Interacciones: Estudio de caso

Discusión en clases: ¿Cómo ocupamos esta información para deducir asociaciones entre una regulación más rígida sobre el consumo de cerveza y los accidentes de tránsito?

Discusión en clases: ¿Es la asociación entre el impuesto a la cerveza y la tasas de accidentes igual para lugares con distintas sentencias de cárcel obligatoria?

Discusión en clases: ¿Cómo testeamos la hipótesis de que la asociación entre impuesto y fatalidad es mayor en lugares con *alcohotent*?

Regresión no lineal

- Otra forma de regresión no lineal es cuando especificamos Y o X en **logaritmos**.
- Para efectos de esta clase utilizaremos siempre el logaritmo natural
- La utilidad está en que la interpretación ahora es en cambios porcentuales
 - Por ejemplo, en vez de decir que ΔX (escolaridad) genera un cambio en ΔY (invresos) de \$10000 pesos vamos a poder inferir en qué porcentaje cambian los ingresos

Logaritmos

- Así la interpretación es en cambios porcentuales
- La relación entre logaritmo y cambios porcentuales viene de que para pequeños Δx :

$$\ln(x + \Delta x) - \ln(x) = \frac{\Delta x}{x}$$

- Ejemplo: $x = 100$; $\Delta x = 1$

$$\frac{\Delta x}{x} = 0.01$$

$$\ln(101) - \ln(100) = 0.00995 = \frac{1}{100} = 0.01 = 1\%$$

Logaritmos

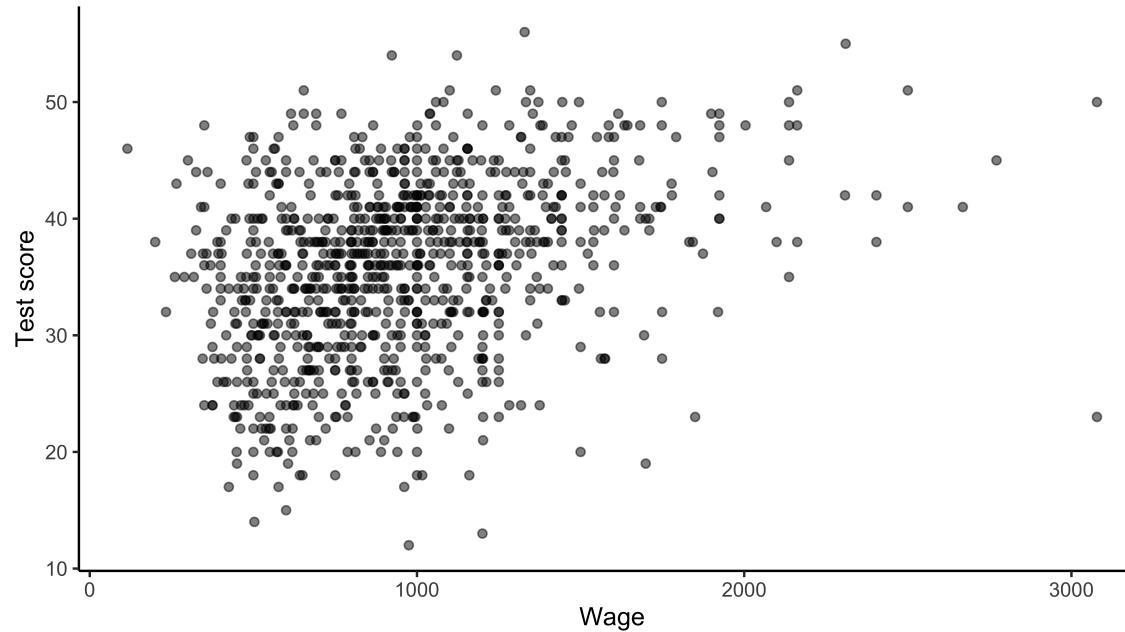
- Hay 3 casos para los cuales se ocupan logaritmos en regresión lineal
 - Caso 1: X está en logs, Y no
 - Caso 2: Y está en logs, X no
 - Ambas están en logs
- Cuando usamos logaritmos:
 - La mecánica del OLS no cambia
 - La interpretación de los coeficientes cambia
- Caso 1: X en log

$$Y = \beta_0 + \beta_1 \ln(X_{1i}) + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \mu_i$$

- Interpretación:
 - β_1 : El cambio en Y asociado a un cambio de in 1% en X_{1i} , manteniendo constante X_{2i}, \dots, X_{ki}

Logaritmos: Ejemplo

Relación entre puntaje académico y salarios



El gráfico sugiere que la relación entre ingreso y puntaje académico no es lineal **¿por qué?**

Logaritmos: Ejemplo

- Una forma de especificar la regresión es:

$$Y = \beta_0 + \beta_1 \ln(X_i) + \mu_i$$

$$ptje_acad_i = \beta_0 + \beta_1 \ln(\text{ingreso}_i) + \mu_i$$

- De esta forma no asumimos que el cambio en pesos es el mismo en todo el rango de ingreso. Lo llevamos a cambios procentuales. Un cambio de un 1% no es lo mismo para ingresos de 200 mil que para ingresos de 1 MM.

Logaritmos: Ejemplo

En R

```
dw2 ← dw2 %>% mutate(lwage=log(wage)) # Logaritmo natural  
m6 ← lm_robust(KWW ~ lwage, data=dw2, se_type = "stata")
```

```
summary(m6)
```

Call:

```
lm_robust(formula = KWW ~ lwage, data = dw2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	-1.91	4.391	-0.436	0.663	-10.53	6.70	933
lwage	5.56	0.644	8.631	0.000	4.29	6.82	933

Multiple R-squared: 0.0938 , Adjusted R-squared: 0.0928

F-statistic: 74.5 on 1 and 933 DF, p-value: <0.0000000000000002

Logaritmos: Ejemplo

Call:

```
lm_robust(formula = KWW ~ lwage, data = dw2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	-1.91	4.391	-0.436	0.663	-10.53	6.70	933
lwage	5.56	0.644	8.631	0.000	4.29	6.82	933

Multiple R-squared: 0.0938 , Adjusted R-squared: 0.0928

F-statistic: 74.5 on 1 and 933 DF, p-value: <0.0000000000000002

- ¿Cómo interpretamos el coeficiente asociado a `lwage` en esta regresión? Dividir por 100 β_1
- El supuesto aquí es que un aumento en un 1% en ingresos tiene el mismo efecto indiferente del nivel de ingreso. ¿Cómo cambia Y con?
 1. un aumento de 1% de un ingreso de \$10.000 es de 10.100
 2. un aumento de 1% de un ingreso de \$100.000 es de 101.000

Logaritmos

Cambio en ptje_acad con un aumento del ingreso de:

	10.000 a 10.100	100.000 a 101.000
--	------------------------	--------------------------

$$ptje_{acad} = \beta_0 + \beta_1 \log(ingreso_i) + \mu_i$$

$$ptje_{acad} = \alpha_0 + \alpha_1 \log(ingreso_i) + \mu_i$$

Logaritmos

- $ptje_{acad_i} = \alpha_0 + \alpha_1 ingreso_i + \mu_i$ asume que el ingreso y el puntaje están linealmente asociados
- $ptje_{acad_i} = \beta_0 + \beta_1 \log(ingreso_i) + \mu_i$ asume que el puntaje y el log de ingresos están linealmente asociados:
 - Puntaje académico y cambio porcentual en ingreso están linealmente asociados
 - Si β_1 es positivo, el puntaje académico aumenta pero a tasas decreciente: ¿los datos sugieren esto? Para un mismo cambio en ingresos (\$100) ¿es el cambio en puntaje igual o menor/mayor dependiendo del nivel de ingreso?

Logaritmos

Caso 2: Y en log

$$\ln(Y) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots \beta_k X_{ki} + \mu_i$$

- Interpretación:
 - β_1 : El cambio porcentual Y asociado a un cambio de X_{1i} en una unidad manteniendo constante X_{2i}, \dots, X_{ki}
 - β_2 : El cambio porcentual Y asociado a un cambio de X_{2i} en una unidad manteniendo constante X_{1i}, \dots, X_{ki}
 - Y así para todos los X s
- Ejemplo: regresión de salarios

$$salario_i = \beta_0 + \beta_1 educ_i + \mu_i$$

- Asume que la asociación entre salario y educación es la misma sin importar el nivel de educación
- Quizás es más realista asumir que el cambio porcentual en salarios es el mismo sin importar el nivel de educación

Logaritmos

$$\rightarrow \ln(\text{salario}_i) = \beta_0 + \beta_1 \text{educ}_i + \mu_i$$

```
m7 ← lm_robust(lwage ~ educ + exper + tenure + female, data=dw, se_type = "stata")
```

```
summary(m7)
```

Call:

```
lm_robust(formula = lwage ~ educ + exper + tenure + female, data = dw,
se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.50135	0.11526	4.35	0.000	0.275	0.728	521
educ	0.08746	0.00799	10.95	0.000	0.072	0.103	521
exper	0.00463	0.00162	2.86	0.004	0.001	0.008	521
tenure	0.01737	0.00355	4.89	0.000	0.010	0.024	521
female	-0.30115	0.03767	-7.99	0.000	-0.375	-0.227	521

Multiple R-squared: 0.392 , Adjusted R-squared: 0.388

F-statistic: 73.8 on 4 and 521 DF, p-value: <0.0000000000000002

Logaritmos

```
summary(m7)
```

Call:

```
lm_robust(formula = lwage ~ educ + exper + tenure + female, data = dw,  
          se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.50135	0.11526	4.35	0.000	0.275	0.728	521
educ	0.08746	0.00799	10.95	0.000	0.072	0.103	521
exper	0.00463	0.00162	2.86	0.004	0.001	0.008	521
tenure	0.01737	0.00355	4.89	0.000	0.010	0.024	521
female	-0.30115	0.03767	-7.99	0.000	-0.375	-0.227	521

Multiple R-squared: 0.392 , Adjusted R-squared: 0.388

F-statistic: 73.8 on 4 and 521 DF, p-value: <0.0000000000000002

¿Cómo interpretamos estos coeficientes?

- Cuando la variable dependiente de un modelo es $\log(Y)$ hay que multiplicar todos los coeficientes por 100 e interpretarlos como el cambio porcentual

Logaritmos

Caso 3: Y en log y X en log

$$\ln(Y) = \beta_0 + \beta_1 \ln(X_{1i}) + \beta_2 X_{2i} + \dots \beta_k X_{ki} + \mu_i$$

- Interpretación:
 - β_1 : El cambio porcentual Y asociado a un cambio de X_{1i} en un 1%, manteniendo constante X_{2i}, \dots, X_{ki}
- Ejemplo: contaminación y precios de las casas

$$\ln(\text{precio}_i) = 9.23 - 0.178 \ln(\text{conta}_i) + 0.306 \text{piezas}_i \mu_i$$

- Interpretemos estos coeficientes

Especificación cuadrática

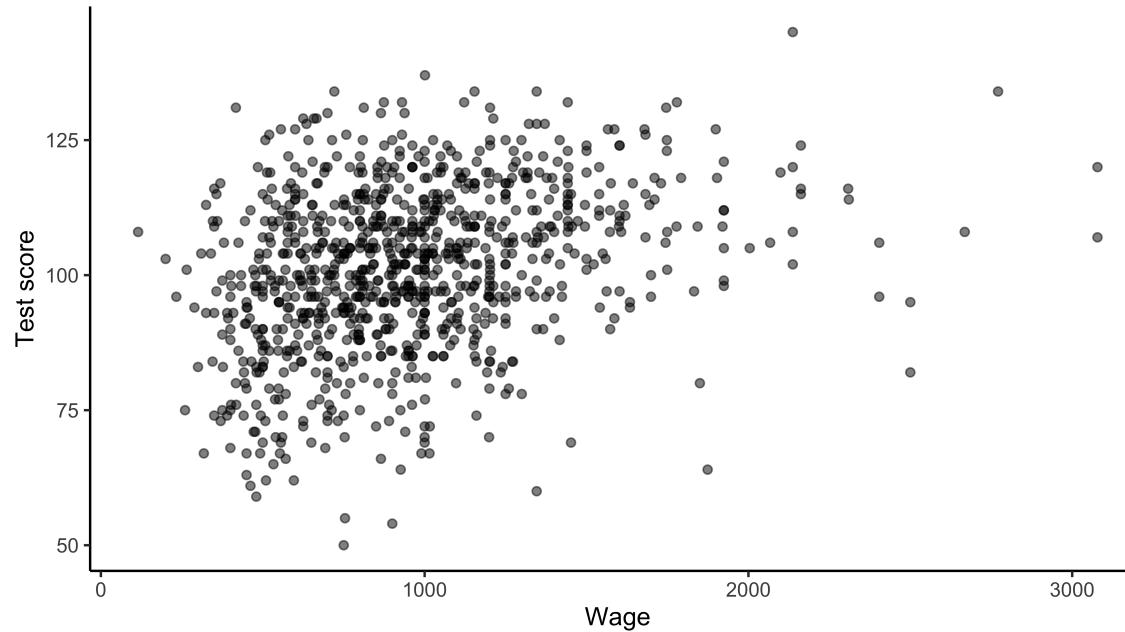
- Intuición: El cambio en Y (salario) respecto a X (escolaridad) depende del valor de X

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \mu_i$$

- Mismo método de estimación pero distintas interpretación
 - β_1 : ya no es el cambio en Y con un cambio en X_{1i}
 - Antes teníamos que: $\frac{\Delta Y}{\Delta X_1} = \beta_1$
 - Ahora: $\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2 \times \beta_2 X_1$
 - Esto viene de tomar una derivada pero lo importante es que ahora el cambio en Y asociado a un cambio en X_1 en una unidad depende del nivel de X_1
 - No es la misma asociación para personas con menores niveles de escolaridad que para personas con mayores niveles de escolaridad

Especificación cuadrática

Relación entre puntaje académico y salarios



- El gráfico sugiere que la relación entre ingreso y puntaje académico no es lineal **¿porqué?**

Especificación cuadrática

- Hagamos la regresión

$$puntaje_i = \beta_0 + \beta_1 \text{ingreso}_i + \beta_2 \text{ingreso}_i^2 + \mu_i$$

```
dw2 ← dw2 %>% mutate(wage2=wage^2)
m8 ← lm_robust(IQ ~ wage + wage2, data=dw2, se_type = "stata")
```

```
summary(m8)
```

Call:

```
lm_robust(formula = IQ ~ wage + wage2, data = dw2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	84.83192627	2.31863776	36.59	0.000	80.282	89.38	932
wage	0.02219924	0.00390765	5.68	0.000	0.015	0.03	932
wage2	-0.00000445	0.00000153	-2.92	0.004	0.000	0.00	932

Multiple R-squared: 0.103 , Adjusted R-squared: 0.101

F-statistic: 52.4 on 2 and 932 DF, p-value: <0.0000000000000002

Especificación cuadrática

- Modelo propuesto:

$$puntaje_i = \beta_0 + \beta_1 \text{ingreso}_i + \beta_2 \text{ingreso}_i^2 + \mu_i$$

$$puntaje_i = 84.83 + 0.022 \text{ingreso}_i + (-0.000005) \text{ingreso}_i^2$$

Ingreso	Predicción del Modelo	Cambio en el score
100	87.0	
101	87.0	0.021
200	89.0	
201	89.0	0.020
400	92.8	
401	92.8	0.018

Especificación cuadrática

- Hagamos la regresión

$$puntaje_i = \beta_0 + \beta_1 \text{ingreso}_i + \mu_i$$

```
m8b <- lm_robust(IQ ~ wage, data=dw2, se_type = "stata")
```

```
summary(m8b)
```

Call:

```
lm_robust(formula = IQ ~ wage, data = dw2, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	90.2602	1.28079	70.47	0	87.747	92.774	933
wage	0.0115	0.00121	9.54	0	0.009	0.014	933

Multiple R-squared: 0.0955 , Adjusted R-squared: 0.0946

F-statistic: 91.1 on 1 and 933 DF, p-value: <0.0000000000000002

Especificación cuadrática

- Modelo propuesto:

$$puntaje_i = \beta_0 + \beta_1 \text{ingreso}_i + \mu_i$$

$$puntaje_i = 90.26 + 0.012 \text{ingreso}_i$$

Ingreso	Predicción del Modelo	Cambio en el score
100	92.6	
101	92.6	0.023
200	94.9	
201	94.9	0.023
400	99.5	
401	99.5	0.023

Variable dependiente binaria (VDB)

- Antes ocupamos variables dummy como variables independientes en una regresión
- Pero también se utilizan ampliamente como variables dependientes
 - Trabaja no trabaja
 - Tiene o no tiene seguro
 - Entre o no a la Universidad
 - Es pobre o no
- Hay tres formas de estimar una regresión con VDB
 - MCO: Modelo de probabilidad lineal
 - Probit
 - Logit

Modelo de probabilidad lineal

- El MPL es una regresión múltiple del tipo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \mu_i$$

- Done la única diferencia es que ahora Y_i es binaria: $\in \{0, 1\}$
- $E[Y|X] = 1Pr(Y = 1|X) + 0Pr(Y = 0|X) = Pr(Y = 1|X)$
- $Pr(Y = 1|X) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$
- β_k : Cambio predicho en la probabilidad de éxito ($Y_i = 1$) cuando X_k aumenta en una unidad, dejando todas las demás variables constante.

Modelo de probabilidad lineal: ejemplo

Se utiliza el mismo método que hemos usado hasta ahora: MCO

Ejemplo: Determinantes del trabajo administrativo

$$PRF : CLEROCC_i = \beta_0 + \beta_1 EDUC_i + \beta_2 FEMALE_i + \beta_3 NONWHITE_i + \mu_i$$

```
dw %>% select(clerocc, educ, female, nonwhite) %>% head()
```

	clerocc	educ	female	nonwhite
1	0	11	1	0
2	0	12	1	0
3	0	11	0	0
4	1	8	0	0
5	0	12	0	0
6	0	16	0	0

Modelo de probabilidad lineal: ejemplo

Ejemplo: Determinantes del trabajo administrativo

$$PRF : CLEROCC_i = \beta_0 + \beta_1 EDUC_i + \beta_2 FEMALE_i + \beta_3 NONWHITE_i + \mu_i$$

```
m9 ← lm_robust(clerocc ~ educ + female + nonwhite, data=dw, se_type = "stata")
```

```
summary(m9)
```

Call:

```
lm_robust(formula = clerocc ~ educ + female + nonwhite, data = dw,
          se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.01009	0.0578	0.175	0.861	-0.103	0.124	522
educ	0.00264	0.0044	0.600	0.549	-0.006	0.011	522
female	0.26642	0.0315	8.448	0.000	0.204	0.328	522
nonwhite	-0.03516	0.0471	-0.747	0.455	-0.128	0.057	522

Multiple R-squared: 0.127 , Adjusted R-squared: 0.122

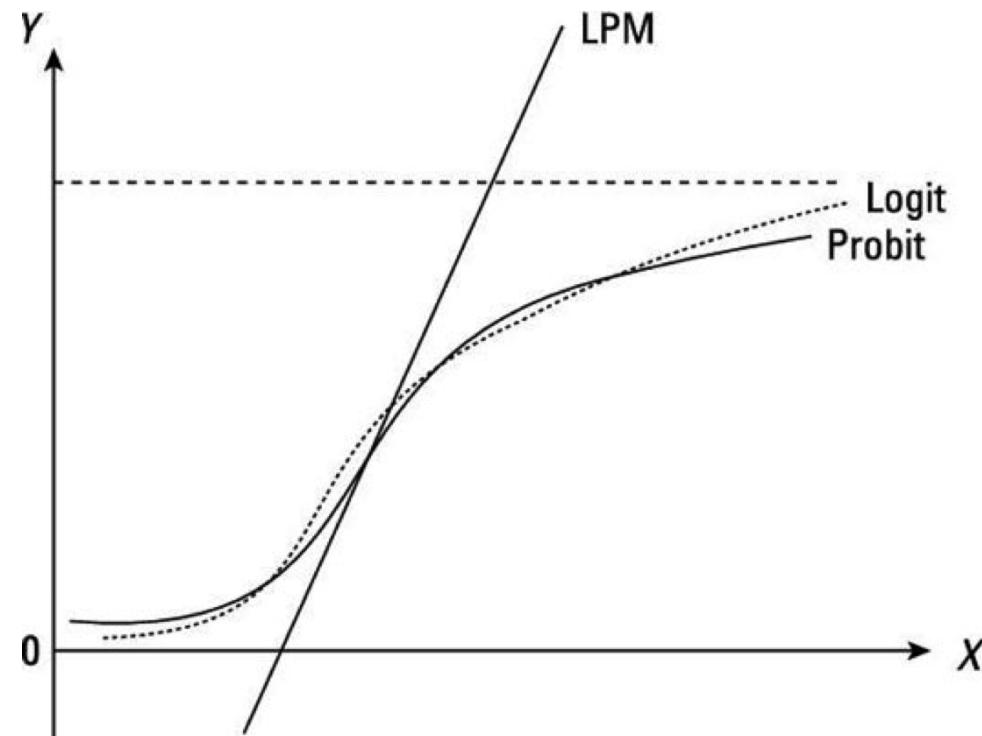
F-statistic: 23.9 on 3 and 522 DF, p-value: 0.000000000000162

Modelo de probabilidad lineal

- La principal ventaja del MPL es que no requiere aprender un nuevo método
- La interpretación de los coeficientes es directa
- Un problema es que podemos obtener valores predichos fuera del intervalo $[0, 1]$
 - Esta es una de las críticas principales a usar MCO para un modelo de variable dependiente binaria y abogan por *logit* o *probit*
 - Pero, si no estamos interesados(as) en \hat{Y} pero en cambio en ΔY MPL puede hacerlo bastante bien
 - Además MPL tiene un buen desempeño cuando queremos hacer predicciones evaluando las variables en el promedio
- En la práctica MPL se usa como primera aproximación a la estimación y luego se utiliza algo más sofisticado.

Ejemplo

Graficamente



Probit

- Modelo no lineal
- Valores predichos entre 0 y 1
- Valores predichos similares a MPL cuando evaluamos en el promedio, pero muy distintos para otros valores de las Xs
- El modelo tiene la siguiente representación:
 - $Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})$
 - Donde Y es binaria y Φ es la función de distribución acumulada de una normal con regresores X_1, X_2, \dots
- Los coeficientes del *probit*, β_0, β_1 , etc no tienen interpretación directa. El modelo se interpreta mejor computando los valores predichos de Y_i con una cambio en el regresor X
- La probabilidad de que $Y = 1$ dado valores de X_1, X_2, \dots, X_k se computan calculando los valores $z = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ y luego buscando este valor z en la tabla de distribución normal

Logit

- Modelo no lineal
- Valores predichos entre 0 y 1
- Valores predichos similares a MPL cuando evaluamos en el promedio, pero muy distintos para otros valores de las Xs
- El modelo tiene la siguiente representación:

$$Pr(Y = 1|X) = F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})$$

$$= \frac{1}{1 - exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}$$

- Los coeficientes del *logit*, β_0 , β_1 , etc no tienen interpretación directa.

Modelos de VDB

- Salvo en el caso del MPL los coeficientes no tienen interpretación directa cuando salen en el output de stata
- Esto es igual que en modelos no lineales
- El signo y la significancia estadística si salen directo el output de Stata
- En R: `glm(... , family=binomial(link="logit"))` reporta coeficientes; `logitmfx()` cambios en probabilidad; `predict()` para probabilidades predichas.
- En R: `glm(... , family=binomial(link="probit"))` reporta coeficientes; `probitmfx()` cambios en probabilidad; `predict()` para probabilidades predichas.

Modelos de VDB

- ¿Cómo pensar en el cambio en Y que viene de un cambio en X ?
- Para cada modelo podemos hacer los siguientes pasos:
 1. Computar la probabilidad predicha para el valor inicial de los regresores
 2. Computar la probabilidad predicha para el nuevo valor de los regresores
 3. Computar diferencia en (2) - (1)

Ejemplo

Participación laboral femenina.

Datos

- *participation*: 1=participa; 0=no participa
- *youngkids*: número de hijos
- *education*: años de educación formal

Variable	Obs	Media	Std	Min	Max
participation	872	0.46	0.499	0	1
kids	872	1.29	1.113	0	6
education	872	9.31	3.036	1	21

Ejemplo: Participación laboral femenina.

```
m10 <- lm_robust(participation ~ kids, data=dl, se_type = "stata")  
  
summary(m10)
```

Call:

```
lm_robust(formula = participation ~ kids, data = dl, se_type = "stata")
```

Standard error type: HC1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.46968	0.0260	18.069	0.000	0.419	0.521	870
kids	-0.00758	0.0152	-0.498	0.619	-0.037	0.022	870

Multiple R-squared: 0.000287 , Adjusted R-squared: -0.000863

F-statistic: 0.248 on 1 and 870 DF, p-value: 0.619

MPL

- ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 3 hijos relativo a una mujer con 2 hijos?
- Mujer con dos 2 hijos:

```
m10$coefficients[1] + m10$coefficients[2]*2
```

(Intercept)
0.455

- Mujer con dos 3 hijos:

```
m10$coefficients[1] + m10$coefficients[2]*3
```

(Intercept)
0.447

Discusión en clases: ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 6 hijos relativo a una mujer con 4 hijos?

Probit

Ejemplo: Participación laboral femenina.

```
m11 <- glm(participation ~ kids, data=dl, family = binomial(link="probit")); summary(m11)
```

Call:

```
glm(formula = participation ~ kids, family = binomial(link = "probit"),
     data = dl)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.13	-1.11	-1.09	1.24	1.31

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.0761	0.0652	-1.17	0.24
kids	-0.0191	0.0382	-0.50	0.62

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1203.2 on 871 degrees of freedom
Residual deviance: 1203.0 on 870 degrees of freedom
AIC: 1207

Number of Fisher Scoring iterations: 3

Probit

Ejemplo: Participación laboral femenina.

- ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 3 hijos relativo a una mujer con 2 hijos?
- Mujer con 2 hijos:
- Mujer con 3 hijos:

$$\begin{aligned}F(\hat{\beta}_0 + \hat{\beta}_1 X) &= F(-0.0761 + -0.0191 * kids) = \\F(-0.0761 + -0.0191 * kids)0 &= \\F(-0.0761 + -0.0191 * 2) &= F(-0.114) = \\F(-0.114) &= F(z \leq -0.114) = 0.545\end{aligned}$$

$$\begin{aligned}F(\hat{\beta}_0 + \hat{\beta}_1 X) &= F(-0.0761 + -0.0191 * kids) = \\F(-0.0761 + -0.0191 * kids) &= \\F(-0.0761 + -0.0191 * 3) &= F(-0.133) = \\F(-0.133) &= F(z \leq -0.133) = 0.553\end{aligned}$$

```
pnorm(-0.114, mean = 0, sd = 1,  
      lower.tail = TRUE)
```

[1] 0.455

```
pnorm(-0.133, mean = 0, sd = 1,  
      lower.tail = TRUE)
```

[1] 0.447

Discusión en clase: ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 6 hijos relativo a una mujer con 4 hijos?

Logit

Ejemplo: Participación laboral femenina.

```
m12 <- glm(participation ~ kids, data=dl, family = binomial(link="logit"))
summary(m12)
```

Call:

```
glm(formula = participation ~ kids, family = binomial(link = "logit"),
     data = dl)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.13	-1.11	-1.09	1.24	1.31

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1214	0.1042	-1.16	0.24
kids	-0.0306	0.0611	-0.50	0.62

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1203.2 on 871 degrees of freedom
Residual deviance: 1203.0 on 870 degrees of freedom
AIC: 1207

Number of Fisher Scoring iterations: 3

Logit

Ejemplo: Participación laboral femenina.

- ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 3 hijos relativo a una mujer con 2 hijos?
- Mujer con 2 hijos:
- Mujer con 3 hijos:

$$\frac{1}{1-exp(\hat{\beta}_0+\hat{\beta}_1X)} = \frac{1}{1-exp(-0.1214+-0.0306*2)}$$

```
1/(1+exp(-1*(-0.1214 + -0.0306*2)))
```

```
[1] 0.454
```

```
predict(m12, data.frame(kids=2), type = "response")
```

```
1  
0.455
```

$$\frac{1}{1-exp(\hat{\beta}_0+\hat{\beta}_1X)} = \frac{1}{1-exp(-0.1214+-0.0306*3)}$$

```
1/(1+exp(-1*(-0.1214 + -0.0306*3)))
```

```
[1] 0.447
```

```
predict(m12, data.frame(kids=3), type = "response")
```

```
1  
0.447
```

Discusión en clase: ¿Cuál es la diferencia en la probabilidad predicha de participar en el mercado laboral para una mujer con 6 hijos relativo a una mujer con 4 hijos?

Comparando modelos

Probabilidad Predicha	MPL	Probit	Logit
Kids=2	0.4545	0.4545	0.4545
Kids=3	0.4469	0.447	0.4469
Diff	-0.0076	-0.0076	-0.0076
Kids=4	0.4393	0.4394	0.4394
Kids=6	0.4242	0.4244	0.4244
Diff	-0.0152	-0.015	-0.015

Discusión en clases: ¿Qué nos dicen las similitudes entre las diferencias estimadas para cada modelo?



Semana 2

Especificaciones de Regresión Lineal

26 de septiembre, 2022

 **Pablo A. Celhay** |  pacelhay@uc.cl

Diseño y formato de la presentación:  José Daniel Conejeros |  jdconejeros@uc.cl |  JDConjeros