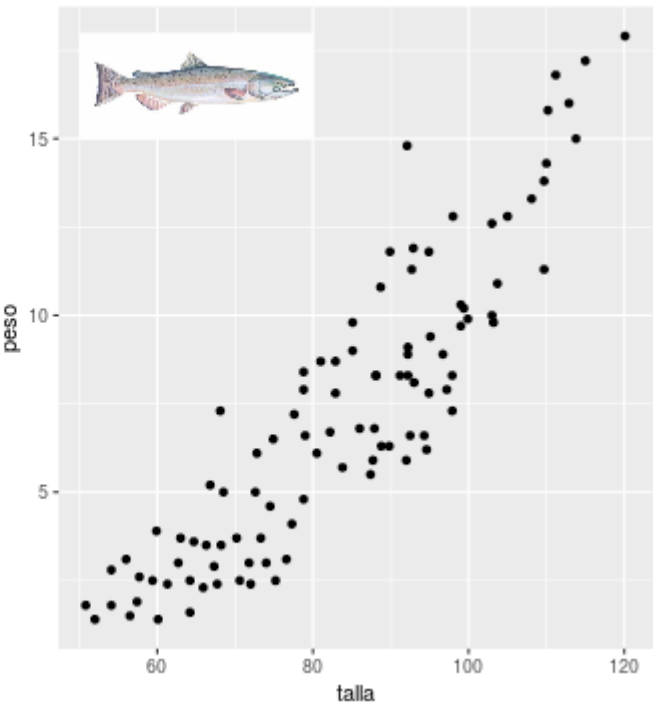
Tema 7: Correlación y Regresión Lineal

Estadística. Grado en Ciencias del Mar

Asociación Lineal

Se dispone de datos de 100 salmones capturados en tres zonas costeras de Argentina y Chile. Para cada ejemplar se ha medido su talla (en cm) y su peso (en kg). A continuación se muestran los datos de algunos de los salmones de esta muestra, y la nube de puntos talla-peso:

talla	peso	loc
68.2	3.5	Puyehue
76.6	3.1	Puyehue
103.2	9.8	Petrohue
113.8	15.0	Petrohue
112.9	16.0	Petrohue
92.0	5.9	Argentina
92.2	8.3	Petrohue
92.1	14.8	Argentina
60.1	1.4	Puyehue
92.9	11.9	Argentina
73.3	3.7	Puyehue
88.1	8.3	Petrohue
50.8	1.8	Puyehue
70.6	2.5	Puyehue
63.0	3.7	Puyehue



Asociación Lineal

Parece razonable modelar la relación talla-peso mediante una recta:

Asociación Lineal

Esta recta se denomina **recta de regresión** y su ecuación es de la forma:

$$y = b_0 + b_1 x$$

Si tomásemos otra muestra de 100 salmones en las mismas localizaciones, podríamos esperar una nube de puntos *parecida*, y por tanto unos valores *parecidos* de b_0 y b_1

Si dispusiéramos de datos de la *población* de salmones podríamos calcular la recta de regresión ajustada a la población:

$$y = \beta_0 + \beta_1 x$$

Si nuestra muestra es representativa, b_0 es un estimador de β_0 (ordenada) y b_1 es un estimador de β_1 (pendiente)

¿Cómo estimar β_0 y β_1?, es decir, ¿cómo calculamos b_0 y b_1 a partir de una muestra de puntos?

El modelo de regresión lineal simple

La recta de regresión es, en realidad, un modelo aproximado de la relación entre x e y. Para cada sujeto de la población la relación exacta es de la forma:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde ε representa la distancia entre el punto observado (x, y) y la recta.

Si podemos asumir que el valor de ε es un valor aleatorio consecuencia de *múltiples* pequeñas causas *independientes* que se suman y contribuyen a apartar el punto de la recta, por efecto del Teorema Central del Límite es razonable modelar ε como una variable aleatoria con distribución normal:

$$\varepsilon \approx N(0, \sigma_\varepsilon)$$

Regresión Lineal: Estimación por Máxima Verosimilitud

Supondremos que:

- Se dispone de n observaciones de dos variables $\{(X_i, Y_i)\,,\, i = 1, \ldots, n\}$
- Los valores de Y_i se ajustan al modelo:

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

- Los valores ε_i son $N(0, \sigma_\varepsilon)$ e independientes.
- Por tanto:
  - Para cada $i = 1, \ldots, n$:

$$Y_i \approx N\left(\beta_0 + \beta_1 X_i, \sigma_\varepsilon\right)$$

  - Para cada valor $X_i = x$ fijo:

$$E\left[Y \,|X_i = x\right] = \beta_0 + \beta_1 x$$

Es decir, los valores individuales de Y se distribuyen alrededor la recta $y = \beta_0 + \beta_1 x$, centrados en ella, y con varianza constante σ_ε^2.

Regresión Lineal: Estimación por Máxima Verosimilitud

- Como $Y_i \approx N(\beta_0 + \beta_1 X_i, \sigma_\varepsilon)$, la función de densidad de Y cuando $X = x_i$ es:

$$f_{\beta_0,\beta_1,\sigma_\varepsilon}(y\,|X = x_i) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-(\beta_0+\beta_1 x_i)}{\sigma_\varepsilon}\right)^2\right)$$

- La función de verosimilitud cuando se ha observado la muestra $\{(x_i, y_i)\,,\ i = 1, \ldots, n\}$ es entonces:

$$L(\beta_0, \beta_1, \sigma_\varepsilon) = \prod_{i=1}^{n} f_{\beta_0,\beta_1,\sigma_\varepsilon}(y_i) = \left(\frac{1}{\sigma_\varepsilon\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=i}^{n}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_\varepsilon}\right)^2\right)$$

- Tomando logaritmos se obtiene la log-verosimilitud:

$$\ell(\beta_0, \beta_1, \sigma_\varepsilon) = -n\log(\sigma_\varepsilon) - n\log\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=i}^{n}(y_i - (\beta_0+\beta_1 x_i))^2$$

Regresión Lineal: Estimación por Máxima Verosimilitud

Para obtener los valores de β_0, β_1 y σ_ε que maximizan la log-verosimilitud derivamos e igualamos a 0:

$$\frac{\partial}{\partial \beta_0} \ell\left(\beta_0, \beta_1, \sigma_\varepsilon\right) = \frac{1}{\sigma_\varepsilon^2} \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right) = 0 \Rightarrow \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right) = 0$$

$$\frac{\partial}{\partial \beta_1} \ell\left(\beta_0, \beta_1, \sigma_\varepsilon\right) = \frac{1}{\sigma_\varepsilon^2} \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right) x_i = 0 \Rightarrow \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right) x_i = 0$$

$$\frac{\partial}{\partial \sigma_\varepsilon} \ell\left(\beta_0, \beta_1, \sigma_\varepsilon\right) = -\frac{n}{\sigma_\varepsilon} + \frac{1}{\sigma_\varepsilon^3} \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2 = 0 \Rightarrow \sum_{i=i}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2 = n\sigma_\varepsilon^2$$

(Ecuaciones normales de la regresión)

Regresión Lineal: Estimación por Máxima Verosimilitud

De la primera ecuación se obtiene:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_1 x_i = 0 \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \Rightarrow \beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Regresión Lineal: Estimación por Máxima Verosimilitud

Sustituyendo en la segunda ecuación:

$$\sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \Rightarrow \sum_{i=1}^{n} (y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i) x_i = 0 \Rightarrow$$

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right) x_i - \beta_1 \sum_{i=1}^{n} \left(x_i - \bar{x}\right) x_i = 0 \Rightarrow$$

$$\beta_1 = \frac{\sum\limits_{i=1}^{n} \left(y_i - \bar{y}\right) x_i}{\sum\limits_{i=1}^{n} \left(x_i - \bar{x}\right) x_i} = \frac{\sum\limits_{i=1}^{n} \left(y_i - \bar{y}\right) \left(x_i - \bar{x}\right)}{\sum\limits_{i=1}^{n} \left(x_i - \bar{x}\right) \left(x_i - \bar{x}\right)} = \frac{S_{xy}}{S_x^2}$$

NOTA: Se ha usado que $\sum\limits_{i=1}^{n} \left(y_i - \bar{y}\right) \bar{x} = \sum\limits_{i=1}^{n} \left(x_i - \bar{x}\right) \bar{x} = 0$

Regresión Lineal: Estimación por Máxima Verosimilitud

Como $\beta_0 = \bar{y} - \beta_1 \bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_x^2} \bar{x}$$

Por último, de la tercera ecuación se obtiene:

$$\sigma_\varepsilon^2 = \frac{1}{n} \sum_{i=i}^{n} \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2$$

Sustituyendo β_0 por $\bar{y} - \beta_1 \bar{x}$, tras operar y simplificar, queda:

$$\sigma_\varepsilon^2 = \frac{1}{n} \left(\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 - \beta_1^2 \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right) = \frac{1}{n} \left((n-1) S_y^2 - \beta_1^2 (n-1) S_x^2 \right)$$

y por tanto:

$$\hat{\sigma}_\varepsilon^2 = \frac{n-1}{n} \left(S_y^2 - \hat{\beta}_1^2 S_x^2 \right) = \frac{n-1}{n} \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right)$$

Regresión Lineal: Estimación por Máxima Verosimilitud

En resumen:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Además teniendo en cuenta que el *coeficiente de correlación lineal* es:

$$r = \frac{S_{xy}}{S_x S_y}$$

se tiene finalmente:

$$\hat{\sigma}_\varepsilon^2 = \frac{n-1}{n}\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) = \frac{n-1}{n} S_y^2 \left(1 - r^2\right)$$

Regresión Lineal: Estimación por Máxima Verosimilitud

Nótese que maximizar la verosimilitud:

$$\ell\left(\beta_0, \beta_1, \sigma_\varepsilon\right) = -n\log(\sigma_\varepsilon) - n\log\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=i}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

es equivalente a minimizar:

$$D\left(\beta_0, \beta_1, \sigma_\varepsilon\right) = \sum_{i=i}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Por tanto:

Si $\varepsilon \approx N(0, \sigma_\varepsilon)$ los estimadores MV de β_0 y β_1 son los que minimizan la suma de cuadrados de las distancias de los puntos a la recta; en definitiva, producen una recta que pasa por el *centro* de la nube de puntos.

Ejemplo

Volvemos a los datos de los salmones que vimos al principio. En este caso x= *Talla* e y= *Peso*. Queremos, por tanto, ajustar la recta:

$$Peso = \beta_0 + \beta_1 \cdot Talla$$

Utilizando R obtenemos:

```
##     mean(talla)      mean(peso)      var(talla)      var(peso)
##          83.520           7.258         285.548         17.107
## cov(talla,peso)
##          62.792
```

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{62.792}{285.548} = 0.22$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.258 - 0.22 \cdot 83.52 = -11.108$$

Por tanto la recta es:

$$Peso = -11.108 + 0.22 \cdot Talla$$

Ejemplo: Regresión lineal con R

R dispone de la función `lm` para ajustar la recta de regresión:

```
recta <- lm(peso~talla,data=salmones)
recta
```

```
##
## Call:
## lm(formula = peso ~ talla, data = salmones)
##
## Coefficients:
## (Intercept)        talla
##    -11.1082       0.2199
```

Ejemplo: Regresión lineal con R

La varianza del error es:

$$\hat{\sigma}^2_\varepsilon = \frac{n-1}{n}\left(S^2_y - \frac{S^2_{xy}}{S^2_x}\right) = \frac{99}{100}\left(17.107 - \frac{62.792^2}{285.548}\right) = 3.333$$

Con R puede calcularse mediante:

```
summary(recta)$sigma^2
```

```
## [1] 3.332849
```

Interpretación de los coeficientes de la regresión:

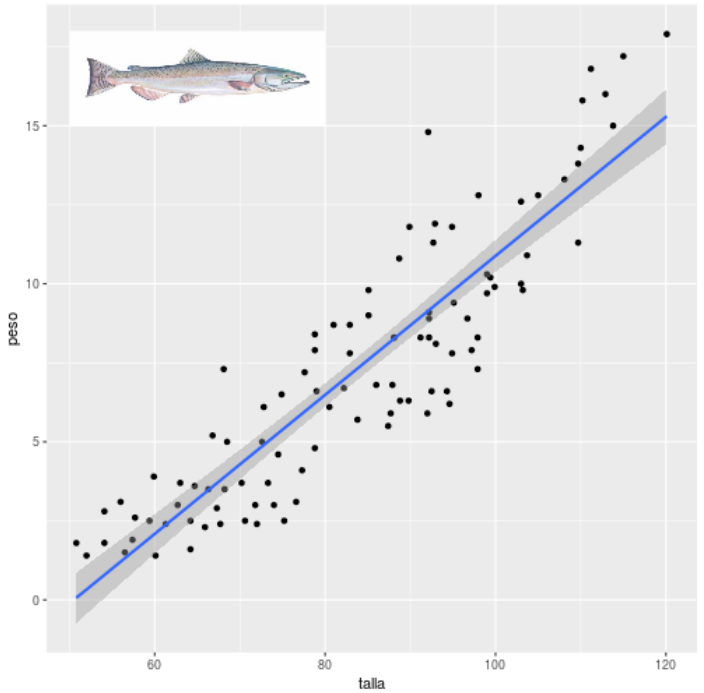
$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \approx N(0, \sigma_\varepsilon)$$

- **Pendiente** (β_1): Representa el cambio que se produce en y por cada unidad de incremento en el valor la variable x.
- **Ordenada** (β_0): Representa el valor esperado de Y cuando la $x = 0$. Sólo tiene sentido interpretar así este coeficiente cuando los puntos observados realmente pasan por $x = 0$. En caso contrario β_0 debe entenderse como un simple coeficiente de ajuste sin mayor interpretación.
- **Predicción** ($\beta_0 + \beta_1 x_i$): representa el valor esperado de Y cuando la X vale x_i.
- **Desviación típica residual** (σ_ε): representa la variabilidad de Y en torno a su valor esperado $\beta_0 + \beta_1 x_i$ cuando $X = x_i$. Se asume que es constante a lo largo de todo el recorrido de la recta.

Ejemplo: Interpretación de los coeficientes.

En el caso de los salmones:

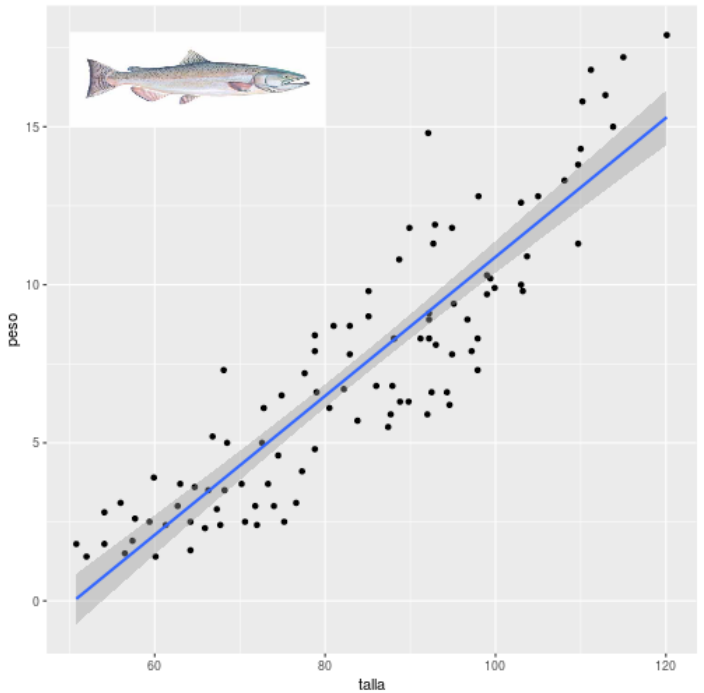
$$Peso = -11.108 + 0.22 \cdot Talla$$



- $\hat{\beta}_1 = 0.22$: significa que por cada centímetro que se incrementa la longitud de un salmón, su peso esperado se incrementa en 0.22 kg.
- $\hat{\beta}_0 = -11.108$: si quisiéramos interpretarlo como el valor de Y cuando $x = 0$, significaría que un salmón de 0 cm de longitud pesaría -11.108 kg; dado que no se han observado salmones de 0 cm de longitud (ni siquiera existen), el valor $\hat{\beta}_0$ solo puede interpetarse como un coeficiente de ajuste, necesario para que la recta pase por la nube de puntos.

Ejemplo: Interpretación de los coeficientes.

Nunca debe utilizarse una recta de regresión para extrapolar ya que, en general, no podemos estar seguros de que la relación entre X e Y sea la misma fuera del rango observado.



- Para $talla = 80\ cm$, la recta predice un peso esperado de:

$$peso = -11.108 + 0.22 \cdot 80 = 6.484\ kg$$

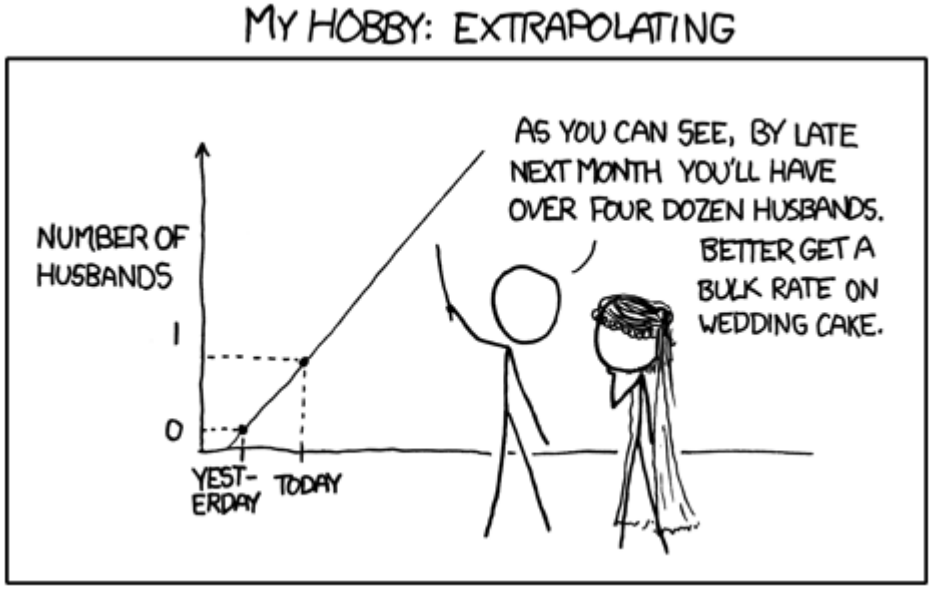
- Para $talla = 100\ cm$, la recta predice un peso esperado de:

$$peso = -11.108 + 0.22 \cdot 100 = 10.882\ kg$$

- Para $talla = 200\ cm$: no deben hacerse predicciones, pues no se han observado salmones en ese rango de talla

Interpretación de los coeficientes: no extrapolar

Los riesgos de extrapolar:



Enlace a la fuente original de la viñeta

Predicción usando la recta de regresión con R

Para realizar predicciones de la recta de regresión utilizando R se utiliza la función `predict`.

Ejemplo:

Para prededir, con la recta anterior, el peso esperado de salmones para tallas de 80 y 100 cm:

```
predict(recta,newdata=data.frame(talla=c(80,100)))
```

```
##          1          2
##  6.483947 10.881974
```

Inferencia en regresión lineal

- Cuando ajustamos una recta a una nube de puntos observados obtenemos unos valores estimados $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma}^2_\varepsilon$
- Si de la misma población observamos una nueva muestra de puntos, obtendremos otros valores de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma}^2_\varepsilon$ distintos (aunque seguramente parecidos) a los anteriores.

- ¿Cuánto se aproximan estos valores estimados a los verdaderos valores de β_0, β_1 y σ^2_ε en la población?
- Cuando se realiza una predicción utilizando los valores estimados $\hat{\beta}_0$ y $\hat{\beta}_1$, ¿qué margen de error cabe esperar en esta predicción?

Inferencia en regresión lineal simple

Si las observaciones $\{(x_i, y_i)\ i = 1, \ldots, n\}$ se ajustan al modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, siendo los $\varepsilon_i \approx N(0, \sigma_\varepsilon)$ e **independientes**, se puede demostrar que:

$$\frac{\beta_1 - \hat{\beta}_1}{\hat{\sigma}_\varepsilon \sqrt{\frac{1}{S_x^2 (n-1)}}} \approx t_{n-2}$$

$$\frac{\beta_0 - \hat{\beta}_0}{\hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2 (n-1)}}} \approx t_{n-2}$$

$$\frac{(n-2)\,\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \approx \chi^2_{n-2}$$

Inferencia en regresión lineal simple: Intervalos de confianza

De las distribuciones anteriores es fácil deducir los siguientes intervalos de confianza:

- Pendiente:

$$\beta_1 \in \left[\hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}_\varepsilon}{\sqrt{n-1}\, S_x}\right]$$

- Ordenada:

$$\beta_0 \in \left[\hat{\beta}_0 \pm t_{n-2,\alpha/2} \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, S_x^2}}\right]$$

Inferencia en regresión lineal simple: Intervalos de confianza

- Varianza residual σ_ε^2

$$\sigma_\varepsilon^2 \in \left(\frac{(n-2)\,\sigma_\varepsilon^2}{\chi^2_{n-2,\alpha/2}}, \frac{(n-2)\,\sigma_\varepsilon^2}{\chi^2_{n-2,1-\alpha/2}} \right)$$

Inferencia en regresión lineal simple: Intervalos de confianza

- También puede probarse que un intervalo de confianza para la **predicción** de los posibles valores de y cuando $X = x$ es:

$$y\left(x\right) \in \left[\hat{y}\left(x\right) \pm t_{n-2,\alpha/2}\hat{\sigma}_{\varepsilon}\sqrt{1+\frac{1}{n}+\frac{\left(x-\bar{x}\right)^{2}}{\left(n-1\right)S_{X}^{2}}}\right] \qquad \hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Si se desea un intervalo de confianza para el **valor medio de todas las y que se pueden observar para un x fijo**, éste es de la forma:

$$\bar{y}\left(x\right) \in \left[\hat{y}\left(x\right) \pm t_{n-2,\alpha/2}\hat{\sigma}_{\varepsilon}\sqrt{\frac{1}{n}+\frac{\left(x-\bar{x}\right)^{2}}{\left(n-1\right)S_{X}^{2}}}\right] \qquad \hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ejemplo: Intervalos de confianza para la regresión en R.

En R es muy fácil obtener los intervalos de confianza. Una vez ajustada la recta mediante `lm`, los intervalos para los coeficientes (ordenada y pendiente) se obtienen mediante `confint`:

```
recta <- lm(peso~talla,data=salmones)
recta
```

```
##
## Call:
## lm(formula = peso ~ talla, data = salmones)
##
## Coefficients:
## (Intercept)        talla
##    -11.1082       0.2199
```

```
confint(recta)
```

```
##                   2.5 %     97.5 %
## (Intercept) -12.943903 -9.2724186
## talla         0.198354  0.2414487
```

Ejemplo: Intervalos de confianza para la regresión en R.

- Los intervalos para las predicciones individuales se obtienen mediante:

```
predict(recta, newdata=data.frame(talla=c(80,100)), interval="predict
```

```
##          fit      lwr      upr
## 1  6.483947 2.842226 10.12567
## 2 10.881974 7.223767 14.54018
```

- Los intervalos para la predicción de valores medios se obtienen mediante:

```
predict(recta, newdata=data.frame(talla=c(80,100)), interval="confide
```

```
##          fit       lwr       upr
## 1  6.483947  6.113807  6.854088
## 2 10.881974 10.374679 11.389269
```

Nótese que la predicción del valor medio tiene un intervalo más estrecho que la predicción de valores individuales; es lógico que sea así, pues el valor medio es siempre menos variable que los valores individuales.

Coeficiente de correlación

El coeficiente de correlación lineal de Pearson es:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

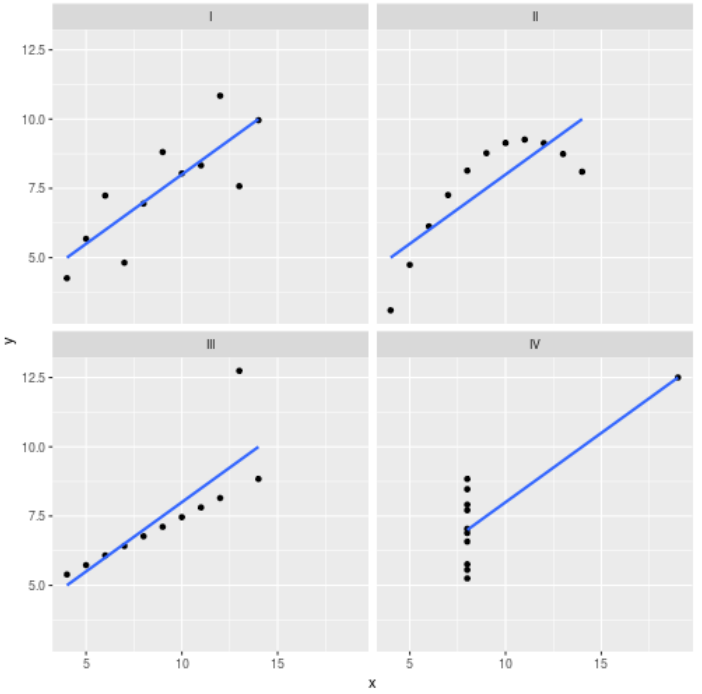
Se estima a partir de la muestra mediante:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

- Valores próximos a 1 (o a -1) indican *habitualmente* un buen ajuste a una recta de pendiente positiva (o negativa, respectivamente)
- Valores próximos a cero indican que la nube de puntos no se parece a una recta (aunque podría adoptar alguna otra forma geométrica, lo que indicaría algún otro tipo de asociación entre x e y).

Coeficiente de correlación

¡¡ Siempre conviene representar gráficamente la nube de puntos. !!



Cuarteto de Anscombe: En los cuatro casos $r = 0.82$; sin embargo las nubes de puntos son completamente diferentes y en los casos II, III y IV se apartan notablemente de la linealidad.

Datasaurus dozen: Otro ejemplo de nubes de puntos completamente diferentes y con la misma correlación aquí

Intervalo de confianza para el coeficiente de correlación

Si definimos:

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

se puede probar (Fisher) que:

$$z_r \approx N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{\sqrt{n-3}}\right)$$

Por tanto:

$$P\left(\left|z_r - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right| \leq z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right) = 1-\alpha$$

de donde:

$$\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \in \left[z_r \pm z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right]$$

Intervalo de confianza para el coeficiente de correlación

A partir de la expresión anterior, llamando:

$$z_{inf} = z_r - z_{\alpha/2}\frac{1}{\sqrt{n-3}}, \quad z_{sup} = z_r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}$$

se obtiene el siguiente intervalo de confianza a nivel $1-\alpha$ para ρ:

$$\rho \in \left[\frac{e^{2z_{inf}}-1}{e^{2z_{inf}}+1}, \frac{e^{2z_{sup}}-1}{e^{2z_{sup}}+1}\right]$$

Ejemplo: Coeficiente de correlación en R

- Cálculo del coeficiente de correlación entre la talla y el peso de los salmones:

```
with(salmones,cor(talla,peso))
```

```
## [1] 0.898414
```

- Intervalo de confianza:

```
with(salmones,cor.test(talla,peso)$conf.int)
```

```
## [1] 0.8524176 0.9306119
## attr(,"conf.level")
## [1] 0.95
```

Correlación y causalidad

Que dos variables X e Y tengan una fuerte asociación lineal (valor alto de correlación, próximo a 1 ó a -1) **no implica** que X sea la **causa** de Y, ni que Y sea la **causa** de X.

Ejemplo:

Se ha observado que cuando aumenta la venta de helados, aumenta de forma prácticamente lineal el número de personas que son atacadas por tiburones: ¿significa eso que podemos disminuir el número de ataques de tiburón simplemente vendiendo menos helados?

Correlación y causalidad

Correlación y causalidad

Obviamente la respuesta es no; la venta de helados no es la causa de los ataques de los tiburones (ni al revés), por muy fuerte que sea la asociación entre estas dos variables.

En este caso es una tercera variable (factor de confusión), la temperatura en este caso, la que dispara simultáneamente la compra de helados y el número de bañistas en la playa (que tiene como efecto colateral que hay más posibilidades de que alguien sea atacado por un tiburón)

Correlación y causalidad

- Se pueden encontrar muchos más ejemplos de correlaciones espurias (relación entre variables sin conexión lógica entre ellas) en la web tylervigen

Para establecer que la asociación entre dos variables es real y no espuria (inducida por un tercer factor oculto, o factor de confusión) es preciso **encontrar un mecanismo plausible** que explique la causalidad, y **ponerlo a prueba** mediante la realización de experimentos adecuadamente diseñados para descartar la intervención de factores ocultos.

Coeficiente de determinación R^2

La **variabilidad total** presente en la variable respuesta Y puede descomponerse en la **variabilidad explicada** por (o debida al efecto de) la variable explicativa X, más la **variabilidad residual** (variabilidad no explicada por X):

VARIABILIDAD TOTAL = VARIABILIDAD EXPLICADA + VARIABILIDAD RESIDUAL

donde:

VARIABILIDAD TOTAL: $V_T = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$

VARIABILIDAD EXPLICADA: $V_E = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$

VARIABILIDAD RESIDUAL: $V_R = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$

Es decir:

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Coeficiente de determinación R^2

Se denomina **Coeficiente de determinación** al cociente:

$$R^2 = \frac{V_E}{V_T} = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$$

Este coeficiente mide la proporción de la variabilidad en la variable respuesta Y que es explicada (o está determinada) por la variable explicativa X.

Cuanto más se aproxime su valor a 1, mejor es el modelo; cuanto más se aproxime a cero tanto peor.

- Se puede probar que $R^2 = r^2$
- R^2 se suele expresar en porcentaje.