{desafío}
latam_

Regresión Lineal_

Sesión Presencial 1

Alcances de la lectura asignada

- *Reconocer la terminología asociada a la modelación estadística.*
- *Conocer la regresión lineal y sus fundamentos.*
- *Interpretar los parámetros estimados en la regresión.*
- *Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y coeficientes.*
- *Reconocer los supuestos en los que la regresión tiene sustento teórico.*
- *Implementar un modelo de regresión con* `statsmodels`
- *Utilizar transformaciones simples en las variables independientes.*
- *Implementar un modelo predictivo con* `scikit-learn`

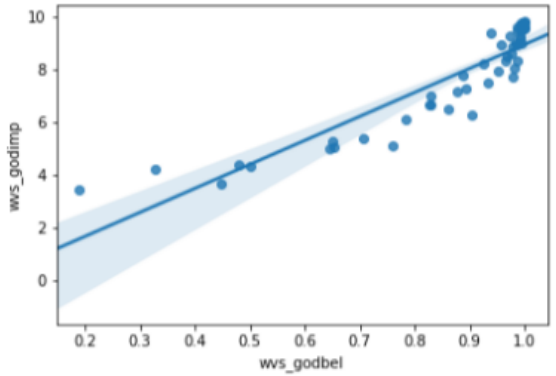
Activación de Conceptos

- En la unidad anterior aprendimos sobre la inferencia estadística y correlación.
- ¡Pongamos a prueba nuestros conocimientos!

Aproximadamente, ¿Cuál es la correlación entre ambas variables?

```
In [2]: sns.regplot(df['wvs_godbel'], df['wvs_godimp'])
```

```
Out[2]: <matplotlib.axes._subplots.AxesSubplot at 0x103bb09e8>
```



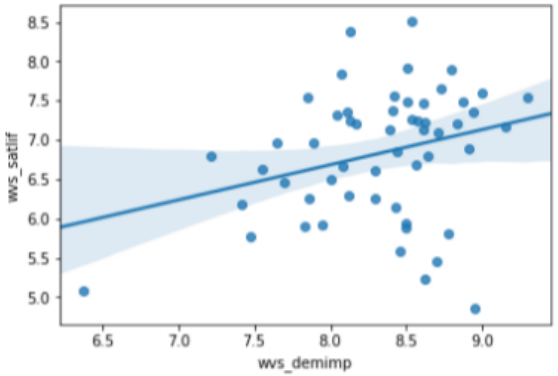
1. 0.14

- -0.3
- 0.9
- No hay suficiente información.

Aproximadamente, ¿Cuál es la correlación entre ambas variables?

In [3]:
```
sns.regplot(df['wvs_demimp'], df['wvs_satlif'])
```

Out[3]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a0b6a7a90>
```

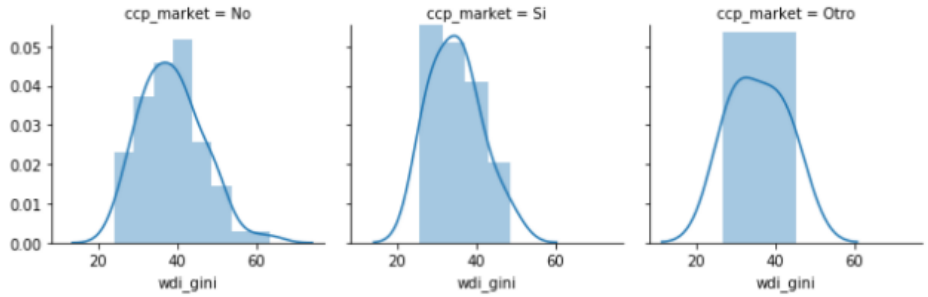


- -0.7
- .6
- No hay suficiente información.

1. 0.20

¿Cuál es la mejor manera para realizar el siguiente gráfico?

In [5]:
```
question()
```



{desafío}
latam_

A

```
for i, e in enumerate(ccp_market.unique()):
    plt.subplot(1, 3, i+1)
    plt.hist(df[df['ccp_market']==e]['wdi_gini'])
```

B

```
g = sns.PairPlot(df['ccp_market'], df['wdi_gini'])
```

C

```
g = sns.FacetGrid(df, col='ccp_market',col_wrap=3)
g = g.map(sns.distplot, 'wdi_gini')
```

D

Ninguna de las anteriores

¿Qué son los P-Value?

1. Es nuestro nivel de confianza en encontrar una hipótesis alternativa

- Es la probabilidad de encontrar resultados observados más extremos cuando H_0 es verdadera.
- Es el puntaje de corte entre la región de aceptación y rechazo
- Ninguna de las anteriores.

Si comparamos la media del índice de Gini frente a una constante de 38 y obtenemos el siguiente resultado. ¿Qué podemos decir?

In [6]:
```
stats.ttest_1samp(df['wdi_gini'].dropna(), 38)
```

Out[6]:
```
Ttest_1sampResult(statistic=-0.7331483806078863, pvalue=0.464972858236
86796)
```

1. La media del índice de Gini no es estadísticamente distinta de 38

- La media del índice de Gini es estadísticamente distinta de 38

Si comparamos el nivel de creencia en Dios entre países del medio oriente y el resto del mundo, obtenemos el siguiente resultado. ¿Qué podemos decir?

In [7]:
```
df['middle_east'] = np.where(df['ht_region']==3, 1, 0)
stats.ttest_ind(df.query('ht_region == 3')['wvs_godbel'].dropna(),
                df.query('ht_region !=3')['wvs_godbel'].dropna())
```

Out[7]:
```
Ttest_indResult(statistic=2.054023852143011, pvalue=0.0454412759545386
6)
```

---

1. Las medias entre las regiones son estadísticamente distintas al 95% de confianza

- La media de los países del medio oriente es mayor que la media del resto del mundo.
- Tenemos un 95% de confianza en que las medias son distintas.

Regresión Lineal (Desde la Econometría)

{desafío}
latam_

Rudimentos de la Regresión

- Desde la econometría, la regresión lineal implica atributos explicativos a un fenómeno en específico. Su nombre viene de la *ecuación de la recta*:

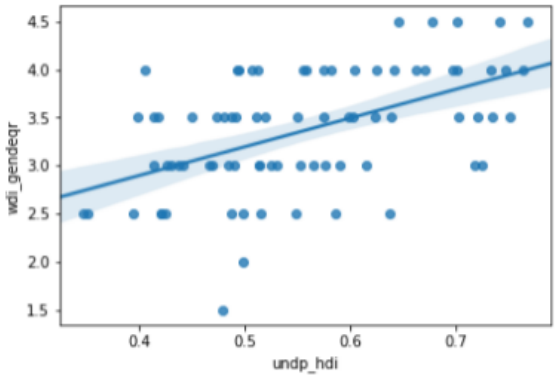
$$y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

- Donde buscamos un punto de partida para nuestra función lineal (β_0).
- Y una pendiente que indica su movimiento cuando X incrementa en 1 unidad ($\beta_1 \cdot X$)
- A diferencia de la ecuación de la recta, acá agregamos un error (ε_i)

La recta de ajuste cuando $\beta_1 > 0$

In [8]:
```
sns.regplot(df['undp_hdi'], df['wdi_gendeqr'])
```

Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a17b40438>
```

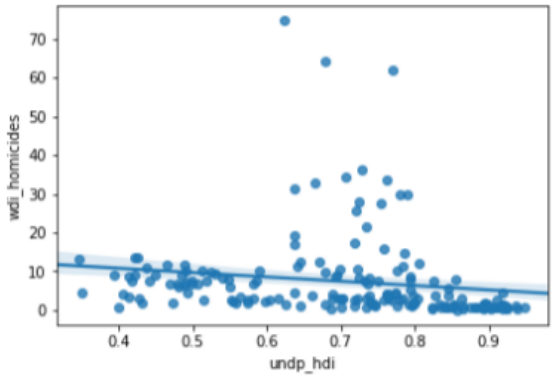


{desafío}
latam_

La recta de ajuste cuando $\beta_1 < 0$

In [9]:
```
sns.regplot(df['undp_hdi'], df['wdi_homicides'])
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a178a54e0>
```



{desafío}
latam_

La regresión en Python

- Para implementar la regresión, utilizamos `statsmodels`. Si deseamos ver el efecto del desarrollo humano en las tasas de equidad de género:

$$\texttt{wdi\_genderqr} = \beta_0 + \beta_1 \cdot \texttt{undp\_hdi} + \varepsilon_i$$

Importamos la librería

In [10]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

Generamos un objeto con nuestra ecuación

In [11]:
```
model1 = smf.ols('wdi_gendeqr ~ undp_hdi', df)
```

Ordenamos estimar el modelo

In [12]:
```
model1 = model1.fit()
```

¿Qué significa todo esto?: Bondad de ajuste

```
In [13]: results = model1.summary()
```

```
In [14]: results.tables[0]
```

Out[14]: OLS Regression Results

Dep. Variable:	wdi_gendeqr	**R-squared:**	0.269
Model:	OLS	**Adj. R-squared:**	0.259
Method:	Least Squares	**F-statistic:**	27.92
Date:	Tue, 30 Oct 2018	**Prob (F-statistic):**	1.17e-06
Time:	16:53:08	**Log-Likelihood:**	-63.048
No. Observations:	78	**AIC:**	130.1
Df Residuals:	76	**BIC:**	134.8
Df Model:	1		
Covariance Type:	nonrobust		

¿Qué significa todo esto?: Parámetros estimados

In [15]:
```
results.tables[1]
```

Out[15]:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.6987	0.320	5.303	0.000	1.061	2.337
undp_hdi	2.9911	0.566	5.284	0.000	1.864	4.118

¿Y cómo llegamos a esto?

- Para estimar los parámetros utilizamos el **método de mínimos cuadrados**, donde buscamos resolver:

$$\boldsymbol{\beta} = \arg\min_{\beta} \mathbb{E}\left[(y_i - X_i'\beta)^2\right]$$
$$= \sum_{i=0}^{N} (Y_i - (\beta_0 + \beta_1 X))^2$$

- Este método será **ELIO** cuando se satisfagan las condiciones *Gauss-Markov*.