Estadística III para Ingenieros de Sistemas

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

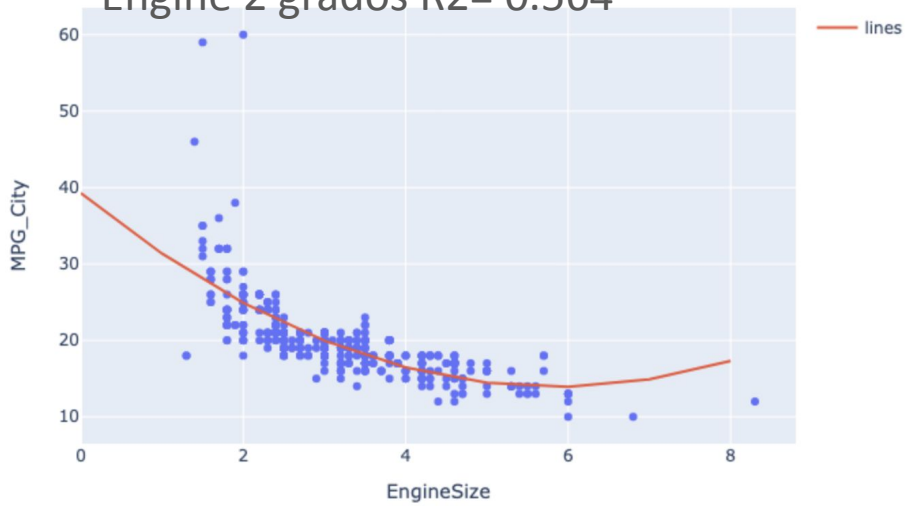
Agenda

- **anuncios varios**
  - https://forms.office.com/r/LeFfxyg4rQ
  - Parcial próxima clase (se enviará algunos ejemplos el viernes).
  - 2 tarea se enviará el Sábado.
- **modelos de analitica (machine learning-ML) Supervisado**
  - **Regresión (Resumen)**
  - **Regresión logística**
    - **K-fold**
    - **Regularización**
- **Práctica de regresión en Python**

Supervisado, Regresión. La flexibilidad vs overfitting

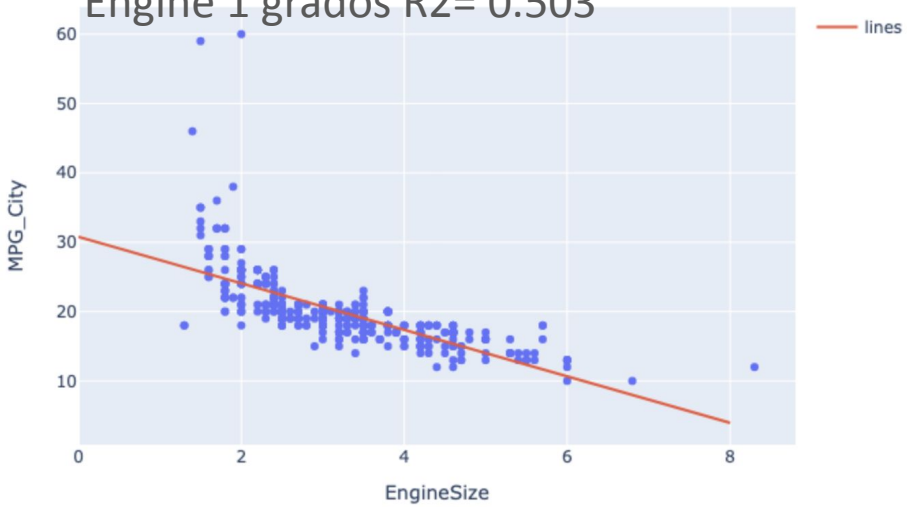
Autos Engine vs MPG

Engine 2 grados R2= 0.564



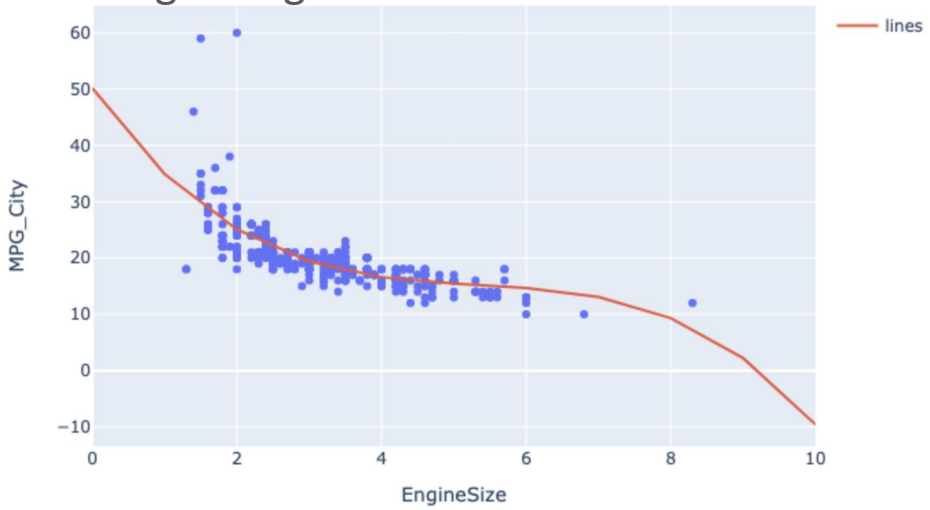
Autos Engine vs MPG

Engine 1 grados R2= 0.503



Autos Engine vs MPG

Engine 3 grados R2= 0.586



Python code

```
#define response variable
y = df_3['MPG_City']

#define predictor variables
x = df_3[['EngineSize','EngineSize_pwr2','EngineSize_pwr3']]

#add constant to predictor variables
x = sm.add_constant(x)

#fit linear regression model
model = sm.OLS(y, x).fit()

#view model summary
model.summary()
```

Supervisado, Regresión con más variables

Calcular Betas o w para muchas variables

$$\hat{y} = w^T \mathbf{x} + b = \sum_{i=1}^{p} w_i x_i + b$$

$$\min_{w} \quad \mathcal{L}(w) = \frac{1}{2} ||\mathbf{X}w - \mathbf{Y}||^2 + \frac{\lambda}{2} ||w||^2$$

$$w = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

$$e = (Xw - y)^2$$

$$\frac{de}{dw} = 2X^T(Xw - y)$$

$$\frac{de}{dw} = 2\left[X^T X w - X^T y\right] = 0$$

$$X^T X w = X^T y$$

$$(X^T X)^{-1}(X^T X) w = (X^T X)^{-1} X^T y$$

$$w = (X^T X)^{-1} X^T y$$

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión

Python code, calculate the regression by hand

```
# Calcular el mismo modelo a mano
df_3["const"] = 1
col_names = ["const",'EngineSize','EngineSize_pwr2','EngineSize_pwr3']
X = df_3[col_names]
y = df['MPG_City']
w = np.linalg.inv(X.T @ X)@(X.T@y)
for n,v in zip(col_names,w):
    print(f"{n}:{v}")
y_hat = X @ w
print(f"training mse error : {np.mean(np.power(y-y_hat,2))}")
```

```
const:50.174588047306315
EngineSize:-18.38827826767556
EngineSize_pwr2:3.3353708038930563
EngineSize_pwr3:-0.2094433106810243
training mse error : 11.329773684672682
```

=

Result using statsmodels

OLS Regression Results

Dep. Variable:	MPG_City	**R-squared:**	0.586
Model:	OLS	**Adj. R-squared:**	0.583
Method:	Least Squares	**F-statistic:**	200.2
Date:	Thu, 16 Mar 2023	**Prob (F-statistic):**	7.54e-81
Time:	23:48:37	**Log-Likelihood:**	-1126.8
No. Observations:	428	**AIC:**	2262.
Df Residuals:	424	**BIC:**	2278.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	50.1746	2.578	19.460	0.000	45.107	55.243
EngineSize	-18.3883	2.155	-8.531	0.000	-22.625	-14.152
EngineSize_pwr2	3.3354	0.553	6.035	0.000	2.249	4.422
EngineSize_pwr3	-0.2094	0.044	-4.768	0.000	-0.296	-0.123

Omnibus:	435.139	**Durbin-Watson:**	1.346
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31812.491
Skew:	4.196	**Prob(JB):**	0.00
Kurtosis:	44.394	**Cond. No.**	1.46e+03

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión con más variables

Universidad Católica de Oriente

Pasos para utilizar crear un modelo de regresión:

- Explorar los datos e identificar las variables que necesitan transformación o polinomios
- Remover outliers
- Si la variable es categórica se debe crear una dummy variable (pd.get_dummies).
- Estandarizar y escalar si existe mucha diferencia en la escala de las variables
- Dividir los datos en train (entrenamiento) y test (pruebas) (train_test_split)
- Entrenar el modelo utilizando training data
- Identificar variables que no son útiles y removerlas
- Medir el error en entrenamiento y test, identificar overfitting o underfitting

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión con más variables (Python code)

```
df_x = pd.get_dummies(df[['Type','Origin','DriveTrain']],
                      prefix=['Type','Origin','DriveTrain'],
                      dummy_na=True, drop_first=True)
df_x = pd.concat((df_x, df), axis=1)
```

```
X_train,X_test, y_train,y_test = train_test_split(x,y,test_size=0.2)
```

```
X_train = sm.add_constant(X_train)
model = sm.OLS(y_train, X_train).fit()
model.summary()
```

```
y_hat_train = model.predict(X_train)
y_hat = model.predict(X_test)
mse_train = mean_squared_error(y_train, y_hat_train)
mse_test = mean_squared_error(y_test, y_hat)
print(f"Erro calculado en train, test: {mse_train}, {mse_test} ")
```

```
Erro calculado en train, test: 5.830516328464864, 4.944361369677291
```

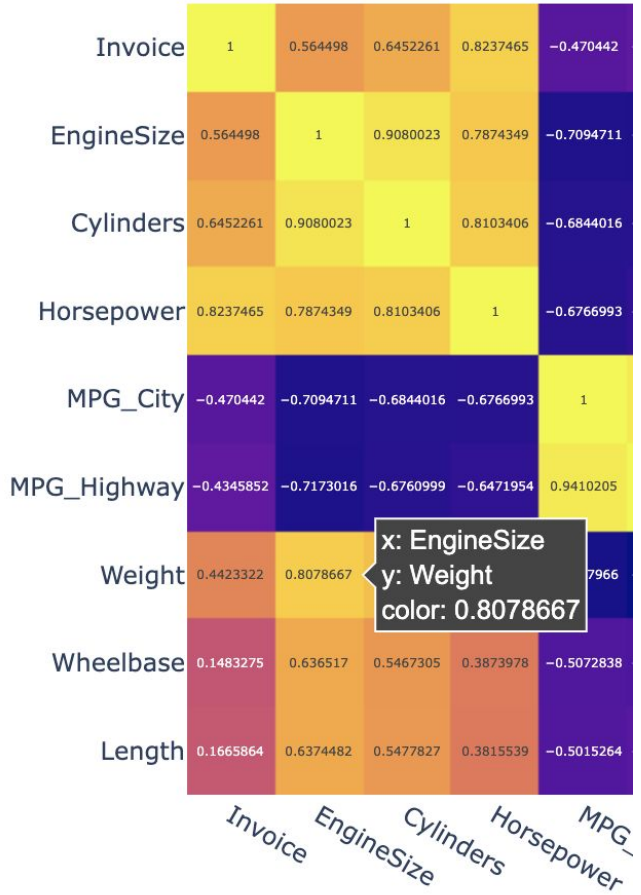
* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión con más variables (resultado)

Dep. Variable:	MPG_Highway	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.826
Method:	Least Squares	**F-statistic:**	127.6
Date:	Thu, 23 Mar 2023	**Prob (F-statistic):**	3.52e-148
Time:	13:45:34	**Log-Likelihood:**	-972.49
No. Observations:	428	**AIC:**	1979.
Df Residuals:	411	**BIC:**	2048.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.9382	3.042	19.704	0.000	53.958	65.918
Type_SUV	-24.6225	1.506	-16.353	0.000	-27.582	-21.663
Type_Sedan	-21.7343	1.434	-15.155	0.000	-24.553	-18.915
Type_Sports	-22.9680	1.528	-15.030	0.000	-25.972	-19.964
Type_Truck	-25.1097	1.565	-16.045	0.000	-28.186	-22.033
Type_Wagon	-22.0636	1.498	-14.731	0.000	-25.008	-19.119
Type_nan	2.329e-14	1.17e-15	19.951	0.000	2.1e-14	2.56e-14
Origin_Europe	0.5041	0.387	1.304	0.193	-0.256	1.264
Origin_USA	0.3023	0.314	0.963	0.336	-0.315	0.919
Origin_nan	8.239e-15	4.46e-16	18.482	0.000	7.36e-15	9.12e-15
DriveTrain_Front	1.3181	0.377	3.497	0.001	0.577	2.059
DriveTrain_Rear	-0.1537	0.421	-0.365	0.715	-0.982	0.675
DriveTrain_nan	2.765e-15	3.54e-16	7.808	0.000	2.07e-15	3.46e-15
Invoice	3.337e-05	1.41e-05	2.367	0.018	5.65e-06	6.11e-05
EngineSize	-0.0212	0.369	-0.058	0.954	-0.746	0.703
Cylinders	-0.2012	0.195	-1.030	0.304	-0.585	0.183
Horsepower	-0.0199	0.004	-4.617	0.000	-0.028	-0.011
Weight	-0.0040	0.000	-9.100	0.000	-0.005	-0.003
Wheelbase	0.0496	0.040	1.254	0.210	-0.028	0.127
Length	0.0105	0.020	0.512	0.609	-0.030	0.051

EngineSize no es una variable importante porque existen otras variables que tienen una alta correlación weight.



* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión y regresión Logística o Clasificación

Regresión logística, es una regresión en la que la variable objetivo es categórica. Por ejemplo, p(y=”yes”) or p(y=“no”)

Utiliza el mismo concepto de la regresión pero utiliza la función logit para q sin importar los valores de X, y siempre este en valores entre 0 y 1.

Utilizando el mismo dataset de Carros, vamos a definir la variable y=1, para todos los carros q un gasto menor a “23 MPG combined city and highway”

sigmoid =1/(1+e^(-wx))

MPG City	10	60	-10
logística	0.99	1	0.000045

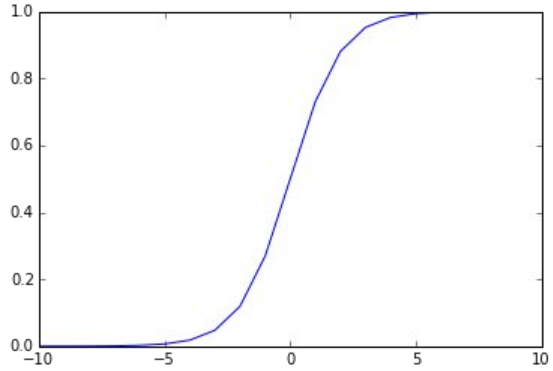
* A Course of Machine Learning http://ciml.info/
https://ecology.wa.gov/Issues-and-local-projects/Education-training/What-you-can-do/Reducing-car-pollution#:~:text=Vehicle%20pollutants%20harm%20our%20health,common%20human%2Dcaused%20greenhouse%20gas.

Supervisado, Regresión y regresión Logística

Regresión logística, encontrar los coeficientes W de una función, objetivo de reducir los errores en la clasificación.

$$\tilde{y} = \frac{1}{1 + e^{-\sum_{i=0}^{p} w_i x_i} + b}$$

La función **Sigmoid**, tiene forma de “S” y valores entre [0,1]. Por defecto retorna 1 si ỹ>0.5



Ejemplo: utilizando datos históricos del año anterior. Predecir si un alumno pasará el curso este año. ¿Cuál es el porcentaje de personas que el modelo predice que pasaron el curso?

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión y regresión Logística

Regresión logística, los resultados de la regresión logística son similares a los de la regresión. Sin embargo, en los problemas de clasificación se utilizan otras métricas para medir el error:

Labels \ Predicción	No Paso	Paso
No Paso	4 (True Negative)	1 (False Positive)10
Paso	1 (False Negative)	14 (True Positive)

accuracy = (TN + TP)/(TN+FP+FN+TP) = 18/20=0.9
precisión = TP/(TP+FP)=14/15 (cols)=0.933 (false positive)
recall = TP/(FN + TP)=14/15 (rows)=0.933 (false negative)

* A Course of Machine Learning http://ciml.info/