

Overview  
 DV (y)  
 IV (x)

V457

**V458 PACE**

Plan - understand problem context  
 Analyze - examine closely  
 perform EDA (check graphs)  
 Constant - select variables  
 transform data, code model  
 evaluate - how good?  
 Execute - prep formal results  
 + visualize

**V459 Linear Regression**

Continuous DV (y)  
 response extreme variable  
 Causation  
 N(x)  
 Explanatory predictor  
 V460 Math for Linear  
 slope beta  
 Intercept  $y = \beta_0 + \beta_1 x$   
 Correlation  
 Causation  
 loss function

**V461 Logistic Regression**

DV = Categorical (join/not join)  
 may need 2 discrete values  
 mean of y given x  
 equal to probability  
 y=1 given x  
 YES observation  
 total observations (N)

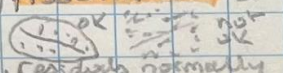
link function

V462 wrap up

**Simple Linear Regression**

V463 Continuous DV +  $y = 1$  IV  
 V465 Ordinary least squares (OLS)  
 best fit = minimize loss  
 predicted = estimated y for each value x calc by model  
 residual = observed - predicted  
 (observed - predicted) residual  
 $\sum (y_i - \hat{y}_i) = 0$  for OLS  
 SSR = Sum squared residuals

**V467 ASSUMPTIONS**

linearity:   
 Normality: residuals normally distributed  
 independent observations  
 constant variance  
 homoscedasticity  
 random error  
 V468 Code - focus on continuous  
 #sns.boxplot  
 ols\_data = df[['IV', 'DV']]  
 ols\_formula = DV ~ IV  
 from statsmodels.formula.api import ols  
 OLS = ols(ols\_formula, data=ols\_data)  
 model = OLS.fit()  
 model.summary()  
 residual (actual - fitted)  
 random? homoscedastic  
 QQ plot = normality - ggplot

**V472 uncertainty**

p-value & coefficient  
 95% CI, 95% chance  
 confidence band = confidence interval for each point on regression line

**V473 Evaluation Metrics**

$R^2$  MSE MAE (mean abs. error)  
 coeff. of determination  
 proportion of variance explained in response variable by x  
 hold out sample - test to see how well it performs

**V474 Communicate Results**

**V475 wrap up**

**Multiple Linear Regression**

V476 Intro  
 V477 1 continuous DV +  $\geq 2$  IV  
 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$   
 V478 one-hot encode  
 #dummy per category minus 1

**V480 MLR ASSUMPTIONS**

linearity, normality  
 ind. obs, homoscedasticity  
 no multicollinearity  
 VIF = variance inflation factor  
 shows correlated each IV with other IVs  
 min = 1, larger VIF more multicollinearity in the model  
 V481 Interaction terms  
 interaction term - represents how relationship between IVs is modified changes in the mean of the DV  
 V482 Interpretation Multiple Regression Results  
 data = df[['IV1', 'IV2', 'DV']]  
 X\_train, X\_test, y\_train, y\_test = train\_test\_split(df[['IV1', 'IV2']], df['DV'], test\_size=0.2, random\_state=42)  
 OLS formula = DV ~ IV1 + IV2  
 model = OLS.fit(X\_train, y\_train)  
 OLS = ols(ols\_formula, data=X\_train, y=y\_train)  
 model.summary()  
 small p-value = stat. significant  
 < has explanation

**V483 overfitting**

not all IV equally important to  $R^2$   
 fit = observed or training data too specifically unable to generalize to new data  
 hold out sample to test  
 Adj.  $R^2$  - compare multiple models

**V484 Variable selection**

selection - which include

**ANOVA, MANOVA**

V487  
 V489 Hypothesis testing  $X^2$  test of independence  
 t-test = mean of 2 groups significantly different  
 $X^2$  goodness of fit - whether observed categorical follows expected distribution  
 $H_0$  - expected count  
 $H_1$  - variable does not follow expected

$X^2$  for independence - are 2 categorical variables related  
 $H_0$  - variables independent  
 $H_1$  - variables are associated  
 observed / expected

**V490 ANOVA**

test mean difference between  $\geq 3$  groups  
 compare on continuous DV  
 all groups are similar  
 $H_0 = \mu_A = \mu_B = \mu_C$

**V491 One continuous DV to 2 categorical variables**

One-way - same as linear regression with one categorical variable  
 two-way - same as linear regression with 2 categorical variables

**V492 Code One-way, Two-way ANOVA**

sns.boxplot  
 import statsmodels.formula.api as sm  
 from statsmodels.formula.api import ols  
 data = df[['IV', 'DV']]  
 Now:  $H_0$  price = color  
 $H_1$  A  
 $H_0$  B price = cost  
 $H_1$  B  
 $H_0$  C interaction cost & color

**ANOVA MANOVA**

$X^2$  Dependent & Salary  
 $H_0$  variables independent  
 $H_1$  variables not independent

One-way = distribution level (C Salary)  
 $H_0$  - distribution equal  
 $H_1$  - distribution not equal

Two-way = same linear regression with 2 categorical variables  
 $H_0$  - distribution equal by salary  
 $H_1$  - distribution not equal by salary  
 $H_0$  - distribution equal by dept  
 $H_1$  - distribution not equal by dept  
 $H_0$  - distribution equal by dept & salary  
 $H_1$  - distribution not equal by dept & salary

Interaction

**PACE: PLAN types of test**

One-way ANOVA -  $\geq 3$  groups 1 categorical IV, 1 continuous DV