**Harold’s Statistics**

**Linear Regression Analysis**

**Cheat Sheet**

24 June 2022

**Simple Linear Regression (SLR)**

|  |  |  |
| --- | --- | --- |
| **Term** | **Formula** | **Description** |
| **Problem Statement** | Predictive Analytics: *How do we make predictions on quantitative variables from historical data using a single predictor variable?* |
| **Response Variable** | $$Y$$ | Output, outcome, dependent variable |
| **Predictor Variable** | $$X$$ | Input, covariate, independent variable |
| **Least-Squares Regression Line** | $$E\left(Y\right)=β\_{0}+β\_{1}X$$$$\hat{Y}=b\_{0}+b\_{1}X$$ | $\hat{Y}$ is the sample estimate$β\_{0} and b\_{0}$ are y-intercepts (population vs. sample)$β\_{1} and b\_{1}$ are slopes (population vs. sample)$\left(\overbar{x},\overbar{y}\right)$ is always a point on the line |
| **Regression Error** | $$ε=\hat{Y}-E(Y)$$$$Y=b\_{0}+b\_{1}X+ε$$ | ε is a random variable with:1) a **normal** distribution2) that has a **zero** mean, 3) **constant** variance, and 4) the values are **independent**.Scatterplot of five points along an increasing regression line, illustrating regression error assumptions |
| **Method of Absolute Errors** | $$\sum\_{i=1}^{n}\left|Y\_{i}-β\_{0}-β\_{1}X\_{i}\right|$$ | Minimize the sum of the magnitude of errors |
| **Method of Least Squares** | $$\sum\_{i=1}^{n}\left(Y\_{i}-β\_{0}-β\_{1}X\_{i}\right)^{2}$$ | Minimize the sum of squared errors |
| **Regression Coefficient (Slope)** | $$b\_{1}=\frac{\sum\_{}^{}\left(X\_{i}-\overbar{X}\right)\left(Y\_{i}-\overbar{Y}\right)}{\sum\_{}^{}\left(X-\overbar{X}\right)^{2}}$$$$b\_{1}=R\frac{s\_{y}}{s\_{x}}$$ | $b\_{1}$ is the slope |
| **Regression Slope Intercept** | $$b\_{0}=\overbar{Y}-b\_{1}\overbar{X}$$ | $b\_{0}$ is the y-intercept |
| **Residual Standard Error** | $$\hat{e}\_{i}=Y\_{i}-\hat{Y}\_{i}$$$$\sum\_{}^{}e\_{i}=\sum\_{}^{}\left(Y\_{i}-\hat{Y}\_{i}\right)=0$$$$\hat{e}\_{i}=Y\_{i}-\left(b\_{0}+b\_{1}X\right)$$ | Linear Regression Residual = Observed – Predictedhttp://mtweb.mtsu.edu/stats/dictionary/images/defimag/regerrors.gif |
| **Python** | **import** numpy **as** np**import** scipy.stats **as** stx = np.**array**([0, 3, 7, 10])y = np.**array**([5, 5, 27, 31])**print**(st.**linregress**(x,y)) | LinregressResult (slope=3.0, intercept=2.0, rvalue=0.9454288003008773, pvalue=0.054571199699122705, stderr=0.7310832774866965, intercept\_stderr=4.594787151274503) |
| **Standard Error of Regression Slope (s)** | $$s\_{b\_{1}}=\frac{\sqrt{\frac{\sum\_{}^{}e\_{i}^{2}}{n-2}}}{\sqrt{\sum\_{}^{}\left(X\_{i}-\overbar{X}\right)^{2}}}= \frac{\sqrt{\frac{\sum\_{}^{}\left(Y\_{i}-Y\_{i}\right)^{2}}{n-2}}}{\sqrt{\sum\_{}^{}\left(X\_{i}-\overbar{X}\right)^{2}}}$$ | Measures how spread out the Y variables are around the mean, μ.The smaller the “s” value, the closer the values are to the regression line. |

**SLR Correlation and Coefficient Determination**

|  |  |  |
| --- | --- | --- |
| **Term** | **Formula** | **Description** |
| **Problem Statement** | *How well does our regression line predict the actual data?* |
| **Correlation** | Describes the association or dependence between two variables |
| **Perfect Positive Correlation** | **Strong Positive Correlation** | **Weak Positive Correlation** |
|  |  |  |
| **Linear Correlation Coefficient (Sample)** | $$R=\frac{1}{n-1}\sum\_{}^{}\left(\frac{X-\overbar{X}}{s\_{x}}\right)\left(\frac{Y-\overbar{Y}}{s\_{y}}\right)$$$$R=\frac{g}{s\_{x}s\_{y}}$$ | Strength and direction of linear relationship / dependence between x and y.$R=\pm 1$ Perfect correlation$R=+0.9$ Positive linear relationship$R=-0.9$ Negative linear relationship$R=\~0$ No relationship**Correlation Strength |R|:**$0.80<|R|\leq 1.00$ Strong$0.40<|R|\leq 0.80$ Moderate$0<|R|\leq 0.40$ WeakCorrelation DOES NOT imply causation. |
| **Pearson Correlation Coefficient** | $$R=\frac{n\sum\_{}^{}xy-\left(\sum\_{}^{}x\right)\left(\sum\_{}^{}y\right)}{\sqrt{n\left(\sum\_{}^{}x^{2}\right)-\left(\sum\_{}^{}x\right)^{2}} \sqrt{n\left(\sum\_{}^{}y^{2}\right)-\left(\sum\_{}^{}y\right)^{2}}}$$ |
| **Correlation Matrix** | A table that shows the correlation coefficients between each pair of variables. |
| **Python** | **import** pandas **as** pdscores = pd.read\_csv("ExamScores.csv")**print**(scores[['Exam1','Exam2']].**corr**()) |
|

|  |  |  |
| --- | --- | --- |
|  | Exam1 | Exam2 |
| Exam1 | 1.000000 | 0.078613 |
| Exam2  | 0.078613 | 1.000000 |

 |
| **t-test for the Population Correlation Coefficient (*ρ*)** | $$t=\frac{R\sqrt{n-2}}{\sqrt{1-R^{2}}}$$ | A t-distribution with n-2 degrees of freedom. Hypotheses:H0 : *ρ* = 0Ha : *ρ* > 0 (right-tailed)Ha : *ρ* < 0 (left-tailed)Ha : *ρ* ≠ 0 (two-tailed) |
| **Python** | **import** pandas **as** pd**import** scipy.stats **as** stscores = pd.read\_csv("ExamScores.csv")st.**pearsonr**(scores['Exam1'],scores['Exam4']) |
| (R=0.2613,two-tailed-p-value=0.06681) |
| **Coefficient of Determination (R2)** | $$R^{2}=\frac{explained variance}{total variance}$$$$R^{2}=\frac{\sum\_{}^{}\left(\hat{Y}\_{i}-\overbar{Y}\right)^{2}}{\sum\_{}^{}\left(Y\_{i}-\overbar{Y}\right)^{2}}$$$$0\leq R^{2}\leq 1$$ | A measure of how closely the regression line follows the pattern of the data, or how well the line ﬁts the data.Measures the amount of variation in the dependent variable that is explained by the model.Represents the percent of the data that is the closest to the line of best fit. Determines how certain we can be in making predictions. |
| **Python** | **import** pandas **as** pd**import** statsmodels.api **as** sm**from** statsmodels.formula.api **import** olsscores = pd.read\_csv('ExamScores.csv')# Creates a linear regression modelresults = **ols**('Exam4 ~ Exam1', data=scores).fit()**print**(results.**summary**())NOTE: Exam4 is the response variable, Exam1 is the predictor variable |

**Analysis of Variance (ANOVA)**

|  |  |  |
| --- | --- | --- |
| **Term** | **Formula** | **Description** |
| **Problem Statement** | *How do we measure both the explained and unexplained variances?* |
| **Residual Sum of Squares (SSE)** | $$SSE=\sum\_{}^{}e\_{i}^{2}=\sum\_{}^{}\left(Y\_{i}-\hat{Y}\_{i}\right)^{2}$$ | Estimator errors |
| **Residual Degrees of Freedom (df)** | $$df=n-p$$ | Number of regression parameters |
| **Residual Mean Square (MSE)** | $$MSE=\frac{SSE}{n-p}$$ | Measures the amount of error in statistical models. 0 = no error.Mean Squared Error (MSE) - Statistics By Jim |
| **Residual Standard Error (s)** | $$s=\sqrt{MSE}$$ | Estimates the standard deviation of the residuals |
| **Python** | **import** pandas **as** pd**import** statsmodels.api **as** sm**from** statsmodels.formula.api **import** olsscores = pd.read\_csv('ExamScores.csv')# Creates a linear regression modelresults = **ols**('Exam4 ~ Exam1', data=scores).fit()**print**(results.**summary**())# The explained and unexplained variance can be obtained from the analysis of variance tableaov\_table = sm.stats.**anova\_lm**(results, typ=2)**print**(aov\_table) |
|  sum\_sq df F PR(>F)Exam1 217.166351 1.0 3.517655 0.066808Residual 2963.333649 48.0 NaN NaN |

**Testing Simple Linear Regression Parameters**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Formula** | **Description** |
| **Problem Statement** | *Does the predictor variable actually contribute to accurate regression line response predictions?* |
| **Slope Parameter (b1)** | b1 | b1 estimates β1.If β = 0, then no linear relationship exists.Sampling uncertainty could lead to a b1 ≠ 0. |
| **Test** | 1. HypothesesH0 : β1 = 0Ha : β1 ≠ 02. t-statistic (t)3. p-value (P>|t|)4. Significance level α = 0.05 | If p-value < α, then reject H0.If p-value > α, then favor H0. |
| **Python** | **import** pandas **as** pd**import** statsmodels.formula.api **as** smfdf = pd.read\_csv('Disease.csv')model = smf.**ols**('Disease ~ Time', df).**fit**()**print**(model.**summary**()) |
|  coef std err t P>|t|-------------------------------------------------------Intercept 21.0000 2.286 9.187 0.000Time -2.0000 0.406 -4.924 0.001 |
| **Intercept Parameter (b0)** | Same as above. | This test is rare. |
| **ANOVA F-test** | The association between two variables can be tested using the ANOVA -test.Since the population regression line E(Y) = β0 + β1X, determining whether an association exists between X and Y is equivalent to determining whether β ≠ 0.  |
| **Regression Sum of Squares (SSR)** | $$SSR=\sum\_{}^{}e\_{i}^{2}=\sum\_{}^{}\left(\hat{Y}\_{i}-\overbar{Y}\right)^{2}$$ | A measure that describes how well our line fits the data. |
| **Regression Degrees of Freedom (df)** | $$df=p-1$$ | Number of regression parameters.SLR has *p* = 2. |
| **Regression Mean Square (MSR)** | $$MSR=\frac{SSR}{p-1}$$ | Predicted mean-squared-anomaly. |
| **Total Sum of Squares (SSTO)** | $$SSTO=\sum\_{}^{}e\_{i}^{2}=\sum\_{}^{}\left(Y\_{i}-\overbar{Y}\right)^{2}$$$$SSTO=SSR+SSE$$ | Quantifies how much the data points, Yi, vary around their mean, $\overbar{Y}$. |
| **Total Degrees of Freedom** | $$df=n-1$$$$n-1=\left(p-1\right)+(n-p)$$ | Total degrees of freedom = regression degrees of freedom + residual degrees of freedom. |
| **Coefficient of Determination (R2)** | $$R^{2}=\frac{SSTO-SSE}{SSTO}=\frac{SSR}{SSTO}$$ | Can also be calculated using ANOVA table. |

**Multiple Linear Regression (MLR)**

|  |  |  |
| --- | --- | --- |
| **Term** | **Formula** | **Description** |
| **Problem Statement** | *How do we make predictions on quantitative variables from historical data using multiple predictor variables?* |
| **Response Variable** | $$Y$$ | Output, outcome, dependent variable |
| **Predictor Variables** | $$X\_{1}, X\_{2}, …, X\_{n}$$ | Inputs, covariates, independent variables |
| **Model Assumptions** | 1. Mean of zero2. Independence3. Normality4. Constant variance |
| **Least-Squares Regression Model** | Population: $Y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+…+β\_{n}X\_{n}+ε$Sample: $\hat{Y}=b\_{0}+b\_{1}X\_{1}+b\_{2}X\_{2}+…+b\_{n}X\_{n}+ε$ |
| **Python** | **import** pandas **as** pd**import** statsmodels.formula.api **as** smscars = pd.read\_csv('cars.csv')Y = cars['Quality']model = sms.**ols**('Y ~ speed + angle', data = cars).**fit**()**print**(model.**summary**())**print**(model.**fittedvalues**)**print**(model.**resid**) |
|  OLS Regression Results ==============================================================================Dep. Variable: Y **R-squared: 0.978**Model: OLS **Adj. R-squared: 0.975**Method: Least Squares F-statistic: 332.2Date: Mon, 15 Jul 2019 Prob (F-statistic): 3.80e-13Time: 20:48:21 Log-Likelihood: -21.142No. Observations: 18 AIC: 48.28Df Residuals: 15 BIC: 50.95Df Model: 2 Covariance Type: nonrobust ============================================================================== **coef** std err t P>|t| [0.025 0.975]------------------------------------------------------------------------------**Intercept 0.5382** 0.473 1.137 0.273 -0.471 1.547**Speed -1.9046** 0.176 -10.834 0.000 -2.279 -1.530**Angle 4.0280** 0.178 22.574 0.000 3.648 4.408==============================================================================Omnibus: 4.358 Durbin-Watson: 2.121Prob(Omnibus): 0.113 Jarque-Bera (JB): 1.414Skew: 0.082 Prob(JB): 0.493Kurtosis: 1.637 Cond. No. 14.4==============================================================================Prediction Equation: $\hat{Y}=0.5382-1.9046 X\_{1}+4.0280 X\_{2}$ |
| **Coefficient of Multiple Determination (R2)** | $$R^{2}=\frac{SSR}{SSTO}$$ | Measures the ratio of total variance in the response variable, Y, that is explained by the predictor variables X1, … ,Xn. |
| **Adjusted Coefficient of Multiple Determination (Radj2)** | $$R\_{adj}^{2}=1-(1-R^{2})\left[\frac{N-1}{N-(k+1)}\right]$$ | Allows alternative models for the same response variable to be compared.k = # predictor variables. |

**Testing Multiple Linear Regression Parameters**

|  |  |  |
| --- | --- | --- |
| **Test** | **Hypotheses** | **Research Question** |
| **Overall F-test** | Multiple regression overall F-test.Determines whether a linear relationship exists with at least one predictor variable. |
| 1. Hypotheses H0 : β1 = β2 = … = βn = 0 Ha : At least one βi ≠ 0 for i = 1, 2, …, n2. F-test (F-statistic)3. p-value (Prob (F-statistic))4. Significance level α = 0.05 | If p-value < α, then reject H0.If p-value > α, then favor H0. |
| **Individual t-test** | Multiple regression individual t-test.Determines whether a single variable has an effect. |
| 1. HypothesesH0 : βi = 0Ha : βi ≠ 02. t-statistic (t)3. p-value (P>|t|)4. Significance level α = 0.05 | If p-value < α, then reject H0.Ha : A significant linear relationship does exist.If p-value > α, then favor H0.H0 : A significant linear relationship does not exist. |