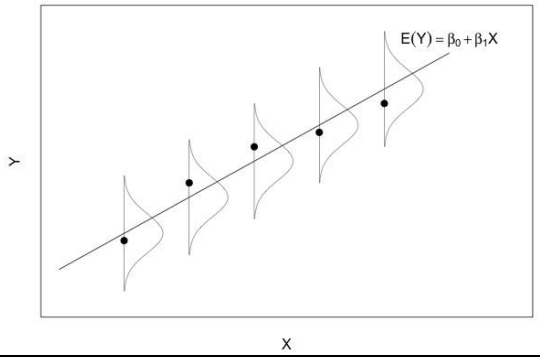
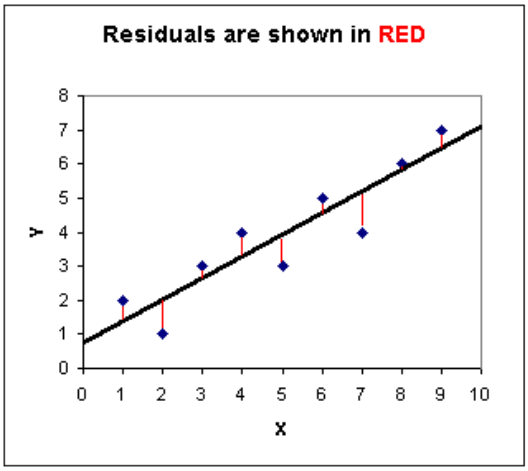
Harold's Statistics
Linear Regression Analysis
Cheat Sheet

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	$\boldsymbol{E(Y) = \beta_0 + \beta_1 X}$ $\boldsymbol{\hat{Y} = b_0 + b_1 X}$	$\hat{Y}$ 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) $(\bar{x}, \bar{y})$ is always a point on the line
Regression Error	$\boldsymbol{\varepsilon = \hat{Y} - E(Y)}$ $\boldsymbol{Y = b_0 + b_1 X + \varepsilon}$	ε is a random variable with: 1) a **normal** distribution 2) that has a **zero** mean, 3) **constant** variance, and 4) the values are **independent**. 
Method of Absolute Errors	$\boldsymbol{\sum_{i=1}^{n} \lvert Y_i - \beta_0 - \beta_1 X_i \rvert}$	Minimize the sum of the magnitude of errors
Method of Least Squares	$\boldsymbol{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}$	Minimize the sum of squared errors
Regression Coefficient (Slope)	$\boldsymbol{b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X - \bar{X})^2}}$ $\boldsymbol{b_1 = R\frac{s_y}{s_x}}$	b_1 is the slope
Regression Slope Intercept	$\boldsymbol{b_0 = \bar{Y} - b_1 \bar{X}}$	b_0 is the y-intercept

<table>
<tr>
<td>Residual Standard Error</td>
<td>$$\hat{e}_i = Y_i - \hat{Y}_i$$
$$\sum e_i = \sum (Y_i - \hat{Y}_i) = 0$$
$$\hat{e}_i = Y_i - (b_0 + b_1 X)$$</td>
<td colspan="2">Linear Regression Residual = Observed – Predicted
 
</td>
</tr>
<tr>
<td>Python</td>
<td colspan="2">

```
import numpy as np
import scipy.stats as st
x = np.array([0, 3, 7, 10])
y = np.array([5, 5, 27, 31])
print(st.linregress(x,y))
```

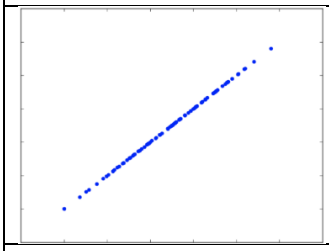
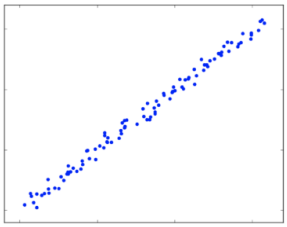
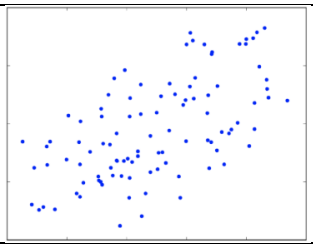
</td>
<td>

```
LinregressResult (slope=3.0,
intercept=2.0,
rvalue=0.9454288003008773,
pvalue=0.054571199699122705,
stderr=0.7310832774866965,
intercept_stderr=4.594787151274503)
```

</td>
</tr>
<tr>
<td>Standard Error of Regression Slope (s)</td>
<td colspan="2">$$s_{b_1} = \frac{\sqrt{\frac{\sum e_i^2}{n-2}}}{\sqrt{\sum (X_i - \bar{X})^2}} = \frac{\sqrt{\frac{\sum (Y_i - Y_i)^2}{n-2}}}{\sqrt{\sum (X_i - \bar{X})^2}}$$</td>
<td>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.</td>
</tr>
</table>

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)	$\boldsymbol{R = \frac{1}{n-1}\sum\left(\frac{X-\bar{X}}{s_x}\right)\left(\frac{Y-\bar{Y}}{s_y}\right)}$ $R = \frac{g}{s_x s_y}$	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 $R = \sim 0$ No relationship **Correlation Strength |R|:** $0.80 < \|R\| \leq 1.00$ Strong $0.40 < \|R\| \leq 0.80$ Moderate $0 < \|R\| \leq 0.40$ Weak Correlation DOES NOT imply causation.
Pearson Correlation Coefficient	$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\ \sqrt{n(\sum y^2) - (\sum y)^2}}$	
Correlation Matrix	A table that shows the correlation coefficients between each pair of variables.	

Python	`import pandas as pd`
	`scores = 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: $H_0: \rho = 0$ $H_a: \rho > 0$ (right-tailed) $H_a: \rho < 0$ (left-tailed) $H_a: \rho \neq 0$ (two-tailed)

<table>
<tr><td rowspan="2">Python</td><td colspan="2">

```
import pandas as pd
import scipy.stats as st
scores = pd.read_csv("ExamScores.csv")
st.pearsonr(scores['Exam1'],scores['Exam4'])
```

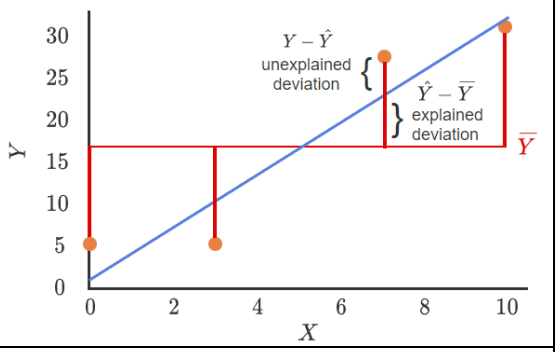
</td></tr>
<tr><td colspan="2">

```
(R=0.2613,
two-tailed-p-value=0.06681)
```

</td></tr>
<tr><td>Coefficient of Determination (R^2)</td><td>$$R^2 = \frac{explained\ variance}{total\ variance}$$
$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$
$$0 \leq R^2 \leq 1$$</td><td>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.

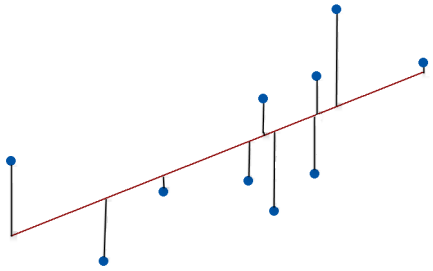
 
</td></tr>
<tr><td>Python</td><td colspan="2">

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
scores = pd.read_csv('ExamScores.csv')

# Creates a linear regression model
results = ols('Exam4 ~ Exam1', data=scores).fit()
print(results.summary())
```

NOTE: Exam4 is the response variable, Exam1 is the predictor variable</td></tr>
</table>

Analysis of Variance (ANOVA)

Term	Formula	Description
Problem Statement	*How do we measure both the explained and unexplained variances?*	
Residual Sum of Squares (SSE)	$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$	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. 
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 ols`
	`scores = pd.read_csv('ExamScores.csv')`
	
	`# Creates a linear regression model`
	`results = ols('Exam4 ~ Exam1', data=scores).fit()`
	`print(results.summary())`
	
	`# The explained and unexplained variance can be obtained from the analysis of variance table`
	`aov_table = sm.stats.anova_lm(results, typ=2)`
	`print(aov_table)`	

```
                sum_sq     df          F    PR(>F)
Exam1       217.166351    1.0   3.517655  0.066808
Residual   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 (b_1)	b_1	b_1 estimates β_1. If $\beta = 0$, then no linear relationship exists. Sampling uncertainty could lead to a $b_1 \neq 0$.
Test	1. Hypotheses $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$ 2. t-statistic (t) 3. p-value (P>|t|) 4. Significance level $\alpha = 0.05$	If p-value < α, then reject H_0. If p-value > α, then favor H_0.

Python	`import pandas as pd`
	`import statsmodels.formula.api as smf`
	`df = 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 0.000 Time -2.0000 0.406 -4.924 0.001	
Intercept Parameter (b_0)	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 $E(Y) = \beta_0 + \beta_1 X$, determining whether an association exists between X and Y is equivalent to determining whether $\beta \neq 0$.	
Regression Sum of Squares (SSR)	$SSR = \sum e_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$	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)	$SSTO = \sum e_i^2 = \sum (Y_i - \bar{Y})^2$ $SSTO = SSR + SSE$	Quantifies how much the data points, Y_i, vary around their mean, $\bar{Y}$.
Total Degrees of Freedom	$df = n - 1$ $n - 1 = (p - 1) + (n - p)$	Total degrees of freedom = regression degrees of freedom + residual degrees of freedom.
Coefficient of Determination (R^2)	$R^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO}$	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	$X_1, X_2, \ldots, X_n$	Inputs, covariates, independent variables
Model Assumptions	1. Mean of zero 2. Independence 3. Normality 4. Constant variance	
Least-Squares Regression Model	Population: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$ Sample: $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n + \varepsilon$	

Python

```
import pandas as pd
import statsmodels.formula.api as sms

cars = 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:                     332.2
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.80e-13
Time:                        20:48:21   Log-Likelihood:                -21.142
No. Observations:                  18   AIC:                             48.28
Df Residuals:                      15   BIC:                             50.95
Df 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:                   2.121
Prob(Omnibus):                  0.113   Jarque-Bera (JB):                1.414
Skew:                           0.082   Prob(JB):                        0.493
Kurtosis:                       1.637   Cond. No.                         14.4
==============================================================================
```

Prediction Equation: $\hat{Y} = 0.5382 - 1.9046\, X_1 + 4.0280\, X_2$

Coefficient of Multiple Determination (R^2)	$R^2 = \frac{SSR}{SSTO}$	Measures the ratio of total variance in the response variable, Y, that is explained by the predictor variables X_1, … ,X_n.
Adjusted Coefficient of Multiple Determination (R_{adj}^2)	$R_{adj}^2 = 1 - (1 - R^2)\left[\frac{N-1}{N-(k+1)}\right]$	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 <u>at least one</u> predictor variable.	
	1. Hypotheses $H_0: \beta_1 = \beta_2 = \ldots = \beta_n = 0$ H_a : At least one $\beta_i \neq 0$ for i = 1, 2, …, n 2. F-test (F-statistic) 3. p-value (Prob (F-statistic)) 4. Significance level $\alpha = 0.05$	If p-value < α, then reject H_0. If p-value > α, then favor H_0.
Individual t-test	Multiple regression individual t-test. Determines whether a <u>single</u> variable has an effect.	
	1. Hypotheses $H_0: \beta_i = 0$ $H_a: \beta_i \neq 0$ 2. t-statistic (t) 3. p-value (P>|t|) 4. Significance level $\alpha = 0.05$	If p-value < α, then reject H_0. H_a : A significant linear relationship does exist. If p-value > α, then favor H_0. H_0 : A significant linear relationship does not exist.