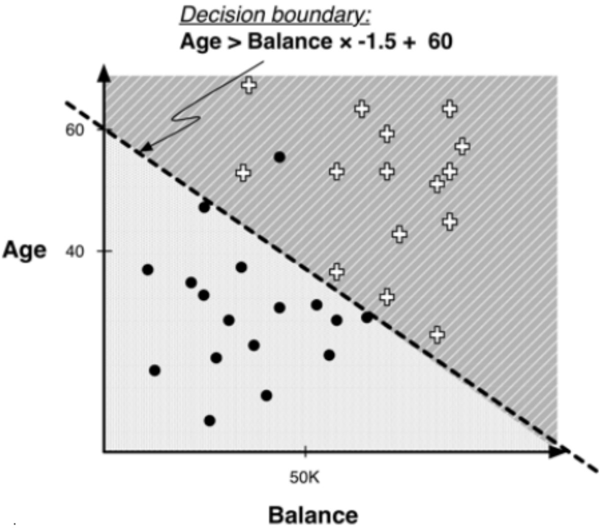
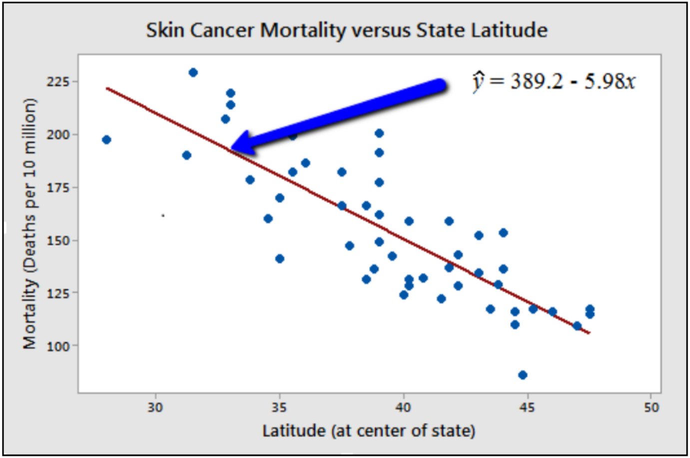
Fundamental Methods of Data Science

Class 7

Linear Classifiers



Linear Regression



Linear Regression in Python

```python
import statsmodels.formula.api as smf
est = smf.ols(formula="TRB ~ AST + STL + BLK", data=nba_data).fit()
est.summary()
```

Linear Regression in Python

```python
import statsmodels.formula.api as smf
est = smf.ols(formula="TRB ~ AST + STL + BLK", data=nba_data).fit()
est.summary()
```

```python
from sklearn import linear_model
X = wt_ht_data[['Height']]
Y = wt_ht_data['Weight']
lm = linear_model.LinearRegression()
lm.fit(X, Y)
print('Intercept is ' + str(lm.intercept_) + '\n')
print('Coefficient value of the height is ' + str(lm.coef_) + '\n')
print(pd.DataFrame(list(zip(X.columns,lm.coef_)),
                   columns = ['features', 'estimatedCoefficients']))
```

Linear Regression in Python

```python
import statsmodels.formula.api as smf
est = smf.ols(formula="TRB ~ AST + STL + BLK", data=nba_data).fit()
est.summary()
```

```python
from sklearn import linear_model
X = wt_ht_data[['Height']]
Y = wt_ht_data['Weight']
lm = linear_model.LinearRegression()
lm.fit(X, Y)
print('Intercept is ' + str(lm.intercept_) + '\n')
print('Coefficient value of the height is ' + str(lm.coef_) + '\n')
print(pd.DataFrame(list(zip(X.columns,lm.coef_)),
                   columns = ['features', 'estimatedCoefficients']))
```

- After cleaning and normalizing the data

Evaluating the Model

Dep. Variable:	TRB	**R-squared:**	0.634
Model:	OLS	**Adj. R-squared:**	0.632
Method:	Least Squares	**F-statistic:**	272.9
Date:	Fri, 06 Oct 2017	**Prob (F-statistic):**	1.10e-102
Time:	09:48:10	**Log-Likelihood:**	-853.73
No. Observations:	476	**AIC:**	1715.
Df Residuals:	472	**BIC:**	1732.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.0288	0.128	8.020	0.000	0.777	1.281
AST	0.0884	0.054	1.633	0.103	-0.018	0.195
STL	1.3464	0.221	6.100	0.000	0.913	1.780
BLK	3.7348	0.154	24.179	0.000	3.431	4.038

Omnibus:	110.206	**Durbin-Watson:**	1.716
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	314.747
Skew:	1.100	**Prob(JB):**	4.50e-69
Kurtosis:	6.321	**Cond. No.**	9.70

- Higher (Adj.) R-squared is normally better

Evaluating the Model

Dep. Variable:	TRB	**R-squared:**	0.634
Model:	OLS	**Adj. R-squared:**	0.632
Method:	Least Squares	**F-statistic:**	272.9
Date:	Fri, 06 Oct 2017	**Prob (F-statistic):**	1.10e-102
Time:	09:48:10	**Log-Likelihood:**	-853.73
No. Observations:	476	**AIC:**	1715.
Df Residuals:	472	**BIC:**	1732.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.0288	0.128	8.020	0.000	0.777	1.281
AST	0.0884	0.054	1.633	0.103	-0.018	0.195
STL	1.3464	0.221	6.100	0.000	0.913	1.780
BLK	3.7348	0.154	24.179	0.000	3.431	4.038

Omnibus:	110.206	**Durbin-Watson:**	1.716
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	314.747
Skew:	1.100	**Prob(JB):**	4.50e-69
Kurtosis:	6.321	**Cond. No.**	9.70

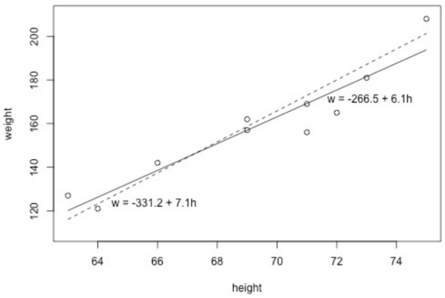
- Higher (Adj.) R-squared is normally better
- What does it mean?

Statistics

- Understanding the R-squared measure
- Understanding how to improve the model
  - Selecting correct features
- Following material is based on `https://onlinecourses.science.psu.edu/stat501/`
  - For better understanding, please follow further the online material

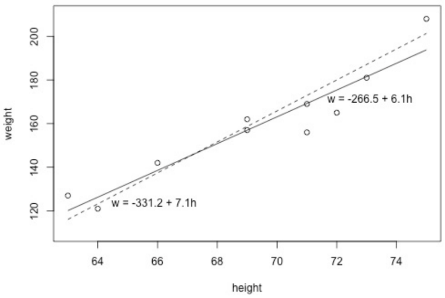
Choosing the Best Line

- Looking for correlation between Height and Weight



Choosing the Best Line

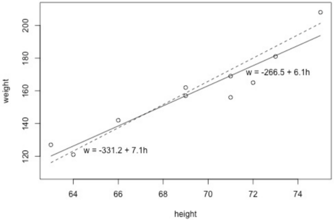
- Looking for correlation between Height and Weight



- Some notation
  - y_i - observed response for instance i
  - x_i - predictor value for instance i
  - $\hat{y}_i$ - predicted response for instance i
  - $\hat{y}_i = b_0 + b_1 x_i$ - linear formula

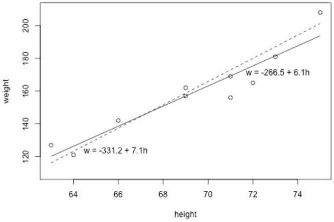
Least Square Error

- For each instance, the residual error is $e_i = y_i - \hat{y}_i$



Least Square Error

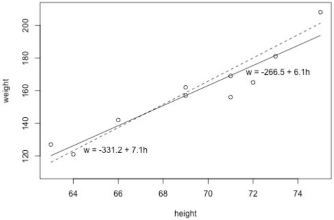
- For each instance, the residual error is $e_i = y_i - \hat{y}_i$



- Least square error - find b_0 and b_1 which minimize
  - $\Sigma_{i=0}^{n} e_i^2$

Least Square Error

- For each instance, the residual error is $e_i = y_i - \hat{y}_i$



- Least square error - find b_0 and b_1 which minimize
  - $\Sigma_{i=0}^{n} e_i^2$

- Assuming we have:
  - dashed - $\Sigma_{i=0}^{n} e_i^2 = 766$
  - solid - $\Sigma_{i=0}^{n} e_i^2 = 597$
- Which line is better?

Correlation

- Assume we've found the best line, can we now safely predict values?

Correlation

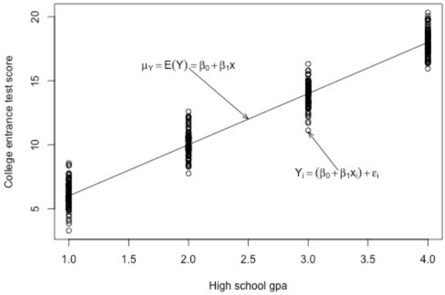
- Assume we've found the best line, can we now safely predict values?
  - We don't know if our sample match the population (later)
  - We don't know if there is a correlation between the dependent and the independent variables at all

Population Regression Line

- To know if our regression line is accurate, we can compare it against the "population" regression line

Population Regression Line

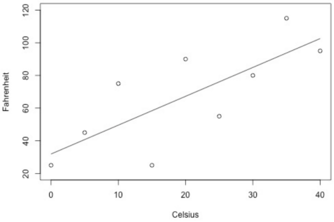
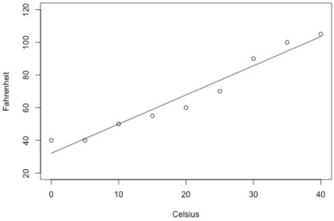
- To know if our regression line is accurate, we can compare it against the "population" regression line



- μ_y - the mean of the dependent variable for the whole population
- Each sample has an error ϵ_i
- We can see the errors ϵ_i have equal variance (σ^2)

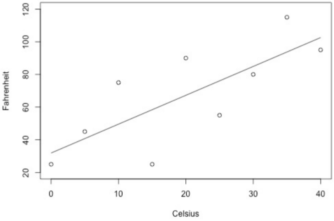
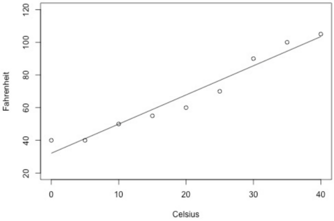
Correlation in Population and Variance

- Assume we are comparing two thermometers

Thermometer A	Thermometer B
	

Correlation in Population and Variance

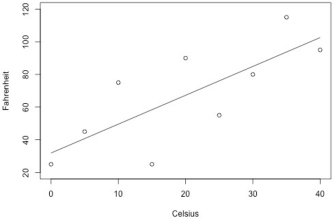
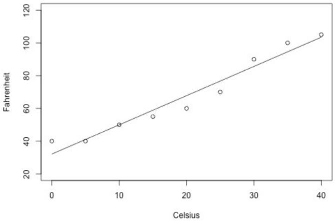
- Assume we are comparing two thermometers

Thermometer A	Thermometer B
	

- We know that $\sigma^2 = 0$ in this case, which one is more precise?

Correlation in Population and Variance

- Assume we are comparing two thermometers

Thermometer A	Thermometer B
	

- We know that $\sigma^2 = 0$ in this case, which one is more precise?

- But if we didn't know $\sigma^2 = 0$?

Estimating the Variance

- In order to compute the variance, we need to take into account the whole population
  - Normally it is impossible, what can we do?

Estimating the Variance

- In order to compute the variance, we need to take into account the whole population
  - Normally it is impossible, what can we do?
- We can estimate the variance
  - $s^2 = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$
    - n - number of samples
    - y_i - response of sample i
    - $\overline{y}$ - estimated mean

Estimating the Variance

- In order to compute the variance, we need to take into account the whole population
  - Normally it is impossible, what can we do?
- We can estimate the variance
  - $s^2 = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$
    - n - number of samples
    - y_i - response of sample i
    - $\overline{y}$ - estimated mean
- Why $n - 1$?

Estimating the Variance

- In order to compute the variance, we need to take into account the whole population
  - Normally it is impossible, what can we do?
- We can estimate the variance
  - $s^2 = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$
    - n - number of samples
    - y_i - response of sample i
    - $\overline{y}$ - estimated mean
- Why $n - 1$?
  - Since we only estimated the mean, we lose 1 “degree of freedom” and increase the variance

Mean Square Error

- How can we estimate the mean - $\overline{y}$?

Mean Square Error

- How can we estimate the mean - $\overline{y}$?
- We can estimate the mean for the set of responses for x_i using our model
  - $\hat{y}_i = b_0 + b_1 x_i$

Mean Square Error

- How can we estimate the mean - $\overline{y}$?
- We can estimate the mean for the set of responses for x_i using our model
  - $\hat{y}_i = b_0 + b_1 x_i$
- The estimated variance is
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

Mean Square Error

- How can we estimate the mean - $\overline{y}$?
- We can estimate the mean for the set of responses for x_i using our model
  - $\hat{y}_i = b_0 + b_1 x_i$
- The estimated variance is
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
- Why $n - 2$?

Mean Square Error

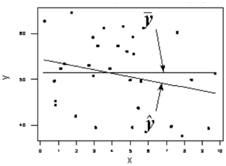
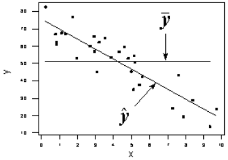
- How can we estimate the mean - $\overline{y}$?
- We can estimate the mean for the set of responses for x_i using our model
  - $\hat{y}_i = b_0 + b_1 x_i$
- The estimated variance is
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
- Why $n-2$?
  - We are estimating two values now, b_0 and b_1

Correlation Between Variables

- How can we check if our model capture a possible correlation between the variables?
  - We check if it explains the variance in the sample

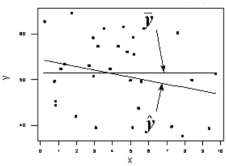
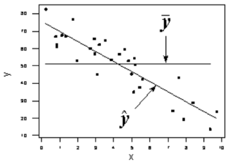
Correlation Between Variables

- How can we check if our model capture a possible correlation between the variables?
  - We check if it explains the variance in the sample
- Below there are two examples containing a regression function, which one can be useful for prediction?

Example A	Example B
	

Correlation Between Variables

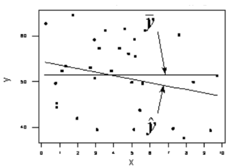
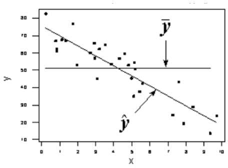
- How can we check if our model capture a possible correlation between the variables?
  - We check if it explains the variance in the sample
- Below there are two examples containing a regression function, which one can be useful for prediction?

Example A	Example B
	

- How can we determine that it is useful?

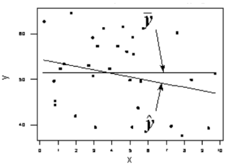
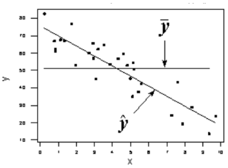
Correlation Between Variables

- How can we check if our model capture a possible correlation between the variables?
  - We check if it explains the variance in the sample
- Below there are two examples containing a regression function, which one can be useful for prediction?

Example A	Example B
	

- How can we determine that it is useful?
  - We compare it against another model

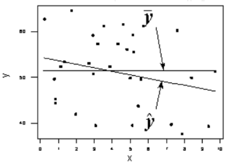
The Null Model

Example A	Example B
	

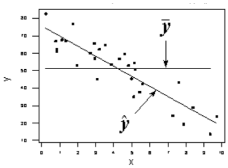
- In the above examples, we compare both functions against a constant model
  - Such a model is called the null model and it always predict $\hat{y}_i = \overline{y}$

R-squared

Example A

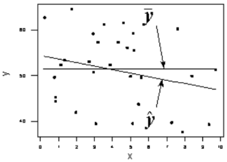
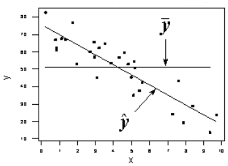


Example B



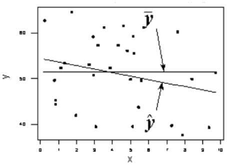
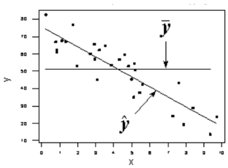
- We can now compute for each example and model the following three values
  - $SSR = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-2}$
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
  - $Tot = \frac{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}{n-2}$

R-squared

Example A	Example B
	

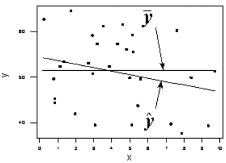
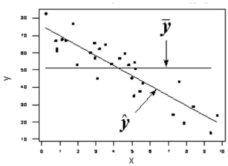
- We can now compute for each example and model the following three values
  - $SSR = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{n-2}$
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
  - $Tot = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-2}$
- $SSR_A = 12, \; MSE_A = 170, \; Tot_A = 182$

R-squared

Example A	Example B
	

- We can now compute for each example and model the following three values
  - $SSR = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{n-2}$
  - $MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
  - $Tot = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-2}$
- $SSR_A = 12, \ MSE_A = 170, \ Tot_A = 182$
- $SSR_B = 670, \ MSE_B = 170, \ Tot_B = 840$

R-squared

Example A	Example B
	

- $SSR_A = 12, \; MSE_A = 170, \; Tot_A = 182$
- $SSR_B = 670, \; MSE_B = 170, \; Tot_B = 840$
- We can now define $R - squared$
  - $R - squared = \frac{SSR}{Tot} = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}$

R-squared and Pearson Correlation Coefficient

- Remember that Pearson correlation coefficient is denoted by R
- What is the relationship between R and $R - squared$?

R-squared and Pearson Correlation Coefficient

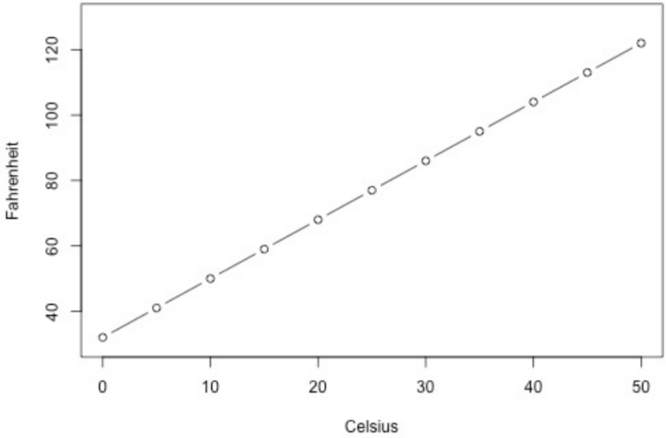
- Remember that Pearson correlation coefficient is denoted by R
- What is the relationship between R and $R - squared$?

- Couldn't we just square R then?

R-squared and Pearson Correlation Coefficient

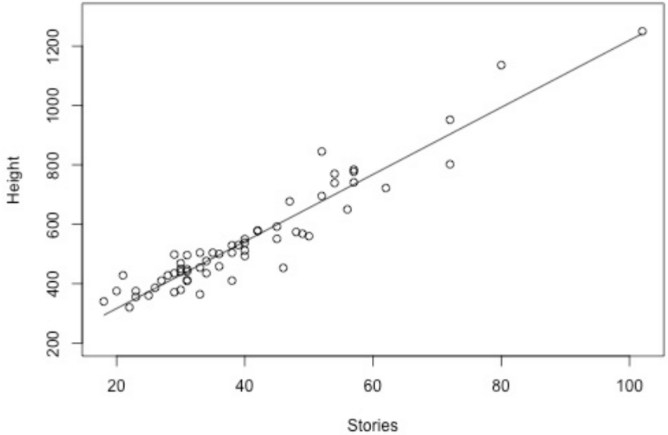
- Remember that Pearson correlation coefficient is denoted by R
- What is the relationship between R and $R - squared$?
- Couldn't we just square R then?
  - Only for simple regression functions

Examples



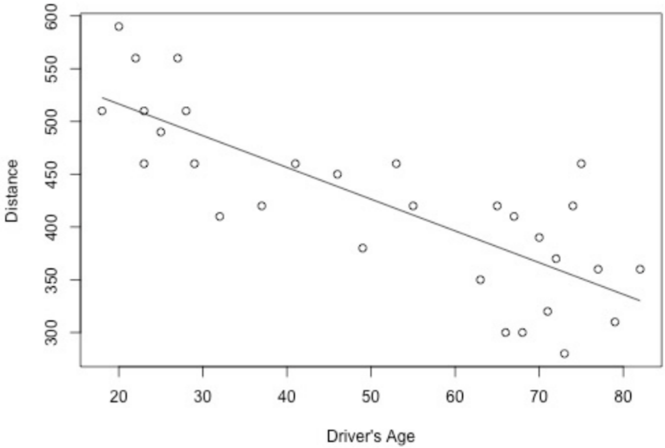
- Can you estimate R-squared? R?

Examples



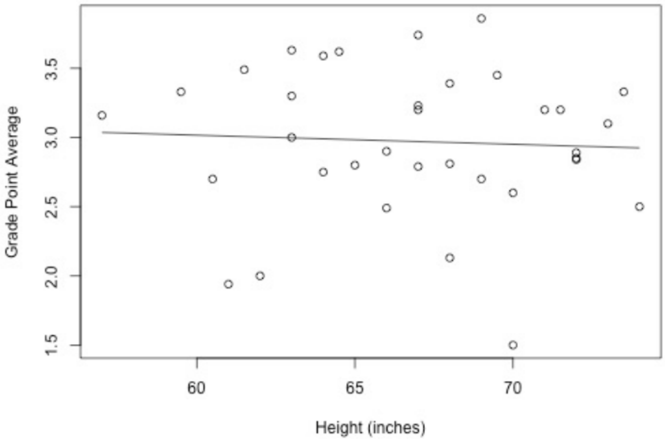
- Can you estimate R-squared? R?

Examples



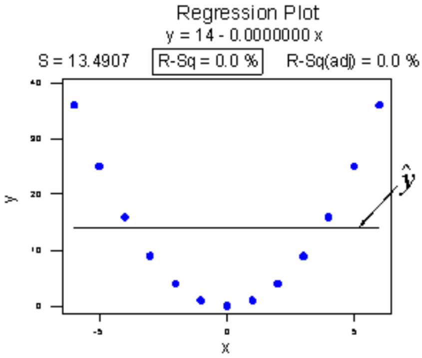
- Can you estimate R-squared? R?

Examples



- Can you estimate R-squared? R?

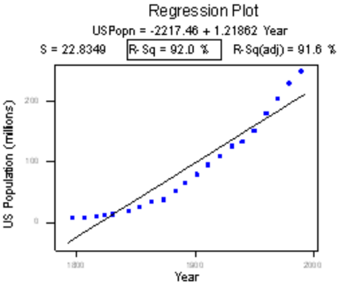
R-squared Warnings




```
Pearson correlation of x and y = 0.000
```

- R-squared relates to linear relationship

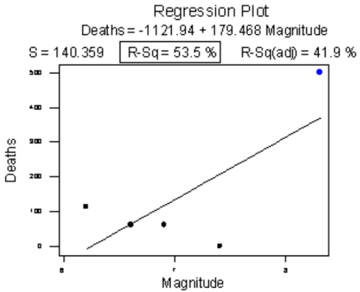
R-squared Warnings



`Pearson correlation of Year and USPopn = 0.959`

- There might be a better function

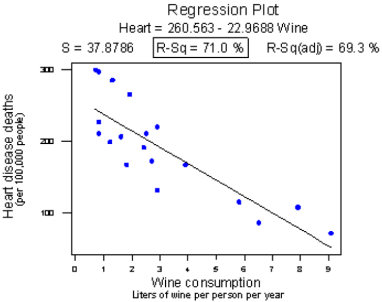
R-squared Warnings




```
Pearson correlation of Deaths and Magnitude = 0.732
```

- Sensitive to outliers

R-squared Warnings



Pearson correlation of Wine and Heart = -0.843

- Correlation does not imply causation

Hypothesis Test for the Population Correlation Coefficient

- All our computations so far were based on sample data
- How can we generalize our observations to the whole population?

Hypothesis Test for the Population Correlation Coefficient

- All our computations so far were based on sample data
- How can we generalize our observations to the whole population?

- We test our hypothesis that our data behaves in a certain way

Criminal Trial Analogy

- Null hypothesis (H_0) - Defendant is not guilty
- Alternative hypothesis (H_1) - Defendant is guilty

Criminal Trial Analogy

- Null hypothesis (H_0) - Defendant is not guilty
- Alternative hypothesis (H_1) - Defendant is guilty

- Jury uses evidence (sample data) to make a decision
  - If there is sufficient evidence to refute the assumption of innocence, they deem the defendant as guilty (they reject the null hypothesis)
  - If there is insufficient evidence, they do not reject the null evidence and the defendant is deemed innocent

Test Statistic and P-values

- How do we make decision?
  - We obtain the evidence (sample data) as a value denoting the behavior of the data
    - This value is called the **test statistic**
  - We check the probability of the test statistic to be this value given the null hypothesis
    - This is the **P-value**
  - If it is very low, we reject the null hypothesis and accept the alternative one

Hypothesis Test for the Population Correlation Coefficient

- When testing for population correlation
  - Test statistic: $t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
  - Null hypothesis: there is no correlation
  - Alternative hypothesis: there is some correlation
  - Compute the probability (P-value) that we have t^* given the null hypothesis
  - If the P-value is sufficiently small, reject the null hypothesis

Hypothesis Test for the Population Correlation Coefficient

- Our dependent variable is total rebounds

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0288	0.128	8.020	0.000	0.777	1.281
AST	0.0884	0.054	1.633	0.103	-0.018	0.195
STL	1.3464	0.221	6.100	0.000	0.913	1.780
BLK	3.7348	0.154	24.179	0.000	3.431	4.038

Adjusted R-squared in Multiple Linear Regression

- For every additional feature added to the model, the R-squared increases
  - Our model can never explain less variance
- In addition, having more features increases the chance of over-fitting

Adjusted R-squared in Multiple Linear Regression

- For every additional feature added to the model, the R-squared increases
  - Our model can never explain less variance
- In addition, having more features increases the chance of over-fitting

- Adjusted R-squared takes the number of used features into account
  - $R^2_{adj} = 1 - (\frac{n-1}{n-p})(1 - R^2)$

Having “Wrong” Predictors

- By including features which do not improve our model we incur several issues
  - We reduce the degree of freedom, which increases the estimated variance and lowers the power of our tests
  - Visualization and understanding are harder
  - Longer computation time

Example - IQ and Physical Characteristics

- Are a person's brain size and body size predictive of his or her intelligence?
- MLR Model: $IQ = b_0 + b_1 * Br + b_2 * Hht + b_3 * Wht$

Example - IQ and Physical Characteristics

- Are a person's brain size and body size predictive of his or her intelligence?
- MLR Model: $IQ = b_0 + b_1 * Br + b_2 * Hht + b_3 * Wht$

Model Summary

S	R-sq	R-sq(adj)
19.7944	29.49%	23.27%

Example - IQ and Physical Characteristics

- Are a person's brain size and body size predictive of his or her intelligence?
- MLR Model: $IQ = b_0 + b_1 * Br + b_2 * Hht + b_3 * Wht$

Model Summary

S	R-sq	R-sq(adj)
19.7944	29.49%	23.27%

Term	Coef	SE Coef	T-Value	P-Value
Constant	111.4	63.0	1.77	0.086
Brain	2.060	0.563	3.66	0.001
Height	-2.73	1.23	-2.22	0.033
Weight	0.001	0.197	0.00	0.998