CS5805 : Machine Learning I
Lecture #8

Reza Jafari, Ph.D

Collegiate Associate Professor
rjafari@vt.edu

COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH.

Simple Linear Regression

Simple Linear Regression

- Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X
- It assumes that there is approximately a linear relationship between X and Y .

$$Y \approx \beta_0 + \beta_1 X$$

- We will sometimes describe above equation by saying that we are regressing Y on X (or Y onto X).
- For example $X \rightarrow$ TV and $Y \rightarrow$ Sales

$$Sales \approx \beta_0 + \beta_1 \times TV$$

Simple Linear Regression

- Both β_0(intercept) & β_1(slope) are unknown.
- β_0 & β_1 are known as the model coefficients or parameters.
- Once we have used our training data to produce $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{y}$ indicates the prediction for Y on the basis of $X = x$.
- We use hat symbol to denote the estimated value for an unknown parameter or coefficient or to denote the predicted value of the response.

Estimating the Coefficients

- In practice β_0 & β_1 are unknown. Let $(x_1, y_1), ..., (x_n, y_n)$ represent n observation pairs. measurements of X & Y.
- Out goal is to obtain coefficients $\hat{\beta}_0$ & $\hat{\beta}_1$ such that the linear model, fit the available data so that

  $$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  for $i = 1, 2, ...n$
- There are a number of ways of measuring closeness. The most common approach involves minimizing the least squares criterion.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on i^{th} value of X.

  $$e_i = y_i - \hat{y}_i$$

Residual & Least Square Estimate (LSE)

- The e_i represents the i^{th} residuals.
- The residual sum of squares (RSS) is defined as:

$$
\begin{aligned}
RSS &= e_1^2 + e_2^2 + \ldots + e_n^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n) \\
&= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2
\end{aligned}
$$

- Taking the derivative of above equation with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and equalizing them to zero yields:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}
$$

  where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample mean.
- The above equation defines the least squares coefficient estimates for simple linear regression.

Standard Error

- How accurate is the $\hat{\beta}_0$ as an estimate of β_0? How accurate is the $\hat{\beta}_1$ as an estimate of β_1?How accurate is the sample mean of $\hat{\mu_y}$ as an estimate of μ_y?

Standard Error

- The answer to above questions is by computing standard error[1].

$$SE(\hat{\beta}_0)^2 = \sigma_\epsilon^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma_\epsilon^2}{n}$$

where $\sigma_\epsilon^2 = Var(\epsilon)$. Deviations shrink with more observations.

[1] This formula holds provided that the n observations are uncorrelated

Standard Error

- In general σ_ϵ^2 is not known, but can be estimated from the data.

Residual standard error

- This estimate of σ_ϵ is known as the residual standard error

$$\hat{\sigma_\epsilon} = RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

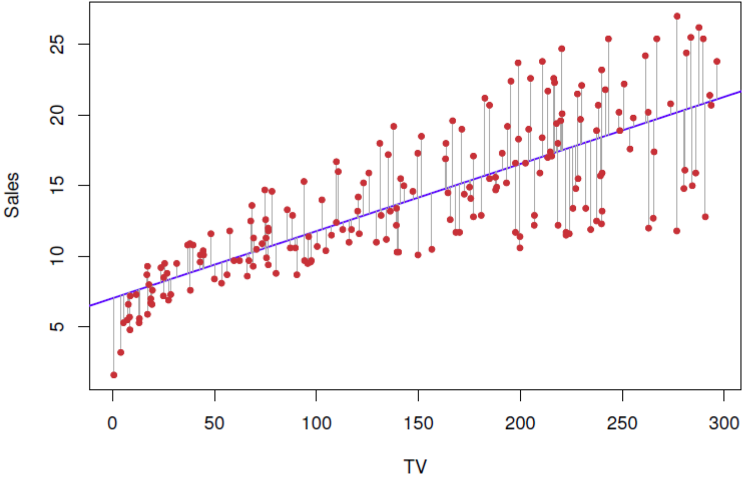
Confidence Interval

- Standard errors can be used to compute confidence interval
- For linear regression, the 95% confidence interval resides in the following range:

$$\boxed{\hat{\beta}_i \pm 2 \times SE(\hat{\beta}_i)}$$

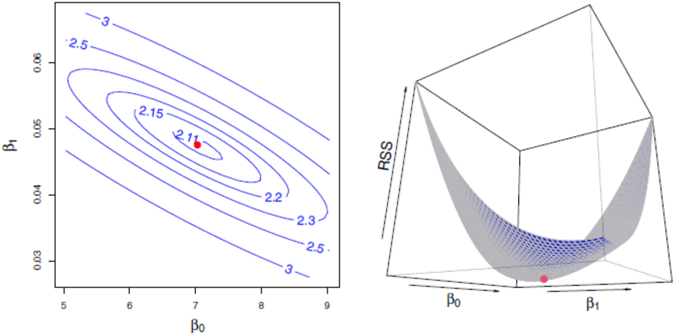
Example

- TV the predictor and Sales the response.



Example

- The red dots corresponds to the least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.



Hypothesis Test

- Standard errors can be used to perform hypothesis test on the coefficients.

Hypothesis Test

- The most common hypothesis test involves the null test hypothesis of

$$H_0 : \textit{There is } \underline{\textit{no}} \textit{ relationship between } X \textit{ and } Y (\beta_1 = 0)$$

versus the alternative hypothesis:

$$H_a : \textit{There is } \underline{\textit{some}} \textit{ relationship between } X \textit{ and } Y (\beta_1 \neq 0)$$

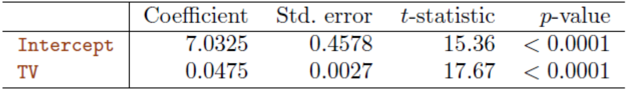
t-statistics

- If $\beta_1 = 0$ then $Y = \beta_0 + \epsilon$ and X is not associated with Y.
- To test the null hypothesis, we need to determine if $\hat{\beta}_1$ (estimate of β_1) is sufficiently far from zero that we can confident that β_1 is non-zero.
- How far is **far enough**?
- If $SE(\hat{\beta}_1)$ is small, then even relatively small value of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0 \Rightarrow$ there is a relationship between X & Y.
- In contrast, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis.
- In practice, we compute a t-statistic given bellow and compare with the t-statistics from the t-distribution from the t-table with $DOF = n - 2$.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

t-test example

- If the p-value of the t-test is less than threshold (5%) then we reject the null hypothesis. (Declare relationship between X & Y).
- If the p-value of the t-test is more than threshold (5%) then we fail to reject the null hypothesis. (Declare no relationship between X & Y).



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

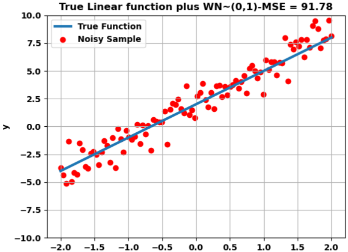
- In above example we declare $\hat{\beta}_0 \neq 0$ and $\hat{\beta}_1 \neq 0$

In class assignment

- Create a True linear function as :

$$Y = 2 + 3X + \epsilon$$

where ϵ defined as WN $\sim (0, 1)$ with the size $N = 100$.
- Plot the true line versus the noisy samples.
- Estimate the mean, intercept and the slope. Calculate the standard error of mean, the intercept and the slope.
- Plot the line that minimizes the sum square errors and display the MSE.



Accuracy of the Model

- Once we have rejected the null hypothesis, it is natural to want to quantify the extent to which the model fits the data.
- The quality of a linear regression fit is typically assessed using two related quantities: the
  1. Residual standard error (RSE)
  2. Coefficient of determination R^2 statistics.

Residual standard error

- RSE is an estimate of the standard deviation of ϵ.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- The RSE is the lack of fit of the regression model to the data.
- If $\hat{y}_i \approx y_i$ for i=1,...,n then RSE will be small and we can conclude that the model fits the data very well.

Accuracy of the Model

Coefficient of determination - R^2 statistics

- Since RSE is measured in the units of Y , it is not always clear what constitutes a good RSE.
- The R^2 statistics provides an alternative measure - proportion of variance explained.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where $TSS = \sum(y_i - \overline{y})^2$ is the total sum of squares.

- TSS-RSS measures the amount of variability in the response that is explained by performing regression.
- R^2 statistics measures the proportion of variability in Y than be explained using X.

R^2 Interpretation

- $0 < R^2 < 1$: If R^2 near 0 this means regression does not explain much of the variability in Y.
- The R^2 statistics is a measure of the linear relationship between X and Y.

$$r = Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- It can be shown that in the simple linear regression setting, $R^2 = r^2$.
- The concept of correlation between the predictors and the response does not extend automatically to multiple linear regression, since correlation quantifies the association between a single pair of variables rather than between a larger number of variables.

Multiple Linear Regression

- Let suppose we have p distinct predictor. Then multiple linear regression model takes the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

- For example :

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

Estimating the Regression Coefficients

- The regression coefficients are $\beta_0, \beta_1, ..., \beta_p$ which are unknown and must be predicted.
- Given $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ we can make a predictions using

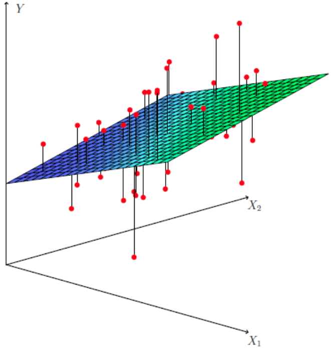
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

- The parameters are estimated using least squares approach. This means find parameters that minimizes the RSS:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

In class assignment

- A three dimensional setting two predictors and one response.
- The least squares regression lines becomes a plane.
- The plan minimizes the sum of the squared distance between each observations & the plan.



Least squares estimation

- In the regression model, we have a collection of observations (x_i, y_i) but coefficients $\beta_0, \beta_1, \ldots \beta_k$ are unknowns which needs to be estimated.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p + \epsilon$$

Least squares estimation

- In the regression model, we have a collection of observations (x_i, y_i) but coefficients $\beta_0, \beta_1, \ldots \beta_k$ are unknowns which needs to be estimated.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p + \epsilon$$

- Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$. Therefore:

$$\epsilon = y - \hat{y}$$

Least squares estimation

- In the regression model, we have a collection of observations (x_i, y_i) but coefficients $\beta_0, \beta_1, \ldots \beta_k$ are unknowns which needs to be estimated.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p + \epsilon$$

- Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$. Therefore:

$$\epsilon = y - \hat{y}$$

- If $\epsilon \to 0$ then $\hat{y} \to y$. One way to achieve this goal is by minimizing the sum square of errors by taking the derivative of the SSE w.r.t parameters and equalizing it to 0. This is called Least squares estimation.

$$\sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \ldots - \hat{\beta}_p x_p)^2$$

Normal Equations for linear regression

- Let suppose the number of observations to be n and p be the number of predictors. Then the multiple linear regression model can be written as system of linear equations:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$$

Normal Equations for linear regression

- Let suppose the number of observations to be n and p be the number of predictors. Then the multiple linear regression model can be written as system of linear equations:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$$

- where :

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{2,T} & \cdots & x_{n,p} \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Normal Equations for linear regression

- It can be proved that the solution to minimize the sum square error $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ is given as :

$$\boxed{\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}$$

Normal Equations for linear regression

- It can be proved that the solution to minimize the sum square error $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ is given as :

$$\boxed{\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}$$

- This is a least square estimator.

Normal Equations for linear regression

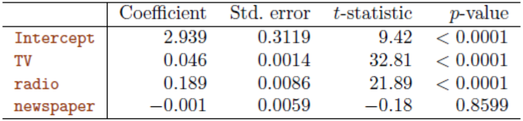
- It can be proved that the solution to minimize the sum square error $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ is given as :

$$\boxed{\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}$$

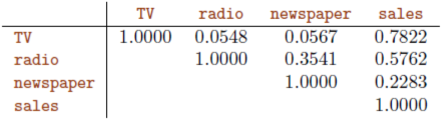
- This is a least square estimator.
- It is also called Normal equation.

Example

- Does it make sense for the multiple linear regression to suggest no association between sales and newspaper?



	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599



	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Some Important Questions

Important Question

When we perform multiple linear regression, we usually are interested in answering a few important questions:

1. *Is at least one of the predictors* $X_1, X_2, ..., X_p$ *useful in predicting the response?*
2. *Do all the predictors help to explain* Y *, or is only a subset of the predictors useful?*
3. *How well does the model fit the data?*
4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

1-Is there a relationship between the response and predictors?

F-statistics

- F-test in regression compares the fits of different linear model.
- Unlike the t-test that can access only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously.
- **Null hypothesis** : The fit of the intercept-only model and your model are equal.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

- **Alternative hypothesis**: The fit of the intercept-only model is significantly reduced compared to your model.

$$H_a : at\ least\ one\ \beta_j\ is\ non - zero$$

F-statistics

- If the p-value $<$ threshold, then you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model.
- In the intercept-only model, all of the fitted values equal the mean of the response variable.
- Therefor if the p-value $>$ threshold, your regression model predicts the response variable better than the mean of the response.
- R^2 provides an estimate of the strength of the relationship between y and $\hat{y}$. The overall F-test determines whether this relationship is statistically significant.
- If the P value for the overall F-test is less than your significance level, you can conclude that the R-squared value is significantly different from zero.

Coefficient of determination

- **Coefficient of determination** or **R-squared** is a measure used in statistical analysis that assesses how well a model explains and predicts future outcome.
- It is indicative of the level of explained variability in the data set or it is a guideline to measure the accuracy of the model.
- For example, $R^2 = 0.5$ means approximately half of the observed variation can be explained by the model.
- $0 \leq R^2 \leq 1$, the closer the R^2 is to 1 the better the fit, or relationship. A value of 1.0 indicates a perfect fit indicating very reliable model that model explains all of the variations observed. A value of 0, indicates that the model fails to accurately model the data at all.

Coefficient of determination

- **Coefficient of determination** or **R-squared** is a measure used in statistical analysis that assesses how well a model explains and predicts future outcome.
- It is indicative of the level of explained variability in the data set or it is a guideline to measure the accuracy of the model.
- For example, $R^2 = 0.5$ means approximately half of the observed variation can be explained by the model.
- $0 \leq R^2 \leq 1$, the closer the R^2 is to 1 the better the fit, or relationship. A value of 1.0 indicates a perfect fit indicating very reliable model that model explains all of the variations observed. A value of 0, indicates that the model fails to accurately model the data at all.

Coefficient of determination

- **Coefficient of determination** or **R-squared** is a measure used in statistical analysis that assesses how well a model explains and predicts future outcome.
- It is indicative of the level of explained variability in the data set or it is a guideline to measure the accuracy of the model.
- For example, $R^2 = 0.5$ means approximately half of the observed variation can be explained by the model.
- $0 \leq R^2 \leq 1$, the closer the R^2 is to 1 the better the fit, or relationship. A value of 1.0 indicates a perfect fit indicating very reliable model that model explains all of the variations observed. A value of 0, indicates that the model fails to accurately model the data at all.

Coefficient of determination

- **Coefficient of determination** or **R-squared** is a measure used in statistical analysis that assesses how well a model explains and predicts future outcome.
- It is indicative of the level of explained variability in the data set or it is a guideline to measure the accuracy of the model.
- For example, $R^2 = 0.5$ means approximately half of the observed variation can be explained by the model.
- $0 \leq R^2 \leq 1$, the closer the R^2 is to 1 the better the fit, or relationship. A value of 1.0 indicates a perfect fit indicating very reliable model that model explains all of the variations observed. A value of 0, indicates that the model fails to accurately model the data at all.

Coefficient of determination

- **Coefficient of determination** or **R-squared** is a measure used in statistical analysis that assesses how well a model explains and predicts future outcome.
- It is indicative of the level of explained variability in the data set or it is a guideline to measure the accuracy of the model.
- For example, $R^2 = 0.5$ means approximately half of the observed variation can be explained by the model.
- $0 \leq R^2 \leq 1$, the closer the R^2 is to 1 the better the fit, or relationship. A value of 1.0 indicates a perfect fit indicating very reliable model that model explains all of the variations observed. A value of 0, indicates that the model fails to accurately model the data at all.
- R^2 can be calculated as **the square of the correlation between the observed y values and the predicted $\hat{y}$.**

Coefficient of determination

- $\boldsymbol{R}^2$ is an intuitive measure of how well your linear model fits a set of observations. Fore example $R^2 = 81\%$ means that 81% of the variation in y can be explained by the relationship between x and y. The remaining 19% of the variation is unexplained and it is due to error.
- Validating a model's forecasting performance on the **test data** is much better than measuring the $\boldsymbol{R}^2$ on **training data**.
- Are high $\boldsymbol{R}^2$ values inherently good?
  - No! A high $\boldsymbol{R}^2$ does not necessary indicate that the model has a good fit.
- Are low $\boldsymbol{R}^2$ values inherently bad?
  - No! not necessary.
- R-squared can be misleading when you assess the goodness of fit for linear regression analysis

Adjusted R-squared

- Problems with R-squared:
  - R-squared cannot determine whether the coefficient estimates and predictions are **biased**, which is why you must assess the residuals plot.
  - Every time a predictor is added to a model, the **R-squared increases and not necessarily improve the model performance**.
  - If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as over-fitting the model and produces misleadingly high R^2.
- The **adjusted R-squared** is a modified version of the R-squared that has been adjusted for the number of predictors in the model. The **adjusted R-squared** increases only of the new term improves the model.
- The adjusted R-squared can be negative, but it is usually not. It is always lower than R-squared.

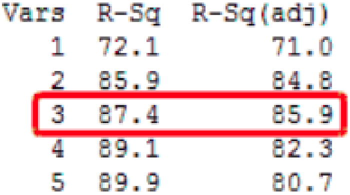
Adjusted R-squared

- Adjusted R-squared can be calculated as :

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

  where n is the number of observations and p is the number of predictors and R^2 is the coefficient of determination.

- The best model will be the one with the largest $\overline{R}^2$.



Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

- Maximizing $\overline{R}^2$ is equivalent to minimizing the standard error $\hat{\sigma}_e$ given in the next slide.

Prediction Interval

- In order to calculate the prediction interval for the regression model,

$$\hat{\sigma}_e^2 = \frac{1}{n-p-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

- Let $\boldsymbol{x}^*$ be a row vector containing the values of the predictors, $[1, x_1, x_2, \ldots, x_p]$
- Then the forecast is given by : $\hat{y} = \boldsymbol{x}^*\hat{\boldsymbol{\beta}} = \boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
- And its estimated variance is given by :

$$\hat{\sigma}_e^2[1 + \boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{x}^*)^T]$$

- A 95% prediction interval can be calculated (assuming normally distributed errors) as

$$\boxed{\hat{y} \pm 1.96\hat{\sigma}_e\sqrt{1 + \boldsymbol{x}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{x}^*)^T}}$$

Example

- Let consider the case where you are asked to predict the tip amount($) by knowing the bill amount($). Here the independent variable is Bill amount and the dependant variable is Tip amount. Find the regression model for this data set (make-up dataset). Calculate R^2 for this estimator and the confident interval for this estimate.

Bill($)	Tip($)
1	2
2	4
3	5
4	4
5	5

Selecting predictors

- When there are many possible predictors, we need some strategy for selecting the best predictors to use in a regression.
- A common approach that is **not recommended** is to *plot the forecast variable against a particular predictor and if there is no noticeable relationship, drop that predictor from the model*.
- This is *invalid* because it is not always possible to see the relationship from a scatter plot, especially when the effects of other predictors have not been accounted for.
- Another **invalid** approach is to do a multiple linear regression on all predictors and disregard all variables whose p-values are greater than 0.05.
- The p-value can be **misleading** when two or more predictors are correlated.

Selecting predictors

- Instead, you can use a measure of predictive accuracy. Five such measures are introduced.
- Compare these values against the corresponding values from other models.
- For the **CV, AIC, AICc and BIC** measures, we want to find the model with lowest values; for **adjusted R- squared** , we seek the model with highest value.

Predictive Accuracy

1. Corss-validation
2. AIC
3. AICc
4. BIC
5. Adjusted R-squared

Akaike's Information Criterion

- Akaike's Information Criterion (AIC) lets you test how well your model fit the data set without over-fitting it.
- The AIC score rewards, models that achieve a high goodness-of-fit score and penalizes them if the become overly complex.
- The AIC by itself, is not much of use unless it is compared with the AIC score of a competing model. The AIC can be calculated as :

  $$\boxed{AIC = n\log(\frac{SSE}{n}) + 2(p+2)}$$

  where n is the number if observations and p is the number of predictors.
- The model with the **minimum value of AIC is often the best model** for forecasting.

Corrected Akaike's Information Criterion

- For small values of T, the AIC tends to select too many predictors , so a bias-corrected version of the AIC has been developed:

$$AIC_c = AIC + \frac{2(p+2)(k+3)}{n-p-3}$$

As with AIC, the AIC_c should be minimized.

Schwarz's Bayesian Information Criterion (BIC)

A related measure to AIC is BIC defined as :

$$BIC = T\log(\frac{SSE}{n}) + (p+2)\log(n)$$

- Minimizing BIC is intended to give the best model.
- BIC penalizes the number of parameters more heavily than AIC

Predictor selection

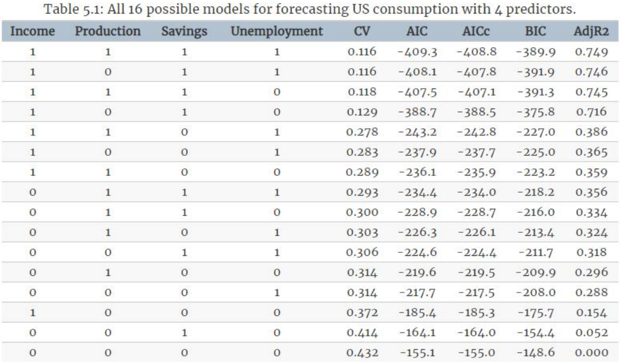
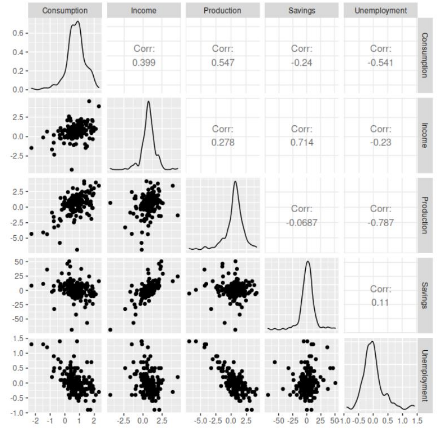


Table 5.1: All 16 possible models for forecasting US consumption with 4 predictors.

Income	Production	Savings	Unemployment	CV	AIC	AICc	BIC	AdjR2
1	1	1	1	0.116	-409.3	-408.8	-389.9	0.749
1	0	1	1	0.116	-408.1	-407.8	-391.9	0.746
1	1	1	0	0.118	-407.5	-407.1	-391.3	0.745
1	0	1	0	0.129	-388.7	-388.5	-375.8	0.716
1	1	0	1	0.278	-243.2	-242.8	-227.0	0.386
1	0	0	1	0.283	-237.9	-237.7	-225.0	0.365
1	1	0	0	0.289	-236.1	-235.9	-223.2	0.359
0	1	1	1	0.293	-234.4	-234.0	-218.2	0.356
0	1	1	0	0.300	-228.9	-228.7	-216.0	0.334
0	1	0	1	0.303	-226.3	-226.1	-213.4	0.324
0	0	1	1	0.306	-224.6	-224.4	-211.7	0.318
0	1	0	0	0.314	-219.6	-219.5	-209.9	0.296
0	0	0	1	0.314	-217.7	-217.5	-208.0	0.288
1	0	0	0	0.372	-185.4	-185.3	-175.7	0.154
0	0	1	0	0.414	-164.1	-164.0	-154.4	0.052
0	0	0	0	0.432	-155.1	-155.0	-148.6	0.000

Predictor selection



Stepwise regression

- For large number of predictors, it is not possible to fit all possible models.i.e 40 predictors leads to $2^{40} > 1$ trillion possible models!

Backwards stepwise regression

1. Start with the model containing all potential predictors.
2. Remove one predictor at a time. Keep the model of it improves the measure of predictive accuracy.
3. Iterate until no further improvement.

Stepwise regression

- If the number of potential predictors is too large, then the backwards stepwise regression will not work.

Forward stepwise regression

1. Start with the model that includes only the intercept.
2. Predictors are added at a time and the one that most improves the measure of predictive accuracy is retained.
3. The procedure is repeated until no further improvement can be achieved.

- It is important to realise that any stepwise approach is not guaranteed to lead to the best possible model, but it almost always leads to a good model.

Singular Value Decomposition

- Reducing the number of input variables for a predictive model is referred to as dimensionality reduction.
- Fewer input variables(features) can result in a simpler predictive model that may have better performance when making predictions on new data.
- Input variables that are correlated make the least square estimate to be in accurate.
- Singular Value Decomposition or SVD the popular technique for dimensionality reduction. This is a linear algebra technique to create a projection of a sparse dataset prior to fitting a model.

Singular Value Decomposition

- Let Matrix $X \in \mathbb{R}^{n \times d}$ where n is the number of observations and d is the dimension of the dataset. Then has a singular value decomposition of the form :

$$X = U \Sigma V^T$$

- where $U \in \mathbb{R}^{n \times n}$ orthogonal matrix whose columns are eigenvectors of XX^T. The columns of U is called left singular vectors of X.
- where $V \in \mathbb{R}^{d \times d}$ orthogonal matrix whose columns are eigenvectors of $X^T X$. The columns of V is called right singular vectors of X.

Singular Value Decomposition

- $\Sigma \in \mathbb{R}^{n \times d}$

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \\ 0 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & 0 \end{pmatrix}$$

- where $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n \geq 0$ are the square roots of the eigenvalues values of $X^T X$.
- The diagonal entries are called the singular values of X.

SVD and Python

- to find the singular values of feature matrix, create a matrix X where the columns are the feature and the rows are the samples(observations).
- Construct a matrix $H = (X^T X)$.
- The root square of the eigenvalues of the H is singular values of X.
- Or you can simply use python numpy linear algebra package to compute the U, Σ, V.
- $U, \Sigma, V = np.linalg.svd(X)$
- Any singular values close to zero means that one or more features are correlated. The correlated feature(s) needs to be detected and removed from the feature space.

Condition Number and co-linearity

- Another popular way to detect multi-collinearity in the dataset is called "Eigensystem Analysis" which uses the concept of Condition number. The condition number is defined as :

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \frac{\sigma_1}{\sigma_d}$$

- where λ is the eigenvalue of $H = X^T X$ and σ_i is i^{th} singular values of X.
- Condition number can also be derived using python and numpy linear algebra package **numpy.linalg.cond(X)**
- The $\kappa < 100 \implies$ Weak Degree of Co-linearity(DOC)
- The $100 < \kappa < 1000 \implies$ Moderate to Strong DOC
- The $\kappa > 1000 \implies$ Severe DOC

Nonlinear regression

- A multiple linear regression model follows a very particular form.

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

  Everything else is called **nonlinear regression**.
- The above equation is linear in the parameters, but it is possible to model curvature with this type of model.
- You can raise an independent variable by an exponent to fit a curve (quadratic-regression). For a model with one independent variable

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$
- In above example, the I.V. is squared (non-linear) but the model is still linear in the parameters.
- Quadratic regression with two variables:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + \beta_5 X_2^2$$

Multivariate quadratic regression

- Let the two features to be notated as u and v. Hence, the multivariate quadratic equation can be written in the matrix form as follow:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \Rightarrow \boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}$$

- where :

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \boldsymbol{X} = \begin{pmatrix} 1 & u_1 & v_1 & u_1^2 & u_1 v_1 & v_1^2 \\ 1 & u_2 & v_2 & u_2^2 & u_2 v_2 & v_2^2 \\ 1 & u_3 & v_3 & u_3^2 & u_3 v_2 & v_3^2 \\ \vdots & \vdots & \vdots & & \vdots & \\ 1 & u_n & v_n & u_n^2 & u_n v_n & v_n^2 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0, \epsilon_1 \ldots, \epsilon_T \end{pmatrix}^T$$

Multi-collinearity and forecasting

- Multi-collinearity occurs when **two predictors are highly correlated** with each other. (correlation coefficient close to +1 or -1).
- In this case, knowing the value one of the variables tells you a lot about the value of the other variable.
- For example, foot size can be used to predict height, but including the size of the both left and right feet in the same model is not going to make the forecast better, although it wont make them worse either.
- If there is a **high correlation**, then the estimation of the regression coefficients is **computationally difficult**.
- When multi-collinearity is present, the uncertainty associated with individual regression coefficients will be large. Consequently, **statistical test** (i.e. t-test) are **unreliable**.

Multi-collinearity and forecasting

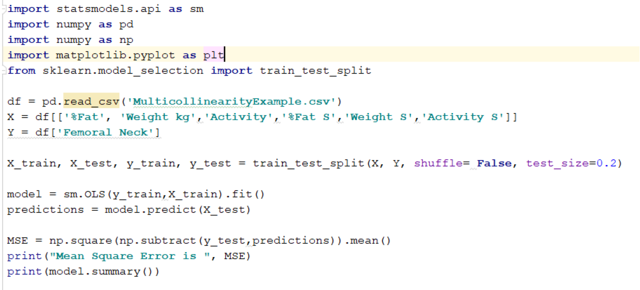
- Forecasts will be unreliable if the values of the future predictors are **outside the range of historical values of the predictors**.
- For example, suppose you have fitted a regression model with predictors x_1 and x_2 which are highly correlated with each other and $0 \leq x_1 \leq 100$ in the fitting data.
- Any forecast based on $x_1 > 100$ and $x_1 < 0$ will be **unreliable**.
- It is always dangerous when future values of the predictors lie much outside of the historical range, but is **especially problematic when multi-collinearity** is present.

Summary

If the future values of predictor variables are within their historical ranges, there is nothing to worry about. Multi-collinearity is not a **problem** except when there is a **perfect correlation**.

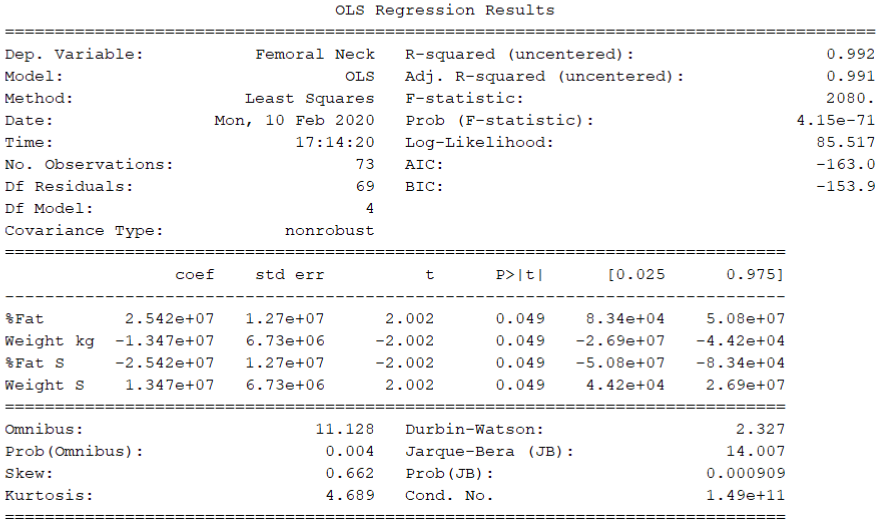
Simple and Multiple Regression Python

- There are two main ways to perform linear regression in Python- with statsmodels and scikit-learn.
- **statsmodels** is a python module that provides classes and functions for the estimation of many different models, as well as for conducting statistical test and statistical data exploration.



```python
import statsmodels.api as sm
import numpy as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

df = pd.read_csv('MulticollinearityExample.csv')
X = df[['%Fat', 'Weight kg','Activity','%Fat S','Weight S','Activity S']]
Y = df['Femoral Neck']

X_train, X_test, y_train, y_test = train_test_split(X, Y, shuffle= False, test_size=0.2)

model = sm.OLS(y_train,X_train).fit()
predictions = model.predict(X_test)

MSE = np.square(np.subtract(y_test,predictions)).mean()
print("Mean Square Error is ", MSE)
print(model.summary())
```

OLS Regression Results



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:           Femoral Neck   R-squared (uncentered):                   0.992
Model:                            OLS   Adj. R-squared (uncentered):              0.991
Method:                 Least Squares   F-statistic:                              2080.
Date:                Mon, 10 Feb 2020   Prob (F-statistic):                    4.15e-71
Time:                        17:14:20   Log-Likelihood:                          85.517
No. Observations:                  73   AIC:                                     -163.0
Df Residuals:                      69   BIC:                                     -153.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
%Fat        2.542e+07   1.27e+07      2.002      0.049    8.34e+04    5.08e+07
Weight kg  -1.347e+07   6.73e+06     -2.002      0.049   -2.69e+07   -4.42e+04
%Fat S     -2.542e+07   1.27e+07     -2.002      0.049   -5.08e+07   -8.34e+04
Weight S    1.347e+07   6.73e+06      2.002      0.049    4.42e+04    2.69e+07
==============================================================================
Omnibus:                       11.128   Durbin-Watson:                   2.327
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               14.007
Skew:                           0.662   Prob(JB):                     0.000909
Kurtosis:                       4.689   Cond. No.                     1.49e+11
==============================================================================
```

Regression Analysis

Hypotheses tests

- t-tests
- F-test

t-tests

- t-tests : are used to conduct hypothesis tests on the *regression coefficients* obtained in multiple regression where the H_0 & H_a defines as:

$$\begin{aligned} H_0 :& \beta_i = 0 \\ H_a :& \beta_i \neq 0 \end{aligned}$$

Regression Analysis F-test

F-test

- F-test: is used to compares the fits of different linear models. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously.
- F-test compares a model with no predictors (intercept-only) to the model that you specify.
  - **Null hypothesis**: The fit of the intercept-only model and your model are equal.
  - **Alternative hypothesis**: The fit of the intercept-only model is significantly reduced compared to your model.
- If the p-value for the F-test of overall significance test is less than your significant level, you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model.

Regression Analysis F-test

- Typically, if you don't have any significant p-values for the individual coefficients in your model, the overall F-test won't be significant.
- In the intercept-only model, all of the fitted values equal the mean of the response variable.
- If the p-value for the F-test of overall significance test is less than your significant level, you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model.

Regression Analysis F-test

- While R-squared provides an estimate of the strength of the relationship between your model and the response variable, it does not provide a formal hypothesis test for this relationship.
- The overall F-test determines whether this relationship is statistically significant.
- If the p-value of the overall F-test is less than your significant level, you can conclude that R-squared value is significantly different than zero.