Linear Regression and Assessing Model Accuracy

Sections 2.2 & 3.1

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Beginning Example

The goal is to predict the *stock_index_price* (the dependent variable) of a fictitious economy based on two independent/input variables:

- *Interest_Rate*
- *Unemployment_Rate*

The data is in the *stock_price.csv* data set in BlackBoard. This is from `https://datatofish.com/multiple-linear-regression-in-r/`

Questions We Want To Answer

1. Is there a relationship between *stock index price* and *interest rate*?
2. How strong is the relationship between *stock index price* and *interest rate*?
3. Is the relationship linear?
4. How accurately can we predict the *stock index price*?
5. Do both *interest rate* and *unemployment rate* contribute to the *stock index price*?

General Approach

- Let Y be the response (dependent variable).
- Let $X = (X_1, X_2, \ldots, X_p)$ be p different predictors (independent) variables.
- We assume there is some sort of relationship between X and Y, which can be written in the general form

$$Y = f(X) + \epsilon$$

- Statistical leaning refers to a set of approaches for estimating f.

How Do We Estimate f?

- The goal is to apply a statistical learning method to the training data in order to estimate the unknown function of f.

- Using a model-based approach, called **parametric**, with assumptions about the model.
  1. We make an assumption about the function form or shape of f.
  2. We need a procedure that uses the training data to fit or train the model.

- No assumptions about the model is called a **non-parametric** method.
  - Non-parametric method seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
  - **Advantage**: they have the potential to accurately fit a wider range of possible shapes for f.
  - **Disadvantage**: a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Parametric Method

Parametric methods involve a two-step model-based approach.

1. We make an assumption about the functional form, or shape, of f. Then determine a model.
2. After a model has been selected, we need a procedure that uses the *training* data to fit or train the model.
    - The training data are observations used to train or teach our method how to estimate f.
    - Let x_{ij} represent the value of the jth predictor for observation i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.
    - Let y_i be the response variable for the ith observation.
    - Then the training data consist of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Training, test, and validation sets

- The model is initially fit on a **training data set**, that is a set of observations used to fit the parameters.
- Successively, the fitted model is used to predict the responses for the observations in a second data set called the **validation data set**.
- Finally, the **test data set** is a data set used to provide an unbiased evaluation of a final model fit on the training data set

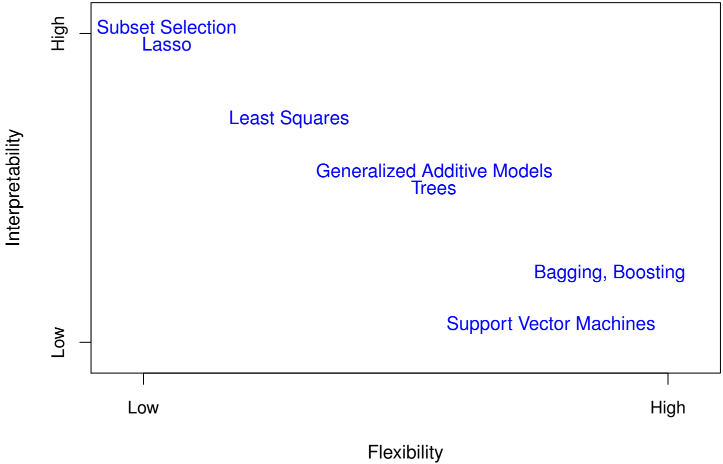
Confusingly the terms test data set and validation data set are sometimes used with swapped meaning. As a result it has become commonplace to refer to the set used in iterative training as the test/validation set and the set that is used for hyper parameter tuning as the **holdout set**.

Flexibility vs. Interpretation

Why choose to use a more restrictive method instead of a very flexible approach?

- If we are mainly interested in inference, then restrictive models are much more interpretable. For example, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and $X_1, X_2, \ldots, X_p$.
- Very flexible approaches, such as the splines and the boosting methods can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

Flexibility vs. Interpretation



Reasons for Estimating f

- Prediction: we want to predict Y, using $\hat{Y} = \hat{f}(X)$.
  - $\hat{f}$ is often treated as a **black box**.
  - The black box means that we are not typically concerned about the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.
- Inference: we want to know how Y is affected as X changes.
  - In this situation we wish to estimate f.
  - Thus $\hat{f}$ cannot be considered as a black box because we do want to know the exact form of $\hat{f}$.

Lab Question 1

Recall the Stock Price questions. What type of statistical learning problem will we use?

a) Regression, inference
b) Regression, prediction
c) Classification
d) Clustering

For now we are going to look at predicting the *Stock_Index_Price* based on its relationship with *Interest_Rate*.

Response => Stock_Index_Price

$Y = f(X) + \boxed{\varepsilon}$ $\qquad f(X) \stackrel{?}{=}$

$f(x) = \beta_0 + \beta_1 X$

Simple Linear Regression Model

- The data are n observations on an explanatory variable x and a response variable y,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- The statistical model for simple linear regression states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- $\mu_y = \beta_0 + \beta_1 x_i$ is the mean response for y when $x = x_i$ a specific value of x.
- ϵ_i are the error terms for predicting y_i for each value of x_i.
- Notice in our general form that $f(X) = \beta_0 + \beta_1 X$.

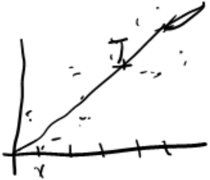
Parameters of the Regression Model

$Y = f(X) + \varepsilon$

irreducible error

- The intercept: β_0.
- The slope: β_1.
- The variability: σ^2 of the response y about this line. More precisely, σ is the standard deviation of the deviations of the errors, ϵ_i in the regression model.
- Each ϵ_i are independent and Normally distributed with mean 0 and standard deviation σ.

$E(\varepsilon) = 0$ $\quad$ $Var(\varepsilon) = \sigma^2$ assume this to be constant for all values of X.



$\hat{f}(x)$

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

$\hat{y}_i$ = the predicted value for each x_i

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Conditions for Regression Inference

- The sample is an SRS from the population.
- There is a linear relationship in the population.
- The standard deviation of the responses about the population line is the same for all values of the explanatory variable.
- The response varies Normally about the population regression line.

LINE

- Linear
- Independent
- Normal
- Equal variance

Principle of Least Squares

The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{hieght of point} \; - \; \text{height of line} \; = \; y_i - (b_0 + b_1 x_i)$$

The sum of the square vertical deviations from the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $f(b_0, b_1)$.

Estimating the Regression Parameters

- In the simple linear regression setting, we use the slope b_1 and intercept b_0 of the least-squares regression line to estimate the slope β_1 and intercept β_0 of the population regression line.
- The standard deviation, σ, in the model is estimated by the regression standard error

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum \text{all residuals}^2}{n-2}}$$

  Recall that y_i is the observed value from the data set and $\hat{y}_i$ is the predicted value from the equation.
- In R s is the called the **Residual Standard Error** in the last paragraph of the summary.

$$\left.\begin{array}{l}\text{Residual Sum of Squares (RSS)}\\ \text{Sum of Squares Error (SSE)}\end{array}\right\} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The Least - Squares Estimates

- Recall $e_i =$ observed Y - predicted Y is the i^{th} residual. Think of it as an estimate of the unobservable true random error ϵ_i.

- The method of **least squares** selects estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the **residual sum of squares**:

$$SS_{(resid)} = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

- Where the estimate of the slope coefficient β_1 is:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\,\frac{s_y}{s_x}$$

- The estimate for the intercept β_0 is:

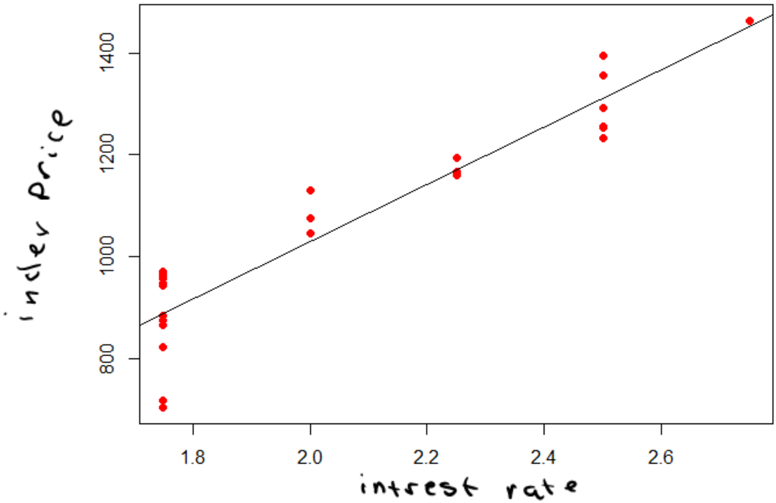
The point $(\bar{x}, \bar{y})$ is a point on the line

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Stock Prices Example

- Use the *stock_price.csv* data.
- We want to predict *stock index price* based on *interest rate*.
  1. Determine if it is a linear relationship. How can we tell?
  2. Get an estimate of the model.
  3. Is this a good fit for the data?

Do We Have A Linear Relationship?



The Estimate of the Model

$x_1 + x_2$

$y \sim x$

```
> stock.lm <- lm(Stock_Index_Price~Interest_Rate,data = stock_price)
> summary(stock.lm)

Call:
lm(formula = Stock_Index_Price ~ Interest_Rate, data = stock_price)

Residuals:
     Min       1Q   Median       3Q      Max
-183.892  -30.181    4.455   56.608  101.057

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -99.46      95.21  -1.045    0.308
Interest_Rate   564.20      45.32  12.450 1.95e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.96 on 22 degrees of freedom
Multiple R-squared:  0.8757,Adjusted R-squared:  0.8701
F-statistic:   155 on 1 and 22 DF,  p-value: 1.954e-11
```

$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

RSE

Equation: index_price = -99.46 + 564.20 × interest rate

Confidence Intervals for β_1

If we want to know a range of possible values for the slope we can use a confidence interval. The confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times SE(\hat{\beta}_1)$$

where

$$SE(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and $s^2 = \hat{Var}(\epsilon)$.

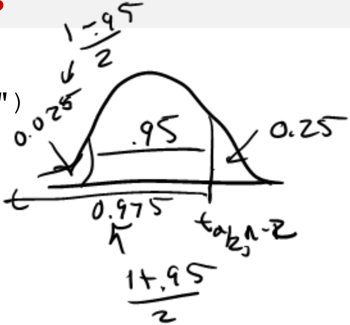
Given the following excerpt from the `R` output, determine a 95% confidence interval for the slope.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -99.46      95.21  -1.045    0.308
Interest_Rate   564.20      45.32  12.450 1.95e-11 ***
```

$\frac{564.20}{45.32}$

$564.20 \pm t_{\alpha/2,n-2}(45.32)$

```
> confint(stock.lm,"Interest_Rate")
                  2.5 %   97.5 %
Interest_Rate 470.2214 658.1864
```




```
> qt(.975,22)
[1] 2.073873
> qt(1.95/2,22)
[1] 2.073873
> 564.20-2.0739*45.32 #lower limit
[1] 470.2109
> 564.20+2.0739*45.32 #upper limit
[1] 658.1891
```

[470.21, 658.19]

$\hat{\beta}_1 = 564.20$

Interpretation:

For 1% increase in interest rate the stock index price will increase by \$564.20 on average.

95% CI: [470.21, 658.19]

Interpretation:

For 1% increase in interest rate the stock index price will increase between \$470.21 and \$658.19 with 95% confidence.

t Test for Significance of β_1

$y = \beta_0 + \beta_1 x + \varepsilon$

- Hypothesis

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

  Or we can think about it in this way

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

  versus

$$H_0 : \text{There is a relationship between } X \text{ and } Y$$

- Test statistic

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

$$\text{standard error} = SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

  With degrees of freedom $df = n - 2$.

- P-value: based on a t distribution with $n - 2$ degrees of freedom.
- Decision: Reject H_0 if p-value $\leq \alpha$.
- Conclusion: If H_0 is rejected we conclude that the explanatory variable x can be used to predict the response variable y.

Given the following excerpt from the R output, Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -99.46      95.21  -1.045    0.308
Interest_Rate   564.20      45.32  12.450 1.95e-11 ***
```

Test statistic: $t = 12.45$

p-value ≈ 0

Decision RH$_0$

There is evidence of a relationship between Stock index price and intrest rate.

Is this good at predicting the response?

R^2 is the percent (fraction) of variability in the response variable (Y) that is explained by the least-squares regression with the explanatory variable.

- This is a measure of how successful the regression equation was in predicting the response variable.
- The closer R^2 is to one (100%) the better our equation is at predicting the response variable.
- We will look later at how this is calculated.
- In the R output it is the **Multiple R-squared** value.

Calculating R^2

1. The **error sum of squares**, denoted by *SSE* is

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Calculating R^2

1. The **error sum of squares**, denoted by *SSE* is
$$SSE = \sum (y_i - \hat{y}_i)^2$$

2. The **regression sum of squares**, denoted *SSR* is the amount of total variation that *is* explained by the model
$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Calculating R^2

1. The **error sum of squares**, denoted by SSE is

$$SSE = \sum (y_i - \hat{y}_i)^2$$

2. The **regression sum of squares**, denoted SSR is the amount of total variation that *is* explained by the model

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

3. A quantitative measure of the total amount of variation in observed values is given by the **total sum of squares**, denoted by SST.

$$SST = \sum (y_i - \bar{y})^2$$

Note: SST = SSR + SSE

Calculating R^2

1. The **error sum of squares**, denoted by SSE is
$$SSE = \sum (y_i - \hat{y}_i)^2$$

2. The **regression sum of squares**, denoted SSR is the amount of total variation that *is* explained by the model
$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

3. A quantitative measure of the total amount of variation in observed values is given by the **total sum of squares**, denoted by SST.
$$SST = \sum (y_i - \bar{y})^2$$

   Note: SST = SSR + SSE

4. The **coefficient of determination**, r^2 is given by
$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

Information from the Summary in R

```
Residual standard error: 75.96 on 22 degrees of freedom
Multiple R-squared:  0.8757,Adjusted R-squared:  0.8701
F-statistic:   155 on 1 and 22 DF,  p-value: 1.954e-11
```

RSE and R^2

- The RSE is considered a measure of the *lack of fit* of the model to the data. Recall this is the estimate of the standard deviation of the residuals $y_i - \hat{y}_i$.
  - If $\hat{y}_i$ s very far from y_i, then the RSE may be quite large.
  - This measurement depends on the units of the original values.

RSE and R^2

- The RSE is considered a measure of the *lack of fit* of the model to the data. Recall this is the estimate of the standard deviation of the residuals $y_i - \hat{y}_i$.
  - If $\hat{y}_i$ s very far from y_i, then the RSE may be quite large.
  - This measurement depends on the units of the original values.

- The R^2 takes the form of a proportion of variance in y that is explained.
  - R^2 thus always takes on a value between 0 and 1.
  - If R^2 is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
  - *Note*: For a simple linear regression $R^2 = Cor(X, Y)^2$.

Assumptions about the Model

1. The error term ε is a random variable with a mean or expected value of zero, that is $E(\varepsilon) = 0$, an estimate for ε is the residuals for each value of the X-variable.

$$\text{residual} = \text{observed y} - \text{predicted y}$$

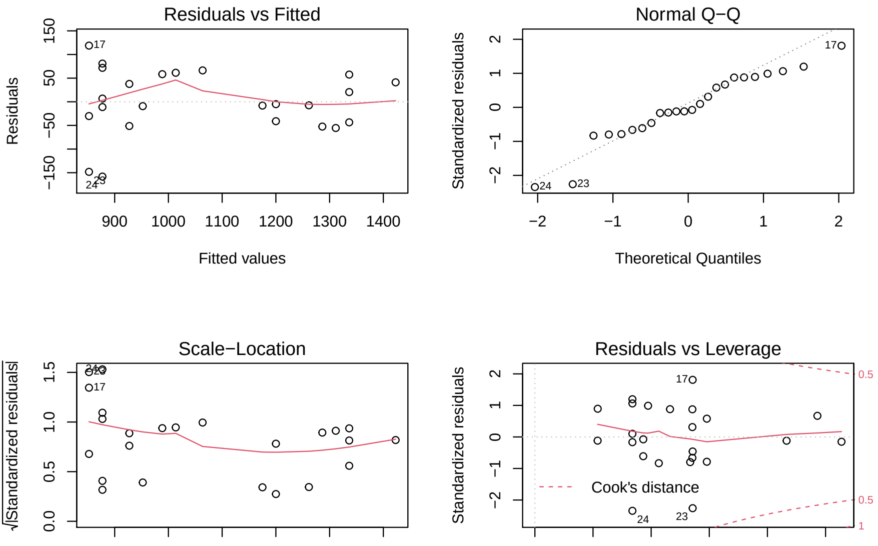
2. The variance of ε, denoted by σ^2, is the same for all values of x. The estimate for σ^2 is $s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$.

3. The values of ε are independent.

4. The error term ε is a normally distributed random variable.

5. The **residual plots** help us assess the fit of a regression line and determine if the assumptions are met.

Plots to Check Assumptions



Introduction Stuff

We will be using two packages for this lab. We will need to load the `MASS` and `ISLR` package by the following code:

```
#Install the packages (only have to do  once)
install.packages("MASS")
install.packages("ISLR")
#Load the packages (Have to do every time you open R)
library(MASS)
library(ISLR)
```

We will use a data set in the `MASS` library called Boston. To know information about this data set you can type in `?Boston`.

Lab Questions

We will start by using the `lm()` function to fit a simple linear regression model, with medv as the response and lstat as the predictor. That is, we will seek to predict medv (median house value per $1000) using lstat (percent of households with low socioeconomic status).

2. Type in R, `lm.fit = lm(medv~lstat)`, what happens?
   a) Nothing
   b) I get an error
   c) The model appears

3. Type in `lm.fit = lm(medv~lstat,data = Boston)`, what happens?
   a) Nothing
   b) I get an error
   c) The model appears

Lab Questions

Type in `R`, `summary(lm.fit)`.

4. What is the estimate of the model, $f(X)$?
   a) $-15.168 - 3.990x$
   b) $34.55384 - 0.95005x$
   c) $34.55384 + 0.56263x$
   d) $-0.95005 + 0.03873x$

5. What percent of the variation in medv can be explained by lstat?
   a) 6.2%
   b) 54.44%
   c) 60.1%
   d) 2.2

Lab Questions

Type in R, `confint(lm.fit)`

6. What is the confidence level?
   a) 99%
   b) 95%
   c) 2.5%
   d) 97.5%

7. Interpret the confidence interval for the lstat line.
   a) For each unit increase in percent of households with low socioeconomic status, the median house value will decrease on average between $0.87 and $1.03 with 95% confidence.
   b) For each unit increase in percent of households with low socioeconomic status, the median house value will decrease on average between $873.95 and $1026.15 with 95% confidence.
   c) For each unit increase in percent of households with low socioeconomic status, the median house value will increase on average between $0.87 and $1.03 with 95% confidence.
   d) For each unit increase in percent of households with low socioeconomic status, the median house value will increase on average between $873.95 and $1026.15 with 95% confidence.
   e) There is a 95% chance that for each unit increase in percent of households with low socioeconomic status, the median house value will decrease on average between $873.95 and $1026.15.

Lab Questions

The `predict()` function can be used to produce confidence intervals and prediction intervals for the prediction of medv for a given value of lstat.

8. Type in `R`, `predict(lm.fit,data.frame(lstat = c(5,10,15)),interval = "confidence")`. If the percent of households with low socioeconomic status is 10% what is the predicted median house value?
   a) \$29.80
   b) \$25.05
   c) \$20.30
   d) \$29,803.59
   e) \$25,053.35

Plots and Plotting Symbols

In R type

```
plot(Boston$lstat ,Boston$medv,
      xlab="lstat",ylab = "medv")
abline(lm.fit,lwd=3,col = "red")
plot(Boston$lstat,Boston$medv,pch = 20)
```

The pch option creates different plotting symbols. You could also do:

```
plot(Boston$lstat,Boston$medv,pch = "+")
```

To look at some of the symbols do:

```
plot(1:20,1:20,pch=1:20)
```

Lab Questions

In `R` type

```
par(mfrow=c(2,2))
plot(lm.fit)
```

9. Looking at the plots are there assumptions that are not met?
   a) It appears that all of the assumptions are met.
   b) The data appears not linear.
   c) The residuals do not have a Normal distribution.
   d) The residuals do not have a constant variance.
   e) Answers b, c and d are all true.

10. Based on all of this information should we use this simple linear model to predict medv based on lstat?
    a) Yes
    b) No