Linear Regression

Links: [MATH 4322](), [Inference for Regression Parameters]()

(data science & machine learning lecture 3)
corresponds to chapter 3.1 in the textbook (pages 59 - 71, pdf pages 70 - 82)

Beginning Example & General Approach

> ### Stock Price 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`
> - `Unempoloyment_Rate`
> - (The data is in the stock_price.csv data set in Canvas)
>
> #### 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?
> 6. What is the statistical learning problem?

> ## General Approach
>
> (see also: Definitions & Intro > General Approach For Supervised Learning)
>
> - *Stock index price* is the response or output. We refer to the response usually as Y.
> - *Interest rate* is an input or predictor, we will name it X_1
> - Also, *Unemployment rate* is an input, we will name is X_2
> - Let $X = (X_1, X_2, \ldots, X_p)$ be *p* different predictors (independent) variables.
> - For this example we will have an input vector as
>
> $$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$
>
> - We assume there is some sort of relationship between *X* and *Y*, which can be written in the general form, thus our model is
>
> $$Y = f(X) + \epsilon$$
>
> - Where ϵ captures the measurement errors and other discrepancies
> - Statistical leaning refers to a set of approaches for estimating *f*.

(to answer those questions we want to figure out/estimate this function *f* and also concern ourselves with the ϵ (error term)).

Estimators

A statistic $\hat{\theta}$ used to estimate and unknown population parameter θ is called an estimator. We desire a uniformly minimum variance unbiased estimator.

- Properties of an estimator $\hat{\theta}$
  - Accuracy - measured by bias

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Precision/variability - measured by its variance, $Var(\hat{\theta})$. The estimated standard deviation of an estimator θ is referred to as its standard error (SE).
- The mean squared error (MSE) combines both measures.

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

- In MATH 3339 (Statistics) we studied estimators for population mean (μ) and proportions (p). In Data Science and Machine Learning we will want estimators for f(X).

Example

Suppose we take a random sample (samples have independence) of 4 from a Normal distribution with $\mu = 2$ and $\sigma = 2$ (population mean 2, and population standard deviation 2 'so we may be off by 2').

- Let $\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i$ be an estimator of μ. What is the expected value, bias, variance, and MSE of $\bar{x}$?

solution: >

Expected Value

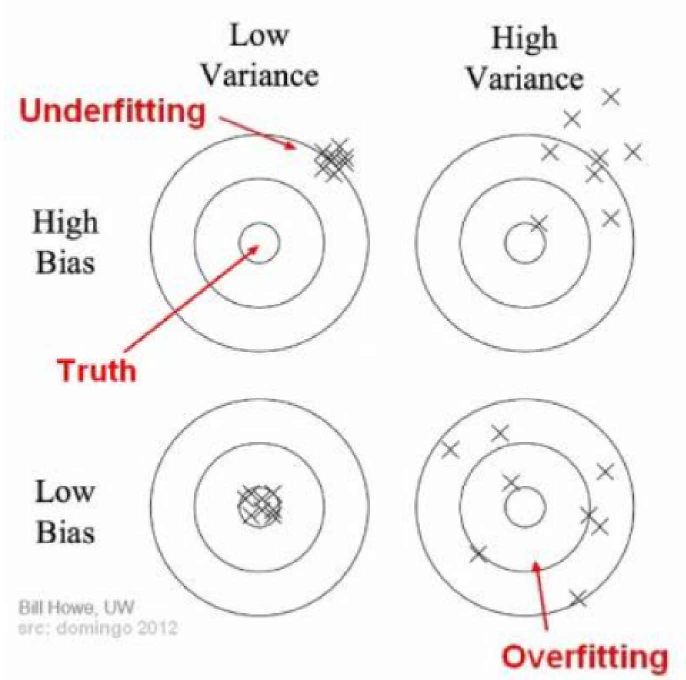
$$E(\bar{x}) = E\left(\frac{1}{4}\sum_{i=1}^{4} x_i\right) = \frac{1}{4}E\left(\sum_{i=1}^{4} x_i\right)$$

$$= \frac{1}{4}\left(E(x_1) + E(x_2) + E(x_3) + E(x_4)\right)$$

$$= \frac{1}{4}(10 + 10 + 10 + 10) = \frac{40}{4} = 10$$

see lecture slides (6 & 7) for answer to all the questions

understanding variance:



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

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

(see also: [Stats Exam 1 Notes > least squares regression line (LSRL)])

- $\mu_y = \beta_0 + \beta_1 x_i$ 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 Simple Regression Model

- The intercept: β_0
- The slope: β_1
- The goal is to obtain coefficient estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ such that for each observed y_i, $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$, for $i = 1, 2, \ldots, n$
- The most common approach is by minimizing the least squares criterion.

Least Squares

Principle of Least Squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for *Y* based on the *i*th value of *X*.
- Then $e_i = y_i - \hat{y}_i$ (aka the residuals) be the *i*th residual, the difference between the *i*th observed response value and the *i*th predicted value by our linear equation.
- The **residual sum of squares** (RSS) is defined by

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

- 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 the *RSS*.

we want to minimize the error sum of squares:

$$\text{minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

(we're trying to find values for the intercept ($\hat{\beta}_0$) and values for the slope ($\hat{\beta}_1$) that minimizes that above expression)

The Least-Squares Estimates

- The method of **least squares** selects estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the residual sum of squares (RSS).
- Where the estimate of the slope coefficient β_1 is:

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

  The estimator for the slope is equal to the correlation of x and y multiplied by the standard deviation of *y* over the standard deviation of *x*. (See the lecture for the super long ~~and boring~~ proof)
- The estimate for the intercept β_0 is:

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

- Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \sum_{i=1}^{n} x_i$

Stock Prices Example

```
see pdf pages 13 to 17 of the lecture 3 slides
```

The R commands used for the example were:

```
stock.lm <- lm(Stock_Index_Price~Interest_Rate,data = stock_price)
# creating the model

summary(stock.lm) # displays the info
```

This has $n - 2$ degrees of freedom since we are doing two predictions

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

```
confint(name.lm, "predictor_name")
```

(see pdf page 20 of lecture 3 slides for example)

t Test for Significance of β_1

(See also Inference for Regression Parameters > t Test for Significance of Slope)

- Hypothesis:

$$H_0 : \beta_1 = 0 \ versus \ H_a : \beta_1 \neq 0$$

  Or we can think about it in this way

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

  versus

$$H_a : There\ is\ a\ relationship\ between\ X\ and\ Y$$

- Test Statistic:

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

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

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*.

(see also: Hypothesis Testing > Components of a Significance Test)

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: $t = 12.450 = \frac{564.2}{45.32}$

P-value $= P(t \leq -12.45 \text{ or } t \geq 12.45) \approx 0$ RH_0

There is very strong evidence of a relationship between interest rate and stock price.

(this test is two tailed, it concluded with the null hypothesis being rejected).

Is this good at predicting the response?

- Once we have said that this model can help predict the **output**: we want to quantify how well the model fits the data.
- Two quantities that we use is the residual standard error (RSE) and the coefficient of determination (R^2).
- These quantities are in the summary output of the `lm()` function in R.

Residual Standard Error

- The RSE is an estimate of the standard deviation if the error term, ϵ.
- We can think about it as the average amount that the response will deviate from the true regression line.

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_1^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y} - i)^2}$$

(*just pretend those square root symbols rendered correctly* 🙃)

- The RSE is really the standard deviation of our residuals (remember that standard deviation is just spread from center, so RSE is measuring how spread apart the residuals are from 0)
- The lower the RSE the better our model fits the data (The smaller the better, because that means the residuals are not spread out that much).

R^2 Statistic

R^2 is the percent (fraction) of variable in the response variable (*Y*) that is explained by the least-squares regression with the explanatory variable (see also: Stats Exam 1 Notes > coefficient of determination)

- This is a measure of how successful the regression equation was in predicting the response variable (how much of the variation can we account for).
- The closer R^2 is to one (100%) the better our equation is at predicting the response variable.
- In the R output it is the **Multiple R-squared** value.

Calculating R^2

1. The **residual sum of squares**, denoted by *RSS* is

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

(*RSS* is also called *SSE*: Sum Squares Error)

2. The **regression sum of squares**, denoted by *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*

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

Note: $TSS = SSR + RSS$

4. The **coefficient of determination**, R^2 is given by

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

The R^2 value can be found in the `summary` output in `R`.

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 $\hat{y}_i$ is 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 this always takes on a value between 0 and 1.
  - If R^2 is close to 1 indicates that a large proportion of 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

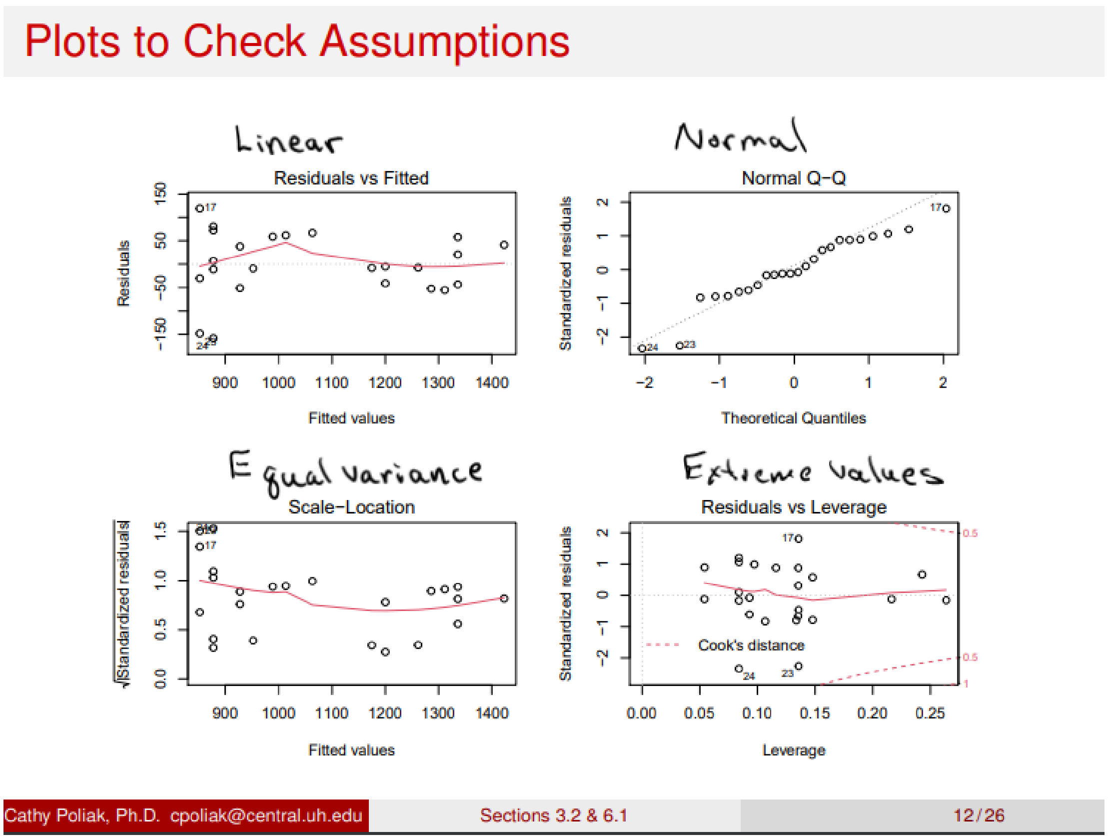
The linear regression model has assumptions that we need to prove is true. We use the acronym LINE to remember these assumptions.

- **L**inear relationship: can we determine a linear relationship between the response an other variables?
- **I**ndependent observations: are the observations a result of a simple random sample?

- **N**ormal distribution: for any fixed value of X, Y is [normally distributed](#).
- **E**qual variance: the variance of the residual is the same for any value of X.

Be careful of extreme values.

~~delete this: [550](#)~~



(values beyond "Cook's distance" can be considered extreme values)

In R to get these plots:

```
par(mfrow = c(2,2)) # displays four plots in one window
plot(name.lm) # diagnostic plots

par(mfrow = c(1,1)) # puts back to one plot in window
```