Linear Regression

Section 3.1

Dr. Cathy Poliak, cpoliak@uh.edu

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*?
6. What is the statistical learning problem?

General Approach

- *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 it 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

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

- 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.

Estimators

A statistic $\hat{\theta}$ used to estimate an unknown population parameter θ is called an **estimator**.

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

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

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

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

- In MATH 3339 we studied estimators for μ and p. In this class we will we will want estimators for $f(X)$.

Example, Estimate of μ

Suppose we take a random sample of 4 from a Normal distribution with $\mu = 10$ and $\sigma = 2$.

- 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 $\bar{x}$.

$E(\bar{x}) = \mu = 10$, $Bias(\bar{x}) = E(\bar{x}) - \mu = 10 - 10 = 0$

$Var(\bar{x}) = \frac{\sigma^2}{n} = \frac{4}{4} = 1$

$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

$MSE(\bar{x}) = var(\bar{x}) + Bias(\bar{x})^2 = 1$

$MSE(\bar{x}) = \sigma^2/n$

- Let 8 be an estimator of μ. What is the expected value, bias, variance, and MSE of 8?

$E(8) = 8$ $\quad bias(8) = 8 - 10 = -2$

$var(8) = 0$

$MSE(8) = 0 + (-2)^2 = 4$ $\quad MSE(8) = (8 - \mu)^2$

$\hat{p}_1 = \frac{X}{n} = \frac{\text{\# of successes in "n" trials}}{\text{\# of trials}}$

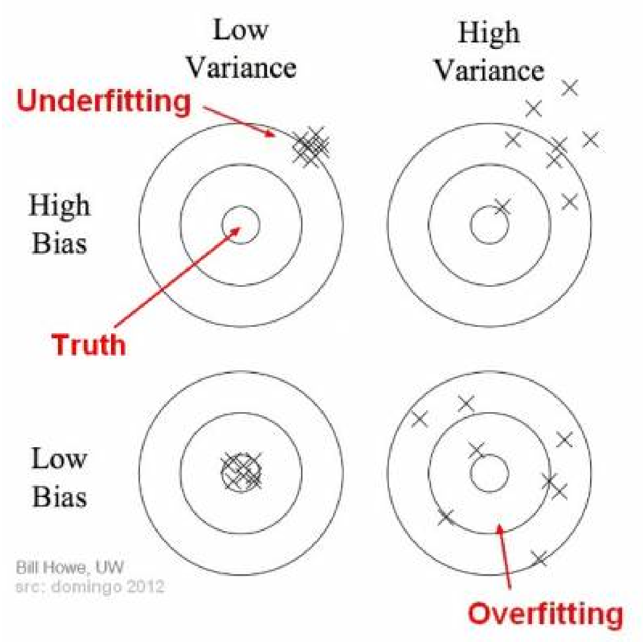
$X \sim Bin(n, p)$ $E(X) = np$ $var(X) = np(1-p)$

$E(\hat{p}_1) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X)$

$Var(\hat{p}_1) = var\left(\frac{X}{n}\right) = \frac{1}{n^2} var(X)$

p = seq(0, 1, by = 0.01)

plot(p, mse1)


Low
Variance
High
Variance
Underfitting
High
Bias
Truth
Low
Bias
Bill Howe, UW
src: domingo 2012
Overfitting

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 Simple Regression Model

- The intercept: β_0.
- The slope: β_1.
- The goal is to obtain coefficient estimates $\hat{\beta}_0$ 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 the minimizing the least squares criterion.

Principle of Least Squares

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

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

- 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.

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

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:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{(n-1)\,\mathrm{cov}(x,Y)}{(n-1)\,\mathrm{var}(x)}$$

$$= \frac{\mathrm{cor}(x,Y)\, s_x s_y}{s_x^2}$$

$$= \mathrm{cor}(x,Y)\,\frac{s_y}{s_x}$$

- 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 $\bar{x} = \sum_{i=1}^{n} x_i$.

minimize

$$Q = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]$$

$$0 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right]$$

$$0 = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \left[\frac{1}{n} \sum_{i=1}^{n} x_i \right]$$

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

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)$$

$$0 = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i y_i - \left[\frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \right] \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

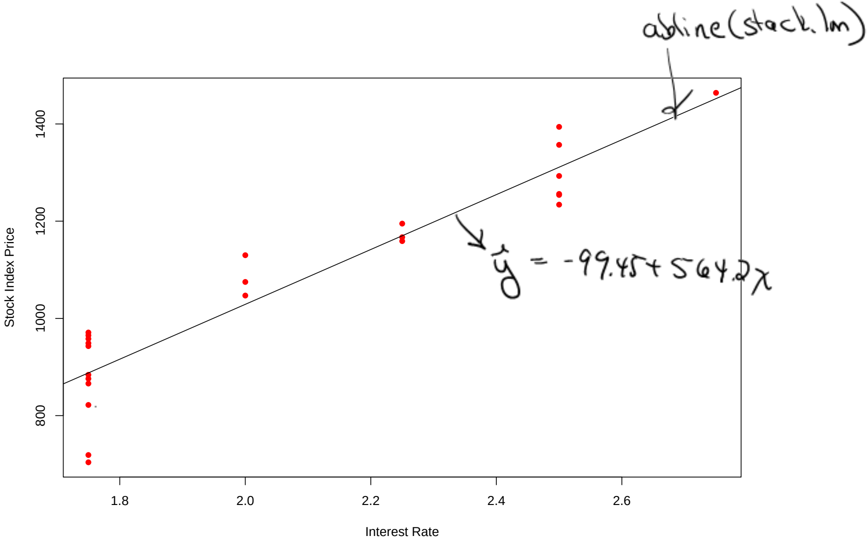
$$\hat{\beta}_1 \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{y} \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

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

tilde

```r
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
```

$-99.46 = \hat{\beta}_0$

$564.20 = \hat{\beta}_1$

$$\hat{y}_i = -99.46 + 564.2x_i$$

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

$$= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}$$

Confidence Intervals for β_1

$n=24$

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.

```
               Estimate Std. Error  t value     Pr(>|t|)
(Intercept)   -99.46432   95.21031 -1.04468 3.075145e-01
Interest_Rate 564.20389   45.31738 12.45006 1.953975e-11
```

$qt(1.95/2, 22) = 2.074$

$CI: \hat{\beta}_1 \pm t_{\alpha/2,n-2} SE(\hat{\beta}_1) = 564.204 \pm 2.074 (45.3174)$
$\approx [470.21, 658.19]$

R Function for Confidence Intervals

```r
confint(stock.lm,"Interest_Rate")
```

```
                 2.5 %   97.5 %
Interest_Rate 470.2214 658.1864
```

t Test for Significance of β_1

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

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

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

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- 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

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

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

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.

Example of Hypothesis Test

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-99.46432	95.21031	-1.04468	3.075145e-01
Interest_Rate	564.20389	45.31738	12.45006	1.953975e-11

Test statistic : $T = \dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \dfrac{564.20389 - 0}{45.31738} = 12.45$

p-value $= P(T_{22} \leq -12.45 \text{ or } T \geq 12.45)$

$= 2 * pt(-12.45, 22)$

$= 1.95397e^{-11} \approx 0$

Decision: Reject H_0, since p-value ≤ 0.05

Conclusion: There is strong evidence between the relationship of interest rate & stock price

Is this good at predicting the response?

- Once we have said that this model can help predict the **output** we want to quantify at 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.

Residual Standard Error

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

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

- The lower the RSE the better our model fits the data.

R^2 Statistic

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.
- In the R output it is the **Multiple R-squared** value.

Calculating R^2

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

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

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.

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

Note: $\text{TSS} = \text{SSR} + \text{RSS}$

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

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

Information from the Summary in R

```
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
```

About 87.57% of the variation in stock index price can be accounted for by this model.

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

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.

Plots to Check Assumptions