Inference in Regression Analysis

Predictive Modeling & Statistical Learning

Gaston Sanchez

So Far ...

So far

We've talked about Linear Regression (simple and multiple via OLS) from a purely mathematical standpoint:

- Optimization problem: minimization by Least Squares
- Geometric interpretation: orthogonal projection
- Algebraic (vector-matrix) notation

But we haven't talked about statistical assumptions (e.g. distributions)

Introduction

- Suppose we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$
- We assume there is some relationship between the response and the predictors
- $Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$
- f represents the systematic information that the predictors provide about Y
- ε represents an *error* term that is a catch-all for what we miss with the model

Introduction

Simple assumptions so far about $f()$

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Introduction

Simple assumptions so far about $f()$

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- We assume that $X_1, \ldots, X_p$ help us predict/explain Y
- i.e. knowing $X_1, \ldots, X_p$ change our uncertainty about Y
- We assume a linear dependence of Y on the predictors
- We assume the error is additive
- The linearity is in the parameters (i.e. coefficients)

What's next

- We haven't assumed much about the error terms ε
- Regression Modeling entails more than just estimating β_j's and computing RSS and R^2
- There's usually interest in answering questions that involve hypothesis tests and confidence intervals

Let's keep discussing how statisticians think about regression analysis by wrapping things within an inferential layer.

Fundamental Idea

How statisticians think about regression

Statisticians commonly define regression so that the goal is to understand as far as possible with the available data how the conditional distribution of some response Y varies with the possible values of one or more predictors X_j

Richard Berk (2008)

What does this mean?

Formal Framework

Let's formalize things

- We will place ourselves in territory of random variables and probability spaces.
- I will consider simple regression: one predictor X and one response Y (although things can be generalized to multiple predictors).
- This framework involves some concepts from Statistical Decision Theory.

A bit of Decision Theory

- Let $X \in \mathbb{R}$ denote a real valued random input variable.
  Actually X does not have to be necessarily real; it can be qualitative
- Let $Y \in \mathbb{R}$ denote a real valued random output variable.
- Let $Pr(X, Y)$ be the joint distribution.
- We seek a function $f(X)$ for predicting Y given the values of the input X.

Loss Function

This theory requires a **loss function** for penalizing errors in prediction:

$$L(Y, f(X))$$

Loss Function

This theory requires a **loss function** for penalizing errors in prediction:

$$L(Y, f(X))$$

By far the most common and convenient loss function is the **squared error loss** a.k.a. quadratic loss:

$$L(Y, f(X)) = (Y - f(X))^2$$

Loss Function

The criterion for choosing $f()$ is the so-called **Expected Prediction Error** (EPE):

$$EPE(f) = E(Y - f(X))^2$$

We look for $f()$ that minimizes EPE

Loss Function

By conditioning on X, we can write EPE as

$$\begin{aligned} EPE(f) &= E(Y - f(X))^2 \\ &= E_X \left\{ E_{Y|X} \left([Y - f(X)]^2 | X \right) \right\} \end{aligned}$$

What function $f()$ minimizes EPE?

Conditional Mean

It turns out that the solution that minimizes $EPE(f)$ is

$$f(x) = E(Y|X = x)$$

which is the conditional expectation, also known as the **regression** function.

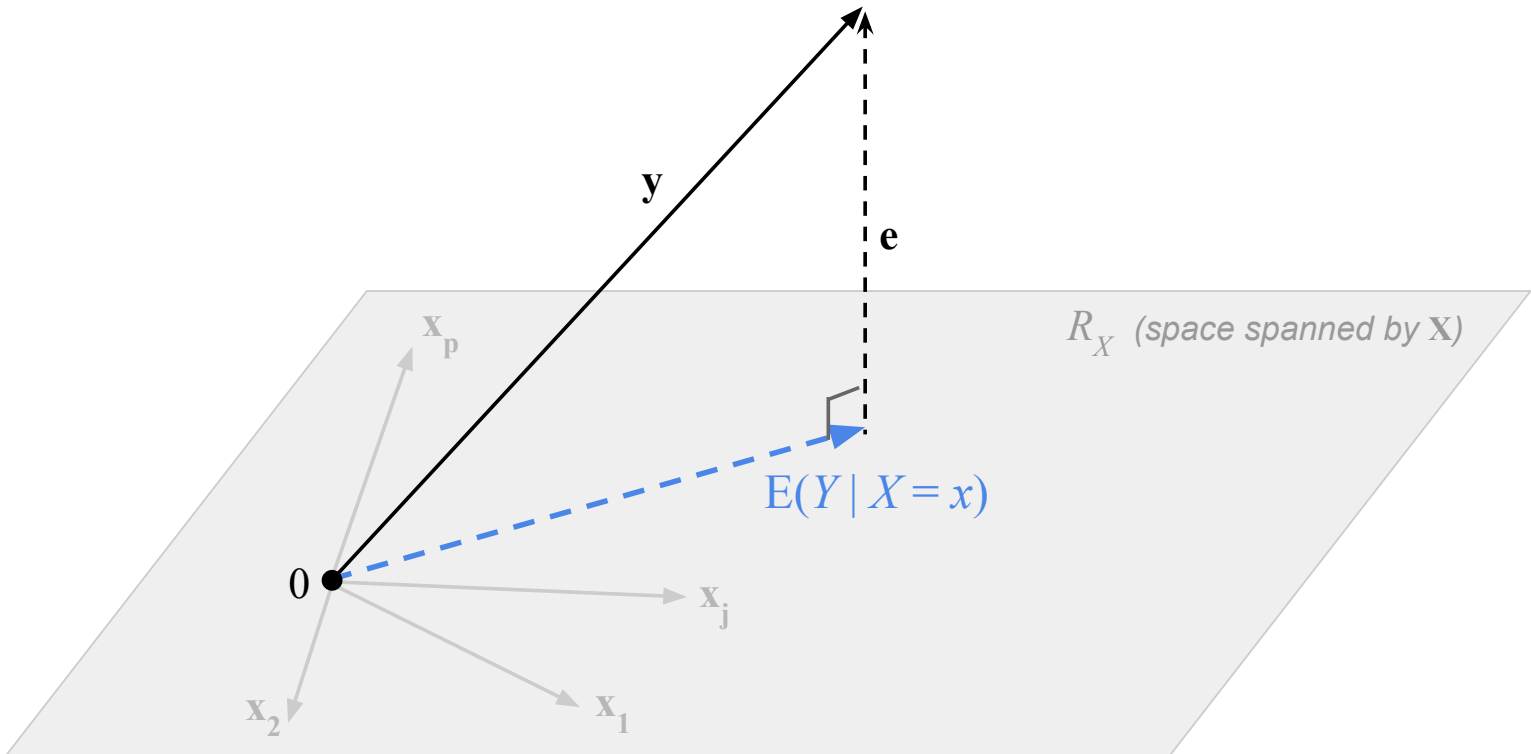
Regression Function

$$f(x) = E(Y|X = x)$$

The conditional expectation is the best prediction of Y at any point $X = x$, when best is measured by average squared error.

It can be shown that, geometrically, $E(Y|X)$ is the orthogonal projection of Y onto the space V_X of variables of type $f(X)$.

Regression Function: conditional mean



How does Least Squares regression fit into this framework?

Conditional Mean and LS

We assume that the regression function $f(x)$ is approximately linear in its arguments:

$$f(x) \approx x^{\mathsf{T}}\beta$$

Plugging this linear model for $f(x)$ into EPE and differentiating, we can solve for β theoretically:

$$\beta = [E(XX^{\mathsf{T}})]^{-1}E(XY)$$

The least squares solution amounts to replacing the previous expectation by averages over the observed data.

Toward the Classic Linear Model

About the error term

Consider the linear model:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

We assume that ε is a term that results from a large number of unobserved causes. Therefore we think of it as a random "noise".

Basic Assumptions about the Errors

- $E(\varepsilon_i) = 0$
- Errors should have the same variance (*homoscedastic*): $var(\varepsilon_i) = \sigma^2$
- Any pair of errors $\varepsilon_i, \varepsilon_j$ $(i \neq j)$ should be uncorrelated: $cor(\varepsilon_i, \varepsilon_j) = 0$
- Errors must be uncorrelated with the predictors: $cor(x_i, \varepsilon_i) = 0$
- We'll make an additional assumption about the explicit distribution of the errors in a few slides.

Basic Assumptions about the Errors

In vector-matrix notation we have:

- $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

Equivalently:

- $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$
- $Var(\mathbf{y}) = \sigma^2\mathbf{I}$

Basic Assumptions about the errors

These assumptions are simple, however:

When the spread around large values of y_i is greater than the spread around small values, assuming homoscedasticity σ^2 is unrealistic.

Assuming that errors are uncorrelated with predictors when there is a time component (autocorrelation) is often unrealistic.

Gauss-Markov Theorem

Gauss-Markov Theorem

- One of the most important results in OLS Regression Modeling.
- Requires previous basic assumptions about the error term ("random noise") ε.
- Guarantees that OLS estimates are *BLUE*:
  Best Linear Unbiased Estimator.

Gauss-Markov Theorem

The Gauss-Markov Theorem (GMT) states that OLS estimators are optimal in the following sense:

Among all linear unbiased estimators, OLS estimators have minimum variance.

In other words, if you have any other linear unbiased estimator, they will have more spread around β.

We refer to this property with the acronym **BLUE**:
Best Linear Unbiased Estimator

Gauss-Markov Theorem

OLS estimates $\hat{\beta}_j$ are unbiased

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}] \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}E(\mathbf{y}) \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} \end{aligned}$$

Gauss-Markov Theorem

What is the variance of OLS estimates $\hat{\beta}_j$?

$$Var(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}]$$

Note that:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\varepsilon}
\end{aligned}$$

Then:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\varepsilon}$$

Gauss-Markov Theorem

What is the variance of OLS estimates $\hat{\beta}_j$?

$$Var(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathsf{T}}]$$

Consequently:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

Gauss-Markov Theorem

GMT says the $\hat{\beta}_j$ have minimum variance

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$

For a given $\hat{\beta}_j$, its variance is:

$$Var(\hat{\beta}_j) = \sigma^2 \left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right]_{jj}$$

where $\left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right]_{jj}$ is the diagonal element of $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$

The issue is that we don't know σ^2!

Gauss-Markov Theorem

Be careful ...

GMT does NOT exclude the possibility of obtaining estimates that are nonlinear or biased that have less variance.

GMT is concerned only with *linear unbiased estimators*.

Classic Linear Model

Advertising Data

```
# file in folder data/ of github repo
Advertising <- read.csv("data/Advertising.csv", row.names = 1)
```

```
      TV   Radio   Newspaper   Sales
1  230.1    37.8        69.2    22.1
2   44.5    39.3        45.1    10.4
3   17.2    45.9        69.3     9.3
4  151.5    41.3        58.5    18.5
5  180.8    10.8        58.4    12.9
6    8.7    48.9        75.0     7.2
7   57.5    32.8        23.5    11.8
8  120.2    19.6        11.6    13.2
```

(first 8 rows)

Data set `Advertising`

Response:

- Y: `Sales`

Predictors:

- X_1: `TV`
- X_2: `Radio`
- X_3: `Newspaper`

Linear model:

$$\texttt{Sales} = \beta_0 + \beta_1\ \texttt{TV} + \beta_2\ \texttt{Radio} + \beta_3\ \texttt{Newspaper} + \epsilon$$

Example: Advertising Data

```
reg <- lm(Sales ~ ., data = Advertising)
reg_sum <- summary(reg)

reg

##
## Call:
## lm(formula = Sales ~ ., data = Advertising)
##
## Coefficients:
## (Intercept)           TV        Radio    Newspaper
##    2.938889     0.045765     0.188530    -0.001037
```

Example: Advertising Data

```
reg$coefficients

##  (Intercept)            TV         Radio     Newspaper
##  2.938889369  0.045764645  0.188530017 -0.001037493
```

$$\texttt{Sales} = 2.9389 + 0.0458\ \texttt{TV} + 0.1885\ \texttt{Radio} + \text{-}0.001\ \texttt{Newspaper}$$

Stochastic Assumptions about the Errors

The *classic* Linear Regression Model (via OLS) takes the Gauss-Markov assumptions, and adds to them a Normal distribution for the error terms $\boldsymbol{\varepsilon}$.

- $\varepsilon_i \sim N(0, \sigma^2)$
- ε_i are independent

In vector-matrix notation we have:

- $\boldsymbol{\varepsilon} \sim$ i.i.d. $N(\mathbf{0}, \sigma^2 \mathbf{I})$

Stochastic Assumptions about the Errors

Keep in mind that these assumptions are formalisms that allow us to develop inferential tasks.

Whether these assumptions are met in practice (real-world data) is another story.

Assumptions

- The main assumption is to further characterize the distribution of the error term.
- We assume that errors follow a Normal distribution.
- This assumption is for convenient reasons (many data sets in practice don't follow it).
- It provides a simple formalism that allows us to perform inference: hypothesis tests and confidence intervals

Derived Distributions of OLS Estimates

$$b_1 = \hat{\beta}_1 \sim N\left(\beta_1;\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$b_0 = \hat{\beta}_0 \sim N\left(\beta_0;\ \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

$$\hat{y} \sim \left(\beta_0 + \beta_1 x;\ \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

Derived Distributions of OLS Estimates

In the multiple regression case we have that the density of the response $\mathbf{y}$ is:

$$\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)$$

$\mathbf{y}$ is multinormally distributed: $\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

Variance of Estimates

Variance of Estimates

$$Var(\mathbf{b}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$

The variance of a particular coefficient b_j is given by:

$$Var(b_j) = \sigma^2 \left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right]_{jj}$$

where $\left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right]_{jj}$ is the diagonal element of $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$

Variance of Estimates

- Recall again that we don't know σ^2. How can we find an estimator $\hat{\sigma}^2$?
- We don't observe the error terms $\boldsymbol{\varepsilon}$ but we do have the residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$
- Perhaps the most "natural" option to estimate σ^2 is using the Residual Sum of Squares (RSS)

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

Question: is RSS an unbiased estimator of σ^2?

Estimate of σ^2

Is RSS an unbiased estimator of σ^2?

$$E\left[\sum_{i=1}^{n} e_i^2\right] = E[\mathbf{e}^\mathsf{T}\mathbf{e}] = (n-p-1)\sigma^2$$

It turns out that RSS is a **biased** estimator for σ^2

How can we get an unbiased estimator $\hat{\sigma}^2$?

Unbiased Estimate of σ^2

Instead of using RSS as an estimator of σ^2 we can slightly modify it to obtain an unbiased estimator as:

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

The square root $\hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}}$ is also known as the **Residual Standard Error** (reported by most software)

```
reg_sum
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Unbiased Estimate of σ^2

In R, the function `summary()` returns `"sigma"`, which is displayed as the `Residual standard error`:

```
reg_sum$sigma


## [1] 1.68551
```

What `summary()` actually returns is $\hat{\sigma} = \sqrt{\frac{RSS}{n-p}}$

instead of: $\hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}}$

recall that I'm using p for the number of predictors, NOT the number of columns in the model matrix

Distribution of σ^2

What can we say about the distribution of s^2?

$$\frac{RSS}{\sigma^2} = \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

Recall that a chi-square random variable is the sum of squared independent standard normal random variables.

Inference for $\hat{\beta}_1 = b_1$

Note that:

$$\frac{(b_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}{\sigma} \sim N(0, 1)$$

and that:

$$\frac{(n - p - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

Thus:

$$\frac{(b_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}{\sigma^2} \sim t_{n-p-1}$$

Confidence Interval for $\hat{\beta}_1 = b_1$

We can use the t-distribution to build confidence intervals for $\hat{\beta}_1 = b_1$

$$\hat{\beta}_1 - t_{1-\alpha/2}(n-p-1)s_j, \;\; \hat{\beta}_1 + t_{1-\alpha/2}(n-p-1)s_j$$

In particular, a 95% confidence interval is given by: $\hat{\beta}_1 \pm 2\hat{SE}(\hat{\beta}_1)$ where:

$$\hat{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_1 - \bar{x})^2}$$

Confidence Interval for coefficients

```
confint(reg)
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## TV           0.04301371 0.04851558
## Radio        0.17154745 0.20551259
## Newspaper   -0.01261595 0.01054097
```

Inference for $\hat{\beta}_1 = b_1$

We can also test the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.
Use the following test statistic:

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

that follows a t-distribution with $n - p - 1$ degrees of freedom.

Inference for coefficients

```
reg_sum$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## Radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## Newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

Anova

We've seen that RSS/σ^2 follows a χ^2_{n-p-1} distribution.

The quotient:

$$(n-p-1)\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \sim F_{(1,n-p-1)}$$

References

- **The Elements of Statistical Learning** by Hastie et al (2009). Springer.
- **Linear Models with R** by Julian Faraway (2015).
- **Modern Regression Methods** by Thomas Ryan (1997).
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*.

References (French Literature)

- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 17: La regression multiple et le modele lineaire general*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 5: La Regression Multiple*. Dunod, Paris.
- **Regression avec R** by Cornillon and Matzner-Lober (2011). Springer.
- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3.2: Regression multiple, modele lineaire*. Dunod, Paris.
- **Traitement des donnees statistiques** by Lebart et al. (1982). *Unit 3: Modele Lineaire*. Dunod, Paris.