Lecture 10: Linear Regression

Jing Ma, Statistics, TAMU

6 November, 2019

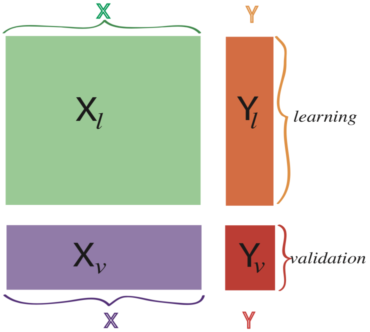
Recap

- Multivariate analysis
- Dimension reduction tools (PCA, MDS, CA)
- Visualization

This lecture

- Linear regression
- Training and test errors
- Analysis of variance

Supervised learning



- In supervised learning, we assign two different roles to our variables.
- We have labeled the explanatory variables X and the response variable(s) Y.
- There are also two different sets of observations: the training set (X_l, Y_l) and the validation set (X_v, Y_v).

Regression versus Classification

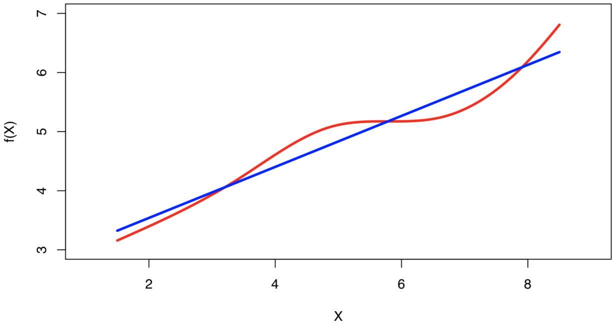
- Regression: Predict a **quantitative** response, such as
  - blood pressure
  - cholesterol level
  - tumor size
- Classification: Predict a **categorical** response, such as
  - tumor versus normal tissue
  - heart disease versus no heart disease
  - subtype of glioblastoma
- This lecture: **Regression**.

Linear models

- It assumes linear relationship between the response Y and the predictors in X.
- Simple linear regression: Y is univariate, X is univariate.
- Multiple linear regression: Y is univariate, X is multivariate.

Linear models

- True model may not be linear!
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.



Example of linear models

A typical linear model assumes

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- y_i: the response from individual i (e.g. the gene expression value).
- x_i: value of the predictor (e.g. disease status) from individual i.
- β_0: the unknown intercept.
- β_1: the unknown slope.
- ϵ_i: unobservable error variable from individual i.

The diabetes data

```r
diabetes = read_csv("../Lecture12/data/diabetes.csv", col_names = TRUE)
diabetes
```

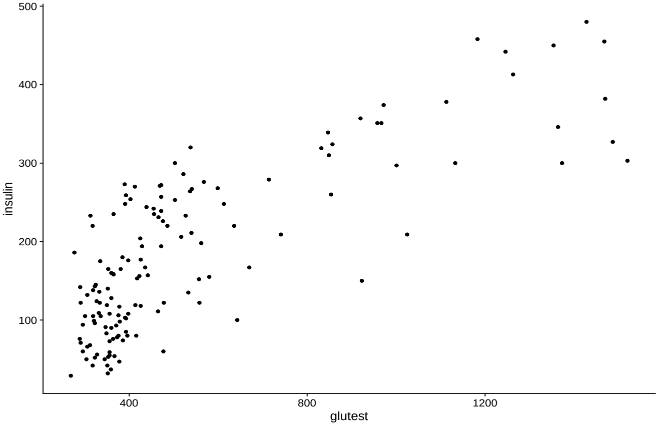
```
## # A tibble: 144 x 7
##       id relwt glufast glutest steady insulin group
##    <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl> <dbl>
##  1     1  0.81      80     356    124      55     3
##  2     3  0.94     105     319    143     105     3
##  3     5  1         90     323    240     143     3
##  4     7  0.91     100     350    221     119     3
##  5     9  0.99      97     379    142      98     3
##  6    11  0.9       91     353    221      53     3
##  7    13  0.96      78     290    136     142     3
##  8    15  0.74      86     312    208      68     3
##  9    17  1.1       90     364    152      76     3
## 10    19  0.83      85     296    116      60     3
## # ... with 134 more rows
```

Linear regression for the diabetes data

Questions we might ask:

- Is there a relationship between *glutest* and *insulin*?
- How strong is the relationship between *glutest* and *insulin*?
- How accurately can we predict future insulin responses?
- Is the relationship linear?

Linear regression for the diabetes data



Simple linear regression with one predictor

- We assume a model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- After fitting the model, we get estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. We predict future insulin response

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

The *hat* symbol denotes an estimated value.

The fitted linear model

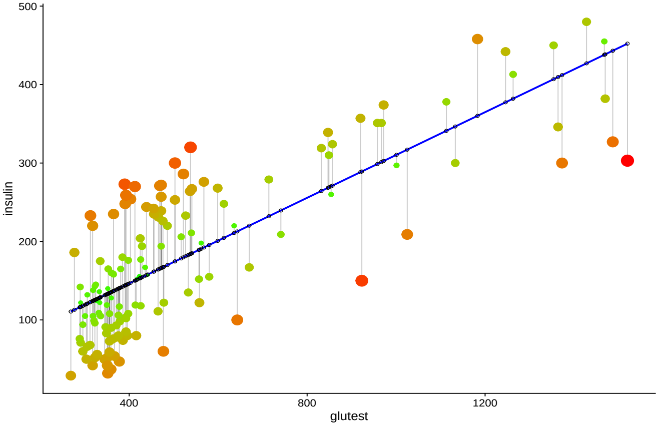


Fig. 1: The least square fit for the regression of insulin onto glutest. Blue line is positioned to minimize the sum of squared lengths of the gray lines.

What makes a model linear?

- A linear model is **linear in the regression coefficients**!

What makes a model linear?

- A linear model is **linear in the regression coefficients**!
- This is a linear model:

$$y_i = \beta_1 \sin(x_{i1}) + \beta_2 x_{i2} x_{i3} + \epsilon_i .$$

What makes a model linear?

- A linear model is **linear in the regression coefficients**!
- This is a linear model:

$$y_i = \beta_1 \sin(x_{i1}) + \beta_2 x_{i2} x_{i3} + \epsilon_i.$$

- This is not a linear model:

$$y_i = \beta_1^{x_{i1}} + \sin(\beta_2 x_{i2}) + \epsilon_i.$$

Fitting the model by 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 $e_i = y_i - \hat{y}_i$ represents the ith **residual**.
- We define the **residual sum of squares** (RSS) as

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

  or equivalently as

  $$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Fitting the model by least squares

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

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

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

Interpreting regression coefficient

Since there is only one predictor, we can interpret β_1 in the following sense:

- a unit change in glucose level is associated with a β_1 change in insulin response.

Interpretation with multiple linear regression can be more challenging, especially when predictors are correlated.

Assessing the accuracy of the estimated coefficient

- Total sum of squares

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

where $\bar{y}$ refers to the mean of the response y.

- Residual sum of squares

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

- R^2 or proportion of variance explained is

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

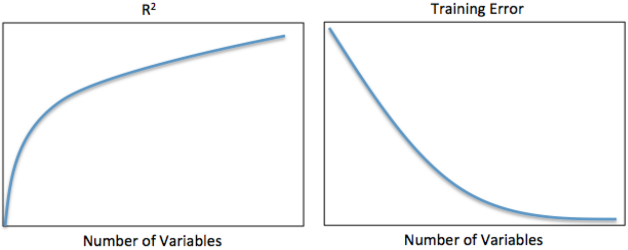
- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

Training error

- Once we fit the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$, we can evaluate the **training error**, i.e. the extent to which the model fits the observations used to train it.
- The training error is closely related to the R^2 for a linear model, that is, the **proportion of variance explained**.
- Big $R^2 \Leftrightarrow$ small traning error.

The problem

As we add more variables into the model. . .



. . . the training error decreases and the R^2 increases!

Why is this a problem?

- We really care about the model's performance on **o**bservations not used to fit the model!
  - We want a model that will predict the survival time of a new patient who walks into the clinic!
  - We want a model that can be used to diagnose cancer for a patient not used in model training!
  - We want to predict risk of diabetes for a patient who wasn't used to fit the model!

Why is this a problem?

- What we really care about:

  $$(y_{test} - \hat{y}_{test})^2,$$

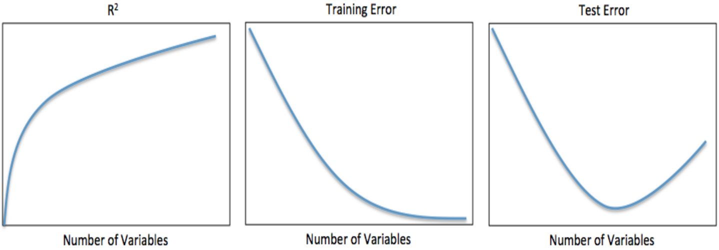
  where

  $$\hat{y}_{test} = \hat{\beta}_0 + \hat{\beta}_1 x_{test},$$

  and (x_{test}, y_{test}) **was not used to train the model**.
- The **test error** is the average of $(y_{test} - \hat{y}_{test})^2$ over a bunch of test observations.

Training error versus test error

As we add more variables into the model. . .

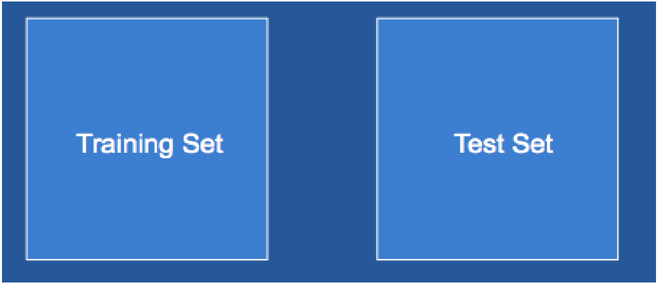


. . . the training error decreases and the R^2 increases!

But the test error might not!

How to estimate the test error?

- Split samples into *training* and *test* sets.
- Fit the model on the training set, and evaluate on test set.

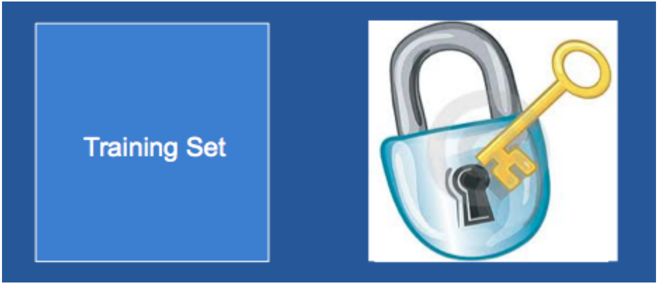


Q: Can there ever, under any circumstance, be sample overlap between the training and test sets?

A: Absolutely Not!

How to Estimate the Test Error?

- We fit a model on the training set, but we must evaluate its performance on a test set.
- Split samples into training set and test set.
- Fit model on training set, and evaluate on test set.



You can't peek at the test set until you are completely done all aspects of model-fitting on the training set!

What if our predictor is categorical?

Questions we may ask:

- Is there a relationship between diabetes group and insulin response?
- How strong is the relationship between diabetes group and insulin response?

Linear model with categorical predictor

Suppose

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

- x_i is group indicator
- We first look at the case where there are only two groups.

```
diabetes$g2 = sapply(diabetes$group, function(s) ifelse(s>2, 2, 1))
diabetes$g2 %<>% factor
```

- We used the forward-backward operator %<>% to convert the group column into a factor.

Linear model with categorical predictor

We can create the **design** matrix as follows:

```
dm = model.matrix(diabetes$insulin ~ diabetes$g2 - 1)
head(dm)
```

```
##   diabetes$g21 diabetes$g22
## 1            0            1
## 2            0            1
## 3            0            1
## 4            0            1
## 5            0            1
## 6            0            1
```

Linear model with categorical predictor

- In this case, there will be two predictors in the model, hence two regression coefficients β_1, β_2.
- We used −1 in creating the design matrix, which effectively removes the intercept.
- The coefficients reflect the group means. For example,
  - since observation 1 belongs to group 2, $y_1 = \mu_2 + \epsilon_1$;
  - observation 30 belongs to group 1, so $y_{30} = \mu_1 + \epsilon_{30}$.

How to fit the model

```
myfit = lm(diabetes$insulin ~ diabetes$g2 - 1)
summary(myfit)
```

- The output in the first column gives the estimated mean per group.
- The second gives the standard error of each mean
- The third gives the t-value (the estimate divided by the standard error)
- The last gives the corresponding p-values.
- From the p-values, do you reject or accept the null hypotheses $H_0 : \mu_1 = 0$?

What if we have three groups?

Try the code yourself? How is this model different from our previous one?

```
diabetes$group %<>% factor
myfit = lm(diabetes$insulin ~ diabetes$group - 1)
```

One-way ANOVA

- A frequent problem is that of testing the null hypothesis that three or more population means are equal.
- By comparing two types of variances, this is made possible by a technique called analysis of variance (ANOVA).
- Again, consider the insulin response from the *diabetes* data example. We use the `group` column which has three groups.
- The null-hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$.

One-way ANOVA

- Let data from group j be $y_{1j}, y_{2j}, \ldots, y_{nj}$, for $j = 1, 2, 3$.
- We have assumed the number of observations to be equal in each group for notational convenience, but this is not required in general.
- The three group means are

$$\bar{y}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}, \quad \bar{y}_2 = \frac{1}{n}\sum_{i=1}^{n} y_{i2}, \quad \bar{y}_3 = \frac{1}{n}\sum_{i=1}^{n} y_{i3}.$$

- The overall mean is

$$\bar{y} = \frac{1}{3n}\left(\sum_{i=1}^{n} y_{i1} + \sum_{i=1}^{n} y_{i2} + \sum_{i=1}^{n} y_{i3}\right).$$

One-way ANOVA

- The *sum of squares within (SSW)* is the sum of the squared deviation of the measurements to their group mean, i.e.

$$SSW = \sum_{j=1}^{g}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2.$$

- The *sum of squares between (SSB)* is the sum of squares of the deviances of the group mean with respect to the total mean, that is

$$SSB = \sum_{j=1}^{g}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2.$$

- The f-value is defined as

$$f = \frac{SSB/(g-1)}{SSW/(N-g)}.$$

One-way ANOVA

- The idea behind the test is that, under the null-hypothesis of equal group means, the value for SSB will tend to be small, so that the observed f-value will be small and $H_0 : \mu_1 = \mu_2 = \mu_3$ is accepted.
- If the data are normally distributed, then this f-value follows an F distribution with degrees of freedom $(g - 1, N - g)$. If $P(F > f) \geq \alpha$, then H_0 is not rejected, and otherwise it is.

ANOVA in R

```r
diabetes$group %<>% factor
anova(lm(diabetes$insulin ~ diabetes$group))
```

```
## Analysis of Variance Table
##
## Response: diabetes$insulin
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## diabetes$group   2 994900  497450  113.27 < 2.2e-16 ***
## Residuals      141 619231    4392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

- Simple linear regression
- Training error versus test error
- ANOVA