

# Data Mining

## W4240 Sections 001, 003/004

Lauren A. Hannah

Columbia University, Department of Statistics

September 25, 2014

# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

Connecting Linear Regression to PCA

Linear Regression Examples

# Video for Sections 003/004

There have been some issues with the technology for hybrid:

- ▶ frequent crashes of Adobe Connect
- ▶ unfamiliar systems for instructor
- ▶ non-functioning smart boards
- ▶ etc

**We will be moving from Adobe Connect to passive video capture for this course.**

Implications:

- ▶ remote students cannot connect during lecture (but do not need to be awake)
- ▶ more online office hours will be offered (AM EST as per Doodle poll)
- ▶ tests will still be administered at standard time

# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

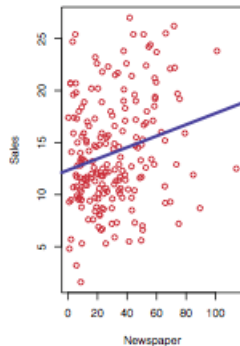
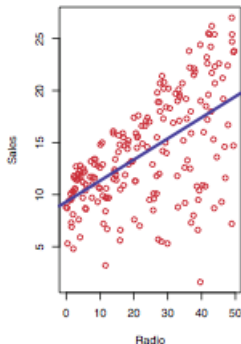
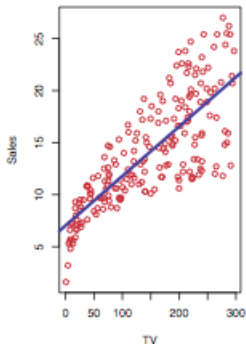
Connecting Linear Regression to PCA

Linear Regression Examples

# Advertising Data

Recall from before:

$$\text{sales} = f(\text{TV}, \text{radio}, \text{newspaper}) + \epsilon$$



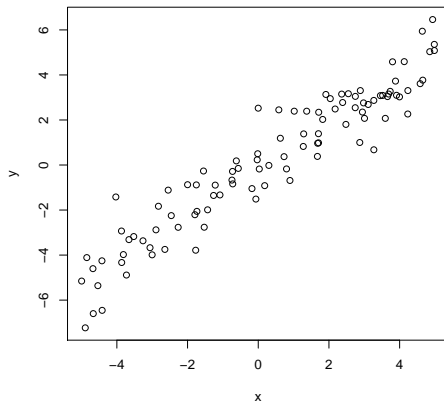
# Advertising Data

Questions:

1. Is there a relationship between the advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effects of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy between media? (Synergy means  $f(a \text{ and } b) > f(a) + f(b)$ )

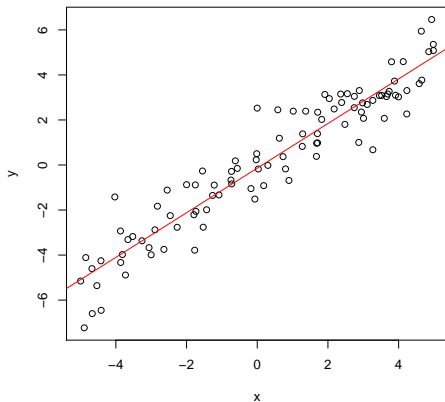
Can we use kNN to answer these questions? What type of model can answer these questions?

# Linear Regression



Training data are the set of inputs and outputs,  $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$

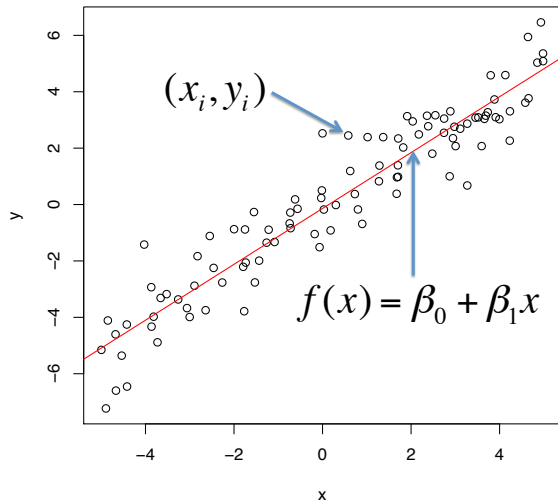
# Linear Regression



In *linear regression*, the goal is to predict  $Y$  from  $X$  using a linear function



# Linear Regression



Why am I now showing  $\beta$  and not  $w$  (a notational apology)?

# Linear Regression

Model:

$$f(X) \approx \beta_0 + \beta_1 X$$

...but we don't know  $\beta_0$  and  $\beta_1$ . How can we estimate them from the data?

Let's begin by looking at the *residual sum of squares* (RSS):

$$\begin{aligned} RSS(\beta_0, \beta_1) &= \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

The  $i^{th}$  *residual* is the difference between the observed value and the predicted value,  $y_i - \beta_0 - \beta_1 x_i$ . (Compare to MSE...)

# Linear Regression

Notice that RSS is a *function* of  $(\beta_0, \beta_1)$ . Let's pick the pair that **minimizes RSS**:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- ▶ Hold on! Last lecture you told us that picking a model that minimizes training MSE can lead to overfitting!
- ▶ Does this *model* minimize training MSE?

# Linear Regression

Let's take the derivative of  $RSS(\beta_0, \beta_1)$  with respect to  $\beta_0$  and set it equal to 0:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \quad (= 0)$$

$$n\beta_0 = -\beta_1 \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

What does this last line mean?

# Linear Regression

Now let's take the derivative of  $RSS(\beta_0, \beta_1)$  with respect to  $\beta_1$  and set it equal to 0:

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) \quad (= 0)$$

$$\beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i$$

$$\beta_1 = \frac{\sum_{i=1}^n y_i x_i - n \beta_0 \bar{x}}{\sum_{i=1}^n x_i^2}$$

# Linear Regression

Now we have two equations and two unknowns, so let's solve for  $\beta_0$  and  $\beta_1$ :

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

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

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

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

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

# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

Connecting Linear Regression to PCA

Linear Regression Examples

# How Accurate are the Coefficients?

We can approximate

$$Y = f(X) + \epsilon_1$$

by a linear model

$$Y = \beta_0 + \beta_1 X + \epsilon_2.$$

This is called the **population regression line**.

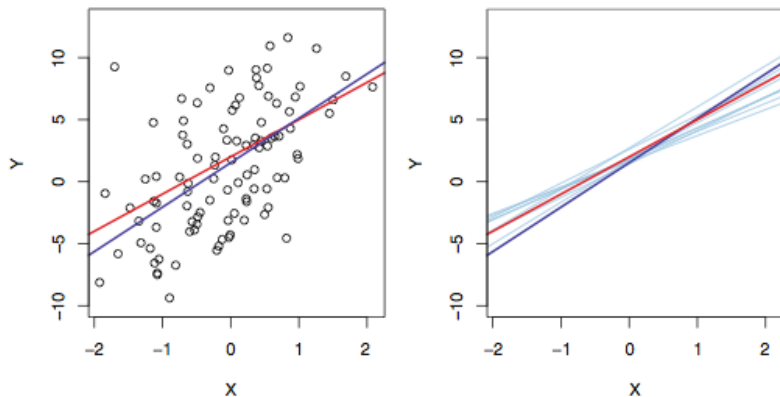
However,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  produce a **least squares line**,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

How different are these?



# How Accurate are the Coefficients?



Different data produce different  $(\hat{\beta}_0, \hat{\beta}_1)$ .<sup>1</sup>

---

<sup>1</sup>Some images are taken from *An Introduction to Statistical Learning* by James, Witten, Hastie and Tibshirani.

# How Accurate are the Coefficients?

Population regression line:

$$Y = \beta_0 + \beta_1 X + \epsilon_2$$

Least squares line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

What can we say about  $\hat{\beta}_0$  vs  $\beta_0$  and  $\hat{\beta}_1$  vs  $\beta_1$ :

- ▶  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are *unbiased* estimators of  $\beta_0$  and  $\beta_1$

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathbb{E}[\hat{\beta}_1] = \beta_1$$

- ▶ (prove that for at least one of these)
- ▶ What about  $\text{Var}(\hat{\beta}_0)$ ? (Read James 3.1 or take W4315)
- ▶ This allows us to do hypothesis testing... why?

# How Accurate are the Coefficients?

We can use these ideas for **hypothesis testing**. Suppose that we would like to know whether there is a relationship between  $X$  and  $Y$ . We can have the null hypothesis

$H_0$  : there is no relationship between  $X$  and  $Y$

and the alternative hypothesis

$H_1$  : there is a relationship between  $X$  and  $Y$

In terms of our model,

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

(Can you think of a situation where there is a relationship but  $\beta_1 = 0$ ?)

# How Accurate are the Coefficients?

We can use a ***t*-statistic** to determine whether we should reject the null hypothesis:

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

If there is no relationship, we would expect  $\hat{\beta}_1$  to have a  $t$  distribution with  $n - 2$  degrees of freedom.

- ▶ if the  $t$ -statistic has really large or really small values, it is unlikely that the observed value came from a  $t$ -distribution with mean 0, spread  $SE(\hat{\beta}_1)$  and  $n - 2$  degrees of freedom
- ▶ if it is not that large, it could have come from that distribution and we cannot reject  $H_0$
- ▶ we use  $p$ -values and a rejection threshold to determine whether we should reject  $H_0$

If you want to know more, take W4315: Linear Regression Models.

# Assessing Model Accuracy

So how well does this model represent the data?

- ▶ let's try to see how much of the data spread this model captures
- ▶  $RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  is the total amount of error left in the model
- ▶ ...but there is some noise to begin with, measured by the *total sum of squares*  $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$
- ▶ the  $R^2$  statistic is the proportion of error reduced by the model,

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

# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

Connecting Linear Regression to PCA

Linear Regression Examples

# Multiple Linear Regression

So what happens if we have more than one predictor? Well, we can just include those in the model as well,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & & & \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_p \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \epsilon$$

$$Y_i = w_0 + w_1 X_{i1} + \dots + w_p X_{ip}$$

# Multiple Linear Regression

How do we find the least squares coefficients? Same as always...

- ▶ Find  $RSS(\mathbf{w})$
- ▶ Take the gradient with respect to  $\mathbf{w}$ , set equal to 0
- ▶ Do some algebra ( $p + 1$  equations,  $p + 1$  unknowns)



# Multiple Linear Regression

To find the best weights  $\mathbf{w}$ :

$$\begin{aligned}\hat{\mathbf{w}} &= \arg \min \sum_{i=1}^n \left( y_i - \mathbf{w}^\top \mathbf{x}_i \right)^2 \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2y_i \mathbf{w}^\top \mathbf{x}_i + \left( \mathbf{w}^\top \mathbf{x}_i \right)^2 \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2\mathbf{w}^\top \left( \sum_{i=1}^n y_i \mathbf{x}_i \right) + \mathbf{w}^\top \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w} \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}\end{aligned}$$

Now take the gradient in  $\mathbf{w}$ , set to 0, and...

$$\hat{\mathbf{w}} = \left( \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

# Is Multivariate Linear Regression Linear Regression?

- ▶ What happens when  $\mathbf{X} \in \mathbb{R}^{n \times 2}$  ? (Remember the ones!)
- ▶ Important crosscheck (and mechanics practice)

Some caveats:

- ▶ we only have a solution if  $(\mathbf{X}^T \mathbf{X})^{-1}$  exists
- ▶ this happens when  $\mathbf{X}^T \mathbf{X}$  has full rank (here,  $p + 1$ ):
  - ▶  $\mathbf{X}$  has rank  $p + 1$  ( $p + 1$  linearly independent covariate observations)
  - ▶ happens with random sampling from a noisy continuous distribution
- ▶ if  $(\mathbf{X}^T \mathbf{X})^{-1}$  does not exist, there are infinitely many solutions that are optimal
- ▶ Example:  $\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ ,  $\mathbf{Y} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$  vs

# Multivariate Linear Regression

So what about hypothesis tests with multiple  $\beta$ 's?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 : \text{at least one } \beta_j \neq 0$$

We can perform this hypothesis test by computing the  $F$ -statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

- ▶ if linear assumption is true,  $\mathbb{E}[RSS/(n - p - 1)] = \sigma^2$
- ▶ if  $H_0$  is true,  $\mathbb{E}[(TSS - RSS)/p] = \sigma^2$
- ▶ so if  $H_0$  is true,  $F$ -statistic should be close to 1, otherwise it should be greater
- ▶ rejection threshold depends on  $n$  and  $p$  (again, W4315)

# Linear Regression: (Almost) All You Need To Know

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

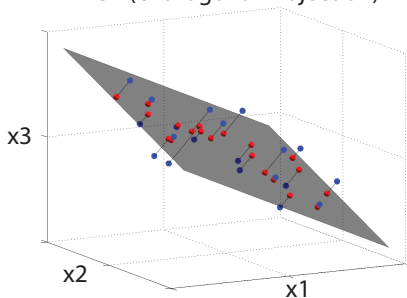
Connecting Linear Regression to PCA

Linear Regression Examples

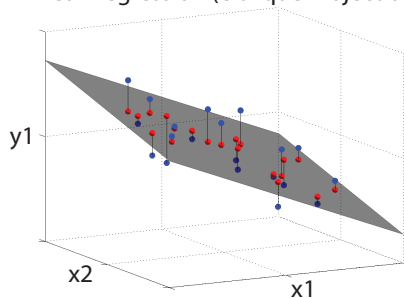
# Big Picture: Supervised vs Unsupervised

- ▶ Unsupervised learning seeks explanatory factors
- ▶ Supervised learning asserts explanatory factors

PCA (Orthogonal Projection)



Linear Regression (Oblique Projection)



# Multiple Linear Regression

To find the best weights  $\mathbf{w}$ :

$$\begin{aligned}\hat{\mathbf{w}} &= \arg \min \sum_{i=1}^n \left( y_i - \mathbf{w}^\top \mathbf{x}_i \right)^2 \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2y_i \mathbf{w}^\top \mathbf{x}_i + \left( \mathbf{w}^\top \mathbf{x}_i \right)^2 \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2\mathbf{w}^\top \left( \sum_{i=1}^n y_i \mathbf{x}_i \right) + \mathbf{w}^\top \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w} \\ &= \arg \min \sum_{i=1}^n y_i^2 - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}\end{aligned}$$

Now take the gradient in  $\mathbf{w}$ , set to 0, and...

$$\hat{\mathbf{w}} = \left( \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

# Principal Component Analysis

To find the best weights  $\mathbf{w}$ :

$$\begin{aligned}\hat{\mathbf{w}} &= \arg \min \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{w}\mathbf{w}^\top \mathbf{x}_i\|^2 \\ &= \arg \min \sum_{i=1}^n \left( \mathbf{x}_i - \mathbf{w}\mathbf{w}^\top \mathbf{x}_i \right)^\top \left( \mathbf{x}_i - \mathbf{w}\mathbf{w}^\top \mathbf{x}_i \right) \\ &= \arg \min \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - 2(\mathbf{w}^\top \mathbf{x}_i)^2 + (\mathbf{w}^\top \mathbf{x}_i)^2 \\ &= \arg \min \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}\end{aligned}$$

Now take the gradient in  $\mathbf{w}$  (with the constraint  $\|\mathbf{w}\| = 1$ ), set to 0, and...

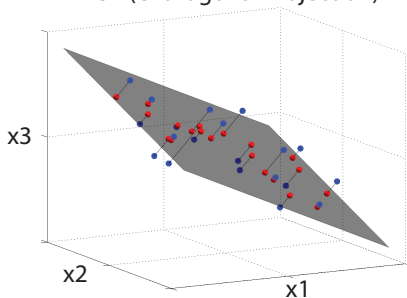
$$\hat{\mathbf{w}} = \text{svd} \left( \mathbf{X}^\top \mathbf{X} \right)$$



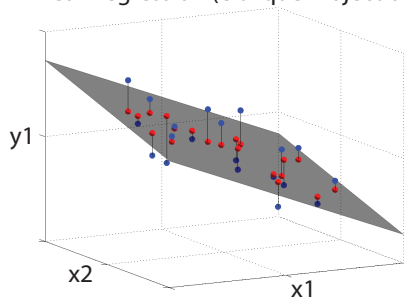
# Big Picture: Supervised vs Unsupervised

- ▶ Unsupervised learning seeks explanatory factors
- ▶ Supervised learning asserts explanatory factors

PCA (Orthogonal Projection)



Linear Regression (Oblique Projection)



# Outline

Announcements

Basic Linear Regression

Accuracy of Linear Regression

Multiple Linear Regression

Connecting Linear Regression to PCA

Linear Regression Examples

# Advertising Data

## Questions:

1. Is there a relationship between the advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effects of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy between media?

**Is there a relationship between the advertising budget and sales?**

Use *hypothesis testing*!

Regress sales on TV + newspaper + radio = advertising;  
 $\text{sales} = \beta_0 + \beta_1 \text{advertising}$

$$H_0 : \beta_1 = 0 \quad (1)$$

$$H_1 : \beta_1 \neq 0$$

**How strong is the relationship between advertising budget and sales?**

*Use a confidence interval!*

Regress sales on TV + newspaper + radio = advertising;  
 $\text{sales} = \beta_0 + \beta_1 \text{advertising}$

- ▶  $\beta_1$  tells us the relationship between advertising budget and sales
- ▶ a confidence interval gives us more certainty about where the true parameter lies

## Which media contribute to sales?

Use coefficient  $p$ -values!

- ▶ from hypothesis testing against coefficient being equal to 0
- ▶ it is highly likely that coefficients with high  $p$ -values have a linear relationship with the output

**How accurately can we estimate the effects of each medium on sales?**

Use a coefficient *confidence interval*!

**How accurately can we predict future sales?**

Use  $R^2$ !

# Advertising Data

More questions:

1. Is the relationship linear?
2. Is there synergy between media?

Well, we can't answer these yet...



## Example: Old Faithful

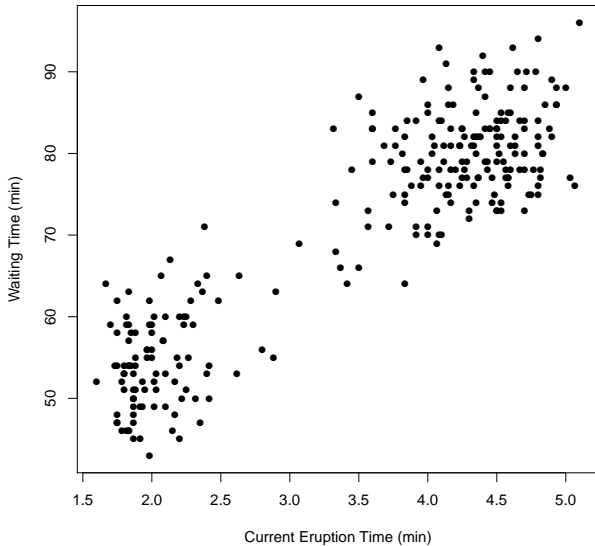


## Example: Old Faithful

We will predict the time that we will have to wait to see the next eruption given the duration of the current eruption

```
> library(datasets)
> names(faithful)
[1] "eruptions" "waiting"
> attach(faithful)
> plot(eruptions,waiting,xlab="Current Eruption Time (min)",
+ ylab="Waiting Time (min)",pch=16)
```

# Example: Old Faithful



## Example: Old Faithful

To fit a linear model in R, use the `lm( )` function, which stands for “linear model”

```
> fit.lm <- lm(waiting ~ eruptions)
> fit.lm
```

Call:

```
lm(formula = waiting ~ eruptions)
```

Coefficients:

(Intercept)	eruptions
33.47	10.73

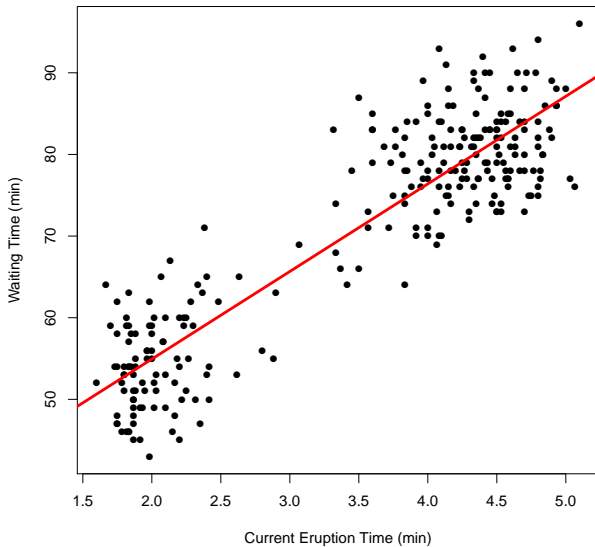
```
> names(fit.lm)
[1] "coefficients" "residuals"      "effects"
[4] "rank"         "fitted.values"  "assign"
[7] "qr"           "df.residual"    "xlevels"
[10] "call"         "terms"          "model"
```

## Example: Old Faithful

We can plot our data and make a function for new predictions

```
> # Plot a line on the data
> abline(fit.lm,col="red",lwd=3)
>
> # Make a function for prediction
> fit.lm$coefficients[1]
(Intercept)
  33.4744
> fit.lm$coefficients[2]
eruptions
 10.72964
> faithful.fit <- function(x) fit.lm$coefficients[1] +
fit.lm$coefficients[2]*x
> x.pred <- c(2.0, 2.7, 3.8, 4.9)
> faithful.fit(x.pred)
[1] 54.93368 62.44443 74.24703 86.04964
```

# Example: Old Faithful



## Example: Prostate Data

Data in Prostate.txt (also available on ESL website)

Predictors (columns 1–8): lcavol (log cancer volume), lweight (log weight), age, lbph (log amount of benign prostatic hyperplasia), svi (seminal vesicle inversion), lcp (log capsular penetration), gleason, pgg45 (percentage of Gleason scores 4 or 5)

outcome (column 9): lpsa (level of prostate-specific antigen)

train/test indicator (column 10)

```
> prostate <- read.table("Prostate.txt",header=TRUE, sep="\t")
> names(prostate)
[1] "X"          "lcavol"     "lweight"    "age"
[5] "lbph"       "svi"        "lcp"        "gleason"
[9] "pgg45"      "lpsa"       "train"
> prostate.train <- prostate[prostate$train==T,2:10]
> prostate.test <- prostate[prostate$train==F,2:10]
```

## Example: Prostate Data

```
> prostate.lm <- lm(lpsa ~ lcavol + lweight + age + lbph
  + svi + lcp + gleason + pgg45, data=prostate.train)
> # Other way:
> # prostate.lm <- lm(lpsa ~., data=prostate.train)
> # Exclude intercept by:
> # prostate.lm <- lm(lpsa ~ lcavol + lweight + age + lbph
  + svi + lcp + gleason + pgg45 - 1, data=prostate.train)
> y.pred.lm <- predict(prostate.lm, prostate.test)
> mean((y.pred.lm - prostate.test$lpsa)^2)
[1] 0.521274
```

Note: the data in ESL was scaled before use, so  $\hat{\beta}$  differs