Contents

1 Fitting a Model in R

Let's create some artificial data that fit a simple linear regression model. We'll

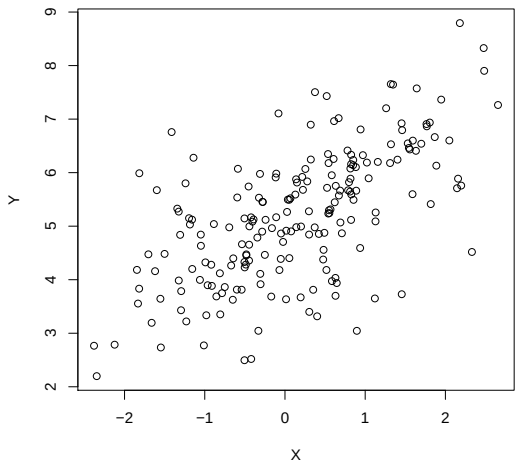
- make the X variable normal;
- create an ϵ that is also normal and independent;
- then create Y from the linear model equation $Y = b_0 + b_1 X + \epsilon$

In the code below, I set b_1 and b_0 to 0.7 and 5, respectively.

```
> set.seed(12345) ## seed the random generator
> X ← rnorm(200)
> epsilon ← rnorm(200)
> b1 ← 0.7
> b0 ← 5
> Y ← b0 + b1 * X + epsilon
```

Displaying the scatterplot in R uses the `plot` command:

```
> plot(X,Y)
```



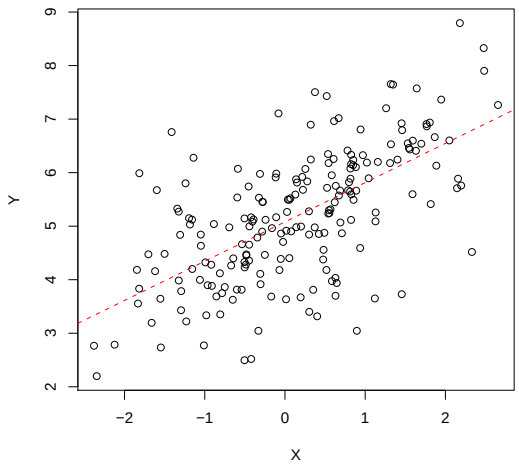
Adding the best-fitting straight line is easy. We obtain a linear model "fit object" with the command:

```
> fit.1 ← lm(Y~X)
> fit.1


Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     5.0796       0.7337
```

These fitted coefficients have a sampling error connected to them, of course. We know that the population values are 0.7 and 5, while the sample estimates here are 0.73 and 5.08.

Adding the best fitting straight line is very simple in R, if you've saved your fit object. You just type:

```
> plot(X,Y)
> #  make the fit line dotted and red
> #  lty=2      dotted
> abline(fit.1,lty=2,col='red')
```



`abline` is a versatile function that plots lines with a given slope and intercept. It also operates directly on a fit object.

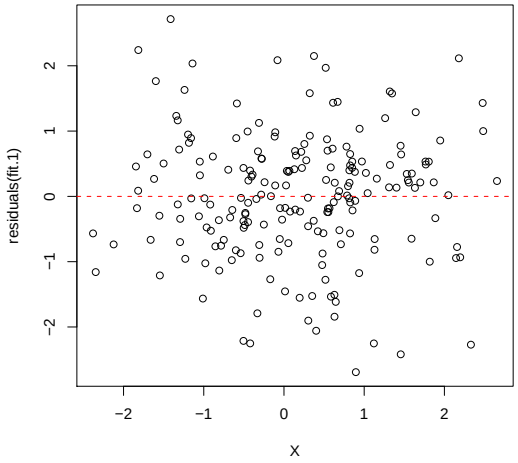
2 Examining Residuals

Recall that the residuals should have a conditional mean of zero for any conditional value of X, and the residual variance should not vary as a function of X.

A quick visual examination with a *residual plot* can reveal major departures from these assumptions.

Computing residuals is really simple in R. You simply apply the `residuals` function to the fit object.

```
> plot(X,residuals(fit.1))
> #add a red dotted line at zero
> #to aid evaluation
> abline(0,0,lty=2,col='red')
```



Things look just as they should.

The residuals are centered on the zero line, and show about the same level of variability as we move across the zero line from left to right.

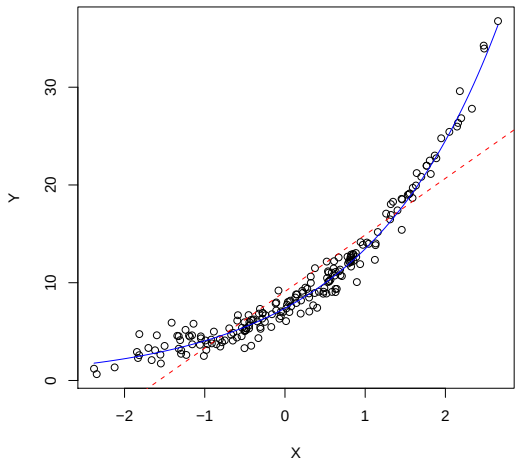
3 Diagnosing Problems

Let's create data that systematically deviate from the linear regression model.

```
> set.seed(12345) ## seed the random generator
> X ← rnorm(200)
> epsilon ← rnorm(200)
> b1 ← .6
> b0 ← 2
> Y ← exp(b0 + b1 * X) + epsilon
```

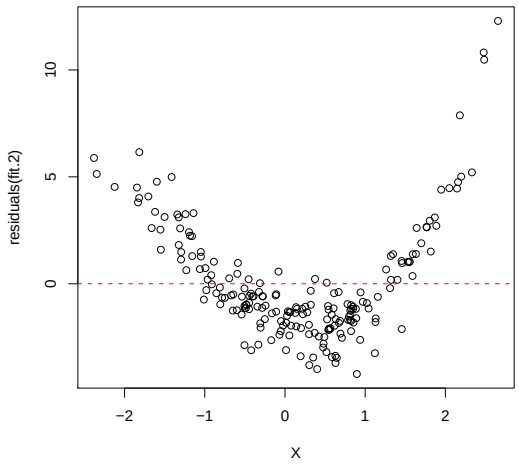
Plotting the data, we see a nonlinear relation. I've added a blue line showing the conditional mean of Y given X. You can see why, when you fit a straight line to these data, the residuals will have a positive mean at low values of X, a negative mean in the middle values, and a positive mean at high values of X.

```
> plot(X,Y)
> fit.2 ← lm(Y~X)
> abline(fit.2,lty=2,col='red')
> curve(exp(b1*x + b0),add=TRUE,col='blue')
```



This shows up very clearly in the residual plot

```
> plot(X,residuals(fit.2))
> abline(0,0,lty=2,col='red')
```



Diagnosing Nonlinearity

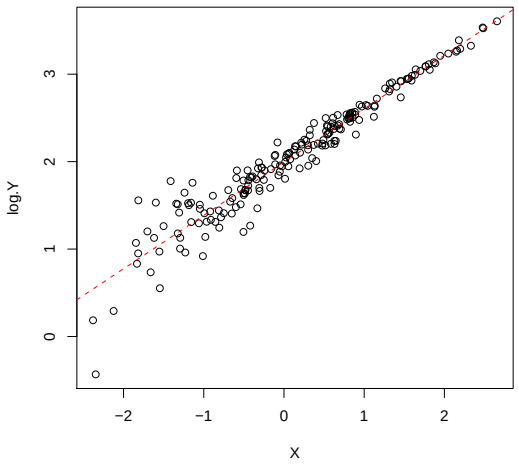
In this case, the solution is straightforward (especially since we know the precise rule that generated the data!).

Since the systematic part of the graph is an exponential function of a linear function of X, the log of Y should have a linear relationship with X. So simply

transforming Y should make a linear regression work much better.

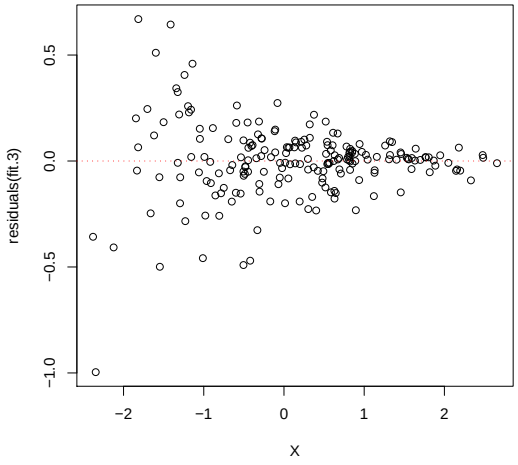
Diagnosing Nonlinearity

```
> log.Y <- log(Y)
> fit.3 <- lm(log.Y ~ X)
> plot(X,log.Y)
> abline(fit.3,lty=2,col='red')
```



Diagnosing Nonlinearity

```
> plot(X,residuals(fit.3))
> abline(0,0,lty=3,col='red')
```



4 Measures of Model Fit

4.1 R^2

The multiple correlation R for a linear regression model is the correlation between the predicted scores $\hat{y}$ and the actual criterion scores y.

Statistical texts have a long tradition of using R to stand for both the population and sample multiple correlation. I'm not sure if it would help for me to depart from this tradition.

Consequently, I'll refer loosely to "the population R^2" or, when confusion is more likely, R^2_{pop}. Some authors will use ρ^2, and you'll have to stay alert for these notational variations.

Denoting the variance of the y's by σ^2_y, the variance of the predicted scores by $\sigma^2_{\hat{y}}$, and the variance of the residuals by σ^2, we can show that

$$R^2_{pop} = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y} \tag{1}$$

That is, R^2_{pop} is the proportion of the variance of the criterion that is predictable from the linear regression equation.

As a simple consequence of the geometry of linear regression, the predicted and residual scores in linear regression are always precisely uncorrelated, and consequently,

$$\sigma^2_y = \sigma^2_{\hat{y}} + \sigma^2_\epsilon \tag{2}$$

Consequently, we may also write

$$R^2_{pop} = 1 - \frac{\sigma^2_\epsilon}{\sigma^2_y} \tag{3}$$

or, equivalently

$$R^2_{pop} = \frac{\sigma^2_y - \sigma^2_{\hat{y}}}{\sigma^2_y} \tag{4}$$

If the regression equation is of no use at all, the residual variance is simply σ^2_y. So another way of describing R^2 is "the proportion of potential error variance saved by using the regression equation."

4.2 The residual standard deviation $\hat{\sigma}$

The residual standard deviation $\hat{\sigma}$

The *conditional distribution* of y given $x = a$ is the mean of the distribution of y for those pairs of values for which x takes on the value a. Conditional distributions are of fundamental importance in statistics, because many distributions are interpretable primarily when conditionalized.

Consider, for example, height and weight. We are used to interpreting weight conditionalized on a value of height. If a person weighs 210 pounds, our interpretation is quite different for a height of 56 inches than for a height of 84 inches.

The residual standard deviation $\hat{\sigma}$

A key result in linear regression theory is that *conditional means follow the regression line*. That is, the conditional mean of the distribution of y given $x = a$ is normal, with mean $\hat{y}$ evaluated at a. Moreover, the standard deviation of the conditional distribution is σ_ϵ.

Consequently, σ_ϵ tells us how the observations in the conditional distribution are spread out around the conditional mean, and is a direct measure of how closely we can predict a value of y from the value of x.

The residual standard deviation $\hat{\sigma}$

For example, suppose the regression line relating weight to height is $y = 6x - 270$, and $\sigma_\epsilon = 20$. What is the conditional distribution of y given $x = 75$?

Let's write an R function to compute $\hat{y}$

```
> yhat ← function(a){6*a-270}
> yhat(75)
```

```
[1] 180
```

The conditional mean is 180. The conditional distribution has a standard deviation of 20.

The residual standard deviation $\hat{\sigma}$

The preceding result tells us that men who are 75 inches tall show a distribution of weights. This distribution is centered on 180, but also shows considerable variation around it.

How likely is a 75 inch tall man's weight to be within ± 20 pounds of the predicted value 180?

From our basic knowledge of the normal curve, we know that values within ± 1 standard deviation of the mean occur about 68% of the time. We can also see that men who are 75 inches tall are above 200 pounds in weight about 16% of the time.

```
> 1-pnorm(200,180,20)
```

```
[1] 0.1586553
```

The residual standard deviation $\hat{\sigma}$ vs. R^2

Gelman and Hill make the point that $\hat{\sigma}$ can convey information different from R^2. Suppose we look at the relationship between height and weight for a large group of men, and assume that, in the general population, weights have a mean of 150 and a standard deviation of 25, and heights have a mean of 70 and a standard deviation of 2.5, and, moreover, heights and weights correlate 0.60. Then $R^2 = .36$, and $\sigma_\epsilon = 20$.

The residual standard deviation $\hat{\sigma}$ vs. R^2

We can create data in R to simulate this situation:

```
> set.seed(12345)
> height <- rnorm(1000,70,2.5)
> error <- rnorm(1000,0,20)
> weight <- 6 * height -270 + error
> height.data <- data.frame(height,weight)
> attach(height.data)
```

```
        The following object(s) are masked _by_ .GlobalEnv :

         height weight
```

The residual standard deviation $\hat{\sigma}$ vs. R^2

Here is the output:

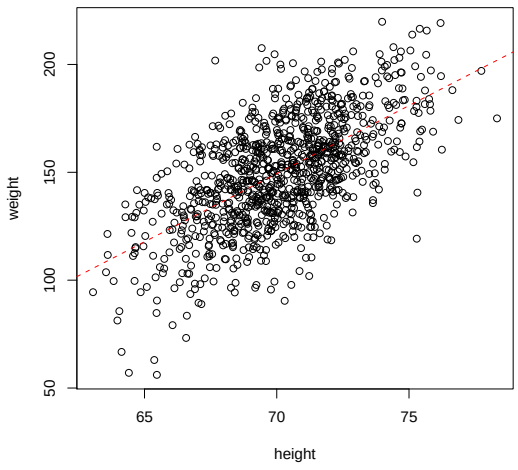
```
> fit <- lm(weight~height)
> summary(fit)
```

```
Call:
lm(formula = weight ~ height)

Residuals:
      Min        1Q    Median        3Q       Max
-64.63010 -13.95044   0.02194  14.31202  67.15312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -292.5670    17.9244  -16.32   <2e-16 ***
height         6.3132     0.2555   24.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.16 on 998 degrees of freedom
Multiple R-squared: 0.3796,        Adjusted R-squared: 0.379
F-statistic: 610.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

The residual standard deviation $\hat{\sigma}$ vs. R^2

Here is the plot:

```
> plot(height,weight)
> abline(fit,lty=2,col='red')
```



The residual standard deviation $\hat{\sigma}$ vs. R^2

We can restrict ourselves to a narrower range of heights

```
> restricted.range <-height.data[which( (height > 68) & (height < 72)),]
> fit <- lm(restricted.range$weight ~ restricted.range$height)
> summary(fit)
```

```
Call:
lm(formula = restricted.range$weight ~ restricted.range$height)

Residuals:
     Min       1Q   Median       3Q      Max
-60.7019 -14.2128   0.2260  14.2996  61.6950

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -273.6950    52.9370  -5.170 3.22e-07 ***
restricted.range$height    6.0431     0.7556   7.998 6.82e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.3 on 584 degrees of freedom
Multiple R-squared: 0.09872,        Adjusted R-squared: 0.09718
F-statistic: 63.97 on 1 and 584 DF,  p-value: 6.816e-15
```

The residual standard deviation $\hat{\sigma}$ vs. R^2

Here is the picture

```
> plot(restricted.range$height,restricted.range$weight)
> abline(fit,lty=2,col='red')
```