

# Lecture Notes for 36-707

## Linear Regression

Larry Wasserman  
Fall 2007

### 1 Prediction

Suppose  $(X, Y)$  have a joint distribution  $f(x, y)$ . You observe  $X = x$ . What is your best prediction of  $Y$ ?  
Let  $g(x)$  be any prediction function. The prediction error (or risk) is

$$R(g) = \mathbb{E}(Y - g(X))^2.$$

Let

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dy \quad \text{the regression function.}$$

Key result: for any  $g$ ,

$$R(r) \leq R(g).$$

Let  $\epsilon = Y - r(X)$ . Then,  $\mathbb{E}(\epsilon) = 0$  and we can write

$$Y = r(X) + \epsilon. \tag{1}$$

But we don't know  $r(x)$ . So we estimate from the data.

### 2 Some Terminology

Given data  $(X_1, Y_1), \dots, (X_n, Y_n)$  we have two goals:

- estimation:** Find an estimate  $\hat{r}(x)$  of the regression function  $r(x)$ .
- prediction:** Given a new  $X$ , predict  $Y$ ; we use  $\hat{Y} = \hat{r}(X)$  as the prediction.

At first we assume that  $Y_i \in \mathbb{R}$ . Later in the course, we consider other cases such as  $Y_i \in \{0, 1\}$ .

	$r$ linear	$r$ arbitrary
$X$ scalar	$r(x) = \beta_0 + \beta_1 x$ simple linear regression	$r(x)$ is some smooth function nonparametric regression
$X$ vector	$r(x) = \beta_0 + \sum_j \beta_j x_j$ multiple linear regression	$r(x_1, \dots, x_p)$ is some smooth function multiple nonparametric regression

### 3 Simple Linear Regression: $X$ scalar and $r(x)$ linear

Suppose that  $Y_i \in \mathbb{R}, X_i \in \mathbb{R}$  and that

$$\textcircled{\times} \quad r(x) = \beta_0 + \beta_1 x. \tag{2}$$

**This model is wrong. There is no reason to assume that  $r$  is linear. We make this assumption tentatively but we will drop it later. I use the  $\textcircled{\times}$  symbol to alert you to model-based statements.**

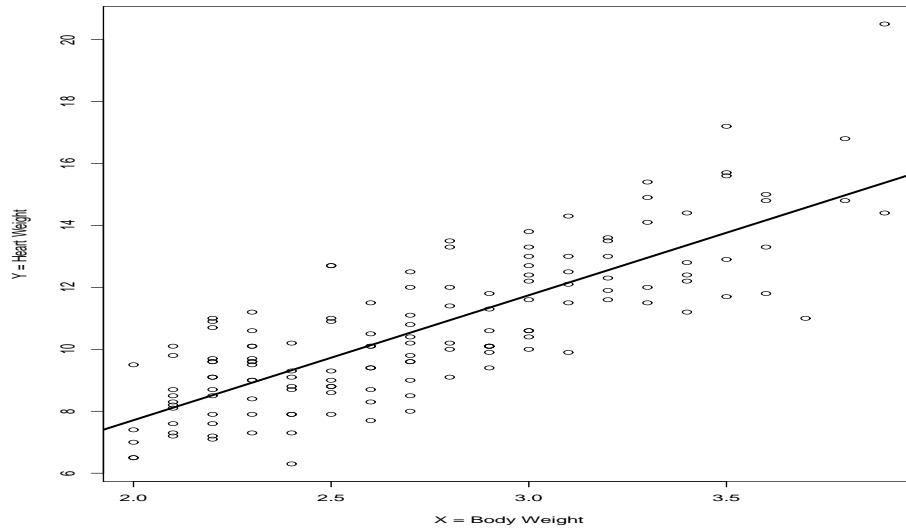


Figure 1: Cat example

We can write

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (3)$$

where  $\mathbb{E}(\epsilon_i) = 0$  and  $\epsilon_1, \dots, \epsilon_n$  are independent. We also assume that  $\mathbb{V}(\epsilon_i) = \sigma^2$  does not depend on  $x$ . (Homoskedasticity.) The unknown parameters are:  $\beta_0, \beta_1, \sigma^2$ . Define the residual sums of squares

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n \left( Y_i - (\beta_0 + \beta_1 X_i) \right)^2. \quad (4)$$

The least squares estimators (LS) minimize:  $\text{RSS}(\beta_0, \beta_1)$ .

**3.1 Theorem.** *The LS estimators are*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (5)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (6)$$

where  $\bar{X} = n^{-1} \sum_{i=1}^n X_i$  and  $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ .

We define:

- The fitted line:**  $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- The predicted or fitted values:**  $\hat{Y}_i = \hat{r}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- The residuals:**  $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- The residual sums of squares:**  $\text{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2$

An unbiased estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-2} \quad (7)$$

The estimators are random variables and have the following properties (**conditional on**  $X_1, \dots, X_n$ ):

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \quad \mathbb{E}(\hat{\beta}_1) = \beta_1, \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} \frac{1}{s_x^2}$$

where  $s_x^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . Also,  $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ . The **standard error**:

$$\text{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{n}} \frac{1}{s_x}, \quad \widehat{\text{se}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n}} \frac{1}{s_x}.$$

### Approximate Normality

$$\hat{\beta}_0 \approx N\left(\beta_0, \widehat{\text{se}}^2(\hat{\beta}_0)\right), \quad \hat{\beta}_1 \approx N\left(\beta_1, \widehat{\text{se}}^2(\hat{\beta}_1)\right) \quad (8)$$

If  $\epsilon_i \sim N(0, \sigma^2)$  then:

1. Equation (8) is exact.
2. The least squares estimators are the maximum likelihood estimators.
3. The variance estimator satisfies:

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi_{n-2}^2}{n-2}$$

## 4 Inference

It follows from (8) that an approximate  $1 - \alpha$  confidence interval for  $\beta_1$  is

$$\hat{\beta}_1 \pm z_{\alpha/2} \widehat{\text{se}}(\hat{\beta}_1) \quad (9)$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$  quantile of a standard Normal:

$$\mathbb{P}(Z > z_{\alpha/2}) = \frac{\alpha}{2}, \quad \text{where } Z \sim N(0, 1).$$

For  $\alpha = .05$ ,  $z_{\alpha/2} = 1.96 \approx 2$ , so, an approximate 95 per cent confidence interval for  $\beta_1$  is

$$\hat{\beta}_1 \pm 2\widehat{\text{se}}(\hat{\beta}_1). \quad (10)$$

**4.1 Remark.** If the residuals are Normal, then an exact  $1 - \alpha$  confidence interval for  $\beta_1$  is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \widehat{\text{se}}(\hat{\beta}_1) \quad (11)$$

where  $t_{\alpha/2, n-2}$  is the upper  $\alpha/2$  quantile of a  $t$  with  $n - 2$  degrees of freedom. This interval is **bogus**. If  $n$  is large,  $t_{\alpha/2, n-2} \approx z_{\alpha/2}$  so just use the Normal interval. If  $n$  is so small, that  $t_{\alpha/2, n-2}$  is much different than  $z_{\alpha/2}$ , the  $n$  is too small to be doing statistical inference. (Do you really believe that the residuals are exactly Normal anyway?) ■

To test

$$\mathcal{C} \quad H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0 \quad (12)$$

use the test statistic

$$z = \frac{\hat{\beta}_1 - 0}{\widehat{\text{se}}(\hat{\beta}_1)}. \quad (13)$$

Under  $H_0$ ,  $z \approx N(0, 1)$ . The p-value is

$$\text{p-value} = \mathbb{P}(|Z| > |z|) = 2\Phi(-|z|) \quad (14)$$

where  $Z \sim N(0, 1)$ . Reject  $H_0$  if p-value is small.

**4.2 Example.** Here is an example. The plots are shown in Figure 2.

```
### Cat example
###
library(MASS)
attach(cats); help(cats)
names(cats)
  [1] "Sex" "Bwt" "Hwt"
summary(cats)
  Sex           Bwt           Hwt
F:47           Min.      :2.000   Min.      : 6.30
M:97           1st Qu.:2.300   1st Qu.:  8.95
              Median :2.700   Median :10.10
              Mean   :2.724   Mean   :10.63
              3rd Qu.:3.025   3rd Qu.:12.12
              Max.   :3.900   Max.   :20.50

postscript("cat.ps",horizontal=F)
par(mfrow=c(2,2))
boxplot(cats[,2:3])
plot(Bwt,Hwt)
out = lm(Hwt ~ Bwt,data = cats)
summary(out)
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.3567      0.6923  -0.515   0.607
Bwt          4.0341      0.2503  16.119 <2e-16 ***
---

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-Squared: 0.6466, Adjusted R-squared: 0.6441
F-statistic: 259.8 on 1 and 142 DF, p-value: < 2.2e-16

abline(out,lwd=3)
names(out)
  [1] "coefficients" "residuals" "effects" "rank"
  [5] "fitted.values" "assign" "qr" "df.residual"
```

```

[9] "xlevels"      "call"          "terms"         "model"
r = out$residuals
plot(Bwt, r, pch=19)
lines(Bwt, rep(0, length(Bwt)), lty=3, col=2, lwd=3)
qqnorm(r)
dev.off()

```

How a qq-plot works. If you are not familiar with qq-plots, read this. Order the data:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Let

$$z_j = \Phi^{-1}(j/n)$$

(Actually, we don't quite use  $j/n$  but never mind.) Plot  $X_{(j)}$  versus  $z_j$ . If  $X_i \sim N(\mu, \sigma^2)$  then this plot should a straight line with slope  $\sigma$  and intercept  $\mu$ . Why? Let

$$\hat{F}(x) = \frac{\text{number of observations } \leq x}{n}.$$

So  $\hat{F}(x) \approx F(x) = \mathbb{P}(X \leq x)$ . Note that

$$\hat{F}(X_{(j)}) = \frac{j}{n}$$

and so

$$X_{(j)} = \hat{F}^{-1}\left(\frac{j}{n}\right) \approx F^{-1}\left(\frac{j}{n}\right) \approx \sigma\Phi^{-1}\left(\frac{j}{n}\right) + \mu = \sigma z_j + \mu.$$

We used the fact that  $F^{-1}(q) = \sigma\Phi^{-1}(q) + \mu$ . Here is a proof of this fact. Let  $x_q = F^{-1}(q)$  be the  $q^{\text{th}}$  quantile. Then

$$q = F(x_q) = \mathbb{P}(X \leq x_q) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x_q - \mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{x_q - \mu}{\sigma}\right) = \Phi\left(\frac{x_q - \mu}{\sigma}\right)$$

and hence,  $F^{-1}(q) = x_q = \sigma\Phi^{-1}(q) + \mu$ .

**4.3 Example (Example: Election 2000).** Figure 3 shows the plot of votes for Buchanan (Y) versus votes for Bush (X) in Florida. The least squares estimates (omitting Palm Beach County) and the standard errors are

$$\begin{aligned} \hat{\beta}_0 &= 66.0991 & \widehat{\text{se}}(\hat{\beta}_0) &= 17.2926 \\ \hat{\beta}_1 &= 0.0035 & \widehat{\text{se}}(\hat{\beta}_1) &= 0.0002. \end{aligned}$$

The fitted line is

$$\text{Buchanan} = 66.0991 + 0.0035 \text{ Bush}.$$

Figure 3 also shows the residuals. The inferences from linear regression are most accurate when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. If we repeat the analysis replacing votes with  $\log(\text{votes})$  we get

$$\begin{aligned} \hat{\beta}_0 &= -2.3298 & \widehat{\text{se}}(\hat{\beta}_0) &= 0.3529 \\ \hat{\beta}_1 &= 0.730300 & \widehat{\text{se}}(\hat{\beta}_1) &= 0.0358. \end{aligned}$$

This gives the fit

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

The residuals look much healthier. Later, we shall address the following questions: how do we see if Palm Beach County has a statistically plausible outcome?

On the log scale, a 95 per cent confidence interval is  $.7303 \pm 2(.0358) = (.66, .80)$ . The statistic for testing  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$  is  $|Z| = |.7303 - 0|/.0358 = 20.40$  with a p-value of  $\mathbb{P}(|Z| > 20.40) \approx 0$ . This is strong evidence that that the true slope is not 0.

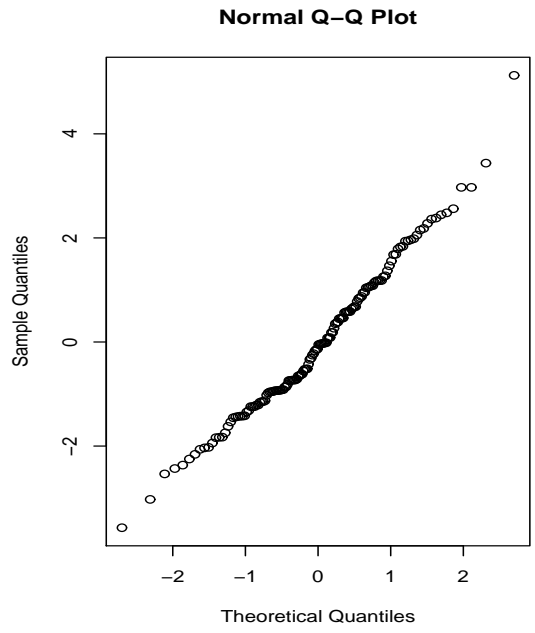
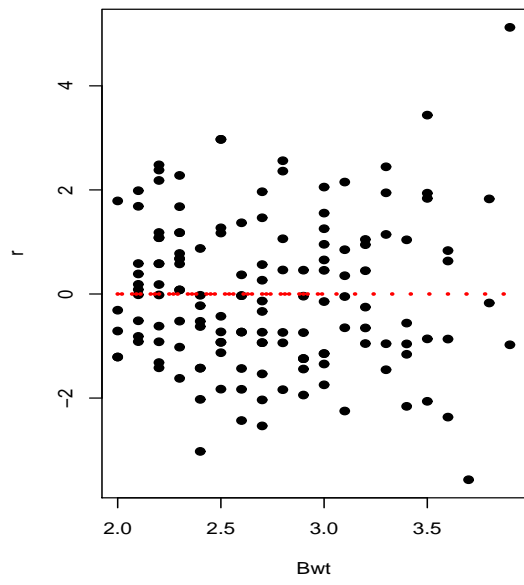
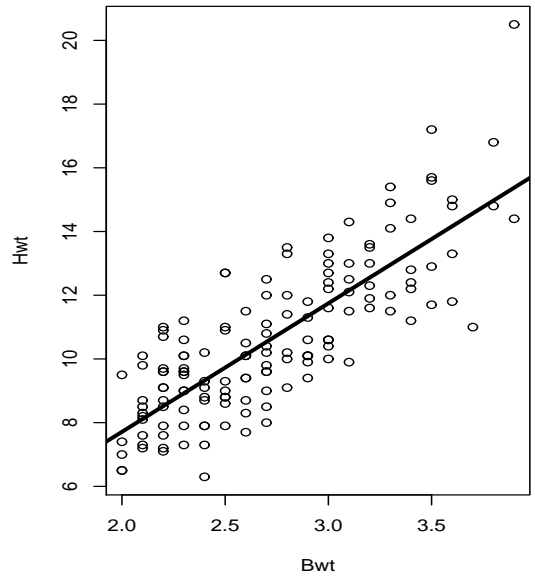
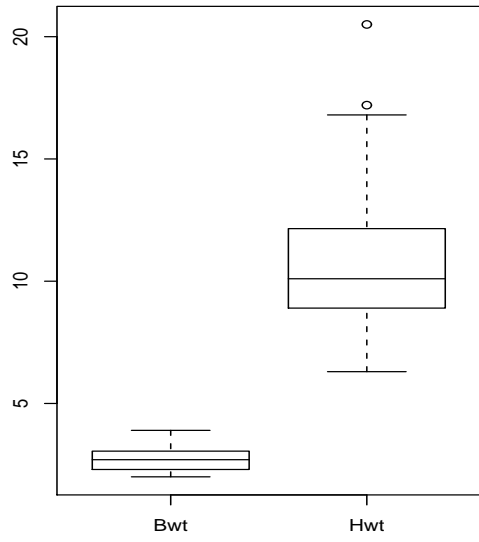


Figure 2: Cat example

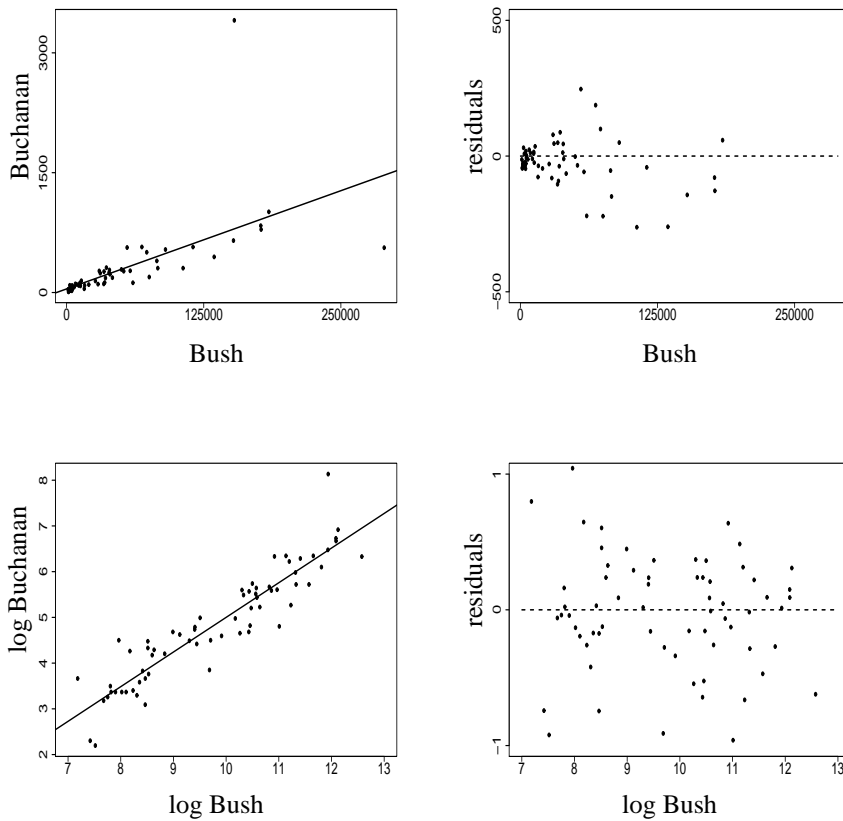


Figure 3: Voting Data for Election 2000.

## 5 ANOVA and $R^2$

In the olden days, statisticians were obsessed with summarizing things in analysis of variance (ANOVA) tables. It works like this. We can write

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$SS_{total} = RSS + SS_{reg}$$

Then we create this table:

Source	df	SS	MS	F
Regression	1	$SS_{reg}$	$SS_{reg}/1$	$MS_{reg}/MSE$
Residual	n-2	RSS	$RSS/(n-2)$	
Total	n-1	$SS_{total}$		

Under  $H_0 : \beta_1 = 0$ ,  $F \sim F_{1,n-2}$ . This is just another (equivalent) way to test this hypothesis.

The **coefficient of determination** is

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{RSS}{SS_{tot}} \quad (15)$$

Amount of variability in  $Y$  explained by  $X$ . Also,  $R^2 = r^2$  where

$$r = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (X_i - \bar{X})^2}}$$

is the **sample correlation**. This is an estimate of the **correlation**

$$\rho = \frac{\mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right)}{\sigma_X \sigma_Y}.$$

Note that

$$-1 \leq \rho \leq 1.$$

**5.1 Remark.** What happens to  $R^2$  if we move the minimum  $x_i$  further to the left and we move the maximum  $x_i$  further to the right? ■

## 6 Prediction Intervals

Given new value  $X_*$ , we want to predict

$$Y_* = \beta_0 + \beta_1 X_* + \epsilon.$$

The prediction is

$$\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 X_*. \quad (16)$$

Define

$$\text{se}_{\text{pred}}(\hat{Y}_*) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}. \quad (17)$$

A confidence interval for  $Y_*$  is

$$\hat{Y}_* \pm z_{\alpha/2} \text{se}_{\text{pred}}(\hat{Y}_*).$$

**6.1 Remark.** This is not really the standard error of the quantity  $\hat{Y}_*$ . It is the standard error of  $\hat{\beta}_0 + \hat{\beta}_1 X_* + \epsilon_*$ . Note that  $\text{se}_{\text{pred}}(\hat{Y}_*)$  does not go to 0 as  $n \rightarrow \infty$ . Why? ■

**6.2 Example (Election Data Revisited).** On the log-scale, our linear regression gives the following prediction equation:

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

In Palm Beach, Bush had 152,954 votes and Buchanan had 3,467 votes. On the log scale this is 11.93789 and 8.151045. How likely is this outcome, assuming our regression model is appropriate? Our prediction for log Buchanan votes - 2.3298 + .7303 (11.93789)=6.388441. Now, 8.151045 is bigger than 6.388441 but is “significantly” bigger? Let us compute a confidence interval. We find that  $\text{se}_{\text{pred}} = .093775$  and the approximate 95 per cent confidence interval is (6.200,6.578) which clearly excludes 8.151. Indeed, 8.151 is nearly 20 standard errors from  $\hat{Y}_*$ . Going back to the vote scale by exponentiating, the confidence interval is (493,717) compared to the actual number of votes which is 3,467. ■

## 7 Why Are We Doing This If The Model is Wrong?

The model  $Y = \beta_0 + \beta_1 x + \epsilon$  is certainly false. There is no reason why  $r(x)$  should be exactly linear. Nonetheless, the linear assumption might be adequate. But how do we assess whether the linear assumption is adequate? There are three ways.

1. We can do a **goodness-of-fit test**.
2. Second, we can do a **nonparametric regression** that does not assume linearity.
3. We can take a purely predictive point and view  $\hat{\beta}_0 + \hat{\beta}_1 x$  as an estimate of the **best linear predictor** not as an estimate of the true regression function.

We will return to these points later.

## 8 Association Versus Causation

There is much confusion about the difference between causation and association. Roughly speaking, the statement “ $X$  causes  $Y$ ” means that changing the value of  $X$  will change the distribution of  $Y$ . When  $X$  causes  $Y$ ,  $X$  and  $Y$  will be associated but the reverse is not, in general, true. Association does not necessarily imply causation.

For example, there is a strong linear relationship between death rate due to breast cancer and fat intake. So,

$$\text{RISK OF DEATH} = \beta_0 + \beta_1 \text{FAT} + \epsilon \quad (18)$$

where  $\beta_1 > 0$ . Does that mean that FAT causes breast cancer? Consider two interpretations of (18).

**ASSOCIATION** (or correlation). Fat intake and breast cancer are associated. Therefore, if I observe someone’s fat intake, I can use equation (18) to predict their chance of dying from breast cancer.

**CAUSATION**. Fat intake causes Breast cancer. Therefore, if I observe someone’s fat intake, I can use equation (18) to predict their chance of dying from breast cancer. Moreover, if I change someone’s fat intake by one unit, their risk of death from breast cancer changes by  $\beta_1$ .

If the data are from a **randomized study** ( $X$  is randomly assigned) then the causal interpretation is correct. If the data are from an **observational study**, ( $X$  is not randomly assigned) then the association interpretation is correct. To see why the causal interpretation is wrong in the observational study, suppose that people with high fat intake are the rich people. And suppose, for the sake of the example, that rich people smoke a lot. Further, suppose that smoking does cause cancer. Then it will be true that high fat intake predicts high cancer rate. But changing someone’s fat intake will not change their cancer risk.

How can we make these ideas precise? The answer is to use either **counterfactuals** or **directed acyclic graphs**.

Look at the top left plot in Figure 4. These are observed data on vitamin C ( $X$ ) and colds ( $Y$ ). You conclude that increasing vitamin C decreases colds. You tell everyone to take more vitamin C but the prevalence of colds stays the same. Why? Look at the second plot. The dotted lines show the **counterfactuals**. The counterfactual  $y_i(x)$  is the value  $Y$  person  $i$  would have had if they had taken dose  $X = x$ . Note that

$$Y_i = y_i(X_i). \quad (19)$$

In other words:

$Y_i$  is the function  $y_i(\cdot)$  evaluated at  $X_i$ .

The causal regression is the average of the counterfactual curves  $y_i(x)$ :

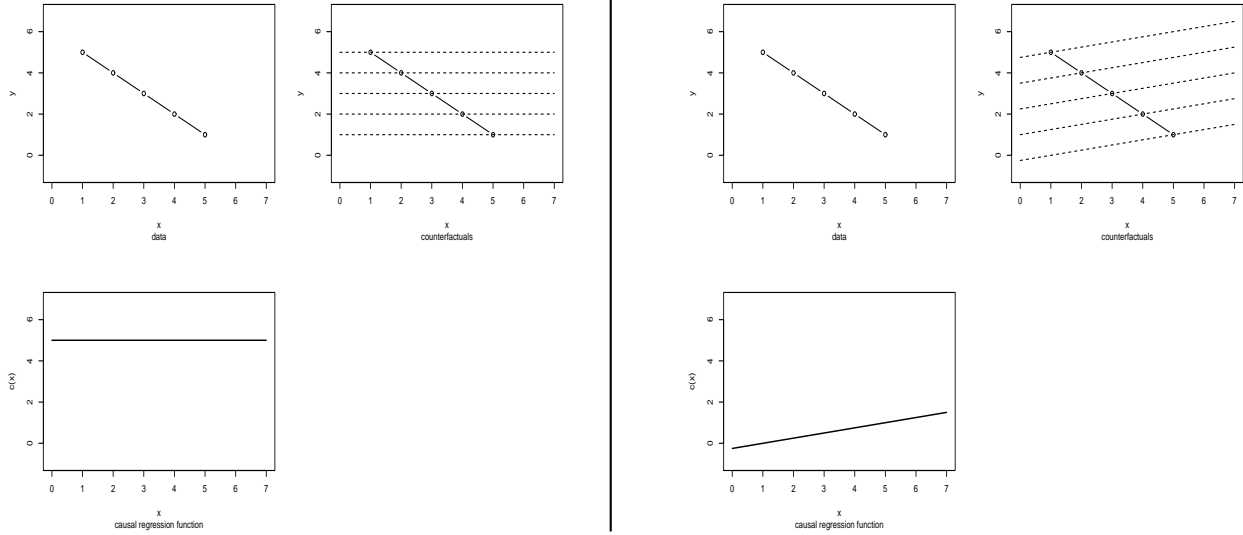


Figure 4: Causation

$$c(x) = \mathbb{E}(y_i(x)). \quad (20)$$

The average is over the population. In other words, fix a value of  $x$  then average  $y_i(x)$  over all individuals. In general:

$$r(x) \neq c(x) \quad \text{association does not equal causation} \quad (21)$$

In this example, changing everyone's dose does not change the outcome. The causal regression curve  $c(x)$  is shown in the third plot. In the second example (right side of Figure 4) it is worse. You tell everyone to take more vitamin C but the prevalence of colds increases.

Suppose that we randomly assign dose  $X$ . Then  $X_i$  is independent of the counterfactuals  $\{y_i(x) : x \in \mathbb{R}\}$ . In that case:

$$c(x) = \mathbb{E}(y(x)) \quad (22)$$

$$= \mathbb{E}(y(x)|X = x) \quad \text{since } X \text{ is indep of } \{y(x) : x \in \mathbb{R}\} \quad (23)$$

$$= \mathbb{E}(Y|X = x) \quad \text{since } Y = c(X) \quad (24)$$

$$= r(x). \quad (25)$$

Thus, if  $X$  is randomly assigned then association is equal to causation.

In an observational (non randomized) study, the best we can do is try to measure **confounding variables**. These are variables that affect both  $X$  and  $Y$ . If we can find all the confounding variables  $Z$  then  $\{y(x) : x \in \mathbb{R}\}$  is independent of  $X$  given  $Z$ . Hence,

$$c(x) = \mathbb{E}(y(x)) \quad (26)$$

$$= \int \mathbb{E}(y(x)|Z = z)f(z)dz \quad (27)$$

$$= \int \mathbb{E}(y(x)|Z = z, X = x)f(z)dz \quad \text{since } X \text{ is indep of } \{y_i(x) : x \in \mathbb{R}\}|Z \quad (28)$$

$$= \int \mathbb{E}(Y|X = x, Z = z)f(z)dz \quad (29)$$

$$= \int \left( \beta_1 x + \beta_2 z \right) f(z) dz \quad \text{if linear} \quad (30)$$

$$= \beta_1 x + \beta_2 \mathbb{E}(Z). \quad (31)$$

This is called **adjusting for confounders**. of course, we can never be sure we have included all confounders. This is why observational studies have to be treated with caution.

Note the following difference:

$$c(x) = \int \mathbb{E}(Y|Z = z, X = x)f(z)dz \quad (32)$$

$$\mathbb{E}(Y|X = x) = \int \mathbb{E}(Y|Z = z, X = x)f(z|x)dz. \quad (33)$$

## 9 Confidence Bands

**9.1 Theorem** (Scheffe, 1959). *Let*

$$I(x) = \left( \hat{r}(x) - c\hat{\sigma}, \hat{r}(x) + c\hat{\sigma} \right) \quad (34)$$

where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$c = \sqrt{2F_{\alpha,2,n-2}} \sqrt{\frac{1}{n} + \frac{(\bar{x} - x)^2}{\sum_i (x_i - \bar{x})^2}}.$$

Then,

$$\mathbb{P} \left( r(x) \in I(x) \text{ for all } x \right) \geq 1 - \alpha. \quad (35)$$

**9.2 Example.** Let us return to the cat example. The R code is:

```
library(MASS)
attach(cats)
plot(Bwt,Hwt)
out = lm(Hwt ~ Bwt,data = cats)
abline(out,lwd=3)
r = out$residuals

n      = length(Bwt)
x      = seq(min(Bwt),max(Bwt),length=1000)
d      = qf(.95,2,n-2)
beta   = out$coeff
xbar   = mean(Bwt)
ssx    = sum( (Bwt-xbar)^2 )
sigma.hat = sqrt(sum(r^2)/(n-2))

stuff = sqrt(2*d)*sqrt( (1/n) + ((x-xbar)^2/ssx))*sigma.hat
### Important: Note that these are all scalars except that x is a vector.
```

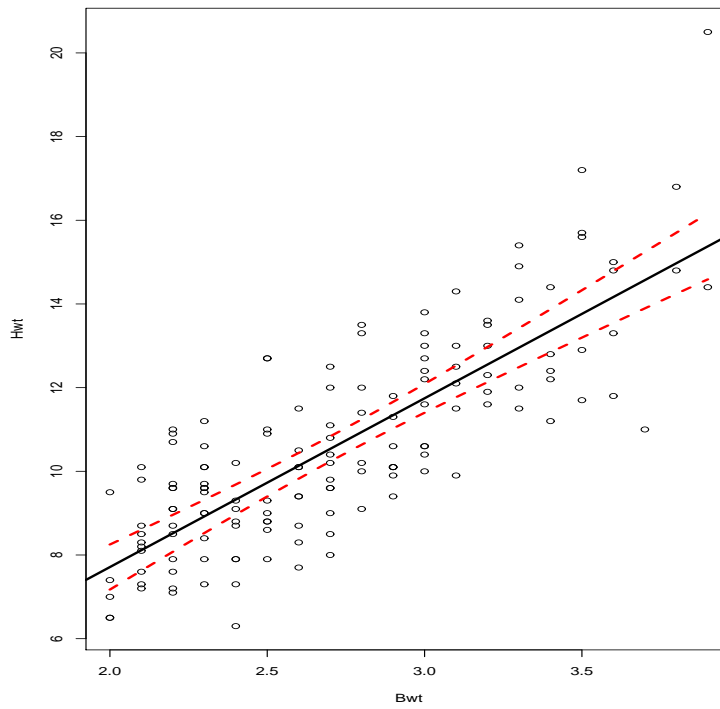


Figure 5: Confidence Band for Cat Example

```

r.hat = beta[1] + beta[2]*x
upper = r.hat + stuff
lower = r.hat - stuff

lines(x,upper,lty=2,col=2,lwd=3)
lines(x,lower,lty=2,col=2,lwd=3)

```

The bands are shown in Figure 5. ■

## 10 Review of Linear Algebra

Before starting multiple regression, we will briefly review some linear algebra. **Read pages 278-287 of Weisberg.** The **inner product** of two vectors  $x$  and  $y$  is

$$\langle x, y \rangle = x^T y = \sum_j x_j y_j.$$

Two vectors are orthogonal if  $\langle x, y \rangle = 0$ . We then write  $x \perp y$ . The **norm** of a vector is

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_j x_j^2}.$$

If  $A$  is a matrix, denote its inverse by  $A^{-1}$  and its transpose by  $A^T$ . The trace of a square matrix, denoted by  $\text{tr}(A)$  is the sum of its diagonal elements.

**PROJECTIONS.** We will make extensive use of projections. Let us start with a simple example. Let

$$e_1 = (1, 0), \quad e_2 = (0, 1)$$

and note that  $\mathbb{R}^2$  is the linear span of  $e_1$  and  $e_2$ : any vector  $(a, b) \in \mathbb{R}^2$  is a linear combination of  $e_1$  and  $e_2$ . Let

$$\mathcal{L} = \{ae_1 : a \in \mathbb{R}\}$$

be the set of vectors of the form  $(a, 0)$ . Note that  $\mathcal{L}$  is a linear subspace of  $\mathbb{R}^2$ . Given a vector  $x = (a, b) \in \mathbb{R}^2$ , the projection  $\hat{x}$  of  $x$  onto  $\mathcal{L}$  is the vector in  $\mathcal{L}$  that is closest to  $x$ . In other words,  $\hat{x}$  minimizes  $\|x - \hat{x}\|$  among all vectors in  $\mathcal{L}$ . Another characterization of  $\hat{x}$  is this: it is the unique vector such that (i)  $\hat{x} \in \mathcal{L}$  and (ii)  $x - \hat{x} \perp y$  for all  $y \in \mathcal{L}$ .

It is easy to see, in our simple example, that the projection of  $x = (a, b)$  is just  $(a, 0)$ . Note that we can write

$$\hat{x} = Px$$

where

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

This is the projection matrix.

In general, given a vector space  $\mathcal{V}$  and a linear subspace  $\mathcal{L}$  there is a projection matrix  $P$  that maps any vector  $v$  into its projection  $Pv$ .

The projection matrix satisfies these properties:

- $Pv$  exists and is unique.
- $P$  is linear: if  $a$  and  $b$  are scalars then  $P(ax + by) = aPx + bPy$ .
- $P$  is symmetric.
- $P$  is **idempotent**:  $P^2 = P$ .
- If  $x \in \mathcal{L}$  then  $Px = x$ .

Now let  $\mathbb{X}$  be some  $n \times q$  matrix and suppose that  $\mathbb{X}^T \mathbb{X}$  is invertible. The **column space** is the space  $\mathcal{L}$  of all vectors that can be obtained by taking linear combinations of the columns of  $\mathbb{X}$ . It can be shown that the projection matrix for the column space is

$$P = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T.$$

Exercise: check that  $P$  is idempotent and that if  $x \in \mathcal{L}$  then  $Px = x$ .

**RANDOM VECTORS.** Let  $Y$  be a random vector. Denote the mean vector by  $\mu$  and the covariance matrix by  $\mathbb{V}(Y)$  or  $\text{Cov}(Y)$ . If  $a$  is a vector then

$$\mathbb{E}(a^T Y) = a^T \mu, \quad \mathbb{V}(a^T Y) = a^T \Sigma a. \quad (36)$$

If  $A$  is a matrix then

$$\mathbb{E}(AY) = A\mu, \quad \mathbb{V}(AY) = A\Sigma A^T. \quad (37)$$

## 11 Multiple Linear Regression

The multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon = \beta^T X + \epsilon \quad (38)$$

where  $\beta = (\beta_0, \dots, \beta_p)^T$  and  $X = (1, X_1, \dots, X_p)^T$ . The value of the  $j^{\text{th}}$  covariate for the  $i^{\text{th}}$  subject is denoted by  $X_{ij}$ . Thus

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i. \quad (39)$$

At this point, it is convenient to use matrix notation. Let

$$\mathbb{X}_{n \times q} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

Each subject corresponds to one row. **The number of columns of  $\mathbb{X}$  will be denoted by  $q$ .** Now define,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \quad (40)$$

We can then rewrite (38) as

$$Y = \mathbb{X}\beta + \epsilon \quad (41)$$

Note that  $Y_i = X_i^T \beta + \epsilon_i$  where  $X_i^T$  is the  $i^{\text{th}}$  row of  $\mathbb{X}$ . The RSS is given by

$$\text{RSS}(\beta) = \sum (Y_i - X_i^T \beta)^2 = (Y - \mathbb{X}\beta)^T (Y - \mathbb{X}\beta). \quad (42)$$

**11.1 Theorem.** *The least squares estimator is*

$$\hat{\beta} = SY \quad (43)$$

where

$$S = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \quad (44)$$

assuming that  $(\mathbb{X}^T \mathbb{X})$  is invertible.

The fitted values are  $\hat{Y} = \mathbb{X}\hat{\beta}$  and the residuals are  $\hat{\epsilon} = Y - \hat{Y}$ . Thus,  $\text{RSS} = \|\hat{\epsilon}\|^2 = \hat{\epsilon}^T \hat{\epsilon}$ . The variance is estimated by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1} = \frac{\text{RSS}}{n-q}. \quad (45)$$

**11.2 Theorem.**  $\odot$  *The estimators satisfy the following properties.*

1.  $\mathbb{E}(\hat{\beta}) = \beta$ .
2.  $\mathbb{V}(\hat{\beta}) = \sigma^2(\mathbb{X}^T \mathbb{X})^{-1} \equiv \Sigma$ .
3.  $\hat{\beta} \approx MN(\beta, \sigma^2(\mathbb{X}^T \mathbb{X})^{-1})$ .
4. An approximate  $1 - \alpha$  confidence interval for  $\beta_j$  is

$$\hat{\beta}_j \pm z_{\alpha/2} \widehat{\text{se}}(\hat{\beta}_j) \quad (46)$$

where  $\widehat{\text{se}}(\hat{\beta}_j)$  is the square root of the appropriate diagonal element of the matrix  $\hat{\sigma}^2(\mathbb{X}^T \mathbb{X})^{-1}$ .

Let's prove the first two assertions. Note that

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(SY) = S\mathbb{E}(Y) = S\mathbb{X}\beta = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} \beta = \beta.$$

Also, since  $\mathbb{V}(Y) = \sigma^2 I$ , where  $I$  is the identity matrix,

$$\begin{aligned} \mathbb{V}(\hat{\beta}) &= \mathbb{V}(SY) = S\mathbb{V}(Y)S^T = \sigma^2 S S^T = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \left( (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \right)^T \\ &= \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}. \end{aligned}$$

The ANOVA table is

Source	df	SS	MS	F
Regression	q-1	$SS_{reg}$	$SS_{reg}/p$	$MS_{reg}/MSE$
Residual	n-q	RSS	RSS/(n-p-1)	
Total	n-1	$SS_{total}$		

The  $F$  test has  $F \sim F_{p, n-p-1}$ . This is testing the hypothesis

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

Testing this hypothesis is of limited value.

## 12 Example: Crime Data

```
### multiple linear regression crime data
x = scan("~/classes/stat707/data/crime.dat", skip=1)
x = matrix(x, ncol=14, byrow=T)
names = c("Crime", "Age", "Southern", "Education",
```

```

      "Expenditure", "Ex1", "Labor", "Males",
      "pop", "NW", "U1", "U2", "Wealth", "X")
crime.dat = as.data.frame(x)
names(crime.dat) = names

postscript("crime.ps",horizontal=F)
boxplot(crime.dat)
out      = lm(Crime ~ Age + Southern + Education +
      Expenditure + Labor + Males +
      pop + U1 + U2 + Wealth, data=crime.dat)
print(summary(out))
r = out$residuals
qqnorm(r)
dev.off()

> ### multiple linear regression crime data
> x      = scan("~/=classes/=stat707/=data/crime.dat",skip=1)
Read 658 items
> x      = matrix(x,ncol=14,byrow=T)
> names = c("Crime", "Age", "Southern", "Education",
+          "Expenditure", "Ex1", "Labor", "Males",
+          "pop", "NW", "U1", "U2", "Wealth", "X")
> crime.dat = as.data.frame(x)
> names(crime.dat) = names
>
> postscript("crime.ps",horizontal=F)
> boxplot(crime.dat)
> out      = lm(Crime ~ Age + Southern + Education +
+          Expenditure + Labor + Males +
+          pop + U1 + U2 + Wealth, data=crime.dat)
> print(summary(out))

Call:
lm(formula = Crime ~ Age + Southern + Education + Expenditure +
    Labor + Males + pop + U1 + U2 + Wealth, data = crime.dat)

Residuals:
    Min       1Q   Median       3Q      Max
-43.6447 -13.3772  0.5895  12.1430  55.4624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -589.39985   167.59057  -3.517 0.001201 **
Age           1.04058     0.44639   2.331 0.025465 *
Southern     11.29464    13.24510   0.853 0.399441
Education     1.17794     0.68181   1.728 0.092620 .
Expenditure   0.96364     0.24955   3.862 0.000451 ***
Labor         0.10604     0.15327   0.692 0.493467
Males         0.30353     0.22269   1.363 0.181344
pop           0.09042     0.13866   0.652 0.518494
U1           -0.68179     0.48079  -1.418 0.164774
U2            2.15028     0.95078   2.262 0.029859 *

```

```

Wealth      -0.08309    0.09099   -0.913  0.367229
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.56 on 36 degrees of freedom
Multiple R-Squared:  0.6845,    Adjusted R-squared:  0.5968
F-statistic: 7.809 on 10 and 36 DF,  p-value: 1.704e-06

> r = out$residuals
> qqnorm(r)
> dev.off()

```

### 13 Testing Subsets of Coefficients

Suppose you want to test if a set of coefficients is 0. Use,

$$F = \frac{(RSS_{small} - RSS_{big}) / (df_{small} - df_{big})}{RSS_{big} / df_{big}} \quad (47)$$

which has a  $F_{a,b}$  distribution under  $H_0$ , where  $a = df_{small} - df_{big}$  and  $b = df_{big}$ .

**13.1 Example.** Let's try dropping the unemployment and labor variables.

```

##Fit full model

out      = lm(Crime ~ Age + Southern + Education +
              Expenditure + Labor + Males +
              pop + U1 + U2 + Wealth, data=crime.dat)
anova(out)

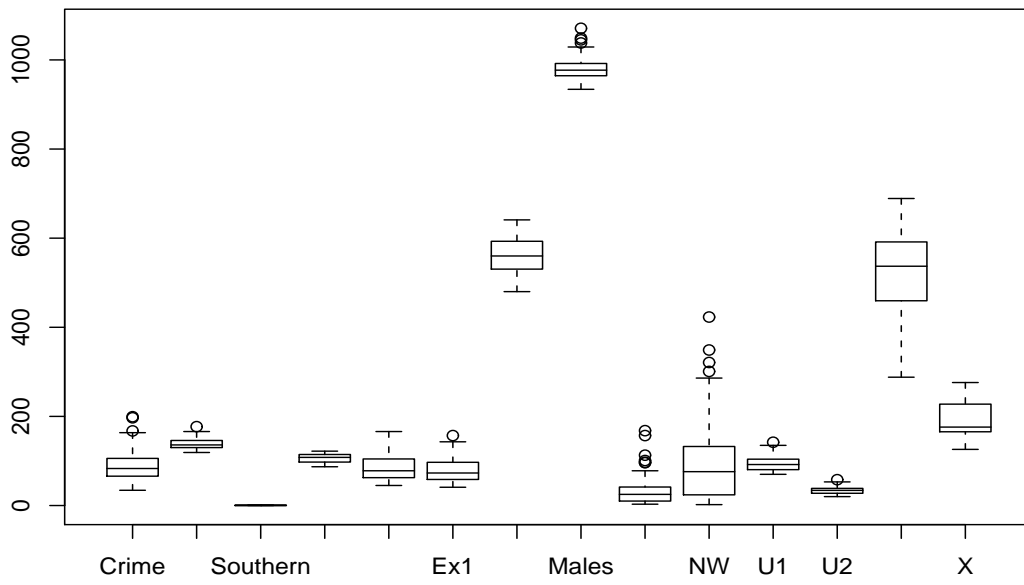
>Analysis of Variance Table
>
>Response: Crime
>
>      Df Sum Sq Mean Sq F value    Pr(>F)
>Age      1   550.8   550.8   0.9133 0.3456065
>Southern  1   153.7   153.7   0.2548 0.6167591
>Education  1  9056.7  9056.7  15.0166 0.0004333 ***
>Expenditure  1 30760.3 30760.3  51.0027 2.142e-08 ***
>Labor       1  1207.0  1207.0   2.0012 0.1657635
>Males       1  1381.5  1381.5   2.2906 0.1388888
>pop         1   528.8   528.8   0.8768 0.3553240
>U1          1   198.7   198.7   0.3295 0.5695451
>U2          1  2756.8  2756.8   4.5710 0.0393779 *
>Wealth      1   502.9   502.9   0.8339 0.3672287
>Residuals  36 21712.0   603.1

```

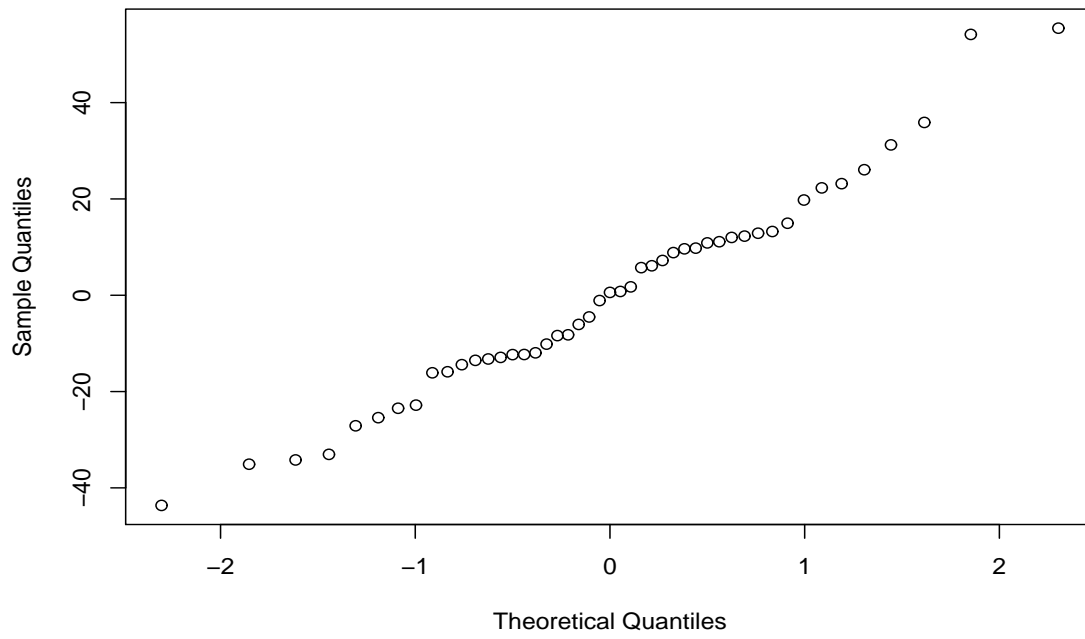
```

##drop Labor and U1 and U2
out      = lm(Crime ~ Age + Southern + Education +
              Expenditure + Males +
              pop + Wealth, data=crime.dat)

```



**Normal Q-Q Plot**



**Figure 6: crime example**

```

anova(out)
>Analysis of Variance Table
>
>Response: Crime
>
      Df  Sum Sq Mean Sq F value    Pr(>F)
>Age      1    550.8   550.8  0.8493 0.3624211
>Southern  1    153.7   153.7  0.2370 0.6291234
>Education  1   9056.7  9056.7 13.9636 0.0005963 ***
>Expenditure  1 30760.3 30760.3 47.4262 3.067e-08 ***
>Males      1   2092.7  2092.7  3.2265 0.0802032 .
>pop        1    667.6   667.6  1.0294 0.3165618
>Wealth     1    232.2   232.2  0.3581 0.5530417
>Residuals 39 25295.1   648.6

top    = (25295.1-21712)/(39-36)
bottom = 21712/36
f      = top/bottom
pvalue = 1-pf(f,3,36)
print(f)
> 1.980343
print(pvalue)
> 0.1343155

```

We conclude that these variables are not important in the regression. However, we should only do this test if there is some a priori reason to test those variables. This is not a variable selection strategy. ■

## 14 The Hat Matrix

Recall that

$$\hat{Y} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY = HY \quad (48)$$

where

$$H = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \quad (49)$$

is called the **hat** matrix. The hat matrix is the projector onto the column space of  $\mathbb{X}$ .

The residuals are

$$\hat{\epsilon} = Y - \hat{Y} = Y - HY = (I - H)Y. \quad (50)$$

The hat matrix will play an important role in all that follows.

**14.1 Theorem.** *The hat matrix has the following properties.*

1.  $H\mathbb{X} = \mathbb{X}$ .
2.  $H$  is symmetric and idempotent:  $H^2 = H$
3.  $H$  projects  $Y$  onto the column space of  $\mathbb{X}$ .
4.  $\text{rank}(\mathbb{X}) = \text{tr}(H)$ .

**14.2 Theorem.** *Properties of residuals:*

1. *True residuals:*  $\mathbb{E}(\epsilon) = 0, \mathbb{V}(\epsilon) = \sigma^2 I$ .
2. *Estimated residuals:*  $\mathbb{E}(\hat{\epsilon}) = 0, \mathbb{V}(\hat{\epsilon}) = \sigma^2(I - H)$ .
3.  $\sum_i \hat{\epsilon}_i = 0$ .
4.  $\mathbb{V}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$  where  $h_{ii}$  is diagonal element of  $H$ .

Let's prove a few of these. First,

$$\begin{aligned} \mathbb{E}(\hat{\epsilon}) &= (I - H)\mathbb{E}(Y) = (I - H)\mathbb{X}\beta \\ &= \mathbb{X}\beta - H\mathbb{X}\beta \\ &= \mathbb{X}\beta - \mathbb{X}\beta \quad \text{since } H\mathbb{X} = \mathbb{X} \\ &= 0. \end{aligned}$$

Next,

$$\begin{aligned} \mathbb{V}(\hat{\epsilon}) &= (I - H)\mathbb{V}(Y)(I - H)^T \\ &= \sigma^2(I - H)(I - H)^T \\ &= \sigma^2(I - H)(I - H) \\ &= \sigma^2(I - H - H + H^2) \\ &= \sigma^2(I - H - H + H) \quad \text{since } H^2 = H \\ &= \sigma^2(I - H). \end{aligned}$$

To see that the sum of the residuals is 0, note that  $\sum_i \hat{\epsilon}_i = \langle 1, \hat{\epsilon} \rangle$  where  $1$  denotes a vector of ones. Now,  $1 \in \mathcal{L}$ ,  $\hat{Y}$  is the projection onto  $\mathcal{L}$  and  $\hat{\epsilon} = Y - \hat{Y}$ . By the properties of the projection,  $\hat{\epsilon}$  is perpendicular to every vector in  $\mathcal{L}$ . Hence,  $\sum_i \hat{\epsilon}_i = \langle 1, \hat{\epsilon} \rangle = 0$ .

**14.3 Example.** Let

$$\mathbb{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Then

$$H = \frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

The column space is

$$\mathcal{V} = \left\{ a \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} : a \in \mathbb{R} \right\}$$

and

$$HY = \begin{pmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix} \blacksquare$$

**14.4 Example.** Suppose that the  $\mathbb{X}$  matrix has two-columns. Denote these columns by  $x_1$  and  $x_2$ . The column space is  $\mathcal{V} = \{a_1x_1 + a_2x_2 : a_1, a_2 \in \mathbb{R}\}$ . The hat matrix projects  $Y \in \mathbb{R}^n$  onto  $\mathcal{V}$ . See Figure 7. ■

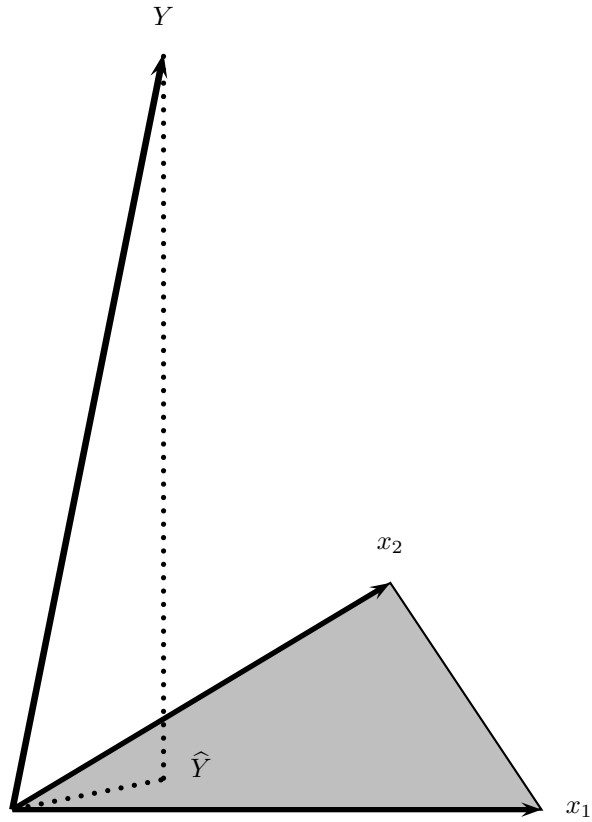


Figure 7: Projection

## 15 Weighted Least Squares

So far we have assumed that the  $\epsilon_i$ 's are independent and have the same variance. What happens if this is wrong?

Suppose that

$$Y = \mathbb{X}\beta + \epsilon$$

where

$$\mathbb{V}(\epsilon) = \Sigma.$$

Suppose we use the usual least squares estimator  $\hat{\beta}$ . Then,

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}((\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y) \\ &= (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{E}(Y) \\ &= (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} \beta = \beta. \end{aligned}$$

So  $\hat{\beta}$  is still unbiased. Also, under weak conditions, it can be shown that  $\hat{\beta}$  is consistent (converges to  $\beta$  as we get more data). The usual estimator has reasonable properties. However, there are two problems.

First, with constant variance, the usual least squares estimator is not just unbiased, it is an optimal estimator in the sense that it is they are the *minimum variance, linear, unbiased estimator*. This is no longer true with non-constant variance. **Second, and more importantly**, the formula for the standard error of  $\hat{\beta}$  is wrong. To see this, recall that  $\mathbb{V}(AY) = A\mathbb{V}(Y)A^T$ . Hence,

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}((\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y) = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{V}(Y) \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \Sigma \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1}$$

which is different than the usual formula.

It can be shown that minimum variance, linear, unbiased estimator is obtained by minimizing

$$\text{RSS}(\beta) = (Y - \mathbb{X}\beta)^T \Sigma^{-1} (Y - \mathbb{X}\beta).$$

The solution is

$$\hat{\beta} = SY \tag{51}$$

where

$$S = (\mathbb{X}^T \Sigma^{-1} \mathbb{X})^{-1} \mathbb{X}^T \Sigma^{-1}. \tag{52}$$

This is unbiased with variance

$$\mathbb{V}(\hat{\beta}) = (\mathbb{X}^T \Sigma^{-1} \mathbb{X})^{-1}.$$

This is called **weighted least squares**.

Let  $B$  denote the square root of  $\Sigma$ . Thus,  $B$  is a symmetric matrix that satisfies

$$B^T B = B B^T = \Sigma.$$

It can be shown that  $B^{-1}$  is the square root of  $\Sigma^{-1}$ . Let  $Z = B^{-1}Y$ . Then we have

$$\begin{aligned} Z &= B^{-1}Y = B^{-1}(\mathbb{X}\beta + \epsilon) \\ &= B^{-1}\mathbb{X}\beta + B^{-1}\epsilon \\ &= M\beta + \delta \end{aligned}$$

where

$$M = B^{-1}\mathbb{X}, \quad \text{and, } \delta = B^{-1}\epsilon.$$

Moreover,

$$\mathbb{V}(\delta) = B^{-1}\mathbb{V}(\epsilon)B^{-1} = B^{-1}\Sigma B^{-1} = B^{-1}B B B^{-1} = I.$$

Thus we can simply regress  $Z$  on  $M$  and do ordinary regression.

Let us look more closely at a special case. If the residuals are uncorrelated then

$$\Sigma = \begin{pmatrix} \frac{\sigma^2}{w_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{\sigma^2}{w_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & \frac{\sigma^2}{w_n} \end{pmatrix}.$$

In this case,

$$\text{RSS}(\beta) = (Y - \mathbb{X}\beta)^T \Sigma^{-1} (Y - \mathbb{X}\beta) \propto \sum_{i=1}^n w_i (Y_i - x_i^T \beta)^2.$$

Thus, in weighted least squares we are simply giving lower weight to the more variable (less precise) observations.

Now we have to address the following question: where do we get the weights? Or equivalently, how do we estimate  $\sigma_i^2 = \mathbb{V}(\epsilon_i)$ ? There are four approaches.

(1) Do a transformation to make the variances approximately equal. Then we don't need to do a weighted regression.

(2) Use external information. There are some cases where other information (besides the current data) will allow you to know (or estimate)  $\sigma_i$ . These cases are rare but they do occur. I am working on such a problem right now, in fact. It is a problem from physics and the  $\sigma_i^2$  are from instrument error which is known to a good approximation.

(3) Use replications. If there are several  $Y$  values corresponding to each  $x$  value, we can use the sample variance of those  $Y$  values to estimate  $\sigma_i^2$ . However, it is rare that you would have so many replications.

(4) Estimate  $\sigma(x)$  as a function of  $x$ . Just as we can estimate the regression line, we can also estimate the variance, thinking of it as a function of  $x$ . We could assume a simple model like

$$\sigma(x_i) = \alpha_0 + \alpha_1 x_i$$

for example. Then we could try to find a way to estimate the parameters  $\alpha_0$  and  $\alpha_1$  from the data. In fact, we will do something more ambitious. We will estimate  $\sigma(x)$  assuming only that it is a smooth function of  $x$ . We will do this later in the course when we discuss nonparametric regression.