Introduction to Linear Models: Two Variables I

The linear model is $Y = \beta_0 + \beta_1 x + \epsilon$,

and we observe pairs (x_i, Y_i). For each observation:

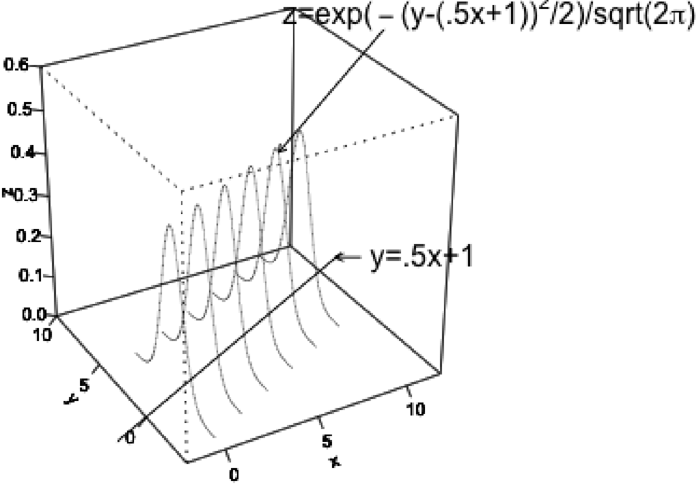
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i\ , \\ &\quad \epsilon_i \sim \text{ independent } N(0, \sigma) \qquad i = 1, 2, \ldots, n \\ \Rightarrow &\quad Y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i = \text{error or ``residual''} \end{aligned}$$

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) \quad = \quad \beta_0 + \beta_1 x_i + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

$$\begin{aligned} var(Y_i) &= var(\beta_0 + \beta_1 x_i + \epsilon_i) \quad = \quad var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

This is a considerable assumption: constant variance for the response variable (Y) for every value of predictor variable (x).

Introduction to Linear Models: Two Variables II



Introduction to Linear Models: Two Variables III

Estimate for $\mu(x)$ by the "least-squares" method: minimize the **sum of squared errors** with respect to the parameters included in the model specified for the mean:

$$\text{SSE} \;=\; \sum_{i=1}^{n} (\;Y_i \;-\widehat{Y}_i\;)^2 \qquad \text{where} \qquad \widehat{Y}_i = \widehat{\mu}(x_i) = b_0 + b_1 x_i$$

LS estimates b_0 for β_0 (the intercept) and b_1 for β_1 (the slope):

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad b_0 = \bar{Y} - b_1\overline{x}\;.$$

In the regression model, errors have distribution $\epsilon_i \sim N(0,\sigma)$. So

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i-\bar{x})^2}}\right),\;\; b_0 \sim N\left(\beta_0, \sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)}\right)$$

Introduction to Linear Models: Two Variables IV

σ is unkown: calculate an unbiased estimate of the error variance: SSE/df

An unbiased estimate for σ^2 has the form

$$\widehat{\sigma}^2 \;=\; s^2 \;=\; \frac{\mathrm{SSE}}{\mathrm{df}} \;=\; \frac{\mathrm{SSE}}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2},$$

where each **residual** $r_i = Y_i - \hat{\mu}(x_i) = Y_i - (b_0 + b_1 x_i)$.

The degrees of freedom are df $= n - 2$ since n is the total sample size, and 2 is the number of parameters estimated in the model component for the mean (β_0, β_1).

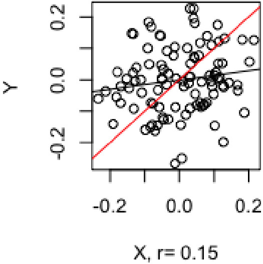
Correlation I

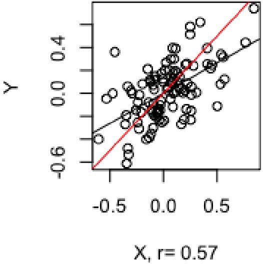
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{s_y}\right)\left(\frac{x_i-\overline{x}}{s_x}\right)\ .$$

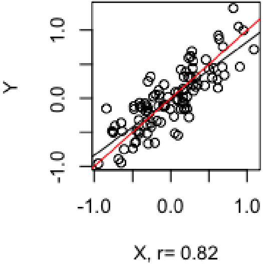
If you standardize both variables, the regression line goes through the origin and the slope is r.

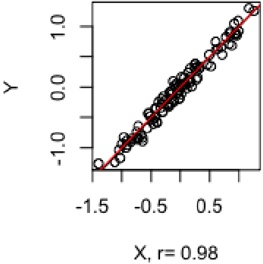
Simple algebra will yield another expression for $b_1 = r\frac{s_y}{s_x}$

Correlation II









Linear models, example I

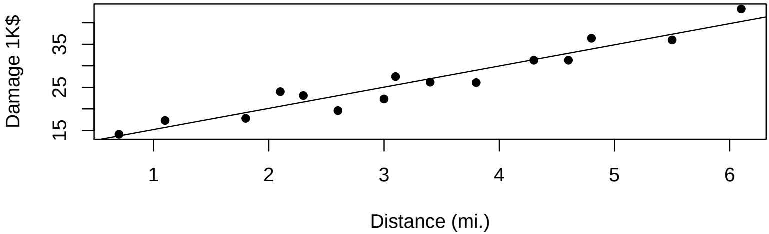
```
fire <- read.table(file="fire.dat")
names(fire) <- c("obs","dist","damage")
glimpse(fire)


## Rows: 15
## Columns: 3
## $ obs    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## $ dist   <dbl> 0.7, 1.1, 1.8, 2.1, 2.3, 2.6, 3.0, 3.1, 3.4, 3.8, 4.3, ...
## $ damage <dbl> 14.1, 17.3, 17.8, 24.0, 23.1, 19.6, 22.3, 27.5, 26.2, 2...


fire.lm <- lm(damage˜dist,data=fire) # Estimate the model
```

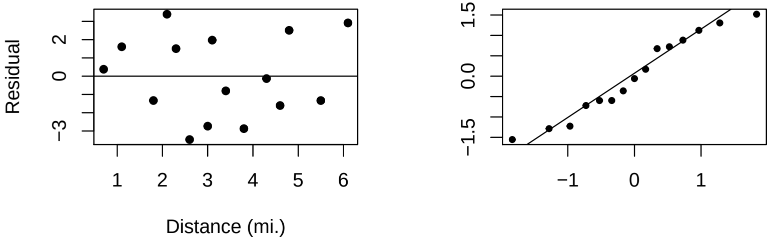
```
# Plot the data and the regression line
plot(fire$dist,fire$damage,pch=16,xlab="Distance (mi.)",
     ylab="Damage 1K$")
abline(fire.lm)
```

Linear models, example II



Linear models, example III

```r
par(mfrow=c(1,2))
plot(fire$dist,fire.lm$res,xlab="Distance (mi.)",
     ylab="Residual",pch=16)
abline(h=0)
# standardize residuals
res=(fire.lm$res-mean(fire.lm$res))/sd(fire.lm$res)
qqnorm(res,main="",pch=20,xlab="",ylab="") # Check qqplot
qqline(res)
```

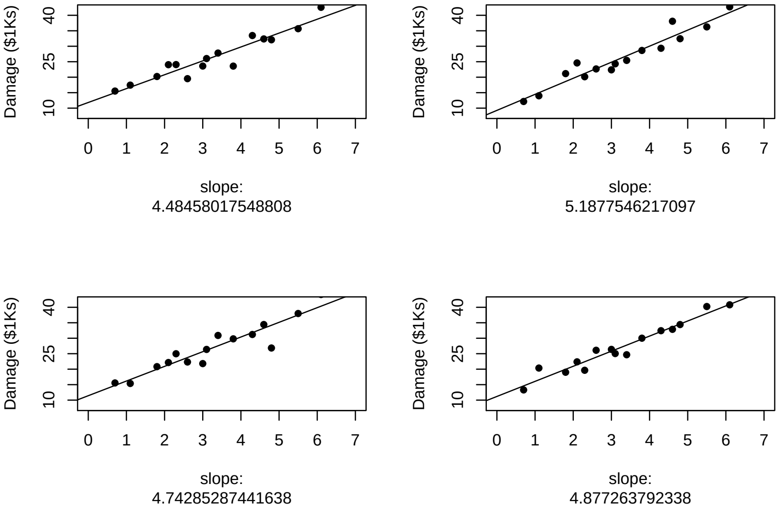


Simulate from model I

```r
sig=sigma(fire.lm) # Get estimate of sigma

N=4
nn=rnorm(N*15,0,sig)
dim(nn)=c(15,N)
nlm = list()
damage=list()
for (j in (1:N)){
damage[[j]]=fire$dist*fire.lm$coefficients[2]+
  fire.lm$coefficients[1]+nn[,j]
mod=lm(damage[[j]]~fire$dist)
nlm[[j]]=mod
}
```

Simulate from model II



Prediction - Confidence intervals I

The fitted line is:

$$\text{damage} = 10.28 + 4.92\,\text{dist}$$

Suppose we want to predict the **mean amount of damage** for fires 2 miles from the nearest fire station. In this case, $x^* = 2$ and our prediction is

$$10.28 + 4.92 \times 2 = 20.12$$

Inference about Prediction

What if we want to predict the amount of damage of a burning house which is 2 miles from the nearest fire station? Still, the prediction is:

$$10.28 + 4.92 \times 2 = 20.12$$

Prediction - Confidence intervals II

The predicted values are the same, but they have different standard errors. Individual burning houses which are 2 miles away from the fire station don't have the same amount of damage, so the prediction for individual amount of damage has larger standard error than the prediction for mean amount of damage.

Prediction - Confidence intervals III

CIs for the Mean Response

For a specific value of x, say x^*, the assumption is that y comes from a $N(\mu(x^*), \sigma)$ distribution, where

$$\mu(x^*) = \beta_0 + \beta_1 x^*$$

Plugging in our estimates of β_0 and β_1, $\mu(x^*)$ is estimated by $\hat{\mu}(x^*) = b_0 + b_1 x^*$, and a level $(1 - \alpha)$ confidence interval for the mean response $\mu(x^*)$ is given by

$$\hat{\mu}(x^*) \pm t^* \mathrm{SE}(\hat{\mu}(x^*))$$

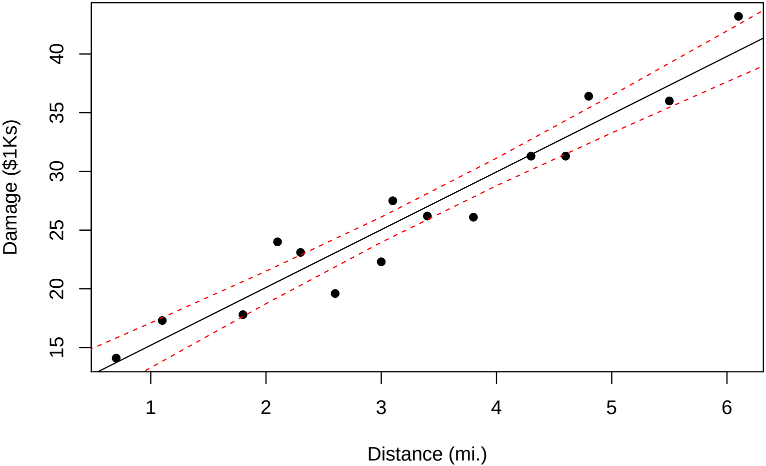
where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$\mathrm{SE}(\hat{\mu}(x^*)) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Prediction - Confidence intervals IV

```r
newdist<-seq(0,7)
# Get predictions for new values in newdist
prd<-predict(fire.lm,newdata=data.frame(dist=newdist),
             interval = c("confidence"),
             level = 0.90,type="response")
plot(fire$dist,fire$damage,pch=16,xlab="Distance (mi.)",
     ylab="Damage ($1Ks)")
abline(fire.lm) # Regression line
# Confidence bounds for prediction
lines(newdist,prd[,2],col="red",lty=2)
lines(newdist,prd[,3],col="red",lty=2)
```

Prediction - Confidence intervals V



Prediction - Confidence intervals VI

Prediction Interval for a Future Observation

Suppose we want to predict a specific observation value at $x = x^*$.
At each x^*, $y \sim N(\mu(x^*), \sigma)$ We want to predict a y drawn from this distribution.
Our best guess is the estimated mean of the distribution,

$$\hat{Y} = \hat{\mu}(x^*) = b_0 + b_1 x^*$$

How accurate is this estimate?
The error here will be larger than the error for the mean response, SE$(\hat{\mu}(x^*))$, because there is error in estimating $\mu(x^*)$ as well as error in drawing a value from the normal distribution $N(\mu(x^*), \sigma)$.

Prediction - Confidence intervals VII

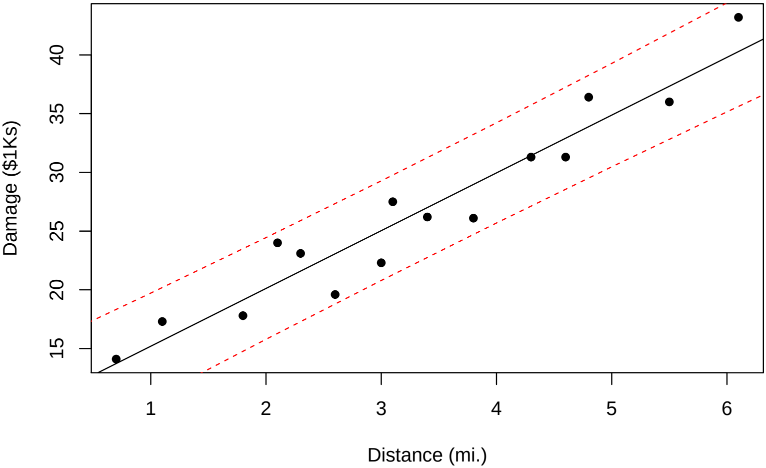
A **level** $(1-\alpha)$ **prediction interval** for a future observation y corresponding to x^* is given by

$$\hat{y} \pm t^* s_{\hat{y}}$$

where t^* is the upper $\alpha/2$ critical value of the t_{n-2} distribution and

$$s_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Prediction - Confidence intervals VIII



Analysis of variance I

Analysis of variance is the term for statistical analyses that break down the variation in data into separate pieces that correspond to different sources of variation. In the regression setting, the observed variation in the responses comes from two sources.

- As the explanatory variable x changes, it “pulls” the response with it along the regression line. This is the **variation along the line** or **regression sum of squares**:

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

- When x is held fixed, y still varies because not all individuals who share a common x have the same response y. This is the **variation about the line** or **error (residual) sum of squares**:

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

Analysis of variance II

The ANOVA Equation

It turns out that SSE and SSR together account for *all* the variation in y (i.e. S_{yy}):

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}$$

The degrees of freedom break down in a similar manner:

$$\underbrace{n-1}_{\text{SST}} = \underbrace{1}_{\text{SSR}} + \underbrace{n-2}_{\text{SSE}}$$

Dividing a sum of squares by its degrees of freedom gives a **mean square (MS)**.

Analysis of variance III

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{\mathrm{df(Error)}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = s^2$$

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_i(\hat{Y}_i - \bar{Y})^2}{\mathrm{SST}} = \frac{\hat{b}_1^2 S_{xx}}{\mathrm{SST}} = r^2$$

The fraction of the total variation in Y explained by the line

Analysis of variance IV

The ANOVA F Statistic

As an alternative test of the hypothesis: $H_0 : \beta_1 = 0$, we use the F statistic:

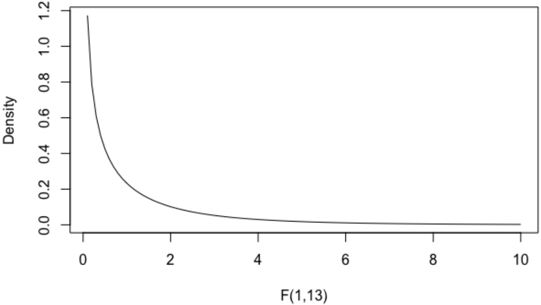
$$\begin{aligned} F &= \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR/dfR}}{\mathrm{SSE/dfE}} \\ &= \frac{b_1^2 S_{xx}}{s^2} \\ &= \left(\frac{b_1}{s/\sqrt{S_{xx}}} \right)^2 \\ &= \left(\frac{b_1}{SE(\hat{\beta}_1)} \right)^2 \\ &= t^2 \end{aligned}$$

Analysis of variance V

Under H_0,

$$F \sim F_{1,n-2}$$

where $F_{1,n-2}$ is an F distribution with 1 and $n-2$ degrees of freedom.



Analysis of variance VI

```
##
## Call:
## lm(formula = damage ~ dist, data = fire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4682 -1.4705 -0.1311  1.7915  3.3915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
## dist          4.9193     0.3927  12.525 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 13 degrees of freedom
## Multiple R-squared:  0.9235,Adjusted R-squared:  0.9176
## F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Residual standard error is the estimate s of σ.
Multiple R-squared is $\frac{\mathrm{SSR}}{\mathrm{SST}}$.
F-statistic is the square of the t-statistic for the slope.