# My Statistics Note

## Contents



# 1 Sampling Distributions Related to the Normal Distribution

**Theorem 1.1.** Let $Y_1, Y_2, ..., Y_n$ be a random sample of size $n$ from a normal distribution with mean $\mu$ and variance $\sigma^2$. Then

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$.

**Theorem 1.2.** Let $Y_1, Y_2, ..., Y_n$ be defined as in Theorem 1.1. Then $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent, standard normal random variables, $i = 1, 2, ..., n$, and

$$\sum_{i=1}^{n} Z^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2$$

has a $\chi^2$ distribution with $n$ degrees of freedom (df).

A good estimator of $\sigma^2$ is the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

**Theorem 1.3.** Let $Y_1, Y_2, ..., Y_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

has a $\chi^2$ distribution with $(n-1)$ df. Also, $\overline{Y}$ and $S^2$ are independent random variables.

**Definition 1.1.** Let $Z$ be a standard normal random variable and let $W$ be a $\chi^2-$distributed variable with $\nu$ df. Then if $Z$ and $W$ are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a $t$ distribution with $\nu$ df.

If $Y_1, ..., Y_n$ constitue a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. Then,

$$Z = \sqrt{n}(\overline{Y} - \mu)/\sigma$$

has a standard normal distribution (Theorem 1.1).

$$W = (n-1)S^2/\sigma^2$$

has a $\chi^2$ distribution with df $\nu = n - 1$ (Theorem 1.3). And $Z$ and $W$ are independent. Therefore,

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\overline{Y} - \mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{n}\left(\frac{\overline{Y} - \mu}{S}\right)$$

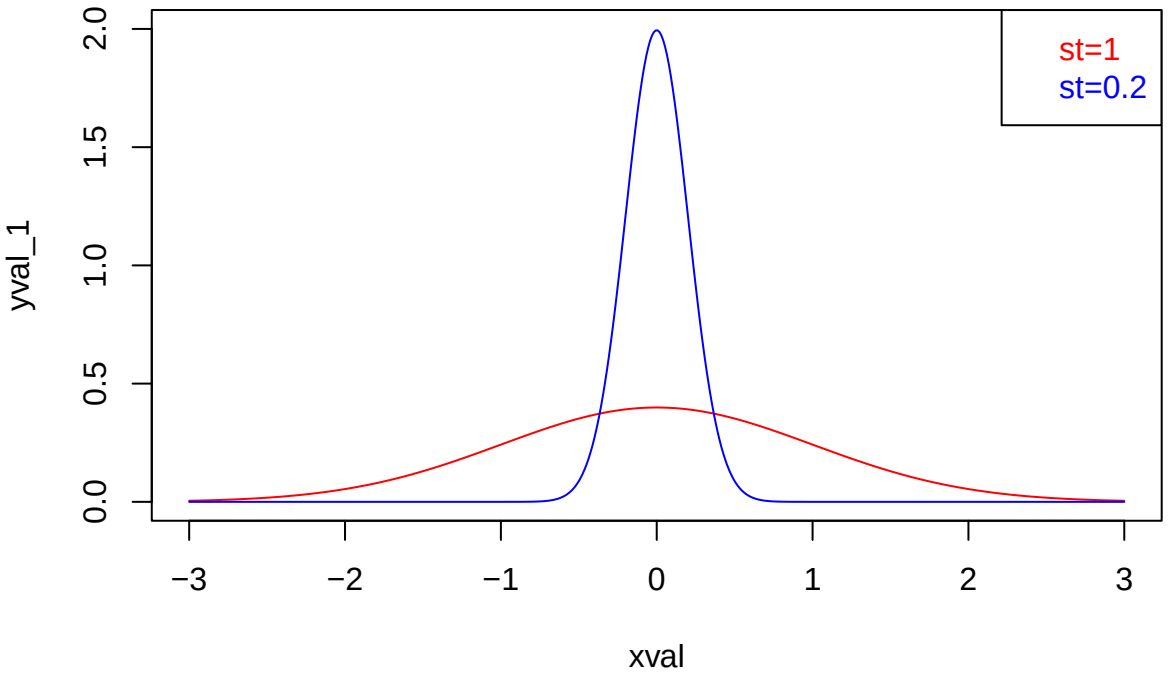
has a $t$ distribution with $(n-1)$ df.

**Definition 1.2.** $F$ distriution...

Graph of Gaussian distribution

```
xval <- seq(-3, 3, length.out = 500)
yval_1 <- dnorm(xval, mean=0, sd=1)
yval_0_2 <- dnorm(xval, mean=0, sd=0.2)
```

```
plot(xval, yval_1, type='l', ylim = c(0,2), col="red")
points(xval, yval_0_2, type='l', col="blue")
legend("topright", legend = c("st=1", "st=0.2"), text.col=c("red", "blue"))
```




```
?legend
```

# 2 Hypothesis Testing

## 2.1 Testing Proportions

### 2.1.1 Single Proportion

Consider $n$ binary trials, in which the results are success (1) and failure (0). Denote by $\pi$ the true proportion, $\hat{p}$ is the sample estimate of $\pi$. A rule of thumb: $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 5.

**One-Sample Z-Test**

$$H_0 : \pi = \pi_0$$
$$H_A : \pi \neq \pi_0$$

Test statistic with the following

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

We can assume $Z \sim N(0,1)$.

**Two Proportions**

$$H_0 : \pi_2 - \pi_1 = 0$$
$$H_A : \pi_2 - \pi_1 > 0.$$

Let $\hat{p}_1 = x1/n1$, $\hat{p}_2 = x2/n2$, the test statistic is given by

$$Z = \frac{\hat{p}_2 - \hat{p}_1 - \pi_0}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}.$$

We can treat $Z \sim N(0,1)$.

**Example 2.1.** Consider the case:

$$H_0 : \pi = 0.9$$
$$H_A : \pi < 0.9$$

$n = 89$ and $x = 71$.

1. Compute the test statistic and the $p-$value and state your conclusion for the test using a significance level of $\alpha = 0.1$.
2. Using your estimated sample proportion, construct a two-sided 90 percent confidence interval for the true proportion of women who would recommend the skin cream.

```
# Q1
prop.test(71, 89, p=0.9, alternative="less", conf.level = 0.9, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  71 out of 89, null probability 0.9
```

```
## X-squared = 10.338, df = 1, p-value = 0.0006515
## alternative hypothesis: true p is less than 0.9
## 90 percent confidence interval:
##  0.0000000 0.8466949
## sample estimates:
##         p
## 0.7977528
```

```
n <- 89
p_hat <- 71/89
pi_0 <- 0.9
Z <- (p_hat - pi_0) / (sqrt(pi_0*(1-pi_0)/n))
# The p-value = pnorm(Z)
pnorm(Z)
```

```
## [1] 0.0006514802
```

$p-$value very small; less than 0.1. There is evidence to reject $H_0$ and conclude the true proportion of women who would recommend in samples of size 89 is less than 0.9.

```
# Q2
qnorm(c(0.05, 0.95))  * sqrt(p_hat * (1-p_hat)/n) + p_hat
```

```
## [1] 0.7277190 0.8677866
```

**Example 2.2.** Consider the case:

$$x_1 = 97, \quad n_1 = 445,$$
$$x_2 = 90, \quad n_2 = 419,$$

$$H_0 : \pi_2 - \pi_1 = 0$$
$$H_A : \pi_2 - \pi_1 \neq 0.$$

```
prop.test(x=c(97, 90), n=c(445, 419), alternative = "two.sided", conf.level = 0.95, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(97, 90) out of c(445, 419)
## X-squared = 0.012871, df = 1, p-value = 0.9097
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05175429  0.05811507
## sample estimates:
##    prop 1    prop 2
## 0.2179775 0.2147971
```

```
x_1 <- 97
x_2 <- 90
n_1 <- 445
n_2 <- 419

p_hat_1 <- x_1 / n_1
p_hat_2 <- x_2 / n_2
p_star <- (x_1 + x_2) / (n_1 + n_2)
```

```r
#The 95% confidence interval
qnorm(c(0.025, 1-0.025)) * sqrt(p_star*(1-p_star)*(1/n_1+1/n_2)) + p_hat_2 - p_hat_1
```

```
## [1] -0.05812427  0.05176349
```

```r
# Write a function Z.test that can perform a one- or two-sample Z-test
Z.test <- function (p1, n1, p2=NULL, n2=NULL, p0, alternative="two.sided", conf.level=0.95){
  if (!is.null(p2) & !is.null(n2)) {
    cat("Two-sample Z-test.\n")
    p.star <- (p1 * n1 + p2 * n2) / (n1 +n2)
    z <- (p1 - p2 - p0) / sqrt(p.star*(1-p.star)*(1/n1 + 1/n2))
    ci <- qnorm(c( (1 - conf.level) / 2, 1 - (1-conf.level) / 2)) * sqrt(p.star*(1-p.star)*(1/n1 + 1/n2)
     if (alternative == "two.sided") {
       if (z<0) {
         p <- 2 * pnorm(z)
       } else if (z>=0) {
           p = 2*(1-z)
         }
       } else if (alternative == "greater") {
         p <- 1 - pnorm(z)
         } else if (alternative == "less") {
           p <- pnorm(z)
         }
  } else if (!is.null(p2) | !is.null(n2)) {
     cat("One-sample Z-test.\n")
    z <- (p1 - p0) / sqrt(p0*(1-p0) / n1)
    if (alternative == "two.sided") {
      if (z<0) {
         p <- 2 * pnorm(z)
       } else if (z>=0) {
           p = 2*(1-z)
         }
       } else if (alternative == "greater") {
         p <- 1 - pnorm(z)
         } else if (alternative == "less") {
           p <- pnorm(z)
         }

    ci <- qnorm(c( (1 - conf.level) / 2, 1 - (1-conf.level) / 2)) * sqrt(p0*(1-p0) / n1) + p1
  }
return(list(Z=z,P=p,CI=ci))
}
```

```r
x1 <- 180
n1 <- 233
p.hat1 <- x1/n1
x2 <- 175
n2 <- 197
p.hat2 <- x2/n2
```

```r
Z.test(p.hat2,n2,NULL,n1,p0=0.2,conf.level=0.95, alternative = "less") # ...or you could flip the order
```

```
## One-sample Z-test.
```

```
## $Z
## [1] 24.15275
```

```
##
## $P
## [1] 1
##
## $CI
## [1] 0.8324682 0.9441815
```

## 2.2 Testing Categorical Variables

### 2.2.1 Single Categorical Variable

Consider the following example. Suppose a researcher in sociology is interested in the dispersion of rates of facial hair in men of his local city and whether they are uniformly represented in the male population. He defines a categorical variable with three levels: clean shaven (1), beard only or moustache only (2), and beard and moustache (3). He collects data on 53 randomly selected men and finds the following outcomes:

```
hairy <- c(2,3,2,3,2,1,3,3,2,2,3,2,2,2,3,3,3,2,3,2,2,2,1,3,2,2,2,1,2,2,3,
           2,2,2,2,1,2,1,1,1,2,2,2,3,1,2,1,2,1,2,1,3,3)
```

The research question asks whether the proportions in each category are equally represented. Let $\pi_1$, $\pi_2$, and $\pi_3$ represent the true pro- portion of men in the city who fall into groups 1, 2, and 3, respectively. You therefore seek to test these hypotheses:

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$$
$$H_A : H_0 \text{ is incorrect}$$

For this test, use a standard significance level of 0.05.

**Calculation: Chi-Squared Test of Distribution**

The quantities of interest are the proportion of $n$ observations in each of $k$ categories, $\pi_1, ..., \pi_k$, for a single mutually exclusive and exhaustive categorical variable. The null hypothesis defines hypothesized null values for each proportion; label these respectively as $\pi_{0(1)}, ..., \pi_{0(k)}$. The test statistic $\chi^2$ is given as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where $O_i$ is the observed count and $E_i$ is the expected count in the $i$th cateogry, $i = 1, ..., k$.

1. $O_i$ are obtained directly from the raw data.
2. $E_i = n\pi_{0(i)}$ are merely the product of the overall samplesize $n$ with the respective null proportion for each category.

The result of $\chi^2$ follows a $\chi^2-$distribution with $\nu = k - 1$ degree of freedom.

```
n <- length(hairy)
hairy.tab <- table(hairy)
hairy.tab / n
```

```
## hairy
##         1         2         3
## 0.2075472 0.5283019 0.2641509
```

```
expected <- 1/3 * n
hairy.matrix <- cbind(1:3,hairy.tab,expected,
                      (hairy.tab-expected)^2/expected)
dimnames(hairy.matrix) <- list(c("clean","beard OR mous.",
```

```
                                   "beard AND mous."),
                                 c("i","Oi","Ei","(Oi-Ei)^2/Ei"))
hairy.matrix
```

```
##                  i Oi       Ei (Oi-Ei)^2/Ei
## clean            1 11 17.66667    2.5157233
## beard OR mous.   2 28 17.66667    6.0440252
## beard AND mous. 3 14 17.66667    0.7610063
```

```
X2 <- sum(hairy.matrix[,4])
X2
```

```
## [1] 9.320755
```

```
1-pchisq(X2, df=2)
```

```
## [1] 0.009462891
```

```
chisq.test(x=hairy.tab)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  hairy.tab
## X-squared = 9.3208, df = 2, p-value = 0.009463
```

This small $p-$value provides evidence to suggest that the true frequencies in the defined categories of male facial hair are not uniformly distributed in a $1/3, 1/3, 1/3$ fashion.

```
chisq.test(x=hairy.tab, p=c(0.25, 0.5, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  hairy.tab
## X-squared = 0.50943, df = 2, p-value = 0.7751
```

The high $p-$value suggests there is no evidence to reject $H_0$ in this scenario. In other words, there is no evidence to suggest that the proportions hypothesized in $H_0$ are incorrect.

# 3 Simple Linear Regression

## 3.1 An Example of Linear Relationship

Consider the example `survey` in `MASS`.

```
library("MASS")
library("ISLR2")
```

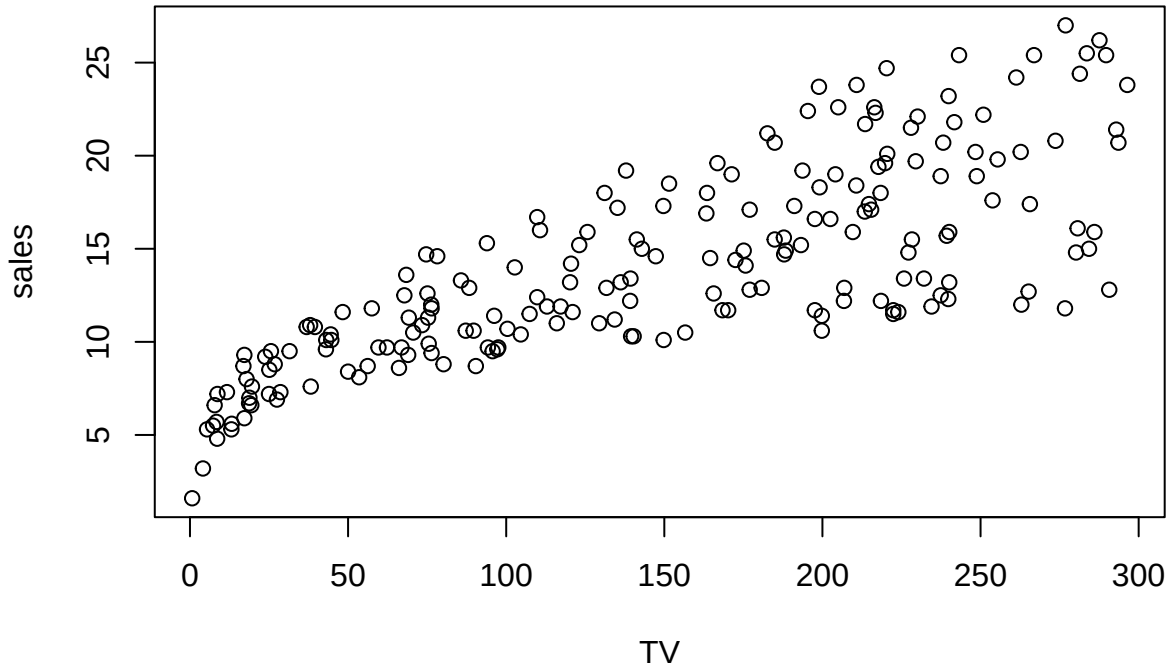
```
##
## Attaching package: 'ISLR2'
```

```
## The following object is masked from 'package:MASS':
##
##     Boston
```

```
data <- read.csv("dataset/Advertising.csv")
head(data)
```

```
##   X    TV radio newspaper sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

```
plot(data$TV, data$sales, xlab="TV", ylab="sales")
```




Consider the simple varible linear model

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

where $x$ is the virable `data$TV`, $\hat{y}$ is the virable `data$sales`.

1. The *resudual sum of squares* (RSS) is

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

Using some calculus one can show that to minimize RSS we have

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

The standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right),$$
$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\sigma$ is known as the *residual standard error* (RSE), can be estimated by

$$\text{RSE} = \sqrt{\text{RSS}/(n-2)}.$$

```
lm.fit <- lm(sales ~ TV, data = data)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

## 3.2 The coefficients:

```
x <- data$TV
y <- data$sales
x.mean <- mean(x)
y.mean <- mean(y)
beta.1 <- sum((x-x.mean)*(y-y.mean))/sum((x-x.mean)^2)
beta.0 <- mean(y) - beta.1*mean(x)
sprintf("beta.1 = %f, beta.0 = %f", beta.1, beta.0)
```

```
## [1] "beta.1 = 0.047537, beta.0 = 7.032594"
```

As is shown,

$$\hat{\beta}_0 = 7.032594,$$
$$\hat{\beta}_1 = 0.047537.$$

## 3.3 Estimating the Coefficients

```
n <- length(x)
rss <- sum((y - (beta.0 + beta.1 * x))^2)
rse <- sqrt(rss/(n-2))
se.beta.0 <- rse * sqrt(1/n + x.mean^2/sum((x-x.mean)^2))
se.beta.1 <- rse / sqrt(sum((x-x.mean)^2))
sprintf("The standard error of beta.0 is : %f.", se.beta.0)
```

```
## [1] "The standard error of beta.0 is : 0.457843."
```

```
sprintf("The standard error of beta.1 is : %f.", se.beta.1)
```

```
## [1] "The standard error of beta.1 is : 0.002691."
```

The parameters follow $t-$distributions with degrees of freedom $(n-2)$. The standarized $t$ value and $p-$value are reported for each parameter:

$$7.032594/0.457843 = 15.36,$$

$$30.047537/0.002691 = 17.67.$$

These represent the results of a two-tailed hypothesis test formally defined as

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0.$$

## 3.4 Confidence Intervals

In this case, the 95% confidence interval for $\beta_0$ is $[6.130, 7.935]$, the 95% confidence interval for $\beta_1$ is $[0.042, 0.053]$.

```
confint(lm.fit, level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

## 3.5 Assessing the Accuracy of the Model

The RSE is considered a measure of the lack of fit of the model to the data.

```
sprintf("The RSE is equal to %f.", rse)
```

```
## [1] "The RSE is equal to 3.258656."
```

## 3.6 The Multiple $R^2$

By definition,

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

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*. The multiple $R^2$ tells that about 61.2% of the variation in the TV can be attributed to sales.

```
tss <- sum((y - y.mean)^2)
1 - rss/tss
```

```
## [1] 0.6118751
```

```
rho_xy <- cor(data$sales, data$TV, use="complete.obs")
rho_xy^2
```

```
## [1] 0.6118751
```

## 3.7 Prediction

### 3.7.1 Confidence Intervals for Mean Heights

```
xvals <- data.frame(TV=c(14.5, 24))
mypred.ci <- predict(lm.fit,newdata = xvals, interval="confidence", level=0.95)
mypred.ci
```

```
##        fit      lwr      upr
## 1 7.721875 6.884587 8.559163
## 2 8.173473 7.378052 8.968894
```

### 3.7.2 Prediction Intervals for Individual Observations

```
mypred.pi <- predict(lm.fit, newdata = xvals, interval="prediction",level=0.95)
mypred.pi
```

```
##        fit      lwr      upr
## 1 7.721875 1.241430 14.20232
## 2 8.173473 1.698304 14.64864
```