**Variability with the model**

Properties of the least squares estimators

One can show  and , where I am treating  and  as random variables. This means that if many, many samples were taken from a population and the sample regression model was calculated each time, the average of the  and  values would be approximately β1 and β0. This property is often referred to as “unbiasedness”. Why is this desirable?

Also, one can show

 & 

Again, this means if many, many samples were taken from a population and the sample regression model was calculated each time, the variance of the  and  would be approximately β1 and β0. Because  usually will be unknown, we can state estimates of the these quantities as

 & 

where



This value of  is often called the mean square error (MSE), which is basically the same as it was in the previous course notes.

Notes:

*  estimates .
* The sample variance presented prevouisly was . Notice that the denominator has n – 1 in it which is actually the degrees of freedom of the numerator. The numerator had 1 estimator of μ in it which caused a “loss” of a degree of freedom. Without going into too much detail, the degrees of freedom of the numerator in  is n – 2 because it includes the estimates of β0 and β1.
* Some books will denote  by  and  by . Also, some books will refer to  as MS for residual, sample standard deviation around the regression line, the standard error of estimate, and the residual standard deviation.
* What do  and  represent?

 measures the variability of points from the
E(Y) = β0 + β1x line in the graph below.



 measures the variability of the sampled points from the  linein the graph below.



* It is also helpful to think about  as follows. Remember that Y has a normal distribution with E(Y) = β0 + β1x and Var(Y) = . For each x, we can use the normal distribution in the same way as we did when it was first introduced. For example,

P[E(Y) - 2σε < Y < E(Y) + 2σε] = 0.954

⇒ P[β0 + β1x - 2σε < Y < β0 + β1x + 2σε] = 0.954

Please remember that Y is dependent on the particular value of x here. Below is a plot illustrating how Y has a normal distribution at each x.




## Practical interpretation of :

## Rule of thumb: All observed values lie within ±2 to 3 standard deviations from the mean.

## Here,  plays the role of the mean and sε plays the role of the standard deviation

## So all of the Y values for a given x should lie between  ± 2×sε

Question: Which plot below is associated with the larger ? Note that the same x values are used in both plots.

1)

2)

If your goal was to produce a sample regression model which predicted the Y values as good as possible, which plot would you prefer?

## Example: Sales and advertising (sales\_advertising.R)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | y |  | y- | (y-)2 |
| 1 | 1 | 0.6 | 0.4 | 0.16 |
| 2 | 1 | 1.3 | -0.3 | 0.09 |
| 3 | 2 | 2 | 0 | 0 |
| 4 | 2 | 2.7 | -0.7 | 0.49 |
| 5 | 4 | 3.4 | 0.6 | 0.36 |
| Σ |  |  |  | **1.1** |



Note: 2×sε = 1.2111



For example, with x = 4,

 ± 2×sε = 2.7 ± 1.2111 = (1.49, 3.91)

For all observations in the data set with x = 4, one would expect all the corresponding Y values to be between 1.49 and 3.91.

Please see the corresponding program for the code.

Example: HS and College GPA (gpa\_regression.R, gpa.csv)

From earlier,

Call:

lm(formula = College.GPA ~ HS.GPA, data = gpa)

Residuals:

 Min 1Q Median 3Q Max

-0.55074 -0.25086 0.01633 0.24242 0.77976

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.0869 0.3666 2.965 0.008299 \*\*

HS.GPA 0.6125 0.1237 4.953 0.000103 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3437 on 18 degrees of freedom

Multiple R-squared: 0.5768, Adjusted R-squared: 0.5533

F-statistic: 24.54 on 1 and 18 DF, p-value: 0.0001027

The “Residual standard error” part of the output gives sε = 0.3437.

Below is additional code and output showing  ± 2×sε

> plot(x = gpa$HS.GPA, y = gpa$College.GPA, xlab = "HS

 GPA", ylab = "College GPA", main = "College GPA vs. HS

 GPA", xlim = c(0,4.5), ylim = c(0,4.5), col = "red",

 pch = 1, cex = 1.0, panel.first = grid(col = "gray",

 lty = "dotted"))

> curve(expr = mod.fit$coefficients[1] +

 mod.fit$coefficients[2]\*x, from = min(gpa$HS.GPA), to =

 max(gpa$HS.GPA), col = "blue", add = TRUE, n = 1000,

 lwd = 2)

> sum.fit <- summary(mod.fit)

> names(sum.fit)

 [1] "call" "terms" "residuals"

 "coefficients"

 [5] "aliased" "sigma" "df"

 "r.squared"

 [9] "adj.r.squared" "fstatistic" "cov.unscaled"

> sum.fit$sigma^2 #s^2\_epsilon

[1] 0.1181225

> low <- mod.fit$fitted.values - 2\*sum.fit$sigma

> up <- mod.fit$fitted.values + 2\*sum.fit$sigma

> segments(x0 = gpa$HS.GPA, y0 = low, x1 = gpa$HS.GPA, y1 =

 up, col = "darkgreen")



Notice that almost all the y values are within  ± 2\*sε.