10

LINEAR REGRESSION

In Chapter 6 we discussed the concepts of covariance and correlation – two ways of measuring the extent to which two random variables, X and Y were related to each other. In many cases we would like to take this a step further and try to use information from one variable to make predictions about the outcome of the other. For instance .....

10.1 SAMPLE COVARIANCE AND CORRELATION

We have so far considered summarizing a set of observations where one measurement is made on each individual or unit, but often in real-life random experiments we make multiple measurements on each individual. For example, during a health check-up a doctor might record the height, weight, age, sex, pulse rate, and blood pressure.

Just as we did for single measurements, we can represent the observed data by their empirical distribution, which is now a function of multiple arguments. For example, if we measure two random variables (X_i, Y_i) for the ith individual (say weight and blood pressure), then the empirical distribution function is given by

$$f(t,s) = \frac{1}{n}\#\{X = t, Y = s\}.$$

We can now use this to estimate population features by the corresponding feature of the empirical distribution. For example, the population covariance $Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$ gives a measure of how X and Y relate to each other. The sample version of this is the sample covariance

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}. \tag{10.1.1}$$

The sample correlation coefficient is defined similarly to population correlation coefficient $\rho[X,Y]$ as

$$r[X,Y] = \frac{S_{XY}}{S_X S_Y}, \tag{10.1.2}$$

where S_X and S_Y are the sample standard deviations of X and Y respectively. As with $\rho[X,Y]$, $r[X,Y]$ is bounded between -1 and 1, and is invariant to scale and location transformations, that is, for real numbers a, b, c, d,

$$r[aX + b, cY + d] = r[X,Y]$$

10.2 SIMPLE LINEAR MODEL

We will assume that the variable Y depends on X in a linear fashion, but that it is also affected by random factors. Specifically we will assume there is a regression line $y = \alpha + \beta x$ and that for given x-values $X_1, X_2, \ldots, X_n$ the corresponding y-values $Y_1, Y_2, \ldots, Y_n$ are given by

$$Y_j = \alpha + \beta X_j + \epsilon_j, \tag{10.2.1}$$

for $j = 1, 2, \ldots, n$ and where each of the ϵ_j are independent random variables with $\epsilon_j \sim \text{Normal}(0, \sigma^2)$. Equation (10.2.1) is referred to as the simple linear model. In particular ϵ_j are the (random) vertical distance of the point (X_j, Y_j) from the regression line. For all results below we assume $\sigma^2 > 0$ is the variance of the errors, assumed to be the same for every data point. We also assume that not all of the X_j quantities are the same so that the variance of these quantities is non-zero. In particular this means $n \geq 2$.

10.3 THE LEAST SQUARES LINE

The values of $(X_1, Y_1), \ldots, (X_n, Y_n)$ are collected data. Though we assume that this data is produced via the simple linear model, we typically do not know the actual values of the slope β or the y-intercept α. The goal of this section is to illustrate a way to estimate these values from the data.

For a line $y = a + bx$ the "residual" of a data point (X_j, Y_j) is defined to be the quantity $Y_j - (a + bX_j)$. This is the difference between the actual y-value of the data point and the location where the line predicts the y-value should be. In other words, it may be viewed as the error of the line when attempting to predict the y-value corresponding to the X_j data point. Among all possible lines through the data, there is one which minimizes the sum of these squared residual errors. This is called the "least squares line".

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be points on the plane. Suppose we wish to find a line that minimises the sum of squared residual errors. That is, let $g : \mathbb{R}^2 \to \mathbb{R}$ be defined as

$$g(a, b) = \sum_{j=1}^{n} [Y_j - (a + bX_j)]^2.$$

The objective is to minimize g. So using calculus,

$$0 = \frac{\partial g}{\partial a} = -2 \sum_{j=1}^{n} [Y_j - a - bX_j] \tag{10.3.1}$$

and

$$0 = \frac{\partial g}{\partial b} = -2 \sum_{j=1}^{n} X_j [Y_j - a - bX_j]. \tag{10.3.2}$$

From equation (10.3.1) we have

$$0 = \sum_{j=1}^{n} [Y_j - a - bX_j] = \sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} a - b \sum_{j=1}^{n} X_j = n\overline{Y} - na - bn\overline{X} = n(\overline{Y} - (a + b\overline{X}))$$

Therefore[1]

$$\overline{Y} = a + b\overline{X}, \tag{10.3.3}$$

which shows that the point $(\overline{X}, \overline{Y})$ must lie on the least squares line. The point $(\overline{X}, \overline{Y})$ is known as the point of averages. Similarly from equation (10.3.2),

$$0 = \sum_{j=1}^{n} X_j [Y_j - a - bX_j] = \sum_{j=1}^{n} (X_j Y_j - aX_j - bX_j^2) = \sum_{j=1}^{n} X_j Y_j - an\overline{X} + b \sum_{j=1}^{n} X_j^2$$

so that

$$\sum_{j=1}^{n} X_j Y_j = an\overline{X} + b \sum_{j=1}^{n} X_j^2. \tag{10.3.4}$$

We now use the system of two equations (given by (10.3.3) and (10.3.4)) solve for a, b to get

$$b = \frac{(\sum_{j=1}^{n} X_j Y_j) - n\overline{X}\,\overline{Y}}{(\sum_{j=1}^{n} X_j^2) - n\overline{X}^2} \tag{10.3.5}$$

$$\tag{10.3.6}$$

[1] We shall use the notation $\overline{X}, \overline{Y}, S_X, S_Y, r[X, Y]$(below), even though they are not necessarily random quantities. This is to simplify notation and will allow us to use known properties, in the event they are random.

Recall that the sample variance of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} S_X^2 &= \frac{1}{n-1}[\sum_{j=1}^{n}(X_j - \overline{X})^2] \\ &= \frac{1}{n-1}[\sum_{j=1}^{n} X_j^2 - 2X_j\overline{X} + \overline{X}^2] \\ &= \frac{1}{n-1}[(\sum_{j=1}^{n} X_j^2) - 2\overline{X}(\sum_{j=1}^{n} X_j) + n\overline{X}^2] \\ &= \frac{1}{n-1}[(\sum_{j=1}^{n} X_j^2) - 2n\overline{X}^2 + n\overline{X}^2] \\ &= \frac{1}{n-1}[(\sum_{j=1}^{n} X_j^2) - n\overline{X}^2] \end{aligned}$$

Therefore, the denominator of (10.3.5) is simply $(n-1)S_X^2$. The numerator may be written more simply by using the notation of sample covairance and correlation defined in (10.1.1) and (10.1.2). So from (10.3.5) we have

$$b = \frac{(\sum_{j=1}^{n} X_jY_j) - n\overline{X}\,\overline{Y}}{(\sum_{j=1}^{n} X_j^2) - n\overline{X}^2} = \frac{(n-1)S_{XY}}{(n-1)S_X^2} = \frac{r[X,Y]S_Y}{S_X}$$

Using the above and (10.3.3), we also now can write a nice formula for a , which is

$$a = \overline{Y} - \frac{r[X,Y]S_Y}{S_X}\overline{X} \tag{10.3.7}$$

By the above calcuation we have show that the least squares line minimizing the sum of the squared residual errors is the line passing through the point of averages $(\overline{X}, \overline{Y})$ and having a slope equal to $b = \dfrac{r[X,Y]S_Y}{S_X}$. We state this precisely in the Theorem below.

THEOREM 10.3.1. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be given data points. Then the least squares line passes through $(\overline{X}, \overline{Y})$ and has slope given by $\dfrac{r[X,Y]S_Y}{S_X}$.*

We illustrate the use of these formulas with two examples given below.

EXAMPLE 10.3.2. Consider the following five data points:

X	Y
3	6
4	5
5	6
6	4
7	2

These points are not colinear, but suppose we wish to find a line that most closely approximates their trend in the least squares sense described above. Viewing these as samples, it is routine to calculate that the formulas above yield $a = 9.1$ and $b = -0.9$. Of all of the lines in the plane, the one that minimizes the sum of squared residual errors for the data set above is the line $y = 9.1 - 0.9x$.

The R software also has a feature to perform a regression directly. To obtain this result using R we could first create vectors that represent the data:

```
> x <- c(3,4,5,6,7)
> y <- c(6,5,6,4,2)
```

And then instruct R to perform the regression using the command "lm" indicating the linear model.

```
> lm(y ~ x)
```

The order of the variables in this command is important with this $y \sim x$ indicating that the y variable is being predicted using the x variable as input.

The resulting output from R is

```
(Inercept)       x
       9.1    -0.9
```

the values of the intercept and slope of the least squares line respectively.

■

EXAMPLE 10.3.3. Suppose as part of a health study, a researcher collects data for weights and heights of sixty adult men in a population. The average height of the men is 174 cm with a sample standard deviation of 8.0 cm. The average weight of the men is 78 kg with a sample standard deviation of 10 kg. The correlation between the variables in the sample was 0.55.

This information alone is enough to find the least squares line for predicting weight from height. The reader may use the formulas above to verify that $b = 0.6875$ and $a = -41.625$. Therefore, among all lines, $y = -41.625 + 0.6875x$ is the one which minimizes the sum of squared residuals.

This does not necessarily mean this line would be appropriate for predicting new data points. To make such a declaration, we would want to have some evidence that the two variables had a linear relationship to begin with, but regardless of whether or not the data was produced from a simple linear model, the line above minimizes error in the least squares sense. ■

EXERCISES

Ex. 10.3.1. Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be data produced via the simple linear model and suppose $y = a + bx$ is the least squares line for the data. Recall from above that the residual for any given data point is $Y_j - (a + bX_j)$, the error the line makes in predicting the correct y-value from the given x-value. Show that the sum of the residuals over all n data points must be zero.

Ex. 10.3.2. Suppose that instead of using the simple linear model, we assume the regression line is known to pass through the origin. That is, the regression line has the from $y = \beta x$ and for given x-values $X_1, X_2, \ldots, X_n$ the corresponding y-values $Y_1, Y_2, \ldots, Y_n$ are given by

$$Y_j = \beta X_j + \epsilon_j, \tag{10.3.8}$$

for $j = 1, 2, \ldots, n$. As with the simple linear model, we assume each of the ϵ_j are independent random variables with $\epsilon_j \sim \text{Normal}(0, \sigma^2)$. (We will refer to this as the "linear model though the origin" and will have several exercises investigating how several formulas from this chapter would need to be modified for such a model.)

Assuming data $(X_1, Y_1), \ldots (X_n, Y_n)$ was produced from the linear model through the origin, find the least squares line through the origin. That is, find a formula for b such that the line $y = bx$ minimizes the sum of squared residual errors.

10.4 a AND b AS RANDOM VARIABLES

In this section (and the remainder of this chapter) we will assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ follow the simple linear model (10.2.1). In other words, there is a regression line $y = \alpha + \beta x$ and that for given x-values $X_1, X_2, \ldots, X_n$ the corresponding y-values $Y_1, Y_2, \ldots, Y_n$ are given by (10.2.1). In the previous section this data was used to produce a mean squared error-minimizing least squares line $y = a + bx$. In this section we investigate how well the random quantities a and b approximate the (unknown) values α and β.

THEOREM 10.4.1. *Under the assumptions of the simple linear model (10.2.1), the slope b of the least squares line is a linear combination of the Y_j variables. Further it has a normal distribution with mean β and variance $\frac{\sigma^2}{(n-1)S_X^2}$.*

Proof - First recall that the $X_1, X_2, X_3, \ldots X_n$ are assumed to be deterministic, so will be treated as known constants. The data points $Y_1, Y_2, \ldots, Y_n$ are assumed to follow the simple linear model (10.2.1). So for $j = 1, \ldots, n$,

$$\begin{aligned} E[Y_j] &= E[\alpha + \beta X_j + \epsilon_j] = \alpha + \beta X_j + E[\epsilon_j] = \alpha + \beta X_j \\ &\quad \text{and} \\ Var[Y_j] &= Var[\alpha + \beta X_j + \epsilon_j] = Var[\epsilon_j] = \sigma^2. \end{aligned}$$

Using the formula, (10.3.5), we derived for b and the above we have

$$\begin{aligned} E[b] &= E\left[\frac{(\sum_{j=1}^{n} X_j Y_j) - n\overline{X}\,\overline{Y}}{(n-1)S_X^2}\right] \\ &= \frac{1}{(n-1)S_X^2}\left[(\sum_{j=1}^{n} X_j E[Y_j]) - n\overline{X}E[\overline{Y}]\right] \\ &= \frac{1}{(n-1)S_X^2}\left[(\sum_{j=1}^{n} X_j(\alpha + \beta X_j)) - n\overline{X}(\alpha + \beta\overline{X})\right] \\ &= \frac{1}{(n-1)S_X^2}\left[n\alpha\overline{X} + \beta(\sum_{j=1}^{n} X_j^2) - n\alpha\overline{X} - \beta n\overline{X}^2\right] \\ &= \frac{\beta}{(n-1)S_X^2}\left[(\sum_{j=1}^{n} X_j^2) - n\overline{X}^2\right] = \beta. \end{aligned}$$

Similarly,

$$\begin{aligned} Var[b] &= Var\left[\frac{(\sum_{j=1}^{n} X_j Y_j) - n\overline{X}\,\overline{Y}}{(n-1)S_X^2}\right] \\ &= \frac{1}{[(n-1)S_X^2]^2}\left[(\sum_{j=1}^{n} X_j^2 Var[Y_j]) - n^2\overline{X}^2 Var[\overline{Y}]\right] \\ &= \frac{1}{[(n-1)S_X^2]^2}\left[(\sum_{j=1}^{n} X_j^2\sigma^2) - n^2\overline{X}^2(\sigma^2/n)\right] \\ &= \frac{\sigma^2}{[(n-1)S_X^2]^2}\left[(\sum_{j=1}^{n} X_j^2) - n\overline{X}^2\right] \\ &= \frac{\sigma^2}{(n-1)S_X^2}. \end{aligned}$$

The algebra below justifies that b is a linear combination of the Y_j variables.

$$b = \frac{(\sum_{j=1}^{n} X_j Y_j) - n\overline{X}\,\overline{Y}}{(n-1)S_X^2} = \frac{1}{(n-1)S_X^2}\left[(\sum_{j=1}^{n} X_j Y_j) - (\sum_{j=1}^{n} \overline{X} Y_j)\right] = \sum_{j=1}^{n}\left[\frac{X_j - \overline{X}}{(n-1)S_X^2}\right] Y_j$$

Since b is a linear combination of independent, normal random variables Y_j, b itself is also a normal random variable (Theorem 6.3.13). ■

As noted above, the least squares line can be defined as the line of slope b passing through the point of averages. The following lemma is a useful fact about how these quantities relate to each other.

LEMMA 10.4.2. *Let b be the slope of the least squares line and let $\overline{Y}$ be the sample average of the Y_j variables. Then b and $\overline{Y}$ are independent.*

Proof - By Theorem 6.3.13, $\overline{Y}$ has a normal distribution and so does b by Theorem 10.4.1. By Theorem 6.4.3, all we have ro show is that $\overline{Y}$ and b are uncorrelated. Note that the Y_j variables are all independent of each other and so $Cov[Y_j, Y_k]$ will be zero if $j \neq k$ and will equal the variance σ^2 otherwise. So,

$$\begin{aligned}
Cov[b, \overline{Y}] &= Cov\left[\sum_{j=1}^{n} \frac{X_j - \overline{X}}{(n-1)S_X^2} Y_j, \frac{1}{n}\sum_{k=1}^{n} Y_k\right] \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} Cov\left[\frac{X_j - \overline{X}}{(n-1)S_X^2} Y_j, \frac{1}{n} Y_k\right] \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{X_j - \overline{X}}{n(n-1)S_X^2} Cov[Y_j, Y_k] \\
&= \sum_{j=1}^{n} \frac{X_j - \overline{X}}{n(n-1)S_X^2}\sigma^2 \\
&= \frac{\sigma^2}{n(n-1)S_X^2}\sum_{j=1}^{n} X_j - \overline{X} = 0.
\end{aligned}$$

■

We conclude this section with a result on the distribution of a.

THEOREM 10.4.3. *Under the assumptions of the simple linear model (10.2.1), The y-intercept a (given by (10.3.7) of the least squares line is a linear combination of Y_j variables. Further it has a normal distribution with mean α and variance $\sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2})$.*

Proof- See Exercise 10.4.1.

EXERCISES

Ex. 10.4.1. Prove Theorem 10.4.3. (Hint: Make use of the fact that $\overline{Y} = a + b\overline{X}$ and what has previously been proven about $\overline{Y}$ and b).

Ex. 10.4.2. Show that, generally speaking, a and b are not independent. Find necessary and sufficient conditions for when the two varaibles are independent.

Ex. 10.4.3. Show that a and $\overline{Y}$ are never independent.

Ex. 10.4.4. Continuing from Exercise 10.3.2, assuming the regression line $y = \beta x$ passes through the origin and b is the lest squares line of the form $y = bx$, do the following:

(a) Find the expected value of b.

(b) Find the variance of b.

(c) Determine whether or not b has a normal distribution.

(d) Determine if b and $\overline{Y}$ are independent.

10.5 PREDICTING NEW DATA WHEN σ^2 IS KNOWN

In this section we return to question of using data for prediction. We continue to assume the simple linear model (10.2.1). We further assume that α and β are estimated by a and $b($ as calculated from the data $(X_1, Y_1), \ldots, (X_n, Y_n))$ and parameter σ^2 describing the variability of data around the regression line is a known quantity.

First suppose for a particular deterministic x-value X^* that we want to use the data to estimate the corresponding y-value $Y^* = \alpha + \beta X^*$ on the regression line by $Y = a + bX^*$.

THEOREM 10.5.1. *The quantity $Y = a + bX^*$ has a normal distribution with mean $Y^* = \alpha + \beta X^*$ and variance $\sigma^2(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{(n-1)S_X^2})$.*

Proof - Recall from Theorem 10.4.3 and Theorem 10.4.1 that a and b are both linear combination of the random variables Y_j normal distribution. So Y has normal distribution by Theorem 6.3.13. We need to calculate only its mean and variance. The expected value is simple to calculate.

$$\begin{aligned} E[Y] &= E[a + bX^*] \\ &= E[a] + E[b]X^* \\ &= \alpha + \beta X^* \ = \ Y^* \end{aligned}$$

If a and b were independent, then calculating the variance of Y would also be a simple task, but this is typically this is not the case. However, from Lemma 10.4.2, we know that b and $\overline{Y}$ are independent. To make use of this, using (10.3.3), we may rewrite the line in point-slope form around the point of averages: $Y = \overline{Y} + b(X^* - \overline{X})$. From this we have,

$$\begin{aligned} Var[Y] &= Var[\overline{Y} + b(X^* - \overline{X})] \\ &= Var[\overline{Y}] + Var[b](X^* - \overline{X})^2 \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{(n-1)S_X^2}(X^* - \overline{X})^2 \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{(n-1)S_X^2} \right). \end{aligned}$$

■

Note that for various values of X^* this variance is minimal when X^* is $\overline{X}$, the average value of the x-data. In this case $Var[Y] = \frac{\sigma^2}{n} = Var[\overline{Y}]$ as expected. The further X^* is from the average of the x-values, the more variance there is in predicting the point on the regression line.

Next suppose that, instead of trying to estimate a point on the regression line, we are trying to predict a new data point produced from the linear model. Let X^* now represent the x-value of some new data point and let $Y^* = \alpha + \beta X^* + \epsilon^*$ where $\epsilon^* \sim \text{Normal}(0, \sigma^2)$ where the random variable ϵ^* is assumed to be independent of all prior ϵ_j which produced the original data set. The following theorem addresses the distribution of the predictive error made when estimating Y^* by the quantity $Y = a + bX^*$.

THEOREM 10.5.2. *If (X^*, Y^*) is a new data point, as described in the previous paragraph, then the predictive error in estimating Y^* using the least square line is $(a + bX^*) - Y^*$ which is normally distributed with mean 0 and variance $\sigma^2(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{(n-1)S_X^2})$.*

Proof - The expected value of the predictive error is zero since

$$\begin{aligned} E[(a + bX^*) - Y^*] &= E[a] + E[b]X^* - E[\alpha + \beta X^* + \epsilon^*] \\ &= \alpha + \beta X^* - \alpha - \beta X^* - E[\epsilon^*] \ = \ 0. \end{aligned}$$

Both quantities a and b are linear combinations of the Y_j variables and so

$$\begin{aligned} a + bX^* - Y^* &= a + bX^* - \alpha - \beta X^* - \epsilon^* \\ &= (-\alpha - \beta X^*) + \text{a linear combination of } Y_1, Y_2, \ldots, Y_n, \epsilon^*. \end{aligned}$$

All $(n+1)$ of the variables, $Y_1, Y_2, \ldots, Y_n, \epsilon^*$, are independent and have a normal distribution. As $(-\alpha - \beta X^*)$ is a constant, from the above $(a + bX^* - Y^*)$ has a normal distribution.

Finally, to calculate the varaince, we again rewrite $a + bX^*$ in point-slope form and exploit independence.

$$\begin{aligned} Var[(a + bX^* - Y^*)] &= Var[(\overline{Y} + b(X^* - \overline{X}) - (\alpha + \beta X^* + \epsilon^*)] \\ &= Var[\overline{Y}] + Var[b](X^* - \overline{X})^2 + Var[\epsilon^*] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{(n-1)S_X^2}(X^* - \overline{X})^2 + \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{(n-1)S_X^2}\right). \end{aligned}$$

Example 10.5.3. A mathematics professor at a large university is studying the relationship between scores on a preparation assessment quiz students take on the first day of class and their actual pecentage score at the end of class. Assuming the simple linear model with $\sigma = 6$, he takes a random sample of 30 students and discovers their average score on the quiz is $\overline{X} = 54$ with a sample standard deviation of $S_X = 12$, while the averae percentage score in the class is $\overline{Y} = 68$ with a sample standard deviation of $S_Y = 10$. The sample correlation is $r[X, Y] = 0.6$. So according to the results above, the least squares line for predicting the course percentage from the preliminary quiz will be $y = 0.5x + 41$.

If we wish to use the line to predict the course percentage for someone who scores a 54 on the preliminary quiz, we would find $y = 0.5(54) + 41 = 68$, as expected since somone who gets an average score on the quiz is likely to get around the average percentage in the class.

Similiarly if we wish to use the line to predice the course percentage for someone who scores a 80 on the preliminary quiz, we would find $y = 0.5(80) + 41 = 81$. Also not surprising. Due to the positive correlation, a student scoring above average on the quiz is also likely to score higher in the course as well.

The previous theorem allows us to go further an calculate a standard deviation associated with these estimates. For the student who scores a 54 on the preliminary quiz, let Y^* be the actual course percentage and let $a + bX^* = 68$ be the least squares line estimate we made above. Then,

$$Var[a + bX^* - Y^*] = 36(1 + \frac{1}{30} + 0) = 37.2$$

and so the standard deviation in the predictive error is $SD[a + bX^* - Y^*] \approx 6.1$. This means that students who make an average score of 54 on the preliminary quiz will have a range of percentages in the course. This range will have a normal distribution with mean 68 and standard deviation 6.1. We could then use normal curve computations to make further predictions about how likely such a student may be to reach a certain benchmark.

Next take the example of a student who scores 80 on the preliminary quiz. The least squares line predicts the course percentage for such a student will be $a + bX^* = 81$, but now

$$Var[a + bX^* - Y^*] = 36(1 + \frac{1}{30} + \frac{(80 - 54)^2}{29 \cdot 12^2}) \approx 43.0$$

and so $SD[a + bX^* - Y^*] \approx 6.6$. Student who score an 80 on the preliminary exam will have a range of course percentages with a normal distribution of mean 81 and standard deviation 6.6.

Thinking of the standard deviation as the likely error associated with prediction this example suggests that predictions of data further from the mean will tend to have less accuracy than predictions near to the mean. This is true in the simple linear model and will be explored in the exercises. ■

EXERCISES

Ex. 10.5.1. Using the figures from Example 10.5.3 do the following. Two students are selected independently at random. The first scored a 50 on the preliminary quiz while the second scored 60. Determine how likely it is that the student who scored the lower grade on the quiz will score a higher percentage in the course.

Ex. 10.5.2. Explain why $Var[a + bX^* - Y^*]$ is minimized when $X^* = \overline{X}$.

10.6 HYPOTHESIS TESTING AND REGRESSION

As a and b both have a normal distriubtion under the assumption of the simple linear model, it is possible to perform tests of significance concerning the values of α and β. Of particular importance is a test with a null hypothesis that $\beta = 0$ and an alternate hypothesis $\beta \neq 0$. This is commonly called a "test of utility". The reason for this name is that if $\beta = 0$, then the simple linear model produces output values $Y_j = \alpha + \epsilon_j$ which do not depend on the corresponding input X_j. Therefore knowing the value of X_j should not be at all helpful in predicting the corresponding Y_j result. However, if $\beta \neq 0$ then knowing X_j should be at least somewhat useful in predicting Y_j value.

EXAMPLE 10.6.1. Suppose $(X_1, Y_1), \ldots, (X_{16}, Y_{16})$ follows the simple linear model with $\sigma = 5$ and produces a least squares line $y = 0.3 + 1.1x$. Suppose the sample average of the X_j data is 20 and the sample variance is $S_X^2 = 10$. What is the conclusion of a test of utility at a significance level of $\alpha = 0.05$? ■

From the given least squares line, $b = 1.1$. As noted above, a test of utility compares a null hypothesis that $\beta = 0$ to an alternate hypothesis $\beta \neq 0$, so this will be a two-tailed test. If the null were true, then $E[b] = 0$ and we can use the normal distribution to determine whether the 1.1 value is so far from zero that the null seems unreasonable. Using the same sample mimicing idea introduced in Chapter 9 we let $Z_1, \ldots, Z_{16}$ be random variables produced from $X_1, \ldots X_{16}$ via the simple linear model. From Theorem 10.4.1, the slope of the least squares line for the $(X_1, Z_1), \ldots, (X_{16}, Z_{16})$ data has a normal distribution with mean $\beta = 0$ and variance $\frac{\sigma^2}{(n-1)S_X^2} = \frac{1}{6}$. Therefore we can calculate

$$\begin{aligned} P(|\text{ slope of the least squares line }| \geq 1.1) &= P(|Z| \geq \frac{1.1}{\sqrt{1/6}}) \\ &= 2P(Z < -\frac{1.1}{\sqrt{1/6}}) \approx 0.007 \end{aligned}$$

where $Z \sim Normal(0, 1)$. As this P-value is less than the significance level, the test rejects the null hypothesis. That is, the test concludes that the slope of 1.1 is far enough from 0 that it demonstrates a true relationship between the X_j input values and the Y_j output values.

EXERCISES

Ex. 10.6.1. Continuing with Example 10.6.1, use Theorem 10.4.3 to devise a hypothesis test for determining whether or not the regression line goes through the origin. That is, determine whether or not $\alpha = 0$ is a plausible assumption.

10.7 ESTIMAING AN UNKNOWN σ^2

In many cases the variance σ^2 of the points around the regression line will be an unknown quantity and so, like α and β, it too will need to be approximted using the $(X_1, Y_1), \ldots, (X_n, Y_n)$ data. The following theorem provides an unbiased estimator for σ using the data.

THEOREM 10.7.1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be data following the simple linear model with $n > 2$. Let $S^2 = \frac{1}{n-2} \sum_{j=1}^{n} (Y_j - (a + bX_j))^2$. Then S^2 is an unbiased estimator for σ^2. (That is, $E[S^2] = \sigma^2$).*

Proof - Before looking at $E[S^2]$ in its entirety, we look at three quantities that will be helpful in computing this expected value.

First note,

$$\begin{aligned} Var[(Y_j - \overline{Y})] &= Var[\frac{nY_j + (Y_1 + Y_2 + \cdots + Y_n)}{n}] \\ &= \frac{1}{n^2}\left(Var[(n-1)Y_j + \sum_{i=1, i\neq j}^{n} Y_i]\right) \\ &= \frac{1}{n^2}\left([(n-1)^2\sigma^2 + \sum_{i=1, i\neq j} \sigma^2]\right) \\ &= \frac{1}{n^2}[(n-1)^2\sigma^2 + (n-1)\sigma^2] \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

and therefore,

$$\begin{aligned} \sum_{j=1}^{n} E[(Y_j - \overline{Y})^2] &= \sum_{j=1}^{n} Var[Y_j - \overline{Y}] + (E[Y_j - \overline{Y}])^2 \\ &= \sum_{j=1}^{n} \frac{n-1}{n}\sigma^2 + ((\alpha + \beta X_j) - (\alpha + \beta\overline{X}))^2 \\ &= \sum_{j=1}^{n} \frac{n-1}{n}\sigma^2 + \beta^2(X_j - \overline{X})^2 \\ &= (n-1)\sigma^2 + \beta^2 \sum_{j=1}^{n}(X_j - \overline{X})^2 \\ &= (n-1)\sigma^2 + \beta^2(n-1)S_X^2. \end{aligned} \tag{10.7.1}$$

Next,

$$\begin{aligned} \sum_{j=1}^{n} E[b^2(X_j - \overline{X})^2] &= E[b^2] \sum_{j=1}^{n}(X_j - \overline{X})^2 \\ &= (Var[b] + (E[b])^2)((n-1)S_X^2) \\ &= (\frac{\sigma^2}{(n-1)S_X^2} + \beta^2)((n-1)S_X^2) \\ &= \sigma^2 + \beta^2(n-1)S_X^2. \end{aligned} \tag{10.7.2}$$

Also,

$$\begin{aligned} E[bY_j] &= Cov[b, Y_j] + E[b]E[Y_j] \\ &= Cov[\sum_{i=1}^{n} \frac{X_i - \overline{X}}{(n-1)S_X^2} Y_i, Y_j] + \beta(\alpha + \beta X_j) \\ &= \sum_{i=1}^{n} \frac{X_i - \overline{X}}{(n-1)S_X^2} Cov[Y_i, Y_j] + \beta(\alpha + \beta X_j) \\ &= \frac{X_i - \overline{X}}{(n-1)S_X^2} Var[Y_j] + \beta(\alpha + \beta X_j) \\ &= \frac{X_i - \overline{X}}{(n-1)S_X^2} \sigma^2 + \beta(\alpha + \beta X_j) \end{aligned}$$

from which we may determine that

$$\begin{aligned} \sum_{j=1}^{n} E[(Y_j - \overline{Y})b(X_j - \overline{X})] &= \sum_{j=1}^{n} (X_j - \overline{X})E[Y_j b] - \sum_{j=1}^{n} (X_j - \overline{X})E[\overline{Y} b] \\ &= \sum_{j=1}^{n}\sum_{j=1}^{n} (X_j - \overline{X})(\frac{X_i - \overline{X}}{(n-1)S_X^2}\sigma^2 + \beta(\alpha + \beta X_j)) - \sum_{j=1}^{n} (X_j - \overline{X})E[\overline{Y}]E[b] \\ &= \sum_{j=1}^{n} \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\sigma^2 + \sum_{j=1}^{n} (X_j - \overline{X})\beta(\alpha + \beta X_j) - \sum_{j=1}^{n} (X_j - \overline{X})(\alpha + \beta\overline{X})\beta \\ &= \sigma^2 + \sum_{j=1}^{n} (X_j - \overline{X})\beta^2(X_j - \overline{X}) \\ &= \sigma^2 + \beta^2(n-1)S_X^2 \end{aligned} \tag{10.7.3}$$

Finally, putting together the results from equations 10.7.1, 10.7.2, and 10.7.3 we find

$$\begin{aligned} E[\sum_{j=1}^{n} (Y_j - (a + bX_j))^2] &= E[\sum_{j=1}^{n} (Y_j - (\overline{Y} + b(X_j - \overline{X})))] \\ &= E[\sum_{j=1}^{n} ((Y_j - \overline{Y}) - b(X_j - \overline{X}))^2] \\ &= E[\sum_{j=1}^{n} (Y_j - \overline{Y})^2 - 2(Y_j - \overline{Y})b(X_j - \overline{X}) + b^2(X_j - \overline{X})^2] \\ &= \sum_{j=1}^{n} E[(Y_j - \overline{Y})^2] - 2E[(Y_j - \overline{Y})b(X_j - \overline{X})] + E[b^2(X_j - \overline{X})^2] \\ &= \left((n-1)\sigma^2 + \beta^2(n-1)S_X^2\right) - 2\left(\sigma^2 + \beta^2(n-1)S_X^2\right) + \left(\sigma^2 + \beta^2(n-1)S_X^2\right) \\ &= (n-2)\sigma^2 \end{aligned}$$

Hence $E[S_X^2] = E[\frac{1}{n-2}\sum_{j=1}^{n} (Y_j - (a + bX_j))^2] = \sigma^2$ as desired.