

CHAPTER 10

INFERENCE FOR REGRESSION

In Stat I we saw that we can use a straight line to describe the linear relationship between two variables. A function is a mathematical relationship that allows us to predict what values of one variable (Y) correspond to given values of another variable (X).

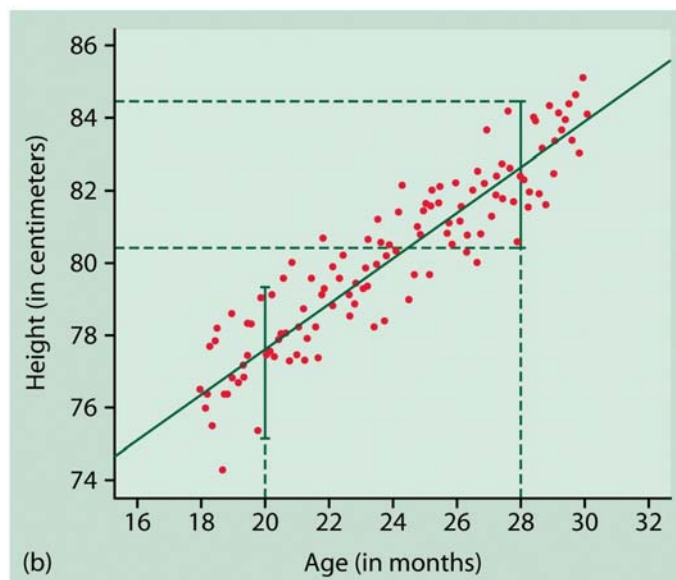
Y: is referred to as the dependent, response or predicted variable.

X: is referred to as the independent, explanatory or predictor variable.

Some examples where Y is a function of X:

- Cholesterol level and Drug dosage
- GPA and average number of hours studied per week
- Braking time and Speed of the car
- Blood Alcohol Content and number of beers consumed in previous hour

When we studied these types of relationships in Stat I, we fit a least-squares regression line and chose the line which minimized the sum of the squared distances from the data points to the predicted points.



SIMPLE LINEAR REGRESSION

In order to be more accurate, when using an equation of a line to describe the true relationship between x and y , we will use an equation that includes the possibility of error. Instead of deterministic models, we will use **probabilistic models**, which will essentially be made by using the deterministic model and allowing for error. In other words, the equation is expressed by **DATA = FIT + RESIDUAL**, where fit is the equation of the line like that given in Stat I and residual is considered to represent the deviations of the data from that line.

The equation for a probabilistic model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Here the expected value of y (or the mean response of y) is

$$\mu_y = E(y) = \beta_0 + \beta_1 x_i.$$

The deviations ε_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ .

This model represents the true and accurate relationship between a response and predictor (or explanatory) variable for a given population. However, just like problems dealing with the true population mean or proportion, we don't have enough time or resources to investigate the true population relationship. **So we pick a SRS and use the simple linear regression line of our sample to estimate the true simple linear regression line of our population.**

For example, suppose we wanted to determine the linear relationship between weight and height of Americans. Obviously we wouldn't be able to get information from every single American, therefore like we did with problems about the mean or proportion of a population, we pick an SRS of Americans and we record their weights along with their heights. We then use the methods of simple linear regression (SLR) to determine the relationship between height and weight for our sample. The equation line that we would get for our sample using simple linear regression would be much more accurate than the line produced by least-squares methods since it would account for some unknown errors and people who did not follow the pattern exactly. Once we have the relationship between height and weight for our

sample, we use it to estimate the true relationship between height and weight for all Americans.

As with the case of the population mean, μ , we use Greek letters to symbolize our **population parameters** for simple linear regression.

They are:

β_0 - the population intercept,

β_1 - the population slope and

σ the population standard deviation of the residuals.

As with the sample mean, \bar{X} we use Latin letters to symbolize our **sample statistics**:

b_0 - the sample intercept,

b_1 - the sample slope and

s - the sample standard deviation of the residuals.

Example #1 (Exercise 10.1)

Returns on common stocks in the United States and overseas appear to be growing more closely correlated as economies become more interdependent. Suppose that this population regression line connects the total annual returns (in percent) of two indexes of stock prices:

$$\text{MEAN OVERSEAS RETURN} = 4.7 + 0.66 \times \text{US RETURN}$$

a) What is β_0 in this line? What does this number say about overseas returns when the US market is flat (0% return)?

b) What is β_1 in this line? What does this number say about the relationship between US and overseas returns?

c) We know that overseas returns will vary during years that have the same return on US common stocks. Write the regression model based on the population regression line given above. What part of this model allows overseas returns to vary when US returns remain the same?

ESTIMATING THE REGRESSION PARAMETERS

Since our model for simple linear regression is made up of our least-squares equation and random error, the equations for our sample statistics, b_0 and b_1 , are the same as those we used in Stat I.

$$\hat{y} = b_0 + b_1x$$

$$b_1 = r \frac{s_y}{s_x} \quad b_0 = \bar{y} - b_1\bar{x}$$

$$\hat{\mu} = b_0 + b_1x$$

Where r is the correlation between y and x and s_y is the sample standard deviation of y and s_x is the sample standard deviation of x . \hat{y} is the predicted value of y for a given x and $\hat{\mu}$ is the predicted value of the mean response (or mean/expected value of y) for a given value of x .

We use ϵ as the symbol of the deviations in our model, but when we talk about residuals we use e as their symbol.

The equation of the **residuals** is **$e_i = \text{observed response} - \text{predicted response}$**

$$= y_i - \hat{y}_i$$

$$= y_i - b_0 - b_1x_i$$

The residuals always sum to 0, $\sum e_i = 0$

The remaining parameter to be estimated is σ , which measures the variation of y about the regression line or the standard deviation of the model deviations. Therefore, not surprisingly, we use the standard deviation of the residuals to estimate it.

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \quad \& \quad s = \sqrt{s^2}$$

STANDARD ERRORS FOR REGRESSION ESTIMATES

We use the term standard error instead of standard deviation when our formula for a given estimate involves an estimate of the population standard deviation, that is the formula uses s .

If s is the estimated standard deviation about our regression line, then

$SE_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$ is the standard error of the **intercept (b_0)** of our LSR line.

$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$ is the standard error of the **slope (b_1)** of our LSR line.

$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$ is the standard error of the **mean response ($\hat{\mu}$)** for a given x^*

$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$ is the standard error of a **predicted y (\hat{y})** for a given x^*

CONFIDENCE INTERVALS

In order to calculate a confidence interval for a population parameter we need 3 things: its estimate, its t score (we use the t distribution instead of the z distribution since our population standard deviation is unknown and estimated by the sample standard deviation) and the standard error of the estimate.

A level C CI: **estimate** $\pm t^*$ (SE_{estimate}), where t^* is the value for the t distribution with n-2 degrees of freedom and that has an area of a between $-t^*$ and t^* .

Level C Confidence intervals for predicted values of \hat{y} are instead called Prediction Intervals, (PI).

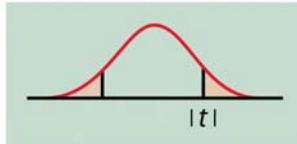
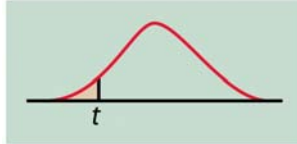
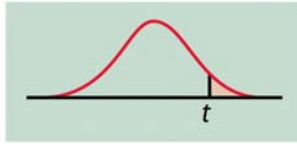
$$\text{CI for } \beta_0: \quad b_0 \pm t^* SE_{b_0} \quad \text{CI for } \beta_1: \quad b_1 \pm t^* SE_{b_1}$$

$$\text{CI for } \hat{\mu}: \quad \hat{\mu} \pm t^* SE_{\hat{\mu}} \quad \text{PI for } \hat{y}: \quad \hat{y} \pm t^* SE_{\hat{y}}$$

SIGNIFICANCE TESTS

If y does not vary with x, that is y remains constant no matter the value of x, then the equation for y does not contain a slope. Therefore, if we suspect that y does not vary with x, we can perform a significance test to determine if a slope is a necessary part of our equation. To do so, we perform a significance test of $H_0: \beta_1 = 0$. If we fail to reject our null hypothesis, then we can conclude that the true relationship between our response and predictor is constant, i.e. $\mu_y = E(y) = \beta_0$.

To test the hypothesis $H_0: \beta_1 = 0$ we use the test statistic $t = \frac{b_1}{SE_{b_1}}$, where t has the t distribution with $n-2$ degrees of freedom.



p-values for t

For a test of $H_0: \beta_1 = 0$ against:

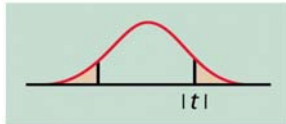
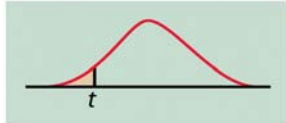
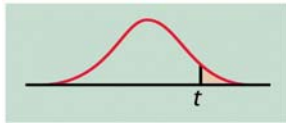
- i) $H_a: \beta_1 > 0$ our p-value is $P(T \geq t)$
- ii) $H_a: \beta_1 < 0$ our p-value is $P(T \leq t)$
- iii) $H_a: \beta_1 \neq 0$ our p-value is $2P(T \geq |t|)$
- iv)

Where T is a random variable from the t distribution with $n-2$ degrees of freedom.

INFERENCE FOR CORRELATION

The correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. We can use the sample correlation, r , to estimate the population correlation, whose symbol is ρ . When ρ is zero, this implies that there is no linear association in the population.

To test the hypothesis $H_0: \rho = 0$, compute the t statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where n is the sample size and r is the sample correlation.



For a test of $H_0: \rho = 0$ against:

- i) $H_a: \rho > 0$ our p-value is $P(T \geq t)$
- ii) $H_a: \rho < 0$ our p-value is $P(T \leq t)$
- iii) $H_a: \rho \neq 0$ our p-value is $2P(T \geq |t|)$

Where T is a random variable from the t distribution with $n-2$ degrees of freedom.

ANOVA

If we wanted to understand more about the different types and different sources of variation in our model, we could perform an **Analysis of Variation**, or ANOVA. Analysis of Variation **summarizes information about the sources of variation** in the data based on the model type $\text{DATA} = \text{FIT} + \text{RESIDUAL}$.

There are two types of deviations in this model. The first is due to the **fit** of the model, or choosing a straight line as the way to describe the relationship between our data. This type of deviation is represented as $\hat{y}_i - \bar{y}$. The second type of deviation is due to the **residuals** in our model. It is represented as $y_i - \hat{y}_i$. So the total variation in the response y is expressed as the sum of these two types of deviations. Therefore, the overall deviation is:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad (*)$$

If we square each of the three deviations in (*) and then sum them over all n observations, it is an algebraic fact that the sums of squares add, and so it holds that:

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

We can rewrite this equation as **SST = SSM + SSE**

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \text{ is the sum of squares due to the } \mathbf{model}$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ is the sum of squares due to } \mathbf{error} \text{ (from the residuals)}$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \text{ is the } \mathbf{total} \text{ sum of squares}$$

As with the t-distribution, when performing an analysis of variance, we need to talk about degrees of freedom, DF. If there are n observations in our

study, the variation of the y's would be $s_y^2 = \frac{\sum (y_i - \bar{y})}{n - 1}$ (the sample

variance). The numerator in this expression is SST and the denominator is the total degrees of freedom, DFT = n-1. Since our model (SLR) has one explanatory variable x, there is only one degree of freedom for our model. So DFM = 1. As with SST, the total degrees of freedom DFT is made up of the sum of the degrees of freedom of the model and the error (DFM + DFE). So the degrees of freedom for error is (n - 1) - 1 = n - 2, DFE = n-2.

$$\mathbf{DFT = DFM + DFE}$$

For each source of variation, the **ratio** of the sum of squares to the degrees of freedom is called the mean square, MS.

$$\text{MST} = s_y^2 = \frac{SST}{n-1} = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

this is the sample variance that we would calculate if all of the data came from a single population.

$$\text{MSM} = \frac{SSM}{1} = \frac{\sum (\hat{y}_i - \bar{y})^2}{1}$$

$$\text{MSE} = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

this is the variance about the population regression line. We use this to estimate σ^2 .

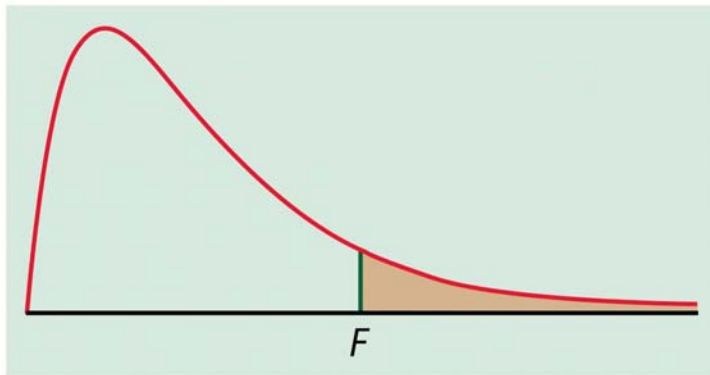
We previously said that r^2 , the coefficient of determination, is the fraction of the variation in y that is explained by the least squares regression of y on x . Since SST is the total variation in y , and SSM is the variation due to the model or regression of y on x , we can rewrite an equation for r^2 in terms of our sums of squares.

$$\mathbf{r^2} = \frac{SSM}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

This equation is another way of writing the **fraction of the variability in y explained by x**.

When we want to perform a two-sided significance test of $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$, we can either use a t test statistic or we can use an F statistic. Both statistics have many forms and either can be used for 2-sided significance tests since the t distribution and F distribution are related from the fact that $t^2 = F$. For ANOVA purposes, the F statistic compares MSM with MSE, and when $\beta_1 \neq 0$, MSM tends to be large relative to MSE. Therefore, **large values of F are evidence against H_0** in favor of the two-sided alternative,

where
$$F = \frac{MSM}{MSE} .$$



The p-value for the ANOVA F test is equal to the shaded area in the figure on the left. It is the probability that a random variable having the $F(1, n-2)$ distribution is greater than or equal to the calculated value of the F statistic.

$$p\text{-value} = P(F_{\text{calc}} \geq F_{1,n-2})$$

When we do ANOVA calculations, we usually display our results in a table called an ANOVA table. Here is the general setup, where column titles and sources of variations are always labeled, but actual calculations replace their symbols.

ANOVA table

Source	Degrees Of Freedom	Sums of Squares	Mean Square	F
Model	1	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	n-2	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	SSE/DFE	
Total	n-1	$\sum_{i=1}^n (y_i - \bar{y})^2$	SST/DFT	