9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown

Lecture 14 Linear Model I: Simple Regression

I. Objectives

Understand the simple linear regression model and its assumptions.

Understand the relation between fitting the simple linear regression model by maximum likelihood and the method-of-least squares.

Understand the properties of the parameter estimates, how we compute confidence intervals for them and how we test hypotheses about them.

Understand how we assess model goodness-of-fit.

Understand the Pythagorean formulation of regression analysis.

The linear model is the most used tool in modern statistical analysis. It also provides an important conceptual framework for constructing more complex models. For example, in neuroscience these methods are at the heart of most functional neuroimaging data analysis paradigms. In the next three lectures we will present the linear model by studying simple regression, multiple regression and analysis of variance (ANOVA).

II. Simple Linear Regression Model

A. Motivation for the Simple Linear Regression Model

To motivate the simple linear regression problem, we consider the following examples from the neuroscience literature.

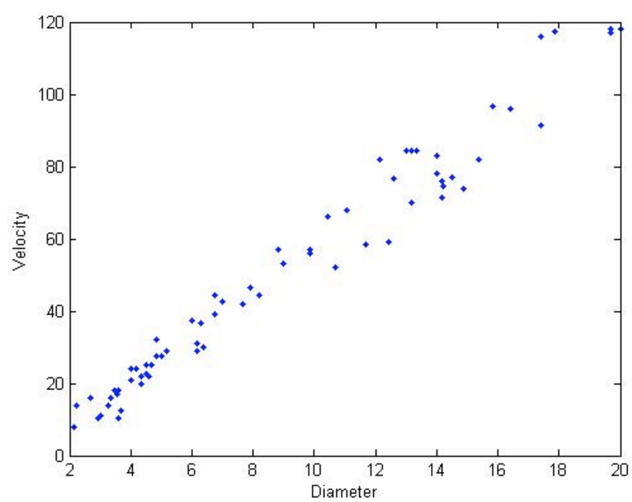

Figure 14.1. Relation between Conduction Velocity and Axon Diameter. Replotted from Hursh (1939).

Example 14.1. The Relation between Axon Conduction Velocity and Axon Diameter. Hursh (1939) presented data from an adult cat relating neuron conduction velocities to the diameters of axons. Hursh reported 67 velocity and diameter measurement pairs. He measured maximal velocity among fibers in several nerve bundles and then measured the diameter of the largest fiber in the bundle. The data are replotted in Figure 9.1. Diameter is reported in microns and velocity is reported in meters per second. We see that there is a strong positive relation in that as the axon diameter increases, so does the conduction velocity. Possible questions we may wish to answer include: How strong is the relation between conduction velocity and axon diameter? Can we quantify how conduction velocity changes as a function of diameter?

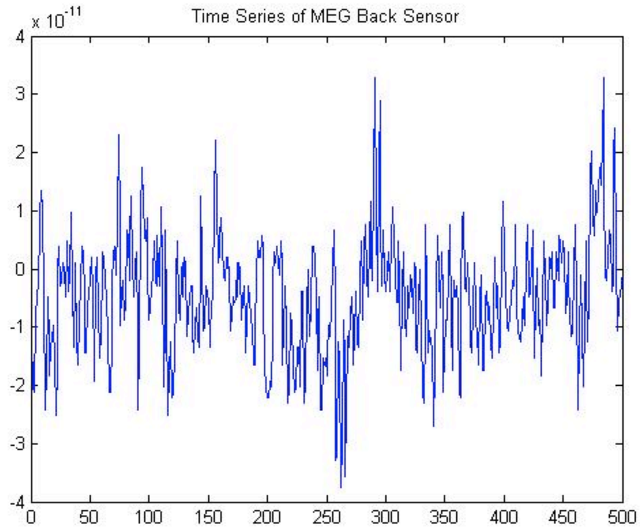

Figure 14.2. Time-Series Plot of first 500 observations of the MEG sensor background noise measurements.

Example 3.2 (continued). In this example, we have until now only considered the distributional properties of the MEG measurements. In so doing, we have treated these observations as if they were independent. If we plot the first 500 observations of the time-series (~ 833 msec), we see that the measurements do not appear to be independent (Figure 14.2). Indeed, there seems to be almost an oscillatory pattern in the time-series. Furthermore, if we plot x_t versus x_{t-1}, we see that there is a strong linear relation between adjacent measurements (Fig. 14.3). When x_{t-1} is large, x_t is also large and when x_{t-1} is small, x_t also tends to be small. In fact, the correlation coefficient between x_t and x_{t-1}, which we define below, is 0.65. Is this relation really there and if so, could this represent a systematic distortion in the local magnetic field?

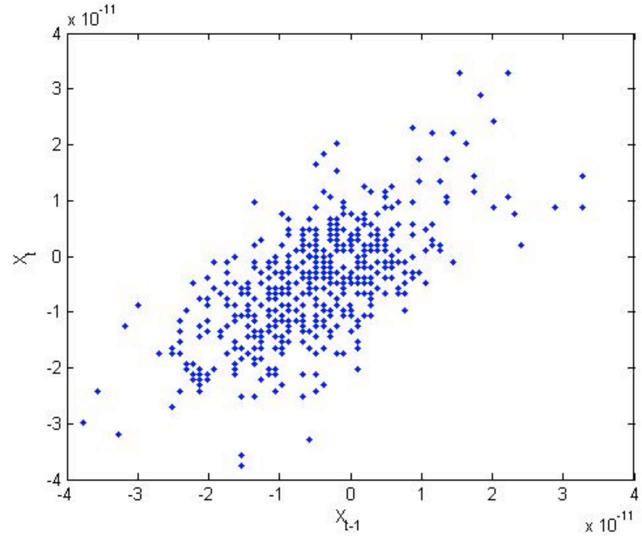

Figure 14.3. Plot of MEG background noise values x_t **versus** x_{t-1}.

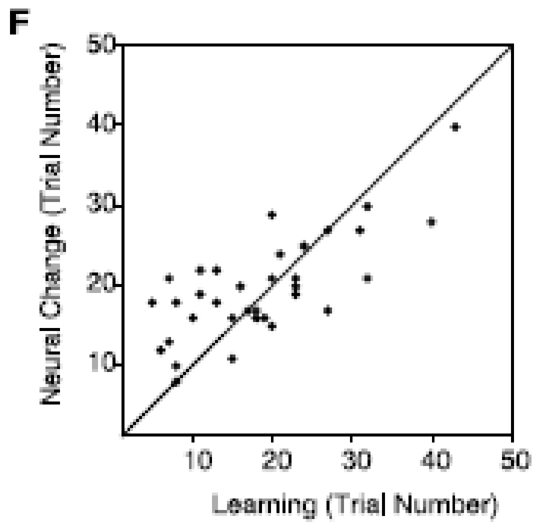

Figure 14.4. Plot of Change in Neural Activity during Learning Experiments versus Learning Trial (Wirth et al. 2003).

Example 14.2. Defining Neural Correlates of Behavioral Learning. Wirth et al. (2003) studied the relation between a monkey's performance learning a location-scene association task and changes in neural activity in the animal's hippocampus (Figure 14.4). They were interested in the question of how the timing of changes in the animal's neural activity during the task related to the trial in the experiment when the animal learned the task. Change in neural activity was defined as a significant increase in neural activity above baseline. Learning was defined by performance better than chance with a high degree of certainty. The 45 degree line in Figure 14.4 is the line which would mean that the neural change and the behavioral change occurred on the same trial. Is this the most accurate description of the relation between neural change and learning for these data?

B. Simple Linear Regression Model Assumptions

To formulate a statistical framework to study these problems, we assume we have data consisting of pairs of observations that we denote as $(x_1, y_1), \ldots, (x_n, y_n)$. For example, in **Example 14.1**, the variable y_i is axon velocity and the variable x_i is axon diameter. Let us assume that there is a linear relation between x_i and y_i and write it as

$$y_i = \alpha + \beta x_i. \tag{14.1}$$

To make the relation in Eq. 14.1 into a statistical model, we will make the following assumptions:

i) $E[y_i \mid x_i] = \alpha + \beta x_i$ for $i = 1, \ldots, n$.
ii) The x_i's are fixed non-random measurements termed **covariates**, **regressors** or **carriers**.
iii) The y_i's are independent Gaussian random variables with mean $\alpha + \beta x_i$ and variance σ^2.

Equation 14.1 and these three assumptions are often summarized as

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{14.2}$$

where the ε_i's are independent Gaussian random variables with mean 0 and variance σ^2. Equation 14.2 defines a **simple linear regression model.** It is simple because only one variable x_i is being used to describe or predict y_i and it is linear because the relation between y_i and x_i is assumed to be linear.

C. Model Parameter Estimation

Our objective is to estimate the parameters α, β and σ^2. Because y_i is assumed to have a Gaussian distribution conditional on x_i, a logical approach is to use maximum likelihood estimation. For these data the joint probability density (likelihood) is

$$\begin{aligned} L\left(\alpha, \beta, \sigma^2 \mid y, x\right) &= f\left(y \mid \alpha, \beta, \sigma^2, x\right) \\ &= \prod_{i=1}^{n} f\left(y_i \mid \alpha + \beta x_i, \sigma^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2}\right\} \end{aligned} \tag{14.3}$$

where $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. The log likelihood is

$$logf\left(y \mid x, \alpha, \beta, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2}. \tag{14.4}$$

To compute the maximum likelihood estimates of α, β and σ^2 we differentiate Eq. 14.4 with respect to α and β and σ^2

$$\frac{\partial \log f\left(y \mid x,\alpha,\beta,\sigma^2\right)}{\partial \alpha} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)}{\sigma^2} \tag{14.5}$$

$$\frac{\partial \log f\left(y \mid x,\alpha,\beta,\sigma^2\right)}{\partial \beta} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)x_i}{\sigma^2} \tag{14.6}$$

$$\frac{\partial logf\left(y \mid x,\alpha,\beta,\sigma^2\right)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{(\sigma^2)^2}. \tag{14.7}$$

Setting the derivatives equal to zero in Eqs. 14.5 and 14.6 yields the **normal equations**

$$\alpha n + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{14.8}$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{14.9}$$

In matrix form Eqs. 14.8 and 14.9 become

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}. \tag{14.10}$$

Solving for α and β yields

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}. \tag{14.11}$$

The solutions for $\hat{\beta}$ and $\hat{\alpha}$, which are the maximum likelihood estimates of β and α are respectively

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{14.12}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \tag{14.13}$$

We may write the estimate of y_i as

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}), \tag{14.14}$$

for $i = 1, \ldots, n.$ If we return to Eq. 14.7 we can compute the maximum likelihood estimate of σ^2 by first substituting in $\hat{\beta}$ and $\hat{\alpha}$ for β and $\alpha.$ Setting the left hand side of Eq. 14.7 to zero we obtain the maximum likelihood estimate of σ^2

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{14.15}$$

which is what we would have predicted based on our analyses in **Lecture 9**.

Remark 14.1. Estimating α and β by maximum likelihood under an assumption of Gaussian errors is equivalent to the **method of least-squares**. In least-squares estimation, we find the values of α and β which minimize the sum of the squared deviations of the y_i's from the estimated regression line. The method of least-squares is a form of method-of-moments estimation as we will show below.

Remark 14.2. Other metrics could be used to estimate α and $\beta,$ such as minimizing the sum of the absolute deviations. This would be defined as $n^{-1}\sum_{i=1}^{n}| y_i - \alpha - \beta x_i |.$

Remark 14.3. The estimate $\hat{\alpha}$ and Eq. 14.14 show that every regression line goes through the point $(\bar{x}, \bar{y}).$

Remark 14.4. The **residuals**, $y_i - \hat{y}_i,$ are the components in the data which the model does not explain. They are estimates of the ε_i's and we often write $\hat{\varepsilon}_i = y_i - \hat{y}_i.$ We note that

$$\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x}), \tag{14.16}$$

and summing we have

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x}) = 0. \tag{14.17}$$

Remark 14.5. There is an important Pythagorean relation between the sum of squared deviation in the data about its mean, the sum of squared deviations of the regression estimates about the mean of the data, and the sum of squared residuals. We derive it now. By **Remark 14.4**, the i^{th} residual is

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}). \tag{14.18}$$

Squaring and summing both sides of Eq. 14.18, we obtain

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \{(y_i - \bar{y}) - (\hat{y}_i - \bar{y})\}^2 \tag{14.19}$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}). \tag{14.20}$$

Now if we analyze the last term in Eq. 14.20 using Eq. 14.16 we find that

$$-2\sum_{i=1}^{n} (y_i - \bar{y})\hat{\beta}(x_i - \bar{x}) = -2\hat{\beta}\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \tag{14.21}$$

$$= -2\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

and Eq. 14.20 becomes

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{14.22}$$

which yields the desired **Pythagorean Relation**

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{14.23}$$

It states that

Total sum of squares (TSS)	=	**Explained sum of squares (ESS)**	+	**Residual sum of squares (RSS)**

We will make extensive use of this relation in our analyses of our regression fits.

D. The Relation between Correlation and Regression
Let us define the **sample correlation coefficient** between x and y as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}. \tag{14.24}$$

It is the method-of-moments estimate of the **theoretical correlation coefficient** between x and y which is defined as

$$\rho_{xy} = \frac{Cov(X,Y)}{[Var(X)Var(Y)]^{\frac{1}{2}}}, \tag{14.25}$$

where $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$ and where the expectation is taken with respect to $f(X,Y)$, the joint distribution of X and Y. If X and Y have a bivariate Gaussian distribution, then it is easy to show that r_{xy} is also the maximum likelihood estimate of ρ_{xy}. It is also straight forward to show that we always have $-1 < r_{xy} < 1$. Recall we showed in **Lecture 5** $-1 < \rho_{xy} < 1$. As we suggested in **Lecture 4**, when we use the term correlation we will be using it in the technical sense defined by either Eqs. 14.24 or 14.25. To work out the relation between the sample correlation coefficient and the slope parameter estimate, we recall that the slope parameter estimate (Eq. 14.12) can be expressed as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{14.26}$$

Hence,

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} \\
&= \left\{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{\frac{1}{2}} r_{xy} = r_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}
\end{aligned} \tag{14.27}$$

where $\hat{\sigma}_{xy} = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$. We see that the regression coefficient is a scaled version of the sample correlation coefficient or simply the sample covariance of x and y divided by the sample variance of x. In this way, we see that the least-squares estimates are method-of-moments estimates.

E. Distributions of the Parameter Estimates

The parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are functions of the data. Therefore, they are random variables and have distributions. It is straight forward to show that these distributions are Gaussian. Therefore, it suffices to specify their means and variances because this is the minimal description needed to define Gaussian random variables. It is straight forward to show that $E(\hat{\beta}) = \beta$ and $E(\hat{\alpha}) = \alpha$. The variances of the regression parameter estimates are

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{14.28}$$

$$Var(\hat{\alpha}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{14.29}$$

If we estimate σ^2 by $\hat{s}^2 = n(n-2)^{-1}\hat{\sigma}^2$ (Eq. 14.15) then following the logic we used in **Lecture 8,** the $100\%(1-\delta)$ confidence intervals for the parameters based on the $t-$distribution with $n-2$ degrees of freedom are

$$\hat{\beta} \pm \frac{t_{n-2,1-\delta/2}\hat{s}}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{\frac{1}{2}}} \tag{14.30}$$

$$\hat{\alpha} \pm t_{n-2,1-\delta/2}\left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{\frac{1}{2}}\hat{s} \tag{14.31}$$

There are $n-2$ degrees of freedom for this *t*-distribution because we estimated two parameters in this analysis.

Remark 14.6. We can also invert the above confidence intervals as we discussed in **Lecture 12** to test hypotheses about the regression coefficients by constructing a $t-$test.

Remark 14.7. The variance of $\hat{\beta}$ decreases as $\sum_{i=1}^{n}(x_i - \bar{x})^2$ increases. Hence, if we can design an experiment in which we can choose the value of the x_k's, we will spread them out as far as possible across the relevant range to decrease the variance of the estimated slope.

To construct a confidence interval for a predicted y_k value at a given value, x_k, we note that the

$$Var(y_k) = \frac{\sigma^2}{n} + \frac{(x_k - \bar{x})^2 \sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{14.32}$$

and hence, the $100\%(1-\delta)$ confidence interval is defined by

$$\hat{y}_k \pm t_{n-2,1-\delta/2}\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{\frac{1}{2}} \hat{s}. \tag{14.33}$$

Example 14.1. (continued). We fit the model in Eq. 14.1 to the axon diameter and velocity data in Fig. 14.1. The estimated regression line, along with the local 95% confidence intervals (Eq. 14.33), is shown in Fig. 14.5. The estimated regression line is consistent with the strong linear relation seen in the original data in Fig. 14.1. Most of the data lie close to the estimated regression line.

Parameter Estimate	**Standard Error**	**t-statistic**
$\hat{\alpha} =$ **-3.77**	**1.44**	**-2.61**
$\hat{\beta} =$ **6.07**	**0.14**	**43.5**

Table 14.1. Parameter Estimate Summary

The parameter estimates and their corresponding standard errors and $t-$statistics are shown in **Table 14.1**. We reject the null hypothesis that there is no linear relation since the t-statistic for $\hat{\beta}$ is 43.5 and has a $p-$value<<0.05. As there are 67 observations this $t-$statistic is essentially a $z-$statistic which has a value of 43.5!! Hence, this result is highly significant. More importantly, the 95% confidence interval for $\hat{\beta}$ is approximately [5.79, 6.35]. The estimate of $\hat{\beta}$ has units of meters per second per micron. This means that for every one micron increase in diameter, there is approximately a 6.07 meter per second increase in velocity. We also reject the null hypothesis that there is a zero intercept for the regression line since the $t-$statistic for $\hat{\alpha}$ is -2.61 and has a p-value<0.05. More importantly, the 95% confidence interval for α is approximately [-5.21, -2.33]. We conclude that the intercept is most likely negative given the limits of the confidence intervals. The intercept has units of meters per second. If this were a physical model, it would mean that at a zero diameter the velocity would be negative. This does not make physical sense and shows the importance of not using a regression model beyond the range of the data where it is estimated. The smallest diameter in our sample is approximately 2 microns and the largest is approximately 20 microns.

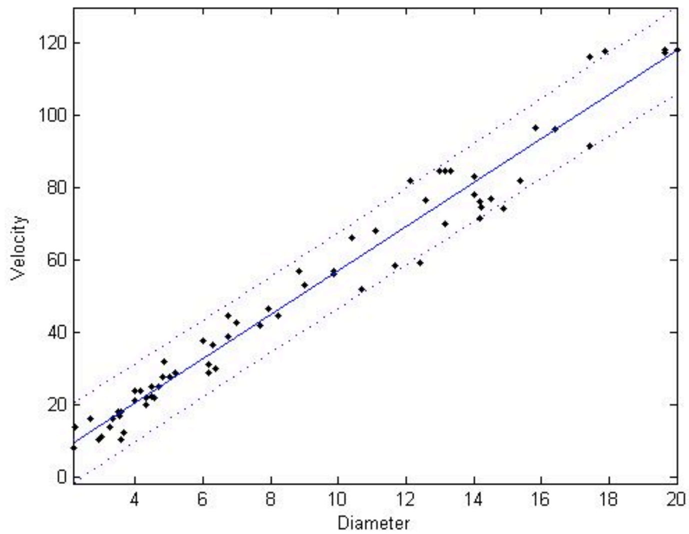

Figure 14.5. Fit and Confidence Intervals for a Simple Linear Regression Model of the Axon Conduction Velocity as a function of Axon Diameter.

F. Model Goodness-of-Fit

A crucial—if not the most crucial—step in a statistical analysis is measuring **goodness-of fit**. That is, how well does the model agree with the data. We previously used Q-Q plots to assess the degree to which elementary probability densities described sets of data. We discuss two statistical measures and three graphical measures of goodness-of-fit.

1. F-test

Given the null hypothesis $H_0 : \beta = 0,$ i.e. that there is no linear relation in x and $y,$ we can test this hypothesis explicitly using an $F-$test. The $F-$statistic with $p-1$ and $n-p$ degrees of freedom is

$$F_{p-1,n-p} = \frac{(p-1)^{-1} ESS}{(n-p)^{-1} RSS} = \frac{(p-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-p)^{-1} \sum_{i=1}^{n} (y_i - \hat{y})^2}, \tag{14.34}$$

where ESS is the explained sum of squares, RSS is the residual sum, p is the number of parameters in the model and n is the number of data points. We reject this null hypothesis for large values of the $F-$statistic. This suggests that if the amount of the variance in the data that the regression explains is large relative to the amount which is unexplained, then we reject the null hypothesis of no linear relation.

2. R-Squared

Another measure of goodness-of-fit is the **Square of the Multiple Correlation Coefficient** or R^2. It is defined as

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{14.35}$$

The R^2 measures the fraction of variance in the data explained by the regression equation. By the Pythagorean relation in Eq. 14.23, we see that $0 < R^2 \leq 1$. For the simple linear regression model $R^2 = r_{xy}^2$. The R^2 and the $F-$statistic are related as

$$R^2 = \frac{F(p-1)/(n-p)}{1+F(p-1)/(n-p)} \tag{14.36}$$

As the $F-$statistic increases, the R^2 increases. Hence, the greater the magnitude of the $F-$statistic, the greater the R^2.

Example 14.1 (continued). We have from the Pythagorean relation in Eq. 14.23

ESS	**65,041**
RSS	**2,233**
TSS	**67,274**

and that $F_{1,65}$ = 1,895 and R^2 = 0.967. Therefore, we reject the null hypothesis of no linear relation in the data with a *p*-value << 0.05. Furthermore, the R^2 suggests that the regressor, axon diameter can explain 97% of the variance in the axon conduction velocity.

Remark 8. The square of the $t-$statistic with $n-p$ degrees of freedom is the $F-$statistic with 1 and $n-p$ degrees of freedom. To verify this we find in **Table 1** that the ($t-$statistic)^2 = (43.5)^2 = 1,893 which is the value of $F_{1,65}$.

3. Graphical Measures of Goodness-of-Fit

a. Plot the raw data. Is the relation linear? (Fig. 14.1).
b. Plot the residuals versus the covariate x. Is there lack of fit of the model? (Fig. 14.6)
c. Plot the residuals versus the predicted values $\hat{y}$. Is there a relation? (Fig. 14.7)

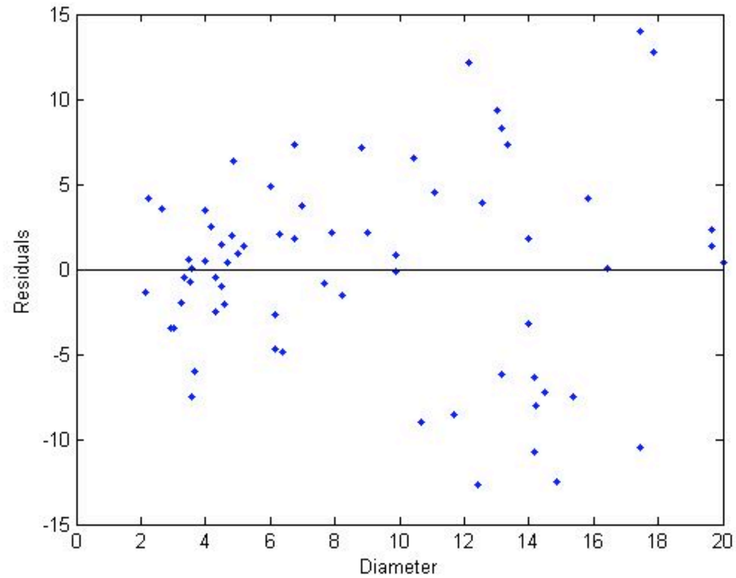

Figure 14.6. Plot of Residuals versus Axon Diameter for the Simple Linear Regression Model of the Axon Diameter and Conduction Velocity.

The residuals have no real discernible structure except that they seem to grow with the diameter of the axon suggesting that the assumption of **homoscedasticity,** i.e., that all of the observations have the same variance may be incorrect. When observations have different variances they are said to be **heteroscedastic**. We might model this by making the variance proportional to the diameter.

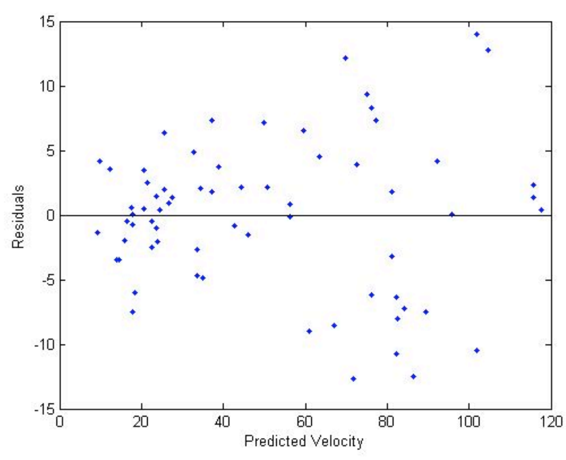

Figure 14.7. Plot of the Residuals against the Predicted Velocity.
The plot of the residuals versus the predicted velocity also shows the heteroscedasticity in the observations. This plot is performed because by the Pythagorean relation, the residuals and the predicted velocity estimates are orthogonal. Hence, if there is any relation between the two it suggests a way in which the model may be misspecified.

Inference. Based on this regression analysis axon conduction velocity is strongly associated with (predicted by) axon diameter. This analysis also reveals that a one micron increase in *diameter is associated with a 6 meter per second change in velocity for axon diameters between 2 and 20 microns.*

Remark 14.9. Correlation does not mean causation.

G. The Geometry of Regression Analysis (Method of Least-Squares)
As we have shown, regression analysis has an intuitively appealing geometric interpretation defined primarily be Eq. 14.23.

In particular, we can understand this geometry by considering two cases:

i) Small Residual Error
If the residual error is small it follows from the Pythogorean relation that the geometry of the analysis must have a small RSS relative to ESS.

ii) Large Residual Error
Similarly, if the residual error is large, it follows from the Pythogorean relation that the geometry of the analysis must have a large RSS relative to ESS.

III. Summary
The simple linear regression model can be used to study how well one variable may be described as a linear function of another variable. The Pythagorean relation is a useful construct for understanding how the analysis framework for the simple linear regression model is constructed. This Pythagorean relation will also be central to our analyses of multiple linear regression and ANOVA models.

Acknowledgments
I am grateful to Uri Eden for making the figures, to Julie Scott for technical assistance and to Jim Mutch for careful proofreading and comments. The data from the Hursh (1993) were provided by Rob Kass at Carnegie Mellon.

Text References
DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Draper NR, Smith H. *Applied Regression Analysis*, 2nd ed. New York: Wiley, 1981.

Rice JA. *Mathematical Statistics and Data Analysis*, 3rd edition. Boston, MA, 2007.

Literature References
Hursh JB. The properties of growing nerve fibers. *American Journal of Physiology*, 1939, 127: 131-39.

Wirth S, Yanike M. Frank LM, Smith AC, Brown EN, Suzuki WA. Single neurons in the monkey hippocampus and learning of new associations. *Science*, 2003, 300: 1578-81.