Density	Gain									
0.686	17.6	17.3	16.9	16.2	17.1	18.5	18.7	17.4	18.6	16.8
0.604	24.8	25.9	26.3	24.8	24.8	27.6	28.5	30.5	28.4	27.7
0.508	39.4	37.6	38.1	37.7	36.3	38.7	39.4	38.8	39.2	40.3
0.412	60.0	58.3	59.6	59.1	56.3	55.0	52.9	54.1	56.9	56.0
0.318	87.0	92.7	90.5	85.8	87.5	88.3	91.6	88.2	88.6	84.7
0.223	128	130	131	129	127	129	132	133	134	133
0.148	199	204	199	207	200	200	205	202	199	199
0.080	298	298	297	288	296	293	301	299	298	293
0.001	423	421	422	428	436	427	426	428	427	429

The Data

The data are from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs, California. The run consists of placing polyethylene blocks of known densities between the two poles of the snow gauge (Figure 8.1) and taking readings on the blocks. The polyethylene blocks are used to simulate snow.

For each block of polyethylene, 30 measurements were taken. Only the middle 10, in the order taken, are reported here. The measurements recorded by the gauge are an amplified version of the gamma photon count made by the detector. We call the gauge measurements the "gain."

The data available for investigation consist of 10 measurements for each of 9 densities in grams per cubic centimeter (g/cm^3) of polyethylene. The complete data appear in Table 8.1.

Background

Location

The snow gauge is a complex and expensive instrument. It is not feasible to establish a broad network of gauges in the watershed area in order to monitor the water supply. Instead, the gauge is primarily used as a research tool. The snow gauge has helped to study snow-pack settling, snow-melt runoff, avalanches, and rain-on-snow dynamics.

At one time, gauges were located on Mt. Baldy, Idaho, on Mt. Hood, Oregon, in the Red Mountain Pass, Colorado, on Mt. Alyeska, Alaska, and in the Central Sierra Nevada, California. The Central Sierra snow gauge provided the data to be analyzed in this lab. It is located in the center of a forest opening that is roughly 62 meters in diameter. The laboratory site is at 2099 meters elevation and is subject to

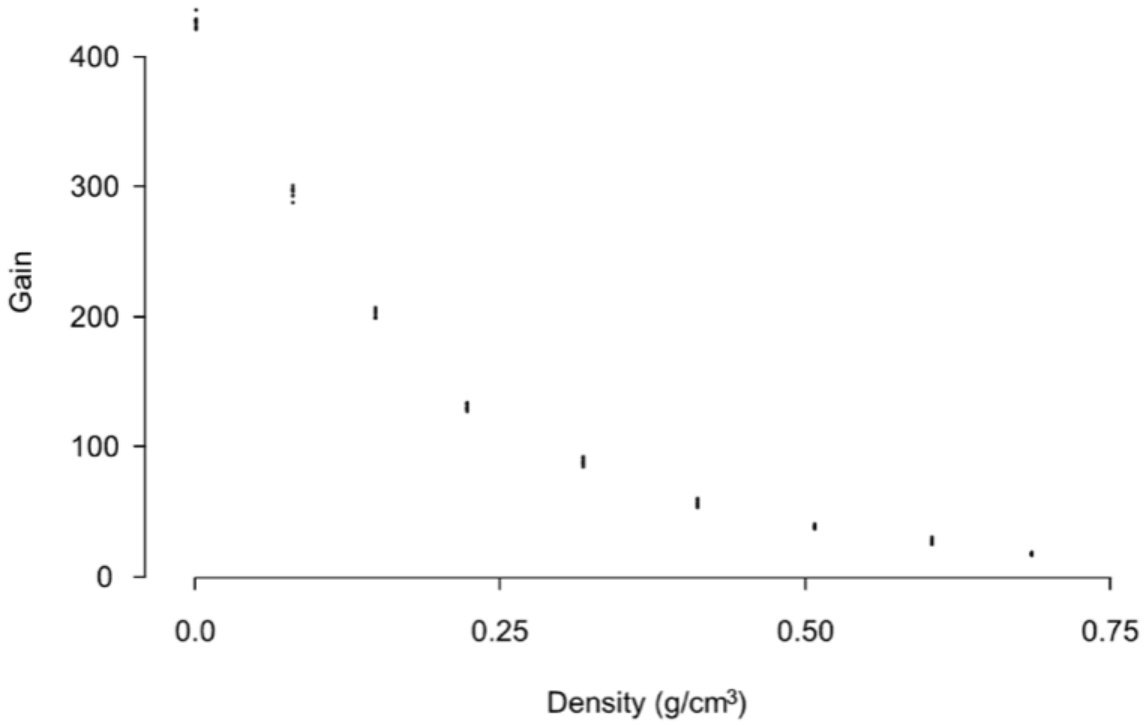

FIGURE 8.2. Plot of gain by density in g/cm^3 for ten blocks of polyethylene (USDA Forest Service).

major high-altitude storms which regularly deposit 5–20 centimeters of wet snow. The snow-pack reaches an average depth of 4 meters each winter.

The Gauge

The snow gauge consists of a cesium-137 radioactive source and an energy detector, mounted on separate vertical poles approximately 70 centimeters apart (Figure 8.1). A lift mechanism at the top of the poles raises and lowers the source and detector together. The radioactive source emits gamma photons, also called gamma rays, at 662 kilo-electron-volts (keV) in all directions. The detector contains a scintillation crystal which counts those photons passing through the 70-centimeter gap from the source to the detector crystal. The pulses generated by the photons that reach the detector crystal are transmitted by a cable to a preamplifier and then further amplified and transmitted via a buried coaxial cable to the lab. There the signal is stabilized, corrected for temperature drift, and converted to a measurement we have termed the "gain." It should be directly proportional to the emission rate.

The densities of the polyethylene blocks used in the calibration run range from 0.001 to 0.686 grams per cubic centimeter (g/cm^3). The snow-pack density is never actually as low as 0.001 or as high as 0.686. It typically ranges between 0.1 and 0.6 g/cm^3.

of 426.7, what is the density of the snow-pack? These two numeric values, 38.6 and 426.7, were chosen because they are the average gains for the 0.508 and 0.001 densities, respectively. Develop a procedure for adding bands around your least squares line that can be used to make interval estimates for the snow-pack density from gain measurements. Keep in mind how the data were collected: several measurements of gain were taken for polyethylene blocks of known density.

- To check how well your procedure works, omit the set of measurements corresponding to the block of density 0.508, apply your calibration procedure to the remaining data, and provide an interval estimate for the density of a block with an average reading of 38.6. Where does the actual density fall in the interval? Try this same test of your procedure for the set of measurements at the 0.001 density.
- Consider the log-polynomial model:

$$\log(\text{gain}) = a + b \times \text{density} + c \times \text{density}^2.$$

How well does it fit the data? Can you provide interval estimates here as well? Compare this model to the simple log-linear model in terms of its fit, predictive ability, physical model, and any other factors that you think are relevant.

Write a short instruction sheet for a lab technician on how the snow gauge should be calibrated for use in the winter months. Include an easy-to-use graph for determining snow density for a given gain. Accompany the instructions with a short technical appendix that explains the method you have chosen.

Theory

The Simple Linear Model

The simple linear model is that the expectation $\mathbb{E}(Y|x)$ of a random response Y at a known design point x satisfies the relation

$$\mathbb{E}(Y|x) = a + bx.$$

The *Gauss measurement model* supposes that measurement errors E have mean 0, constant variance (say σ^2), and are uncorrelated. A common practice is to express the response in the form

$$Y = a + bx + E,$$

with the understanding that the Es have the properties noted above. Notice that the Y is capitalized to denote that it is a random variable, and x is in lowercase to denote that it is fixed. Also, we use capitals when we explicitly represent Y as a random variable, and we use lowercase to represent an observed response, say y_1. Sometimes it is necessary to deviate from this convention when, for example, we take a function of the observations, such as the residual sum of squares $\sum(y_i - \hat{y}_i)^2$, and we also want to take the expectation of that quantity.

If we observe pairs $(x_1, y_1), \ldots (x_n, y_n)$, then the method of least squares can be used to estimate a and b. As in Chapter 7 on Dungeness crabs, the least squares estimates of a and b are

$$\hat{a} = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2},$$
$$\hat{b} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}. \tag{8.1}$$

Note that $\hat{a}$ and $\hat{b}$ are linear functions of the responses y_i, so they are linear functions of the errors, even though we don't get to see them. It is left to the Exercises to show that $\hat{a}$ and $\hat{b}$ are unbiased and to find their variances and covariance under the Gauss measurement model. The residuals $y_i - (\hat{a} + \hat{b}x_i)$ are also unbiased, and we can think of the residuals as estimates of the errors. It is left to the Exercises to show that the residual sum of squares has expectation

$$\mathbb{E}(\sum [y_i - (\hat{a} + \hat{b}x_i)]^2) = (n-2)\sigma^2.$$

The residual sum of squares can thus provide an estimate of the variance in the Gauss measurement model.

Model Misfit

An alternative to the simple linear model is a polynomial model. For example, a quadratic model for the expectation $\mathbb{E}(Y|x)$ is

$$\mathbb{E}(Y|x) = c + dx + ex^2.$$

This model can also be expressed as a two-variable linear model if we rewrite x^2 as u;

$$\mathbb{E}(Y|x) = c + dx + eu. \tag{8.2}$$

The topic of fitting quadratics and other multivariable linear models is covered in Chapter 10.

Regardless of the model for the expected value of the response, we can fit a line to the data. For example, say that (8.2) is the true model for our data. If we observe pairs $(x_1, y_1), \ldots, (x_n, y_n)$, we can fit a line by the method of least squares to these observations: minimize, with respect to a and b, the sum of squares

$$\sum_{i=1}^{n}[y_i - (a + bx_i)]^2.$$

The solutions for $\hat{a}$ and $\hat{b}$ remain as shown in equation (8.1). However, the model has been misfitted. These sample coefficients $\hat{a}$ and $\hat{b}$ may be biased under the Gauss measurement model,

$$Y_i = c + dx_i + eu_i + E_i,$$

We might use a multiplicative alternative to the Gauss measurement error model, which often describes data of this kind. In the multiplicative error model, random proportional error factors are nonnegative, have a common mean and variance, and are independent across measurements:

$$G = ce^{bx}W$$
$$\log(G) = \log(c) + bx + \log(W).$$

Now $\log(W)$ is like an error term in the Gauss model, except that its mean may not be zero (i.e., there may be bias).

Residual Plots

Graphical representations of the residuals are important diagnostic tools for detecting curvilinear relationships. If the simple linear model holds, then plots of the residuals against the explanatory variable —i.e., scatter plots of the pairs $(x_i, y_i - \hat{a} - \hat{b}x_i)$— should show a horizontal blur of points about the horizontal axis. Residuals can also be plotted against fitted values $\hat{y}_i$, another variable not used in the regression, or a function of x such as x^2.

Consider the residual plot (Figure 8.3) from the least squares fit of gain to density for the data in this lab. An obvious problem appears: the curvature in the residual plot indicates that it may be appropriate to include x^2 in the least squares fit or to transform G to $\log(G)$. The physical model for the relationship between gain and density suggests proceeding with the log transformation. In addition, the residual plots for the fit of log gain to density can help determine whether the errors are multiplicative or additive. If they are multiplicative, then the transformation should equalize the variance of the residuals. That is, for each value of x, the spread in the residuals should be roughly equal. However, any systematic error in the residuals indicates a problem other than whether the error may be multiplicative.

Figure 8.4 gives hypothetical examples of residual plots for different types of departures from the simple linear model. They show that:

- Patterns in the residuals may indicate a nonlinear relationship (top left).
- Unusually large residuals point to outliers that may have a large effect on the least squares line (bottom left).
- Funneling of the residuals may indicate that the variability about the least squares line may change with the values of x (top right).
- A linear trend in the residuals plotted against some variable not used in the regression may indicate that this variable should be included (bottom right).

Additionally, a normal-quantile plot of the residuals may indicate that they are non-normal, or more importantly that there is a long tail or skewness in the distribution of the residuals.

is an unbiased estimate of σ^2, where $\bar{y}_1 = k^{-1} \sum_j y_{1j}$. There are m such estimates $s_1^2, \ldots, s_m^2$ of σ^2 which can be pooled to form a single unbiased estimate:

$$s_p^2 = \frac{(k-1)s_1^2 + \cdots + (k-1)s_m^2}{mk - m}$$

$$= \frac{1}{mk-m}\sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij}-\bar{y}_i)^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}s_i^2.$$

The subscript p stands for "pooled."

With replicate measurements, the residual sum of squares (RSS) from the least squares fit of a simple linear model can be split into a measurement error sum of squares and a model misfit sum of squares :

$$\sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij}-\bar{y}_i)^2 + k\sum_{i=1}^{m}(\hat{y}_i-\bar{y}_i)^2.$$

Note that $\hat{y}_{ij} = \hat{a} + \hat{b}x_i$, which we abbreviate $\hat{y}_i$. The first sum on the right side is the measurement error sum of squares. It is an unnormalized pooled estimate of σ^2, with $m(k-1)$ degrees of freedom. The second term is the lack of fit sum of squares. When the model is correct, the second term provides another estimate of σ^2, with $m-2$ degrees of freedom.

Dividing a sum of squares by the degrees of freedom yields what is called a *mean square*. The pure error mean square and the model misfit mean square can be compared to help determine whether there is a lack of fit. If the errors are mutually independent and normally distributed, and there is no model misfit, then the ratio

$$\frac{k\sum_i(\hat{y}_i-\bar{y}_i)^2/(m-2)}{\sum_i\sum_j(y_{ij}-\bar{y}_i)^2/m(k-1)}$$

follows an F distribution with $m-2$ and $m(k-1)$ degrees of freedom. On the other hand, if there is model misfit, then the numerator should be larger than the denominator, with values of the ratio bigger than 3 to 5 indicative of misfit. We can use this ratio as a test of how well the data fit the model.

Confidence and Prediction Bands

We can predict the response y at any value x by $\hat{y} = \hat{a} + \hat{b}x$. An interval for this prediction can be based on the following variance for $R = 1$:

$$\mathrm{Var}(y-\hat{y}) = \mathrm{Var}(y-\hat{a}-\hat{b}x) = \sigma^2\left[1 + \frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].$$

On the right side, the first term in square brackets is the contribution from the variation of y about the line; the second and third terms are the contributions from the uncertainty in the estimates of a and b, respectively.

An approximate 99% *prediction interval* for y, at x, is formed from the following two bounds:

$$(\hat{a} + \hat{b}x) + z_{0.995}\,\hat{\sigma}\left[1 + \frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]^{1/2}$$

and

$$(\hat{a} + \hat{b}x) - z_{0.995}\,\hat{\sigma}\left[1 + \frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]^{1/2},$$

where $z_{0.995}$ is the 0.995 quantile of the standard normal distribution, and $\hat{\sigma}$ is the standard deviation of the residuals. The prediction interval for y differs from a confidence interval for $a + bx$,

$$(\hat{a} + \hat{b}x) \pm z_{0.995}\,\hat{\sigma}\left[\frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]^{1/2}.$$

The confidence interval is smaller because it does not include the variability of y about the line $a + bx$. It tells us about the accuracy of the estimated mean of y at x, whereas a prediction interval is for an observation, or the mean of a set of observations, taken at design point x.

Figure 8.5 displays these prediction intervals for all values of x over a range. These curves are called *prediction bands*. Notice that the size of an interval depends on how far x is from $\bar{x}$. That is, the interval is most narrow at $\bar{x}$ and gets wider the further x is from $\bar{x}$.

If y_0 is the average of r measurements, all taken for the same density x_0 and R replicate measurements are taken at each x_{-i}, then the prediction variance becomes

$$\mathrm{Var}(y_0 - \hat{y}_0) = \sigma^2\left[\frac{1}{r} + \frac{1}{mR} + \frac{(x_o-\bar{x})^2}{R\sum(x_i-\bar{x})^2}\right].$$

As before, the first term in the variance is the contribution from the variation about the line, which is now σ^2/r because we have an average of r observations. The bands must be adjusted accordingly.

The bands shown here are pointwise bands. This means that they provide an interval for one future reading, or the average of a set of future readings.

Calibration

There is an important difference between the calibration run and how the gauge will be used in practice. In the calibration run, polyethylene blocks of known densities are placed in the gauge, and the gain is measured. At each density level, several measurements were taken and were not all the same. This variability is due to measurement error. In the future, the gain will be measured for an unknown snow density, and the gain measurement will be used to estimate snow density; that is,

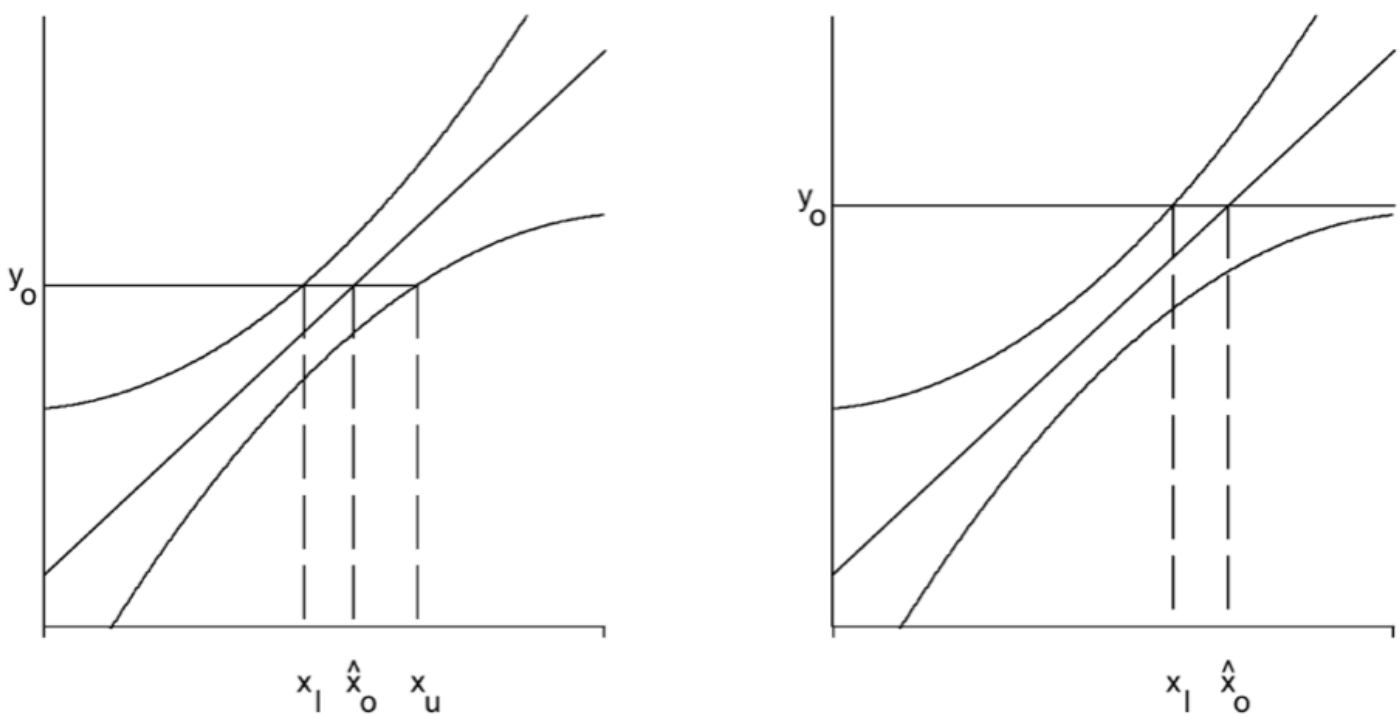

FIGURE 8.5. Prediction bands and inverse prediction estimates for x_0.

in the calibration run, we take the relationship between log gain and snow density to follow a standard linear model:

$$Y = a + bx + E,$$

where Y represents the log gain, x denotes the known polyethylene density, and the errors follow the Gauss measurement model. In the future, when the gauge is in operation, we will observe the new measurement y_0 for a particular snow-pack and use it to estimate the unknown density, say x_0.

One procedure for estimating the density first finds the least squares estimates $\hat{a}$ and $\hat{b}$ using the data collected from the calibration run. Then it estimates the density as follows:

$$\hat{x}_0 = \frac{y_0 - \hat{a}}{\hat{b}}.$$

This is called the *inverse estimator*.

Just as the least squares line can be inverted to provide an estimate for snow density, the bands in Figure 8.5 can be inverted to make interval estimates for snow density. In the left plot, read across from y_0, to the top curve, then read down to find the corresponding x-value, which we call x_l. This is the lower bound for the interval estimate. We similarly find the upper bound by reading across from y_0 to the lower curve and then down to the x-value, which we call x_u. The interval (x_l, x_u) is not symmetric about $\hat{x}_0$, and in some cases it may be a half-line, as in the example shown in the right plot in Figure 8.5.

5. Explain why the confidence interval obtained in Exercise 3, for $\mathbb{E}(Y)$, is smaller than the prediction interval for y, using the prediction variance