

March 5, 2009

Chapter 8: Regression

A very common problem in scientific studies is to determine how one variable depends on one or more other variables. For example, is a hurricane's wind speed related to the ocean water temperature? If so, what is the relationship?

Think back to the pelican egg example where data was collected on eggshell thickness and the level of PCB's in the birds. A question of interest is whether or not the eggshell thickness depends on the PCB level and if so, how does the PCB level affect eggshell thickness? Again, regression analysis can be used to answer this question.

Francis Galton is credited with originating the term regression when he studied how heights of children depended on the heights of their parents. Can we predict a child's height based on the height of the child's parents? Or, for parents of a particular height, what is the average height of their offspring? Linear regression is used to answer these types of questions.

Regression analysis is a broad field of statistics that deals with the problem of modelling how a variable depends on one or more other variables. Consequently, regression analysis is one of the most heavily used branches of statistics. Regression models can become quite complicated. We begin this chapter with a simple model.

1 Simple Linear Regression

Figure 1 shows a scatterplot of the time to run two miles for a sample of middle-aged males versus the maximum volume of oxygen uptake in the blood ($\text{VO}_2 \text{ max}$). Clearly there appears to be a relation between how fast one can run two miles and the maximum VO_2 uptake. One of the goals of the current chapter is to learn how to model this type of data where one variable depends on another variable. From Figure 1, a linear relationship seems reasonable. The general term used to model how one variable depends on another variable is *regression analysis*.

In regression settings we have a response variable y (also known as the dependent variable) and a regressor variable x (also known as the independent or predictor variable). Data is collected on pairs $(x_1, y_1), \dots, (x_n, y_n)$. The set up is that the average response depends on the value of x which defines a *conditional expectation of y given x* , denoted

$$E[y|x].$$

The way to think of this conditional expectation is: what is the average value of y for a given value of x ? This conditional expectation is regarded as a function of x . In the simple linear regression, we model this conditional expectation as a linear function:

$$E[y|x] = \beta_0 + \beta_1 x. \quad (1)$$

This is the equation of a line with y -intercept β_0 and slope β_1 .

We can write (1) in another form:

$$y = \beta_0 + \beta_1 x + \epsilon, \quad (2)$$

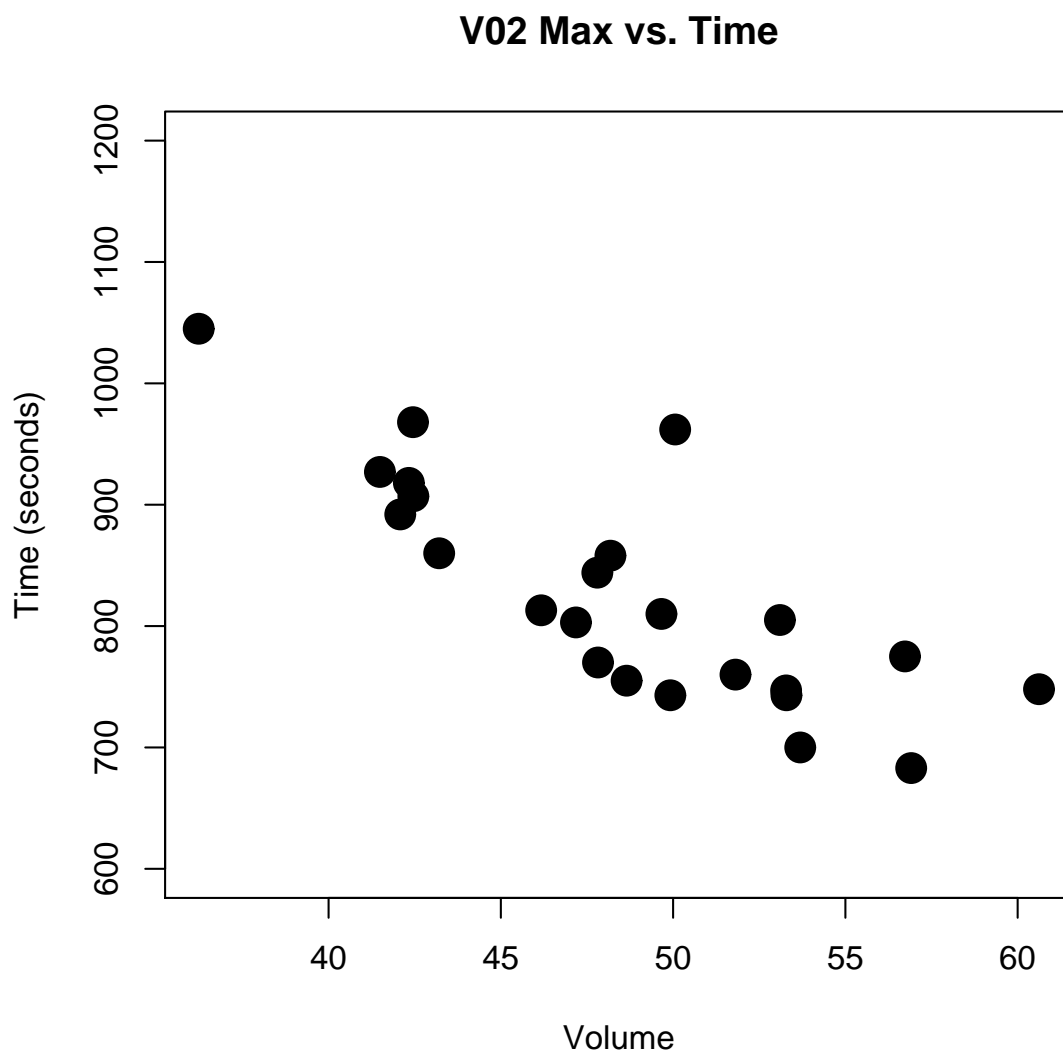


Figure 1: Scatterplot of time (to run 2 miles) versus Maximum volume of O_2 uptake for a sample of middle-aged men.

where ϵ is a random error which is needed because the points will not all lie exactly on a line.

In previous chapters, we have concentrated on estimating and testing hypotheses about the mean of a population. The mean is a parameter denoted by μ . In the regression setting, we also have parameters. In particular, the parameters are β_0 , the y -intercept, and the slope β_1 .

There is another parameter for the model (2), but it does not show up directly in the formula. That is, the random error is assumed to have a mean zero and a variance σ^2 . Thus, σ^2 , the error variance, is the other parameter of the model. For small values of σ^2 , the points in the scatterplot will be tightly clustered about the true regression line; for large values of σ^2 , the points will be spread out more from the regression line.

In regression analysis, the common statistical problems of interest are:

- To test if the regressor variable x affects the response variable y .
- Predict a response for a given value of x .
- Estimate an average response for a given value of x .
- Model the relation between x and y .

2 Fitting Regression Lines: Least Squares

In Figure 1, the points show a roughly linear relation. If the simple linear regression model is the correct model, then there is a linear function defined in terms of β_0 and β_1 that models the relation. However, β_0 and β_1 are model parameters and are unknown. We must use the data to estimate the intercept and slope of the line.

Question: How should we estimate the regression line?

Least-Squares. Looking back at Figure 1, the goal is to find a line that runs through the middle of the data. There are several choices for “fitting” a line. The most common method of estimating the regression line is to use *least-squares*. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least-squares estimators of β_0 and β_1 respectively. Given a data point (x_i, y_i) , the corresponding point on the least-squares regression line is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The least-squares estimators are determined by minimizing (i.e. finding the *least* value) of the sum of *squares*:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (3)$$

Calculus can be used to solve this problem and the solution is given by the following equations:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (\text{Least-Squares Estimator of the intercept}) \quad (4)$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (\text{Least-Squares Estimator of the slope}) \quad (5)$$

Given a set of data $(x_1, y_1), \dots, (x_n, y_n)$, one can plug the data values into (4) and (5) to get the least-squares regression line.



Figure 2: Scatterplot of time (to run 2 miles) versus Maximum volume of O_2 uptake for a sample of middle-aged men. The line is the least-squares regression line $\hat{y} = 1450.86 - 12.86x$.

Predicted Value. For a given data point (x_i, y_i) , the predicted value or predicted response from the estimated regression line is denoted by \hat{y}_i and is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (6)$$

One reason for the popularity of the least-squares estimator is that if the error ϵ in (2) has a normal distribution, then the least-squares estimators of the intercept and slope correspond to maximum likelihood estimators.

Figure 2 shows the scatterplot of the VO_2 max data again along with the estimated least-squares regression line that was computed using the above formulas. Using these formulas by hand is quite tedious. Typically statistical software such as SAS is used to do the computation (which we illustrate below) The least-squares regression coefficients are found to be $\hat{\beta}_0 = 1450.86$ and $\hat{\beta}_1 = -12.86$. Therefore, the least-squares regression line is given by the equation

$$\hat{y} = 1450.86 - 12.86x,$$

where x is the maximum VO_2 uptake. Note that the estimated slope is negative indicating that as VO_2 max increases, the time to run the two miles decreases. In particular, we estimate that for each additional unit of VO_2 max, the time it takes to run the two miles decreases on average by 12.86 seconds. In Section 3.2 we show how to compute a confidence interval for the slope coefficient.

Note that the intercept in the VO_2 max example is estimated to be $\hat{\beta}_0 = 1450.86$. In the context of this example, the y -intercept is not interpretable. This value tells us where the regression line crosses the y -axis. If the VO_2 max value is $x = 0$, the predicted response is 1450.86. However, a value of zero for x is

not meaningful here since a value of $x = 0$ is not possible for a living middle aged man. Also, we do not have data collected for values of x near zero. To predict a response near zero would require *extrapolation* which is always a very dangerous thing to do in statistics.

3 Inference for Regression Parameters

In simple linear regression there is only a single predictor variable x . Inference in simple linear regression is usually focused on the slope β_1 which relates how the response variable y depends on the predictor x . Recall that the slope tells us how much of a change one can expect to see in the response y for a unit change in x . For instance, in the VO_2 max example, how much faster would we expect a man to run 2 miles if his VO_2 max is increased by one unit? The slope answers this question. Thus, our interest is to not only estimate the slope β_1 but to obtain some measure of the reliability of the estimate.

In other examples, it may not be clear if there is a relation between x and y . For example, does exposure to lead in children lead to diminished IQ scores? Studies have been done that looked at IQ levels and blood lead levels in children. If lead does not affect IQ, then we would expect the slope of the regression line relating IQ to lead levels to be zero (i.e. a horizontal line). Therefore, we have an interest in testing hypotheses about the slope of a regression line. We can also test hypotheses about the y -intercept of a line, but this is often not of interest.

Inference procedures for the slope are based on the following fact about the least-square estimator of the slope:

FACT 1: If $(x_1, y_1), \dots, (x_n, y_n)$, is a random sample and the error term ϵ in (2) has a normal distribution, then the least square estimator of the slope $\hat{\beta}_1$ has a normal distribution with mean β_1 and variance

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

That is, $\hat{\beta}_1$ has a normal distribution and is unbiased for the true regression slope β_1 . The least-square estimator of the intercept $\hat{\beta}_0$ is also unbiased and follows a normal distribution. Even if the error ϵ is not exactly normal, the least-square estimators will still have approximately a normal distribution due to the central limit theorem effect, provided the sample size is sufficiently large.

Standard Error of the Estimated Slope. The square root of the variance of $\hat{\beta}_1$ is the standard error of the estimated slope, denoted by $SE(\hat{\beta}_1)$:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Of course σ^2 is unknown and must be estimated. Therefore in practice we use the estimated standard error for the slope:

$$\hat{SE}(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

where MSE stands for *mean squared error* and is given by

$$\text{MSE} = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 / (n - 2).$$

We will discuss MSE in greater detail below.

The next fact is used for statistical inference:

FACT 2: If $(x_1, y_1), \dots, (x_n, y_n)$, is a random sample and the error term ϵ in (2) has a normal distribution, then

$$t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)},$$

follows a t -distribution with $n - 2$ degrees of freedom.

Here are a couple important points:

- Note that the degrees of freedom for t is $n - 2$, not $n - 1$. The reason why is that we lose two degrees of freedom for estimating the intercept and slope of the line.
- The sampling distribution of t will have an approximate t distribution if the error ϵ deviates from normality, provided the deviation is not too severe or the sample size is large.

3.1 Hypothesis Testing for the Slope

A null hypothesis about the slope β_1 of the regression line can be stated as

$$H_0 : \beta_1 = \beta_{10},$$

where β_{10} is some hypothetical value of the slope of interest. The alternative hypothesis can be one-sided or two-sided: depending on the application:

$$H_a : \beta_1 \neq \beta_{10}, \text{ or } H_a : \beta_1 > \beta_{10}, \text{ or } H_a : \beta_1 < \beta_{10}.$$

If $\hat{\beta}_1$ deviates substantially from the hypothesized value β_{10} , then we would reject the null hypothesis. The benchmark for making this determination is the t -distribution based on Fact 2 above.

$$\text{Test Statistic: } t = \frac{\hat{\beta}_1 - \beta_{10}}{\widehat{SE}(\hat{\beta}_1)}. \quad (7)$$

If H_0 is true, then the t -test statistic follows a t -distribution on $n - 2$ degrees of freedom.

When performing a test about the slope at a significance level α , the following summarizes how hypothesis testing is done:

- If the alternative hypothesis is $H_a : \beta_1 \neq \beta_{10}$, we reject H_0 when $t > t_{n-2, \alpha/2}$ or $t < -t_{n-2, \alpha/2}$, where t is the t -test statistic in (7).
- If the alternative hypothesis is $H_a : \beta_1 > \beta_{10}$, we reject H_0 when $t > t_{n-2, \alpha}$.
- If the alternative hypothesis is $H_a : \beta_1 < \beta_{10}$ we reject H_0 if $t < -t_{n-2, \alpha}$.

One of the most common tests in regression is to test if the slope differs from zero in which case the hypothesized value of the slope is $\beta_{10} = 0$ and the test statistic becomes

$$t = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}. \quad (8)$$

In SAS (and many other statistical software packages), this test is done automatically when a regression is run. The VO_2 max example will be used to illustrate ideas:

```

/*****
One measure of cardiovascular fitness is the maximum volume of oxygen
uptake during exercise. 24 max. vol. of oxygen (column 1) was measured
on 24 middle aged men who ran 2 miles on a treadmill. Column 2 of the
data gives the time to run 2 miles in seconds.
(Reference: Ribisl and Kachadorian, "Maximal Oxygen Intake
Prediction in Young and Middle Aged Males," J. of Sports Medicine, 9,
1969, 17-22).
*****/
data vo2;
input vol time;
datalines;
42.33 918
53.10 805
42.08 892
50.06 962
42.45 968
42.46 907
47.82 770
49.92 743
36.23 1045
49.66 810
41.49 927
46.17 813
48.18 858
43.21 860
51.81 760
53.28 747
53.29 743
47.18 803
56.91 683
47.80 844
48.65 755
53.69 700
60.62 748
56.73 775
;

```

```
run;
proc reg;
  model time = vol;
run;
quit;
```

Here is the SAS output from PROC REG:

```

                                The REG Procedure
                                Model: MODEL1
                                Dependent Variable: time

                                Number of Observations Read      24
                                Number of Observations Used       24

                                Analysis of Variance

Source                            DF          Sum of Squares          Mean Square          F Value          Pr > F
Model                              1           129720                129720                42.01           <.0001
Error                              22           67938                 3088.10214
Corrected Total                    23           197658

                                Root MSE          55.57070
                                Dependent Mean    826.50000
                                Coeff Var       6.72362
                                R-Square          0.6563
                                Adj R-Sq       0.6407

                                Parameter Estimates

Variable        DF          Parameter Estimate      Standard Error      t Value          Pr > |t|
Intercept       1           1450.86476              96.99988            14.96            <.0001
vol             1           -12.86113                1.98437             -6.48            <.0001

```

We will only discuss the last portion of this SAS output for now. The other portions of the output will be explained in Section 6.

From the SAS output, we see that the least-squares regression line is given by

$$\hat{y} = 1450.865 - 12.861\text{vol.}$$

SAS's PROC REG also automatically prints out the estimated standard errors for the slope and intercept estimates. The estimated standard error for the slope is $\hat{SE}(\hat{\beta}_1) = 1.984$. SAS also automatically prints the results of testing $H_0 : \beta_1 = 0$. The test statistic $t = -6.48$ from the output is computed from (8) as

$$t = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} = \frac{-12.86113}{1.98437} = -6.48.$$

The p -value reported in SAS for this test (i.e. $p < 0.0001$) is for a two-sided alternative $H_a : \beta_1 \neq 0$. Because the p -value is so small, we have very strong evidence that the slope differs from zero. In other words, the time it takes to run two miles depends on the VO_2 max of the runner. The result of this hypothesis test is not too interesting because people familiar with VO_2 max on conditioning would already know there would be a relation. The next subsection shows how to compute a confidence interval for the slope which is more informative in this example.

SAS also prints out the results for testing if the intercept is zero or not, but as mentioned earlier, this hypothesis test is not of interest in this example.

3.2 Confidence Intervals for Regression Coefficients

A $(1 - \alpha)100\%$ confidence interval for the slope β_1 of a regression line is

$$\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \hat{SE}(\hat{\beta}_1). \quad (9)$$

In the VO_2 max example, a 95% confidence interval for the slope is given by

$$\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \hat{SE}(\hat{\beta}_1) = -12.861 \pm (2.074)(1.984) = -12.861 \pm 4.116,$$

where 2.074 is the 97.5th percentile of the t -distribution on $n - 2 = 24 - 2 = 22$ degrees of freedom. This gives a 95% confidence interval for the slope of $(-16.978, -8.746)$. Thus, with 95% confidence we estimate that for every additional increase in a runner's VO_2 max, we would expect to see a *decrease* in the time to run two miles of between 8.75 and 16.98 seconds on average. The confidence interval procedure here is interpreted the same as it was for confidence intervals for the mean of a population. The 95% confidence interval for the slope is a procedure that works 95% of the time. If the experiment were to be repeated over and over and a confidence interval for the slope was obtained for each experiment, we would expect roughly 95% of these intervals to contain the true slope β_1 .

Caution: For the inference procedures described here to be valid, the underlying regression model needs to be (approximately) correct. That is, the relation between x and y should be approximately linear. From Figure 2, the linear relationship assumption seems reasonable. However, in many examples similar this, a non-linear relationship would become evident if data were collected over a larger range of x values. For instance, one may expect that the time it takes to run 2 miles may plateau out for large values of VO_2 max. After all, it is physically impossible to see a continued linear improvement in times to run 2 miles for increasing VO_2 max values (for example, negative times are not possible). Thus, in regression problems, it is very important not to *extrapolate* outside the range where the data is collected because we cannot determine if the observed relation seen in the data would continue to hold outside the range where data is available. Exceptions to this rule require high confidence that the linear relation between x and y remains the same outside the range where the data is collected.

A confidence interval for intercept is given by

$$\hat{\beta}_0 \pm t_{n-2, 1-\alpha/2} \hat{SE}(\hat{\beta}_0),$$

where $\widehat{SE}(\hat{\beta}_0)$ can be read off the SAS output.

3.3 Estimating and Predicting Responses

A couple common uses of an estimated regression model are:

1. For a given x value, estimate the mean response $E[y|x]$.
2. For a given x value, predict a new response.

In both cases, the mean response and predicted response at x are each given by the same expression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (10)$$

The goal in this section is to provide a confidence interval for an estimated mean response and a *prediction* interval for a predicted response.

It may seem at first that because the estimated mean response and the predicted response are both given by the same value \hat{y} in (10), the confidence interval for a mean response and the prediction interval for a predicted response would be the same. However, they are not the same. We shall use the VO_2 max example to illustrate the difference. Suppose we want to predict the mean response for runners whose VO_2 max is 50. Then the estimated mean response is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(50) = 1450.86476 - 12.86113(50) = 807.81 \text{ seconds.}$$

We estimate that the average time it would take all all male runners with a VO_2 max of 50 to run two miles is 807.81 seconds. In this sense, the 807.81 is the mean of a population: it represents the mean time to run two miles for all males whose VO_2 max is 50. Therefore, it makes sense to compute a confidence interval for this parameter.

On the other hand, suppose a male runner is enrolled into the experiment whose VO_2 max is 50. What would we predict this runner's time would be for two miles? The answer is 807.81. However, here we are trying to predict the time of an individual runner selected at random. We would like a *prediction* interval to give a range of plausible values for the time. The idea of say a 95% prediction interval is to find an interval that will contain 95% of the times of all runners whose VO_2 max is 50. Because a prediction interval is attempting to capture the time of a single randomly selected runner, it necessarily needs to be wider than a confidence interval for a mean response that is attempting to capture a mean of an entire population.

A $100(1 - \alpha)\%$ **prediction interval** for y at a given x value is

$$\hat{y} \pm t_{n-2, 1-\alpha/2} SE_1(\hat{y}), \quad (11)$$

where

$$SE_1(\hat{y}) = \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]}.$$

A $100(1 - \alpha)\%$ **confidence interval** for a mean response $E[y|x]$ at a given x value is

$$\hat{y} \pm t_{n-2, 1-\alpha/2} SE_2(\hat{y}), \quad (12)$$

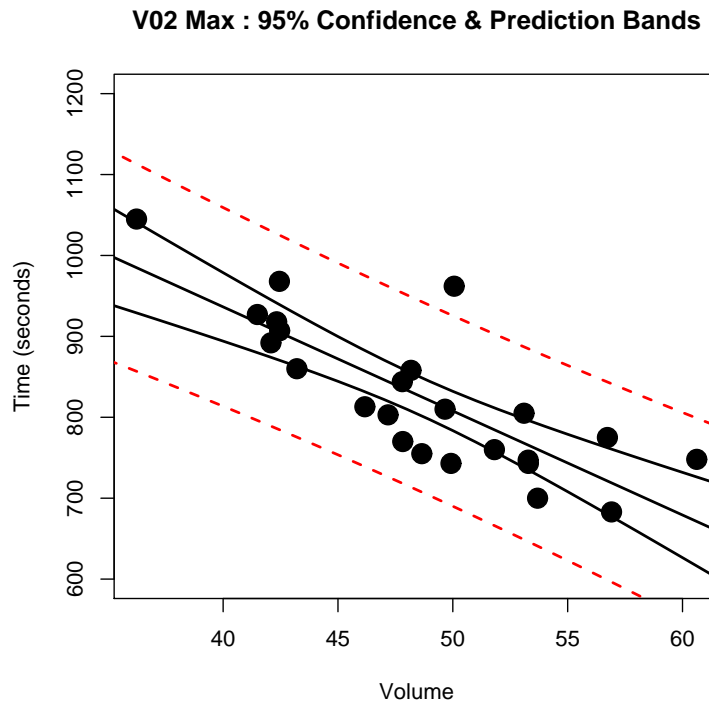


Figure 3: 95% Confidence (solid curves) and prediction bands (dashed curves) for the VO₂ max data.

where

$$SE_2(\hat{y}) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]}.$$

From (11) and (12), it is clear that the prediction interval will be wider than the confidence interval due to the extra “1” underneath the radical sign in (11). This additional term is needed for predicting a single random response as opposed to estimating a mean response.

Figure 3 shows a scatterplot of the VO₂ max data with 95% confidence bands for the mean response (solid curves) and 95% prediction bands for predicted responses (dashed curves). Note that the prediction bands are considerably wider than the confidence bands as expected. Also note that out of the 24 observations, only one observation falls outside the prediction band which is consistent with what one would expect for a 95% prediction interval.

From (11) and (12) the 95% prediction and confidence intervals for a VO₂ max of 50 are (783.54, 832.08) and (690.03, 925.58) respectively.

Note also that both intervals (11) and (12) are narrowest at $x = \bar{x}$ (due to the $(x - \bar{x})^2$ term in both formulas). Thus, estimation of the mean response is most precise when the estimation is done for $x = \bar{x}$.

4 Diagnostics

In a simple linear regression, the model assumes that the response y is a linear function of x (plus a random error). This linear relationship assumption should be checked in practice. Often in a simple linear regression setting, a simple scatterplot of y versus x will indicate if a linear relationship seems reasonable.

4.1 Residual Analysis

We can write the error ϵ in the simple linear regression model (2) as

$$\epsilon = y - (\beta_0 + \beta_1 x).$$

The sample counterpart to the error are the *residuals*

Definition. For the i th observation (x_i, y_i) , the i th **residual** r_i is defined as

$$r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

Because the error ϵ is assumed to be totally random in a correctly specified model, the residuals should also look like a random scatter of points. A useful diagnostic to perform with any regression analysis is to plot the residuals r_i versus the predicted values \hat{y}_i . If the model is correctly specified, then the residual plot should not show any structure. One can also plot the residuals versus the predictor variable. It is also useful to draw reference horizontal line through $r = 0$ since residuals should scatter randomly about zero.

Even if the model is correctly specified, the variance of the residuals depends on the x value. To remove this dependence, a re-scaled residual called the *Studentized residual* is often used:

$$\text{Studentized Residual} = \frac{\hat{r}_i}{\sqrt{\text{MSE} \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right]}}.$$

Normal Quantile Plot. The inference procedures discussed above require that the random error ϵ in (2) is approximately normal. In order to assess the normality of the random error, diagnostics can be performed on the residuals. For instance, a formal *goodness-of-fit* test for normality can be performed on the residuals, such as the Shapiro-Wilks test computed by PROC UNIVARIATE in SAS. One can also plot the ordered residuals versus their corresponding normal quantiles (the so-called Q-Q plot). If the error is approximately normal, then the points in the normal quantile plot should lie approximately in a straight line.

4.2 Heteroscedasticity

Recall that for the error ϵ in the regression model we assumed it had a constant variance σ^2 for all x . This assumption is often violated in practice. The violation of this is called **heteroscedasticity**. Figure 4 illustrates a classic case of heteroscedasticity. The left panel shows the data y versus x . The

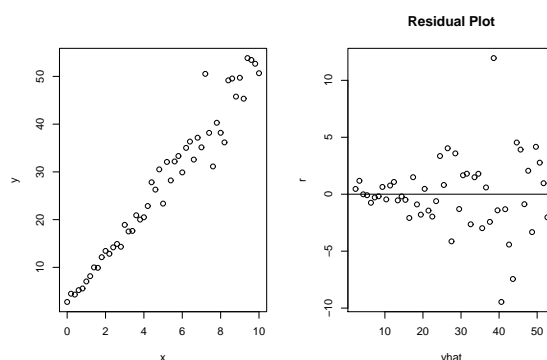


Figure 4: An illustration of heteroscedasticity. Left panel shows raw regression data of y versus x . Right panel shows the corresponding residual plot with a “fan” shape indicating that the equal error variance assumption is violated.

right panel shows the corresponding residual plot versus \hat{y} . The horizontal line is plotted for reference. For larger predicted values, the scatter of the residuals about the horizontal line increases. This residual plot shows the classic “funnel” shape indicative of an unequal error variance. If the equal variance assumption for the error is violated, then the validity of the statistical inference procedures described above will be compromised.

4.3 Dealing with Model Violations

We have illustrated some common ways to determine if there are model violations in a regression setting. The question arises: what next? If there is a problem with the model, how do we fix it? Here are some pointers on how to deal with model violations:

- If the residual plot shows structure, then the model may be misspecified. For instance, a common problem is that the residual plot will show a strong “U”-shape. This is indicative of the need for a quadratic term in the model to accommodate a nonlinear relation between x and y . Other types of terms could also be added to the model to account for nonlinear relationships. This will take us into the realm of multiple regression – see Section 6.

Note: Humans are very adept at finding patterns in data. The problem is that there is often a temptation to see a pattern in a truly random scatter of points. There is an art to residual analysis and determining if there really is structure or not in a residual plot.

- **Dealing with heteroscedasticity.** Perhaps the simplest way to solve an unequal error variance problem is to try a transformation of y and/or x to stabilize the variance. The most common type of transformation used in practice is the *logarithm* transformation which is often very effective in *stabilizing the variance*. However, a transformation will not always solve the problem. Another

common approach is to use a regression model with an unequal error variance built into the model. *Weighted least-squares* can be used to directly model an unequal error variance (details on this can be found in advanced regression textbooks). One problem with weighted least squares is that the weights have to be estimated.

If the response variable represents a count, then a *Poisson* distribution can be considered for modelling the response. The probability function for the Poisson distribution is

$$e^{-\lambda}\lambda^y/y!, \quad y = 0, 1, 2, \dots$$

In a Poisson regression, the unequal variance is expected due to the nature of the count data. Details on Poisson regression can be found in textbooks on *generalized linear models*.

4.4 Outliers and Influential Points

One of the shortfalls of least-squares regression is that the fit of the line can be highly influenced by a few *influential points*. Figure 5 illustrates the problem. The left panel of this figure shows a scatterplot of copper (CU) concentration versus aluminum (Al) concentration in 21 sediment samples from a harbor in Alexandria, Egypt (Mostafa et al. 2004). The solid line is the least-squares regression line for the full data. The far-right point in this plot is quite influential. This point is pulling the regression line towards it. The dashed line in the left panel is the least-squares regression line fit to the data without the influential point. After the influential point is removed, the least-squares regression line changes quite a bit with a smaller slope. The right panel of Figure 5 shows the residual plot for the full data. Also note that there is an outlier in this data set – the point with the corresponding largest positive residual is an outlier. This point does not fall along the same linear pattern as the other points. Perhaps this point corresponds to a location in the harbor where the heavy metal concentration exceeds what is expected in nature, perhaps due to some contamination. Another possibility is that this data point represents a typographical error when the data was recorded.

In multiple regression with several predictor variables, Section 6, it is more difficult to find outliers and influential points due to the high dimensionality of the data. SAS's PROC REG has many built in *options* that allow for the investigation of influential points in multiple regression.

As a final illustration in this section, Figure 6 shows the famous simulated Anscombe data. Each data set in Figure 6 produces an identical simple linear regression least squares line, but the four data sets are vastly different. This data set illustrates the importance of looking at your data in a regression analysis and realizing that different data sets can produce the same regression line.

5 The Correlation Coefficient

The sample correlation coefficient was introduced and discussed in Chapter 4. Pearson's sample correlation coefficient is

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/(n - 1)}{s_x s_y},$$

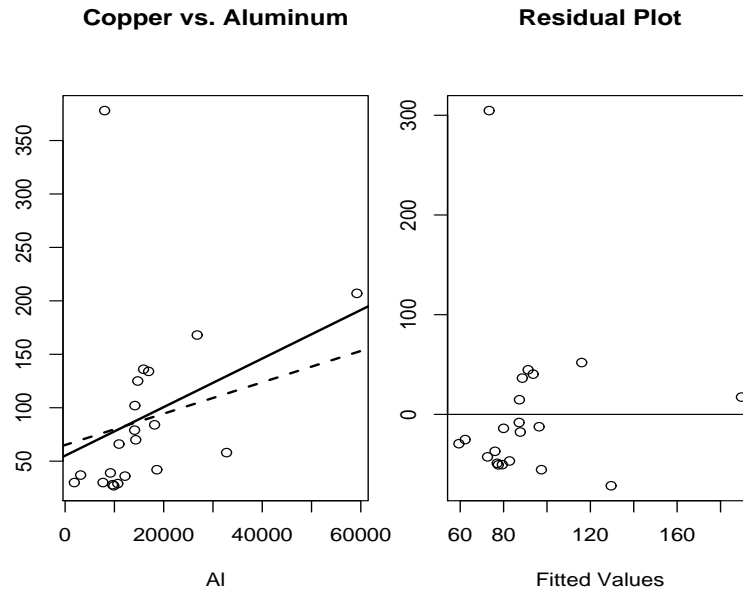


Figure 5: Left Panel: Copper concentration versus Aluminum in 21 sediment samples in a harbor off Alexandria, Egypt. One point is an outlier and the other point to the far right is an influential point. The solid line is the least-squares regression line for the full data and the dashed line is the least-squares regression line after omitting the influential observation. The right panel shows the residual plot for the full data.

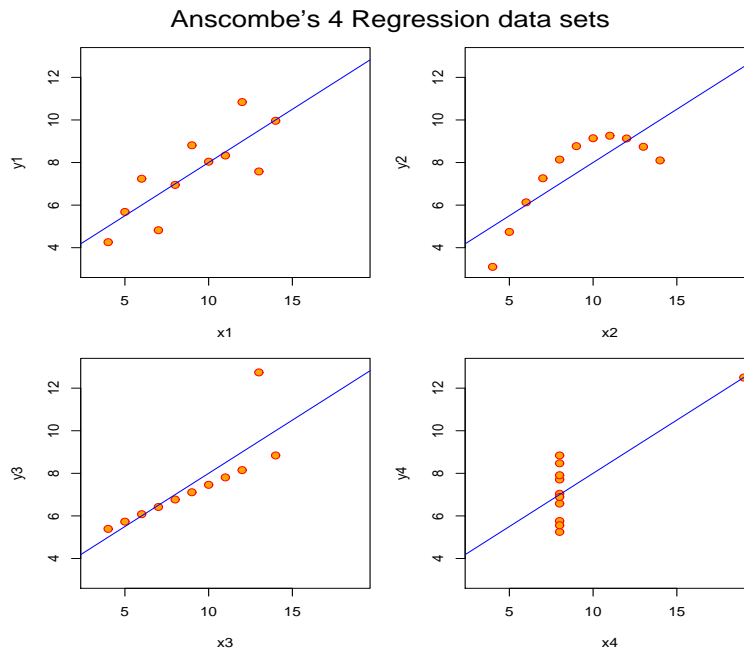


Figure 6: Anscombe's famous simple linear regression data: 4 vastly different data sets all yielding the exact same fitted regression line.

where s_x and s_y are the sample standard deviations of x and y respectively. This relation shows that the least squares estimate of the slope $\hat{\beta}_1$ is related to the correlation by:

$$\hat{\beta}_1 = r \frac{s_y}{s_x}.$$

Thus, if $r = 0$, then $\hat{\beta}_1 = 0$. Just as the estimated slope $\hat{\beta}_1$ estimates the true regression slope β_1 , the sample correlation coefficient r is an estimate for the population correlation coefficient, typically denoted by the Greek letter ρ (“rho”). The formal definition of ρ involves a double integration of the joint density of x and y . The correlation coefficient is a measure of the strength of the *linear* relation between x and y .

Recall however that the correlation coefficient is not really meaningful if the relationship between x and y is nonlinear.

Testing $\beta_1 = 0$ is equivalent to testing if $\rho = 0$. A common approach for testing if the correlation equals a value other than zero is to use *Fisher’s z-transformation*. However, these notes do not cover this topic.

The square of the correlation coefficient r^2 is called the *coefficient of determination* which is discussed in Section 7.

6 Multiple Regression

Now we consider models where there can be more than one predictor variable. In particular, consider models with predictor variables x_1, \dots, x_k :

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon. \quad (13)$$

The method of least-squares can be used in the multiple regression setting to estimate the regression coefficients as in the simple linear regression setting. The idea is to find the values of the coefficients that minimize the sum of squares:

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})^2.$$

The solution can be found using multivariate calculus. The formulas are quite complicated to write out unless matrix notation is used. Statistical software packages such as PROC REG in SAS compute the estimated regression coefficients and their standard errors.

We will illustrate with the follow example:

Example: Predicting Body Fat Percentage. Can accurate body fat predictions be made using easily obtained body measurements? An experiment was conducted where the body fat percentage of $n = 252$ men was measured (using Brozek’s equation). (SOURCE: Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, Provo, Utah 84602). We shall use predictor variables weight (lbs), abdomen (cm), thigh circumference (cm), and wrist circumference (cm) to predict body fat percentage in a multiple regression model:

$$\text{body fat \%} = \beta_0 + \beta_1(\text{Abdomen}) + \beta_2(\text{Weight}) + \beta_3(\text{Thigh}) + \beta_4(\text{Wrist}) + \text{Error}.$$

The SAS code to run this multiple regression is

```
proc reg;
  model fat1 = abdomen wt thigh wrist;
run;
```

The SAS output follows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	11009	2752.22165	167.02	<.0001
Error	247	4070.13002	16.47826		
Corrected Total	251	15079			

Root MSE	4.05934	R-Square	0.7301
Dependent Mean	18.93849	Adj R-Sq	0.7257
Coeff Var	21.43435		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-31.73071	7.79205	-4.07	<.0001
abdomen	1	0.90523	0.05193	17.43	<.0001
wt	1	-0.13380	0.02867	-4.67	<.0001
thigh	1	0.15370	0.10114	1.52	0.1299
wrist	1	-1.00421	0.41367	-2.43	0.0159

The regression models are part of a large class of models called **general linear models**. This class of models includes simple linear regression, multiple linear regression, analysis of variance (ANOVA), analysis of covariance (ANCOVA). The classic inference procedure for general linear models is to partition the variability in the response y into parts that depend on the predictor variable(s) and the random error.

The *total* variability in the response y can be computed by squaring the all the deviations $y_i - \bar{y}$ and adding them up to form the **total sum of squares** or **SS(Total)**:

$$SS(\text{Total}) = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (14)$$

Next, we re-write each deviation as

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \quad (15)$$

If we square both sides of (15) and sum up from $i = 1$ to n , we get the analysis of variation partition:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (16)$$

which can be expressed as

$$\text{SS(Total)} = \text{SS(Regression)} + \text{SS(Error)}.$$

Here the regression sum of squares (SS(Regression)) is defined as

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

and the Error (or Residual) sum of squares (SS(Error)) is defined as

$$\text{SS(Error)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Mean Squares The regression mean square is obtained by taking the regression sum of squares and dividing it by its degrees of freedom:

$$\text{MS(Regression)} = \frac{\text{SS(Regression)}}{k},$$

where k is the number of predictor variables. In simple linear regression, $k = 1$. However, in multiple regression where we can have more than one predictor, k can be greater than one.

The **mean square error** (MSE) is the error sum of squares divided by its degrees of freedom:

$$\text{MSE} = \text{MS(Error)} = \frac{\text{SS(Error)}}{n - k - 1}.$$

In simple linear regression, the denominator is $n - 2$ since $k = 1$: we lose two degrees of freedom for estimating the y -intercept and slope. The mean square error is an unbiased estimator of the error variance σ^2 . Therefore, the mean square error (MSE) is also denoted by $\hat{\sigma}^2$.

The ANOVA table at the top is used to test the overall hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

versus

$$H_a : \text{at least one of the } \beta_1\text{'s differ from zero.}$$

The F Probability Distribution. In order to test this hypothesis, the F probability distribution needs to be introduced. In this setting, the F -test statistic is defined as

$$F = \frac{\text{MS(Regression)}}{\text{MSE}}.$$

If the null hypothesis is true, then both mean squares are estimating the same quantity which is the variance σ^2 of the error of the multiple regression model (13). Thus, if H_0 is true, the F -test statistic takes the value $F = 1$ on average. However, if H_0 is false, then the F -test statistic tends to be larger than one.

The F -probability distribution is our reference distribution to be used to determine if the F -test statistic is too big to be attributed to chance and hence leading to a rejection of H_0 . From the SAS output, we see that the observed F -test statistic is considerably larger than one: $F = 167.02$ and the corresponding p -value is less than 0.0001 according to SAS. In other words, we have very strong evidence that at least one of the regression coefficients is different from zero which means that at least one of the predictors explains variability in fat percentage.

To find out which predictor variables are “significant” we can look at the partial t -test statistics at the bottom of the PROC REG output. If we let $\hat{\beta}_j$ denote the least-squares estimator of β_j from (13), then the partial t -test statistic used to test

$$H_0 : \beta_j = 0 \text{ versus } H_a : \beta_j \neq 0$$

is

$$t = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)},$$

which, under the null hypothesis will follow a t -distribution on $n - k - 1$ degrees of freedom (where k equals the number of predictor variables). The formulas for $\widehat{SE}(\hat{\beta}_j)$ are complicated to express without the use of matrix algebra and we omit them here. SAS’s PROC REG automatically prints out the results of these tests for each predictor variable. It is important to note because multiple t -tests are being performed, the type I error rate will tend to be inflated – this problem is discussed in more detail in the next chapter.

Interpretation of Regression Coefficients. In the simple linear regression setting, the slope β_1 represents the mean change in the response variable for a unit change in the predictor variable. In the multiple regression setting, a similar interpretation can be used for the regression coefficients β_j ’s:

β_j represents the mean change in the response variable for a unit change in the predictor variable x_j provided all the other predictor variables are held fixed.

Here is the problem: typically in a multiple regression, the predictor variables will be correlated with one another. Thus, it usually does not make sense to consider the change in the response for a unit change in a predictor while holding the other predictors fixed because if one predictor changes, then the other predictors tend to change as well. For instance, in the body fat example, as a person’s weight increases, their abdomen tends to increase as well. Therefore, it is usually very difficult to interpret the estimated slope coefficients in a multiple regression setting.

Assessing Goodness of Fit. In a multiple regression setting, assessing the fit of the model is often much more difficult than in the simple linear regression setting due to the high dimensionality of the model. We cannot generate a single plot to observe the response and all the predictors at once. Instead, we can plot residuals versus fitted values and also versus the individual predictor variables to look for problems with the fit of the model. If any of the residual plots show some structure, then the model is not specified correctly. In a multiple regression, it is also more difficult to find outliers and influential points using only plots of the data because the plots generally only allow us to view two or three variables at a time. SAS’s PROC REG has many influential diagnostic tools to help find influential points and their effects on the fitted regression equation.

Collinearity. The fitted model can become very unstable if the predictor variables are highly correlated. This problem is known as (multi)collinearity. One of the simplest solutions to the problem is to delete predictor variables from the model. The question then becomes: which predictor variables should we throw out? There are many model building algorithms that help deal with this problem (e.g. forward selection, backwards elimination, stepwise regression). The *lasso* (Tibshirani, 1996) is a more modern approach to model building which does not suffer from many of the problems of the older model selection algorithms. Instead of throwing out predictor variables, there are other ways to deal with collinearity. A couple common examples are ridge regression and principal component regression which are discussed in more advanced regression textbooks.

7 Coefficient of Determination

A very popular statistic used in regression analysis is the coefficient of determination, often referred to as the R^2 (“R-squared”). In a simple linear regression, R^2 is just the squared correlation coefficient. The coefficient of determination R^2 measures the proportion of the variation in the response that is explained by the regressor variable(s):

$$R^2 = \frac{\text{SS(Regression)}}{\text{SS(total)}} = 1 - \frac{\text{SS(Error)}}{\text{SS(Total)}}.$$

R^2 is usually reported whenever a regression analysis is performed and generally investigators like to see large values of R^2 . However, what is considered “large” varies from application to application. For instance, in many engineering applications, one may expect R^2 ’s around 0.95 or higher. On the other hand, in an environmental study of a plant’s size, an R^2 of around 0.40 may be considered large. In an engineering example, most of the factors that affect a response may be accounted for in an experiment. However, the growth of a plant can depend on many influential factors and an experiment may only account for a tiny fraction of these factors.

The following list highlights some important points about R^2 :

- $0 \leq R^2 \leq 1$.
- $R^2 = 1$ implies all the points lie exactly on a line in simple linear regression.
- R^2 equals the square of the sample correlation coefficient between x and y in simple linear regression.
- R^2 becomes larger (or does not get smaller) as you add more regressors to the model. Choosing a regression model only on the basis of a large R^2 will typically lead to unstable models with too many predictors. For instance, we can increase the R^2 in a simple linear regression of y on x by adding the quadratic term x^2 to the model. We can also add a cubic term x^3 to make the R^2 even bigger. However, these *polynomial* models can become very unstable.

Here are a couple common misconceptions about R^2 :

- A large R^2 does not necessarily mean the regression model will provide useful predictions.
- A large R^2 does not necessarily mean that the model is correctly specified. Figure 7 illustrates this problem. The scatterplot shows a nonlinear relation between x and y but yet the $R^2 = 0.93$ is quite high.

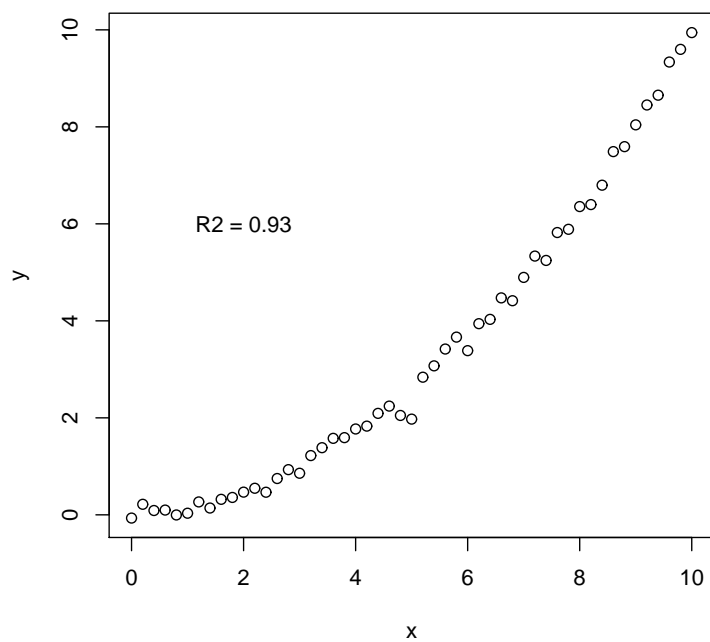


Figure 7: Scatterplot showing a nonlinear relationship between x and y but yet the $R^2 = 0.93$. Thus, a high R^2 does not necessarily mean the relation is linear.

8 Conclusion

This chapter presented a brief overview of regression analysis. As pointed out in the notes, there are several more advanced topics in regression analysis. Regression analysis can handle more complicated models than those presented here. For instance, nonlinear regression analysis can be used to model complicated (or non-complicated) nonlinear relations. These nonlinear models typically require numerical algorithms to find estimators. Often the relation between a response variable and a predictor is completely unknown. A branch of regression analysis called nonparametric regression can be used to allow the data to determine the shape of a regression curve. Multivariate regression is used to model several response variables as a function of one or more predictor variables. Binary indicator variables can be added to regression models to allow the model to distinguish between different groups (e.g. male & female) in the population – this leads to *analysis of covariance*. Functional data analysis is another branch of statistics where each data point is a curve. For instance, in a clinical study, subjects may be observed over time and a curve can be estimated for each individual subject to model the subject's response to treatment over time.

9 Problems

1. It is well known that the height of a plant depends on the amount of fertilizer the plant receives. A study was done by varying the amount of fertilizer plants received and the heights of each of the plants were the measured. The correlation between the amount of fertilizer and the heights of the plants was found to not differ significantly from zero (this is equivalent to testing if the slope in a regression analysis differs significantly from zero). The experimenter was expecting to see a significant positive correlation. Can you suggest two possible explanations why the correlation did not differ significantly from zero?
2. The Federal Trade Commission rates different brands of domestic cigarettes. In their study, they measured the amount of carbon monoxide (co) in mg. and the amount of nicotine (mg.) produced by a burning cigarette of each brand. A simple linear regression model was run in SAS using co as the dependent variable and nicotine as the independent (or predictor) variable. The output is shown below – use this output to answer the questions below.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	462.25591	462.25591	138.27	<.0001
Error	23	76.89449	3.34324		
Corrected Total	24	539.15040			
	Root MSE	1.82845	R-Square	0.8574	
	Dependent Mean	12.52800	Adj R-Sq	0.8512	
	Coeff Var	14.59493			

Parameter Estimates

Parameter	Standard
-----------	----------

Variable	DF	Estimate	Error	t Value	Pr > t
Intercept	1	1.66467	0.99360	1.68	0.1074
nicotine	1	12.39541	1.05415	11.76	<.0001

- a) How many cigarette brands did the Federal Trade Commission evaluate?
 - b) What percentage of variation in carbon monoxide is explained by the amount of nicotine?
 - c) Write down the least-squares regression equation.
 - d) Does the amount of carbon monoxide produced by a burning cigarette depend on how much nicotine is in the cigarette? Perform an appropriate hypothesis test to answer this question. Be sure to define any parameters you use. Base your conclusion on the p -value of the test.
 - e) Give an estimate of how much carbon monoxide increases on average for each additional milligram of nicotine in the cigarette.
 - f) What is the predicted amount of carbon monoxide produced by a cigarette containing 1.5 mg of nicotine?
3. A recent study investigated the concentration y of arsenic in the soil (measured in mg/kg) and the distance x from a smelter plant (in meters). A simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$ was used to model the relationship between these two variables. The data from the study was used to test the hypothesis of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 < 0$ at a significance level $\alpha = 0.05$. Do the following parts:
- a) In the context of this problem, what would be a type I error?
 - b) In the context of this problem, what is a type II error?
 - c) The least-squares regression line was found to be $\hat{y} = 60.2 - 5.1x$ and the standard error of the estimated slope was found to be 1.32. What is the value of the t -test statistic for testing the hypothesis in this problem?
 - e) The null hypothesis was rejected in this test at $\alpha = 0.05$. Write a sentence interpreting the estimated slope in the context of this problem.
4. The SAS program chlorophyll.sas in the Appendix has data on the concentration of chlorophyll-a in a lake along with the concentration of phosphorus and nitrogen in the lake (Source: Smith and Shapiro 1981). Chlorophyll-a is used as an indicator of water quality that measures the density of algal cells. Phosphorus and nitrogen stimulate algal growth. The purpose of this homework is to use a simple linear regression model to model the relationship between $\log(\text{chlorophyll-a})$ and $\log(\text{phosphorus})$ in the lake. To obtain answers to the questions, run the SAS program.
- a) Write down the equation for the estimated least-squares regression line relating the logarithm of chlorophyll-a to $\log(\text{phosphorus})$.
 - b) Write a sentence that interprets the estimated slope in the context of this problem.
 - c) What is the estimated standard error of the estimated slope of the regression line?
 - d) SAS reports a t -statistic value of $t = 10.81$ associated with $\log(\text{phosphorus})$. Write down the regression model and write out the null and alternative hypothesis that is being tested with this t -test statistic.
 - e) Write a sentence interpreting the p -value that is associated with the test statistic in part (d).

- f) Compute a 95% confidence interval for the slope of the regression line and write a sentence interpreting this confidence interval.
5. Data on the velocity of water versus depth in a channel were acquired at a station below the Grand Coulee Dam at a distance of 13 feet from the edge of the river. Modeling water velocity in a channel is important when calculating river discharge. (Reference: Savini, J. and Bodhaine, G. L. (1971), Analysis of current meter data at Columbia River gaging stations, Washington and Oregon; USGS Water Supply Paper 1869-F.) Run the SAS program ColumbiaRiver.sas (which contains the data) in the Appendix to do the following parts:
- Write down the equation for the estimated least-squares regression line relating water velocity to depth.
 - What is the estimated water velocity at the surface of the channel?
 - Find a 95% confidence interval for the slope of the regression line. Write a sentence that interprets this confidence interval in the context of this problem.
 - Generate a plot of the data along with the least-squares regression line and attach it to this assignment. Does the relation between water depth and water velocity appear linear?

References.

Alaa r. Mostafa, Assem O. Barakat, Yaorong Qian, Terry L. Wade, Dongxing Yuan, (2004), "An Overview of Metal Pollution in the Western Harbour of Alexandria, Egypt", *Soil & Sediment Contamination*, **13**, 299311.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288)

10 Appendix

```

/*****
      Chlorophyll-a in Lakes
Source: Smith and Shapiro (1981) in Environmental Science and
      Technology, volume 15, pages 444-451.
Data:
Column 1: Chlorophyll-a
Column 2: Phosphorus
Column 3: Nitrogen
*****/
data lake;
input chlor phos nitro;
logc=log(chlor);
logp=log(phos);
logn=log(nitro);
datalines;
95.0 329.0 8
39.0 211.0 6

```

```

27.0 108.0 11
12.9 20.7 16
34.8 60.2 9
14.9 26.3 17
157.0 596.0 4
5.1 39.0 13
10.6 42.0 11
96.0 99.0 16
7.2 13.1 25
130.0 267.0 17
4.7 14.9 18
138.0 217.0 11
24.8 49.3 12
50.0 138.0 10
12.7 21.1 22
7.4 25.0 16
8.6 42.0 10
94.0 207.0 11
3.9 10.5 25
5.0 25.0 22
129.0 373.0 8
86.0 220.0 12
64.0 67.0 19
;
proc print;
run;
proc reg;
    model logc = logp;
    output out=a p=p r=r;
run;
proc plot;
    plot r*p/vref=0;
quit;

```

```

/*****
COLUMBIA RIVER VELOCITY-DEPTH
Data on the velocity (ft/sec) of water versus depth (feet)
in a channel were acquired at a station below
the Grand Coulee Dam at a distance of 13 feet from the edge of
the river. Modeling water velocity in a channel is important when
calculating river discharge. (Reference: Savini, J. and Bodhaine, G. L.
(1971), Analysis of
current meter data at Columbia River gaging stations, Washington and
Oregon; USGS Water Supply Paper 1869-F.)
*****/

```

```
options ls=76;
data water;
input depth velocity;
datalines;
0.7 1.55
2.0 1.11
2.6 1.42
3.3 1.39
4.6 1.39
5.9 1.14
7.3 0.91
8.6 0.59
9.9 0.59
10.6 0.41
11.2 0.22
;
run;
proc reg;
    model velocity = depth;
run;
quit;
```