MODULE 4

SIMPLE LINEAR REGRESSION

Module Objectives:

1. Describe the equation of a line including the meanings of the two parameters.
2. Describe how the best-fit line to a set of bivariate data is derived.
3. Understand how to construct hypothesis tests and confidence intervals for parameters and predictions.
4. Describe the difference between confidence and prediction intervals related to predictions.
5. Describe the importance of the default hypothesis tests for parameters.
6. Show all interrelationships among coefficient, ANOVA, and summary computations.
7. Describe why meeting the assumptions of a regression analysis is important.
8. Describe what a residual is.
9. Describe how to assess the four major assumptions of linear regression.
10. Describe the importance of transforming variables.
11. Describe three methods for choosing an appropriate variable transformation.
12. Understand the concept of competing models and their relationship to hypothesis tests in SLR.
13. Understand how to interpret the results in an ANOVA table.

A simple linear regression (SLR) is used when a single quantitative response and a single quantitative explanatory variable are considered.[1] The goals of SLR are to use the value of the explanatory variable to (1) predict future values of the response variable and (2) explain the variability of the response variable. A simple linear regression would be used in each of these situations:

1. Predict annual consumption for a species from a biomass estimate.
2. Evaluate the variability of porcupine body mass based on days since the beginning of winter.
3. Evaluate the variability in clutch size relative to length of female spiders.
4. Predict daily energy consumption from the body weight of penguins.
5. Predict change in duck abundance from the loss of wetlands.
6. Predict plant shoot dry mass from total leaf length.

◇ **The two major goals of SLR are to use the explanatory variable to (1) predict a future value of the response variable and (2) explain the variability of the response variable.**

4.1 Foundational Review

4.1.1 Variable Definitions

In SLR, the variable that will be predicted or have its variability explained, is called the *response variable*.[2] The other variable that will be used to make predictions and to help explain the variability of the response variable is called the *explanatory variable*.[3] The response variable is always plotted on the y-axis.

△ **Response Variable**: The variable to be predicted or have its variability explained.

△ **Explanatory Variable**: The variable used to predict or explain the variability of the response variable.

4.1.2 Line Equation

Both goals of SLR are accomplished by finding the model[4] that best fits the relationship between the response and explanatory variables.[5] When examining statistical regression models, the most common expression for the equation of the best-fit line is

$$\mu_{Y|X} = \alpha + \beta_1 X \tag{4.1.1}$$

where Y represents the response variable, X the explanatory variable, α the y-intercept and β_1 the slope. The left-hand-side (LHS) of Equation (4.1.1) is read as "the mean of Y at a given value of X." This terminology is used because the best-fit line actually models the mean values of Y at each value of X rather than the

[1] Indicator variable regression (IVR; see Module 5 is used when multiple explanatory variables are present and some of those are factor variables. Multiple linear regression (MLR) will be used when multiple quantitative explanatory variables are present.

[2] Some call this the *dependant variable*.

[3] Some call this the *independent variable*.

[4] "Model" is used here instead of "line" because in this class, models other than lines will be fit to some data. However, all models will be fit on a "scale" where the form of the relationship between the response and explanatory variable is linear.

[5] It is assumed that you covered the basics of simple linear regression in your introductory statistics course. Thus, parts of this section will be review, although the nomenclature used may differ somewhat.

individuals themselves. The model in Equation (4.1.1) is for the population.[6] The equation of the best-fit line based on the information in the sample is written as

$$\hat{\mu}_{Y|X} = \hat{\alpha} + \hat{\beta}_1 X \tag{4.1.2}$$

In other words, Equation (4.1.1) is essentially a parameter (i.e., it represents the population line) and Equation (4.1.2) is a statistic (i.e., it estimates a parameter based only on the information in a sample). Thus, α and β_1 represent the population y-intercept and slope, respectively; whereas $\hat{\alpha}$ and $\hat{\beta}_1$ represent the sample y-intercept and slope.

4.1.3 Best-Fit Line

The statistics $(\hat{\alpha}, \hat{\beta}_1)$ in Equation (4.1.2) are determined by finding the values for (α, β_1) that minimize the residual sum-of-squares (RSS). A residual is the difference between an observed value of the response variable for an individual, y_i, and a predicted value of the response variable based on Equation (4.1.1) and the individual's value of the explanatory variable, x_i (Figure 4.1). Thus, a residual is $y_i - \hat{\mu}_{Y|X=x_i}$.

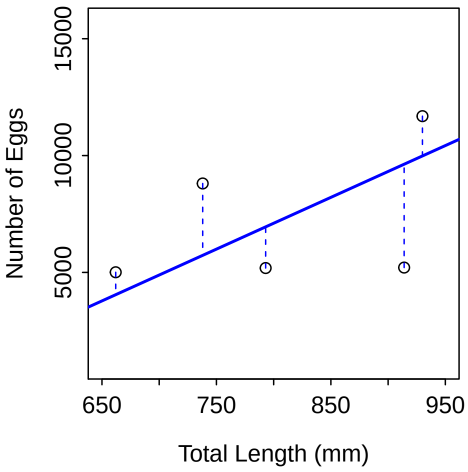

Figure 4.1. Scatterplot with best-fit line illustrating five residuals.

△ **Residual**: The difference between the observed value of the response variable for an individual and the predicted value of the response variable using the best-fit line; i.e., $y_i - \hat{\mu}_{Y|X=x_i}$.

The RSS is the sum of the squares of these residuals, or

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\mu}_{Y|X=x_i}\right)^2$$

It can be shown with calculus that the RSS is minimized with a slope given by

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

[6] Note that Greek letters are usually reserved for parameters.

and a y-intercept given by

$$\hat{\alpha} = \overline{Y} - \hat{\beta}_1\overline{X} \tag{4.1.3}$$

where r is the sample correlation coefficient, $\overline{Y}$ and $\overline{X}$ are the sample means, and s_Y and s_X are the sample standard deviations of the response and explanatory variables, respectively.

Δ **RSS**: Residual Sum-of-Squares

◇ **The best-fit line is the line of all possible lines that minimizes the RSS.**

4.1.4 Best-Fit Line in R

The best-fit regression line is obtained with `lm()` with a formula of the form `response∼explanatory`. As usual, the results of this function should be assigned to an object that can be given to other functions to extract specific results. For example, the simple linear regression for predicting the number of eggs based on the total length of a Lake Superior female lake trout is computed below.

```
> LT <- read.csv("LakeTroutEggs.csv")

> ( lt.lm <- lm(eggs~tl,data=LT) )
Coefficients:
(Intercept)           tl
  -10620.52        22.15
```

These results show that the slope of the best-fit line is 22.15 and the intercept is -10620.52. Thus, it is predicted that the number of eggs will increase by about 22.15, on average, for each 1 mm increase in total length of a female lake trout. Because these data are not longitudinal this result is best stated for a 1 mm difference in total length of two female lake trout; i.e., a female lake trout that is 1 mm longer than another female lake trout will have an average of 22.15 more eggs.

A fitted-line plot is made by super-imposing the best-fit line onto a scatterplot of the data (Figure 4.2). Fitted-line plots are constructed with `fitPlot()` in essentially the same way it was used in previous modules.

```
> fitPlot(lt.lm,ylab="Number of eggs",xlab="Total Length")
```

4.2 Inferences

4.2.1 Slope & Intercept

Just like every other statistic, $\hat{\alpha}$ and $\hat{\beta}_1$ are subject to sampling variability and, thus, have sampling distributions. The sampling distributions of $\hat{\alpha}$ and $\hat{\beta}_1$ are normally distributed (if the assumptions of SLR are met; see Section 4.3) and unbiased.[7] The standard error of the sample y-intercept is given by

$$SE_{\hat{\alpha}} = \sqrt{s^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s^2_X}\right)}$$

[7] Think back to your introductory statistics course to remind yourself what it means for a statistic to be unbiased.

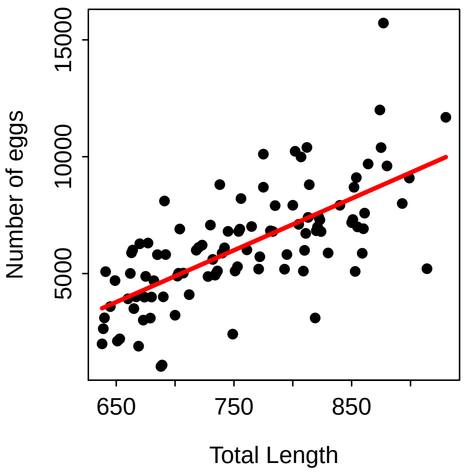

Figure 4.2. Number of eggs versus total length for Lake Superior lake trout with best-fit line superimposed.

and the standard error of the sample slope is given by

$$SE_{\hat{\beta}_1} = \sqrt{\frac{s^2_{Y|X}}{(n-1)s^2_X}}$$

where $s^2_{Y|X}$ is the variance that measures the natural variability of individuals around the line.[8]

◇ **The sample slope and y-intercept are statistics that have sampling distributions and standard errors.**

More complete results from fitting the SLR model to the lake trout egg data are shown in Table 4.1. All results for the y-intercept are in the row labeled with "(Intercept)." All results for the slope are in the row labeled with the variable name of the explanatory variable (*tl* in this example). The estimated y-intercept and slope values are in the column labeled "Estimate." The standard errors for the sample y-intercept and slope are in the column labeled "Std. Error."

Table 4.1. Least-squares regression results for the model $\mu_{eggs|tl} = \alpha + \beta_1 tl$

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10620.525   1787.520  -5.941 4.22e-08
tl              22.155      2.346   9.442 1.81e-15
---
Residual standard error: 1795 on 99 degrees of freedom
Multiple R-squared: 0.4738,Adjusted R-squared: 0.4685
F-statistic: 89.15 on 1 and 99 DF,  p-value: 1.806e-15
```

Two common hypothesis tests in SLR are to determine whether or not the population y-intercept or the population slope is equal to a particular value. These hypotheses are written in the typical format as

$$H_A : \alpha \neq \alpha_0$$
$$H_A : \beta_1 \neq \beta_{10}$$

[8] The best-fit line does not perfectly represent every individual. This statistic measures how individuals scatter around the line.

where α_0 and β_{10} represents a specific value for α and β_1. As an example,[9] the hypothesis for the slope is tested with the following test statistic

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{SE_{\hat{\beta}_1}} \tag{4.2.1}$$

with $df = n - 2$. Familiarly, a confidence interval for β is constructed with

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1}$$

The test statistic and confidence interval for the intercept test is constructed similarly.

> ◇ **Hypothesis tests and confidence intervals for the population slope and population y-intercept are performed with the same general formulas for t- test statistics and confidence intervals learned in your introductory statistics course. The major difference being that $df = n - 2$.**

As an example, suppose that interest is in determining whether or not a significant portion of the variability in the number of eggs in mature female lake trout can be explained by the total length of the fish. This question translates into a hypothesis to determine if the response and explanatory variable are significantly related. This in turn translates into a simple hypothesis test to determine if the population slope is equal to zero or not.[10] Thus, the statistical hypotheses to be tested are

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

The test statistic for testing this hypothesis is $t = \frac{22.15-0}{2.35} = 9.44$.[11] Thus, with $df = 99$, the p-value is exceptionally small and the null hypothesis is soundly rejected resulting in a conclusion that the number of eggs per female is, in fact, significantly related to the length of the female lake trout. A 95% confidence interval for the population slope gives an estimate of the direction and magnitude of the relationship. A 95% confidence interval for the slope is $22.15 \pm 1.984 * 2.35$ or $(17.50, 26.81)$. Thus, one is 95% confident that the true change in the mean number of eggs with a one mm increase in total length is between 17.50 and 26.81.

A careful examination of Table 4.1 shows that the test statistic and p-value for this hypothesis test have already been calculated.[12] In fact, these t- and p-values are for testing the very common specific hypotheses of whether the model parameter equals zero or not; e.g., tests if the population slope is different from zero.

It is important to note that the t- and p-values printed in this output are only useful for testing the particular hypotheses that the corresponding parameter is equal to zero; all other hypothesis tests are NOT computed automatically by common software packages. However, it is also important to note that testing that the slope is equal to zero is of great importance in linear regression because it determines whether the explanatory and response variable are significantly related or not.

> ◇ **The default p-values printed by most softwares are ONLY for the specific H_0 that the corresponding parameter is equal to zero (vs. that it is not).**

[9] It is assumed that you are familiar with the following test statistic and confidence interval formula from the 1-sample t-test taught in your introductory statistics course.

[10] You should convince yourself that a test of whether or not the slope is equal to zero, is also a test of whether or not the response and explanatory variable are significantly related

[11] The values used in this test statistic come from Table 4.1.

[12] As an example, compare the results in the slope row from Table 4.1 to the results in the previous paragraph.

◇ The test of whether the population slope equals zero or not is also a test of whether the response and explanatory variable are significantly related.

Slope & Intercept Inferences in R

The information for computing hypothesis tests about the slope and y-intercept (i.e., the results shown in Table 4.1) is obtained by submitting the fitted `lm()` object to `summary()`. In addition, confidence intervals for each parameter in a linear model can be obtained by submitting the saved `lm()` object to `confint()`. The 95% confidence interval for the population slope is found in the row labeled with the explanatory variable. Thus, in this example, one is 95% confident that the population slope is between 17.5 and 26.8.[13]

```
> confint(lt.lm)
                   2.5 %      97.5 %
(Intercept) -14167.35175 -7073.69789
tl              17.49882    26.81051
```

A simultaneous confidence "region" for both parameters of an SLR forms an ellipse (Figure 4.3). The region shown in Figure 4.3 is set to contain the point (α, β_1) with 95% confidence. In other words, one is 95% confident that *both* α and β_1 are simultaneously contained within the ellipse shown. Interestingly, projections of the extremes of the ellipse onto the "Intercept" (i.e., "X") and "slope" (i.e., "Y") axes form univariate 95% confidence intervals for the intercept and slope parameters, respectively.[14]

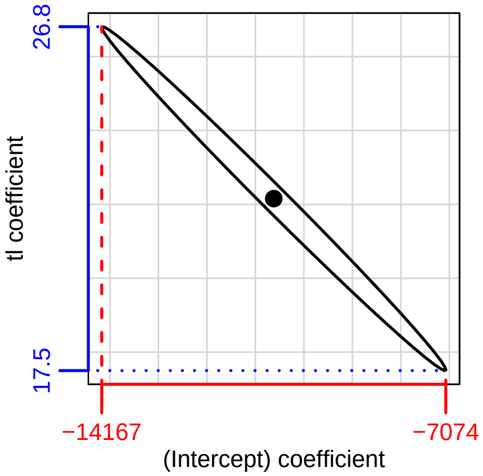

Figure 4.3. Confidence ellipse for α and β_1 from the regression of number of eggs versus total length for Lake Superior lake trout.

Another interesting result illustrated with Figure 4.3 is that the regression parameters are highly correlated. In this instance, as the intercept increases the slope decreases. Not all regression parameters are highly correlated, but many are. This may cause some difficulty in some situations; corrections will be addressed in subsequent sections.

[13] Note that this is the same interval computed previously "by hand."

[14] Note how these projections have the same endpoints as the results from `confint()`.

The testing of hypotheses comparing the slope and intercept to values other than zero can be efficiently computed with `hoCoef()`.[15] This function requires the saved `lm()` object as its first argument, a number representing the term in the model to be tested (in SLR, `term=1` is the intercept and `term=2` is the slope), and the null hypothesized value (i.e., α_0 or β_{10}) in `bo`. In addition, the direction of the alternative hypothesis can be defined with `alt=`. For example, the code below is used to test whether the true slope is greater than 20 eggs (i.e., $H_A : \beta_1 > 20$). Thus, there is very little evidence ($p = 0.180$) that the increase in number of eggs for a 1 mm increase in length is significantly greater than 20.[16]

```
> hoCoef(lt.lm,term=2,bo=20,alt="greater")
 term Ho Value Estimate Std. Error         T df   p value
    2       20 22.15467    2.34644 0.9182716 99 0.1803542
```

4.2.2 Centering & Intercept

Most hypothesis tests in SLR are related to the slope, though interest is in the intercept in some instances. However, the interpretation of the y-intercept is often nonsensical because $X = 0$ is not within the domain of the explanatory variable. In other words, the value of the response variable when the explanatory variable is zero is often an extrapolation and can result in exceptionally odd statements. This problem is further exacerbated because, as will be shown in the next section, the variability around the model increases further away from the mean of the explanatory variable. Thus, if $X = 0$ is considerably outside the range of the data (thus, far from $\bar{x}$) then the variability at the intercept will be large and statistical power will be low.

One method for constructing an intercept that is meaningful is to center the explanatory variable to zero. Variables are re-centered to zero by subtracting the mean of that variable from each observed value of that variable. In other words, the new variable, X^*, is formed from $X - \bar{x}$.

Centering the explanatory variable shifts the distribution from being centered on $\bar{x}$ to being centered on 0 (Figure 4.4). The interpretation of the y-intercept changes with this shift from representing the average value of Y when $X = 0$ to representing the average value of Y when $X = \bar{x}$. However, the slope and its interpretation is unchanged as are all predictions.[17]

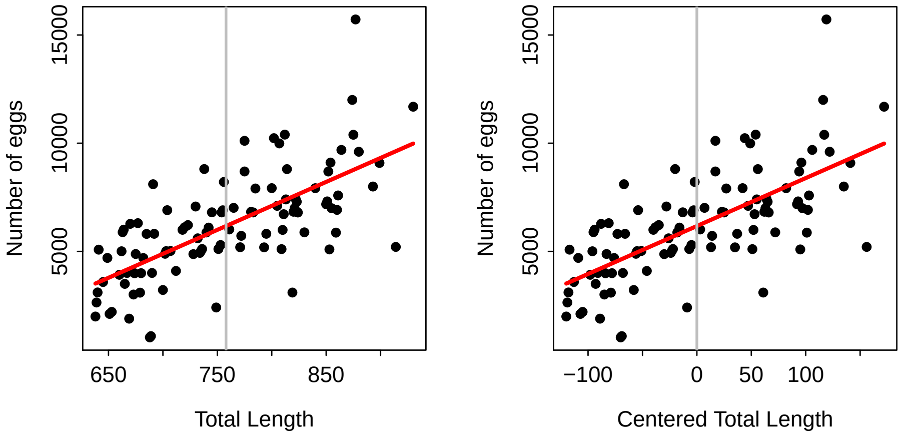

Figure 4.4. Number of eggs versus total length (Left) and centered total length (Right). The vertical gray lines are the mean total lengths (left) and mean centered total length (=0; right).

[15] This function is a very simple function that computes the test statistic as defined in Equation (4.2.1) and then computes the p-value from the t-distribution with the `pt()` function.

[16] This is not a surprising result given the confidence interval for the slope computed above.

[17] The researcher must remember to center the value of X though before using the model to predict the mean value of Y.

As an example, the linear regression of number of eggs in female lake trout on total length and centered total length is shown in Tables 4.1 and 4.2, respectively. Note that the estimated intercept terms in each model are dramatically different as is their corresponding standard errors. The intercept term from the un-centered model represents the mean number of eggs in a lake trout with a total length of 0. In contrast, the intercept term from the centered model represents the mean number of eggs in a lake trout with an average total length. Further note that all other results are exactly the same between the two models.

Table 4.2. Least-squares regression results for the model $\mu_{eggs|tl} = \alpha + \beta_1(tl - \overline{tl})$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6172.495    178.568  34.567  < 2e-16
c.tl          22.155      2.346   9.442 1.81e-15
---
Residual standard error: 1795 on 99 degrees of freedom
Multiple R-squared: 0.4738,Adjusted R-squared: 0.4685
F-statistic: 89.15 on 1 and 99 DF,  p-value: 1.806e-15
```

In general, centering the explanatory variable is not critical unless the interpretation of the intercept term is important. However, a side effect of centering the explanatory variable is that the estimated slope and intercept are then orthogonal – which, for all practical purposes, means uncorrelated (Figure 4.5). This characteristic allows the centering of explanatory variables to have other uses and positive impacts in multiple linear regressions.

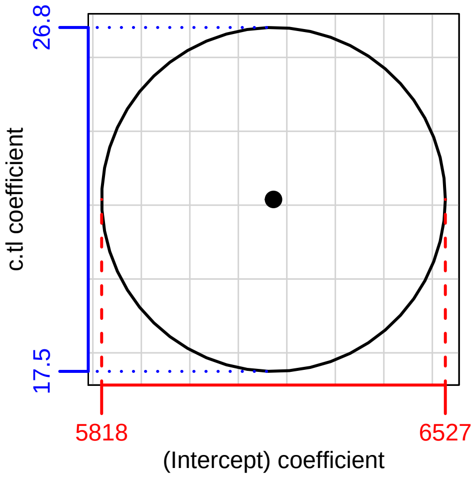

Figure 4.5. Confidence ellipse for α and β_1 from the regression of number of eggs versus CENTERED total length for Lake Superior lake trout.

Centering Variables in R

A variable is centered in R by subtracting the mean value from all measurements of the explanatory variable. The mean of all measurements is computed with `mean()`. For example, the centered total length of lake trout in the lake trout eggs example is obtained from the original *tl* as shown below. This new variable can then be used in `lm()` to perform a regression using the centered explanatory variable.

```
> LT$ctl <- LT$tl - mean(LT$tl)
```

4.2.3 Predicting Mean & Individual Values

One of the major goals of linear regression is to use the best-fit line and a known value of the explanatory variable to predict a future value of the response variable. This prediction is easily made by plugging the known value of the explanatory variable (generically labeled as x_0) into the equation of the best-fit line for X. Generically, this is

$$\hat{\mu}_{Y|X=x_0} = \hat{\alpha} + \hat{\beta}_1 x_0$$

For example, the predicted number of eggs for a 700-mm female lake trout is computed by plugging 700 into $\hat{\mu}_{eggs|tl=700}$=-10620.52+22.15*700 = 4887.74 or 4888 eggs.

This prediction is the best estimate of the **mean** number of eggs for all 700-mm individuals and is, thus, also the best guess at the number of eggs for a 700-mm long individual. So, in essence, this one calculation accomplishes two things: (1) predicts the **mean** value of the response variable for **all** individuals with a given value of the explanatory variable and (2) predicts the value of the response variable for **an** individual with a given value of the explanatory variable. To keep these two items separate, the first objective (predict the mean) is called finding a *fitted value* because the best-fit line actually "fits" the mean values of Y at a given value of X. The second objective (predict the individual) is called finding a *predicted value* because "predictions" are generally made for individuals.

> Δ **Fitted value**: Predicted mean value of the response variable for all individuals with a given value of the explanatory variable.

> Δ **Predicted value**: Predicted value of the response variable for an individual with a given value of the explanatory variable.

Both calculations – of the fitted value and the predicted value – are statistics that are subject to sampling variability. Thus, both results have sampling distributions that are normally distributed[18] with a mean equal to $\hat{\mu}_{Y|X=x_0}$. The standard error for the fitted value is labeled as SE_{fits} and is given by

$$\sqrt{s^2_{Y|X}\left(\frac{1}{n}+\frac{\left(x_0-\bar{X}\right)^2}{(n-1)s^2_x}\right)} \tag{4.2.2}$$

The standard error for the predicted value is labeled as SE_{pred} and is given by

$$\sqrt{s^2_{Y|X}+s^2_{Y|X}\left(\frac{1}{n}+\frac{\left(x_0-\bar{X}\right)^2}{(n-1)s^2_x}\right)} \tag{4.2.3}$$

Both standard errors can be used to construct intervals. For example, the interval of $\hat{\mu}_{Y|X=x_0} \pm t^* SE_{fits}$ gives a confidence interval for the mean value of Y when X is equal to x_0; whereas the interval of $\hat{\mu}_{Y|X=x_0} \pm t^* SE_{pred}$ gives a prediction interval for the value of Y when X is equal to x_0.

> ◇ **A confidence interval is for the mean value of the response variable at a given value of the explanatory variable. A prediction interval is for the value of the response variable at a given value of the explanatory variable.**

[18] If the regression assumptions are met; see Section 4.3.

The width of these intervals depends on the value of x_0. The SE_{fits} and SE_{pred} are minimized when $x_0 = \bar{X}$. Thus, these intervals are narrowest at $\bar{X}$ and progressively wider as x_0 is further away from $\bar{X}$.[19] This widening is intuitive because the bulk of the data (or information) is near the "middle" of the range of the explanatory variable; thus, confidence in fitted or predicted values is greatest near the middle and is less near the margins of the range of the explanatory variable.

◇ **Both confidence and prediction intervals are narrowest at the mean of the explanatory variable and get wider further from the mean of the explanatory variable.**

It is critically important to understand the interpretational difference between intervals made using SE_{fits} and those using SE_{pred}. Intervals using SE_{fits} are for estimates of the mean value of Y at a given value of X. Thus, SE_{fits} is a measure of sampling variability or a measure of how different the mean value of Y at a given X would be if different samples were taken. In contrast, intervals using SE_{pred} are for estimates of the value of Y at a given value of X for an individual. Thus, SE_{pred} contains two types of variability; (1) sampling variability associated with estimating the mean and (2) natural variability associated with individuals. In other words, there is variability associated with estimating the mean (as measured by SE_{fits}) and there is natural variability among individuals (as measured by $s^2_{Y|X}$).

Graphically, sampling variability is illustrated in Figure 4.6. Thus, the dashed lines in Figure 4.6 illustrate "confidence bands" for the mean value of Y at all given values of X. To make "prediction bands" that will contain the predicted value of Y for an individual at each given value of X, additional natural variability would need to be added onto the ends of each of the confidence bands in Figure 4.6. This additional variability is illustrated by the dashed blue lines in Figure 4.7 and is also evident by noting the extra $s^2_{Y|X}$ when comparing Equations (4.2.2) and (4.2.3).

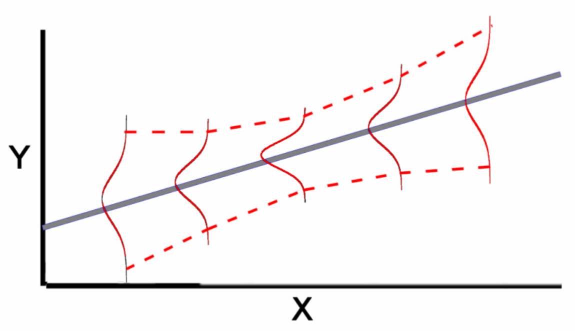

Figure 4.6. Idealistic conceptualization of a best-fit line surrounded by confidence intervals for $\mu_{Y|X}$.

◇ SE_{fits} **represents one type of error - sampling variability related to predicting the mean value of the response variable at a given value of the explanatory variable.** SE_{pred} **represents two types of error - sampling variability related to predicting the mean value of the response variable and natural variability related to predicting an individual's difference from that mean.**

For example, suppose that one wants to predict the mean number of eggs for all 700-mm female lake trout (i.e., "fitted value"). This requires the calculation of a confidence interval using SE_{fits} because it is related to a *mean* for *all* 700-mm lake trout. Thus, the mean number of eggs for all 700-mm female lake trout is

[19] Make sure you can see why this is true by looking at Equations (4.2.2) and (4.2.3)

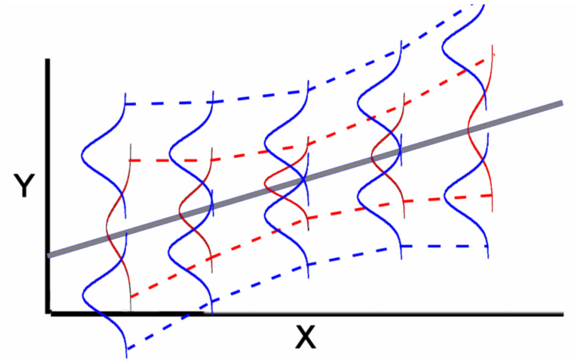

Figure 4.7. Idealistic conceptualization of a best-fit line surrounded by confidence intervals for $\mu_{Y|X}$ (in red) and prediction intervals for Y (in blue).

between 4442 and 5333 (Table 4.3). Further suppose that one wants to predict the number of eggs for a 700-mm female lake trout (i.e., "predicted value"). This requires a prediction interval using SE_{pred} because it is about *an individual* 700-mm lake trout. Thus, the number of eggs predicted for a 700-mm female lake trout is between 1299 and 8476 (Table 4.3).

Table 4.3. Fitted values, confidence intervals (top row), and prediction intervals (bottom row) for number of eggs for 700-mm female lake trout.

```
      fit      lwr      upr
 4887.744 4442.281 5333.206
 4887.744 1299.143 8476.345
```

Predictions in R

Future means and individuals can be predicted using the results of a simple linear regression with `predict()`. This function requires the `lm()` object as the first argument, a data.frame that consists of values of the explanatory variable at which to make predictions as the second argument, and a string in `interval=` that indicates whether to construct a confidence (`interval="confidence"`) or prediction (`interval="prediction"`) interval. For example, the confidence interval for the mean number of eggs in all 700-mm total length lake trout is obtained with

```
> predict(lt.lm,data.frame(tl=700),interval="confidence")
       fit      lwr      upr
1 4887.744 4442.281 5333.206
```

As another example, prediction intervals for the number of eggs in a 700-mm and in a 770-mm lake trout are obtained with

```
> predict(lt.lm,data.frame(tl=c(700,770)),interval="prediction")
       fit      lwr       upr
1 4887.744 1299.143  8476.345
2 6438.570 2859.704 10017.437
```

The most difficult aspect of making predictions in R appears to be the use of the data.frame in the second argument of `predict()`. The main thing to remember is that within `data.frame()` the name of the explanatory variable must be used exactly as it appeared in the saved `lm()` object. In other words, if *tl* was used to fit the linear model, then *tl* must be used in `data.frame()` within `predict()`.

The confidence and prediction bands are plotted by adding the `interval=` argument to `fitPlot()` (Figure 4.8). The `interval=` argument can be set to `"confidence"` to construct a confidence band, to `"prediction"` to construct a prediction band, or `"both"` to construct both confidence and prediction bands.

```
> fitPlot(lt.lm,interval="both",xlab="Total Length",ylab="Number of Eggs")
```

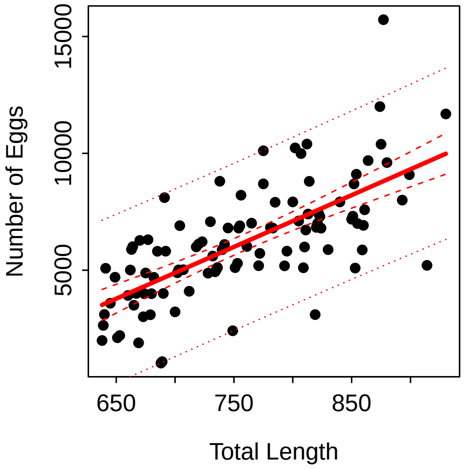

Figure 4.8. Scatterplot of number of eggs versus total length for Lake Superior lake trout with best-fit line and 95% confidence and prediction bands superimposed.

Finally, a visual of fitted or predicted values with intervals superimposed on to a fitted-line plot with the confidence and prediction bands can be constructed with `predictionPlot()`. This function takes the exact same arguments as `predict()` and returns the predicted values, with intervals, and a visual plot. For example, the predicted values for 700 and 770 mm lake trout are obtained with

```
> predictionPlot(lt.lm,data.frame(tl=c(700,770)),interval="prediction",
                 xlab="Total Length",ylab="Number of Eggs")
  obs  tl      fit      lwr       upr
1   1 700 4887.744 1299.143  8476.345
2   2 770 6438.570 2859.704 10017.437
```

and visualized in Figure 4.9. Inasmuch as the results of `predict()` are included in `predictionPlot()`, it is suggested that you use `predictionPlot()` to make predictions as this should help you avoid making extrapolations with the model.

4.2.4 Models & SS

As with all linear models (see Module 1), the important hypothesis tests of SLR can be reduced to comparing two models' lack-of-fit to data. Previously it was seen that the hypothesis test to determine if the response and explanatory variable are significantly related is

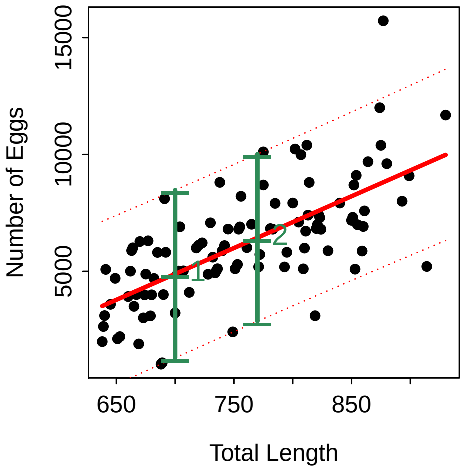

Figure 4.9. Scatterplot of number of eggs versus total length for Lake Superior lake trout with best-fit line, 95% confidence and prediction bands superimposed, and 95% prediction intervals shown for a 700 (interval 1) and 710 (interval 2) mm lake trout.

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

This can be written in terms of models as

$$\begin{aligned} H_0 &: \mu_{Y|X} = \alpha \\ H_A &: \mu_{Y|X} = \alpha + \beta_1 X \end{aligned}$$

Furthermore, by substituting Equation (4.1.3) into Equation (4.1.2), the sample best-fit line can be rewritten as follows

$$\begin{aligned} \hat{\mu}_{Y|X} &= \overline{Y} - \hat{\beta}_1\overline{X} + \hat{\beta}_1 X \\ &= \overline{Y} + \hat{\beta}_1\left(X - \overline{X}\right) \end{aligned}$$

Thus, if $\hat{\beta}_1 = 0$, then the model reduces to $\hat{\mu}_{Y|X} = \overline{Y}$.

Therefore, testing the hypothesis that the slope is equal to zero is equivalent to testing whether the simple model of $\overline{Y}$ (or, equivalently, the constant $\hat{\alpha}$) adequately fits the data versus a more complicated model where a slope is needed (Figure 4.10). Said another way, this is the same as testing whether the mean value of Y is the same for all Xs (i.e., no slope, no relationship) or whether the mean value of Y depends on the value of X (i.e., a slope, there is a relationship).

◇ **The simple model in SLR represents a flat line at the mean of the response variable. The full model in SLR represents a line with a significant slope.**

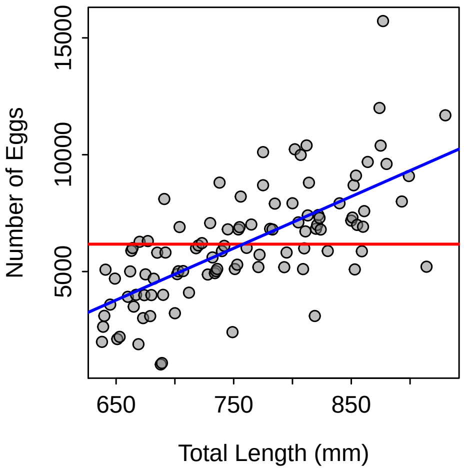

Figure 4.10. Scatterplot illustrating two competing models for describing the relationship between number of eggs and total length. The horizontal red line is placed at the mean number of eggs and represents the simple model, whereas the blue line is the best-fit line and represents the full model.

Of course, the lack-of-fit of the two models is calculated by summing the squared residuals using predictions from the two models. Specifically, the lack-of-fit of the simple model is computed from the mean value of the response variable (Figure 4.11-Left), or

$$SS_{Total} = \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 \tag{4.2.4}$$

As expected, this SS is called SS_{Total} because it is the basis of the measure of the "total" variability in the response variable as calculated by the variability around the overall mean (and is the same as that discussed for the one- and two-way ANOVAs).

The lack-of-fit of the full model is computed from the best-fit regression line (Figure 4.11-Right), or

$$SS_{Residual} = \sum_{i=1}^{n}\left(y_i - \hat{\mu}_{Y|X}\right)^2 = \sum_{i=1}^{n}\left(y_i - \left(\hat{\alpha} + \hat{\beta}_1 x_i\right)\right)^2$$

This SS is termed $SS_{Residual}$ in SLR, but it is exactly analogous to SS_{Within} from Modules 1 and 2.

◇ **SS_{Total} measures the "variability" in the simplest model, which is just the mean of the response variable. Thus, SS_{Total} measures the maximum total "variability" in the response variable.**

As always, SS_{Total} can be partitioned into two parts

$$SS_{Total} = SS_{Residual} + SS_{Regression}$$

or, more specifically,

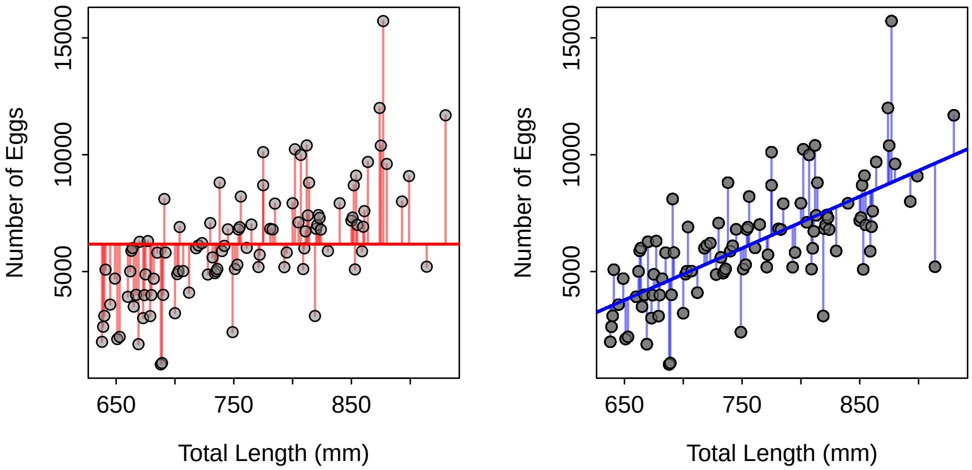

Figure 4.11. Scatterplots illustrating two competing models for describing the relationship between number of eggs and total length. The horizontal red line is placed at the mean number of eggs and represents the simple model (Left). The blue line is the best-fit line and represents the full model (Right). Residuals for each model are shown on the respective graphs.

$$\sum_{i=1}^{n}\left(y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(y_i-\hat{\mu}_{Y|X}\right)^2+\sum_{i=1}^{n}\left(\hat{\mu}_{Y|X}-\overline{Y}\right)^2$$

$SS_{Residual}$ represents the part of the total variability in the response variable that is not explained by the full model. The difference between SS_{Total} and $SS_{Residual}$ is called $SS_{Regression}$ and represents the part of the total variability in the response variable that *is* explained by the full model. This part is exactly analogous to SS_{Among} from Modules 1 and 2, but is called $SS_{Regression}$ in SLR because it is the amount of variability explained by using the best-fit regression line. Thus, as would be expected, $SS_{Regression}$ measures how much "better" the full model fits compared to the simple model. However, as with SS_{Among}, this statistic must be converted to an MS and then, ultimately, to an F test statistic.

◇ $SS_{Total} = SS_{Residual} + SS_{Regression}$.

◇ **The residual SS ($SS_{Residual}$) is the measure of the "variability" in the response variable that is unexplained when an explanatory variable is incorporated into the model.**

◇ **The regression SS ($SS_{Regression}$) is the measure of the "variability" in the response variable that is explained when an explanatory variable is incorporated into the model.**

ANOVA Table

The three SS just discussed are converted to MS by dividing by their respective df. As in Module 1, $df_{Total} = n-1$. The $df_{Residual}$ is equal to the number of individuals minus the number of parameters in the full model – i.e., $n-2$.[20] Thus, using the rule that dfs partition in the same way as SS, the $df_{Regression}$ is

[20] This is a general rule for the calculation of $df_{Residual}$.

$(n-1)$–$(n-2)$= 1.[21]

◇ The regression df are always 1 in SLR.

With these df, the MS are computed as

$$MS_{Total} = \frac{SS_{Total}}{df_{Total}} = \frac{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}{n-1} = s_Y^2 \tag{4.2.5}$$

$$MS_{Residual} = \frac{SS_{Residual}}{df_{Residual}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_{Y|X}\right)^2}{n-2} = s_{Y|X}^2 \tag{4.2.6}$$

$$MS_{Regression} = \frac{SS_{Regression}}{df_{Regression}} = \frac{\sum_{i=1}^{n}\left(\hat{\mu}_{Y|X} - \overline{Y}\right)^2}{1}$$

The F test statistic is computed similarly to what was described in Module 1. Specifically,

$$F = \frac{MS_{Regression}}{MS_{Residual}}$$

with $df_{Regression}$ numerator and $df_{Residual}$ denominator df.

As usual, the degrees-of-freedom (df), sum-of-squares (SS), mean-squares (MS), F test statistic (F), and corresponding p-value are summarized in an analysis of variance table (Table 4.4).

Table 4.4. Analysis of variance table for the regression of $\mu_{eggs|tl} = \alpha + \beta_1 tl$.

```
          Df    Sum Sq   Mean Sq F value    Pr(>F)
tl         1 287104252 287104252  89.148 1.806e-15
Residuals 99 318832897   3220534
```

The results in Table 4.4 indicate that there is a significant relationship between *eggs* and *tl* ($p < 0.0005$). This same result indicates that a full model with a slope term on the *tl* variable is significantly "better" at fitting the observed data then a simple model that does not contain a slope term.

In addition to the primary objective of comparing the full and simple models, several items of interest can be identified from an analysis of variance table. Using Table 4.4 as an example, the following items are identified:

- The variance of individuals about the regression line ($s_{Y|X}^2$) is given by $MS_{Residual}$ (e.g., =3220534).
- The variance of individuals about the mean (s_Y^2) is given by MS_{Total} (e.g., SS_{Total}=287104252+318832897 divided by df_{Total}=1+99 or 6059371.5).
- The F test statistic is equal to the square of the t test statistic from testing $H_0: \beta_1 = 0$.[22]

[21] It is also common that the $df_{Regression}$ is the difference in number of parameters between the full and simple models.

[22] This is a general rule between the T and F distributions. An F with 1 numerator df and ν denominator df is equal to the square of a T with ν df.

Regression ANOVA Table in R

The ANOVA table for a SLR is obtained by submitting the saved `lm()` object to `anova()` (e.g., the ANOVA table results for the lake trout egg data shown in Table 4.4 were obtained with `anova(lt.lm)`).

Coefficient Of Determination

The coefficient of determination (R^2) is a measure of the proportion of the total variability in the response variable that is explained by knowing the value of the explanatory variable.[23] From the SS definitions,

$$R^2 = \frac{SS_{Regression}}{SS_{Total}}$$

In essence, R^2 is a measure of the strength for predictions that can be made. In other words, R^2 values near one indicate a relationship that is very strong and will lead to precise predictions, whereas R^2 values near zero indicate a very weak relationship with correspondingly weak predictions.

Coefficient of Determination in R

The coefficient of determination is shown in the output following “multiple R-squared“ when a saved `lm()` object is submitted to the `summary()` function. An example is shown in Table 4.2. This value can also be isolated by submitting the saved `lm()` object to `rSquared()`.

4.3 Assumptions

Simple linear regression has five major assumptions,

1. Each individual is independent of each and every other individual (“independence” assumption).
2. The mean values of Y at each given value of X fall on a straight line (“linearity” assumption).
3. The variances of Y at each given value of X are all equal (to $\sigma^2_{Y|X}$) (“homoscedasticity” assumption).
4. The values of Y at each given value of X are normally distributed (“normality” assumption).
5. No outliers.

These five assumptions lead to the idealistic model illustrated in Figure 4.12.

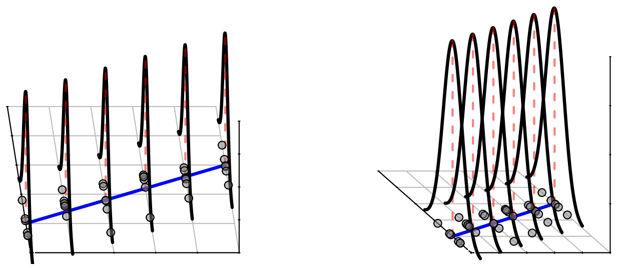

Figure 4.12. Two depictions of the assumptions of the simple linear regression model. The blue “best-fit” line intersects the mean of each normal distribution. Each normal distribution represents the distribution of the response variable for all individuals at a particular value of the explanatory variable.

[23] It is assumed that you learned this statistic in your introductory statistics course.

The first assumption states that the **means** fall on a straight line, not that the individuals do. This is why the left-hand-sides of Equations (4.1.1) and (4.1.2) contain $\mu_{Y|X}$ rather than just Y. This is also demonstrated by the observation that the line drawn on Figure 4.12 intersects the means of the individual normal distributions. Thus, don't forget that Equation (4.1.1) represents the mean values of the response variable at a given value of the explanatory variable and not the individual values of the response variable.

The model in Equation (4.1.1) can be modified to represent each individual (rather than the mean of individuals) by adding an error term (ϵ). Thus, the model to represent individuals is written

$$Y|X = \alpha + \beta_1 X + \epsilon \tag{4.3.1}$$

From the assumptions (as illustrated in Figure 4.12), the errors will be normally distributed with a mean of 0 (because the line passes through the $\mu_{Y|X}$ and the residuals, or errors, are computed from that point) and a variance of $\sigma^2_{Y|X}$. Thus, in shorthand, the $\epsilon \sim N(0, \sigma_{Y|X})$.

◇ **The errors about the best-fit line are $N(0, \sigma_{Y|X})$.**

The $\sigma^2_{Y|X}$ is called "the common variance about the model" and represents the natural variability about the model (i.e., "how much does each individual naturally vary from the model"). This common variance is exactly analogous to MS_{Within} discussed in Modules 1 and 2 and is the basis of nearly all inferences in SLR (as seen in Section 4.2). The common variance is a parameter that is estimated using the residuals from the individuals in a sample with

$$s^2_{Y|X} = MS_{Residual} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\mu}_{Y|X=x_i}\right)^2}{n-2}$$

Thus, $\sigma^2_{Y|X}$ is a population variance and $s^2_{Y|X}$ is a sample variance.

4.3.1 Diagnostics

The assumption of the independence of individuals is generally assessed with common sense and is controlled through proper sample design as was described in Modules 1-3.[24]

The linearity assumption is the most important assumption in SLR; i.e., the form of the bivariate relationship must be linear in order to fit a line to it. Problems with the linearity assumption are diagnosed by close examination of the fitted-line plot. In some instances, departures from linearity may be subtle or the strength of relationship so strong and the range of the explanatory variable so large that departures from linearity are difficult to discern. In these instances, one should look at a residual plot from the model fit. The residual plot effectively "zooms in" on the best-fit line such that subtle departures from linearity can be more easily identified. The most common departures from linearity look like parabolas, but more "complicated" structures may also be noted (Figure 4.13).

◇ **The linearity assumption can be addressed with a fitted-line or a residual plot.**

[24] If the individuals are ordered by time or space then the Durbin-Watson statistic can be used to determine if the individuals are serially correlated or not. Generally, the H_0 : "not serially correlated" and H_A : "is serially correlated" are the hypotheses for the Durbin-Watson test. Thus, p-values $< \alpha$ result in the rejection of H_0 and the conclusion of a lack of independence. In this case, the regression assumption would be violated and other methods, primarily time-series methods, should be considered.

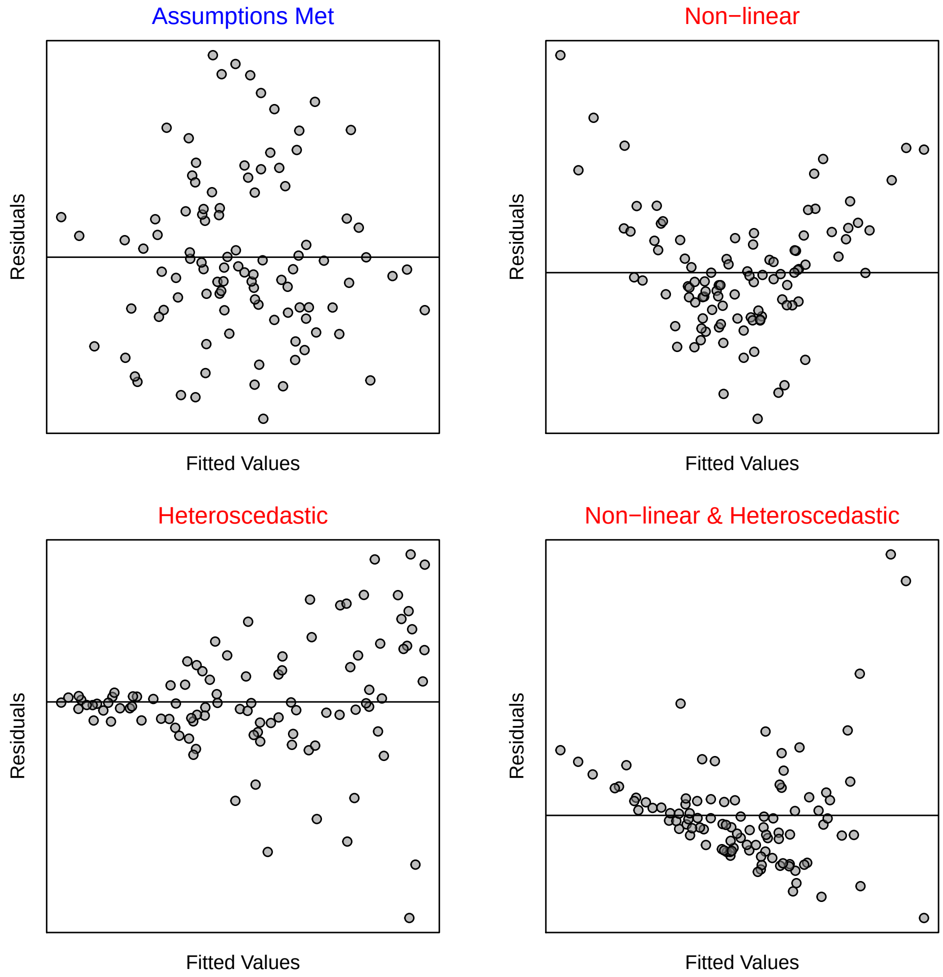

Figure 4.13. Residual plots illustrating when the regression assumptions are met (upper-left) and three common assumption violations.

◇ **A fitted-line or residual plot that exhibits no obvious curvature is evidence that the linearity assumption has been met.**

The homoscedasticity assumption is also vital because all inferences in SLR depend on $s^2_{Y|X}$. The homoscedasticity assumption assures that the variability is constant around the line and that $s^2_{Y|X}$ estimates a constant quantity. If this assumption is not met, then a common variance does not exist, $s^2_{Y|X}$ measures a quantity that does not exist, and all of the SE calculations from Section 4.2 will not work properly. Difficulties with the homoscedasticity assumption are diagnosed by close examination of a residual plot. If the points on the residual plot show the same degree of scatter from left-to-right then the homoscedasticity assumption is likely met. A common violation of the assumption appears as a funnel shape from left-to-right (Figure 4.13).

◇ **The homoscedasticity assumption is most often addressed with a residual plot.**

◇ A residual plot that exhibits no vertical compression of points is evidence that the homoscedasticity assumption has been met.

Interpreting residual plots requires some practice and experience. The trick to examining residual plots is to look for distinctive patterns and shapes. Most novice statisticians find too much detail in residual plots, identifying every subtle curvature and change in variability. If distinct patterns do not exist then the assumptions are probably adequately met. Remember, a residual plot with random scatter and no discernible pattern is an indication that the linearity and homoscedasticity assumptions have been met.

The normality assumption is dealt with exactly as it was in Modules 1-3. It is virtually impossible in most SLR to test the normality of residuals at each given value of X because there typically is very few replicates of each value of X. Thus, the assumption is assessed with the Anderson-Darling normality test of the residuals from the full model.

The assumption of no outliers is also tested exactly as described in Modules 1-3; i.e., with a hypothesis test using the Studentized residuals and a Bonferroni correction. However, further discussion of outliers in SLR is warranted. Fox (1997) describes "unusual" observations the best with,

> Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis, and because their presence may signal that the model fails to capture important characteristics of the data.

In its simplest sense, a regression outlier is an individual for which the response variable is unusual given the value of the explanatory variable and the overall "fit" of the model. Following the arguments of Fox (1997), regression outliers are shown in Figure 4.14-left and Figure 4.14-middle. In contrast, a univariate outlier is an individual for which the value of a variable is unusual relative to the mean of that single variable. An outlier for the response variable and for the explanatory variable, but not a regression outlier is shown in Figure 4.14-right. Note that in Figure 4.14-right that the outlying individual is "extreme" along both the x- and y-axes but not relative to the linear model fit. Univariate outliers are not necessarily regression outliers. However, sometimes the two types of outliers are found in the same individual – e.g., Figure 4.14-middle.

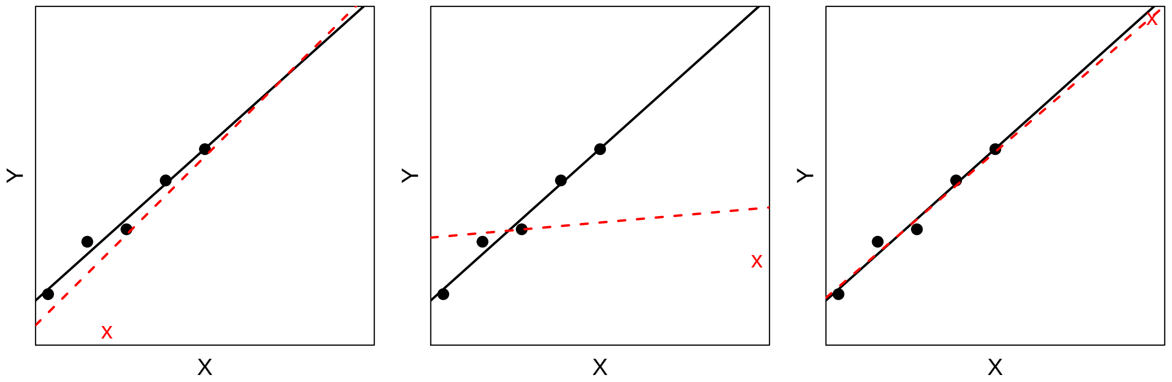

Figure 4.14. Illustration of the concepts of outliers and influence in a simple linear regression. Each plot consists of one SLR fit to the "original" data (i.e., black solid line and black dots) and another SLR fit to the "original" data with the addition of one "unusual" point (point shown by red "x" and fitted line shown by red dashed line).

Individuals that substantially impact the fitted line (i.e., result in substantially different values for the regression coefficients; Figure 4.14-middle) are called *influential points*. The "influence" of an individual is related to both it's horizontal distance from the center of the explanatory variable and its vertical distance from the best-fit line. Intuitively then, highly influential points are points that have a combined "large" distance from left-to-right and top-to-bottom relative to the best-fit line. It is possible that a highly influential point will not appear as an outlier in `outlierTest()` because of the high influence it has on the position of the line. Thus, influential points are diagnosed with careful attention to the fitted-line and residual plots.

> △ **Influential Point**: An individual whose inclusion in the data set substantially impacts the coefficients of the fitted line.

Assumption Checking in R

The fitted-line plot, residual plot, Anderson-Darling normality test, and the outlier test are constructed with `transChooser` as described in Modules 1-3, with the exception that the residual plot plots the residuals versus the fitted values from the full model rather than a boxplot of residuals versus level names.

4.4 Transformations

If one or more of the linearity, homoscedasticity, normality, or outlier assumptions are violated, then the data may be transformed to a different scale where the assumptions are met. Either the response or explanatory variable can be transformed, although transformation of the response variable is generally more effective. In general, the family of power transformations (see Section 2.6) will be considered for both the response and explanatory variables, although some special transformations can be used in specific situations (e.g., $\sin^{-1}\sqrt{Y}$ for proportions or percentage data).

> ⋄ **If the normality, linearity, or homoscedasticity assumption is violated then a transformation of variables should be considered.**

4.4.1 Selecting Power Transformations

Power transformations for the variables in a SLR can be selected from a variety of methods – (1) based on theory, (2) from past experience, or (3) trial-and-error with dynamic graphics. Specifics of these methods are discussed in the following sections.

Theoretical Transformations

Transformations may be chosen if a theoretical functional form for the relationship between the response and explanatory variables can be identified. For example, many relationships follow a non-linear power function – $Y = aX^b$ – where X and Y are variables (as always) and a and b are scalars (i.e., constant numbers). An example of this type of data is shown in Figure 4.15.

If logarithms are taken of both sides of a power function then the form reduces to

$$\begin{aligned}
log(Y) &= log(aX^b) \\
log(Y) &= log(a) + log(X^b) \\
log(Y) &= log(a) + blog(X)
\end{aligned}$$

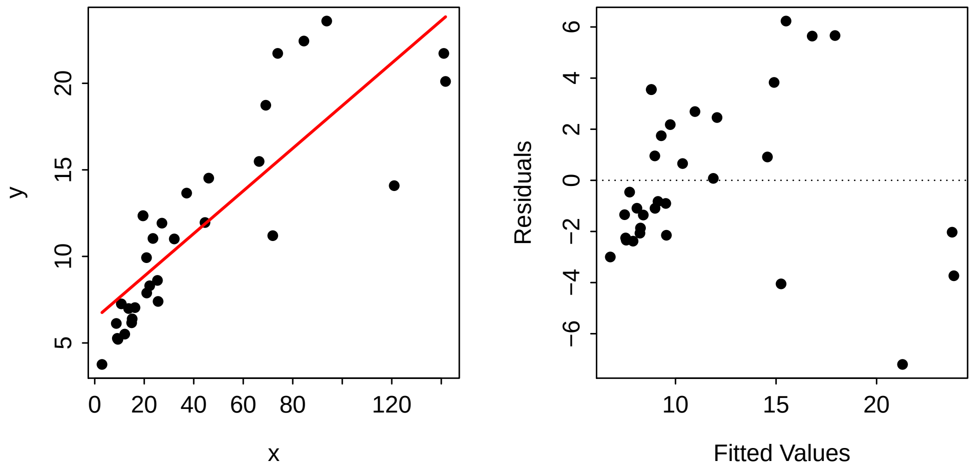

Figure 4.15. Fitted line plot (Left) and residual plot (Right) for data simulated from a power function.

which, because $log(Y)$ and $log(X)$ are still variables and $log(a)$ is still a constant, is in a linear form. Thus, transforming both the response and explanatory variable to the logarithm scale will "linearize" a relationship that is known to theoretically follow a power function (Figure 4.16; note the lack of curvature in both plots).

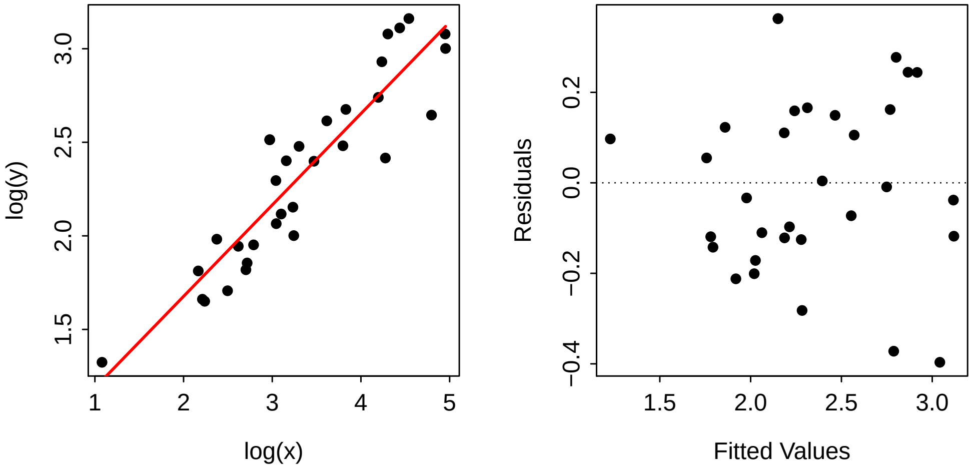

Figure 4.16. Scatterplot (Left) and residual plot (Right) from a log-log transformation of the data simulated from power function and shown in Figure 4.15.

Another common form is the non-linear exponential form – $Y = ae^{bX}$ – where e is the base of the natural log and is a constant. An example of this type of data is shown in Figure 4.17. Again, taking the natural log of both sides reduces this functional form to

$$
\begin{aligned}
log(Y) &= log(ae^{bX}) \\
log(Y) &= log(a) + log(e^{bX}) \\
log(Y) &= log(a) + bX
\end{aligned}
$$

Thus, transforming the response, but not the explanatory, variable to the logarithm scale will “linearize” a relationship that is known to theoretically follow an exponential function (Figure 4.18).

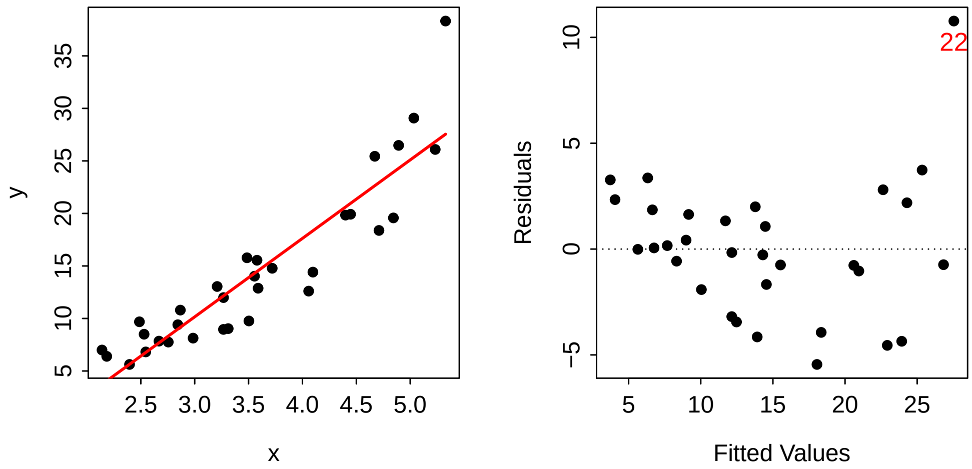

Figure 4.17. Scatterplot (Left) and residual plot (Right) for data simulated from an exponential function.

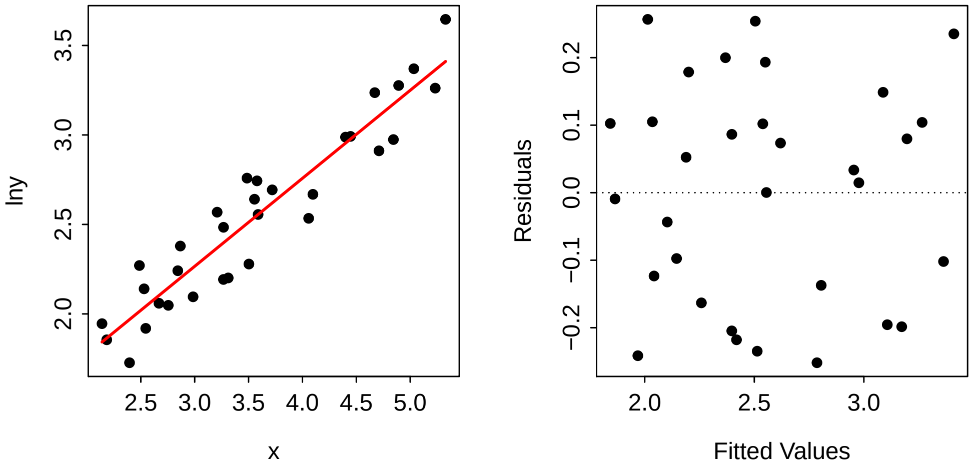

Figure 4.18. Scatterplot (Left) and residual plot (Right) for log-transformed data simulated from an exponential function and shown in Figure 4.17.

Other common equation types, their transformations to a linear form, and estimates of the model parameters are shown in Table 4.5. Many of these models are important in specific scientific fields.

Table 4.5. Common mathematical models, their transformations to a linear form, and parameter estimates..

Model Forms

Model Name	Standard	Linear
Exponential	$Y = ae^{bX}$	$log(Y) = log(a) + bX$
Power Function	$Y = aX^b$	$log(Y) = log(a) + blog(X)$
Modified Power Function	$Y = aX^b + c$	$log(Y - c) = log(a) + blog(X)$
Sigmoid	$Y = \frac{c}{1+aX^b}$	$log\left(\frac{c}{Y} - 1\right) = log(a) + blog(X)$
Exponential Sigmoid	$Y = \frac{c}{1+ae^{bX}}$	$log\left(\frac{c}{Y} - 1\right) = log(a) + bX$
Exponential Saturation	$Y = a\left(1 - e^{bX}\right)$	$log(a - Y) = log(a) + bX$
Maxima Function	$Y = aXe^{bX}$	$log\left(\frac{Y}{X}\right) = log(A) + bX$
Modified Inverse	$Y = \frac{a}{b+X}$	$\frac{1}{Y} = \frac{b}{a} + \frac{1}{a}X$
Hyperbola	$Y = \frac{aX}{b+X}$	$\frac{X}{Y} = \frac{b}{a} + \frac{1}{a}X$

	Transformations		Parameter Estimates		
Model Name	Response	Explanatory	a	b	c
Exponential	$log(Y)$	X	$e^{intercept}$	$slope$	–
Power Function	$log(Y)$	$log(X)$	$e^{intercept}$	$slope$	–
Modified Power Function	$log(Y - c)$	$log(X)$	$e^{intercept}$	$slope$	estimated
Sigmoid	$log\left(\frac{c}{Y} - 1\right)$	$log(X)$	$e^{intercept}$	$slope$	estimated
Exponential Sigmoid	$log\left(\frac{c}{Y} - 1\right)$	X	$e^{intercept}$	$slope$	estimated
Exponential Saturation	$log(a - Y)$	X	$e^{intercept}$	$slope$	–
Maxima Function	$log\left(\frac{Y}{X}\right)$	X	$e^{intercept}$	$slope$	–
Modified Inverse	$\frac{1}{Y}$	X	$\frac{1}{slope}$	$\frac{intercept}{slope}$	–
Hyperbola	$\frac{X}{Y}$	X	$\frac{1}{slope}$	$\frac{intercept}{slope}$	–

Transformations from Experience

Finally, transformations for the response variable may be determined based on common transformations for a particular type of data. Transformations for some common types of data are shown in Table 4.6.

Table 4.6. Common response transformations and their typical usage.

$Y^* = Y^{0.5}$	Often used for discrete count data (from Poisson distribution theory).
$Y^* = Y^{0.5} + (Y+1)^{0.5}$	Same as above except for use when some values are 0 or very small.
$Y^* = log(Y+1)$	When a log transformation is warranted but 0s exist in the data.
$Y^* = Y^{-1}$	Often used when a response time is recorded. This changes the units from the scale of time per response to number of responses per time.
$Y^* = sin^{-1}\left(Y^{0.5}\right)$	Commonly used when the data recorded are proportions or percentages.

It should also be noted that Weisberg (2014) suggested that if the ratio of maximum to minimum value for the explanatory variable is greater than 10, then it should be transformed to the natural log scale.

Trial-and-Error

Transformations for the response and explanatory variable can be identified by trying a variety of powers for each variable and exploring the results of each. Of course, this is a tedious process. Fortunately, computer programs exist for dynamically determining the effect of transforming each variable. When using these programs the user should attempt to transform just the response variable first.

Back-Transformation Issues

As discussed in Section 2.6.3, back-transformation is the process of reversing the results found on the transformed scale to the original scale for ease of interpretation. In regression analyses, the coefficients and predicted values can be thought of as being one of two types. The intercept, fitted values, and predicted values (along with the endpoints of their confidence intervals) are estimates of average values. These values are back-transformed as discussed in Section 2.6.3, including the use of the correction factor when a log transformation is used.[25]

In contrast, the slope is an estimate of a rate of change and can only be interpreted if back-transformed from a logarithm. For example, back-transforming a slope from the natural log scale produces an estimate of the **multiplicative change** in the mean value of the response variable for a one unit change of the explanatory variable. Without transformation, the slope is an estimate of how much is **added** to the mean of Y for a unit change in X; however, back-transforming the slope from the log scale is an estimate of how much the mean of Y is **multiplied** by for a unit change of X.

This interpretation can be illustrated by first looking at the difference in the log-transformed means for a unit change in X,

$$log\left(\mu_{Y|X+1}\right) - log\left(\mu_{Y|X}\right) = [\alpha + \beta_1(X+1)] - [\alpha + \beta_1 X] = \beta_1$$

Thus, not surprisingly, on the log-transformed scale a unit change in X results in a β_1 unit change in the natural logarithm mean of Y. Now, raise both sides of this equation to the power of e and simplify,

[25] Note, however, that the MS_{Within} is replaced with $MS_{Residual}$ so that the correction factor is $e^{\frac{MS_{Residual}}{2}}$.

$$e^{\beta_1} = e^{log(\mu_{Y|X+1}) - log(\mu_{Y|X})}$$
$$e^{\beta_1} = e^{log\left(\frac{\mu_{Y|X+1}}{\mu_{Y|X}}\right)}$$
$$e^{\beta_1} = \frac{\mu_{Y|X+1}}{\mu_{Y|X}}$$

Thus, e^{β_1} does NOT represent the difference in two means, rather it represents the ratio of two means on the original scale.

◇ **The value of a slope back-transformed from a natural log transformation represents the ratio of the two means separated by one unit of the explanatory variable on the original scale. In other words, this back-transformed values it he multiplicative change in the mean response for a unit change in the explanatory variable.**

Note that the correction factor discussed in Section 2.6.3 is NOT used when back-transforming rates. Also, note that this interpretation on the original scale is only true if a log transformation was used. Useful interpretations on the original scale can not be found for most other transformations. Thus, the slope should *NOT* be back-transformed if a transformation other than a logarithm was used.

◇ **Do NOT back-transform slopes if other than a log transformation was used.**

4.4.2 Polynomial Regression

The power transformations discussed above have a tendency to simultaneously “fix” violations of the linearity, homoscedasticity, and normality assumptions. This can be troublesome if only one of these assumptions is violated. For example, it is possible to have data that meet the normality and homoscedasticity assumptions, but are non-linear. In this case, a power transformation may linearize the data, but may then also make the residuals non-normal or heteroscedastic. In this example, the meeting of two assumptions was “traded for” the meeting of just one assumption.

In this specific situation – non-linear data with homoscedastic, normal residuals – it may be appropriate to fit a polynomial regression. A polynomial regression consists of a generic model of the form

$$\mu_{Y|X} = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \ldots$$

Generally speaking, the polynomial regression model will have one more term (i.e., highest exponent) than the number of “turns” in the data. So, if the form of the relationship looks like a parabola (i.e., one “turn”) then a model that stops with X^2 would be tried. If the data has two “turns” then a model that stops with X^3 would be tried. Models with more than three or four terms are rarely used because of their complexity.

An example of this type of data is shown in Figure 4.19. Clearly a linear model does not fit these data, but the residuals also do not show an increasing variance around the relationship. Thus, these data are a good candidate to be fit by a polynomial regression. In fact, a polynomial regression with three terms seems to fit the data well (Figure 4.20).

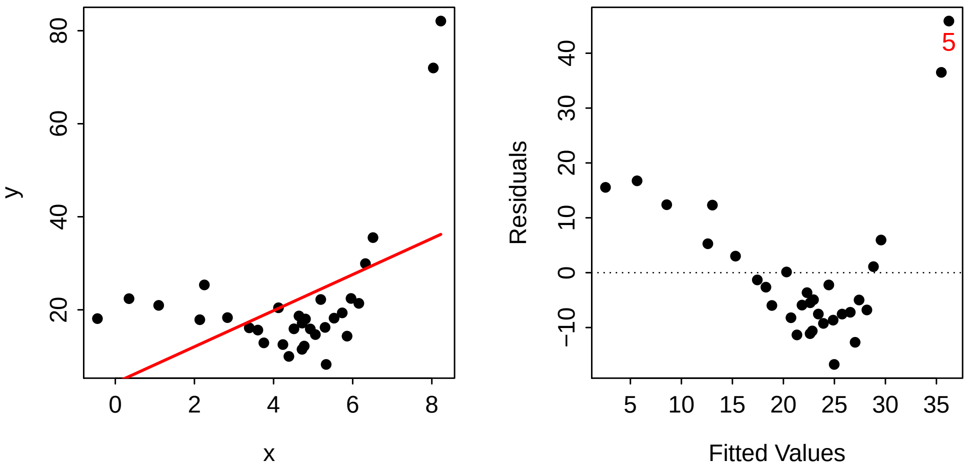

Figure 4.19. Fitted line plot (Left) and residual plot (Right) for the fit of a linear model to data simulated from a cubic model.

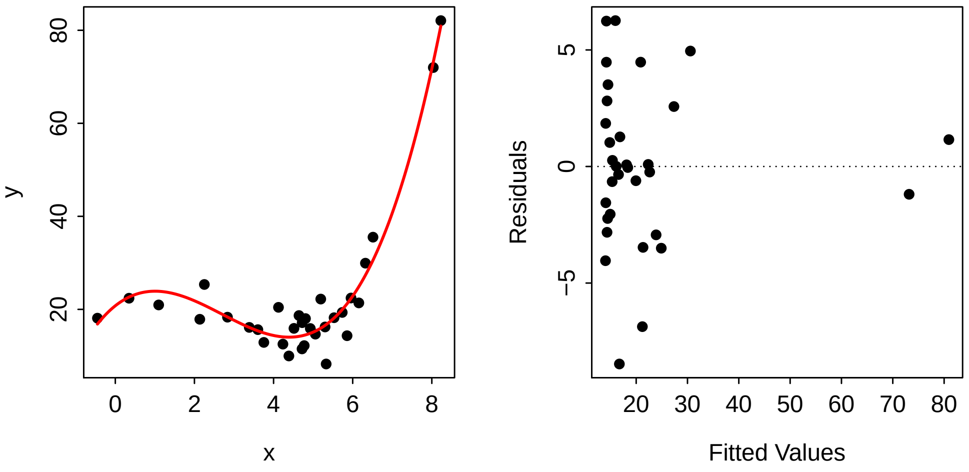

Figure 4.20. Fitted line plot (Left) and residual plot (Right) for the fit of $\mu_{Y|X} = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ to data simulated from a cubic model.

Polynomial regression generally falls under the heading of multiple linear regression and will not be discussed further here. However, at this point, note that polynomial regression is a useful technique if only the linearity assumption has been violated.

◇ **Polynomial regressions are useful when ONLY the linearity assumption has been violated. If more than one assumption is violated then polynomial regression is unlikely to solve the problem (and, thus, a power transformation should be considered).**

4.4.3 Transformations in R

The trial-and-error method of finding a transformation uses `transChooser()` with the only difference being that a start value for the explanatory value can be included in the `startx=` argument and that a slider bar will be provided for both the explanatory and response variables. Once a transformation has been identified the variable is transformed exactly as described in Modules 1-3.

4.5 Example Analyses

4.5.1 Lake Trout Fecundity

Schram (1993) examined the relationship between the total length and number of eggs found in female lake trout from Lake Superior. Schram's primary goal was to develop a model that could be used to predict the number of eggs produced by a female lake trout from the total length of that fish. This relationship could be used in subsequent models used to manage the population of lake trout.

Data Collection

Lake trout were collected during the spawning season from set netting areas in the Apostle Islands, Wisconsin. A sample of 101 ripe but not spent female lake trout over the range of observed lengths were sacrificed and the eggs removed by dissection. For each fish, the total length (TL; mm) and weight (g) was recorded and the total number of eggs counted. The data are stored in **LakeTroutEggs.csv** (view, download, meta).

EDA & Assumption Checking

The initial fit of the untransformed simple linear regression model of the number of eggs on total length indicated that the linear model was adequate for these data (Figure 4.21-Right) and the residuals were largely homoscedastic (Figure 4.21-Right) and weakly, but not significantly, non-normal (Anderson-Darling p = 0.059; Figure 4.21-Left). However, observation 96 was a significant outlier (p = 0.005; Figure 4.21-Right). A closer examination of observation 96 showed that (1) the number of eggs was 31% larger than the next highest number of eggs for similarly sized fish, (2) the number of eggs was 78% larger than predicted by the linear model, and (3) five fish were larger but none produced more eggs. There is no indication that the length or the number of eggs for this fish were measured in error. Thus, in an effort to produce a model that represents "typical" female lake trout this "unusual" observation was removed from further analysis.

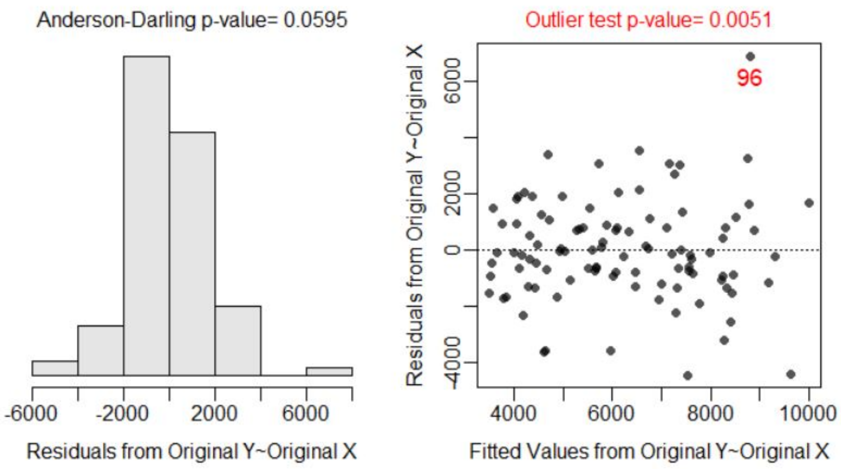

Figure 4.21. Histogram of residuals (Left) and residual plot (Right) from the fit of a SLR model to the raw Lake Superior lake trout data.

The fit of the untransformed simple linear regression model of the number of eggs on total length without observation 96 indicated that the linear model still adequately fit the data (Figure 4.22-Right), the residuals were largely homoscedastic (Figure 4.22-Right) and approximately normal (Anderson-Darling $p = 0.168$; Figure 4.22-Left), and there were no significant outliers left in these data ($p = 0.005$). Thus, a simple linear regression, with no transformations, will be fit to the data with observation 96 removed.

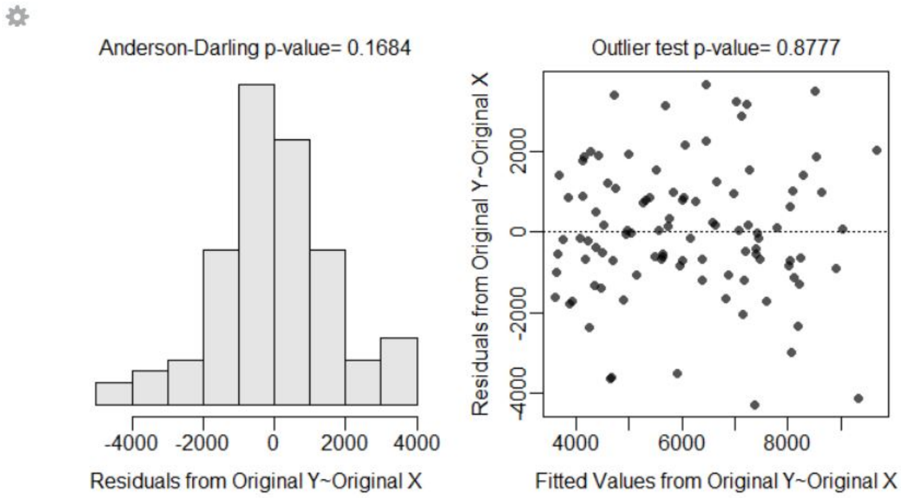

Figure 4.22. Histogram of residuals (Left) and residual plot (Right) from the fit of a SLR model to the raw Lake Superior lake trout data with observation 96 removed.

Results

There is a significant relationship between the number of eggs produced and the total length of female Lake Superior lake trout (ANOVA $p < 0.0005$). The relationship is only moderately strong (R^2 =0.476) and is characterized by the linear equation $\mu_{eggs|tl}$ =20.699tl-9588.084 (Table 4.7). Thus, a 1-mm increase in the total length of a female lake trout corresponds to an average increase of approximately 20.7 eggs with a 95% confidence interval of between 16.3 and 25.1 eggs.

Table 4.7. Coefficient results from the fit of eggs produced on total length of the raw Lake Superior lake trout data with observation 96 removed.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9588.084   1669.328  -5.744 1.04e-07
tl             20.699      2.195   9.431 2.08e-15
---
Residual standard error: 1658 on 98 degrees of freedom
Multiple R-squared: 0.4758,Adjusted R-squared: 0.4704
F-statistic: 88.94 on 1 and 98 DF,  p-value: 2.083e-15
```

As an example, **a** female lake trout with a total length of 650 mm would be predicted to have 3866 eggs with a 95% prediction interval of 527 to 7206 eggs. Examples for fish of other lengths are shown in Table 4.8.

Conclusion

A significant, but only moderately strong, relationship was found between the number of eggs produced and the total length of “typical” female lake trout from Lake Superior. A model was produced that can

Table 4.8. Predicted number of eggs and 95% prediction intervals for female Lake Superior lake trout with total lengths of 650, 750, and 850 mm.

```
  tl      fit       lwr       upr
 650 3866.361  527.1333  7205.589
 750 5936.276 2629.4759  9243.076
 850 8006.190 4674.6984 11337.683
```

be used to predict the number of eggs in female lake trout of different total lengths but, because the fit is only moderately strong, these predictions exhibit substantial variability. Further, it must be noted that this model was fit with the exclusion of one observed fish that had an "unusually" large number of eggs given its total length. Thus, these results are only appropriate for individuals that don't have an "unusually" large number of eggs given their total length.

Appendix – R commands

```
LakeTroutEggs <- read.csv("data/LakeTroutEggs.csv")
lt.lm1 <- lm(eggs~tl,data=LakeTroutEggs)
fitPlot(lt.lm1)
transChooser(lt.lm1)
LakeTroutEggs[tl>870 & tl<880,]        # finding fish with similar length to #96
(eggs[96]-eggs[94])/eggs[94]
lt.lm1$residuals[96]/lt.lm1$fitted.values[96]
rank(tl)[96]; rank(eggs)[96]
LakeTroutEggs1 <- LakeTroutEggs[-96,]  # remove #96
lt.lm2 <- lm(eggs~tl,data=LakeTroutEggs1)
fitPlot(lt.lm2)
transChooser(lt.lm2)
anova(lt.lm2)
summary(lt.lm2)
confint(lt.lm2)
new <- c(650,750,850)
predictionPlot(lt.lm2,data.frame(tl=new),interval="prediction")
```

4.5.2 Forest Allometrics

Understanding the carbon dynamics in forests is important to understanding ecological processes and managing forests. The amount of carbon stored in a tree is related to tree biomass. Measuring the biomass of a tree is a tedious process that also results in the harvest (i.e., death) of the tree. However, biomass is often related to other simple metrics (e.g., diameter-at-breast-height (DBH) or tree height) that can be made without harming the tree. Thus, forest scientists have developed a series of equations (called allometric equations) that can be used to predict tree biomass from simple metrics for a variety of trees in a variety of locations.

The Miombo Woodlands is the largest continuous dry deciduous forest in the world. It extends across much of Central, Eastern and Southern Africa including parts of Angola, the Democratic Republic of Congo, Malawi, Mozambique, Tanzania, Zambia and Zimbabwe. The woodlands are rich in plant diversity and have the potential to contain a substantial amount of carbon. There is, however, significant uncertainty in the amount of biomass carbon in the Miombo Woodlands. The objective of this study (Kuyah *et al.* 2016) is to develop allometric equations that can be used to reliably estimate biomass of trees in the Miombo Woodlands so that biomass carbon can be more reliably estimated.

Data Collection

Trees for building allometric models were sampled from three 10 km by 10 km sites located in the districts of Kasungu, Salima, and Neno. A total of 88 trees (33 species) were harvested from six plots in Kasungu, seventeen plots in Salima, and five plots in Neno. The DBH (cm) of each tree was measured using diameter tape. Each tree was felled by cutting at the lowest possible point using a chainsaw. The length (m) of the felled tree was measured along the longest axis with a measuring tape. This measurement was used as the total tree height in the analysis. Felled trees were separated into stem, branches, and twigs (leaves and small branches). Total biomass (kg) of the tree (above-ground biomass; AGB) and the separate biomasses (kg) of the stems, branches, and twigs was recorded. The data are stored in **TreesMiombo.csv** (view, download, meta). This analysis will attempt to predict AGB from DBH.

EDA & Assumption Checking

An initial EDA of the data indicate highly skewed distributions, high variability, and univariate outliers in both AGB and DBH. In addition, the residual plot (Figure 4.23-Right) indicated a lack of linearity, homoscedasticity, and normality (Anderson-Darling $p = 0.004$) and the presense of outliers (outlier test $p < 0.0005$). These results all indicate that a transformation should be explored.

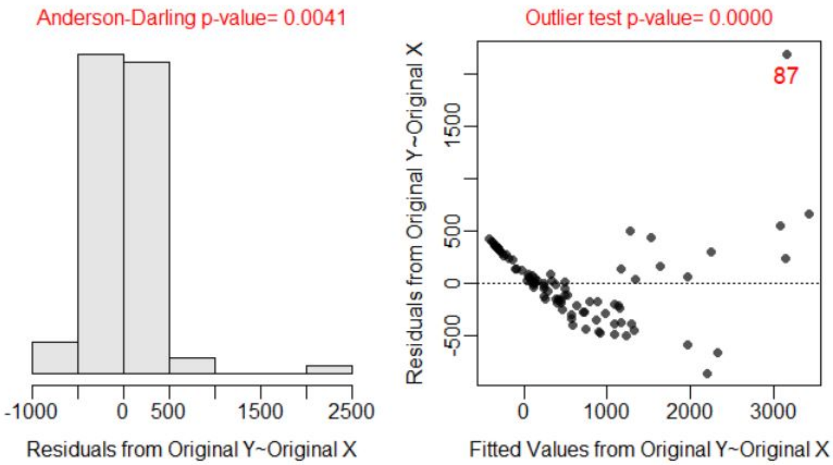

Figure 4.23. Histogram of residuals (Left) and residual plot (Right) for the regression of above-ground biomass (AGB) on diameter-at-breas-height (DBH).

Allometric relationships between weights (i.e., biomass) and lengths (i.e., height) tend to follow power functions, which can be linearized with the log of both variables. The relationship between log(AGB) and log(DBH) was largely linear and homoscedastic (Figure 4.24-Right), the residuals appear to be normal (Anderson-Darling $p = 0.446$; Figure 4.24-Left), and no outliers are present (outlier test $p > 1$). The linear regression model was fit to the log-log transformed data as all assumptions were met.

Results

The relationship between log(AGB) and log(DBH) was significantly positive (ANOVA $p < 0.0005$; Figure 4.25). In fact, it appears that as the log(DBH) increases by one unit that the average log(AGB) increases by approximately 2.27 (Table 4.9) units with a 95% confidence interval between 2.17 and 2.37 units (Table 4.10). Alternatively, a one unit increase in DBH results in an average 8.80 to 10.65 **times** increase in AGB.

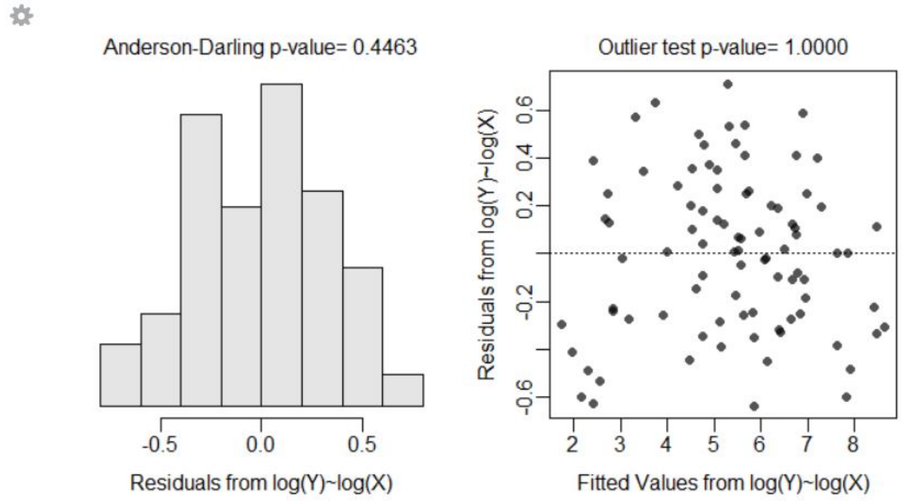

Figure 4.24. Histogram of residuals (Left) and residual plot (Right) for the regression of natural-log transformed above-ground biomass (AGB) on natural-log transformed diameter-at-breast-height (DBH).

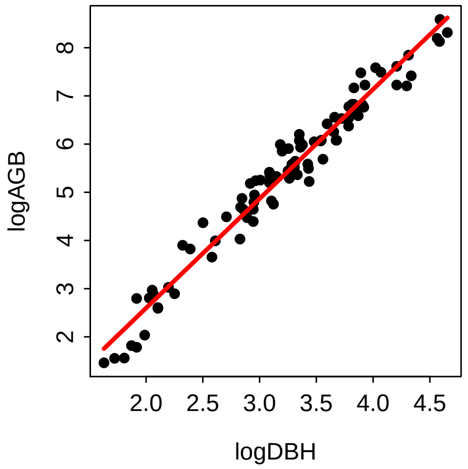

Figure 4.25. Fitted line plot for the regression of natural-log transformed above-ground biomass (AGB) on natural-log transformed diameter-at-breast-height (DBH).

As an example, suppose that interest is in predicting the mean above-ground biomass for all trees with a DBH of 50 cm. This prediction is accomplished by first converting the given DBH to log(DBH) and plugging that value into the equation of the best-fit line to predict the log(AGB) (Table 4.11). The predicted value and the end points of the confidence interval are then used as the powers of e to “back-transform” the results to the original scale. These back-transformed values are 935.15 and 1134.56 kg. However, because the natural log transformation was used for the response variable, these back-transformed values should be corrected for back-transformation bias by multiplying each value by $e^{\frac{s^2_{Y|X}}{2}} = e^{\frac{0.3088^2}{2}} = e^{0.109} = 1.116$.[26] With this correction, the mean above-ground biomass for all systems with a DBH of 50 cm would be expected to be between 1043.29 and 1265.76 kg.

[26] Note that $s^2_{Y|X}$=0.3308 as shown in the “Residual standard error” portion of Table 4.9

Table 4.9. Summary results for the regression of $\mu_{log(AGB)|log(DBH)} = \alpha + \beta_1 log(DBH)$.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.94331    0.15822  -12.28   <2e-16
logDBH       2.27010    0.04798   47.31   <2e-16
---
Residual standard error: 0.3308 on 86 degrees of freedom
Multiple R-squared: 0.963,Adjusted R-squared: 0.9626
F-statistic:  2238 on 1 and 86 DF,  p-value: < 2.2e-16
```

Table 4.10. Confidence intervals for parameters in $\mu_{log(AGB)|log(DBH)} = \alpha + \beta_1 log(DBH)$.

```
                2.5 %    97.5 %
(Intercept) -2.257842 -1.628781
logDBH       2.174707  2.365484
```

Table 4.11. Results for the prediction of the mean log(AGB) for a DBH of 50 cm.

```
      fit      lwr      upr
 6.937355 6.840709 7.034002
```

Conclusion

A significant, and very strong, relationship was found between the natural log above-ground biomass and the natural log diameter-at-breast-height for trees in the Miombo Woodlands. A model was developed from this relationship that can be used to predict above-ground biomass of a tree from the measured DBH of the tree.

Appendix – R commands

```
tm <- read.csv("TreesMiombo.csv")
tm.lm1 <- lm(AGB~DBH,data=tm)
transChooser(tm.lm1)
tm$logDBH <- log(tm$DBH)
tm$logAGB <- log(tm$AGB)
tm.lm2 <- lm(logAGB~logDBH,data=tm)
transChooser(tm.lm2)
anova(tm.lm2)
fitPlot(tm.lm2)
summary(tm.lm2)
confint(tm.lm2)
pred2 <- predict(tm.lm2,data.frame(logDBH=log(50)),interval="confidence")
exp(pred2)
```

4.6 Summary Process

The following is a template for a process of fitting a simple linear regression model. Consider this process as you learn to fit one-way ANOVA models, but don't consider this to be a concrete process for all models.

1. Perform a thorough EDA.
   - Pay close attention to the form, strength, and outliers on the scatterplot [`plot()`] of the response and explanatory variables.
2. Fit the untransformed ultimate full model [`lm()`].
3. Check the assumptions of the fit of the model [`transChooser()`].
   - Check the linearity of the relationship a residual plot.
   - Check homoscedasticity with a fitted-line plot and residual plot.
   - Check normality of residuals with an Anderson-Darling test and histogram of residuals.
   - Check for outliers and influential points with the outlier test and residual plot.
4. If an assumption or assumptions are violated, then attempt to find a transformation where the assumptions are met.
   - Use the trial-and-error method [`transChooser()`], theory, or experience to identify possible transformations for the response variable and, possibly, for the explanatory variable.
   - If only an "unusual" or influential observation exists (i.e., linear, homoscedastic, and normal residuals) and no transformation corrects the "problem," then consider removing that observation from the data set.
5. Fit the ultimate full model with the transformed variable(s) or reduced data set.
6. Construct an ANOVA table for the full model [`anova()`] and interpret the overall F-test.
7. Summarize findings with coefficient results [`summary()`, `confint()`] and fitted-line plot [`fitPlot()`].
8. Make predictions if desired [`predictionPlot()`].