STAT 462

Applied Regression Analysis

Lesson 6: MLR Assumptions, Estimation & Prediction

Overview of this Lesson

This lesson extends the methods from Lesson 4 to the context of multiple linear regression. How do we evaluate a model? How do we know if the model we are using is good? One way to consider these questions is to assess whether the assumptions underlying the multiple linear regression model seem reasonable when applied to the dataset in question. Since the assumptions relate to the (population) prediction errors, we do this through the study of the (sample) estimated errors, the residuals.

Next, we focus our efforts on *using a multiple linear regression model* to answer two specific research questions, namely:

- What is the average response for a given set of values of the predictors $x_1, x_2, ...$?
- What is the value of the response likely to be for a given set of values of the predictors $x_1, x_2, ...$?

In particular, we will learn how to calculate and interpret:

- A confidence interval for estimating the mean response for a given set of values of the predictors $x_1, x_2, ...$
- A prediction interval for predicting a new response for a given set of values of the predictors $x_1, x_2, ...$

Key Learning Goals for this Lesson:
▪ Understand why we need to check the assumptions of our model. ▪ Know the things that can go wrong with the linear regression model. ▪ Know how we can detect various problems with the model using a residuals vs. fits plot. ▪ Know how we can detect various problems with the model using a residuals vs. predictor plot. ▪ Know how we can detect a certain kind of dependent error terms using a residuals vs. order plot. ▪ Know how we can detect non-normal error terms using a normal probability plot. ▪ Apply some numerical tests for assessing model assumptions. ▪ Distinguish between estimating a mean response (confidence interval) and predicting a new observation (prediction interval). ▪ Understand the various factors that affect the width of a confidence interval for a mean response. ▪ Understand why a prediction interval for a new response is wider than the corresponding confidence interval for a mean response. ▪ Know the formula for a prediction interval depends strongly on the condition that the error terms are normally distributed, while the formula for the confidence interval is not so dependent on this condition for large samples.

- Know the types of research questions that can be answered using the materials and methods of this lesson.

6.1 - MLR Model Assumptions ›
(/stat462/node/145)

STAT 462

Applied Regression Analysis

6.1 - MLR Model Assumptions

The four conditions ("**LINE**") that comprise the multiple linear regression model generalize the simple linear regression model conditions to take account of the fact that we now have multiple predictors:

- The mean of the response , $\mathrm{E}(Y_i)$, at each set of values of the predictors, $(x_{1i}, x_{2i}, \ldots)$, is a **Linear function** of the predictors.
- The errors, ε_i, are **Independent**.
- The errors, ε_i, at each set of values of the predictors, $(x_{1i}, x_{2i}, \ldots)$, are **Normally distributed**.
- The errors, ε_i, at each set of values of the predictors, $(x_{1i}, x_{2i}, \ldots)$, have **Equal variances** (denoted σ^2).

An equivalent way to think of the first (linearity) condition is that the mean of the error, ϵ_i, at each set of values of the predictors, $(x_{1i}, x_{2i}, \ldots)$, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2.

As in simple linear regression, we can assess whether these conditions seem to hold for a multiple linear regression model applied to a particular sample dataset by looking at the estimated errors, i.e., the residuals, $e_i = y_i - \hat{y}_i$.

‹ Lesson 6: MLR Assumptions, Estimation & Prediction (/stat462/node/84) | up (/stat462/node/84) | 6.2 - Assessing the Model Assumptions › (/stat462/node/146)

STAT 462

Applied Regression Analysis

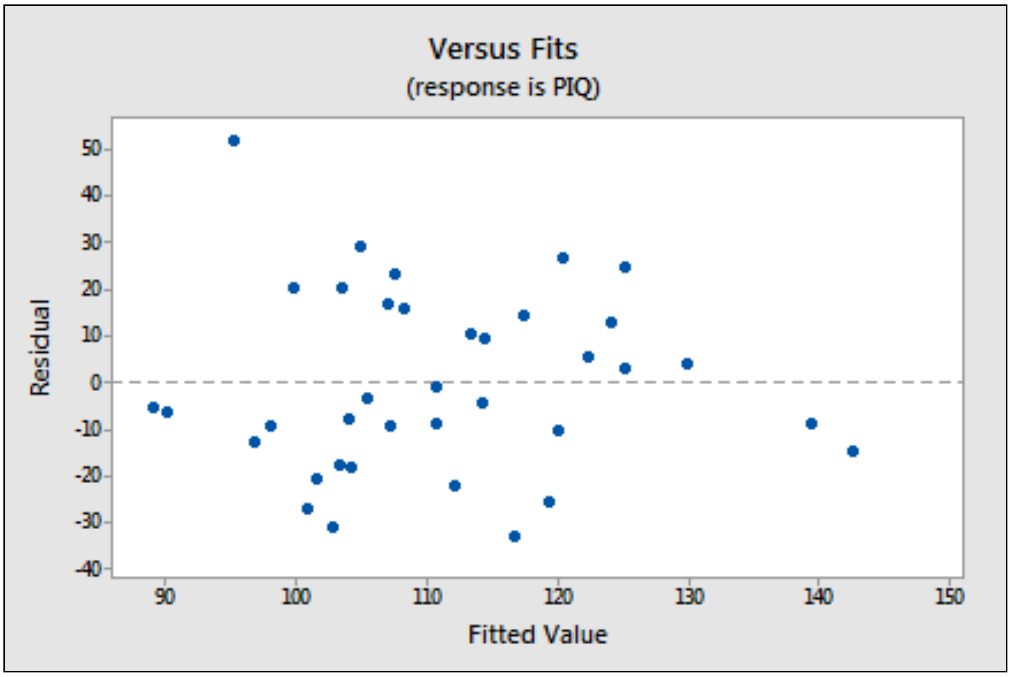
6.2 - Assessing the Model Assumptions

We can use all the methods we learnt about in Lesson 4 to assess the multiple linear regression model assumptions:

- Create a scatterplot with the residuals, e_i, on the vertical axis and the fitted values, $\hat{y}_i$, on the horizontal axis and visual assess whether:
  - the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
  - the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
  - there are no excessively outlying points (we'll explore this in more detail in Lesson 9).
  - violation of any of these three may necessitate remedial action (such as transforming one or more predictors and/or the response variable), depending on the severity of the violation (we'll explore this in more detail in Lesson 7).
- If the data observations were collected over time (or space) create a scatterplot with the residuals, e_i, on the vertical axis and the time (or space) sequence on the horizontal axis and visual assess whether there is no systematic non-random pattern (this affirms the "I" condition).
  - Violation may suggest the need for a time series model (a topic that is not considered in this course).
- Create a series of scatterplots with the residuals, e_i, on the vertical axis and each of the predictors in the model on the horizontal axes and visual assess whether:
  - the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
  - the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
  - violation of either of these for at least one residual plot may suggest the need for transformations of one or more predictors and/or the response variable (again we'll explore this in more detail in Lesson 7).
- Create a histogram, boxplot, and/or normal probability plot of the residuals, e_i to check for approximate normality (the "N" condition). (Of these plots, the normal probability plot is generally the most effective.)
- Create a series of scatterplots with the residuals, e_i, on the vertical axis and each of the available predictors that have been omitted from the model on the horizontal axes and visual assess whether:
  - there are no strong linear or simple nonlinear trends in the plot;
  - violation may indicate the predictor in question (or a transformation of the predictor) might be usefully added to the model.
  - it can sometimes be helpful to plot functions of predictor variables on the horizontal axis of a residual plot, for example interaction terms consisting of one quantitative predictor multiplied by another quantitative predictor. A strong linear or simple nonlinear trend in the resulting plot may indicate the variable plotted on the horizontal axis might be usefully added to the model.

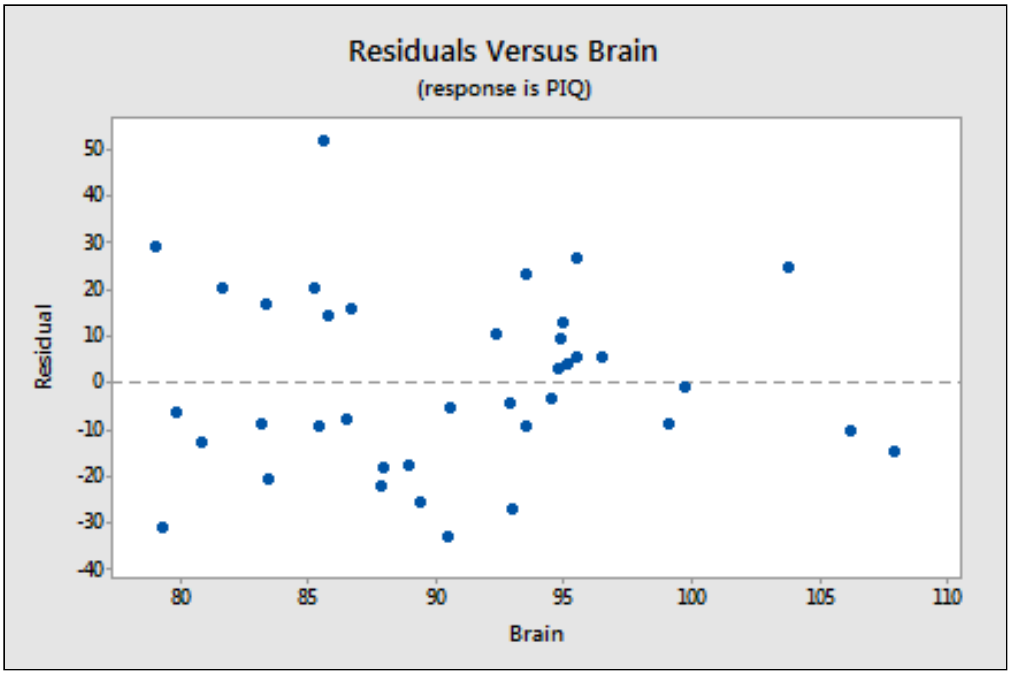
As you can see, checking the assumptions for a multiple linear regression model comprehensively is not a trivial undertaking! But, the more thorough we are in doing this, the greater the confidence we can have in our model. To illustrate, let's do a residual analysis for the example on IQ and physical characteristics from Lesson 5 (iqsize.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/iqsize.txt)), where we've fit a model with *PIQ* as the response and *Brain* and *Height* as the predictors:

- First, here's a residual plot with the residuals, e_i, on the vertical axis and the fitted values, $\hat{y}_i$, on the horizontal axis:



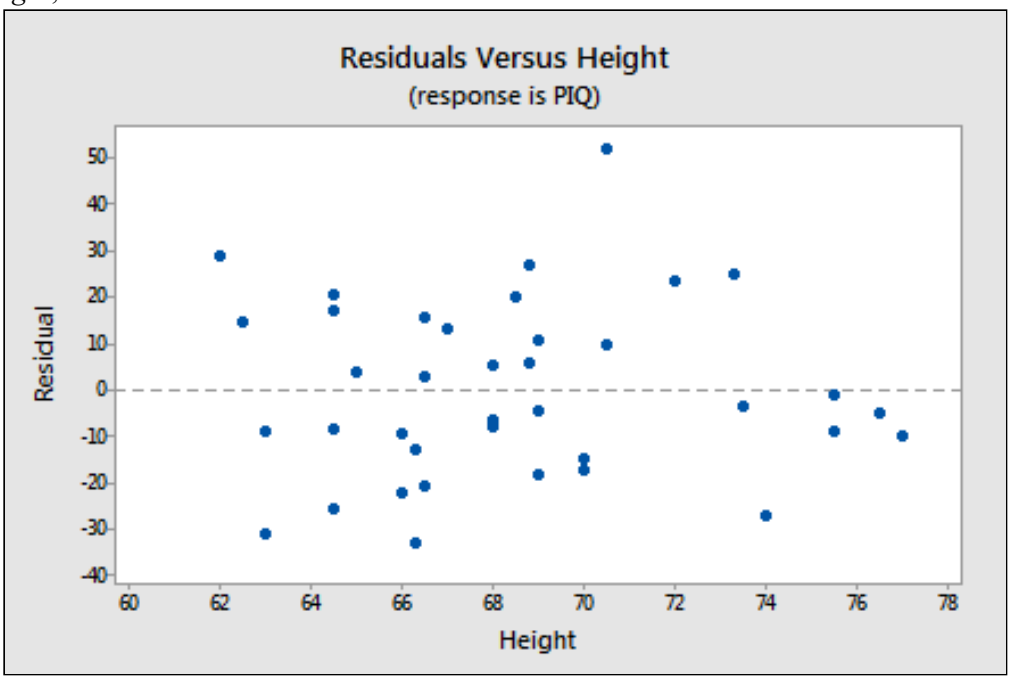
  As we scan the plot from left to right, the average of the residuals remains approximately 0, the variation of the residuals appears to be roughly constant, and there are no excessively outlying points (except perhaps the observation with a residual of about 50, which might warrant some further investigation but isn't too much of a worry). Note that the two observations on the right of the plot with fitted values close to 140 are of no concern with respect to the model assumptions. We'd be reading too much into the plot if we were to worry that the residuals appear less variable on the right side of the plot (there are only 2 out of a total of 38 points here and hence there is little information on residual variability in this region of the plot).

- There is no time (or space) variable in this dataset so the next plot we'll consider is a scatterplot with the residuals, e_i, on the vertical axis and one of the predictors in the model, *Brain*, on the horizontal axis:



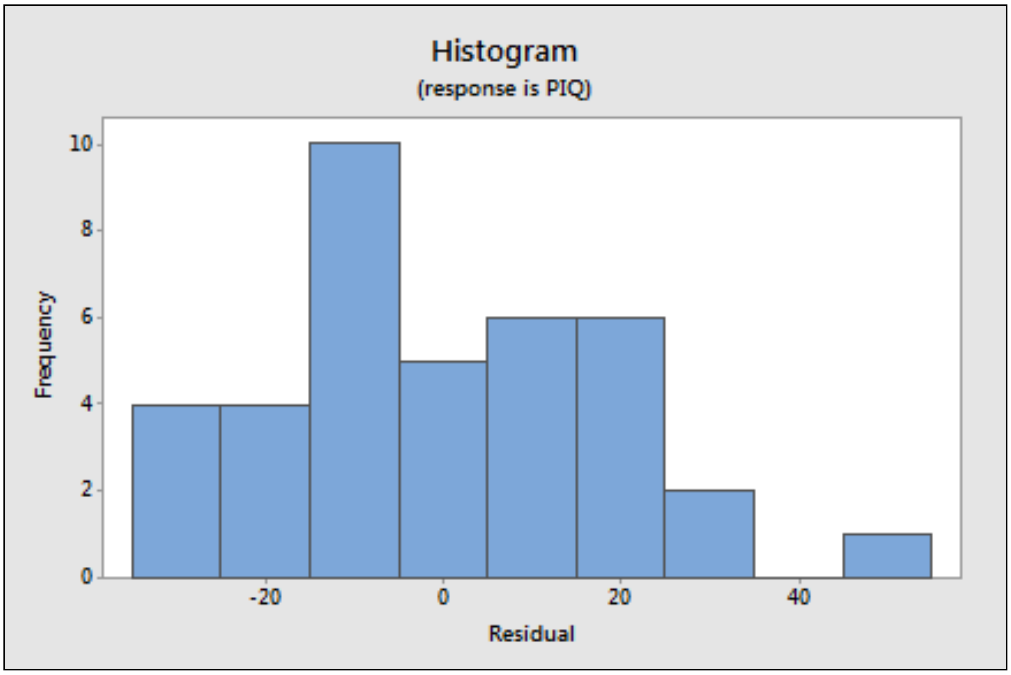
Again, as we scan the plot from left to right, the average of the residuals remains approximately 0, the variation of the residuals appears to be roughly constant, and there are no excessively outlying points. Also, there is no strong nonlinear trend in this plot that might suggest a transformation of *PIQ* or *Brain* in this model.

- The next plot we'll consider is a scatterplot with the residuals, e_i, on the vertical axis and the other predictor in the model, *Height*, on the horizontal axis:



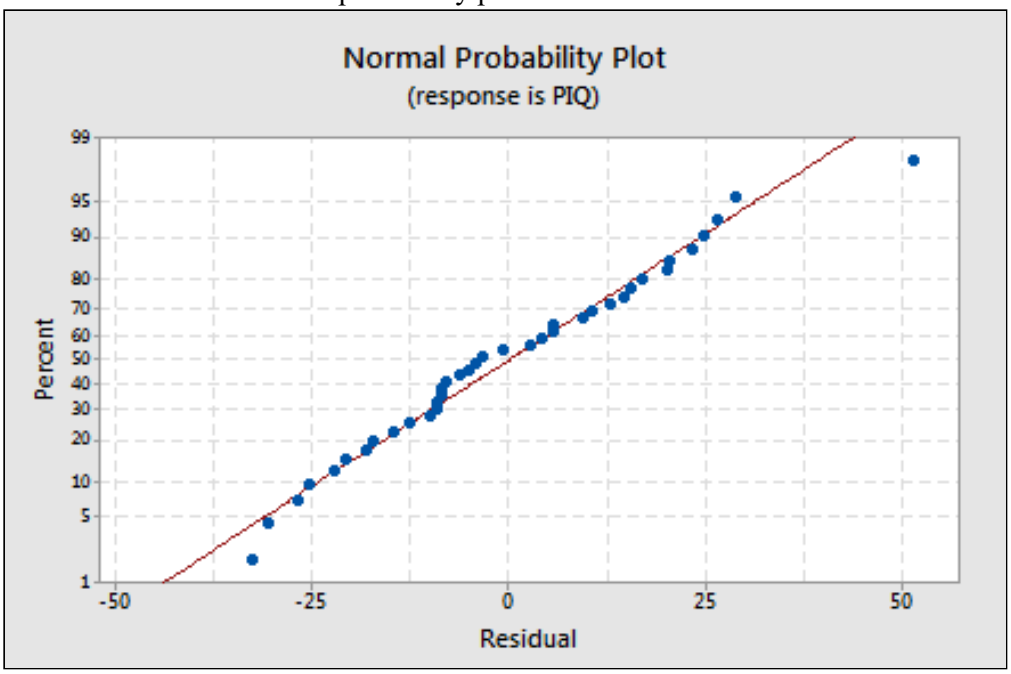
Again, as we scan the plot from left to right, the average of the residuals remains approximately 0, the variation of the residuals appears to be roughly constant, and there are no excessively outlying points. Also, there is no strong nonlinear trend in this plot that might suggest a transformation of *PIQ* or *Height* in this model.

- The next plot we'll consider is a histogram of the residuals:



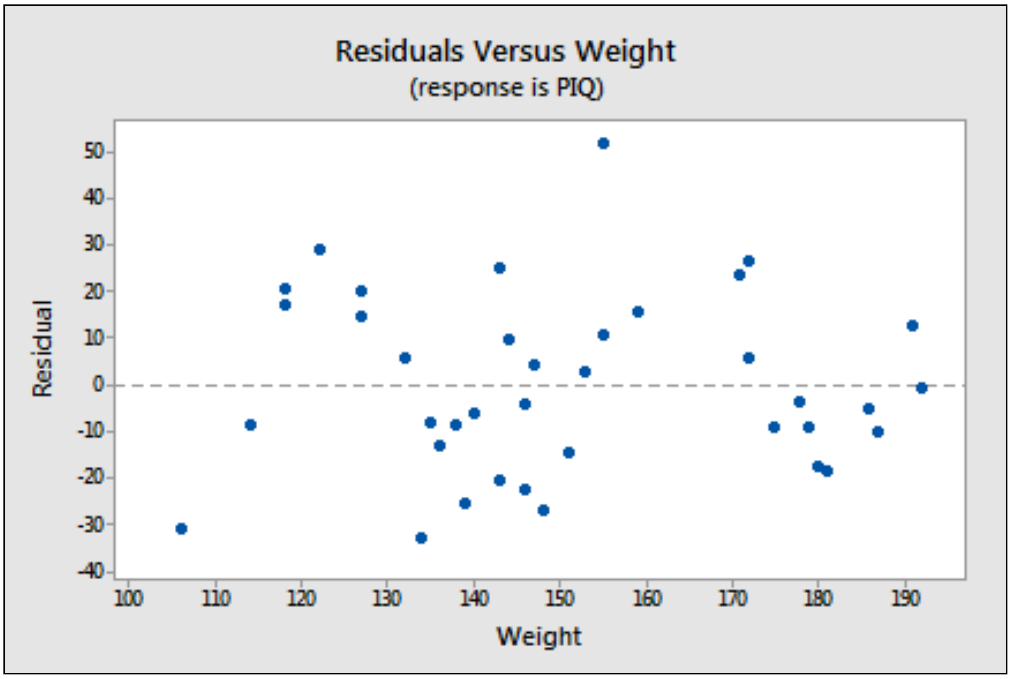
Although this doesn't have the ideal bell-shaped appearance, given the small sample size there's little to suggest violation of the normality assumption.

- Since the appearance of a histogram can be strongly influenced by the choice of intervals for the bars, to confirm this we can also look at a normal probability plot of the residuals:



Again, given the small sample size there's little to suggest violation of the normality assumption.

- The final plot we'll consider for this example is a scatterplot with the residuals, e_i, on the vertical axis and the only predictor excluded from the model, *Weight*, on the horizontal axis:



Since there is no strong linear or simple nonlinear trend in this plot, there is nothing to suggest that *Weight* might be usefully added to the model. Don't get carried away by the apparent "up-down-up-down-up" pattern in this plot. This "trend" isn't nearly strong enough to warrant adding some complex function of *Weight* to the model - remember we've only got a sample size of 38 and we'd have to use up at least 5 degrees of freedom trying to add a fifth-degree polynomial of *Weight* to the model. All we'd end up doing if we did this is over-fitting the sample data and ending up with an over-complicated model that predicts new observations very poorly.

One key idea to draw from this example is that if you stare at a scatterplot of completely random points long enough you'll start to see patterns even when there are none! Residual analysis should be done thoroughly and carefully but without over-interpreting every slight anomaly. Serious problems with the multiple linear regression model generally reveal themselves pretty clearly in one or more residual plots. If a residual plot looks "mostly OK," chances are it is fine.

‹ 6.1 - MLR Model Assumptions (/stat462/node/145) up (/stat462/node/84) 6.3 - Tests for Error Normality › (/stat462/node/147)

STAT 462

Applied Regression Analysis

6.5 - Confidence Interval for the Mean Response

In this section, we are concerned with the confidence interval for the mean response μ_Y when the predictor values are $\mathbf{X}_h = (1, X_{h,1}, X_{h,2}, \ldots, X_{h,k})^{\mathrm{T}}$. The general formula in words is as always:

Sample estimate ± (*t*-multiplier × standard error)

First we define the standard error of the fit at $\mathbf{X}_h$ given by:

$$\text{se}(\hat{y}_h) = \sqrt{\text{MSE}(\mathbf{X}_h^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_h)}.$$

and the confidence interval is:

$$\hat{y}_h \pm t_{(\alpha/2, n-(k+1))} \times \text{se}(\hat{y}_h)$$

where:

- $\hat{y}_h$ is the "**fitted value**" or "**predicted value**" of the response when the predictor values are $\mathbf{X}_h$.
- $t_{(\alpha/2, n-(k+1))}$ is the "***t*-multiplier**." Note that the *t*-multiplier has *n*-(*k*+1) degrees of freedom because the confidence interval uses the mean square error (*MSE*) whose denominator is *n*-(*k*+1).

Fortunately, we won't have to use the formula to calculate the confidence interval, since statistical software will do the dirty work for us. Here is software output for the example on IQ and physical characteristics from Lesson 5 (iqsize.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/iqsize.txt)), where we've fit a model with *PIQ* as the response and *Brain* and *Height* as the predictors:

```
Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI            95% PI
      1  105.64    3.65  (98.24, 113.04)  (65.35, 145.93)


Values of Predictors for New Observations

New Obs  Brain  Height
      1   90.0    70.0
```

Here's what the output tells us:

- The section labeled "**Values of Predictors for New Observations**" reports the values for $\mathbf{X}_h$ (brain size = 90 and height = 70) for which we requested the confidence interval for μ_Y.

- The section labeled "**Predicted Values for New Observations**" reports the 95% confidence interval. We can be 95% confident that the average performance IQ score of all college students with brain size = 90 and height = 70 is between 98.24 and 113.04 counts per 10,000.
- The section labeled "**Predicted Values for New Observations**" also reports the predicted value $\hat{y}_h$, ("**Fit**" = 105.64), the standard error of the fit ("**SE Fit**" = 3.65), and the 95% prediction interval for a new response (which we discuss in the next section).

Factors affecting the width of the *t*-interval for the mean response μ_Y

As always, the formula is useful for investigating what factors affect the width of the confidence interval for μ_Y.

- **As the mean square error (*MSE*) decreases, the width of the interval decreases.** Since *MSE* is an estimate of how much the data vary naturally around the unknown population regression hyperplane, we have little control over *MSE* other than making sure that we make our measurements as carefully as possible.
- **As we decrease the confidence level, the *t*-multiplier decreases, and hence the width of the interval decreases.** In practice, we wouldn't want to set the confidence level below 90%.
- **As we increase the sample size *n*, the width of the interval decreases.** We have complete control over the size of our sample — the only limitation beting our time and financial constraints.
- The closer $\mathbf{X}_h$ is to the average of the sample's predictor values, the narrower the interval.

Let's see this last claim in action for our IQ example:

```
Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI           95% PI
      1  104.81    6.60  (91.42, 118.20)  (63.00, 146.62)


Values of Predictors for New Observations

New Obs  Brain  Height
      1   79.0    62.0
```

The width of the first confidence interval we calculated earlier (113.04 - 98.24 = 14.80) is shorter than the width of this new interval (118.20 - 91.42 = 26.78), because 90 and 70 are much closer than 79 and 62 are to the sample means (90.7 and 68.4).

When is it okay to use the formula for the confidence interval for μ_Y?

- When $\mathbf{X}_h$ is within the "**scope of the model**." But, note that $\mathbf{X}_h$ does not have to be an actual observation in the data set.
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. The formula works okay even if the error terms are only approximately normal. And, if you have a large sample, the error terms can even deviate substantially from normality.

‹ 6.4 - Tests for Constant Error Variance (/stat462/node/148) up (/stat462/node/84) 6.6 - Prediction Interval for a New Response › (/stat462/node/151)

STAT 462

Applied Regression Analysis

6.6 - Prediction Interval for a New Response

In this section, we are concerned with the prediction interval for a new response y_{new} when the predictor values are $\mathbf{X}_h = (1, X_{h,1}, X_{h,2}, \ldots, X_{h,k})^{\mathrm{T}}$. Again, let's just jump right in and learn the formula for the prediction interval. The general formula in words is as always:

Sample estimate ± (*t*-multiplier × standard error)

and the formula in notation is:

$$\hat{y}_h \pm t_{(\alpha/2, n-(k+1))} \times \sqrt{MSE + [\text{se}(\hat{y}_h)]^2}$$

where:

- $\hat{y}_h$ is the "**fitted value**" or "**predicted value**" of the response when the predictor values are $\mathbf{X}_h$.
- $t_{(\alpha/2, n-(k+1))}$ is the "***t*-multiplier**." Note again that the *t*-multiplier has *n*–(*k*+1) degrees of freedom, because the prediction interval uses the mean square error (*MSE*) whose denominator is *n*–(*k*+1).
- $\sqrt{MSE + [\text{se}(\hat{y}_h)]^2}$ is the "**standard error of the prediction**," which is very similar to the "standard error of the fit" when estimating μ_Y. The standard error of the prediction just has an extra MSE term added that the standard error of the fit does not. (More on this a bit later.)

Again, we won't use the formula to calculate our prediction intervals. We'll let statistical software do the calculation for us. Let's look at the prediction interval for our IQ example(iqsize.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/iqsize.txt)):

```
Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI            95% PI
      1  105.64    3.65  (98.24, 113.04)  (65.35, 145.93)


Values of Predictors for New Observations

New Obs  Brain  Height
      1   90.0    70.0
```

The output reports the 95% prediction interval for an individual college student with brain size = 90 and height = 70. We can be 95% confident that the performance IQ score of an individual college student with brain size = 90 and height = 70 will be between 65.35 and 145.93 counts per 10,000.

When is it okay to use the prediction interval for y_{new} formula?

The requirements are similar to, but a little more restrictive than, those for the confidence interval. It is okay:

- When $\mathbf{X}_h$ is within the "scope of the model." Again, $\mathbf{X}_h$ does not have to be an actual observation in the data set.
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. Unlike the case for the formula for the confidence interval, the formula for the prediction interval depends **strongly** on the condition that the error terms are normally distributed.

Understanding the difference in the two formulas

In our discussion of the confidence interval for μ_Y, we used the formula to investigate what factors affect the width of the confidence interval. There's no need to do it again. Because the formulas are so similar, it turns out that the factors affecting the width of the prediction interval are identical to the factors affecting the width of the confidence interval.

Let's instead investigate the formula for the prediction interval for y_{new}:

$$\hat{y}_h \pm t_{(\alpha/2,n-(k+1))} \times \sqrt{MSE + [\text{se}(\hat{y}_h)]^2}$$

to see how it compares to the formula for the confidence interval for μ_Y:

$$\hat{y}_h \pm t_{(\alpha/2,n-(k+1))} \times \text{se}(\hat{y}_h)$$

Observe that the only difference in the formulas is that the standard error of the prediction for y_{new} has an extra *MSE* term in it that the standard error of the fit for μ_Y does not. If you're not sure why this makes sense, re-read Section 4.11 on "Prediction Interval for a New Response" in the context of simple linear regression.

What's the practical implications of the difference in the two formulas?

- Because the prediction interval has the extra *MSE* term, a (1-α)100% confidence interval for μ_Y at $\mathbf{X}_h$ will always be narrower than the corresponding (1-α)100% prediction interval for y_{new} at $\mathbf{X}_h$.
- By calculating the interval at the sample means of the predictor values and increasing the sample size *n*, the confidence interval's standard error can approach 0. Because the prediction interval has the extra *MSE* term, the prediction interval's standard error cannot get close to 0.

‹ 6.5 - Confidence Interval for the Mean Response (/stat462/node/150) up (/stat462/node/84)