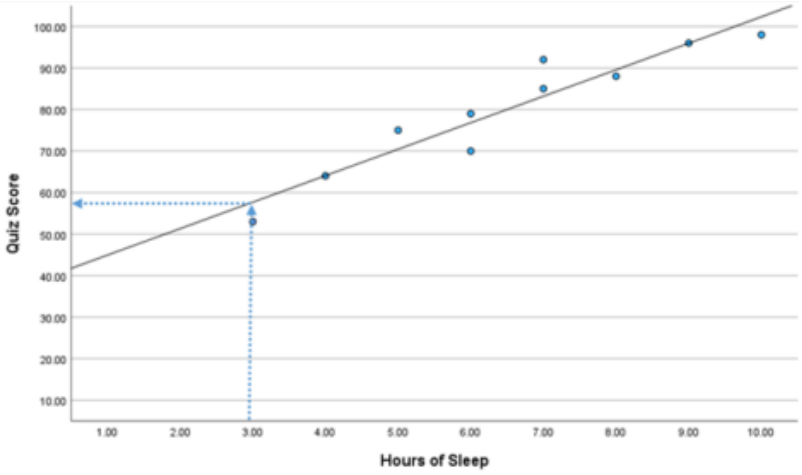
13.8: Visualizing Predictions and Residuals in Regression

ANOVA is used to test whether the regression model is a good fit to the data and, thus, is useful for predicting Y. It does this by computing and comparing the errors in predictions when using the regression model to an alternative model using the mean of Y. The computations can be easier to understand if we can visualize what each model does and what their errors represent before learning how to compute the corresponding parts of the ANOVA. Thus, this section will review how each model can be represented in graphical form.

Visualizing Predictions Using the Regression Equation

Predictions are made based on the regression line. The regression line summarizes where scores are expected to be. Later in this chapter we will learn how to precisely calculate the predicted Y-value. However, for this section we are reviewing how to estimate values visually to help us understand the logic of regression. With that in mind, let's take a look at an example using Graph 13.1 below. Suppose we wanted to predict the exam score for someone who slept 3 hours.

To locate the prediction on the graph, we would look over to 3.00 on the x-axis because that corresponds to getting 3 hours of sleep (i.e. 3 units for the predictor variable). We would then find where the regression line crosses over X = 3.00; the height of the regression line where it crosses over X = 3.00 indicates the predicted Y-value. We can see in the graph that when X = 3.00, Y is around 58 units high on the y-axis. This is the approximate predicted score for someone who gets 3.00 hours of sleep. To state it another way, someone who gets 3.00 hours of sleep is predicted to get a quiz score of *approximately* 58. Notice the imprecision of the estimate when looking visually. This is why a formula must be used later to get a more precision. For now, however, we will stick to the visual estimate.



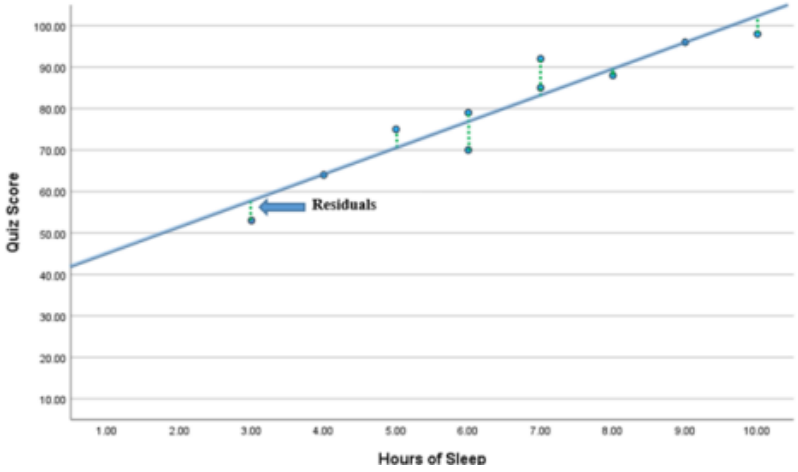
Graph 13.1 Regression Line Used to Predict Y when X = 3.00.

We can compare our predicted Y-value to an actual, known Y-value. Specifically, we can see that there is a data point for which X was 3.00. That data poi falls at X = 3.00 and Y = 53.00. This data point is a known value from Data Set 12.1 and corresponds to participant number 9. Notice that the known data point is not on the regression line. It is close to the prediction (i.e. it is close to the regression line) but not exact which means that the prediction was pretty good but did have some error. This error is known as a *residual* and is an important part of estimating how accurate and, thus, how useful a regression is for making predictions.

Residuals Using the Regression Line

Residuals represent the amount of *inaccuracy* in the regression predictions. Specifically, **residuals** are the errors in locating actual Y-values when using the regression line and represent the vertical distances between the known bivariate data points and the regression line. Another way say this is that residuals represent the amount of variation in the Y-variable that is not accounted for using the X-variable. Below is Graph 13.2 with the residuals shown for all data points. Notice that the residuals are all shown vertically. This is because we are assessing error in predicting Y and, thus, are concerned with error in locating the data on the y-axis (which is vertical). Recall from Chapter 12 that the correlation between sleep hours and quiz scores for Data Set 12.1 was summarized as r = .95 (See Chapter 12 for review of the computations). This is a very strong, positive correlation. When a correlation is stronger, the average residuals will be lower. Consistent with this, the residuals for Graph 13.2 are fairly small,

overall, as we should expect because the correlation was very strong. Thus, predictions of Y using X will be fairly accurate (though imperfect) for Data Set 12.1.

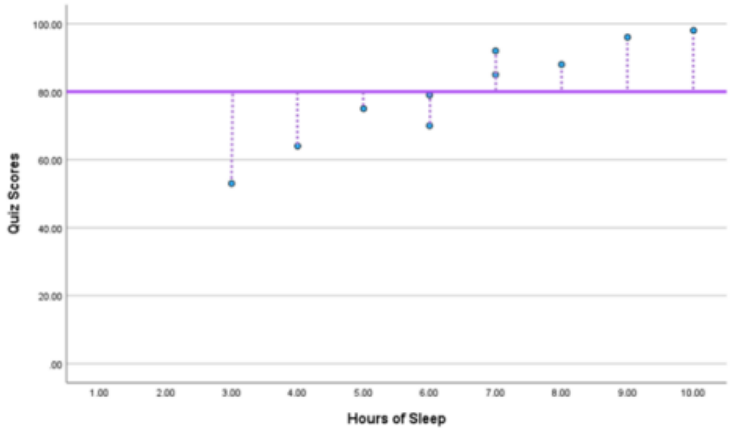


Graph 13.2 Regression Line with Residuals Shown.

Visualizing the Alternative Prediction Model: Using $\bar{Y}$ as the Prediction

To know whether using X to predict Y significantly improves predictions, it needs to be compared to another model. The alternative way to predict Y is simply to predict that all Y-values are equal to the mean of Y. This alternative is based on the fact that means are summaries of what tends to be true of data for a given variable. When data are normally distributed, scores closer to the mean are more common and scores farther from the mean are rarer. Thus, it is logical to use the summary of what tends to be true (which is the mean for the variable) as an alternative way to estimate values.

Let's take a look at the graph using this alternative prediction model (see Graph 13.3). The mean of Y is 80.00 (i.e. $\bar{Y}$ = 80.00) for Data Set 12.1. The prediction line is, therefore, set as $\hat{Y}$ = 80.00 and all Y-values are predicted to be 80.00. This is represented by a non-sloping, horizontal line. Just as we did with the regression line in Graph 13.2, we can draw vertical residual lines from each data point to the prediction line (which in this case is the $\bar{Y}$ line) to represent error in predictions. The alternative model residuals are depicted in Graph 13.3.

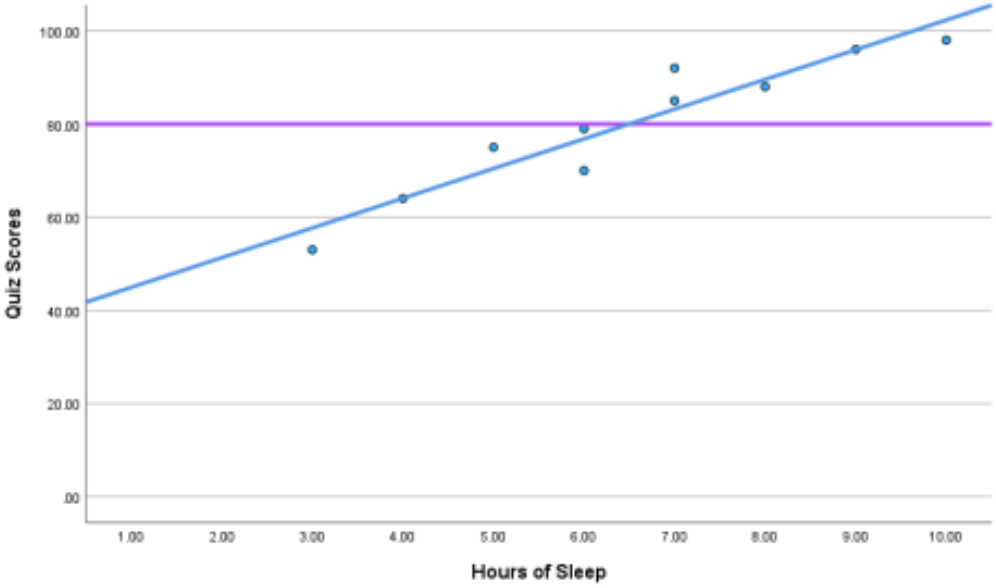


Graph 13.3 Mean of Y Line with Residuals Shown.

Visually Comparing Model Fit: Comparing Predictions with the Regression Line vs. the $\bar{Y}$ Line

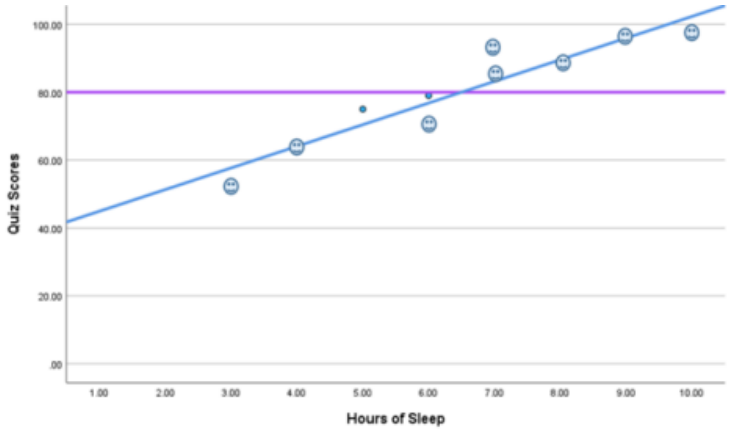
Model fit refers to how good a model is at making predictions. It is called *model fit* because we are asking how well the model fit the data. For regression this means we want to know how close the prediction lines were to the actual data. The lesser the residuals, the better the fit and the greater the residuals, the poorer the fit. The best way to assess the fit of a regression model is by pitting it against another prediction model. This is precisely what is done in simple linear regression. The residuals from two models are compared: 1.The model predicting Y using X (the regression model) and 2. The model predicting each Y is equal to the $\bar{Y}$.

Let's compare the two models visually to see which is a better fit to the data and, thus, is more useful for predicting Y-values. Graph 13.4 shows both models and their residuals. This version overlays the regression line (which is blue and has a positive slope) to the $\bar{Y}$ line (which is purple and has no slope). Whichever model's line is closer to a data point is the one which is better at predicting that data point.



Graph 13.4 Comparison of Prediction Models

Let's make it more overt by marking how many of the data points were closer to the regression line than the $\bar{Y}$ line with a smiley face. Because our hypothesis is that the regression line will be the better predictor, Graph 13.5 shows smiling faces over the dots which are better predicted by the regression line than the alternative model. We can see that for 8 of the 10 data points, the regression model was better. This means the residuals were lower using the regression model for 8 of the 10 data points than when using the alternative model. For one of the remaining data points, the two models were about equally accurate in their prediction and for the other the alternative model was more accurate. Overall, we can see that the residuals are lower for the regression model, in general, than for the alternative model. Thus, though the regression model was not always more accurate, it tended to be the better model for predicting.



Graph 13.5 Data Points Better Predicted by the Regression Model than $\bar{Y}$

In this section we have taken the time to visually understand:

1. the two models used to predict Y-values,
2. what residuals represent in the models, and
3. how the two models are being compared for fit based on their residuals.

In a regression, the residuals of the two models are calculated and used to test whether the regression model provides significantly improved predictions over the alternative model. We will now to turn to those calculations. You may find it helpful to refer back to the visuals in this section as we cover those calculations to remind yourself what is being represented in each part of the regression ANOVA.

1. How is a predicted Y-value found on the graph of the regression model?
2. What does a residual represent when using a regression line to predict a Y-value?
3. How is a predicted Y-value found in the alternative prediction model?
4. Which model has a sloping prediction line?
5. What is compared to test whether the regression model improves predictions compared to the alternative model?
