Prediction for Simple Linear Regression

Simple linear regression – Stat 230

In this class, you will work with your group to review inference for simple linear regression models with a quantitative predictor, and also learn how to create intervals for predictions. While you work through this activity, make sure that all group members are engaged and contribute ideas, and also follow the code. The R Manual has useful R code for today's activities.

Review

We have already thought quite a bit about simple linear regression, but we were using a categorical predictor variable with two levels. To begin, let's explore the `Cars` data set and regression model discussed in Section 3.1 of your textbook.

Recall that the simple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Task 1. (Ch. 3 Activity 1) Load in the `Cars` data set and create a scatterplot to display the relationship between `Mileage` and `Price`. Be sure that the x- and y-axis labels are informative. Describe the relationship between the two variables.

Task 2 (Adaptation of Activity 2) Fit the simple linear regression model where `Mileage` is used to predict `Price`. Report the following information from the fitted model: the fitted regression equation, R^2, the correlation between `Price` and `Mileage`, and the coefficient table containing the t-statistics, standard errors, and p-values for the y-intercept and sloe. (Note that R^2 is called `Multiple R-squared` in the summary of a regression model.)

Task 3 Provide interpretations of the estimate y-intercept and slope in context.

Task 4 Provide an interpretation of the R^2 value.

Task 5 State the hypotheses and draw a conclusion in context based on the t-statistic and p-value from the coefficient table.

Task 6 To add the fitted regression line to your scatterplot from Task 1, add a *layer* to your plot. Specifically, fill in your plotting code from Task 1 in the below code chunk.

```
___ |> gf_lm()
```

Task 7. (Adaptation of Activity 3) The first car in the data set is a Buick Century with 8221 miles. Calculate the residual value for this car (the observed retail price minus the expected price calculated from the regression line). Based on this value, did the model over- or under-predict the sales price?

Prediction intervals

You just thought about how the model made a prediction for a Buick Century, now it's time to think about how to provide ranges of plausible values for predictions made by the regression model. As you can see from the plot of your fitted model, there is plenty of variability of the observed points around the regression line. This variability needs to be considered when we are making predictions—providing only a single value for a prediction will almost surely be wrong, but an interval will capture a range of plausible values for our prediction!

There are two types of prediction problems:

- We can predict the mean response at a specific value of x, that is we could be interested in $E(Y|X)$. For example, we might be interested in the average sales price of a car with 40,000 miles.
- We can predict the response for a specific future observation. For example, you might be interested in predicting the sales price of your car, which has 40,000 miles on it. (Yes, I know the data from from 2005, but you get the point.)

Task 8. Think of two additional examples of each type of prediction. Come up with new hypothetical situations here, so don't use a car price example here.

Once you have determined which type of prediction you want to make, it's time to construct a confidence interval for your prediction. If you are predicting the mean response, then we'll call this a *confidence interval*. If you are predicting the response for a specific future observation, then we'll call this a *prediction interval*. Both intervals have the same familiar form:

$$\text{estimate} \pm \text{multiplier} \times \text{SE}$$

Both intervals use the same prediction, $\hat{y}_i$ the value on the regression line, and both intervals also use the same multiplier, the $1-\alpha/2$ quantile from a t-distribution with $n-2$ degrees of freedom, which we usually label as t^* (and is the same t^* we use for the CI for the slope). The difference in the intervals shows up in the standard errors.

Standard errors for predictions:

- The standard error for the confidence interval for $\widehat{E}(Y|X)$ is $\text{SE}(\widehat{E}(Y|X)) = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}$.
- The standard error for the prediction interval of a new observation $\hat{y}$ is $\text{SE}(\hat{y}) = \hat{\sigma}\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}$

In the above standard error formulas, $\hat{\sigma}$ is the estimated residual standard deviation and x_0 is the value of x at which we are making a prediction.

Task 9. Looking at the standard errors for the two intervals, which interval will be wider? Why does this make sense? (Try to give an intuitive explanation.)

Task 10. As x_0 gets farther from $\overline{x}$ what happens to the standard errors?

Predictions in R

The standard error formulas are quite tedious, so we'll use R to perform all of the computation. Once you have a fitted regression model, the `predict()` function can use used to make predictions and calculate the intervals.

As an example, let's return to making predictions for the Buick Century with 8221 miles. The below code can be used to calculate the $\hat{y}$ value for this observation:

```
predict(car_lm, newdata = data.frame(Mileage = 8221))
```

Notice that `newdata` expects a data frame with a column of x values, and needs to have the same column name as the predictor in the original data set.

We can also use predict get the confidence and prediction intervals (default is 95%):

```
predict(car_lm, newdata = data.frame(Mileage = 8221), interval = "confidence")
predict(car_lm, newdata = data.frame(Mileage = 8221), interval = "prediction")
```

To change the confidence level, add the `level` argument. For example, we can calculate 89% intervals using the commands:

```
predict(car_lm, newdata = data.frame(Mileage = 8221), interval = "confidence", level = 0.89)
predict(car_lm, newdata = data.frame(Mileage = 8221), interval = "prediction", level = 0.89)
```

We can also ask the predict function to return the standard errors necessary for the by-hand calculations. To do this, we add the argument `se.fit = TRUE` to our `predict()` command.

```
predict(car_lm, newdata = data.frame(Mileage = 8221), se.fit = TRUE)
```

Here, the `se.fit` entry corresponds to $\text{SE}(\widehat{E}(Y|X))$ and `residual.scale` corresponds to $\hat{\sigma} = \sqrt{\text{MSE}}$.

Task 11. Calculate a 90% confidence interval for the average sales price for a car with 40,000 miles. Interpret this interval in context.

Task 12. Calculate a 90% prediction interval for the predicted sales price of **your** car, which has 40,000 miles on it. Interpret this interval in context.

Plotting predictions in R

In Task 6 you added the fitted regression line to the scatterplot of `Price` against `Mileage`. This plot helped to communicate what the regression line is telling us. An even more informative plot would include either the confidence intervals for $\widehat{E}(Y|X)$ or the prediction intervals for $\hat{y}$ (or both).

You can add these intervals to your plot from Task 6 by adding an `interval = <type>` argument to the `gf_lm()` command.

Task 13. Add the confidence intervals for $\widehat{E}(Y|X)$ or the prediction intervals for $\hat{y}$ by filling in the blank with your scatterplot code.

```
___ |>
  gf_lm(interval = "prediction") |>
  gf_lm(interval = "confidence", alpha = 0.6)
```

Note: `alpha = 0.6` in the above code makes the confidence interval for the mean response darker than the prediction interval.

Task 14. When we are making predictions, our model assumptions become even more important than when we are conducting inference for our model coefficients. Create our usual set of diagnostic plots and comment on whether each of the necessary conditions/assumptions for the simple linear regression model are valid.