4/29/2023

Continue with discussion of linear regression:
Residuals
Outliers
Hypothesis Testing
Review for final exam

Linear regression: inference

The population model we are assuming is of the form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

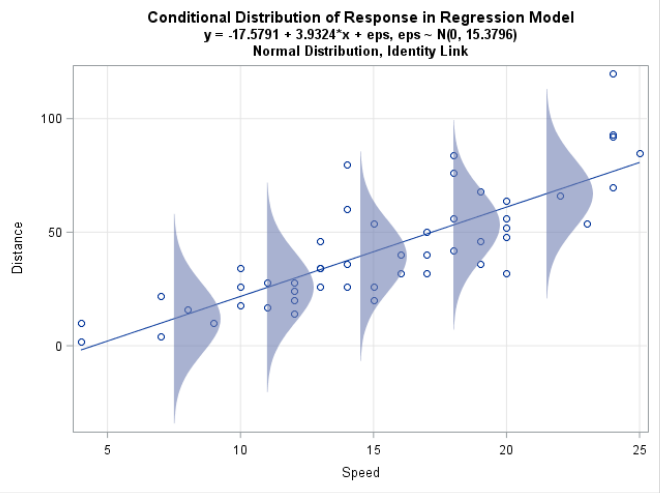
Our sample regression equation:

$$\hat{y} = b_0 + b_1 x$$

The b-coefficients are estimates of the beta coefficients from the population, and as with other samples, there is some wiggle room between the value we measure and the true value. How far off is the true value likely to be? How good is our estimate?

The assumption on the errors is that the mean error is zero, and there is a fixed standard deviation (variance). And the errors are normally distributed.

Suppose you measure a sample from the population at a fixed input value (call it x_0). And measure the y-value repeatedly. The measurements we get with have a center (a mean) on the population mean at that x-value (based on the regression), and a standard deviation that matches the error term.

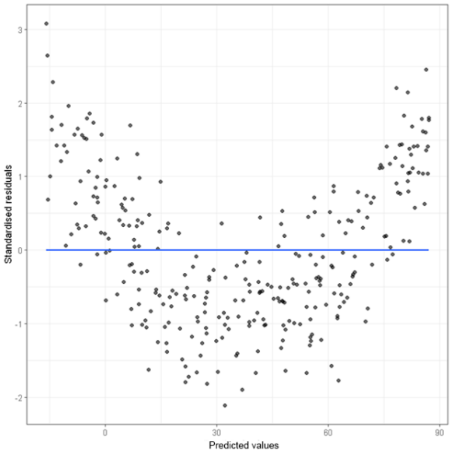


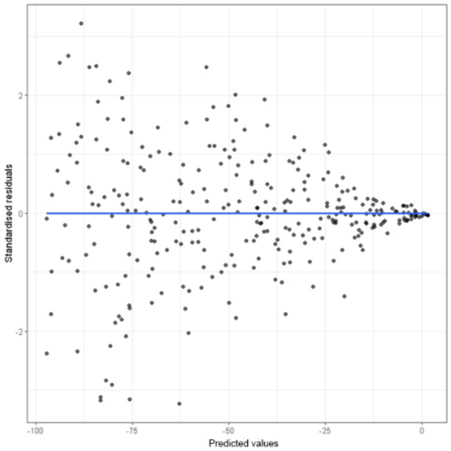
We want to measure these errors: they are called residuals. We calculate them by using our regression to predict the mean y-value at each of our observed x-values, and then subtracting the observations at those points.

These residuals can then be analyzed, numerically and graphically.

We calculate the residuals as $\hat{y} - y$ (the difference between the predictions and the observed values).

What we want to avoid in the residual graph (plotting usually the residuals vs. the x-value, or residuals vs. y-values) non-constant variance (trumpet effect), or other patterns.





The first one is a bad residual graph because the model is not linear.
The second one is bad because the variance in the residuals is not constant.

We can verify that the mean is near zero (rounding errors aside). And the standard deviation of the residuals is called the residual standard error. We can use this to put bounds on predictions (prediction intervals). Predicting from the line is predicting the mean (of a normal distribution), and the residual standard error is the variability measure of that distribution.

In general, the smaller the residual standard error, the better the predictive power of the model.

Outliers.
Outliers in a regression model are measured by their distance from the regression line. We can use the residual standard error we calculated to determine the relative distance to spot outliers.

Influential points:
Are values (that may be different from the rest of the data in some respect) and can have an outsized influence on the values of the regression coefficients.

For outliers, we are basically going to look at the residual standard error, and then determine which (if any) residuals are 2 or 3 standard deviations from the mean.

Extreme outliers are 3 standard deviations. In some cases, you may want to remove these points—when there isn't a lot of data. If you have a lot of data, you should expect some outliers.

Unusual values are considered the most extreme 5% of the data. In our example there are 208 observations. That's around 10.4…. so we would expect that around 10 values would be more than 2 standard deviations from the mean, just because it's random.

If you have very small dataset, outliers can be removed because they have a bigger impact on the results.

For influential points: the test would be to remove the point and recalculate the regression line. If the values don't change significantly, then the point influential. If the coefficients do change a lot, then the point influential (and typically would be left of the final analysis).

Hypothesis testing

Option 1: Test whether the correlation is 0 or not
Option 2: Test whether the slope of the regression line is 0 or not
Option 3: Test whether any variable in the model is alternating the mean at different points (model test)

Option 1: $H_0: \rho = 0, H_a: \rho \neq 0$

Option 2: $H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$

Option 3: $H_0: \beta_i = 0, H_a: \beta_i \neq 0, for\ some\ i$

Can do hypothesis tests on the intercept value, but this will not test the correlation and only determines whether when x is 0, y is also 0 or not.

Done with t-distributions, for simple linear regression, the degrees of freedom are $n - 2$

We can also calculate confidence intervals for the coefficients.

Final Exam is next week!!!!!

Since the last exam: we talked about hypothesis testing

- One sample tests
- Two sample tests
- ANOVA
- Test of independence
- Regression

For these topics, look at the quizzes for the best model of the kinds of questions to expect; or last year's final.

The rest of the material is from the first two exams: look at your last two exams.