Inference for Linear Regression

November 6, 2019

Regression Example

Asking R for a summary of the regression model, we get the following:

```
lm(formula = eruptions ~ waiting)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Let's pick this apart piece by piece.

Regression Example

```
Call:
lm(formula = eruptions ~ waiting)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329
```

- The first line shows the command used in `R` to run this regression model.
- The `Residuals` item shows a quartile-based summary of our residuals.

Regression Example

```
Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

The `F-statistic` and `p-value` give information about the model overall.

- These are based on an F-distribution.
- The null hypothesis is that all of our model parameters are 0 (the model gives us no good info).
- Since p-value$< 2.2 \times 10^{-16} < \alpha = 0.05$, at least one of the parameters is nonzero (the model is useful).

Regression Example

```
Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

- `Multiple R-squared` is our squared correlation coefficient R^2.
- This tells us how good our fit is.
- Ignore the adjusted R-squared and residual standard error for now.

Regression Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16
waiting      0.075628   0.002219   34.09   <2e-16
```

Finally, the `Coefficients` section gives us several pieces of information:

1. `Estimate` shows the estimated parameters for each value.
2. `Std. Error` gives the standard error for each parameter estimate.
3. The `t valuess` are the test statistics for each parameter estiamte.
4. Finally, `Pr(>|t|)` are the p-values for each parameter estimate.

Regression Example

The hypothesis test for each regression coefficient has hypotheses

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

where $i = 0$ for the intercept and $i = 1$ for the slope.

Regression Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16
waiting      0.075628   0.002219   34.09   <2e-16
```

- $p - value < 2 \times 10^{-16}$ for b_0 so we can conclude that the intercept is nonzero.
- $p - value < 2 \times 10^{-16}$ for b_1 so we conclude that the intercept is also nonzero.
- This means that the intercept and slope both provide useful information when predicting values of $y =$ `eruptions`.

Confidence Intervals for a Coefficient

We can construct confidence intervals similar to those for hypothesis tests. A $(1-\alpha)100\%$ confidence interval for β_i is

$$b_i \pm t_{\alpha/2}(df) \times SE(b_i)$$

where the model df and SE can be found in the regression output.

Aside: ANOVA for Regression Models

- ANOVA will also play a role in regression.
- We can get the ANOVA table for a regression.

Aside: ANOVA for Regression Models

The ANOVA table in regression will look something like this:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
faithful$waiting	1	286.478	286.478	1162.1	< 2.2e-16
Residuals	270	66.562	0.247		

Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16
waiting      0.075628   0.002219   34.09   <2e-16
```

Find 95% confidence intervals for β_0 and β_1.

Estimation and Prediction Using a Regression Line

We now know

- how to examine if a model is useful.
- how to confirm that our regression assumptions are satisfied.

Estimation and Prediction Using a Regression Line

Given a useful regression line, we want to

- estimate an average value of y for a given value of x.
- estimate a particular value of y for a given value of x.

Estimation and Prediction Using a Regression Line

We've already talked about using a regression line to make predictions.

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

Plug in x and we get a good estimate for the *average* value of y at that point.

Estimation and Prediction Using a Regression Line

Point estimates are useful, but we want to consider variability!

- Recall: one of our regression assumptions is normally distributed errors.
- This means that the variability around the regression line should be approximately normal
  - with mean $\beta_0 + \beta_1 x$
  - and standard deviation σ.

The Variability of $\hat{y}$

- Notice that $\hat{y}$ is an estimator.
- The variability of an estimator is its standard error.
- Then σ is well-approximated by

$$SE(\hat{y}) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}\right)}$$

The Variability of $\hat{y}$

Since we are working with a normal distribution, estimation and testing can be based on the test statistic

$$t = \frac{\hat{y} - y_0}{SE(\hat{y})}$$

which corresponds to a $t(n-2)$ distribution.

Confidence Intervals for y

A $(1-\alpha)100\%$ confidence interval for the average value of y (measured by $\beta_0 + \beta_1 x$) when $x = x_0$ is

$$\hat{y} \pm t_{\alpha/2}(n-2) \times SE(\hat{y})$$

or

$$\hat{y} \pm t_{\alpha/2}(n-2) \times \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}\right)}$$

Prediction Intervals for y

- So far, we've only considered *average* values of the outcome variable y.
- What if we wanted to predict a *particular* value of y?

Prediction Intervals for y

For a residual,

$$e = \epsilon + \text{error in estimating line}$$

- We don't know the true breakdown between these components.
- ...but we can use this concept to build a new standard error formula.

Prediction Intervals for y

The standard error of $(y - \hat{y})$ is

$$SE(y - \hat{y}) = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}\right)}$$

Prediction Intervals for y

A $(1-\alpha)100\%$ **prediction interval** for a specific value of y when $x = x_0$ is

$$\hat{y} \pm t_{\alpha/2}(n-2) \times SE(y - \hat{y})$$

or

$$\hat{y} \pm t_{\alpha/2}(n-2) \times \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}\right)}$$