Inference for Prediction

Stat 230: Applied Regression Analysis

Warm up

- Work with a neighbor
- Answer the questions associated with the warm up on the worksheet
- Note that the explanatory variable is standardized (mean 0, SD 1)

Prediction

There are two types of predictions in regression

1. Predicting the **mean response** at a specific value of x

   e.g., the average starting salary for some with a B.A. in statistics

2. Predicting the response for a **specific future observation**

   e.g., predicting **your** starting salary (if you have a B.A. in statistics)

🤜 Think of two additional examples of each type of prediction.

Inference for prediction

Best estimate $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_0$

Interval formula: estimate $\pm$ $q \times$ SE

- $\hat{y}$ is our estimate
- Use a t-distribution with $df = n - 2$ to find q

We'll need different SEs depending on if we are building a

- confidence interval for $\widehat{\mu}(Y|X_0)$
- prediction interval for $\hat{y}$ or Pred$(Y|X_0)$

Standard errors

$$\text{SE}(\widehat{\mu}(Y|X)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$\text{SE}(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

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

R: Point estimate

We'll let R do the computational work

`predict` allows you to quickly calculate the value of $\hat{y}$ for a given x (or vector of xs)

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

```
       1
23346.27
```

R: Intervals

The `interval` argument allows you to specify the type of interval you want

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

```
       fit      lwr      upr
1 23346.27 22170.67 24521.86
```

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

```
       fit      lwr      upr
1 23346.27 4094.689 42597.85
```

R: SEs

Adding `se.fit = TRUE` returns the necessary standard errors for "by hand" calculations

```
$fit
       1
23346.27

$se.fit
[1] 598.8985

$df
[1] 802

$residual.scale
[1] 9789.288
```

Activity

- Work with a neighbor
- Work through the inference for prediction example on the worksheet
- The R tutorial is linked on Moodle, also can follow the QR code

Conditions required for inference

Our model must be valid for inference to be valid

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

Conditions to check:

- Linear relationship is appropriate
- Errors are independent and identically distributed (iid)
- Errors are normally distributed
- Variance of the errors doesn't depend on x

Regression conditions

What happens if our assumptions aren't valid?

- **Linearity:** if nonlinear, everything breaks!
- **Independence:** estimates are still unbiased (i.e. we fit the right line) but measures of the accuracy of those estimates (the SEs) are typically too small
- **Normality:** estimates are still unbiased (i.e. we fit the right line), SEs are correct BUT confidence/prediction intervals are wrong (we can't use t-distribution)
- **Constant error variance:** estimates are still unbiased but standard errors are wrong (and we don't know how wrong)

RMarkdown demo