Inference in regression

Brian Caffo, Jeff Leek and Roger Peng

May 19, 2016

Recall our model and fitted values

- Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- $\epsilon \sim N(0, \sigma^2)$.
- We assume that the true model is known.
- We assume that you've seen confidence intervals and hypothesis tests before.
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$.

Review

- Statistics like $\frac{\hat\theta - \theta}{\hat\sigma_{\hat\theta}}$ often have the following properties.
  1. Is normally distributed and has a finite sample Student's T distribution if the variance is replaced with a sample estimate (under normality assumptions).
  2. Can be used to test $H_0 : \theta = \theta_0$ versus $H_a : \theta >, <, \neq \theta_0$.
  3. Can be used to create a confidence interval for θ via $\hat\theta \pm Q_{1-\alpha/2}\hat\sigma_{\hat\theta}$ where $Q_{1-\alpha/2}$ is the relevant quantile from either a normal or T distribution.
- In the case of regression with iid sampling assumptions and normal errors, our inferences will follow very similarily to what you saw in your inference class.
- We won't cover asymptotics for regression analysis, but suffice it to say that under assumptions on the ways in which the X values are collected, the iid sampling model, and mean model, the normal results hold to create intervals and confidence intervals

Results

- $\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (X_i - \bar{X})^2$
- $\sigma^2_{\hat{\beta}_0} = Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2$
- In practice, σ is replaced by its estimate.
- It's probably not surprising that under iid Gaussian errors

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

  follows a t distribution with $n - 2$ degrees of freedom and a normal distribution for large n.
- This can be used to create confidence intervals and perform hypothesis tests.

Example diamond data set

```r
library(UsingR); data(diamond)
```

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'
```

Example continued

```
coefTable
```

```
##                Estimate Std. Error   t value      P(>|t|)
## (Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
## x           3721.0249   81.78588  45.49715 6.751260e-40
```

```
fit <- lm(y ~ x);
summary(fit)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
## x           3721.0249   81.78588  45.49715 6.751260e-40
```

Getting a confidence interval

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[1
```

```
## [1] -294.4870 -224.7649
```

```
(sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef
```

```
## [1] 355.6398 388.5651
```

With 95% confidence, we estimate that a 0.1 carat increase in diamond size results in a 355.6 to 388.6 increase in price in (Singapore) dollars.

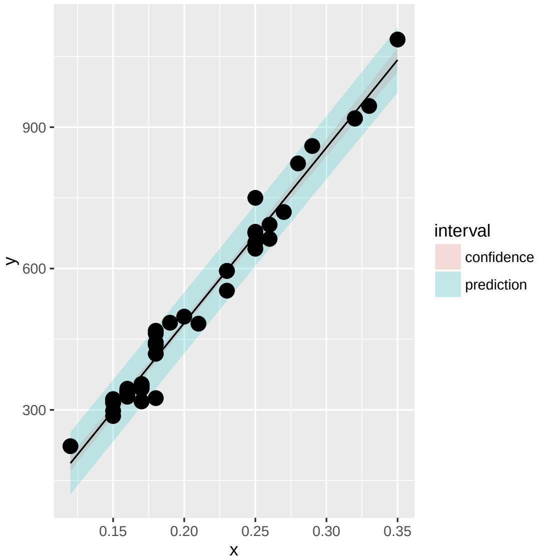
Prediction of outcomes

- Consider predicting Y at a value of X
- Predicting the price of a diamond given the carat
- Predicting the height of a child given the height of the parents
- The obvious estimate for prediction at point x_0 is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

- A standard error is needed to create a prediction interval.
- There's a distinction between intervals for the regression line at point x_0 and the prediction of what a y would be at point x_0.
- Line at x_0 se, $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$
- Prediction interval se at x_0, $\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$

Plotting the prediction intervals



Discussion

- Both intervals have varying widths.
- Least width at the mean of the Xs.
- We are quite confident in the regression line, so that interval is very narrow.
- If we knew β_0 and β_1 this interval would have zero width.
- The prediction interval must incorporate the variabilibity in the data around the line.
- Even if we knew β_0 and β_1 this interval would still have width.

In R

```
newdata <- data.frame(x = xVals)
p1 <- predict(fit, newdata, interval = ("confidence"))
p2 <- predict(fit, newdata, interval = ("prediction"))
plot(x, y, frame=FALSE,xlab="Carat",ylab="Dollars",pch=21,c
abline(fit, lwd = 2)
lines(xVals, p1[,2]); lines(xVals, p1[,3])
lines(xVals, p2[,2]); lines(xVals, p2[,3])
```