Agenda

1. Bootstrap for Regression

Regression Diagnostics

Warmup

1. (EOCE 5.17) The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year. The scatterplot below shows the relationship between these two variables along with the least squares fit.

(a) Describe the relationship between number of tourists and spending.

(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

Bootstrap for Regression  Recall that a slope coefficient is an average or expected change in the response variable as a function of a unit change in that explanatory variable, holding the other explanatory variables constant. Like the sample mean, the estimated coefficient ($b_1$) is a deterministic
calculation based on a single sample of data, but it too has a sampling distribution. Thus, we can
use the bootstrap percentile method to construct a confidence interval for it. The default confidence
interval is constructed using the \( t \)-distribution.

```r
require(mosaicData)
mod <- lm(wt ~ age, data=Gestation)
coef(mod)
#> (Intercept) age
#> 116.6834606 0.1062233

confint(mod)
#> 2.5 % 97.5 %
#> (Intercept) 111.77014517 121.596776
#> age -0.07012632 0.282573
```

The bootstrap percentile method should give us a similar interval:

```r
bstrap <- do(1000) * coef(lm(wt ~ age, data = resample(Gestation)))
qdata(~age, p = c(0.025, 0.975), data = bstrap)
#> quantile p
#> 2.5% -0.07163802 0.025
#> 97.5% 0.28668962 0.975
# qplot(x = age, data = bstrap, geom = "density")
```

Inference for Multiple Linear Regression

Inference for MLR is in many ways just a direct
extension of inference for SLR. Recall the Italian restaurants from NYC.

```r
NYC <- read.csv("http://www.science.smith.edu/~bbaumer/mth241/nyc.csv")
fm <- lm(Price ~ Food + Service + Decor + East, data = NYC)
maummary(fm)
#> Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -24.023800 4.708359 -5.102 9.24e-07 ***
#> Food 1.538120 0.368951 4.169 4.96e-05 ***
#> Service -0.002727 0.396232 -0.007 0.9945
#> Decor 1.910087 0.217005 8.802 1.87e-15 ***
#> East 2.068050 0.946739 2.184 0.0304 *
#>
#> Residual standard error: 5.738 on 163 degrees of freedom
#> Multiple R-squared: 0.6279, Adjusted R-squared: 0.6187
#> F-statistic: 68.76 on 4 and 163 DF, p-value: < 2.2e-16
```

1. Interpret the results of the individual \( t \)-tests for the significance of each coefficient. Which
coefficients that are statistically significantly different from 0?

2. Drop the variable with the highest \( p \)-value from the regression model, and refit the model with
the remaining two variables. What happens to the \( R^2 \)? What about the adjusted \( R^2 \)?

3. What happens to the values of the coefficients? Does their statistical significance change?

4. Check the conditions for inference. Are they met?