Agenda

1. Conditions for Regression
2. Bootstrap for Regression

Example: Gestation The `Gestation` data set contains birth weight, date, and gestational period collected as part of the Child Health and Development Studies in 1961 and 1962. Information about the baby's parents—age, education, height, weight, and whether the mother smoked is also recorded.

1. Fit a linear regression model for birthweight (wt) as a function of the mother's age (age).

2. Use the `summary` command to find a 95% confidence interval and p-value for the slope coefficient.

3. Now run the command `confint(mod)` to find a 95% confidence interval for the slope coefficient. Does it agree with what you calculated above?

4. What do you conclude about the association between a mother's age and her baby's birthweight?

Conditions for Regression The inferences we made above were predicted upon our assumption that the slope coefficient followed a t-distribution. Recall also that when we fit the regression model

$$Y = b_0 + b_1 \cdot X + e$$

we assumed that $e \sim N(0, \sigma)$, for some constant σ. Our inferences will only be valid if the following assumptions are reasonable:

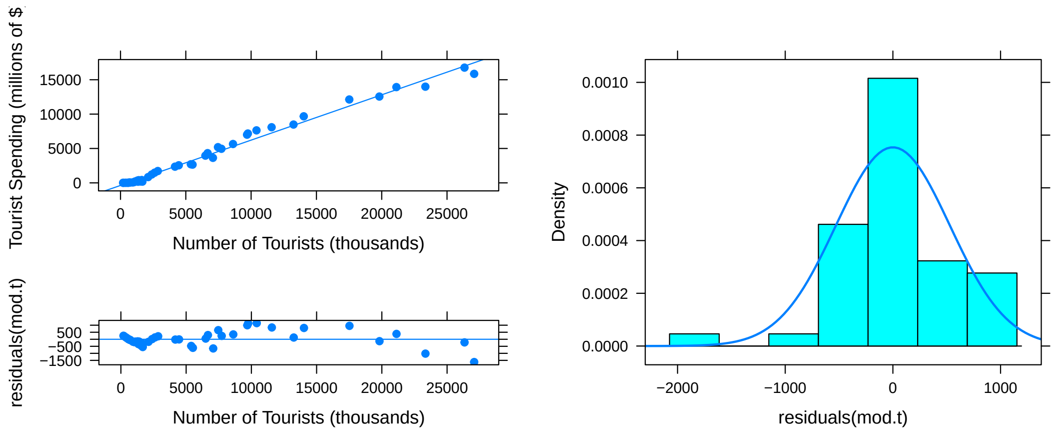
- **L**inearity:
- **I**ndependence:
- **N**ormality of Residuals:
- **E**qual Variance of Residuals:

These conditions are usually verified using diagnostic plots.

```
plot(mod)
```

Practice Problems

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.

2. Verify the conditions for the Gestation model above.

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.

```
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:

```
bstrap <- do(1000) * coef(lm(wt ~ age, data = resample(Gestation)))
qdata(~age, p = c(0.025, 0.975), data = bstrap)

##          quantile     p
## 2.5%  -0.07563621 0.025
## 97.5%  0.29453164 0.975
```

1. Get the confidence interval for the intercept using this bootstrap percentile method.

2. Use the bootstrap method to evaluate the null hypothesis that $\beta_1 = 0$. Do this by first resampling, then recalculating the estimate for each bootstrap sample (but subtracting off the original slope estimate), and then, finally, using the `pdata()` command to find the bootstrap p-value.

Inference for Multiple Linear Regression Inference for MLR is in many ways just a direct extension of inference for SLR. Recall the poverty and graduation rates example. Now, we might also be interested in the relationship between the racial make-up of the state and poverty rates, as well as the relationship between the percentage of the state population living in metropolitan areas and poverty rates. With multiple regression, we can build a model of poverty rates as a function of all three things, graduate percentages, percentage of residents who are White, and percentage of residents who live in metropolitan areas.

```
poverty <- read.csv("http://math.smith.edu/~bbaumer/mth241/poverty.txt", sep = "\t")
mod <- lm(Poverty ~ Graduates + White + Metropolitan.Residence, data = poverty)
summary(mod)
confint(mod)
```

We can again check the **LINE** assumptions for this model using the plot command.

```
plot(mod)
```

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. Write a sentence providing an interpretation of the coefficient for *Graduates* in the context of this multiple regression model.

3. Write a sentence providing an interpretation of the coefficient for $White$.

4. Write a sentence providing an interpretation of the coefficient for $Metropolitan.Residence$.

5. What is the R^2 value for this model? Write a sentence interpreting this value in the context of the problem.

6. 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?

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

8. Are the regression assumptions for this model satisfied?