

Agenda

1. Inference for Regression
2. Conditions for Regression

Warmup Consider the following data about US states. Previously, we fit a simple linear regression line for the poverty rate in each state as a function of the high school graduation rate.

```
library(mosaic)
poverty <- read.csv("http://math.smith.edu/~bbaumer/mth241/poverty.txt", sep = "\t")
mod <- lm(Poverty ~ Graduates, data = poverty)
coef(mod)

## (Intercept)  Graduates
## 64.7809658  -0.6212167
```

1. Write a sentence providing an interpretation of the coefficient for *Graduates* in the context of the problem.

Inference for Regression We can use our understanding of the t -distribution to make *inferences* about the true (unknown) value of regression coefficients. In particular, we can test the hypothesis that $\beta_1 = 0$ and find a confidence interval for β_1 .

```
summary(mod)
```

1. Find a 95% confidence interval and p -value for the slope coefficient.
2. What do you conclude about the association between poverty rates and high school graduation rates among US states?

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 birth-weight?

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

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon,$$

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

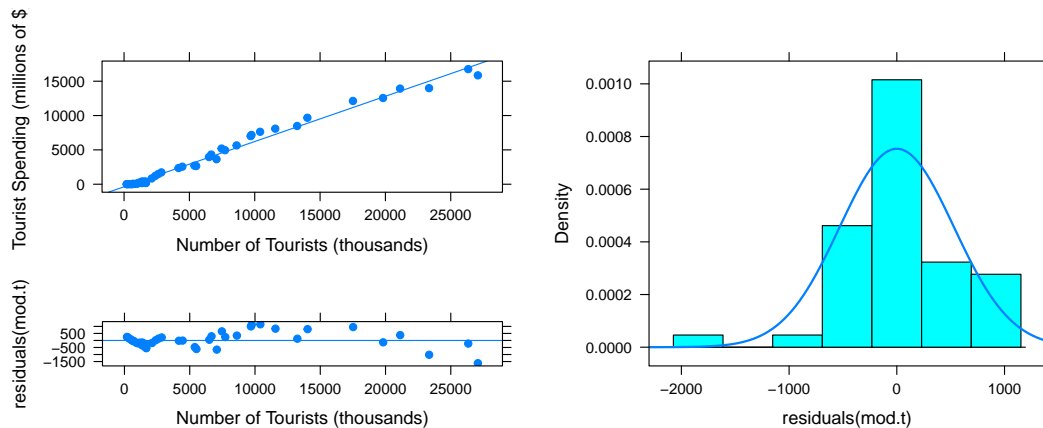
- **Linearity:**
- **Independence:**
- **Normality of Residuals:**
- **Equal Variance of Residuals:**

These conditions are usually verified using diagnostic plots.

```
plot(mod)
```

Practice Problems

1. Verify the conditions for the US states model above.
2. Verify the conditions for the Gestation model above.
3. (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.

Multiple Regression 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.

```
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. Write a sentence providing an interpretation of the coefficient for *Graduates* in the context of this multiple regression model.
2. Write a sentence providing an interpretation of the coefficient for *White*.
3. Write a sentence providing an interpretation of the coefficient for *Metropolitan.Residence*.
4. What is the R^2 value for this model? Write a sentence interpreting this value in the context of the problem.
5. Are the regression assumptions for this model satisfied?