Agenda

1. Inference for Regression
2. Conditions for Regression

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

```
require(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)

##
## Call:
## lm(formula = Poverty ~ Graduates, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1624 -1.2593 -0.2184  0.9611  5.4437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
## Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 49 degrees of freedom
## Multiple R-squared:  0.5578,Adjusted R-squared:  0.5488
## F-statistic: 61.81 on 1 and 49 DF,  p-value: 3.109e-10
```

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. 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

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

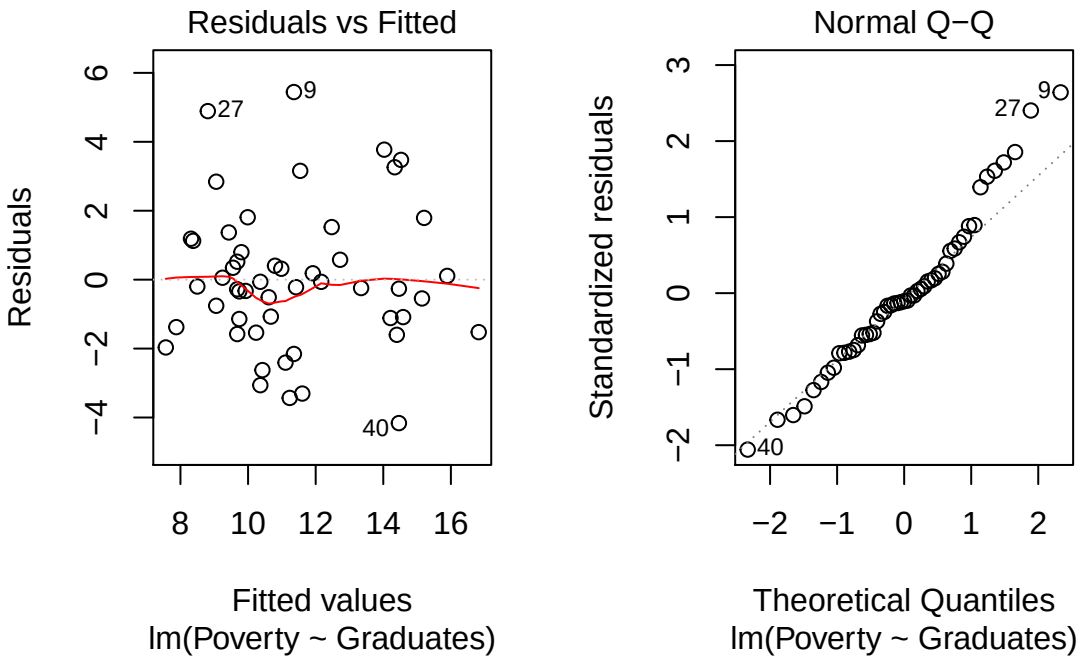
we assumed that $\epsilon \sim N(0, \sigma)$, for some constant σ. Our inferences will only be valid if the following assumptions are reasonable:

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

These conditions are usually verified using diagnostic plots.

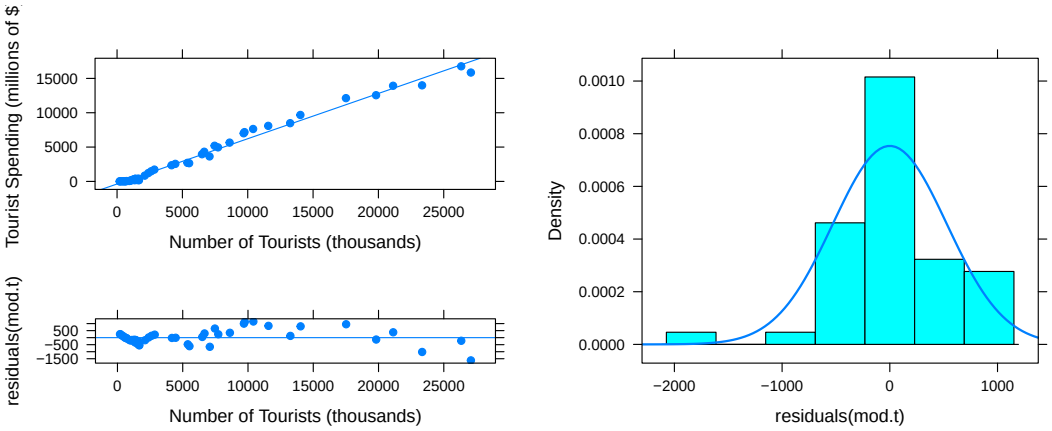
```
plot(mod, which=c(1,2))
```

```
## NULL
## NULL
```



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.