Agenda

1. Bootstrap for Regression
2. Inference for Multiple Regression
3. More on Regression Assumptions

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.

```
library(mosaic)
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%  -0.08375802 0.025
## 97.5%  0.29302681 0.975
```

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?

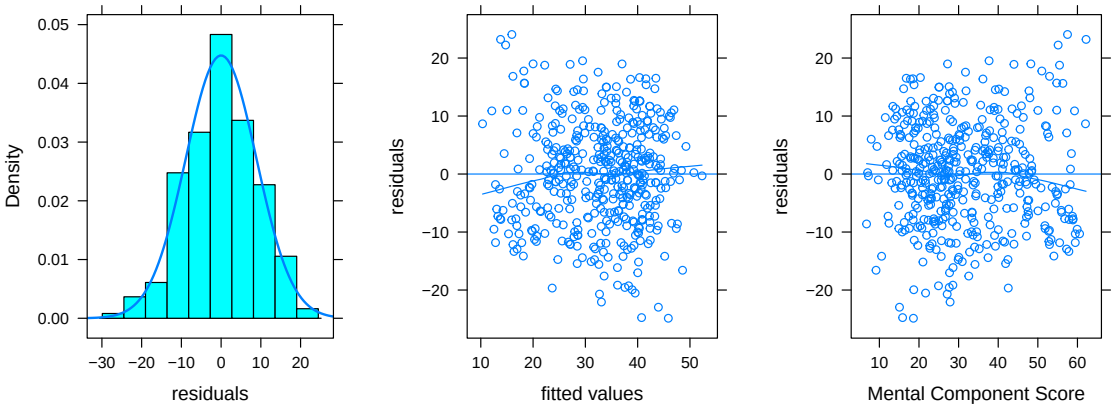
Case study: Predictors of depressive symptoms In the HELP (Health Evaluation and Linkage to Primary Care) study, investigators were interested in determining predictors of severe depressive symptoms (measured by the Center for Epidemiologic Studies - Depression scale, aka `cesd`) amongst a cohort enrolled at a substance abuse treatment facility. These includes `substance` of abuse (alcohol, cocaine, or heroin), `mcs` (a measure of mental well-being), gender and housing status (housed or homeless). Consider the following multiple regression model.

```
library(mosaic)
fm <- lm(cesd ~ substance + mcs + sex + homeless, data = HELPrct)
msummary(fm)

##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.77942    1.46640  39.402  < 2e-16 ***
## substancecocaine -3.54056    1.01013  -3.505 0.000503 ***
## substanceheroin  -1.68181    1.07310  -1.567 0.117766
## mcs              -0.64073    0.03377 -18.971  < 2e-16 ***
## sexmale          -3.32387    1.00749  -3.299 0.001047 **
## homelesshoused   -0.83270    0.86864  -0.959 0.338265
##
## Residual standard error: 8.973 on 447 degrees of freedom
## Multiple R-squared:  0.4915,Adjusted R-squared:  0.4859
## F-statistic: 86.43 on 5 and 447 DF,  p-value: < 2.2e-16

confint(fm)

##                      2.5 %     97.5 %
## (Intercept)      54.8975311 60.6613125
## substancecocaine -5.5257585 -1.5553529
## substanceheroin  -3.7907660  0.4271435
## mcs              -0.7071036 -0.5743498
## sexmale          -5.3038731 -1.3438759
## homelesshoused   -2.5398149  0.8744190
```



1. Write out the linear model

2. Calculate the predicted CESD for a female homeless cocaine-involved subject with an MCS score of 20.

3. Interpret the 95% confidence interval for the `substancecocaine` coefficient

4. Make a conclusion and summarize the results of a test of the `homeless` parameter

5. Report and interpret the R^2 (coefficient of determination) for this model

6. What do we conclude about the distribution of the residuals?

7. What do we conclude about the relationship between the fitted values and the residuals?

8. What do we conclude about the relationship between the MCS score and the residuals?

9. What other things can we learn from the residual diagnostics?

10. Which observations should we flag for further study?

```
help_mod <- broom::augment(fm)
help_mod %>%
  slice(c(40, 351, 433, 450))
```