Observation = Model + Error \quad Observation = Fitted + Residual

1. More than one predictor. \( k \) explanatory variables, \( p = k + 1 \) coefficients.

\[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \]

2. The Chicago redlining example.

3. Sums of squares.

\[
\begin{align*}
SS_{\text{Residual}} &= \sum (y - \hat{y})^2 \\
SS_{\text{Regression}} &= \sum (\hat{y} - \bar{y})^2 \\
SS_{\text{Total}} &= \sum (y - \bar{y})^2 \\
R^2 &= \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Residual}}}{SS_{\text{Total}}}
\end{align*}
\]

4. Mean squares. A mean square is a sum-of-squares divided by degrees of freedom.

\[
\begin{align*}
MS_{\text{Total}} &= \frac{SS_{\text{Total}}}{n - 1} \\
MS_{\text{Residual}} &= \frac{SS_{\text{Residual}}}{n - p} \\
MS_{\text{Regression}} &= \frac{SS_{\text{Regression}}}{p - 1}
\end{align*}
\]

5. Adjusted \( R^2 \). \( R^2 \) is artificially high if we use a large number of even useless predictors. Adjust downwards, counting predictors.

\[
\text{adj-} R^2 = 1 - \frac{MS_{\text{Residual}}}{MS_{\text{Total}}}
\]

6. \( F \)-tests. An \( F \)-test is a test of model utility. \( F \) tests compare explained variation to unexplained variation.

\[
F = \frac{\text{variation explained by model}}{\text{variation not explained explained by model}}
\]

In our case, \( A \) is the model with regression coefficients and \( B \) is the model without (the constant model).

7. \( F \)-test first and then \( t \)-test on individual coefficients.

(a) Null hypothesis: \( H_0 : \beta_1 = \cdots = \beta_k = 0 \).

(b) Statistic: \( F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} \).

(c) Reject \( H_0 \) if \( F \) is large.

8. Checking conditions: plots.

(a) Plot residuals against fitted values:

\[
\text{plot(residuals(1)~fitted(1))}
\]

(b) Plot a histogram of residuals:

\[
\text{hist(residuals(1))}
\]

(c) A normal probability plot of residuals:

\[
\text{qqnorm(residuals(1))}
\]

**Homework**

1. Read Chapter 30, pages 792–801.

2. For Tuesday, May 4, do the following problem. The course website has a dataset called gala that has data on several Galapagos Islands. The variable **species** is the number of species on the island. We would like to write a model that predicts the number of species from several geographic variables: Area, Elevation, Nearest (distance to nearest island, km), Scruz (distance to Santa Cruz Island), and Adjacent (size of closest island in km\(^2\)). Write such a model, evaluate the model, and check the assumptions of the model. You can get the data by using get243() or at the course webpage.
```r
> l = lm(involact ~ race, data = chicago)
> summary(l)

Coefficients:
        Estimate Std. Error t value Pr(>|t|) 
(Intercept)  0.129218   0.096611  1.338  0.188 
race         0.013882   0.002031  6.836 1.78e-08 *** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.4488 on 45 degrees of freedom
Multiple R-squared: 0.5094, Adjusted R-squared: 0.4985
F-statistic: 46.73 on 1 and 45 DF,  p-value: 1.784e-08

> anova(l)

Analysis of Variance Table

Response: involact
 Df Sum Sq Mean Sq F value    Pr(>F) 
race   1 9.4143 9.4143 46.733 1.784e-08 *** 
Residuals 45 9.0653 0.2015

> ltwo = lm(involact ~ race + fire, data = chicago)
> summary(ltwo)

Coefficients:
        Estimate Std. Error t value Pr(>|t|) 
(Intercept) -0.057774   0.098116 -0.589  0.558987
race         0.008906   0.002214  4.023  0.000223 ***
fire         0.029407   0.007756  3.792  0.000452 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.3941 on 44 degrees of freedom
Multiple R-squared: 0.6303,    Adjusted R-squared: 0.6135
F-statistic: 37.5 on 2 and 44 DF,  p-value: 3.117e-10

> anova(ltwo)

Analysis of Variance Table

Response: involact
 Df Sum Sq Mean Sq F value    Pr(>F) 
race   1 9.4143 9.4143 60.625  8.195e-10 ***
fire   1 2.2326 2.2326 14.377  0.000452 ***
Residuals 44 6.8326 0.1553

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```