Simple linear regression:

Idea: The probability distribution of a random variable $Y$ may depend on the value $x$ of some predictor variable.

Ingredients:

- $Y$ is a continuous response variable (dependent variable).
- $x$ is an explanatory or predictor variable (independent variable).
- $Y$ is the variable we’re mainly interested in understanding, and we want to make use of $x$ for explanatory reasons.

Examples:

- **Predictor**: First year mileage for a certain car model; **Response**: Resulting maintenance cost.
- **Predictor**: Height of fathers; **Response**: Height of sons.

Motivating example: Allergy medication

Ten patients given varying doses of allergy medication and asked to report back when the medication seems to wear off.

Data:

<table>
<thead>
<tr>
<th>Dosage of Medication (mg)</th>
<th>Hours of Relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9.1</td>
</tr>
<tr>
<td>3</td>
<td>5.5</td>
</tr>
<tr>
<td>4</td>
<td>12.3</td>
</tr>
<tr>
<td>5</td>
<td>9.2</td>
</tr>
<tr>
<td>6</td>
<td>14.2</td>
</tr>
<tr>
<td>6</td>
<td>16.8</td>
</tr>
<tr>
<td>7</td>
<td>22.0</td>
</tr>
<tr>
<td>8</td>
<td>18.3</td>
</tr>
<tr>
<td>8</td>
<td>24.5</td>
</tr>
<tr>
<td>9</td>
<td>22.7</td>
</tr>
</tbody>
</table>
Assumptions for simple linear regression:

- \( Y \sim N(\mu_y, \sigma) \)
- \( \mu_y = \beta_0 + \beta_1 x \) (linearity assumption)
- Values of \( x \) are treated as given information.

“Simple” refers to a single predictor variable.

The line \( y = \beta_0 + \beta_1 x \) is the “population regression line.”

Interpretation of \( \beta_0 \) and \( \beta_1 \):

- The value of \( \beta_0 \), the \( y \)-intercept, is the mean of \( Y \) for a population with \( x = 0 \).
- The value of \( \beta_1 \), the slope, is the increase in the mean of \( Y \) for a unit increase in \( x \).

We’ll see these in the allergy example.

Our goals for simple linear regression:
• Make inferences about $\beta_0, \beta_1$ (and $\sigma$). Most importantly, we might want to find out if $\beta_1 = 0$. 

  **Reason:** $\beta_1 = 0$ is equivalent to the $x$-values providing no information about the mean of $Y$.

• Make inferences about $\mu_y$ for a specific $x$ value.

• Predict potential values of $Y$ given a new $x$ value.

**Comparison to correlation:**

Regression and correlation are related ideas.

• Correlation is a statistic that measures the linear association between two variables. It treats $y$ and $x$ symmetrically.

• Regression is set up to predict $y$ from $x$. The relationship is not symmetric.

**Inference for the population regression line:**

Observe pairs of data $(x_i, y_i)$ for $i = 1, \ldots, n$ are obtained independently. Want to make inferences about $\beta_0$ and $\beta_1$ from the sample of data.

One common criterion for estimating $\beta_0$ and $\beta_1$ (there are many others!):

Select values of $\beta_0$ and $\beta_1$ that minimize

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$ 

Let $b_0$ and $b_1$ be the estimates of $\beta_0$ and $\beta_1$ that achieve the minimum. These are called the least-squares estimates of the regression coefficients.

**Formulas for estimates $b_0$ and $b_1$:**

We need the sample means $\bar{x}$ and $\bar{y}$, the sample standard deviations $s_x$ and $s_y$, and the correlation coefficient $r$. Then

$$b_1 = \frac{rs_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$ 

These formulas can be derived using differential calculus.

**Estimate of $\sigma$:**

$$s = s_y \sqrt{\frac{n - 1}{n - 2(1 - r^2)}}$$

Notice that the quantity in the square root is usually less than 1 (typically when $r$ is not close to 1 or $-1$).
Standard errors of $b_0$ and $b_1$:

\[
s.e.(b_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}
\]

\[
s.e.(b_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1)s_x^2}}
\]

100C% CI for $\beta_0$ and $\beta_1$:

\[
b_0 \pm t^*(s.e.(b_0))
\]

\[
b_1 \pm t^*(s.e.(b_1))
\]

where $t^*$ is the critical value from a $t(n-2)$-distribution.

**Hypothesis testing for $\beta_1$:**

In most cases, $H_0$: $\beta_1 = 0$. This is equivalent to hypothesizing that the predictor variable has no explanatory effect on the response.

The alternative hypothesis $H_a$ depends on the context of the problem.

**Test statistic:**

\[
t = \frac{b_1 - \beta_1}{s.e.(b_1)} = \frac{b_1}{s.e.(b_1)}
\]

Under $H_0$, the test statistic has a $t(n-2)$-distribution.

**Example:** Allergy medication

Suppose the R data frame `allergy` contains the variables `dosage` and `relief`.

```r
> mean(allergy$dosage)
[1] 5.9
> mean(allergy$relief)
[1] 15.46
> sd(allergy$dosage)
[1] 2.131770
> sd(allergy$relief)
[1] 6.473742
> cor(allergy$dosage,allergy$relief)
[1] 0.9142966
```

**Least-squares estimates:**

\[
b_1 = \frac{r_s y}{s_x} = \frac{0.914(6.47)}{2.13} = 2.776
\]

\[
b_0 = \bar{y} - b_1 \bar{x} = 15.46 - 2.776(5.9) = -0.92
\]

\[
s = s_y \sqrt{\frac{n-1}{n-2}(1 - r^2)} = 6.47 \sqrt{\frac{10-1}{10-2}(1 - 0.914^2)} = 2.78
\]
Standard errors:

\[
\text{s.e.}(b_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = 2.78 \sqrt{\frac{1}{10} + \frac{5.9^2}{(10-1)2.13^2}} = 2.71
\]

\[
\text{s.e.}(b_1) = \frac{s}{\sqrt{(n-1)s_x^2}} = \frac{2.78}{(10-1)2.13^2} = 0.435
\]

Example: Allergy medication in R

```r
> allergy.lm = lm(relief ~ dosage, data=allergy)
> summary(allergy.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9215  2.7124  -0.340   0.7428
dosage      2.7765  0.4349   6.385  0.000213 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 2.781 on 8 degrees of freedom
Multiple R-Squared: 0.8359,  Adjusted R-squared: 0.8154
F-statistic: 40.76 on 1 and 8 DF,  p-value: 0.0002126
```
Example: Computation of a 95% CI for the intercept, $\beta_0$

We have $n = 10$ values, and the critical value from a $t(8)$ distribution is $t^* = 2.306$. We have

$$-0.92 \pm 2.306(2.71) = (-7.17, 5.33)$$

We are 95% confident that $\beta_0$ is between $-7.17$ and $5.33$.

**Example: Hypothesis test:** Does dosage *increase* period of relief (at the 0.05 level)?

We have

$H_0$: $\beta_1 = 0$

$H_a$: $\beta_1 > 0$

*Test statistic:*

$$t = \frac{2.777 - 0}{0.435} = 6.38$$

(One-sided) p-value is less than 0.0005 (from Table D, examining the $t(8)$ distribution). Can reject $H_0$ at the 0.05 level.
Also, from the R output, can read that the 2-sided $p$-value is approximately 0.0002. Divide by 2 to get one-sided $p$-value.

**Inference for $\mu_y$ for a specific $x = x^*$:**

Estimate of $\mu_y$ for $x = x^*$:

$$\hat{\mu}_y = b_0 + b_1 x^*.$$  

Standard error of $\hat{\mu}_y$:

$$\text{s.e.}(\hat{\mu}_y) = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$= s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s^2_x}}$$

**100C% CI for $\mu_y$ (for specific $x^*$):**

$$\hat{\mu}_y \pm t^* (\text{s.e.}(\hat{\mu}_y))$$

where $t^*$ is the critical value from a $t(n - 2)$-distribution.

**Hypothesis testing:**

For testing $H_0: \mu_y = \mu_0$, use the test statistic,

$$t = \frac{\hat{\mu}_y - \mu_0}{\text{s.e.}(\hat{\mu}_y)}.$$  

Under $H_0$, the test statistic has a $t(n - 2)$-distribution.

**Prediction interval for future response:**

Point prediction: \(\hat{y} = b_0 + b_1 x^*\)

Standard error of the prediction:

$$\text{s.e.}(\hat{y}) = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s^2_x}}$$

**100C% Prediction Interval (PI) for future response:**

$$\hat{y} \pm t^* (\text{s.e.}(\hat{y}))$$

where $t^*$ is the critical value from a $t(n - 2)$-distribution.

This is not a confidence interval because it’s not a statement about a parameter of a probability distribution.

**Example: 95% confidence interval of $\mu_y$ for $x = 5.5$**

Find a 95% confidence interval for the mean time of relief for people taking 5.5 mg of medicine.
Can calculate $\bar{x} = 5.9$ and $s_x = 2.132$.

We have that $(n-1)s_x^2 = 9(2.132^2) = 40.9$.

Also, for 95% confidence and df=8, $t^* = 2.306$ from Table D.

\[
s.e.(\hat{\mu}_y) = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}
\]
\[
= 2.781 \sqrt{1 + \frac{(5.5 - 5.9)^2}{40.9}} = 0.896
\]

\[
\hat{\mu}_y \pm t^* s.e.(\hat{\mu}_y) = (-0.9215 + 2.7765(5.5)) \pm 2.306(0.896)
\]
\[
= 14.349 \pm 2.066
\]
\[
= (12.283, 16.415)
\]

We are 95% confident that the mean number of hours of relief for people taking 5.5 mg of the medication is between 12.283 and 16.415 hours.

What about a 95% prediction interval for $y$ with $x = 5.5$?

\[
s.e.(\hat{y}) = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}
\]
\[
= 2.781 \sqrt{1 + \frac{1}{10} + \frac{(5.5 - 5.9)^2}{40.9}} = 2.922
\]

\[
\hat{y} \pm t^* s.e.(\hat{y}) = (-0.9215 + 2.7765(5.5)) \pm 2.306(2.922)
\]
\[
= 14.349 \pm 6.738
\]
\[
= (7.611, 21.087)
\]

We are 95% confident that a new subject given 5.5 mg of the medication will have relief between 7.611 and 21.087 hours.

“Goodness of fit” for regression:

Tightness of points to a regression line can be described by $R^2$, where

\[
R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\]

where $y_i$ is an observed value, $\bar{y}$ is the average of the $y_i$, and $\hat{y}_i$ is the predicted value (on the line).
The numerator of the second term will be small when the predicted responses are close to the observed responses.

For simple linear regression, $R^2$ is the same as the correlation coefficient, $r$, squared.

**Interpretation:**

Proportion of variance in the response explained by the predictor values.

- $R^2 = 100\% \Rightarrow$ All the data lie on the regression line.
- $R^2 > 70\% \Rightarrow$ The response is very well-explained by the predictor variable.
- $30\% < R^2 < 70\% \Rightarrow$ The response is moderately well-explained by the predictor variable.
- $R^2 < 30\% \Rightarrow$ The response is weakly explained by the predictor variable.
- $R^2 = 0\% \Rightarrow$ The $x_i$ have no explanatory effect.

**Example:** Value of $R^2$ from regression of Hours on Dose: $83.6\%$

This is a very strong fit.

**Comments on linearity:**

- Can only assume (and verify) linearity holds for the range of the data.
- For inference about $\mu_y$ or future $Y$, do not use a value of $x^*$ that is beyond the range of the data (do not extrapolate beyond the data).

**Robustness:**

- Inference for least-squares regression is very sensitive to underlying assumptions (normality, common standard deviation). There are other ways to estimate the regression coefficients (robust regression methods) if you are concerned about outliers, etc.

**Multiple linear regression:**

**Idea:**

- Instead of just using one predictor variable as in simple linear regression, use several predictor variables.
- The probability distribution of $Y$ depends on several predictor variables $x_1, x_2, \ldots, x_p$ simultaneously.
- Rather than perform $p$ simple linear regressions on each predictor variable separately, use information from all $p$ predictor variables simultaneously.
Motivating example:

Predicting graduation rates at random sample of 100 universities, with predictor variables

- Percentage of new students in top 25% of their HS class,
- Instructional expense ($) per student,
- Percentage of faculty with PhDs,
- Percentage of Alumni who donate,
- Public/private institution (Public = 1, Private = 0).


Some of the data (100 schools):

70 8328 80 20 0 80
22 6318 89 8 1 79
38 10139 66 23 0 47
45 7344 64 24 0 69
58 9060 92 17 0 72
44 7392 75 17 1 53
53 8944 85 20 1 73
82 13388 90 47 0 77
38 6391 77 13 1 38
44 9114 78 30 0 65
55 5955 70 2 1 64
79 5527 72 4 1 50
...

Assumptions for multiple regression:

\[ Y \sim N(\mu_y, \sigma) , \]

\[ \mu_y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \]

- \( Y \) is the response,
- For \( j = 1, \ldots, p \), \( x_j \) is the value of the \( j \)-th predictor variable, and
- \( \beta_j \) is the (unknown) regression coefficient to the \( j \)-th predictor variable.

\( \beta_j \) is the increase in \( \mu_y \) for a unit change in \( x_j \), holding the other predictors at their current values.

Data summary in R:
> summary(schools)

<table>
<thead>
<tr>
<th>Gradrate</th>
<th>PctTop25</th>
<th>Expense</th>
<th>PctPhds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>30.00</td>
<td>Min.</td>
<td>16.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>52.75</td>
<td>1st Qu.</td>
<td>34.75</td>
</tr>
<tr>
<td>Median</td>
<td>67.00</td>
<td>Median</td>
<td>51.50</td>
</tr>
<tr>
<td>Mean</td>
<td>65.36</td>
<td>Mean</td>
<td>53.73</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>78.25</td>
<td>3rd Qu.</td>
<td>69.25</td>
</tr>
<tr>
<td>Max.</td>
<td>100.00</td>
<td>Max.</td>
<td>99.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PctDonate</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>1.0</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>11.0</td>
</tr>
<tr>
<td>Median</td>
<td>19.0</td>
</tr>
<tr>
<td>Mean</td>
<td>22.3</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>31.0</td>
</tr>
<tr>
<td>Max.</td>
<td>63.0</td>
</tr>
</tbody>
</table>

R multiple regression output:

> schools.lm = lm(Gradrate ~ PctTop25+Expense+PctPhds+PctDonate+Public, data=schools)
> summary(schools.lm)

Coefficients:                   Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                    39.55168   6.349332   6.229   1.31e-08 ***
PctTop25                       0.16749    0.084900   1.973   0.05146 .
Expense                        0.00083    0.000430   1.937   0.05571 .
PctPhds                        0.13919    0.104169   1.336   0.18469
PctDonate                      0.08148    0.108376   0.752   0.45405
Public                         -10.8898    3.398053  -3.205   0.00185 **
---                                Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 13.41 on 94 degrees of freedom
Multiple R-Squared: 0.388, Adjusted R-squared: 0.3554
F-statistic: 11.92 on 5 and 94 DF, p-value: 6.082e-09

Indicator variables (or Dummy variables):

If a predictor variable is a binary categorical variable (e.g., male/female), then can arbitrarily define the first category to have the value 0 and the second category to have the value 1.

**Interpretation:** The coefficient to an indicator variable is the effect on the response of the binary variable, holding the other variables constant.

**Example:** The coefficient $-10.89$ of Public (public versus private school) implies that public
colleges have a graduation rate of 10.89% less, on average, than private colleges (controlling for the other variables in the model).

**Indicator variables in R:**

- Can have the variable appear as 0’s and 1’s, as in the current example.
- Alternatively, the variable can be read in as a 2-level categorical variable with names (say) “public” and “private”.
  By default, R will assign 0 to the category that comes first in alphabetical order (though you can override this assignment).

**Least-squares estimation in multiple regression:**

Observe $y_i, x_{i1}, x_{i2}, \ldots, x_{ip}$ for $i = 1, \ldots, n$ ($i$ is indexing an observation).

**Least-squares criterion:**

Select estimates of $\beta_0, \beta_1, \ldots, \beta_p$ to minimize

$$
\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2.
$$

Denote $b_0, b_1, \ldots, b_p$ the values of $\beta_0, \beta_1, \ldots, \beta_p$ that minimize this expression. These values are called the least-squares estimates. Just as in simple linear regression, these can be derived using differential calculus.

But the formulas are a mess! (we’ll perform all the computation on the computer)

**Inference in multiple regression:**

We’ll concentrate on

- Inference (separately) for $\beta_0, \beta_1, \ldots, \beta_p$ (i.e., what is the effect of each predictor variable?)
- Inference for whether $\beta_1 = \cdots = \beta_p = 0$ (i.e., are any of the predictor variables important?).
- Inference for $\mu_y$ given specific predictor values.
- Prediction interval for a future value of the response given the observation’s predictor values.

**Estimating the standard deviation $\sigma$:**

Once the least squares estimates of the $\beta_j$ are obtained, can estimate $\sigma$ by

$$
s = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2}.
$$
R will do this for you – you would never compute this away from the computer.

**Inference for \( \beta_j \):** From running a multiple regression in R, you are provided

- Least squares estimates \( b_j \) of the \( \beta_j \),
- Standard errors of the \( b_j \).
- \( p \)-values for testing the significance of each predictor (2-sided test)

**Inference for regression coefficients:**

- Use \( t \)-distribution (Table D) for inference about a regression coefficient
- Degrees of freedom for \( t \)-distribution is \( n - p - 1 \).
- Same formulas for CI’s and same test statistics for hypothesis tests as simple linear regression.

(you need to construct confidence intervals manually, but the \( p \)-values for testing \( H_o: \beta_j = 0 \) are given)

**Example:**

Is the percent of faculty with PhD’s an important predictor (at the \( \alpha = 0.05 \) level) of graduation rate beyond the information in the other variables?

**Solution:** Want to test

\[
H_o: \beta_3 = 0 \\
H_a: \beta_3 \neq 0
\]

From R output, we have that the \( p \)-value for this test is 0.185.

Thus, we cannot reject the null hypothesis at the 0.05 significance level. There is not strong evidence to conclude that percent faculty with PhD’s is an important predictor beyond the other variables.

**IMPORTANT:**

When you are able to reject \( H_o: \beta_j = 0 \), it would appear that the variable \( x_j \) is a significant predictor of \( Y \).

One main interpretation difference from simple linear regression:

If \( H_o: \beta_j = 0 \) is rejected, then

**WRONG:** The variable \( x_j \) is a significant predictor of \( Y \).
The variable $x_j$ is a significant predictor of $Y$, beyond the effects of the other predictor variables.

**Inference for $\mu_Y$ or future $Y$:**

Point estimation for $\mu_Y$ and prediction for future $Y$ given specific predictor variables $x_1, \ldots, x_p$ — plug in $x$ values into estimated regression equation.

**Examples:**

Harvard University: $x_1 = 100$, $x_2 = 37219$, $x_3 = 97$, $x_4 = 52$, $x_5 = 0$.

Boston University: $x_1 = 80$, $x_2 = 16836$, $x_3 = 80$, $x_4 = 16$, $x_5 = 0$.

Lamar University: $x_1 = 27$, $x_2 = 5879$, $x_3 = 48$, $x_4 = 12$, $x_5 = 1$.

Could ask one of two questions:

1. What are estimates of graduation rates at these schools?
2. What are estimates of the mean graduation rates at schools just like these three (with the same predictor information)?

In the first case, we want a prediction interval for a future $Y$. In the second case, we want a confidence interval for $\mu_Y$.

We can carry this out in R.

First, create three new data frames containing the predictor values:

```r
> harvard=data.frame(PctTop25=100, Expense=37219, PctPhds=97, PctDonate=52, Public=0)
> bu=data.frame(PctTop25=80, Expense=16836, PctPhds=80, PctDonate=16, Public=0)
> lamar=data.frame(PctTop25=27, Expense=5879, PctPhds=48, PctDonate=12, Public=1)
```

The fitted value can be computed directly from the estimated regression. For Harvard:

\[
\hat{y} = 39.6 + 0.167(100) + 0.000833(37219) + 0.139(97) + 0.081(52) - 10.9(0)
\]

\[
\hat{y} = 105.00
\]

Now use the `predict` function to construct 95% confidence intervals and 95% prediction intervals:
> predict(schools.lm, harvard, interval="confidence")
  fit   lwr   upr
[1,] 105.0423 85.59586 124.4888

> predict(schools.lm, harvard, interval="prediction")
  fit   lwr   upr
[1,] 105.0423 72.07601 138.0086

Actual response value for Harvard was 100.

# Boston University

> predict(schools.lm, bu, interval="confidence")
  fit   lwr   upr
[1,] 79.4142 72.6324 86.19601

> predict(schools.lm, bu, interval="prediction")
  fit   lwr   upr
[1,] 79.4142 51.94412 106.8843

# Lamar University

> predict(schools.lm, lamar, interval="confidence")
  fit   lwr   upr
[1,] 45.74036 38.55972 52.921

> predict(schools.lm, lamar, interval="prediction")
  fit   lwr   upr
[1,] 45.74036 18.1691 73.31161

Observed graduation rates were 72 for BU, and 26 (?) for Lamar.

**F-test for regression:** Purpose – to carry out the hypothesis test

\[ H_0: \beta_1 = \cdots = \beta_p = 0 \]
\[ H_a: \text{At least one } \beta_j \neq 0, (j = 1, \ldots, p). \]

Use line corresponding to “F-statistic” to answer this question.

Under \( H_0 \), the F-statistic has an F-distribution on \( p \) and \( n - p - 1 \) degrees of freedom, with the \( p \)-value as given.

Residual standard error: 13.41 on 94 degrees of freedom
Multiple R-Squared: 0.388,  Adjusted R-squared: 0.3554
F-statistic: 11.92 on 5 and 94 DF,  p-value: 6.082e-09

15
**Goodness of fit:**

Analogous to $R^2$ in simple linear regression, can also calculate $R^2$ for multiple regression (sometimes called “multiple $R^2$”).

**Formula for $R^2$:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$  

As in simple linear regression, $R^2$ is the proportion of variability in the $Y$ values explained by the predictor variables.

In the graduation rate example, 38.8% of the variability in graduation rates is explained by the five predictors.

**Categorical predictor variables:**

- We know how to address binary predictor variables: Use indicator variables.

- But what if a predictor variables is categorical variable with more than two levels?

**Example:** Predicting pulse rates

Pulse rates were measured for a sample of 83 people, along with a participant’s

- Age (in years)

- Self-reported exercise level (high, moderate, low)

Want to perform least-squares regression estimating mean pulse rate from age and exercise level.

**Data summary:**

<table>
<thead>
<tr>
<th>Age</th>
<th>exercise</th>
<th>Pulse1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. :18.00</td>
<td>high : 9</td>
<td>Min. : 47.00</td>
</tr>
<tr>
<td>1st Qu.:19.00</td>
<td>moderate:45</td>
<td>1st Qu.: 66.00</td>
</tr>
<tr>
<td>Median :20.00</td>
<td>low :29</td>
<td>Median : 75.00</td>
</tr>
<tr>
<td>Mean :20.69</td>
<td></td>
<td>Mean : 75.29</td>
</tr>
<tr>
<td>3rd Qu.:21.00</td>
<td></td>
<td>3rd Qu.: 81.50</td>
</tr>
<tr>
<td>Max. :45.00</td>
<td></td>
<td>Max. :145.00</td>
</tr>
</tbody>
</table>
Converting categorical predictors into numerical data:

If a categorical predictor variable has $K$ levels, then create $K - 1$ indicator (binary) variables to use in the regression in its place.

**Application to the “exercise” variable:** Define $x_1$ to be a person’s age, and then let

- $x_2 = 1$ if a person reports “moderate” exercise, 0 otherwise
- $x_3 = 1$ if a person reports “low” exercise, 0 otherwise

Leave out “high” exercise category.

**A few observations of the original data:**

<table>
<thead>
<tr>
<th>Age</th>
<th>exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>high</td>
</tr>
<tr>
<td>18</td>
<td>moderate</td>
</tr>
<tr>
<td>19</td>
<td>low</td>
</tr>
<tr>
<td>20</td>
<td>moderate</td>
</tr>
<tr>
<td>18</td>
<td>moderate</td>
</tr>
</tbody>
</table>
The “exercise” variable recoded:

<table>
<thead>
<tr>
<th>Age</th>
<th>exerciselow</th>
<th>exercisemoderate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The regression model:

\[ \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]

The contribution of exercise to the above expression is \( \beta_2 x_2 + \beta_3 x_3 \).

Specifically, the contribution of

- “high” exercise is \( \beta_2(0) + \beta_3(0) = 0 \).
- “moderate” exercise is \( \beta_2(1) + \beta_3(0) = \beta_2 \).
- “low” exercise is \( \beta_2(0) + \beta_3(1) = \beta_3 \).

This implies that

- \( \beta_2 \) is the effect of “moderate” exercise relative to “high” exercise, and
- \( \beta_3 \) is the effect of “low” exercise relative to “high” exercise.

In summary, choose one category to be the “reference level” (like “high” exercise), and create indicator variables for the other levels. The effect of each indicator variable is then the effect of the level relative to the reference level.

Luckily, R does the conversion to indicator variables for you!

Implementation in R:

```r
> x1.fit = lm(Pulse1 ~ Age + exercise, data=x)
> summary(x1.fit)
```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.5055  8.8820  8.501  9.18e-13 ***
Age        -0.5942  0.3764 -1.579  0.11839
exercisemoderate  12.2766  5.0952  2.409  0.01830 *
exerciselow    15.5114  5.3270  2.912  0.00467 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 13.94 on 79 degrees of freedom
Multiple R-squared: 0.1176, Adjusted R-squared: 0.08414
F-statistic: 3.511 on 3 and 79 DF, p-value: 0.019

**Interpretation for exercise predictor variable:**

- A person’s resting pulse rate is about 12.28 beats per minute faster if the person engages in moderate exercise relative to high levels of exercise.
- A person’s resting pulse rate is about 15.51 beats per minute faster if the person engages in low levels of exercise relative to high levels of exercise.

In both cases, these conclusions are significant at the $\alpha = 0.05$ level.

**Categorical variables and statistical significance:**

- The two tests we performed were comparing the effect of one level of the categorical variable to the reference level.
- We cannot tell based on the output whether moderate exercise has a significantly different effect on pulse rates than low exercise levels.
- Even more subtle: We cannot tell whether “exercise” as an entire variable is significant beyond the effect of age based on the output.

However, we can answer this by performing the “anova” command on the final model:

```r
> anova(x1.fit)
Analysis of Variance Table

Response: Pulse1
           Df Sum Sq Mean Sq F value Pr(>F)
Age         1 395.3  395.30  2.0339 0.15776
exercise    2 1651.9  825.93  4.2497 0.01766 *
Residuals   79 15353.9 194.35
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

The categorical variable “exercise” does appear to be significant at the $\alpha = 0.05$ level.