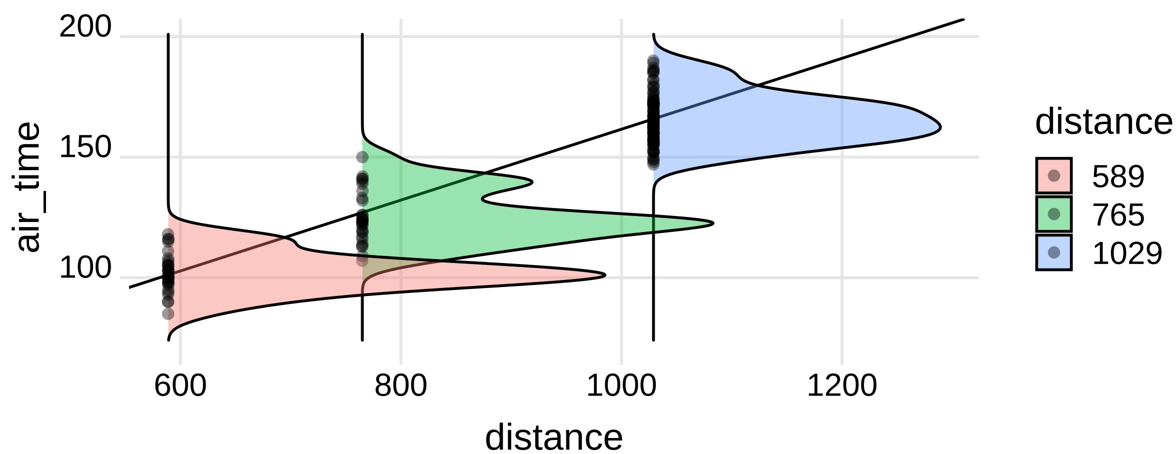
Residuals for "Simple" Linear Regression

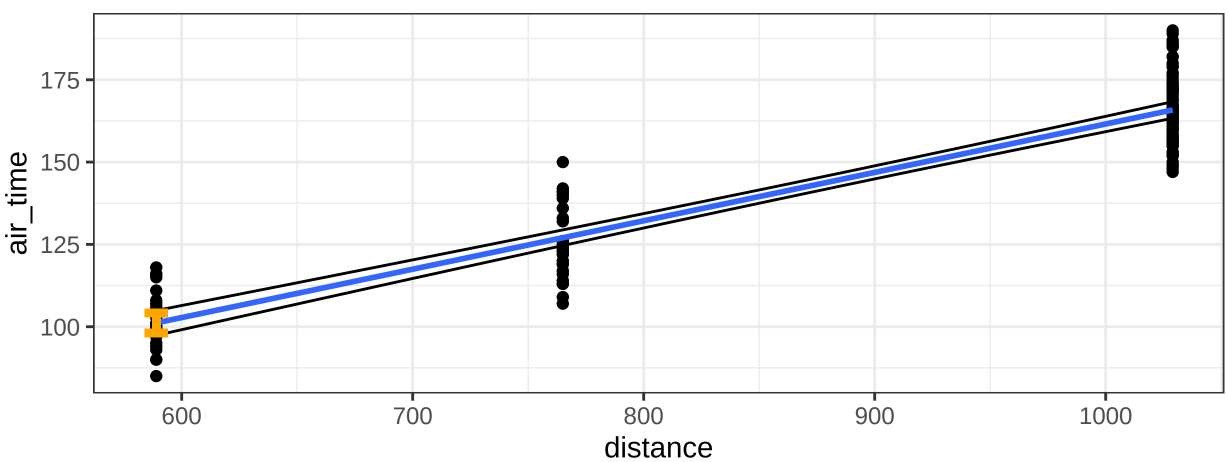
Sleuth 3 Sections 7.3.1, 7.3.4, and 7.4.3

Previously

- Example: flight air times (response) as a function of distance (explanatory)



- Observations follow a normal distribution with mean that is a linear function of the explanatory variable
- A few ways of writing this:
  - Y follows a normal distribution with mean $\mu = \beta_0 + \beta_1 X$
  - $Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma)$
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$ where $\varepsilon_i \sim \text{Normal}(, \sigma)$
- The last topic we covered was confidence intervals for the mean response at a given value of X:
  - We are 95% confident that the mean air time for flights travelling 589 miles is between 98.1 min and 104.2 min.
  - We are 95% confident that at every distance, the population mean air time at that distance is within the Scheffe-adjusted confidence bands.



Today

- Individual responses don't fall exactly at the mean. We can quantify how far from the line observations tend to fall
- After today, you should be able to:
  - Calculate a residual from a simple linear regression model fit
  - Know that the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found by minimizing the sum of squared residuals
  - Use the residual standard error to get a rough sense of how close points tend to fall to the line

- Find and interpret a prediction interval using R commands
- Understand why prediction intervals are wider than confidence intervals

Example Data Set: US News and World Reports 2013 College Statistics

Across colleges in the US, we have measurements of (among other variables):

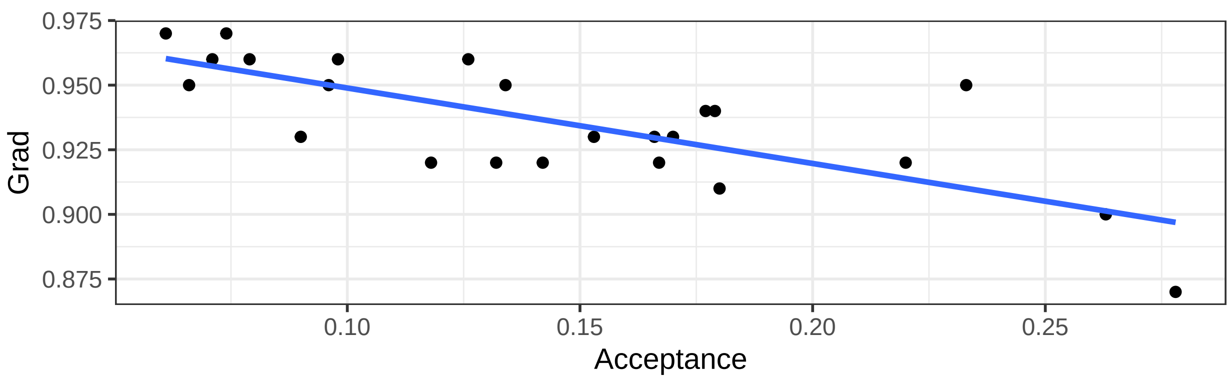
- Acceptance rate (what proportion of applicants are admitted)
- Graduation rate (what proportion of students graduate within 6 years)

Let's study the association between the acceptance rate (explanatory) and graduation rate (response).

```
library(readr)
colleges <- read_csv("http://www.evanlray.com/data/sdm4/Graduation_rates_2013.csv")
head(colleges)
```

```
## # A tibble: 6 x 5
##   Tuition Enrollment Acceptance Retention  Grad
##     <dbl>      <dbl>      <dbl>     <dbl> <dbl>
## 1   40170       8010      0.079      0.98  0.96
## 2   42292      19726      0.061      0.98  0.97
## 3   44000      11906      0.071      0.99  0.96
## 4   49138      23168      0.074      0.99  0.97
## 5   43245      18217      0.066      0.98  0.95
## 6   46386      12508      0.132      0.99  0.92
```

```
ggplot(data = colleges, mapping = aes(x = Acceptance, y = Grad)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

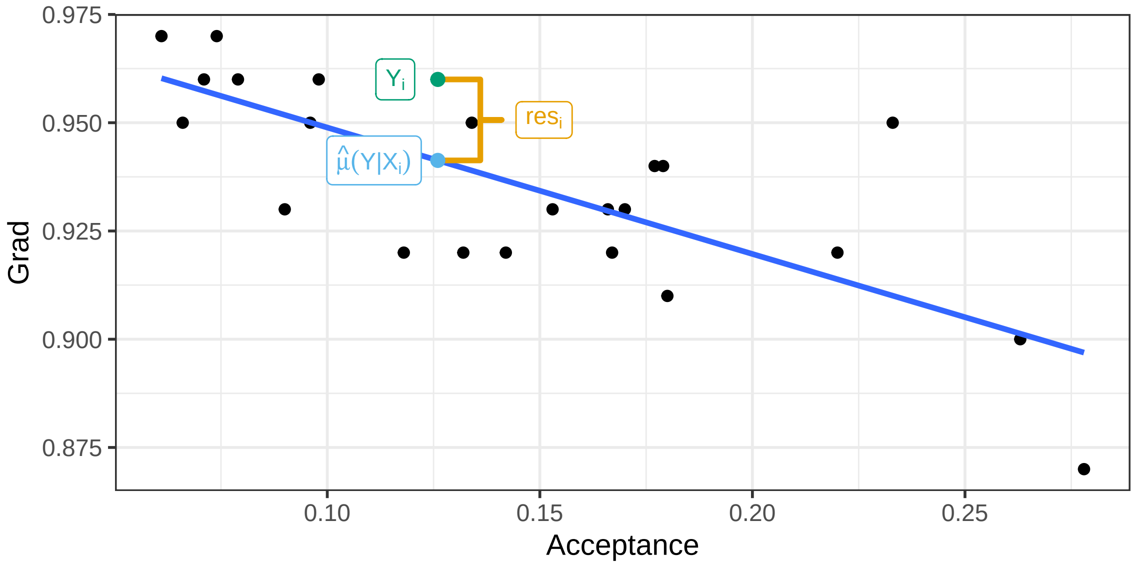



```
linear_fit <- lm(Grad ~ Acceptance, data = colleges)
summary(linear_fit)
```

```
##
## Call:
## lm(formula = Grad ~ Acceptance, data = colleges)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.026914 -0.010876  0.000968  0.010656  0.039947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.978086   0.008582 113.966  < 2e-16 ***
## Acceptance  -0.291986   0.054748  -5.333 2.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01617 on 22 degrees of freedom
## Multiple R-squared:  0.5639, Adjusted R-squared:  0.544
## F-statistic: 28.44 on 1 and 22 DF,  p-value: 2.36e-05
```

Residuals

- Residual = Observed Response - Predicted Response
- res_i = Y_i - $\widehat{\mu}(Y|X_i)$
- res_i = Y_i - $(\hat{\beta}_0 + \hat{\beta}_1 X_i)$



1. The college highlighted in the figure above had an acceptance rate of 0.126, and a graduation rate of 0.96. Find the predicted graduation rate for colleges with acceptance rates of 0.126 and the residual for this college.

Find the predicted value:

```r
0.978 - 0.292 * 0.126
```

```
## [1] 0.941208
```

```r
predict(linear_fit, newdata = data.frame(Acceptance = 0.126))
```

```
##         1
## 0.9412959
```

Find the residual:

Model fit by least squares

- In general, smaller residuals are better (but not always – to be discussed in more depth later?)
- Most common strategy for estimating β_0 and β_1 is by minimizing the Residual Sum of Squares:

$$\hat{\beta}_0 \text{ and } \hat{\beta}_1 \text{ minimize } \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

- There are also other approaches (to be discussed later?)

Accessing the Residuals in R

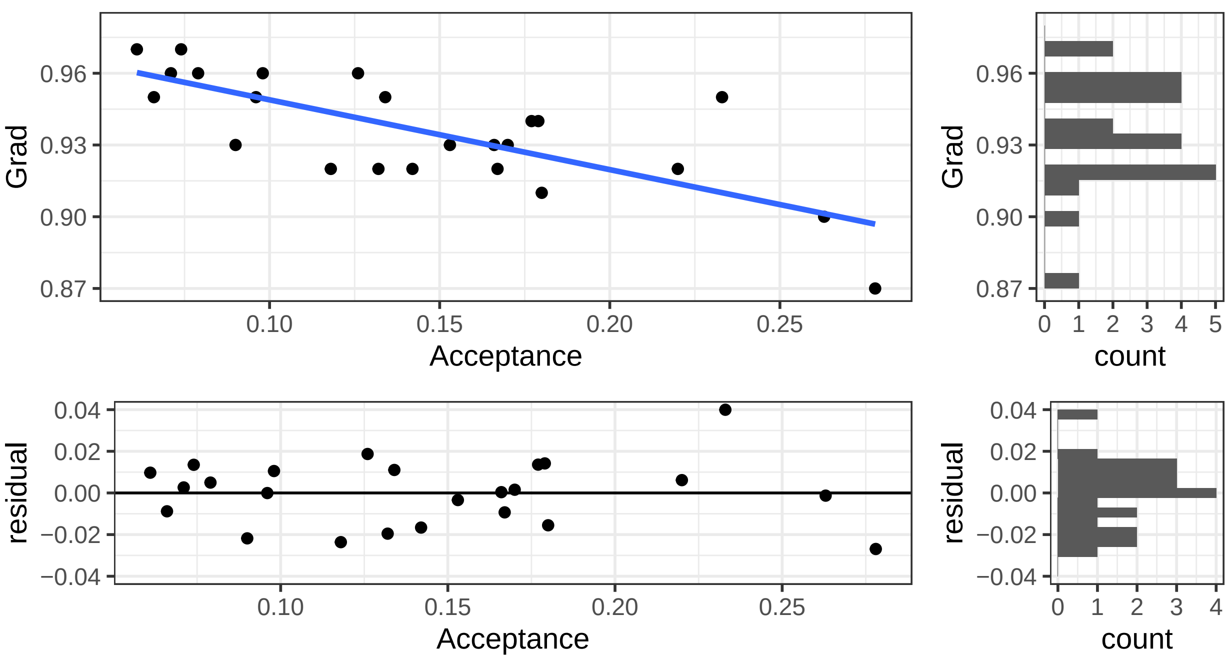
```r
colleges <- colleges %>%
  mutate(
    fitted = predict(linear_fit),
    residual = residuals(linear_fit)
  )
head(colleges)
```

```
## # A tibble: 6 x 7
##   Tuition Enrollment Acceptance Retention  Grad fitted residual
##     <dbl>      <dbl>      <dbl>     <dbl> <dbl>  <dbl>    <dbl>
## 1   40170       8010      0.079      0.98  0.96  0.955  0.00498
## 2   42292      19726      0.061      0.98  0.97  0.960  0.00972
## 3   44000      11906      0.071      0.99  0.96  0.957  0.00264
## 4   49138      23168      0.074      0.99  0.97  0.956  0.0135
## 5   43245      18217      0.066      0.98  0.95  0.959 -0.00882
## 6   46386      12508      0.132      0.99  0.92  0.940 -0.0195
```

```r
# Verifying the first residual calculation: observed response - fitted response
0.96 - 0.955
```

```
## [1] 0.005
```

We can then make plots (more next class):



- **Question of the day:** How far do the points tend to be from the line?
  - **Answer 1:** $\pm 2 \times$ (Standard deviation of residuals) (quick and approximate)
  - **Answer 2:** Prediction intervals (formal)

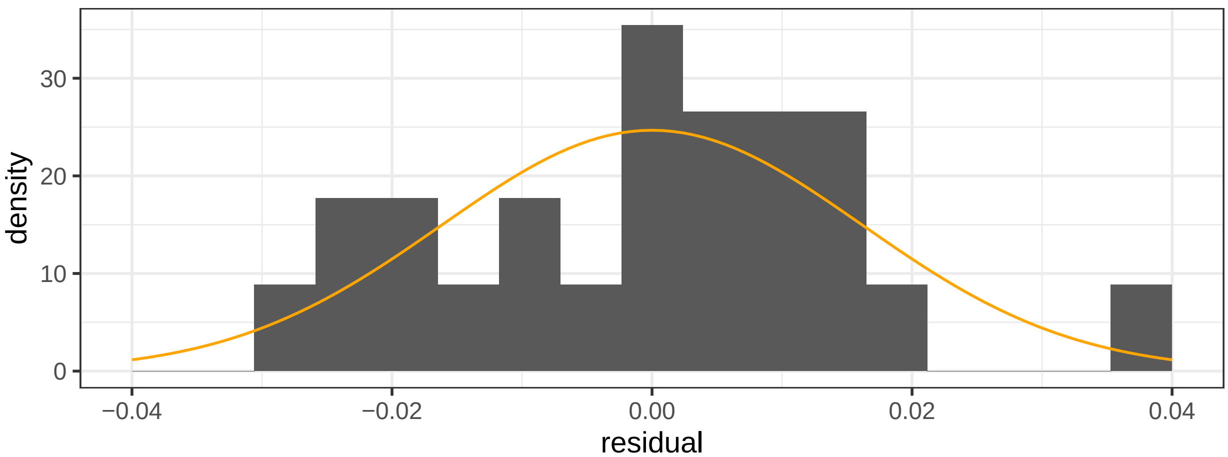
Answer 1: $\pm 2\times$ Standard Deviation of Residuals (Approximate)

- Model: $Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma)$
- Parameter σ (unknown!!) describes standard deviation of the normal distribution **in the population**
- Estimate it by

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of Squared Residuals}}{n - (\text{number of parameters for the mean})}} = \sqrt{\frac{\sum_{i=1}^{n}\{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\}^2}{n-2}}$$

- This is listed in the `summary` output as the "Residual standard error": 0.01617
    - (this is reasonable terminology but not quite in agreement with our definition of standard error)

Here is the histogram of the residuals from the last page with a Normal(0, 0.01617) distribution overlaid:



- Fact 1: If a variable follows a normal distribution, about 95% of observations will fall within ± 2 standard deviations of the mean
- Fact 2: The mean of the residuals is 0

2. Based on the residual standard deviation, about how close are the observed responses to the fitted mean responses?

```
2 * 0.01617
```

```
## [1] 0.03234
```

Prediction Intervals

Our Goal

- An interval that will contain the response y_0 for a new observation at a value x_0 of the explanatory variable
- Our best guess is the estimated mean $\hat{\mu}$
- The amount by which our guess is wrong is the residual for the new observation:

$$\text{Observed Response - Estimated Mean}$$

Two Contributions to Prediction Error

$$\text{Observed Response - Estimated Mean} = \text{(Observed Response - Actual Mean) - (Estimated Mean - Actual Mean)}$$

1. Variability of observed response around true population mean: σ, estimated by $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$
2. Variability of estimated mean around true population mean: estimated by $SE(\hat{\mu}) = \sqrt{\hat{\sigma}^2 \frac{1}{n} + \hat{\sigma}^2 \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$

- We put those two pieces together to get:
  - $SE(\hat{\mu} - y_0) = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \frac{1}{n} + \hat{\sigma}^2 \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$

Prediction Intervals

- Prediction intervals for a new response are based on the error of the estimated mean from the response y for a new individual observation
  - $[\hat{\mu} - t^* SE(\hat{\mu} - y_0), \hat{\mu} + t^* SE(\hat{\mu} - y_0)]$
  - For 95% of samples and 95% of new observations with the specified value of x, a CI calculated using this formula will contain the response for those new observations, $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$.

Compare to Confidence Intervals (from last class)

- Confidence intervals for the mean were based on the error of the estimated mean from the actual population mean
  - $[\hat{\mu} - t^* SE(\hat{\mu}), \hat{\mu} + t^* SE(\hat{\mu})]$ where
  - For 95% of samples, a CI calculated using this formula will contain the population mean response at x_0, $\mu = \beta_0 + \beta_1 x_0$

3. Find and interpret a 95% prediction interval for the graduation rate of a college that was not in our data set before, and has an acceptance rate of 0.1.

```r
predict_df <- data.frame(
  Acceptance = 0.1
)
predict(linear_fit, newdata = predict_df, interval = "prediction", se.fit = TRUE)
```

```
## $fit
##         fit       lwr     upr
## 1 0.9488876 0.9142951 0.98348
##
## $se.fit
## [1] 0.004108595
##
## $df
## [1] 22
##
## $residual.scale
## [1] 0.01616618
```

Compare to a confidence interval for the mean:

```r
predict(linear_fit, newdata = predict_df, interval = "confidence", se.fit = TRUE)
```

```
## $fit
##         fit       lwr       upr
## 1 0.9488876 0.9403669 0.9574083
##
## $se.fit
## [1] 0.004108595
##
## $df
## [1] 22
##
## $residual.scale
## [1] 0.01616618
```

No easy way to get Scheffe adjusted simultaneous intervals, but we can plot the individual prediction intervals at each value of x in our data set as follows:

```
intervals <- predict(linear_fit, interval = "prediction") %>%
  as.data.frame()
```

```
## Warning in predict.lm(linear_fit, interval = "prediction"): predictions on current data refer to _future_
```

```
head(intervals)
```

```
##         fit       lwr       upr
## 1 0.9550193 0.9199975 0.9900411
## 2 0.9602750 0.9247617 0.9957884
## 3 0.9573552 0.9221287 0.9925817
## 4 0.9564792 0.9213321 0.9916263
## 5 0.9588151 0.9234494 0.9941808
## 6 0.9395440 0.9052956 0.9737924
```

```
colleges <- colleges %>%
  bind_cols(
    intervals
  )
head(colleges)
```

```
## # A tibble: 6 x 10
##   Tuition Enrollment Acceptance Retention  Grad fitted residual   fit   lwr   upr
##     <dbl>      <dbl>      <dbl>     <dbl> <dbl>  <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1   40170       8010      0.079      0.98  0.96  0.955  0.00498 0.955 0.920 0.990
## 2   42292      19726      0.061      0.98  0.97  0.960  0.00972 0.960 0.925 0.996
## 3   44000      11906      0.071      0.99  0.96  0.957  0.00264 0.957 0.922 0.993
## 4   49138      23168      0.074      0.99  0.97  0.956  0.0135  0.956 0.921 0.992
## 5   43245      18217      0.066      0.98  0.95  0.959 -0.00882 0.959 0.923 0.994
## 6   46386      12508      0.132      0.99  0.92  0.940 -0.0195  0.940 0.905 0.974
```

```
ggplot(data = colleges, mapping = aes(x = Acceptance, y = Grad)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_line(mapping = aes(y = lwr), linetype = 2) +
  geom_line(mapping = aes(y = upr), linetype = 2) +
  theme_bw()
```