Elementary Regression 2

Paul E. Johnson[1] [2]

[1] Department of Political Science

[2] Psychology, University of Kansas

Sept 28, 2020

Outline

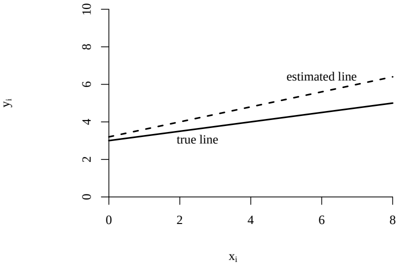
1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

How Does Uncertainty Manifest Itself?

- The Truth is $\beta_0 = 3$, $\beta_1 = 0.25$
- Suppose sample estimate $\hat{\beta}_0 = 3.2$ and $\hat{\beta}_1 = 0.4$.
- We are a little bit off the mark, but we are not doing too badly to formulate a "prediction" thusly

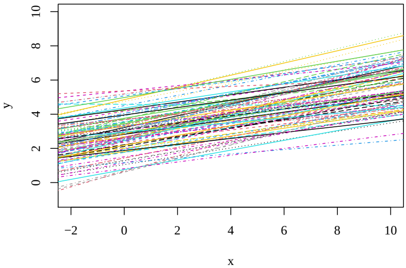
$$\hat{y}_i = 3.2 + 0.4 \cdot x_i$$

- If we did not know true β_0 and β_1, could we guess "how far wrong" our estimates are?



Repeat that Exercise Hundreds of Times

- Draw samples, estimate a line for each
- Vital observations
  - Estimates do seem to "hover" around the correct values
  - More predictive fluctuation on edges than in the middle
  - If $\hat{\beta}_0$ is "off" by a larger amount, the $\hat{\beta}_1$ will generally be off as well (that's $Cov(\hat{\beta}_0, \hat{\beta}_1)$.



How Did I Manufacture the Data?

- Sample size N=100
- Draw one sample of input variables, $x_i \sim \mathit{Normal}(50, 10^2)$
- The "true" parameter values: $\beta_0 = 3$, $\beta_1 = 0.25$, $\sigma_e = 10$
- Repeatedly draw sets of errors, estimate regressions (leaving x_i vector the same)
- This is what it means when textbooks say "x is fixed" across repeated samples

The Simulation Confirms The Theory

- The expected values of the estimators are:

$$\begin{aligned} E[\hat{\beta}_0] &= 3 \\ E[\hat{\beta}_1] &= 0.25 \\ E[RMSE] &= 10 \end{aligned}$$

- According to results derived below:
  - Variance of $\hat{\beta}_1$: $Var[\hat{\beta}_1] = \sigma_e^2 / E[\sum(x - \bar{x})^2] = 1/100 = 0.01$
  - Standard deviation of $\hat{\beta}_1$: $std.dev(\hat{\beta}_1) = 0.1$
  - $\hat{\beta}_1$ is Normally distributed if
    - Sample is large (Recall the Central Limit Theorem)
    - Or we assume e_i is Normal. Then $\hat{\beta}$ will be Normal.
  - $\hat{t} = (\hat{\beta}_1 - \beta_1)/s.e.(\hat{\beta}_1)$ is distributed according to a t distribution with N-2 degrees of freedom.

Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

Variance Results

- $Var[\hat{\beta}]$: Theoretical "True Variance" of estimate across repeated samples

$$Var(\hat{\beta}_1) = \sigma_e^2 \left(\frac{1}{\sum (x_i - \bar{x})^2} \right) \tag{1}$$

- $\widehat{Var}[\hat{\beta}]$: Estimate of $Var[\hat{\beta}]$ From one sample. Replace σ_e^2 with $\hat{\sigma}_e^2$ (MSE).

$$\widehat{Var}(\hat{\beta}_1) = \hat{\sigma}_e^2 \left(\frac{1}{\sum (x_i - \bar{x})^2} \right) \tag{2}$$

- $std.err.(\hat{\beta}) = \sqrt{\widehat{Var}[\hat{\beta}]}$: Standard error of $\hat{\beta}_1$. We don't call it a "standard deviation" because it is based on an estimate of the variance, rather than the true variance.

See Appendix for Derivation

- The derivation begins by applying the Var operator to both sides of the formula for the slope estimate

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right)$$

- Appendix shows derivation.
- Demonstrates role played by assumptions $E[e_i] = 0$ and $Var[e_i] = \sigma_e^2$.

Similar formulas for variances of other parameter estimates

- Variance of Intercept:

$$\widehat{Var(\hat{\beta}_0)} = \hat{\sigma}_e^2 \frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}$$

- Covariance of estimates of Intercept and Slope:

$$\widehat{Cov(\hat{\beta}_0, \hat{\beta}_1)} = \frac{-\bar{x}\hat{\sigma}_e^2}{\sum (x_i - \bar{x})^2}$$

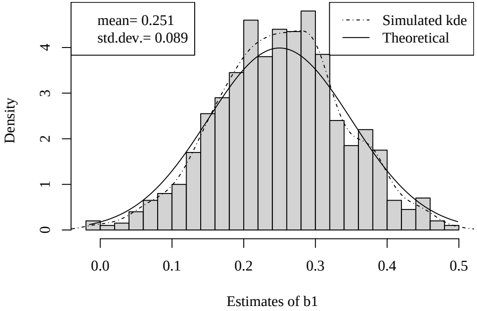
They Fit into a Variance/Covariance matrix

$$Var/Covar(\hat{\beta}) : \begin{bmatrix} \widehat{Var(\hat{\beta}_0)} & \widehat{Cov(\hat{\beta}_0, \hat{\beta}_1)} \\ \widehat{Cov(\hat{\beta}_0, \hat{\beta}_1)} & \widehat{Var(\hat{\beta}_1)} \end{bmatrix}$$

- The square roots of the diagonal appear in the standard regression output
- They are the 2nd column, the standard errors of parameter estimates.
- $\widehat{Cov(\hat{\beta}_0, \hat{\beta}_1)}$ is not presented in the standard regression output, must be obtained separately
- Note: I am lazy and don't put a giant hat over the matrix on the left.

The Sampling Distribution of $\hat{\beta}_1$ is Normal

- Simulation draws similar to theoretical Normal distribution
- Recall, the true value of $\beta_1 = 0.25$
- Variation we expect (theoretical) is observed in simulation

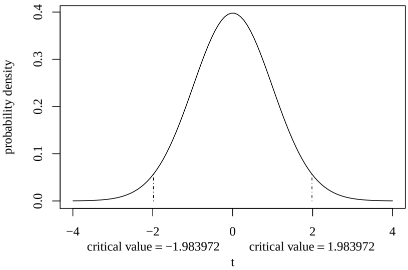


1000 Simulated Samples, N=100, x sample fixed

Sampling Distribution of $(\hat{\beta}_1 - \beta_1)/s.e.(\hat{\beta}_1)$ follows a t Distribution

- When we studied estimating the average from a sample, we found the ratio $\bar{x}/s.e.(\bar{x})$ is distributed as a t statistic.
- The same idea applies here: because
  - $\hat{\beta}_1$ follows a Normal distribution, and
  - $s.e.(\hat{\beta}_1)$ follows a Chi-Square
  - therefore, $(\hat{\beta}_1 - \beta_1)/s.e.(\hat{\beta}_1)$ follows a t distribution
- Often, people simply refer to that ratio as a “t statistic”, but I’m calling it $\hat{t}$ because it varies from sample to sample, just like $\hat{\beta}$ and $s.e.(\hat{\beta})$.

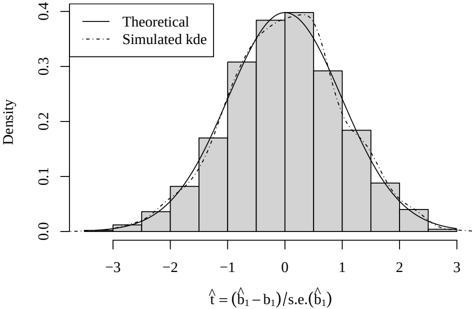
Recall a t Distribution with 100 df



- The estimate from a sample, $\hat{t} = \hat{\beta}_1 - \beta_1 / s.e.(\hat{\beta}_1)$ will take on a range of values around 0
- Only infrequently, with probability (2x0.025), will $\hat{t}$ be in the "tails", the critical regions.

The Sampling Distribution of $(\hat{\beta}_1 - \beta_1)/s.e.(\hat{\beta}_1)$

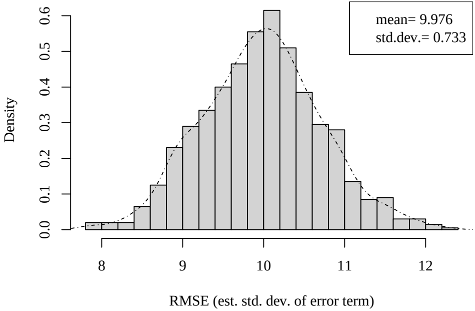
- Note similarity of sample estimates $(\hat{\beta}_1 - \beta_1)/s.e.(\hat{\beta}_1)$ with the theoretical t distribution



1000 Simulated Samples, N=100, x sample fixed

The Sampling Distribution of RMSE

distribution of estimated root mean square error is centered on the true value of the standard deviation of the error term.



1000 Simulated Samples, N=100, x sample fixed

Outline

1 $\hat{\beta}$ Uncertainty
- Visualize Uncertainty

2 Sampling Distribution $\hat{\beta}$

3 t-test Hypotheses about β_j

4 Confidence Interval of $\hat{\beta}_j$

5 Afterthought: Simulated Distribution of R^2

6 Prediction CI

7 Re-scale variables
- Multiply x_i to Re-Scale It
- Subtract from x_i to Make Intercept Easier to Interpret
- Standardize Variables

8 Appendix 1. Variance of $\hat{\beta}$

9 Appendix 2: Proofs of CI and PI

10 Practice Problems

Check the Standard Regression Output

```
require(car)
incedmod1 <- lm(income~education, data=Prestige)
summary(incedmod1)
```

```
Call:
lm(formula = income ~ education, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-5493.2 -2433.8   -41.9  1491.5 17713.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2853.6     1407.0  -2.028   0.0452 *
education      898.8      127.0   7.075 2.08e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3483 on 100 degrees of freedom
Multiple R^2:  0.3336, Adjusted R^2:  0.3269
F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10
```

Ask R for the Covar Matrix

```
incedvcov <- vcov(incedmod1)
incedvcov
```

```
            (Intercept)  education
(Intercept)   1979759.4  -173290.4
education     -173290.4    16138.0
```

- Note the square root of the diagonals is same as “standard error” in regression table

```
sqrt(diag(incedvcov))
```

```
(Intercept)   education
  1407.0392    127.0354
```

The Standard Error of $\hat{\beta}_1$ Leads to a T-test

- The regression output has columns

Estimate of b	std. error of b	t=$\hat{\beta}/s.e.(\hat{\beta})$	prob t more extreme than $\hat{t}$

- t column is meaningful only if NULL is $\beta_j = 0$ (j means either 0 or 1 in β_0 and β_1)
- β_j does not always have to be 0!. More generally

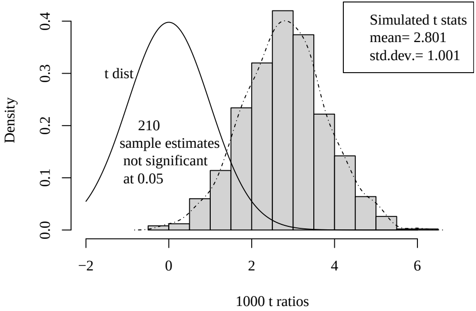
$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{std.err.(\hat{\beta}_j)} \qquad (3)$$

Compare that against a t distribution.

- Rule of Thumb: if $|\hat{t}| \leq 2$, the difference between the estimate $\hat{\beta}_j$ and β_j is not “statistically significant”

The Simulated Sampling Distribution of t

- The t-stats reported by regression models assume null, $H_0 : \beta_1 = 0$
- Many estimated $\hat{\beta}_1 / s.e.(\hat{\beta}_1)$ are greater than 1.983, as they should be!
- Many are not. This is an example of Type II error (β error), failing to reject an incorrect null hypothesis.

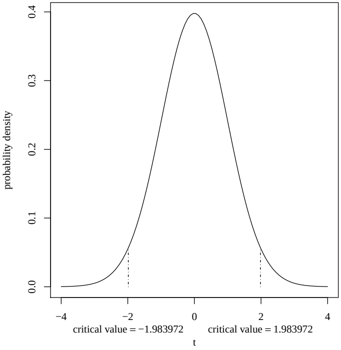


Solid line: t would follow this if $\beta_1 = 0$, df=98
Dotted line: Simulated estimates of $\hat{\beta}_1 / s.e.(\hat{\beta})$ when $\beta_1 = 0.25$

Two-Tailed Versus One Tailed

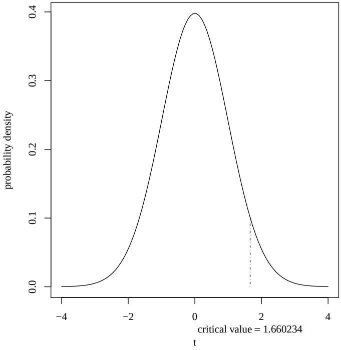
If Null Hypothesis is Correct, the estimate of t will be distributed like this:

Two Tailed Test



Can reject null on either high or low side

One Tailed Test



Can reject null only on high side

4 Steps of T-test: The Prestige Regression Slope

1. State Theoretical model to define terms:
   $income_i = \beta_0 + \beta_1 \cdot education_i + e_i$, $(E[e_i] = 0,\ E[e_i^2] = \sigma_e^2)$
2. State Null Hypothesis (for example): $H_0 : \beta_1 = 0$.
3. Define decision guideline: with 100 df, the 0.05 critical value of t is 1.983 (two-tailed test).
4. Calculate $\hat{t} = (898.8 - 0)/127 = 7.075$
   which is far greater than 1.983, so the null is rejected.

Estimated $s.e.(\hat{\beta}_1)$ Will be High if

- Recall, the estimated variance of an estimated slope:

$$\widehat{Var(\hat{\beta}_1)} = \widehat{\sigma_e^2}\left[\frac{1}{\sum(x_i - \bar{x})^2}\right] = \frac{RMSE^2}{\sum(x_i - \bar{x})^2}$$

- $se(\hat{\beta}_1)$ will be high if
  - $\widehat{\sigma_e^2}$ is high (so, Big error variance -> Big $\hat{\beta}$ Variance)
  - $\sum(x_i - \bar{x})^2$ is small (Low variance of x_i -> Big $\hat{\beta}$ Variance).

Outline

1 $\hat{\beta}$ Uncertainty
- Visualize Uncertainty

2 Sampling Distribution $\hat{\beta}$

3 t-test Hypotheses about β_j

4 Confidence Interval of $\hat{\beta}_j$

5 Afterthought: Simulated Distribution of R^2

6 Prediction CI

7 Re-scale variables
- Multiply x_i to Re-Scale It
- Subtract from x_i to Make Intercept Easier to Interpret
- Standardize Variables

8 Appendix 1. Variance of $\hat{\beta}$

9 Appendix 2: Proofs of CI and PI

10 Practice Problems

Confidence Interval Reminder

- Build a Confidence Interval around $\hat{\beta}$.

$$CI: \quad \hat{\beta} - t \cdot std.err(\hat{\beta}) \leq \beta \leq \hat{\beta} + t \cdot std.err(\hat{\beta})$$

We believe that the probability is 95% that the "true value of b" will be in the CI. The t value will depend on the degrees of freedom available (Sample Size minus parameters estimated, or $N-2$ in this case).

- Result was derived in previous lecture on Confidence Intervals. With probability 0.95, the estimated t ratio will lie in a range,

$$Prob(-t \leq \frac{\widehat{\beta} - \beta}{std.err.(\hat{\beta})} \leq t) = 0.95$$

- That implies this, with probability 0.95, the interval includes the "true" β

$$-t \cdot std.err(\hat{\beta}) \leq \hat{\beta} - \beta \leq t \cdot std.err(\hat{\beta})$$

CI($\hat{\beta}$) Example: The Prestige Regression

- Confidence Interval for slope estimate: $CI(\hat{\beta}_1)$

$$\hat{\beta}_1 \pm t \cdot std.err.(\hat{\beta}_1) =$$

$$898.8 \pm 1.98 \times 127$$

- Or

$$[646.7792, \;\; 1150.84748]$$

- Result: We believe that the probability is 0.95 that the "true β_1" would be between 646.7 and 1150.8.
- Or, 95% of the time, when we conduct this sampling experiment, the CI calculated according to this formula would include the true value.

In R, ask for the Confidence Intervals of all coefficients

```
confint(incedmod1)
```

```
                 2.5 %      97.5 %
(Intercept) -5645.1114   -62.05979
education     646.7782  1150.84748
```

Outline

1 $\hat{\beta}$ Uncertainty
- Visualize Uncertainty

2 Sampling Distribution $\hat{\beta}$

3 t-test Hypotheses about β_j

4 Confidence Interval of $\hat{\beta}_j$

5 Afterthought: Simulated Distribution of R^2

6 Prediction CI

7 Re-scale variables
- Multiply x_i to Re-Scale It
- Subtract from x_i to Make Intercept Easier to Interpret
- Standardize Variables

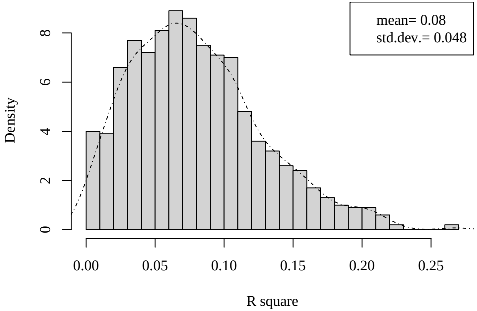
8 Appendix 1. Variance of $\hat{\beta}$

9 Appendix 2: Proofs of CI and PI

10 Practice Problems

The Sampling Distribution of R^2

- R^2
- Cov(xy)/Sd(x)sd(y)
- not so encouraging



1000 Simulated Samples, N=100, x sample fixed

Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. **Prediction CI**
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

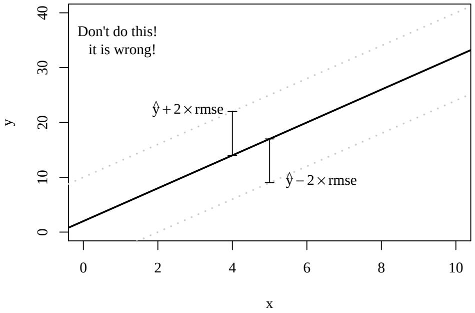
Does $\hat{y}_i$ estimate $E[y_i|x_i]$ or y_i?

- We are asking “How meaningful is $\hat{y}_i$” What is it good for?
- 2 possibilities.
  - estimate the “true value” of y_i, which is $E[y_i|x_i]$
  - estimate a particular case’s outcome, y_i
- That leads to 2 different confidence intervals we can place around our prediction.

Recall root MSE: estimated std.dev. of error term

- If we knew β_o and β_1 for sure, then we could draw lines$\pm 2 \times RMSE$ to predict 95% of the observations (supposing e_i is Normal, of course).
- That would be wrong: We don't know β_o and β_1 for sure.
- It is not wide enough to include our uncertainty!

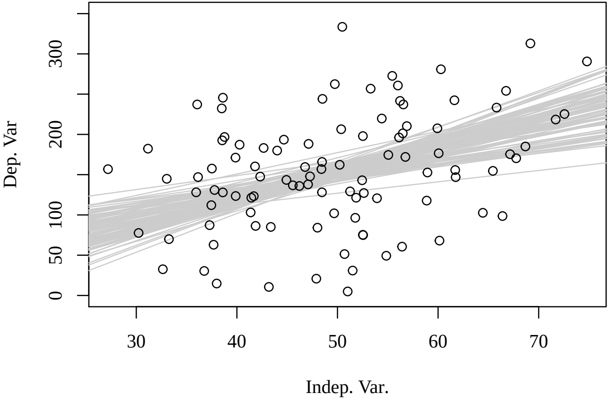
Danger, This is Wrong



Including uncertainty about $\hat{b}_0$ and $\hat{b}_1$ leads to an hour glass shaped region

Example: 100 regression lines

$\beta_0 = 2, \beta_1 = 3, \sigma_e^2 = 80^2$



Note: Points represent one "sample", lines represent 100 "sample fits".

Confidence and Prediction Intervals

- Confidence Interval:
  - Given x_i and predicted value $\hat{y}_i$, how wide must an interval be to include the "true (error free) y_i" with probability.
  - Summarizes our uncertainty about $\hat{y}_i$ as an estimate of $E[y_i|x_i]$
  - $\hat{y}_i$ should be "pretty close" to $E[y_i|x_i]$
- Prediction interval:
  - Given x_i and predicted value $\hat{y}_i$, how wide must an interval be to include a randomly drawn observation
  - Our uncertainty about $\hat{y}_i$ as an estimate of a particular observation
  - Intuition: PI must be wider then CI because y_i is less certain than $\hat{y}_i$

Confidence Interval for estimating $E[y|x]$

Please remember: The format of these CIs is symmetric, like $[\hat{y} - something, \hat{y} + something]$.

- A 95% "confidence interval" includes the true $E[y_i|x_0]$ with probability 0.95

$$Confidence\ Interval = \hat{y}_0 \pm \ t \times \widehat{\sigma_e} \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]^{1/2} \tag{4}$$

- To work with that, select some "example" values of the predictor. Call them $x_0 \in \{0, 2, 4, 6\}$, for example
- $\hat{y}_o$ is the predicted value for a particular x, $\hat{\beta}_0 + \hat{\beta}_1 x_0$.
- For 95% CI, set t 1.98
- $\widehat{\sigma_e}$ is "RMSE," the "estimated standard deviation of the error term"

Use predictOMatic to see some for some values of X

```
predictOMatic(incedmod1, predVals = list(education = "quantile"), n
     = 10, interval = "confidence")
```

```
   education       fit       lwr       upr
1      6.380  2880.840  1586.748  4174.933
2      7.522  3907.285  2846.511  4968.058
3      8.128  4451.965  3502.770  5401.160
4      8.766  5025.408  4179.669  5871.147
5      9.708  5872.089  5140.216  6603.963
6     10.540  6619.902  5933.801  7306.003
7     11.172  7187.951  6494.982  7880.921
8     12.209  8120.020  7341.762  8898.279
9     13.620  9388.245  8390.331 10386.160
10    14.703 10361.660  9150.520 11572.799
11    15.970 11500.456 10014.844 12986.067
```

- Steps across 10 values of education, showing fitted (predicted) values and lower and upper 95% CI
- Will Plot below. Is “hour glass shaped”.

Prediction Interval

- A 95% “prediction Interval” includes a randomly drawn outcomes y_i with probability 0.95

$$Prediction\ Interval = \hat{y}_0 \pm t \times \widehat{\sigma_e}\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]^{1/2} \tag{5}$$

- Notice that the *something* in the PI is equal to the *something* in the CI with an additional amount that depends directly on the standard deviation of the error term $\widehat{\sigma_e}$
- This interval is larger because we think of “the line bouncing about”, and the random draws are added on after that.

Derivation of CI and PI is presented in Appendix 2

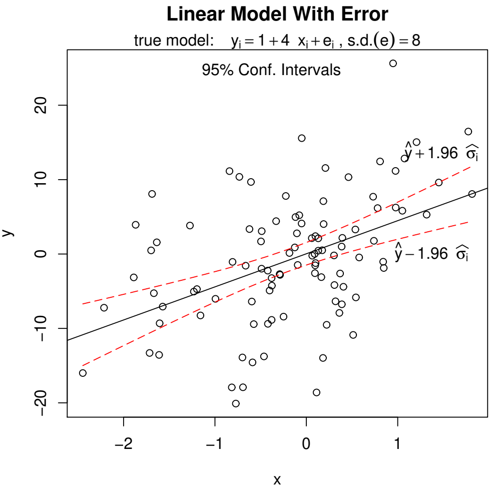
Use predictOMatic to see some for some values of X

```
predictOMatic(incedmod1, predVals = list(education = "quantile"), n
    = 10, interval = "prediction")
```

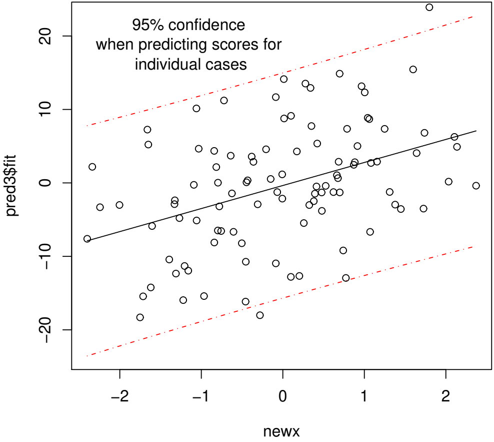
```
   education       fit        lwr      upr
1      6.380  2880.840 -4150.1994  9911.88
2      7.522  3907.285 -3084.5739 10899.14
3      8.128  4451.965 -2523.8370 11427.77
4      8.766  5025.408 -1937.0716 11987.89
5      9.708  5872.089 -1077.4777 12821.66
6     10.540  6619.902  -324.9942 13564.80
7     11.172  7187.951   242.3737 14133.53
8     12.209  8120.020  1165.4153 15074.63
9     13.620  9388.245  2405.6470 16370.84
10    14.703 10361.660  3345.4139 17377.91
11    15.970 11500.456  4431.6587 18569.25
```

Steps across 10 values of education, showing fitted (predicted) values and lower and upper 95% PI

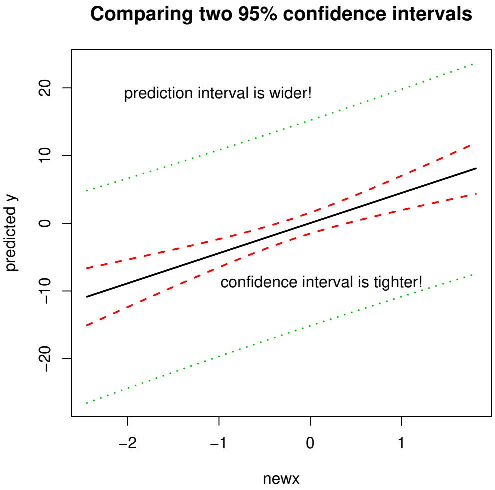
The Hour Glass Shaped Confidence Interval



Prediction Interval also Hour-Glass Shaped, but Curvature Gradual



Compare PI and CI



Too Many Intervals Floating About?

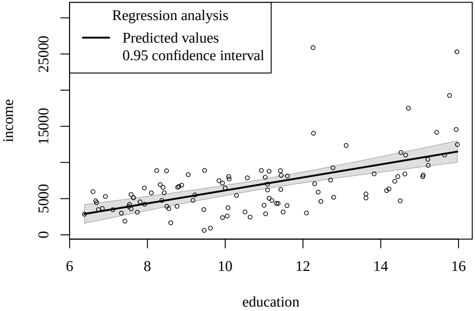
- I am always surprised that students don't see a difference between these confidence intervals.
- The CI around $\hat{\beta}_1$ says we believe the true β_1 lies in here: $[\hat{\beta}_1 - \textit{something}, \hat{\beta}_1 + \textit{something}]$
- We use $\hat{\beta}_0$ and $\hat{\beta}_1$ and the predicted value $\hat{y}_i$. The value $E[y_i|x_i] = \beta_0 + \beta_1 x_i$ is the "true" expected value of y_i, what would happen if there were no random error. $E[y_i|x_i]$ is likely in $[\hat{y}_i - \textit{something else}, \hat{y}_i + \textit{something else}]$. The *something else* includes our uncertainty about $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The prediction interval is a statement that, for a particular x_i, the observed $y_i$would be in: $[\hat{y}_i - \textit{something bigger}, \hat{y}_i + \textit{something bigger}]$. That's bigger because it includes uncertainty about $\hat{\beta}_0$, $\hat{\beta}_1$ and $\widehat{\sigma_e}$.

Canadian Prestige: plotSlopes illustrates predictOMatic

- Receive a fitted regression, plot one predictor and the desired interval

```
plotSlopes(incedmod1, plotx =
    "education", interval = "
    confidence")
```

- Argument “plotx”: name of predictor on x axis
- Run example(plotSlopes) to get the big idea

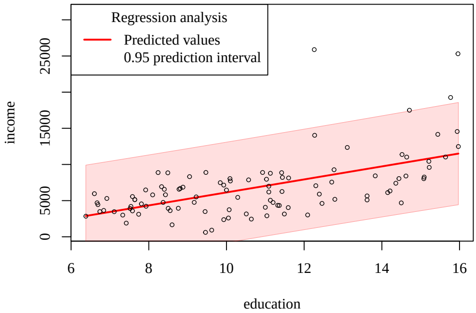


Prediction Intervals are Wider

- Receive a fitted regression, plot one predictor and the desired interval

```
plotSlopes(incedmod1, plotx =
    "education", interval = "
    prediction", col = "red")
```

- Argument "plotx": name of predictor on x axis
- Run example(plotSlopes) to get the big idea



Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

Here are the main points

Re-scaling predictors has predictable effects on the intercept and slope.

- Multiply x_i by a factor k implies
  - new $\hat{\beta}_1$ will be $1/k$ times old $\hat{\beta}_1$
  - new $std.err(\widehat{\beta_1})$ will be $1/k$ times old $std.err.(\hat{\beta}_1)$
  - $\hat{t}$ ratio $\frac{\hat{\beta}_1}{std.err.(\hat{\beta}_1)}$ is thus UNCHANGED.
- Add k to x_i,
  - new $\hat{\beta}_1$ exactly same as old $\hat{\beta}_1$. Same standard error, same $\hat{t}$
  - new intercept estimate $\hat{\beta}_0$ and its standard error will be changed
  - t statistic will be changed for $\hat{\beta}_0$.

Your Data is in Pesos?

- Problem. Income is a predictor, but it is coded in a small denomination
- Example: Chile data on status quo support

	M1
	Estimate
	(S.E.)
(Intercept)	-0.042
	(0.026)
income	0.000*
	(0.000)
N	2591
RMSE	1.001
R^2	0.002

$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

Its Not Really Zero

It's 0.000000098

```
Call:
lm(formula = statusquo ~ income, data = Chile)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71792 -1.00329 -0.06181  0.97407  1.74119

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.235e-02  2.588e-02  -1.636   0.1019
income       9.809e-07  4.971e-07   1.973   0.0486 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 2589 degrees of freedom
  (109 observations deleted due to missingness)
Multiple R^2:  0.001502, Adjusted R^2:  0.001116
F-statistic: 3.894 on 1 and 2589 DF,  p-value: 0.04858
```

Your Data is in 1000000's of Pesos

- Solution. Divide Income by 1,000,000

	M1 Estimate (S.E.)
(Intercept)	-0.042
	(0.026)
income2	0.981*
	(0.497)
N	2591
RMSE	1.001
R^2	0.002

$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

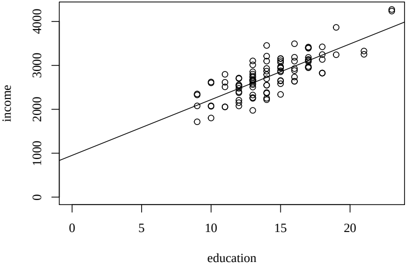
- Looks as if we had taken original $\hat{\beta}$ and $std.err.(\hat{\beta})$ and chopped off 7 0's at the beginning of the fraction.
- Same
    - t-ratio
    - R^2
    - intercept

Here's the Full Printout, Just For the Record

```
Call:
lm(formula = statusquo ~ income2, data = Chile)

Residuals:
     Min        1Q    Median        3Q       Max
-1.71792  -1.00329  -0.06181   0.97407   1.74119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04235    0.02588  -1.636   0.1019
income2      0.98091    0.49712   1.973   0.0486 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 2589 degrees of freedom
  (109 observations deleted due to missingness)
Multiple R^2:  0.001502, Adjusted R^2:  0.001116
F-statistic: 3.894 on 1 and 2589 DF,  p-value: 0.04858
```

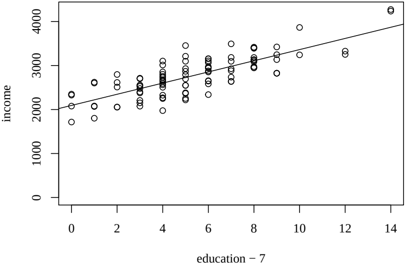
Problem: Estimated Intercept Seems Meaningless

- The Y axis is placed at $x_i = 0$, but there are no observations near there
- Your x data puts the "data cloud" out in the "middle of nowhere"
- Example: Predict income from education. Nobody has education equal to 0.
- Seems silly to interpret the intercept in this case.
- You'd rather discuss the lowest observed education level.



Solution: Push the y axis to the Edge of the Data Cloud

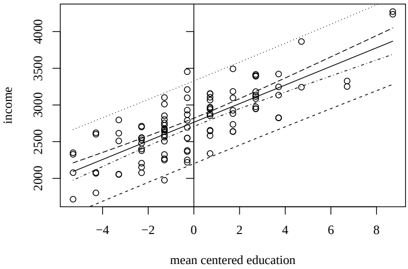
- Subtract 8 or 10 (or whatever you like) from x
- The "y axis" will "move" 8 or 10 (or whatever) to the right.
- Subtract smallest value of x, then you have the "lowest educated person" as a baseline



- Education = Education - 7

Mean-Center x_i: Then the y axis is at the mean of x_i.

- Rescale $x_i = x_i - \bar{x}$, where $\bar{x}$ is the sample mean of x
- Pushes "y axis" into middle of data.
- Benefit of being in the middle! Remember the "hourglass" shape of the CI?



- "mean centered education" = Education - mean(education)

Rescaling By Subtraction (or Addition)...

- leaves the slope estimate EXACTLY the same (1 unit increase in x_i causes a $\hat{\beta}_1$ change in y_i)
- changes the intercept estimate
- Changes the t-ratio

	M1
	Estimate
	(S.E.)
(Intercept)	951.121***
	(143.784)
oldx	126.882***
	(9.866)
N	100
RMSE	282.609
R^2	0.628

$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

	M1
	Estimate
	(S.E.)
(Intercept)	2764.258***
	(28.261)
x	126.882***
	(9.866)
N	100
RMSE	282.609
R^2	0.628

$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

Rescale by Standardizing

- Recall, to Standardize means subtract sample mean and divide by sample standard deviation

$$x_i^{st} = \frac{x_i - \bar{x}}{\widehat{Std.Dev.}[x]} \qquad y_i^{st} = \frac{y_i - \bar{y}}{\widehat{Std.Dev.}[y]} \tag{6}$$

- If we knew the “true” standard deviations, we could call these “Z scores”, Z_{x_i} or Z_{y_i}.
- But we don’t know true standard deviations, so these are just “standardized variables”.
- For standardized data, ALWAYS,
  - mean equals 0: $\widehat{E[x^{st}]} = 0$
  - variance=standard deviation=1: $\widehat{Var[x^{st}]} = 1$

Note How this Changes Parameter Estimates

Recall the OLS estimator for the slope is

$$\hat{\beta}_1^{OLS} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{7}$$

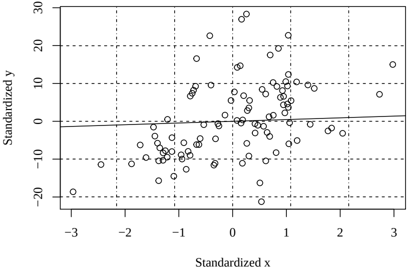
- Insert the standardized variables x_i^{st} and y_i^{st} in place of x_i and y_i, and what do you get?
- Lets call this "standardized regression coefficient," $\hat{\beta}^{st}$. Note how the math simplifies,
$$\hat{\beta}_1^{st} = \sum x_i^{st} \cdot y_i^{st} \tag{8}$$
- And the intercept "disappears", it becomes 0, denominator becomes 1.

The Standardized Regression Line

- The predicted value is $\hat{y}_i^{st} = \hat{\beta}_1^{st} x_i^{st}$
- The "units of measurement" become standard deviation units
- Dotted lines mark "one standard deviation" units

	M1	
	Estimate	(S.E.)
1	0.455***	(0.089)
N	100	
RMSE	0.890	
R^2	0.207	

$*p \leq 0.05 ** p \leq 0.01 *** p \leq 0.001$



"An increase in x of one of its standard deviations causes a $\hat{\beta}_1$ standard deviation increase in y"

Interesting Tidbits

- If there is just one independent variable, the R^2 reported with regression equals the r squared.
- If both the indep. and dependent variables are standardized, the slope coefficient of the fitted model equals the Pearson r.
- In rockchalk package, there are functions standardize() and meanCenter() that can make this more convenient.

Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. **Appendix 1. Variance of $\hat{\beta}$**
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

Fundamentals about Variance.

Recall the rules of working with Variance. Suppose k and m are constants and x_i and y_i are variables.

1. $V(k \cdot x_i) = k^2 V(x_i)$
2. $V(k \cdot x_i + m \cdot y_i) = k^2 V(x_i) + m^2 V(y_i) + 2 \cdot k \cdot m \cdot Cov(x_i, y_i)$

V is variance
Cov is covariance

Derive $Var(\hat{b})$

- Make our lives simpler by beginning with the OLS estimator for "data in deviations" form:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{9}$$

  That means we have pre-scaled $x_i = x_i observed - \bar{x}$ and $y_i = y_i observed - \bar{y}$. That leaves $\hat{\beta}_1$ and the $Var[\hat{\beta}_1]$ unchanged, but math is easier.

- Start by trying to figure out the "true variance" of $\hat{\beta}_1$. Apply the Var() operator to both sides of (9)

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum x_i y_i}{\sum x_i^2}\right) \tag{10}$$

Derive $Var(\hat{b})$ ...

- The values of x_i are not thought of as random variables. Instead, they are variables that are "fixed" attributes of the observations. (If you want x_i to be a random variable, you can do that, but the math is slightly different).
- With fixed x_i, the sum of x_i^2, $\sum x_i^2$, is a constant, "just some number."
  Applying Variance rule 1, we take $1/\sum x_i^2$ outside the parentheses in 10.
  $$Var(\hat{\beta}_1) = \left(\frac{1}{\sum x_i^2}\right)^2 Var\left(\sum x_i y_i\right) \tag{11}$$
- Replace y_i by $\beta_1 x_i + e_i$ (recall β_0=0 with deviations form data)
  $$Var(\hat{\beta}_1) = \left(\frac{1}{\sum x_i^2}\right)^2 Var\left(\beta_1 \sum x_i^2 + \sum x_i e_i\right) \tag{12}$$

Derive $Var(\hat{b})$ ...

- Apply the Variance rule 2:

$$\begin{aligned} Var(\hat{\beta}) &= \left(\frac{1}{\sum x_i^2}\right)^2 \left(Var\left(\beta_1 \sum x_i^2\right) + Var\left(\sum x_i e_i\right)\right. \qquad (13) \\ &\left. +2Cov\left(\beta_1 \cdot \sum x_i^2, \sum x_i e_i\right)\right) \end{aligned}$$

- Expression (13) is our focal point. We want to simplify that.
  - Use a sneaky trick to make $Var(\beta_1 \sum x_i^2)$ go away. Obviously, that is equal to:

    $$\beta_1^2 Var(\sum x_i^2).$$

    Now, here is the trick. Observe:

    $$Var(\sum x_i^2) = 0.$$

    How? Recall, Var() refers to variance across experiments. Since we are thinking of x_i as "fixed", then across experiments there is no variation in the sum of squared x's. That sum of squared x's is a constant. So its variance is 0.

Derive $Var(\hat{b})$ ...

- Make $2Cov\left(\beta_1 \cdot \sum x_i^2, \sum x_i e_i\right)$ disappear. If k is any constant, and x is a variable, then $Cov(k,x) = 0$. Covariance between a constant and a variable equals 0.
- Thus (13) reduces to:

$$Var(\hat{\beta}_1) = \left(\frac{1}{\sum x_i^2}\right)^2 Var\left(\sum x_i e_i\right) \tag{14}$$

Be verbose about it. There's a constant times the variance of a sum:

$$Var(\hat{\beta}_1) = \left(\frac{1}{\sum x_i^2}\right)^2 Var\left(x_1 e_1 + x_2 e_2 + x_3 e_3 + ... x_n e_n\right) \tag{15}$$

Now apply rule #2 about variance. After a little thought, one must realize that $Cov(x_i e_i, x_j e_j) = 0$ because all the error terms are 'stochastically independent' of each other and the x's are fixed. (If you don't assume the x's are fixed, you have to assume instead that the x's are uncorrelated with the e's).

Derive $Var(\hat{b})$ ...

- After applying rule #2 and throwing away all those Covariances (which are 0), we find:

$$Var(\hat{\beta}_1) = \left(\frac{1}{\sum x_i^2}\right)^2 \left(x_1^2 Var(e_1) + x_2^2 Var(e_2) + ... + x_n^2 Var(e_n)\right) \tag{16}$$

Since we assumed above that $Var(e_i) = \sigma^2$, then this becomes:

$$Var(\hat{\beta}) = \left(\frac{1}{\sum x_i^2}\right)^2 \left(x_1^2\sigma^2 + x_2^2\sigma^2 + ... + x_n^2\sigma^2\right) \tag{17}$$

$$Var(\hat{\beta}) = \left(\frac{1}{\sum x_i^2}\right)^2 \left(\sum x_i^2\sigma^2\right) \tag{18}$$

$$Var(\hat{\beta}) = \left(\frac{1}{\sum x_i^2}\right)^2 \left(\sum x_i^2\right) * \sigma^2 \tag{19}$$

Derive $Var(\hat{b})$ ...

$$Var(\hat{\beta}) = \left(\frac{1}{\sum x_i^2}\right) * \sigma^2 \tag{20}$$

Whew. As Batman says, “my work is done.”

Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

Show My Work: Derive Confidence Interval

- We want to fill in “something” in this expression:

$$Pr[\hat{y}_0 - something \ \leq E[y_o|x_o] \leq \hat{y}_0 + something] = 0.95 \qquad (21)$$

- “Something” depends on the sampling distribution of $\hat{y}_0 - E[y_0|x_0]$.

$$\begin{aligned} Var(\hat{y}_0 - E[y_0|x_0]) &= Var[\hat{\beta}_o + \hat{\beta}_1 x_0 - E[y_0|x_0]) \qquad (22) \\ &= Var[\hat{\beta}_0] + x_0^2 Var[\hat{\beta}_1] + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) \end{aligned}$$

- Put in the estimated variances and covariances, and rearrange, and we end up with

$$Var(\hat{y}_0 - E[y_0|x_0]) = \hat{\sigma}^2_{CI} = \sigma^2_e \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \qquad (23)$$

- Replace the unknown σ_e^2 with the estimated *MSE*
- The square root of that is the “standard error” *SE* that can be used to create the CI:

$$\hat{y}_0 \pm \ t_{\alpha/2, df} \times \hat{\sigma}_{CI} \qquad (24)$$

Show My Work: Derive the Prediction Interval

- Go back to the basics of Confidence Intervals. We want to fill in "something":

$$Pr[\hat{y}_0 - something \ \leq y_o \leq \hat{y}_0 + something] = 0.95 \qquad (25)$$

  - y_o is the score that "will be observed" in a case.
  - $\hat{y}_o$ is the predicted value for that case (point on regression line)

- "Something" ends up being a standard error for many types of estimators (including regression coefficients), so we need the sampling distribution of $\hat{y}_0 - y$.

$$\begin{aligned} Var[\hat{y}_0 - y_0] &= Var[\hat{\beta}_o + \hat{\beta}_1 x_0 - y_0] \qquad (26) \\ &= Var[\hat{\beta}_0] + x_0^2 Var[\hat{\beta}_1] + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) + Var[e_i] \end{aligned}$$

- Put in the estimated variances and covariances, and rearrange, and we end up with

Show My Work: Derive the Prediction Interval ...

$$Var(\hat{y}_0 - y_0) = \sigma_e^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \qquad (27)$$

- Replace the unknown σ_e^2 with the estimated *MSE*
- The square root of that is the "standard error" *SE*.

Outline

1. $\hat{\beta}$ Uncertainty
   - Visualize Uncertainty
2. Sampling Distribution $\hat{\beta}$
3. t-test Hypotheses about β_j
4. Confidence Interval of $\hat{\beta}_j$
5. Afterthought: Simulated Distribution of R^2
6. Prediction CI
7. Re-scale variables
   - Multiply x_i to Re-Scale It
   - Subtract from x_i to Make Intercept Easier to Interpret
   - Standardize Variables
8. Appendix 1. Variance of $\hat{\beta}$
9. Appendix 2: Proofs of CI and PI
10. Practice Problems

Problems

1. Run any regression and print out a summary of the estimates. Circle and label the following elements in the output:
    1. point estimate of the intercept
    2. point estimate of the slope
    3. estimate of the standard deviation of the estimated intercept
    4. estimate of the standard deviation of the estimated slope
    5. estimate of the standard deviation of the error term
    6. estimate of the coefficient of determination

    If you can't find any data to experiment with, I suggest one of these (in R's base distribution):

    ```
     library(datasets)
     library(help=datasets)
    ?Orange
    m1 <- lm( circumference ~ age, data = Orange)
    summary(m1)
    ?cars
    m2 <- lm(dist ~ speed, data=cars)
    summary(m2)
    ```

Problems ...

2. Write down the formula for the slope estimate, $\hat{b}_1$. Suppose your assignment is to make sure that value is as large as possible.
    1. Would you rather have the variable x_i scattered far-and-wide across the horizontal axis?
    2. If you could get all of your x_i observations at a single value, wouldn't that help you pinpoint the predicted value of y_i, and hence make for a better slope estimate?
    3. Would you rather that the true variance of the error is really small, or really big?
3. Did you ever want to make up your own data? Here's the chance. I want you to see the effect of changes in the standard deviation of the error term. Run this:

```r
stde <- 1
dat <- data.frame(x=rpois(500, lambda=200))
dat$y <- 3 + 0.08 * dat$x + stde * rnorm(500)
m1 <- lm(y ~ x, data=dat); summary(m1)
plot(y ~ x, data=dat); abline(m1)
```

Problems ...

We want to see what happens as stde is made larger, so re-set that to 2, and create $y2$ from it, and run again:

```
stde <- 2
dat$y2 <- 3 + 0.08 * dat$x + stde * rnorm(500)
m2 <- lm(y2 ~ x, data=dat); summary(m2)
plot(y2 ~ x, data=dat); abline(m2)
```

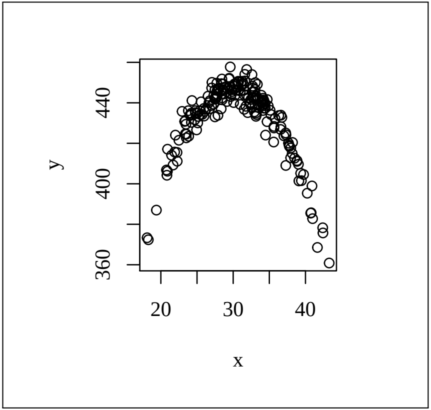
Hopefully 2 examples are enough to give you the idea. Adjust stde again, create y3, run m3, and compare.

Here are the questions:

1. What changes in the regression estimates result from tuning up stde?
2. What parameter estimate in the output is supposed to represent the variable "stde"?
3. How well does the OLS procedure do at estimating stde?

Problems ...

4. Here's a scatterplot that I found. When I fit the regression line, I can't understand the estimates. The numbers seem to say there is no relationship, but it is plain to the eye that there is! How would you explain it? I think you are more likely to get full credit if you sketch the OLS fitted line on the scatter.



	M1
	Estimate
	(S.E.)
(Intercept)	455.787***
	(7.333)
x	-0.741**
	(0.235)
N	200
RMSE	17.686
R^2	0.048

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

Problems ...

5. Here is a fitted OLS regression model with standard errors in parentheses

$$\widehat{autism_i} = 8.0 + 6.3 \cdot iron_i \quad (28)$$

$$(1.0) \quad (2.1) \quad (29)$$

The RMSE (or "sigma" or "residual standard error" is 10.0), the $R^2 = 0.32$, and the sample size is 1000. The $autism_i$ score represents a child's placement on the 100 point autism spectrum disorder scale and $iron_i$ is the number of iron molecules per billion in the child's blood. If the $autism_i$ score is greater than 20, the child is deemed to be in the "moderate autism" range, and if it is greater than 50, the child is deemed to be in the "severe autism" range. The $iron_i$ variable ranges from 1 to 6 in this sample (so the lowest observed score represents 1 part per billion).

1. What do the results allow us to conclude about the impact of exposure to iron on the rate of autism? I mean "interpret the slope and intercept."
2. Conduct a "null hypothesis test" for the estimated slope. That means go through the "four steps" outlined previously in this lecture.

Problems ...

3. Construct a "confidence interval" for the estimate of the slope and write a brief discussion of what the results indicate to you.
4. Create a plot representing the predicted score on the autism scale for iron exposure rates from 1 to 10.
5. What meaning does the estimated intercept have in this case? Where does it appear in the plot you created?
6. You don't have all of the information you need to construct the "hour glass" shaped confidence and prediction intervals for that plot. But, if I told you the mean of iron exposure is 3, then you can sketch the general shape of those intervals. So draw them in and label them (its vital to know which one is outside the other).
7. Refer to your sketch of the confidence interval. Suppose one family lives in a house with lead paint equal to 3, while another has a score of 6. For which family are we most confident about our estimate $\widehat{autism}_i$? Explain.

Problems ...

8. Suppose we measured iron value for another family at 10.0. What would the linear model lead us to predict about their prospects for autism (moderate or severe)? Note there are 2 problems here. One is that the value of $iron_i$ is far from the mean. Another is that our observations range from 0 to 6, and so in a sense 10 is "out of the range" of our experience. .
9. Suppose you just found out that your research assistant does not understand numbers. He uses the word "billion" when he really means "million". What changes will you need to make in your handling of this data and the result?