Week 4
Bivariate Regression 2

Applied Statistical Analysis/Quantitative Methods I

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2024

Class Business

- Problem set #2 is out right now!
  - Due before **Monday October 14**
  - Answer key for problem set #1 is up on GitHub
- Questions from last time?

Roadmap through Stats Land

Where we've been:

- We're learning how to make inferences about a population from a sample
- How to determine if two samples are different or independent (diff-in-means, contingency tables)
- Last week: We learned about bivariate correlation and regression (assumptions, estimation)

Outline for today:

- Inference about...
  - Correlation
  - Parameters
  - Prediction

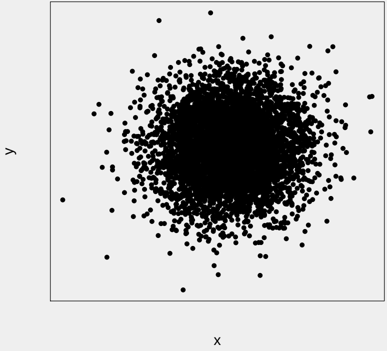
Inference: Correlation (r vs. ρ)

- So far, we've only established some correlation coefficient r from our sample
- We should ask: is there any correlation in the population, or if the population has no correlation, is this sample r just the result of normal sample variability?
- How can we check? Conduct a hypothesis test!
- What are we actually testing?
  - Is there any (linear for now) relationship between the two variables
- More formally, H_0: $\rho = 0$ and H_a: $\rho \neq 0$
- And you can even calculate a CI for ρ!

Inference for "true" population correlation ρ

- Pearson correlation assumes that both x and y are approximately normally distributed
- Note: A significant result does not imply a strong linear relationship
- And a strong linear relationship does not imply statistical reliability

Is there an association?



- Not really, so we cannot reject the null hypothesis that $H_0 : \rho = 0$ because $p < 0.05$, so we can say the result is "statistically reliable"
- However, the correlation is very weak ($r = 0.04$)
- As $n \uparrow$, correlation can be statistically different from 0 even though linear association is weak

Steps: Inference for ρ

What do we need?

1. The 2 quantitative variables (Y, X)
2. Hypothesis test (H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$)
3. α = 95% for true population correlation ρ
4. Estimate of sample correlation r
5. Test statistic = t = $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
6. P-value

By hand: Inference for ρ

Pearson correlation for a sample:

$$r_{xy} = \frac{\text{covariance}_{xy}}{\text{SD}_x \text{SD}_y}$$

so...

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

OR

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}}$$

By hand: Inference for ρ

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

```r
# bivariate correlation by hand
regressMat <- matrix(c(3.8, 3.5, 3.0, 3.3, 3.5, 4.0, 2.8,
    2.3, 2.4, 1.8, 2.7, 2.7), nrow=6, byrow = T)
```

```r
r <- cov(regressMat)[1,2]/(sd(regressMat[,1]) * sd(
    regressMat[,2]))
```

```
[1] 0.8708898
```

By hand: Inference for ρ

Now, that we have r, let's calculate our TS

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

```r
n <- dim(regressMat)[1]

# test statistic
t_stat <- (r*sqrt(n-2))/sqrt(1-r^2)
```

```
[1] 3.54396
```

```r
# get p-value
2*pt(t_stat, n-2, lower.tail=FALSE)
```

```
[1] 0.02392807
```

Easy way: Inference for ρ

```r
# check correlation coefficient (r)
cor(regressMat[,1], regressMat[,2], method="pearson")
```

```
[1] 0.87
```

```r
# check if null p = 0
cor.test(regressMat[,1], regressMat[,2])
```

```
t = 3.544, df = 4, p-value = 0.02393
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2023348 0.9857455
sample estimates:
cor
0.8708898
```

Inference: Parameters in regression

Remember from last time, we want to minimize our squared residuals

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- We want to do this to estimate μ (pop. mean) from Y (data)

$$\mu \equiv \text{arg min}_\mu \sum_{i=1}^{N} (y_i - \mu)^2 = \frac{\sum y_i}{N} = \bar{Y}$$

- Since least squares (LS) traces a conditional mean (of a normal distribution), makes sense to minimize same quadratic "loss function":

$$\text{Sum of squared errors} = S = \sum (Y_i - \alpha - \beta X_i)^2$$

Review: LS as point estimates for parameters

Differentiate S w.r.t. α and β and set derivatives equal to "0",

$$\frac{\partial S}{\partial \alpha} = \frac{\partial \sum (Y_i - \alpha - \beta X_i^2)}{\partial \alpha} = 0$$

$$\frac{\partial S}{\partial \beta} = \frac{\partial \sum (Y_i - \alpha - \beta X_i)^2}{\partial \beta} = 0$$

to arrive at "LS Normal Equations" (2 equations with 2 unknowns):

$$n\hat{\alpha} + \hat{\beta} \sum X_i = \sum Y_i$$

$$\hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 = \sum X_i Y_i$$

▸ Explaination of derivation

OLS estimators of α and β:

$$\hat{\alpha} = \frac{\sum(Y_i - \hat{\beta}\sum X_i)}{n} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{n\sum(X_iY_i) - \sum X_i \sum Y_i}{n\sum(X_i^2) - \sum(X_i)^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

where $x_i = X_i - \bar{X}$ (mean deviate form)

Are these the best estimators? Gauss-Markov

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- I said these estimators were most efficient, unbiased estimators, why?
- **Gauss-Markov theorem:** Assuming errors are normal, the LS estimators are most efficient among all unbiased estimators, not just among linear estimators
  - I'll prove this when we go over inference in multiple regression (same idea, just easier with matrix notation)

Uncertainty of $\hat{\beta}$: Sampling Distributions of $\hat{\beta}$

Under the assumptions we've outlined, the sampling distributions of the least squares estimates are themselves normally distributed, remember:

- Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a sample, estimators themselves are random variables with a probability distribution
- So, we make an assumption that $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$ (they're unbiased estimates of those parameters)

$$\hat{\beta}_0 \sim N\Big(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})\Big)$$

$$\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\Big)$$

Uncertainty of $\hat{\beta}$: Standard error

- Don't worry too much about how to derive the variance of the sampling distributions, but note that...
- The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on unknown parameter σ^2 (joint variance)
- So, we need to estimate

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

- Remember, an estimate of the standard deviation is called the standard error (SE) of the estimate

$$se_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$se_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

On further inspection: Variability of $\hat{\beta}_0$

$$se_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

- Standard error of $\hat{\beta}_0$ depends on
  - Estimated standard deviation $\hat{\sigma}$
  - Sample size n
  - How closely clustered the x's are
  - Center of the x values
    - If x's are centered at zero, $\rightarrow se_{\hat{\beta}_0}$ = standard error of $\bar{X}$ (which is the smallest possible $se_{\hat{\beta}_0}$)

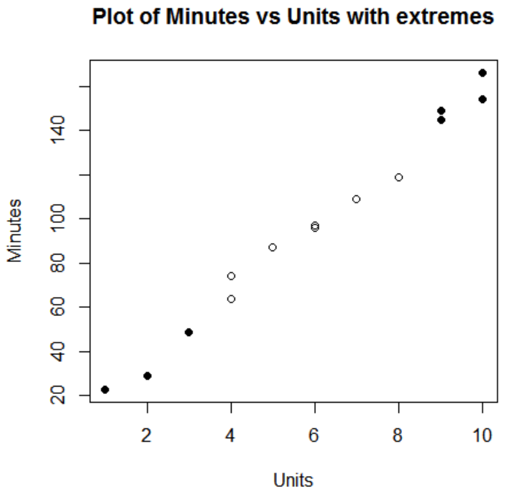
On further inspection: Variability of $\hat{\beta}_1$

$$se_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- Standard error of $\hat{\beta}_1$ could be small
  - When the estimated standard deviation $\hat{\sigma}$ is small
  - When n is large
  - When x values are spread out (I'll explain in next slide)

What data we have impacts variability of $\hat{\beta}_1$

Suppose we only had a subset of 7 extreme observations, 3 at the top and 4 at the bottom



Ex: What data we have impacts Variability of $\hat{\beta}_1$

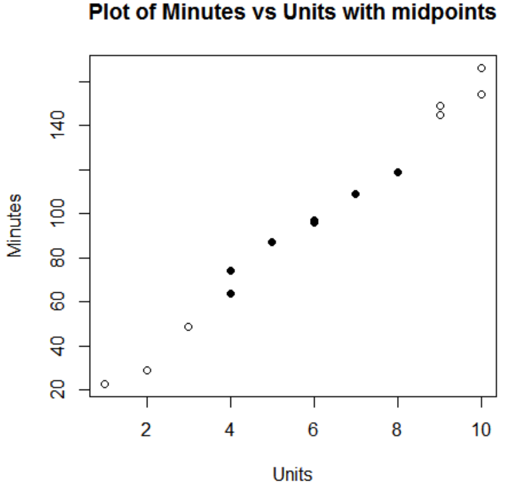
Let's run a regression

```
> lm1<-lm(subset$Minutes~subset$Units)
> summary(lm1)

Call:
lm(formula = subset$Minutes ~ subset$Units)

Residuals:
      1       2       3       4       5       6       7 
 4.8362 -5.0517 -0.9397  3.7328 -0.2672 -7.1552  4.8448 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.2759     3.8679   0.588    0.582    
subset$Units  15.8879     0.5278  30.105 7.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.262 on 5 degrees of freedom
Multiple R-squared:  0.9945,    Adjusted R-squared:  0.9934 
F-statistic: 906.3 on 1 and 5 DF,  p-value: 7.586e-07
```

$\hat{\beta}_1 = 15.89$, and $se_{\hat{\beta}_1} = 0.53$

Ex: What data we have impacts Variability of $\hat{\beta}_1$

Now, suppose we only had a subset of 7 points in the middle



Do you think the results will change? How?

Ex: What data we have impacts Variability of $\hat{\beta}_1$

Let's re-run that regression

```
> lm2<-lm(midset$Minutes~midset$Units)
> summary(lm2)

Call:
lm(formula = midset$Minutes ~ midset$Units)

Residuals:
      1       2       3       4       5       6       7 
-6.7660  3.2340  3.6809  0.1277  1.1277  0.5745 -1.9787 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    20.553      6.227    3.30   0.0215 *  
midset$Units   12.553      1.059   11.85 7.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.881 on 5 degrees of freedom
Multiple R-squared:  0.9656,	Adjusted R-squared:  0.9588 
F-statistic: 140.5 on 1 and 5 DF,  p-value: 7.529e-05
```

Now, $\hat{\beta}_1 = 12.55$, and $se_{\hat{\beta}_1} = 1.06$

Take-away: Both subsets are from the same set of data (with the same underlying relationship), but when x values we observe are more spread out, we get a better estimate of the slope

We can use the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ to generate hypothesis tests:

- **Hypotheses:**
  - $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$
  - $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
- How do we do this?

$$\frac{\hat{\beta}_0 - \beta_0}{se_{\hat{\beta}_0}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}} \sim t_{n-2}$$

$n - 2$ = df (# of observations - # of estimated coefficients)

CIRCLE BACK: CONFIDENCE INTERVALS OF β_1

- We also use the standard error of $\hat{\beta}_1$ to form a CI around the "true" slope β_1
- A $(1-\alpha)$% confidence interval for the true slope β_1:

$$\beta_1 \pm t_{\alpha/2} se_{\hat{\beta}_1}$$

df of t distribution is $n-2$

- If CI includes the value 0, we can't reject the null hypothesis of $H_0 : \beta_1 = 0$ at the α significance level
- When providing a point estimate of β_1, report variability on your estimate with a CI or SE

Ex: Interpreting R output, β_0

```
> lm<-lm(computer$Minutes~computer$Units)
> summary(lm)

Call:
lm(formula = computer$Minutes ~ computer$Units)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2318 -3.3415 -0.7143  4.7769  7.8033 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.162      3.355    1.24    0.239    
computer$Units   15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874,	Adjusted R-squared:  0.9864 
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

Ex: Hypothesis test of β_0 from R output

- **Hypotheses:** $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$
- **Test statistic:** $t = \dfrac{\hat{\beta}_0 - 0}{se_{\hat{\beta}_0}} = \dfrac{4.162 - 0}{3.355} = 1.24$
- **p-value:** two-tailed probability from t distribution with $df = n - 2$

  p-value$= 2 \times Pr(t_{12} > |1.24|) = 0.239$
- or we can create a **CI for** β_0: $\hat{\beta}_0 \pm t \times se$

  95% CI for β_0: $4.162 \pm 2.18 \times 3.355$ = (-3.19, 11.47)

Ex: Conclusions about β_0

- The 95% CI for β_0 is (-3.19, 11.47) - therefore it is plausible that the length of a service call is 0 when there is no component must be repaired
- At $\alpha = 0.05$, fail to reject H_0, there is not sufficient evidence to conclude that the intercept differs significantly from zero
- Inference for the intercept is not always of interest - usually the focus is on the slope
- *We can evaluate this issue sensibly only if we have collected data around the value $X = 0$*

Ex: Interpreting R output, β_1

```
> lm<-lm(computer$Minutes~computer$Units)
> summary(lm)

Call:
lm(formula = computer$Minutes ~ computer$Units)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2318 -3.3415 -0.7143  4.7769  7.8033

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.162      3.355    1.24    0.239
computer$Units   15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874,    Adjusted R-squared:  0.9864
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

Ex: Hypothesis test of β_1 from R output

- **Hypotheses:**
  - $H_0 : \beta_1 = 0$
  - $H_a : \beta_1 \neq 0$
- **Test statistic:** $t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}} = \frac{15.509 - 0}{0.505} = 30.71$
- **p-value:** two-tailed probability from t distribution

  $df = n - 2$

  p-value$= 2 \times Pr(t_{12} > |30.71|) \approx 0$
- **Confidence interval for β_1:** $\hat{\beta}_1 \pm t_{score} \times se$

  95% CI for β_1: $15.509 \pm 2.180 \times 0.505$ = (14.41, 16.61)

Ex: Conclusions about β_1

- 95% CI for β_1 is (14.41, 16.61)
  - *Positive* association between the length of a service call and the number of components that need to be fixed
- At the 95% confidence level, for each one additional component that has to be repaired, the length of a service increases by as few as $\approx$ 14.4 minutes or as many as 16.6 minutes
- Reject H_0, and conclude that β_1 is statistically differntiable from zero
  - "Statistically reliable" and positive association between the length of a service call and the number of components
- β_1 (slope) is usually quantity of interest in linear regression

Ex: Standard Errors and inference by hand

First, create the data and run a regression:

```
# create linearly dependent data
X <- runif(100, 0, 1)
Y <- 2 + X*1.5 + rnorm(100, 0, 1)
reg_DF <- as.data.frame(cbind(X, Y))
# calculate estimates of:
# beta_1
beta <- sum((reg_DF$Y - mean(reg_DF$Y)) * (reg_DF$X - mean(
    reg_DF$X)))/
  sum((reg_DF$X - mean(reg_DF$X))^2)
# beta_0
alpha <- mean(reg_DF$Y) - beta*mean(reg_DF$X)
```

```
> beta
[1] 1.269
> alpha
[1] 1.956
```

Check that our estimates are correct:

```r
# check regression
reg1 <- lm(Y~X, data=reg_DF)
```

```
Call:
lm(formula = Y ~ X, data = reg_DF)

Coefficients:
(Intercept)            X
1.956        1.269
```

Ex: Standard Errors and inference by hand

First, we need $\hat{\sigma}^2$:

```
sd_estimate <- sqrt(sum(resid(reg1)^2)/(dim(reg_DF)[1]-2))
# another way to get it
sigma(reg1)
```

```
[1] 0.9850476
```

Ex: Standard Errors and inference by hand

Now let's calculate the SE of our $\hat{\beta}$s:

```
# SE for beta_1
beta_se <- sd_estimate/sqrt(sum((reg_DF$X-mean(reg_DF$X))^2))
# SE for beta_0
alpha_se <- sd_estimate * sqrt((1/dim(reg_DF)[1]) + (mean(reg_DF$X)^2/sum((reg_DF$X-mean(reg_DF$X))^2)))
```

```
> beta_se
[1] 0.3409386
> alpha_se
[1] 0.1987247
```

Ex: Standard Errors and inference by hand

And now the test statistics and p-values:

```
# beta
2*pt((beta-0)/beta_se, dim(reg_DF)[1]-2, lower.tail = F)
# alpha
2*pt((alpha-0)/alpha_se, dim(reg_DF)[1]-2, lower.tail = F)
```

```
[1] 0.0003288147
[1] 2.663569e-16
```

Which looks correct:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9559     0.1987   9.842 2.66e-16 ***
X             1.2693     0.3409   3.723 0.000329 ***
---
Residual standard error: 0.985 on 98 degrees of freedom
```

Interpretation: What if slope = zero?

- **Hypotheses:** $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
- Failure to reject $H_0 : \beta_1 = 0$
  - X provides little help in predicting Y
  - X and Y could be related in a non-linear way, but this isn't what we looked for
  - Possible Type II error (relationship exists, just failed to find it)
- Reject $H_0 : \beta_1 = 0$
  - X provides significant help in predicting Y
  - There could also be a non-linear relationship exists and would be a better fit, even though this model was OK
  - Type I error (relationship doesn't exist, but we found it)?

New task: Create prediction estimates

- We can use the estimated linear regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ to predict a value of y for a specific value of x

- This prediction is not perfect because there is variability about regression line, so we'll construct a confidence interval for our prediction estimate

Purposes of prediction

Prediction can serve two purposes:

(1) Predict an individual's response at any chosen value x_0 of the predictor variable (a **prediction interval**)

- *Ex: What is the predicted birth weight of an individual baby who has a gestational period of 275 days?*
- *Ex: What is the length of a service call in which four components had to be repaired?*

(2) Predict an average response when $X = x_0$ (a **confidence interval** for the mean)

- *Ex: What is the predicted average birth weight of babies who have a gestational period of 275 days?*
- *Ex: What is the estimated average service time for calls that needed four components repaired?*

Note, there's more variability in individual responses than in average responses, so these intervals take different forms

#2: Estimate $\bar{Y}$ for a given value of x

- The estimated mean of Y for a given x-value falls on the fitted line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for a given x_0
- The standard error of the estimated conditional mean $\hat{E}(Y|x_0)$ is:

$$se_{\hat{E}(Y|x_i)} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

  when estimating $\bar{Y}$ for $x_i = 0$, the standard error is $se_{\hat{\beta}_0}$
- We have a better prediction for conditional mean of $Y|x_0$ when x_0 is near $\bar{x}$ compared to when x_0 is far from $\bar{x}$, why?
  - We have a lot more information near the bulk of the x values

Ex: Estimate $\bar{Y}$ for a given value of x

Estimate the conditional mean of $Y|x_0$ for $x_0 = 0$

```
# create new data
new_DF <- reg_DF; new_DF$X <- 0
# run predict on new data
predict(lm(Y~X), newdata=new_DF, se.fit=T)
```

1.955883

Does this look familiar?

This is the same as the $\hat{\beta}_0$

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.9559     0.1987   9.842 2.66e-16 ***
X              1.2693     0.3409   3.723 0.000329 ***
---
Residual standard error: 0.985 on 98 degrees of freedom
```

Which makes sense, why?

SE changes for a given value of x

Now, let's choose a value of x that's closer to $\bar{X}$, which ≈ 0.5

```r
# closer to mean
new_DF$X <- 0.5
predict(lm(Y~X), newdata=new_DF, se.fit=T)
```

These are the estimate SEs around the estimated predicted value

```
Before: 0.1987247
Now: 0.09852764
```

- Prediction doesn't really change, just smaller SE. Why?
- We have better confidence in predicting the conditional mean of Y value at $x = 5$, because this is closer to the middle of the data

CONFIDENCE INTERVAL FOR $se_{\hat{E}(Y|x_0)}$

- A $(1-\alpha)\%$ confidence interval for $E(Y|x_0)$ is given by

$$\hat{y_0} \pm t_{\alpha/2} se_{\hat{E}(Y|x_0)}$$

df of t distribution is $n-2$

```r
# confidence interval
predict(lm(Y~X), newdata=new_DF, interval = "confidence",
    level=0.95)
```

```
2.590534 2.39501 2.786059
```

#1: Predict a new value Y for a given value x_i

- Sometimes we want to predict the value of a single new response Y_{new} for a given x_0 value (not estimating the conditional mean)
- Estimate is $\hat{Y}_{\text{new}} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_0$, but the standard error is different, why?
- The standard error for a prediction of Y_{new}:

$$se(\hat{Y}_{\text{new}}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- There are two sources of variability in prediction on an individual Y
  - Uncertainty in group mean $E(Y|x_0)$
  - Variability of individual response around group mean

Prediction intervals for new Y values

Prediction intervals from R:

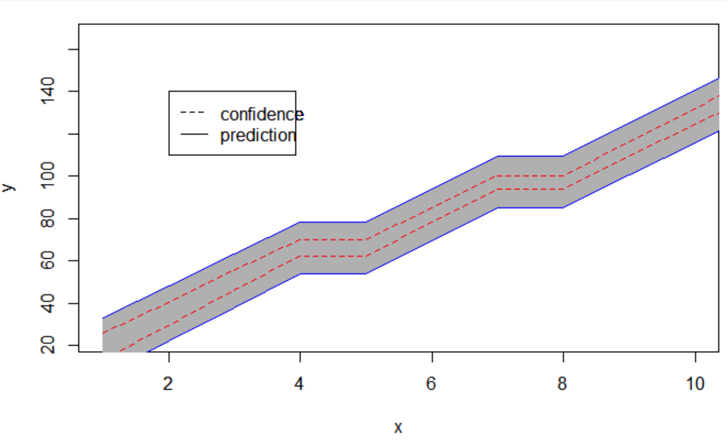
```
# prediction interval
predict(lm(Y~X), newdata=new_DF, interval = "prediction",
    level=0.95)
```

```
2.590534 0.6259854 4.555083
```

- Much wider than the 95% confidence intervals of the means we computed earlier at these same x values, why?

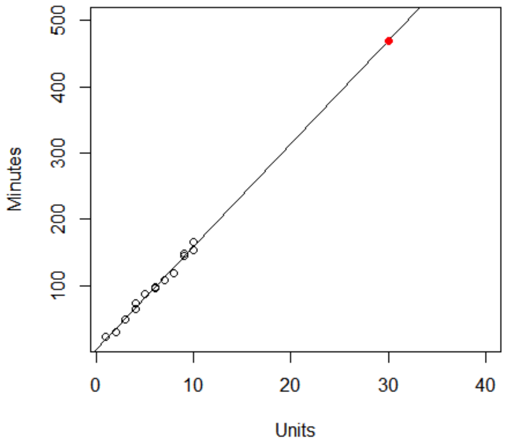
Visualizing the two intervals

Prediction intervals for an individual response are wider than a confidence interval for a mean



Don't extrapolate!

Worry about outliers



Why?

Don't extrapolate!

- Extrapolation is using a regression line to predict y-values for x-values outside the observed range of the data
- Extrapolation gets riskier the further we move from the range of the given x-values
- There is no guarantee that the relationship given by the regression equation holds outside the range of sampled x-values

Some reminders about bivariate regression

- A steeper slope (large absolute value β_1) does not mean there is a stronger linear relationship between Y and X, it does not mean you have a larger R^2 or r is closer to -1 or 1
- A strong relationship is when the model (fitted line) explains a lot of the variation in Y
- A strong linear relationship is when the observations fluctuate tightly around fitted line
- We can do a better job of predicting near the 'bulk' of the data
- We can do a better job of predicting a mean than in predicting an individual new Y value
- Inference of the parameters relies on the assumptions of linear regression

Wrap-up

Today we learned about...

- Correlation
- Parameters
- Prediction

LOOKING AHEAD

- **Today**: Regression with one explanatory variable
- **After reading week** we will learn how to:
  - Draw the best (hyper)plane through the data
  - Interpret multivariate regression results

Derivation: Parameters in regression

Differentiate S w.r.t. α and β and set derivatives equal to "0",

$$
\begin{aligned}
\frac{\partial S}{\partial \alpha} &= \frac{\partial \sum (Y_i - \alpha - \beta X_i)^2}{\partial \alpha} && (1)\\
&= 2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-1) && (2)\\
\text{Set to 0} : &= \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-1) = 0 && (3)\\
\text{Set to } \sum Y_i : \sum Y_i &= n\hat{\alpha} + \hat{\beta} \sum X_i = 0 && (4)\\
\text{Divide by } n : \frac{1}{n} \sum Y_i &= \hat{\alpha} + \frac{1}{n} \hat{\beta} \sum X_i && (5)\\
\text{Simplify} : \bar{Y} &= \hat{\alpha} + \hat{\beta} \bar{X} && (6)\\
\hat{\alpha} &= \bar{Y} - \hat{\beta} \bar{X} && (7)
\end{aligned}
$$

▸ Back to slides

Derivation: Parameters in regression

Differentiate S w.r.t. β_1 and set derivatives equal to "0",

$$\frac{\partial S}{\partial \beta_1} = \frac{\partial \sum (Y_i - \alpha - \beta_1 X_i)^2}{\partial \beta_1} \tag{8}$$

$$= 2 \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i) x_i (-1) \tag{9}$$

$$\text{Set to 0 and sub in } \hat{\beta}_0 : 0 = \sum_i^N x_i y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) x_i - \hat{\beta}_1 x_i^2 \tag{10}$$

$$\text{Distribute sum } : 0 = \sum_i^N x_i y_i - \bar{Y} \sum_i^N x_i + \hat{\beta}_1 \bar{X} \sum_i^N x_i - \hat{\beta}_1 \sum_i^N x_i^2 \tag{11}$$

$$\sum Y_i = n\bar{Y} : \hat{\beta}_1 = \frac{\sum_i^N x_i - N\bar{Y}\bar{X}}{\sum_i^N x_i^2 - N\bar{X}^2} \tag{12}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{X})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{X})^2} \tag{13}$$

Back to slides