Univariate regression II

Last time...

- Introduction to univariate regression
- Calculation and interpretation of b_0 and b_1
- Relationship between X, Y, $\hat{Y}$, and e

Today...

Statistical inferences with regression

- Partitioning the variance
- Testing b_{xy}

Statistical Inference

- The way the world is = our model + error
- How good is our model? Does it "fit" the data well?

To assess how well our model fits the data, we simply take all the variability in our outcome and partition it into different categories. For now, we will partition it into two categories: the variability that is predicted by (explained by) our model, and variability that is not.

Partitioning variation

- We formally test how well we are doing with our guesses by partitioning variation
- To the extent that we can generate different predicted values of Y, based on the values of the predictors, we are doing well in our prediction

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$

Partitioning variation

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$

Each of these is the sum of a squared deviation from an expected value of Y. We can abbreviate the sum of squared deviations:

$$SS_Y = SS_{\text{Model}} + SS_{\text{Residual}}$$

$$\frac{s^2_{regression}}{s^2_y} = \frac{SS_{regression}}{SS_Y} = R^2$$

Partitioning Variance

The relative magnitude of sums of squares, especially in more complex designs, provides a way of identifying particularly large and important sources of variability. In the future, we can further partition SS_{Model} and SS_{Residual} into smaller pieces, which will help us make more specific inferences and increase statistical power, respectively.

$$s_Y^2 = s_{\hat{Y}}^2 + s_e^2$$

Partitioning variance in Y

Consider the case with no correlation between X and Y

$$\hat{Y} = \bar{Y} + r_{xy}\frac{s_y}{s_x}(X - \bar{X})$$

$$\hat{Y} = \bar{Y}$$

To the extent that we can generate different predicted values of Y based on the values of the predictors, we are doing well in our prediction

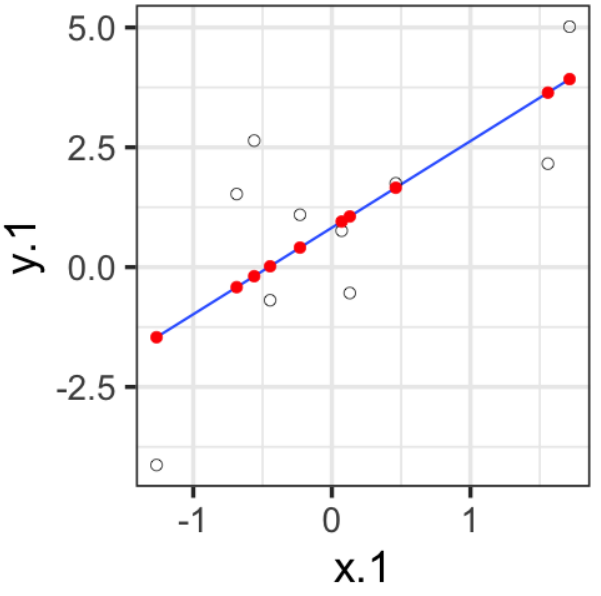
Coefficient of Determination

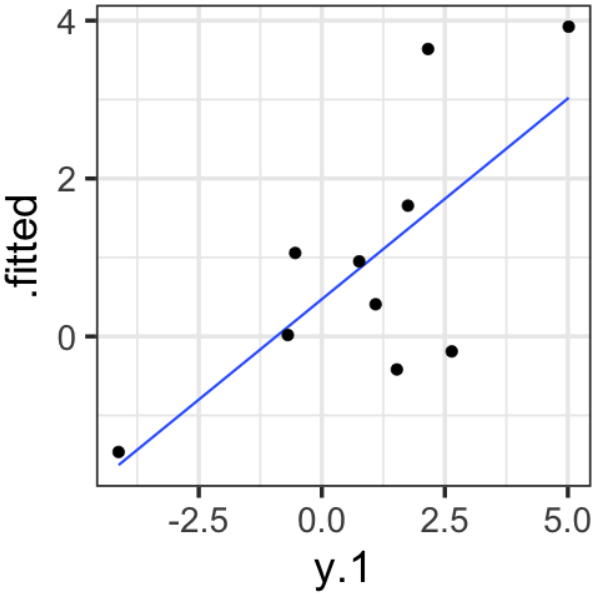
$$\frac{s^2_{Model}}{s^2_y} = \frac{SS_{Model}}{SS_Y} = R^2$$

R^2 represents the proportion of variance in Y that is explained by the model.

$\sqrt{R^2} = R$ is the correlation between the predicted values of Y from the model and the actual values of Y

$$\sqrt{R^2} = r_{Y\hat{Y}}$$


5.0
2.5
0.0
-2.5
y.1
-1
0
1
x.1


4
2
0
.fitted
-2.5
0.0
2.5
5.0
y.1

```
galton.data <- psychTools::galton
fit.1 = lm(child ~ parent, data = galton.data)
summary(fit.1)
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Example

```
cor(galton.data$parent, galton.data$child, use = "pairwise")
```

```
## [1] 0.4587624
```

```
cor(galton.data$parent, galton.data$child)^2
```

```
## [1] 0.2104629
```

```
galton.fits = augment(fit.1)
cor(galton.fits$child, galton.fits$.fitted)
```

```
## [1] 0.4587624
```

```
summary(fit.1)$r.squared
```

```
## [1] 0.2104629
```

Computing Sum of Squares

$$\frac{SS_{Model}}{SS_Y} = R^2$$

$$SS_{Model} = R^2(SS_Y)$$

$$SS_Y = SS_{Model} + SS_{residual}$$

$$SS_{residual} = SS_Y - R^2(SS_Y)$$

$$SS_{residual} = (1 - R^2)SS_Y$$

Using R To Check Yourself

$$SS_{residual} = (1 - R^2)SS_Y$$

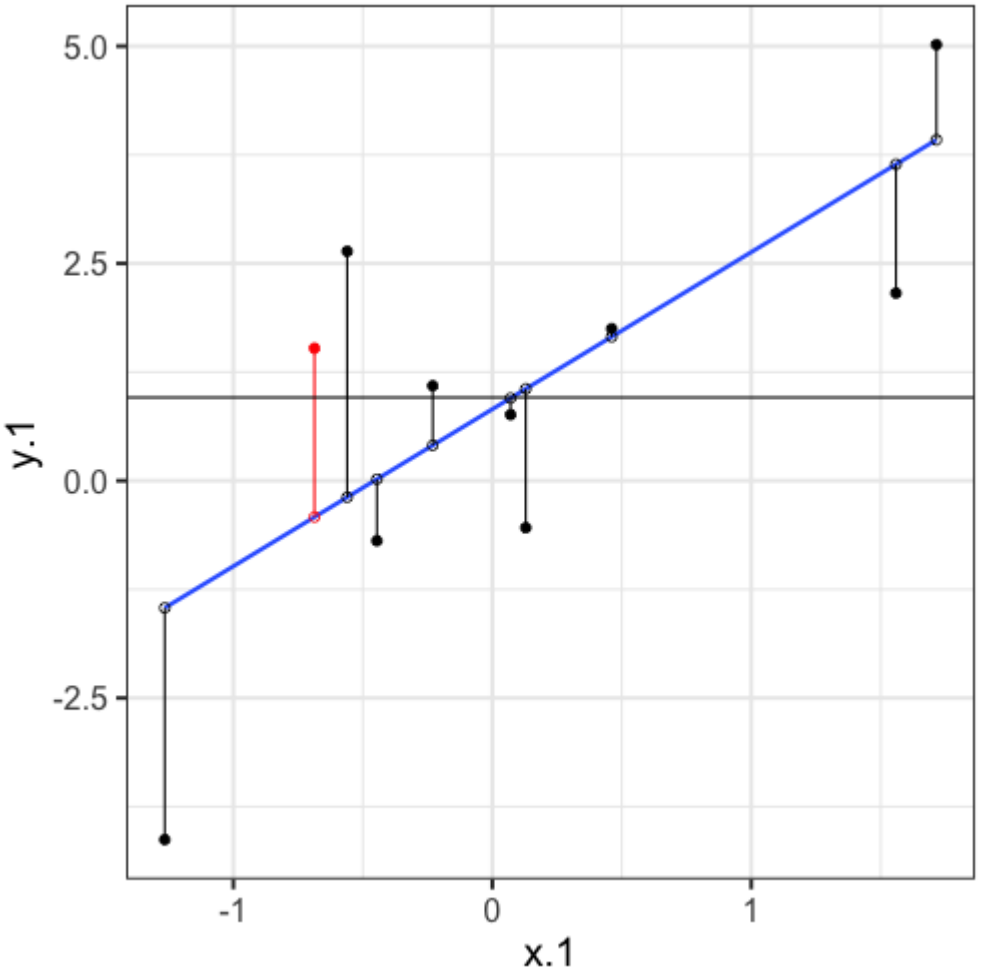
```
r2 = summary(fit.1)$r.squared
fit.1.anova = summary(aov(fit.1))
ssTotal = fit.1.anova[[1]]$`Sum Sq`[1] + fit.1.anova[[1]]$`Su
ssResidual = (1 - r2) * ssTotal
fit.1.anova
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## parent         1   1237    1237   246.8 <2e-16 ***
## Residuals    926   4640       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ssResidual
```

```
## [1] 4640.273
```

What is an "error"?



Mean Square Error (MSE) (msw/msr)

- AKA square of residual standard error/deviation (σ^2)
- *Unbiased* estimate of error variance
- Measure of discrepancy between the data and the model
- Variance of data around the fitted regression line (same as MSwithin for group means)
- It is the mean of the square of the residuals

```
head(fit.1$residuals)
```

```
##         1         2         3         4         5         6
## -7.805016 -6.512435 -4.573563 -3.927273 -3.604127 -5.366144
```

MSE

- It is the square of the residual standard error/deviation (sigma) aka RSE^2

```
mse = round((summary(fit.1)[["sigma"]]) ^2, digits = 2)
mse
```

```
## [1] 5.01
```

Residual Standard Error

```
##
## Call:
## lm(formula = child ~ parent, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Residual standard error/deviation

- aka standard deviation of the residual
- aka **standard error of the estimate**
- aka σ

$$\hat{\sigma} = \sqrt{\frac{SS_{\text{Residual}}}{df_{\text{Residual}}}} = s_{Y|X} = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}}$$

- interpreted in original units (unlike R^2)
- standard deviation of Y not accounted by model

Residual standard error/deviation or standard error of the estimate

```
summary(fit.1)$sigma
```

```
## [1] 2.238547
```

```
galton.data.1 = broom::augment(fit.1)
psych::describe(galton.data.1$.resid)
```

```
##    vars   n mean   sd median trimmed  mad   min  max range  skew kurto
## X1    1 928    0 2.24   0.05    0.06 2.26 -7.81 5.93 13.73 -0.24    -0
```

```
sd(galton.data$child)
```

```
## [1] 2.517941
```

$\hat{\sigma}$ is in original units of Y, so you need to compare to the sd of Y!

RSE vs MSE

Residual standard error = square root of the mean square error

Both measuring error, but RSE is a little more useful

```
sqrt(mse)
```

```
## [1] 2.238303
```

Residual Standard Error and σ

- So many names to represent the spread of data around the regression line
- Standard deviation of the residual, standard error of the estimate, MSE...
- We will refer to this as sigma, and use estimated sigma, as we do not know the population value ($\hat{\sigma}$)
- It is interpreted in original units (unlike R^2)
- It is the standard deviation of Y not accounted by the model (i.e., residuals)

Why do we care about σ?

- Let's simulate!
- Data generating process:

$$Y_i \sim \mathcal{N}(\mu, \sigma)$$

- This describes how we think our DVs are generated, and the parameters of interest
- For normal, μ gets all the focus but σ is just as important

Simulation

Our plan is to fix μ and then vary σ to see what happens

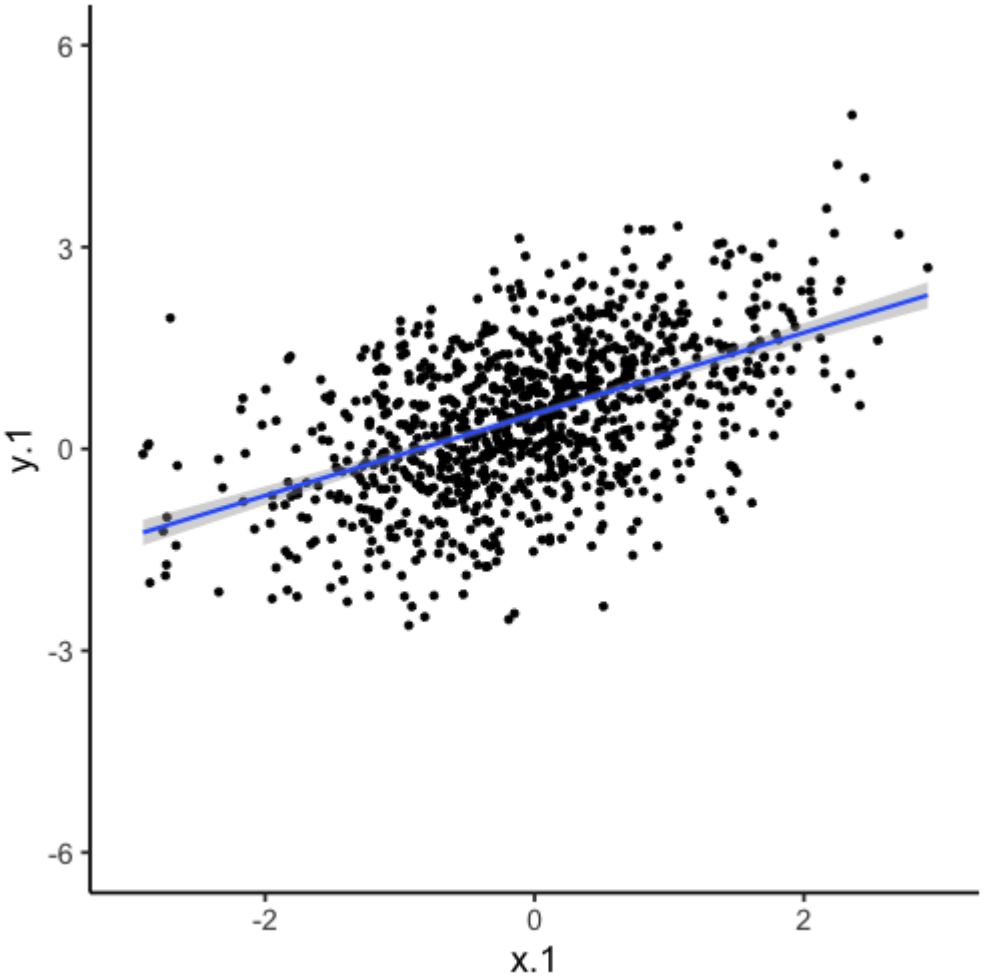
```
set.seed(1234)
x.1 <- rnorm(1000, 0, 1) # randomly select 1000 numbers for x
e.1 <- rnorm(1000, 0, 1) # randomly select 1000 numbers for e
y.1 <- .5 + .55 * x.1 + e.1 # create our y
d.1 <- data.frame(x.1,y.1) # combine x and y into a data.fram
m.1 <- lm(y.1 ~ x.1, data = d.1) # use x to predict y with th
```

Simulation

```
summary(m.1)
```

```
##
## Call:
## lm(formula = y.1 ~ x.1, data = d.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1661 -0.6439  0.0145  0.6537  3.0684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51599    0.03100   16.64   <2e-16 ***
## x.1          0.60571    0.03109   19.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9801 on 998 degrees of freedom
## Multiple R-squared:  0.2755,    Adjusted R-squared:  0.2748
## F-statistic: 379.5 on 1 and 998 DF,  p-value: < 2.2e-16
```

Simulation



Simulation

Again, but with a larger sigma

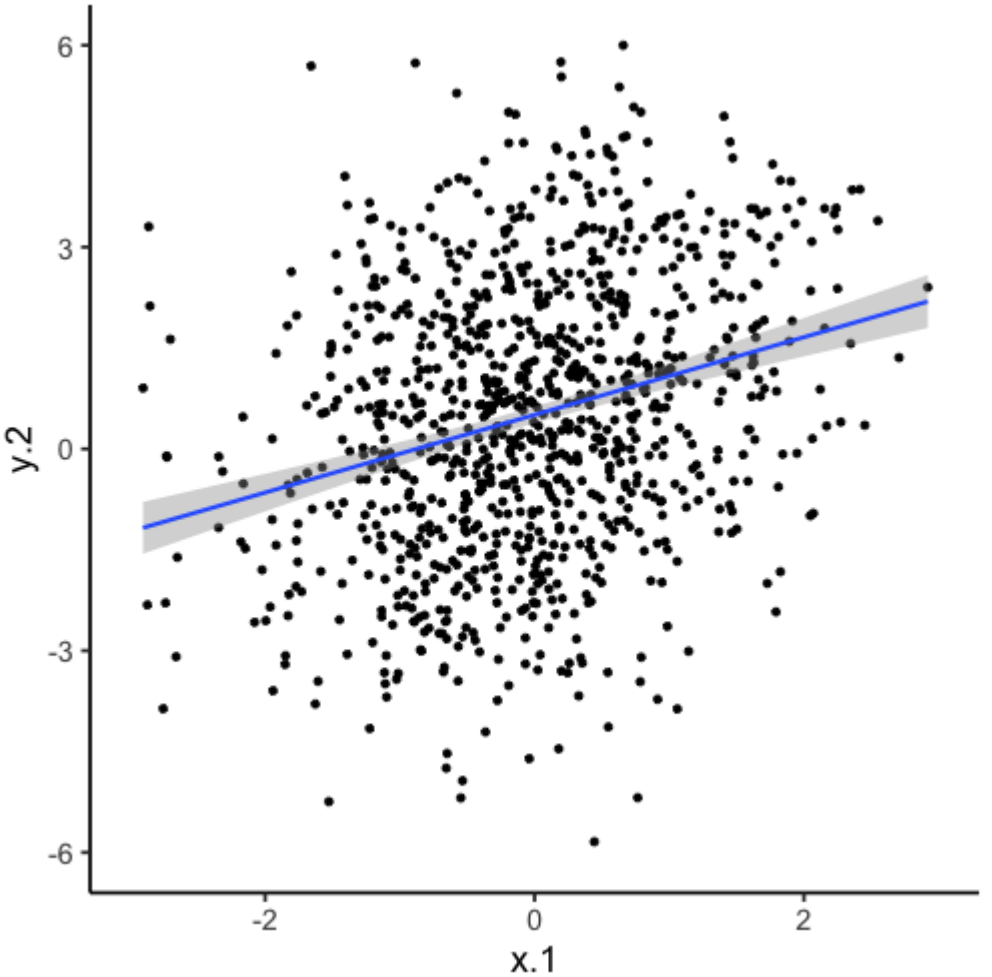
```
set.seed(987)
e.2 <- rnorm(1000, 0, 2) # larger sigma
y.2 <- .5 + .55 * x.1 + e.2 # same Xs, same mu (.5)
d.2 <- data.frame(x.1,y.2)
m.2 <- lm(y.2 ~ x.1, data = d.2)
```

Simulation

```
summary(m.2)
```

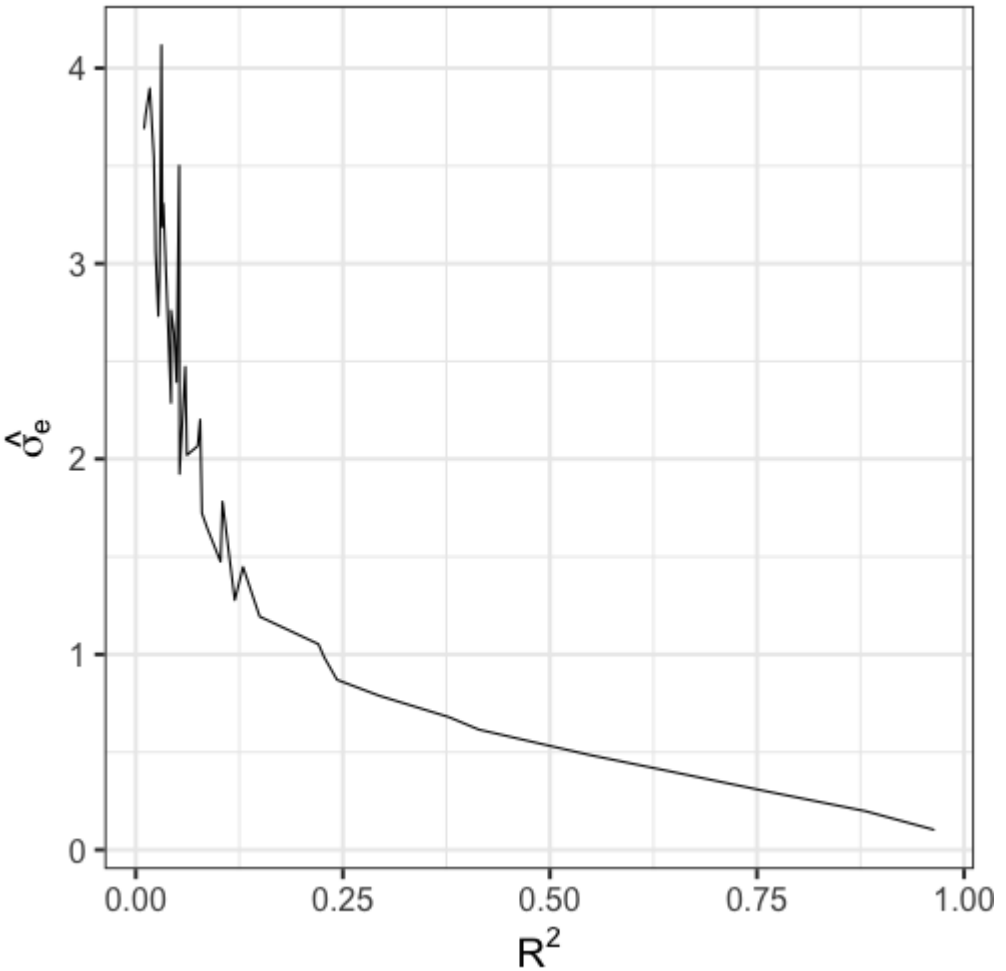
```
##
## Call:
## lm(formula = y.2 ~ x.1, data = d.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6267 -1.4359 -0.0192  1.4480  6.3439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52137    0.06345   8.217 6.43e-16 ***
## x.1          0.59823    0.06363   9.402  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.006 on 998 degrees of freedom
## Multiple R-squared:  0.08136,    Adjusted R-squared:  0.08044
## F-statistic: 88.39 on 1 and 998 DF,  p-value: < 2.2e-16
```

Simulation



R^2 and residual standard deviation

- two sides of same coin
- one in original units, the other standardized
- R^2 can be tricky because the numerator and denominator can be changed in different ways.
- for example if variance in Y is changed but with the same regression model and residual standard error, R^2 could increase or decrease


4
3
2
1
0
$\hat{\sigma}_e$
0.00
0.25
0.50
0.75
1.00
R^2

Inferential tests

NHST is about making decisions:

- these two means are/are not different
- this correlation is/is not significantly different from 0
- the distribution of this categorical variable is/is not different between these groups

In regression, there are several inferential tests being conducted at once. The first is called the **omnibus test** -- this is a test of whether the model fits the data.

Omnibus test

$$H_0 : \rho^2_{XY} = 0$$

$$H_0 : \rho^2_{XY} \neq 0$$

It is possible to calculate the significance of your regression with a correlation test.

However, we use **the *F* distribution** to estimate the significance of our model. Methods are mathematically equivalent! But when have we seen the **F** before?

F Distribution review

The F probability distribution represents all possible ratios of two variances:

$$F \approx \frac{s_1^2}{s_2^2}$$

Each variance estimate in the ratio is χ^2 distributed, if the data are normally distributed. The ratio of two χ^2 distributed variables is F distributed. It should be noted that each χ^2 distribution has its own degrees of freedom.

F Distributions and regression

Recall that when using a *z* or *t* distribution, we were interested in whether one mean was equal to another mean. In this comparison, we compared the *difference of two means* to 0 (or whatever), and if the difference was not 0, we concluded significance.

F statistics are not testing the likelihood of differences; they test the size of a *ratio*. Is the variance explained by our model larger in magnitude than another variance.

Which variance?

$$F_{\nu_1\nu_2} = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

$$F_{\nu_1\nu_2} = \frac{\frac{\text{Variance}_{\text{Model}}}{\nu_1}}{\frac{\text{Variance}_{\text{Residual}}}{\nu_2}}$$

$$F = \frac{MS_{Model}}{MS_{residual}}$$

The degrees of freedom for our model are

$$DF_1 = k$$

$$DF_2 = N - k - 1$$

Where *k* is the number of IV's in your model, and *N* is the sample size.

Mean squares are calculated by taking the relevant Sums of Squares and dividing by their respective degrees of freedom.

- SS_{Model} is divided by DF_1
- SS_{Residual} is divided by DF_2

```
anova(fit.1)
```

```
## Analysis of Variance Table
##
## Response: child
##            Df Sum Sq Mean Sq F value    Pr(>F)
## parent      1 1236.9 1236.93  246.84 < 2.2e-16 ***
## Residuals 926 4640.3    5.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit.1)
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Model Comparisons

- Omnibus F-statistic is the ratio of MSregression to MSresidual
- Test of overall significance. Does your model give a better fit to the data than a model that contains no independent variables?

Model Comparisons

- How much variance remains unexplained in our model. This "left over" variance can be contrasted with an alternative model/hypothesis. Does adding a new predictor variable help explain more variance or should we stick with the most parsimonious (simplest) model?

- Every model you report implies that you are favoring that model over an alternative model, typically the null. Taking a more formal model comparison approach allows you to be more flexible and explicit.

Full vs. Restricted Models

```
fit.1 <- lm(child ~ parent, data = galton.data)
fit.0 <- lm(child ~ 1, data = galton.data)

summary(fit.0)
```

```
##
## Call:
## lm(formula = child ~ 1, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3885 -1.8885  0.1115  2.1115  5.6115
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.08847    0.08266   823.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.518 on 927 degrees of freedom
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

```
anova(fit.0)
```

```
## Analysis of Variance Table
##
## Response: child
##            Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 927 5877.2    6.34
```

```
anova(fit.1)
```

```
## Analysis of Variance Table
##
## Response: child
##            Df Sum Sq Mean Sq F value Pr(>F)
## parent      1 1236.9 1236.93  246.84
## Residuals 926 4640.3    5.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.
```

The comparison!

```
anova(fit.1, fit.0)
```

```
## Analysis of Variance Table
##
## Model 1: child ~ parent
## Model 2: child ~ 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    926 4640.3
## 2    927 5877.2 -1   -1236.9 246.84 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparisons

- Model comparisons are redundant with nil/null hypotheses and coefficient tests right now, be they'll be more flexible down the road
- Key is to start thinking about your **implicit** alternative models
- The ultimate goal would be to create two models that represent two equally plausible theories
- A model embodies your hypothesis! It is the mathematical expression of your hypothesis!

Regression coefficient

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

We conduct a one-sample *t*-test! What do you need to do so?

Regression coefficient

$$se_b = \frac{s_Y}{s_X}\sqrt{\frac{1 - r_{xy}^2}{n-2}}$$

$$t(n-2) = \frac{b_1}{se_b}$$

SE_b

- standard errors for the slope coefficient
- represent our uncertainty (noise) in our estimate of the regression coefficient
- different from residual standard error/deviation (but proportional to)
- much like previously we can take our estimate (b) and put confidence regions around it to get an estimate of what could be "possible" if we ran the study again

What does the regression coefficient test?

- Does X provide any predictive information?
- Does X provide any explanatory power regarding the variability of Y?
- Is the the average value the best guess (i.e., is Y bar equal to the predicted value of Y?)
- Is the regression line flat?
- Are X and Y correlated?

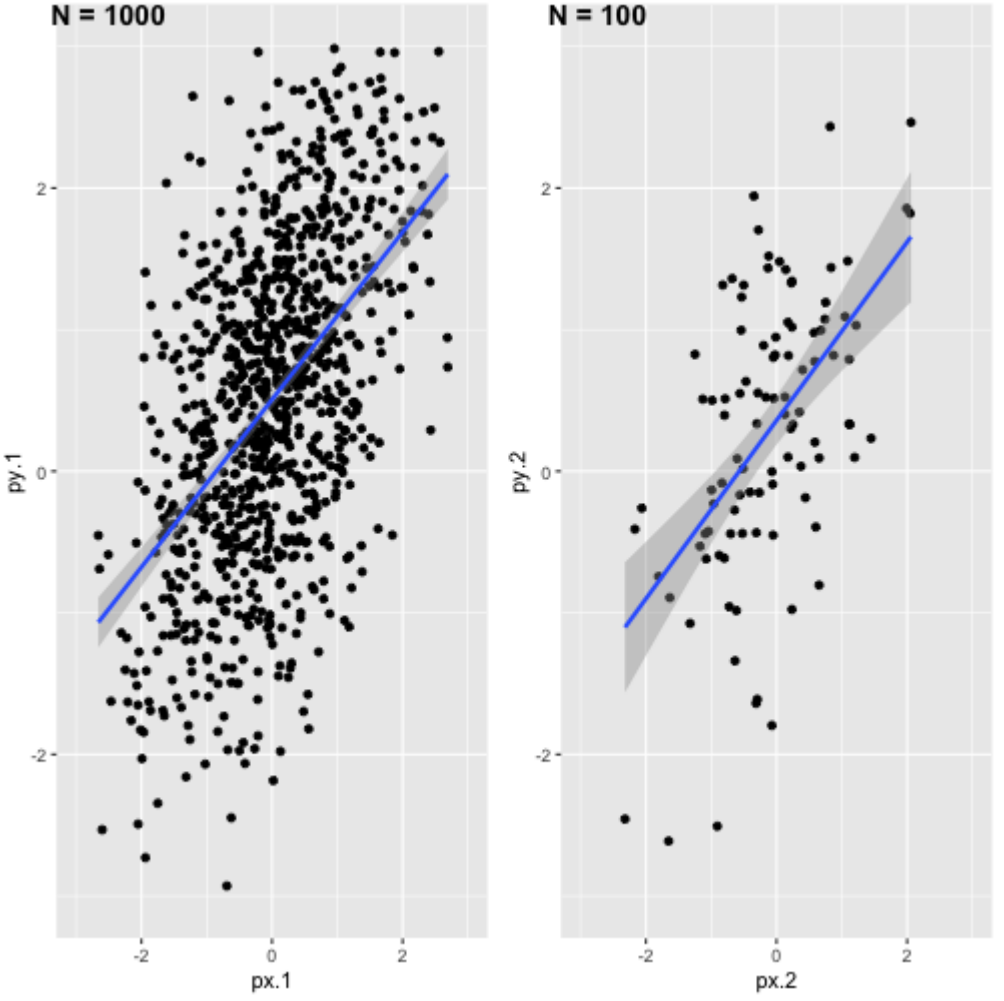
Intercept

- The same but different
- More complex standard error calculation as the calculation depends on how far the X value (here zero) is away from the mean of X
    - farther from the mean, less information, thus more uncertainty

Confidence interval for coefficents

- Same equation as we've been working with
- Estimate plus minus 1.96*se

Confidence Bands



Confidence Bands

$$\hat{Y} \pm t_{critical} * se_{residual} * \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_X^2}}$$

Prediction Bands

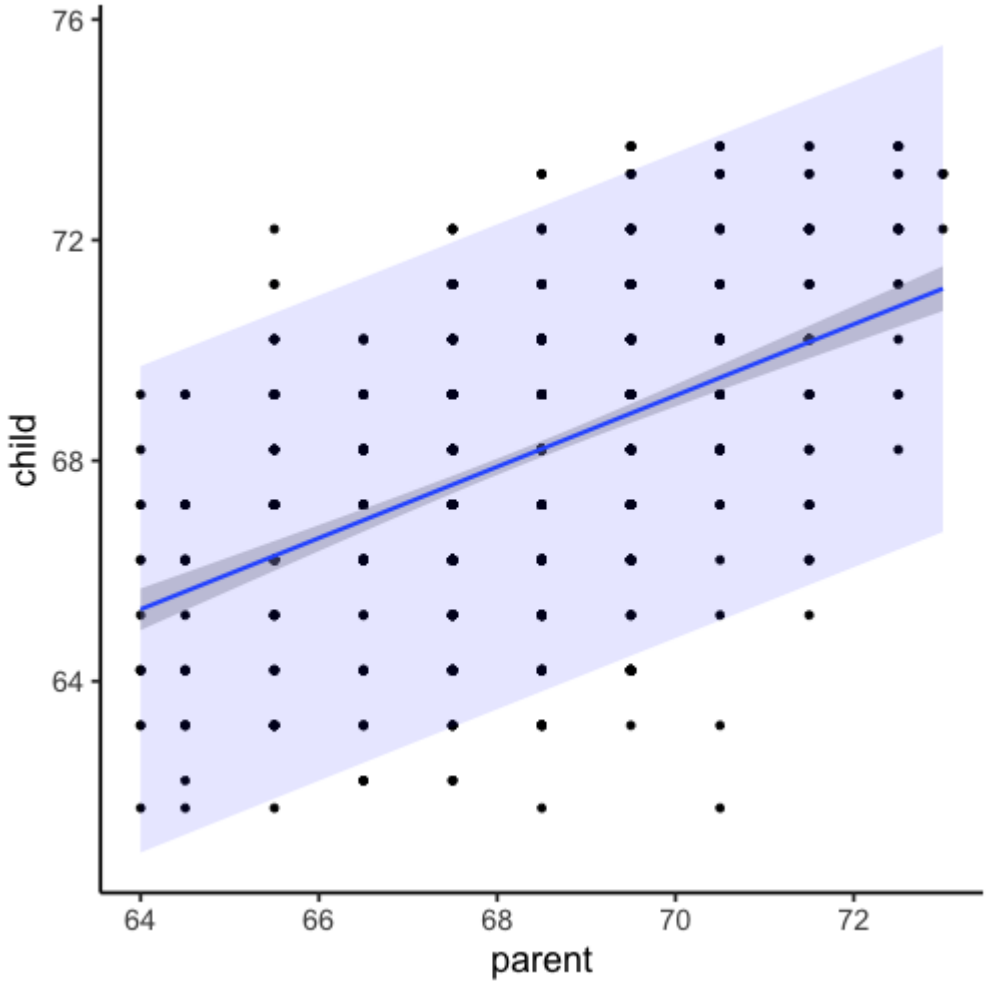
- We are predicting and individual score, not the $\hat{Y}$ for a particular level of X.
- Because there is greater variation in predicting an individual value rather than a collection of individual values (i.e., the mean) the prediction band is greater
- Combines unknown variability in 1) the estimated mean (as reflected in se of b) 2) peoples scores around mean (residual standard error)

$$\hat{Y} \pm t_{critical} * se_{residual} * \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_X^2}}$$

Prediction Bands

```
temp_var <- predict(fit.1, interval = "prediction")
new_df <- cbind(galton.data, temp_var)
pred <- ggplot(new_df, aes(y = child, x = parent)) +
  geom_point() +
  geom_smooth(method = lm, se = TRUE) +
  geom_ribbon(aes(ymin = lwr, ymax = upr),
              fill = "blue",
              alpha = 0.1) +
  theme_classic(base_size = 18)
```

Prediction Bands



Things you should be able to do (based on this lecture and the one before):

- Can you get an individual's predicted score & residual from an equation?
- Can you calculate b0 & b1?
- Can you take the output from R and write a regression equation, being very specific?
- Can you tell me if you have a "good" model?
- Is a coefficient meaningful? How do you know and can you calculate it?
- Can you plot a regression, confidence band, and prediction band?

Next time...

Partial Correlations