STA302/STA1001, Week 4

Mark Ebden, 28 September - 3 October 2017

With grateful acknowledgment to Alison Gibbs and Becky Lin

Today's class

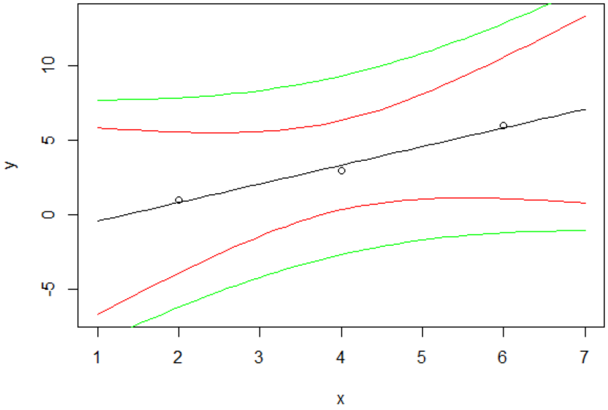
- Reinforcing Week 3 concepts
- Qualitative Predictors: Dummy variable regression
- Exploring Correlation
- Reference: Simon Sheather §2.6, Ch 3

Before we begin

- StatsCan
- DataCamp
- TA office hours on Tuesdays: 5-7 pm $\rightarrow$ 4-6 pm?

(A) Comparing the CIs to the PIs

In Week 3, some of you were interested to see that the red-green vertical distance (slide 64) isn't constant.



(A) Comparing the CIs to the PIs

The confidence interval from slide 14 is:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) \pm t(\alpha/2, n-2)\, S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

The prediction interval from Week 3, slide 62 is:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) \pm t(\alpha/2, n-2)\, S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Comparing the CIs to the PIs

Explanation: As $x^* - \bar{x}$ increases, the "1 +" term in the square root has less relative effect. As $(x^* - \bar{x}) \to \pm\infty$, the confidence- and prediction intervals will become indistinguishable.

In other words: The prediction intervals (PIs) are wider than the CIs because of the e_i model error term on slide 60, associated with drawing an actual observation and not just considering the population regression line $\beta_0 + beta_1 x^*$ as CIs do. This leads to the "1 +" term. So there are two sources of uncertainty rather than one: the uncertainty in $E(Y|X = x^*)$ (affecting both the CIs and PIs) and the uncertainty in e_i (affecting the PIs). After they are added you take the square root, so the overall uncertainty cannot be expressed as simply their sum.

(B) What's in a name? That which we call a R^2...

Recall from Week 3, slide 43:

$$\frac{\text{SSReg}}{\text{SST}} = R^2, \quad 0 \leq R^2 \leq 1$$

R^2 gives the percent of variation in $\mathbf{y}$ that is explained by the regression line

Why is $\frac{\text{SSReg}}{\text{SST}}$ referred to as R^2? Take the square root of the above and see!

(B) continued

$$\sqrt{R^2} = \sqrt{\frac{\text{SSReg}}{\text{SST}}} = \sqrt{\frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}}}$$

In Week 2, slide 19, we found that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Substituting into the above gives:

$$\sqrt{\frac{S_{xy}^2}{S_{xx}S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

This is Pearson's correlation coefficient by the definition given in Week 3, slide 7. The coefficient is often referred to as ρ for unobserved random variables and as r or R when calculated on an actual data sample.

This is why we refer to the coefficient of determination, $\frac{\text{SSReg}}{\text{SST}}$, as R^2.

We'll return to r later in the week.

(C) The expected mean square of regression

In Week 3, recall from slide 44 that $\mathbb{E}(\text{MSReg}) = \sigma^2 + \beta_1^2 S_{xx}$. To begin the proof:

$$\mathbb{E}(\text{MSReg}) = \mathbb{E}\left(\hat{\beta}_1^2\right) \sum_{i=1}^{n} (X_i - \bar{X})^2 = \mathbb{E}\left(\hat{\beta}_1^2\right) S_{xx}$$

Where can we find an expression for $\mathbb{E}\left(\hat{\beta}_1^2\right)$? Here:

$$\text{var}(\hat{\beta}_1) = \mathbb{E}\left(\hat{\beta}_1^2\right) - \left[\mathbb{E}\left(\hat{\beta}_1\right)\right]^2$$

where we'd calculated the other two terms in Week 2, slides 32 and 34:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \qquad \mathbb{E}\left(\hat{\beta}_1\right) = \hat{\beta}_1$$

Strategy: Solve for $\mathbb{E}\left(\hat{\beta}_1^2\right)$ and multiply that expression by S_{xx} to give you the equation on slide 44.

(D) Cochran's theorem

On slide 45 we discussed the F distribution. You may be wondering why the quantities MSReg and MSE are χ^2 distributions. This is due to *Cochran's theorem*, a proof of which is beyond the scope of this course. For the purposes of STA302, Cochran's theorem says:

- Let n observations of $Y \sim \mathcal{N}(\mu, \sigma^2)$ be used to calculate an SS_T
- Decompose SS_T into k sums of squares SS_r each with degrees of freedom df_r — e.g. in our simple linear regression analysis, $k = 2$ and $r \in \{\mathrm{Reg}, E\}$
- If $\sum_{r=1}^{k} \mathrm{df}_r = n - 1$ then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom

Our SLR analysis example was:

$$\mathrm{SST} = \mathrm{SSReg} + \mathrm{RSS}$$

This meets the criteria for Cochran's theorem — e.g. recall that for the degrees of freedom, $n - 1 = 1 + n - 2$

Therefore, MSReg and MSE are χ^2 distributions.

(E) Exercise from Week 3 slide 50: Prove $t^2 = F$

Our observed test statistic is

$$F^* = \frac{\text{SSReg}/1}{\text{SSE}/(n-2)} = \frac{\hat{\beta}_1^2 S_{xx}}{\text{MSE}}$$

In Week 3, slides 17–19, we saw that:

- The s.e. of the regression is $\widehat{\text{var}(\hat{\beta}_1)} = \frac{S^2}{S_{xx}}$ where S^2 is the MSE
- $t^* = \hat{\beta}_1 \,/\, \text{se}(\hat{\beta}_1)$ is our t-statistic

Therefore,

$$F^* = \frac{\hat{\beta}_1^2 S_{xx}}{\widehat{\text{var}(\hat{\beta}_1)} S_{xx}} = \left[\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}\right]^2 = (t^*)^2$$

Today's class

- Reinforcing Week 3 concepts
- **Qualitative Predictors: Dummy variable regression**
- Exploring Correlation
- Reference: Simon Sheather §2.6, Ch 3

Coding Qualitative Predictors in Regression Models

- A coded qualitative variable in a regression takes on a finite number of values so that different categories of a nominal variable can be identified
- The term *qualitative* reflects the fact that the values taken on by such variables (e.g., 0, 1, -1) do not indicate meaningful measurements but rather categories of interest.

Coding Qualitative Predictors in Regression Models

- Consider a regression model: $y_i = \beta_0 + \beta_1 x_i + e_i$
- Examples of coded qualitative variables are:

$$x = \begin{cases} 1 & \text{for one category of data} \\ 0 & \text{for the other category} \end{cases}$$

$$x = \begin{cases} 1 & \text{if subject is male} \\ -1 & \text{if subject is female} \end{cases}$$

- The first is a qualitative predictor which includes a 0 among its choices, and it is known as a "dummy variable"

Dummy variables

So a common choice, when using simple linear regression to compare two models, is to set $x = 0$ for one model and $x = 1$ for the other, in the following:

$$Y = \beta_0 + \beta_1 x + e$$

In the CFC example, suppose we were interested in comparing the mean of the data before and after the Montreal Protocol. Let

$$x_i = \begin{cases} 1 & \text{if the } i\text{th observation is before the MP} \\ 0 & \text{if the } i\text{th observation is after the MP} \end{cases}$$

Montreal Protocol example continued

Thus in the model $Y_i = \beta_0 + \beta_1 x_i + e_i$ we have that

$$\mathbb{E}(Y_i) = \begin{cases} \beta_0 + \beta_1 & \text{if the } i\text{th observation is before the MP} \\ \beta_0 & \text{if the } i\text{th observation is after the MP} \end{cases}$$

We can estimate $\mathbb{E}(Y_i)$ by

$$\mathbb{E}(Y_i) = \begin{cases} b_0 + b_1 & \text{if the } i\text{th observation is before the MP} \\ b_0 & \text{if the } i\text{th observation is after the MP} \end{cases}$$

We can test whether the mean of Y_i is the same before and after the Montreal Protocol by testing:

$$H_0 : \beta_1 = 0 \quad \textit{versus} \quad H_a : \beta_1 \neq 0$$

This is equivalent to the two-sample t-test assuming equal variances. (Consistent with the Gauss-Markov conditions.)

A cousin of dummy-variable regression

The two-sample t-test can give the same result as dummy-variable regression.

Building on Week 3, slides 35–36:

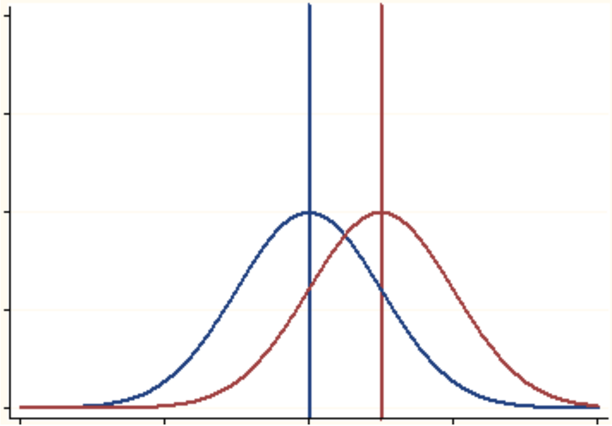
We can estimate σ^2 via the pooled variance, s^2 :

$$s^2 = \frac{\sum_{i=1}^{n_A}(y_{iA} - \bar{y}_A)^2 + \sum_{i=1}^{n_B}(y_{iB} - \bar{y}_B)^2}{n_A + n_B - 2}$$

Then the following is the **two-sample t-statistic**:

$$\frac{\bar{y}_A - \bar{y}_B}{s\sqrt{1/n_A + 1/n_B}} \sim t_{n_A+n_B-2}$$

Example: Comparing dummy-variable regression to the t-test and F-test



Generate a dataset for the Example

```
set.seed(3)
y1 <- rnorm(12,1,1); y2 <- rnorm(10,2,1)
n1 <- length(y1); n2 <- length(y2)
y <- c(y1,y2) # All response variables
x <- c(rep(0,n1),rep(1,n2)) # Indicator variables
print(t(matrix(sort(round(y1,2)),ncol=2))) # Just to show y1 (!)
```

```
##       [,1]  [,2]  [,3] [,4] [,5] [,6]
## [1,] -0.22 -0.15 -0.13 0.04 0.26 0.71
## [2,]  1.03  1.09  1.20 1.26 2.12 2.27
```

```
print(round(sort(y2),2)) # Show y2
```

```
##  [1] 1.05 1.06 1.28 1.35 1.42 1.69 2.15 2.20 2.25 3.22
```

Method 1(a): Two-sample t-test (by hand)

```r
s <- sqrt(((n1-1)*var(y1)+(n2-1)*var(y2))/(n1+n2-2)) # Pooled var.
tstar <- (mean(y1)-mean(y2))/(s*sqrt(1/n1+1/n2)); round(tstar,2)
```

```
## [1] -2.92
```

```r
pval <- 2*pt(tstar,n1+n2-2); round(pval,5)
```

```
## [1] 0.00853
```

Method 1(b): Two-sample *t*-test (by R function)

```r
t.test(y1,y2,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  y1 and y2
## t = -2.9164, df = 20, p-value = 0.008535
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6818907 -0.2792216
## sample estimates:
## mean of x mean of y
## 0.7877203 1.7682764
```

Method 2(a): Dummy-variable regression (by hand)

```
n <- length(x)
mx <- mean(x); my <- mean(y)
Sxx <- sum((x-mx)^2); Sxy <- sum((x-mx)*(y-my))
b1 <- Sxy/Sxx; b0 <- mean(y) - b1*mean(x)
yhat <- b0 + b1*x
RSS <- sum((y-yhat)^2); S <- sqrt(RSS/(n-2))
seB0 <- S*sqrt(1/n+mx^2/Sxx) # standard error; Wk 3 slide 18
seB1 <- S/sqrt(Sxx)
t0 <- b0/seB0 # the test statistic for the intercept
t1 <- b1/seB1 # and for the slope
pval0 <- 2*pt(-abs(t0),n-2) # pvalue for the intercept
pval1 <- 2*pt(-abs(t1),n-2) # and the slope
print(c(b0,b1,pval0,pval1))
```

```
## [1] 0.787720254 0.980556125 0.002388998 0.008534651
```

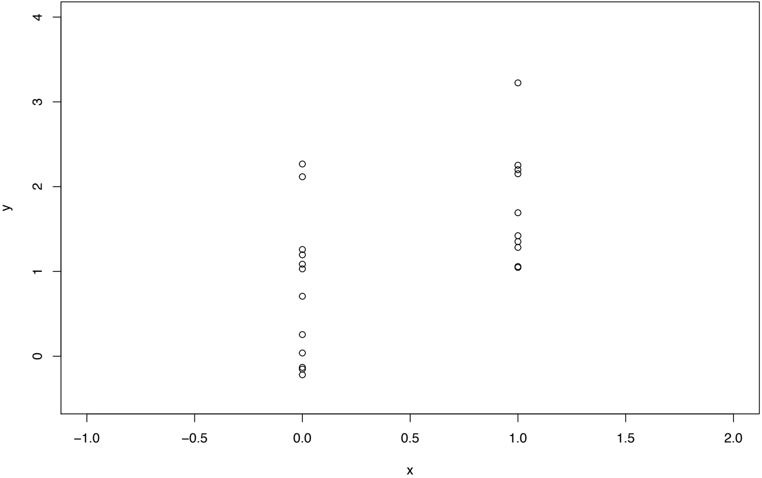
Method 2(b): Dummy-variable regression (by R function)

```r
myFit <- lm(y~x); summary(myFit) # (Free F-test too!)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00658 -0.66606 -0.07809  0.42567  1.47965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7877     0.2267   3.475  0.00239 **
## x             0.9806     0.3362   2.916  0.00853 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7852 on 20 degrees of freedom
## Multiple R-squared:  0.2984, Adjusted R-squared:  0.2633
## F-statistic: 8.506 on 1 and 20 DF,  p-value: 0.008535
```

The Mystery of the Trinity — Explained

```
plot(x,y,xlim=c(-1,2),ylim=c(-.5,4))
```



The Mystery of the Trinity — Explained

We have shown already that F and T^2 are equivalent.

The t-test considers the difference in means of the y's, ignoring x. The test statistic is then based on $\bar{y}_2 - \bar{y}_1$.

Linear regression estimates a $\hat{\beta}_0$ and $\hat{\beta}_1$. It can be shown that $\hat{\beta}_1 = \Delta y / \Delta x = \frac{\bar{y}_2 - \bar{y}_1}{\Delta x}$. Recall that Δx is a fixed scalar (based on your arbitrary choice of dummy variable).

Today's class

- Reinforcing Week 3 concepts
- Qualitative Predictors: Dummy variable regression
- **Exploring Correlation**
- Reference: Simon Sheather §2.6, Ch 3

Exploring Correlation

Example use:

- Suppose you're interested in the relationship between two r.v.'s X and Y, believing it's linear
- If there is a clear choice for Y being the response and X being the explanatory variable, you may carry out a regression
- An option when X and Y don't have a clear choice for dependent/independent variables: you can **summarize the strength** of the linear relation by correlation

Exploring Correlation

Correlation is a symmetric measure. Given n observations (x_i, y_i) of the r.v.'s, Pearson's Product-Moment Correlation is:

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

Compare to Week 3, slide 7: r is a good estimate for ρ. Under certain conditions (X and Y have a bivariate normal distribution), r is the maximum likelihood estimate of ρ.

Facts about r

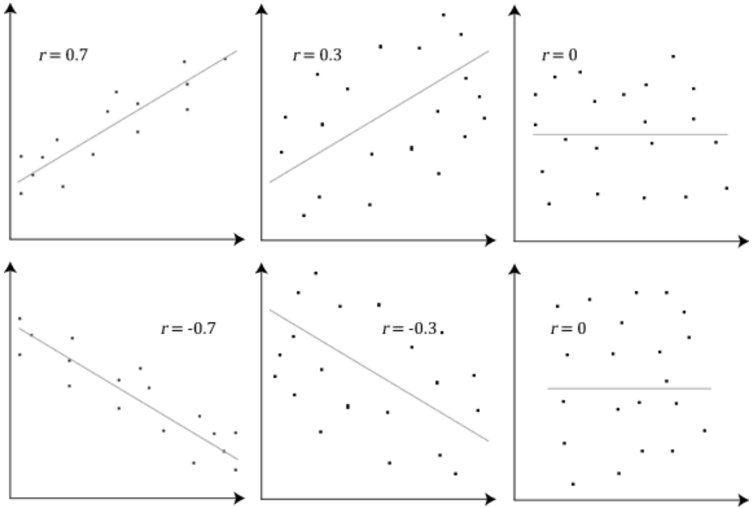
It's a measure of the degree of *linear* association between X and Y.

It's dimension-free.

It's always between -1 and $+1$, inclusive:

- $r = 1$ means (x_i, y_i) fall exactly on a straight line, with positive relationship
- $r = -1$ means (x_i, y_i) fall exactly on a straight line with negative relationship
- $r = 0$ means no linear relationship

Examples



r in R

Use the `cor()` function, e.g.: `cor(x,y)` where `x` and `y` are vectors.

You can also write simply `cor(Z)` if `Z` is a matrix or data frame. This will calculate the r values among all possible combinations of the columns.

An advantage of using R is that it's signed: positive- and negative correlations are differentiated. An advantage of using R^2 is that the ANOVA framework extends well to multiple linear regression (with two or more independent variables rather than a single x).

Correlation exercise

We know that $\sum_{i=1}^{n} \hat{e}_i x_i = 0$ and $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$.

Show why this implies that

$$r_{\hat{e}_i, \hat{y}_i} = 0, \quad r_{\hat{e}_i, x_i} = 0$$

Next steps

- Solutions to the seven questions in Chapter 2 of our textbook will be posted on the weekend – last chance to try them without peaking!
- Next TA office hours: tomorrow morning