HUMBEHV 3ST3

Linear Regression

Prof. Patrick Bennett

Correlation vs. Regression

• Correlation: a measure of the strength & direction of linear association
• Linear Regression: calculates best-fitting line
  - predicts average change in Y associated with change in X

What is the equation for the line shown below?

\[ \hat{Y} = bX + a \]

• \( a \) (intercept) = ?
• \( b \) (slope of line) = ?

\[ a = 2 \]
\[ b = \frac{\Delta Y}{\Delta X} = \frac{2}{4} = 0.5 \]

\( \hat{Y} = bX + a \)

\( \hat{Y} \): predicted value of Y
\( b \): slope of regression line; average change in Y with 1-unit change in X
\( a \): intercept of regression line (\( \hat{Y} \) when X=0)

An increase of 1 cm of Petal Width is associated with an average increase of 0.888 cm in Sepal Length.
A Petal Width of zero is associated with a Sepal Length of 4.78 cm. (Is Petal Width of zero meaningful?)
Regression slope ≠ correlation strength
Pearson r is NOT a good measure of regression slope

- 2 methods of test preparation
- For both methods:
  - test mark (Y) is correlated with study time (X)
  - correlations are approx equal & same sign
  - linear X,Y association very strong
- Regression can help to answer such questions as:
  - Which method yields a greater increase in Marks per hour of study time?
  - How long should you study to get a predicted mark of 85?

Where does the regression line come from?

Example: Relation between mental health & stress
Permutation test on (Stress, Symptom) data

- Dashed lines form 99.9% interval:
  - 99.9% of simulated r values fall inside interval
  - 0.1% (p=0.001) of simulated r values fall outside interval
- Use interval to define “unusual” r values
- Observed correlation r=0.506 falls outside interval
  - permutation test suggests 0.506 is a very unusual sample r when population r is zero
- We may decide to reject null hypothesis that population r=0 in favour of alternative hypothesis population r≠0
  - Our decision might be wrong!
- We state conclusion thusly:
  - the correlation was significant (r=.506, p<.001)

N.B. This permutation test illustrates the logic of null hypothesis testing. In fact, other statistical methods provide better tests of the null hypothesis r=0 by estimating the interval boundaries more accurately. But the logic behind those methods is similar to that described here.

Least-squares regression line

- find values of slope & intercept that minimize sum of squared residuals
  - this criterion is reasonable because it minimizes differences between observed and predicted values
- linear regression provides values of regression coefficients for best-fitting line
  - best-fitting line minimizes the sum of squared residuals

What is “best” fitting line?

- residual = difference between observed and predicted value of Y
  \[ e_i = Y_i - \hat{Y}_i \]
- “best” fit: line that minimizes the sum of squared residuals
  \[ \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 \]
- “least-squares” fit

\[ \hat{Y} = bX + a \]

Y-hat: predicted value of Y
- \( b = \text{slope} \) of regression line
  - average change in Y associated with a 1-unit change in X
- \( a = \text{intercept} \) of regression line
  - value of Y-hat when \( X=0 \)

An increase of 1 on the Stress measure is associated with an increase of 0.78 (on average) on the Symptom measure

A Stress score of zero is associated with a Symptom score of 73.89.
### Regression Table

#### Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 73.88    | 3.2714     | 22.587  | < 2e-16 *** |
| Stress     | 0.78     | 0.1303     | 6.012   | 2.69e-08 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Standard Error of Estimate:** 17.56 on 105 degrees of freedom

Multiple R-squared: 0.2561, Adjusted R-squared: 0.249

F-statistic: 36.14 on 1 and 105 DF, p-value: 2.692e-08

---

Standard Error of Estimate = Standard Deviation of Y around the regression line

Standard Error of Estimate = Standard Deviation of Residuals

Variance(Residuals) = (Standard Error of Estimate)^2

A measure of uncertainty about the predicted value of Y

---

```r
> lm.stress <- lm(Symptoms~Stress,data=tab10.2.dat)  # calculate regression line
> var(residuals(lm.stress))  # variance of residuals
[1] 305.5288
> sqrt(var(residuals(lm.stress)))  # standard deviation of residuals
[1] 17.47938
```

---

### Standard Error of Coefficients

Coefficients estimated with high or low precision

#### Effect of Varying Slope on Goodness of Fit

- In other cases, varying the slope has a large effect on goodness of fit, so a very small range of slopes fit the data well... leading to a small standard error.
- In some cases, varying the slope has relatively small effects on goodness of fit, so a range of slopes fit the data nearly equally well... leading to a large standard error.

---

Multiple R-squared: 0.2561, Adjusted R-squared: 0.249

F-statistic: 36.14 on 1 and 105 DF, p-value: 2.692e-08

Multiple R-squared: proportion of the variance of Y that is "accounted for" by regression
- regression line shows how variation in X accounts for variation in Y
- if fit is very good, then variation in X should account for most of variation in Y
- For linear regression, Multiple R^2 equals the squared correlation (r^2 = 0.506^2 = 0.256)

\[
R^2 = \frac{\text{VAR}(Y) - \text{VAR}(Y - \hat{Y})}{\text{VAR}(Y)} = \frac{\text{VAR}(Y) - \text{VAR(Residuals)}}{\text{VAR}(Y)}
\]
Var(residuals) < Var(Scores)

Var(Symptoms) = 410.7; SD(Symptoms) = 20.26
Var(Residuals) = 305.5; SD(Residuals) = 17.48

\[ R^2 = \left( \frac{410.7 - 305.5}{410.7} \right) = 0.256 \]

\[ r = \sqrt{R^2} = \sqrt{0.256} = 0.506 \]

Our regression model accounts for 25% of variance of Symptoms scores

---

Standardized Scores

- Any set of numbers can be standardized:
  - if \( z = \frac{Y - \text{mean}(Y)}{\text{SD}(Y)} \), then mean(z)=zero & SD(z) = 1
  - transformed scores like \( z \) often called standard scores
- Standard scores are unit-less:
  - represent # of SDs above/below the mean
- Regression on standardized scores:
  - correlation, \( r \), is unchanged
  - but values of regression coefficients are different
  - slope coefficient equals \( r \)

---

Regression Table

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 73.88 | 3.2714 | 22.587 | < 2e-16 *** |
| Stress | 0.78 | 0.1303 | 6.012 | 2.69e-08 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Standard Error of Estimate: 17.56 on 105 degrees of freedom
Multiple R-squared: 0.2561, Adjusted R-squared: 0.249

F-statistic: 36.14 on 1 and 105 DF, p-value: 2.692e-08

F-statistic evaluates Null Hypothesis Y & X are not associated in population
p-value is probability of obtaining \( R^2 \) at least as large as ours when Null Hypothesis is true
For linear regression: if \( r \) is statistically significant, then F-statistic and regression coefficient for Stress (X) will be statistically significant, too

---

Regression table for raw scores

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.00 | 0.08378 | 0.000 | 1 |
| z.stress | 0.506 | 0.08417 | 6.012 | 2.69e-08 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8666 on 105 degrees of freedom
Multiple R-squared: 0.2561, Adjusted R-squared: 0.249
F-statistic: 36.14 on 1 and 105 DF, p-value: 2.692e-08

---

Regression table for standardized scores

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.0 | 0.08378 | 0.000 | 1 |
| z.stress | 0.506 | 0.08417 | 6.012 | 2.69e-08 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8666 on 105 degrees of freedom
Multiple R-squared: 0.2561, Adjusted R-squared: 0.249
F-statistic: 36.14 on 1 and 105 DF, p-value: 2.692e-08

---

• \( R^2 \) is unchanged
• But values of regression coefficients differ
• Note value of z.stress coefficient:
  - Standardized coefficient referred to as “Beta”
  - Beta = \( \beta = r = 0.506 \)
Regression On **Standard Scores**

Standard Scores (Symptoms vs. Stress)

![Graph showing regression line with slope Beta = r = 0.506](image)

Quantifying (X,Y) Linear Association

- $R^2 = r^2$ equals the proportion of Y variance “accounted for” by regression
- $R = \text{correlation ($r$) between } Y \text{ and } \hat{Y} \text{ (i.e., predicted value of } Y)$
- $b = \text{regression slope coefficient}$
- change in $Y$ associated with 1-unit change in $X$
- **Beta ($\beta$):**
  - let $Z_Y$ and $Z_X$ represent standardized scores of $Y$ & $X$
  - Beta ($\beta$) = number of SDs of $Z_Y$ associated with 1 SD change in $Z_X$
  - useful when comparing coefficients for variables with different units and/or variances

Multiple Regression: A Generalization of Linear Regression

- Linear Regression can be generalized to:
  - cases that use more than 1 predictor variable
  - accounting for curvilinear relations between $Y$ and predictors

Multiple Regression

- Dotted line fit by Linear Regression
- Solid lines fit by **Multiple Regression**
  - this multiple regression analysis includes Dosage and Gender
Multiple Regression: extension to curvilinear association

- **Not just linear fits!**
- Curvilinear associations can be estimated & evaluated with Multiple Regression
- **two predictor variables:** $X$ & $X^2$

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 2.967    | 0.092      | 32.39   | <2e-16   *** |
| $x$           | 0.513    | 0.037      | 13.69   | 2.82e-16 *** |
| $x^2$         | 0.101    | 0.013      | 7.86    | 1.72e-09 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4124 on 38 degrees of freedom
Multiple R-squared: 0.8349, Adjusted R-squared: 0.8262
F-statistic: 96.06 on 2 and 38 DF, p-value: 1.375e-15

Comments on Multiple Regression

- Linear Regression can be **generalized to:**
  - more than 1 predictor variable
  - curvilinear relations between $Y$ and regression equation
- Like Linear Regression, Multiple Regression uses **Least-Squares:**
  - finds coefficients that minimize sum of squared residuals
- $R$: still represents correlation between $Y$ and predicted-$Y$ values
- $R^2$: still is the proportion of $Y$ variance accounted for by predictor variables
- regression coefficients represent the change in $Y$ (on average) associated with 1-unit change in one predictor variable **when all other predictors are held constant**

Linear Regression Summary

- Least-squares criterion to find best-fitting line
  - line defined by intercept & slope coefficients
- Regression slope ≠ Correlation
- Regression, like $r$, is sensitive to extreme (high-leverage) data points
- $R^2 =$ proportion of $Y$ variance accounted for by regression
- **Beta:** slope coefficient for regression on **standardized scores**
  - Beta = $r$
- Multiple Regression is a generalization of Linear Regression
  - can include multiple predictors
  - can estimate curvilinear association btwn $Y$ & predictor variables