Chapter 5: Linear regression

Last lecture: Ch 4

- Transformations (Ch 4)
- Advantage: transformations can help satisfy the assumptions of linearity, constant variance and normality.
- Disadvantage: interpretation is more difficult.
- The family of powers and roots (X^p or $(X^p - 1)/p$):
  - Ascending the ladder of powers ($p > 1$) spreads out large values and compresses small values.
  - Descending the ladder of powers ($p < 1$) does the opposite.
- The family of folded powers (and in particular the logit transformation $\log(X/(1 - X))$) can be used for proportions. It remedies stacking up of the data against the boundaries.

Next: Ch 5

- We've seen that linear regression has its limitations. However, it is worth studying linear regression because:
  - Sometimes data (nearly) satisfy the assumptions.
  - Sometimes the assumptions can be (nearly) satisfied by transforming the data.
  - There are many useful extensions of linear regression: weighted regression, robust regression, nonparametric regression, and generalized linear models.
- How does linear regression work? We start with one independent variable.

Simple linear regression

Linear model (1)

- Linear statistical model: $Y = \alpha + \beta X + \epsilon$.
- α is the intercept of the line, and β is the slope of the line. One unit increase in X gives β units increase in Y. (see figure on blackboard)
- ϵ is called a statistical error. It accounts for the fact that the statistical model does not give an exact fit to the data.
- Statistical errors can have a fixed and a random component.
  - Fixed component: arises when the true relation is not linear (also called lack of fit error, bias) - we assume this component is negligible.
  - Random component: due to measurement errors in Y, variables that are not included in the model, random variation.

Linear model (2)

- Data $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Then the model gives: $Y_i = \alpha + \beta X_i + \epsilon_i$, where ϵ_i is the statistical error for the ith case.
- Thus, the observed value Y_i equals $\alpha + \beta X_i$, except that ϵ_i, an unknown random quantity is added on.
- The statistical errors ϵ_i cannot be observed. Why?
- We assume:
  - $E(\epsilon_i) = 0$
  - Var$(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$
  - Cov$(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

Linear model (3)

- The *population parameters* α, β and σ are unknown. We use lower case Greek letters for population parameters.
- We compute *estimates* of the population parameters: A, B and S_E. We use capital case roman letters for estimates.
- $\hat{Y}_i = A + BX_i$ is called the *fitted value*. (see figure on blackboard)
- $E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$ is called the *residual*.
- The residuals are observable, and can be used to check assumptions on the statistical errors ϵ_i.
- Points above the line have positive residuals, and points below the line have negative residuals.
- A line that fits the data well has small residuals.

Small residuals

- We want the residuals to be small in *magnitude*, because large negative residuals are as bad as large positive residuals.
- So we cannot simply require $\sum E_i = 0$.
- In fact, any line through the means of the variables - the point $(\bar{X}, \bar{Y})$ - satisfies $\sum E_i = 0$ (derivation on board).
- Two immediate solutions:
  - Require $\sum |E_i|$ to be small.
  - Require $\sum E_i^2$ to be small.
- We consider the second option because working with squares is mathematically easier than working with absolute values (for example, it is easier to take derivatives). However, the first option is more resistant to outliers.
- Eyeball regression line (see overhead).

Minimize $\sum E_i^2$

- We call $RSS(A,B) = \sum E_i^2$ the *Residual Sum of Squares*.
- We want to find the pair (A,B) that minimizes $RSS(A,B) = \sum E_i^2 = \sum (Y_i - A - BX_i)^2$.
- Thus, we set the partial derivatives of $RSS(A,B)$ with respect to A and B equal to zero:
  - $\frac{\partial RSS(A,B)}{\partial A} = \sum(-1)(2)(Y_i - A - BX_i) = 0$
    $\Rightarrow \sum(Y_i - A - BX_i) = 0.$
  - $\frac{\partial RSS(A,B)}{\partial B} = \sum(-X_i)(2)(Y_i - A - BX_i) = 0$
    $\Rightarrow \sum X_i(Y_i - A - BX_i) = 0.$
- We now have two *normal equations* with two unknowns A and B. The solution is (derivation on board):
  - $A = \bar{Y} - B\bar{X}$
  - $B = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

Properties of residuals

- $\sum E_i = 0$, since the regression line goes through the point $(\bar{X}, \bar{Y})$.
- $\sum X_i E_i = 0$ and $\sum \hat{Y}_i E_i = 0$. $\Rightarrow$ The residuals are uncorrelated with the independent variables X_i and with the fitted values $\hat{Y}_i$ (homework).
- Least squares estimates are uniquely defined as long as the values of the independent variable are not all identical. In that case the numerator $\sum(X_i - \bar{X})^2 = 0$ (draw figure).

Regression in R

- `model <- lm(y ~ x)`
- `summary(model)`
- Coefficients: `model$coef` or `coef(model)`
  (Alias: `coefficients`)
- Fitted mean values: `model$fitted` or `fitted(model)`
  (Alias: `fitted.values`)
- Residuals: `model$resid` or `resid(model)`
  (Alias: `residuals`)

R output - Davis data

```
> model <- lm(weight ~ repwt)
> summary(model)

Call: lm(formula = weight ~ repwt)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5248 -0.7526 -0.3654  0.6118  6.3841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.77750    1.74441   1.019    0.311
repwt        0.97722    0.03053  32.009   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.057 on 99 degrees of freedom
Multiple R-Squared: 0.9119,     Adjusted R-squared: 0.911
F-statistic: 1025 on 1 and 99 DF,  p-value: < 2.2e-16
```


How good is the fit?

S_E

- Residual standard error: $S_E = \sqrt{RSS/(n-2)} = \sqrt{\frac{\sum E_i^2}{n-2}}$.
- $n-2$ is the degrees of freedom (we lose two degrees of freedom because we estimate the two parameters α and β).
- For the Davis data, $S_E \approx 2$. Interpretation:
    - on average, using the least squares regression line to predict weight from reported weight, results in an error of about 2 kg.
    - If the residuals are approximately normal, then about 2/3 is in the range ± 2 and about 95% is in the range ± 4.

R^2 (1)

- We compare our fit to a *null model* $Y = \alpha' + \epsilon'$, in which we don't use the independent variable X.
- We define the fitted value $\hat{Y}_i' = A'$, and the residual $E_i' = Y_i - \hat{Y}_i'$.
- We find A' by minimizing $\sum E_i'^2 = \sum (Y_i - A')^2$. This gives $A' = \bar{Y}$.
- Note that $\sum (Y_i - \hat{Y}_i)^2 = \sum E_i^2 \leq \sum E_i'^2 = \sum (Y_i - \bar{Y})^2$ (why?).

R^2 (2)

- Total sum of squares: $TSS = \sum E_i'^2 = \sum (Y_i - \bar{Y})^2$.
- Residual sum of squares: $RSS = \sum E_i^2 = \sum (Y_i - \hat{Y}_i)^2$.
- Regression sum of squares: $RegSS = TSS - RSS$ gives *reduction* in squared error due to the linear regression.
- $R^2 = RegSS/TSS = 1 - RSS/TSS$ is the *proportional reduction* in squared error due to the linear regression.
- Thus, R^2 is the proportion of the variation in Y that is explained by the linear regression.

16 / 41

Analysis of variance

- $\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$ (homework)
- $RegSS = \sum (\hat{Y}_i - \bar{Y})^2$ (derivation on board, use $TSS = RegSS + RSS$)
- Hence, $\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$. This decomposition is called *analysis of variance*.

17 / 41

r

- Correlation coefficient $r = \pm\sqrt{R^2}$ (take positive root if $B > 0$ and take negative root if $B < 0$).
- r gives the strength and direction of the relationship.
- Alternative formula: $r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$.
- Using this formula, we can write $B = r\frac{SD_Y}{SD_X}$ (derivation on board).
- In the 'eyeball regression', the steep line had slope $\frac{SD_Y}{SD_X}$, and the other line had the correct slope $r\frac{SD_Y}{SD_X}$.
- r is symmetric in X and Y.
- r has no units $\Rightarrow$ doesn't change when scale is changed (homework).

18 / 41

≥ 2 independent variables

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
- This describes a plane in the 3-dimensional space $\{X_1, X_2, Y\}$ (see figure):
  - α is the intercept
  - β_1 is the increase in Y associated with a one-unit increase in X_1 when X_2 is held constant
  - β_2 is the increase in Y for a one-unit increase in X_2 when X_1 is held constant.

Statistical error

- Data: $(X_{11}, X_{12}, Y_1), \ldots, (X_{n1}, X_{n2}, Y_n)$.
- $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$, where ϵ_i is the statistical error for the ith case.
- Thus, the observed value Y_i equals $\alpha + \beta_1 X_{i1} + \beta_2 X_{i2}$, except that ϵ_i, an unknown random quantity is added on.
- We make the same assumptions about ϵ as before:
  - $E(\epsilon_i) = 0$
  - Var$(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$
  - Cov$(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

Estimates and residuals

- The *population parameters* α, β_1, β_2, and σ are unknown.
- We compute *estimates* of the population parameters: A, B_1, B_2 and S_E.
- $\hat{Y}_i = A + B_1 X_{i1} + B_2 X_{i2}$ is called the *fitted value*.
- $E_i = Y_i - \hat{Y}_i = Y_i - (A + B_1 X_{i1} + B_2 X_{i2})$ is called the *residual*.
- The residuals are observable, and can be used to check assumptions on the statistical errors ϵ_i.
- Points above the *plane* have positive residuals, and points below the plane have negative residuals.
- A plane that fits the data well has small residuals.

Computing estimates

- The triple (A, B_1, B_2) minimizes $RSS(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1 X_{i1} - B_2 X_{i2})^2$.
- We can again take partial derivatives and set these equal to zero.
- This gives three equations in the three unknowns A, B_1 and B_2. Solving these *normal equations* gives the regression coefficients A, B_1 and B_2.
- Least squares estimates are unique unless one of the independent variables is invariant, or independent variables are perfectly collinear.
- The same procedure works for k independent variables $X_1, \ldots, X_k$. However, it is then easier to use matrix notation.
- In R: `model <- lm(y ~ x1 + x2)`

23 / 41

Properties of residuals

- $\sum E_i = 0$
- The residuals E_i are uncorrelated with the fitted values $\hat{Y}_i$ and with each of the independent variables $X_1, \ldots, X_k$.
- The standard error of the residuals $S_E = \sqrt{\sum E_i^2/(n-k-1)}$ gives the "average" size of the residuals.
- $n - k - 1$ is the *degrees of freedom* (we lose $k+1$ degrees of freedom because we estimate the $k+1$ parameters α, $\beta_1, \ldots, \beta_k$).

24 / 41

R^2 and $\tilde{R}^2$

- $TSS = \sum (Y_i - \bar{Y})^2$.
- $RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum E_i^2$.
- $RegSS = TSS - RSS = \sum (\hat{Y}_i - \bar{Y})^2$.
- $R^2 = RegSS/TSS = 1 - RSS/TSS$ is the proportion of variation in Y that is captured by its linear regression on the X's.
- R^2 can never decrease when we add an extra variable to the model (homework).
- Corrected sum of squares: $\tilde{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$ penalizes R^2 when there are extra variables in the model.
- R^2 and $\tilde{R}^2$ differ very little if sample size is large.

25 / 41

Ozone example

- Data from Sandberg, Basso, Okin (1978):
  - SF = Summer quarter maximum hourly average ozone reading in parts per million in San Francisco
  - SJ = Same, but then in San Jose
  - YEAR = Year of ozone measurement
  - RAIN = Average winter precipitation in centimeters in the San Francisco Bay area for the preceding two winters
- Research question: How does SF depend on YEAR and RAIN?
- Think about assumptions: Which one may be violated?

27 / 41

Ozone data

```
YEAR RAIN SF  SJ
1965 18.9 4.3 4.2
1966 23.7 4.2 4.8
1967 26.2 4.6 5.3
1968 26.6 4.7 4.8
1969 39.6 4.1 5.5
1970 45.5 4.6 5.6
1971 26.7 3.7 5.4
1972 19.0 3.1 4.6
1973 30.6 3.4 5.1
1974 34.1 3.4 3.7
1975 23.7 2.1 2.7
1976 14.6 2.2 2.1
1977  7.6 2.0 2.5
```

28 / 41

R output

```
> model <- lm(sf ~ year + rain)
> summary(model)

Call: lm(formula = sf ~ year + rain)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61072 -0.20317  0.06129  0.16329  0.51992

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 388.412083  49.573690   7.835 1.41e-05 ***
year         -0.195703   0.025112  -7.793 1.48e-05 ***
rain          0.034288   0.009655   3.551  0.00526 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3224 on 10 degrees of freedom
Multiple R-Squared: 0.9089,     Adjusted R-squared: 0.8906
F-statistic: 49.87 on 2 and 10 DF,  p-value: 6.286e-06
```


Standardized coefficients

Standardized coefficients

- We often want to compare coefficients of different independent variables.
- When the independent variables are measured in the same units, this is straightforward.
- If the independent variables are not commensurable, we can perform a *limited* comparison by rescaling the regression coefficients in relation to a measure of variation:
  - using hinge spread
  - using standard deviations

Using hinge spread

- Hinge spread = interquartile range (IQR)
- Let IQR_Y be the IQR of Y, and let $IQR_1, \ldots, IQR_k$ be the IQRs of $X_1, \ldots, X_k$.
- We start with $Y_i = A + B_1X_{i1} + \ldots B_kX_{ik} + E_i$.
- This can be rewritten as (derivation on board):
  $Y_i = A + (B_1IQR_1)\frac{X_{i1}}{IQR_1} + \cdots + (B_kIQR_k)\frac{X_{ik}}{IQR_k} + E_i$.
- Let $Z_{ij} = \frac{X_{ij}}{IQR_j}$, for $j = 1, \ldots, k$ and $i = 1, \ldots, n$.
- Let $B_j^* = B_jIQR_j$, $j = 1, \ldots, k$.
- Then we get $Y_i = A + B_1^*Z_{i1} + \cdots + B_k^*Z_{ik} + E_i$.
- $B_j^* = B_jIQR_j$ is called the *standardized regression coefficient*.

Interpretation

- Interpretation: Increasing Z_j by 1 and holding constant the other Z_ℓ's ($\ell \neq j$), is associated, on average, with an increase of B_j^* in Y.
- Increasing Z_j by 1, means that X_j is increased by one IQR of X_j.
- So increasing X_j by one IQR of X_j and holding constant the other X_ℓ's ($\ell \neq j$), is associated, on average, with an increase of B_j^* in Y.
- Ozone example:

Variable	Coefficient B_j	Hinge spread	Stand. coeff. B_j^*
Year	-0.196	6	-1.176
Rain	0.034	11.6	0.394

Using st.dev.

- Let S_Y be the standard deviation of Y, and let $S_1, \ldots, S_k$ be the standard deviations of $X_1, \ldots, X_k$.
- We start with $Y_i = A + B_1 X_{i1} + \ldots B_k X_{ik} + E_i$.
- This can be rewritten as (derivation on board):
$\frac{Y_i - \bar{Y}}{S_Y} = \left(B_1 \frac{S_1}{S_Y}\right) \frac{X_{i1} - \bar{X}_1}{S_1} + \cdots + \left(B_k \frac{S_k}{S_Y}\right) \frac{X_{ik} - \bar{X}_k}{S_k} + \frac{E_i}{S_Y}$.
- Let $Z_{iY} = \frac{Y_i - \bar{Y}}{S_Y}$ and $Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}$, for $j = 1, \ldots, k$.
- Let $B_j^* = B_j \frac{S_j}{S_Y}$ and $E_i^* = \frac{E_i}{S_Y}$.
- Then we get $Z_{iY} = B_1^* Z_{i1} + \cdots + B_k^* Z_{ik} + E_i^*$.
- $B_j^* = B_j \frac{S_j}{S_Y}$ is called the *standardized regression coefficient*.

Interpretation

- Interpretation: Increasing Z_j by 1 and holding constant the other Z_ℓ's ($\ell \neq j$), is associated, on average, with an increase of B_j^* in Z_Y.
- Increasing Z_j by 1, means that X_j is increased by one SD of X_j.
- Increasing Z_Y by 1 means that Y is increased by one SD of Y.
- So increasing X_j by one SD of X_j and holding constant the other X_ℓ's ($\ell \neq j$), is associated, on average, with an increase of B_j^* SDs of Y in Y.

Ozone example

- Ozone example:

Variable	Coeff.	$\frac{\text{St.dev(variable)}}{\text{St.dev(Y)}}$	Stand. coeff.
Year	-0.196	3.99	-0.783
Rain	0.034	10.39	0.353

- Both methods (using hinge spread or standard deviations) only allow for a *very limited* comparison. They both assume that predictors with a large spread are more important, and that does not need to be the case.

36 / 41

Added variable plots

37 / 41

Added variable plots

- Suppose we start with SF $\sim$ YEAR
- We want to know whether it is helpful to add the variable RAIN
- We want to model that part of SF that is not explained by YEAR (residuals of `lm(SF ~ YEAR)` with the part of RAIN that is not explained by YEAR (residuals of `lm(RAIN ~ YEAR)`
- Plotting these residuals against each other is called an *added variable plot* for the effect of RAIN on SF, controlling for YEAR.
- Regressing residuals of `lm(SF ~ YEAR)` on the residuals of `lm(RAIN ~ YEAR)` gives the coefficient for RAIN when controlling for YEAR.

38 / 41

Summary

39 / 41

Summary (1)

- Linear statistical model: $Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$.
- We assume that the statistical errors ϵ have mean zero, constant standard deviation σ, and are uncorrelated.
- The *population parameters* α, $\beta_1, \ldots, \beta_k$ and σ cannot be observed. Also the statistical errors ϵ cannot be observed.
- We define the *fitted value* $\hat{Y}_i = A + B_1 X_{i1} + \cdots + B_k X_{ik}$ and the residual $E_i = Y_i - \hat{Y}_i$. We can use the residuals to check the assumptions about the statistical errors.
- We compute estimates $A, B_1, \ldots, B_k$ for $\alpha, \beta_1, \ldots, \beta_k$ by minimizing the *residual sum of squares* $RSS = \sum E_i^2 = \sum (Y_i - (A + B_1 X_{i1} + \cdots + B_k X_{ik}))^2$.
- Interpretation of the coefficients?

40 / 41

Summary (2)

- To measure how good the fit is, we can use:
  - the residual standard error $S_E = \sqrt{RSS/(n-k-1)}$
  - the multiple correlation coefficient R^2
  - the adjusted multiple correlation coefficient $\tilde{R}^2$
  - the correlation coefficient r
- Analysis of variance (ANOVA): $TSS = RegSS + RSS$
- Standardized regression coefficients
- Added variable plots (partial regression plots)