Linear Regression II

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

Topics

Introduction

- We have covered regression models with a single independent (explanatory) variable.
- In practice, more than one independent variable may explain changes in the dependent variable.
- We are going to focus on the case of multiple numeric explanatory (independent) variables.

Multiple Linear Regression

- **Multiple linear regression models** contain more than one independent variable.
- Multiple linear regression models are used in many different real-world applications.
- Example: There may be multiple variables that explain someone's income level.

Nature of the Relationship

- One variable (Y) is the dependent (response) variable and other variables play the role of independent (explanatory) variables (X_1, X_2, ..., X_k)
- The relationship is not deterministic (functional) but is **statistical** (stochastic).
- There is a (conditional) distribution of the dependent variable associated with various combinations of independent (explanatory) variables.
- Initially we will focus on **linear** relationships.

Multiple Linear Regression Models

- Linear regression is based on estimating the linear relationship between the dependent and independent variables **plus an error term**.
- **Multiple Linear Regression Model** (Expected value of Y):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_2 X_2 + \epsilon. \quad (1)$$

  - Y is the dependent variable.
  - β_0 is the intercept.
  - X_1, X_2, ..., X_k are the independent variables.
  - β_1, β_2, ..., β_k are the regression coefficients for the independent variables.
  - ϵ is the random error term.
    - Follows an assumed distribution with $E[\epsilon] = 0$ and constant variance σ_ϵ^2

Estimation

- We do not know the true values of β_0, β_1, β_2, ..., β_k because we do not have the entire population.
- We need to *estimate* these parameters the best we can using the data we do have.
- We want to minimize the sum of the squared residuals (observed errors).

Estimated Regression Model

- **The estimated linear regression equation is**:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k. \quad (2)$$

  - b_0, b_1, b_2, ..., b_k, are estimates of β_0, β_1, β_2, ..., β_k.

- Where $\hat{Y}_i$ is the fitted (expected) value of Y_i.

Interpretation

- We can use expected value $\hat{Y}$ to *predict* the value of the dependent variable for any combination of the independent variables.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k.$$

- **The regression coefficients represent the expected change in the dependent variable when the associated independent variable is increased by one unit.**
  - *While the values of other independent variables are held constant.*

Parameter Estimates in R

- We can estimate linear regression models in R using:
  - `lm1 <- lm(formula = y.variable ~ x.variable1 + x.variable2 + ... + x.variablek, data = data.frame)`
- *We add more independent variables to the equation using +*
- To view the results of your model:
  - `summary(lm1)`

Example 1

- Import the *wage2.csv* file into R and take some time to understand what the variables are. Then, complete the following tasks (*wage* will always be out dependent variable):
  1. Visualize the relationships between *wage* (Y) and *age* (X_1), *educ* (X_2), *hours* (X_3), and *exper* (X_4).
  2. Create **simple** linear regression models for all of the pairs of variables from part 1.
  3. Create a multiple linear regression model with all of the independent variables mentioned in part 1.
- Comment on your findings.

Model Summary

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25316 -0.29087  0.03779  0.36510  1.16111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09559    0.07781   1.229    0.225
X            1.00891    0.04054  24.885   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5415 on 53 degrees of freedom
Multiple R-squared:  0.9212,    Adjusted R-squared:  0.9197
F-statistic: 619.3 on 1 and 53 DF,  p-value: < 2.2e-16
```

Coefficient Estimates

Significance of coefficient estimates

Percentage of variance in Y explained by X

Model significance. Is this model better than an empty model (no explanatory variables)

Coefficient of Determination

- The **coefficient of determination (R^2)** is the proportion of the variation in the dependent variable that is *accounted for* by the independent variables.

$$R^2 = \frac{SS_{res}}{SS_{tot}}. \tag{3}$$

Where

$$SS_{res} = \sum_{i=1}^{n} e_i^2 \quad \text{Residual sum of squares}$$

and

$$SS_{res} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{Total sum of squares}$$

Adjusted R^2

- The R^2 value will increase as we add more independent variables to regression model even if they do not account for any of the variability.

- The **adjusted** R^2 value only increases when potential explanatory variables explain part of the variability in the dependent variable.

$$\bar{R}^2 = 1 - \frac{SS_{res}/df_{res}}{SS_{tot}/df_{tot}}. \tag{4}$$

Where $df_{res} = n - p$ $(p = k + 1)$ and $df_{tot} = n - 1$

- It can be re-written as:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Example 2

- Comment on the R^2 and adjusted R^2 ($\bar{R}^2$) values in the models you created in Example 1.
- Does adding the variables one at a time improve these values?

ANOVA Test

- **In the context of linear regression we use an ANOVA test to test the significance of an entire model.**

- We compute an F-statistic to test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \text{at least one } \beta_j \text{ is not } 0.$$

- Under the null hypothesis, no linear relationship exists between the dependent and any of the independent variables.

Coefficient Significance

- We use two-tailed one sample t-tests to determine the significance of individual independent variables:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

- Under the null hypothesis, the slope associated with variable j is equal to zero (no linear relationship).

Example 3

- Comment on the results of the ANOVA test and the individual t-test values in the models you created in Example 1.
- Do you think that all of these variables are appropriate?

Correlation Review

- The **correlation** is a measure of the direction and strength of the *linear association* between two random variables.
  - The correlation coefficient takes on values between -1 and 1.
  - If $\rho = -1$ then there is a perfect (negative) linear relationship between X and Y.
  - If $\rho = 1$ then there is a perfect (positive) linear relationship between X and Y.
  - If ρ is close to 0 then there is a weak linear relationship between X and Y.
- The strength of the linear association increases as ρ moves away from 0.

Sample Correlation in R

- R function: `cor(x, y, method = "pearson")`
  - `x`: a numeric vector, matrix, or data frame.
  - If `x` is a vector we must give a vector `y`.
  - `method = "pearson"` is the default method.

- `method = "kendall"` to measure the ordinal association between two measured quantities.
- `method = "spearman"` (rank correlation) to assess monotonic relationships between two measured quantities.

Correlation Matrix

Variable	X	Y	Z
X	$\rho_{XX} = 1$	ρ_{XY}	ρ_{XZ}
Y	ρ_{XY}	$\rho_{YY} = 1$	ρ_{YZ}
Z	ρ_{XZ}	ρ_{YZ}	$\rho_{ZZ} = 1$

Multicollinearity

- **Multicollinearity** occurs when two or more independent variables contain high levels of the same information.
- It becomes difficult to isolate the effect of one independent variable on the dependent variable.
- Can result in incorrect conclusions being drawn from model summaries.
- Remember: The fourth assumption about our regression model is independence.

Variance Inflation Factor VIF

- We can use the Durbin-Watson or Ljung-Box test to test for independence.
- We can also measure multicollinearity using the variance inflation factor (VIF) for each independent variable.
- If the independent variables are not correlated then $\text{VIF}_j = 1$ (approximately).
- Rule of thumb suggests that VIF values should be less than 5 to indicate that multicollinearity is not an issue. **We may remove variables that exceed this threshold.**

VIF in R

- We can use the `vif(lm1)` function from the *car* package.
- *Rule of thumb suggests that VIF values should be less than 5 to indicate that multicollinearity is not an issue.*
- Visualize VIF:

```
barplot(vif(lm1), main = "VIF Values", horiz = TRUE,
col = "steelblue")
abline(v = 5, lwd = 3, lty = 2)
```

Example 4

- Comment on the VIF values from the multiple linear regression model you created in Example 1.
- Do you think that all of these variables are appropriate?

Multiple Linear Regression Assumptions

- **The multiple linear regression models must satisfy the same assumptions as the simple linear regression models!**

Model Assumptions

- **The validity of the significance of our regression model estimates depends on some key assumptions:**

1. **Linearity**
   - The relationship between the dependent and independent variable(s) needs to be linear.

2. **Normality** (multivariate normal for multiple independent variables)
   - In linear regression, all variables must be normally distributed (can be fixed).

3. **Homoscedasticity**
   - The variation about the regression line is constant for all values of the independent variable(s) (can be fixed).

4. **Independence**
   - There is little or no multicollinearity in the data (independent variables are too highly correlated with each other).

Diagnostics

- Due to the assumptions imposed on the error term (closely related to the residuals) we can use the residuals to help us check the model assumptions.
- We can also *standardize* the residuals to help with outlier identification and assumption checks.
- **Standardized residuals**: $\text{Residual}_i / \text{Standard Deviation of Residual}_i$
- In R:
  - `rstandard(linear.model)` from the *car* package

Linearity Check

- The **linearity** assumption may be checked in a few different ways:
  1. Examine the scatterplot(s) of the dependent and independent variable(s) (linear relationship?).
  2. Plot the residuals, they should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(lm1$residuals)`
     - Add line at 0: `abline(h=0,col="blue")`
  3. Plot residuals vs fitted values, again should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(model, 1)`

Normality Check

- The **normality** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the histogram of the *standardized residuals* for approximate normality.
  2. Q-Q plot of the *standardized residuals*.
     - In R we can also use: `plot(model, 2)`
  3. We can use a K-S test or a Shapiro–Wilk test on the *standardized residuals*.
     - K-S: `ks.test(standardized.residuals, "pnorm")`
     - Shapiro–Wilk: `shapiro.test(standardized.residuals)`
- *Slight departures from normality may be acceptable and we can also take steps to fix departures from normality.*

Homoscedasticity Check

- The **Homoscedasticity** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the scatterplot(s) of the *standardized residuals* for a constant variance.
  2. Examine the scatterplot(s) of the fitted values and the square root of the *standardized residuals* (*scale-location plot*). (It's good if you see a horizontal line with equally spread points)
     - In R: `plot(model, 3)`
  3. We can use the `ncvTest()` similar to a (Breusch–Pagan test) in R with the null hypothesis being a constant variance (Homoscedasticity).
     - In R (*car* package): `ncvTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume Homoscedasticity.
- *We can also take steps to fix departures from Homoscedasticity.*

Independence Check

- *This assumption may be violated if we have time-dependent independent (explanatory) variables.*

- The **Independence** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the scatterplot(s) of the *standardized residuals* for patterns or obvious clusters.

  2. We can use the Durbin Watson test in R with the null hypothesis being that the residuals are independent.
     - In R (*car* package): `durbinWatsonTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume independence at one lag.

- *There are other tests like the Ljung-Box test that may be used for multiple lags.*

Example 5

- Using the *wage2.csv* data estimate the following regression models, complete all of the required diagnostics, remove any variables as required to improve your models, comment on the estimates, and comment on the fit ($\bar{R}^2$):
  1. `wage` $\sim$ `age + IQ + educ + meduc`
  2. `wage` $\sim$ `hours + tenure + sibs + educ`
  3. `wage` $\sim$ `KWW + IQ + age + meduc + exper`
  4. `wage` $\sim$ `age + IQ + educ + meduc + hours + tenure + KWW + meduc + feduc + exper`

Exercise 1

- Using the *wage2.csv* data estimate a regression model with all of the numeric variables included (exclude *lwage*), complete all of the required diagnostics, remove any variables as required to improve your models, comment on the estimates, and comment on the fit ($\bar{R}^2$).

Exercise 2

- Take some time to estimate some multiple linear regression models using numeric variables from your project data (keep in mind that causality matters). Complete all of the required diagnostics, remove any variables as required to improve your models, comment on the estimates, and comment on the fit ($\bar{R}^2$).

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm
- https://www.rdocumentation.org/packages/car/versions/1.2-6/topics/ncv.test
- https://www.rdocumentation.org/packages/car/versions/3.1-2/topics/durbinWatsonTest
- https://cran.r-project.org/web/packages/car/index.html
- https://www.rdocumentation.org/packages/car/versions/3.1-2/topics/vif