The Linear Regression Model – Estimation and Inference

Philip Leifeld

GV903: Advanced Research Methods, Week 5

1. Graphical Intuition of the Linear Model

The Line of Best Fit

- A line can be drawn to show the association between variables X and Y.
- The "line of best fit" indicates how much more of Y is associated with a one-unit increase in X.
- "Best fit" because it minimises the distance between the line and the observations.
- Can be used to express association between IV and DV.

Regression Tables

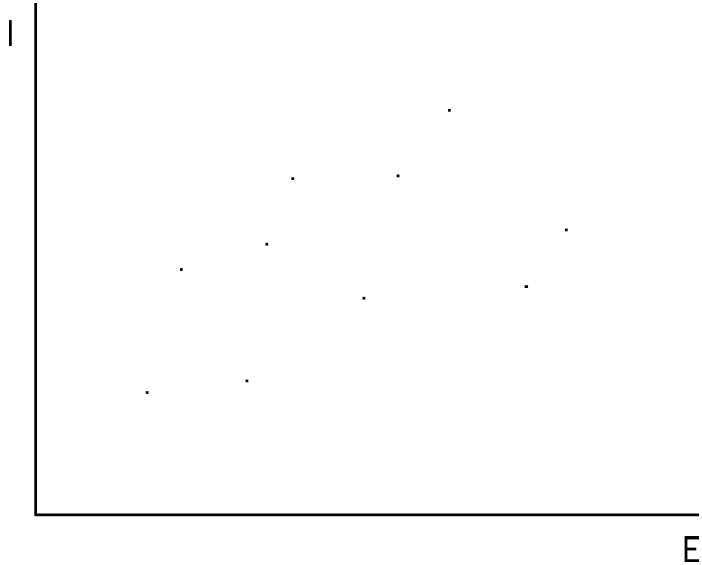
	Effect on income
(Intercept)	$7.61\ (0.15)^{***}$
education	$0.13\ (0.01)^{***}$
women	$-0.01\ (0.00)^{***}$
R^2	0.61
Adj. R^2	0.60
Num. obs.	102

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

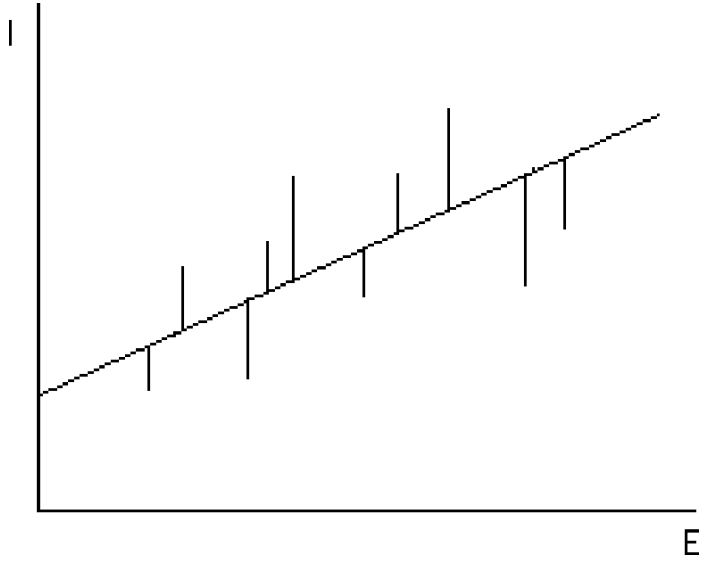
This table shows the influence of education (the IV) on income (the DV), controlling for the share of women in a given profession.

The next few slides will give an intuition of these numbers.

Scatterplot



Minimising Distances Between Line and Observations

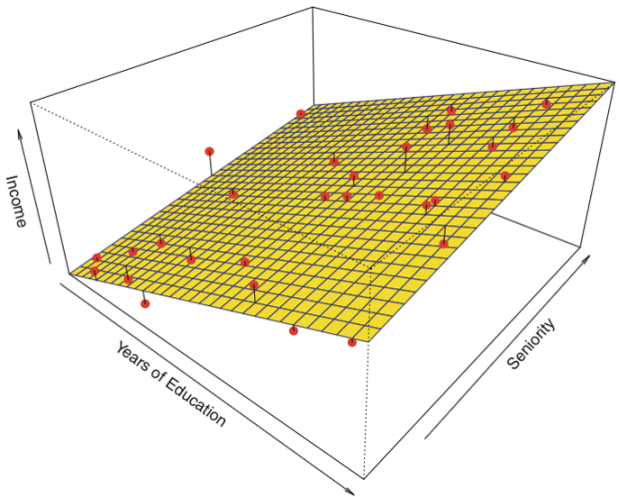


Ordinary Least Squares

- Any line in two dimensions can be expressed as $\alpha + \beta \cdot x$.
- α is the offset (= the height at which the line starts).
- β is the slope of the line (= the amount by which the DV increases when the IV is increased by one unit).
- More formally: $Y = \alpha + \beta X + \epsilon$.
- Minimise the distances to fit the line.
- `https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html`

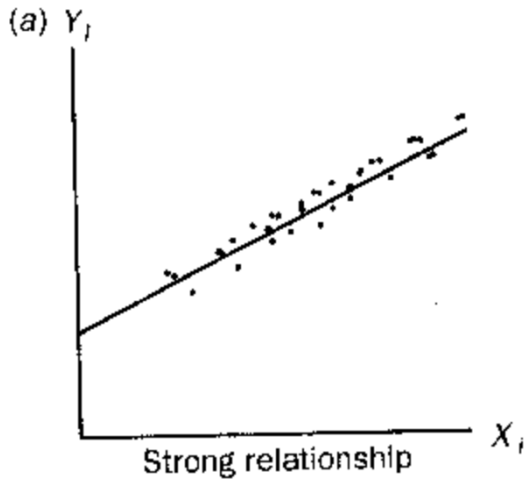
Ordinary Least Squares with Two IVs

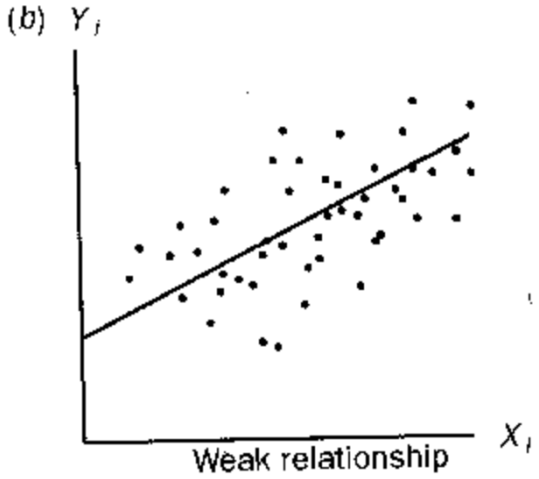
Source: James et al. (2013): An Introduction to Statistical Learning. Springer.



Strength of a Relationship

Sanders and Brynin (1988)





Steps in Computing a Linear Model

1. Draw line or plane through the points.
2. Move the line or plane to achieve best fit.[1]
3. Extract the β values as "estimates" of the effects of each IV on the DV.
4. Assess how certain we are in the β relationships by looking at the strength of the relationship and the number of observations.

[1] There are several types of estimation to find the best fit.

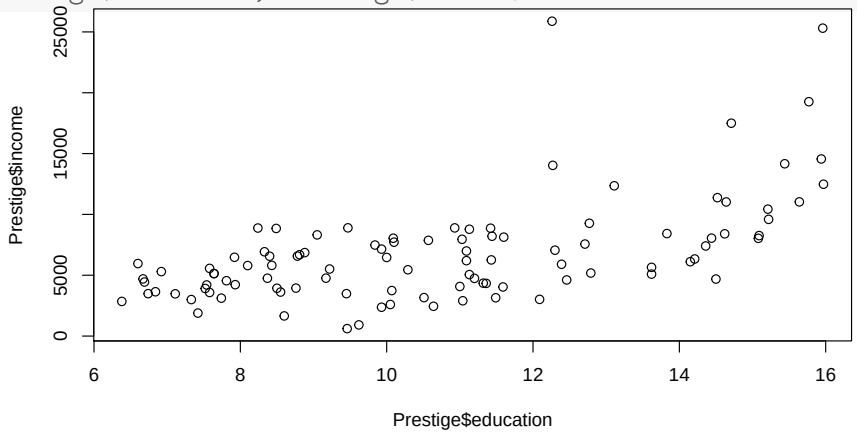
Example: Linear Model in R

```
library("car")

## Loading required package:  carData

head(Prestige)
##                       education income women prestige census type
## gov.administrators        13.11  12351 11.16     68.8   1113 prof
## general.managers          12.26  25879  4.02     69.1   1130 prof
## accountants               12.77   9271 15.70     63.4   1171 prof
## purchasing.officers       11.42   8865  9.11     56.8   1175 prof
## chemists                  14.62   8403 11.68     73.5   2111 prof
## physicists                15.64  11030  5.13     77.6   2113 prof
```

Income, percentage of women, and average years of education in 102 occupations in Canada.

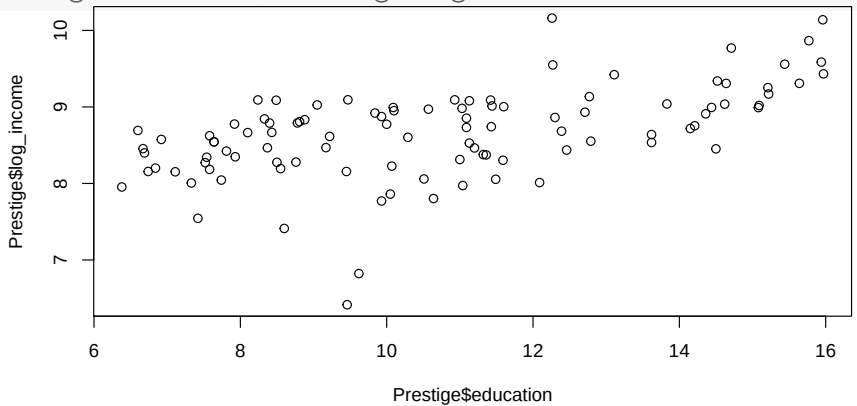
Example: Linear Model in R

```
plot(Prestige$education, Prestige$income)
```



Example: Linear Model in R

```
Prestige$log_income <- log(Prestige$income)
plot(Prestige$education, Prestige$log_income)
```



Transformation of the DV because it is skewed.

Example: Linear Model in R

```
model <- lm(log_income ~ education + women, data = Prestige)
summary(model)
##
## Call:
## lm(formula = log_income ~ education + women, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92598 -0.10273  0.06927  0.15239  1.04854
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.606040   0.152498  49.876  < 2e-16 ***
## education    0.126303   0.013610   9.280 4.07e-15 ***
## women       -0.010414   0.001171  -8.897 2.78e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3725 on 99 degrees of freedom
## Multiple R-squared:  0.6112,Adjusted R-squared:  0.6034
## F-statistic: 77.83 on 2 and 99 DF,  p-value: < 2.2e-16
```

Explanation of Results

- *Estimate* contains the slopes.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.
- If the estimate is large (i. e., different from zero) and the standard error is small (i. e., the effect is strongly visible), then our confidence in the finding is high.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.
- If the estimate is large (i. e., different from zero) and the standard error is small (i. e., the effect is strongly visible), then our confidence in the finding is high.
- Here, education has a positive slope (0.13), which means one additional year of education increases logged income by 0.13.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.
- If the estimate is large (i. e., different from zero) and the standard error is small (i. e., the effect is strongly visible), then our confidence in the finding is high.
- Here, education has a positive slope (0.13), which means one additional year of education increases logged income by 0.13.
- We can be confident that this effect is really there given the slope, the strength of relationship, and the sample size.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.
- If the estimate is large (i. e., different from zero) and the standard error is small (i. e., the effect is strongly visible), then our confidence in the finding is high.
- Here, education has a positive slope (0.13), which means one additional year of education increases logged income by 0.13.
- We can be confident that this effect is really there given the slope, the strength of relationship, and the sample size.
- Share of women in the profession has a negative effect: the more women, the lower the salary in a given occupation.

Explanation of Results

- *Estimate* contains the slopes.
- *Std. Error* is the square root of the variance of the residual distances/errors.
- If the estimate is large (i. e., different from zero) and the standard error is small (i. e., the effect is strongly visible), then our confidence in the finding is high.
- Here, education has a positive slope (0.13), which means one additional year of education increases logged income by 0.13.
- We can be confident that this effect is really there given the slope, the strength of relationship, and the sample size.
- Share of women in the profession has a negative effect: the more women, the lower the salary in a given occupation.
- There is very little statistical uncertainty around this finding.

This Results in the Regression Table Shown Before

```
library("texreg")
texreg(model, dcolumn = TRUE, booktabs = TRUE, use.packages = FALSE,
       single.row = TRUE, table = FALSE)
```

	Model 1
(Intercept)	7.61 (0.15)***
education	0.13 (0.01)***
women	−0.01 (0.00)***
R^2	0.61
Adj. R^2	0.60
Num. obs.	102

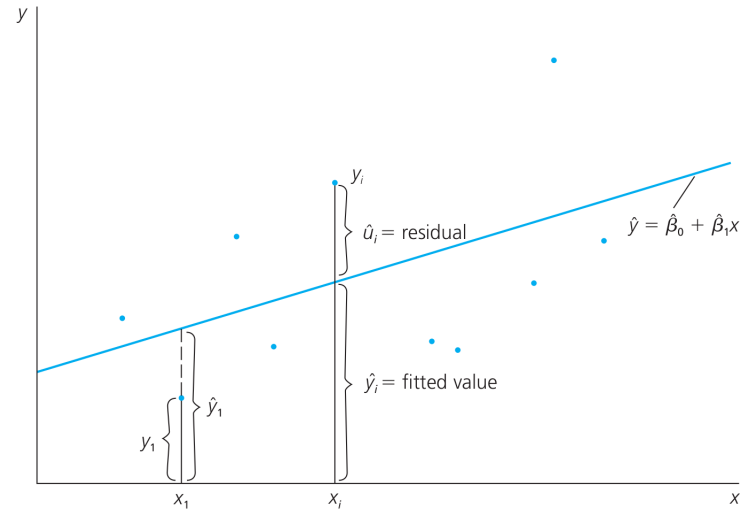
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

This table shows the influence of education (the IV) on income (the DV), controlling for the share of women in a given profession.

2. Unpacking Ordinary Least Squares

Fitted Values and Residuals

FIGURE 2.4 Fitted values and residuals.



Note how the error term u absorbs all unmodeled variables that affect y:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

Note how the error term u absorbs all unmodeled variables that affect y:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

Factors in the error term must not be correlated with any of the explanatory variables:

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$$

That is, the residuals in all directions of the line cancel each other out and have a mean of 0.

Note how the error term u absorbs all unmodeled variables that affect y:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

Factors in the error term must not be correlated with any of the explanatory variables:

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$$

That is, the residuals in all directions of the line cancel each other out and have a mean of 0.

Exercise: Can you come up with examples where this assumption is violated?

Omitted Variable Bias

Example: explaining final marks in a course

- Y: final marks
- X_1: attendance
- X_2: intelligence/skills
- X_3: study

If X_3 is omitted, marks are wrongly attributed to attendance.

If an omitted variable X_3 is correlated both with Y and X_1, too much variance in Y is explained by X_1, leading to a biased estimate for X_1 (type I error).

Exception: do not include intervening variables as controls.

Ordinary Least Squares

Ordinary least squares form of the model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots + \hat{\beta}_k x_k$$

In order to obtain the estimates, we must minimise the sum of squared residuals:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

This results in $k + 1$ simultaneous equations. They can be solved jointly using matrix algebra (where $\mathbf{X}$ is the $n \times k$ matrix of IVs):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Fitted Values and Residuals

Fitted (= predicted) value for observation i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

Fitted Values and Residuals

Fitted (= predicted) value for observation i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

The residual for observation i is the difference between the line and the observed value:

$$\hat{u}_i = y_i - \hat{y}_i$$

Exercise[2]

We model the number of deaths in riots (Y), which is a proxy variable for political instability, as a linear function of the magnitude of IMF loans as a percentage of GDP (X_1), controlling for drought conditions as measured in days per year (X_2). Suppose $\hat{\beta}_0 = 27.45$, $\hat{\beta}_1 = 230.32$, and $\hat{\beta}_2 = -2.95$.

1. Interpret the effect sizes.
2. How many deaths should we expect in a country with 8.45 % IMF loans and 45 drought days per year?
3. Discuss the conditional zero mean assumption in this case. What could cause it to fail?
4. Is the linear model the best choice for these data? Can you see any potential problems?

[2] Based on fictitious estimates.

Solution

1. If there are no loans or droughts, the model predicts 27.45 deaths. An additional percentage point of loans leads to 240.32 additional deaths, at constant levels of droughts. An additional drought day leads to 2.95 fewer deaths.
2. $\hat{y}_i = 27.45 + 230.32(8.45) - 2.95(45) = 1840.904$ deaths.
3. There could be omitted variables that explain both IMF loans and deaths, such as regime type; drought is an exogenous variable and thus not affected by spuriousness. Another reason could be that the observations are not independent: if, for example, countries learn from neighbouring countries and this is not modelled, it is a form of omitted variable bias.
4. Independence of observations potentially not given; perhaps a network model might be better. Moreover, the DV is a count; a model for Poisson-distributed data might be better because we might otherwise assume the wrong functional form and predict negative numbers of deaths.

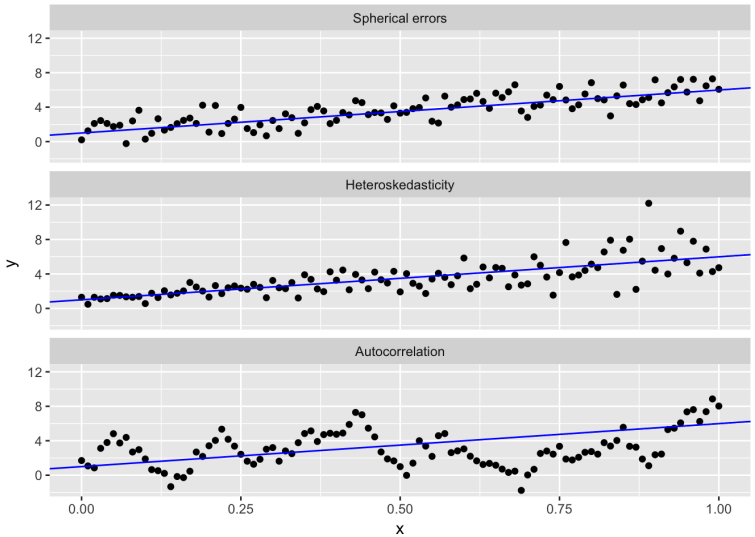
Assumptions of OLS

The Gauss–Markov assumptions:

1. The parameters are linear. That is, we add the terms via $+$.
2. Random sampling. If we choose observations deliberately, the model does not represent the population.
3. No perfect collinearity. If a model term is a linear function of at least one other model term, the system of first-order equations cannot be solved.
4. Zero conditional mean. No omitted variables that explain an IV and the DV.
5. Homoskedasticity. Across the range of Y, the residuals u_i have the same variance.

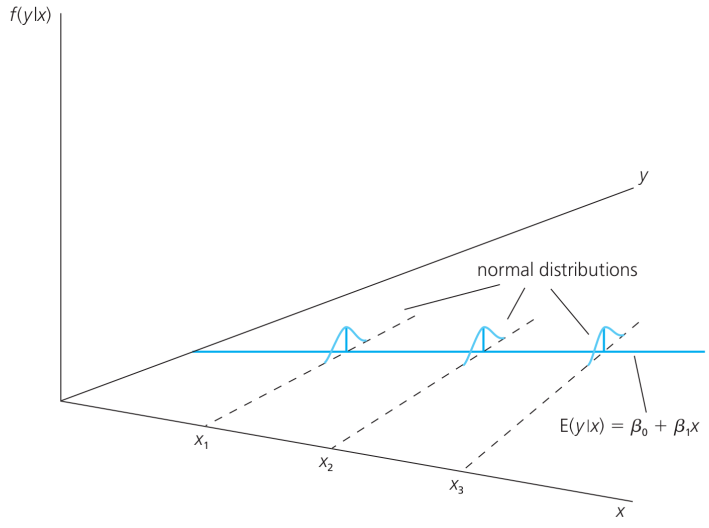
Spherical Errors, Heteroskedasticity, Autocorrelation

http://bkenkel.com/pdaps/nonspherical.html



Homoskedasticity and Normality

Figure 4.1 The homoskedastic normal distribution with a single explanatory variable.



Normally distributed residuals: $\mu = 0$ and constant σ across Y.

Exercise

1. What happens if we include irrelevant variables in a model?
2. What is multicollinearity? How is it a problem? How can we fix it?
3. Can you imagine an empirical example in which there is heteroskedasticity?

Solutions

1. The variance of the residuals around the predicted values increases, leading to more uncertainty and larger standard errors.
2. If two IVs are highly correlated, there are fewer observations that vary between the variables and from which something can be learned about the relationship betwen IV and DV; increase sample size to fix this problem or change model specification!
3. Years since PhD $\rightarrow$ higher salary.
   Household income $\rightarrow$ household expenditures.

3. Assessing Uncertainty

Goodness of Fit

Total sum of squares (of differences between actual and mean values on DV):

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Goodness of Fit

Total sum of squares (of differences between actual and mean values on DV):

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares (of differences between predicted and mean values):

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Goodness of Fit

Total sum of squares (of differences between actual and mean values on DV):

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares (of differences between predicted and mean values):

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares:

$$\text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2. \text{ Note that } \text{SST} = \text{SSE} + \text{SSR}.$$

Goodness of Fit

Total sum of squares (of differences between actual and mean values on DV):

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares (of differences between predicted and mean values):

$$\text{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares:

$$\text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2. \text{ Note that } \text{SST} = \text{SSE} + \text{SSR}.$$

R-squared:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

Computing Standard Errors of Estimates

Standard deviation of an IV:

$$\text{sd}(x_j) = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n}}$$

Estimated variance in the residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\text{SSR}}{n-k-1}$$

Standard error of an estimate:

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\text{SST}_j(1-R^2)}} = \frac{\hat{\sigma}}{\sqrt{n}\text{sd}(x_j)\sqrt{1-R^2}}$$

Computing Standard Errors of Estimates

Standard deviation of an IV:

$$\text{sd}(x_j) = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n}}$$

Estimated variance in the residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{\text{SSR}}{n-k-1}$$

Standard error of an estimate:

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\text{SST}_j(1-R^2)}} = \frac{\hat{\sigma}}{\sqrt{n}\text{sd}(x_j)\sqrt{1-R^2}}$$

Note how the sample size directly influences the SEs!

Significance Testing

See also last week's slides.

Test statistic, a.k.a. t statistic or t ratio:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)}$$

Critical value from quantile function of t distribution:

$$c = F^{-1}(\alpha_{\frac{1}{2}}; df = n - k - 1)$$

Two-tailed hypothesis test:

$$|t_{\hat{\beta}}| > c$$

p value:

$$p = P(c > |t_{\hat{\beta}}|) = 2P(c > t_{\hat{\beta}}) = 2(1 - F_t(t_{\hat{\beta}}))$$

Confidence interval:

$$\hat{\beta}_j \pm c \cdot \text{se}\left(\hat{\beta}_j\right)$$

Computing the t Statistic Using the Covariance Matrix

Variance-covariance matrix:

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{pmatrix} \widehat{\text{Var}}(\hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_k) \\ \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_0) & \widehat{\text{Var}}(\hat{\beta}_1) & \dots & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\hat{\beta}_k, \hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_k, \hat{\beta}_1) & \dots & \widehat{\text{Var}}(\hat{\beta}_k) \end{pmatrix}$$

Take the diagonal values from $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ to compute t statistic:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)}$$

Then proceed like before...