3. Statistical Inference, Simulations, and the Linear Regression Model

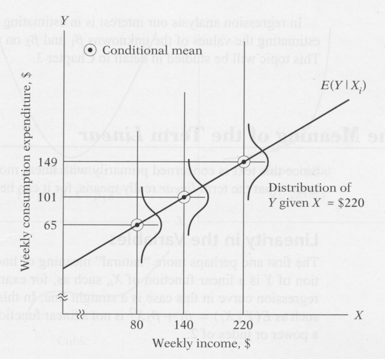
POL 212: Quantitative Analysis I

Winter 2024

Some quick background and terminology

- The method of “least squares” has its origins in astronomy, specifically Legendre’s (1805) use of the method to estimate a set of coefficients for planetary orbits by minimize the sum of squared errors.
- Gauss and Laplace later added a probabilistic component by including a random error term (using, naturally, the Gaussian/normal distribution).
- The methods of correlation and regression came later with Galton and Pearson in the mid/late 1800s.
- Yule (1897) finally combined the two, showing that least squares could be used to estimate regeression models.

The population regression model



- Regression models give us the *conditional expectation* of Y given X, $E(Y|X)$. This should be more informed than the *unconditional expected value* $E(Y)$.
- Generally speaking, we try to model the *population regression function*, $E(Y|X_i) = f(X_i)$.

A linear population regression function

- It falls on the researcher to specify the functional form of the population regression function. (Remember: the attributes of a population are unknown.)
- A common specification is the linear population regression function: $E(Y|X_i) = \alpha + \beta_1 X_i$.
- Equivalent representation: $Y_i = \alpha + \beta_1 X_i + u_i$.

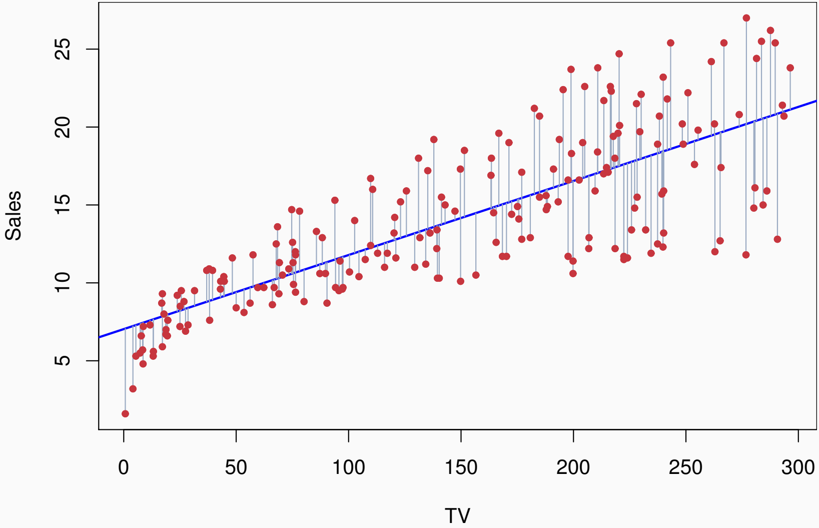
The meaning of the term linear

- A model is linear in the variables if Y is a linear function of every X variable.
- A model is linear in the parameters if each parameter is only raised to the power 1 and is not multiplied or divided by any other parameter.
- The **linear regression model** is linear in the parameters.
- **BUT**, a linear regression model need not be linear in the variables!
- Hence, the linear regression model can in fact produce a variety of nonlinear relationships.

The sample regression model

- We usually have to estimate our models with a sample from the population.
- Hence, the sample regression function is: $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i$.
- $\hat{\alpha}$ & $\hat{\beta}_1$ are used as estimates of the population parameters.
- The estimate for $\hat{\alpha}$ is the *sample intercept* and the estimate for $\hat{\beta}_1$ is the *sample slope coefficient*.
- $\hat{u}_i = y_i - \hat{y}_i$. This is, as before, the residual or error term.
- We don't really estimate u_i as much as we predict it.

Is this the best line?



The method of Ordinary Least Squares (OLS)

- Want to minimize residuals.
- The OLS approach: fit a line that minimizes *squared* residuals.
- For the two-variable population regression model, $Y_i = \alpha + \beta_1 X_i + u_i$, we do this by choosing $\hat{\alpha}$ and $\hat{\beta}_1$ to minimize the sum of squared residuals.
- The sum of squared residuals is: $\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i)^2$.

The method of Ordinary Least Squares (OLS)

- The sum-of-squares function, then, is simply: $S(\hat{\alpha}, \hat{\beta}_1) = \sum(Y_i - \hat{\alpha} - \hat{\beta}_1 X_i)^2$
- With a bit of differential calculus (taking the partial derivatives $\frac{\partial S(\hat{\alpha}, \hat{\beta}_1)}{\partial \hat{\alpha}}$ and $\frac{\partial S(\hat{\alpha}, \hat{\beta}_1)}{\partial \hat{\beta}_1}$, setting them to 0, and solving), we create simultaneous linear equations: the normal equations for simple OLS regression.

Calculating the slope

- From this, the equation to compute $\hat{\beta}_1$ is: $\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$
- In other words, the slope is computed by taking the difference of each value of the dependent variable from the mean of the dependent variable multiplied by the difference of each value of independent variable from the mean of the independent variable.
- Then a sum of all those products, divided by the sum of squared deviations of the independent variable.
- **Intuition**: how much do the two variables go together (covariance) divided by how much does the independent variable vary (variance).

Calculating the intercept

- From this point we can also compute the sample intercept as well.
- $\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}$, where:
  - $\bar{Y}$=mean of the dependent variable.
  - $\bar{X}$=mean of the independent variable.

The Gauss-Markov Theorem

- When the Gauss-Markov assumptions are satisfied, the OLS estimator is BLUE:
  - (B)est (efficient)
  - (L)inear
  - (U)nbiased
  - (E)stimator
- Bias and efficiency (minimum variance of the sampling distribution).

The Gauss-Markov assumptions

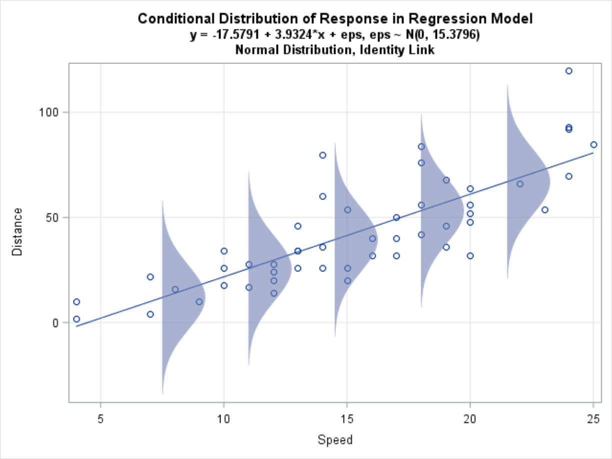
- There is a strong set and a weak set of Gauss-Markov assumptions.
- The weak set is sufficient to satisfy the Gauss-Markov theorem.
- To form the strong set, we add a final assumption: that the disturbances are normally distributed.
- This allows us to draw inferences about our estimates (for example, building confidence intervals and conducting hypothesis tests).

The Gauss-Markov assumptions

The weak set:

1. **Linearity**: The expected (mean) value of the disturbance term is 0.
2. **Nonstochastic regressors**: X values are independent of the error term (or X is exogenous). $cov(X_i, u_i) = 0$.
3. **Homoscedasticity**: constant error variance across values of X_i. Means OLS no longer efficient.
4. **Independence**: No autocorrelation between disturbances. $cov(u_i, u_j) = 0$ for $i \neq j$. Means OLS no longer efficient.

Gauss-Markov visually



Recall: Pearson's *r*, the correlation coefficient

- Pearson's *r* is a symmetric measure of bivariate association. It shows how well the independent variable predicts the dependent variable. This measure will range between -1 and 1.
- A correlation coefficient of 0 would suggest the absence of any relationship between the two variables. A value of 1 would imply a perfect positive relationship, and a value of -1 would imply a perfect negative relationship. (Do we see perfect linear relationships in the social sciences?)
- The square of a Pearson's *r* calculates the amount of variance explained by the predictor.

A side note on r^2

- *Generally speaking*, larger r^2 are preferred (bounded at 1).
- However, as the number approaches 1, concerns about the data emerge. There are no perfect relationships in the social sciences and so r^2 over .90 or so are immediately suspect. (Unless we're in the time series world.)
- Generally, it is best to compare your r^2 against scholars doing similar work (for example, political psychology research rarely reaches .25).
- Good research asks good questions. So finding a high r^2 on a question that has been decided in the literature is less important than research with a low r^2 on an interesting question or a brand new idea.

Our plan for calculating r^2

- The total variance in Y can be split into explained and residual variance: $TSS = ESS + RSS$.
- Compute the proportional reduction in error (r^2).

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

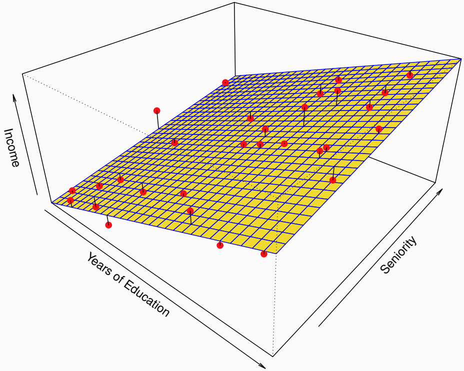
- This is the square of the linear coefficient or Pearson's r.
- There are a few formulas we can use for r^2 (one predictor) or R^2 (multiple predictors), but we generally stick to the sum of squares based formulae when we start estimating multiple regression models.

Multiple linear regression

- The central difference between simple and multiple linear regression: The slope coefficients for the explanatory variables in multiple regression are **partial** coefficients.
- That is, it represents the ’effect’ on the response variable of a one-unit increment in the corresponding explanatory variable, holding constant the value of the other explanatory variable.
- The simple-regression slope simply ignores the other explanatory variable.

Multiple linear regression

Because the regression plane is flat, its slope in the direction of 1, holding 2 constant, does not depend upon the specific value at which 2 is fixed.



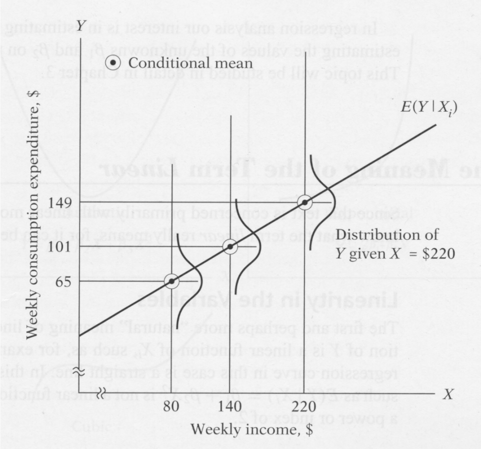
The regression model

Recall the linear population regression function:
$Y_i = \alpha + \beta_1 X_i + u_i$.

- The coefficients α and β_1 are the population regression parameters to be estimated.
- The error u_i represents the aggregated, omitted causes of Y, whether it's other explanatory variables that have been omitted, measurement error in Y, or whatever component of Y is inherently random.

Assumptions of regression models

The key assumptions of the regression models concern the distribution of *Y* conditional on *X* or, equivalently, the behavior of the **errors**.



What are these assumptions?

The weak set of Gauss-Markov assumptions (enough to prove OLS is BLUE):

1. **Linearity**: The expected (mean) value of the disturbance term is 0. Why important?
   - $E(Y_i) = E(Y|x_i) = E(\alpha + \beta_1 x_i + u_i)$
   - $\alpha + \beta_1 x_i + E(u_i)$
   - $\alpha + \beta_1 x_i$

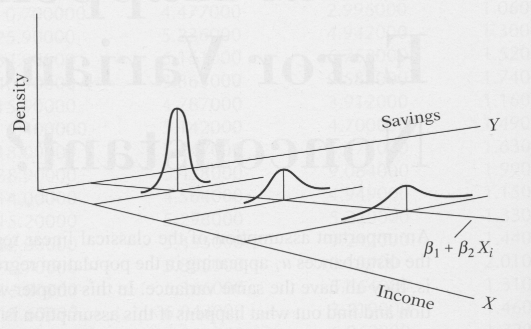
What are these assumptions?

2. **Nonstochastic regressors, fixed X, or X independent of the error**: X values are independent of the error term (or X is **exogenous**). $cov(X_i, u_i) = 0$.
   - Can arise because of measurement error on X, omitted confounder(s), or simultaneous causation.
   - Example: economic conditions and civil conflict.
   - Instrumental variables: correlated with X, but not Y (e.g., rainfall).

What are these assumptions?

3. **Homoskedasticity**: constant error variance across values of X_i.
    - $Var(\varepsilon) = \sigma^2$
    - When violated? Situations where some values of X have greater uncertainty.
    - Robust standard errors.

Heteroskedasticity



What are these assumptions?

4. **Independence**: No autocorrelation between disturbances. $cov(u_i, u_j) = 0$ for $i \neq j$. Means OLS no longer efficient.
    - Influences our understanding of the error term.
    - Time series data.

Properties of the OLS estimator

- Our weak set of assumptions: **linearity**, **nonstochastic regressors**, **homoskedastic errors**, and **independent errors**.
- With these assumptions, the Gauss-Markov theorem tells us that the OLS estimates are BLUE.
- α and β will be linear estimators of the form:

Simple regression:

$\beta = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$ and $\alpha = \bar{Y} - \beta\bar{X}$

Properties of the OLS estimator

Multiple regression:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ X'Y &= X'X\beta + X'\varepsilon \\ X'Y &= X'X\beta + 0 \\ (X'X)^{-1}X'Y &= \beta + 0 \\ \beta &= (X'X)^{-1}X'Y \end{aligned} \tag{1}$$

OLS in geometric terms

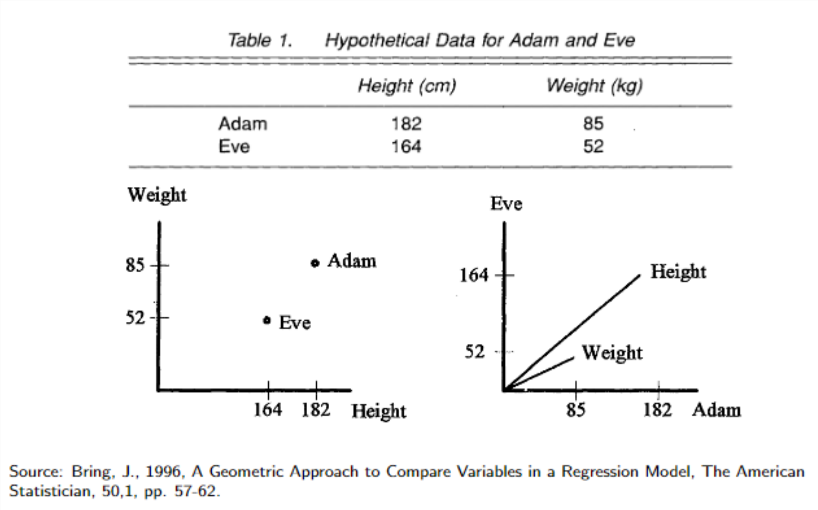


Table 1. Hypothetical Data for Adam and Eve

	Height (cm)	*Weight (kg)*
Adam	182	85
Eve	164	52

Source: Bring, J., 1996, A Geometric Approach to Compare Variables in a Regression Model, The American Statistician, 50,1, pp. 57-62.

Sampling variances

- With the assumptions of linearity, constant variance ($Var(Y|x_i) = \sigma^2$), and independence, α and β have simple sampling variances:

$$\begin{aligned} Var(\alpha) &= \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} \\ Var(\beta) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \end{aligned} \tag{2}$$

- Note how increasing variation in X reduces $Var(\beta)$.

Sampling variances

- Let's add a final (fifth) assumption to complete the strong set: **Normality**: Errors are distributed normally ($\varepsilon \sim N(0, \sigma^2)$).
- So what? For one, means that the OLS estimator is the most efficient among *all* unbiased estimators, not just linear unbiased estimators.

Sampling variances

- We like the normal distribution:
  1. By the **central limit theorem** (CLT), the sum of a large number of independent and identically distributed random variables converges to a normal distribution.
  2. Ease of inference: The sum of normally distributed variables has a normal distribution. Hence, $\hat{\alpha}$ and $\hat{\beta}$ have normal sampling distributions.
  3. Many natural phenomena follow the normal distribution, it is well known, and it is relatively simple with only two parameters.
  4. With a small sample size, the normality assumption allows us to use t, F, and χ^2 tests in regression models.
  5. In large data sets, deviations from the normality assumption are not so critical.
  6. Quick aside: how does this translate into assumptions about the distribution of the observed y values?

Calculating standard errors

- The variance of the residuals provides an unbiased estimator of σ^2, or more precisely the **standard error** σ, where 2 represents the number of coefficients estimated by the simple regression model:

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$$

Calculating standard errors

- From this, we compute the standard error of each sample coefficient:

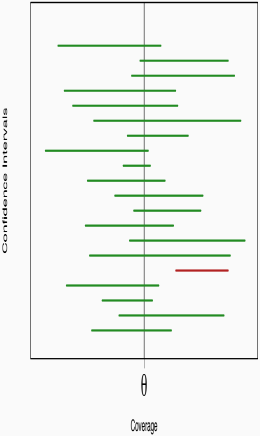
$$se(\hat{\alpha}) = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum(X_i - \bar{X})^2}}$$

$$se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

- And, with the normality assumption, we can use these standard errors to construct confidence intervals and perform hypothesis testing.

Interpreting a confidence interval

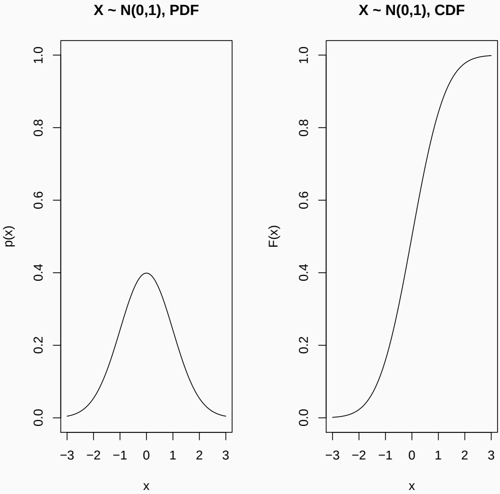
- $\alpha =$Type I error rate. What is a Type II error?
- Which of these is the correct interpretation of a $(1 - \alpha)$ confidence interval?
    - An interval that has a $1 - \alpha\%$ chance of containing the true value of the parameter.
    - An interval that over $1 - \alpha\%$ of replications contains the true value of the parameter, *on average*.



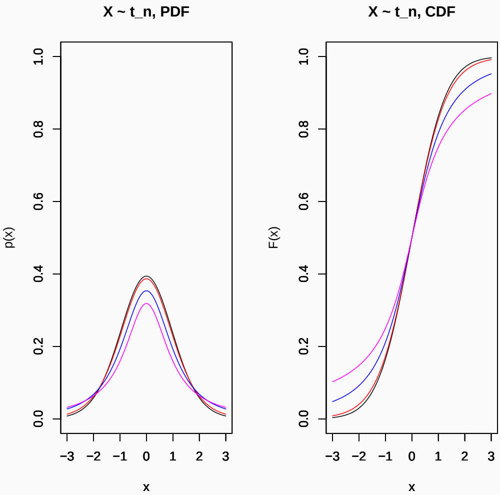
Why the t distribution is critical for inference

- From the sampling distribution of $\hat{\beta}$, we know that $Z = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}}$ is distributed $Z \sim \mathscr{N}(0,1)$.
- This can be re-written as: $Z = \frac{(\hat{\beta} - \beta)\sqrt{\sum(X_i - \bar{X})^2}}{\sigma}$ for σ the standard error of regression.
- Since we don't know σ, we often must substitute our estimate $\hat{\sigma}$, yielding $t = \frac{(\hat{\beta} - \beta)\sqrt{\sum(X_i - \bar{X})^2}}{\hat{\sigma}}$.

The normal distribution



Student's *t*-distribution



Confidence intervals for regression coefficients

- Therefore, for the right t distribution, we can use the following fact:

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} \leq t_{\alpha/2}\right] = 1-\alpha$$

- Doing some algebra yields:

$$Pr\left[\hat{\beta} - t_{\alpha/2}\hat{\sigma}_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_{\alpha/2}\hat{\sigma}_{\hat{\beta}}\right] = 1-\alpha$$

- Thus, our $100(1-\alpha)$ percent confidence interval for β is $\hat{\beta} \pm t_{\alpha/2}\hat{\sigma}_{\hat{\beta}}$.

Hypothesis testing

- Presumably we estimate a regression model to test some theory. We would like to make a statement about whether our theoretical relationship holds in the larger population.
- Hence, hypothesis tests about population slope coefficients are conducted daily in political science and are a major workhorse in our research.
- By default, software conducts the following hypothesis test:

$$\begin{aligned} H_0: &\quad \beta = 0 \\ H_1: &\quad \beta \neq 0 \end{aligned}$$

- H_0 is the null hypothesis, and H_1 (also called H_A) is the alternative hypothesis.
- This is a two-tailed alternative hypothesis.

One-tailed hypothesis tests

- A one-tailed alternative hypothesis would be:

$$\begin{aligned} H_0 : \quad & \beta = 0 \\ H_1 : \quad & \beta < 0 \end{aligned}$$

or

$$\begin{aligned} H_0 : \quad & \beta = 0 \\ H_1 : \quad & \beta > 0 \end{aligned}$$

- Recall: Confidence intervals and hypothesis tests have a close relationship.

Conducting a t-test

- Based on what we know, we can compute a t-distributed test statistic with $n-k-1$ degrees of freedom:

$$t = \frac{\hat{\beta} - \beta}{se(\hat{\beta})}$$

- Our value of β is drawn from the null hypothesis.
- Once we calculate t, we compare it to our *critical value*. Use the following rules based on the nature of your alternative hypothesis.
  - $H_1 : \beta \neq \beta^*$: If $|t| > t_{\alpha/2}$, then we reject the null hypothesis.
  - $H_1 : \beta < \beta^*$: If $t < -t_{\alpha}$, then we reject the null hypothesis.
  - $H_1 : \beta > \beta^*$: If $t > t_{\alpha}$, then we reject the null hypothesis.

Statistical inference: the *F*-ratio

- Besides our *t*-ratios, we may want to evaluate the model as a whole.
- Recall that the residual sum of squares is the squared set of errors: $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the fitted values. These are found with `fitted()` in `R` or `predict` in Stata. *RSS* has $n-k-1$ degrees of freedom. ($k =$number of predictors.)

Statistical inference: the *F*-ratio

- Now consider the *explained sum of squares*, which has k degrees of freedom:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- Define the total sum of squares as the total variation in the Y values about the mean, which has $n-1$ degrees of freedom:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Statistical inference: the F-ratio

- Recall $TSS = ESS + RSS$
- When we divide any of these quantities by their associated degrees of freedom, it is called a *mean sum of squares*, and is a type of variance.
- Since ESS is “good” and RSS is “bad,” we can create a comparison:

$$F = \frac{ESS/(k)}{RSS/(n-k-1)}$$

  which is distributed “F” with k and $n-k-1$ degrees of freedom.
- Thus we have an omnibus test of whether the whole model fits well:
  - H_0: model does not fit at the 1-α level
  - H_1: model fits at the 1-α level

Predictions and uncertainty

- A nice feature of regression models is they allow us to make predictions. The formula is simple: $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta} X_0$.
- Beware, though, predictions that are outside of the domain of the inputs may be dicey.
- We would like to measure how uncertain we are about a prediction.

Uncertainty for means or individuals

- The key distinction: Are we expressing our uncertainty about the mean of Y given X_0? If so:

$$\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

- Alternatively, are we expressing our uncertainty about the value of a specific individual's Y given X_0? If so:

$$\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$