Estimating σ^2

● **We can do simple prediction of *Y* and estimation of the mean of *Y* at any value of *X*.**

● **To perform inferences about our regression line, we must estimate σ^2, the variance of the error term.**

● **For a random variable *Y*, the estimated variance is:**

● **In regression, the estimated variance of *Y* (and also of ε) is:**

$\sum (Y - \hat{Y})^2$ **is called the error (residual) sum of squares (SSE).**

● **It has *n* – 2 degrees of freedom.**

● **The ratio MSE = SSE / df is called the mean squared error.**

● **MSE is an unbiased estimate of the error variance σ^2.**

● **Also, $\sqrt{MSE}$ serves as an estimate of the error standard deviation σ.**

<u>Partitioning Sums of Squares</u>

● **If we did not use *X* in our model, our estimate for the mean of *Y* would be:**

Picture:

For each data point:

● $Y - \bar{Y}$ **= difference between observed *Y* and sample mean *Y*-value**

● $Y - \hat{Y}$ **= difference between observed *Y* and <u>predicted</u> *Y*-value**

● $\hat{Y} - \bar{Y}$ **= difference between predicted *Y* and sample mean *Y*-value**

● **It can be shown:**

● **TSS = overall variation in the *Y*-values**
● **SSR = variation in *Y* accounted for by regression line**
● **SSE = extra variation beyond what the regression relationship accounts for**

Computational Formulas:

TSS = S_{YY} = $\sum Y^2 - \frac{(\sum Y)^2}{n}$

SSR = $(S_{XY})^2 / S_{XX}$ = $\hat{\beta}_1 S_{XY}$

SSE = $S_{YY} - (S_{XY})^2 / S_{XX}$ = $S_{YY} - \hat{\beta}_1 S_{XY}$

Case (1): If SSR is a large part of TSS, the regression line accounts for a lot of the variation in *Y*.

Case (2): If SSE is a large part of TSS, the regression line is leaving a great deal of variation unaccounted for.

ANOVA test for β_1

● **If the SLR model is useless in explaining the variation in *Y*, then $\bar{Y}$ is just as good at estimating the mean of *Y* as $\hat{Y}$ is.**

=> true β_1 is zero and *X* doesn't belong in model

● **Corresponds to case (2) above.**

• But if (1) is true, and the SLR model explains a lot of the variation in Y, we would conclude $\beta_1 \neq 0$.

• How to compare SSR to SSE to determine if (1) or (2) is true?

• Divide by their degrees of freedom. For the SLR model:

• We test:

• If MSR much bigger than MSE, conclude H_a. Otherwise we cannot conclude H_a.

The ratio F* = MSR / MSE has an F distribution with df = (1, *n* – 2) when H_0 is true.

Thus we reject H_0 when

where α is the significance level of our hypothesis test.

t-test of H_0: $\beta_1 = 0$

• Note: β_1 is a parameter (a fixed but unknown value)

• The estimate $\hat{\beta}_1$ is a random variable (a statistic calculated from sample data).

• Therefore $\hat{\beta}_1$ has a sampling distribution:

• $\hat{\beta}_1$ is an unbiased estimator of β_1.

• $\hat{\beta}_1$ estimates β_1 with greater precision when:

- **the true variance of *Y* is small.**
- **the sample size is large.**
- **the *X*-values in the sample are spread out.**

Standardizing, we see that:

Problem: σ^2 is typically unknown. We estimate it with MSE. Then:

To test H_0: $\beta_1 = 0$, we use the test statistic:

<u>Advantages of t-test over F-test</u>:
(1) Can test whether the true slope equals <u>any</u> specified value (not just 0).
Example: To test H_0: $\beta_1 = 10$, we use:

(2) Can also use t-test for a one-tailed test, where:
H_a: $\beta_1 < 0$ or H_a: $\beta_1 > 0$.

H_a	**<u>Reject H_0 if</u>:**

(3) The value $\sqrt{\frac{MSE}{S_{XX}}}$ **measures the precision of** $\hat{\beta}_1$ **as an estimate.**

Confidence Interval for β_1

● **The sampling distribution of** $\hat{\beta}_1$ **provides a confidence interval for the** **true slope** β_1**:**

Example (House price data):

Recall: $S_{YY} = 93232.142$, $S_{XY} = 1275.494$, $S_{XX} = 22.743$

Our estimate of σ^2 **is MSE = SSE / (*n* – 2)**

SSE =

MSE =

and recall

● **To test** H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$ (at $\alpha = 0.05$)

Table A2: $t_{.025}(56) \approx 2.004$.

● With 95% confidence, the true slope falls in the interval

Interpretation:

Inference about the Response Variable

● We may wish to:

(1) Estimate the mean value of *Y* for a particular value of *X*. Example:

(2) Predict the value of *Y* for a particular value of *X*. Example:

The point estimates for (1) and (2) are the same: The value of the estimated regression function at *X* = 1.75.

Example:

● **Variability associated with estimates for (1) and (2) is quite different.**

$Var[\hat{E}(Y \mid X)] =$

$Var[\hat{Y}_{pred}] =$

● **Since σ^2 is unknown, we estimate σ^2 with MSE:**

CI for E(*Y* | *X*) at *x:**

Prediction Interval for *Y* value of a new observation with *X* = *x:**

Example: 95% CI for mean selling price for houses of 1750 square feet:

Example: 95% PI for selling price of a new house of 1750 square feet:

Correlation

● $\hat{\beta}_1$ **tells us something about whether there is a linear relationship between** ***Y*** **and** ***X*****.**
● **Its value depends on the** **<u>units of measurement</u>** **for the variables.**

● **The** **<u>correlation coefficient</u>** ***r*** **and the** **<u>coefficient of determination</u>** r^2 **are** **<u>unit-free</u>** **numerical measures of the** **<u>linear association</u>** **between two variables.**

● $r =$

(measures strength and direction of linear relationship)

● ***r*** **always between -1 and 1:**
● $r > 0 \rightarrow$

- $r < 0$ →

- $r = 0$ →
- r near -1 or 1 →

- r near 0 →

- Correlation coefficient (1) makes no distinction between independent and dependent variables, and (2) requires variables to be numerical.

Examples:
House data:

Note that $r = \hat{\beta}_1 \left(\frac{s_X}{s_Y} \right)$ so r always has the same sign as the estimated slope.

- The population correlation coefficient is denoted ρ.
- Test of H_0: $\rho = 0$ is equivalent to test of H_0: $\beta_1 = 0$ in SLR (p-value will be the same)

- Software will give us r and the p-value for testing H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$.

- To test whether ρ is some nonzero value, need to use transformation – see p. 318.

● **The square of r, denoted r^2, also measures strength of linear relationship.**

● **Definition: r^2 = SSR / TSS.**
<u>Interpretation of r^2</u>: It is the proportion of overall sample variability in Y that is explained by its linear relationship with X.

Note: In SLR, $F = \frac{(n-2)r^2}{1-r^2}$ **.**

● **Hence: large r^2 → large F statistic → significant linear relationship between Y and X.**

<u>Example</u> (House price data):

Interpretation:

<u>Regression Diagnostics</u>

● **We assumed various things about the random error term. How do we check whether these assumptions are satisfied?**

● **The (unobservable) error term for each point is:**

● **As "estimated" errors we use the <u>residuals</u> for each data point:**

● **<u>Residual plots</u> allow us to check for four types of violations of our assumptions:**

(1) The model is misspecified
(linear trend between *Y* and *X* incorrect)
(2) Non-constant error variance
(spread of errors changes for different values of *X*)
(3) Outliers exist
(data values which do not fit overall trend)
(4) Non-normal errors
(error term is not (approx.) normally distributed)

● **A residual plot plots the residuals** $Y - \hat{Y}$ **against the predicted values** $\hat{Y}$**.**
● **If this residual plot shows <u>random scatter</u>, this is <u>good</u>.**
● **If there is some notable pattern, there is a possible violation of our model assumptions.**

<u>Pattern</u>	**<u>Violation</u>**

● We can verify whether the errors are approximately normal with a Q-Q plot of the residuals.

● If Q-Q plot is roughly a straight line → the errors may be assumed to be normal.

<u>Example</u> (House data):

<u>Remedies for Violations – Transforming Variables</u>

● When the residual plot shows megaphone shape (non-constant error variance) opening to the right, we can use a <u>variance-stabilizing transformation</u> of *Y*.

● Picture:

● Let $Y^* = \log(Y)$ or $Y^* = \sqrt{Y}$ and use *Y as the dependent variable.**

● These transformations tend to reduce the spread at high values of $\hat{Y}$.

● Transformations of *Y* may also help when the error distribution appears <u>non-normal</u>.

● Transformations of *X* and/or of *Y* can help if the residual plot shows evidence of a <u>nonlinear trend</u>.

● Depending on the situation, one or more of these transformations may be useful:

● Drawback: Interpretations, predictions, etc., are now in terms of the <u>transformed variables</u>. We must reverse the transformations to get a meaningful prediction.

<u>Example</u> (Surgical data):